Memorandum COSOR 96-33
Modeling stochastic lead times in multi-echelon systems
  
  E.B. Diks
  M.C. van der Heijden

Eindhoven, November 1996
The Netherlands
Modeling stochastic lead times in multi-echelon systems

E.B. Diks
Department of Mathematics and Computing Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: diks@win.tue.nl

M.C. van der Heijden
Department of Technology and Management
University of Twente
P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: m.c.vanderheijden@sms.utwente.nl

Abstract
In many multi-echelon inventory systems the lead times are random variables. A common and reasonable assumption in most models is that replenishment orders do not cross, which implies that successive lead times are correlated. However, the process which generates such lead times is usually not well-defined, which is especially a problem for simulation modeling. In this paper we use results from queuing theory to define a set of simple lead time processes guaranteeing that (a) orders do not cross, and (b) prespecified means and variances of all lead times in the multi-echelon system are attained.

Keywords: stochastic lead time, multi-echelon, simulation, inventory.

1 Introduction

The fact that the lead time in an inventory system is rarely constant is widely recognized. It is important to account for lead time variability when analyzing multi-echelon inventory systems, since ignoring it may lead to a high cost and a poor performance (cf. Gross & Soriano [1969]). An important issue when incorporating stochastic lead times is whether successive lead times are independent, since in that case orders can cross. A cross over occurs when a quantity that was ordered in a latter period arrives before one that was ordered in an earlier period. In practice successive lead times are usually dependent (cf. Yano [1987]), therefore the 'no crossing' assumption needs to be incorporated in the model. Various modeling approaches have been used to circumvent this complication. When the interval between successive orders is large enough, the probability of cross over is negligible and can be omitted (cf. Hadley & Whitin [1963]). Friedman [1984] circumvents the problem by assuming that there is always only one order outstanding. In Sphicas [1982] and Sphicas & Nasri [1984] the problem of crossovers is eliminated by assuming that unit demands are non-interchangeable. It is clear that these approaches are not generally valid.

To our knowledge Kaplan [1970] was the first which explicitly incorporated the no crossing assumption. A mechanism of the arrival of orders is used, to ensure that orders never cross, while in general lead times will be dependent. Kaplan (see also Nahmias [1979] and Anupindi, Morton & Pentico [1996]) developed this mechanism for a periodic review system, which places an order at the beginning of every period. This mechanism is based on a stationary, discrete-time Markov process \( U(t) \) with
nonnegative-integer states, representing the number of outstanding orders after the possible delivery at time $t$, with

$$Pr(U(t + 1) = j | U(t) = i) = p_j, \quad j = 0, 1, \ldots, i.$$  

(1)

Ehrhardt [1984] derives the distribution of the lead time $L$, for this specification, namely

$$Pr(L = i) = \begin{cases} p_0 & i = 0 \\ (1 - p_0)(1 - p_0 - p_1) \cdots (1 - \sum_{j=0}^{i-1} p_j) \sum_{j=0}^{i} p_j & i = 1, 2, \ldots, m, \end{cases}$$

where $m$ denotes the maximum lead time. Like Zipkin [1986], we think it is hard to imagine a physical process giving rise to this scenario. Moreover, many parameters need to be estimated from data, and furthermore, Zipkin has shown that not any lead time distribution (e.g. the geometric distribution) can be attained. Zipkin extended the arrival mechanism of Kaplan [1970] by relaxing (1) (along with the discrete time assumption). However, it is not clear how to choose an order order arrival process that fits certain lead time characteristics (the mean and variance as observed in practice).

Heuts & De Klein [1995] introduced another model to deal with stochastic lead times. They distinguish between a start-up and a follow-up order. For a start-up order holds $Pr(L = l_j) = p_j$, where the discrete lead time $L$ is defined at the points $l_0, l_1, \ldots, l_{M-1}$ (in ascending order). A follow-up order will be delivered simultaneously with the start-up order not yet arrived. However, their model requires a lot of information (i.e., all the possible lead times $l_j$ of the start-up order and its probabilities $p_j$) whereas our model only requires the first two moments of the lead time (and possibly the first-order autocorrelation).

In this paper we develop a simple model for a lead time process in which orders do not cross. We show that the combination of a single server queue and a deterministic pipeline yields a versatile and simple lead time process. We use results from queuing theory to choose the parameters of this process such that the successive throughput times (approximately) match a prespecified lead time mean and variance. Although the approach is general applicable, we focus on divergent networks with a periodic review, order-up-to policy. The results of this paper facilitates building a simulation model of multi-echelon inventory systems, to be used for scenario analysis and/or validation of approximate analytical methods. In fact, we developed this approach for the latter purpose, see Van der Heijden, Diks & De Kok [1996].

The remainder of this paper is structured as follows. First we consider a single location inventory system. Section 2 deals with the model and approach for this situation. The analysis of the model is discussed in Section 3-4. In Section 5 we show how the lead time model can be used in a general multi-echelon setting. In Section 6 the approximations are validated for divergent networks under periodic review by a simulation study. Finally, we give some concluding remarks in Section 7.

2 Single location inventory system

Consider a single location inventory system in which the lead time $L$ is a random variable with mean $\mu_L$ and variance $\sigma_L^2$. Denote by $A_n$ the time period between issuing the $(n - 1)^{th}$ and the $n^{th}$ replenishment order. $A_n$ can be a random variable, e.g. under continuous review, or deterministic, e.g. under periodic review. Since it is assumed that orders cannot cross, the subsequent lead times faced in the inventory system are correlated for most lead time distributions. Therefore it becomes rather cumbersome to model a lead time process for which the mean and variance coincide with the target values $\mu_L$ and $\sigma_L^2$, respectively.
To illustrate this we consider a single location inventory system under periodic review. Figure 1 depicts the arrival times of two subsequent orders. At time $t$ an order is placed at the supplier with lead time $L_1$. This order arrives at $t + L_1$. At time $t + R$ the stockpoint again inspects the inventory position, and places an order with lead time $L_2$. This order arrives at $t + R + L_2$. Since subsequent orders do not cross, there has to hold $R + L_2 \geq L_1$. If the lead time distribution of $L$ is such that this always holds, the successive lead times can be independent. An example of such a distribution is a uniform distribution on $[\alpha, \alpha + \beta]$ with $\alpha \geq 0$ and $0 \leq \beta \leq R$. For the more general case where not always $R + L_2 \geq L_1$, it is obvious that $L_2$ depends on $L_1$. In the remainder of this paper we develop an approach to model the lead time process of interrelated lead times, such that the mean and variance of the lead time equals the predetermined $\mu_L$ and $\sigma^2_L$, respectively.

2.1 Model and approach

The lead time is modeled by the sojourn time in a $G_1/G/1$-queue plus a deterministic pipeline time. The arrival of a customer at this queue corresponds to the stockpoint placing an order, and since the server adopt the FCFS-discipline, customers (i.e. orders) cannot cross. For our convenience we introduce the following notation:

- $\mu_X := \text{The mean of random variable } X.$
- $\sigma_X := \text{The variance of random variable } X.$
- $c_X := \text{The coefficient of variation of a random variable } X, c_X = \sigma_X/\mu_X.$
- $\mu_3[X] := \text{The third central moment of random variable } X, \mu_3[X] = E[X - \mu_X]^3.$
- $F_X := \text{The cumulative distribution function of a random variable } X.$
- $\lambda := \text{The arrival rate at the queuing system.}$
- $B := \text{The service time of a customer in the queuing system.}$
- $W := \text{The waiting time of a customer in the queuing system.}$
- $S := \text{The sojourn time of a customer in the queuing system, } S = W + B.$

The lead time is not modeled solely by a queuing system for the following reason. In practice the mean lead time can be large in contrast with its small variance. In order to obtain a lead time process for which the sojourn time has a large mean and a small variance, a very high utilization (> 0.999) is required. So, using the sojourn time of the $G_1/G/1$-queue solely is inadequate. Therefore we suggest to model the lead time as the sum of the sojourn time in a $G_1/G/1$-queue plus a fixed time in a pipeline (see Figure 2).

![Figure 2: Model of the lead time process.](image-url)
By setting this fixed time properly we avoid lead time processes with an extremely high utilization. Thus, the lead time \( L \) is given by

\[
L := l + S,
\]

where \( l \) denotes a fixed nonnegative time and the random variable \( S \) denotes the stochastic part of the lead time (in general it depends on preceding lead times). As one possible interpretation one might think of \( l \) as a fixed handling & transportation time, and \( S \) as the production time which is subjected to capacity constraints.

The sojourn time of the GI/G/1-queue plus the deterministic pipeline time has mean \( \mu_W + \mu_B + l \) and variance \( \sigma_B^2 + \sigma_W^2 \). So, we have to find the parameters of the lead time process, such that \( \mu_L = \mu_W + \mu_B + l \) and \( \sigma_L^2 = \sigma_B^2 + \sigma_W^2 \) for given \( \mu_L \) and \( \sigma_L^2 \). The process by which replenishment orders are generated is given, so the remaining degrees of freedom are \( l \) and the service distribution \( B \). Using only the first two moments of \( B \) to analyze the GI/G/1-queue, this means that we have to find the right values for \( l, \mu_B \) and \( \sigma_B^2 \). So we have one excess degree of freedom. We deal with this as follows. First, we show how to determine the values of \( \mu_B \) and \( \sigma_B^2 \) for a given value of \( l \) (Section 3). Next, we discuss a proper choice of \( l \) (Section 4).

### 3 Determination of \( \mu_B \) and \( \sigma_B^2 \)

The objective of many papers in the queuing literature is to determine performance characteristics (e.g., mean and variance of sojourn time) given information regarding the arrival process and the service process. In this section we have to deal with the 'inverse' problem: determine the service process characteristics given information regarding the arrival process, and the mean and variance of the sojourn time. Specifically, determine \( \mu_B \) and \( \sigma_B^2 \) such that \( \mu_B + \mu_W = \mu_L - l \) and \( \sigma_B^2 + \sigma_W^2 = \sigma_L^2 \). To solve this problem we developed an algorithm which uses some techniques developed in the literature. First, we briefly address these queuing techniques in Section 3.1. Next, we discuss a nested bisection algorithm to determine \( \mu_B \) and \( \sigma_B^2 \).

#### 3.1 Queuing approximations

In this section we address several techniques which determine \( \mu_W \) and \( \sigma_W^2 \) from \( \mu_B \) and \( \sigma_B^2 \). We distinguish between the \( D/G/1 \)-queue (Section 3.1.1) and the GI/G/1-queue (Section 3.1.2), since for the former queue good approximations have been developed, and since it represents a specific case namely the periodic review.

##### 3.1.1 \( D/G/1 \)-model

Fredericks [1982] and Bhat [1993] (F&B)

In Fredericks [1982] an excellent approximation is given for the mean waiting time, denoted by \( \mu_W \), in a \( D/G/1 \)-queue.

\[
\mu_W = \frac{\int_{R}^{\infty} (t - R) dF_B(t)}{F_B(R) - 1 + \int_{R}^{\infty} e^{\delta(R-y)} dF_B(t)},
\]

where \( \delta \) is the unique solution to the equation

\[
e^{-\delta R} \int_{0}^{\infty} e^{\delta y} dF_B(y) = 1.
\]
In Whitt [1982] also several diffusion approximations are given for $\mu_W$ in case of heavy traffic (say $\rho = \lambda \mu_B > 0.95$). Bhat [1993] gives an approximation for $\sigma^2_w$:

$$\sigma^2_w = z_1^2 + \frac{z_2}{3(1 - \rho)}, \quad \text{where} \quad z_1 := \frac{\lambda \sigma^2_B}{2(1 - \rho)}, \quad z_2 := \lambda \mu_3[B]. \quad (3)$$

$\mu_3[B]$ in (3) is determined by first fitting a mixture of two Erlang distributions to the first two moments of $B$ (cf. Tijms [1986]), and next computing its third moment.

De Kok [1989]
Let, for $n = 0, 1, 2, \ldots$, the random variables $B_n$ and $W_n$ denote the service time and the waiting time of the $n$th customer, respectively. Further, suppose that the $n$th customer arrives at an empty system, so that $W_0 = 0$. Then, it is easily seen that the following relation holds

$$W_n = (W_{n-1} + B_{n-1} - R)^+, \quad n = 1, 2, \ldots, \quad (4)$$

where $x^+ := \max(0, x)$. The moment-iteration method developed by De Kok [1989] determines the distribution of $W_n$ by an iterative procedure. It yields excellent results. Only for the case of heavy traffic ($\rho > 0.95$) we suggest the use of the F&B approach.

In case of discrete lead times one might use the simple method of Adan, Van Eenige & Resing [1995] which fits a discrete distribution to $B$. Next, they used the moment-iteration method for the $D/G/1$-queue to obtain $\mu_W$ and $\sigma^2_w$. Numerical results show excellent performance of the method.

3.1.2 GI/G/1-model
As a $GI/G/1$ analogue of the F&B-approach of Section 3.1.1 we suggest to use the approximation of Krämer & Langenbach-Belz [1978] to approximate $\mu_W$ and again approximate $\sigma^2_w$ by the formula of Bhat [1993] ($z_1$ and $z_2$ of (3) need to be adapted). As mentioned by Tijms [1986] the approximation of Krämer & Langenbach-Belz [1978] is a very useful and quick estimate when $c_A$ is not too large. Also we could use the moment-iteration method of De Kok [1989] for the $GI/G/1$-queue to determine $\mu_W$ and $\sigma^2_w$ (eq. (4) needs to be adapted).

3.2 Algorithm
Without loss of generality it is assumed that the review period $R$ equals 1, i.e., $\lambda = 1$. By applying either the F&B-approximations or the moment-iteration algorithm of De Kok [1989] we are able to compute $\mu_S(\mu_B, \sigma^2_B)$ and $\sigma^2_S(\mu_B, \sigma^2_B)$, denoting the mean and variance of the sojourn time given $\mu_B$ and $\sigma^2_B$, respectively. In order to explain our algorithm (which determines $\mu_B$ and $\sigma^2_B$ such that $\mu_S = \mu_L - 1$ and $\sigma^2_S = \sigma^2_L$) we depicted the behavior of both functions in Figure 3 (using the moment-iteration method).

Figure 3a shows for which $(\mu_B, \sigma^2_B)$ a certain mean sojourn time $\mu_S$ is achieved. It indicates that when $\mu_B$ decreases, $\sigma^2_B$ at first increases linearly, but after a certain point increases more than linear. Figure 3b depicts the behavior of the variance $\sigma^2_S$ for all the points depicted in Figure 3a.

Let us explain the algorithm by considering an example. Suppose $\mu_S = 1$ and $\sigma^2_S = 1$. First, we choose a mean service time, say 0.4, and determine by bisection how to set $\sigma^2_B$ such that $\mu_S = 1$. In Figure 3 this point is depicted by $b1$. Since $\sigma^2_S$ in $b1$ exceeds 1 we decide to increase $\mu_B$, say to 0.7. For this service time again the $\sigma^2_B$ is determined such that $\mu_S = 1$. This yields point $b2$. However, since $\sigma^2_S$ in $b2$ is too small now decrease $\mu_B$. Proceed this procedure till the point $(\mu_B, \sigma^2_B)$ for which $\mu_S = 1$ and $\sigma^2_S = 1$ is found. This nested bisection procedure is formalized in Appendix A.
Figure 3: The behavior of $\sigma_B^2$ (fig. a) and $\sigma_S^2$ (fig. b) in a $D/G/1$-queue as a function of $\mu_B$.

4 Determination of $l$

As described in Section 2.1 it is important to choose a proper $l$. In this section we give two methods to split the lead time into the fixed part $I$ and random part $S$. Again we distinguish between the $D/G/1$-model (Section 4.1) and the $GI/G/1$-model (Section 4.2).

4.1 $D/G/1$-model

The first method, assumes that there is some knowledge on the correlation of successive lead times. Specifically, this method determines how $l$ should be set such that the first-order autocorrelation equals $\rho_L$. Using (2) it follows

$$\rho_L = \frac{\text{cov}(S_1, S_2)}{\sigma_S^2},$$

where $S_1$ and $S_2$ are two successive sojourn times of the $D/G/1$-queue. By conditioning on $S_1$ it can be shown that

$$\text{cov}(S_1, S_2) = \sigma_S^2 - (R - \mu_B)\mu_S + \int_0^R (Rx - x^2)dF_S(x).$$

Substituting (6) in (5) yields

$$\rho_L = 1 - \frac{(R - \mu_B)\mu_S - \int_0^R (Rx - x^2)dF_S(x)}{\sigma_S^2}.$$  

By fitting a suitable distribution on $S$ we are able to actually compute $\rho_L$. As described in Tijms [1986] we fit a mixture of two Erlang distributions on its mean and variance.

From (7) it follows that when $\sigma_S^2$ is fixed and $\mu_S$ converges to 0, then $\rho_L$ eventually converges to 1. This is intuitively clear. From $\mu_B < \mu_S$, it follows that when $\mu_S$ converges to 0 also $\mu_B$ converges to 0. In order to still meet the fixed variance $\sigma_S^2$, the $\sigma_B^2$ increases considerably when $\mu_B$ converges to 0. So, for a small $\mu_S$ the $c_B$ is very large. This results in a 'lumpy' service time distribution, i.e., most of the time $B$ is small, but on rare occasions $B$ is huge. Hence the sojourn time of a customer becomes highly correlated with the sojourn time of its preceding customer.
On the other hand when \( \sigma_S^2 \) is fixed and \( \mu_S \) is very large (with respect to \( \sigma_S^2 \)), then \( \rho_L \) is also large. Also this is intuitively clear. When \( \mu_S \) is large compared to \( \sigma_S^2 \) the \( c_S \) is small. Hence the behavior of \( S \) is more or less deterministic.

Figure 4 depicts the behavior of \( \rho_L \) as a function of \( \mu_S \) in a periodic review system with \( R = 1 \).

![Figure 4: The behavior of \( \rho_L \) as a function of \( \mu_S \), when \( R = 1 \).](image)

we expected, this figure shows that \( \rho_L \) is approximately one for a small \( \mu_S \) and large for \( \mu_S \) sufficiently large. Furthermore, this figure shows that not any \( \rho_L \) can be obtained for given \( \mu_L \) and \( \sigma_L^2 \). The smaller \( \sigma_L^2 \) is the larger the range of \( \rho_L \) becomes. For some values of \( \rho_L \) there even are two ways of splitting the lead time. In advance it is not clear which of these two possibilities yields the best way of splitting the lead time. Finally, we note that \( \rho_L \) is independent of \( \mu_L \), but only depends on \( \mu_S (\leq \mu_L) \) and \( \sigma_L^2 \).

The second method which yields a suitable way of splitting \( L \) into the fixed \( I \) and the random \( S \), is just by taking an exponential distribution for \( B \). Then from queuing theory it follows that the sojourn time in a \( D/M/1 \)-queue is exponentially distributed as well. So \( c_S = 1 \). After some elementary algebra it follows

\[
l := (1 - c_L) \mu_L, \quad c_L \leq 1.
\]

Figure 3a shows indeed that all points \((\mu_B, \sigma_B^2)\) with \( c_S = 1 \) are on the line for which \( c_B = 1 \). So, by using the splitting rule of (8) the service process easily follows from the analysis of the \( D/M/1 \)-queue: \( B \) is exponentially distributed with

\[
\mu_B = \mu_S (1 - e^{-\frac{1}{\mu_S}}).
\]

4.2 GI/G/1-model

As in Section 4.1 we address the same two methods of splitting \( L \). For applying the first method we need to have an expression for the covariance of two successive sojourn times. Unlike the \( D/G/1 \)-queue, the variability of the arrival process of the \( GI/G/1 \)-queue has to be taken into account. So,
equation (6) slightly changes.

\[
\text{cov}(S_1, S_2) = \sigma_S^2 - (\mu_A - \mu_B) \mu_S + \mu_S^2 + \int_0^\infty \int_0^\infty x(1 - F_A(r + x)) dr \ dF_S(x).
\]  

(10)

By a fitting a mixture of two Erlang distributions on the mean and variance of \( A \) we can obtain a more tractable expression for (10). This means that the interarrival time \( A \) follows an \( E_{k_1, \theta_1} \) distribution with probability \( \alpha \), and an \( E_{k_2, \theta_1} \) distribution with probability \( 1 - \alpha \). The parameters \( \alpha, k_1, k_2, \theta_1 \) and \( \theta_2 \) can easily be obtained from \( \mu_A \) and \( \sigma_A^2 \). For more details regarding the fitting procedure we refer to Tijms [1986]. After some algebra (10) is elaborated as

\[
\text{cov}(S_1, S_2) = \sigma_S^2 - (\mu_A - \mu_B) \mu_S + \mu_S^2 + \sum_{i=1}^2 \frac{\alpha_i}{\theta_i} \sum_{j=0}^{k_i-1} \sum_{l=0}^{i-j} \frac{\theta_i^j}{j!} \int_0^\infty x^{l+1} e^{-\theta_i x} dF_S(x).
\]

(11)

Also on \( S \) we fit a mixture of two Erlang distributions so as to be able to compute \( \rho_L \).

Again, the second method can be applied. Similar results hold, since the distribution of the sojourn time in a \( G1/M/1 \)-queue is exponentially distributed. So the service time \( B \) is exponentially distributed with

\[
\mu_B = \mu_S \left( 1 - A^* \left( \frac{1}{\mu_S} \right) \right),
\]

(12)

where \( A^*(s) \) denotes the Laplace-Stieltjes transform of \( A \). Notice that for the aforementioned mixture of two Erlang distributions holds

\[
A^*(s) = \alpha \left( \frac{\theta_1}{\theta_1 + s} \right)^{k_1} + (1 - \alpha) \left( \frac{\theta_2}{\theta_2 + s} \right)^{k_2}.
\]

5 Multi-echelon systems

In this section we extend the results of Section 2 to arbitrary large networks of stockpoints. We distinguish between push systems (Section 5.1) and pull systems (Section 5.2). With 'push' we mean that each stockpoint ships replenishment orders to its successors immediately upon arrival of a shipment, whereas 'pull' reflects that the receiving stockpoint issues replenishment orders.

5.1 Push systems

In a convergent system every stockpoint has a unique successor (but may have several predecessors), whereas in a divergent system every stockpoint has a unique predecessor (but may have several successors). Several papers concerning divergent systems use a push mechanism (see Van der Heijden, Diks & De Kok [1996]). Upon arrival of an order at a stockpoint, the products are shipped toward its successors. The \( D/G/1 \)-analysis performed in Section 3–4 applies to the most upstream stockpoint if a periodic review policy is used. When the upstream stockpoint uses a continuous review policy the \( GI/G/1 \)-model may be applied. For a downstream stockpoint holds that the arrival process is the departure process of its upstream stockpoint. Hence, the interarrival times generally will be correlated. As an approximation we suggest to model the lead time process in a downstream stockpoint by the sojourn time in a \( GI/G/1 \)-queue plus a fixed handling & transportation time. So we disregard the dependency between the interarrival times. In Section 6 this assumption is verified by simulation.

For our convenience we describe how to model the lead time process of the downstream stockpoint in a two-stage serial system (see Figure 5). The extension to larger serial systems, or even divergent systems...
systems, is straightforward since for all cases the arrival process of the $GI/G/1$-queue equals the departure process of its predecessor. We use a two moment approximation for the interdeparture time process. It is clear that the mean interdeparture time equals the mean interarrival time at stockpoint 1. The variance of the interdeparture time of stockpoint 1 can be determined from the three approximations for the squared coefficient of variation of the interdeparture times of a $GI/G/1$-queue given by Buzacott & Shanthikumar [1993].

For convergent systems the push mechanism is hardly used, since it is mainly used to model an assembly system. Then material coordination is required, since it does not make sense to ship parts to its successor if not all the parts are available. Hence a push mechanism is not appropriate. Moreover, in a push mechanism the arrival process of a stockpoint equals the superposition of all the departure processes of its predecessors. It is doubtful whether this arrival process can be described appropriately by independent arrivals where its interarrival times are only based on the first two moments. Hence, for these systems probably an alternative model of the lead times is required.

5.2 Pull systems

In pull systems a stockpoint typically places an order at all its upstream stockpoints when its inventory position drops below a certain reorder-point. When one of these upstream stockpoints is not able to satisfy the order immediately, then (1) this order is lost, or (2) it is backordered and delivered as soon as possible. In the first case some lead times may represent an order of amount zero. Like Kaplan [1970] we adopt the convention that such an order can be treated as a pseudo order or infinitesimal amount, but to be considered as outstanding until the pseudo order arrives. Hence, for every stockpoint the $D/G/1$-model may be applied in case of periodic review, and the $GI/G/1$-model may be used as an approximation in case of continuous review. A subject for further research is to determine for which inventory policies this yields satisfactory results. However, for the second case the lead time process of a stockpoint also depends on the waiting times at its upstream stockpoints. In that case it is hard to characterize the arrival process at this stockpoint.

6 Numerical results

Let us first address the numerical results with respect to a single inventory system (Section 6.1). Next, we examine the approximation of assuming independent arrivals of orders at downstream stockpoints (see Section 6.2).

6.1 Single location inventory system

Consider a single location inventory system. Since splitting rule (8) yields exact results, there is no need for testing. For the case where the lead time is split so as to attain a predetermined first-order autocorrelation $p_L$, we examined several instances of a single inventory system under periodic review. Again, without loss of generality $R := 1$.

Table 1 shows the results when the target mean lead time $\mu_L$ equals 2. We considered three different $\sigma^2_L$. For $\sigma^2_L = 0.25$ we require $\rho \in [0.5, 0.65]$, for $\sigma^2_L = 0.5$ we require $\rho_L \in [0.7, 0.8]$, and for $\sigma^2_L = 1$ we require $\rho_L = 0.8$. For every instance we determined both possibilities (see Figure 4) of attaining $p_L$, denoted by 1 and 2. Furthermore, we compare the F&B-technique and the moment-iteration method.
<table>
<thead>
<tr>
<th>Policy</th>
<th>Possibility</th>
<th>$l$</th>
<th>$\mu_L$</th>
<th>$\hat{\mu}_L$</th>
<th>$\sigma^2_L$</th>
<th>$\hat{\sigma}_L^2$</th>
<th>$\rho_L$</th>
<th>$\hat{\rho}_L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F&amp;B</td>
<td>1</td>
<td>1.744</td>
<td>2.000</td>
<td>2.000</td>
<td>0.250</td>
<td>0.192</td>
<td>0.50</td>
<td>0.38</td>
</tr>
<tr>
<td>F&amp;B</td>
<td>2</td>
<td>1.315</td>
<td>2.000</td>
<td>2.002</td>
<td>0.250</td>
<td>0.208</td>
<td>0.50</td>
<td>0.42</td>
</tr>
<tr>
<td>Kok</td>
<td>1</td>
<td>1.764</td>
<td>2.000</td>
<td>1.996</td>
<td>0.250</td>
<td>0.243</td>
<td>0.50</td>
<td>0.47</td>
</tr>
<tr>
<td>Kok</td>
<td>2</td>
<td>1.215</td>
<td>2.000</td>
<td>2.004</td>
<td>0.250</td>
<td>0.260</td>
<td>0.50</td>
<td>0.52</td>
</tr>
<tr>
<td>F&amp;B</td>
<td>1</td>
<td>1.863</td>
<td>2.000</td>
<td>2.002</td>
<td>0.250</td>
<td>0.195</td>
<td>0.65</td>
<td>0.53</td>
</tr>
<tr>
<td>F&amp;B</td>
<td>2</td>
<td>1.060</td>
<td>2.000</td>
<td>2.010</td>
<td>0.250</td>
<td>0.224</td>
<td>0.65</td>
<td>0.62</td>
</tr>
<tr>
<td>Kok</td>
<td>1</td>
<td>1.863</td>
<td>2.000</td>
<td>1.997</td>
<td>0.250</td>
<td>0.218</td>
<td>0.65</td>
<td>0.58</td>
</tr>
<tr>
<td>Kok</td>
<td>2</td>
<td>0.967</td>
<td>2.000</td>
<td>2.018</td>
<td>0.250</td>
<td>0.297</td>
<td>0.65</td>
<td>0.70</td>
</tr>
<tr>
<td>F&amp;B</td>
<td>1</td>
<td>1.742</td>
<td>2.000</td>
<td>2.000</td>
<td>0.500</td>
<td>0.402</td>
<td>0.70</td>
<td>0.60</td>
</tr>
<tr>
<td>F&amp;B</td>
<td>2</td>
<td>1.039</td>
<td>2.000</td>
<td>2.005</td>
<td>0.500</td>
<td>0.428</td>
<td>0.70</td>
<td>0.66</td>
</tr>
<tr>
<td>Kok</td>
<td>1</td>
<td>1.742</td>
<td>2.000</td>
<td>1.995</td>
<td>0.500</td>
<td>0.439</td>
<td>0.70</td>
<td>0.63</td>
</tr>
<tr>
<td>Kok</td>
<td>2</td>
<td>0.945</td>
<td>2.000</td>
<td>2.010</td>
<td>0.500</td>
<td>0.543</td>
<td>0.70</td>
<td>0.72</td>
</tr>
<tr>
<td>F&amp;B</td>
<td>1</td>
<td>1.853</td>
<td>2.000</td>
<td>1.999</td>
<td>0.500</td>
<td>0.422</td>
<td>0.80</td>
<td>0.73</td>
</tr>
<tr>
<td>F&amp;B</td>
<td>2</td>
<td>0.841</td>
<td>2.000</td>
<td>2.017</td>
<td>0.500</td>
<td>0.468</td>
<td>0.80</td>
<td>0.78</td>
</tr>
<tr>
<td>Kok</td>
<td>1</td>
<td>1.853</td>
<td>2.000</td>
<td>1.993</td>
<td>0.500</td>
<td>0.368</td>
<td>0.80</td>
<td>0.71</td>
</tr>
<tr>
<td>Kok</td>
<td>2</td>
<td>0.675</td>
<td>2.000</td>
<td>2.032</td>
<td>0.500</td>
<td>0.662</td>
<td>0.80</td>
<td>0.86</td>
</tr>
<tr>
<td>F&amp;B</td>
<td>1</td>
<td>1.673</td>
<td>2.000</td>
<td>2.000</td>
<td>1.000</td>
<td>0.825</td>
<td>0.80</td>
<td>0.73</td>
</tr>
<tr>
<td>F&amp;B</td>
<td>2</td>
<td>0.790</td>
<td>2.000</td>
<td>2.010</td>
<td>1.000</td>
<td>0.902</td>
<td>0.80</td>
<td>0.79</td>
</tr>
<tr>
<td>Kok</td>
<td>1</td>
<td>1.673</td>
<td>2.000</td>
<td>1.992</td>
<td>1.000</td>
<td>0.780</td>
<td>0.80</td>
<td>0.72</td>
</tr>
<tr>
<td>Kok</td>
<td>2</td>
<td>0.669</td>
<td>2.000</td>
<td>2.021</td>
<td>1.000</td>
<td>1.157</td>
<td>0.80</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Table 1: The (simulated) mean, variance and correlation of the lead time observed at a single stockpoint, using the F&B en Kok-technique.
of De Kok [1989]. In Table 1 $\hat{\mu}_L$, $\hat{\sigma}_L$, and $\hat{\rho}_L$ denote the simulated values. From Table 1 it is clear that for 8 out of the 20 cases the results are satisfactory (i.e., $\hat{\rho}_L$ deviates from $\rho_L$ maximally by 0.04). However, for the worst 7 cases the results are very poor (i.e., $\hat{\rho}_L$ deviates from $\rho_L$ between 0.08 and 0.12). This may be explained as follows. The $c_L^2$ for these 7 cases vary from 3.2 to 32.7, therefore both the F&B and the moment-iteration method does not approximate the variance of the sojourn time accurately. Notice that for the cases where $\hat{\sigma}_L^2$ almost coincides with $\sigma_L^2$, also $\hat{\rho}_L$ almost coincides with $\rho_L$. From these results we conclude the performance of the suggested model is moderate. Hence, we recommend to concentrate on the first two moments of the lead time, and use splitting rule (8).

6.2 Two-stage serial system
Consider the two-serial system of Figure 5. Suppose splitting rule (8) is used in both stockpoints. Stockpoint 1 inspects the inventory at the start of every period, i.e., $R := 1$, and upon arrival of such order stockpoint 2 inspects its inventory position and places an order. For stockpoint 1 holds $\mu_L \in \{0.5, 0.75, 1, 1.5, 2, 3\}$, and $c_L \in \{0.25, 0.5, 0.75\}$. For stockpoint 2 holds $\mu_L = 1$ and $\sigma_L^2 = 0.50$. We only varied the lead time parameters of stockpoint 1, since these influence the behavior of the departure process of stockpoint 1, and hence the arrival process of stockpoint 2. To examine the impact of neglecting the dependency of arrivals at stockpoint 2 we simulated $\hat{\mu}_L$ and $\hat{\sigma}_L$ of stockpoint 2 for the aforementioned 18 instances. This is done for the three approximations derived in Buzacott & Shanthikumar [1993] (see Table 2). For every instance we simulated 400,000 periods. Approximation 1

<table>
<thead>
<tr>
<th>Stockpoint 1 $\mu_L$</th>
<th>Approx. 1 $\hat{\mu}_L$</th>
<th>Approx. 1 $\hat{\sigma}_L^2$</th>
<th>Approx. 2 $\hat{\mu}_L$</th>
<th>Approx. 2 $\hat{\sigma}_L^2$</th>
<th>Approx. 3 $\hat{\mu}_L$</th>
<th>Approx. 3 $\hat{\sigma}_L^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.50</td>
<td>0.50</td>
<td>0.25</td>
<td>0.25</td>
<td>0.125</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>0.25</td>
<td>0.50</td>
<td>0.245</td>
<td>0.245</td>
<td>0.187</td>
<td>0.989</td>
<td>0.991</td>
</tr>
<tr>
<td>0.75</td>
<td>0.50</td>
<td>0.349</td>
<td>0.349</td>
<td>0.173</td>
<td>0.981</td>
<td>0.987</td>
</tr>
<tr>
<td>0.75</td>
<td>0.50</td>
<td>0.187</td>
<td>0.187</td>
<td>0.349</td>
<td>0.994</td>
<td>0.994</td>
</tr>
<tr>
<td>0.25</td>
<td>0.50</td>
<td>0.467</td>
<td>0.467</td>
<td>0.349</td>
<td>0.973</td>
<td>0.981</td>
</tr>
<tr>
<td>0.75</td>
<td>0.50</td>
<td>0.425</td>
<td>0.425</td>
<td>0.349</td>
<td>0.975</td>
<td>0.975</td>
</tr>
<tr>
<td>1.00</td>
<td>0.50</td>
<td>0.432</td>
<td>0.432</td>
<td>0.552</td>
<td>0.972</td>
<td>0.979</td>
</tr>
<tr>
<td>0.25</td>
<td>0.50</td>
<td>0.349</td>
<td>0.349</td>
<td>0.552</td>
<td>0.982</td>
<td>0.986</td>
</tr>
<tr>
<td>0.75</td>
<td>0.50</td>
<td>0.552</td>
<td>0.552</td>
<td>0.349</td>
<td>0.972</td>
<td>0.979</td>
</tr>
<tr>
<td>1.50</td>
<td>0.50</td>
<td>0.552</td>
<td>0.552</td>
<td>0.552</td>
<td>0.972</td>
<td>0.979</td>
</tr>
<tr>
<td>0.25</td>
<td>0.50</td>
<td>0.632</td>
<td>0.632</td>
<td>0.552</td>
<td>0.970</td>
<td>0.977</td>
</tr>
<tr>
<td>0.75</td>
<td>0.50</td>
<td>0.730</td>
<td>0.730</td>
<td>0.632</td>
<td>0.968</td>
<td>0.973</td>
</tr>
<tr>
<td>2.00</td>
<td>0.50</td>
<td>0.552</td>
<td>0.552</td>
<td>0.730</td>
<td>0.971</td>
<td>0.977</td>
</tr>
<tr>
<td>0.25</td>
<td>0.50</td>
<td>0.730</td>
<td>0.730</td>
<td>0.552</td>
<td>0.968</td>
<td>0.974</td>
</tr>
<tr>
<td>3.00</td>
<td>0.50</td>
<td>0.730</td>
<td>0.730</td>
<td>0.730</td>
<td>0.972</td>
<td>0.974</td>
</tr>
<tr>
<td>0.25</td>
<td>0.50</td>
<td>0.850</td>
<td>0.850</td>
<td>0.730</td>
<td>0.966</td>
<td>0.973</td>
</tr>
<tr>
<td>0.75</td>
<td>0.50</td>
<td>0.897</td>
<td>0.897</td>
<td>0.850</td>
<td>0.974</td>
<td>0.977</td>
</tr>
<tr>
<td>6.00</td>
<td>0.50</td>
<td>0.730</td>
<td>0.730</td>
<td>0.897</td>
<td>0.974</td>
<td>0.973</td>
</tr>
</tbody>
</table>

Table 2: The (simulated) mean and variance of the lead time observed at stockpoint 2 ($\mu_L = 1$ and $\sigma_L^2 = 0.5$), for the three approximations of Buzacott & Shanthikumar [1993].
and 2 results in lead times for which both the mean and the variance are too low, while approximation 3 generally results in lead times for which both the mean and the variance are too high. It turns out that approximation 1 is always worse than approximation 2. Furthermore, the deviation of the mean lead time of approximation 2 and 3 are almost equal. However, approximation 3 results in a much better performance of the lead time variance. Hence, from this experiment we suggest to use approximation 3. As can be seen from Table 2 the performance of the approximations depend on the utilization, \( \rho \) say, of the queue of stockpoint 1. When \( \rho \) is small customers hardly interact with each other (since almost every customer arrives at an empty queue of stockpoint 1). When \( \rho \) is large, almost every customer sees a non-empty queue. Hence most of the times the interdeparture time of two successive customers equals a service time. Since the service times are independent, also the interarrival times at the queue of stockpoint 2 are independent. Table 2 shows that the performance of the approximations are the worse when \( \rho \) is around 0.8. However, for small \( \rho \) and high \( \rho \) the approximations yield very satisfactory results.

Finally, we notice that it takes quite some time before the simulator 'stabilizes'. This issue was already discussed in Vinson [1972]. Take for example the instance (\( \mu_L = 3 \) and \( c_L = 0.75 \)). Since the parameters of the queuing system of stockpoint 1 can be determined by the exact expression (9) we know that \( \hat{\sigma}_T^2 \) obtained by simulation has to converge to \( \sigma_T^2 = 5.0625 \). In Table 3 this variance is depicted for several different sizes of the simulation run. That it converges very slowly is due to the high correlation between successive lead times. When this correlation is not so large the convergence is much quicker.

<table>
<thead>
<tr>
<th>Cycle time</th>
<th>( \hat{\sigma}_T^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>50,000</td>
<td>5.655</td>
</tr>
<tr>
<td>100,000</td>
<td>5.421</td>
</tr>
<tr>
<td>200,000</td>
<td>5.339</td>
</tr>
<tr>
<td>300,000</td>
<td>5.273</td>
</tr>
<tr>
<td>400,000</td>
<td>5.077</td>
</tr>
</tbody>
</table>

Table 3: \( \hat{\sigma}_T^2 \) for several cycle times.

7 Conclusions

In this paper we develop a simple model for a lead time process in which orders do not cross. We show that the combination of a single server queue and a deterministic pipeline yields a versatile and simple lead time process. To model the lead time process for a single location inventory system we analyzed two methods such that: (1) its first two moments match the moments observed in practice, and (2) besides its first two moments also the first-order correlation matches. For the first method exact expressions are derived ((8) and (9) for \( D/M/1 \)-queue, (12) for the \( GI/M/1 \)-queue). For the second method we developed an algorithm which determines approximate values for the parameters of the queuing model (using (5)-(7),(10) and (11)). It turns out that not every coefficient of correlation can be attained. An interesting question, of course, is whether in practice there are lead time processes which do attain such coefficients of correlation. Since in our opinion the performance of the second method is moderate we recommend to use the first method. Furthermore, the results are extended to arbitrary large divergent multi-echelon systems. Now the performance of the first method depends on the impact of neglecting the dependency between arrivals (if there is any). Numerical results are quite satisfying (see Table 2).
References


EHRHARDT, R. [1984], (s,S) policies for a dynamic inventory model with stochastic lead times, *Operations Research* 32, 121–132.


HEUTS, R., AND J. DE KLEIN [1995], An (s,q) inventory model with stochastic and interrelated lead times, *Naval Research Logistics* 42, 839–859.


A Nested bisection algorithm

max := 10²⁰;
Initialize ε₁ > 0, ε₂ > 0;
Choose μ_b ∈ (0, μ_A);
Determine σ₂ ≥ 0 by bisection such that |μ_S(μ_b, σ₂) − μ_S| < ε₂;
if σ₂(μ_b, σ₂) < σ₂ then
    begin
        μ := 0; σ² := max;
        μ := μ_b; σ² := σ₂;
    end
else
    begin
        μ := μ_A; σ² := 0;
        μ := μ_b; σ² := σ₂;
    end;
while |σ₂(μ_b, σ₂) − σ₂| ≥ ε₁ do
    begin
        μ_b := (μ + μ_b)/2;
        Determine σ₂ ∈ (σ², σ²) by bisection such that |μ_S(μ_b, σ₂) − μ_S| < ε₂;
        if σ₂(μ_b, σ₂) < σ₂ then
            begin
                μ := μ_b; σ² := σ₂;
            end
        else
            begin
                μ := μ_b; σ² := σ₂;
            end;
        end;
    μ_b := μ_b; σ₂ := σ₂;
end;

Figure 6: Nested bisection algorithm to determine μ_b and σ₂ such that the sojourn time has mean μ_S and variance σ₂.