A multi-item multi-echelon inventory system with quantity-based order consolidation

A multi-item multi-echelon inventory system with quantity-based order consolidation

G. P. Kiesmüller, A. G. de Kok
Faculty of Technology Management,
Technische Universiteit Eindhoven,
P.O. Box 513,
5600 MB Eindhoven,
The Netherlands

Abstract

The efficient management of a distribution system requires coordination between transportation planning and inventory control. Small and frequent shipments can reduce inventory levels while maintaining high customer service levels, but they increase unit transportation costs due to inappropriate utilization of the vehicles and more handling costs. On the other hand shipment consolidation policies reduce unit shipping cost because of economies of scale but they require higher inventory levels to maintain high service levels. Hence, it is of considerable interest to understand the trade-off between shipment consolidation policies and inventory levels.

In this paper we consider a distribution network under a quantity based shipment consolidation policy, which means that all orders for a particular destination are held and shipped when a predetermined weight or volume is reached. The studied system is a divergent multi-item multi-echelon network with stochastic demand at the lowest echelon. The stockpoints are controlled by continuous review (s, Q) installation stock policies where the replenishment leadtime is dependent on the shipment consolidation policy. We derive approximations for the first two moments of the leadtimes, the reorder levels, the average inventory levels and the service levels which enables the evaluation of the relevant performance characteristics as well as the allocation of safety stocks for given target service levels. In a numerical study we test our approximations using computer simulation and we illustrate that the estimated service and inventory levels are close to the actual ones.

keywords: stochastic, inventory, multi-echelon, multi-item, shipment consolidation

1 Introduction

The management of multi-echelon inventory systems is a crucial part of supply chain management. In order to reduce costs and to improve efficiency an integrated approach is needed. For instance, decreasing order quantities in order to lower inventory costs induce more frequent shipments, which in general increases transportation unit costs. We believe that careful coordination of shipment and replenishment policies can lead to substantial cost savings.

In this paper we investigate a multi-item multi-echelon distribution network where several deliveries to a single warehouse are dispatched as a single load, realizing economies

of scale in transportation costs. We consider a quantity-based shipment policy which means that orders for a destination are held and shipped when a minimum consolidated weight or volume is reached (Higginson and Bookbinder [13]). The required amount to be consolidated has an influence on the truck departure time and therefore on the replenishment leadtime of the supplied stockpoints. We model this interaction considering the replenishment leadtime as an endogenous variable. Moreover, we explicitly model the waiting time due to shipment consolidation and the waiting time due to a lack of stock at the higher echelon.

We consider a divergent network of warehouses keeping stock of different items. Inventories are controlled using an (s, Q) installation stock policy, which means a replenishment order of size Q is placed when the local inventory position, which is defined as the physical stock plus the stock on order minus backorders, is equal or smaller than the reorder level s. Customer demand is stochastic, only observed at the lowest echelon, and modeled as a compound renewal process which means that interarrival time of orders as well as demand sizes are modeled as random variables. Demand which cannot be satisfied is backordered.

It is assumed that demand sizes and batch sizes are measured in volume and that the batch sizes Q are given and identical for the same stockpoints at an echelon, which can be observed when orders have to be a multiple of a box or a pallet. This standardization of the replenishment sizes enables the use of automated material handling equipment and results in lower handling costs. But this assumption also has an impact on the analysis because, all stockpoints, besides the stockpoints at the lowest echelon, observe a discrete demand size distribution and, therefore, we have to derive different formulae for the first echelon and all other echelons.

Due to the general network structure, the general demand model and the integration of shipment and replenishment policies the study of the system is very challenging and it is very unlikely to obtain exact expressions for relevant performance measures, like service levels and average inventory levels. Therefore, we derive in this paper approximations for the moments of the leadtimes, the customer service levels (in this paper the fill rates) and the average inventory levels. The approximate expressions are derived using fitting techniques (Tijms [23]), a superposition technique proposed by Whitt [24], and approximations based on renewal theory (Cox [10]). Moreover, we provide a fast algorithm, based on our approximations, to determine reorder levels such that given target service levels, are met. Our proposed algorithm determines recursively in a first step the aggregate demand processes at all upper echelons starting at the lowest echelon, and in a second step the waiting times and the leadtimes are determined successively, starting at the highest echelon, which enables the calculation of the reorder levels.

Using a recursive algorithm for the evaluation of a multi-echelon network is an idea, which is already implemented in the famous METRIC model (Sherbrooke [20]) of a multi-item multi-echelon inventory system for spare parts. However, a demand process for spare parts is often modeled by a Poisson or a compound Poisson process and appropriate replenishment policies for spare parts are one-for-one replenishment policies, denoted by $(s-1, s)$. A lot of results are obtained for stochastic multi-item multi-echelon networks for spare parts, see for example Muckstadt [17], but they cannot be used in general systems due to the special characteristics of spare parts.

A general multi-echelon distribution system for a single item with a general demand model and an (s, nQ) replenishment policy is studied in Kiesmüller et al. [15]. In that paper it is assumed that each replenishment order induces an immediate shipment, which is the general assumption in most of the literature on single item multi-echelon inventory models, see for example Axsäter [1]. Since the amount of literature in the field of single item multi-echelon inventory systems is enormous, we only refer to surveys as given in Axsäter [1], Federgruen [12], Axsäter [3], Diks et al. [11] and van Houtum [14]. In this paper we extend the model studied in Kiesmüller et al. [15] and we consider multiple items and explicitly model the shipment policy, resulting in an integrated transportation and inventory model.

A large body of literature on integrated inventory and transportation models is available for logistic networks under deterministic demands. We can distinguish between two main modeling approaches. The first approach has its basis in the EOQ model (see for example Bruns et al. [6] or Çetinkaya and Lee [8]) while the other approach is using mathematical programming formulations of the problem (see for example Popken [18]). We also refer to an excellent overview in this field by Bertazzi [5]. However, all these deterministic models cannot cope with the effects of stochastic demands and consolidation that arise in the problem setting studied in this paper.

An integrated approach for the coordination of inventory and transportation decisions in VMI systems with stochastic demands can be found in Çetinkaya and Lee [9]. A renewal theoretic model is developed for a single item, following a Poisson demand process, to compute the optimal replenishment quantity and dispatch frequency simultaneously. In contrast to our paper, the authors consider a time-based policy, where an accumulated load is shipped every T periods. Moreover, their model only includes the vendor and inventory replenishment leadtime for the vendor is assumed to be negligible. Under these assumptions exact analytical results can be obtained. Dropping these restrictive assumptions, as we do in this paper, leads to a challenging problem where tractable exact solutions will be very difficult to obtain if not impossible. Axsäter [4] provides a note on the model of Çetinkaya and Lee [9] and he suggests an approximation and an adjustment that can be used to improve the proposed solution of Çetinkaya and Lee.

Cachon [7] balances transportation, inventory and shelf space costs for a single link between one retail store and one warehouse. He allows for multiple items and models demand as Poisson processes. The replenishment orders of the retailers are generated for each product by a one-for-one replenishment policy where the order-up-to level is equal to the shelf space. Besides a time based consolidation policy he also investigates a quantity based policy and provides exact expressions for the average costs per time unit. But since the cost functions are not well behaved, he proposes heuristics for the computation of the optimal policy parameters.

The model under study in this paper is an extension of the above mentioned models since we consider a more general network structure and allow for a divergent multi-echelon systems. Moreover, we model a demand process as a compound renewal process, which comprises the Poisson process as a special case. We additionally consider (s, Q) replenishment policies which take into account that order sizes are often restricted to material handling units, such as pallets and boxes. Due to this general assumptions we have to rely on approximations and cannot derive exact formulas to enable a very fast

evaluation of a network. We will illustrate that our approximations are very accurate.

The sequel of the paper is organized as follows. In section 2 a detailed model and problem description is given. In section 3 we explain the used approximation techniques and in section 4 we derive approximated expressions for the service level, the average inventory level, the first two moments for the waiting time due to a lack of stock and the waiting time due to shipment consolidation. In section 5 we propose an algorithm to compute reorder levels such that specified target service levels can be met. In a numerical study we test the performance of the approximations and the performance of the distribution network when the reorder levels are determined by our computational procedure. The paper ends with a short summary and conclusions.

2 Model and problem description

2.1 The distribution network

In this paper we study a divergent multi-echelon distribution network consisting of several warehouses and external suppliers. Each warehouse stocks several items and we call an item at a warehouse a stockpoint. While a replenishment order is sent from a stockpoint to another stockpoint, a shipment is done from a warehouse (a set of stockpoints) to another warehouse. We assume that each stockpoint is supplied exactly by one stockpoint and without loss of generality a warehouse is supplied by exactly one warehouse or an external supplier.

Stockpoints are uniquely numbered with arabic numbers, while the warehouses are numbered with roman numbers. The set of warehouses is denoted by $\mathcal{W}$, the set of all stockpoints by $\mathcal{M}$, and the set of stockpoints located at warehouse $m \in \mathcal{W}$ is denoted by $\mathcal{L}_m$.

Let k be an arbitrary stockpoint, $k \in \mathcal{M}$. Then $\mathcal{S}_k$ is defined as the set of all immediate successors of k, where $j \in \mathcal{M}$ is a successor of k if and only if j and k represent the same item and replenishment orders for j are part of the demand process of k. Further, $\mathcal{P}_k$ is the set of immediate predecessors of $k \in \mathcal{M}$, i.e. replenishment orders of k are part of the demand process of $j \in \mathcal{P}_k$. Note that $|\mathcal{P}_k| = 1$ or 0 for all $k \in \mathcal{M}$ due to the divergent network structure.

We denote N as the number of echelons in the network and $\mathcal{E}_n$ as the set of stockpoints of the n-th echelon.

$$\mathcal{E}_1 := \{k \in \mathcal{M} | \mathcal{S}_k = \emptyset\}$$

$$\mathcal{E}_n := \{k \in \mathcal{M} | \mathcal{S}_k \subset (\bigcup_{i=1}^{n-1} \mathcal{E}_i)\} \backslash \{k \in \mathcal{M} | \mathcal{S}_k \subset (\bigcup_{i=1}^{n-2} \mathcal{E}_i)\} \qquad n = 2, ..., N$$

where $\bigcup_{i=1}^{0} \mathcal{E}_i = \emptyset$. Further, we assume that customers are replenished by $k \in \mathcal{E}_1$, a stockpoint $j \in \mathcal{E}_{n-1}$ by $k \in \mathcal{E}_n$, $(n = 2, ..., N)$ and $k \in \mathcal{E}_N$ by an external supplier.

In Figure 1 we illustrate a three echelon distribution network with four warehouses and

four different items.

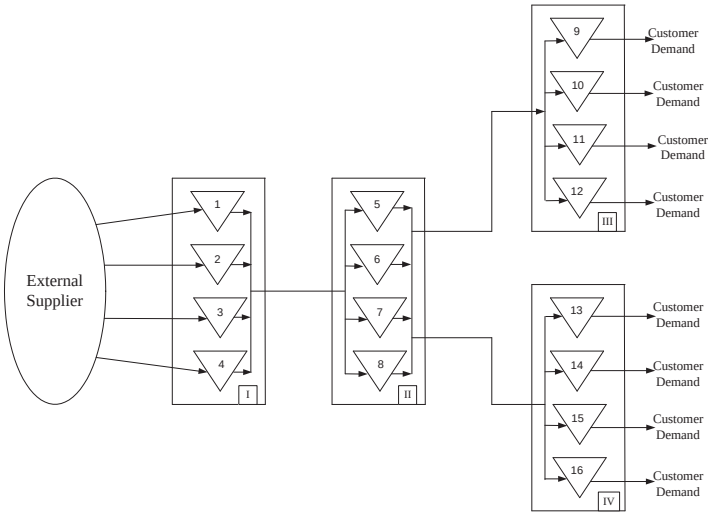

Figure 1: A typical distribution network

Transportation rates are often not only dependent on the origin-destination distance but also on the quantity shipped. In many situations increased per-shipment volume results in economies of scale and lower per unit transportation cost. Therefore, orders from the same warehouse are consolidated and shipped together. In this paper we assume a quantity-based consolidation policy (Higginson and Bookbinder [13]) which means orders for a warehouse $m \in \mathcal{W}$ are dispatched until a target quantity Q_m^c is reached. Furthermore, transportation times, defined as the time between the truck departure at a warehouse m and the moment that the items are available on stockpoint k at the supplied warehouse, are denoted by $T_{m,k}, m \in \mathcal{W}, k \in \mathcal{S}_j, j \in \mathcal{L}_m$. We model the transportation time as a random variable and assume the first two moments to be known.

The stockpoints in the warehouses are continuously reviewed and replenishment orders are triggered according to an (s_k, Q_k)-installation stock policy. This means that as soon as the local inventory position, which is expressed as the physical inventory plus the stock on order minus backorders, is equal to s_k or drops below s_k an amount Q_k is ordered. We call s_k the reorder level and Q_k the batch size of stockpoint $k \in \mathcal{M}$.

We assume that the replenishment size is standardized which means that all stockpoints at one echelon use the same size batch sizes, i.e. $Q_k = \tilde{Q}_n, \forall k \in \mathcal{E}_n$, and that the batch size of a stockpoint k is a multiple of the batch size of its successor, i.e. $Q_k = m \cdot \tilde{Q}_n, k \in \mathcal{P}_j, j \in \mathcal{E}_n$, which is a common assumption (integer-ratio assumption) in literature, see for example Axsäter [1]. As a consequence we also restrict the values of Q_m^c to multiples of $Q_j, (j \in \mathcal{S}_k, k \in \mathcal{L}_m, m \in \mathcal{W})$ as well as the reorder levels s_k.

Customer demand is only observed at the most downstream stockpoints at the first echelon, $k \in \mathcal{E}_1$ and can be modeled as a compound renewal process. Let A_k, $k \in \mathcal{E}_1$, denote the length of an arbitrary interarrival time and D_k, $k \in \mathcal{E}_1$, the demand size of an arbitrary customer at the stockpoint. We assume the first two moments of these random variables to be known and fit a mixed-Erlang distribution on these moments. Further, we assume that all unsatisfied demand is backordered.

Summarizing, the following random variables are exogenous to the system under consideration and we assume the first two moments to be known.

D_k	Demand size for $k \in \mathcal{E}_1$
A_k	Time between two subsequent arrivals of orders $k \in \mathcal{E}_1$
$T_{m,k}$	Transportation time from warehouse $m \in \mathcal{W}$ to stockpoint k, $(k \in \mathcal{S}_j, j \in \mathcal{L}_m)$

Additionally we assume that the following parameters are given.

Q_k	Batchsize at $k \in \mathcal{M}$
Q_m^c	Predetermined consolidation quantity towards $m \in \mathcal{W}$

2.2 The problem description

Since forecasted demand in general differs from actual demand safety stock is held in order to be able to satisfy customer demand. One of the major aims of inventory management is to balance service requirements with the costs related to inventory availability. Therefore, it is important to be able to estimate the service levels and average inventory levels under given policy parameters or to determine policy parameters such that prespecified service requirements can be met. In this paper we provide formulae to evaluate a general distribution network under a quantity based consolidation policy and we show how this formula can be applied in a computational procedure to determine reorder levels such that given target service levels can be met.

It is well known that the leadtime L_k $(k \in \mathcal{M})$, defined as the time between the placement and the arrival of an order at stockpoint k, has an influence on the required safety stock at stockpoint k. However, the leadtime of a stockpoint in a multi-echelon network in turn depends on the stock availability of its predecessor. If there is not enough stock at $\mathcal{P}_k$ to fulfill the entire order, then the order has to wait until $\mathcal{P}_k$ is replenished and afterwards the order becomes available for consolidation. If more than one order is waiting due to a lack of stock a FIFO (first in, first out) rule is applied. The time between the arrival of the replenishment order at the stockpoint $\mathcal{P}_k$ and the time where the order is available for consolidation is called the waiting time due to a lack of stock and is denoted with W_k^s. When an order is available for consolidation it has to wait until the target quantity Q_m^c is dispatched and the truck departs. We call this part W_k^c of the leadtime the waiting time due to shipment consolidation. We explicitly model these components of the leadtime via relation (1) which enables us to study the relationship between safety stocks and the shipment consolidation policy.

$$L_k = T_{m,k} + W_k^s + W_k^c, \quad m \in \mathcal{W}, j \in \mathcal{L}_m, k \in \mathcal{S}_j \tag{1}$$

We assume that crossing of orders cannot occur and that the waiting times for replenishment orders at the external suppliers are equal to zero $(W_k^s = 0, W_k^c = 0, \forall k \in \mathcal{E}_N)$.

The following notation is used:

s_k	Reorder level of stockpoint $k \in \mathcal{M}$
$E[I_k]$	Long run expected physical inventory level at $k \in \mathcal{M}$
β_k	Actual fill rate at stockpoint k

Below we define the used functions and operators.

$E[Y]$	Expectation of the random variable Y
$\sigma^2(Y)$	Variance of the random variable Y
c_Y	Coefficient of variation of Y ($c_Y := \frac{\sigma(Y)}{E[Y]}$)
$(a)^+$	Maximum of the real number a and zero
$P(A)$	Probability of event A

2.3 The processes

In order to be able to compute the reorder levels, the service levels and the average inventory levels for all stockpoints in the distribution network we have to consider several processes and we need some variables to describe these processes.

The demand processes at higher echelons are also modeled as compound renewal processes with random interarrival times of orders A_k and fixed demand sizes D_k equal to the batchsizes of the successors.

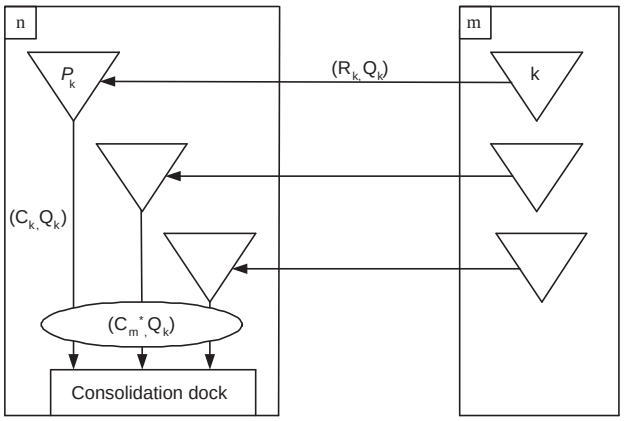

Figure 2: The replenishment processes

In the following we will distinguish between two types of arrival processes (see Figure 2): the arrival process (R_k, Q_k) of replenishment orders from stockpoint k at stockpoint $\mathcal{P}_k$ and the arrival process of replenishment orders from stockpoint k at the consolidation dock. We define C_k as the stationary interarrival time of the latter process. Both processes are equivalent when there are enough items on stock. Otherwise the replenishment orders at the consolidation dock are delayed and also clustering of the orders is possible, which means that more than one replenishment order of the same item can arrive simultaneously at the consolidation dock. The interarrival time of arbitrary replenishment orders from warehouse m at the consolidation dock is denoted with C_m^* and the size of an arbitrary order is denoted with B_m^*.

The waiting time of an order due to shipment consolidation is dependent on the amount of goods already consolidated when the order arrives at the consolidation dock Y_k. We are further interested in the consolidated quantity $\Delta_m(t)$ for warehouse m at time t and the number of arrivals of replenishment orders from warehouse m between the arrival of a

replenishment order at the consolidation dock and the departure of the truck $N(Q_m^c - Y_k)$.

The distribution of the following random variables is unknown:

R_k	Time between subsequent replenishment orders of stockpoint $k \in \mathcal{M}$
A_k	Time between two subsequent arrivals of orders at stockpoint $k \in \mathcal{M} \backslash \mathcal{E}_1$
D_k	Demand sizes at stockpoints $k \in \mathcal{M} \backslash \mathcal{E}_1$
C_k	Interarrival time of replenishment orders from stockpoint $k \in \mathcal{M}$ at the consolidation dock
C_m^*	Interarrival time of an arbitrary replenishment order from warehouse m at the consolidation dock for $m \in \mathcal{W}$
$\Delta_m(t)$	Consolidated quantity for warehouse $m \in \mathcal{W}$ at time t
Y_k	Amount consolidated between the last truck departure towards $m \in \mathcal{W}$ and the arrival of an arbitrary order from $k \in \mathcal{L}_m$
$N(Q_m^c - Y_k)$	Number of arrivals of replenishment orders from m between the arrival of a replenishment order at the consolidation dock from $k \in \mathcal{L}_m$ and the departure of the truck towards $m \in \mathcal{W}$.
$D_k(T)$	Demand at stockpoint $k \in \mathcal{M}$ during a time interval with length T
W_k^s	Waiting time due to a lack of stock for a replenishment order from stockpoint $k \in \mathcal{M}$
W_k^c	Waiting time due to shipment consolidation for a replenishment order from stockpoint $k \in \mathcal{M}$

3 The approximation techniques

In general, the exact expressions of the distribution functions or of the moments of the above mentioned random variables are difficult to obtain or intractable. Therefore, we rely on approximated expressions. Before we will derive these expressions we will explain the techniques used. Three different approaches are used.

3.1 Fitting mixed Erlang Distributions

When exact expressions for the distribution function of a random variable are difficult to obtain or intractable, then we can resort to the following approximation technique. We evaluate the first two moments of the random variable and we approximate the distribution of it by a mixed-Erlang distribution with the same first two moments.

A mixed-Erlang distribution, denoted as $E_{k_1,k_2}((\mu_1,\mu_2),(p_1,p_2))$, is here defined as the mixture of two-Erlang distributions. This means a mixed-Erlang distributed random variable X is with probability p_1 (resp. $p_2 = 1 - p_1$) the sum of $k_1 - 1$ (resp. $k_2 - 1$) independent exponentials with mean $\frac{1}{\mu_1}$ (resp. $\frac{1}{\mu_2}$). The density of an $E_{k_1,k_2}((\mu_1,\mu_2),(p_1,p_2))$ distribution is therefore given by

$$f_X(x) := \sum_{i=1}^{2} p_i \mu_i^{k_i} \frac{x^{k_i-1}}{(k_1-1)!} e^{-\mu_i x} \qquad x > 0 \tag{2}$$

The parameters p_i, k_i and μ_i $(i = 1, 2)$ can be found in the following way if the first two moments of X are known.

- If $c_X^2 < 1$ then a mixture of two Erlang distributions with the same scale parameter is used. Hence,

  $k_1 := \lfloor \frac{1}{c_X^2} \rfloor$, $k_2 := k_1 + 1$

  $p_1 := \frac{1}{1+c_X^2} \left(k_2 c_X^2 - \sqrt{k_2(1 + c_X^2) - k_2^2 c_X^2} \right)$, $p_2 := 1 - p_1$

  $\mu_1 := \frac{k_2 - p_1}{E[X]}$, $\mu_2 := \mu_1$

- If $c_X^2 \geq 1$ then a mixture of two exponential distributions used.

  $k_1 := 1$, $k_2 := 1$

  $\mu_1 := \frac{2}{E[X]} \left(1 + \sqrt{\frac{c_X^2 - \frac{1}{2}}{c_X^2 + 1}} \right)$, $\mu_2 := \frac{4}{E[X]} - \mu_1$

  $p_1 := \frac{\mu_1(\mu_2 E[X] - 1)}{\mu_2 - \mu_1}$, $p_2 := (1 - p_2)$

This fitting technique is common, see for example Tijms [23].

3.2 Superposition of renewal processes

The demand process of a stockpoint k is the superposition of the replenishment order processes of the stockpoints $j \in \mathcal{S}_k$. How to compute approximations of the first two moments of the interarrival times of replenishment orders is given in Appendix A. For a detailed derivation see Kiesmüller et al. [15]. We assume that the superposition process is again a renewal process and for the computation of the first two moments of the interarrival times of arbitrary orders we use the stationary interval method developed by Whitt [24]. But instead of superposing hyper-exponential and shifted exponential distributions, we superpose mixtures of Erlang distributions. A detailed description of the algorithm is given in Appendix B.

3.3 Approximations based on renewal theory

One of the problems when analyzing the distribution network is to determine the number of events $N(L)$ in a given time interval with length L, when the events follow a renewal process. If the moments of the interarrival time X of the events are known, then the first two moments of $N(L)$ can be computed using asymptotic relations from standard renewal theory and we obtain (see, for example Cox [10])

$$E[N(L)] \approx \frac{L}{E[X]} \tag{3}$$

and

$$E[N(L)^2] \approx \frac{L^2}{E[X]^2} + L\Big(\frac{E[X^2]}{E[X]^3} - \frac{1}{E[X]}\Big) + \frac{E[X^2]^2}{2E[X]^4} - \frac{E[X^3]}{3E[X]^3} \tag{4}$$

if time epoch zero is an arbitrary point in time. Otherwise, if at time zero an arrival occurs, we get:

$$E[N(L)] \approx \frac{L}{E[X]} + \frac{E[X^2]}{2E[X]^2} - 1 \tag{5}$$

and

$$E[N(L)^2] \approx \frac{L^2}{E[X]^2} + L\Big(\frac{2E[X^2]}{E[X]^3} - \frac{3}{E[X]}\Big) + \frac{3E[X^2]^2}{2E[X]^4} - \frac{2E[X^3]}{3E[X]^3} - \frac{3E[X^2]}{2E[X]^2} + 1 \tag{6}$$

If L is a random variable then we take the expectations of the right hand side of the equations (3) - (6). The above approximation technique works well if $P(L < Cond(X)) \leq \epsilon$ for ϵ small enough (cf. Tijms [23]) and

$$Cond(X) := \begin{cases} \frac{3}{2}c_X^2 E[X] & : \quad c_X^2 > 1 \\ E[X] & : \quad 0.2 < c_X^2 \leq 1 \\ \frac{1}{2c_X} E[X] & : \quad 0 < c_X^2 \leq 0.2 \end{cases} \tag{7}$$

4 Analytical approximations

4.1 The service level

A reasonable service level for measuring the availability of stock is the fill rate, defined as the fraction of demand delivered directly from shelf. We denote the fill rate at a stockpoint $k \in \mathcal{M}$ by β_k. The following relationship between the reorder level s_k and the fill rate $\beta_k, (k \in \mathcal{M})$ is well-known (see Tijms [23])

$$\beta_k = 1 - \frac{E[(D_k(L_k) + U_k - s_k)^+] - E[(D_k(L_k) + U_k - s_k - Q_k)^+]}{Q_k} \tag{8}$$

where U_k is the undershoot, defined as the difference between the reorderlevel and the inventory position just before the placement of an order. Additionally, we have to distinguish between stockpoints $k \in \mathcal{E}_1$ at the first echelon where demand sizes are continuously distributed and the stockpoints $k \in \mathcal{M} \backslash \mathcal{E}_1$ at upper echelons where demand sizes are discrete.

For the stockpoints k at the first echelon $\mathcal{E}_1$ we use the approximation technique as described in section 3.1 and determine the first two moments of the demand during the leadtime $D_k(L_k)$ and the undershoot U_k and fit a mixed Erlang distribution. The expectations in (8) can easily be computed due to the relation given in (9) which holds under the assumption that the random variable X is mixed Erlang distributed $E_{k_1,k_2}((\mu_1,\mu_2),(p_1,p_2))$ and $z \in \mathbb{R}$.

$$\begin{aligned} E[(X-z)^+] &= \int_z^\infty (x-z) dF_X(x) \\ &= \sum_{j=0}^{1} (-z)^j \left(\sum_{i=1}^{2} p_i \frac{\mu_i^{k_i}}{(k_i-1)!} \sum_{l=0} \frac{z^l (k_i-j)!}{l! \mu_i^{k_j-l-j+1}} e^{-\mu_i z} \right) \end{aligned} \tag{9}$$

Formulas for the moments of the demand during the leadtime $D_k(L_k)$ and the undershoot U_k are obtained using asymptotic relations based on renewal theory. They are given in Appendix C.

Stockpoints k at upper echelons $k \in \mathcal{M}\backslash\mathcal{E}_1$ only observe a demand size $\tilde{Q}_j, (j \in \mathcal{S}_k)$. Since the reorderlevel is also a multiple of this demand size the undershoot U_k at these stockpoints is always equal to zero. Further, the demand during the leadtime is a discrete random variable and the expectations in (8) can be computed as

$$E[(D_k(L_k) - c)^+] = \sum_{i=r}^{\infty}(i-r)\tilde{Q}_j P(D_k(L_k) = i\tilde{Q}_j) = \sum_{i=r}^{\infty}(i-r)\tilde{Q}_j P(N(L_k) = i) \quad (10)$$

with $c = r\tilde{Q}_j$. How to compute the probability function of $N(L_k)$ can be found in Appendix D.

It remains to explain how to compute the first two moments of the leadtime. As mentioned in section 2.2 the leadtime is composed of the transportation time $T_{m,k}$, the waiting time due to a lack of stock W_k^s and the waiting time due to shipment consolidation W_k^c. In order to be able to use the approximation technique as described in section 3.1 we have to determine the first two moments of W_k^s and W_k^c.

4.2 The waiting time due to a lack of stock

From standard probability theory it is well-known that the first two moments of a continuous random variable W which can only have values with positive probability on the interval $[0, \ell]$ is given as

$$E[W] = \int_0^{\ell} P(W \geq x)\, dx \qquad E[W^2] = \int_0^{\ell} 2xP(W \geq x)\, dx \quad (11)$$

Suppose that the system is stationary at time 0 and let us consider stockpoint k which is supplied by stockpoint $j \in \mathcal{P}_k$. The cumulative probability function $P(W_k(t) \leq x)$ for the waiting time of an arbitrary order of stockpoint k arriving at time t can be computed as follows:

$$P(W_k(t) \leq x) = \int_0^{\infty} P(W_k(t) \leq x \mid L_j(t) = z)\, f_{L_j}(z)\, dz \quad (12)$$

where $L_j(t)$ denotes the leadtime of the last outstanding order. We first compute the conditional probability function $P(W_k(t) \leq x \mid L_j(t) = z)$.

In the sequel we assume $s \geq 0$. If the leadtime $L_j(t)$ is equal to z then the waiting time $W_k(t)$ can never be larger than z. Further, the waiting time of an order of size Q_k arriving at time t at stockpoint $j \in \mathcal{P}_k$ is smaller than x if the inventory position $Z_j(t-z+x)$ at time $t-z+x$ minus the demand $D(t+x-z,t)$ during the interval $(t+x-z,t)$ is larger than the ordersize Q_k.

$$Y_j(t-z+x) - D_j(t+x-z,t) \geq Q_k \Leftrightarrow W_k(t) \leq x \quad (13)$$

Since the inventory position is uniformly distributed on $s_j + iQ_k, i = 1, 2, \ldots \frac{Q_j}{Q_k}$ we get

$$
\begin{aligned}
&P(W_k(t) \leq x \mid L_j(t) = z) \\
&= \sum_{i=1}^{\frac{Q_j}{Q_k}} P(Y_j(t - z + x) - D_j(t + x - z, t) \geq Q_k \mid Y_j(t - z + x) = s_j + iQ_k) \\
&\qquad \cdot P(Y_j(t - z + x = s_j + iQ_k) \\
&= \frac{Q_k}{Q_j} \sum_{i=1}^{\frac{Q_j}{Q_k}} P\Big(N(z - x) \leq \frac{s_j + (i-1)Q_k}{Q_k}\Big) \qquad (14)
\end{aligned}
$$

How to obtain the distribution function of $N(z)$ can be found in the Appendix D and the integrals in (11) and (12) are computed numerically.

4.3 The consolidation process

Since the arrival process of replenishment orders at the consolidation dock has a large impact on the truck departure process and therefore also on the waiting time due to consolidation, we investigate in the sequel the process in more detail and we derive approximated formulae for the first to moments of the interarrival time C_k of an individual replenishment process at the consolidation dock.

As already mentioned in Section 2.3 the interarrival times C_k of the replenishment orders from stockpoint k at the consolidation dock are different from the interarrival times R_k of the replenishment processes at the stockpoint $\mathcal{P}_k$ due to the clustering of arrivals at the consolidation dock. See for an illustration Figure 3.

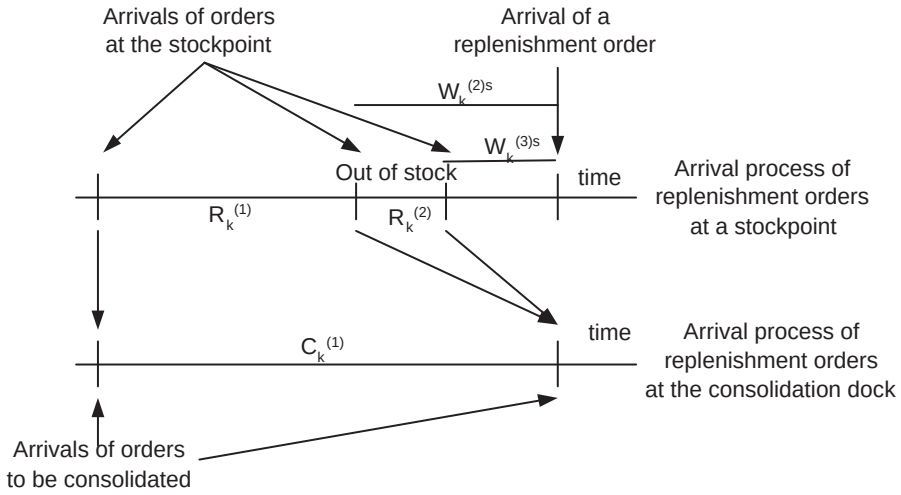

Figure 3: The clustering of arrivals at the consolidation dock

The following equation holds, denoting the interarrival time between the n-th and the $(n+1)$-th order of stockpoint k at the consolidation dock with $C_k^{(n)}$:

$$C_k^{(n)} = R_k^{(n)} - W_k^{(n)s} + W_k^{(n+1)s} \qquad \forall k \in \mathcal{M} \qquad (15)$$

Assuming that all random variables in (15) are identically distributed for all n leads to

$$E[C_k] = E[R_k] \qquad \forall k \in \mathcal{M} \tag{16}$$

The second moment of C_k is difficult to obtain because $W_k^{(n)s}$ and $W_k^{(n+1)s}$ are dependent as illustrated in Figure 3. But if the fill rate at $\mathcal{P}_k$ is not too small then the probability that an order has to wait is negligible. Therefore, we neglect the waiting time due to a lack of stock and use the following formula for the variance of C_k:

$$\sigma^2(C_k) = \sigma^2(R_k) \qquad \forall k \in \mathcal{M} \tag{17}$$

Further, we are not only interested in the arrival process of replenishment orders of one stockpoint k at the consolidation dock but in the aggregated process of arbitrary orders from warehouse m. The first two moments of C_m^* are derived using the superposition algorithm as given in Appendix B.

4.4 The waiting time due to shipment consolidation

Replenishment orders at the consolidation dock for warehouse $m \in \mathcal{W}$ are collected until the amount is equal to a predetermined quantity Q_m^c. Then the consolidation process starts all over again. The collected quantity at the consolidation dock to be shipped to warehouse m at time t is denoted with $\Delta_m(t)$. The departure times of the trucks are random, because they depend on the arrival process of replenishment orders at the consolidation dock (see Figure 4).

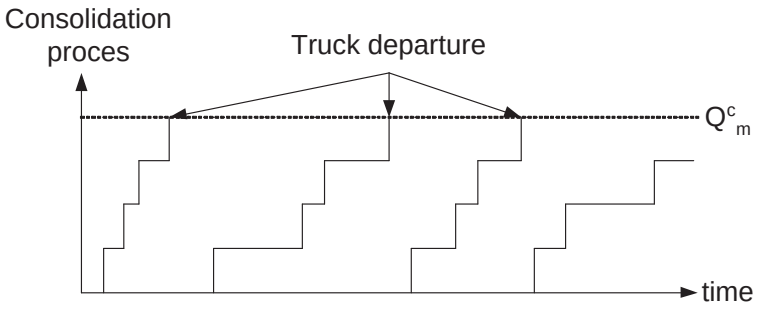

Figure 4: Evolution of the consolidated quantity

The waiting time due to shipment consolidation for an order of size Q_k from stockpoint $k \in \mathcal{L}_m$ is dependent on the order process after the arrival of this order and the amount already consolidated. We define Y_k as the collected amount immediately after a replenishment order of stockpoint k has arrived at the consolidation dock and $N(Q_m^c - Y_k)$ as the number of arrivals between an arrival of a replenishment order of stockpoint k at the consolidation dock and the departure of the truck.

Then the waiting time can be calculated as

$$W_k^c = \sum_{i=1}^{N(Q_m^c - Y_k)} C_{m,i}^* \tag{18}$$

where $C^*_{m,i}$ is the interarrival time between the i-th and $i+1$-th order at the consolidation dock.

Assuming that, $N(Q^c_m - Y_k)$ and $C^*_{m,i}$ are independent random variables, and $C^*_{m,i}$ is identically distributed for all i, it follows from (18) that

$$E[W^c_k] = E[N(Q^c_m - Y_k)]E[C^*_m] \tag{19}$$

$$E[(W^c_k)^2] = E[N(Q^c_m - Y_k)]\sigma^2(C^*_m) + E[N(Q^c_m - Y_k)^2]E^2[C^*_m] \tag{20}$$

Since Q^c_m is a multiple of Q_k the truck consolidation process, $\Delta_m(t)$, follows a discrete distribution. Moreover, in the steady state $\Delta_m(t)$ is uniformly distributed over $\{0, Q_k, 2Q_k, ..., (\frac{Q^c_m}{Q_k} - 1)Q_k\}$ with the different possibilities having the same probability namely $\frac{Q_k}{Q^c_m}$ and it easily follows that

$$E[N(Q^c_m - Y_k)] = \frac{Q_k}{Q^c_m} \sum_{j=0}^{(\frac{Q^c_m}{Q_k}-1)} j = \frac{1}{2}\Big(\frac{Q^c_m}{Q_k} - 1\Big) \tag{21}$$

$$E[N(Q^c_m - Y_k)^2] = \frac{Q_k}{Q^c_m} \sum_{j=0}^{(\frac{Q^c_m}{\tilde{Q}}-1)} j^2 = \frac{1}{3}\Big(\frac{Q^c_m}{\tilde{Q}} - 1\Big)\Big(\frac{Q^c_m}{\tilde{Q}} - \frac{1}{2}\Big) \tag{22}$$

We can substitute these expressions for the first two moments of $N(Q^c_m - Y_k)$ together with the expressions for the first two moments of C^*_m in equations (19) and (20) to get approximations for the first two moments of the waiting time due to shipment consolidation.

4.5 The average inventory level

In order to measure the performance of the distribution network we additional compute the average inventory levels. We have to distinguish between the situation with continuously distributed demand sizes, as given at the first echelon, and the situation with discrete demand sizes as given at the upper echelons. For the first echelon, which means for all $k \in \mathcal{E}_1$, we can use formula (23) and the approximation technique described in Section 3.1 to fit the first two moments of the undershoot and the demand during the leadtime on a mixed Erlang distribution. Denoting with $\tilde{A}_k$ the asymptotic residual lifetime of A_k and with $c_k := E[D_k]\frac{E[\tilde{A}_k]-E[A_k]}{E[A_k]}$ we obtain

$$\begin{aligned} E[I_k] \;=\; & \frac{1}{2Q_k}\Big(E[\big((s_k + Q_k - D_k(L_k))^+\big)^2] - E[\big((s_k - D_k(L_k))^+\big)^2]\Big) \\ & + c_k \frac{E\Big[\Big(s_k + Q_k - U_k - D_k(L_k)\Big)^+\Big] - E\Big[\Big(s_k - U_k - D_k(L_k)\Big)^+\Big]}{Q_k} \end{aligned} \tag{23}$$

(for a proof of (23) see Kiesmüller et al. [15]).

For the upper echelons we have to take into account that the demand sizes are discrete and that the inventory level can only have discrete values. Then formula (24) can be used for computing the average inventory level for the stockpoints $k \in \mathcal{M}\backslash\mathcal{E}_1$.

$$\begin{aligned} E[I_k] &= \frac{Q_j^2}{Q_k E[A_k]}\left(\sum_{i=0}^{i_{s+Q}}\left((i_{s+Q}-i)E[\tilde{A}_k]+\frac{1}{2}(i_{s+Q}-i)(i_{s+Q}-i-1)E[A_k]\right)P(N(L_k)=i)\right) \\ &\quad +\frac{Q_j^2}{Q_k E[A_k]}\left(\sum_{i=0}^{i_s}\left((i_s-i)E[\tilde{A}_k]+\frac{1}{2}(i_s-i)(i_s-i-1)E[A_k]\right)P(N(L_k)=i)\right) \end{aligned} \tag{24}$$

with $j \in \mathcal{S}_k$, $i_s = \frac{s}{Q_j}$ and $i_{s+Q} = \frac{s+Q_k}{Q_j}$. For a proof of (24) see Appendix E.

5 An algorithm to determine reorder levels

For given policy policy parameters $s_k, Q_k, (k \in \mathcal{M})$ and $Q_m^c, (m \in \mathcal{W})$ it is possible to compute an estimation of fill rates using (8) and an estimation of the average inventory levels using (23) and (24). Since many inventory managers work with target service levels in order to be able to satisfy customer demand we now assume that for all stockpoints $k \in \mathcal{E}_n, (n = 1, 2, \ldots, N)$ a target fill rate β_k^{target} is given and we provide an algorithm, based on our approximations, to compute the reorder levels such that the service requirements are met.

The algorithm can be split in two main parts. In the first part we determine the replenishment order processes, the aggregate demand processes and the consolidation processes. The demand process at a stockpoint can be translated in the replenishment order process independently of the value of the reorder level. The superposition of all replenishment order processes of the stockpoints $j \in \mathcal{S}_k$ describes the aggregate demand process at stockpoint k, and the superposition of the replenishment order processes of all stockpoints $j \in \mathcal{L}_m$ describes the consolidation process. We start at the lowest echelon and compute successively the moments of the processes. A sketch of the first part of the algorithm is given below.

1. Set $n = 1$.
2. Compute the first two moments of the interarrival time of replenishment orders $R_k, \forall k \in \mathcal{E}_n$ using the formulas (31) and (32) as given in Appendix A.
3. Compute the first two moments of the interarrival time of orders A_j for all $j \in \mathcal{E}_{n+1}$ by superposing the replenishment processes of all stockpoints $k \in \mathcal{S}_j$ using the algorithm given in Appendix B.
4. Compute the first two moments of the interarrival time C_m^* of an arbitrary replenishment order from warehouse m at the consolidation dock by superposing the replenishment processes of all stockpoints $j \in \mathcal{E}_n, j \in \mathcal{L}_m$ using the algorithm as given in Appendix B.

5. Set $n := n + 1$ and repeat step 2,3, and 4 until $n = N - 1$

After determining the moments of the waiting time due to a lack of stock and of the waiting time due to shipment consolidation the moments of the leadtime can be computed. Since the demand process is already determined in the first part of the algorithm and target service levels are given, the reorderlevels can be computed such that a service level of at least β_k^{target} can be met. The second part of the algorithm starts at the highest echelon and computes successively the reorder levels echelon by echelon until all values are determined. We outline the second part of the algorithm as follows:

6. Set $n := N, W_k^s := 0$ and $W_k^c := 0 \; \forall k \in \mathcal{E}_n$.
7. Compute the first two moments of the leadtime $L_k \; \forall k \in \mathcal{E}_n$ using relation (1).
8. Compute the reorder levels $s_k \forall k \in \mathcal{E}_n$ using equation (8).
9. Set n := n-1
10. Compute the moments of the waiting time due to a lack of stock $W_k^s \; \forall k \in \mathcal{E}_n$ as described in section 4.2.
11. Compute the moments of the waiting time due to shipment consolidation $W_k^c \; \forall k \in \mathcal{E}_n$ using the formulas (19) and (20).
12. Compute the first two moments of the leadtime $L_k \forall k \in \mathcal{E}_n$ using relation (1).
13. Compute the reorder levels $s_k \; \forall k \in \mathcal{E}_n$ using equation (8).
14. Repeat step 9-13 until $n = 1$

For the application of the algorithm data about the demand process at the lowest echelon $(A_k, D_k), (k \in \mathcal{E}_1)$, the transportation times $T_{m,k}$, the batchsizes Q_k and the target fillrates β_k^{target} at all stockpoints and the target dispatch quantity Q_m^c for all warehouses are needed as input parameters.

6 Numerical study

In this paragraph we investigate the performance of the approximations by means of discrete event simulation. The performance of the approximations for a single item distribution network without order consolidation is tested in Kiesmüller [15]. There it is illustrated for the single item distribution network that approximations for the moments of the waiting time due to a lack of stock as well as the approximations of the performance measures perform very well and best in heterogeneous structures, i.e. non-identical demand streams and batch sizes. Therefore, in this paper we only study the performance of the approximations of the moments of the interarrival times of the replenishment orders at the consolidation dock and the moments of the waiting time due to shipment consolidation. We further report on the performance of the distribution network in terms of

actual service level and average inventory when reorder levels are used, computed with our algorithm.

As a base case we consider a two echelon distribution network with one central warehouse I and 4 decentral warehouses $\{II, III, IV, V\}$. At $k \in \mathcal{E}_1$ demand arrives according to a compound renewal process. The interarrival time and the order size of $k \in \mathcal{E}_1$ are mixed-Erlang distributed with known first two moments. In the first part of the numerical study we assume all customers to be identical. For the simulation of the base case the input parameters are chosen as given in Table 1.

Parameter	for	value
$E[D_k]$	$k \in \mathcal{E}_1$	50
$c^2_{D_k}$	$k \in \mathcal{E}_1$	1
$E[A_k]$	$k \in \mathcal{E}_1$	1
$c^2_{A_k}$	$k \in \mathcal{E}_1$	1
Q_k	$k \in \mathcal{E}_1$	500
Q_k	$k \in \mathcal{E}_2$	4000
L^d_k	$k \in \mathcal{E}_1$	2
L_k	$k \in \mathcal{E}_2$	4
β^{target}_k	$k \in \mathcal{E}_1$	0.95
β^{target}_k	$k \in \mathcal{E}_2$	0.75
Q^c_m	$m \in \{II, III, IV, V\}$	2000

Table 1: Input parameter values for the base case

To investigate the effects of the input parameters on the output parameters we vary one input parameter while the other parameters are fixed. If not mentioned otherwise, we have chosen the input parameters as given in Table 1.

For measuring the performance of the approximations, we define $\delta^{(Z_k)}$ the relative error of Z_k where $k \in \mathcal{W}$ or $\mathcal{M}$ depending of the context.

$$\delta^{(Z_k)} := \frac{|Z^{simu}_k - Z_k|}{Z^{simu}_k} \cdot 100\% \tag{25}$$

Z_k is the calculated and Z^{simu}_k is the simulated value. Further, we define $\bar{\delta}^{(Z_k)}$ as the average relative error

$$\bar{\delta}^{(Z_k)} := \frac{\sum_{k=1}^{\mathrm{K}} \delta^{(Z_k)}_k}{\mathrm{K}} \tag{26}$$

and $\max(\delta^{(Z_k)})$ as the maximum relative error

$$\max(\delta^{(Z_k)}) := \max_{k=1,...,\mathrm{K}} \delta^{(Z_k)}_k \tag{27}$$

6.1 Approximations for the second moment of C_m^*

In order to compute the second moment of the interarrival times of replenishment orders at the consolidation dock we have used the approximation technique based on the superposition of renewal processes as developed by Whitt [24]. We first investigate the influence of the number of renewal processes which are superposed on the quality of the approximations. Therefore, we vary the number of different items at a warehouse m ($|\mathcal{L}_m| \in \{4, 8, 12, 16, 24, 32\}, m \in \{II, III, IV, V\}$) and analyze the errors. In order to avoid large waiting times due to a lack of stock we have assumed a high service level at the upper stockpoints and we have set $\beta_k^{target} = 0.90\ \forall k \in \mathcal{E}_2$. In Figure 5 we illustrate the average relative error $\overline{\delta}^{(E[(C_m^*)^2])}$ computed over all stockpoints at the first echelon for different values for the squared coefficient of variation of the interarrival times of customer demand at the first echelon and for different number of items.

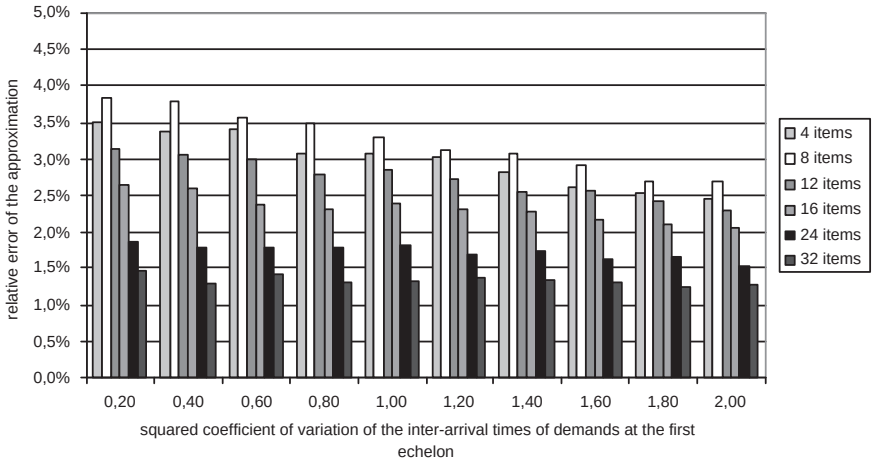

Figure 5: The influence of the number of different items on the performance of the approximation of $E[(C_m^*)^2]$

It can be seen that the average relative error is quite small for all examples and decreases with increasing squared coefficient of variation of the interarrival time of the demands at the first echelon. We can further observe that after a short increase the error is decreasing with increasing number of items.

In Figure 6 the influence of the demand size variability at the first echelon is illustrated. Different values of the coefficient of variation lead to different average errors, but we can see, that the differences are quite small and that other parameters have a larger impact on the performance, such as the number of different items.

In order to investigate the influence of the assumption that the waiting time due to a lack of stock is negligible when computing the variance of C_k, we have computed the second moment $E[(C_m^*)^2]$ of the interarrival time of replenishment orders at the consolidation dock for different target service levels $\beta_k^{target} = betaW, k \in \mathcal{E}_2$. We have chosen a small number of items (4) to have a high probability of a stockout situation and we study the

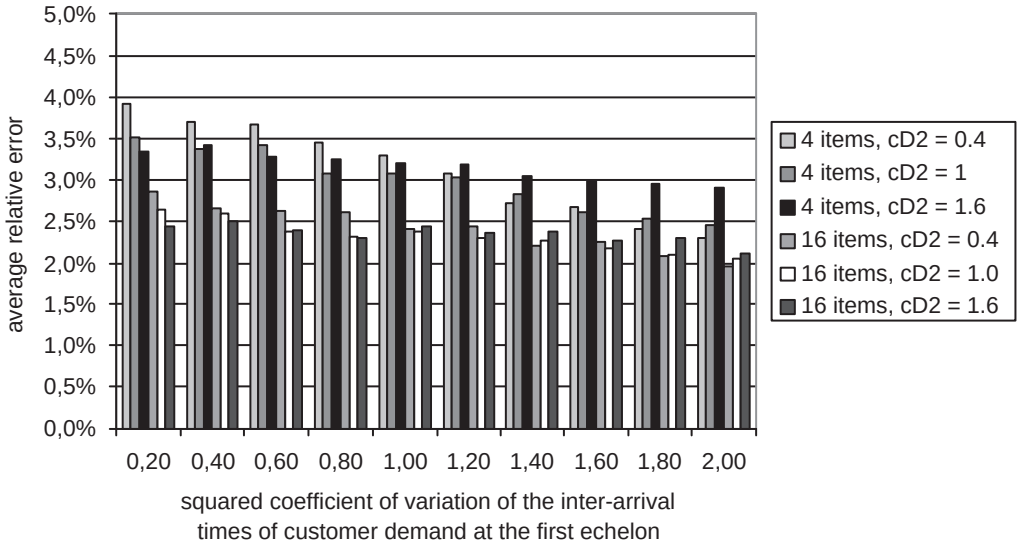

Figure 6: The influence of the demand size variation on the performance of the approximation of $E[(C_m^*)^2]$

performance of the approximations as a function of the target service level at the second echelon $betaW$. The results are illustrated in Figure 7.

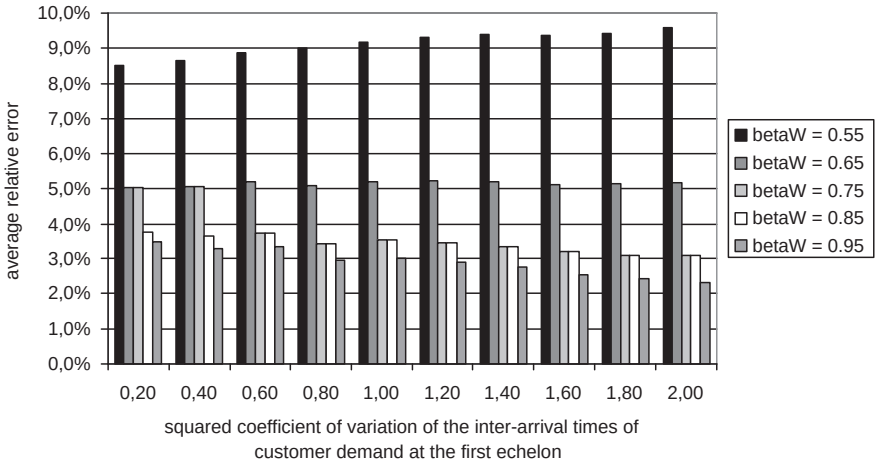

Figure 7: The influence of the target service level at the upper echelon on the performance of the approximation of $E[(C_m^*)^2]$

It can be seen that for smaller target service levels at the upper echelon larg errors of the approximation can be observed due to larger waiting times due to a lack of stock. But for large target service levels at the upper echelon the waiting time due to a lack of stock can be neglected.

6.2 The waiting time due to shipment consolidation

We are interested in the influence of the shipment consolidation process on the leadtime, respectively on the waiting time due to shipment consolidation. Although formulae (21) and (22) are exact, formulae (19) and (20) only provide approximations since the moments of the interarrival times of the replenishment orders at the consolidation dock are approximations. In the following we study the impact of these approximations on the moments of the waiting time due to shipment consolidation.

We can observe a similar behavior as for the approximations for the interarrival time for the replenishment orders at the consolidation dock. With increasing number of different items the average relative errors are decreasing and the impact of the variability of the demand process is quite small as well as the impact of the target service level at the second echelon (see Table 2 in Appendix E).

6.3 The performance of the distribution network

In the last part of the numerical study we investigate the performance of the distribution network when the reorder levels are computed with our algorithm as given in section 5. Moreover, we compute the actual service levels observed by the customers and the average inventory levels with the formulae (8),(23), and (24) and compare our results with results obtained by simulation. Note that due to the discrete demand distribution at the upper echelon we have to find

$$s_k^{min} := min\{s_k \mid \beta_k(s_k) \geq \beta_k^{target}, \quad k \in \mathcal{E}_2\} \tag{28}$$

to determine the reorder level for stockpoint $k \in \mathcal{M}\backslash\mathcal{E}_1$. We compare the computed service level $\beta_k(s_k^{min})$ and the service level obtained by simulation and measure the absolute error by

$$\delta_k^{(\beta_k)} := |\beta_k^{simu} - \beta_k(s_k^{min})| \cdot 100\% \qquad k \in \mathcal{M}\backslash\mathcal{E}_1 \tag{29}$$

and

$$\delta_k^{(\beta_k)} := |\beta_k^{simu} - \beta_k^{target}| \cdot 100\% \qquad k \in \mathcal{E}_1 \tag{30}$$

$\bar{\delta}^{(\beta_k)}$ and $max(\delta^{(\beta_k)})$ are computed similar to (26) and (27) .

We have considered one leadtime scenario $L_k^d = 2, k \in \mathcal{E}_1$ and $L_k = 4, k \in \mathcal{E}_2$ and two scenarios of the lotsizes and the capacity of the vehicles.

1. $Q_k = 200, k \in \mathcal{E}_1$, $Q_k = 800, k \in \mathcal{E}_2$ and $Q_m^c = 800, m \in \{II, III, IV, V\}$
2. $Q_k = 500, k \in \mathcal{E}_1$, $Q_k = 4000, k \in \mathcal{E}_2$ and $Q_m^c = 2000, m \in \{II, III, IV, V\}$

The other parameters are chosen as given in Table 2. The number of different items at each warehouse are equal and we have computed all possible combinations, which leads in total to 532 examples.

Additionally, we considered another set of 532 examples where customers are non-identical. For the average demand sizes we have drawn values randomly from a uniform distribution

Parameter	for	values
$E[D_k]$	$k \in \mathcal{E}_1$	50
$c^2_{D_k}$	$k \in \mathcal{E}_1$	0.4,1,1.6
$E[A_k]$	$k \in \mathcal{E}_1$	1
$c^2_{A_k}$	$k \in \mathcal{E}_1$	0.4,1,1.6
β_k^{target}	$k \in \mathcal{E}_1$	0.90,0.95, 0.99
β_k^{target}	$k \in \mathcal{E}_2$	0.75, 0.9
$\mathcal{L}_m$	$m \in I, II, III, IV$	4,8,12,16

Table 2: Input parameter values

of the interval (30,70) and for the average arrival times we have chosen randomly from a uniform distribution of the interval (0.5,2.5)

In Figure 8 the average of the absolute deviations of the service level at the first echelon is illustrated. It can be seen that our procedure for allocation the safety stock leads to a performance very closed to the required. Deviations of around 3% are only observed for 4 identical demand streams. We can further observe smaller deviations in case of non-identical demand streams. This is in line with the results obtained in Kiesmüller et al. [15] where it is also illustrated that the approximations perform best in heterogeneous distribution structures, i.e. non-identical demand streams.

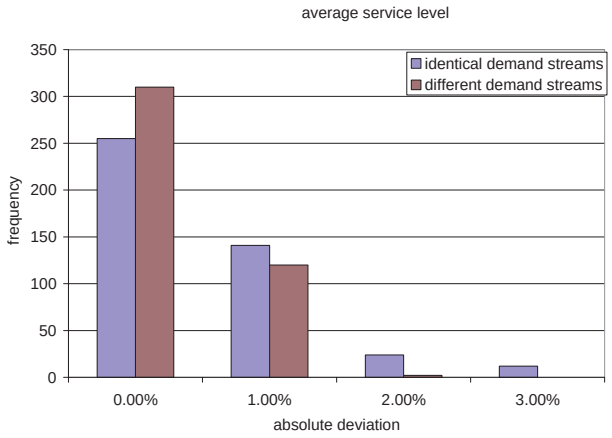

Figure 8: The absolute error of the service level at the first echelon

The performance of the approximations for the average inventory levels are illustrated in Figure 9. On the left hand side a histogram for the relative error of the approximation at the first echelon is given while on the right hand sight the relative errors for the approximations for the second echelon are given. Again, we can observe a very good performance and only quite small deviations between the approximations and the values obtained by simulation.

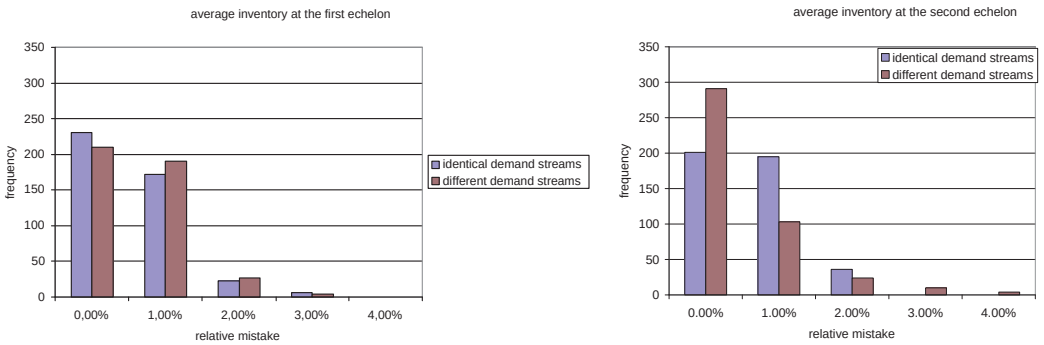

Figure 9: The relative error of the approximation of the inventory levels

7 Summary and conclusions

In this paper we provided approximations for the evaluation of the service performance of a multi-item multi-echelon distribution network under a quantity based shipment policy. Moreover, we have shown how the formulae can be applied to compute reorder levels such that target service levels can be met. We have tested the quality of the approximations and our computational procedure by means of simulation and we have illustrated that the performance of a distribution network, where policy parameters are determined with our algorithm, is closed to the required one. Moreover, the algorithm enables us to analyze the trade-off between inventory levels and transportation frequency. In our future research we focus on managerial insights obtained from such trade-offs. This should deepen the understanding of the trade-offs between transportation cost and holding cost i. e. between efficiency and effectiveness.

References

[1] Axsäter, S., 2003. Supply Chain Operations: Serial and Distribution Inventory. Handbooks in Operations Research and Management Science, Vol. 11, Supply Chain Management: Design, Coordination and Operation. Editors: de Kok, A.G., and Graves, S.C. Elsevier North-Holland.

[2] Axsäter, S., 1993. Continuous review policies for multi-level inventory systems with stochastic demand. Handbooks in Operations Research and Management Science, Vol. 4, Logistics of Production and Inventory. Editors: Graves, S.C., Rinnooy Kan, A.H.G. and Zipkin, P.H. Elsevier North-Holland.

[3] Axsäter, S., 1998. Exact and approximate evaluation of batch-ordering policies for two-level inventory systems. Operations Research 41, 777-785.

[4] Axsäter, S., 2001. A Note on Stock Replenishment and Shipment Scheduling for Vendor-Managed Inventory Systems. Management Science 47/9, 1306-1310.

[5] Bertazzi, L., and Speranza, M.G., 1999. Models and Algorithms for the Minimization of Inventory and Transportation Costs: A survey, Lecture Notes in Economics and Mathematical Systems vol 480, New Trends in Distribution Logistics. Editors: Speranza, M.G. and Staehly, P., eds., Springer-Verlag, Berlin Heidelberg.

[6] Burns, L. D., Hall, R.W., Blumenfeld, D.E. and Daganzo, C.F. 1985. Distribution Strategies that Minimize Transportation and Inventory Costs. Operations Research 33, 469-490.

[7] Cachon, G., 2002. Managing a Retailer's Shelf Space, Inventory, and Transportation. Manufacturing and Service Operations Management 3/3, 211-229.

[8] Çetinkaya, S. and Lee, C.Y., 2002. Optimal Outbound Dispatch Policies: Modeling Inventory and Cargo Capacity. Naval Research Logistics 49, 531-556.

[9] Çetinkaya, S. and Lee, C.Y., 2000. Stock Replenishment and Shipment Scheduling for Vendor-Managed Inventory Systems. Management Science 46/2, 217-232.

[10] Cox, D.R., 1962. Renewal theory. Methuen, London

[11] Diks, E.B., Kok, A.G. de and Lagodimos, A.G., 1996. Multi-echelon systems: a measure perspective. European Journal of Operational Research 95, 241-264.

[12] Federgruen, A., 1993. Centralized planning models for multi-echelon inventory systems under uncertainty, Handbooks in Operations Research and Management Science, Vol. 4, Logistics of Production and Inventory. Editors: Graves, S.C., Rinnooy Kan, A.H.G. and Zipkin, P.H. Elsevier North-Holland.

[13] Higginson, J.K. and Bookbinder, J.H., 1994. Policy Recommendations for a shipment consolidation program. Journal of Business Logistics 15, 87-112.

[14] Houtum, van G.J., Inderfurth, K., and Zijm, W.H.M, 1996. Material coordination in stochastic multi-echelon systems. European Journal of Operational Research 95, 1-23.

[15] Kiesmüller, G.P., de Kok, A.G., Smits, S.R., Van Laarhoven, P.J.M., 2004. Analysis of Divergent N-echelon (s,nQ)-policies Under Compound Renewal Demand, OR Spectrum 26/4, 547-577.

[16] Masters, J.M., 1993. Determination of near optimal stock levels for multi-echelon distribution inventories, Journal of Business Logistics 14/2, 165-195.

[17] Muckstadt, J.A., 1973. A model for a multi-item, multi-echelon, multi-identure inventory sytem, Management Science 20/4, 472-481.

[18] Popken, D.A., 1994. An algorithm for the multiattribute, multicommodity flow problem with freight consolidation and inventory costs. Operations Research 42/2, 274-286.

[19] Ross, S.M., 1993. Introduction to probability models, Fifth edition, Academic Press.

[20] Sherbrooke,C.C., 1968. METRIC: A Multi-Echelon Technique for Recoverable Item Control, Operations Research 16, 122-141.

[21] Silver, E.A., Pyke, D.F., and Peterson, R.,1998. Inventory Management and Production Planning and Scheduling, third edition, Wiley, New York.

[22] Smits, S.R., de Kok, A.G., 2002. Approximations for the waiting time in (s,nQ)-inventory models for different types of consolidation policies. Lecture notes Economics and Mathematical Systems 519, Springer Verlag.

[23] Tijms, H.C., 1994. Stochastic models: an algorithmic approach, John Wiley & Sons.

[24] Whitt, W., 1982. Approximating a point process by a renewal process, I: Two basic methods, Operations Research 30, 125-147.

Appendix A: The first two moments of the time between replenishments R_k

$$E[R_k] = \frac{Q_k}{E[D_k]} E[A_k] \quad \forall k \in \mathcal{M} \tag{31}$$

$$E[R_k^2] = \frac{Q_k}{E[D_k]}\sigma^2[A_k] + E^2[A_k]\Big(\frac{Q_k^2}{E[D_k]^2} + Q_k \frac{c_{D_k}^2}{E[D_k]} + \frac{E[D_k^2]^2}{E[D_k]^4} - \frac{2E[D_k^3]}{3E[D_k]^3}\Big) \quad \forall k \in \mathcal{M} \tag{32}$$

Appendix B: Algorithm to compute the first two moments of a superposed compound renewal process

The interarrival time of the single processes is denoted with X_i and of the superposed process it is denoted with X_0. The following iterative algorithm can be used to compute the first two moments of X_0 (cf. Whitt [24]).

1. Order the N interarrival processes from the largest to the smallest first moment.
2. Compute
$$E[X_0^{(1)}] := \frac{1}{\sum_{i=1}^{2} \frac{1}{E[X_i]}} \tag{33}$$
$$E[(X_0^{(1)})^2] := 2E[X_0^{(1)}] \int_0^\infty \left(\prod_{i=1}^{2} \frac{1}{E[X_i]}\right) \left(\prod_{i=1}^{2} \int_x^\infty (1 - F_{X_i}(y))dy\right) dx \tag{34}$$
3. Fit a mixed-Erlang distribution to the first two moments of $X_0^{(1)}$.
4. Initially set n:=2 and i:=3.
5. Compute
$$E[X_0^{(n)}] := \frac{1}{\frac{1}{E[X_0^{(n-1)}]} + \frac{1}{E[X_i]}} \tag{35}$$
$$E[(X_0^{(n)})^2] := 2E[X_0^{(n)}] \int_0^\infty \frac{1}{E[X_0^{(n-1)}]E[X_i]} \left(\int_x^\infty (1 - F_{X_0^{(n-1)}}(y))dy \int_x^\infty (1 - F_{X_i}(y))dy\right) dx \tag{36}$$
Fit a mixed-Erlang distribution to the first two moments of $X_0^{(n)}$
6. If $n < N$ then $n := n + 1$, $i := i + 1$ and go to step 5 else $E[X_0] := E[X_0^{(N)}]$ and $E[X_0^2] := E[(X_0^{(N)})^2]$

Appendix C: The demand during the leadtime for continuously distributed demand sizes

The moments of the demand during the leadtime can be computed using the following equations:

$$E[D_k(L_k)] = E[N(L_k)]E[D_k], \qquad k \in \mathcal{E}_1 \tag{37}$$

$$E[D_k^2(L_k)] = E[N(L_k)]\sigma^2(D_k) + E[N(L_k)^2]E[D_k]^2, \qquad k \in \mathcal{E}_1 \tag{38}$$

The moments of the number of demand arrivals can be determined using the approximations based on renewal theory.

$$E[N(L_k)] = \frac{E[L_k]}{E[A_k]} + \frac{E[A_k^2]}{2E[A_k]^2} - 1, \qquad k \in \mathcal{E}_1 \tag{39}$$

$$\begin{aligned} E[N(L_k)^2] &= \frac{E[L_k^2]}{E[A_k]^2} + E[L_k]\left(\frac{2E[A_k^2]}{E[A_k]^3} - \frac{3}{E[A_k]}\right) \\ &\quad + \frac{3E[A_k^2]^2}{2E[A_k]^4} - \frac{2E[A_k^3]}{3E[A_k]^3} - \frac{3E[A_k^2]}{2E[A_k]^2} + 1, \qquad k \in \mathcal{E}_1 \end{aligned} \tag{40}$$

The moments for the undershoot are given as

$$E[U_k] = \frac{E[D_k^2]}{2D[D_k]} \qquad E[U_k^2] = \frac{E[D_k^3]}{3E[D_k]}, \qquad k \in \mathcal{E}_1 \tag{41}$$

Appendix D: The probability function of $N(L_k)$

$$\begin{aligned} P(N(L_k) = i \mid L_k = z) &= P(N(z) \geq i) - P(N(z) \geq i+1) \\ &= P\Big(\sum_{j=1}^{i} A_j \leq z\Big) - P\Big(\sum_{j=1}^{i+1} A_j \leq z\Big) \end{aligned} \tag{42}$$

We use the approximation technique described in 3.1 to fit a mixed Erlang distribution on the moments of $\sum_{j=1}^{i} A_j$.

$$P(N(L_k) = i) = \int_0^\infty P(N(L_k) = i \mid L_k = z) f_{L_k}(z)\, dz \tag{43}$$

Appendix E: The average inventory level

Define $H(nQ_j)$ as the expected area between the physical inventory level and the zero level, given that the physical level is nQ_j at time zero, there are no outstanding orders and no orders are placed in the future. Further we consider the situation that a customer arrives at time zero. Then we get

$$H(nQ_j) = E[A_k]Q_j \cdot \frac{n(n+1)}{2}, \quad k \in \mathcal{M}\backslash\mathcal{E}_1, j \in \mathcal{S}_k \tag{44}$$

For the situation that time zero is an arbitrary point in time we get for the expected area between the physical inventory level and the zero level $\tilde{H}(nQ_j)$

$$\tilde{H}(nQ_j) = Q_j(nE[\tilde{A}_k] + \frac{n(n-1)}{2}E[A_k]) \tag{45}$$

where $\tilde{A}_k$ is the residual lifetime of A_k.

Now we consider the average stock on hand per time unit in an arbitrary replenishment cycle, defined as the time between two successive arrivals of replenishment orders. It is given as

$$E[I] = \frac{E[\tilde{H}(I_b)] - E[\tilde{H}(I_e)]}{E[\text{length of a replenishment cycle}]} \tag{46}$$

where I_b (I_e) denotes the stock on hand at the beginning (end) of a replenishment cycle. We get

$$E[I_k] = \frac{E[\tilde{H}(s_k + Q_k - D(L_k))] - E[\tilde{H}(s_k - D(L_k))]}{\frac{Q_k E[A_k]}{Q_j}} \tag{47}$$

If i_{s+Q} is defined as

$$i_{s+Q} := \frac{s_k + Q_k}{Q_j} \tag{48}$$

then we get

$$\begin{aligned} E[\tilde{H}(s_k + Q_k - D(L_k))] &= \sum_{i=0}^{i_{s+Q}} \tilde{H}(i_{s+Q}Q_j - iQ_j)P(N(L_k) = i) \\ &= Q_j \sum_{i=0}^{i_{s+Q}} \Big((i_{s+Q} - i)E[\tilde{A}_k] + \frac{1}{2}(i_{s+Q} - i)(i_{s+Q} - i - 1)\Big)P(N(L_k) = i) \end{aligned} \tag{49}$$

Similarly, we get with $i_s := \frac{s_k}{Q_j}$

$$\begin{aligned} E[\tilde{H}(s_k - D(L_k))] &= \sum_{i=0}^{i_s} \tilde{H}(i_s Q_j - iQ_j)P(N(L_k) = i) \\ &= Q_j \sum_{i=0}^{i_s} \Big((i_s - i)E[\tilde{A}_k] + \frac{1}{2}(i_s - i)(i_s - i - 1)\Big)P(N(L_k) = i) \end{aligned} \tag{50}$$

The formulas (49) and (50) together with (46) proof our formula (24).

Appendix F:

$\beta_k^{target}, k \in \mathcal{E}_2$	items	c_D^2	c_A^2	$max(\delta(E[W^c])$	$\overline{\delta}^{(E[W^c])}$	$max(\delta(E[(W^c)^2])$	$\overline{\delta}^{(E[(W^c)^2])}$
0.75	4	0.4	0.4	1.08	0.44	7.28	6.41
			1.0	0.57	0.19	5.24	4.77
			1.6	0.70	0.30	5.13	3.90
		1.0	0.4	0.76	0.39	5.39	4.84
			1.0	1.01	0.41	5.23	3.86
			1.6	0.65	0.31	4.41	3.17
		1.6	0.4	0.61	0.25	4.62	3.83
			1.0	0.61	0.22	3.70	2.89
			1.6	0.74	0.32	3.18	2.41
	16	0.4	0.4	1.16	0.36	2.97	0.91
			1.0	1.13	0.36	2.68	1.02
			1.6	1.30	0.42	3.10	0.97
		1.0	0.4	1.67	0.44	2.98	0.87
			1.0	1.14	0.36	2.21	0.86
			1.6	1.09	0.34	2.09	0.75
		1.6	0.4	1.23	0.36	2.46	0.79
			1.0	0.97	0.36	1.87	0.68
			1.6	1.31	0.36	2.20	0.68
0.90	4	0.4	0.4	0.91	0.41	7.54	6.75
			1.0	0.53	0.22	6.08	5.26
			1.6	0.66	0.26	4.95	4.39
		1.0	0.4	0.74	0.38	6.42	5.30
			1.0	0.92	0.37	5.64	4.24
			1.6	0.72	0.26	4.86	3.75
		1.6	0.4	0.49	0.25	4.97	4.31
			1.0	1.10	0.24	4.21	3.41
			1.6	0.63	0.34	3.87	2.92
	16	0.4	0.4	1.12	0.43	2.91	1.00
			1.0	1.25	0.41	3.20	1.08
			1.6	1.43	0.49	3.20	1.18
		1.0	0.4	1.56	0.43	2.86	1.03
			1.0	0.94	0.36	2.64	1.10
			1.6	1.21	0.36	2.59	0.95
		1.6	0.4	1.18	0.25	4.97	4.31
			1.0	0.90	0.24	4.21	3.41
			1.6	1.54	0.34	3.87	2.92

Table 3: Performance of the approximations for the first and second moment of the waiting time due to shipment consolidation