Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Capacitated Multiechelon Inventory Systems: Policies and Bounds

Woonghee Tim Huh, Ganesh Janakiraman, Mahesh Nagarajan

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 18, No. 4, Fall 2016, pp. 570–584
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs®
http://dx.doi.org/10.1287/msom.2016.0588
© 2016 INFORMS

Capacitated Multiechelon Inventory Systems: Policies and Bounds

Woonghee Tim Huh
Operations and Logistics Division, Sauder School of Business, University of British Columbia, Vancouver, Canada V6T 1Z2,
tim.huh@sauder.ubc.ca

Ganesh Janakiraman
Naveen Jindal School of Management, The University of Texas at Dallas, Richardson, Texas, 75080,
ganesh@utdallas.edu

Mahesh Nagarajan
Operations and Logistics Division, Sauder School of Business, University of British Columbia, Vancouver, Canada V6T 1Z2,
mahesh.nagarajan@sauder.ubc.ca

We study a periodically reviewed multiechelon serial inventory system with a capacity constraint on the order quantity at each stage. The cost criterion we use to evaluate inventory policies for this system is the sum of the expected long-run average holding and shortage costs. It is well known that for this problem, characterizing the structure of the optimal policy and computing it are very difficult. We consider the use of echelon base-stock policies for our system (even though they are known to be suboptimal) and propose algorithms for finding base-stock levels that are easy to understand and implement. We derive bounds on the ratios between the costs achieved by our algorithms and the optimal costs (over all policies). For light-tailed demand distributions, our algorithms are shown to be asymptotically optimal in the sense that our bounds are close to one in high service-level environments. Our computational investigations reveal that our algorithms perform well even under modest service levels.

Keywords: multiechelon inventory; policies; capacity
History: Received: April 5, 2012; accepted: January 25, 2016. Published online in *Articles in Advance* September 13, 2016.

1. Introduction

One of the fundamental problems in supply chain management is that of managing the flows of inventories through large-scale supply chains "optimally." We study this problem using a prototypical model of a supply chain. In this model, a multiechelon serial inventory system with capacity constraints at every stage and stochastic external demands at the downstream stage is reviewed periodically and production/procurement/shipping decisions are made with the goal of minimizing holding and shortage costs.

The seminal paper by Clark and Scarf (1960) presented an elegant solution to this problem by ignoring capacity constraints: an echelon base-stock policy was shown to be optimal and an algorithm to compute that policy was presented. Recognizing the practical importance of these constraints, Glasserman and Tayur (1995) sought to include them in their influential paper. With this inclusion, an elegant solution (i.e., an easily describable optimal policy) was no longer possible. But this is merely a theoretical issue. Practicing managers are more likely to use simple policies, in this case echelon base-stock policies, even though they are theoretically suboptimal. What these managers need is well grounded guidance on how the echelon base-stock levels should be chosen. Glasserman and Tayur (1995) focus on this problem. They present a simulation-based gradient estimation procedure (*Infinitesimal Perturbation Analysis* or IPA, in short) and use it in a *gradient search* technique to prescribe the base-stock levels. However, there are no known theoretical or computational results on the cost performance of the base-stock policy resulting from this procedure relative to the optimal policy. As a starting point for our paper, we view Clark and Scarf (1960) and Glasserman and Tayur (1995) as the theoretical state of the art and practical state of the art for multiechelon problems, respectively. The former ignores the practical reality of capacity constraints but presents an algorithm to find the optimal policy. The latter incorporates capacity constraints and presents an algorithm that can be implemented in practice (i.e., it runs in a reasonable amount of time) but offers no insights or guarantees on the optimality gap (i.e., cost difference between this policy and the optimal policy).

Our objective in this paper is to find intuitively appealing approximate polices that are based on theoretical fundamentals, are easy to implement, and perform well in practice. Toward this end, we develop echelon base-stock policies that are canonically constructed in that they organically arise from the literatures on single-stage capacitated systems and multistage *uncapacitated* serial systems, *à la* Clark and Scarf (1960). In addition, we show that our approximations are asymptotically optimal as the backorder cost parameter grows large (or equivalently, the system operates at high service levels). The choice of such an asymptotic regime is motivated by the observation that many supply chains maintain very high service levels at the external-customer-facing stage. Numerous studies (for example, the Coca-Cola Retailing Research Councils 1996 and Gruen et al. 2002) suggest that service levels in retail, as measured by in-stock percentages tend to be in the 85%–99% range for various product categories. In particular, service levels for most products in grocery supply chains tend to be in the 90%–99% range. See also the discussion on service levels in the text, (Cachon and Terwiesch 2009). Using an extensive set of numerical experiments, we demonstrate empirically that our policies perform well even at moderately high service levels. In performing this latter task, we compare our policy to lower bounds on the optimal cost that we develop in this paper. We note that for capacitated serial systems, to the best of our knowledge, useful lower bounds on the optimal cost are not available in the literature. Thus, our construction of such bounds may be of use to future research on this topic.

1.1. Related Literature

We now present a brief review of the related literature, starting with single-stage systems, then multiechelon uncapacitated serial systems, and finally multiechelon capacitated serial systems.

For single-stage inventory systems with capacity constraints, Federgruen and Zipkin (1986a, b) establish the optimality of base-stock policies and Tayur (1992) provides an algorithm to compute the optimal base-stock level.

For multiechelon systems without capacity constraints, Federgruen and Zipkin (1984) extend the results of Clark and Scarf (1960) to infinite horizon models and provide a simple algorithm for computing the optimal policy. Shang and Song (2003) derive upper and lower bounds that are easy to calculate, expressed using newsvendor formulae, on the optimal base-stock levels. They show through numerical experiments that the heuristic of using the averages of these bounds as echelon base-stock levels is near optimal.

Focusing on multiechelon serial inventory systems with capacity constraints, Parker and Kapuscinski (2004) is the first paper to provide structural results on the optimal policy. They study a two-echelon system with identical capacities at both stages and an upstream lead time of one period. They show that a *modified echelon base-stock* (MEBS) policy is optimal for this system. An MEBS policy is nothing but an echelon base-stock or order-up-to policy in which orders by any stage are truncated by the capacity limit at that stage. It is known that the MEBS policy need not be optimal for more general systems; see, for instance, Speck and van der Wal (1991) and Janakiraman and Muckstadt (2009). Angelus and Zhu (2009) also study the multiechelon system with identical capacities at all stages; they derive the optimal policies in two modified DPs and use them as heuristics for the original problem and computationally test them.

A third stream of literature that is directly relevant to this paper is the one that provides approximation algorithms for serial capacitated inventory systems. For reasons mentioned earlier, this literature confines itself to finding heuristics within the class of base-stock policies. Glasserman and Tayur (1995) use IPA in a gradient descent procedure to find *locally optimal* echelon base-stock levels. They initiate their procedure from different starting solutions and pick the terminating solution with the least cost. Glasserman (1997) studies both single and multiechelon inventory systems, all operated using echelon base-stock policies. For single-stage systems, he develops easily computable bounds on the optimal base-stock level and shows that the difference between the optimal base-stock level and one of these bounds converges to zero as the backorder cost parameter grows infinitely large. His analysis is based on approximations of the tail probabilities of the *shortfall* process, where the shortfall is the difference between the base-stock level and the inventory level. In our paper, we also demonstrate asymptotic bounds that are a function of the backorder cost. However, we note that our asymptotic analysis is very different from the above work. For one, Glasserman's bounds only pertain to single-echelon systems and cannot be extended to multiechelon systems. In a subsequent paper, Glasserman and Tayur (1996) use an approximation for the cost function in multiechelon systems to develop a heuristic procedure for determining echelon base-stock levels. We discuss this paper in some detail in our next section.

1.2. Our Contribution

Two special cases of capacitated serial inventory systems have been solved satisfactorily in the literature, i.e., the *uncapacitated* serial system and the *single-stage* capacitated system. Although the optimal policy of this system cannot be inferred from the two special cases, since echelon base-stock policies are easy to understand and implement, the idea of using an algorithm for computing echelon base-stock levels by combining the approach of Tayur (1992) with that of Federgruen and Zipkin (1984) or Shang and Song (2003) is intuitively appealing. We propose such algorithms for finding echelon base-stock levels.

We contrast our work to the existing numerical approaches in the literature. The first approach is the one in Glasserman and Tayur (1995). This involves simply finding the "best" base-stock policy and using this as a heuristic. The limitation of this method is that as discussed earlier, there is no real guarantee that this procedure will converge to a globally optimal base-stock vector. The second approach, the one in Glasserman and Tayur (1996), uses asymptotic tail approximations of the shortfall distribution and derives an approximation of the cost function. They use this to derive their heuristic. If one looks carefully at the construction of their heuristic, its apparent ease strictly depends on the underlying demand distribution. Two crucial parameters that are needed in their construction are nontrivial to calculate. First, the parameter γ that involves solving for the root of a function that is related to a moment generating function, is not guaranteed to exist for all demand distributions. Second, the parameter $\mathbf{C}$ requires at least a numerical integration for most demand distributions. Therefore, their heuristic is perhaps not too intuitive, may not be computable in many instances, and not easy to implement as the ones we propose in our paper. For these reasons, we do not compare our results with the heuristic they provide.

The merits of our approach are the following: (i) From a theoretical perspective, we derive bounds on the ratios between the costs achieved by our algorithms and the optimal costs over all policies (not just base-stock policies). Our bounds are close to one in high service-level environments, i.e., when we consider a sequence of problems in which the backorder cost parameter keeps growing large, our algorithms are asymptotically optimal for this sequence. We note that the construction of our algorithms do not involve any asymptotic approximations. The only role high service levels play is that they provide a theoretical limiting regime, which also has practical value because many firms do operate at high service levels, under which optimality can be demonstrated. (ii) We computationally evaluate the performance of our algorithms by evaluating their costs and comparing them with the cost of the best base-stock policy and with a lower bound on the optimal cost on a large test bed of problem samples; the service levels in our test bed range from 67% to 99%, i.e., our computational investigation is not restricted to high service level environments. Our computational investigation reveals that the cost of the best base-stock policy is merely 0.6% higher, on an average, than our *lower bound* on the optimal cost. This formally confirms a perhaps optimistic suspicion that researchers have cherished that base-stock policies are effective (even though suboptimal) for managing capacitated serial inventory systems. Moreover, on an average, the best of our policies for finding the base-stock levels achieves a cost that is only about 1% more, on an average, than the cost of the best base-stock policy. Our algorithms only involve newsvendor computations on the convolutions of appropriate demand random variables and certain random variables called shortfalls, which are necessary to capture even for capacitated *single-location* problems. (iii) Finally, our heuristics, as will be clear, are easier to calculate than existing approaches. As an illustration, in contrast to a standard search procedure, the cost function in our approach does not need to be evaluated at all. Moreover, because of the intuitive construction of the base-stock levels, it is reasonably easy to implement in any real setting. (iv) Our heuristics are robust in that if any simplification or advance is made in the computation of policies for uncapacitated serial systems, we can immediately construct corresponding heuristic policies for the capacitated serial systems, and that any computational method or analytical properties for the shortfall distributions of capacitated single-location systems can be applied to our heuristics for the capacitated serial systems.

We note that, in conducting this numerical exercise, we make use of two lower bounds on the optimal cost of capacitated serial systems that we develop in this paper. To our knowledge, these are the first lower bounds documented for these systems. Our computational results show that they are close to the optimal cost. These lower bounds are important because computing the optimal policy using dynamic programming is computationally prohibitive for these systems, in general; thus, our lower bounds serve as an easily computable benchmark for measuring the quality of the algorithms we develop in this paper as well as any other algorithms that may be developed in the future.

The remainder of the paper is organized as follows. In Section 2, we formally describe our model in detail. We propose three new policies in Section 3, all of the echelon order-up-to type, and then in Section 4 we introduce two lower bounds on the cost. In Section 5 we describe our results on the asymptotic behavior of these policies as the unit backorder cost approaches infinity. We present our numerical experiments in Section 6 and conclude in Section 7.

2. Model

We describe our model in detail in this section. Since the same model is also studied in Huh et al. (2010), our presentation and notation here closely follows the corresponding material in their paper.

We consider a capacitated multiechelon serial inventory system under periodic review. Let $N \geq 2$ be the number of stages in the system. Stages are numbered from downstream to upstream, i.e., stage 1 is the most downstream stage and stage N is the most upstream stage. External consumer demand occurs in the most downstream stage (stage 1), and the most upstream stage (stage N) orders from an outside source with ample supply, which we refer to as stage $N+1$. Each stage $j \in \{1, \ldots, N\}$ orders from its immediate upstream stage $j+1$, and has an ordering capacity of CAP^j units. We allow the possibility that CAP^j is infinite. We assume throughout that any stage has a higher capacity than the stage immediately above it, i.e.,

$$\text{CAP}^1 \geq \text{CAP}^2 \geq \cdots \geq \text{CAP}^N. \tag{1}$$

We note that when lead times are deterministic, this assumption is without loss of generality when the inventory holding costs at downstream stages are higher than those at upstream stages—this is because it is never optimal for an upstream stage to produce/order more than the capacity of any downstream stage. We also assume that all the lead times in this system are exactly one period in length. Again, for deterministic lead times, this is without loss of generality as a lead time of k periods between two stages can be modeled using $k-1$ fictitious stages between these two stages with holding costs identical to the downstream stage. Please see Huh et al. (2010). Let $(D_t \mid t = 0, 1, \ldots)$ denote the sequence of nonnegative random variables representing demand. We assume that this sequence is independently and identically distributed, with the same distribution as some random variable D (unless otherwise noted). Let μ denote the mean demand, i.e., $\mu = E[D]$. As in all such studies, in order for the system to be reasonably maintained with finite cost, we assume that the bottleneck capacity exceeds the expected demand in a period, i.e.,

$$\text{CAP}^N > \mu. \tag{2}$$

In what follows, we describe the dynamics of the system and some notation that we use. The sequence of events in each period t is given as follows. (1) Each echelon j receives the delivery due to arrive in period t, which is the same as the quantity ordered in the previous period, q^j_{t-1}. Any backlogged demand at stage 1 is immediately satisfied to the extent possible. (2) The manager observes the amount of inventory in each echelon j. Let $I^1_t \in (-\infty, \infty)$ denote the net inventory (on-hand inventory minus backorders) at stage 1, and let $I^j_t \in [0, \infty)$ denote the inventory available at stage j, for $j = 2, \ldots, N$. (3) The manager makes the ordering decision q^j_t for each stage $j = 1, \ldots, N$. We require that q^j_t can neither exceed the ordering capacity CAP^j nor exceed the amount of inventory available in the immediately upstream stage, which is I^{j+1}_t. (For simplicity, we let $I^{N+1}_t = \infty$ for all t.) (4) Demand D_t is realized and is satisfied to the extent possible. Any excess demand is backordered to the next period. The dynamics of the system can be represented as follows:

$$I^j_{t+1} = \begin{cases} I^j_t + q^j_t - q^{j-1}_t & \text{if } j = 2, \ldots, N, \\ I^j_t + q^j_t - D_t & \text{if } j = 1. \end{cases}$$

It is convenient to introduce some more definitions about the dynamics of the system that are commonly used in our exposition. Let IP^j_t and e^j_t denote the echelon-j inventory position after ordering in period t and the echelon-j net-inventory at the end of period t after demand is realized, respectively, i.e.,

$$\text{IP}^j_t = (I^1_t + \cdots + I^j_t) + q^j_t \tag{3}$$

and

$$e^j_t = (I^1_t + \cdots + I^j_t) - D_t. \tag{4}$$

In addition, let

$$B_t = [e^1_t]^- = [I^1_t - D_t]^- \tag{5}$$

denote the amount of backordered demand at the end of period t.

Next, we discuss the costs we consider in our model. Let $b > 0$ denote the unit backlogging cost, and let $H^j > 0$ denote the unit holding cost at stage j. We assume $H^1 \geq \cdots \geq H^N$, which is a standard assumption in serial inventory systems. For each $j = 1, \ldots, N$, let $h^j = H^j - H^{j+1}$ denote the echelon-j holding cost, where $H^{N+1} = 0$. Then, the period-t cost is given by

$$c_t = b \cdot [I^1_t - D_t]^- + H^1 \cdot [I^1_t - D_t]^+ + \sum_{j=2}^{N} H^j \cdot I^j_t \ .$$

We can rewrite this expression as

$$\begin{aligned} c_t &= b \cdot B_t + \sum_{j=1}^{N} h^j \cdot [e^j_t + B_t] \\ &= (b + H^1) \cdot B_t + \sum_{j=1}^{N} h^j \cdot e^j_t. \end{aligned} \tag{6}$$

To prove the first equality of (6), we use $H^j = \sum_{j'=1}^{j} h^{j'}$ and Equations (3) and (4); note that the amount of physical inventory held in echelon j (i.e., not considering backorders, if any) at the end of period t is $e^j_t + B_t$. We now introduce a modification to the definition of the single-period cost by performing a shift of indices to the second term in (6), which is common in multiechelon inventory theory (see, for example, Chen and Zheng 1994). Let

$$C_t = (b+H^1)\cdot B_t + \sum_{j=1}^{N} h^j \cdot e^j_{t-(j-1)}. \tag{7}$$

The intuition behind defining the cost C_t in this way is that the backorders in period t are most directly influenced by the ending inventory in echelon j in period $t-(j-1)$. Thus, by capturing the holding cost for j in period $t-(j-1)$, we take into account all the holding costs associated with the echelon inventory levels that affect the backorder cost in period t. This accounting proves to be useful.

Next, we define the performance measure we use to evaluate any policy, π. We use the superscript π to be explicit about the dependence of the quantities of interest on the policy π that is used. Our performance measure is the long-run average cost of this system, i.e., $\limsup_{T\to\infty} \sum_{t=0}^{T} E[C_t^\pi \mid \mathbf{I}_0]/(T+1)$, where $\mathbf{I}_0$ represents the starting state of the system, that is, $\mathbf{I}_0 = (I_0^1, \ldots, I_0^N)$. In practice, for any given policy π, this long-run average cost is computed by simulating a large number of sample paths or realizations of demands over thousands of periods and taking the average cost per period averaged over all the simulated sample paths.

The optimal cost of this system is

$$C^*(\mathbf{I}_0) = \inf_{\pi} \limsup_{T\to\infty} \frac{\sum_{t=0}^{T} E[C_t^\pi \mid \mathbf{I}_0]}{T+1}. \tag{8}$$

An *echelon base-stock policy* is an ordering policy characterized by a vector of echelon base-stock levels $\mathbf{S} = (S^1, \ldots, S^N)$, and the ordering quantity at each stage is chosen to bring the echelon inventory position as close to its echelon base-stock level as possible. (Recall that the ordering quantity is constrained by its ordering capacity and also by the amount of available inventory at its immediate upstream stage.) Thus,

$$q_t^j = \min\big([S^j - (I_t^1 + \cdots + I_t^j)]^+, \mathrm{CAP}^j, I_t^{j+1}\big).$$

We also call this the *echelon order-up-to* **S** *policy*.

When this policy is used, we introduce an additional assumption that the initial inventory levels $(I_0^1, \ldots, I_0^N)$ satisfy, for each j,

$$I_0^1 + \cdots + I_0^j = S^j. \tag{9}$$

This assumption is without loss of generality because the choice of the initial inventory levels does not affect the long-run average cost. (This follows from the ergodicity properties of this system, which are discussed in Glasserman and Tayur 1994 and Huh et al. 2010. For example, we can use the results of these papers to show that we can reach the assumed state in a random amount of time with finite expectation. Thus, we can ignore the costs incurred from period zero until the first period in which the assumed state is reached. This effectively makes the assumed state the starting state.)

3. Three Algorithms for Finding Base-Stock Levels

For uncapacitated systems, Clark and Scarf (1960) have shown the optimality of echelon order-up-to policies for the finite horizon model, and they have also provided an algorithm to compute the optimal policy. This algorithm requires the use of the cost-to-go functions of the dynamic program. Federgruen and Zipkin (1984) have extended this work to infinite horizon models with stationary demands. Even though the infinite horizon model is computationally simpler to handle than the finite horizon, the algorithm however involves recursive computations that are somewhat difficult to implement on a simple tool such as a spreadsheet. Recently, Shang and Song (2003) have proposed two heuristic policies that use only the newsvendor formula for computing the echelon order-up-to levels. These heuristics can easily be implemented on spreadsheets and their dependence on the parameters of the system is transparent. Most importantly, Shang and Song (2003) have shown computationally that the cost performance of these newsvendor-based heuristic policies is quite close to that of the optimal policy. All of the above discussion is only applicable to uncapacitated systems. Motivated by the above policies for uncapacitated systems, we propose in this section, three echelon order-up-to policies for capacitated serial systems. In each of the three policies we propose, we use the shortfall process to make adjustments to the demand distribution in order to account for the fact that our system is capacitated.

Let V_t^j denote the shortfall in period t in a single-stage system with capacity CAP^j, i.e., the process $\{V_t^j\}$ is defined by the following dynamics: for each j,

$$V_{t+1}^j = \max\big(0, V_t^j + D_t - \mathrm{CAP}^j\big), \tag{10}$$

where $V_0^j = 0$. Note that this shortfall process is a reflected random walk where each increment is given by $D_t - \mathrm{CAP}^j$. For simplicity of exposition, we adopt the convention that all shortfalls and demands are zero for $t<0$, i.e., $D_t = 0$ and $V_t^j = 0$ for all $t<0$ and all j. We use the following notation:

$$D[t_1, t_2] = \begin{cases} D_{t_1} + D_{t_1+1} + \cdots + D_{t_2} & \text{if } t_1 \le t_2, \\ 0 & \text{otherwise.} \end{cases}$$

We briefly discuss how the single-stage shortfall processes $\{V_t^j\}$ discussed above are related to the true shortfall processes at each echelon. The echelon-j shortfall is the difference between the echelon order-up-to level (target), S^j, and the echelon inventory position after ordering (actual) at the beginning of each period. Formally, let $\tilde{V}_t^j$ denote the echelon-j shortfall in period t, defined as

$$\tilde{V}_t^j = S^j - (I_t^1 + \cdots + I_t^j) - q_t^j = S^j - \mathrm{IP}_t^j. \qquad (11)$$

From Huh et al. (2010), the echelon-j shortfall in period t is related to the single-stage shortfalls at stages $j, \dots, N$ and the base-stock levels at those echelons, as follows:

$$\tilde{V}_t^j = \max_{i=j,\dots,N} \{V_{t-i+j}^i + (D_{t-i+j} + \cdots + D_{t-1}) - (S^i - S^j)\}. \qquad (12)$$

Note that for the topmost echelon ($j = N$), the above expression is simply V_t^N, i.e., the echelon shortfall process is the same as the single-stage shortfall process. The echelon-$(N-1)$ shortfall is the greater of the single-stage shortfall V_t^{N-1} and $V_{t-1}^N + D_{t-1} - (S^N - S^{N-1})$. Thus, the echelon-$(N-1)$ shortfall is at least V_t^{N-1}. Since $V_t^{N-1} = \max(0, V_{t-1}^{N-1} + D_t - \mathrm{CAP}^{N-1})$ by (10), the echelon-$(N-1)$ shortfall and V_t^{N-1} are more likely to coincide with V_t^{N-1} if CAP^{N-1} is small, and the base-stock level difference $S^N - S^{N-1}$ is large.

MFZ Policy. Our first policy is an extension of the algorithm of Federgruen and Zipkin (1984). We refer to this policy as the *Modified Federgruen-Zipkin* (MFZ) policy. We need the following definitions to describe this policy. Recall that $h^j = H^j - H^{j+1}$, where $H^{N+1} = 0$, is the echelon holding cost. Let the random variable $\mathcal{D}_k$ be the convolution $D_1 + \cdots + D_k$ for any $2 \le k \le N+1$. The optimal echelon base-stock vector of Federgruen and Zipkin (1984) for the uncapacitated system is $(S^{1*}, \dots, S^{N*})$, which is defined by the following recursion: for $j \in \{1, \dots, N\}$,

$$g^j(y) = \begin{cases} h^1 \cdot (y - 2 \cdot E[D]) + (b + H^1) \cdot E[(\mathcal{D}_2 - y)^+], \\ \quad \text{if } j = 1; \\ h^j \cdot (y - 2 \cdot E[D]) + E[g^{j-1}(\min(y - D, S^{(j-1)*}))], \\ \quad \text{if } j = 2, \dots, N, \end{cases} \qquad (13)$$

and

$$S^{j*} = \arg\min_y g^j(y).$$

Note that each of the g^j functions is convex.

To describe the MFZ policy, we introduce for $j \in \{1, \dots, N\}$,

$$\tilde{S}^j = \arg\min_y E[g^j(y - V_\infty^j)], \qquad (14)$$

where V_∞^j is the stationary distribution of the single-stage shortfall process, $\{V_t^j\}$, defined in (10). This distribution can be computed, in practice, by simulating the demands and using the recursion in (10)—see also discussions in the literature on single-stage capacitated systems (Tayur 1992 and 1997). Roundy and Muckstadt (2000) provide excellent approximations of this distribution—these approximations can be used in our policies directly. Let $\tilde{\mathbf{S}} = (\tilde{S}^1, \dots, \tilde{S}^N)$. We call the echelon order-up-to $\tilde{\mathbf{S}}$ policy as the MFZ policy. Since each g^j is a convex function, the function $E[g^j(y - V_\infty^j)]$ is also convex with respect to y; therefore, it is easy to find $\tilde{S}^j$ numerically via simulation. In an uncapacitated system, since the shortfall random variable V_∞^j is zero for every j, the vector $(\tilde{S}^1, \dots, \tilde{S}^N)$ is identical to the vector $(S^{1*}, \dots, S^{N*})$, which is the optimal echelon base-stock vector for an uncapacitated system.

MSS-U Policy and MSS-L Policy. We propose two other echelon order-up-to policies, which are motivated by Shang and Song (2003), where the echelon base-stock levels are given as solutions to suitably defined newsvendor problems. We term these policies as the *Modified Shang-Song* (MSS) policies. Let

$$G^j(y; \hat{h}, \hat{b}) = \hat{h} \cdot E[y - \mathcal{D}_{j+1}] + \hat{b} \cdot E[(\mathcal{D}_{j+1} - y)^+]. \qquad (15)$$

Note that G^j is a representation of the newsvendor cost function where the distribution of demand is the $(j+1)$-fold convolution of D, and the overage and the underage costs are $\hat{h}$ and $\hat{b} - \hat{h}$, respectively. (We caution the reader that $\hat{b}$ does *not* represent the underage cost parameter, but rather corresponds to the sum of the overage and underage cost parameters.) The second and the third policies we propose are based on the G^j functions. The second policy, called the MSS-U policy, is the echelon order-up-to $\bar{\mathbf{S}}$ policy, where $\bar{\mathbf{S}} = (\bar{S}^1, \dots, \bar{S}^N)$ is given by

$$\bar{S}^j = \arg\min_S E\big[G^j(S - V_\infty^j; h^j, b + h^j + \cdots + h^N)\big]. \qquad (16)$$

The third policy, called the MSS-L policy, is the echelon order-up-to $\underline{\mathbf{S}}$ policy, where $\underline{\mathbf{S}} = (\underline{S}^1, \dots, \underline{S}^N)$ is given by

$$\underline{S}^j = \arg\min_S E\big[G^j(S - V_\infty^j; h^1 + \cdots + h^j, \\ b + h^1 + \cdots + h^N)\big]. \qquad (17)$$

For uncapacitated systems, each $V_\infty^j = 0$ and these two policies are identical to the two newsvendor heuristics of Theorem 3(b) of Shang and Song (2003). Note that each component of $\underline{\mathbf{S}}$ can equivalently be expressed as

$$\underline{S}^j = \arg\min_S E\left[G^j\left(S - V_\infty^j; h^j, \\ h^j \frac{b + h^1 + \cdots + h^N}{h^1 + \cdots + h^j}\right)\right]. \qquad (18)$$

Remark. The three algorithms we propose above require the computation of the N *single-stage* steady-state shortfall distributions, $\{V_\infty^j\colon j=1,2,\dots,N\}$. Note that this computation of the shortfall distribution (exactly or approximately, analytically or through simulation) is necessary even to find the optimal base-stock level in a single-stage capacitated system. Therefore, any practical methods for computing or approximating the shortfall distribution that have been developed (or will be developed) in the context of single-stage systems serve as equally practical methods for executing our algorithms for serial systems.

We now establish a relationship between the echelon order-up-to vectors used in the three policies we have proposed. In particular, we show that the vector $\tilde{\mathbf{S}}$ is sandwiched between $\underline{\mathbf{S}}$ and $\bar{\mathbf{S}}$, componentwise. The proof of Lemma 1 appears in the appendix. This is a generalization of a similar result by Shang and Song (2003) for uncapacitated systems.

Lemma 1. *For every $j\in\{1,\dots,N\}$,*

$$\underline{S}^j \le \tilde{S}^j \le \bar{S}^j.$$

It should be noted that this result only establishes upper and lower bounds on the base-stock vector used by the MFZ policy; it is unclear if similar bounds apply to the best base-stock vector. The computational investigation of these three policies is discussed in Section 6.

4. Two Lower Bounds on the Optimal Cost

In this section, we develop two lower bounds on the optimal cost of the capacitated serial system. The first lower bound, appearing in Section 4.1, is based on the sum of the optimal costs of suitably defined single-stage capacitated systems. This is a generalization to capacitated systems of a lower bound developed by Chen and Zheng (1994) for uncapacitated systems. This bound, as we shall see later in Section 5, plays a pivotal role in establishing the asymptotic optimality of echelon base-stock policies. The computational investigation (reported in Section 6) shows that this lower bound is close to the optimal cost in many problem instances, but not always. Motivated by this, we develop another lower bound in Section 4.2 based on a completely different approach.

4.1. Lower Bound 1 (LB1)

We provide the first lower bound on $C^*(\mathbf{I}_0)$, which is defined in (8)—it is the optimal long-run average cost for the capacitated N-echelon serial system that we study, given a starting state $\mathbf{I}_0$. We denote the possible dependence of the optimal cost on the starting state explicitly. Recall the definition of $G^j(y;\hat{h},\hat{b})$ from (15), which represents the expected newsvendor cost with the inventory level y, the overage cost $\hat{h}$, the underage cost $\hat{b}-\hat{h}$, and the demand distribution $\mathcal{D}_{j+1}$.

Theorem 2. *Let $(\alpha^1,\dots,\alpha^N)$ be such that $\alpha^j\in[0,1]$ for all j and $\sum_j \alpha^j=1$. Then, for all $\mathbf{I}_0$,*

$$\begin{aligned} C^*(\mathbf{I}_0) &- \sum_{j=1}^{N}(j-1)\cdot h^j\cdot\mu \\ &\ge \sum_{j=1}^{N}\min_S E_{V_\infty^j}\big[G^j(S-V_\infty^j;\,h^j,\alpha^j\cdot(b+H^1))\big].\end{aligned}$$

The intuition for this result is identical to the cost decomposition arguments in Chen and Zheng (1994) for uncapacitated systems; the idea is that while the original system has constraints that link the decision at every echelon to the inventory level in the previous echelon, the lower bound is constructed by considering a set of single-stage systems (without these constraints) and ensuring that, for any given state vector, the cost in the original system equals the sum of the costs in the single-stage systems. The proof of Theorem 2 is based on the comparison of the capacitated serial system to a set of capacitated single-stage systems, each of which is optimized by a base-stock policy. (See the online appendix (available as supplemental material at http://dx.doi.org/10.1287/msom.2016.0588) for the proof.) Each minimization on the right side expression is a newsvendor problem, for which a critical fractile solution is optimal. If the system is uncapacitated (i.e., $\mathrm{CAP}^j=\infty$ for each j), then it follows that $V_\infty^j=0$ deterministically for all j's, and we recover the lower bound of Chen and Zheng (1994). Also, notice that the above theorem gives us a lower bound for every nonnegative vector $(\alpha^1,\dots,\alpha^N)$ such that they sum to 1; consequently, the maximum of all these bounds over this space of vectors is also a lower bound. The exercise of maximizing this lower bound by optimizing over $(\alpha^1,\alpha^2,\dots,\alpha^N)$ could exploit the fact that the bound is a concave function of this vector.)

4.2. Lower Bound 2 (LB2)

We now proceed to derive another lower bound on the optimal cost of our system. To do this, we introduce another capacitated serial inventory system, which is identical in all ways to our original system with the exception that the capacities at stages $1,\dots,N-1$ are infinite (the capacity at stage N (the topmost stage) is CAP^N, as in the original system). Let $\tilde{C}^*$ denote the optimal long-run average cost of this new system. (Here we simply denote the cost as $\tilde{C}^*$ instead of $\tilde{C}^*(\mathbf{I}_0)$ because as will be shown in Theorem 3, this optimal cost does not depend on the starting state of the system.) Clearly, this system is less constrained than the original system, and therefore its optimal cost is smaller than that of the original system. We state this result formally as a theorem. In this theorem, we also state the optimal policy for the new system—this is Parker and Kapuscinski (2004, Theorem 6(b)); this result makes it easy to compute this system's optimal cost, i.e., the lower bound on $C^*(\mathbf{I}_0)$.

Theorem 3. *For any* $\mathbf{I}_0$, $C^*(\mathbf{I}_0)$ *is bounded below by* $\tilde{C}^*$, *the optimal cost of the new system. Furthermore, the optimal cost* $\tilde{C}^*$ *of the new system, is independent of* $\mathbf{I}_0$, *and is attained by the MFZ policy for this system.*

We conclude this section by remarking that we will use only the lower bound of Theorem 2 (LB1) for our asymptotic analysis in Section 5. Both lower bounds in Theorems 2 and 3 (LB1 and LB2) are used in our computational investigation in Section 6, where it is shown that a combination of these lower bounds provides a good approximation of the optimal cost.

5. Bounds and Strong Asymptotic Optimality

In this section, we show that all the three policies we propose (in Section 3) display a desirable property called *strong asymptotic optimality* as b approaches ∞. We now define and discuss the concept of *strong asymptotic optimality* in our context. A policy is *asymptotically optimal* if the relative difference between the cost of this policy and the optimal cost, as a percentage of the optimal cost, approaches 0 as the backorder cost parameter b approaches ∞. Moreover, a policy is *strongly asymptotically optimal* if in addition the difference between its cost and the optimal cost is uniformly bounded for all values of b. We will show later (in Lemma 4) that the optimal cost of the system approaches ∞ as b approaches ∞. Therefore, strong asymptotic optimality of a policy implies asymptotic optimality but the reverse implication is not necessarily true. Moreover, the convergence rate of the cost of a strongly asymptotically optimal policy to the optimal cost is the same as the rate at which the reciprocal of the optimal cost converges to zero.

Our proof of the strong asymptotic optimality of our policies consists of two main elements. First, one of the lower bounds that we have derived in Section 4 (LB1) approaches ∞ as $b \to \infty$; thus, the cost of the optimal policy also approaches ∞ (Section 5.1). Second, we provide an upper bound on the cost of each of the three policies (Section 5.2). Finally, using these ingredients, we will show that the difference between the lower and upper bounds is uniformly bounded with respect to b (Section 5.3), implying the strong asymptotic optimality of our policies.

A reasonable question that might arise is whether a simpler policy such as one that the base-stock level for each echelon is computed by replacing the single-stage shortfalls by their means is also asymptotically optimal. We show in the online appendix that this policy is *not* asymptotically optimal. It is easy to understand why this is the case by considering the example of a single-stage system with a bounded demand distribution—even in this case, the shortfall is an unbounded random variable under our assumptions. The base-stock level computed by replacing the shortfall with its mean is the newsvendor level for a distribution that is the original demand distribution shifted to the right by the mean shortfall—since this distribution is bounded, the base-stock level remains bounded for all backorder costs. However, the truly optimal base-stock level approaches infinity as the backorder cost parameter approaches infinity because the shortfall is an unbounded random variable. Thus, the heuristic incurs a heavy amount of backorder cost when the backorder cost parameter grows.

5.1. Asymptotic Behavior of the Optimal Cost

In this section, we show that the optimal long-run average cost for the capacitated N-echelon serial system grows arbitrarily large as the backorder cost parameter b approaches ∞. We achieve this result (stated in Lemma 4) by showing the same property for the first lower bound (LB1) on the optimal cost, introduced in Section 4.1. We restrict our attention to the case that the system is truly capacitated (i.e., there is a strictly positive probability of the demand in some period exceeding the bottleneck capacity), since a capacitated system is otherwise equivalent to an uncapacitated system. We formally state this assumption.

Assumption 1. *Demand distribution* D *satisfies* $P(D > \mathrm{CAP}^N) > 0$.

We denote the dependence of costs on b explicitly when it is convenient to do so; for example, $C^*(\mathbf{I}_0; b)$ denotes the optimal cost of the system. The following lemma is an implication of Theorem 2, and its proof is presented in the online appendix.

Lemma 4. *Under Assumption 1,* V_∞^N *is an unbounded random variable, i.e.,* $P(V_\infty^N > x) > 0$ *for all* x. *Moreover,* $C^*(\mathbf{I}_0; b) \to \infty$ *as* $b \to \infty$ *for any* $\mathbf{I}_0$.

5.2. Upper Bounding the Cost of an Echelon Order-Up-to Policy

Lemma 5. *The expected long-run average cost of the order-up-to* $\mathbf{S}$ *policy,* $E[C_\infty(\mathbf{S})]$, *satisfies the following inequality:*

$$E[C_\infty(\mathbf{S})] - \sum_{j=1}^{N} h^j \cdot (j-1) \cdot \mu \le \sum_{j=1}^{N} E_{V_\infty^j}\big[G^j(S^j - V_\infty^j; h^j, b + h^j + \cdots + h^N)\big].$$

To understand this upper bound intuitively, recall that the cost for a given base-stock vector, $\mathbf{S}$, is the sum of the holding costs at echelons $1, 2, \ldots, N$ and the backorder cost at echelon 1. In Lemma 9 of Appendix A, we use (12) to express the inventory level at any echelon as a minimum over several quantities; thus, any quantity within the minimum can be used to obtain an upper bound on the holding cost at an echelon. Similarly, the backorder level is expressed as a maximum over several quantities; thus, the backorder cost can be upper bounded using the sum of these quantities. The upper bound on $E[C_\infty(\mathbf{S})]$ is obtained by summing the bounds indicated in the preceding two arguments. The formal proof is in Appendix C.

Remark. For the uncapacitated system, where the ordering capacity CAP^j at every stage j is infinite, we know that $V^j_\infty = 0$ for all j. In this case, the upper bound of the cost in Lemma 5 is the bound given in Shang and Song (2003).

5.3. Strong Asymptotic Optimality

The main result of this section is to show the strong asymptotic optimality of the three echelon base-stock policies that have been introduced in Section 3. We first show that the echelon order-up-to $\bar{\mathbf{S}}$ policy (MSS-U policy) and the echelon order-up-to $\underline{\mathbf{S}}$ policy (MSS-L policy) are both strongly asymptotically optimal. Then, we argue that the echelon order-up-to $\tilde{\mathbf{S}}$ policy (MFZ policy) is strongly asymptotically optimal using the fact that the order-up-to vector $\tilde{\mathbf{S}}$ is sandwiched between **ULS** and $\bar{\mathbf{S}}$ (Lemma 1). Since the backorder cost parameter b is a variable in this analysis, we will use the notation $\underline{\mathbf{S}}(b)$, $\tilde{\mathbf{S}}(b)$ and $\bar{\mathbf{S}}(b)$ to represent the dependence of $\underline{\mathbf{S}}$, $\tilde{\mathbf{S}}$ and $\bar{\mathbf{S}}$, respectively, on b.

MSS-U Policy. To show the strong asymptotic optimality of the MSS-U policy, we will show that the difference $E[C_\infty(\bar{\mathbf{S}}(b); b)] - C^*(\mathbf{I}_0; b)$ is uniformly bounded above with respect to b.

To state our results, we introduce a definition and an assumption on the demand distribution. A nonnegative random variable X is said to be *light tailed* if $\exists\, t > 0$ such that $E[e^{tX}] < \infty$. It is well known that a random variable is light tailed if and only if its tail, $P(X > x)$, is bounded above by an exponentially decreasing function (Foss et al. 2011). The family of light-tailed distributions includes several popular demand distributions including BMRL, the class of distributions with a bounded mean residual life; see Theorem 2.1 of Su and Tang (2003). It is well known that BMRL contains the class of distributions with decreasing mean residual life, which, in turn, contains the popular class of distributions with increasing failure rate. All bounded distributions are also light tailed.

We make the following assumption throughout the section.

Assumption 2. *The random variable, D, representing single-period demand is light tailed.*

Also, define for any $j \in \{1, \ldots, N\}$,

$$\Gamma^j = \sup_{x \geq 0} E\big[\mathscr{D}_{j+1} + V^j_\infty - x \mid \mathscr{D}_{j+1} + V^j_\infty > x\big];$$

that is, Γ^j is the maximum mean residual life of $\mathscr{D}_{j+1} + V^j_\infty$. When D is light tailed, a famous result on the tail of the steady-state waiting time distribution in a $GI/G/1$ queue called the *Cramér-Lundberg* approximation (please see Asmussen 2000) can be used to show (Glasserman 1997) that as x becomes large, $P(\mathscr{D}_{j+1} + V^j_\infty > x)$ approaches an exponential function of $(-x)$, which then implies the finiteness of Γ^j.

We formally state the strong asymptotic optimality result for MSS-U in Theorem 6. Its proof, which appears in Appendix D, first establishes the finiteness of Γ^j for all j. We use it to provide an upper bound on $E[C_\infty(\bar{\mathbf{S}}(b); b)] - C^*(\mathbf{I}_0; b)$ using a sum of the differences of the optimal costs of certain capacitated single-stage systems with suitably defined parameters.

Theorem 6. *Suppose Assumption 2 holds. Then, Γ^j is finite for all j, and*

$$E[C_\infty(\bar{\mathbf{S}}(b); b)] - C^*(\mathbf{I}_0; b) \leq H^1 \cdot \sum_{j=1}^{N} \Gamma^j.$$

Furthermore, if Assumption 1 also holds, then, as $b \to \infty$ $E[C_\infty(\bar{\mathbf{S}}(b); b)]/C^*(\mathbf{I}_0; b) \to 1$.

Intuitively, the main reason why such a uniform bound is possible on the gap between the cost of our policy and the optimal policy is the following. As b increases, the optimal levels of safety stock also increase. This increase is required to reduce the expected amount of backorders. That expected amount depends on the conditional expectation of the backordered amount given that there is a backorder. The property that the random variable $\mathscr{D}_{j+1} + V^j_\infty$ is asymptotically exponential implies that this conditional expectation is bounded. An immediate corollary of Theorem 6 is that the MSS-U policy is strongly asymptotically optimal under Assumptions 1 and 2.

MSS-L Policy. Next, we consider the echelon order-up-to $\underline{\mathbf{S}}$ policy. Analogous to Theorem 6, we show the strong asymptotic optimality of this policy under Assumptions 1 and 2. This result is formally stated in Theorem 7. The proof of this result follows the same general structure as the proof of Theorem 6, but uses a slightly different approach to bound $E[C_\infty(\underline{\mathbf{S}}(b); b)] - C^*(\mathbf{I}_0; b)$. See the online appendix for the proof.

Theorem 7. *The result of Theorem 6 holds where $\bar{\mathbf{S}}(b)$ is replaced with $\underline{\mathbf{S}}(b)$.*

MFZ Policy. Finally, we establish the strong asymptotic optimality of the order-up-to $\tilde{\mathbf{S}}(b)$ policy. This result is stated in the following theorem. Its proof is based on the fact that $\underline{\mathbf{S}}(b) \le \tilde{\mathbf{S}}(b) \le \bar{\mathbf{S}}(b)$ (Lemma 1) and the bounds established in the proofs of the asymptotic optimality of the MSS-U and MSS-L policies (Theorems 6 and 7). See the online appendix for the proof.

THEOREM 8. *The result of Theorem 6 holds where* $\bar{\mathbf{S}}(b)$ *is replaced with* $\tilde{\mathbf{S}}(b)$.

REMARK. When the demand distribution is unbounded, the optimal cost of the uncapacitated serial system approaches ∞ as b approaches ∞. Thus, when we replace Assumption 1 with the assumption that the demand distribution is unbounded, Theorems 6 and 7 automatically imply that the MSS-L and MSS-U policies are strongly asymptotically optimal for uncapacitated serial systems. This provides a new argument in support of the newsvendor bounds of Shang and Song (2003). As for the MFZ policy, it is optimal (not just asymptotically) for these uncapacitated systems (Federgruen and Zipkin 1984).

6. Computational Results

In this section, we report on computational experiments to investigate the performance of the three order-up-to policies introduced in Section 3 and the quality of the lower bounds developed in Section 4. In particular, we investigate the performance of our heuristics over a wide range of the backorder cost parameter.

We use the Monte Carlo simulation method with a common random number generator to compute our three heuristic policies, their expected costs, and the two lower bounds on the optimal cost as follows.

Step (a). *Computation of the heuristics*. For each of our three policies, the echelon base-stock vectors can be computed using the steady state distributions of the single-stage shortfall processes $\{V_t^j\}$. To obtain approximations of these distributions, we generate a simulation run composed of demands in 50,000 periods and generate 100 such simulation runs. (As will be clear soon, the first 10,000 periods are used for warming up the system and the remaining 40,000 periods are used for actual cost calculations.) For each run, we use the recursion (10) to simulate the $\{V_t^j\}$ processes. Then, for each stage j, we use the values of the single-stage shortfall V_t^j in periods $t = 10{,}001 - 50{,}000$ over all 100 simulation runs to compute (approximately) the distribution of the steady state shortfall V_∞^j. Once this distribution is available, the echelon base-stock vector for each of our policies is computed using the definitions of the policies in Section 3.

Step (b). *Computation of the expected costs of the heuristics*. For any echelon base-stock policy with echelon base-stock vector $(S^1, S^2, \ldots, S^N)$, the costs incurred by the system through time depend on the dynamics (i.e., inventory levels, backorder amounts) of the system. For each of the 100 simulation runs, these dynamics can be simulated using the single-stage shortfall processes $\{V_t^j\}$ and the relation in (12) to obtain the echelon shortfall processes $\{\tilde{V}_t^j\}$. Consider any one of our three heuristic policies. The base-stock vector used by this policy has already been computed in Step (a). Thus, by definition of $\tilde{V}_t^j$ in (11), we obtain the echelon-j inventory position IP_t^j. Now that all the echelon inventory positions are known, we follow the mechanics detailed in Section 2 and the initialization defined in (9) to compute the inventory level in each echelon in period t and the backordered amount in that period. These quantities (inventory levels, backordered amounts) are used to compute the cost incurred in each period, as defined in Section 2. The long-run average cost incurred by the system under the chosen heuristic is approximated by taking the average cost incurred by the system from periods $10{,}001 - 50{,}000$ and averaged over the 100 simulation runs.

Step (c). *Computation of the lower bounds on the optimal cost*. Recall from the definition of our first lower bound, $LB1$, that it can be computed using the distributions of the steady state, single-stage shortfalls, $\{V_\infty^j\}$. These distributions are available from Step (a) and $LB1$ can be computed using them. As for $LB2$, recall that it is the optimal cost for a system in which the capacity constraint applies only to stage N and that the MFZ policy is optimal for this "lower bounding system." The MFZ policy for this system, its long-run average cost, and thus that system's optimal cost, are all computed (approximately) using the method detailed in Step (b). This optimal cost for the lower bounding system is the desired lower bound, $LB2$.

We test two sets of examples that are adapted from the paper by Glasserman and Tayur (1995) in which they use IPA to find locally optimal base-stock vectors (within the class of base-stock policies). The first set corresponds to a two-echelon system and the next one corresponds to a four-echelon system.

- *Two-Echelon Example*. We fixed the echelon holding cost parameters at $h^1 = h^2 = 5$, and vary the backorder cost parameter among $b \in \{20, 30, 90, 190, 990\}$. (These backorder cost parameters are chosen such that the corresponding "implied service levels," $b/(b + H^1)$ values, are from the set $\{0.67, 0.75, 0.90, 0.95, 0.99\}$.) The capacities are the same in both stages, and vary among $\mathrm{CAP}^1 = \mathrm{CAP}^2 \in \{55, 60, 65, 70, 75\}$. Demand per period is sampled from an Erlang distribution that has a mean of 50 and a squared coefficient of variation (SCV) of $\{0.25, 0.5, 1.0\}$. Thus, there are $5 \times 5 \times 3 = 75$ possible combinations of parameter values.

- *Four-Echelon Example*. The capacities in all four stages are identical and vary among {55, 60, 65, 70, 75}. We test four scenarios of echelon holding costs (h^1, h^2, h^3, h^4): (10, 10, 10, 10), *uniform value added* case; (1, 1, 28, 10), *early value added* case; (28, 1, 1, 10), *late value added* case; and (1, 28, 1, 10), *middle value added* case. The backorder cost parameter varies among {80, 120, 360, 760, 3,960} such that the corresponding implied service levels belong to the set {0.67, 0.75, 0.90, 0.95, 0.99}. Demand has mean 50 and SCV 0.25. Thus, there are $5 \times 5 \times 4 = 100$ possible combinations of parameter values.

In summary, we test a total of $75 + 100 = 175$ problem instances. (Note that while we only report the results for a system with identical capacities, we have also performed experiments with systems having nonidentical capacities. The computational results for those cases are similar to the results reported here.)

The three heuristics that we have tested are the MFZ policy, the MSS-U policy, and the MSS-L policy. We also evaluate the best echelon base-stock policy, which we obtain by optimizing over the space of echelon base-stock vectors, which is an N-dimensional search. (We use Matlab's implementation of a global optimization routine called patternsearch.) The routine patternsearch works by designing sequential meshes in which points are randomly polled and the value of the objective function is compared to values of previously polled points. We repeat the advantage of our heuristics over using global optimization to find the best base-stock policy. Any such routine requires the evaluation of the long-run average cost achieved at numerous base-stock vectors (i.e., at numerous points in N-dimensional space), while our algorithms involve "one-shot" computations. More precisely, once the steady state, single-stage shortfall distributions $\{V_\infty^j\}$ are computed by simulation (note that this distribution would have to be computed even if one were to solve the single-stage, capacitated inventory problem), our algorithms take only as much effort as the algorithms of Federgruen and Zipkin (1984) and Shang and Song (2003) for uncapacitated serial systems.

To measure the performance of each policy, we compute the difference between the cost of the policy and the cost of the best echelon base-stock policy, and divide it by the latter. For each configuration, we denote the best performing policy by the *best heuristic* (BH) policy. We have also tested the two lower bounds given in Theorems 2 and 3, which we refer to as LB1 and LB2, respectively. To obtain LB1, we needed to perform maximization over the parameter vector $(\alpha^1, \ldots, \alpha^N)$ (see Section 4.1), for which we again used patternsearch. We denote the maximum of these bounds as the *better lower bound* (BL). We evaluate the performance of the BL by dividing the difference between the cost of the best echelon base-stock policy and the BL by the latter.

Aggregate summary of the performances of the heuristic policies and the lower bound appear in Tables 1 and 2. For the two-echelon example in Table 1, each column under "Capacity" represents the average and worst performances of the heuristics and the lower bounds over 5×3 combinations of b and SCV for the given capacity. The last column, titled "All," corresponds to the average and worst performances over all $5 \times 5 \times 3$ experiments. Similarly, for the four-echelon example, our $5 \times 5 \times 4$ experiments are summarized in Table 2.

For the two-echelon example, the cost of the BH policy is, on average, 1.1% higher than the cost of the best echelon base-stock policy, and never more than 2.7% in our 75 experiments. For the four-echelon example, the average and maximum errors were 1.3% and 4.0%, respectively. Based on our numerical results on the two sets of examples, the MSS-L policy performs the best in most instances, achieving an average error gap of 1.7% and a worst gap of 4.4%.

The lower bounds that we developed are also good; the better lower bound (BL) is on average 0.5% away from the cost of the best base-stock policy in the two-echelon examples, and 0.6% away in the four-echelon examples. Since the optimal cost lies between the lower bound and the cost of the best base-stock policy, this observation indicates that our lower bound

Table 1 **Aggregate Summary: The Two-Echelon Example**

		Capacity (%)					
Two-echelon example	Error	55	60	65	70	75	All (%)
MFZ	Avg	4.2	3.2	2.8	2.1	1.6	2.8
	Max	6.3	4.3	4.5	3.8	3.3	6.3
MSS-U	Avg	5.9	5.4	5.2	4.4	3.9	5.0
	Max	9.2	7.8	8.5	7.7	6.6	9.2
MSS-L	Avg	0.0	1.1	1.9	1.7	1.5	1.1
	Max	0.0	2.6	2.7	2.7	2.7	2.7
Best heuristic (BH)	Avg	0.0	1.1	1.9	1.7	1.4	1.1
	Max	0.0	2.6	2.7	2.7	2.7	2.7
Better lower bound (BL)	Avg	0.9	0.9	0.4	0.3	0.2	0.5
	Max	2.0	2.9	1.5	0.8	1.1	2.9

Table 2 **Aggregate Summary: The Four-Echelon Example**

		Capacity (%)					
Two-echelon example	Error	55	60	65	70	75	All (%)
MFZ	Avg	7.3	3.5	1.8	0.9	0.4	2.8
	Max	10.9	6.4	4.5	2.6	1.3	10.9
MSS-U	Avg	10.6	7.5	5.5	4.1	3.4	6.2
	Max	18.4	12.4	9.5	8.2	7.6	18.4
MSS-L	Avg	1.5	2.3	2.2	2.0	1.8	2.0
	Max	3.6	4.4	4.2	4.1	4.2	4.4
Best heuristic (BH)	Avg	1.5	2.1	1.5	0.8	0.4	1.3
	Max	3.6	4.0	3.3	2.5	1.3	4.0
Better lower bound (BL)	Avg	0.8	1.0	0.7	0.5	0.3	0.6
	Max	2.4	2.6	2.2	2.0	1.6	2.6

is close to the optimal cost and that the best base-stock policy is almost optimal over all policies.

Finally, we report some numbers on computational speed. In terms of the computation environment, we used a Lenovo Thinkpad Notebook with an Intel Core Duo Processor (CPU of 1.86 GHz and memory of 4 GB). The main software we used was Matlab 7. Based on a representative sample of hard instances with $L=4$, the times in seconds are as follows: MFZ 291.7, MSSU 20.5, MSSL 20.6, and OPT 1536.2. We note that for MSSU and MSSL the computatonal times include shortfall generation times (which takes most of the time).

7. Conclusions

In this paper, we have provided heuristic policies for capacitated serial inventory systems that are intuitive and have a theoretical foundation in the sense that these policies are optimal both for capacitated, single-stage systems as well as uncapacitated serial systems. Further, we showed analytically that these policies are asymptotically optimal in high service level environments and showed empirically that they perform well by comparing them to a lower bound on the optimal cost, which we developed in this paper. Moreover, these policies are easy to implement and intuitive to understand. A valuable future research exercise on this problem is that of developing approximation algorithms with provable bounds for these systems. It is our hope that the lower bounds developed in this paper will be useful in such endeavors. Another promising direction is to extend our policies and results to the discounted cost performance measure—here, results on uncapacitated serial systems and discounted costs such as Dong and Lee (2003) and Chao and Zhou (2007) would likely play a crucial role. On a more practical note, discount factors over inventory review periods are typically very close to 1 in practice because of the magnitude of interest rates and review cycles. Since discounted cost dynamic programs converge to their average cost counterparts (Schäl 1993, Huh et al. 2011) as the discount factor approaches 1, it is possible to use our results to derive good bounds on the performance of our policies even under discounting.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2016.0588.

Acknowledgments

The authors gratefully acknowledge two M.Sc. students at Columbia University, Suckyoung Oh for generating the computational results reported in this paper, and Soohyun Cho for identifying several typographical errors. The research was partially funded by Natural Sciences and Engineering Research Council of Canada.

Appendix. Proofs

A. A Preliminary Result

Recall the definitions of IP^j_t, e^j_t and B_t in (3)–(5). Lemma 9 provides expressions for these quantities. We make their dependence on the echelon base-stock vector $\mathbf{S}$ explicit by noting it as their argument.

Lemma 9. *In the N-echelon system operated under an echelon base-stock policy with the order-up-to vector* $\mathbf{S}$, *the following equations hold for all $j\in\{1,\dots,N\}$ and $t\ge 0$:*

$$\mathrm{IP}^j_t(\mathbf{S}) = \min(S^i-(V^i_{t+j-i}+D[t+j-i,t-1])\,|\,i=j,\dots,N),$$
$$e^j_{t+1}(\mathbf{S}) = \min(S^i-(V^i_{t+j-i}+D[t+j-i,t+1])\,|\,i=j,\dots,N),$$
$$B_{t+1}(\mathbf{S}) = \max(0,V^i_{t+1-i}+D[t+1-i,t+1]-S^i\,|\,i=1,\dots,N).$$

Proof. The expression for $\mathrm{IP}^j_t(\mathbf{S})$ follows immediately from (12)—this is Theorem 1 of Huh et al. (2010)—and the definition of the shortfall process $\tilde{V}^j_t$. The expression for $e^j_{t+1}(\mathbf{S})$ follows from the identity

$$e^j_t(\mathbf{S}) = \mathrm{IP}^j_{t-1}(\mathbf{S})-D_{t-1}-D_t, \tag{19}$$

which is straightforward to verify since the lead time between consecutive stages is 1. The expression for B_{t+1} follows from the fact that $B_t(\mathbf{S})=[e^1_t(\mathbf{S})]^-$. □

B. Proof of Lemma 1

For each $j \in \{1, \dots, N\}$, a standard argument shows that g^j is a convex function. In the proof, we assume that each g^j is differentiable. (The case where g^j is not differentiable can be argued similarly.) We first provide an upper bound and a lower bound on the derivative of $g^j(y)$.

Proposition 10. *For any $j \in \{1, \dots, N\}$, if g^j is differentiable, then*

$$h^j - \left(b + \sum_{k=j}^{N} h^k\right) \cdot P(y \le \mathcal{D}_{j+1}) \le \frac{d}{dy}[g^j(y)] \le \sum_{k=1}^{j} h^k - (b + H^1) \cdot P(y \le \mathcal{D}_{j+1}).$$

Proof. By differentiating g^1 in (13), we can see that the required result is true for $j = 1$; in fact, the relationship between the left and right sides holds with an equality in this case. Let us now assume that the statement is true for some $j \in \{1, \dots, N-1\}$ and proceed to prove the statement for $j+1$. By differentiating g^{j+1} in (13), we obtain

$$\frac{d}{dy}[g^{j+1}(y)] = h^{j+1} + E\left[\frac{d}{dy}[g^j(y-D)] \cdot \mathbf{1}[y - D < S^{j*}]\right]. \quad (20)$$

Using the fact that $(d/dy)(g^j(y))$ crosses zero from below at S^{j*} (from the definition of S^{j*}), we obtain

$$\begin{aligned}\frac{d}{dy}[g^{j+1}(y)] &\le h^{j+1} + E\left[\frac{d}{dy}(g^j(y-D))\right]\\ &\le h^{j+1} + E\left[\sum_{k=1}^{j} h^k - (b + H^1) \cdot P(y - D_{j+2} \le \mathcal{D}_{j+1})\right]\\ &= \sum_{k=1}^{j+1} h^k - (b + H^1) \cdot P(y \le \mathcal{D}_{j+2}),\end{aligned}$$

where the second inequality follows from the induction hypothesis.

Also, applying the induction hypothesis to (20),

$$\begin{aligned}\frac{d}{dy}[g^{j+1}(y)] &\ge h^{j+1} + E\big[[h^j - (b + h^j + \dots + h^N) \cdot P(y \le \mathcal{D}_{j+1})]\\ &\qquad \cdot \mathbf{1}[y - D_{j+2} < S^{j*}]\big]\\ &\ge h^{j+1} - E\big[[(b + h^{j+1} + \dots + h^N) \cdot P(y \le \mathcal{D}_{j+1})]\\ &\qquad \cdot \mathbf{1}[y - D_{j+2} < S^{j*}]\big]\\ &\ge h^{j+1} - (b + h^{j+1} + \dots + h^N) \cdot P(y \le \mathcal{D}_{j+2}).\end{aligned}$$

This completes the induction step, and thus we complete the proof.

Lemma 1. *For every $j \in \{1, \dots, N\}$,*

$$\underline{S}^j \le \tilde{S}^j \le \bar{S}^j.$$

Proof. Since $\tilde{S}^j$ minimizes $E[g^j(S - V_\infty^j)]$, we obtain the following relations using the first-order condition for (16) and the upper bound on $(d/dy)g^j(y)$ from Proposition 10:

$$0 = E\left[\frac{d}{dy} g^j(\tilde{S}^j - V_\infty^j)\right] \le \sum_{k=1}^{j} h^k - (b + H^1) \cdot P(\tilde{S}^j - V_\infty^j \le \mathcal{D}_{j+1}).$$

By the definition of $\underline{S}^j$ in (17) and also from (15), we obtain

$$0 = \sum_{k=1}^{j} h^k - (b + H^1) \cdot P(\underline{S}^j - V_\infty^j \le \mathcal{D}_{j+1}).$$

This equation, along with the previous inequality and the convexity of $E[g^j(S - V_\infty^j)]$ in S, implies that $\underline{S}^j \le \tilde{S}^j$, which is the first desired result.

Now, we prove $\tilde{S}^j \le \bar{S}^j$. From the definition of $\tilde{S}^j$ in (16) and the lower bound inequality in Proposition 10,

$$0 = \frac{d}{dS}(E[g^j(S - V_\infty^j)])|_{S=\tilde{S}^j} \ge h^j - \left(b + \sum_{k=j}^{N} h^k\right) \cdot P(\mathcal{D}_{j+1} + V_\infty^j \ge \tilde{S}^j).$$

However, since $\bar{S}^j$ minimizes $E[G^j(S - V_\infty^j; h^j, b + h^j + \dots + h^N)]$ by (16), we obtain

$$0 = h^j - \left(b + \sum_{k=j}^{N} h^k\right) \cdot P(\mathcal{D}_{j+1} + V_\infty^j \ge \bar{S}^j).$$

Thus, we obtain $\tilde{S}^j \le \bar{S}^j$, which is the desired result. □

C. Proof of Lemma 5

Lemma 5. *The expected long-run average cost of the order-up-to* **S** *policy, $E[C_\infty(\mathbf{S})]$, satisfies the following inequality*:

$$E[C_\infty(\mathbf{S})] - \sum_{j=1}^{N} h^j \cdot (j-1) \cdot \mu \le \sum_{j=1}^{N} E_{V_\infty^j}[G^j(S^j - V_\infty^j; h^j, b + h^j + \dots + h^N)].$$

Proof. From the definition of the single-period cost C_t in (7), we obtain the following:

$$C_t(\mathbf{S}) = b \cdot B_t(\mathbf{S}) + \sum_{j=1}^{N} h^j \cdot [e_{t+1-j}^j(\mathbf{S}) + B_t(\mathbf{S})]. \quad (21)$$

Recall, from the characterization of $B_{t+1}(\mathbf{S})$ and $e_{t+1}^j(\mathbf{S})$ in Lemma 9 that

$$\begin{aligned}&e_{t-j+2}^j(\mathbf{S})\\ &\quad = \min\big(S^i - V_{t-i+1}^i - D[t-i+1, t-j+2] \,|\, i = j, \dots, N\big) \quad (22)\\ &\quad \le S^j - V_{t-j+1}^j - D[t-j+1, t-j+2], \quad \text{and} \quad (23)\\ &B_{t+1}(\mathbf{S})\\ &\quad = \max\big([S^i - V_{t-i+1}^i - D[t-i+1, t+1]]^- \,|\, i = 1, \dots, N\big) \quad (24)\\ &\quad \le \sum_{i=1}^{N} [S^i - V_{t-i+1}^i - D[t-i+1, t+1]]^-. \quad (25)\end{aligned}$$

We make the following claim: for $j \in \{1, \dots, N\}$,

$$\begin{aligned}&B_{t+1}(\mathbf{S}) + e_{t-j+2}^j(\mathbf{S}) - D[t-j+3, t+1]\\ &\quad \le \sum_{i=1}^{j-1} [S^i - V_{t-i+1}^i - D[t-i+1, t+1]]^-\\ &\qquad + [S^j - V_{t-j+1}^j - D[t-j+1, t+1]]^+.\end{aligned}$$

This claim is used later in the proof to bound $C_{t+1}(\mathbf{S})$ from above. We prove this claim by considering the following cases separately.

Case. $B_{t+1}(\mathbf{S}) = \max([S^i - V^i_{t-i+1} - D[t-i+1,t+1]]^- \mid i = 1,\dots,j)$.

Then, (24) implies

$$B_{t+1}(\mathbf{S}) \le \sum_{i=1}^{j}[S^i - V^i_{t-i+1} - D[t-i+1,t+1]]^-.$$

By adding the above inequality with (23), we obtain

$$\begin{aligned} &B_{t+1}(\mathbf{S}) + e^j_{t-j+2}(\mathbf{S}) \\ &\quad\le \sum_{i=1}^{j}[S^i - V^i_{t-i+1} - D[t-i+1,t+1]]^- \\ &\qquad + [S^j - V^j_{t-j+1} - D[t-j+1,t-j+2]]. \end{aligned}$$

Subtracting both sides by $D[t-j+3,t+1]$ and using the identity $x = x^+ - x^-$, we obtain the required result in the claim.

Case. $B_{t+1}(\mathbf{S}) > \max([S^i - V^i_{t-i+1} - D[t-i+1,t+1]]^- \mid i = 1,\dots,j)$. Then, (24) implies

$$\begin{aligned} B_{t+1}(\mathbf{S}) &= \max([S^i - V^i_{t-i+1} - D[t-i+1,t+1]]^- \mid i = j,\dots,N) \qquad (26)\\ &= [e^j_{t-j+2}(\mathbf{S}) - D[t-j+3,t+1]]^-, \qquad (27) \end{aligned}$$

where the last inequality follows from (22) and the fact that $\max(a^-, b^-) = [\min(a,b)]^-$ for any real a and b. Then,

$$\begin{aligned} &B_{t+1}(\mathbf{S}) + e^j_{t-j+2}(\mathbf{S}) \\ &\quad = [e^j_{t-j+2}(\mathbf{S}) - D[t-j+3,t+1]]^+ + D[t-j+3,t+1] \\ &\quad \le [S^j - V^j_{t-j+1} - D[t-j+1,t+1]]^+ + D[t-j+3,t+1], \end{aligned}$$

where the equality follows from (27) and the fact that $[a-b]^- + a = \max(a,b) = [a-b]^+ + b$ for any real a and b, and the inequality follows from (23). Thus, we complete the proof of the claim.

Now, the claim implies the following bound for $C_{t+1}(\mathbf{S})$ defined in (21):

$$\begin{aligned} &C_{t+1}(\mathbf{S}) - \sum_{j=1}^{N} h^j \cdot D[t-j+3,t+1] \\ &\quad = b \cdot B_{t+1}(\mathbf{S}) + \sum_{j=1}^{N} h^j \cdot [B_{t+1}(\mathbf{S}) + e^j_{t-j+2}(\mathbf{S}) - D[t-j+3,t+1]] \\ &\quad \le b \cdot \sum_{i=1}^{N} [S^i - V^i_{t-i+1} - D[t-i+1,t+1]]^- \\ &\qquad + \sum_{j=1}^{N} h^j \cdot \sum_{i=1}^{j-1} [S^i - V^i_{t-i+1} - D[t-i+1,t+1]]^- \\ &\qquad + \sum_{j=1}^{N} h^j \cdot [S^j - V^j_{t-j+1} - D[t-j+1,t+1]]^+ \\ &\quad = \sum_{k=1}^{N} (b + h^{k+1} + \cdots + h^N) \cdot [S^k - V^k_{t-k+1} - D[t-k+1,t+1]]^- \\ &\qquad + \sum_{k=1}^{N} h^k \cdot [S^k - V^k_{t-k+1} - D[t-k+1,t+1]]^+, \end{aligned}$$

where the first equality follows from (21), the middle inequality follows from both (25) and the above claim, and finally the last equality follows from rearranging terms. Replacing the shortfall random variables with their steady state versions and taking the expectation, we get the desired result from the definition of G^j in (15). □

D. Proof of Theorem 6

Before we present the proof of Theorem 6, it is convenient to introduce some preliminary results.

Lemma 11. *Let $S^j_{\hat b} = \operatorname{argmin}_S E[G^j(S - V^j_\infty; h^j, \hat b)]$, for each j. Then, for any $\theta > h^j/\hat b$ and any $j \in \{1,\dots,N\}$,*

$$\hat b \cdot E[(\mathcal{D}_{j+1} + V^j_\infty - S^j_{\theta\cdot\hat b})^+] \le \frac{1}{\theta} \cdot h^j \cdot \Gamma^j.$$

Furthermore, for any $\alpha \in (h^j/\hat b, 1]$ and any $j \in \{1,\dots,N\}$,

$$\begin{aligned} &\min_S E[G^j(S - V^j_\infty; h^j, \hat b)] - \min_S E[G^j(S - V^j_\infty; h^j, \alpha\cdot\hat b)] \\ &\quad \le \frac{1-\alpha}{\alpha} \cdot h^j \cdot \Gamma^j. \end{aligned}$$

Proof. From the definition of $S^j_{\theta\cdot\hat b}$ and the fact that $G^j(\cdot\,; h^j, \theta\cdot\hat b)$ defined in (15) is a newsvendor cost with the overage and underage parameters h^j and $\theta\cdot\hat b - h^j$, we know that

$$P(\mathcal{D}_{j+1} + V^j_\infty > S^j_{\theta\cdot\hat b}) \le \frac{h^j}{\theta\cdot\hat b}.$$

Therefore, from the definition of Γ^j, we obtain

$$\begin{aligned} &\hat b \cdot E[(\mathcal{D}_{j+1} + V^j_\infty - S^j_{\theta\cdot\hat b})^+] \\ &\quad = \hat b \cdot P(\mathcal{D}_{j+1} + V^j_\infty > S^j_{\theta\cdot\hat b}) \cdot E[\mathcal{D}_{j+1} + V^j_\infty - S^j_{\theta\cdot\hat b} \mid \mathcal{D}_{j+1} + V^j_\infty > S^j_{\theta\cdot\hat b}] \\ &\quad \le \frac{h^j}{\theta} \cdot \Gamma^j. \end{aligned}$$

To prove the second part of the lemma, observe that

$$\begin{aligned} &\min_S E[G^j(S - V^j_\infty; h^j, \hat b)] - \min_S E[G^j(S - V^j_\infty; h^j, \alpha\cdot\hat b)] \\ &\quad \le E[G^j(S^j_{\alpha\cdot\hat b} - V^j_\infty; h^j, \hat b)] - E[G^j(S^j_{\alpha\cdot\hat b} - V^j_\infty; h^j, \alpha\cdot\hat b)] \\ &\quad = (\hat b - \alpha\cdot\hat b) \cdot E[(\mathcal{D}_{j+1} + V^j_\infty - S^j_{\alpha\cdot\hat b})^+], \end{aligned}$$

where the inequality follows since $S^j_{\alpha\cdot\hat b}$ is optimal solution to the second minimization problem, and the equality follows from the definition of G^j. The second part now follows by a direct application of the first part of the lemma. □

Theorem 6. *Suppose Assumption 2 holds. Then, Γ^j is finite for all j, and*

$$E[C_\infty(\bar{\mathbf{S}}(b); b)] - C^*(\mathbf{I}_0; b) \le H^1 \cdot \sum_{j=1}^{N} \Gamma^j.$$

Furthermore, if Assumption 1 also holds, then

$$E[C_\infty(\bar{\mathbf{S}}(b); b)]/C^*(\mathbf{I}_0; b) \to 1, \quad \text{as } b \to \infty.$$

Proof. The finiteness of Γ^j under Assumption 2 follows from the fact that the distribution of $\mathcal{D}_{j+1} + V^j_\infty$ at x approaches an exponentially decreasing function as $x \to \infty$.

This fact is due to Glasserman (1997), whose assumptions are implied by ours—in that paper, please see Theorem 5, the paragraph following it, and also the last paragraph of page 247.

We proceed to bound $E[C_\infty(\bar{\mathbf{S}}(b);b)] - C^*(\mathbf{I}_0;b)$. Since $\bar{S}^j(b)$ is the minimizer of $E[G^j(S - V_\infty^j; h^j, b+h^j+\cdots+h^N)]$, we can bound using Lemma 5 the cost of the echelon order-up-to $\bar{\mathbf{S}}(b)$ policy from above as follows:

$$\begin{aligned} E[C_\infty(\bar{\mathbf{S}}(b))] &\le \sum_{j=1}^{N} E_{V_\infty^j}\big[G^j(\bar{S}^j(b) - V_\infty^j; h^j, b+h^j+\cdots+h^N)\big] \\ &\quad + \sum_{j=1}^{N} h^j \cdot (j-1)\cdot \mu \\ &= \sum_{j=1}^{N} \min_S E_{V_\infty^j}\big[G^j(S - V_\infty^j; h^j, b+h^j+\cdots+h^N)\big] \\ &\quad + \sum_{j=1}^{N} h^j \cdot (j-1)\cdot \mu, \end{aligned} \tag{28}$$

where the equality follows from the definition of $\bar{S}^j(b)$. Combining this inequality with Theorem 2, we obtain, for any $\mathbf{I}_0$ and $(\alpha^1,\ldots,\alpha^N)$ satisfying $\alpha^j \in (h^j/(b+H^1),1]$ for all j and $\sum_j \alpha^j = 1$,

$$\begin{aligned} &E[C_\infty(\bar{\mathbf{S}}(b);b)] - C^*(\mathbf{I}_0;b) \\ &\quad \le \sum_{j=1}^{N}\Big(\min_S E_{V_\infty^j}[G^j(S - V_\infty^j; h^j, b+h^j+\cdots+h^N)] \\ &\qquad - \min_S E_{V_\infty^j}[G^j(S - V_\infty^j; h^j, \alpha^j\cdot(b+H^1))]\Big). \end{aligned} \tag{29}$$

Since $G^j(y;\hat{h},\hat{b})$ is an increasing function of $\hat{b}$, we know that

$$\begin{aligned} &\min_S E_{V_\infty^j}[G^j(S - V_\infty^j; h^j, b+h^j+\cdots+h^N)] \\ &\quad \le \min_S E_{V_\infty^j}[G^j(S - V_\infty^j; h^j, b+H^1)]. \end{aligned}$$

Using this inequality in (29), we obtain

$$\begin{aligned} &E[C_\infty(\bar{\mathbf{S}}(b);b)] - C^*(\mathbf{I}_0;b) \\ &\quad \le \sum_{j=1}^{N}\Big(\min_S E_{V_\infty^j}[G^j(S - V_\infty^j; h^j, b+H^1)] \\ &\qquad - \min_S E_{V_\infty^j}[G^j(S - V_\infty^j;\ h^j,\ \alpha^j\cdot(b+H^1))]\Big). \end{aligned} \tag{30}$$

Applying the second part of Lemma 11 gives us the following inequality:

$$E[C_\infty(\bar{\mathbf{S}}(b);b)] - C^*(\mathbf{I}_0;b) \le \sum_{j=1}^{N}\left(\frac{1-\alpha^j}{\alpha^j}\right)\cdot h^j \cdot \Gamma^j. \tag{31}$$

The above inequality implies the required inequality in the statement of the theorem if we choose for each j, $\alpha^j = h^j/H^1$. Furthermore, the asymptotic result follows directly from Lemma 4, which implies that $C^*(\mathbf{I}_0;b)\to\infty$ as $b\to\infty$ under Assumption 1. □

References

Angelus A, Zhu W (2009) Managing capacitated multiechelon systems with domain-optimal policies. Working paper, Singapore Management University, Singapore.

Asmussen S (2000) *Applied Probability and Queues 2/e* (Springer-Verlag, New York).

Cachon G, Terwiesch C (2009) *Matching Supply with Demand*, Vol. 2 (McGraw-Hill, Singapore).

Chao X, Zhou SX (2007) Probabilistic solution and bounds for serial inventory systems with discounted and average costs. *Naval Res. Logist.* 54(6):623–631.

Chen F, Zheng Y (1994) Lower bounds for multiechelon stochastic inventory systems. *Management Sci.* 40(11):1426–1443.

Clark A, Scarf H (1960) Optimal policies for a multiechelon inventory problem. *Management Sci.* 6(4):475–490.

Coca-Cola Retailing Research Councils, The (1996) Where to look for incremental sales gains: The retail problem of out-of-stock merchandise. Accessed September 9, 2016, http://www.ccrrc.org/1996/02/24/where-to-look-for-incremental-sales-gains-the-retail-problem-of-out-of-stock-merchandise/.

Dong L, Lee HL (2003) Optimal policies and approximations for a serial multiechelon inventory system with time-correlated demand. *Oper. Res.* 51(6):969–980.

Federgruen A, Zipkin P (1984) Computational issues in an infinite-horizon, multiechelon inventory model. *Oper. Res.* 32(4): 818–836.

Federgruen A, Zipkin P (1986a) An inventory model with limited production capacity and uncertain demands I: The average-cost criterion. *Math. Oper. Res.* 11(2):193–207.

Federgruen A, Zipkin P (1986b) An inventory model with limited production capacity and uncertain demands II: The discounted-cost criterion. *Math. Oper. Res.* 11(2):208–215.

Foss S, Korshunov D, Zachary S (2011) *An Introduction to Heavy-Tailed and Subexponential Distributions* (Springer, New York).

Glasserman P (1997) Bounds and asymptotics for planning critical safety stocks. *Oper. Res.* 45(2):244–257.

Glasserman P, Tayur S (1994) The stability of a capacitated, multi-echelon production-inventory system under a base-stock policy. *Oper. Res.* 42(5):913–925.

Glasserman P, Tayur S (1995) Sensitivity analysis for base-stock levels in multiechelon production-inventory systems. *Management Sci.* 41(2):263–281.

Glasserman P, Tayur S (1996) A simple approximation for a multistage capacitated production-inventory system. *Naval Res. Logist.* 43(1):41–58.

Gruen TW, Corsten D, Bharadwaj S (2002) Retail out of stocks: A worldwide examination of causes, rates, and consumer responses. Technical report, Grocery Manufacturers of America, Washington, DC.

Huh W, Janakiraman G, Nagarajan M (2010) Capacitated serial inventory systems: Sample path and stability properties under base-stock policies. *Oper. Res.* 58(4):1017–1022.

Huh W, Janakiraman G, Nagarajan M (2011) Average cost inventory models: An analysis using a vanishing discount approach. *Oper. Res.* 59(1):143–155.

Janakiraman G, Muckstadt J (2009) A decomposition approach for a class of capacitated serial systems. *Oper. Res.* 57(6):1384–1393.

Parker R, Kapuscinski R (2004) Optimal policies for a capacitated two-echelon inventory system. *Oper. Res.* 52(5): 739–755.

Roundy RO, Muckstadt JA (2000) Jan. Heuristic computation of period-review base stock inventory policies. *Management Sci.* 46(1):104–109.

Schäl M (1993) Average optimality in dynamic programming with general state space. *Math. Oper. Res.* 18(1):163–172.

Shang K, Song J (2003) Newsvendor bounds and heuristic for optimal policies in serial supply chains. *Management Sci.* 49(5): 618–638.

Speck C, van der Wal J (1991) The capacitated multi-echelon inventory system with serial structure: 1. The “Push-Ahead”—Effect. Memorandum COSOR 91-39, Eindhoven University of Technology, Netherlands.

Su C, Tang Q (2003) Characterizations on heavy-tailed distributions by means of hazard rate. *Acta Mathematicae Applicatae Sinica* 19(1):135–142.

Tayur S (1992) Computing the optimal policy for capacitated inventory models. *Comm. Statist.: Stochastic Models* 9(4):585–598.

Tayur S (1997) Recent developments in single product, discrete-time, capacitated production-inventory systems. *Sadhana* 22(1): 45–67.