Strategic Safety-Stock Placement in Supply Chains with Capacity Constraints

Strategic safety stock placement in supply chains with capacity constraints

May 2008, revised November 2015

Stephen C. Graves • Tor Schoenmeyr

A. P. Sloan School of Management,
Massachusetts Institute of Technology, Massachusetts Institute of Technology, Cambridge,
Massachusetts 02139-4307, sgraves@mit.edu; tor@sloan.mit.edu

We generalize the guaranteed-service (GS) model for safety stock placement in supply chains to include capacity constraints. We first examine the guaranteed-service model for a capacitated single-stage system with bounded demand. We characterize the optimal inventory policy, which depends on the entire demand history. Due to this complexity, we develop a heuristic, namely a constant base-stock policy with censored ordering. This is an order-up-to policy but with its replenishment orders censored by the capacity constraint. We refer to this heuristic as the modified constant base-stock policy (MCBS). We use a numerical experiment to compare the performance of the heuristic policy to the optimal policy. We find that the performance of the heuristic relative to the optimal policy improves with a tighter capacity constraint. We also observe that the performance of the heuristic itself can sometimes be improved by tightening the capacity constraint. We then use the results for a single-stage system to model a multi-stage serial system that operates with a constant base-stock policy with censored ordering, i.e., a MCBS policy. We show how to adapt the existing dynamic programming algorithm for the unconstrained case to solve for the safety stock locations and base-stock levels in a capacitated serial system. We describe how to extend this model to other supply chain topologies. We report on numerical tests for serial systems, and find that the best MCBS policy in a capacitated system can outperform the best constant base-stock policy in an identical but uncapacitated system.

Key Words: guaranteed service; multi-echelon inventory system; safety stock optimization

Acknowledgements
This research has been funded in part by the Singapore MIT Alliance (SMA) and by Dr. Marcus Wallenberg's Foundation for Education in International Industrial Management. The authors thank the guest editor Alan Scheller-Wolf, and the anonymous associate editor and referees for their helpful feedback and valuable suggestions that have improved the paper.

1. Introduction

A central question in supply chain management is how to coordinate activities and inventories over a large number of stages and locations, while providing a high level of service to end-item customers. Simpson (1958) found that if the individual stages in a serial-system supply chain operate according to base-stock policies with service guarantees, then the optimal safety stock strategy is to concentrate inventory at a few key locations, effectively decoupling the supply chain into independent segments. Once the optimal safety stock strategy has been determined, each segment of the supply chain, consisting of a set of adjacent stages, can operate with a simple base-stock policy, with a minimum need for communication and coordination between the different segments of the supply chain. Graves and Willems (2003) term this framework for supply chain management as the guaranteed service (GS) model.

In reality, there are often limits to the quantity of goods that can be transported, processed or stored in a given time frame. If a stage is unable to process a large order in a short period of time, then this may cause delays and stock-outs that are not anticipated by the existing GS models. The intent of this paper is to extend the GS model for safety stock placement to supply chains with capacity constraints. In particular we strive to develop models and algorithms that have the potential to determine where to place strategic safety stocks in real-world supply chains. We contend that the paper makes the following contributions.

First we characterize the optimal policy for a single-stage capacitated system that is required to provide guaranteed service, assuming bounded demand. As we expect this policy is difficult to implement, we then develop a constant base-stock policy with censored ordering as a heuristic. We refer to this heuristic as the modified constant base-stock policy (MCBS). The MCBS policy is a generalization of Kimball's classic base-stock model (1955, reprinted in 1988) to account for a capacity constraint. For the most common specification of the demand bound, we provide simple analytical expressions for the base stock and safety stock; these expressions show clearly the impact from the capacity constraint. We briefly discuss how this result has been used to guide inventory policy for a solar panel manufacturer.

We then consider a serial-system supply chain, where each stage can have a capacity constraint and operates with a MCBS policy. With the assumption of a concave demand bound and guaranteed service, the optimal placement of safety stocks for a serial-system supply chain satisfies the all-or-nothing property: that is, each stage either holds a decoupling safety stock or no safety stock. We then describe how we can modify the efficient safety-stock optimization methods developed for unconstrained supply chains to determine the safety stock placement for capacitated supply chains for various supply chain topologies.

From a numerical experiment we develop some findings for a serial supply chain on the performance and the structure of the safety stock configuration, as it depends on the location and tightness of the capacity constraint. We also find that the best MCBS policy in a capacitated system can sometimes require less inventory than an identical but uncapacitated system operating with a constant base-stock policy. Finally, for the MCBS policy the orders still depend only on local information.

This paper consists of five sections. In the remainder of this section, we provide a review of related literature. In §2, we treat the single-stage system, characterize the optimal policy and introduce and analyze the MCBS policy. In §3, we consider the optimization of a capacitated serial supply chain, assuming a MCBS policy. In §4, we perform a numerical experiment to examine the structure and performance of these policies. We conclude the paper with a discussion on our findings and suggestions for possible future work in §5.

Literature Review

For a review of work on GS models we cite the overview articles of Inderfurth (1991), Diks et al. (1996) and Graves and Willems (2003), as well as the thesis Eruguz (2014) and the recent survey article by Eruguz et al. (2015). Graves and Willems (2000) extend Simpson's work to supply chains with spanning tree topology, and formulate an efficient dynamic programming algorithm. Optimizing general networks is an NP-hard problem (Lesnaia et al. 2005); nevertheless, Humair and Willems (2006, 2011) and Magnanti et al. (2006) have developed algorithms for optimizing the safety stocks in large-scale real-world supply chains. Sitompul et al. (2008) is the only paper we found that includes a capacity constraint; the paper empirically estimates an approximate correction factor to account for the impact on the required safety stock due to the capacity constraint. We note that the GS framework has been deployed successfully in multiple industries (e.g., Graves and Willems, 2000, Billington et al. 2004, Willems 2008, Farasyn et al. 2011, Hsieh 2011, Masse, 2011, Polak 2014). Some recent extensions include the treatment of non-stationary demand (Graves and Willems 2008), evolving forecasts (Schoenmeyr and Graves 2009) and dual sourcing (Klosterhalfen, Minner and Willems 2014).

There is a larger literature examining capacity constraints for stochastic service (SS) models, in which the delivery or service time between stages can vary depending upon the material availability at the supply stage. The main focus for much of this literature has been to characterize the structure of the optimal ordering policy. Gallego and Scheller-Wolf (2000) find that the optimal policy for a single-stage system with fixed ordering costs and capacity constraints has “an (s,S)-like structure.” Gallego and Toktay (2004) also consider a single-stage system under the assumption that all non-zero orders are full capacity orders, and show that the optimal policy is a threshold policy.

For multi-stage systems, Speck and van der Wal (1991) show by example that the echelon-based ordering policy of Clark and Scarf (1960) is generally not optimal in a two-stage serial system with capacity constraints. Parker and Kapuscinski (2004) also analyze a two-stage serial system, and show that a modified echelon base-stock (MEBS) policy is optimal, when the tightest capacity constraint is at the downstream stage and there is a single period lead-time at the upstream stage. For these assumptions the MEBS policy is the same as the censored ordering policy that we propose and analyze. However, for serial systems with more than two stages and non-unit lead-times, these policies will differ. In particular, the MEBS policy is a modification to an echelon base-stock policy; as such, an upstream stage needs to see the end-item demand and will replenish accordingly, subject to a possible capacity constraint at the stage. Our censored ordering policy is a modification to a local base-stock policy, whereby an upstream stage only sees the orders from its adjacent downstream stage. Furthermore, at each capacitated stage, the stage orders the lesser of its capacity and the shortfall between its inventory position and its local base-stock level. As such, the order signal gets censored at each stage as it is passed upstream.

Janakiraman and Muckstadt (2009) examine the same setup as Parker and Kapuscinski, but allow for the lead time at the upstream supplier to be two periods; they characterize the form of the optimal policy, as a two-tier base-stock policy. The paper discusses how the structure of the optimal policy depends on the lead times and on the number of stages, finding that the number of parameters will grow exponentially.

Glasserman and Tayur (1994) consider the stability properties of multi-echelon systems with capacity constraints. They find that inventories and back-logs converge to unique stationary distributions if the mean demand is less than the capacity constraint. In subsequent work, Glasserman and Tayur (1995, 1996) assume that an echelon base-stock policy is used in a multi-echelon system, and find optimal order points using simulation and perturbation analysis (1995), and analytical approximations (1996).

Huh et al. (2010) consider a serial system with unit lead times and the assumption of an echelon base-stock ordering policy. They develop a recursive expression for determining the shortfall at each stage, which can then be used to re-prove the stability properties given in Glasserman and Tayur (1994). Huh et al. (2014) develop three heuristics for a capacitated serial system, and provide a cost bound; these heuristics have both theoretical and intuitive properties.

Gupta and Selvaraju (2006) develop an approximation for analyzing a serial system as a queuing network, where the stages have exponentially-distributed process times and operate according to echelon base-stock policies. In this way, they can characterize and optimize a two-stage system with capacity constraints. This approach is computationally costly for larger systems, for which the paper proposes some approximations.

Bertsimas and Thiele (2004, 2006) propose a robust optimization model and show that capacity constraints can be incorporated into a tractable, robust optimization problem. This approach can handle general networks, with echelon policies. A similarity between the robust optimization approach and ours is that there is no need to specify a probability distribution for demand. However, our work differs from *all* of the aforementioned work in that we consider *local* base-stock ordering policies, and, as mentioned, guaranteed service constraints rather than backorder costs and stochastic service. Moreover, we will show that existing optimization algorithms (Graves and Willems, 2000) can be generalized to handle capacity constraints. In practice, these methods are fast enough to handle systems with thousands of stages.

2. Single-stage model

A stage represents a processing activity that requires one or more inputs and that converts these inputs into a single output product or final good. The stage might represent the procurement of a raw material, or the production of a component, or the manufacture of a subassembly, or the assembly and test of a finished good, or the transportation of a finished product from a distribution center to a warehouse. We can store the output of the stage as inventory that can then be used to meet customer demand or as input into downstream stages.

The stage has a capacity limit *c*: each period the stage can release into its process any amount up to the capacity limit of *c*, assuming that sufficient inputs are on hand. We assume a deterministic process lead-time T, a non-negative integer, equal to the time between when the process starts and when the process completes and the inventory is available to serve demand. The process lead-time represents any fixed processing time at the stage and does not include any time waiting for input or for capacity. For instance, for a transportation stage, the lead-time is the time to transport inventory (say) from a manufacturing plant to a distant warehouse; for a manufacturing stage, the lead-time might be a batch processing time. The lead-time can be zero, if there is no fixed processing time.

We let $d(t)$ denote the demand in period *t*. We assume that $d(t)$ is non-negative, with average value μ where $\mu < c$. We assume that the stage provides a *guaranteed service time S*: the stage will satisfy the demand $d(t)$ by time $t + S$, where *S* is a non-negative integer.

Furthermore, as in Kimball (1988) and Simpson (1958), for the purposes of setting safety stock levels, we assume that *demand is bounded.* Specifically we assume that there exists a function $D(s)$ that bounds demand over any *s* consecutive periods. That is,

$$D(s) = \max\left\{d(t, t+s)\right\} \quad \forall t, s \geq 0 \tag{1}$$

where we define the notation

$$d(a,b) = \begin{cases} \sum_{i=a+1}^{b} d(i) & \text{for } a < b \\ 0 & \text{for } a = b \\ -\sum_{i=b+1}^{a} d(i) & \text{for } a > b \end{cases}.$$

Guaranteed service and bounded demand constitute the most significant assumptions in the GS framework. Simply put, we assume that as long as demand stays within certain bounds, there should always be enough safety stock to meet that demand within the service time. This general approach applies well to the typical context in which the implicit and explicit costs of stocking out are perceived to be much greater than the costs of holding inventory. We refer to Graves and Willems (2000) for more discussion and motivation of these assumptions.

We define a class of demand bounds and capacity constraints that will be of prime interest to us, as follows.

Defintion 1. *A bound function* $D(s)$ *on* $s \in [0, \infty)$ *is said to be* valid *if* $D(0) = 0$, *and if it is non-decreasing, and concave. For* $s < 0$ *we define* $D(s) = 0$.

Definition 2. *A capacity constraint c is said to be* valid *with respect to a valid bound function* $D(s)$ *if there exist a single point* $\tilde{s} > 0$ *such that* $D(\tilde{s}) = c\tilde{s}$, *and that*

$D(s) > cs \quad \forall s < \tilde{s}$ *and* $D(s) < cs \quad \forall s > \tilde{s}$.

Demand bounds and capacity constraints that arise in practice are likely to be valid. The maximum possible demand over some time period will increase with the length of the period. We expect that it increases with a diminishing rate, due to increased risk pooling of the demand variability over longer periods of time. We also assume demand can exceed the production capacity over some time interval (otherwise the capacity constraint is not relevant), but given sufficiently long time there must be enough capacity to meet any valid demand realization (otherwise guaranteed service is infeasible).

To simplify the presentation we assume the stage has a single supplier that provides a single input. This supplier quotes a guaranteed service time, denoted by a non-negative integer *SI* for the inbound service time. Thus, for an order placed at time *t*, the supplier delivers its input to the stage at time *t* + *SI*.

We assume that the stage operates with a periodic-review policy with a review period being one time period (e.g., one day). The timing of events is as assumed by Kimball (1988) and Simpson (1958). In each period *t*, the stage first observes its demand $d(t)$ and then places an order on its upstream supplier. The stage then receives the earlier order placed at time *t* – *SI* from its upstream supplier, decides the quantity to release into its process, and completes the process on the release quantity from time *t* – *T* and places this quantity into its inventory. Finally the stage serves the demand from period *t* – *S*, namely $d(t-S)$. We illustrate the system in Figure 1.

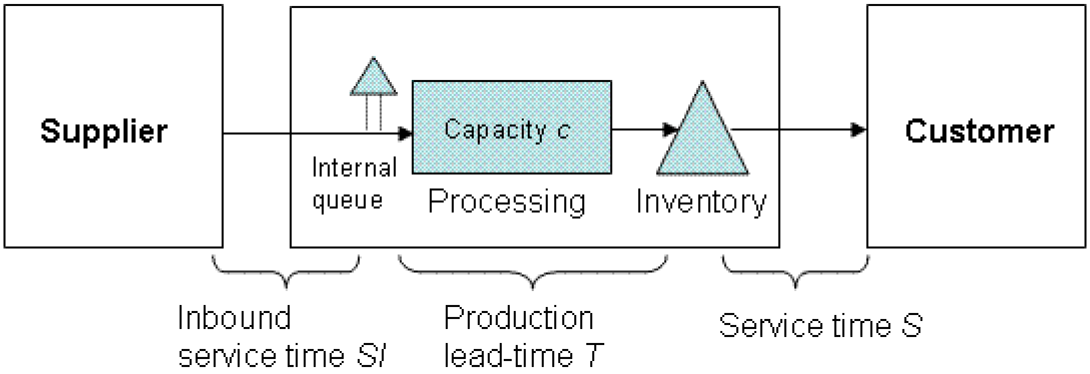

Figure 1: Overview of a single-stage system

Optimal Policy for a Single-stage System

In this section we characterize the optimal inventory policy for the single-stage GS model. Without loss of generality, we assume $SI = 0, T = \tau, S = 0$, as the arguments can easily be extended by translation to other settings for which $SI + T - S = \tau$.

We define $O(t)$ equal to the replenishment order placed at period *t*, which is delivered into inventory at period $t + \tau$. We need to set $O(t)$ to assure guaranteed service; that is we need to set it so that $I(t + \tau + k) \geq 0, \forall k \geq 0,$ where *I(t)* denotes the on-hand inventory at the end of period *t*. As of time *t*, we can express $I(t + \tau + k), \forall k \geq 0$ as:

$$I\left(t+\tau+k\right)=I\left(t\right)+\sum_{i=1}^{\tau-1}O\left(t-\tau+i\right)+\sum_{i=0}^{k}O\left(t+i\right)-d\left(t,t+\tau+k\right)$$

where $\sum_{i=1}^{\tau-1}O\left(t-\tau+i\right)$ are the known orders placed prior to time *t* that have yet to be received, $\sum_{i=0}^{k}O\left(t+i\right)$ represent the unknown orders that will be placed in the time interval $\left[t,t+k\right]$ and will arrive prior to time $t+\tau+k,$ and $d\left(t,t+\tau+k\right)$ is the unknown demand over the interval $\left(t,t+\tau+k\right]$. In order to assure that $I\left(t+\tau+k\right)\geq 0,\forall k\geq 0,$ we require that

$$\sum_{i=0}^{k}O\left(t+i\right)\geq \max\left\{d\left(t,t+\tau+k\right)\right\}-I\left(t\right)-\sum_{i=1}^{\tau-1}O\left(t-\tau+i\right). \qquad (2)$$

That is, the orders placed in the interval $\left[t,t+k\right]$ must exceed the demand over the interval $\left(t,t+\tau+k\right]$, net of the current inventory position. We now find the minimal upper bound on the unknown demand $d\left(t,t+\tau+k\right)$, as of period t. As of period *t*, we have observed the realized demand in the interval $\left(s,t\right]$ for all $s<t$, which we denote as $d\left(s,t\right)$. From the definition of the demand bound, we observe for any non-negative integer *n*:

$$d\left(t-j,t+n\right)=d\left(t-j,t\right)+d\left(t,t+n\right)\leq D\left(j+n\right),\forall j=0,\dots,t.$$

Thus, at any time *t*, we can use the demand history to create a new bound on the unknown $d\left(t,t+n\right)$, namely:

$$d\left(t,t+n\right)\leq \min_{j=0,\dots t}\left\{D\left(j+n\right)-d\left(t-j,t\right)\right\}.$$

We can now use this demand bound to characterize the minimal order quantity. We first re-write the requirement on future orders (2) as:

$$\sum_{i=0}^{k}O\left(t+i\right)\geq \min_{j=0,1\dots t}\left\{D\left(j+\tau+k\right)-d\left(t-j,t\right)\right\}-I\left(t\right)-\sum_{i=1}^{\tau-1}O\left(t-\tau+i\right).$$

We then set $O\left(t\right)$ to the minimal value that satisfies the above constraint for all $k\geq 0$, subject to future orders being within the capacity limit. The minimum for $O\left(t\right)$ is achieved by setting each future order equal to the capacity limit, i.e., $O(t+i)=c, \forall i>0$; hence, we set the current order quantity $O(t)$ to the following quantity:

$$O(t)=\max_{k\geq 0}\left\{\min_{j=0,\ldots t}\left\{D(j+\tau+k)-d(t-j,t)\right\}-kc\right\}-I(t)-\sum_{i=1}^{\tau-1}O(t-\tau+i). \tag{3}$$

Although the maximization is over an infinite set, this is not a problem; if the demand bound and capacity are valid, then there is a finite *k* beyond which we can bound the expression within the maximization.

We claim that this is the optimal order quantity. The argument has two steps. We first need to show feasibility with regard to the capacity constraint. We show this by induction in the Online Supplement, with a required boundary condition on the starting inventory: $I(0)=\max_{k\geq 0}\left\{D(\tau+k)-kc\right\}$. We then argue that this is the minimal order quantity that assures guaranteed service. If one orders less than in (3), then there is a realization for $d(t,t+\tau+k)$ for some *k* that is within the demand bounds and for which there is not guaranteed service. Thus, any smaller value is not feasible given the assumption of guaranteed service.

As a special case we consider the uncapacitated problem; we immediately find the optimal order quantity to be: $O(t)=\min_{j=0,\ldots t}\left\{D(j+\tau)-d(t-j,t)\right\}-I(t)-\sum_{i=1}^{\tau-1}O(t-\tau+i).$

Modified Constant Base-stock Policy

The optimal policy depends upon the entire demand history, and as such could be a challenge to implement. Furthermore, the single-stage model, operating with the optimal ordering policy, is a relatively cumbersome "building block" for modeling a supply chain consisting of multiple stages. For these reasons, we will consider a non-optimal policy, namely a modified constant base-stock policy. Such policies are easy to implement and indeed are pervasive in practice. And as will be seen, by assuming that each stage operates with a modified constant base-stock policy, we can readily model supply chains consisting of multiple stages.

The modified constant base-stock policy is parameterized by a single parameter *B* for the base stock. The operating policy is to order in each period an amount that raises the inventory position (defined as the on-hand inventory plus inventory in process and inventory on order) to as close to the base stock as possible. When there is no capacity constraint, the policy will place a replenishment order, equal to *d(t),* on its upstream supplier in each period *t*. Given the assumption of guaranteed service and no lost sales, this ordering policy maintains the inventory position at *B* in each period. However, when there is a capacity constraint, we can observe inefficiency with this ordering policy. Consider what happens whenever $d(t) > c$. At time *t*, the stage places an order, equal to $d(t)$, on its upstream supplier. This order is filled at time *t+SI*. But since $d(t) > c$, only *c* units of the replenishment will be processed at the stage, due to the stage's capacity constraint. The remainder will sit in an internal queue at the process until capacity becomes available in a subsequent period. We can avoid this inefficiency with a censored ordering policy: the basic idea is that a stage should not propagate its base-stock order upstream, if it knows that it is unable to process such a quantity because of its capacity constraint.

We term this policy with censored ordering as the modified constant base-stock (MCBS) policy. To specify the MCBS policy, we define the order backlog $BL(t)$, which is the difference between the base stock *B* and the inventory position at the end of period *t*. In effect, the backlog is the amount of demand for which a replenishment order has yet to be placed as of the end of period *t*. Then in each period the amount that is ordered is the minimum of the capacity and the remaining backlog plus the current demand:

$$O(t) = \min\{c, BL(t-1) + d(t)\}. \tag{4}$$

We can use this order equation to express the backlog $BL(t)$ recursively:

$$\begin{aligned} BL(t) &= BL(t-1) + d(t) - O(t) = \max\{BL(t-1) + d(t) - c, 0\} \\ &= \max\{\max\{BL(t-2) + d(t-1) - c, 0\} + d(t) - c, 0\} \\ &= \max_{n\in Z}\{d(t-n,t) - cn\}, \end{aligned} \tag{5}$$

where *Z* denotes the set of non-negative integers.

To develop an expression for *I(t)*, we first write the inventory balance equation for the stage. We observe that at time *t* the stage needs to satisfy the demand that occurred in period *t-S*, according to the definition of the service time *S*. Also, at time *t* the stage receives into inventory its order placed at time *t* – *SI* – *T*: the upstream supplier fulfills this order at time *t* – *T*, given its service time of *SI*; the stage enters this order into its process at time *t* – *T* and the order is completed at time *t*, according to the definition of the process lead time *T*. Hence, the balance equation for the finished goods inventory is:

$$I(t) = I(t-1) + O(t - SI - T) - d(t-S),$$

where (4) specifies the MCBS policy that assures feasibility with respect to the capacity constraint. We can substitute $O(t) = BL(t-1) + d(t) - BL(t)$ into the balance equation to get:

$$I(t) + BL(t - SI - T) = I(t-1) + BL(t - SI - T - 1) + d(t - SI - T) - d(t-S). \tag{6}$$

If the system starts at time *t = 0* with $d(t)=0, BL(t)=0$ for $t \le 0$ and $I(0)=B$, then for suitably large *t* we can write the inventory as:

$$I(t) = B - d(t - SI - T, t - S) - BL(t - SI - T) \qquad (7)$$

We substitute (5) into (7) to obtain:

$$I(t) = B - \max_{n \in Z}\left\{d(t - SI - T - n, t - S) - cn\right\}. \qquad (8)$$

For the guaranteed service model, we set *B* to the minimal constant that assures that the inventory $I(t)$ is non-negative.

We now invoke the assumption of bounded demand to develop an explicit expression for the base stock. We can combine (1) and (8) to find that the minimal base stock for $I(t) \ge 0$ is:

$$B(\tau) = \max_{n \in Z}\left\{D(\tau + n) - cn\right\} \quad \text{where} \quad \tau = T + SI - S\ . \qquad (9)$$

In (9) τ denotes the *net replenishment time* for the stage without the capacity limit. We write the base stock in (9) as a function of τ to make explicit its dependence on this parameter. When we optimize the safety stocks across a supply chain, the decision variables will be the service times (*S, SI*) for each stage, which combine with the given lead-time *T* to determine the unconstrained net replenishment time τ.

When the demand bound and capacity constraint are valid, we can solve (9) by enumeration over a finite range, as the demand bound will eventually fall below the capacity, i.e., $D(s) < cs, \forall s > \tilde{s}$.

In order to get an explicit solution to the maximization in (9), we assume the demand bound is differentiable and ignore the integrality restriction on the argument *n*.[1] We define θ by $\frac{dD(\theta)}{d\theta} = c$, i.e., θ is the point at which the derivative of the demand bound equals the capacity. Then the base stock, as a function of the net replenishment time, is:

[1] When we ignore the integrality restriction on *n* in (9) we get an upper bound on the base-stock level, which in theory can be arbitrarily bad. However, our purpose here is to get an analytically-based understanding of the model behavior for reasonable demand bounds; we do not make this relaxation in any of the computational algorithms for finding the actual base stocks.

$$B(\tau)=\begin{cases} 0 & \text{for } \tau<\theta-\dfrac{D(\theta)}{c} \\ c\times(\tau-\theta)+D(\theta) & \text{for } \theta-\dfrac{D(\theta)}{c}\le\tau<\theta. \\ D(\tau) & \text{for } \tau\ge\theta \end{cases} \tag{10}$$

We see from (10) that if the net replenishment time is sufficiently negative (i.e., $\tau<\theta-\dfrac{D(\theta)}{c}<0$), the base stock is zero; in this case, the service time *S* is sufficiently long that the stage can produce to order, even with the capacity constraint. For higher values of τ, the base stock grows linearly at rate *c* until $\tau=\theta$. Beyond $\tau=\theta$, the capacity constraint does not matter: the base stock equals the demand bound function, as is true for the unconstrained base-stock policy.

We note that unlike the unconstrained base-stock model, we permit the net replenishment time to be negative when we have a capacity limit. Thus, the stage might quote a service time *S* that is ***longer*** than its nominal replenishment time $SI+T$; but due to the capacity constraint, the stage may still need a positive base stock in order to provide guaranteed service. For instance, consider $\tau=0$, at which the service time *S* equals the replenishment time $SI+T$; in order for the stage to provide guaranteed service, it must have a positive base stock equal $D(\theta)-c\theta>0$.

In the Online Supplement we show that the base stock is a concave function of τ, namely:

Lemma 1. *Suppose* $D(\tau)$ *is valid and differentiable, and c valid with respect to* $D(\tau)$**.** *Then* $B(\tau)=\max_{n\ge0}\{D(\tau+n)-cn\}$ *is concave on* $\tau\in[\theta-\dfrac{D(\theta)}{c},\infty)$.

In practice we often set the demand bound as

$$D(t)=\mu t+z\sigma\sqrt{t} \tag{11}$$

where σ corresponds to a measure of demand variability per period, such as the standard deviation, and *z* is a safety factor. For instance, if one were to assume i.i.d. normally distributed demand, then one could interpret this demand bound as being the $\Phi^{-1}(z)$ percentile of demand, for $\Phi(\ )$ being the standard normal cdf.

We note that this bound (11) is valid and differentiable, and that any $c > \mu$ will constitute a valid capacity constraint. For illustrative purposes, suppose *z = 2*; then we have $\theta = \left(\frac{\sigma}{c-\mu}\right)^2$ for this demand function and from (10) the base stock is:

$$B(\tau) = \begin{cases} 0 & \text{for } \tau < -\frac{\sigma^2}{c(c-\mu)} \\ c\tau + \frac{\sigma^2}{c-\mu} & \text{for } -\frac{\sigma^2}{c(c-\mu)} \le \tau < \left(\frac{\sigma}{c-\mu}\right)^2 \\ \mu\tau + 2\sigma\sqrt{\tau} & \text{for } \tau \ge \left(\frac{\sigma}{c-\mu}\right)^2 \end{cases}. \qquad (12)$$

Thus the base stock with a capacity constraint takes a rather simple and intuitive form. Again the capacity constraint is immaterial when $\tau \ge \theta = \left(\frac{\sigma}{c-\mu}\right)^2$. Second, when the capacity constraint is relevant, the base stock depends not just on the demand variability σ and net replenishment time τ but also on the amount of headroom or slack capacity: $c-\mu$. Third, in this range the base stock ***increases linearly at rate c*** in the net replenishment time. We illustrate these observations in Figure 2(a), where we plot the capacitated base stock level (12) together with the base stock level for the unconstrained case.

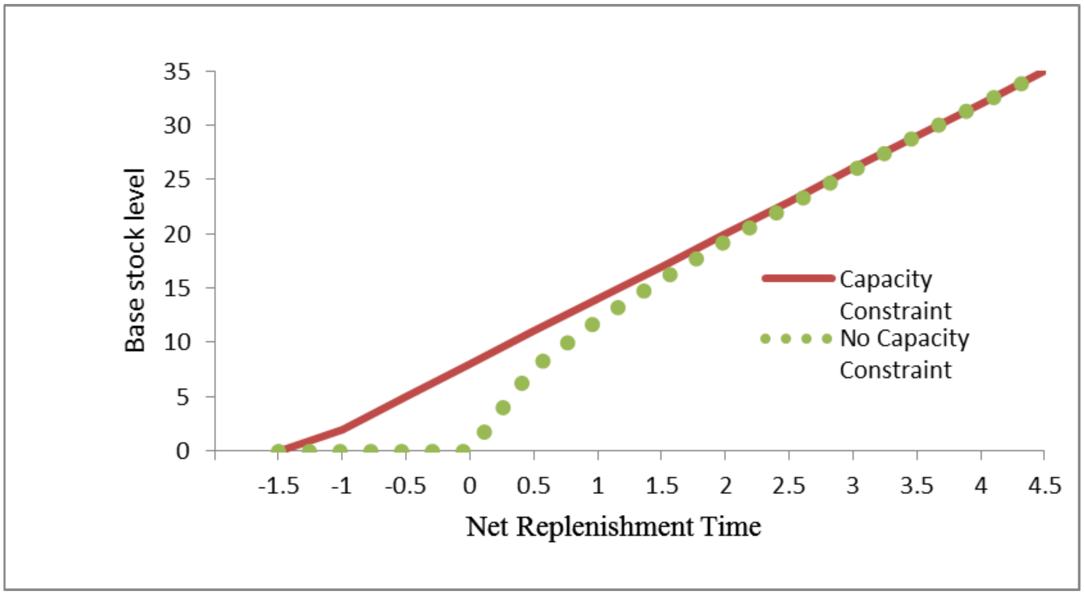

Figure 2(a): Base stock for $\mu = \sigma = 4$, $c = 6$ and $z = 2$

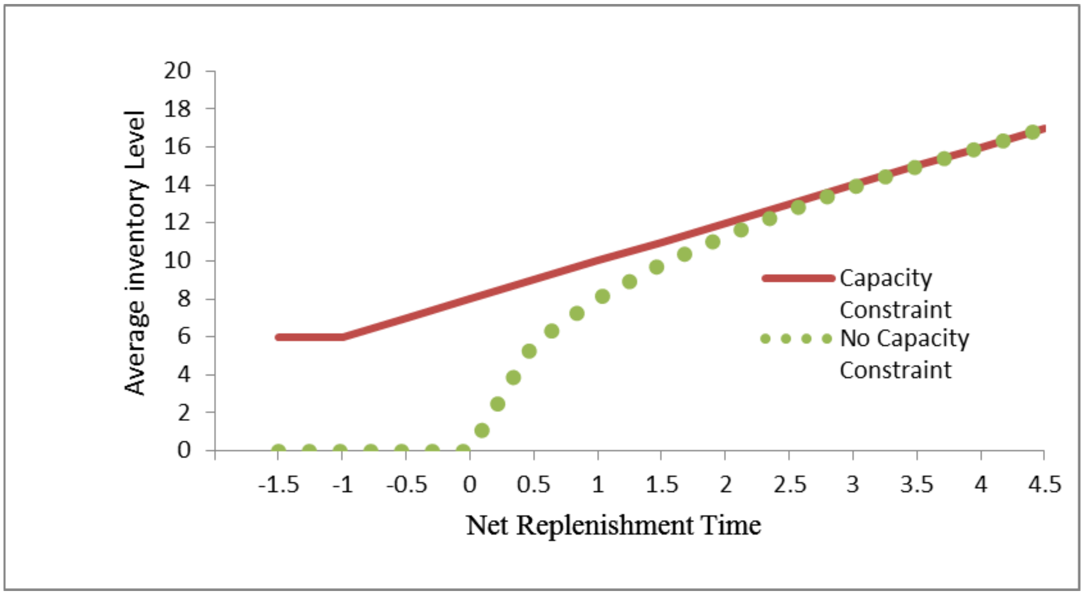

Figure 2(b): Average inventory level for $\mu = \sigma = 4$, $c = 6$ and $z = 2$

To get some additional insight, we use (7) and (12) to find the average inventory plus backlog, $E\left[I(t)\right] + E\left[BL(t - SI - T)\right]$; as we expect the backlog to be modest, we use this to approximate the average safety stock:

$$B(\tau) - \mu\tau = \begin{cases} (c-\mu)\tau + \dfrac{\sigma^2}{c-\mu} & \text{for } -\dfrac{\sigma^2}{c(c-\mu)} \le \tau < \left(\dfrac{\sigma}{c-\mu}\right)^2 \\ 2\sigma\sqrt{\tau} & \text{for } \tau \ge \left(\dfrac{\sigma}{c-\mu}\right)^2 \end{cases}. \qquad (13)$$

In Figure 2(b) we plot (13) for a capacitated example, as it depends on the net replenishment time and compare this to the unconstrained case. We again note that there is a threshold value for the net replenishment time, beyond which the capacity constraint does not matter. When the capacity constraint is relevant, the inventory is a ***fixed amount*** $\dfrac{\sigma^2}{c-\mu}$ plus a ***variable amount that increases linearly*** in the net replenishment time. In contrast, for the unconstrained system, the safety stock is ***proportional to the square root*** of the net replenishment time. Finally, as the slack capacity goes to zero, the fixed amount of safety stock increases hyperbolically; this is analogous to what happens to the waiting time in a queue as utilization goes to one. Similar to the traditional square-root formula for safety stocks in an unconstrained setting, we regard (13) to be a valuable back-of-the-envelope approximation in that it succinctly displays how the safety stock depends on the slack capacity, the demand variability and the net replenishment time.

To illustrate the potential value from this model, we briefly describe a real-world situation for which these results have been applicable. First Solar manufactures thin-film solar panels using a highly automated process. The production lead-time is virtually zero compared to the times needed to replenish materials (e.g., hours compared to weeks). Raw materials are relatively cheap and are held in inventory. The traditional unconstrained base-stock policy suggests that little or no finished goods inventory is needed, as the replenishment time is so short. However, the automated process has significant capacity constraints. If there is an upswing in demand, then there would not be enough capacity to meet that demand. As a consequence, First Solar finds it necessary to keep significant safety stocks of finished goods inventory. The formula (13) has been used to guide these levels as it provides a simple way to see how the safety stock depends on the key drivers, namely, the demand variability, the replenishment lead-time, and the amount of slack capacity or headroom.

Comparison between Optimal Policy and MCBS Policy

We have noted earlier a rationale for considering the MCBS policy: it is relatively easy to implement and it lends itself to modeling more complex, multi-stage systems. Yet these advantages come at a cost, as the policy is not optimal. In this section we report on a representative numerical experiment comparing the MCBS policy with the optimal policy.

We specify the demand bound as follows:

$$D(t)=\begin{cases} \mu t + z\sigma\sqrt{t} \ \ for \ \ 0 \le t \le b \\ \mu b + z\sigma\sqrt{b} + s\times(t-b) \ \ for \ \ t > b \end{cases}$$

That is, the demand bound follows the common form (11) up to a break point *b*; beyond the break point the bound grows linearly. This form was chosen as it facilitated both the generation of random demand that complied with the demand bound, as well as the simulation of the optimal and MCBS policies.

For the experiment, we set: $\mu = 40, \sigma = 10, z = 2, b = 10, s = 42, \tau = 1$. We created 20 test cases by varying the capacity: $c = 42, 43, \dots, 60, \infty$. For each test case, we simulated each policy on a randomly-generated demand series over 50,000 periods, divided into ten batches. In the Online Supplement we detail how the demand series were generated. In Table 1, we report for each test case the average inventory estimate for each policy, along with the larger of the two standard errors for the estimates.

capacity	MCBS policy	optimal policy	% increase	std error
42	28.9	28.9	0.00%	0.44
43	28.0	27.9	0.40%	0.15
44	23.8	23.7	0.24%	0.14
45	21.6	21.3	1.25%	0.10
46	20.1	19.6	2.67%	0.08
47	19.5	18.7	4.22%	0.08
48	19.1	18.1	5.07%	0.05
49	19.0	17.7	7.52%	0.08
50	19.3	17.6	9.77%	0.10
51	19.6	17.6	11.39%	0.05
52	19.7	17.4	13.51%	0.05
53	19.8	17.1	15.31%	0.09
54	19.9	17.1	16.66%	0.18
55	19.8	16.6	19.28%	0.11
56	19.9	16.5	20.31%	0.09
57	20.0	16.6	20.66%	0.11
58	20.0	16.4	21.84%	0.16
59	20.0	16.3	23.11%	0.15
60	20.0	16.2	23.96%	0.16
Infinite	20.0	15.4	30.10%	0.20

Table 1: Comparisons of average inventory for MCBS versus optimal policy

We make two observations from these results. First, we see that the performance of the MCBS policy depends on the tightness of the capacity constraint. When the constraint is tight, the MCBS policy appears to be very good. For instance, when $c \le 1.1\mu$, the average inventory for the MCBS policy is within 0.5% of that for the optimal policy. However, the performance of the heuristic degrades as we loosen the capacity constraint. Indeed for the uncapacitated case, the MCBS policy requires about one third more inventory than for the optimal policy.

As explanation, the optimal policy exploits the demand history to develop a tighter bound on future demand. When demand is very high in one or more periods, then the demand will need to be relatively low in the following period(s) in order to stay within the demand bound. The optimal policy recognizes this and will flex its orders accordingly; in effect, the optimal policy operates with a state-dependent base-stock level, whereby it continually adjusts the base stock to reflect the demand history. In contrast, the MCBS policy ignores the demand history and continues to order up to its constant base stock in each period. A capacity constraint, however, seems to limit the ability of the optimal policy to exploit the demand history. The capacity constraint acts as damper on the optimal policy, limiting its ability to flex its orders in response to the demand history. When demand is high, the optimal policy (and MCBS policy) can only replenish up to the capacity limit; the replenishment of demand beyond the capacity must be postponed until a future period. The resulting orders for both policies will be less variable as the capacity limit becomes tighter; as a consequence, a tighter capacity constraint reduces the advantage that the optimal policy has in terms of using the demand history.

The second observation is that the performance of the MCBS policy is not monotonic in the capacity. Of interest, we see that the average inventory with a capacity constraint can be less than that for the unconstrained system. For instance, when $c = 49$, the average inventory is 19.0, which is 5% less than the average inventory for the unconstrained system. This is surprising. The explanation is similar to that for the first observation. The MCBS policy ignores the demand history. Thus, when demand is high, the MCBS policy will try to order up to the constant base stock; this is unnecessary as future demand must be relatively low, in order to stay within the demand bound. Imposing a capacity constraint can help to alleviate this response, by limiting the policy's ability to order too much, i.e., to order up to the base stock, when demand is high.

One implication from this observation is that for the GS model with bounded demand, imposing a capacity constraint can improve the performance of a MCBS policy. Adding capacity provides a way to adapt the policy in light of the demand history: the capacity limit prevents the policy from over-ordering when demand is high, and as a consequence, allows the policy to adapt to the fact that future demand will be relatively less. In the Online Supplement we provide a small example illustrating this improvement.

3. Multi-stage model

In this section we describe how to extend the single-stage model to a multi-stage system. To simplify the presentation, we consider a serial system, where we number the stages or nodes from downstream to upstream: node 1 is the customer-facing node, and N is the most upstream node. We will briefly discuss the extension to other network structures at the end of this section.

Graves and Willems (2000) find the optimal *constant* base-stock policy for a serial system with bounded demand, with guaranteed service for the external customer and with no capacity constraints. However, as noted above, the constant base-stock policy is not optimal under the assumptions of bounded demand and guaranteed service, even for a single stage. We have not been able to establish the form of the optimal policy for a serial system under assumptions of bounded demand and guaranteed service. The main difficulty arises in characterizing how the demand process gets transformed as it is passed upstream from one stage to another. For instance, we have not been able to see how to propagate the optimal single-stage order process given by (3), which may or may not be the optimal order process for a multi-stage system.

Hence, in this section we will assume that each stage operates with guaranteed service and with a MCBS policy. This will permit tractability and is the natural extension of the prior literature on the GS model. We associate with each stage *k* a service time S_k, a lead-time T_k and an inbound service time SI_k. The customer for node *k* is node *k-1*; hence the stage-*k* service time S_k equals the inbound service time for stage *k-1*, SI_{k-1}, for k= 2,…N. We assume that each stage operates with a *local* base-stock policy with censored ordering as described for the single-stage system, i.e., the MCBS policy. When we censor orders, each node can generate a different series of orders due to its capacity constraint. We denote the order *received by* node *k* at time *t* by $d_k(t)$, where $d_1(t)=d(t)$. Each stage also maintains an order backlog denoted by $BL_k(t)$. As the order received by stage *k* is the order placed by stage *k-1*, we have for *k* = 2,…N:

$$d_k(t)=\min(d_{k-1}(t)+BL_{k-1}(t-1),c_{k-1}). \qquad (14)$$

The order backlog is given by:

$$\begin{aligned} BL_k(t) &= \max\left\{BL_k(t-1)+d_k(t)-c_k,0\right\} \\ &= \max\left\{\max\left\{BL_k(t-2)+d_k(t-1)-c_k,0\right\}+d_k(t)-c_k,0\right\} \\ &= \max_{n\in Z}\left\{d_k(t-n,t)-c_k n\right\}. \end{aligned} \qquad (15)$$

Here we assume that $BL_k\left(t\right)=0, d_k\left(t\right)=0$ for $t\le 0$.

We can show that the inventory for node *k*, $I_k(t)$, is the inventory for the unconstrained problem, net of the backlog at time $t-SI_k-T_k$. Since the replenishment time is SI_k+T_k, anything in the backlog at node *k* at the end of time $t-SI_k-T_k$ cannot reach the inventory by time *t*. Thus, we have:

$$\begin{aligned}I_k(t) &= B_k - d_k(t - SI_k - T_k, t - S_k) - BL_k(t - SI_k - T_k) \\ &= B_k - d_k(t - SI_k - T_k, t - S_k) - \max_{n \in Z}\left\{d_k(t - n - SI_k - T_k, t - SI_k - T_k) - c_k n\right\} \\ &= B_k - \max_{n \in Z}\left\{d_k(t - n - SI_k - T_k, t - S_k) - c_k n\right\}.\end{aligned} \tag{16}$$

Except for the fact that the demand d_k is now stage-specific, this is the same inventory equation as for a single capacitated stage, namely equivalent to (8). To complete the model, we need to define a demand bound $D_k(t)$ for each stage: $D_k(s) = max\left\{d_k(t, t+s)\right\}, \forall t, s \geq 0$, where $D_1(s) = D(s)$.

To facilitate the explanation for how we determine the base stock from the capacity constraint and the demand bound, we introduce operator notation (e.g., Griffel, 1985). We use the symbol ψ_k to denote the continuous and node-specific version of (9) as follows:

$$(\psi_k D_k)(\tau) = \max_{n \geq 0}\left\{D_k(\tau + n) - c_k n\right\}. \tag{17}$$

As before, we set the base stock level to this quantity to ensure guaranteed service:

$$B_k(\tau) = (\psi_k D_k)(\tau). \tag{18}$$

If node *k* does not have a capacity constraint, we can set $c_k = \infty$ in (17) and find that $B_k(\tau) = D_k(\tau)$.

There are a couple of immediate implications from the MCBS policy. First, in comparison to the base-stock policy without censoring, the average inventory will be less (for a fixed base-stock level), because orders never exceed capacity and there is no internal queue. We obtain the total average inventory by taking the average of (16):

$$\overline{I}_k = B_k - (SI_k + T_k - S_k)\mu - \overline{BL_k}. \tag{19}$$

Thus for a MCBS policy, we need to calculate the average backlog $\overline{BL_k}$, in order to determine average inventory levels and costs. The term $\overline{BL_k}$ depends on specific properties of the demand distribution, and is generally difficult to estimate; we discuss this topic in greater detail in the Online Supplement. However, from (14) and (15) we observe that the backlog at any stage does not depend on the service times, which are the decision variables in our safety stock optimization. Hence, we do not need to determine $\overline{BL_k}$ to solve this optimization problem, i.e., to find the optimal safety stocks; rather, the sole purpose for determining $\overline{BL_k}$ is for calculating the average inventory level given by (19).

A second implication of the MCBS policy is that the censored order is bounded by the capacity at stage k, i.e., $\left(d_{k+1}(t) \leq c_k\right)$; thus, a looser capacity constraint at stage *k + 1* $\left(c_{k+1} > c_k\right)$ is irrelevant. Indeed, we can ignore any upstream capacity constraint that is greater than a downstream capacity limit.

A third, more significant implication is that each upstream stage will face a different demand bound, one that is censored by node *k*'s capacity. We describe next how to determine the bound on the censored order. The proofs of all propositions are in the Online Supplement.

Proposition 1. *Suppose* $d_{k+1}(t) = \min(d_k(t) + BL_k(t-1), c_k)$ *where* BL_k *is given by* (15), *and initialized with* $BL_k(t) = 0$ *for* $t \leq 0$. *Assume that* D_k *is a valid bound for* d_k, *and that* c_k *is valid with respect to* $D_k(\tau)$. *Then*

$$d_{k+1}(t, t+\tau) \leq D_{k+1}(\tau) = (\Phi_k D_k)(\tau) \quad \forall t, \tau \geq 0 \tag{20}$$

where Φ_k *is defined by*

$$(\Phi_k D)(\tau) = \min(c_k \tau, D(\tau)) \quad \forall \tau \geq 0 \tag{21}$$

This bound is tight, in that for every $\tau \geq 0$ *there is some* $d_k(t, t+\tau) = D_k(\tau)$.

Thus we have an evaluative model for a serial system with capacity constraints and censored orders. We assume that the demand is propagated up the supply chain by (14), with (15) to account for the backlog of orders. We can then use Proposition 1 to compute the demand bound at each node. Given these demand bounds we can then use (18) to find the necessary base stock for each node. Given the base stock, we can calculate the average inventory from (19). We summarize these iterative steps in Table 2, with a comparison to the unconstrained model.

Node k	**No capacity constraint**	**Capacity constraint** c_k
Orders placed	$d_{k+1}(t) = d_k(t)$	$d_{k+1}(t) = \min(d_k(t) + BL_k(t-1), c_k)$
Bound on orders placed	$D_{k+1}(\tau) = D_k(\tau)$	$D_{k+1}(\tau) = (\Phi_k D_k)(\tau)$ $= \min(c_k \tau, D_k(\tau))$
Base stock level	$B_k(\tau_k) = D_k(\tau_k)$	$B_k(\tau_k) = (\psi_k D_k)(\tau_k)$ $= \max\limits_{n \geq 0} \left\{ D_k(\tau_k + n) - c_k n \right\}$

Average inventory	$\overline{I}_k = B_k(\tau_k) - \mu\tau_k$	$\overline{I}_k(t) = B_k(\tau_k) - \mu\tau_k - \overline{BL}_k$

Table 2: Summary of node properties in serial system supply chain with capacity constraints and MCBS policy; we use τ_k to denote the net replenishment time at node *k*.

In the following proposition we establish that the demand bounds and capacity constraints are valid as defined earlier; these properties are necessary to derive the necessary base stock levels. We also show that the resulting base stock levels $B_k(\tau_k)$ are concave functions.

Proposition 2. *Suppose that end demand* $d(t) = d_1(t)$ *is bounded by* $D(\tau)$ *and that* $D(\tau)$ *is valid. Assume further that some subset of nodes has capacity constraints, and that these are all valid with respect to* $D(\tau)$, *and that these* c_k *are decreasing with increasing k. Finally, suppose that each node k places orders* $d_{k+1}(t)$ *according to* (14). *Then*

a) *All orders* d_k *are bounded by* $D_k(\tau)$, *as specified by* (20)-(21)

b) *All* $D_k(\tau)$ *are valid*

c) c_{k+s} *for nodes with capacity constraints are valid with respect to* $D_k(\tau)$, *for all* $s \geq 0$

d) *The base stock levels* $B_k(\tau)$ *as specified by* (18) *ensure that* $I_k(t) \geq 0$

e) *All* $B_k(\tau)$ *are concave in* τ .

Thus we have shown that in a serial-system supply chain with capacity constraints and MCBS policy, we can compute the demand bounds and base-stock levels by recursively applying a sequence of functional operators, as summarized in Table 2.

We can now embed this model in an optimization to find the best choices for the service times. This optimization model is the natural extension to the unconstrained model, as given in Simpson (1958) and Graves and Willems (2000), in which $\tau_k = SI_k + T_k - S_k$ and the service times SI_k, S_k are the decision variables. In the Online Supplement we provide the formulation for the capacitated system when each stage operates with a MCBS policy. We can use the dynamic programming algorithm in Graves and Willems (2000) to find the optimal service times for a serial-system supply chain with both capacity constraints and a MCBS policy, after only small modifications. As shown by Simpson for the unconstrained serial system, we again have that an *all-or-nothing* result holds here, i.e., for each stage either $S_k = 0$ or $B_k = 0$.

We can extend the results in this section to arborescent (assembly tree) supply chain topologies in which each node supplies a single downstream node. The iterative steps laid out in Table 2 apply directly, but with modest modifications: the order placed by node *k* (given by(14)) is now placed concurrently on each of the suppliers to node *k* and the inbound service time for node *k, SI$_k$*, is the maximum of the service times for its supply nodes. If any of the supply nodes have a strictly smaller service time (i.e., $S_j < SI_k$ for node *j* being a direct supplier to node *k*), then by convention node k delays orders from stage j by $SI_k - S_j$ periods to avoid unnecessary inventory (Graves and Willems, 2000). Again we can use the existing algorithm from Graves and Willems to find the optimal service times and safety stocks.

The extension to supply chains with several end demand nodes (as in a distribution network or spanning tree, for example) is not as immediate. The primary challenge is to determine how to combine multiple demand bounds, each of which may be censored. For unconstrained supply chains, Graves and Willems (2000) propose how to combine bounds when the demand streams are assumed to be independent and each bound is set as in(11). They also propose bounds for larger or smaller measures of risk pooling. If one or more bounds were generated by a MCBS policy, then it is not clear how best to merge these bounds from multiple streams. Of course, one can always obtain a valid and conservative bound by simply adding the bounds of downstream stages; we leave for future research the question of how to improve upon this demand bound for supply chains operating with a MCBS policy.

4. Numerical experiments

To examine the impact of capacity constraints, we perform a set of numerical experiments. We consider a serial system with N = 5 stages or nodes, and with a demand bound given by $D(\tau) = \mu\tau + z\sigma\sqrt{\tau}$, with the parameters $(\mu, z, \sigma) = (40, 2, 20)$ and with customer service time of zero, i.e., $S_1 = 0$. We have three alternatives each for the holding costs and for the lead-times, as shown in Table 3.

	Holding costs $(h_5, h_4, h_3, h_2, h_1)$	Lead-times $(T_5, T_4, T_3, T_2, T_1)$
Upstream-Heavy	(0.36, 0.64, 0.84, 0.96, 1.00)	(36, 28, 20, 12, 4)
Constant	(0.20, 0.40, 0.60, 0.80, 1.00)	(20, 20, 20, 20, 20)
Downstream-Heavy	(0.04, 0.16, 0.36, 0.64, 1.00)	(4, 12, 20, 28, 36)

Table 3: Alternative structures for supply chain lead-time and cost accumulation

In all cases the total lead time is 100 and the holding cost at the customer-facing stage 1 is 1.00. The term “upstream-heavy” means that the upstream stages have the longest lead-times or have the largest echelon holding costs. (The echelon holding cost for stage k is $h_k - h_{k+1}$.) Similarly, downstream-heavy means that the largest lead-times or echelon holding costs are at the downstream stages.

We assume each supply chain has a single capacity constraint. We allow the constraint to be at any of the five stages, and we set the capacity to be one of five values from the set (42, 45, 50, 60, 70), representing headroom between 0.1σ and 1.5σ. Thus, we specify 225 test problems: three choices for holding costs, three choices for lead-times, five locations for the capacity constraint, and five values for the capacity.

For each test problem, we compare the best MCBS policy to an *adapted* unconstrained policy. To determine the adapted unconstrained policy, we solve the unconstrained problem to find its optimal location of the decoupling buffers (i.e., the stages at which safety stock is placed, namely $S_k = 0$). We fix these safety-stock locations. We then determine the MCBS policy that provides guaranteed service with the capacity constraint. In this way we can evaluate the performance of the safety-stock configuration for the unconstrained solution in the presence of a capacity constraint.

When determining the average cost for a MCBS policy, one must calculate the term $\overline{BL}_k$ in (15). This term will depend on specific properties of the demand, properties which have not been needed up to this point. For these numerical experiments we estimated this term assuming that demand is normally distributed and i.i.d. We report the details of this estimation and its accuracy in the Online Supplement. We reiterate that while the value of the term $\overline{BL}_k$ does impact the total cost, it does not affect the determination of the best MCBS policy.

In Table 4 we report the results for the test problems with capacity *c* = 45, which are indicative of the behavior at the other capacity levels. For each of the 45 test problems we report the cost for the unconstrained problem, and then report the costs of the best MCBS policy and the adapted unconstrained policy, as a percentage of the cost for the unconstrained problem. We make two observations.

First, the adapted unconstrained policy performs relatively well as a heuristic for many of the test problems. For 75% of the cases with capacity *c* = 45, the cost for the adapted policy is within 3% of that for the best MCBS policy. However, when the upstream lead-times are long and the constraint is downstream at stage 1 or 2, the adapted unconstrained policy can cost 8 to 20% more.

			Location of Capacity Constraint				
Holding cost profile	**Lead time profile**	**Uncap Cost**	**5**	**4**	**3**	**2**	**1**
Upstream Heavy	**Upstream Heavy**	400	99% 105%	104% 107%	106% 109%	103% 111%	89% 93%
	Constant	400	103% 105%	106% 107%	108% 109%	109% 111%	91% 93%
	Downstream Heavy	400	104% 105%	107% 107%	109% 109%	110% 111%	92% 93%
Constant	**Upstream Heavy**	368	98% 98%	95% 97%	93% 105%	87% 107%	73% 89%
	Constant	394	98% 98%	99% 101%	101% 103%	103% 105%	86% 88%
	Downstream Heavy	400	101% 103%	104% 105%	106% 107%	108% 109%	91% 93%
Downstream Heavy	**Upstream Heavy**	268	100% 100%	97% 97%	91% 91%	81% 97%	65% 73%
	Constant	346	100% 100%	98% 98%	95% 101%	96% 104%	78% 87%
	Downstream Heavy	392	100% 100%	98% 98%	98% 101%	101% 104%	86% 89%

Table 4: Results for test problems with capacity constraint $c = 45$. In each cell, the numbers are the normalized costs for the best MCBS policy (top), and the adapted unconstrained policy (bottom). The normalized costs are given as a percentage of the unconstrained problem, whose cost is in the third column.

The second observation is that the cost impact of the capacity constraint is not great, relative to the unconstrained system with a base-stock policy. For the test problems with capacity c = 45, the incremental cost relative to the cost for the unconstrained supply chain is at most 10%, and is less than 5% for 80% of the test problems. More surprising, though, is the fact that the costs for the constrained system are often less than the costs for identical but unconstrained system. That is, the introduction of a constraint results in lower costs. Indeed, for the 225 test problems, we find that the cost for the constrained system with a MCBS policy is on average 3.6% lower than the cost for the unconstrained system operating with a constant base-stock policy. We note that any feasible policy for a constrained system (i.e., it provides guaranteed service under the assumption of bounded demand) will also be feasible for the analogous unconstrained system, and will operate with the same inventory levels.

To get a better understanding of these observations, we will examine in detail a set of solutions. In Table 5 we provide the solutions for the test problems with a constant holding cost profile and upstream-heavy lead-times and with capacity $c = 45$. In particular we report the safety stock levels and the costs for the best base-stock policy for the unconstrained problem, and then the same for the best MCBS policy, for each constraint location.

	Stage	5	4	3	2	1	
	Holding cost	0.2	0.4	0.6	0.8	1.0	
	Lead-time	36	28	20	12	4	
		Safety stock – units and cost					Total cost
Location of Capacity Constraint	No Constraint	240	0	0	0	320	368
	Stage 5	210	0	0	0	320	362
	Stage 4	180	182	0	0	240	349
	Stage 3	180	140	149	0	160	342
	Stage 2	180	140	100	110	80	320
	Stage 1	180	140	100	60	70	270

Table 5: Average safety stock inventory and cost for test problems with constant holding cost profile and upstream heavy lead-times.

The solution for the unconstrained system holds a decoupling safety stock at stage 1, and then another at the most upstream stage, stage 5. Each safety stock is sized to cover the demand bound over the relevant net replenishment time. For stage 1, the net replenishment time is $\tau = 64$ and the safety stock is $z\sigma\sqrt{\tau} = 2 \times 20 \times \sqrt{64} = 320.$ Similarly for stage 5, we have $\tau = 36, z\sigma\sqrt{\tau} = 240.$ The intermediate stages hold no safety stock; placing an inventory, say, at stage 3 will reduce the safety stock needed at stage 1, but this reduction is not enough to pay for the additional inventory held at stage 3. For these cost and lead-time profiles, though, the inventory at stage 5 can be justified as its cost is less than the benefits from reducing the net replenishment time seen by stage 1.

Suppose we now have a capacity constraint at stage 5. The safety stock locations remain the same as for the unconstrained system. Stage 1 must hold a safety stock to assure its service time commitment, and the costs from placing a safety stock at an upstream stage can only be justified for stage 5. The safety stock at stage 1 is the same as for the unconstrained system as it sees the same demand bound and has the same net replenishment time. However, the safety stock at stage 5 is now less; in this case, the net replenishment time is still $\tau = 36$ and the safety stock is given by $z\sigma\sqrt{\tau} - \overline{BL}_5 = 240 - 30 = 210.$ (For the given demand bound and c = 45, we find the expected backlog to be 30; see the Online Supplement for details of this estimation.) The reason as to why stage 5 requires less safety stock in the constrained system is exactly as found for the single-stage system discussed in section 2: the base-stock policy for the unconstrained system ignores the demand history and will over-order when it experiences high demand; adding a capacity constraint can help to mitigate this, as the MCBS policy will never order more than the capacity and hence will dampen out the high demand events. As a result of needing less safety stock at stage 5, we see that the constrained system with censored ordering has a lower cost than the unconstrained system with a base-stock policy (362 versus 368).

Now consider the system with the capacity constraint placed at stage 1. The structure of the solution completely changes as now all upstream stages hold a safety stock. This is because placing a constraint at stage 1 affects the demand bound seen by each of the upstream stages. The demand bound for each upstream stage is now linear up to $\tau = 64$ with a slope equal to the capacity, $c = 45$. When the demand bound is linear, we find that the safety stock for an unconstrained stage is also linear in the net replenishment time, and given by $(c - \mu)\tau$, where c is the slope of the demand bound. As a consequence there is no value from risk pooling over the lead time, and thus each upstream stage holds a safety stock, proportional to its lead time. As stage 2 now has a safety stock, the net replenishment time for stage 1 is just $\tau = 4$ and its safety stock is given by: $c\tau + \frac{\sigma^2}{c - \mu} - \overline{BL}_1 = 70$, as the backlog for the constrained stage is 30 (as noted above).

We again see that the constrained system with censored ordering at stage 1 has a lower cost than the unconstrained system with a base stock policy. Indeed, the cost of the constrained system is now 73% of that for the unconstrained system (270 versus 368). There are two drivers for this. First, we observe that the capacity constraint at stage 1 dramatically dampens the demand bound seen by all upstream stages. With this demand bound, it is now economical to place a relatively small and inexpensive safety stock at each of the upstream stages. A decoupling safety stock at stage 2 then reduces the net replenishment time for stage 1, and thus the safety stock needed there. Second, for the given safety stock configuration, the constraint at stage 1 does not increase the costs at stage 1, for the reasons discussed in section 2: for $\tau = 4$, an unconstrained stage 1 would require a safety stock of 80; adding the constraint reduces the safety stock at stage 1 to 70 units.

The test problems for which the capacity is placed at an intermediate stage have similar behavior. The best MCBS policy requires an inventory at the capacitated stage, and then places safety stocks at each upstream stage for the reasons noted above. The cost of each solution is less than that for the unconstrained system with a base-stock policy. Finally we see that the cost goes down as the constraint is moved downstream in this system. This seems to be due to more stages being subject to the dampened demand bound and to the fact that the longer lead times are upstream.

The observations from this detailed example are common across this set of test problems; we summarize below the key points:

- The best MCBS policy places a decoupling safety stock at the stage with the constraint. This inventory protects the downstream stages, particularly stage 1, from the effect of the capacity constraint; if there were no decoupling inventory at the constraint, then the safety stock at stage 1 would need to increase substantially to protect against any replenishment delays at the constrained stage.
- The best MCBS policy then places a safety stock at each stage that is upstream from the constraint. This censored-ordering policy dampens the demand bound seen by the upstream stages, making it economical for these stages to hold a safety stock.
- Adding a constraint can actually result in a lower inventory cost than for the corresponding unconstrained solution with a base-stock policy. This is (1) due to the non-optimality of the base-stock policy, in that it does not account for the autocorrelation in the demand, and (2) due to the censoring of the demand signal, which dampens the demand bound seen by the upstream stages.
- From a safety stock standpoint it is better to have the capacity constraint downstream. Indeed, in all the test problems, for a given capacity value, the safety stock cost was always minimized when the constraint is at the customer facing-node (stage 1). Furthermore, the capacity constraint needs to be only slightly higher than average demand; in fact, for all cases with the constraint at the customer-facing stage, the smallest capacity $\mu + 0.1\sigma = 42$ resulted in the lowest cost. However, we also know from (13)

that as the capacity approaches average demand, the necessary base-stock level goes to infinity.

5. Conclusions and discussion

We have analyzed the inclusion of capacity constraints in the context of the GS model for safety stocks in multi-echelon supply chains. We have shown how to extend the single-stage base-stock model to include a capacity constraint. We have used this result to model multi-stage supply chains with capacity constraints and with a modified constant base-stock (MCBS) policy. We have characterized the base-stock level necessary for guaranteed service. We find that the solution methods developed for unconstrained supply chains can be used, with a minor modification, to determine the safety-stock placement for capacitated systems.

From our numerical experiments we find that the best MCBS policy places a decoupling safety stock at the stage with the constraint, and then places a safety stock at each stage that is upstream from the constraint. This configuration is much different than that for the corresponding unconstrained system, in which the best base-stock policy places safety stocks much more sparsely across a system.

We also find from our numerical tests that adding a constraint can actually result in a lower inventory cost than for the corresponding unconstrained solution with a base-stock policy. There are two explanations for this.

First, the base-stock policy is not optimal in that it does not account for the autocorrelation in the demand, due to the assumption of a demand bound; consequently the base-stock policy will order too much when it experiences high demand. In the presence of a capacity constraint, the censored policy can mitigate this as it will not order more than the capacity. As a consequence, adding a constraint to a stage, under MCBS control, is often not too costly and at times, can reduce the inventory cost at the stage. This finding is an artifact of the assumption of bounded demand.

The second reason is that the MCBS policy affects the demand bound for the upstream stages, reducing its magnitude. In effect, the MCBS policy smooths the demand signal that gets passed upstream. As a consequence it can become more economical to hold inventory at upstream stages, which can allow the system to reduce its more-expensive downstream inventory. In our experiments we find that having the capacity constraint at the customer-facing node always resulted in the least inventory costs for a constrained system. This suggests that it may be desirable for the first stage in a supply chain to act as a damper, whereby it absorbs the variability from a demand signal, rather than pass along this variability to the rest of the supply chain.

For future research, one topic might be to try to determine the relative magnitude of these effects. A capacity constraint results in censored ordering which can reduce over-ordering in the presence of an autocorrected demand, and can result in a smoother demand signal being passed upstream. Which is more valuable and for what circumstances?

We find that the demand smoothing from a capacity constraint in the GS model significantly affects the safety stock placement upstream of the constraint. Furthermore, the MCBS policy relies only on local information. Future research might examine how prevalent are the benefits from smoothing, and whether further benefits can accrue from using global information.

In this paper we were able to characterize the optimal policy for the GS model with bounded demand for a single stage. We found numerically that the MCBS policy is near optimal when capacity is tight, but that its performance deteriorates as the constraint is loosened. Possible future research could examine this relationship more deeply. Future research might also determine the optimal policy (or a bound) for a multi-stage system so as to understand better how effective is the MCBS policy as a heuristic.

Another topic is to see if one can improve the characterization of the backlog associated with the MCBS policy. As we have noted, the costs (but not the optimal solution) of systems with censored ordering depend on the term $\overline{BL}_k$, which in turn depends on the demand distribution, and which moreover appears difficult to estimate.

It would be valuable to extend the solution methods for unconstrained supply chains modeled as general networks (Humair and Willems, 2006, 2011) to account for capacity constraints. Another possible extension is to allow for more complex capacity constraints, in which a single capacity is shared over a subset of stages rather than a single stage.

A final line of research might relax the assumption that the customer service time is exogenously set. One might include within the safety-stock optimization the decision on the customer service time, where there is a penalty cost associated with a longer customer service time. Alternatively a firm might segment its customers into different classes, each with a customer service time and a price; the research question might be how to set the segments in conjunction with the supply-chain safety stock optimization.

References

Bertsimas D., Thiele A. 2004. A robust optimization approach to supply chain management. Integer Programming and Combinatorial Optimization, Springer Berlin / Heidelberg, 86-100.

Bertsimas D., Thiele A. 2006. A robust optimization approach to inventory theory. *Operations Research*. **54** 150-168.

Billington C., Callioni G., Crane B., Ruark J. D., Rapp J. U., White T. and Willems S. P. 2004. Accelerating the profitability of Hewlett-Packard's supply chains. *Interfaces.* **34** 59-72.

Clark A.J., Scarf H. 1960. Optimal policies for a multi-echelon inventory problem, *Management Science* **6**. 475-490

Diks, E. B., de Kok A. G., Lagodimos A. G. 1996. Multi-echelon systems: A service measure perspective. *European Journal of Operations Research.* **95** 241–263.

Eruguz, A. S. 2014. *Contributions to the multi-echelon inventory optimisation problem using the guaranteed-service model approach* (Doctoral dissertation, Ecole Centrale Paris).

Eruguz, A. S., Sahin, E., Zied Jemai. Z., and Yves Dallery, Y. 2016.. "A comprehensive survey of guaranteed-service models for multi-echelon inventory optimization." *International Journal of Production Economics* 172, 110-125.

Farasyn, I., Humair, S., Kahn, J. I., Neale, J. J., Rosen, O., Ruark, J., Tarlton, W. Van de Velde, W., Wegryn, G. and Willems, S. P., Inventory Optimization at Procter & Gamble: Achieving Real Benefits Through User Adoption of Inventory Tools, *Interfaces*, January/February 2011 41:66-78.

Gallego G., Scheller-Wolf A. 2000. Capacitated inventory problems with fixed order costs: Some optimal policy structure. *European Journal of Operations Research*. **126** 603-613.

Gallego G., Toktay B.L. 2004. All-or-Nothing ordering under a capacity constraint. *Operations Research.* **52** 1001-1002.

Glasserman P., Tayur S. 1994. The stability of capacitated, multi-echelon production-inventory system under base-stock policy *Operations Research*. **42** 913-925.

Glasserman P., Tayur S.,1995. Sensitivity analysis for base-stock levels in multiechelon production-inventory systems. *Management Science*. **41** 263-281.

Glasserman P., Tayur S. 1996. A simple approximation for a multistage capacitated production-inventory system. *Naval Research Logistics*. **43** 41-58.

Graves S.C., Willems S. P. 2000. Optimizing strategic safety stock placement in supply chains. *Manufacturing & Service Operations Management.* **2** 68-83.

Graves S.C., Willems S. P. 2003. Supply chain design: safety stock placement and supply chain configuration. A. G. de Kok and S. C. Graves, eds, *Handbooks in Operations Research and*

Management Science Vol. 11, Supply Chain Management: Design, Coordination and Operation. North-Holland Publishing Company, Amsterdam, The Netherlands.

Graves, S. C., Willems, S. P. 2008. Strategic inventory placement in supply chains: Nonstationary demand. *Manufacturing & Service Operations Management*, *10*(2), 278-287

Griffel D.H. 1985. Applied functional analysis. John Wiley & Sons, New York.

Gupta D., Selvaraju N. 2006. Performance evaluation and stock allocation in capacitated serial systems. *Manufacturing & Service Operations Management*. **8** 169–191.

Hsieh, Min Fang 2011. "Applying a MEIO approach to manage Intel's VMI hub supply chain," S.M. thesis, MIT, Cambridge MA, June 2011.

Huh, W. T., Janakiraman, G., and Nagarajan, M. 2010. Technical Note-Capacitated serial inventory systems: Sample path and stability properties under base-stock policies. *Operations Research*, *58*(4-part-1), 1017-1022.

Huh, W. T., Janakiraman, G., and Nagarajan, M. 2014. Capacitated multi-echelon inventory systems: Policies and bounds. Working paper, University of British Columbia, Vancouver, British Columbia, Canada.

Humair, S., Willems, S. P. 2006. Optimizing strategic safety stock placement in supply chains with clusters of commonality, *Operations Research*, Vol. 54, No. 4, pp. 725-742.

Humair S., Willems S. P. 2011. Optimizing strategic safety stock placement in general acyclic networks. , *Operations Research*, Vol. 59, No. 3, pp. 781-787.

Inderfurth, K. 1991. Safety stock optimization in multi-stage inventory systems. *International Journal of Production Economics*. **24** 103-113.

Janakiraman, G., Muckstadt, J. 2009. A decomposition approach to a class of capacitated serial systems. *Operations Research,* 57, pp. 1384 - 1393.

Kimball, G. E. 1988. General principles of inventory control. *Journal of Manufacturing and Operations Management*. **1** 119-130.

Klosterhalfen, S. T., Minner, S., and Willems, S. P. 2014. Strategic safety stock placement in supply networks with static dual supply. *Manufacturing & Service Operations Management*, *16*(2), 204-219.

Lesnaia, E. 2004. Optimizing Safety Stock Placement in General Network Supply Chains. PhD Thesis, Massachusetts Institute of Technology

Lesnaia, E., Vasilescu, I., Graves, S. C. 2005. The complexity of safety stock placement in general-network supply chains. *Proceedings of the 2005 SMA Conference*. Singapore.

Magnanti, T. L., Z.-J. M. Shen, J. Shu, D. Simchi-Levi, C.-P. Teo. 2006. Inventory placement in acyclic supply chain networks. Oper. Res. Lett. 34(2) 228–238.

Masse, B. R. 2011. Inventory Optimization in High Volume Aerospace Supply Chains, S.M. thesis, MIT, Cambridge MA, June 2011.

Parker R. P., Kapuscinski R., 2004. Optimal policies for a capacitated two-echelon inventory system. *Operations Research*. **52** 739-755.

Polak, B. M., 2014. Multi-Echelon Inventory Strategies for a Retail Replenishment Business Model, S.M. thesis, MIT, Cambridge MA, June 2014.

Schoenmeyr, T., Graves, S. C. 2009. Strategic safety stocks in supply chains with evolving forecasts. *Manufacturing & Service Operations Management*, *11*(4), 657-673.

Simpson, K. F. 1958. In-process inventories. *Operations Research*. **6** 863–873.

Sitompul, C., Aghezzaf, E., Dullaert, W., Van Landeghem, H. 2008. Safety stock placement problem in capacitated supply chains. *International Journal of Production Research* 46, no. 17, 4709-4727.

Wal, J. van der, Speck, C.J. 1991. *The capacitated multi-echelon inventory system with serial structure. 1. The 'push-ahead'-effect.* Memorandum COSOR No. 91-39, Eindhoven: Technische Universiteit Eindhoven, 10 pp.

Willems S.P. 2008. Real-World multiechelon supply chains used for inventory optimization. *Manufacturing & Service Operations Management . 10* (1), 19-23.

Online Supplement

Appendix 1: Proof for optimal policy for single-stage system

In this appendix we show that the order quantity given by (3) is feasible. That is, we intend to show that $O(t) \le c$ where $O(t)$ is given by:

$$O(t) = \max_{k \ge 0} \left\{ \min_{j=0,\ldots t} \{ D(j+\tau+k) - d(t-j,t) \} - kc \right\} - I(t) - \sum_{i=1}^{\tau-1} O(t-\tau+i). \qquad (22)$$

For ease of presentation we define:

$$g(t,k) = \min_{j=0,\ldots t} \{ D(j+\tau+k) - d(t-j,t) \} - kc\,. \qquad (23)$$

Then we can express (22) as:

$$O(t) = \max_{k \ge 0} \{ g(t,k) \} - I(t) - \sum_{i=1}^{\tau-1} O(t-\tau+i). \qquad (24)$$

To develop our result, we first need to determine a *boundary condition* on the inventory. In order to have guaranteed service, we need that $I(t) \ge 0, \forall t \ge 0$. We assume that the system starts at time 0, and that $O(t) = 0, \forall t \le 0$. Then we can express the inventory as:

$$I(t) = \begin{cases} I(0) - d(0,t) \ \mathit{for}\ 0 < t \le \tau \\ I(0) - d(0,t) + \sum_{i=1}^{t-\tau} O(i) \ \mathit{for}\ t > \tau \end{cases}$$

To assure that $I(t) \ge 0, \forall t \ge 0,$ we see that we need $I(0) \ge d(0,\tau+k) - \sum_{i=1}^{k} O(i), \forall k.$ We can use the demand bound to re-write this condition as $I(0) \ge D(\tau+k) - \sum_{i=1}^{k} O(i), \forall k.$ To get the minimal value, we set the orders to their maximum value, i.e. $O(i) = c, \forall i.$ Thus, we set the initial inventory by:

$$I(0) = \max_{n \in Z} \{ D(\tau+n) - nc \}. \qquad (25)$$

We now assume that we set the initial inventory as given in (25). As a side note, we observe that the initial inventory for the optimal policy is the same as the constant base stock, given by (9).

We can now use (22) and (25) to determine the orders placed in period 1:

$$O(1) \geq \min\left\{D(\tau+k), D(\tau+k+1)-d(1)\right\}-kc-I(1), \forall k \geq 0. \tag{26}$$

Consider now the right-hand-side of the above equation for a given value of *k*:

$$\begin{aligned}
&\min\left\{D(\tau+k), D(\tau+k+1)-d(1)\right\}-kc-I(1) \\
&\leq D(\tau+k+1)-d(1)-kc-I(1) \\
&= D(\tau+k+1)-d(1)-kc-\left(I(0)-d(1)\right) \\
&= D(\tau+k+1)-kc-\max_{n\in Z}\left\{D(\tau+n)-nc\right\} \\
&= \left\{D(\tau+k+1)-(k+1)c\right\}-\max_{n\in Z}\left\{D(\tau+n)-nc\right\}+c \\
&\leq c.
\end{aligned}$$

As explanation, we substitute $I(1)=I(0)-d(1)$ in the third line; and we observe that

$\left\{D(\tau+k+1)-(k+1)c\right\} \leq \max_{n\in Z}\left\{D(\tau+n)-nc\right\}$ in the fifth line. Thus we have shown that the RHS of (26) is at most *c* for any value of *k*, and hence $O(1) \leq c$.

For an induction argument we now assume that $O(s) \leq c,\ for\ s=1,\ldots t-1.$ We now need to show that $O(t) \leq c$ to complete the proof.

The argument will proceed as for the case $t=1.$ We intend to show that

$$g(t,k)-I(t)-\sum_{i=1}^{\tau-1} O(t-\tau+i) \leq c, \forall k. \tag{27}$$

If this is true, then it is clear from (24) that $O(t) \leq c.$

We re-express (27) by using the balance equation for the inventory:

$$\begin{aligned}
&g(t,k)-I(t)-\sum_{i=1}^{\tau-1} O(t-\tau+i) \\
&= g(t,k)-I(t-1)+d(t)-O(t-\tau)-\sum_{i=1}^{\tau-1} O(t-\tau+i) \\
&= \min_{j=0,\ldots t}\left\{D(j+\tau+k)-d(t-j,t-1)\right\}-kc-I(t-1)-\sum_{i=0}^{\tau-1} O(t-\tau+i).
\end{aligned} \tag{28}$$

We note that $d\left(t,t-1\right)=-d\left(t\right)$. We can now substitute for $O\left(t-1\right)$ using(22):

$$\begin{aligned}
&\min_{j=0,\ldots t}\left\{D\left(j+\tau+k\right)-d\left(t-j,t-1\right)\right\}-kc-I\left(t-1\right)-\sum_{i=0}^{\tau-1}O\left(t-\tau+i\right)\\
&=\min_{j=0,\ldots t}\left\{D\left(j+\tau+k\right)-d\left(t-j,t-1\right)\right\}-kc-I\left(t-1\right)-\sum_{i=0}^{\tau-2}O\left(t-\tau+i\right)\\
&-\max_{n\geq0}\left\{g\left(t-1,n\right)\right\}+I\left(t-1\right)+\sum_{i=1}^{\tau-1}O\left(t-1-\tau+i\right)\\
&=\min_{j=0,\ldots t}\left\{D\left(j+\tau+k\right)-d\left(t-j,t-1\right)\right\}-kc-\max_{n\geq0}\left\{g\left(t-1,n\right)\right\}.
\end{aligned}\tag{29}$$

To evaluate this expression we note that:

$$\begin{aligned}
&\min_{j=0,\ldots t}\left\{D\left(j+\tau+k\right)-d\left(t-j,t-1\right)\right\}-kc\\
&\leq\min_{j=1,\ldots t}\left\{D\left(j+\tau+k\right)-d\left(t-j,t-1\right)\right\}-kc\\
&=c+\min_{j=0,\ldots t-1}\left\{D\left(j+\tau+k+1\right)-d\left(t-1-j,t-1\right)\right\}-\left(k+1\right)c\\
&=c+g\left(t-1,k+1\right).
\end{aligned}\tag{30}$$

Thus, by substituting (30) into (29), we now have an inequality on (28):

$$\begin{aligned}
&g\left(t,k\right)-I\left(t\right)-\sum_{i=1}^{\tau-1}O\left(t-\tau+i\right)=\min_{j=0,\ldots t}\left\{D\left(j+\tau+k\right)-d\left(t-j,t-1\right)\right\}-kc-\max_{n\geq0}\left\{g\left(t-1,n\right)\right\}\\
&\leq c+g\left(t-1,k+1\right)-\max_{n\geq0}\left\{g\left(t-1,n\right)\right\}\leq c.
\end{aligned}$$

From (24) this is sufficient to show that $O\left(t\right)\leq c,\forall t,$ which completes the induction argument.

Appendix 2: Proof of Lemma 1

Lemma 1. *Suppose* $D(\tau)$ *is valid and differentiable, and c valid with respect to* $D(\tau)$**.** *Then*

$B\left(\tau\right)=\max_{n\geq0}\left\{D(\tau+n)-cn\right\}$ *is concave on* $\tau\in[\theta-\dfrac{D(\theta)}{c},\infty)$.

Proof of Lemma 1.

We define

$$\theta=\arg\max_{\tau\geq0}\left\{D(\tau)-c\tau\right\}\tag{31}$$

Because $D(\tau)$ is concave (by Definition 1) for nonnegative τ and crosses $c\tau$ at a single point $\tilde{\tau}$ (by Definition 2), there must be some maximizing positive $0 \le \theta < \tilde{\tau}$ (if there are multiple maximizing values, anyone can be picked as θ for this proof). Now

$$\begin{aligned} B(\tau) &= \max_{n\ge 0}\left\{D(\tau+n)-cn\right\} \\ &= c\tau + \max_{n\ge 0}\left\{D(\tau+n)-c(\tau+n)\right\} \\ &= c\tau + \max_{x\ge \tau}\left\{D(x)-cx\right\} \end{aligned} \tag{32}$$

Now if $\tau \le 0$, then either $x=\tau$ or $x=\theta,$ since $D(x)=0,$ for $x \le 0$; that is:

$$B(\tau)=\max\left(0, D(\theta)-c\theta+c\tau\right), if \ \ \tau \le 0$$

If $\tau > 0$, then either $x=\max(\tau,\theta)$ because D is concave, and crosses cx at a unique point. Combining these two results we get equation (10), namely:

$$B(\tau)=\begin{cases} 0 & \text{for } \tau < \theta - \dfrac{D(\theta)}{c} \\ c\times(\tau-\theta)+D(\theta) & \text{for } \theta - \dfrac{D(\theta)}{c} \le \tau < \theta \\ D(\tau) & \text{for } \tau \ge \theta \end{cases} \tag{33}$$

We now prove concavity on the range $\tau \in [\theta - \dfrac{D(\theta)}{c}, \infty)$ by inspection: $B(\tau)$ is continuous; its second derivative is zero on the interval $\theta - \dfrac{D(\theta)}{c} < \tau < \theta$ and its second derivative is non-positive for $\tau > \theta$ due to the assumption that D is a valid demand bound. At $\tau = \theta$, both the left and right derivatives are equal to *c;* the left second derivative is zero and the right second derivative is non-positive, which assures concavity at this point.

Appendix 3: Algorithm for generating random bounded demand

The demand bound for the single-stage numerical experiments is as follows:

$$D(t)=\mu t + z\sigma\sqrt{t} \ \ for \ \ t \le b$$
$$D(t)=\mu b + z\sigma\sqrt{b} + s\times(t-b) \ \ for \ \ t > b$$

For the experiment, we need to generate a demand series that does not violate the bound. We do this by generating normally-distributed demand and then fixing the demand one period at a time, to assure that the demand series is always within the demand bound. In any period any excess demand that needs to be "trimmed" to keep the series within the demand bound is carried over to the next period. In this way we assure that the mean demand remains at μ.

The procedure is as follows:

- $\Delta(0)=0; \varepsilon(0)=D(b).$.
- For t = 1 to N
    - Generate random demand $rd(t)$ from $ND(\mu,\sigma)$, truncated to be non-negative.
    - $d(t)=rd(t)+\Delta(t-1)$ where $\Delta(t)$ is leftover demand
    - For each demand $d(t)$, we filter the demand with the following procedure:
        - For k =1 to b
        - $d(t):=Min(D(k),d(t-k,t))-d(t-k,t-1)$
        - Next k
    - These steps assure that the demand $d(t)$ is within bound up to the break point at b.
    - To assure that it satisfies the bound beyond the break point , we modify $d(t)$ as follows:

$$d(t):=Min(s+\varepsilon(t-1),d(t))$$
$$where$$
$$\varepsilon(t)=Min(D(b)-d(t-b,t),s+\varepsilon(t-1)-d(t)).$$

  Explanation: after time t, $\varepsilon(t)$ is a measure of slack; that is, $\varepsilon(t)=\underset{j\geq b}{Min}(D(j)-d(t-j,t))$, the amount by which the current demand history falls below the bound beyond the break point. The update equation for $d(t)$ assures that the demand history falls within the bound beyond the break point; the update for the slack variable follows the definition.

    - $\Delta(t)=\Delta(t-1)+rd(t)-d(t)$. This represents the amount of demand that is carried over to the next period (i.e, the leftover demand), because it could not fit under the bound.

- Next t

Appendix 4: Single-stage example of non-optimality of base-stock policies

We present here a simple example first to show the non-optimality of the base-stock policy for the uncapacitated problem, and then to show how adding a capacity constraint might improve the performance of a base-stock policy.

Suppose that $\tau = 1$. Suppose demand can take on two values, 5 and 15. If demand in one period is 15, then it must be 5 in the next period. If demand in one period is 5, then it is 5 or 15 with equal probability in the next period. The expected demand is 8.33 per period, and the demand bound is not concave and is given by $D(1) = 15, D(2) = 20, D(3) = 35, D(4) = 40, \ldots$

First, we assume there is no capacity constraint. The best *constant base-stock policy* is $B = D(1) = 15$. The average inventory level is 6.67.

The *optimal ordering policy* is as follows:

$$O(t) + I(t) = \begin{cases} 5 \; if \;\; d(t) = 15 \\ 15 \; if \;\; d(t) = 5 \end{cases} .$$

That is, the optimal policy sets an order-up-to level of 5, if it observes a demand of 15; otherwise the order-up-to level is 15, same as for the base-stock policy. The average inventory is reduced to 4.44. Clearly the optimal policy exploits the fact that a demand of 15 will be followed be a lower demand, namely 5.

Now suppose we have a capacity limit, $c = 10$. The optimal policy is now:

$$O(t) + I(t) = \begin{cases} 10 \; if \;\; d(t) = 15 \\ 15 \; if \;\; d(t) = 5 \end{cases} .$$

The optimal policy now needs an order-up-to level of 10, if it observes a demand of 15; even though demand in period *t+1* is only 5, the optimal policy needs to order more in period *t* due to the limit on what it can order in period *t+1*. In particular if $d(t) = 15,$ the optimal policy will order at capacity for two periods so as to recover the inventory position to the desired level of 15, i.e., $O(t) = O(t+1) = c = 10.$ The average inventory for this policy is 6.11. The inventory does increase from 4.44 due to the capacity constraint; as noted above, whenever there is a high demand, the policy must re-build its inventory over the next two periods, incurring additional inventory holding.

The best MCBS policy will set the base-stock level again as $B = D(1) = 15$. But one can easily see that this is equivalent to the optimal policy given above. Thus, its average inventory is 6.11, which is less than that for the uncapacitated base-stock policy (6.67). The uncapacitated policy does not account for the information in the demand history, namely the fact that a high demand is followed by a low demand. The capacitated policy does better because it is limited in its response to a high demand. It still needs a base stock of 15 in this example. But when there is a high demand it can only replenish its capacity, namely 10 units. Fortunately, this is sufficient due to the negative correlation in demand, and the policy will re-build to its base stock after two periods, when it can again experience and serve a high demand.

As seen in Table 1, this example demonstrates that the average inventory for the MCBS policy is non-monotonic in the capacity level. Thus, in some contexts, one might improve performance (i.e., reduce average inventory) by imposing a (virtual) capacity constraint. One could find the best choice of capacity by a simple search using simulation to evaluate the performance for different values of c.

Appendix 5: The average backlog

Here we consider estimating $\overline{BL}_k$ .We recall that this quantity is necessary to compute the average inventory costs for a node with capacity constraints and censorship. However, $\overline{BL}_k$ does not depend on the decision variables (the service times) and it is not needed in order to obtain or implement the optimal solution.

We note that the backlog described in (15) behaves like a Lindley process for modeling the wait times in a queue. Even though BL_k operates in discrete time, we will approximate it with a continuous-time queuing model. In particular, we propose to model the backlog as the wait time for an M/D/1 queue with arrival rate λ , and deterministic processing time *s*. We set the parameters λ and *s* to correspond to average and standard deviation of demand per period, normalized with respect to capacity; these moments are given as $\frac{\mu}{c}$ and $\frac{\sigma}{\sqrt{c}}$. Specifically we make a second-order, continuous-time approximation and equate the normalized demand moments to those for the offered load for the queuing system, as follows:

$$\begin{aligned} \frac{\mu}{c} &= \lambda s \\ \frac{\sigma}{\sqrt{c}} &= \sqrt{\lambda} s \end{aligned} \tag{34}$$

We solve for the parameters to obtain:

$$\begin{aligned} s &= \frac{\sigma^2}{\mu} \\ \lambda &= \frac{\mu^2}{c\sigma^2} \end{aligned} \tag{35}$$

Having calculated s and λ, we can use the Pollaczek-Khintchine formula (with zero processing time variability) for calculating expected number of jobs and the expected waiting time. This formula is exact for Poisson arrivals in continuous time, but for a discrete time system it is only an approximation. Noting that the utilization is simply $\rho = \frac{\mu}{c}$, we have:

$$\begin{aligned} \overline{BL} &= \overbrace{\left(\rho + \frac{\rho^2}{2(1-\rho)}\right)}^{\text{Expected number of jobs}} \times \overbrace{s}^{\text{Time per job}} \\ &= \left(\frac{\mu}{c} + \frac{\left(\frac{\mu}{c}\right)^2}{2(1-\frac{\mu}{c})}\right)\left(\frac{\sigma^2}{\mu}\right) \\ &= \left(1 + \frac{\mu}{2(c-\mu)}\right)\left(\frac{\sigma^2}{c}\right) \\ &= \left(\frac{2c-\mu}{c-\mu}\right)\left(\frac{\sigma^2}{2c}\right) \end{aligned} \tag{36}$$

Finally, we mention that if one wants to estimate $\overline{BL_k}$ for more complex (not i.i.d.) demand processes or if greater precision is desired, it is easy to estimate $\overline{BL_k}$ numerically. One can simply evaluate (15), $BL_k(t) = \max\{BL_k(t-1) + d_k(t) - c_k, 0\}$, for a real or simulated sequence of demand realizations $d_k(t)$ and calculate the average value $\overline{BL_k}$. In the context of the numerical experiments in §4, we compared the formula (36), with the aforementioned numerical estimate, using a normal distribution.

c	**42**	**45**	**50**	**60**	**70**
Simulated/ Normal	88.5	29.6	10.6	2.5	0.7
PK-formula/ Poisson	104.8	44.4	24.0	13.3	9.5

Table 5: $\overline{BL_k}$ for $\mu = 40, \sigma = 20$, and various censorship values *c*.

The continuous-time formula gives higher values of $\overline{BL_k}$, especially for large capacities. As explanation, the PK formula provides an estimate of the average waiting time for any point in time. But by (15) we measure the backlog at an instant of time, namely at the end of the period just prior to demand arrival in the next period. As such, this epoch is when the backlog is smallest. With a fluid approximation, one can adjust the PK formula to represent the backlog at the end of the period, and substantially reduce the error.

Nevertheless, the term $\overline{BL_k}$ is typically only responsible for a small portion of all costs and its value does not depend upon the safety stock decisions. In the experiments we performed, the total cost difference between the two estimation methods was, on average, only 2.1%.

Appendix 6: Proof of propositions

Proposition 1. *Suppose* $d_{k+1}(t) = \min(d_k(t) + BL_k(t-1), c_k)$ *where* BL_k *is given by* (15), *and initialized with* $BL_k(t) = 0$ *for* $t \le 0$. *Assume that* D_k *is a valid bound for* d_k, *and that* c_k *is valid with respect to* $D_k(\tau)$. *Then*

$$d_{k+1}(t, t+\tau) \le D_{k+1}(\tau) = (\Phi_k D_k)(\tau) \quad \forall t, \tau \ge 0 \tag{37}$$

where Φ_k *is defined by*

$$(\Phi_k D)(\tau) = \min(c_k \tau, D(\tau)) \quad \forall \tau \geq 0 \tag{38}$$

This bound is tight, in that for every $\tau \geq 0$ *there is some* $d_k(t, t+\tau) = D_k(\tau)$**.**

Proof of Proposition 1. We need to show that $d_{k+1}(t,t+\tau) \leq \min\left[c_k \tau, D_k\left(\tau\right)\right], \forall \tau$. Let $\tilde{\tau}$ be the point such that $D_k(\tilde{\tau}) = c_k \tilde{\tau}$, which must exist per Definition 2. Clearly, by the ordering mechanism (14), $d_{k+1}(t)$ can never exceed c_k , and so we must have that

$$d_{k+1}(t,t+\tau) \leq c_k \tau \leq D_k(\tau) \quad \tau \leq \tilde{\tau} \tag{39}$$

In order to investigate $\tau > \tilde{\tau}$ we note that:

$$\begin{aligned} & d_{k+1}(t,t+\tau) = d_k(t,t+\tau) - BL_k(t+\tau) + BL_k(t) \\ & \leq d_k(t,t+\tau) + BL_k(t) \\ & = d_k(t,t+\tau) + \left(d_k(t-\tilde{n},t) - c_k \tilde{n}\right) \\ & = d_k(t-\tilde{n},t+\tau) - c_k \tilde{n} \end{aligned} \tag{40}$$

where

$$\tilde{n} \equiv \min n \geq 0 : BL_k(t-n) = 0 \tag{41}$$

That is, $\tilde{n} \geq 0$ is the number of periods that node *k* has been working at capacity before time *t*. By the assumption that $BL_k(t) = 0$ for t=0, there must always exist such an $\tilde{n}$. We can replace the RHS of (40) by maximizing over *n*; we will still have a valid (although potentially looser) bound:

$$\begin{aligned} & d_{k+1}(t,t+\tau) \\ & \leq d_k\left(t-\tilde{n},t+\tau\right) - c_k \tilde{n} \\ & \leq \max_{n\geq 0}\left\{d_k\left(t-n,t+\tau\right) - c_k n\right\} \end{aligned} \tag{42}$$

Finally, we invoke the bound on d_k :

$$d_{k+1}(t,t+\tau) \leq \max_{n\geq 0}\left\{D_k\left(\tau+n\right) - c_k n\right\} \tag{43}$$

However, for $\tau > \tilde{\tau}$ we have, because D_k is concave and $\tilde{\tau}$ is the equality point, that (43) is maximized for $n = 0$ and hence, for $\tau > \tilde{\tau}$, we have

$$d_{k+1}(t,t+\tau) \leq D_k(\tau) < c_k \tau \quad \tau > \tilde{\tau} \tag{44}$$

Combining (39) and (44) gives us the claimed relation. Finally we note that the bound (37) is tight; for example $d_{k+1}(t,t+\tau)=D_{k+1}(\tau)$ is realized if $BL_k(t-1)=0$ and $d_k(t,t+\tau)=D_k(\tau)$

□

Proposition 2. *Suppose that end demand* $d(t)=d_1(t)$ *is bounded by* $D(\tau)$ *and that* $D(\tau)$ *is valid. Assume further that some subset of nodes has capacity constraints, and that these are all valid with respect to* $D(\tau)$, *and that these* c_k *are decreasing with increasing k. Finally, suppose that each node k places orders* $d_{k+1}(t)$ *according to* (14)*. Then*

a) *All orders* d_k *are bounded by* $D_k(\tau)$, *as specified by* (20)

b) *All* $D_k(\tau)$ *are valid*

c) c_l *for nodes with capacity constraints are valid with respect to* $D_k(\tau)$, *for all* $l\geq k$

d) *The base stock levels* $B_k(\tau)$ *as specified by* (15) *ensure that* $I_k(t)\geq 0$

e) *All* $B_k(\tau)$ *are concave in* τ .

Proof of Proposition 2: We start by proving a)-c) by induction, noting that they are true by assumption for *k* = 1. The inductive step is trivial if there is no capacity constraint; we therefore consider the case when *k* does have a capacity constraint. We make the induction hypothesis, that a)-c) are true for some *k*, and that node *k+1* has a capacity constraint. We can then use Proposition 1 to get that $D_{k+1}(\tau)=\min(c_k\tau,D_k(\tau))$. Thus a) holds for *k+1* as well. Moreover, $D_{k+1}(0)=\min(c_k\times 0,D_k(0))=0$. Both $c_k\tau$ and $D_k(\tau)$ are non-decreasing and concave, and these properties are preserved under minimization. Hence, if $D_k(\tau)$ is valid then $D_{k+1}(\tau)$ is valid as well, and so b) holds for *k+1*.

Suppose now that c) holds for *k*, that is, any c_l $(l\geq k)$ is valid with respect to $D_k(\tau)$. We need to show that any c_l $(l\geq k)$ is valid with respect to $D_{k+1}(\tau)$. By Definition 2 and since $c_l<c_k$, there is a crossing point $\tilde{\tau}$ such that

$$c_l\tilde{\tau}=D_k(\tilde{\tau}) \tag{45}$$

By the inductive assumptions a)-c), we can see from Proposition 1 that

$$\min(c_k\tilde{\tau},D_k(\tilde{\tau}))=D_{k+1}(\tilde{\tau}) \tag{46}$$

Because c_l is decreasing in *l*, we have

$$c_l\tilde{\tau}=\min(c_k\tilde{\tau},c_l\tilde{\tau}) \quad \forall l\geq k \tag{47}$$

Combining (40) – (42) gives us $c_l \tilde{\tau} = D_{k+1}(\tilde{\tau}) \quad \forall l \geq k$. That is, $c_l \tau$ crosses $D_{k+1}(\tau)$ and $D_k(\tau)$ at the same point $\tilde{\tau}$. Furthermore, for $\tau < \tilde{\tau}$ we have $c_l \tau < D_k(\tau)$ and $c_l \tau < c_k \tau$, and thus $c_l \tau < \min(c_k \tau, D_k(\tau)) = D_{k+1}(\tau)$. For $\tau > \tilde{\tau}$ we have $c_l \tau > D_k(\tau) \geq \min(c_k \tau, D_k(\tau)) = D_{k+1}(\tau)$. Thus, c_l is valid with respect to $D_{k+1}(\tau) = \min(c_k \tau, D_k(\tau))$ as well. Thus c) holds for *k+1* as well.

Therefore, we have shown that a)-c) for node *k* imply that a)-c) hold for *k+1* as well. Since the base case $k = 1$ is true by assumption, by the induction axiom a)-c) must hold for all *k*.

The proof of (d) follows immediately from a) – c), and the inventory expression (16).

The proof of (e) also follows from Lemma 1 and a) – c). □

Appendix 7: Multiple stage model with MCBS policy

We show in this appendix how to formulate the safety-stock optimization problem for a supply chain consisting of multiple stages, each of which may potentially have a capacity constraint. In this network, each stage provides guaranteed service to its customers (downstream neighbors), and operates according to a censored base-stock policy, the MCBS policy. Given this formulation we then observe how the results from the uncapacitated problem can be extended to the capacitated case for serial systems and for spanning tree topologies.

In order to describe the network and its characteristics, we index the nodes (or stages) and denote the parameters S_k, T_k, SI_k and c_k specific to node *k*. We specify the topology of the network by the directed edge set *A* where $(j,k) \in A$ indicates that node *j* directly supplies (is upstream of) node *k*. Customer facing nodes are defined by the set *C*, and have service times exogenously specified; $S_k = s_k$ for $k \in C$. Other service times (and inbound service times) are *decision variables.* The demand bound $D_k(\ )$ is a bound on the demand from the customer node(s) downstream of node *k*. Graves and Willems (2000) propose how to combine the bounds from multiple demand streams without censoring; as noted in the paper, it is not clear how to determine the demand bound that is a combination of multiple censored-demand streams. Nevertheless, for the rest of this presentation we will assume that we can determine the demand bound for each node *k*.

To facilitate the derivations to follow, we use operator notation (see e.g. Griffel, 1985) to describe how we determine the base stock from the capacity constraint and the demand bound. We use the symbol ψ_k to denote the continuous and node-specific version of (17) as follows:

$$(\psi_k D_k)(\tau) = \max_{n \geq 0} \left\{ D_k(\tau + n) - c_k n \right\} \tag{48}$$

As before, we set the base stock level to this quantity to ensure guaranteed service:

$$B_k(\tau) = (\psi_k D_k)(\tau) \tag{49}$$

If node *k* does not have a capacity constraint, we can set $c_k = \infty$ in (49) and find that $B_k(\tau) = D_k(\tau)$. At stage *k*, the total inventory that is on hand is $I_k(t)$. We approximate this, using the expectation of (7), by the average of this quantity, $\overline{I}_k + \overline{BL}_k$:

$$\overline{I}_k + \overline{BL}_k = (\psi_k D_k)(T_k + SI_k - S_k) - (T_k + SI_k - S_k)\mu \tag{50}$$

We justify this approximation by noting that the backlog term $\overline{BL}_k$ is typically very small and does not depend upon the decision variables, namely the service times; hence, its inclusion in the objective function has no impact on the optimal solution and is just a constant. We also do not include inventory in process at stage *k*, because the average pipeline inventory is proportional to the lead time T_k and is not affected by the choice of service times. We assume that stage *k* accrues holding costs proportional to $\overline{I}_k + \overline{BL}_k$, with the proportionality constant h_k.

We now formulate an optimization problem to find the service times that minimize the total inventory holding cost, subject to providing guaranteed service at all nodes, for any demand realization within the bounds.

$$\begin{aligned}
&\min_{S_k, SI_k} \sum_{k=1}^{N} h_k \left((\psi_k D_k)(SI_k + T_k - S_k) - (SI_k + T_k - S_k)\mu \right) \\
&S_k, SI_k \geq 0 \quad \forall k \\
&SI_k \geq S_j \quad \forall (j,k) \in A \\
&S_k = s_k \quad k \in C \\
&SI_k + T_k - S_k \geq \theta_k - \frac{D_k(\theta_k)}{c_k} \quad \forall k
\end{aligned} \tag{51}$$

The decision variables are the service times, which are non-negative by the first constraint. The second constraint assures that the inbound service time for each node is greater than or equal to the maximum service time from its supply nodes. The third constraint fixes the service times for customer-facing nodes to the exogenous specifications. The fourth constraint provides a lower bound on the net replenishment time for each node, where θ_k is specified by $\frac{dD_k(\theta_k)}{d\theta} = c_k$. When optimizing the safety stocks in a supply chain, we need not consider any net replenishment time in the range $\tau < \theta - \frac{D(\theta)}{c}$; for any solution in this range we can find another solution with $\tau = \theta - \frac{D(\theta)}{c}$ with less inventory. Simpson (1958) and Graves and Willems (2000) formulate the uncapacitated version of (51) for a serial system and for general networks, respectively. In both cases, they observe that an optimal solution is on a corner point of the solution space, since the problem minimizes a concave objective function over a polyhedral set. We are therefore interested in whether this observation applies here, namely whether the modified function $(\psi_k D_k)(\tau)$ is concave for each node *k*. From Lemma 1, we observe this is in fact the case under some reasonable technical conditions.

Hence, the optimal solution will be at an extreme point of the solution space. One practical implication of this is that an all-or-nothing result holds for the capacitated serial system, analogous to the result proved by Simpson (1958) for the uncapacitated case. Specifically, for a serial system either $S_k = 0$, meaning that the stage holds enough inventory to always provide immediate service, or alternatively, $S_k = SI_k + T_k - \theta_k + \frac{D(\theta_k)}{c_k}$ in which case the base stock level $B_k = 0$.

Simpson (1958) solved the uncapacitated version of (51)for a serial system by enumeration. Graves and Willems (2000) develop a dynamic programming algorithm, which can be used for networks with spanning-tree topology; Lesnaia (2004) shows how to modify and implement this algorithm so that it is polynomial. We can modify the Graves-Willems algorithm to solve(51) for spanning-tree networks with capacity constraints: we just need two changes. First, instead of using $B_k(\tau) = D_k(\tau)$ for the base stock levels, we use the exact characterization of the base stock necessary to handle capacity constraints given by(49). Second, the lower bound on the net replenishment time is no longer zero, but is given by $\theta_k - \frac{D_k(\theta_k)}{c_k}$ for node *k*. To account for this, we just extend the search space in each iteration of the dynamic program; this can be done with no change to the computational complexity.