A Closed-Form Approximation for Serial Inventory Systems and Its Application to System Design

Kevin H. Shang • Jing-Sheng Song

The Fuqua School of Business, Duke University, Durham, North Carolina 27708, USA
khshang@duke.edu • jssong@duke.edu


We analyze a serial base-stock inventory model with Poisson demand and a fill-rate constraint. Our objective is to gain insights into the linkage between the stages so to facilitate optimal system design and decentralized system control. To this end, we develop a closed-form approximation for the optimal base-stock levels. The development consists of two key steps: (1) convert the service-constrained model into a backorder-cost model by imputing an appropriate backorder cost rate and then adapt the single-stage approximation developed for the latter, and (2) use a logistic distribution to approximate the leadtime demand distribution in the single-stage approximation obtained in (1) to yield close-form expressions. We then use the closed-form expressions to conduct sensitivity analysis and establish qualitative properties on system design issues, such as optimal total system stock, stock positioning, and internal fill rates.

The closed-form approximation and most of the qualitative properties apply equally to the model with a backorder cost, although some differences do exist. Other results of this study include a bottom-up recursive procedure to evaluate any given echelon base-stock policy and lower bounds on the optimal echelon base-stock levels.

(Keywords: Multiechelon Inventory System, Service Level, Logistic Distribution, Closed-Form Approximations, System Design, Base-Stock Policies)

1. Introduction

Managing the material flow of a supply chain with a service constraint is a common challenge across industries; see, for example, Lee and Billington (1993), Ettl et al. (2000), Graves and Willems (2000, 2003). The most frequently used approach in both the literature and in practice has been decomposition, which treats each stage as a separate entity with certain preassigned parameters (Cohen and Lee 1988, Lee and Billington 1993, Inderfurth and Minner 1998, Bollapragada et al. 2000, Paschalidis and Liu 2000). The decomposition approach clearly provides a computational advantage. Instead of solving a multidimensional optimization problem, one needs only to solve a series of single-dimensional problems. Another advantage is implementation. This way, each manager needs only to optimize his/her own inventory through access to local information.
Despite this common practice, there has been little theoretical investigation on the linkages of different stages. For example, while commenting on their internal fill rate assumptions, Graves et al. (1998) wrote: “In extending the single-stage model to a multistage setting, we assume that the service levels will be set to assure a high level of service, and in the model analysis we ignore the downstream consequences of an upstream stockout; i.e., starvation of inputs. These assumptions raise two questions. (1) What are the consequences of ignoring internal stockouts? (2) What should the internal service be? Subsequent research shows that when service levels are high, ignoring the internal stockouts in the analysis does not distort the results of the model. But the issue remains as to how to set the service levels.”

Indeed, looking through the literature, we found two seemingly contradictory views for setting internal fill rates. More often than not, researchers and practitioners assume upstream locations should achieve higher fill rates in order to guarantee a desired system fill rate (see Bollapragada et al. 2004, Paschalidis and Liu 2000). However, using a two-stage model Choi et al. (2003) show that this rule of thumb is not necessarily accurate. Moreover, drawing from empirical observations based on the optimal system behavior in multiechelon inventory models, Axsäter (2003b) concludes that “What is typical is that upstream installations should have very low stocks compared to downstream installations.” This implies that internal fill rates for upstream stages may be lower than those for downstream stages. Graves and Willems (2003) also note that “[t]he literature on multi-echelon distribution systems suggests that, from a system perspective, it often may be better to have low levels of internal service.” Likewise, the numerical examples in Bollapragada et al. (2004) contain optimal internal fill rates much lower than the system fill rate.

These contradictory views are all based on empirical observations. We are not aware of any theoretical studies on internal fill rates. Evidently, this is largely due to the complexity of the algorithms for solving the optimal solution (cf., Boyaci and Gallego 2001, Axsäter 2003a, Boyaci et al. 2003). This paper aims to take a first step in developing analytical guidelines for managing service constrained systems, with a special attention to the linkages between stages.

We focus on a very basic model: a serial inventory system with a Poisson demand process, linear holding costs and a fill-rate constraint. An echelon base-stock policy is employed at each stage (see §2 for details). In §3 and §4, we develop closed-form approximations for the optimal base-stock levels and demonstrate that they are effective. Using these closed-form expressions, in §5 we conduct analysis to understand the key determinants of optimal system design. Finally, we outline the key insights of our study and conclude the paper in §6.
The development of our closed-form approximation takes two steps. The first (§3) is to adapt the single-stage approximations for the backorder-cost (BC) model developed by Shang and Song (2003a) to the service-constrained (SC) model. For a single-stage SC model with Poisson demand, there exists an equivalent BC model by choosing an appropriate backorder cost rate (see Equation (9)). Recently, Sobel and Zhang (2004) extend this result to a model with general random demand. For the multi-stage SC model with Poisson demand, Boyaci and Gallego (2001) propose a heuristic backorder cost rate for the corresponding multi-stage BC model. We use the same backorder cost rate. We term the resulting single-stage approximation the *newsvendor heuristic* (NVH). Numerical study indicates that NVH is quite effective.

The second step (§4) involves using a logistic distribution to approximate the leadtime demand distribution in NVH. This approach leads to closed-form expressions for the approximate base-stock levels, which greatly enhance transparency and analytical tractability. (Clearly, by construction, the closed-form approximation is also valid for the BC model.) While the logistic distribution has been used in several other fields because of its tractability, e.g., Balakrishnan (1992), to our knowledge, applications to inventory models have been rare.

Our research also yields several other results for the SC model. First, instead of the well-known top-down procedure to evaluate an installation base-stock policy, we develop an alternative bottom-up recursive procedure for evaluating an echelon base-stock policy. This result unifies the evaluation procedures for the SC and the BC models. Second, we construct easy-to-compute lower bounds for optimal base-stock levels. We summarize these results in §2.

2. Model and Preliminaries

We consider an *N*-stage serial inventory system. Poisson demand with rate \( \lambda \) arises at stage 1 and unsatisfied demands are backlogged. Stage 1 obtains resupply from stage 2, stage 2 obtains resupply from stage 3, and so on. Stage \( N \) obtains replenishment from an outside ample supplier. There is constant leadtime \( L_j \) for stage \( j \) and let \( D_j \) denote the demand during \( L_j \). Let \( IN_j \) denote the net echelon inventory level at stage \( j \) (= inventory on hand + inventory at or in transit to all downstream stages - backorders). Define *echelon* inventory position at stage \( j \) to be \( IN_j \) plus inventory on order at this stage. We assume an echelon base-stock policy \( s = (s_1, ..., s_N) \) is used to control the material flow. That is, for any stage \( j \) if its echelon inventory position is lower than \( s_j \), an order is placed to the upstream stage immediately to bring it back to \( s_j \). There is an echelon
inventory-holding cost rate $h_j$ for each stage $j$. The objective is to find echelon base-stock levels so that the average inventory-holding cost is minimized while meeting a prespecified fill rate $\beta$ at stage 1. We term this problem the service-constrained (SC) model.

Denote $h_{[m,n]} = \sum_{i=m}^{n} h_i$. For any given echelon base-stock policy $s$, let

$$
R(s) = \text{fill rate under } s = P(IN_1 > 0),
$$

$$
H(s) = \text{average holding cost under } s = E[\sum_{j=1}^{N} h_j IN_j + h_{[1,N]}(IN_1^-)],
$$

where $(\cdot)^- = -\min\{\cdot, 0\}$. The problem is to solve

$$
\begin{align*}
\text{(SC)} & \quad \min_s H(s) \\
\text{s.t.} \quad & R(s) \geq \beta.
\end{align*}
$$

Denote $s^* = (s_1^*, \ldots, s_N^*)$ to be the optimal echelon base-stock policy. Boyaci et al. (2003) present an exact algorithm to search for $s^*$.

**Policy Evaluation**

In the literature, the most familiar policy evaluation procedure is a top-down procedure to evaluate installation base-stock policies. For convenience, we introduce the following bottom-up procedure to evaluate $H(s)$ and $R(s)$ for any given echelon base-stock policy $s$. The procedure can be easily adapted to evaluate any given installation base-stock policy (see Shang and Song 2003b).

Define

$$
H_1(s_1) = E[h_1(s_1 - D_1) + h_{[1,N]}(s_1 - D_1^-)],
$$

$$
R_1(s_1) = P(D_1 < s_1).
$$

For $j = 2, \ldots, N$,

$$
H_j(s_j) = E[h_j(s_j - D_j) + H_{j-1}(\min\{s_{j-1}, s_j - D_j\})],
$$

$$
R_j(s_j) = E[R_{j-1}(\min\{s_{j-1}, s_j - D_j\})].
$$

Then $H(s) = H_N(s_N)$ and $R(s) = R_N(s_N)$.

Using (2) and (3), we establish the following results:

**Proposition 1** For any echelon base-stock policy $s$,
(a) $R_j(s_j)$ and $R(s)$ increase in $s_j$ for $j = 1, ..., N$.

(b) $R_j(s_j) \leq R_{j-1}(s_{j-1})$, for $j = 2, ..., N$.

From Proposition 1(a), it is easy to verify that moving one unit of inventory from an upstream stage to a downstream stage improves fill rate. Also, allocating one unit of inventory to a downstream stage has a bigger effect in improving fill rate than to an upstream stage. Proposition 1(b) implies that adding one additional stage to the chain will decrease the fill rate.

**Lower Bounds**

As noted by Boyaci et al. (2003), the exact optimal policy can be difficult to compute. The following simple lower bounds on the optimal echelon base-stock levels can be useful. Define

$$L_{[1,j]} = \sum_{i=1}^{j} L_i = \text{total leadtime for echelon } j = L_1 + \cdots + L_j,$$

$$D_{[1,j]} = \sum_{i=1}^{j} D_i = \text{cumulative demand during } L_{[1,j]},$$

$$F_j(y) = \Pr(D_{[1,j]} \leq y),$$

$$F_j^{-1}(\theta) = \min \{y | F_j(y) \geq \theta\}, \quad 0 \leq \theta \leq 1.$$

Let $\ell_j$ be the minimum echelon stock for echelon $j$ to ensure $\beta$, assuming stage $j$ has ample supply from upstream. Since allocating an additional unit of stock to a downstream stage is a more effective way to raise fill rate, this minimum echelon stock occurs when we allocate all the inventories to stage 1 and leave stages 2 through $j$ empty. Under this construction, the original $j$-stage serial system collapses into a single-stage system with accumulated leadtime demand $D_{[1,j]}$, and $\ell_j$ is exactly the smallest base-stock level that ensures $\beta$. In other words, we have:

**Proposition 2** For an $N$-stage system, a lower bound $\ell_j$ on $s^*_j$ can be obtained from

$$\ell_j = F_j^{-1}(\beta) + 1, \quad j = 1, ..., N.$$

We now present some numerical examples to illustrate the performance of the lower bounds. The base model is a four-stage system with the arrival rate $\lambda = 16$, echelon holding cost rates $h_1 = h_2 = h_3 = h_4 = 0.25$, leadtimes $L_1 = L_2 = L_3 = L_4 = 0.25$. We consider two levels of fill rate, $\beta = 0.99$ and 0.9. For each service level, eight variants of the base model are considered. Variant $j, j = 1, 2, 3, 4$ increases the echelon holding cost rate at stage $j$ from 0.25 to 2.5 while keeping the
Table 1: Optimal solutions and bounds for the base model and eight variants

<table>
<thead>
<tr>
<th>Changed Parameter</th>
<th>( \beta )</th>
<th>( \ell_1 )</th>
<th>( s_1^* )</th>
<th>( \ell_2 )</th>
<th>( s_2^* )</th>
<th>( \ell_3 )</th>
<th>( s_3^* )</th>
<th>( \ell_4 )</th>
<th>( s_4^* )</th>
<th>( R(s^*) )</th>
<th>( H(s^*) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>.99</td>
<td>10</td>
<td>12</td>
<td>16</td>
<td>18</td>
<td>22</td>
<td>23</td>
<td>27</td>
<td>27</td>
<td>.9905</td>
<td>15.47</td>
</tr>
<tr>
<td></td>
<td>.90</td>
<td>8</td>
<td>10</td>
<td>13</td>
<td>15</td>
<td>18</td>
<td>19</td>
<td>22</td>
<td>22</td>
<td>.9006</td>
<td>11.49</td>
</tr>
<tr>
<td>Variant 1</td>
<td>.99</td>
<td>10</td>
<td>10</td>
<td>16</td>
<td>19</td>
<td>22</td>
<td>24</td>
<td>27</td>
<td>30</td>
<td>.9902</td>
<td>29.97</td>
</tr>
<tr>
<td>( h_1 = 2.5 )</td>
<td>.90</td>
<td>8</td>
<td>8</td>
<td>13</td>
<td>15</td>
<td>18</td>
<td>20</td>
<td>22</td>
<td>23</td>
<td>.9013</td>
<td>20.13</td>
</tr>
<tr>
<td>Variant 2</td>
<td>.99</td>
<td>10</td>
<td>13</td>
<td>16</td>
<td>16</td>
<td>22</td>
<td>24</td>
<td>27</td>
<td>29</td>
<td>.9902</td>
<td>42.86</td>
</tr>
<tr>
<td>( h_2 = 2.5 )</td>
<td>.90</td>
<td>8</td>
<td>11</td>
<td>13</td>
<td>13</td>
<td>18</td>
<td>19</td>
<td>22</td>
<td>23</td>
<td>.9002</td>
<td>31.13</td>
</tr>
<tr>
<td>Variant 3</td>
<td>.99</td>
<td>10</td>
<td>13</td>
<td>16</td>
<td>17</td>
<td>22</td>
<td>22</td>
<td>27</td>
<td>28</td>
<td>.9903</td>
<td>55.63</td>
</tr>
<tr>
<td>( h_3 = 2.5 )</td>
<td>.90</td>
<td>8</td>
<td>9</td>
<td>13</td>
<td>15</td>
<td>18</td>
<td>18</td>
<td>22</td>
<td>23</td>
<td>.9045</td>
<td>42.49</td>
</tr>
<tr>
<td>Variant 4</td>
<td>.99</td>
<td>10</td>
<td>12</td>
<td>16</td>
<td>18</td>
<td>22</td>
<td>23</td>
<td>27</td>
<td>27</td>
<td>.9905</td>
<td>67.24</td>
</tr>
<tr>
<td>( h_4 = 2.5 )</td>
<td>.90</td>
<td>8</td>
<td>10</td>
<td>13</td>
<td>15</td>
<td>18</td>
<td>19</td>
<td>22</td>
<td>22</td>
<td>.9006</td>
<td>52.35</td>
</tr>
<tr>
<td>Variant 5</td>
<td>.99</td>
<td>27</td>
<td>30</td>
<td>32</td>
<td>34</td>
<td>37</td>
<td>38</td>
<td>42</td>
<td>42</td>
<td>.9900</td>
<td>28.04</td>
</tr>
<tr>
<td>( L_1 = 1 )</td>
<td>.90</td>
<td>22</td>
<td>26</td>
<td>27</td>
<td>29</td>
<td>31</td>
<td>32</td>
<td>36</td>
<td>36</td>
<td>.9006</td>
<td>22.48</td>
</tr>
<tr>
<td>Variant 6</td>
<td>.99</td>
<td>10</td>
<td>12</td>
<td>32</td>
<td>34</td>
<td>37</td>
<td>39</td>
<td>42</td>
<td>42</td>
<td>.9902</td>
<td>24.05</td>
</tr>
<tr>
<td>( L_2 = 1 )</td>
<td>.90</td>
<td>8</td>
<td>11</td>
<td>27</td>
<td>29</td>
<td>31</td>
<td>32</td>
<td>36</td>
<td>36</td>
<td>.9000</td>
<td>19.07</td>
</tr>
<tr>
<td>Variant 7</td>
<td>.99</td>
<td>10</td>
<td>12</td>
<td>16</td>
<td>18</td>
<td>37</td>
<td>40</td>
<td>42</td>
<td>42</td>
<td>.9901</td>
<td>20.25</td>
</tr>
<tr>
<td>( L_3 = 1 )</td>
<td>.90</td>
<td>8</td>
<td>10</td>
<td>13</td>
<td>15</td>
<td>31</td>
<td>33</td>
<td>36</td>
<td>36</td>
<td>.9000</td>
<td>15.66</td>
</tr>
<tr>
<td>Variant 8</td>
<td>.99</td>
<td>10</td>
<td>11</td>
<td>16</td>
<td>17</td>
<td>22</td>
<td>23</td>
<td>42</td>
<td>44</td>
<td>.9902</td>
<td>16.55</td>
</tr>
<tr>
<td>( L_4 = 1 )</td>
<td>.90</td>
<td>8</td>
<td>10</td>
<td>13</td>
<td>15</td>
<td>18</td>
<td>20</td>
<td>36</td>
<td>36</td>
<td>.9000</td>
<td>12.32</td>
</tr>
</tbody>
</table>

Table 1: Optimal solutions and bounds for the base model and eight variants

holding cost rates at the other three stages at 0.25. Variant \( j, j = 5, 6, 7, 8 \) increases the leadtime at stage \( j - 4 \) from 0.25 to 1 and keeps the leadtimes at the other three stages at 0.25. Table 1 shows the results.

We observed that if there is a “jump” in echelon holding cost at stage \( j \), \( \ell_j \) tends to match \( s_j^* \), as shown in variants 1 to 4. Also, the lower bound is tighter as we move toward upstream when echelon holding costs \( h_j \) are equal. For example, \( s_4^* \) is the same as \( \ell_4 \) in base case and in variants 5 to 8. These observations seem to suggest that, as a quick heuristic for the service-constrained model, we may stock at the lower bound level at the upmost stage if \( h_j \) values are similar across stages or at a stage with a relatively high \( h_j \).

Unfortunately, there seems to be no simple way to obtain upper bounds; see Boyaci et al. (2003) and Shang and Song (2003b).

### 3. The Newsvendor Heuristic

To see “what an optimal solution looks like” (Axsäter 2003b), we aim to develop closed-form approximations for optimal echelon base-stock levels. As a first step, in this section, we adapt...
the single-stage approximation for the serial backorder-cost model developed by Shang and Song (2003a) to the service constrained model.

3.1 Single-Stage Approximation for the Backorder-Cost Model

Instead of the service-level constraint in (1), the backorder-cost model assumes that there is a linear backorder cost rate \( b \) for unsatisfied demands. The objective is to find an echelon base-stock policy that minimizes the average holding and backorder costs, i.e.,

\[
(BC) \quad \min_h H(s) + bB(s),
\]

where \( B(s) = \mathbb{E}[IN_1^-] \). Let \((r^*_1, \ldots, r^*_N)\) denote the optimal solution to this model.

Shang and Song (2003a) show that the optimal echelon base-stock level \( r^*_j \) can be approximated by a single-stage base-stock level \( r^\alpha_j \). Specifically, for any set of weights \( \alpha_{k,j} \), \( k = 1, \ldots, j \), satisfying \( 0 \leq \alpha_{k,j} \leq 1 \) and \( \sum_{k=1}^j \alpha_{k,j} = 1 \), define

\[
h^\alpha_j = \alpha_{j,j} h_j + \alpha_{j-1,j} h_{j-1,j} + \ldots + \alpha_{1,j} h_{1,j}.
\]

Suppose the optimal echelon base-stock policy is followed for all stages \( i < j \). If each stage in echelon \( j \) is charged the same holding cost rate \( h^\alpha_j \), and a backorder cost rate \( b + h_{j+1,N} \) is charged at stage 1, then echelon \( j \) collapses into a single stage with total leadtime \( L_{[1,j]} \), whose optimal base-stock level is

\[
r^\alpha_j = F_j^{-1}\left(\frac{b + h_{j+1,N}}{b + h_{j+1,N} + h^\alpha_j}\right).
\]

Because of its similarity in format to the newsvendor solution, \( r^\alpha_j \) is referred as the newsvendor heuristic solution for the BC model.

Taking \( \alpha_{1,j} = 1 \) and \( \alpha_{j,j} = 1 \) respectively leads to newsvendor-type lower and upper bounds on \( r^*_j \). That is, \( r^\ell_j \leq r^*_j \leq r^u_j \), where

\[
\begin{align*}
r^\ell_j &= F_j^{-1}(\theta^\ell_j), \\
r^u_j &= F_j^{-1}(\theta^u_j), \\
\theta^\ell_j &= \frac{b + h_{[j+1,N]}}{b + h_{[1,N]}}, \\
\theta^u_j &= \frac{b + h_{[j+1,N]}}{b + h_{[j,N]}}.
\end{align*}
\]

Any linear combination of \( r^\ell_j \) and \( r^u_j \) can also be used to approximate \( r^*_j \). Gallego and Özer (2004) and Watson and Zheng (2005) show that the following choice of \( \alpha_{k,j} \) in (5) yields effective approximation:

\[
\alpha_{k,j} = \frac{L_k}{L_{[1,j]}}, \quad 1 \leq k \leq j, \quad j = 1, \ldots, N.
\]
3.2 The Newsvendor Heuristic (NVH)

A natural question arises here: can we adopt the above results for the BC model to serial systems with a service constraint? The key issue here is, of course, whether we can find an imputed backorder cost parameter \( b \) for any given fill rate \( \beta \).

It is well known that for a single-stage system \((N = 1)\), the service-constrained (SC) problem (1) and the backorder-cost (BC) problem (4) are equivalent if we set

\[
b = \frac{\beta h_{[1,N]}}{1 - \beta},
\]

in the sense that they produce the same optimal solution. More specifically, letting \( r_1^* \) be the solution to the BC model, then \( r_1^* + 1 \) is the optimal solution to the SC model. See, for example, Zipkin (2000) page 185-186.

For \( N > 1 \) and a given target fill rate \( \beta \), we use (9) as a heuristic backorder-cost rate for the corresponding BC model and substitute it in (6) to obtain \( r_j^a \). Here, \( h_j^a \) is chosen to be the leadtime-weighted echelon holding cost defined in (8). Finally, we use \( s_j^a = r_j^a + 1 \) to approximate \( s_j^* \), \( j = 1, ..., N \). To summarize, we have

\[
s_j^a = F_j^{-1}(\theta_j^a) + 1, \quad j = 1, ..., N,
\]

where

\[
\theta_j^a = \frac{\beta h_{[1,j]} + h_{[j+1,N]}}{\beta h_{[1,j]} + h_{[j+1,N]} + (1 - \beta) \sum_{i=1}^{j} h_i L_{[1,i]} / L_{[1,j]}},
\]

We call \((s_1^a, ..., s_N^a)\) the newsvendor heuristic (NVH) solution.

3.3 Numerical Study

It turns out that the NVH solution does not always guarantee the target fill rate (see an explanation in §3.3.2). Therefore it is of interest to see (1) the effectiveness of NVH and the level of violation on resulting fill rates, and (2) whether the NVH solution is consistent with the optimal solution when a system parameter changes. The numerical study in this section aims to address these questions.

3.3.1 Comparison with Optimal Solution

We first examine question (1) above. We use the following measures to show the effectiveness of a heuristic:

\[
\text{Percentage Cost Error} = \frac{100|H^a - H^*|}{H^*} \%
\]
Percentage Fill-rate Error \(= \frac{100(\beta - \beta^a)}{\beta} \%\)

Here, \((x)^+ = \max\{x, 0\}\). We measure the percentage cost error in absolute value since the NVH cost may be smaller than the optimal cost when a target fill rate is not achieved. In such a case, we use percentage fill-rate error to measure the gap.

We consider different holding cost structures – linear, affine, kink, and jump holding costs – adopted from Boyaci and Gallego (2001) and Gallego and Zipkin (1999). The choice of a holding cost structure will lead to different stocking amounts at each stage. We assume leadtimes are equal for all stages and total system leadtime equals 1. Also, the total system holding cost \(h_1\) is fixed and equal to 1.

For linear holding costs, \(h_j = 1/N\). For affine holding costs, \(h_N = \alpha + (1 - \alpha)/N\) and \(h_j = (1 - \alpha)/N, j = 1, 2, ..., N - 1\). Here, we compare \(\alpha = 0.25\) case. The kink form is piecewise linear with two pieces. We assume the system changes the holding cost rate in the middle stage, that is, \(h_j = (1 - k)/N, j \leq N/2\) and \(h_j = (1 + k)/N, j > N/2\). We use \(k = 0.75\) in the example. The last one is the jump holding cost form, where cost is incurred at a constant rate, except for one stage with a large cost. We assume the jump occurs at stage \(N/2\), so \(h_j = u + (1 - u)/N, j = N/2\), and \(h_j = (1 - u)/N\) otherwise. Here, we test the case \(u = 0.75\). Our numerical experiment considers a four-stage system. For each holding cost structure, we let for \(\beta = 0.9, 0.99\) and \(\lambda = 16, 32, 64\). There are 24 instances in this study.

Table 2 summarizes the optimal cost \(C^*\) and percentage errors of NVH in all 24 examples tested. Note that we exclude the average in-transit cost from total system cost since it is independent of base-stock levels. The exact optimization algorithm can be found in Boyaci et al. (2003). We use parentheses to indicate the case in which the resulting fill rate is below \(\beta\). In these 7 cases whose resulting fill rate is below \(\beta\), the NVH costs are less than the optimal costs. The average percentage error is 0.08\% among these 24 instances.

The optimal total costs seem sensitive to different holding cost structures. The maximum cost difference, 18.30-14.75 = 3.56 (24.11\% above 14.75), occurs between the affine and kink holding cost forms under \(\lambda = 64\) and \(\beta = 0.99\). On the contrary, the optimal system stocks seem fairly stable. For example, in the same case, the optimal system stock ranges from 85 to 89. This result suggests that the system stock is relatively insensitive to stock positioning, provided that the overall stock level is about right and low-cost stocking points are exploited. This conclusion is consistent with that in the BC model found by Gallego and Zipkin (2001). We will discuss stock positioning...
in detail in §5.

Next, we examine question (2), that is, whether the NVH solution moves in the same direction as the optimal when a system parameter changes. Table 3 shows the optimal and the NVH solutions for the same base model and 8 variants studied in §2. From this table, we observe that the NVH solution is generally close to the optimal solution. In addition, the patterns of change for the optimal and the NVH solution are consistent when system parameters change. For instance, when \( \beta \) decreases, the optimal and the NVH solutions for all stages decrease in all 9 cases. However, we find the optimal solution is rather insensitive to \( h_N \). For example, when \( h_4 \) increases from 0.25 to 2.5 in variant 4, the optimal solution at each stage remains the same as that in the base model. This is because \( h_N \) will be carried over to each unit in the system. Although the echelon holding costs \( h_j, j = 2, \ldots, N \) increase, the marginal increase of two adjacent holding costs \( (h_{j-1} - h_j) \) remains the same. Thus, the optimal solution in the variant 4 will be the same as in the base model.

### 3.3.2 Comparison with Other Heuristics

It is worth mentioning that Boyaci and Gallego (2001) also propose heuristic solutions for the SC model, two of which are relevant to ours: the majorization heuristic (MH) and the backorder-cost heuristic (BCH). The former is their best performing heuristic. The latter uses the solution \((r_1^*, \ldots, r_N^*)\) as an approximation for \( s^* \), where \((r_1^*, \ldots, r_N^*)\) is the optimal echelon base-stock policy for the corresponding BC model with \( b \) in (9).

The MH solution is obtained as follows: for any given system stock level chosen from a feasible

<table>
<thead>
<tr>
<th>Form</th>
<th>( \lambda )</th>
<th>Heuristic</th>
<th>( \beta = .9 )</th>
<th>( \beta = .99 )</th>
<th>Form</th>
<th>( \lambda )</th>
<th>Heuristic</th>
<th>( \beta = .9 )</th>
<th>( \beta = .99 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affine</td>
<td>16</td>
<td>( C^* )</td>
<td>5.66</td>
<td>9.86</td>
<td>Linear</td>
<td>16</td>
<td>( C^* )</td>
<td>5.49</td>
<td>9.47</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>( C^* )</td>
<td>7.75</td>
<td>13.34</td>
<td></td>
<td>32</td>
<td>( C^* )</td>
<td>7.52</td>
<td>12.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NVH</td>
<td>11.18%</td>
<td>5.71%</td>
<td></td>
<td></td>
<td>NVH</td>
<td>4.56%</td>
<td>2.13%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NVH</td>
<td>4.60%</td>
<td>3.25%</td>
<td></td>
<td></td>
<td>NVH</td>
<td>4.54%</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>( C^* )</td>
<td>10.72</td>
<td>18.30</td>
<td></td>
<td>64</td>
<td>( C^* )</td>
<td>10.25</td>
<td>17.40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NVH</td>
<td>7.61%</td>
<td>0.00%</td>
<td></td>
<td></td>
<td>NVH</td>
<td>2.56%</td>
<td>0.11%</td>
</tr>
<tr>
<td>Kink</td>
<td>16</td>
<td>( C^* )</td>
<td>4.80</td>
<td>8.13</td>
<td>Jump</td>
<td>16</td>
<td>( C^* )</td>
<td>4.90</td>
<td>8.45</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>( C^* )</td>
<td>6.48</td>
<td>10.85</td>
<td></td>
<td>32</td>
<td>( C^* )</td>
<td>6.65</td>
<td>11.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NVH</td>
<td>(1.86%)</td>
<td>(1.26%)</td>
<td></td>
<td></td>
<td>NVH</td>
<td>2.02%</td>
<td>(2.71%)</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>( C^* )</td>
<td>8.76</td>
<td>14.75</td>
<td></td>
<td>64</td>
<td>( C^* )</td>
<td>9.15</td>
<td>15.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NVH</td>
<td>1.10%</td>
<td>(2.10%)</td>
<td></td>
<td></td>
<td>NVH</td>
<td>(2.27%)</td>
<td>(3.35%)</td>
</tr>
</tbody>
</table>

Table 2: Performance of heuristics under different holding cost forms
region, we initially place all the stock to stage 1 and then move as much stock as possible to stage 2 while retaining feasibility, and repeat this procedure for stage 3, 4, etc. until stage N.

It is of interest to compare NVH and MH. We consider variants 1 to 4 under $\beta = .99, .975, .95, .925, .90, .875, .85$. There are a total of 35 cases. Let $H^a$ and $\beta^a$ be the heuristic cost and heuristic fill rate, and $H^*$ be the optimal cost. The numbers in parenthesis next to the minimum and maximum percentage cost errors correspond to the $\beta$ values with which the extreme errors occur. The ratio in parenthesis next to each percentage fill rate error is the infeasibility ratio $n/m$, where $m$ is the total number of cases tested and $n$ is the number of cases in which the heuristic fails to satisfy $\beta$. Table 4 is the summary.

From the above numerical study we observe:

1. On average MH performs better, especially when the echelon holding cost at the lower stage is higher. However, when the echelon holding cost at higher stage becomes larger (except the most upstream stage), MH performance worsens. For example, when $h_1 = 2.5$, MH performs best, but its performance becomes worse when $h_3$ increases to 2.5. This is intuitively clear: When $h_1$ is high relative to the other holding costs, the optimal allocation scheme tends to

<table>
<thead>
<tr>
<th>Changed Parameters</th>
<th>$\beta$</th>
<th>$s_1^a$</th>
<th>$s_1^*$</th>
<th>$s_2^a$</th>
<th>$s_2^*$</th>
<th>$s_3^a$</th>
<th>$s_3^*$</th>
<th>$s_4^a$</th>
<th>$s_4^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>.99</td>
<td>12</td>
<td>12</td>
<td>18</td>
<td>17</td>
<td>23</td>
<td>23</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td>Variant 1</td>
<td>.99</td>
<td>10</td>
<td>11</td>
<td>19</td>
<td>17</td>
<td>24</td>
<td>23</td>
<td>30</td>
<td>29</td>
</tr>
<tr>
<td>$h_1 = 2.5$</td>
<td>.90</td>
<td>8</td>
<td>8</td>
<td>15</td>
<td>14</td>
<td>20</td>
<td>19</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>Variant 2</td>
<td>.99</td>
<td>13</td>
<td>13</td>
<td>16</td>
<td>17</td>
<td>24</td>
<td>22</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>$h_2 = 2.5$</td>
<td>.90</td>
<td>11</td>
<td>11</td>
<td>13</td>
<td>13</td>
<td>19</td>
<td>18</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>Variant 3</td>
<td>.99</td>
<td>13</td>
<td>13</td>
<td>17</td>
<td>19</td>
<td>22</td>
<td>22</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>$h_3 = 2.5$</td>
<td>.90</td>
<td>9</td>
<td>11</td>
<td>15</td>
<td>16</td>
<td>18</td>
<td>18</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>Variant 4</td>
<td>.99</td>
<td>12</td>
<td>13</td>
<td>18</td>
<td>19</td>
<td>23</td>
<td>24</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>$h_4 = 2.5$</td>
<td>.90</td>
<td>10</td>
<td>11</td>
<td>15</td>
<td>16</td>
<td>19</td>
<td>21</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>Variant 5</td>
<td>.99</td>
<td>30</td>
<td>29</td>
<td>34</td>
<td>34</td>
<td>38</td>
<td>38</td>
<td>42</td>
<td>43</td>
</tr>
<tr>
<td>$L_1 = 1$</td>
<td>.90</td>
<td>26</td>
<td>25</td>
<td>29</td>
<td>29</td>
<td>32</td>
<td>33</td>
<td>36</td>
<td>37</td>
</tr>
<tr>
<td>Variant 6</td>
<td>.99</td>
<td>12</td>
<td>12</td>
<td>34</td>
<td>34</td>
<td>39</td>
<td>38</td>
<td>42</td>
<td>43</td>
</tr>
<tr>
<td>$L_2 = 1$</td>
<td>.90</td>
<td>11</td>
<td>9</td>
<td>29</td>
<td>30</td>
<td>32</td>
<td>33</td>
<td>36</td>
<td>37</td>
</tr>
<tr>
<td>Variant 7</td>
<td>.99</td>
<td>12</td>
<td>12</td>
<td>18</td>
<td>17</td>
<td>40</td>
<td>39</td>
<td>42</td>
<td>43</td>
</tr>
<tr>
<td>$L_3 = 1$</td>
<td>.90</td>
<td>10</td>
<td>9</td>
<td>15</td>
<td>14</td>
<td>33</td>
<td>34</td>
<td>36</td>
<td>37</td>
</tr>
<tr>
<td>Variant 8</td>
<td>.99</td>
<td>11</td>
<td>12</td>
<td>17</td>
<td>17</td>
<td>23</td>
<td>23</td>
<td>44</td>
<td>44</td>
</tr>
<tr>
<td>$L_4 = 1$</td>
<td>.90</td>
<td>10</td>
<td>9</td>
<td>15</td>
<td>14</td>
<td>20</td>
<td>19</td>
<td>36</td>
<td>38</td>
</tr>
</tbody>
</table>

Table 3: Optimal and NVH policies for the base model and eight variants
<table>
<thead>
<tr>
<th>Holding Cost</th>
<th>Solution</th>
<th>Average Cost Error %</th>
<th>Minimum Cost Error % (β)</th>
<th>Maximum Cost Error % (β)</th>
<th>Average Fill-rate Error % (Infeasibility Ratio)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>NVH</td>
<td>5.07</td>
<td>0.21 (.975)</td>
<td>7.16 (.925)</td>
<td>0.00 (0/7)</td>
</tr>
<tr>
<td></td>
<td>MH</td>
<td>1.08</td>
<td>0.00 (.90)</td>
<td>4.12 (.875)</td>
<td>0.00 (0/7)</td>
</tr>
<tr>
<td>variant 1</td>
<td>NVH</td>
<td>3.97</td>
<td>0.72 (.90)</td>
<td>8.01 (.875)</td>
<td>0.35 (2/7)</td>
</tr>
<tr>
<td>$h_1=2.5$</td>
<td>MH</td>
<td>0.16</td>
<td>0.00 (.90)</td>
<td>0.41 (.85)</td>
<td>0.00 (0/7)</td>
</tr>
<tr>
<td>variant 2</td>
<td>NVH</td>
<td>2.52</td>
<td>0.33 (.85)</td>
<td>4.59 (.875)</td>
<td>0.07 (2/7)</td>
</tr>
<tr>
<td>$h_2=2.5$</td>
<td>MH</td>
<td>5.99</td>
<td>4.09 (.99)</td>
<td>9.14 (.90)</td>
<td>0.00 (0/7)</td>
</tr>
<tr>
<td>variant 3</td>
<td>NVH</td>
<td>2.20</td>
<td>0.00 (.925)</td>
<td>3.92 (.875)</td>
<td>0.01 (1/7)</td>
</tr>
<tr>
<td>$h_3=2.5$</td>
<td>MH</td>
<td>7.14</td>
<td>0.38 (.90)</td>
<td>15.07 (.875)</td>
<td>0.00 (0/7)</td>
</tr>
<tr>
<td>variant 4</td>
<td>NVH</td>
<td>3.08</td>
<td>1.26 (.85)</td>
<td>10.74 (.975)</td>
<td>0.00 (0/7)</td>
</tr>
<tr>
<td>$h_4=2.5$</td>
<td>MH</td>
<td>0.33</td>
<td>0.00 (.90)</td>
<td>1.00 (.875)</td>
<td>0.00 (0/7)</td>
</tr>
<tr>
<td>All Cases</td>
<td>NVH</td>
<td>3.37</td>
<td>0.00</td>
<td>10.74</td>
<td>0.11 (5/35)</td>
</tr>
<tr>
<td></td>
<td>MH</td>
<td>2.94</td>
<td>0.00</td>
<td>15.07</td>
<td>0.00 (0/35)</td>
</tr>
</tbody>
</table>

Table 4: Comparison of the optimal and heuristic solutions: The base model vs. four variants

allocate more stocks to upper stages, which is consistent with the idea of MH. However, if there is an increase in $h_3$, pushing inventory to upper stages will cause a higher cost penalty. On the other hand, the performance of NVH tends to be more stable, which ranges from 2 to 4% on average in these five groups.

2. The average cost error for NVH is 3.37%. Although there are 5 infeasible solutions among 35 cases, the average fill rate error in these 35 instances is only 0.11%. Also, in these 5 infeasible solutions, the corresponding costs are all lower than the optimal cost.

For the 57 instances tested in this section, the average cost error for NVH is 3.13%, which is almost the same as MH of 3.08%.

It is shown by Boyaci and Gallego (2001) that the BCH solution guarantees the target fill rate. This implies that the upper bound newsvendor solution $r_{ij}^u + 1$ in (7) also guarantees the target fill rate. The lower bound newsvendor solution $r_{ij}^l + 1$, on the other hand, may violate the target fill rate. Thus, as a weighted average of these bounds, the NVH solution may not always satisfy the service constraint. For the 57 instances tested in this section, the average cost error for BCH is 5.87%, which is larger than that of NVH. Although NVH cannot guarantee feasibility, the average fill rate error is only 0.1%. A closer look at (10) reveals that when the downstream echelon holding costs ($h_i, i \leq j$) are high, $\theta_j^a$ is closer to $\theta_j^l$, which drives the NVH solution closer to $r_{ij}^l + 1$. It turns out that in such
cases, the optimal solution is also close to \( r^j_f + 1 \). On the other hand, the BCH solution is closer to \( r^u_j + 1 \). For instance, for a four-stage system with \( \lambda = 16 \), \( h_1 = h_2 = 2.5, h_3 = h_4 = 0.01 \), \( L_j = 0.25, j \leq 4 \), and \( \beta = 0.9 \), the BCH solution is (9,13,23,30), the optimal solution is (8,13,21,25) and the NVH solution is (9,13,19,24). Clearly, the NVH solution is closer to the optimal one.

Nonetheless, overall, the NVH solution should be close to the BCH solution in most cases, as tested by Gallego and Özer (2004) for the BC model. Since BCH guarantees feasibility, we expect that the violation of fill rate constraint in NVH should not be significant, as demonstrated in the above numerical results. Given that it is algorithm-free, NVH appears to be a convenient and practical heuristic. More importantly, the good performance of NVH suggests that we can rely on its simple form to gain insights. This will be the focus of the remainder of the paper.

4. The Logistic Distribution and Closed-Form Expressions

To enhance transparency, a common approach is to approximate the Poisson leadtime demand \( D_{[1,j]} \) by a normal distribution with mean \( E[D_{[1,j]}] = \lambda L_{[1,j]} \) and \( \text{Var}[D_{[1,j]}] = \lambda L_{[1,j]} \). This approximation is appropriate when the leadtime demand is large. Let \( \Phi(\cdot) \) denote the c.d.f of the standard normal distribution. Then,

\[
\begin{align*}
    s^a_j &= \lambda L_{[1,j]} + \sqrt{\lambda L_{[1,j]}} \Phi^{-1}(\theta^a_j),
\end{align*}
\]

where \( \theta^a_j \) is defined in (10). This expression allows us to see clearly the key determining factors of the optimal echelon-\( j \) base-stock level.

However, due to the implicit form of \( \Phi(\cdot) \), it is difficult to see the connections between optimal installation base-stock levels. To overcome this shortcoming, we further approximate the normal distribution by a logistic distribution.

A logistic distribution with parameters \( m \) and \( \gamma \) has c.d.f

\[
\begin{align*}
    \Psi(x) &= \frac{1}{1 + \exp\left(-\frac{x-m}{\gamma}\right)},
\end{align*}
\]

and p.d.f.

\[
\begin{align*}
    \psi(x) &= \frac{\exp\left(-\frac{x-m}{\gamma}\right)}{\gamma \left[1 + \exp\left(-\frac{x-m}{\gamma}\right)\right]^2},
\end{align*}
\]

where \( \gamma > 0 \). The mean and variance of this distribution are \( m \) and \( \frac{1}{3}\pi^2\gamma^2 \), respectively. See Balakrishnan (1992) for more information on this distribution.
From our numerical experiment, a normal distribution with mean $\mu$ and standard deviation $\sigma$ can be accurately approximated by a logistic distribution with $m = \mu$ and $\gamma = \sigma / \sqrt{\frac{\pi}{3.75}} = \sigma / 1.62 = 0.617\sigma$. This choice of $\gamma$ minimizes

$$
\int_{-\infty}^{\infty} (\Psi_{\gamma}(x) - \Phi(x))^2 \, dx,
$$

where $\Psi_{\gamma}(x)$ is the logistic distribution with mean $m = 0$ and scale parameter $\gamma$. Consequently, the Poisson leadtime demand $D_{[1,j]}$ can be approximated by a logistic distribution with $m = \lambda L_{[1,j]}$ and $\gamma = 0.617\sqrt{\lambda L_{[1,j]}}$. Figure 1 compares Poisson distribution with mean 16, normal distribution with mean 16 and standard deviation 4, and logistic distribution with $m = 16$ and $\gamma = (0.617)(4) = 2.21$.

With the above logistic approximation, we have

$$
P(D_{[1,j]} \leq s^a_j) = \left[1 + \exp \left(\frac{-s^a_j - \lambda L_{[1,j]}}{0.617\sqrt{\lambda L_{[1,j]}}}\right)\right]^{-1} = \theta^a_j,
$$

This yields the following closed-form expression for $s^a_j$:

$$
s^a_j = \lambda L_{[1,j]} + 0.617\sqrt{\lambda L_{[1,j]}} \ln(\xi_j),
$$

where

$$
\xi_j = \frac{\theta^a_j}{1 - \theta^a_j} = \frac{\beta h_{[1,j]} + h_{[j+1,N]}}{(1 - \beta)(\sum_{i=1}^{j} h_i L_{[i,1,j]} / L_{[1,j]})},
$$

$j = 1, ..., N$. 
Also, the installation base-stock levels are \( s_1^a = s_1^a \) and
\[
s_j^a = s_j^a - s_{j-1}^a = \lambda L_j + 0.617 \sqrt{\lambda} \left[ \sqrt{L_{[1,j]} \ln(\xi_j)} - \sqrt{L_{[1,j-1]} \ln(\xi_{j-1})} \right],
\]

\( j = 2, ..., N \). Note that \( s_j^a \) (or \( s_j^a \)) may not be an integer. In that case, we simply round it up to the next smallest positive integer.

5. System Design Issues

Equipped with the closed-form approximations of the optimal base-stock levels developed in the previous section, we perform analysis in order to gain insight into system design and implementation issues. We situate ourselves as a central planner for a serial supply chain. We are interested in the following questions:

(1) What is an appropriate stocking amount for the entire chain? How do system parameters affect it?

(2) How should system stock be allocated among stages? If the current stock positioning is not appropriate, how should we reallocate stock? Should we consolidate some stages to form a shorter supply chain? (Note that a stage can be eliminated or merged with another stage if no stock is allocated to that stage.)

(3) What is the internal fill rate needed for a stage so that the supply chain satisfies a target fill rate? Is this internal fill rate larger or smaller than the target fill rate?

5.1 Total System Stock

We now examine the key determinants of the optimal total system stock \( s_N^a \), which is approximated by
\[
s_N^a = \lambda L_{[1,N]} + 0.617 \sqrt{\lambda L_{[1,N]}} \ln(\xi_N)
= \lambda L_{[1,N]} + 0.617 \sqrt{\lambda L_{[1,N]}} \ln \left( \frac{\beta}{1 - \beta} \cdot \frac{h_{[1,N]} L_{[1,N]}}{\sum_{i=1}^{N} h_i L_{[1,i]}} \right).
\]

It is interesting to see the effect of the presence of immediate stages on total system stock. We first compare the total system stock in the serial system with that in a single-stage system with demand rate \( \lambda \), leadtime \( L_{[1,N]} \) and holding cost \( h_{[1,N]} \), assuming both systems have the same fill
rate $\beta$. As in a single-stage system, the total system stock in the serial system covers the mean leadtime demand $\lambda L_{[1,N]}$ and a safety stock that is proportional to the standard deviation of the leadtime demand $\sqrt{\lambda L_{[1,N]}}$. The only difference is the magnitude of the safety factor. In the single-stage system, this factor is $0.617\ln\left(\frac{\beta}{1-\beta}\right)$, which is increasing in $\beta$. In the serial system, it is $0.617\ln(\xi_N)$, which is not only increasing in $\beta$ but is also affected by the distribution of the holding costs and the distribution of the leadtimes. Since $(h_{[1,N]}L_{[1,N]})/(\sum_{i=1}^{N} h_{i}L_{[1,i]}) > 1$, the total system stock in the serial system is strictly larger. This is not surprising: In the single-stage system, all stock is located near the customer, so it is more responsive to customer demand. However, due to lower upstream holding costs, the total system cost becomes lower. In other words, upstream stages hold inventory only if their holding costs are lower.

In the following, we fix the total leadtime $L_{[1,N]} = 1$ and the total holding cost $h_{[1,N]} = 1$. This allows us to take a closer look at the effects of the number of stages $N$, the distribution of the leadtimes, and the distribution of the holding costs. Since $\lambda$ and $L_{[1,N]}$ are fixed, it is sufficient to examine the effect on $\xi_N$.

To examine the effect of $N$, we assume echelon holding costs and leadtimes are equal, i.e., $h_j = 1/N$ and $L_j = 1/N$ for all $j$. This leads to

$$\xi_N = \frac{2\beta}{1-\beta} \cdot \frac{N}{N+1},$$

which is increasing in $N$. Hence, we have:

**Proposition 3** In an $N$-stage supply chain with equal leadtime and echelon holding cost for all stages, the system stock increases in $N$. Thus, reducing the number of stages in a supply chain results in lower total system stock.

This property is consistent with our earlier comparison of the series system and the single-stage system. It has also been numerically observed in the BC model by Gallego and Zipkin (1999).

Next, we examine the effect of the leadtime length on the total system stock. We assume that echelon holding costs are equal, i.e., $h_j = 1/N$ and stage $j$ has the longest leadtime, while the other stages have the same leadtimes. That is, $L_j = l + (1-l)/N$ and $L_i = (1-l)/N$, $i \neq j$, $j = 1, \ldots, N$, and $0 < l < 1$. Denote $s^a_N(L_j)$ and $\xi_N(L_j)$ to be the corresponding total system stock and $\xi_N$, respectively. Then

$$\xi_N(L_j) = \frac{2\beta}{1-\beta} \cdot \frac{N}{(1-l)(1+N) + 2l(N-j+1)},$$

which is increasing in $j$. Thus, we have:
**Proposition 4** In an $N$-stage system with fixed total leadtime and equal echelon holding cost for all stages,

$$s_N^a(L_1) \leq s_N^a(L_2) \leq ... \leq s_N^a(L_N).$$

That is, moving an intermediate stage closer to its downstream stage increases total system stock.

Proposition 4 states that, given the same amount of leadtime increment, the required total system stock needed is smaller if the increment happens at a downstream stage. This is intuitive, because the same amount of leadtime increment would imply same increase in safety stock in any stage if we treat each stage as independent. However, in a serial system, the increased safety stock at a downstream stage has a bigger effect on increasing service level than at an upstream stage. This proposition provides some insights into the effect of an intermediate stage location in a chain. For instance, consider a supply chain where the distribution center can be located close to either the manufacturer or the retailer site. The above result indicates that the optimal system stock in the former configuration would be smaller.

Finally, we study the effect of holding costs on system stock. We assume leadtimes are equal, i.e., $L_j = 1/N$ for all $j$, but the echelon holding cost jumps at stage $j$, i.e., $h_j = u + (1 - u)/N$ and $h_i = (1 - u)/N$, $i \neq j, j = 1, ..., N$. Denote $s_N^a(h_j)$ and $\xi_N(h_j)$ to be the corresponding system stock and $\xi_N$, respectively. Note that

$$\xi_N(h_j) = \frac{2\beta}{1 - \beta} \cdot \frac{N}{(1 - u)(1 + N) + 2uj}$$

decreases in $j$. We obtain:

**Proposition 5** In an $N$-stage supply chain with equal leadtime for all stages,

$$s_N^a(h_1) \geq s_N^a(h_2) \geq ... \geq s_N^a(h_N).$$

That is, system stock increases if a substantially value-adding step moves to a downstream stage.

Proposition 5 is particular relevant if there is a substantially value-adding step in a process and it is possible to choose where to locate the step. Locating it at a downstream stage would lead to larger optimal system stock. This is intuitively clear: Adding substantial value at a downstream stage increases the holding cost for this echelon, which in turn leads to lower echelon inventory. However, this would need more inventories upstream in order to achieve the desired fill rate (see Proposition 1(a)). Thus, total system stock increases.
5.2 Stock Positioning and Stage Consolidation

After appropriate total system stock has been determined (say, let it be $s_N$), the next question is how to allocate it among stages (stock positioning). To gain insights into this issue, we examine the behaviors of the local base-stock levels $s_j$, $j = 1, ..., N$.

As noticed by Axsäter (2003b), “In practice it is still common to handle different stocks in a supply chain by single-echelon techniques. Quite often practitioners handle upstream and downstream stock points in a similar way. Generally, the distribution of the total stock between upstream and downstream installations is far from optimal.” Axsäter further points out that “under an optimal policy, what is typical is that upstream installations should have very low stocks compared to downstream installations.” The following proposition shows that our heuristic indeed helps to confirm this observation.

**Proposition 6** In a $N$-stage supply chain with equal holding cost and leadtime for all stages, we have

$$s_{j+1}^a < s_j^a, j = 1, ..., N - 1.$$

That is, the optimal installation base-stock level decreases as we move toward upstream.

For the general case where holding costs and leadtimes are not equal, the optimal installation base stock levels do not necessarily decrease in $j$. For example, in Table 6, $s_2^* < s_3^*$ in both variants 2 and 7. In this case, the sensitivity analysis results on $s_j^a$ can provide a guideline on how to allocate stocks.

**Proposition 7** For any $j = 1, ..., N$, we have:

1. $s_j^a$ increases in $\lambda, \beta, L_j$.
2. $s_j^a$ increases in $h_i, i > j$, but decreases in $h_j$.
3. $s_j^a$ increases in $\lambda, L_j$, but decreases in $h_j$.

Proposition 7 suggests that we should allocate more stock to a stage with longer leadtime and/or lower holding cost.

The issue of stage consolidation is highly related to stock positioning. Graves et al. (1998) pointed out that “In some instances, the best policy may be to remove the inventory between an upstream and downstream stage....

Within a multistage system, depending on the leadtimes and holding costs, it may be optimal to consolidate some of the stages.” We can shed some light on this observation by using NVH
Examining (12) and (13) reveals that \( s_j' \) will be close to zero if \( L_j \) is close to zero and \( \beta \) is sufficiently large.

### 5.3 Internal Fill Rates

In terms of implementation, a decentralized system is often favorable. This is because a centralized system usually requires seamless information systems and synchronized controls which are not always feasible. Whereas in a decentralized system, only local information and local control are needed. Thus, it is important to understand the linkages between stages through viable performance metrics.

One relevant measure is internal fill rates. The internal fill rate of a stage refers to the fill rate of this stage with respect to orders from its immediate downstream stage. This is equivalent to the fill rate of a subsystem, consisting of this stage and all upstream stages. Assuming the system implements the base-stock policy \((s_1, ..., s_N)\), and denoting the internal fill rate at stage \( j \) by \( \beta_j \), then

\[
\beta_j = R(s_j, s_{j+1}, ..., s_N).
\]

As mentioned in §1, there is a contradictory view on setting the internal fill rate needed for a downstream stage. Our goal is to use NVH to gain some insights. That is, we aim to understand how

\[
\beta_j = R(s_j', s_{j+1}', ..., s_N')
\]

relates to \( \beta \).

We first consider a two-stage system. The internal fill rate at stage 2 is

\[
\begin{align*}
\beta_2' &= P(D_2 < s_2') = P(D_2 < s_2' - s_1') \\
&= \left(1 + \exp \left( \sqrt{L_1}/\text{ln} \left( \frac{\beta(h_1 + h_2)}{1 - \beta} \right) - \sqrt{L_1}/\text{ln} \left( \frac{\beta(h_1 + h_2)}{1 - \beta} \right) \right) \frac{\beta(h_1 + h_2)}{1 - \beta} \text{ln} \left( \frac{\beta}{h_1 \text{ln} \left( \frac{\beta}{h_1 + h_2} \right)} \right)^{-1} \\
&= \left(1 + \frac{(\beta h_1 + h_2)}{(1 - \beta) h_1} \sqrt{L_1} \cdot \frac{(1 - \beta)(h_2 + h_1 L_1/L_{[1,2]})}{\beta(h_1 + h_2)} \right)^{-1} \cdot \text{ln} \left( \frac{\beta}{h_1 \text{ln} \left( \frac{\beta}{h_1 + h_2} \right)} \right)^{-1}.
\end{align*}
\]

We observe that the upstream stage fill rate depends on the holding costs of the two stages only through the ratio \( h_1/h_2 \). The formula is also consistent with a finding by Choi et al. (2003). That is, in a decentralized chain, to specify an upstream stage fill rate \( \beta_2 \) in order to guarantee a target system service level \( \beta \) we need to know the cost information at the upstream stage \( h_2 \).
Table 5: Values of $\beta_0$ and $N$: When $\beta > \beta_0$, $\beta_j < \beta$ for $j = 2, ..., N$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>10</th>
<th>100</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>0.40</td>
<td>0.51</td>
<td>0.57</td>
<td>0.61</td>
<td>0.70</td>
<td>0.92</td>
<td>0.98</td>
</tr>
</tbody>
</table>

We now show the following result:

**Proposition 8** In a 2-stage supply chain with equal echelon holding cost and leadtime, i.e., $h_1 = h_2$ and $L_1 = L_2$, we have $\beta_2^a < \beta$, $0 < \beta < 1$. That is, the internal fill rate at the upstream stage is strictly lower than the target fill rate.

More generally, for the $N > 2$ case, we have

**Proposition 9** In an $N$-stage supply chain with equal holding cost and leadtime for all stages, there exists $\beta_0 \geq 0$ such that for $\beta > \beta_0$, $\beta_j^a < \beta$ for all $j \geq 2$. That is, all upstream internal fill rates are lower than the target fill rate, provided the target fill rate is sufficiently high.

We demonstrate several examples of $\beta_0$ and $N$ in Table 5. Propositions 8 and 9 provide analytical justifications that the internal fill rates are smaller than the target fill rate in some special cases.

For general holding cost and leadtimes, it is difficult to obtain closed-form expression for internal fill rates. However, we observe similar results in Table 6 – the optimal internal fill rate is smaller than the desired system fill rate. These analytical studies and numerical observations provide evidence that it is not necessary to set a high internal fill rate from upstream to achieve a desired target fill rate.

Table 6 also reveals several interesting insights about how to set internal fill rates across stages. First, although it is tempting to conjecture that the internal fill rate decreases as $j$ increases, we find that this is not necessarily true. Second, in variants 1 to 4, as $h_j$ increases, $\beta_j^*$ decreases but $\beta_j^{*+1}$ increases. This is because higher holding cost at a stage “pushes” more stock upstream. Third, in variants 5 to 8, when $L_j$ increases, the internal fill rate $\beta_i^*$, $2 \leq i \leq j$ increases. This seems to suggest that when the leadtime at a stage lengthens, it is a better strategy to increase internal fill rates for all stages within the echelon.

6. **Summary and Applications to the Backorder-Cost Model**

In this paper, we have developed a closed-form approximation for the optimal base-stock levels in a serial base-stock system with a service constraint. This result enhances the transparency and
implementation of the optimal policy. In the process of such development, we have also constructed a bottom-up recursive procedure for evaluating echelon base-stock policies and provided lower bounds for the optimal echelon base-stock levels.

Using the approximate formulas we were able to demonstrate many qualitative properties of such systems. Some of these properties were casually observed or numerically demonstrated in the literature, but others appear to be new. For instance, in responding to the questions asked in the literature mentioned in the introduction, we find that (1) all internal fill rates are smaller than the target system fill rate as long as the latter is sufficiently high. Thus, requesting high internal fill rates may lead to significant overstocking. In addition, we show that (2) moving a high value-adding stage to a downstream location may increase the optimal system stock but reduce the optimal system cost; (3) the optimal system stock is larger when an upstream stage has a longer leadtime; (4) the optimal installation base-stock level decreases as we move upstream.

Recall that the closed-form approximation works for the BC model as well. Also, due to the connection of the BC and SC models, most of the insights gained from this study equally apply to the BC model. Likewise, many insights gained from Shang and Song (2003a) for the BC model are also valid here. However, differences between the two models do exist. For example, unlike

<table>
<thead>
<tr>
<th>Variant</th>
<th>$h_1 = 2.5$</th>
<th>$h_2 = 2.5$</th>
<th>$h_3 = 2.5$</th>
<th>$h_4 = 2.5$</th>
<th>$L_1 = 1$</th>
<th>$L_2 = 1$</th>
<th>$L_3 = 1$</th>
<th>$L_4 = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>0.99</td>
<td>0.9905</td>
<td>0.9918</td>
<td>0.6499</td>
<td>0.5705</td>
<td>0.4940</td>
<td>0.7162</td>
<td>0.6288</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.906</td>
<td>0.9111</td>
<td>0.3667</td>
<td>0.4939</td>
<td>0.2506</td>
<td>0.4940</td>
<td>0.4335</td>
</tr>
<tr>
<td>Variant 1</td>
<td>0.99</td>
<td>0.9902</td>
<td>0.9916</td>
<td>0.9467</td>
<td>0.7416</td>
<td>0.5934</td>
<td>0.7523</td>
<td>0.6893</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.9013</td>
<td>0.9040</td>
<td>0.7241</td>
<td>0.6824</td>
<td>0.4051</td>
<td>0.5600</td>
<td>0.4335</td>
</tr>
<tr>
<td>Variant 2</td>
<td>0.99</td>
<td>0.9902</td>
<td>0.9903</td>
<td>0.2297</td>
<td>0.3595</td>
<td>0.9103</td>
<td>0.5934</td>
<td>0.7851</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.9004</td>
<td>0.9004</td>
<td>0.0732</td>
<td>0.0709</td>
<td>0.6543</td>
<td>0.5934</td>
<td>0.6288</td>
</tr>
<tr>
<td>Variant 3</td>
<td>0.99</td>
<td>0.9903</td>
<td>0.9919</td>
<td>0.3595</td>
<td>0.5383</td>
<td>0.5934</td>
<td>0.2177</td>
<td>0.8893</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.9045</td>
<td>0.9149</td>
<td>0.5129</td>
<td>0.2603</td>
<td>0.1976</td>
<td>0.0738</td>
<td>0.7851</td>
</tr>
<tr>
<td>Variant 4</td>
<td>0.99</td>
<td>0.9905</td>
<td>0.9920</td>
<td>0.6499</td>
<td>0.5980</td>
<td>0.4940</td>
<td>0.4051</td>
<td>0.6288</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.9006</td>
<td>0.9089</td>
<td>0.3667</td>
<td>0.3033</td>
<td>0.2506</td>
<td>0.1884</td>
<td>0.4335</td>
</tr>
<tr>
<td>Variant 5</td>
<td>0.99</td>
<td>0.9900</td>
<td>0.9916</td>
<td>0.2656</td>
<td>0.4549</td>
<td>0.3204</td>
<td>0.3711</td>
<td>0.6288</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.9006</td>
<td>0.9220</td>
<td>0.0951</td>
<td>0.2656</td>
<td>0.1660</td>
<td>0.3204</td>
<td>0.6288</td>
</tr>
<tr>
<td>Variant 6</td>
<td>0.99</td>
<td>0.9902</td>
<td>0.9916</td>
<td>0.8398</td>
<td>0.8490</td>
<td>0.4051</td>
<td>0.3711</td>
<td>0.4335</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.9000</td>
<td>0.9149</td>
<td>0.4732</td>
<td>0.7202</td>
<td>0.1659</td>
<td>0.1659</td>
<td>0.6288</td>
</tr>
<tr>
<td>Variant 7</td>
<td>0.99</td>
<td>0.9901</td>
<td>0.9907</td>
<td>0.7120</td>
<td>0.5909</td>
<td>0.7843</td>
<td>0.8678</td>
<td>0.2381</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.9000</td>
<td>0.9072</td>
<td>0.4493</td>
<td>0.5278</td>
<td>0.5319</td>
<td>0.7028</td>
<td>0.4335</td>
</tr>
<tr>
<td>Variant 8</td>
<td>0.99</td>
<td>0.9902</td>
<td>0.9916</td>
<td>0.7362</td>
<td>0.5784</td>
<td>0.7483</td>
<td>0.7483</td>
<td>0.9108</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.9000</td>
<td>0.9123</td>
<td>0.4317</td>
<td>0.5117</td>
<td>0.4236</td>
<td>0.5484</td>
<td>0.5660</td>
</tr>
</tbody>
</table>

Table 6: Optimal and NVH internal fill rates
in the BC model, in the SC model: (a) the optimal solution is rather insensitive to the echelon holding cost rate at the topmost stage \((h_N)\); and (b) the optimal system cost is sensitive to stock positioning.

**Acknowledgments:** We would like to thank Tamer Boyaci, Guillermo Gallego and Paul Zipkin for helpful comments on an earlier version of this paper. We also thank the editor, the senior editor and the referees for constructive suggestions. The research was supported in part by National Science Foundation under grant DMI-0353552 and Award No. 70328001 from the National Natural Science Foundation of China.

**Appendix: Proofs**

**Proposition 6**

Without loss of generality, we assume \(L_{[1,N]} = 1\) and \(h_{[1,N]} = 1\). Thus equal leadtimes and holding costs implies \(L_j = 1/N\) and \(h_j = 1/N\) for all \(j\). In this case, (12) becomes

\[
\xi_j = \frac{2(j\beta + (N - j))}{(1 - \beta)(j + 1)}
\]

and (13) becomes

\[
s_j^{\prime a} = \frac{\lambda}{N} + 0.617 \sqrt{\frac{\lambda}{N}} \left[ \sqrt{j} \ln \left( \frac{2(j\beta + (N - j))}{(1 - \beta)(j + 1)} \right) - \sqrt{j-1} \ln \left( \frac{2(j-1)\beta + (N - j + 1))}{j(1 - \beta)} \right) \right].
\]

To show \(s_j^{\prime a} < s_j^{\prime a+1}\), it is equivalent to show \(f(j) \overset{\text{def}}{=} \sqrt{j} \ln(\xi_j)\) is concave in \(j\).

Define the continuous extension of \(f(j)\) on the real interval \([1, N]\) as

\[
f(x) = \sqrt{x} \ln \left( \frac{2(\beta x + (N - x))}{(1 - \beta)(1 + x)} \right) = \sqrt{x} \left( \ln \left( \frac{2}{1 - \beta} \right) + \ln((\beta - 1)x + N) - \ln(1 + x) \right).
\]

We now take the second derivative of each term in \(f(x)\) with respect to the continuous variable \(x\). Let \(f_1(x) \overset{\text{def}}{=} \sqrt{x} \ln \left( \frac{2}{1 - \beta} \right)\). Then, \(f_1''(j) = -\frac{1}{4}x^{-3/2}\ln(\beta x + N - x) < 0\). Next, let \(f_2(x) \overset{\text{def}}{=} \sqrt{x} \ln(\beta x + N - x)\). Then,

\[
f_2''(x) = -\frac{1}{4}x^{-3/2}\ln(\beta x + N - x) + \frac{1}{2}x^{-1/2} \frac{\beta - 1}{(\beta x + N - x)}

+ (\beta - 1) \left( \frac{1}{2}x^{-1/2}(\beta x + N - x) - (\beta - 1)\sqrt{x} \right) < 0.
\]
Further, let \( f_3(x) \overset{\text{df}}{=} \sqrt{x}ln(x + 1) \). Then
\[
f''_3(x) = \frac{1}{\sqrt{x}} \left( \frac{4x - (1 + x)^2ln(1 + x)}{4x(1 + x)^2} \right).
\]
Thus, \( f''_3(x) < 0 \) when \( x \geq 2 \). Therefore, we have shown that when \( x \geq 2, f(x) \) is concave in \( x \). This implies that \( f(j) - f(j - 1) > f(j + 1) - f(j), j = 2, ..., N - 1 \). Thus, we have \( s_2^a > s_3^a > ... > s_N^a \).

Finally, from the definition,
\[
s_1^a - s_2^a = 0.617 \sqrt{\frac{\lambda}{N}} \left( 2ln \left( \frac{\beta + N - 1}{1 - \beta} \right) - \sqrt{2ln \left( \frac{2(2\beta + N - 2)}{3(1 - \beta)} \right)} \right).
\]
Since \( \frac{\beta + N - 1}{1 - \beta} > \frac{2(2\beta + N - 2)}{3(1 - \beta)} > 1 \), \( s_1^a - s_2^a > 0 \). This completes the proof.

**Proposition 7**

We only prove \( s_j^a \) decreases in \( h_j \); the rest of the results are straightforward from (11) and (13) so the proof is omitted. To show the property, it is sufficient to show \( \xi_j \) decreases in \( h_j \). But this is implied by
\[
\frac{\partial \xi_j}{\partial h_j} = \frac{\beta(1 - \beta) \sum_{i=1}^{j-1} h_i(L_{[1,i]}/L_{[1,j]}) - h_{[1,j-1]} - h_{[j+1,N]}}{[(1 - \beta)h_j + (1 - \beta) \sum_{i=1}^{j-1} h_i(L_{[1,i]}/L_{[1,j]})]^2} < 0.
\]

**Proposition 8**

Without loss of generality, we assume \( h_1 = h_2 = 1 \) and \( L_1 = L_2 = 1 \). From (15), this is equivalent to showing
\[
\beta_2^a - \beta = \left( 1 + \frac{1 + \beta}{1 - \beta} \left( \frac{3(1 - \beta)}{4\beta} \right)^{\sqrt{2}} \right)^{-1} - \beta < 0
\]
for \( 0 < \beta < 1 \). This can be further simplified by showing
\[
\frac{1 + \beta}{\beta^{\sqrt{2} - 1} \cdot (1 - \beta)^{2 - \sqrt{2}}} - \left( \frac{4}{3} \right)^{\sqrt{2}} \geq 0.
\]
is positive. Note that the first term in (16) is a convex function of \( \beta, 0 < \beta < 1 \), and its minimizer is \( \beta = (\sqrt{2} - 1)/(3 - \sqrt{2}) \). At this minimizer, (16) achieves its lowest value \( 2.626 - 1.502 > 0 \). The result follows immediately.
**Proposition 9**

We introduce a concept of *nominal fill rate* before we proceed. The nominal fill rate is the fill rate the stage would achieve if it had ample supply. Denote the nominal fill rate by using the base-stock level \( s_j' \) at stage \( j \) to be \( \beta_j' \), then

\[
\beta_j' = P(D_j < s_j').
\]

Obviously, the nominal fill rate at each stage is larger than the real internal fill rate, i.e., \( \beta_j' \geq \beta_j \), for all \( j \).

From Proposition 6, it is easy to see that \( \beta_2' \geq \beta_3' \geq ... \geq \beta_N' \). Note that \( \beta_j \leq \beta_j' \leq \beta_2' \). It is sufficient to show \( \beta_2' < \beta \) for \( \beta > \beta_0 \). From the expression

\[
\beta_2' = \mathbb{P}(D_2 < s_2') = \left( 1 + \left( \frac{\beta + N - 1}{1 - \beta} \right) \left( \frac{3(1 - \beta)}{2(2\beta + N - 2)} \right)^{\frac{\sqrt{2}}{2}} \right)^{-1},
\]

this is equivalent to showing

\[
\left( \frac{\beta + N - 1}{1 - \beta} \right) \left( \frac{3(1 - \beta)}{2(2\beta + N - 2)} \right)^{\frac{\sqrt{2}}{2}} > \frac{1 - \beta}{\beta}, \quad \beta > \beta_0
\]

or

\[
w(\beta) \overset{\text{def}}{=} \beta(\beta + N - 1) - \left( \frac{2}{3} \right)^{\frac{\sqrt{2}}{2}} (2\beta + N - 2)^{\frac{\sqrt{2}}{2}}(1 - \beta)^{2-\sqrt{2}} > 0, \quad \beta > \beta_0. \quad (17)
\]

We now argue there indeed exists \( \beta_0 \) so that (17) holds. Note that \( \lim_{\beta \to 1} w(\beta) = N \). So, given \( \epsilon = \frac{N}{2} > 0 \), there exists \( \delta_\epsilon > 0 \) such that for all \( |\beta - 1| < \delta_\epsilon \), we have \( |w(\beta) - N| < \epsilon = \frac{N}{2} \). This implies that, for \( \beta > \beta_0 = 1 - \delta_\epsilon, \ w(\beta) > \frac{N}{2} > 0 \), proving (17).

**References**


