JOINT REPLENISHMENT PROBLEM IN TWO ECHELON INVENTORY SYSTEMS WITH TRANSPORTATION CAPACITY

A THESIS
SUBMITTED TO THE DEPARTMENT OF INDUSTRIAL ENGINEERING
AND THE INSTITUTE OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Nasuh Çağdaş Büyükkaramıklı
December, 2006

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof.Ülkü Gürler (Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Osman Alp (Co-advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Alper Şen

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Pelin Bayındır

Approved for the Institute of Engineering and Science:

Prof. Mehmet B. Baray
Director of the Institute

ABSTRACT

JOINT REPLENISHMENT PROBLEM IN TWO ECHELON INVENTORY SYSTEMS WITH TRANSPORTATION CAPACITY

Nasuh Çağdaş Büyükkaramıklı
M.S. in Industrial Engineering
Supervisor: Prof. Ülkü Gürler , Asst. Prof. Osman Alp
December, 2006

In this study, we examine the stochastic joint replenishment problem in the presence of a transportation capacity. We first study the multi-retailer and single-echelon setting under a quantity based joint replenishment policy. A limited fleet of capacitated trucks is used for the transportation of the orders from the ample supplier in our setting. We model the shipment operations of the trucks as a queueing system, where the customers are the orders and trucks are the servers. Consequently, different transportation limitation scenarios and methods of approach for these scenarios are discussed. We then extend our model to a two-echelon inventory system, where the warehouse also holds inventory. We characterize the departure process of the warehouse inventory system, which becomes the arrival process of the queueing system that models the shipment operations between the warehouse and the retailers. This arrival process is then approximated to an Erlang Process. Several numerical studies are conducted in order to assess the sensitivity of the total cost rate to system and cost parameters as well as the performance of the approximation.

Keywords: Stochastic Joint Replenishment Problem, Queueing Theory, Transportation Limitation, Inventory Theory.

ÖZET

ULAŞIM KISITLI İKİ DÜZEYLİ ENVANTER SİSTEMLERİNDE TOPLU SİPARİŞ PROBLEMİ

Nasuh Çağdaş Büyükkaramıklı
Endüstri Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Prof. Dr. Ülkü Gürler , Yard. Doç. Dr. Osman Alp
Aralık, 2006

Bu çalışmada ulaşım kapasiteli envanter sistemlerindeki toplu sipariş problemi incelenmiştir. Önce tek düzeyli çok perakendecili bir ortamda miktar bazlı toplu sipariş politikaları incelenmiştir. Siparişlerin, kapasitesi sınırsız olan bir tedarikçiden perakendeciye taşınmasında kapasiteli kamyonlardan oluşmuş bir filo kullanılmaktadır. Toplu siparişlerin kamyonlar tarafindan taşınması işlemleri, kamyonların birer işgoren, siparişlerin de birer müşteri olduğu bir kuyruk sistemi ile modellenmiştir. Bu model altında sisteme ait toplam maliyet fonksiyonu yazılmış, farklı ulaşım kısıtı senaryoları için çeşitli çözüm yaklaşımları geliştirilmiştir. Daha sonra model, tedarikçide de envanter tutulan iki düzeyli bir envanter sistemine genişletilmiştir. Depo envanter sisteminin çıkış sürecinin, depo ile perakendeciler arasındaki taşıma sisteminin modellenmesinde kullanılan kuyruk sisteminin giriş sürecine eşdeğer olduğu gözlemlenmiş ve bu süreç karakterize edilmiştir. Bu giriş sürecini bir Erlang süreci ile yaklaşıklayarak, sisteme ait toplam maliyet fonksiyonu türetilmiştir. Yapılan yaklaşıklamanın performansını ölçmek ve sistem parametrelerinin duyarlılığını gözlemlemek amacıyla çeşitli sayısal çalışmalar yürütülmüştür.

Anahtar sözcükler: Rassal Toplu Sipariş Problemi, Kuyruk Kuramı, Ulaşım Kısıtlamaları, Envanter Kuramı.

Acknowledgement

First and foremost, I would like to express my sincere gratitude to my supervisors Prof. Ülkü Gürler and Asst. Prof. Osman Alp for their concern and guidance during my M.S. study. They have been always ready to provide help, support and trust. I have learned a lot of things from them, not only in academic but also in personal and intellectual matters. I consider myself lucky to have worked under their supervision.

I would like to thank to Asst. Prof. Alper Şen and Asst. Prof. Pelin Bayındır for accepting to read and review this thesis and their substantial comments and suggestions.

Also, I would like to express my gratitude to TUBITAK for its financial support throughout my Master's study.

I am indebted to Banu Yüksel Özkaya for her academic and morale support at all times. She has always provided help even when she was very far away.

I want to thank Önder Bulut, Ahmet Camcı and Sıtkı Gülten. It would have been very hard without their comradeship.

I indebted to my dear friends Nurdan Ahat, Ayşegül Altın, Çiğdem Ataseven, Didem Ekmekçi, Çağrı Latifoğlu and Berkay Toprak for their morale support. Also, I want to thank to all of my office-mates for their understanding as well as to all of the friends that I failed to mention here.

Finally, I would like to express my deepest gratitude to my family for their everlasting love and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

Current research trend in logistics management stresses the importance of integration of different functional operations within a firm throughout the supply chain. By the integration of the supply chain, many companies have succeeded in reducing costs and increasing service levels. Recent advances in the information technology enable the sharing of available information among different parts of the supply chain more effectively, which facilitates the coordination of the different functional areas within a firm.

Inventory and transportation costs comprise the bulk of the total operating costs of a distribution system. Substantial reduction in total costs is achievable by incorporating transportation and inventory control decisions carefully. In general, there is a trade-off between the inventory and transportation costs in a logistics system. Hence, coordinated planning of the inventory and transportation decisions can greatly reduce the total operating costs of the system.

In this study, we focus on coordinated replenishment policies in single-echelon and two-echelon single-item/multi-location inventory settings under transportation capacity. In particular, we study Stochastic Joint Replenishment Problem (SJRP) in settings, where transportation is capacitated.

SJRP is the determination of replenishment and stocking decisions for different items (or retailers) to minimize total expected operating (i.e. holding, shortage and order setup/transportation) costs per unit time, when demands are random and joint ordering cost structures are present. In most of the real world systems, the ordering cost structure presents an opportunity to benefit from the economies of scale in replenishment by giving orders jointly. This is possible when the items are purchased from the same supplier or they share the same transportation vehicle.

In most of the distribution systems, the transportation of items are capacitated. Most of the firms have their own limited fleet for their transportation operations. (e.g. Shell, BP, etc...), whereas some of the firms contract with a 3PL provider for running of their transportation operations. In both of the cases, the transportation is not unlimited. The fleet size and the capacity of the trucks have their own kind of cost structures. Hence, the size of the fleet that is used in the transportation of the orders and the capacity of the vehicles are also the challenging decisions companies have to make. So, considerable cost savings can be achieved by coordinating the joint replenishment policy decisions with these aforementioned transportation related decisions in the supply chain systems.

The stochastic joint replenishment problem (SJRP) differs from the deterministic joint replenishment problem greatly in modelling methodologies and policy structures. We refer the reader to Aksoy and Erengüc [1] and Goyal and Satir [20] for the extensive review of the works about the deterministic JRP.

To our knowledge, Ignall [23] is the only study that analyzes the optimal joint replenishment policy in the $SJRP$. The optimal policy, even for a two-item case and zero lead-times has a very complicated structure. As the number of items in the system increases, the structure of the optimal policy would be more complicated. Therefore, the control and the implementation of the optimal joint order policies in practice would be even more challenging. This is one of the main reasons why most of the existing studies in the literature is mostly focused on finding and evaluating intuitive heuristic policy classes, which are easier to control and implement in practice.

Balintfy [6], which is the earliest study in the stochastic joint replenishment problem literature, develops a continuous-review joint ordering policy: (**S, c, s**), which is referred to as the *can-order policy*. This policy determines the reorder, $\mathbf{s} = (s_1, s_2, ..., s_N)$ and can-order, $\mathbf{c} = (c_1, c_2, ..., c_N)$ points as well as the order up-to levels, $\mathbf{S} = (S_1, S_2, ..., S_N)$ for each item i. The policy operates as follows: a demand to item i triggers a replenishment order whenever the inventory position of that item drops to its reorder point s_i. After the replenishment order, the inventory position of item i is increased to its order up-to level S_i. At the same time, any other item j whose inventory position is less than or equal to its can-order point, c_j $(s_j < c_j < S_j)$ is also included to the joint replenishment order, and their inventory positions are raised to their order up-to levels S_j. The implementation of the policy seems to be simple, however the calculation of operating characteristics under this policy is very difficult, even in the presence of unit Poisson demands.

Silver [34] studies a special case of the can-order policy. In his study, the lead-time is zero, the items face unit Poisson demands, shortages are not allowed and $\mathbf{c} = \mathbf{S\text{-}1}$ and $\mathbf{s{=}0}$. With the objective of minimizing the expected total cost per unit per time, Silver [34] proves that the can-order policy performs better than the individual control policy if the cost of ordering an item is equal to that of ordering two items jointly. If these ordering costs are not equal, he shows that there exists a critical value for the fixed item ordering cost above which joint replenishment policy becomes more profitable compared to individual replenishment policy.

Silver [36] also develops a new approximation technique for the analysis of the can-order policies. In this approximation technique, the N-item problem is decomposed into N single-item problems. This single-item problem is first analyzed by Silver [35] and solved optimally by Zheng [44]. Subsequently, Thompson and Silver [39], Federgruen et al. [16], Silver [37], Van Eijs [42] and Schultz and Johansen [33] focus on the different aspects of can-order type policies.

The difficulties in the analytical treatment of the can-order policy as well as the size of the optimization problem (for an N-item setting, can-order policy employs $3N$ policy parameters) call for the need for more parsimonious continuous-review control policies.

The continuous-review $(Q, \mathbf{S})$ policy, one of the most frequently used policies in the industry, is first proposed by Renberg and Planche [30]. In this policy, whenever a total demand of Q units accumulate for all items, a joint order of size Q is given to the supplier, and the inventory positions of each item is raised to the vector of order-up-to levels $\mathbf{S}$=$(S_1, S_2, ..., S_N)$. This policy uses $N+1$ policy parameters for an N-item setting. Pantumsinchai [29] subsequently presents the exact analysis of the $(Q, \mathbf{S})$ policy and compares the performance of it with that of $(\mathbf{S},\ \mathbf{c},\ \mathbf{s})$ policy. The numerical results indicate that the $(Q, \mathbf{S})$ policy performs better than the can-order policy if the fixed ordering cost is high and the backorder cost is low, whereas the can-order policy only performs better if the fixed ordering cost is low.

Atkins and Iyogun [2] suggest two periodic review replenishment policies for unit Poisson demands. In the first proposed policy, which is referred as P, inventory positions of all items are raised to their order up-to levels $\mathbf{S}$ at the end of each period of length T. In the second policy, which is represented by MP, the review periods are integer multiples of a base period and review periods can differ for each item. From the numerical results, Atkins and Iyogun [2] assert that P and MP type policies outperform the $(\mathbf{S},\ \mathbf{c},\ \mathbf{s})$ policy as the fixed ordering cost increases.

Nielsen and Larsen [26] suggest a new policy referred to as the $Q(\mathbf{S},\ \mathbf{s})$ policy. This policy functions as follows: whenever a total demand of Q units are accumulated since the last review, a replenishment order is triggered and the items whose inventory positions at or below $\mathbf{s}$ in this review epoch are raised to their order-up-to levels $\mathbf{S}$.

Cachon [10] proposes a new policy for the dispatchment of the trucks in a single-echelon distribution system, and compares the performance of this new policy with those of $(Q, \mathbf{S})$ and P type policies. This new policy, which is referred to as the $(Q, \mathbf{S}|T)$ policy operates as follows: the retailer reviews its inventory every T time units and trucks are dispatched if the accumulated retailer orders can fill the trucks in such a way that one of the trucks has at least Q units and others are full. Note that Q can take a value less than or equal to the truck capacity in this particular policy.

In a recent study, Özkaya et. al [28], suggest a new parsimonious policy for unit Poisson demands as well as for the batch demands. This policy, which is referred to as $(Q, \mathbf{S}, T)$ policy, is a sort of hybrid of the continuous review $(Q, \mathbf{S})$ and periodic review P policies. It performs as follows: the items are reviewed continuously and the inventory positions of all items are raised to $\mathbf{S} = (S_1, S_2, ..., S_N)$ whenever a total Q demands accumulate for the items or T time units have elapsed, whichever occurs first. It is shown numerically that $(Q, \mathbf{S}, T)$ policy performs better than the other joint replenishment policies under most of the settings. Next, we elaborate on the multi-echelon joint replenishment literature.

There is a vast literature on multi-echelon inventory systems. For a general review of the literature, the reader is referred to Federgruen [15]. The analytical treatment of the policies for the SJRP in two-echelon inventory systems is more difficult compared to single-echelon inventory systems. Therefore the studies about SJRP in single-echelon inventory systems outnumber the studies in multi-echelon inventory systems. Among the related works, Axsäter and Zhang [5] and Cheung and Lee [13] study the SJRP for continuous review models in two-echelon divergent inventory systems.

Özkaya [27] provides a modeling methodology for a general policy class for the SJRP in two-echelon inventory systems. Via this modeling methodology, she analyzes most of the joint replenishment policies in the literature in two-echelon distribution systems and compares the performance of these policies in various parameter settings.

Lastly, we mention about Vendor Managed Inventory (VMI) systems, which is a related topic in the multi-echelon inventory management. The recent studies by Çetinkaya and Lee [11] and Kiesmüller and de Kok [24] focus on different aspects of consolidation policies in the VMI systems. In the VMI systems, small retailer orders are combined to larger shipments at the warehouse level according to a consolidation policy. In most of the VMI literature, the problem is analyzed from the warehouse perspective, though the impact of the consolidation policies on the retailers' performance is mostly neglected.

Existing literature in SJRP overlooks some important transportation matters such as the effects of a limited fleet size and cargo capacity constraints. To our knowledge, Cachon [10] and Gürbüz [21] are the only studies that incorporate the truck capacity considerations with the analysis of the joint replenishment policies, but in their studies, they both assume that there is an unlimited availability of transportation vehicles. However, unlimited availability of transportation vehicles is not possible in most of the real life applications. Therefore, the transportation limitation problem has been analyzed in the supply chain literature both in deterministic and stochastic demand cases. In some of these studies, system parameters are optimized for a given truck capacity and/or fleet size, whereas there are some studies, where they are taken as decision variables, too. Next, we briefly review the relevant supply chain literature on the transportation limitation problem.

In both deterministic and stochastic demand cases, there are various models that handle the truck/cargo capacity problem under the assumption of an unlimited fleet size similar to Cachon [10]. In such a case, truck capacity constraint greatly determines the ordering cost structure of the system.

In the deterministic demand case, Toptal et al.[41], Çetinkaya and Lee [12] study different types of coordination problems in the presence of cargo capacity constraints. Also, Benjamin [8] considers a joint production, transportation and inventory problem with deterministic demand, allowing supply capacity constraints.

In the random demand case, Yano and Gerchak [43], Henig et al. [22] and Ernst and Pyke [18] analyze different types of models with the truck capacity consideration. Note that truck capacity is also considered to be one of the decision variables in these aforementioned studies.

In the literature, fleet size consideration is generally taken into account in inventory-vehicle routing problems. Ball et al. [7] address the problem of finding an optimal fleet size as well as determining the vehicle routes where the demand is deterministic. In the stochastic demand case, Federgruen and Zipkin [17] study an integrated vehicle routing and allocation problem with a given fleet of capacitated vehicles. For a broader review of inventory-vehicle routing problems, we refer the reader to Dror and Ball [14] and Ben-Khedher and Yao [9].

In a recent paper, Sindhuchao et al.[38] develop a mathematical programming approach for the multi-item joint replenishment problem in an inventory-routing system with a limited vehicle capacity. In this study, the demand is assumed to be deterministic. Although the motivations behind this work are parallel to ours, the methodology of the approach differs greatly in our problem due to the random demand structure.

In this study we study a specific kind of a control policy for the SJRP, which is (Q,**S**) policy, in the presence of a truck capacity and fleet size limitation in both single and two-echelon divergent multi-retailer systems with unit Poisson demands. To our knowledge, our study is the first one that incorporates the decisions of the truck capacity and fleet size in the stochastic joint replenishment problem.

The shipment operations of the trucks is modelled as a queueing system, where the customers are the orders and trucks are the servers. By using some of the key results in the queueing theory, we derive the operating characteristics of the single-echelon inventory systems in the presence of the limited fleet size of the capacitated trucks analytically. We investigate the characteristics of the total cost rate function and construct methods of approach for different kinds of transportation limitation when there is a maintenance/depreciation cost rate factor per truck. We also present the results of our numerical study to assess the sensitivity of decision variables to the system parameters.

The analysis for the single-echelon inventory system is extended to a two-echelon system, where the upper echelon also holds inventory. In this two-echelon system, a joint retailer order has to be satisfied by the warehouse inventory system before it is dispatched by the trucks. Since the operations of the trucks are modeled as a queueing system, the departure process of the warehouse inventory system constitutes the arrival process for this queueing system. This problem in our setting leads us to a more general problem, and we derive the general characteristics of the departure process of an (S-1, S) inventory system where the arrivals occur according to a renewal process. From these general results, we obtain the first two moments of the inter-departure times of the warehouse inventory system. Working with the original departure process, which has dependent increments is analytically intractable due to its complicated nature, therefore, we approximate the departure process to an Erlang Process, and derive the operating characteristics of the system according to this approximation. We compare the approximation results with the simulation results and present the sensitivity of the system to the decision and system variables in the numerical study part. We observe that our approximation method works fairly well except for very high traffic rate. ($\rho > 98\%$)

The remainder of the thesis is organized as follows:

In Chapter 2, we analyze the SJRP with transportation limitation in the single-echelon inventory systems. In Chapter 3, the analysis is extended to a two-echelon inventory system, where the upper echelon also holds inventory. Finally, in Chapter 4, we conclude the thesis by giving an overall summary of our work, our contribution to the existing literature and its practical implications with future research directions.

Chapter 2

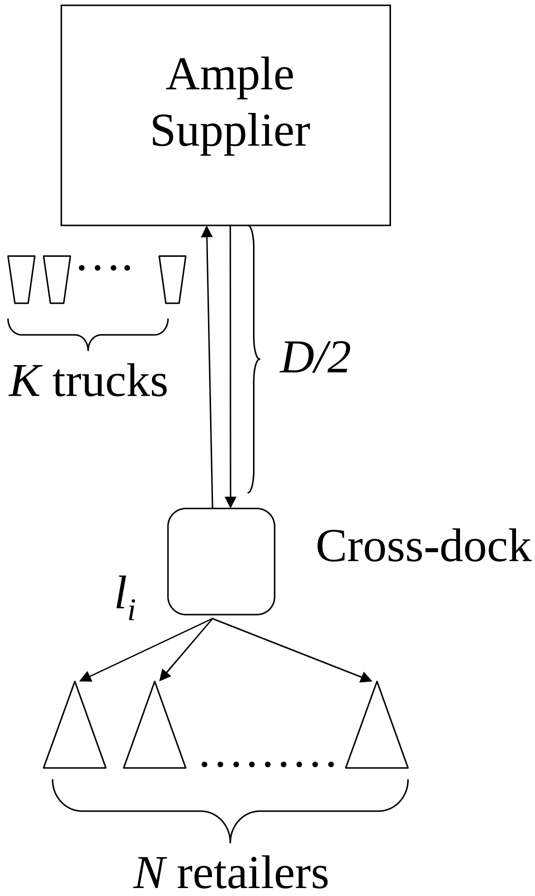
Single-Echelon Environment

In this chapter we present an analytical model for the coordination of inventory and transportation decisions in a single echelon, single item, multi-retailer distribution system under transportation capacity. Main assumptions of the model are presented in Section 2.1. Section 2.2 presents a preliminary analysis, which is followed by the derivation of the expressions for the key operating characteristics and the statement of the optimization problem in Section 2.3. In Section 2.4, we discuss some of the characteristics of the total cost rate function of the system, which we use in Section 2.5, while constructing the search algorithms according to different scenarios of transportation limitation. Finally in Section 2.6, we present the numerical results, where we assess the total cost rate of the system with respect to the system parameters.

2.1 Model Characteristics

We consider a single item, multi-retailer inventory setting under continuous review. (The model presented in this Chapter can be easily adopted to a multiple-items/single-retailer setting). The retailers face stationary and independent unit Poisson demands with rate λ_i ($i = 1, 2, ..N$), and all unmet demands are fully

Figure 2.1: Illustration of the environment



backordered. Retailers are supplied from an ample supplier via a fleet of K identical trucks of capacity C that are used in the transportation of retailer orders. When a truck is available, units are carried from the ample supplier to a cross-dock station, which takes $D/2$ time units. The truck returns back to its base at the ample supplier after unloading at the cross-dock station. At the cross-dock station, items destined to each retailer are transferred to smaller sized vehicles to be conveyed to the retailers. We do not allow for any re-allocation of items to the retailers at the cross-dock station. We assume that the transportation is not capacitated after the cross-dock point. The lead time between retailers and the cross-dock point, which we refer as *minor lead time*, is l_i for retailer i. Therefore, the order delivery lead time for retailer i is L_i, where $L_i = l_i + D/2$. Figure 2.1 illustrates the considered system. Holding cost per unit per time is charged at each retailer with rate h_i. There is a common fixed order setup cost for each order, which is linear to the truck capacity, $A(C) = a \times C$ for $a > 0$. The shortage cost per unit per time at retailer i is incurred at a rate of β_i. Any possible additional costs for monitoring the inventory system continuously are ignored. The joint orders are satisfied based on the first come-first serve (FCFS) rule at the ample supplier. Under this policy and cost structure, the objective is to minimize the expected total cost per unit time.

Due to the ordering cost structure, retailers implement a joint replenishment policy to manage their inventory and replenishment decisions. Since we are considering an ample supplier, the orders received from the retailers constitute a compound renewal process, where inter-order time, Y, and the order quantity, Q_0, have a joint density, $f_{Y,Q_0}(y,q)$. Joint orders received by the ample supplier are shipped immediately by trucks if there are sufficiently many trucks available to hold the existing order. In such a case, any given loaded truck spends a total of D time units to reach to the cross-dock point, unload the item and return back to its base. If there are not sufficiently many trucks available at the time of an order trigger at the base, then the order waits until enough number of trucks become available. Hence, the shipment operations of the retailer orders at the ample supplier can be modeled as a queueing model where the orders are the *customers* in the system and the trucks are the *servers* which are busy while carrying the materials to the cross-dock point and return back.

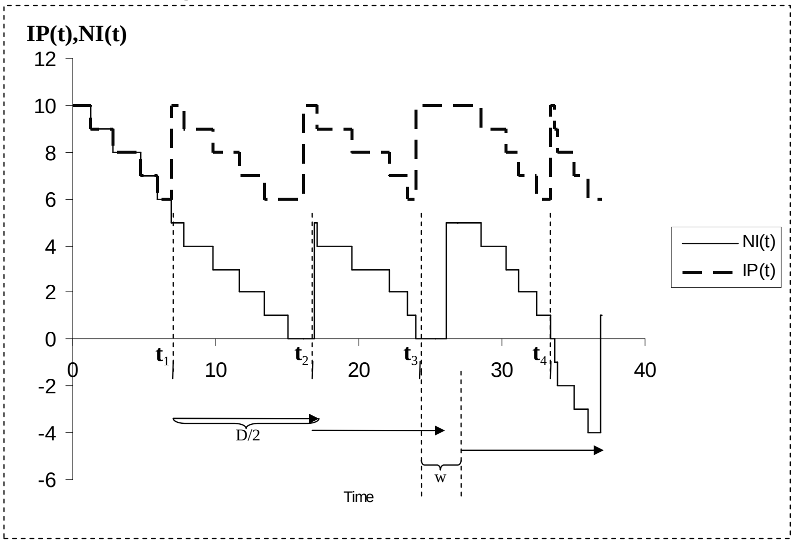
We assume that the retailers use $(Q, \mathbf{S})$ policy of Renberg and Planche [30] as the joint replenishment policy for controlling their inventory in this study. In this continuous review policy, whenever a total demand of Q units are accumulated at the retailer level, a joint order of size Q is given to the supplier, and the inventory positions of the retailers are raised to the vector of order-up-to levels $\mathbf{S}$=$(S_1, S_2, ..., S_N)$. We ignore the truck loading and unloading times, however they could be easily incorporated to our model by modifying the existing parameters.

2.2 Preliminary Analysis

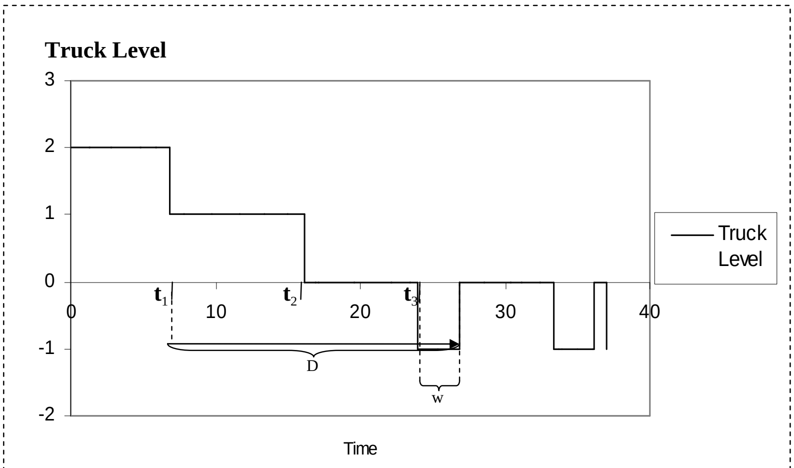
In this section we present the methods that we use in our analysis. First, we introduce our notation. Let r_i be the probability that the demand is for retailer i, given that a demand arrival has occurred. Since demand process is Poisson, we have $r_i = \lambda_i/\lambda_0$, where $\lambda_0 = \Sigma_{i=1}^{N}\lambda_i$. Under the $(Q, \mathbf{S})$ policy, a *cycle* is defined as the time between two consecutive joint order placements, where the inventory positions of all retailers are raised to $\mathbf{S} = (S_1, S_2, ..., S_N)$. Inter-order time, which is denoted by X is the cycle length. Since the last epoch, total of Q retailer demands must be accumulated to place an order again. Hence, $f_{X,Q_0}(x, q) = 0$ for $q \neq Q$, which means that Q_0 is always equal to Q, where f denote the joint density of X and Q_0. Since the inter-arrival times of the demands are exponential, X has an *Erlang* Q distribution with scale parameter λ_0. Let $F(x, k, \lambda)$ and $f(x, k, \lambda)$ denote the probability distribution and density functions of an Erlang random variable with shape and scale parameters k and λ, respectively and $\overline{F}(\cdot)= 1-F(\cdot)$ is the complementary distribution for any distribution function. For clarity and later use, we have the following definitions. At any given time t, $IP_i(t)$ denotes the inventory position at retailer i and $IP(t)$ denotes the total inventory position at the retailer level. $IP(t) = \Sigma_{i=1}^{N} IP_i(t) \leq \Sigma_{i=1}^{N} S_i = S_0$. Also, let $NI_i(t)$ denote the net inventory level at retailer i, and $NI(t) = \Sigma_{i=1}^{N} NI_i(t)$ denote the total inventory level at any given time t. A joint order is placed by the retailers when $IP(t)$ falls to $S_0 - Q$. If there are enough trucks on hand to meet the joint order that is placed at time t, the order is immediately dispatched, otherwise it waits until a truck becomes available.

A typical realization is depicted in Figure 2.2. In this particular realization we ignore minor lead times l_i, and assume there are $K = 2$ trucks, $S_0 = 10$, $C = 8$ and $Q = 5$. Figure 2.2a shows the total inventory position $IP(t)$ and net inventory $NI(t)$ at the retailer level and Figure 2.2b shows the number of available trucks on hand. First joint order occurs at t_1, and it is dispatched immediately since there are enough trucks available. Second order occurs at t_2, and it is dispatched without any delay, too. However the third joint order placed at time t_3 is not dispatched immediately, since there is no available truck at

Figure 2.2: Realization of the model



(a)



(b)

time t_3. After the placement of the third joint order, number of trucks on hand decreases from 0 to -1. When the number of trucks on hand is negative, there is at least one order that is waiting for a truck to be dispatched. The third joint order in this illustration is dispatched at $t_1 + D = t_3 + w$, when a truck returns back to the ample supplier.

Let W denote the time a joint order waits for dispatching. In our particular realization, $W = 0$ for the first and the second joint orders, and $W = w$ for the third one. The lead-time (total transit time) for retailer i, (L_i), with the waiting time for a truck W constitute the effective lead-time for retailer i, which is denoted by $\overline{L_i} = L_i + W$. W is a random variable and its distribution function is denoted by $F_W(w)$ for $w \geq 0$.

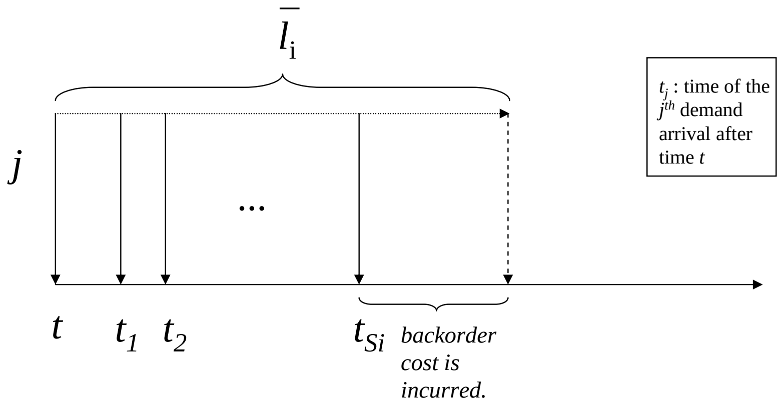
As mentioned above, joint orders are received from the retailers with an $Erlang - Q$ distributed inter-arrival time with scale parameter λ_0 in a $(Q, \mathbf{S})$ policy. If $Q \in (C/2, C]$ and there is an order integrity constraint, each joint order occupies exactly one truck. Hence, dispatching operations act as an $E_Q/D/K$ queue when more than a 50% truck utilization is guaranteed. Enforcing a minimum truck utilization may result in suboptimal policy parameters, however it is a common practice in industry due to the transportation limitations and high order set-up costs. Hence, $F_W(w)$, which is essential for deriving the operating characteristics of our system, is identical to the waiting time distribution of an $E_Q/D/K$ queue.

2.3 Derivation of the Operating Characteristics

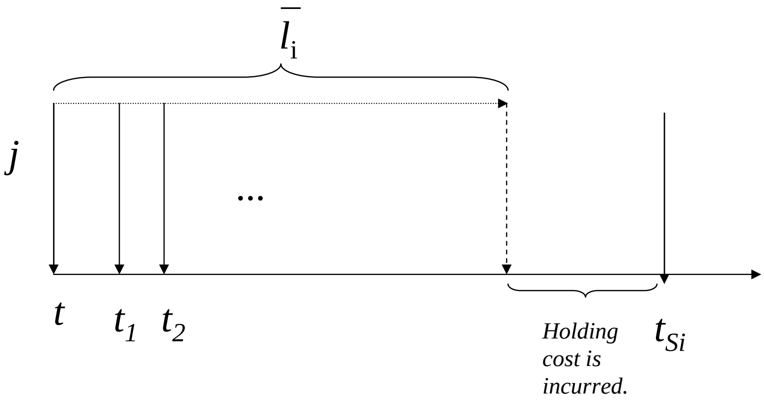
In this section, the operating characteristics of our system are derived, and these expressions are used in calculating the total cost rate. Total cost of the system consists of two parts. The first part is the order setup cost and the latter is the holding and backorder costs. We begin with expected cycle length, $E[X]$. As noted in Section 2.2, X has an *Erlang* Q distribution with scale parameter λ_0. Therefore, expected cycle length is simply Q/λ_0. In each cycle, the fixed ordering cost of a joint retailer order is incurred once, and order setup cost rate is simply $A(C) \times \lambda_0/Q$. The evaluation of the expected costs under the $(Q, \mathbf{S})$ policy with capacitated trucks and unlimited fleet size is analyzed by Cachon [10]. Note that $W = 0$ and the lead-times $(L_1, L_2, ..., L_N)$ of the system are constants if there is no limit on the fleet size. On the contrary, when the fleet size is limited, effective lead-times, $\overline{L_i} = L_i + W$, are random variables which may take any value on the interval $[L_i, \infty)$.

Axsater [3] presents an approach, which can be used to evaluate the expected holding and backorder cost rate of a two-echelon inventory system consisting of

Figure 2.3: Illustrations of how backordering and holding costs per unit are incurred



(a) A case when the time of the S_i^{th} demand after t (t_{Si}) is less than the given lead time($\bar{l}_i$)



(b) A case when the time of the S_i^{th} demand after t (t_{Si}) is greater than the given lead time($\bar{l}_i$)

a single depot and multiple retailers, both following base-stock policies with base stock levels S_i and deterministic lead-times. This approach is based on the observation that in such distribution systems, a unit ordered by retailer i is used to fill the S_i^{th} subsequent demand following this order. Accordingly, expected time for a unit that is held in the inventory and the expected time a unit is backordered can be evaluated by relating the arrival time of the order and the arrival time of the S_i^{th} subsequent demand. The illustrations of how the backordering and holding costs are incurred for a unit are given in Figure 2.3a and Figure 2.3b, respectively. The unit demand that occurs at time t arrives after the S_i^{th} subsequent demand in Figure 2.3a. Therefore backorder cost per time is incurred for that unit. On the contrary, the unit demand that occurs at time t arrives before its S_i^{th} subsequent demand in Figure 2.3b. and holding cost per time is incurred for that unit. Cachon [10] adopts this approach to a similar environment, where the retailers use the (Q, **S**) policy in a coordinated fashion; there is an ample supplier, and an unlimited number of capacitated trucks are used for the distribution of items from the supplier to the retailers. We also use the same approach in our analysis.

We first define the function $g_i(S_i, \overline{l_i})$, which corresponds to the expected holding and backorder costs incurred per time per unit for retailer i with a base-stock level S_i and a given effective lead time $\overline{L_i} = L_i + W = \overline{l_i}$, which is constant. Then,

$$g_i(S_i, \overline{l_i}) = \beta_i \int_0^{\overline{l_i}} (\overline{l_i} - x) f(x, S_i, \lambda_i) dx + h_i \int_{\overline{l_i}}^{\infty} (x - \overline{l_i}) f(x, S_i, \lambda_i) dx. \qquad (2.1)$$

The expected backorder time a unit would face is expressed by the first integral above, whereas the expected time a unit is stored in the inventory is expressed by the second one. From the properties of the gamma distribution, we can rewrite:

$$\int_0^{\overline{l_i}} x f(x, S_i, \lambda_i) dx = S_i/\lambda_i \int_0^{\overline{l_i}} f(x, S_i + 1, \lambda_i) dx \qquad (2.2)$$

which corresponds to the probability that S_i^{th} demand occurs before $\overline{l_i}$. The probability of this event is identical to the probability that S_i or more demands occur in $[0, \overline{l_i}]$. Therefore we can write

$$\int_0^{\overline{l_i}} f(x, S_i, \lambda_i) dx = (1 - F_P(S_i - 1, \lambda_i \overline{l_i})) \qquad (2.3)$$

where $F_P(y, \lambda \overline{l_i})$ denotes the cumulative probability distribution of a Poisson process with rate λ. Hence, we can rewrite Equation (2.1) as follows:

$$g_i(S_i, \overline{l_i}) = \frac{1}{\lambda_i} \left[S_i(h_i + \beta_i) F_P(S_i, \lambda_i \overline{l_i}) - \lambda_i \overline{l_i}(h_i + \beta_i) F_P(S_i - 1, \lambda_i \overline{l_i}) + \beta_i(\lambda_i \overline{l_i} - S_i) \right]. \qquad (2.4)$$

Equation (2.4) gives the expected holding and backorder costs a unit demand from retailer i faces when that unit demand triggers a joint replenishment decision. When $Q = 1$ in a $(Q, \mathbf{S})$ policy, a unit demand always triggers a joint replenishment decision. Now we consider the general case when a unit demand to retailer i is arrived, but a joint replenishment decision is not triggered until a total of $Q \geq 1$ demands are accumulated. Let a unit demand to retailer i occurs at time τ, and the trigger of a joint replenishment decision that contains this unit demand to retailer i is delayed until $\tau + t$. That joint order arrives at the retailer at $\tau + t + \overline{l_i}$. Let M_i denote the total number of the demands occur at retailer i between τ and $\tau + t$. When $M_i = m_i$, the unit demand occured at τ is used to fill the $(S_i - m_i)^{th}$ subsequent demand after $\tau + t$. Therefore, the expected holding and backorder cost that we incur for a unit demand from retailer i is simply $g_i(S_i - m_i, \overline{l_i})$. It must be noted that, when $m_i \geq S_i$, $g_i(S_i - m_i, \overline{l_i})$ still gives the expected holding and backorder cost that is incurred for a unit demand from retailer i. However, Equation (2.4) should be used for the calculation. The reader is referred to Axsäter [4] for a detailed proof.

Let $M_0 \geq M_i$ be the total number of retailer demands (including i) that have occurred in $(\tau, \tau + t]$. When $M_0 = n$, the probability that m_i of these n demands are from retailer i is the probability of having m_i successful draws out of n, where $r_i = \lambda_i / \lambda_0$ is the probability of success. Let $Z(m_i|n)$ be the probability mass function of the number of successful ones from n draws. Then,

$$Z(m_i|n) = Pr(M_i = m_i | M_0 = n) = \binom{n}{m_i}(r_i)^{m_i}(1 - r_i)^{n - m_i}.$$

It is known that M_0 is a uniformly distributed integer on the interval $[0, Q - 1]$ (see Axsäter [4]). Finally, we can derive the expected holding and backorder cost per time per unit demanded from retailer i for a given effective lead-time $\overline{L_i} = \overline{l_i}$ as below:

$$\frac{1}{Q} \sum_{n=0}^{Q-1} \sum_{m_i=0}^{n} Z(m_i|n) g_i(S_i - m_i, \overline{l_i}). \tag{2.5}$$

After the expected holding and backorder cost rate for retailer i is analyzed for a given effective lead-time, we need the distribution function of the effective lead-time in order to calculate the expectation of the holding and backorder cost rate over effective lead-time. Since L_i is constant, the distribution function of the effective lead-time, $\overline{L_i} = L_i + W$ is determined by the distribution function $F_W(w)$ of the waiting time W of an order for a truck to be procured.

2.3.1 Derivation of the Waiting Time Distribution of an Order for a Truck

In this subsection, we derive $F_W(w)$, which is essential for analyzing the operating characteristics of our system. As mentioned in Section 2.2, the dispatching of the orders via trucks operates as a queueing system and $F_W(w)$ is identical to the waiting time distribution of this $E_Q/D/K$ queue. The following theorem provides a basis for the method that we use to derive $F_W(w)$.

Theorem 1 *(Tijms [40], p.321): The waiting time distribution $F_W(w)$ in the multi-server $GI/D/c$ queue is the same as in the single-server $GI^{c*}/D/1$ queue in which the inter-arrival time is distributed as the sum of c inter-arrival times in the $GI/D/c$ queue.*

The theorem has the following important corollary.

Corollary 1 *(Tijms [40], p.321): The waiting time distribution $(F_W(w))$ in the $E_k/D/c$ queue is identical to the waiting time distribution in the $M/D/ck$ queue with the same server utilization.*

The dispatching of the trucks operate as an $E_Q/D/K$ queue. Due to the Corrollary 1, in order to find the distribution of the time an order waits for dispatching, we need the waiting time distribution of an $M/D/c$ queue where $c = Q \times K$. We use the solution method that Franx [19] proposes to find $F_W(w)$ in an efficient manner for such queues. In order to be coherent with the terminology, we use *customer* and *server* instead of joint order and truck, respectively, from now on.

First, let $p_i(t)$ denote the probability of the system holding i customers at time t. Since the service time is deterministic, all the customers in the service will have left the system at time $t + D$. Consequently, customers in the system at time $t + D$ either have arrived during the time interval $(t, t + D]$ or they were waiting for service at time t. Therefore the following expression can be written by conditioning on the number of customers present at time t:

$$p_i(t+D) = \Sigma_{j=0}^{c} p_j(t)\frac{(\lambda_0 D)^i}{i!}e^{-\lambda_0 D} + \Sigma_{j=c+1}^{i+c} p_j(t)\frac{(\lambda_0 D)^{i+c-j}}{(i+c-j)!}e^{-\lambda_0 D}, t \in \mathcal{R}, i \in \mathcal{N}. \tag{2.6}$$

When there are less than c customers in the system at time t, all customers present at time t will be served at time $t+D$. If i customers arrive between t and $t+D$ then the number of customers at time $t+D$ will be i. However, when there are $j > c$ customers in the system at time t, c of the customers will be served by time $t+D$. If there are $i \geq 0$ customers in the system at $t+D$, then $j-c+i$ customers must have arrived between t and $t+D$.

As $t \to \infty$, we can obtain the stationary state probability of the number of customers in the system, which is denoted by $p_i = \lim_{t\to\infty} p_i(t)$ as

$$p_i = \Sigma_{j=0}^{c} p_j \frac{(\lambda_0 D)^i}{i!}e^{-\lambda_0 D} + \Sigma_{j=c+1}^{i+c} p_j \frac{(\lambda_0 D)^{i+c-j}}{(i+c-j)!}e^{-\lambda_0 D}, i \in \mathcal{N}. \tag{2.7}$$

These expressions of p_i's constitute an infinite system of linear equations with the normalization equation $\sum_{i=0}^{\infty} p_i = 1$. This infinite system of linear equations can be reduced to a finite system of linear equations by the following theorem:

Theorem 2 *(Tijms [40],p.289): The state probabilities (p_j) of the $M/D/c$ queue exhibit geometric tail property:*

$$p_j \approx \delta\gamma^{-j}$$

for large j, where $\gamma \in (1,\infty)$ is the unique solution of the equation:

$$\lambda_0 D(1-\gamma) + c\ln(\gamma) = 0.$$

Also, the constant δ is given by

$$\delta = (c - \lambda_0 D\gamma)^{-1}\sum_{i=0}^{c-1} p_i(\gamma^i - \gamma^c).$$

Via this geometric tail property, the infinite system of linear equations for the p_j's is reduced to a finite system by replacing p_j by $p_M(1/\gamma)^{j-M}$ for $j > M$ and an appropriately chosen M. Gaussian Elimination method is one of the possible methods in the literature, which can be used to solve finite systems of linear equations. The computational complexity of this method is $O(M^3)$

We apply an iterative method for choosing the appropriate M with a predetermined error bound, ϵ. The algorithm of the iterative method is demonstrated in the Appendix A. 1. After deriving stationary state probabilities p_i, we can derive the stationary probability q_i of the queue containing i customers, by $q_0 = \Sigma_0^c p_i$ and $q_i = p_{i+c}$ for $i > 0$. Also we define the cumulative probability that there are j or less customers in the queue as $G_j = \sum_{i=0}^{j} q_i$. Finally, the following theorem provides an expression for the Waiting Time Distribution in the queue.

Theorem 3 *(Franx, [19]): For a M/D/c queue,*
$$F_W(w) = e^{-\lambda_0(a_w D - w)} \sum_{j=0}^{a_w c-1} G_{a_w c-j-1} \frac{\lambda_0^j (a_w D - w)^j}{j!}$$
where a_w is the greatest integer less than or equal to the $\frac{w}{D} + 1$ for $w \geq 0$.

A critical point for our analysis is whether this random structure of the effective lead-time $\overline{L_i}$ permits order-crossing or not, because in the stochastic lead time environments, order crossing considerably complicates the situation. Considering that the time a unit-demand is held in the inventory (or backordered) is calculated based on the observation that a unit-demand from retailer i is used to fill the S_i^{th} subsequent demand, *no order crossing* is the sine qua non condition of the approach that is employed by Axsäter [3]. Since the joint orders are served based on the FCFS rule and the service times are deterministic, the *no-order-crossing* condition is satisfied at all times, which enables us to use the approach employed by Axsäter [3] in our system.

2.3.2 Objective Function

After the analysis for a given effective lead-time and the derivation of the distribution $F_W(w)$ of the waiting time W for truck that a joint order encounters, we take the expectation of the holding and backlogging costs over effective lead-times for each retailer i.

The waiting time of an order before dispatching is a mixed distribution. Let $\mathcal{C}$ and $\mathcal{D}$ denote the sets, where $F_W(w)$ is continuous and discrete respectively. Recall that $F_W(w)$ is dependent on K, because the number of servers is $c = K \times Q$ in Theorem 3. By taking the expectation of the cost expression in Equation (2.5) with respect to W, we can derive the expected unit holding and backorder cost of retailer i as below:

$$\begin{aligned} U(Q, \mathbf{S}, K)_i = & \int_{w \in \mathcal{C}} \frac{1}{Q} \sum_{n=0}^{Q-1} \sum_{m_i=0}^{n} Z(m_i|n) g_i(S_i - m_i, L_i + w) dF_W(w) + \\ & \sum_{w \in \mathcal{D}} \frac{1}{Q} \sum_{n=0}^{Q-1} \sum_{m_i=0}^{n} Z(m_i|n) g_i(S_i - m_i, L_i + w) P(W_q = w) \end{aligned} \tag{2.8}$$

Finally, the expected cost rate of the the whole system is given by:

$$AC(Q, \mathbf{S}, K, C) = \lambda_0 \frac{A(C)}{Q} + \sum_{i=1}^{N} \lambda_i U(Q, \mathbf{S}, K)_i. \tag{2.9}$$

The first part of the equation above represents the order setup cost rate and the second part represents the expected holding and backorder costs incurred per unit time of our distribution system that is using a $(Q, \mathbf{S})$ policy with a fleet of K trucks, each having a capacity of C. Considering the truck utilization constraint, the optimization problem of our system can be stated as follows:

$$\begin{aligned} &\text{Min } AC(Q, \mathbf{S}, K, C) \\ &\text{s.t. } Q \in (C/2, C]. \end{aligned}$$

2.4 Analysis of the Total Cost Rate Function

In this section, some characteristics of the function $AC(Q, \mathbf{S}, K, C)$ and the relations between the decision variables are discussed. We use these relations in our search algorithms that are constructed for different problem scenarios.

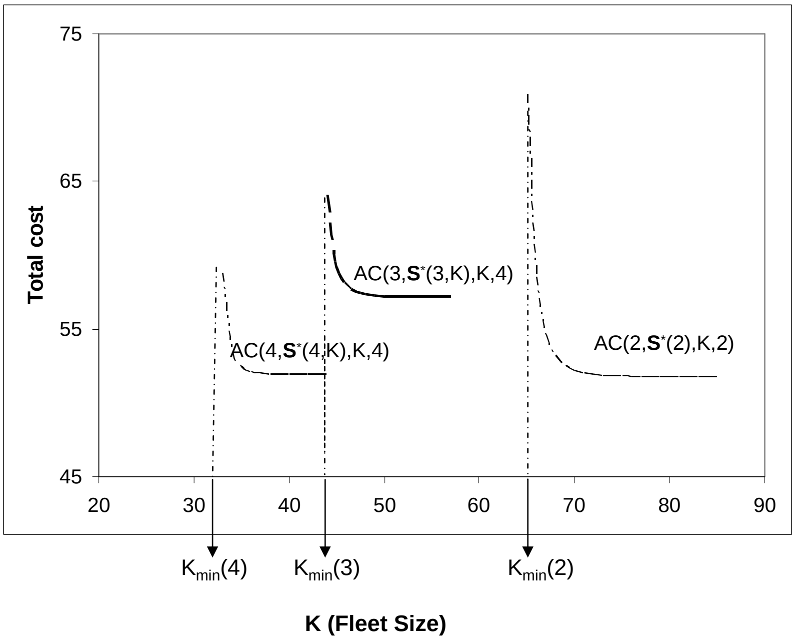
Axsäter [4] demonstrates that the unit holding and backorder cost rate function, which is given in Equation (2.5), is convex in S_i for each retailer. Recall that the $U(Q, \mathbf{S}, K)_i$ is the expectation of the unit holding and backorder cost rate function over effective lead-times. Since expectation is a linear operator, $U(Q, \mathbf{S}, K)_i$ is also convex in S_i. Therefore, the optimal order-up-to levels for each retailer i ($S_i^*(Q, K)$) for a given joint-order quantity Q and a fleet size K can easily be found. Let $\mathbf{S}^*(Q, K) = (S_1^*(Q, K), S_2^*(Q, K), ..., S_N^*(Q, K))$. When the number of the fleet size and the capacity of the trucks are given, the total cost rate function $AC(Q, \mathbf{S}^*(Q, K), K, C)$ is not necessarily convex in Q. Therefore we need to search over the feasible interval $(C/2, C]$ to find the optimal shipment quantity Q^* for given K and C. Next, we analyze the effects of K on the total cost rate of our distribution system.

It is important to note that the total cost rate of the system goes to infinity if the $M/D/KQ$ queue blows up. Hence, there is a minimum number of trucks, say $K_{min(Q)}$, that will guarantee that the queueing system operates at the steady state in the long run for a given Q value. $K_{min}(Q)$ is the smallest positive integer K that satisfies $\rho = (\lambda_0 \times D/K \times Q) < 1$.

Also as K increases, the system begins to behave as if there is no transportation limitation. In our numerical results, we observe that each truck added to $K_{min}(Q)$ brings a lower decrease in the total cost rate. Although this observation is parallel to that of Rolfe [31], who shows that the average waiting time in the queue is a convex decreasing function of the number of servers, we could not prove it analytically in our problem due to the intricacy of the expressions in our problem.

A graphical illustration of the change of $AC(Q, (\mathbf{S}^*(Q, K)), K, C)$ in K is demonstrated in Figure 2.4 for $Q = 2, 3, 4$ when $\lambda_0 = 16$, $D = 8$, $N = 16$, $A(C) = C$, $h = 1$ and $\beta_i = 2$ for all retailers.

Figure 2.4: Illustration of the change of $AC(Q, (\mathbf{S}^*(Q, K)), K, C)$ in K



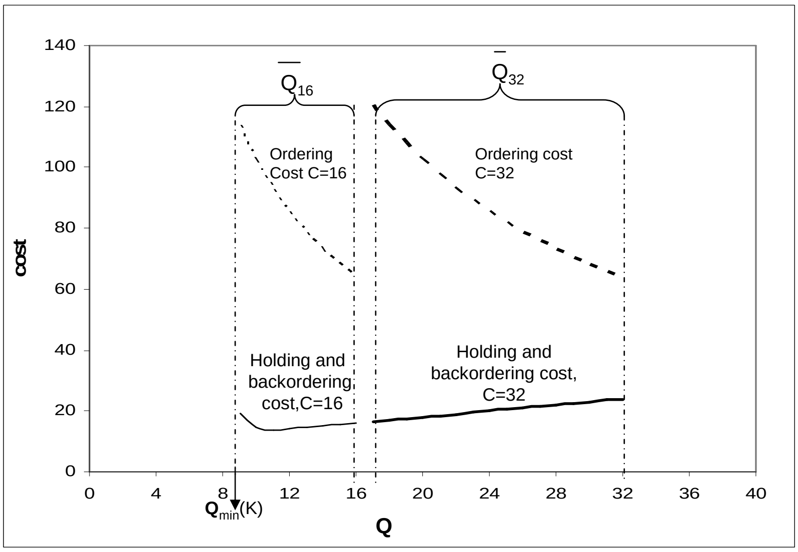
Recall that the trucks of the system operate as an $E_Q/D/K$ queue while the system uses a $(Q, \mathbf{S})$ policy and posseses K trucks. Also, recall that the waiting time distributions of $E_Q/D/K$ queues and $M/D/KQ$ queues are identical. Therefore, in a system that uses $(Q, \mathbf{S})$ policy for a given fleet size K, when one more truck is purchased, the waiting time distribution for a truck $F(w)$ is the waiting time distribution of a $M/D/(K+1)Q$ queue. So, buying one more truck changes the waiting time distribution of the system as if we buy Q more servers to a $M/D/KQ$ queue. On the other hand if the system used $(Q', \mathbf{S})$ policy for $Q' > Q$, buying one more truck would change the distribution of the system as if we bought Q' more servers to a $M/D/KQ'$ queue. Therefore, the traffic rate ρ of the system decreases more slowly with K, when the system uses $(Q, \mathbf{S})$ policy for $Q' > Q$.

Let $AC(Q, (\mathbf{S}^*(Q)), C)$ denote the average total cost rate function, when there is no fleet size limitation and where the joint order quantity is Q and the goods are dispatched with trucks of capacity C. In parallel to the discussion above, we observe in our numerical results that $AC(Q', (\mathbf{S}^*(Q', K)), K, C)$ approaches quicker to $AC(Q', (\mathbf{S}^*(Q')), C)$ as K increases compared to $AC(Q, (\mathbf{S}^*(Q, K)), K, C)$ approaches to $AC(Q, (\mathbf{S}^*(Q)), C)$ when $Q < Q'$.

Note that $K_{min}(4) < K_{min}(3) < K_{min}(2)$ in Figure 2.4. Since $K_{min}(Q)$ is the minimum integer value of K that makes $\rho = (\lambda_0 D/KQ) < 1$, $K_{min}(.)$ is a non-increasing function of Q. Next, we analyze the effects of Q and C on the total cost rate for a given fleet size K.

We now define $Q_{min}(K)$, which is the minimum joint order size to carry on a $(Q, \mathbf{S})$ policy for a given fleet size K. $Q_{min}(K)$ is the smallest positive integer Q that satisfies $\rho = (\lambda_0 \times D/K \times Q) < 1$. We also note that $Q_{min}(K)$ is non-increasing in K, that is $Q_{min}(K) \leq Q_{min}(K')$ if $K' < K$. Figure 2.5 sketches the effects of Q over the total cost rates for a given $K = 2$ and when we have two truck capacity options $C = 16$ and $C = 32$ with $\lambda_0 = 16$, $D = 1$, $N = 4$, $A(C) = 4 \times C$, $h = 1$ and $\beta_i = 4$. When there is no fleet size limitation, the

Figure 2.5: Illustration of the change of $AC(Q, (\mathbf{S}^*(Q, K)), K, C)$ in Q for a given K



holding and backorder cost rates increase as Q increases in $(Q, \mathbf{S})$ policy. However, as we can observe from Figure 2.5, the holding and backorder cost rate faces a decrease, while we increase Q at the beginning. This decline in the holding and backorder costs is due to the decrease in the traffic ratio of the $M/D/KQ$ queue. In general, we notice a decrease in the holding and backorder cost rate at the beginning when the traffic ratio $\frac{\lambda_0 \times D}{K \times Q_{min}(K)}$ is greater than 0.8 in our numerical results. An increase in Q leads to a decrease in the traffic ratio, which leads to a decrease in the holding and backorder cost rate when the traffic ratio is high. In other words, the effect of decreasing the traffic ratio on holding and backorder costs is more dominant than the effect of increasing the joint order quantity when the traffic ratio is high.

Here we observe from Figure 2.5 that the order setup cost decreases as Q increases for a given C, since the order setup cost is $A(C) \times \lambda_0/Q$ and $A(C)$ is a linear function of C. The order setup cost rates are the same for different truck capacities as long as they are utilized with 100% utilization. Let $\overline{Q}_C = (C_1, C_2, ..., C_m)$, with $C_1 < C_2 < ... < C_m$ denote the possible values of the order quantities for a given C. The following proposition suggests that using trucks with a higher utilization is always more profitable without changing other decision variables whenever it is possible.

Proposition 1: Suppose $C < C'$. If the intersection of $\overline{Q}_C$ and $\overline{Q}_{C'}$ is non-empty, it is always more profitable to choose C as a truck capacity for the intersecting Q values.

Proof: Referring to Subsection 2.3.2, $AC(Q, \mathbf{S}, K, C)$ consists of two parts: the order setup cost rate, which is $\lambda_0 \times \frac{A(C)}{Q}$, and the holding and backordering cost rates, which do not change with C when the other parameters are the same. So for $C < C'$, if there is a Q such that $Q \in \overline{Q}_C$ and $Q \in \overline{Q}_{C'}$, then the holding and backorder cost rates of C and C' are the same. However,
$\lambda_0 \times \frac{A(C)}{Q} < \lambda_0 \times \frac{A(C')}{Q}$ since $A(C) < A(C')$.$\square$

2.5 Different Scenarios and Solution Procedures

Firms can face the problem of transportation limitation in different forms. For instance the firm may not have the chance to choose the capacity if there is solely one kind of truck. In this section, we list different transportation scenarios that firms can encounter. In each of these scenarios, firms have to make decisions about the parameters of the joint order policy and about the features of their means of transportation. In order to help this decision-making process of the managers, we develop solution procedures based on the results that we discuss in Section 2.4. In this section, we take a new cost parameter into account: $\phi(C)$, which is the depreciation and maintenance cost rate per time we incur to a truck with a capacity C. We assume that $\phi(C)$ is a linear function of C, and our solution procedures are based upon this assumption.

Single Truck Capacity Option

This case can be frequently encountered in the industry. For instance, suppose that the truck to be used in the transportation is very specific to the unit sold in the retailers and that's why it is not vended often in the market and product diversity does not exist for that truck of interest (*e.g. hazardous materials*). In this case, K, Q and $\mathbf{S}^*(Q, K)$ are jointly optimised for a given truck capacity option C. Hence, built on our discussions in Section 2.4, Search Algorithm 1, given in Appendix A.2 is suggested.

Several Truck Capacity Options

This type of limitation is the most common type that the firms encounter in the market. Suppose that there are many truck producers, and they are providing trucks of different capacities. Therefore, there is a capacity option set, $\overline{C} = (C_1, C_2, ..., C_m)$, which consists of m truck capacity options ($C_1 < C_2 < ... < C_m$) that a firm can choose. Besides the fleet size K, order quantity Q and order-up to levels $\mathbf{S}^*(Q, K)$, the firms also have to decide upon the truck capacity C in order to minimize their total cost rates. Therefore, we suggest the Search Algorithm 2, given in Appendix A.2.

Since the single truck capacity option scenario is a special case of the several truck capacity options scenario, we only mention about Search Algorithm 2 in this section. In this search algorithm, we first obtain the possible values of the order quantities $\overline{Q}_{C_i}$ for each $C_i \in \overline{C}$. Due to the Proposition 1, we delete the intersecting elements of $\overline{Q}_{C_i}$ and $\overline{Q}_{C_j}$ from $\overline{Q}_{C_j}$ for each $i < j$. Then, for each

$C_i \in \overline{C}$, we find the minimum number of the trucks needed $K_{min}(Q)$ as well as the critical number of the fleet size, $K_{max}(Q, \phi(C_i))$ that is given by Equation (A.2), where buying one more truck increases the total cost rate (in the presence of maintenance/depreciation cost rate) for the first time for every $Q \in \overline{Q}_{C_i}$. Since the truck/maintenance cost is assumed to be linear with K, $K_{max}(Q, \phi(C_i))$ would give the optimal fleet size K^* for a given joint order quantity Q if the conjecture on the convex decreasing behavior of the $AC(Q, \mathbf{S}^*(K,Q), K, C_i)$ holds true. Subsequently, we search for the best joint order quantity Q_i^* for a given capacity C_i first, and then we choose the capacity C^* that brings the minimum cost rate among all of the capacity options. Hence, this algorithm finds K^*, C^*, Q^* and $\mathbf{S}^*(K^*, Q^*)$ accordingly.

Next, we consider the case, where the firm decides on the capacity of the trucks for a given number of the fleet size. The motivation behind a given fleet size number can be the area restriction of the hangar as well as the investment constraints. First, suppose that the truck producer guarantees to provide trucks with the capacity that the firm demands. Since the order setup cost rate is $A(C) \times \lambda_0/Q$, the firm aims to use the trucks with 100% utilization. In other words, the firm would like to purchase the trucks with capacity $C = Q$. In addition, we consider that the truck producer firm can have an upper limit for the capacity of the trucks, and no trucks with a capacity more than C_{max} can be produced. The firm has to revise the fleet size (K) decision, if $Q_{min}(K) > C_{max}$. Otherwise, the firm has to decide upon the joint order quantity $Q^* \in [Q_{min}(K), C_{max}]$ and the order-up to levels $\mathbf{S}^*(Q^*, K)$, for a given fleet size K in order to minimize the total cost rate of the system. After deciding the order quantity Q^*, the firm requires trucks with capacity $C = Q^*$ from the truck producer.

Now, rather than full flexibility on the capacity, we consider the case when there is a capacity option set, $\overline{C} = (C_1, C_2, ..., C_m)$, with $(C_1 < C_2 < ... < C_m)$. Suppose that the fleet size is K. Similar to the previous setting, the firm has to revise the fleet size (K) decision, when $Q_{min}(K) > C_m$. We suggest the Search Algorithm 3, given in Appendix A.2 for this limitation scenario.

2.6 Numerical Study

This section details a numerical study that illustrates the general behavior of the optimal policy parameters and the average cost rate with respect to different cost and system parameters. For the sensitivity analysis, all combinations of the following sets are analyzed: λ_0={2, 4, 8, 16, 32}, D={1, 2, 8}, $\overline{C}$={2, 4, 8, 16, 32}, $a = A(C)/C$={0.25, 1, 4}, h_i={1}, β_i={2, 4, 16, 32}, N={2, 4, 16}.

In all of the scenarios, the retailers are identical (same mean demand, holding and backorder cost rates and lead-times). Note that we ignore the truck maintenance and depreciation cost rate $\phi(C)$ as well as the minor lead times, $l_i = 0$, in our numerical analysis part.

Table 2.1: The Effects of the Change in $a = A(C)/C$ and K on Total Cost Rate

A(C):	0.25C			C			4C		
K	AC*	(Q*,S*)	Q*/C%	AC*	(Q*,S*)	Q*/C%	AC*	(Q*,S*)	Q*/C%
5	14.337	(8,7)	100%	17.337	(8,7)	100%	29.337	(8,7)	100%
6	14.180	(8,7)	100%	17.180	(8,7)	100%	29.180	(8,7)	100%
7	14.177	(8,7)	100%	17.177	(8,7)	100%	29.177	(8,7)	100%
8	14.114	(6,6)	75%	17.177	(8,7)	100%	29.177	(8,7)	100%
9	14.088	(5,6)	62.5%	17.177	(8,7)	100%	29.177	(8,7)	100%
10	14.062	(5,6)	62.5%						
11	14.060	(5,6)	62.5%						
12	14.060	(5,6)	62.5%						

The influence of A(C) and K on the performance of the system

In Table 2.1, we tabulate how the average cost rate changes with $A(C) = a \times C$ and fleet size K, where $N = 4$, $\lambda_0 = 4$, $C = 8$, $\beta_i = 4$ for all i and $D = 8$. We observe that for a given capacity C, truck utilization Q^*/C has a non-increasing structure when the fleet size K increases, because the system is enforced to have a higher utilization when there is a scarcity of trucks. In addition, as a increases, Q^*/C increases as well. This is due to the fact that the savings from the order set-up costs dominate the increase in holding and backorder costs as $A(C)$ increases. Hence, the truck utilization increases as $A(C)$ increases. Also, note that S^* has a non-increasing behavior with the fleet size. We can assert that the retailers try to balance the delay due to the absence of enough trucks with higher order up-to levels. Parallel to the conjecture that we made in Section 2.4, we observe the convex decreasing behavior of the total cost rate with K in Table 2.1 as well as in all of our numerical studies for the same system parameters.

The influence of backorder cost rate and K on the performance of the system

Table 2.2 illustrates the impacts of backorder cost rate β_i and fleet size K on the total cost rate where $N = 4$, $\lambda_0 = 4$, $C = 8$, $A(C) = C$ and $D = 8$.

Table 2.2: The Effects of the Change in β_i and K on Total Cost Rate

β_i:	4			16			32		
K	AC*	(Q*,S*)	Q*/C%	AC*	(Q*,S*)	Q*/C%	AC*	(Q*,S*)	Q*/C%
5	17.337	(8,7)	100%	24.041	(8,9)	100%	27.358	(8,10)	100%
6	17.180	(8,7)	100%	23.797	(8,9)	100%	26.908	(8,9)	100%
7	17.177	(8,7)	100%	23.793	(8,9)	100%	26.870	(7,9)	87.5%
8	17.177	(8,7)	100%	23.793	(8,9)	100%	26.862	(7,9)	87.5%
9	17.177	(8,7)	100%	23.793	(8,9)	100%	26.862	(7,9)	87.5%

From the numerical results, we observe that the retailers respond to higher backorder cost rates by increasing their order up-to levels. Also, contrary to the Table 2.1, truck utilization is nonincreasing in β_i, because the savings from the holding and backorder costs begin to dominate the increase in order set-up costs while having a lower truck utilization as β_i increases.

Joint Effects of K and C

Suppose that the truck capacity is exogenous to our system. In such a case, C has a great influence over the performance of our system, since it sets the limits for Q. Note that the order set-up cost rate is identical for all C as long as the trucks are utilized 100%, since $\frac{A(C)\times\lambda_0}{Q} = a \times \lambda_0$ when $Q = C$. Next, we present Table 2.3, which depicts the effects of C and K jointly on total cost rate where $N = 4$, $\lambda_0 = 2$, $\beta_i = 8$, $A(C) = C$ and $D = 8$. As expected, truck utilization percentage $\frac{100\times Q^*}{C}$ decreases whereas order up-to levels increase for bigger C. Notice that in some of the cases in Table 2.3, there can be an insignificant decrease in cost rates with respect to K. This is due to the fact that in these cases the traffic rate $\rho = \frac{\lambda_0 D}{KQ}$ is not that much high for the smallest K that makes $\rho < 1$. In addition, as we have discussed in Section 2.4, the cost rate of the system decreases more quickly to its lower bound for bigger C. Recall that the lower bound of AC is the cost rate of a system with the same parameters, where there is no limitation at all on the fleet size. If truck maintenance and depreciation cost rate $\phi(C)$ is taken into consideration, $K\phi(C)$ must be added to each of the cost rates given in Table 2.3 and the decisions must be made accordingly.

Table 2.3: The Effects of C and K on Total Cost Rate for $C = 2, 4, 8, 16$ and 32

	C=2				C=4		
K	AC*	(Q*,S*)	Q*/C%	K	AC*	(Q*,S*)	Q*/C%
9	13.246	(2,4)	100%	5	11.682	(4,4)	100%
10	11.494	(2,3)	100%	6	11.330	(4,4)	100%
11	10.897	(2,3)	100%	7	11.309	(4,4)	100%
12	10.735	(2,3)	100%	8	11.307	(4,4)	100%
13	10.687	(2,3)	100%	9	11.307	(4,4)	100%
14	10.674	(2,3)	100%				
15	10.670	(2,3)	100%				
16	10.669	(2,3)	100%				
17	10.669	(2,3)	100%				

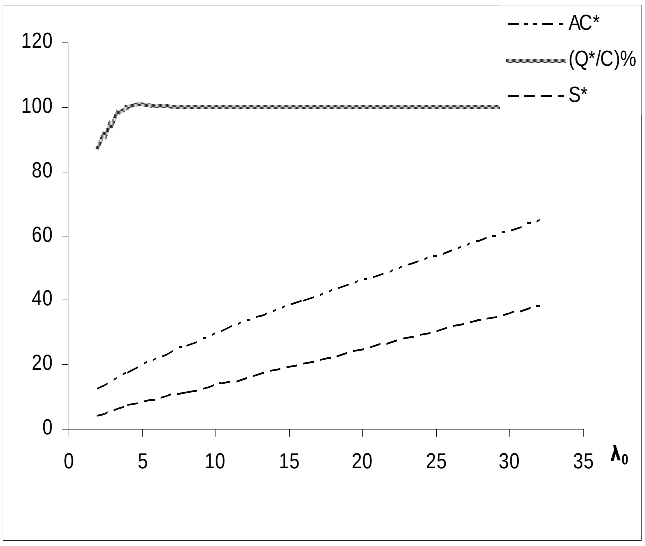
	C=8				C=16		
K	AC*	(Q*,S*)	Q*/C%	K	AC*	(Q*,S*)	Q*/C%
3	12.332	(8,4)	100%	2	14.212	(12,5)	75%
4	12.182	(7,4)	87.50%	3	14.147	(11,5)	68.75%
5	12.179	(7,4)	87.50%	4	14.147	(11,5)	68.75%
6	12.179	(7,4)	87.50%	5	14.147	(11,5)	68.75%

	C=32		
K	AC*	(Q*,S*)	Q*/C%
1	18.129	(21,7)	65.63%
2	17.095	(17,6)	53.13%
3	17.095	(17,6)	53.13%

Effects of D/2 and λ_0

The impacts of the distance between the ample supplier and the cross-dock point $D/2$ on total cost rate AC are similar to those of the total demand rate λ_0. Both D and λ_0 increase the traffic ratio ρ as well as the expected demand during lead-time. Hence, the system's reactions to higher D and λ_0 are higher order up-to levels at the retailer level and a bigger fleet size. Note that only holding and backorder costs at the retailer level increase as D increases, whereas order setup, holding and backorder costs increase as λ_0 increases for the same Q. In our numerical results, we observe that truck utilization Q^*/C increases with both D and λ_0. Our observations are consistent with Cachon's [10] suggestions, which claim that Q^* is increasing in both lead-time L and λ_0. Figure 2.6 depicts how the total cost rate AC^*, order up-to level S^* and truck utilization percentage Q^*/C change with λ_0 where $C = 8$, $A(C) = C$, $\beta_i = 4$, $N = 4$ and $D = 8$ provided that there is an ample number of trucks.

Figure 2.6: Illustration of the change of AC^*, S^* and Q^*/C with λ_0



Effects of the number of retailers N

Table 2.4 tabulates how the optimal system parameters (Q^*, S^*, C^*) change with K for different N, where we have a capacity option set $\overline{C} = 2, 4, 8, 16, 32$, $\lambda_0 = 4$, $\beta_i = 4$, $A(C) = C$ and $D = 8$. In our numerical results, we observe an increase in holding and backordering costs as N increases. This increase is due to the fact that we enjoy the benefits of *risk pooling* when N is smaller. Note that the total demand rate λ_0 is remained fixed for all N. When N is greater, a bigger amount of inventory is held at the retailer level for the same λ_0. When there is no fleet size limitation, we expect that an increase in N lead to a decrease in truck utilization percentage (See Cachon [10]). However, we observe that $(Q^*/C)\%$ increases as N increases when $K = 2$ in Table 2.4. In this specific case, limited fleet size brings about a longer effective lead time, which increases $Q^*/C\%$, and the impact of a longer effective lead-time dominates that of a bigger N on truck utilization percentage.

Table 2.4: The Effects of K on Optimal Cost Parameters for different N

	N=2	N=4	N=16
K	(Q*, S*, C*)	(Q*, S*, C*)	(Q*, S*, C*)
2	(21, 17, 32)	(21,9,32)	(23,3,32)
3	(16, 15, 16)	(16,8,16)	(14,2,16)
4	(16, 15, 16)	(15,8,16)	(11,2,16)
5	(8, 13, 8)	(8,7,8)	(8,2,8)
6	(8, 12, 8)	(8,7,8)	(8,2,8)
7	(8,12,8)	(8,7,8)	(8,2,8)
8	(8,12,8)	(8,7,8)	(8,2,8)
9	(4,11,4)	(4,6,4)	(4,2,4)
10	(4,11,4)	(4,6,4)	(4,2,4)
11	(4,11,4)	(4,6,4)	(4,2,4)
12	(4,11,4)	(4,6,4)	(4,2,4)
13	(4,11,4)	(4,6,4)	(4,2,4)
14	(2,11,2)	(2,6,2)	(2,2,2)
20	(2,11,2)	(2,6,2)	(2,2,2)
21	(2,11,2)	(2,6,2)	(2,2,2)
22	(2,11,2)	(2,6,2)	(2,2,2)
23	(2,11,2)	(2,6,2)	(2,2,2)
24	(2,11,2)	(2,6,2)	(2,2,2)
25	(2,11,2)	(2,6,2)	(2,2,2)
26		(2,6,2)	(2,2,2)
27		(2,6,2)	(2,2,2)

Chapter 3

Two-Echelon Environment

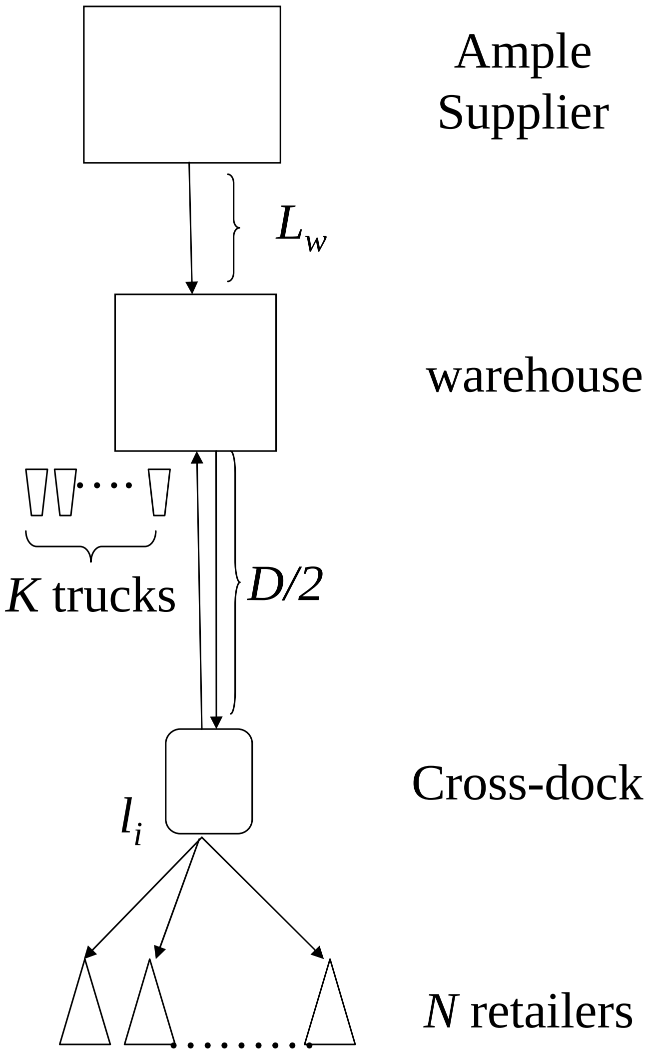
In this chapter we extend our analytical model to a single-item, multi-retailer distribution system, where the upper echelon also holds inventory.

Main characteristics of this extended model and the ordering policies for both of the echelons are given in Section 3.1. In Section 3.2, the framework of our analysis for each installation is explained, and necessary analytical tools for the derivation of the operating characteristics are presented. Afterwards, we derive the key operating characteristics of our two-echelon distribution system in Section 3.3.

3.1 Model Characteristics

We consider a single-item, divergent two-echelon inventory system with a single warehouse and N retailers, where the system is continuously reviewed. The retailers face stationary and independent unit Poison demands with rates λ_i ($i = 1, 2, ..N$), and all unmet demands are fully backlogged. Retailer demands are supplied from the warehouse, which is located at the upper echelon. A cross-dock station is situated at a distance of $D/2$ time units from the warehouse. Similar to the single-echelon case, no order-allocation takes place in this cross-dock. A

Figure 3.1: Illustration of the extended model



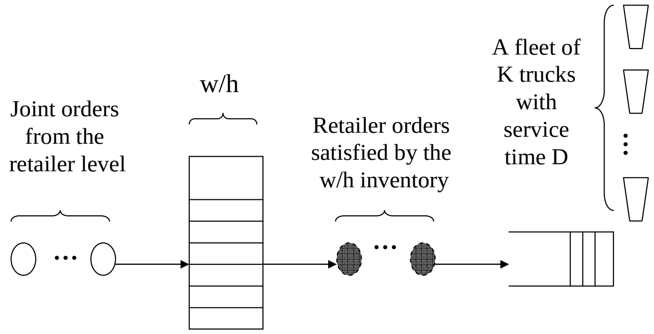
fleet of K identical trucks with capacity C are used in the transportation of retailer orders from the warehouse to the cross-dock station. We assume that the transportation between the cross-dock station and the retailers and between the warehouse and the ample supplier are not capacitated and the lead time between the retailers and the cross-dock point is l_i for retailer i. The warehouse gives orders to an external supplier with ample stock and the lead-time for warehouse deliveries is L_w. Figure 3.1 illustrates the extended model.

Holding cost per unit per time is charged at each retailer with rate h_i, and at the warehouse with rate h_w. The ordering costs associated with the distribution system are the warehouse fixed ordering cost K_{wh} for each warehouse order and the fixed retailer setup cost, $A(C)$, for each retailer order. Fixed retailer order setup cost is linear in truck capacity, that is $A(C) = a \times C$ for $a > 0$. In addition, a shortage cost per unit per time is incurred at a rate of β_i for each retailer i. Any possible additional costs for monitoring the inventory system continuously are ignored. The retailer orders are satisfied based on the first come-first serve, $FCFS$ rule at the warehouse, and the integrity of the orders must be sustained.

As in the single-echelon distribution system that is analyzed in the previous chapter, the retailers implement a joint replenishment policy to manage their inventory and replenishment decisions. The orders received from the retailers constitute a compound renewal process, where inter-order time, X, and the order quantity, Q_0, have a joint density $f_{X,Q_0}(x, q)$. Joint orders received by the warehouse are shipped immediately by trucks if there are both sufficiently many trucks and sufficiently many units available in the inventory. In that case, loaded trucks spend a total of D time units to reach the cross-dock point, unload the item and return back to its base. Therefore, the order delivery lead time for retailer i is constant, $L_i = l_i + D/2$, provided that there is no delay encountered. However, if there are not sufficiently many units available in the warehouse inventory at the time of a retailer order, then the retailer order waits until enough number of units in the warehouse inventory becomes available. After the retailer order is satisfied by the warehouse, it should be loaded onto the trucks for transportation to the cross-dock station. If there are not sufficiently many trucks available at the base, then the retailer order waits until a truck becomes available. Hence, the shipment operations of the retailer orders at the warehouse can be modeled as a queueing model like in the single-echelon case. However, this time the satisfied retailer orders by the warehouse inventory system are the *customers* and the trucks are the *servers* which are busy while carrying the materials to the cross-dock point and returning back. Figure 3.2 depicts the shipment operation of the joint retailer orders in our setting.

The joint replenishment policy that retailers use for controlling their inventory decisions in this setting is the same as the joint replenishment policy that is used in the single-echelon setting, which is the $(Q, \mathbf{S})$ policy. In our setting, the warehouse

Figure 3.2: Illustration of the Shipment Operations of the Retailer Orders



employs an inventory control policy, too. It is assumed that the continuous review installation base-stock, $(S-1, S)$ ordering policy is employed for controlling the warehouse level inventory and ordering decisions. When $K_{wh} > 0$, $(S-1, S)$ is a sub-optimal policy, nevertheless, we consider this policy for analytical tractability. Under this policy, whenever a demand from the lower echelon arrives at the warehouse, a replenishment order is placed at the ample supplier. Hence the inventory position of the warehouse is always equal to its order-up-to level, which is S_w in our setting. Under these model assumptions and the cost structure that are defined above, our objective is to minimize the expected total cost per unit time. Next, we present the framework of our method and give the necessary analytical derivations to obtain the operating characteristics of our system.

3.2 Preliminary Analysis

In this section, we present a general framework of our analysis and explain some of the analytical derivations that we need in order to derive the operating characteristics of our two-echelon distribution system.

We use the same notation as in the previous chapter with a few additions. Let W denote the random delay that a joint retailer order encounters before it is dispatched by the trucks. W consists of two parts. These are the random delay due to the lack of sufficient inventory, W_s at the warehouse and the random delay due to the lack of sufficient trucks, W_q at the transportation base. The lead-time

L_i (total transit time) for retailer i with the random delay, $W = W_s + W_q$, an order encounters constitute the effective lead-time for retailer i, which is denoted by $\overline{L_i} = L_i + W$. In our derivations, we assume that W_s and W_q are independent random variables, however there may be an implicit dependence between W_s and W_q. As mentioned before, the joint retailer orders form a compound renewal process, where inter-order time, X, and the order quantity, Q_0, have a joint density $f_{X,Q_0}(x, q)$. Since the SJRP that the retailers use is the $(Q, \mathbf{S})$ policy, $Q_0 = Q$ and X has an $Erlang(Q, \lambda_0)$ distribution. A joint retailer order that is received at a warehouse is perceived as a batch demand to be satisfied by the warehouse. Next, we focus on the analysis of the operations at the warehouse level.

The warehouse uses an $(S-1, S)$ policy with an order up to level S_w. Since partial shipments are not allowed, the optimal S_w value would be an integer multiple of the batch size Q. Therefore, we assume that the order up to level of the warehouse is $S_w = \triangle \times Q$ for $\triangle \geq 0$. The distribution function of the random delay due to the lack of sufficient inventory at the warehouse is given by Özkaya [27] for various stochastic joint replenishment policies. The following equation provides an expression for the distribution function $F_{W_s}(\tau)$ of W_s, where the retailers use the $(Q, \mathbf{S})$ policy with an order quantity of Q and the warehouse order up to level S_w is equal to $\triangle \times Q$:

$$F_{W_s}(\tau) = \begin{cases} 0 & \tau < 0 \\ (1 - F(L_w - \tau, \triangle \times Q, \lambda_0) & 0 \leq \tau \leq L_w \\ 1 & \tau \geq L_w. \end{cases} \tag{3.1}$$

Note that W_s is always L_w when the warehouse does not hold any inventory; in other words, when it employs cross-docking.

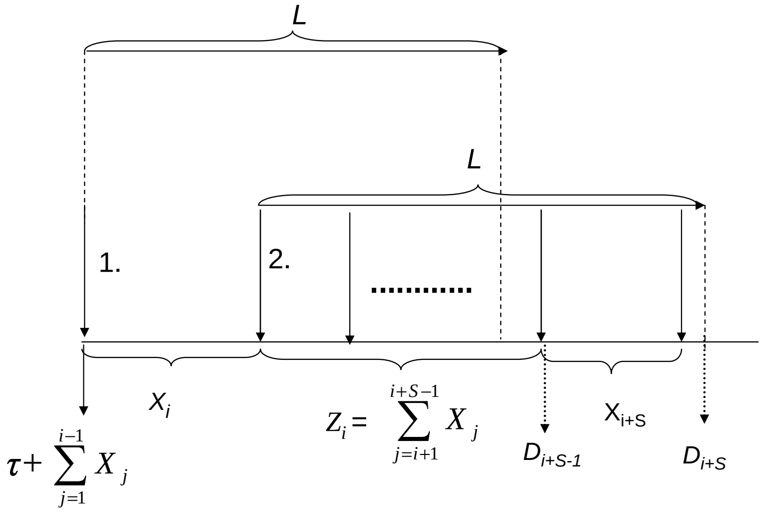
As mentioned before, the shipment of the joint retailer orders can be considered as a queueing system. Therefore, the waiting time W_q a joint retailer order encounters at the transportation base is the waiting time in the queue of a customer in a queueing system, which consists of k servers (trucks) with a deterministic service time $D = 2 \times D/2$. In order to derive the distribution function of W_q, we need to characterize the arrival process of the queueing system. After the joint retailer orders are satisfied by the warehouse, they are perceived as the arriving customers by the trucks, and the departure process of the $(S-1, S)$ inventory system at the warehouse level constitutes the arrival process of the queueing system that is mentioned above. Next, we analyze the characteristics of the departure process of the $(S-1, S)$ inventory systems.

3.2.1 General Characteristics of the Departure Process of a $(S-1, S)$ Inventory System

In this subsection, we analyze the general characteristics of the departure process of $(S-1, S)$ inventory systems with unit renewal demand arrivals. The lead-time, L between the inventory system and the ample supplier is deterministic. Let X_i denote the inter-arrival time between the $(i-1)^{th}$ and i^{th} consecutive demands. Since the inter-arrival times are *i.i.d* random variables, density and distribution functions of X_i, $f_X(\cdot)$ and $F_X(\cdot)$ are same for each i and the n^{th} demand arrival time, $X^{(n)}$ is the sum of the durations of n inter-arrival times and the density function of $X^{(n)}$ is the n^{th} convolution of $F_X(\cdot)$. Let $f_{X^{(n)}}(\cdot)$ and $F_{X^{(n)}}(\cdot)$ be the density and distribution functions of $X^{(n)}$ for any integer $n \geq 0$. Note that $X^{(0)}$ is 0 with probability 1 and we take $F_X(x) = 0$ and $f_X(x) = 0$ when $x < 0$ for convenience. Also, recall that $\overline{F}= 1-F$ connotes the complementary distribution for any distribution function.

A unit demand departs the inventory system immediately if there is enough inventory on hand, otherwise it waits for the inventory on road to arrive. We present an illustration of the departures of consecutive demands in Figure 3.3. First, we analyze the case when $S > 0$. Suppose a demand arrives at the inventory system at time τ. Since the $(S-1, S)$ policy is employed, a demand always triggers an order. The S^{th} subsequent demand is satisfied by this triggered order. If there are sufficient number of items in the stock, it is immediately satisfied, hence its departure time will be : $\tau + \displaystyle\sum_{i=1}^{S} X_i$. Otherwise, it waits for the arrival of the triggered order to be satisfied and hence its departure time will be: $\tau + L$. Now let D_S be the departure time of the S^{th} subsequent demand after τ. According

Figure 3.3: Illustration of the Consecutive Demand Departures



to the discussions above, D_S can be written as follows:

$$D_S = \tau + max\left(\sum_{j=1}^{S} X_j, L\right)$$

Now, consider the i^{th} subsequent demand that arrives after time τ. This demand triggers another order, which arrives at time $\tau + \sum_{j=1}^{i} X_j + L$. Likewise the previously triggered orders, this one meets the S^{th} subsequent demand after its trigger time. Hence we can write the departure time, D_{i+S} of the $(i+S)^{th}$ subsequent demand after τ as

$$D_{i+S} = \tau + max\left(\sum_{j=1}^{i+S} X_j, \sum_{j=1}^{i} X_j + L\right).$$

Similarly, the departure time, D_{i+S-1} of the $(i+S-1)^{th}$ subsequent demand is given below:

$$D_{i+S-1} = \tau + max\left(\sum_{j=1}^{i+S-1} X_j, \sum_{j=1}^{i-1} X_j + L\right).$$

We define the inter-departure time as the time between two consecutive departures from the inventory system. Let Y_{i+S} be the time between the departures D_{i+S-1} and D_{i+S}. Hence, we can write Y_{i+S} as:

$$
\begin{aligned}
Y_{i+S} &= D_{i+S} - D_{i+S-1} \\
= \tau + max\left(\sum_{j=1}^{i+S} X_j, \sum_{j=1}^{i} X_j + L\right) &- \left(\tau + max\left(\sum_{j=1}^{i+S-1} X_j, \sum_{j=1}^{i-1} X_j + L\right)\right) \\
&= max\left(\sum_{j=i}^{i+S} X_j, X_i + L\right) - max\left(\sum_{j=i}^{i+S-1} X_j, L\right)
\end{aligned}
$$

Let $Z_i = \sum_{j=i+1}^{i+S-1} X_j$. Since the inter-arrival times are i.i.d random variables, the distribution $F_Z(\cdot)$ and density $f_Z(\cdot)$ functions of Z_i are identical to the distribution $F_{X^{(S-1)}}(\cdot)$ and density $f_{X^{(n)}}(\cdot)$ functions of $X^{(S-1)}$ for all i.

Hence, the following expression for Y_{i+S} provides a more convenient representation for the derivation of the distribution of inter-departure times:

$$
Y_{i+S} = \begin{cases}
X_{i+S} & \text{if} \quad (L - Z_i) \leq min(X_i, X_{i+S}) \\
X_i + Z_i + X_{i+S} - L & \text{if} \quad X_i < (L - Z_i) \leq X_{i+S} \\
L - Z_i & \text{if} \quad X_{i+S} < (L - Z_i) \leq X_i \\
X_i & \text{if} \quad (L - Z_i) > max(X_i, X_{i+S})
\end{cases} . \tag{3.2}
$$

The expression of the inter-departure times in Equation (3.2) can be generalized for every integer index i, solely by changing of the variables. Note that Y_{i+S} and all Y_j for $j < i$ are independent, because all of the constituents of Y_j and of Y_{i+S} are independent for $j < i$. In the next theorem we present the probability density function $f_{Y_i}(y)$ of the inter-departure times. Recall that the inter-arrival times, X_j are i.i.d random variables and $f_X(.)$ and $F_X(.)$ denote the density and the distribution functions of the inter-arrival times.

Theorem 4 *The probability density function $f_Y(y) = f_{Y_i}(y)$ of the inter-departure time Y_i of an $(S-1, S)$ inventory system with deterministic lead time* *and unit renewal demands is identical for all i, which is given as follows.*

$$\begin{aligned} f_Y(y) = f_X(y) \Big\{ &\overline{F}_Z(L-y)\overline{F}_X(y) + F_Z(L-y)F_X(y) + \int_{x=0}^{y} f_X(x)\overline{F}_Z(L-x)dx \\ &+ \int_{x=y}^{\infty} f_X(x)F_Z(L-x)dx \Big\} + f_Z(L-y)\overline{F}_X(y)F_X(y) \\ &+ \int_{x_2=y}^{L+y} \int_{x_1=0}^{min(y,L+y-x_2)} f_Z(L+y-(x_1+x_2))f_X(x_1)f_X(x_2)dx_1dx_2 \end{aligned} \tag{3.3}$$

for $y \geq 0$

Proof: See Appendix B.

Since the distribution of Y_i is identical for all i, $E[Y_i] = E[Y]$ and $Var[Y_i] = Var[Y]$ are identical for all i, too. From the Equation (3.2), we can see that the departure process has a recursive structure and each departure is dependent on the $S+1$ arrivals before that departure. Hence, the covariance $Cov(Y_i, Y_j)$ of Y_i and Y_j is identical to $Cov(Y_{i+k}, Y_{j+k})$ for every $k \in \mathcal{N}$. Also, $Cov(Y_{i+S}, Y_j) = 0$ for $j < i$, since Y_{i+S} and all Y_j for $j < i$ are independent. Next, we analyze the characteristics of the mean, $E[Y]$ and the variance, $Var[Y]$ of the inter-departure times.

Theorem 5 *Let E[X] and Var[X] be the expectation and the variance of the inter-arrival time X, respectively. Similarly, E[Y] and Var[Y] are the expectation and the variance of the inter-departure time Y.*
Then, $E[Y] = E[X]$, $Var[Y] \leq Var[X]$ and

$$Var[X] - Var[Y] = 2\int_{z=0}^{\infty} \left\{ \left(\int_{x=L-z}^{\infty} \overline{F_X}(x)dx \right) \left(\int_{x=0}^{L-z} F_X(x)dx \right) \right\} f_Z(z)dz$$

Proof: See Appendix B.

Let $N_D(t)$ represent the total number of departures that have occurred up to time t. In the light of the discussions above, we can conclude that the departure process, $(N_D(t), t \geq 0)$ of an $(S-1, S)$ inventory system is a counting process, where the elapsed time between the consecutive departures, Y_i are identical but not independent random variables.

Returning back to the analysis of our model, the warehouse in our distribution system uses $(S-1, S)$ policy, too. An order is placed immediately to the ample supplier as a joint order from the retailer level arrives at the warehouse. Since $S_w = \triangle Q$, the departure process of the warehouse is identical to a $(\triangle - 1, \triangle)$ inventory system where the unit demands arrive in batches of size Q with $Erlang(Q, \lambda_0)$ distributed inter-arrival times X. Thus, $E[X] = \frac{Q}{\lambda_0}$ and $Var[X] = \frac{Q}{\lambda_0^2}$. Also, the probability density function of the inter-departure times of the warehouse with $S_W = \triangle Q$ in our system can be simply found by using Theorem 4 for $y \geq 0$. Next, we analyze the shipment operations of the retailer orders at the truck base.

3.2.2 Analysis of the Shipment Operations of the Retailer Orders at the Truck Base

As mentioned in Section 3.1, the shipment operations of the retailer orders at the truck base operate as a queueing system with multiple servers and deterministic service time, and the arrival process of the customers of this queueing system is the departure process of the warehouse inventory system, whose characteristics are analyzed in the previous subsection.

In order to derive the operating characteristics, we need the distribution of the second constituent of the total random delay, W_q which is the random delay due to the lack of sufficient trucks. W_q is identical to the waiting time a customer encounters in a $G/D/K$ queue with dependent inter-arrival times, where the inter-arrival times of the customers are same as the inter-departure times of the $(S-1, S)$ warehouse inventory system.

The behavior of the distribution function of the waiting time in the queue of this $G/D/K$ queueing system is analytically intractable. Therefore, we approximate the arrival process of the customers to a simpler one so that we can analyze the behavior of the W_q

As we have developed an analytical tool that generates the $F_{W_q}(\tau)$ of any $E_k/D/c$ queue with a predetermined error bound in Subsection 2.3.1, we next examine how to approximate the original arrival process of the customers, $(N_D(t), t \geq 0)$, to a renewal process with Erlang distributed inter-arrival times. First note that $Y_{i+S} = X_{i+S}$, when $L = 0$ and $Y_{i+S} = X_i$ when $L = \infty$ as can be observed from Equation (3.2). In our numerical results, we also observed that when L_w is much longer or shorter than the depletion time of the $\triangle^{th}$ inventory batch, the distribution of the queue inter-arrival time is almost identical to that of an Erlang(Q, λ_0) random variable. If however L_w is close to the depletion time of the $\triangle^{th}$ inventory batch, the distribution function of the queue inter-arrival time deviates from the density of Erlang (Q, λ_0). The following method approximates the arrival process of the queue by a renewal process with Erlang(Q', λ') distributed inter-arrival times using the moments $E[Y] = E[X] = \frac{Q}{\lambda_0}$ and $Var[Y]$, which we derive from Theorem 5:

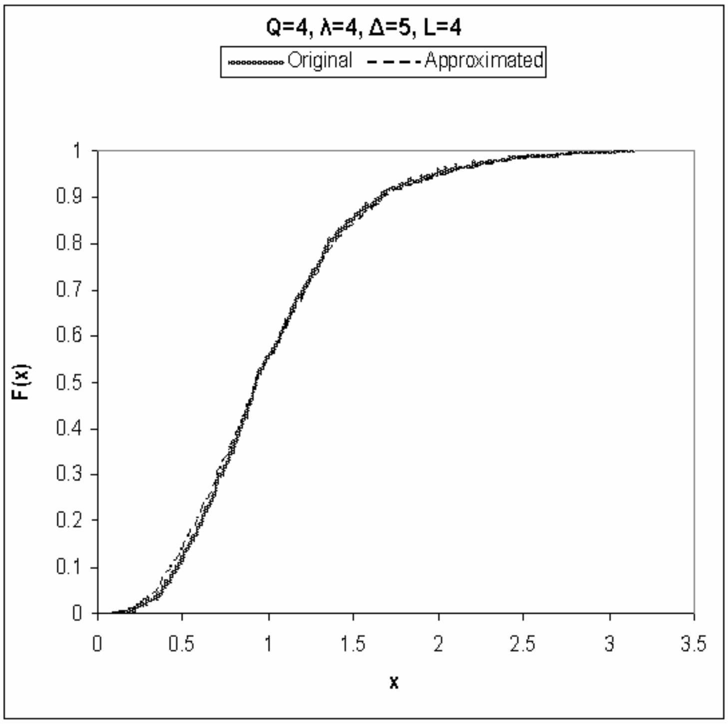
Approximation Method

1. Find $E[Y]$ and $Var[Y]$
2. Set $Q' = [E[Y]^2/Var[Y]]$ where $[x]$ is the nearest integer to x.
3. Set $\lambda' = Q'/E[Y]$.

Using this method, we approximate the arrival process of the queue to a renewal process with Erlang(Q', λ') inter-arrival times with mean $E[Y] = Q/\lambda = Q'/\lambda'$ and variance $Q'/(\lambda')^2$

In Figures 3.4 and 3.5, some examples are presented to illustrate the performance of our Erlang approximation. In Figure 3.4, we present the exact distribution function and the Erlang Approximation of the queue inter-arrival times with $Q = 4$, $\lambda_0 = 4$, $\triangle = 5$ and $L_{wh} = 4$. In this particular example, the approximated scale, $\lambda' = \lambda_0$ and shape $Q' = Q$ parameters are not changed, and parallel to the observation that is mentioned above, the queue inter-arrival time

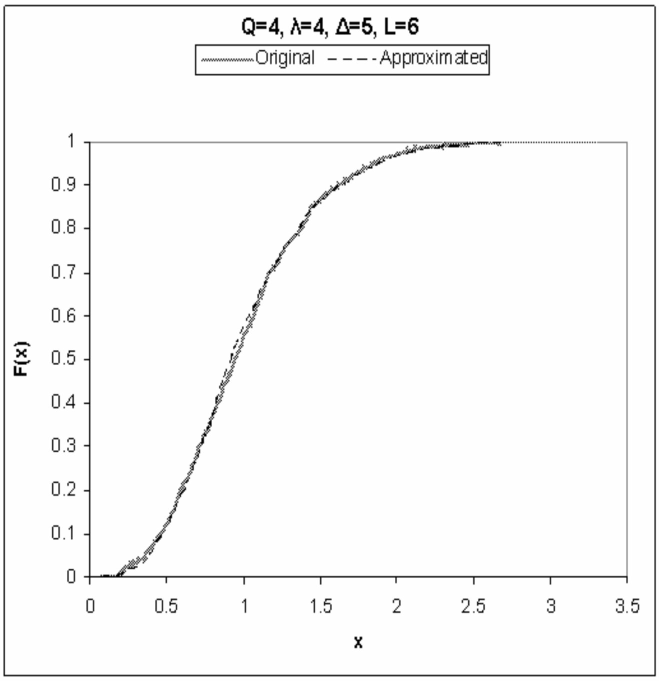
Figure 3.4: Comparison of the Erlang Approximation with the Queue Inter-arrival Times



distribution is almost identical to the distribution of an Erlang(Q, λ_0) random variable. On the contrary, for the instances, where $L_w = 6$, given in Figure 3.5, the adjusted scale, $\lambda' = 5$ and shape $Q' = 5$ parameters are different from λ_0 and Q. In this example, the variance difference, $Var[X] - Var[Y]$ transforms the queue inter-arrival time to a random variable with the same mean and a smaller variance. The approximated scale, $\lambda' = \lambda_0$ and shape $Q' = Q$ parameters are not changed for $L_w > 6$ and $L_w < 4$ as well.

The performance of the Erlang distribution to approximate the queue inter-arrival time distribution increases when the L_w is much more larger or smaller than the depletion time of the $\triangle^{th}$ inventory batch. However, the Erlang distribution is good enough to approximate the queue inter-arrival time distribution with adjusted shape and scale parameters when L_w is around the depletion time of the $\triangle^{th}$ inventory batch, too. In addition, the effect of the approximation on the operating characteristics is very negligible in most of the cases as will be mentioned in the subsequent sections of this chapter.

Figure 3.5: Comparison of the Erlang Approximation with the Queue Inter-arrival Times



After the approximation of the arrival process of the queue to a renewal process with Erlang(Q', λ') inter-arrival times, we can derive the distribution function of the random delay due to the lack of sufficient trucks by following the steps described in the Subsection 2.3.1. Therefore, the distribution $F_W(\cdot)$ of the total random delay $W = W_s + W_q$ can be derived as follows:

$$F_W(x) = \int_{y=0}^{L_w} F_{W_q}(x-y) dF_{W_s}(y). \tag{3.4}$$

After deriving the distribution function of the total random delay a joint retailer order encounters before it is dispatched, we present the derivation of the operating characteristics in the next section.

3.3 Derivation of the Operating Characteristics

In this section, the operating characteristics of the two-echelon distribution system is derived, and these characteristics are used in calculating the total cost rate. Similar to the single-echelon model, total cost of the two-echelon model consists of two parts. The first part is the order setup costs and the latter is the holding and the backorder costs. We begin with the expected cycle length. We define the cycle length as the time between two consecutive joint order replenishments, X for the retailer part, and as the time between two consecutive warehouse order placements for the warehouse part. Since the warehouse uses $(S-1, S)$ policy, it places an order to the ample supplier as a joint retailer order arrives, therefore the expected cycle length of the retailer part and the warehouse part are the same. So, the expectation of X, which is an Erlang(Q, λ_0) distributed random variable is simply Q/λ_0. In each cycle, the fixed ordering cost of a joint retailer order, $A(C)$ and the fixed ordering cost of a warehouse order, K_{wh} are incurred once, and the order setup cost rate is simply $(A(C) + K_{wh}) \times \lambda_0/Q$. Next, we derive the backlogging and holding cost rate at the retailer level.

Note that the *no-order-crossing* condition is also satisfied in this two-echelon model, since both the warehouse and the queue serve based on the $FCFS$ principle. Therefore the backorder and the holding costs incurred at each retailer can be evaluated in a similar fashion to our single echelon model. Equation (2.5) that is given in Chapter 2 can be used to evaluate the expected backorder and holding costs incurred at retailer i for a given effective lead-time $\overline{l_i}$. After deriving the distribution function $F_W(\cdot)$ of the total random delay W, which is given in Equation (3.4), we can take the expectation of the holding and backorder cost rate of a unit demand to retailer i over effective lead-times. Hence, the expected holding and backorder cost expression of a unit demand to retailer i per unit time is given by:

$$U(Q, \mathbf{S}, K, \triangle)_i = \int_w \frac{1}{Q} \sum_{n=0}^{Q-1} \sum_{m_i=0}^{n} Z(m_i|n) g_i(S_i - m_i, L_i + w) dF_W(w). \quad (3.5)$$

Recall that $F_W(w)$ is dependent on K, which affects the $F_{W_q}(\cdot)$ and on $S_{wh} =$

$\triangle \times Q$, which affects $F_{W_s}(\cdot)$. Hence, the holding and backorder cost rate of the whole retailer level is: $\sum_{i=0}^{N} \lambda_i \times U(Q, \mathbf{S}, K, \triangle)_i$. Next, we derive the holding cost rate incurred in the warehouse.

We use a similar method to that of Axsäter [3], while calculating the holding cost rate at the warehouse level. We incur holding cost for a joint retailer demand of size Q that arrives at the warehouse at time τ, if the $\triangle^{th}$ subsequent joint retailer demand arrives before $\tau + L_w$. Therefore, the expected time a joint retailer order is incurred a holding cost at the warehouse is as follows:

$$\int_{x=0}^{L_w} (L_w - x) f(x, \triangle Q, \lambda_0) dx. \tag{3.6}$$

Using Equations (2.2) and (2.3), Equation (3.7) can be reduced to the expression below:

$$L_w - \frac{\triangle Q}{\lambda_0} + \frac{\triangle Q}{\lambda_0} F_P(\triangle Q, \lambda_0 L_w) - L_w F_P(\triangle Q - 1, \lambda_0 L_w). \tag{3.7}$$

In addition to the time expression above, holding cost is incurred during W_q, while it waits for dispatching. Therefore, $E[W_q]$, which can be computed via Theorem 3 with adjusted shape, Q' and scale λ' parameters inserted to Equation (3.7). Hence, we can derive the holding cost rate incurred in the warehouse level, $WH(Q, \triangle, K)$:

$$\begin{aligned} WH(Q, \triangle, K) = h_w Q \frac{\lambda_0}{Q} \Big\{ & E[W_q] + L_w - \frac{\triangle Q}{\lambda_0} + \frac{\triangle Q}{\lambda_0} F_P(\triangle Q, \lambda_0 L_w) \\ & - L_w F_P(\triangle Q - 1, \lambda_0 L_w) \Big\}. \end{aligned}$$

Finally the expected cost rate of the whole distribution system is given by:

$$AC(Q, \mathbf{S}, K, C, \triangle) = \lambda_0 \frac{(A(C) + K_{wh})}{Q} + \sum_{i=1}^{N} \lambda_i U(Q, \mathbf{S}, K, \triangle)_i + WH(Q, \triangle, K). \tag{3.8}$$

The first part of the equation above represents the retailer and warehouse order setup cost rate and the second part represents the holding and backorder cost incurred per time in the retailer level of our distribution system that is using a $(Q, \mathbf{S})$ policy with a fleet of K trucks, each having a capacity of C, and finally the third part represents the holding cost incurred per time in the warehouse level.

Considering the truck utilization constraint the optimization problem of our system can be stated as follows:

$$\begin{gathered}\text{Min. } AC(Q, \mathbf{S}, K, C, \triangle) \\ \text{s.t. } Q \in (C/2, C]\end{gathered}$$

3.4 Numerical Study

This section consists of two parts. The first part details a numerical study that illustrates the general behavior of the optimal policy parameters and the average cost rate with respect to different cost and system parameters. Moreover, we discuss the possible reasons of these behaviors in this subsection. In the second part, we discuss the performance quality of our approximation method that we mention in Subsection 3.2.2.

3.4.1 Sensitivity of the Total Cost Rate to Cost and System Parameters

In this subsection we conduct a sensitivity analysis. The computations and optimization study are carried out according to the approximations we made. We search over a vast interval that consists of positive integers while searching for $\triangle^*$ for a given Q and $S^*(Q)$. For the sensitivity analysis, all combinations of the following sets are analyzed: λ_0={2, 4, 8, 16, 32}, D={8}, $\overline{C}$={2, 4, 8, 16, 32}, $A(C)/C$={0.25, 1, 4}, h_i={1}, β_i={2, 4, 16, 32}, $N = 4$, $h_w = 0.5$, $L_w = \{2\}$ and $K_{wh} = \{1\}$. In all of the scenarios, the retailers are assumed to be identical in their cost, lead-time and demand parameters and $N = 4$, $h_w = 0.5$. Also we neglect the maintenance and depreciation cost factor $\phi(C)$ in our numerical study similar to Section 2.6.

While commenting on the behavior of the total cost rate under different parameters in the two-echelon system, the warehouse holding and order set-up costs must be taken into the consideration as well. In the two echelon system, warehouse order up-to level, $S_{wh} = \triangle Q$ is a decision parameter like the retailer order up-to levels $\mathbf{S}$. Therefore, we also discuss the impacts of the cost and system parameters on $\triangle^*$ in this subsection.

Note that the effective lead-times increase as S_{wh} decreases due to the Equation (3.1), and higher effective lead-times bring about higher S^*. Parallel to the discussion above, we observe that S^* is non-increasing in $\triangle$ in our numerical resulus.

Recall that in the single echelon case, there is a trade-off between the order set-up costs and holding and backorder costs in the retailer level. For a given C, order set-up costs decrease whereas holding and backorder costs increase if the joint order quantity Q is accrued. In the two echelon case, this trade-off can be generalized to systemwide order set-up and systemwide holding and backorder costs.

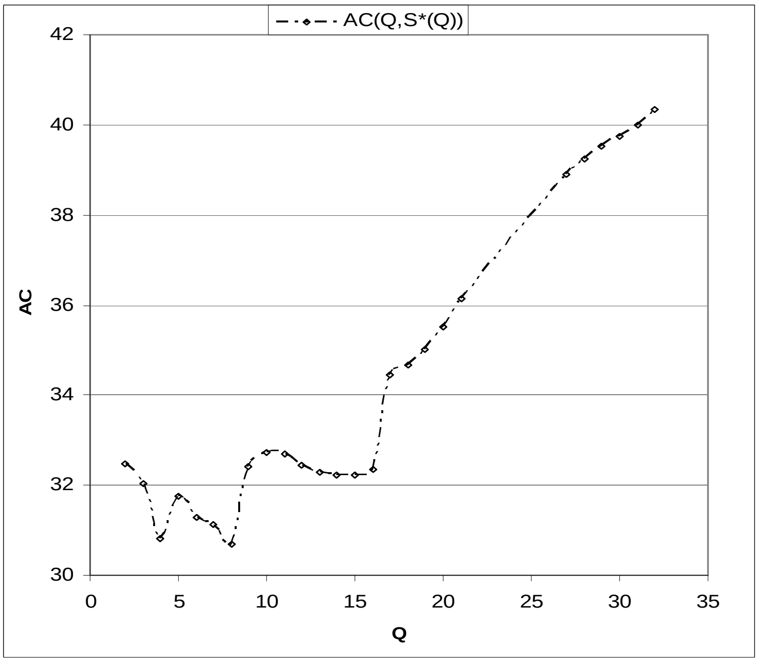
The Impact of A(C) and K on the Performance of the System
Next, we examine the impact of $a = A(C)/C$ and K on the performance of our system. In Table 3.1, we tabulate how the optimal system parameters $(Q^*, S^*, \triangle^*, C^*)$ changes with $A(C)$ and fleet size K jointly, where we have a capacity option set $\overline{C} = \{2, 4, 8, 16, 32\}$, $N = 4$, $\lambda_0 = 4$, $D = 8$, $L_w = 2$ and $\beta_i = 32$ for all i.

Table 3.1: The Effects of the Change in $A(C)$ and K on Total Cost Rate

A(C)=	0.25C	C	4C
K	(Q*,S*, △*, C*)	(Q*,S*, △*, C*)	(Q*,S*, △*, C*)
2	(20,15,0,32)	(21,15,0,32)	(21,15,0,32)
3	(13,11,1,16)	(14,11,1,16)	(16,14,0,16)
4	(11,10,1,16)	(11,10,1,16)	(11,10,1,16)
9	(5,9,2,8)	(8,10,1,8)	(8,10,1,8)
14	(4,9,2,4)	(4,9,2,4)	(4,9,2,4)

Parallel to our observations in Section 2.6, we note that, truck utilization $Q^*/C*\%$ has a non-increasing structure when the fleet size K increases for a given capacity C. Also, the order up-to level S^* decreases with K and increases with $A(C)$. Moreover, $Q^*/C*\%$ increases as the order setup cost rate $A(C)/C$ increases. The reasons of these behaviors are similar to those that are discussed in the single echelon case. Note that with a higher fleet size, trucks with a smaller capacity can be used for the transportation of the goods. Hence, as K increases, Q^* decreases until the increase in the warehouse order setup cost rate $\frac{K_{wh}\lambda_0}{Q^*}$ dominates the savings from systemwide holding and backorder costs. Recall that in the single echelon systems, using the smallest C (which is 2 in our numerical test bed) with a 100% truck utilization is the best choice when we ignore the $\phi(C)$ cost factor. However, this may not hold true for the two-echelon case due to the warehouse order set-up cost. Figure 3.6 depicts the behavior of $AC^* = AC(Q, \mathbf{S}^*(Q))$ with regards to Q when $D = 8$, $N = 4$, $A(C) = 0.25C$ and provided that there is an ample number of trucks in the truck base. Notice that AC^* is minimum when $Q = 8$.

Figure 3.6: Illustration of the change of $AC^*\%$ with Q

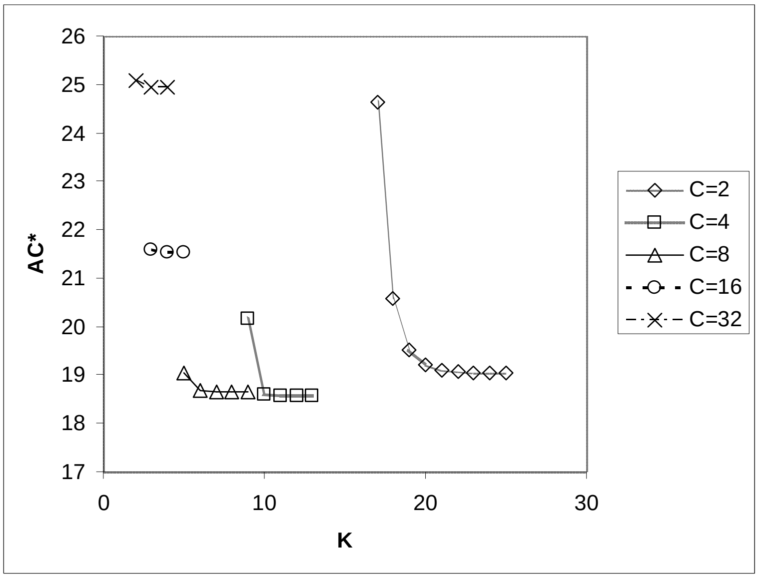


Next, we consider how the warehouse is affected by Q. From the numerical results we observe that the response of the warehouse to bigger Q is to decrease its order up-to level S_{wh}. For instance, consider the case when $K = 3$ in Table 3.1. For $A(C) = 0.25C$, optimal joint order quantity is $Q^* = 13$ with $\triangle^* = 1$, for $A(C) = C$, Q^* is raised to 14 with the same $\triangle^*$. However Q^* becomes 16 when $A(C) = 4C$. In this case, $\triangle^* = 0$. This decrease in $\triangle$ is due to the fact that the savings from the warehouse holding costs dominate the extra costs the higher effective lead-time brings when $\triangle^* = 0$ for $Q = 16$.

Joint Effect of C and K

Next, we consider the case when C is exogenous to the system. Figure 3.7

Figure 3.7: Illustration of the change of AC^* with K for different C



illustrates the impact of the capacity constraint C on AC^* when $\lambda_0 = 4$. In Figure 3.7, AC^* is minimum when $C = 4$ with full truck utilization ($Q = 4$), provided that there are abundant number of trucks. We observe that $\triangle^*$ decreases as C increases, although $S^*_{wh} = Q\triangle^*$ remain same for $C = 2, 4$ and 8. Since the batch demands of size Q arrive at the warehouse, after some Q, it begins to be unprofitable to hold stock in the warehouse. Hence, we can explain why the warehouse serves as a cross-dock point when $C = 16$ and 32. The other impacts of C on AC^* and on other system parameters are similar to those that are mentioned in Section 2.6.

The Impacts of the Change in β_i and K

Next, we present Table 3.2, which tabulates the impacts of backorder cost rate β_i and fleet size K jointly on the optimal system parameters where we have a capacity option set $\overline{C} = \{2, 4, 8, 16, 32\}$, $N = 4$,$\lambda_0 = 2$, $A(C) = C$ and $D = 8$.

Table 3.2: The Effects of the Change in β_i and K on Total Cost Rate

	β_i=2	β_i=4	β_i=16	β_i=32
K	(Q*,S*,△*,C*)	(Q*,S*,△*,C*)	(Q*,S*,△*,C*)	(Q*,S*,△*,C*)
1	(22,7,0,32)	(21,8,0,32)	(21,10,0,32)	(21,11,0,32)
5	(4,3,1,4)	(4,4,1,4)	(4,6,1,4)	(6,6,1,4)
9	(4,3,1,4)	(4,4,1,4)	(4,5,1,4)	(4,6,1,4)
11	(4,3,1,4)	(4,4,1,4)	(2,5,2,2)	(2,5,3,2)
15	(4,3,1,4)	(4,4,1,4)	(2,5,2,2)	(2,5,3,2)

Similar to the single-echelon case, the retailers react to higher β_i with higher order up-to levels S^* and lower truck utilization percentage, $Q^*/C*\%$. Also, we observe that smaller Q^* leads to higher $\triangle^*$ in Table 3.2. As we have a higher β_i, retailer holding and backorder costs gain more importance and the warehouse order setup cost rate ,$\frac{K_{wh}\lambda_0}{Q}$ becomes dominant over the other cost rates for the first time when Q^* is smaller. For instance in Table 3.2, $Q = 4$ with a fleet size of $K = 15$ gives the minimum cost if $\beta_i = 2$. However, $Q = 2$ with the same fleet size brings about the minimum AC^* when $\beta_i = 32$.

The Impact of D on the Performance of the System

Next we present Table 3.3, which tabulates how D affects AC^* and other system parameters when $C = 8$, $\lambda_0 = 8$, $A(C) = C$, $\beta_i = 4$ and $N = 4$. Note that the holding costs during the lead time are not incurred in Table 3.3. Also, it is obvious that as D increases, number of the arrivals during service time increases as well. Hence, the number of servers must be augmented in order to cope up with the increased service time. Notice that K values for $D = 1$ and $D = 8$ are not same in the table, however AC^*, Q^*, S^* and $\triangle^*$ do not change for $K > 3$ where $D = 1$. We observe that an increase in D leads to higher retailer order up-to levels S^* from the table. The reasons of higher S^* for $D = 8$ can be explained by the reasons that are discussed in the single echelon case. Also, we observe that the warehouse order up-to level S^*_{wh} is rather invariant to a change in D compared to other system parameters. For instance in Table 3.3, $\triangle^* = 1$ for both $D = 1$ and $D = 8$.

Table 3.3: The Effects of the Change in D and K on Total Cost Rate and Other System Parameters

D=1				
K	AC*	(Q*,S*)	Q*/C%	△*
1	13.4927	(8,3)	100%	1
2	13.4599	(8,3)	100%	1
3	13.4599	(8,3)	100%	1

D=8				
K	AC*	(Q*,S*)	Q*/C%	△*
5	19.0293	(8,7)	100%	1
6	18.6708	(8,7)	100%	1
7	18.6621	(8,7)	100%	1
8	18.6621	(8,7)	100%	1

The effects of λ_0

Another system parameter which increases the number of the arrivals during service time D is the total demand arrival rate λ_0, hence the number of the servers must be augmented in order to handle an increase in λ_0, too. AC^*, Q^*, S^* and $\triangle^*$ do not change for $K > 9$ where $\lambda_0 = 4$. Table 3.4 tabulates how λ_0 affects AC^* and other system parameters when $C = 8$, $D = 8$, $A(C) = C$, $\beta_i = 4$ and $N = 4$. Similar to the single echelon case, we observe that higher λ_0 leads to a higher S^*. As we can observe from the table, $\triangle^*$ is more sensitive to an increase in λ_0 compared to D.

Table 3.4: The Effects of λ_0 on AC^* and Other System Parameters

λ_0=4					λ_0=16				
K	AC*	(Q*,S*)	Q*/C%	△*	K	AC*	(Q*,S*)	Q*/C%	△*
5	19.0293	(8,7)	100%	1	17	46.8618	(8,22)	100%	4
6	18.6708	(8,7)	100%	1	18	44.0647	(8,21)	100%	4
7	18.6621	(8,7)	100%	1	19	43.7391	(8,21)	100%	4
8	18.6621	(8,7)	100%	1	20	43.7009	(8,21)	100%	4
9	18.6621	(8,7)	100%	1	21	43.6976	(8,21)	100%	4

The effects of N and K_{wh}

The retailer backorder and holding costs increase as N increases due to the fact that the system enjoys the *risk pooling* effect when N is smaller. From the numerical results, we also remark that $\triangle^*$ is rather insensitive to the changes in N, compared to other parameters. Note that the warehouse holding cost h_w affects $\triangle^*$, and as h_w increases, $\triangle^*$ decreases. This leads to an increase in effective lead-times, which brings about higher S^*. Also, the effects of K_{wh} are very similar to those of $A(C)/C$. An increase in K_{wh} leads to an increase in Q^*. In addition, when we have a truck capacity option set $\overline{C}$, we observe that Q^*, (or C^*) that gives the minimum cost AC^* increases with K_{wh}.

The effects of the warehouse lead-time L_w

Table 3.5 tabulates how AC^* and other parameters change with L_w when $C = 8$, $\lambda_0 = 4$, $\beta_i = 4$ and $D = 8$.

Table 3.5: The Impacts of L_{wh} on AC^* and Other System Parameters

$L_w = 1$					$L_w = 4$				
K	AC*	(Q*,S*)	Q*/C%	$\triangle^*$	K	AC*	(Q*,S*)	Q*/C%	$\triangle^*$
5	19.1202	(8,8)	100%	0	5	19.5415	(8,7)	100%	2
6	18.8040	(8,8)	100%	0	6	19.3039	(8,7)	100%	2
7	18.7964	(8,8)	100%	0	7	19.3007	(8,7)	100%	2
8	18.7963	(8,8)	100%	0	8	19.3007	(8,7)	100%	2

From the Table 3.5, we can observe that the warehouse reacts to bigger L_w with higher $\triangle^*$. Since the effective lead-times increase as L_w increases, our expectation of the response of the retailers to a higher L_w would be higher S^*. However, the behavior of S^* contradicts with our expectation in some of the parameter settings. For instance, when $L_w = 1$, which is considerably smaller than $D = 8$, the warehouse employs a cross-docking policy $\triangle^* = 0$, and $S^* = 8$, which is higher than the S^* value when $L_w = 4$. This can be explained by the dominance of the savings from warehouse holding costs on the effects of the higher effective lead-time when $L_w = 1$ and $\triangle^* = 0$. However the complicated structure of the AC^* hinders us to make generalizations on the behaviors of AC^*. In the next subsection, we discuss how robust our approximation is, which is given in Subsection 3.2.2.

3.4.2 Accuracy of the Approximation

In this subsection, we examine the accuracy of the approximation that is made in Subsection 3.2.2. Recall that we approximate the departure process of the

$(S_{wh}-1, S_{wh})$ inventory system by a renewal process with Erlang(Q', λ'_0) distributed inter-arrival times. In order to examine the performance of the approximation on the operating characteristics, we simulated this distribution system to obtain the true operating characteristics.

In our simulations, we used a run length of $100,000$ warehouse ordering instances and 20 replications to obtain the corresponding operating characteristics. The distribution system is simulated for more than 250 different scenarios and AC_{sim} obtained from the simulation are compared with AC_{app} that we obtain from Equation (3.8). Since the simulation of the system takes a long time, we do not optimize the parameters.

Table 3.6: The Accuracy of the Approximation for different ρ

ρ	$\%\mu_{\lvert err \rvert}$	min % err	median % err	max % err
$\rho \leq 0.5$	0.030	-0.059	0.009	0.077
$0.5 < \rho \leq 0.6$	0.035	-0.005	0.030	0.089
$0.6 < \rho \leq 0.7$	0.037	-0.100	0.008	0.121
$0.7 < \rho \leq 0.8$	0.101	-0.012	0.042	0.469
$0.8 < \rho \leq 0.9$	0.269	-0.200	0.013	1.817
$0.9 < \rho \leq 0.95$	0.938	-0.930	0.048	10.775
$0.95 < \rho \leq 0.98$	2.500	-2.211	2.057	7.890
$0.98 < \rho$	10.853	3.532	9.985	18.899

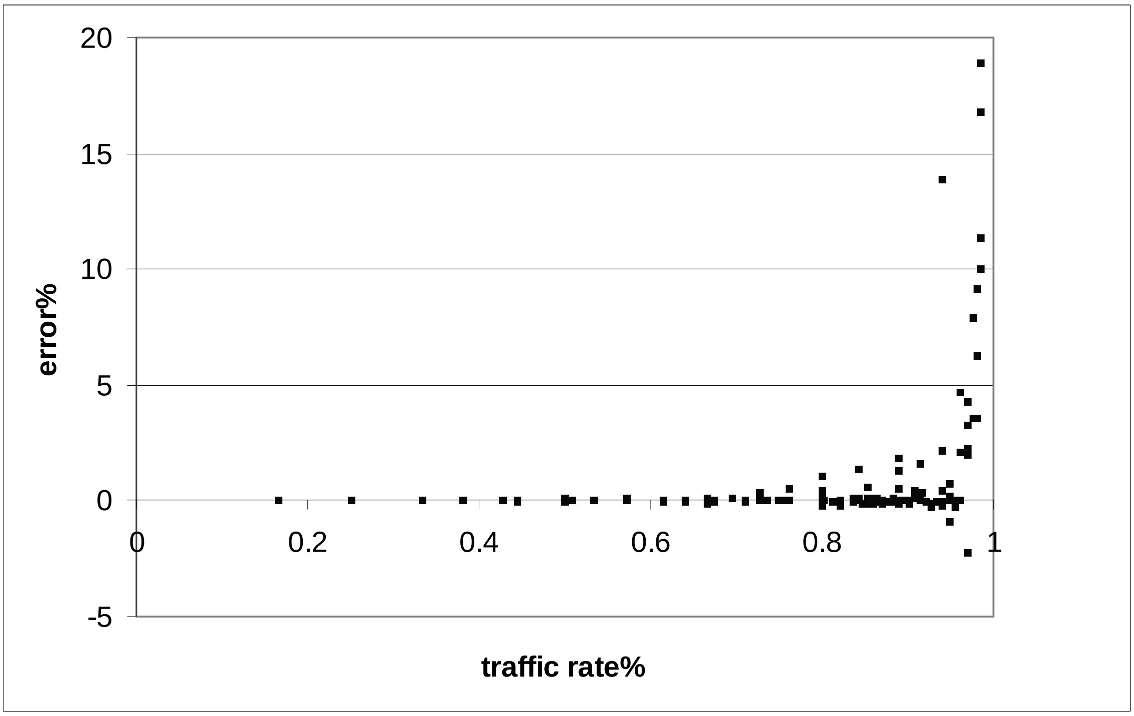
From the simulation results, we observe that the accuracy of the approximation is strongly correlated with the traffic rate ρ of the queue. When $\rho < 0.95$, we observe that the approximation works very well. Let

$$\%err = 100 \times \frac{AC_{sim.} - AC_{app.}}{AC_{sim.}}$$

denote the relative error of the approximation with regards to the AC obtained from the simulation. Note that $\%err$ can take both positive and negative values, that's why we take the average $\%\mu_{|err|}$ of the absolute values of the $\%err$. In Table 3.6, we present how the sample mean of the absolute value of $\%err$, $\%\mu_{|err|}$ and the minimum, median and maximum values of $\%err$ change with ρ.

Notice that our approximation works better for lower ρ. Even when ρ is as high as 0.98, our approximation works reasonably well. Figure 3.8 illustrates how the simulation results deviates from the approximated ones as ρ approaches to 1. The approximation method does not perform well for $\rho > 0.95$, hence we recommend to use the simulation methods to obtain the total cost rate of the system and optimum parameters when the traffic rate is higher than 0.95.

Figure 3.8: Illustration of the change of the difference between simulated AC and approximated AC with ρ



Chapter 4

Conclusion and Future Studies

In this study, we have examined the effect of the transportation capacity on the performance of the joint replenishment policies in single and two echelon inventory systems. Particularly we have focused on quantity based (Q, **S**) replenishment policies. Different from the previous work in the literature, we have considered the fleet size limitation in our analysis. We model the shipment operations of the joint orders at the truck base as a queueing system, and we employed the Axsatër's [3] method to derive the unit holding and backorder cost rate per unit time. We constructed the expected cost rate expression for unit Poisson demands. Consequently, weinvestigated the characteristics of the expected cost rate of the system with other system parameters, and we discussed methods of approach for different kinds of transportation limitation scenarios in the existence of a maintenance/depreciation cost rate factor per a truck. An extensive numerical study has been conducted for the single echelon case in order to assess the sensitivity of the expected cost rate to various system parameters. Afterwards, the analysis is extended to a two-echelon system. In this setting, the fleet is used for the transportation of the orders between the warehouse and the retailers. A joint order placed by the retailers must be first satisfied by the warehouse inventory before being loaded onto the trucks. Therefore the departure process of the warehouse inventory system, which uses (S-1,S) policy constitutes the arrival process of the queueing system where the trucks are the servers and the satisfied joint orders are the customers. Afterwards, we characterize the departure process of an inventory system with deterministic lead times which uses (S-1,S) policy, and where the arrivals occur according to a renewal process. Consequently we apply these results to our setting, and we obtain the arrival process of the queueing system. Then we approximated this arrival process to an Erlang Process and derived the expected cost rate of the system accordingly. A numerical study has been conducted for this two-echelon system, and the sensitivity of the system parameters and the accuracy of our approximation method is assessed.

We observe that the management of the transportation capacity and the fleet size in conjunction with the joint replenishment policies can lead to substantial savings. We believe that our study may have important applications for supply chain design. In addition our study can be extended to contractual design agreements especially for the 3PL provider firms which supply logistic service to retailer chains.

Bibliography

[1] Y. Aksoy and S. Erenguc. Multi-item models with coordinated replenishments: A survey. *International Journal of Production Management*, 8; 63–73, 1988.

[2] D. R. Atkins and P. O. Iyogun. Periodic versus can-order policies for coordinated multi-item inventory systems. *Management Science*, 34; 791–796 1988.

[3] S. Axsäter. Simple solution procedures for a class of two-echelon inventory systems. *Operations Research*, 38(1); 64–69 1990.

[4] S. Axsäter. Continuous review policies for multi-level inventory systems with stochastic demand. In S. C. Graves, A. R. Kan, and P. Zipkin, editors, *Logistics of Production and Inventory*, volume 4 of *Handbooks in OR & MS*. Elsevier, Amsterdam, 1993.

[5] S. Axsäter and W. F. Zhang. A joint replenishment policy for multi-echelon inventory control. *International Journal of Production Economics*, 59; 243–250 1999.

[6] J. L. Balintfy. On a basic class on inventory problems. *Management Science*, 10 ;287–297 1964.

[7] M. O. Ball, B. L. Golden, A. A. Assad and L. D. Bodin. Planning for truck fleet size in the presence of a common carrier option. *Decision Sciences*, 14 ;103–120 1983.

[8] J. Benjamin. An analysis of inventory and transportation costs in a constrained network. *Transportation Science*, 23(3); 177–183 1989.

[9] N. Benkheder and C. A. Yao. The multi-item joint replenishment problem with transportation and container effects. *Transportation Science*, 28(1); 37–57 1994.

[10] G. Cachon. Managing a retailer's shelf space, inventory and transportation. *Manufacturing and Service Operations Management*, 3; 211–229 2001.

[11] S. Çetinkaya and C. Y. Lee. Stock replenishment and shipment scheduling for vendor managed inventory systems. *Management Science*, 16; 217–232 2000.

[12] S. Çetinkaya and C. Y. Lee. Optimal outbound dispatch policies: modelling inventory and cargo capacity. *Naval Research Logistics*, 49; 531–556 2002.

[13] K. L. Cheung and H. Lee. The inventory benefit of shipment coordination and stock rebalancing in a supply chain. *Management Science*, 48; 300–306 2002.

[14] M. Dror and M. Ball. Inventory/routing: reduction from an annual to a short period problem. *Naval Research Logistics*, 34; 891–905 1987.

[15] A. Federgruen. Centralized planning models for multi-echelon inventory systems under uncertainty. vol. 4. North Holland, Amsterdam, The Netherlands; 175–197 1993.

[16] A. Federgruen, H. Greoenevelt and H. Tijms. Coordinated replenishments in a multi-item inventory system with compound poisson demands and constant lead times. *Management Science*, 30; 344–357 1984.

[17] A. Federgruen and P. Zipkin. A combined vehicle routing and inventory allocation problem. *Operations Research*, 32; 1019–1037 1984.

[18] R. Ernst and F. D. Pyke. Optimal base stock policies and truck capacity in a two echelon system. *Naval Research Logistics*, 40; 879–903 1993.

[19] G. J. Franx. A simple solution for the M/D/c waiting time distribution. *Operations Research Letters*, 29(5); 221–229 2001.

[20] S. H. Goyal and A. T. Satir. Joint replenishment inventory control. *European Journal of Operations Research*, 38; 2–13 1989.

[21] M. C. Gürbüz, K. Moinzadeh and Y. P. Zhou. Replenishment and allocation policies for supply chains with crossdocking. *Working Paper*, 2004.

[22] M. Henig, Y. Gerchak, R. Ernst and D. F. Pyke. An inventory model embedded in designing a supply contract. *Management Science*, 43(2); 184–189 1997.

[23] E. Ignall. Optimal continuous review policies for two product inventory systems with joint set-up costs. *Management Science*, 15; 278–283 1969.

[24] G. P. Kiesmüller and A. G. de Kok. Safety stock allocation in a multi-item, multi-echelon inventory system with quantity based order consolidation. *Working Paper*, 2005.

[25] D. Simichi-Levi, P. Kaminsky and E. Simichi-Levi. Designing & Managing the Supply Chain. *McGraw–Hill*, 2003.

[26] C. Nielsen and Larsen C. An analytical study of the Q(s,S) policy applied to the joint replenishment problem. *European Journal of Operational Research*, 163; 721–732 2005.

[27] B. Y. Özkaya. The stochastic joint replenishment problem: a new policy and analysis for single and two echelon inventory systems. *PhD Thesis*, Bilkent University, 2005.

[28] B. Y. Özkaya, Ü. Gürler and E. Berk. The stochastic joint replenishment problem: a new policy, analysis, and insights. *Naval Research Logistics*, 53; 525–546 2006.

[29] P. Pantumsinchai. A comparison of three joint ordering policies. *Decision Sciences*, 23; 111–127 1992.

[30] B. Renberg and R. Planche. Un modle pour la gestion simultane des n articles d'un stock. *Revue Francaise d'Informatique et de Recherche Oprationelle*, 6; 47–59 1967.

[31] A. J. Rolfe. A note on marginal allocation in multiple service systems. *Management System*, 17(9); 656–658 1971.

[32] S. Ross. Introduction to Probability Models. *Academic Press Inc.*, 1997.

[33] H. Schultz and S. G. Johansen. Can-order policies for coordinated inventory replenishment with erlang distributed times between ordering. *European Journal of Operations Research*, 113; 30–41 1999.

[34] E. A. Silver. Some characteristics of a special joint-order inventory model. *Operations Research*, 13; 319–322 1965.

[35] E. A. Silver. Three ways of obtaining the average cost expressions in a problem related to joint replenishment inventory control. *Naval Research Logistics Quarterly*, 20; 241–254 1973.

[36] E. A. Silver. A control system for coordinated inventory replenishment. *International Journal of Production Research*, 12; 647–671 1974.

[37] E. A. Silver. Establishing reorder points in the (S,c,s) coordinated control system under compound poisson demand. *International Journal of Production Research*, 19; 743–750 1981.

[38] S. Sindhuchao, H. E. Romeijn, E. AkÇali and R. Boondiskulchok. An integrated inventory routing system for multi-item joint replenishment with limited vehicle capacity. *International Journal of Production Research*, 19; 743–750 1981.

[39] R. M. Thompstone and Silver E. A. A coordinated inventory control system under compound poisson demand and zero lead time. *Journal of Global Optimization*, 32; 93–118 2005.

[40] H. C. Tijms. Stochastic Models: An algorithmic approach. *John Wiley & Sons*, 1994.

[41] A. Toptal, S. Çetinkaya and C. Y. Lee. The buyer-vendor coordination problem: modeling inbound and outbound cargo capacity and costs. *IIE Transactions*, 35; 987–1002 2003.

[42] M. J. G. Van Eijs. On the determination of the control parameters of the optimal can-order policy. *Zeitschrift of Operations Research*, 39; 289–304 1994.

[43] C. A. Yano and Y. Gerchak Transportation contracts and safety stocks for just-in-time deliveries. *Journal of Manufacturing and Operations Management*, 2; 314–330 1989.

[44] Y. S. Zheng. Optimal conrol policy for stochastic inventory system with markovian discount opportunities. *Operations Research*, 42; 721–738 1994.

Appendix A

Algorithms Part

A.1 Accuracy Check Algorithm

The accuracy check is based on alternative expressions for the mean queue length, ($E[L_q]$). The pseudocode for the search algorithm is presented below:

begin

Set $M_{min} = kQ$

Set $E[L_q](0)$*=0*

Set ϵ *(a small enough value, which is our error bound)*

Set $\triangle$*(a big enough value)*

Set i=1

while($\triangle > \epsilon$) ***do***

Set $M = M_{min} + i \times \lceil 20 + (kQ/10) \rceil$

Compute p_j*'s for* $j \in 0, 1, ...M$

Compute $E[L_q](i)$

$\triangle = |E[L_q](i) - E[L_q](i-1)|$

i++

end

end

Here $E[L_q(i)]$ denotes the expected queue length when $M = M_{min} + i \times kQ/10$.

One can derive $E[L_q]$ in two ways:
1. By choosing a large enough value (Θ), $E[L_q] \approx \Sigma_{j=c+1}^{\Theta}(j-c)p_j$
2. A formula for $E[L_q]$ is suggested in *Tijms,ref. no,p.290* as below:

$$E[L_q] = \frac{1}{2c(1-\rho)}\left[(c\rho)^2 - c(c-1) + \sum_{j=0}^{c-1}\left\{c(c-1) - j(j-1)\right\}p_j\right] \tag{A.1}$$

where $\rho = \lambda_0 D/c$ is the traffic intensity, and $\rho < 1$, otherwise queue would blow up. Both of these expressions can be used for computing $E[L_q(i)]$.

A.2 Search Algorithms for Different Types of Transportation Limitation Scenarios

We define $K_{max}(Q,\epsilon)$, which is frequently used in our algorithms, for an arbitrarily small ϵ as below:

$$\begin{aligned}K_{max}(Q,\epsilon) = min\{K \geq K_{min}(Q) : AC(Q,(\mathbf{S}^*(Q,K)),K,C) - \\ AC(Q,(\mathbf{S}^*(Q,K+1)),K+1,C) \leq \epsilon\}\end{aligned}$$

where K is a positive integer.

Search Algorithm 1: Search Algorithm for Single Truck Capacity

Step 1. Find $\overline{Q}_C = (\lceil C/2\rceil, \lceil C/2\rceil + 1, ..., C)$.
Step 2. Find $K_{min}(Q)$ for every $Q \in \overline{Q}_C$.
Step 3. Find $K_{max}(Q,\phi(C))$.
Step 4. The recommended order quantity, Q^*, is equal to
$\arg\min_{Q\in\overline{Q}_C} AC(Q,(\mathbf{S}^*(Q,K_{max}(Q,\phi(C)))),K_{max}(Q,\phi(C)),C) + K_{max}(Q,\phi(C)) \times \phi(C)$
The recommended fleet size and order-up to levels are $K_{max}(Q^*,\phi(C))$ and $\mathbf{S}^*(Q^*,K_{max}(Q^*,\phi(C))$, respectively.

Search Algorithm 2: Search Algorithm for Multiple Truck Capacities

Step 1. Find $\overline{Q}_{C_i} = (\lceil C_i/2 \rceil, \lceil C_i/2 \rceil + 1, ..., C_i)$ for every $C_i \in \overline{C}$.
Step 2. If $\overline{Q}_{C_i} \cap \overline{Q}_{C_j}$ is not an empty set for every $(i, j) \in (1, 2, ..., m)$ such that $i < j$, eliminate the intersecting elements from $\overline{Q}_{C_j}$.
Step 3. For every $C_i \in \overline{C}$, repeat the following steps (from Step 3.1 to Step 3.3):
Step 3.1. Find $K_{min}(Q)$ for every $Q \in \overline{Q}_{C_i}$.
Step 3.2 Find $K_{max}(Q, \phi(C_i))$ for every $Q \in \overline{Q}_{C_i}$.
Step 3.3 The recommended order quantity for a given capacity C_i, is Q_i^* such that:
$Q_i^* = \arg \min_{Q \in \overline{Q}_{C_i}} AC(Q, (\mathbf{S}^*(Q, K_{max}(Q, \phi(C_i)))), K_{max}(Q, \phi(C_i)), C_i) + K_{max}(Q, \phi(C_i)) \times \phi(C_i)$
For a given C_i, the recommended fleet size and optimal order-up to levels are $K_{max}(Q_i^*, \phi(C_i))$ and $\mathbf{S}^*(Q_i^*, K_{max}(Q_i^*))$ respectively.
Step 4. Decide upon the best capacity option C^* such that
$C^* = \arg \min_{C_i \in \overline{C}} AC(Q_i^*, (\mathbf{S}^*(Q_i^*, K_{max}(Q_i^*, \phi(C_i)))), K_{max}(Q_i^*, \phi(C_i)), C_i) + K_{max}(Q_i^*, \phi(C_i)) \times \phi(C_i)$

Search Algorithm 3: Search Algorithm for a Given Fleet Size (There is a Capacity Option Set)

Step 1. Find $Q_{min}(K)$ for K.
Step 2. Eliminate all C_i such that $C_i < Q_{min}(K)$ from $\overline{C}$.
Now we have a set $\overline{C} = (C_k, C_{k+1}, ..., C_m)$ and C_k is the smallest capacity such that $C_k \geq Q_{min}(K)$
Step 3. Find $\overline{Q}_{C_i} = (\lceil C_i/2 \rceil, \lceil C_i/2 \rceil + 1, ..., C_i)$ for every $C_i \in \overline{C}$, .
Step 4. If $\overline{Q}_{C_i} \cap \overline{Q}_{C_j}$ is not an empty set for every $(i, j) \in (k, k+1, ..., m)$ such that $i < j$, eliminate the intersecting elements from $\overline{Q}_{C_j}$.
Step 5 Find the recommended order quantity (denoted as Q_i^*) for every $C_i \in \overline{C}$ for a given K is as follows:
$Q_i^* = \arg \min_{Q \in \overline{Q}_{C_i}} AC(Q, (\mathbf{S}^*(Q, K)), K, C_i) + K(Q) \times \phi(C_i)$
For a given capacity C_i and fleet size K, the optimal order-up to levels $(\mathbf{S}^*(Q_i^*, K))$

can be found easily.

Step 6. Decide upon the capacity option C^* such that

$C^* = \arg\min_{C_i \in \overline{C}} AC(Q_i^*, (\mathbf{S}^*(Q_i^*, K)), K, C_i) + K \times \phi(C_i)$

Appendix B

Proof Part

Proof of Theorem 4

Recall that we have the following expression for Y_i.

$$Y_i = \begin{cases} X_i & if \quad (L - Z_{i-S}) \leq min(X_{i-S}, X_i) \\ X_{i-S} + Z_{i-S} + X_i - L & if \quad X_{i-S} < (L - Z_{i-S}) \leq X_i \\ L - Z_{i-S} & if \quad X_i < (L - Z_{i-S}) \leq X_{i-S} \\ X_{i-S} & if \quad (L - Z_{i-S}) > max(X_{i-S}, X_i). \end{cases} \tag{B.1}$$

As demonstrated above, Y_i takes different values for different X_{i-S}, X_i and Z_{i-S} realizations. Next we define the disjoint events E_1, E_2, E_3 and E_4 as follows:

$E_1 = \{(L - Z_{i-S}) \leq min(X_{i-S}, X_i)\}$

$E_2 = \{X_{i-S} < (L - Z_{i-S}) \leq X_i\}$

$E_3 = \{X_i < (L - Z_{i-S}) \leq X_{i-S}\}$

$E_4 = \{(L - Z_{i-S}) > max(X_{i-S}, X_i)\}$.

Since the events that are defined above are disjoint, the probability distribution function of Y_i can be calculated as follows:

$F_{Y_i}(y) = P(Y_i \leq y) = \sum_{j=1}^{4} P(Y_i \leq y, E_j)$

Accordingly, we derive $P(Y_i \leq y, E_j)$ for all j in order to derive the $F_{Y_i}(y)$.

1. $P(Y_i \leq y, E_1) =$

$P(X_i \leq y, L - Z_{i-S} \leq min(X_{i-S}, X_i))$

$= P(X_i \leq y, L - Z_{i-S} \leq X_i \leq X_{i-S}) + P(X_i \leq y, L - Z_{i-S} \leq X_{i-S} < X_i)$

$= P(X_i \leq y, L - Z_{i-S} \leq X_i, X_i \leq X_{i-S})$
$+P(X_i \leq y, L - Z_{i-S} \leq X_{i-S}, X_{i-S} < X_i)$

$P(X_i \leq y, L - Z_{i-S} \leq X_i, X_{i-S} \leq X_i)$
$= \int_{x_2=0}^{y} \int_{z=L-x_2}^{\infty} \int_{x_1=x_2}^{\infty} dF_{X_{i-S}}(x_1) dF_{X_i}(x_2) dF_{Z_{i-S}}(z)$
$= \int_{x_2=0}^{y} \overline{F_X}(x_2) \overline{F_Z}(L - x_2) dF_X(x_2)$

$P(X_i \leq y, L - Z_{i-S} \leq X_{i-S}, X_{i-S} < X_i)$
$= \int_{x_2=0}^{y} \int_{x_1=0}^{x_2} \int_{z=L-x_1}^{\infty} dF_{X_i}(x_2) dF_{X_{i-S}}(x_1) dF_{Z_{i-S}}(z)$
$= \int_{x_2=0}^{y} \int_{x_1=0}^{x_2} \overline{F_Z}(L - x_1) dF_X(x_1) dF_X(x_2)$

$$P(Y_i \leq y, E_1) = \int_{x_2=0}^{y} \left\{ \overline{F_X}(x_2) \overline{F_Z}(L - x_2) + \int_{x_1=0}^{x_2} \overline{F_Z}(L - x_1) f_X(x_1) dx_1 \right\} dF_X(x_2)$$

2. $P(Y_i \leq y, E_2) =$
$P(X_{i-S} + Z_{i-S} + X_i - L \leq y, X_{i-S} < L - Z_{i-S} \leq X_i)$
$= P(Z_{i-S} \leq y + L - (X_i + X_{i-S}), X_{i-S} < L - Z_{i-S} \leq X_i)$

$= P(Z_{i-S} \leq y + L - (X_i + X_{i-S}), Z_{i-S} < L - X_{i-S}, Z_{i-S} \geq L - X_i,$
$X_i > X_{i-S}, X_{i-S} \leq y)$

$$= \int_{x_2=0}^{\infty} \int_{x_1=0}^{min(x_2,y)} \int_{z=L-x_2}^{min(L-x_1, L+y-(x_1+x_2))} dF_{Z_{i-S}}(z) dF_{X_{i-S}}(x_1) dF_{X_i}(x_2)$$

$$= P(Y_i \leq y, E_2) = \int_{x_2=0}^{y} \int_{x_1=0}^{x_2} \int_{z=L-x_2}^{L-x_1} dF_Z(z) dF_X(x_1) dF_X(x_2)$$
$$+ \int_{x_2=y}^{\infty} \int_{x_1=0}^{y} \int_{z=L-x_2}^{L-x_1+y-x_2} dF_Z(z) dF_X(x_1) dF_X(x_2)$$

3. $P(Y_i \leq y, E_3) =$
$P(L - Z_{i-S} \leq y, X_i < L - Z_{i-S} \leq X_{i-S})$
$= P(Z \geq L - y, Z < L - X_i, Z \geq L - X_{i-S}, X_i < y)$

$$= \int_{x_1=0}^{\infty} \int_{x_2=0}^{min(x_1,y)} \int_{z=max(y,L-x_1)}^{L-x_2} dF_{Z_{i-S}}(z) dF_{X_i}(x_2) dF_{X_{i-S}}(x_1)$$

$$= \int_{x_1=0}^{y} \int_{x_2=0}^{x_1} \int_{z=L-x_1}^{L-x_2} dF_Z(z) dF_X(x_2) dF_X(x_1)$$
$$+ \int_{x_1=y}^{\infty} \int_{x_2=0}^{y} \int_{z=L-y}^{L-x_2} dF_Z(z) dF_X(x_2) dF_X(x_1)$$

4. $P(Y_i \leq y, E_4) =$
$P(X_{i-S} \leq y, L - Z_{i-S} > max(X_{i-S}, X_i))$
$= P(X_{i-S} \leq y, L - Z_{i-S} > X_{i-S} \geq X_i) + P(X_{i-S} \leq y, L - Z_{i-S} > X_i > X_{i-S})$

$= P(X_{i-S} \leq y, Z_{i-S} < L - X_{i-S}, X_i \leq X_{i-S})$
$+P(X_{i-S} \leq y, L - X_i > Z_{i-S}, X_{i-S} < X_i)$

$P(X_{i-S} \leq y, Z_{i-S} < L - X_{i-S}, X_i \leq X_{i-S})$
$$= \int_{x_1=0}^{y} \int_{z=0}^{L-x_1} \int_{x_2=0}^{x_1} dF_{X_i}(x_2) dF_{Z_{i-S}}(z) dF_{X_{i-S}}(x_1)$$
$$= \int_{x_1=0}^{y} F_X(x_1) F_Z(L - x_1) dF_X(x_1)$$

$P(X_{i-S} \leq y, L - X_i > Z_{i-S}, X_{i-S} < X_i)$
$$= \int_{x_1=0}^{y} \int_{x_2=x_1}^{\infty} \int_{z=0}^{L-x_2} dF_{Z_{i-S}}(z) dF_{X_i}(x_2) dF_{X_{i-S}}(x_1)$$
$$= \int_{x_1=0}^{y} \int_{x_2=x_1}^{\infty} F_Z(L - x_2) dF_X(x_2) dF_X(x_1)$$

$$P(Y_i \leq y, E_4) = \int_{x_1=0}^{y} \left\{ F_X(x_1) F_Z(L - x_1) + \int_{x_2=x_1}^{\infty} F_Z(L - x_2) dF_X(x_2) \right\} dF_X(x_1)$$

Hence we can obtain $F_{Y_i}(y) = P(Y_i \leq y) = \sum_{j=1}^{4} P(Y_i \leq y, E_j)$. Since the inter-arrival times are i.i.d, the distribution function, $F_Y(y)$ of each Y_i are identical. The density of the Y, denoted as $f_Y(y) = \frac{dF_Y(y)}{dy}$ can be obtained by taking the derivation of $F_Y(y)$ with respect to y. Leibnitz's Rule for differentiating the integrals is used when $f_Y(y)$ is derived.

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \sum_{j=1}^{4} \frac{d(P(Y \leq y, E_j))}{dy}$$

1. $\frac{d(P(Y \leq y, E_1))}{dy} = \dfrac{d\left(\int_{x_2=0}^{y} \left\{ \overline{F_X}(x_2) \overline{F_Z}(L - x_2) + \int_{x_1=0}^{x_2} \overline{F_Z}(L - x_1) dF_X(x_1) \right\} dF_X(x_2)\right)}{dy}$

$$= f_X(y)\left\{\overline{F}_Z(L-y)\overline{F}_X(y) + \int_{x=0}^{y}\overline{F}_Z(L-x)dF_X(x)\right\}$$

2. $$\frac{d(P(Y\le y,E_2))}{dy} = \frac{d\left(\int_{x_2=0}^{y}\int_{x_1=0}^{x_2}\int_{z=L-x_2}^{L-x_1} dF_Z(z)dF_X(x_1)dF_X(x_2)\right)}{dy}$$

$$+\frac{d\left(\int_{x_2=y}^{\infty}\int_{x_1=0}^{y}\int_{z=L-x_2}^{L-x_1+y-x_2} dF_Z(z)dF_X(x_1)dF_X(x_2)\right)}{dy}$$

$$\frac{d\left(\int_{x_2=0}^{y}\int_{x_1=0}^{x_2}\int_{z=L-x_2}^{L-x_1} dF_Z(z)dF_X(x_1)dF_X(x_2)\right)}{dy} = f_X(y)\int_{x_1=0}^{y}\int_{z=L-y}^{L-x_1} dF_Z(z)dF_X(x_1)$$

$$\frac{d\left(\int_{x_2=y}^{\infty}\int_{x_1=0}^{y}\int_{z=L-x_2}^{L-x_1+y-x_2} dF_Z(z)dF_X(x_1)dF_X(x_2)\right)}{dy}$$

$$= -f_X(y)\int_{x_1=0}^{y}\int_{z=L-y}^{L-x_1} dF_Z(z)dF_X(x_1)$$

$$+\int_{x_2=y}^{\infty}\frac{d\left(\int_{x_1=0}^{y}\int_{z=L-x_2}^{L-x_1+y-x_2} dF_Z(z)dF_X(x_1)\right)}{dy}dF_X(x_2)$$

$$\frac{d(P(Y\le y,E_2))}{dy} = \int_{x_2=y}^{\infty}\frac{d\left(\int_{x_1=0}^{y} f_X(x_1)\left[F_Z(L-x_1+y-x_2)-F_Z(L-x_2)\right]dx_1\right)}{dy}dF_X(x_2)$$

$$= \int_{x_2=y}^{\infty}\left[\int_{x_1=0}^{y} f_Z(L-x_1+y-x_2)dF_X(x_1) + f_X(y)\left(F_Z(L-x_2)-F_Z(L-x_2)\right)\right]dF_X(x_2)$$

$$= \int_{x_2=y}^{\infty}\int_{x_1=0}^{y} f_Z(L-x_1+y-x_2)dF_X(x_1)dF_X(x_2).$$

Since $f_X(x) = 0$ for $x < 0$ and $f_Z(z) = 0$ for $z < 0$, we obtain the following:

$$\frac{d(P(Y\le y,E_2))}{dy} = \int_{x_2=y}^{L+y}\int_{x_1=0}^{min(y,L+y-x_2)} f_Z(L-x_1+y-x_2)dF_X(x_1)dF_X(x_2)$$

3. $$\frac{d(P(Y\le y,E_3))}{dy} = \frac{d\left(\int_{x_1=0}^{y}\int_{x_2=0}^{x_1}\int_{z=L-x_1}^{L-x_2} dF_Z(z)dF_X(x_2)dF_X(x_1)\right)}{dy}$$

$$+\frac{d\left(\int_{x_1=y}^{\infty}\int_{x_2=0}^{y}\int_{z=L-y}^{L-x_2} dF_Z(z)dF_X(x_2)dF_X(x_1)\right)}{dy}$$

$$\frac{d\left(\int_{x_1=0}^{y}\int_{x_2=0}^{x_1}\int_{z=L-x_1}^{L-x_2} dF_Z(z)dF_X(x_2)dF_X(x_1)\right)}{dy} = f_X(y)\int_{x_2=0}^{y}\int_{z=L-y}^{L-x_2} dF_Z(z)dF_X(x_2)$$

$$\frac{d\left(\int_{x_1=y}^{\infty}\int_{x_2=0}^{y}\int_{z=L-y}^{L-x_2} dF_Z(z)dF_X(x_2)dF_X(x_1)\right)}{dy} =$$

$$-f_X(y)\int_{x_2=0}^{y}\int_{z=L-y}^{L-x_2} dF_Z(z)dF_X(x_2)$$

$$+\int_{x_1=y}^{\infty}\frac{d\left(\int_{x_2=0}^{y}\{F_Z(L-x_2)-F_Z(L-y)\}dF_X(x_2)\right)}{dy}dF_X(x_1)$$

$$\frac{d(P(Y\leq y,E_3))}{dy} = \int_{x_1=y}^{\infty}\left[\int_{x_2=0}^{y} f_Z(L-y)dF_X(x_2) + f_X(y)(F_Z(L-y)-F_Z(L-y))\right]dF_X(x_1)$$

$$= \int_{x_1=y}^{\infty}\int_{x_2=0}^{y} f_Z(L-y)dF_X(x_2)dF_X(x_1) = \overline{F_X}(y)F_X(y)f_Z(L-y)$$

4. $$\frac{d(P(Y\leq y,E_4))}{dy} = \frac{d\left(\int_{x_1=0}^{y}\left\{F_X(x_1)F_Z(L-x_1)+\int_{x_2=x_1}^{\infty}F_Z(L-x_2)dF_X(x_2)\right\}dF_X(x_1)\right)}{dy}$$

$$= f_X(y)\left\{F_Z(L-y)F_X(y)+\int_{x=y}^{\infty}F_Z(L-x)dF_X(x)\right\}$$

Hence we obtain $f_Y(y) = \sum_{j=1}^{4}\frac{dP(Y_i\leq y,E_j)}{dy}$ given in Theorem 4.□

Proof of Theorem5

Firstly, we define $a = L - Z_{i-S}$, and replace a with $L - Z_{i-S}$ in the Equation 3.2.

$$Y_i = \begin{cases} X_i & if \quad a \leq min(X_{i-S}, X_i) \\ X_{i-S}+X_i-a & if \quad X_{i-S} < a \leq X_i \\ a & if \quad X_i < a \leq X_{i-S} \\ X_{i-S} & if \quad a > max(X_{i-S}, X_i). \end{cases} \tag{B.2}$$

After simplifying the expression for Y_i, then we take the expectation of Y_i. Since all Y_i are identically distributed, we use $E[Y]$ for the expectation of Y_i for all i.

$$E[Y] = \int_{a=-\infty}^{L} \int_{x_2=a}^{\infty} \int_{x_1=a}^{\infty} x_2 f_{Z_{i-S}}(L-a) dF_{X_{i-S}}(x_1) dF_{X_i}(x_2) da$$
$$+ \int_{a=-\infty}^{L} \int_{x_2=a}^{\infty} \int_{x_1=0}^{a} (x_2 + x_1 - a) f_{Z_{i-S}}(L-a) dF_{X_{i-S}}(x_1) dF_{X_i}(x_2) da$$
$$+ \int_{a=-\infty}^{L} \int_{x_2=0}^{a} \int_{x_1=a}^{\infty} a f_{Z_{i-S}}(L-a) dF_{X_{i-S}}(x_1) dF_{X_i}(x_2) da$$
$$+ \int_{a=-\infty}^{L} \int_{x_2=0}^{a} \int_{x_1=0}^{a} x_1 f_{Z_{i-S}}(L-a) dF_{X_{i-S}}(x_1) dF_{X_i}(x_2) da$$

Note that all X_i are i.i.d, therefore all $f_{X_i}(.) = f_X(.)$. We can write the following after some algebra:

$$E[Y] = \int_{a=-\infty}^{L} \left\{ \int_{x_2=a}^{\infty} \left(\int_{x_1=0}^{a} x_2 f_X(x_1) dx_1 + \int_{x_1=a}^{\infty} x_2 dF_X(x_1) \right) dF_X(x_2) \right.$$
$$\left. + \int_{x_1=0}^{a} \left(\int_{x_2=0}^{a} x_1 dF_X(x_2) + \int_{x_2=a}^{\infty} x_1 dF_X(x_2) \right) dF_X(x_1) \right\} f_Z(L-a) da$$

$$E[Y] = \int_{a=-\infty}^{L} \left\{ \int_{x_2=a}^{\infty} (x_2 F_X(a) + x_2 \overline{F}_X(a)) dF_X(x_2) \right.$$
$$\left. + \int_{x_1=0}^{a} (x_1 F_X(a) + x_1 \overline{F}_X(a)) dF_X(x_1) \right\} f_Z(L-a) da$$

$$E[Y] = \int_{a=-\infty}^{L} \left\{ (E[X] - \int_{x_2=0}^{a} x_2 dF_X(x_2) + \int_{x_1=0}^{a} x_1 dF_X(x_1) \right\} f_Z(L-a) da$$

$E[Y] = E[X] \int_{a=-\infty}^{L} f_Z(L-a) da$. By changing the variable: $z = L - a$ and $dz = -da$, we have:
$E[Y] = E[X] \int_{z=0}^{\infty} f_Z(z) dz = E[X]$ Hence, we had shown that $E[Y] = E[X]$.

Now we can prove $Var[Y] \leq Var[X]$ $Var[Y_i] = Var[Y] = E[Y^2] - E[Y]^2$. We have shown in Theorem 5 that $E[Y_i] = E[Y] = E[X]$. Therefore we have:
$Var[X] - Var[Y] = E[X^2] - E[Y^2]$. Next we will derive $E[Y^2]$. Note that we replace a with $L - Z_{i-S}$ for convenience.

$$E[Y^2] = \int_{a=-\infty}^{L} \int_{x_2=a}^{\infty} \int_{x_1=a}^{\infty} x_2^2 f_{Z_{i-S}}(L-a) dF_{X_{i-S}}(x_1) dF_{X_i}(x_2) da$$
$$+ \int_{a=-\infty}^{L} \int_{x_2=a}^{\infty} \int_{x_1=0}^{a} (x_2 + x_1 - a)^2 f_{Z_{i-S}}(L-a) dF_{X_{i-S}}(x_1) dF_{X_i}(x_2) da$$
$$+ \int_{a=-\infty}^{L} \int_{x_2=0}^{a} \int_{x_1=a}^{\infty} a^2 f_{Z_{i-S}}(L-a) dF_{X_{i-S}}(x_1) dF_{X_i}(x_2) da$$
$$+ \int_{a=-\infty}^{L} \int_{x_2=0}^{a} \int_{x_1=0}^{a} x_1^2 f_{Z_{i-S}}(L-a) dF_{X_{i-S}}(x_1) dF_{X_i}(x_2) da$$

Note that all X_i are i.i.d, therefore all $f_{X_i}(.)$ are same, and we use $f_X(.)$ for $f_{X_i}(.)$ from now on.

$$E[Y^2] = \int_{a=-\infty}^{L} \left\{ \int_{x_2=a}^{\infty} \int_{x_1=a}^{\infty} x_2^2 dF_X(x_1) dF_X(x_2) \right.$$
$$+ \int_{x_2=a}^{\infty} \int_{x_1=0}^{a} (x_2^2 + x_1^2 + a^2 + 2x_1x_2 - 2(x_1+x_2)a) dF_X(x_1) dF_X(x_2)$$
$$\left. + \int_{x_2=0}^{a} \int_{x_1=a}^{\infty} a^2 dF_X(x_1) dF_X(x_2) + \int_{x_2=0}^{a} \int_{x_1=0}^{a} x_1^2 dF_X(x_1) dF_X(x_2) \right\} f_Z(L-a)da.$$
$$= \int_{a=-\infty}^{L} \left\{ \int_{x_2=a}^{\infty} x_2^2 dF_X(x_2) + \int_{x_1=0}^{a} x_1^2 dF_X(x_1) \right.$$
$$+2 \int_{x_2=a}^{\infty} \int_{x_1=0}^{a} a^2 dF_X(x_1) dF_X(x_2) + 2 \int_{x_2=a}^{\infty} \int_{x_1=0}^{a} x_1 x_2 dF_X(x_1) dF_X(x_2)$$
$$\left. -2 \int_{x_2=a}^{\infty} \int_{x_1=0}^{a} a(x_1+x_2) dF_X(x_1) dF_X(x_2) \right\} f_Z(L-a)da$$
$$= \int_{a=-\infty}^{L} \left\{ E[X^2] + 2a^2 F_X(a)\overline{F}_X(a) - 2a \left(\int_{x_1=0}^{a} \overline{F}_X(a) x_1 dF_X(x_1) + \right. \right.$$
$$\left. \left. \int_{x_2=a}^{\infty} F_X(a) x_2 dF_X(x_2) \right) + 2 \int_{x_2=a}^{\infty} x_2 \int_{x_1=0}^{a} x_1 dF_X(x_1) dx_2 \right\} f_Z(L-a)da$$
$$E[Y^2] = E[X^2] \int_{a=-\infty}^{L} f_Z(L-a)da + \int_{a=-\infty}^{L} \left\{ 2aF_X(a) \left(a\overline{F}_X(a) - \int_{x_2=a}^{\infty} x_2 dF_X(x_2) \right) \right.$$
$$\left. +2 \int_{x_1=0}^{a} x_1 dF_X(x_1) \left(\int_{x_2=a}^{\infty} x_2 dF_X(x_2) - a\overline{F}_X(a) \right) \right\} f_Z(L-a)da$$
$$= E[X^2] + 2 \int_{a=-\infty}^{L} \left\{ \left(\int_{x_2=a}^{\infty} x_2 dF_X(x_2) - a\overline{F}_X(a) \right) \left(\int_{x_1=0}^{a} x_1 dF_X(x_1) - aF_X(a) \right) \right\}$$
$$f_Z(L-a)da$$

Let $z = L-a$ and replace $L-z$ with all a in the above expression, then we obtain:

$$E[Y^2] = E[X^2] + 2 \int_{z=0}^{\infty} \left\{ \left(\int_{x=L-z}^{\infty} x dF_X(x) - (L-z)\overline{F}_X(L-z) \right) \right.$$
$$\left. \left(\int_{x=0}^{L-z} x dF_X(x) - (L-z)F_X(L-z) \right) \right\} dF_Z(z)$$

By integration by parts, we can write:

$$\int_{x=0}^{L-z} x dF_X(x) = (L-z)F_X(L-z) - \int_{x=0}^{L-z} F_X(x)dx.$$

Then, $\int_{x=0}^{L-z} x dF_X(x) - (L-z)F_X(L-z) = -\int_{x=0}^{L-z} F_X(x)dx.$

We can write $\int_{x=L-z}^{\infty} x dF_X(x)$ as $E[X] - \int_{x=0}^{L-z} x dF_X(x)$. Also note that:

$-\displaystyle\int_{x=0}^{L-z} x dF_X(x) = \int_{x=0}^{L-z} F_X(x)dx - (L-z)F_X(L-z)$

By adding and subtracting $(L-z)$, we obtain:

$\displaystyle\int_{x=0}^{L-z} F_X(x)dx - (L-z) + (L-z)(1 - F_X(L-z)) = (L-z)\overline{F}_X(L-z) - \int_{x=0}^{L-z} \overline{F}_X(x)dx$

Since $E[X] = \displaystyle\int_{x=0}^{\infty} \overline{F}_X(x)dx$, we can write as follows:

$\displaystyle\int_{x=L-z}^{\infty} x dF_X(x) - (L-z)\overline{F}_X(L-z) = \int_{x=L-z}^{\infty} \overline{F}_X(x)dx$

So, we obtain the following expression for $E[Y^2]$.

$$E[Y^2] = E[X^2] - 2\int_{z=0}^{\infty}\left[\left(\int_{x=L-z}^{\infty}\overline{F}(x)dx\right)\left(\int_{x=0}^{L-z}F(x)dx\right)\right]dF_Z(z)$$

Since $Var[X] - Var[Y] = E[X^2] - E[Y^2]$, we have:

$Var[X] - Var[Y] = 2\displaystyle\int_{z=0}^{\infty}\left[\left(\int_{x=L-z}^{\infty}\overline{F}(x)dx\right)\left(\int_{x=0}^{L-z}F(x)dx\right)\right]dF_Z(z) \geq 0$

Hence, $Var[X] \geq Var[Y]\Box$

Appendix C

Summary of Notation

λ_i	:	Demand rate of retailer i
λ_0	:	Total demand rate of retailers $(\sum_{i=1}^{N} \lambda_i)$
r_i	:	λ_i/λ_o, The probability that a demand arrives at retailer i
$A(C)$	:	Fixed ordering cost of a joint retailer order
h_i	:	Unit inventory holding cost per time for retailer i, i=1,2,..,N
β_i	:	Unit backorder holding cost per time for retailer i, i=1,2,..,N
S_i	:	Order-up-to level of retailer i, i=1,2,..,N
S_0	:	$(\sum_{i=1}^{N} S_i)$
Q	:	Joint order quantity
l_i	:	Minor lead-time from the cross-dock to the retailer i, i=1,2,..,N
D	:	Time that passes between the dispatch of a truck and that truck's arrival
L_i	:	Total lead-time from the ample supplier to the retailer i, i=1,2,..,N
C	:	Each truck's capacity
k	:	Number of trucks in the fleet
W_q	:	Waiting time of a joint order due to the absence of an available truck.
$F(W_q)$	:	Distribution function of the waiting time of a replaced order for a truck.
EL_i	:	Effective leadtime from the warehouse to the retailer i, i=1,2,..,N
$NI_i(t)$	:	Net inventory level of retailer i at time t, i=1,2,..,N
$NI(t)$	:	Net inventory level of all retailers at time t, $(\sum_{i=1}^{N} NI_i(t))$
$IP_i(t)$	:	Inventory position of retailer i at time t, i=1,2,..,N
$IP(t)$	:	Total inventory position of all retailers at time t, $(\sum_{i=1}^{N} IP_i(t))$
$TL(t)$	:	Truck level at time t
X	:	Cycle Length t
$f(x, y, z)$	:	p.d.f of gamma distribution with shape:y and scale: z.
$F(x, y, z)$	:	Distribution of a gamma distribution with shape:y and scale: z.