

Strategic safety stocks in supply chains with capacity constraints

May 2008, revised January 2009

Tor Schoenmeyr • Stephen C. Graves

Optisolar, Inc. 31302 Huntwood Avenue, Hayward, CA 94544, tschoenmeyr@optisolar.com
Leaders for Manufacturing Program and A. P. Sloan School of Management,
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307, sgraves@mit.edu

We generalize the guaranteed-service (GS) model for multi-echelon safety stock placement to include capacity constraints. We first develop an extension of the single-stage base-stock model to include a capacity constraint. We then use this result to model a multi-stage system with a base-stock operating policy. We establish that we can adapt the existing algorithms for the unconstrained case to solve for the safety stocks in a capacitated system. We then consider a multi-stage system in which stages censor their orders, based on their capacity limits. Again we analytically characterize the necessary base stock levels, and develop an extension to the existing dynamic programming algorithms to find the optimal base stock levels and safety stocks. The censored order policy leads to a better solution compared to that for the base-stock policy. Indeed, we find that the total holding costs for the censored order policy can be less than that for the corresponding base-stock system without capacity constraints.

Subject classifications: Inventory/production: Multi-echelon; policies. Manufacturing: Strategy.

Strategic safety stocks in supply chains with capacity constraints

May 2008, revised January 2009

Tor Schoenmeyr • Stephen C. Graves

Optisolar, Inc. 31302 Huntwood Avenue, Hayward, CA 94544, tschoenmeyr@optisolar.com
Leaders for Manufacturing Program and A. P. Sloan School of Management,
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307, sgraves@mit.edu

We generalize the guaranteed-service (GS) model for multi-echelon safety stock placement to include capacity constraints. We first develop an extension of the single-stage base-stock model to include a capacity constraint. We then use this result to model a multi-stage system with a base-stock operating policy. We establish that we can adapt the existing algorithms for the unconstrained case to solve for the safety stocks in a capacitated system. We then consider a multi-stage system in which stages censor their orders, based on their capacity limits. Again we analytically characterize the necessary base stock levels, and develop an extension to the existing dynamic programming algorithms to find the optimal base stock levels and safety stocks. The censored order policy leads to a better solution compared to that for the base-stock policy. Indeed, we find that the total holding costs for the censored order policy can be less than that for the corresponding base-stock system without capacity constraints.

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 to certain key locations, effectively decoupling different parts

of the supply chain. Once the optimal safety stock strategy has been determined, each stage of the supply chain can operate independently, providing guaranteed service to its downstream customer, and operating according to a simple base stock policy, with a minimum need for communication and coordination between different parts of the supply chain. Graves and Willems (2003) term this framework for supply chain management as the guaranteed service (GS) model.

To our knowledge, all published work on the GS model assumes unlimited capacity for processing and inventory storage at each stage. 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 existing theoretical models.

The intent of this paper is to extend the GS model for the optimal placement of safety stock inventory to supply chains with capacity constraints. In particular we strive to develop models and algorithms that have the potential to determine safety stocks in real-world supply chains. We assert that the paper makes three contributions.

First we determine for a single-stage capacitated system the base-stock level that is required to provide guaranteed service, assuming bounded demand. This is a generalization of Kimball's classic single-stage base-stock model (original manuscript 1955, reprinted in 1988) to account for a capacity constraint.

Second, we use the single-stage model to model a supply chain, namely a multi-stage network, where each stage can have a capacity constraint and where we assume that each stage operates with a base-stock policy. With the assumption of a concave demand bound, 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 establish for networks with spanning tree topologies that we can directly apply the efficient safety-stock optimization methods developed for uncapacitated supply chains to the safety-stock optimization problem for capacitated supply chains.

Third, we analyze a modification of the base-stock policy, in which a node propagates an order which is the lesser of its capacity and the order it receives (plus extra quantities to “catch up”, as necessary). We refer to this as the “censored” ordering policy. We show again that the all-or-nothing property holds for serial systems. We also show that we can optimize the safety stock inventory in supply chains with censored ordering and capacity constraints, with small modifications to the Graves-Willems’ dynamic programming method. We find that the inventory holding costs for the censored ordering policy are less than for the original base-stock policy, and sometimes even less than that for the corresponding system without capacity constraints. Moreover, for the censored policy the orders still depend only on local information.

This paper consists of six sections. In the remainder of this section, we provide a brief review of related literature. In §2, we generalize the base-stock model of Kimball (1988) and Simpson (1958) to include a capacity constraint in a single-stage setting. In §3, we consider optimization of a supply chain with a base-stock policy and potentially any number of capacity constraints. In particular, we find that the optimization procedures and structural results identified by Simpson (1958) and Graves and Willems (2000) carry over to this setting. In §4, we analyze what happens if each stage modifies its order based on its capacity constraint, i.e., censors the order. We find that the structural properties and optimization methods carry over to this setting as well. In §5, 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 §6.

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). 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, 2008) have developed very effective algorithms

for optimizing the safety stocks in large-scale real-world supply chains. We also note that the GS framework has been deployed successfully in industry (e.g., Billington et al 2004, Willems 2008).

There is a somewhat 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) examine a single stage system with fixed ordering costs and capacity constraints, and find that the optimal policy takes an (s,S) form. Gallego and Toktay (2004) also consider a single-stage system under the assumption that all 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 no lead-time at the upstream stage. For these specific 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-zero 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 will 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 order from its adjacent downstream stage. Furthermore, at each capacitated stage, the stage orders the lesser of its capacity and 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.

Glasserman and Tayur (1994) consider the stability properties of multi-echelon systems with capacity constraints. They find that inventories and back-logs are stable (i.e., they converge

to unique stationary distributions from any initial state) if the mean demand is less than the capacity constraint. In subsequent papers, Glasserman and Tayur assume that an echelon-base stock policy is used in a multi-echelon system, and find optimal order points using simulation and perturbation analysis (Glasserman and Tayur, 1995), and analytical approximations (Glasserman and Tayur, 1996).

Gupta and Selvaraju (2006) develop an approximation for analyzing a serial system as a queueing network, where the stages have exponentially distributed service 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 Gupta and Selvaraju propose some approximations.

A markedly different approach is taken by Bertsimas and Thiele (2004, 2006), who show that capacity constraints can be incorporated into a tractable, robust optimization problem. This approach can handle general networks, and uses echelon policies as in the Clark and Scarf (1960) model. 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 back-order 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

In this section we generalize the single-stage model originally developed by Kimball (1988) and Simpson (1958) to include a capacity constraint. A stage represents a processing activity that requires one or more inputs and that converts these inputs into an output product or final good. The output can be stored as inventory at the stage, and is used to meet demand from

multiple customers or as input into downstream stages. 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 let $d(t)$ denote the demand in period t . We assume that the stage provides the same *guaranteed service time* S to each of its customers; this means that the stage guarantees that it will satisfy the demand $d(t)$ by time $t + S$, where S is a non-negative integer.

We also assume that the suppliers to the stage provide a guaranteed service time, which we denote by SI for the inbound service time. Thus, for an order placed at time t , the suppliers will deliver their inputs to the stage at time $t + SI$.

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 all of the inputs are on hand. We assume a deterministic lead-time T equal to the time between when the process starts and when the process completes and the inventory is available to serve demand. The lead-time represents any fixed processing time at the stage (and does not include any time waiting for input; see below). 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.

The stage operates with a periodic review base-stock replenishment policy with a review period being one time unit (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 each of its upstream suppliers. The stage then receives the earlier order placed at time $t - SI$ from each of the upstream suppliers; 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 $d(t - S)$.

For the base-stock policy without a capacity constraint, each period the stage places an order, equal to $d(t)$, on its upstream suppliers. When these inputs are received at time $t + SI$, the stage will then initiate production of $d(t)$ units; that is, the release quantity at time $t + SI$ is $d(t)$, which will complete the process and be placed in inventory at time $t + SI + T$.

We now adapt this policy to account for a capacity constraint. We again assume that each period the stage places an order, equal to $d(t)$, on each of its upstream suppliers. When the stage receives these inputs at time $t + SI$, it places the supplies into an internal queue; the stage will then process as much of the internal queue as possible, subject to its capacity constraint. When the constraint is binding, the stage will release less than $d(t)$ at time $t + SI$, and the delayed production (in the form of the internal queue) is carried over until the next time period. We illustrate the envisioned arrangement in Figure 1.

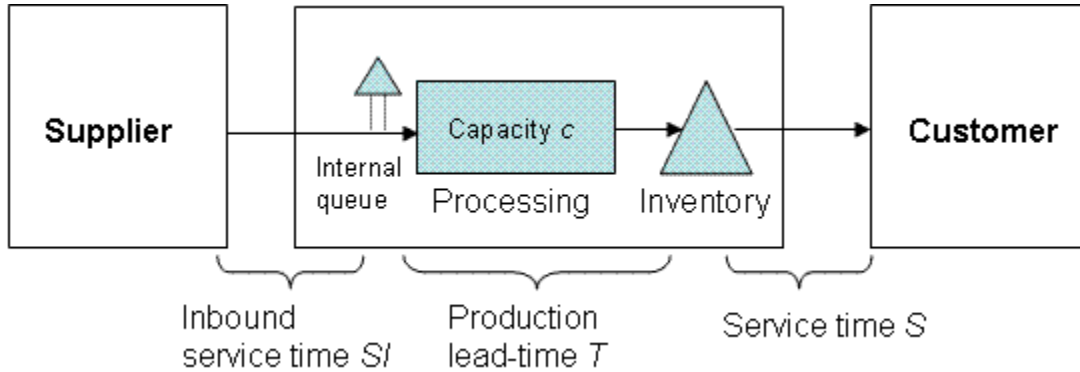


Figure 1: Overview of a single-stage system

We denote the production release at time t by $R(t)$. With capacity constraints, we have

$$R(t) = \min(c, d(t - SI) + IQ(t - 1)), \quad (1)$$

where $IQ(t)$ is the internal queue, given by the equation:

$$IQ(t) = IQ(t - 1) + d(t - SI) - R(t). \quad (2)$$

The balance equation for the final-good inventory $I(t)$ at the stage is now:

$$I(t) = I(t-1) + R(t-T) - d(t-S). \quad (3)$$

Combining (2) and (3) we have

$$I(t) + IQ(t-T) = I(t-1) + IQ(t-T-1) + d(t-T-SI) - d(t-S). \quad (4)$$

If the system starts at time $t = 0$ with $d(t) = 0, IQ(t) = 0$, for $t \leq 0$ and $I(0) = B$, where B is the base-stock level, then for suitably large t we can write the inventory as:

$$I(t) = B - d(t-T-SI, t-S) - IQ(t-T) \quad (5)$$

where we define the notation

$$d(a, b) = \begin{cases} \sum_{i=a+1}^b d(i) & \text{for } a < b \\ 0 & \text{for } a = b \\ -\sum_{i=b+1}^a d(i) & \text{for } a > b \end{cases} \quad (6)$$

In the Appendix we show that we can express the internal queue as:

$$IQ(t) = \max_{n \in Z} \{d(t-SI-n, t-SI) - cn\} \quad (7)$$

where Z denotes the set of non-negative integers. We substitute (7) into (5) to obtain:

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

The base-stock problem is now to determine the minimal value of B that assures that the inventory $I(t)$ is non-negative, which is sufficient to satisfy the guaranteed service commitment.

We assume that $d(t)$ is non-negative and takes the average value μ where $\mu < c$.

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

$$D(\tau) = \max \{d(t, t+\tau)\} \quad \forall t, \tau \geq 0. \quad (9)$$

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 can now combine (8) and (9) to find that the minimal base stock for $I(t) \geq 0$ is:

$$B(\tau) = \max_{n \in \mathbb{Z}} \{D(\tau + n) - cn\} \quad \text{where } \tau = T + SI - S. \quad (10)$$

In (10) τ denotes the *net replenishment time* for the stage without the capacity limits. We write the base stock in (10) 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 uncapacitated net replenishment time τ .

To get some insight into the nature of the solution in (10), we will consider demand bounds and capacity constraints that are valid as defined below:

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

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

These properties hold true for demand bounds and capacity constraints that arise in practice. 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).¹

In order to get an explicit solution to the maximization in (10), we assume the demand bound is differentiable and ignore the integrality restriction on the argument n .² We define θ by $D'(\theta) = c$, i.e., θ is the point at which the derivative of the demand bound equals the capacity. Then the base stock is:

$$B(\tau) = \begin{cases} 0 & \text{for } \tau < \theta - \frac{D(\theta)}{c} \\ c(\tau - \theta) + D(\theta) & \text{for } \theta - \frac{D(\theta)}{c} \leq \tau < \theta \\ D(\tau) & \text{for } \tau \geq \theta \end{cases} . \quad (11)$$

Thus the base stock is zero for $\tau \leq \theta - \frac{D(\theta)}{c}$.³ For higher values of τ , the necessary 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 uncapacitated case.

We note that unlike the uncapacitated base-stock model, we permit the net replenishment time to be negative when we have a capacity limit. That is, the stage quotes a service time S that is **longer** than the nominal time $SI + T$ that it takes for its inventory to be replenished; but due to

¹ These definitions are not needed to solve (10). All that is needed is that eventually the demand bound falls below the capacity, i.e., $D(\tau) < c\tau \quad \forall \tau > \tilde{\tau}$. Nevertheless, imposing these definitions provides tractability that permits us to generate insights into the nature of the solutions.

² When we ignore the integrality restriction on n in (10) 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.

³ 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.

the capacity constraint, the stage may still need a positive base stock in order to provide guaranteed service. In particular, consider $\tau = \theta - \frac{D(\theta)}{c} < 0$, at which the base stock is zero; thus, in order for the stage to provide guaranteed service with a zero base stock, it sets its service time S equal to its replenishment time $SI + T$ plus an increment $\frac{D(\theta)}{c} - \theta$.

In practice we often set the bound function analogous to a probabilistic service level with i.i.d. normally distributed demand. That is, we set

$$D(\tau) = \mu\tau + z\sigma\sqrt{\tau}, \quad (12)$$

where σ corresponds to the standard deviation of demand and z is a safety factor. We note that this bound 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 we can find from (11) the base stock to be:

$$B(\tau) = \begin{cases} 0 & \text{for } \tau < -\left(1 - \frac{\mu}{c}\right)\left(\frac{\sigma}{c - \mu}\right)^2 \\ c\tau + \frac{\sigma^2}{c - \mu} & \text{for } -\left(1 - \frac{\mu}{c}\right)\left(\frac{\sigma}{c - \mu}\right)^2 \leq \tau < \left(\frac{\sigma}{c - \mu}\right)^2 \\ \mu\tau + 2\sigma\sqrt{\tau} & \text{for } \tau \geq \left(\frac{\sigma}{c - \mu}\right)^2 \end{cases} \quad (13)$$

Thus the base stock with a capacity constraint takes a rather simple and intuitive form. Again the

capacity constraint is immaterial when $\tau \geq \theta = \left(\frac{\sigma}{c - \mu}\right)^2$. 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 plot the capacitated base stock level (13) in Figure 2, together with the base stock level for the unconstrained case.

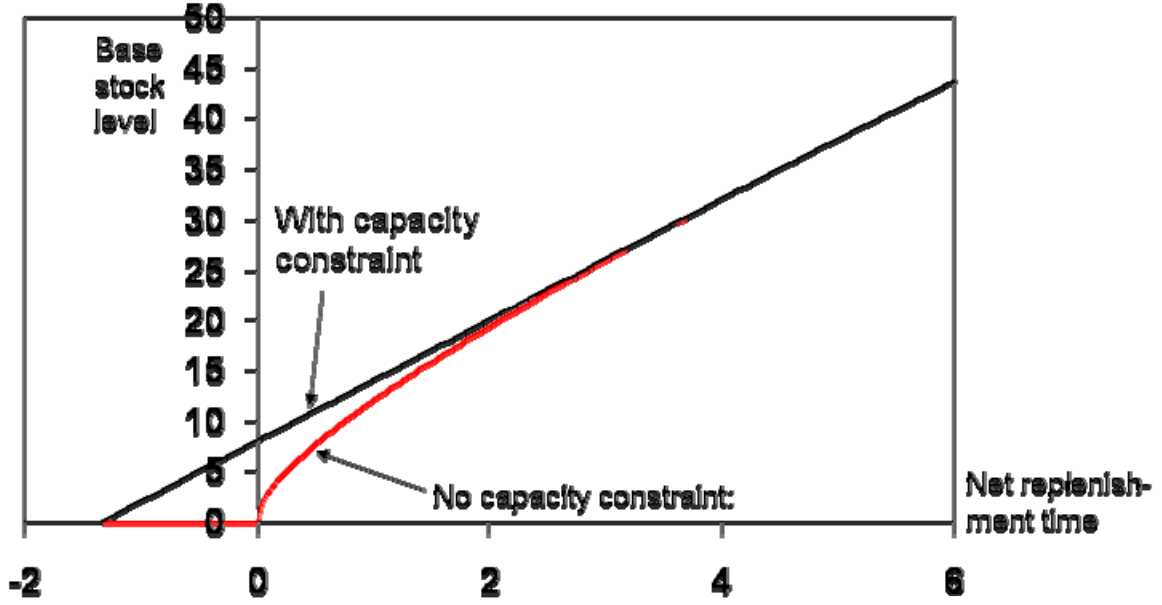


Figure 2: With capacity constraints, for small τ a higher base stock level is necessary, but for $\tau \geq 4$ the capacity constraint does not matter. Plotted parameters are $\mu = \sigma = 4$, $c = 6$ and $z = 2$

To get some additional insight, we use (5) and (13) to find the average inventory in finished goods and the internal queue $E[I(t)] + E[IQ(t-T)]$; as we expect the internal queue to be modest, we use this to approximate the average safety stocks:

$$B(\tau) - \mu\tau = \begin{cases} (c - \mu)\tau + \frac{\sigma^2}{c - \mu} & \text{for } -\left(1 - \frac{\mu}{c}\right)\left(\frac{\sigma}{c - \mu}\right)^2 \leq \tau < \left(\frac{\sigma}{c - \mu}\right)^2 \\ 2\sigma\sqrt{\tau} & \text{for } \tau \geq \left(\frac{\sigma}{c - \mu}\right)^2 \end{cases} \quad (14).$$

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 safety stock is a

fixed amount $\frac{\sigma^2}{c - \mu}$ plus a *variable amount that increases linearly* in the net replenishment

time. In contrast, for the uncapacitated 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 uncapacitated setting, we regard (14) to be a valuable back-of-the-envelope heuristic in that it succinctly displays how the safety stock depends on the slack capacity, the demand variability and the net replenishment time.

3. Multiple stage model with base-stock ordering

We now investigate 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 (local) base-stock policy. We assume that customer demand is immediately propagated up through the system, so that in each period t each node places an order on its suppliers equal to the sum of the customer demand from all its adjacent downstream stages.

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 $D_k(\cdot)$ is a bound on the demand from the customer node(s) downstream of node k ; see Graves and Willems (2000) for how to combine the bounds from multiple demand streams. 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 (10) as follows:

$$(\psi_k D_k)(\tau) = \max_{n \geq 0} \{D_k(\tau + n) - c_k n\} \quad (15)$$

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

$$B_k(\tau) = (\psi_k D_k)(\tau) \quad (16)$$

If node k does not have a capacity constraint, we can set $c_k = \infty$ in (15) and find that

$B_k(\tau) = D_k(\tau)$. At stage k , the total inventory that is on hand (either as finished inventory or delayed in the internal queue) is $I_k(t) + IQ_k(t)$. We use the node-specific version of equation (5) to characterize the average of this quantity, $\bar{I}_k + \overline{IQ}_k$:

$$\bar{I}_k + \overline{IQ}_k = (\psi_k D_k)(T_k + SI_k - S_k) - (T_k + SI_k - S_k)\mu \quad (17)$$

We 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 $\bar{I}_k + \overline{IQ}_k$, with the proportionality constant h_k . This is a simplification that we make for tractability. In many contexts one might expect the holding cost for the internal queue inventory to be less than that for the finished good, as holding costs should increase as we add more value to the product by processing. Nevertheless, we expect this difference in holding costs to be modest; moreover, the average internal queue \overline{IQ}_k does not depend on the service times and thus has no impact on the optimal solution.

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.

$$\begin{aligned}
& \min_{S_k, SI_k} \sum_{k=1}^N h_k ((\psi_k D_k)(SI_k + T_k - S_k) - (SI_k + T_k - S_k)\mu) \\
& S_k, SI_k \geq 0 \quad \forall k \\
& SI_k \geq S_j \quad \forall (j, k) \in A \\
& S_k = s_k \quad k \in C \\
& SI_k + T_k - S_k \geq \theta_k - \frac{D_k(\theta_k)}{c_k} \quad \forall k
\end{aligned} \tag{18}$$

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, as discussed in the prior section, where θ_k is specified by $D_k'(\theta_k) = c_k$. Simpson (1958) and Graves and Willems (2000) formulate the uncapacitated version of (18) for a serial system and for general networks, respectively. In both cases, they observe that an optimal solution is on a corner 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 . Under some reasonable technical conditions, this is in fact the case.

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

$$(\psi D)(\tau) = \max_{n \geq 0} \{D(\tau + n) - cn\} \text{ is concave on } \tau \in [\theta - \frac{D(\theta)}{c}, \infty).$$

The proofs of all propositions are in the Appendix. 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,

$$S_k = SI_k + T_k - \theta_k + \frac{D(\theta_k)}{c_k} \text{ in which case the base stock level } B_k = 0.$$

Simpson (1958) solved the uncapacitated version of (18) for a serial system by enumeration. Graves and Willems (2000) develop an exact 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 (18) for spanning-tree networks with capacity constraints: we just 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 (16). 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 for each iteration of the dynamic program; this can be done with no change to the computational complexity.

4. Multiple stage model with censored order policy

In the prior section we have shown how to generalize the GS framework and associated optimization methods to encompass capacity constraints. In this section, we show that certain improvements are possible, if the stages with capacity constraints modify their orders. The basic idea with the censored policy is that a stage should not propagate a full order upstream, if it knows that it will be unable to process such a quantity because of its capacity constraints. Alternatively, we observe that there is no need for an internal queue at a capacitated stage; rather the stage should place orders on its supplier so that these orders arrive at a rate consistent with the stage's capability to process the work.

To simplify the presentation, we consider a serial system, where we number the nodes from downstream to upstream: node 1 is the customer facing node, and N is the most upstream node. We will briefly discuss censorship in other network structures at the end of this section.

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 $d_k(t)$; we denote its bound by $D_k(t)$, where

$$d_k(t, t + \tau) \leq D_k(\tau) \quad \forall t, \tau \geq 0. \quad (19)$$

Accordingly, we write $d_1(t) = d(t)$ and $D_1(t) = D(t)$ for customer demand and its bound.

We assume that each node k never orders more than its capacity c_k . Thus whenever its demand exceeds its capacity ($d_k(t) > c_k$), it orders less than the demand and creates a shortfall. When demand falls below c_k again, stage k increases its order to reduce the shortfall. To this end, node k keeps a backlog $BL_k(t)$ of this shortfall, equal to the amount of its demand for which it has yet to place a replenishment order. Node k will increase its order whenever it has a backlog and capacity is available. We now specify the orders placed at time t by node k on node $k + 1$ by:

$$d_{k+1}(t) = \min(d_k(t) + BL_k(t-1), c_k). \quad (20)$$

We can specify the backlog $BL_k(t)$ recursively:

$$\begin{aligned} BL_k(t) &= \max \{ BL_k(t-1) + d_k(t) - c_k, 0 \} \\ &= \max \{ \max \{ BL_k(t-2) + d_k(t-1) - c_k, 0 \} + d_k(t) - c_k, 0 \} \\ &= \max_{n \in \mathbb{Z}} \{ d_k(t-n, t) - c_k n \}. \end{aligned} \quad (21)$$

Here we assume that the system starts at $t = 0$ with $BL_k(t) = 0, d_k(t) = 0$ for all $t \leq 0$. In each period, stage k adds the difference between its demand and capacity, $d_k(t) - c_k$, to the backlog, subject to keeping the backlog non-negative.

We can show that the inventory for node k , $I_k(t)$, is the inventory for the uncapacitated 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 time $t - SI_k - T_k$ cannot be available by time t to meet demand at node k . Thus, we have:

$$\begin{aligned}
I_k(t) &= B_k - d_k(t - SI_k - T_k, t - S_k) - BL_k(t - SI_k - T_k) \\
&= B_k - d_k(t - SI_k - T_k, t - S_k) - \max_{n \in Z} \{d_k(t - n - SI_k - T_k, t - SI_k - T_k) - c_k n\} \\
&= B_k - \max_{n \in Z} \{d_k(t - n - SI_k - T_k, t - S_k) - c_k n\}
\end{aligned} \tag{22}$$

Except for the fact that the demand d_k is now stage-specific, this is equivalent to the inventory equation for a capacitated stage that does not censor its orders, namely equivalent to (5). That is, when considering the inventory $I_k(t)$ at some stage, it makes no difference whether items are waiting in the internal queue, or whether the orders placed by that stage were temporarily put into a backlog because of censorship. In both cases the quantities that start and finish their processing in each period are the same. Thus if node k has a capacity constraint, we can use equation (16) to determine the base stock level that guarantees service, regardless of whether a censored order policy is employed or not. Thus we set the base stock level by

$$B_k(\tau) = (\psi_k D_k)(\tau) .$$

where we use the operator ψ_k defined in (15).

There are a couple of immediate implications from the censored order policy. First, in comparison to the base-stock order policy 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 (22):

$$\bar{I}_k = B_k - \mu(SI_k + T_k - S_k) - \overline{BL}_k \tag{23}$$

Thus in the case of censored orders, we need to calculate the term \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 Appendix. However, we note that \overline{BL}_k does not depend on the service times. Thus we do not need to determine \overline{BL}_k to find the optimal safety stocks; rather, the sole purpose for determining \overline{BL}_k is for the determination of the average inventory level given by (23)

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

A third more significant implication of the censorship is that the upstream stages 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 orders.

Proposition 2. *Suppose $d_{k+1}(t) = \min(d_k(t) + BL_k(t-1), c_k)$ where BL_k is given by (21), and initialized with $BL_k(t) = 0$ for $t \leq 0$. Assume that D_k is a valid bound for d_k (that is, $d_k(t, t + \tau) \leq D_k(\tau) \quad \forall t, \tau \geq 0$), and that c_k is valid with respect to $D_k(\tau)$. Then*

$$d_{k+1}(t, t + \tau) \leq D_{k+1}(\tau) = (\Phi_k D_k)(\tau) \quad \forall t, \tau \geq 0 \quad (24)$$

where Φ_k is defined by

$$(\Phi_k D)(\tau) = \min(c_k \tau, D(\tau)). \quad (25)$$

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 (21), with (22) to account for the backlog of orders. We can then use Proposition 2 to compute the demand bound at each node. Given these demand bounds we can then use (16) to characterize the necessary base stock level for each node. Given the base stock level, we can calculate the average

inventory from (23). We summarize these iterative steps in Table 1, with a comparison to the uncapacitated model.

	No capacity constraint at node k	Capacity constraint c_k at node 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	$B_k(\tau_k) = D_k(\tau_k)$	$B_k(\tau_k) = (\Psi_k D_k)(\tau_k)$ $= \max_{n \geq 0} \{D_k(\tau_k + n) - c_k n\}$
Average inventory	$\bar{I}_k = B_k(\tau_k) - \mu \tau_k$	$\bar{I}_k(t) = B_k(\tau_k) - \mu \tau_k - \overline{BL}_k$

Table 1: Summary of node properties in serial system supply chain with capacity constraint(s) and censored base stock policy; we use τ_k to denote the net replenishment time at node k .

We can now embed this model in an optimization, analogous to (18), to find the best choices for the service times. In the following proposition we establish that the demand bounds and capacity constraints are valid as defined earlier; as these properties were 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 3. *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 (20). Then*

- a) *All orders d_k are bounded by $D_k(\tau)$, as specified by (24)*
- 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 (16) 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 censorship, we can compute the demand bounds and necessary base stock levels by recursively applying a sequence of functional operators, as summarized in Table 1. Thus, as for the case of base-stock ordering, we can use the algorithm developed by Graves and Willems to find the optimal service times for a serial-system supply chain with both capacity constraints and censorship, after only small modifications.

We can extend the results in this section to arborescent (assembly tree) supply chain topologies in which each node has a single customer node. The iterative steps laid out in Table 1 apply directly, but with one modest modification: the order placed by node k (given by (21)) is now placed concurrently on each of the suppliers to node k . 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, 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 uncapacitated supply chains, Graves and Willems (2000) propose how to set bounds if demand streams are independent, and the bound is set analogous to a probabilistic service level, as in (12). They also propose bounds for larger or smaller measures of risk pooling. If one or more bound are generated by a censored order policy, then it is not clear how best to merge 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 censored order policy.

5. 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 $(\mu, z, \sigma) = (40, 2, 20)$. We have three alternatives each for the holding costs and for the lead-times, as shown in Table 2.

	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 2: 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.⁴ 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 $\mu + 0.1\sigma$ up to $\mu + 1.5\sigma$. Thus, we specify $3 \times 3 \times 5 \times 5 = 225$ 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 solve for both the base-stock policy and the censored ordering policy. To benchmark the performance of the uncapacitated GS model, we also solve the corresponding uncapacitated problem to find its inventory holding cost and optimal policy. The uncapacitated solution is typically not feasible for the capacitated problem in that it will not provide guaranteed service. However, we can adapt the uncapacitated solution to create a feasible censored ordering policy for the capacitated problem; we assume the decoupling safety stocks are located as given by the uncapacitated solution, and then find the censored ordering policy that

⁴ The echelon holding cost for stage k is $h_k - h_{k+1}$.

provides guaranteed service for these safety stock locations. In this way we can evaluate the performance of the uncapacitated solution in the presence of a capacity constraint.

In Table 3 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 costs of the censored policy, the base-stock policy, and the adapted uncapacitated policy; in each case these costs are given as a percentage of the cost for the corresponding uncapacitated problem, which is also given in the table. We make three observations.

Holding cost profile	Lead time profile	Uncap Cost	Location of Capacity Constraint				
			5	4	3	2	1
Upstream Heavy	Upstream Heavy	400	99%	104%	106%	103%	89%
			102%	111%	116%	114%	100%
			105%	107%	109%	111%	93%
	Constant	400	103%	106%	108%	109%	91%
			106%	112%	116%	118%	100%
			105%	107%	109%	111%	93%
Downstream Heavy	400	104%	107%	109%	110%	92%	
		107%	112%	116%	118%	100%	
		105%	107%	109%	111%	93%	
Constant	Upstream Heavy	368	98%	95%	93%	87%	73%
			100%	100%	102%	102%	100%
			98%	97%	105%	107%	89%
	Constant	394	98%	99%	101%	103%	86%
			100%	104%	112%	115%	100%
			98%	101%	103%	105%	88%
	Downstream Heavy	400	101%	104%	106%	108%	91%
			103%	108%	111%	115%	100%
			103%	105%	107%	109%	93%
Downstream Heavy	Upstream Heavy	268	100%	97%	91%	81%	65%
			100%	100%	100%	100%	100%
			100%	97%	91%	97%	73%
	Constant	346	100%	98%	95%	96%	78%
			100%	100%	102%	109%	100%
			100%	98%	101%	104%	87%
	Downstream Heavy	392	100%	98%	98%	101%	86%
			100%	100%	103%	113%	100%
			100%	98%	101%	104%	89%

Table 3: Results for test problems with capacity constraint $c = 45$. In each cell, the three numbers are the normalized costs for the censored order policy (top), the base-stock policy (middle) and the adapted uncapacitated policy (bottom). The normalized costs are given as a percentage of the corresponding unconstrained problem, whose cost is in the third column.

First, as expected, the censored ordering policy dominates the base-stock policy. Indeed, for the 225 test problems, we found an average cost reduction of 8.0% when we replace the base-stock policy with the censored ordering policy. Thus when capacity constraints are present, a strong case can be made for censoring the orders at that stage.

Second, the adapted uncapacitated policy performs quite well as a heuristic. For 75% of the test problems with capacity $c = 45$, the cost for the adapted policy is within 3% of that for the censored ordering policy. However, when the upstream lead-times are long and the constraint is downstream at stage 1 or 2, the adapted uncapacitated policy can cost 8 to 20% more than the censored ordering policy.

The third observation is that the cost impact of the capacity constraint is not great. For the test problems with capacity $c = 45$, the incremental cost relative to the cost for the uncapacitated supply chain is at most 10%, and is less than 5% for 80% of the cases. More surprising, though, is the fact that under censorship the costs are often less than in the corresponding problem *without capacity constraints*. Indeed, for the 225 test problems, we find that the total cost for the constrained system with a censored ordering policy is on average 3.6% lower than the total cost for the corresponding unconstrained base-stock system.

To get a better understanding of the last two observations, we consider the test problems with a constant holding cost profile and upstream-heavy lead-times. In Table 4, we show the locations of the decoupling safety stocks for the uncapacitated solution and for the censored ordering policy for each location of the constraint. In the unconstrained problem, the optimal solution is to have one large inventory (and a long net replenishment time) at the first, customer-facing stage and a small inventory at stage 5; the inventory at stage 1 is sized to cover the demand

bound over the net replenishment time of 64 days. When we add a capacity constraint, we find that the structure of the solution changes in two ways for these test problems. First we put a decoupling inventory 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 second change is to add a decoupling inventory at each stage that is upstream from the constraint. The censored ordering policy smoothes the order signal that gets sent upstream of the capacitated stage. This reduction in demand variability makes it economical to hold a decoupling safety stock upstream of the constraint; that is, the cost of each upstream inventory is less than the savings from the resulting reduction in inventory at the immediately downstream stage.

	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 Locations				
Location of Capacity Constraint	No Constraint	1	0	0	0	1
	Stage 5	1	0	0	0	1
	Stage 4	1	1	0	0	1
	Stage 3	1	1	1	0	1
	Stage 2	1	1	1	1	1
	Stage 1	1	1	1	1	1

Table 4: Structure of solutions for test problems with constant holding cost profile and upstream-heavy lead-times. A “1” signifies the location of a decoupling safety stock.

We note here that adding a constraint can actually result in a lower inventory cost than in the uncapacitated solution, due to the censoring of the demand signal. The best outcome (in terms of total supply chain costs) occurs when the constraint is downstream in the supply chain, ideally at the customer-facing stage. This is especially true when the customer-facing stage needs to carry inventory to provide a zero service time to its customers. Indeed, in all the cases we investigated, for a given capacity value, the best outcome was always when the constraint is at the customer facing-node. In order to make a meaningful impact, the capacity constraint needs to be

only slightly higher than average demand; in fact, for all the examples with censorship at the customer-facing stage, the lowest tested capacity $\mu + 0.1\sigma = 42$ always gave the best value. However, we also know from (13) that as the capacity approaches average demand, the necessary base stock level goes to infinity.

When determining the average cost for the censored ordering policy, one must calculate the term \overline{BL}_k in (23). We discuss this topic in greater detail in the Appendix. We note here that this term will depend on specific properties of the demand process, properties which up until this point have not been specified. In the experiments listed here we estimated this term using both the formula (A28) and the numerical estimates of \overline{BL}_k based on assuming that demand is normally distributed and i.i.d. The average difference in terms of total costs was only 2.1%, and none of the overall results and conclusions were different from those presented here. We also emphasize that while the value of the term \overline{BL}_k does impact the total cost, it does not affect the optimal solution, nor does a poorly estimated \overline{BL}_k compromise the guaranteed service constraint.

6. 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. We have characterized the base stock level for two cases that depend on how orders are propagated across the supply chain. In both cases we can extend the structural findings and solution methods that have been developed for the uncapacitated supply chains, e.g., by Simpson (1958) and Graves and Willems (2000).

In general we expect to need more safety stock when we have capacity constraints. Indeed, as is clear from (13), the costs associated with capacity constraints can be arbitrarily

large, as the slack capacity goes to zero. However, for stages with sufficiently long net replenishment time, there may be no additional cost associated with capacity constraints.

When we use the censored order policy, the costs are not only lower than in the corresponding capacity-constrained problem with base-stock ordering, but frequently even lower than in the unconstrained problem as well. Intuitively, a capacity constraint typically increases the base stock level for the stage with the constraint (Equation (10); Figure 2); however, the resulting increase in the average inventory may not be as great because some of the increase in the base stock ends up as the positive backlog. Furthermore, the censored orders can be much smoother than the original demand process. As a consequence there can be a substantial reduction in the need for inventory at upstream nodes (Proposition 2; Table 4). The total effect may imply higher or lower costs depending on the specific parameters.

On a more abstract level, it may seem surprising or even paradoxical that adding a constraint can lead to a better solution. The explanation is that we have in effect expanded the space of possible ordering policies. The original GS model from Simpson (1958) is based under the assumption that all stages operate with a base stock policy. The results from our experiments illustrate that this policy need not be optimal in a multi-echelon supply chain with guaranteed service. Indeed, one might want to introduce a censored order policy even in the absence of capacity constraints in order to get these benefits. In our experiments the best outcome was when the customer-facing node was tightly constrained. This suggests that it may be preferable 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.

As we have noted, the costs (but not the optimal solution) of systems with censorship depend on the term \overline{BL}_k , which in turn depends on specific properties of the demand distribution, and which moreover appears difficult to estimate. This suggests opportunities for future research, but also hints at certain fundamental limitations on the ability to predict the

performance of censored systems. It is a rare practical situation in which one can be confident about higher-order properties of the demand distributions (although, in our experiments, the total system costs were quite similar when we estimated \overline{BL}_k in different ways). The challenges of calculating \overline{BL}_k also make it difficult to find “optimal” censorship value(s) and location(s), although surely improvements are possible over the simple brute-force searches we presented in our experiments.

Another opportunity for future research is to examine how to extend the methods for uncapacitated supply chains modeled as general networks (Humair and Willems, 2006, 2008) to account for capacity constraints.

A final line of research might relax the assumption that the customer service times are 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.

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 reviewers for their helpful comments on an earlier version of this paper.

Appendix 1: The average backlog

Here we consider estimating \overline{BL}_k . We recall that this quantity is necessary to understand the expected 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 back log described in (21) behaves rather like a queue - there is a random quantity of arrivals every period, and a fixed, maximum processing rate. Even though BL_k operates in discrete time, we can use continuous-time queuing theory as an approximation. Suppose that we model the internal queue with an M/D/1 queue with arrival rate λ , and deterministic processing time s . We then seek λ and s that agree with the average μ and standard deviation σ of demand per period, or $\frac{\mu}{c}$ and $\frac{\sigma}{\sqrt{c}}$ if we normalize with respect to capacity. That is, we model demand using a continuous-time model with Poisson arrivals, but we ensure that the probability distribution of the number of arrivals agrees in first and second moments to our original process. We are making a second order, continuous-time, approximation:

$$\begin{aligned}\frac{\mu}{c} &= \lambda s \\ \frac{\sigma}{\sqrt{c}} &= \sqrt{\lambda} s\end{aligned}\tag{A26}$$

Conversely, if we already have a mean and standard deviation for demand, we can invert the relations (A26) solve for s and λ .

$$\begin{aligned}
s &= \frac{\mu}{\lambda c} = \frac{\mu}{\left(\frac{\sigma^2}{s^2}\right)} = \frac{\mu s^2}{\sigma^2} \\
s &= \frac{\sigma^2}{\mu} \\
\lambda &= \frac{\mu}{s c} = \frac{\mu^2}{c \sigma^2}
\end{aligned} \tag{A27}$$

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:

$$\begin{aligned}
\overline{BL} &= \overbrace{\left(\rho + \frac{\rho^2}{2(1-\rho)}\right)}^{\text{Expected number of jobs}} \times \overbrace{\frac{1}{s}}^{\text{Time per job}} \\
&= \left(\frac{\mu}{c} + \frac{\left(\frac{\mu}{c}\right)^2}{2\left(1 - \frac{\mu}{c}\right)}\right) \left(\frac{\sigma^2}{\mu}\right) \\
&= \left(1 + \frac{\mu}{2(c-\mu)}\right) \left(\frac{\sigma^2}{c}\right) \\
&= \left(\frac{2c-\mu}{c-\mu}\right) \left(\frac{\sigma^2}{2c}\right)
\end{aligned} \tag{A28}$$

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 numerically or using historical data. One can simply evaluate (21), $BL_k(t) = \max\{BL_k(t-1) + d_k(t) - c_k, 0\}$, 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 §5, we compared the formula (A28), 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 appears to give much higher values of \overline{BL}_k for large capacities, but for smaller capacities the results are reasonably similar. An explanation for this is the continuous time assumption in the Pollaczek-Khintchine formula; there will frequently (and on average) be a queue just after each job arrival, but in the discrete-time system processing effectively happens “after” all the job arrivals in each period. This effect becomes less and less significant as the capacity constraint becomes smaller and smaller and the busy periods become longer and longer.

While these methods yield rather different results, the term \overline{BL}_k is typically only responsible for a small portion of all costs. Fortunately, the different methods are increasingly similar as the cost contribution from the \overline{BL}_k term grows and makes a significant contribution. Thus in the experiments we performed, the average total cost difference between the two methods was, on average, only 2.1%.

Appendix 2: Proofs and derivations

Derivation of (7). First we apply (2) recursively on $IQ_k(t-1)$, $IQ_k(t-2)$, and so on:

$$\begin{aligned} IQ(t) &= \max \{IQ(t-1) + d(t-SI) - c, 0\} \\ &= \max \{ \max \{IQ(t-2) + d(t-SI-1) - c, 0\} + d(t-SI_k) - c, 0 \} \\ &= \max \{IQ(t-2) + d(t-SI) + d(t-SI-1) - 2c, d(t-SI) - c, 0\} \\ &= \dots = \\ &= \max_{n \in \mathbb{Z}} \{d(t-SI-n, t-SI) - cn\} \end{aligned} \tag{A29}$$

Proof of Proposition 1. We define

$$\hat{\tau} = \arg \max_{\tau} \{D(\tau) - c\tau\} \quad (\text{A30})$$

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

$$\begin{aligned} B(\tau) &= (\psi D)(\tau) \\ &= \max_{n \geq 0} \{D(\tau + n) - cn\} \\ &= c\tau + \max_{n \geq 0} \{D(\tau + n) - c(\tau + n)\} \\ &= c\tau + \max_{x \geq \tau} \{D(x) - cx\} \end{aligned} \quad (\text{A31})$$

Now if $\tau \leq \hat{\tau}$, then the constraint is irrelevant and x can take the maximizing value $\hat{\tau}$ from (A30). However, if $\tau > \hat{\tau}$, then x will take the smallest value possible (because D is concave, and crosses $c\tau$ at a unique point), which is τ . Thus:

$$B(\tau) = \begin{cases} c\tau + D(\hat{\tau}) - c\hat{\tau} & \text{for } \tau \leq \hat{\tau} \\ D(\tau) & \text{for } \tau > \hat{\tau} \end{cases} \quad (\text{A32})$$

Now we are ready to prove concavity, by definition. For $\tau_1 < \tau_2 \leq \hat{\tau}$ or $\hat{\tau} < \tau_1 < \tau_2$ we must clearly have that

$$B(\lambda\tau_1 + (1-\lambda)\tau_2) \geq \lambda B(\tau_1) + (1-\lambda)B(\tau_2), \quad (\text{A33})$$

since both $c\tau + D(\hat{\tau}) - c\hat{\tau}$ and $D(\tau)$ are individually concave. Now for $\tau_1 < \hat{\tau} < \tau_2$ let us first suppose that $\lambda\tau_1 + (1-\lambda)\tau_2 \leq \hat{\tau}$. Then

$$\begin{aligned} B(\lambda\tau_1 + (1-\lambda)\tau_2) &= c(\lambda\tau_1 + (1-\lambda)\tau_2) + D(\hat{\tau}) - c\hat{\tau} \\ &= \lambda(c\tau_1 + D(\hat{\tau}) - c\hat{\tau}) + (1-\lambda)(c\tau_2 + D(\hat{\tau}) - c\hat{\tau}) \\ &= \lambda B(\tau_1) + (1-\lambda)(c\tau_2 + D(\hat{\tau}) - c\hat{\tau}) \end{aligned} \quad (\text{A34})$$

However, by definition (A30) of $\hat{\tau}$ we must have that $c\tau + D(\hat{\tau}) - c\hat{\tau} \geq D(\tau)$, and hence

$$\begin{aligned} &B(\lambda\tau_1 + (1-\lambda)\tau_2) \\ &\geq \lambda B(\tau_1) + (1-\lambda)D(\tau_2) \\ &= \lambda B(\tau_1) + (1-\lambda)B(\tau_2) \end{aligned} \quad (\text{A35})$$

On the other hand, if $\lambda\tau_1 + (1-\lambda)\tau_2 = \tau_3 > \hat{\tau}$, then

$$\begin{aligned}
B(\tau_3) &\geq \\
&= B(\hat{\tau}) + \frac{\tau_3 - \hat{\tau}}{\tau_2 - \hat{\tau}} (B(\tau_2) - B(\hat{\tau})) \\
&= B(\hat{\tau}) \left(1 - \frac{\tau_3 - \hat{\tau}}{\tau_2 - \hat{\tau}}\right) + \frac{\tau_3 - \hat{\tau}}{\tau_2 - \hat{\tau}} B(\tau_2)
\end{aligned} \tag{A36}$$

This holds because $B(\tau)$ is concave by assumption for $\tau > \hat{\tau}$. Now we note that

$$B(\hat{\tau}) = \left((c\tau_2 + D(\hat{\tau}) - c\hat{\tau}) - B(\tau_1) \right) \frac{\hat{\tau} - \tau_1}{\tau_2 - \tau_1} + B(\tau_1) \tag{A37}$$

This is simply describing $B(\hat{\tau})$ as a point on a line between $B(\tau_1)$ and $c\tau_2 + D(\hat{\tau}) - c\hat{\tau}$.

Combining (A36) and (A37) we have

$$\begin{aligned}
&B(\tau_3) \\
&\geq \left(\left((c\tau_2 + D(\hat{\tau}) - c\hat{\tau}) - B(\tau_1) \right) \frac{\hat{\tau} - \tau_1}{\tau_2 - \tau_1} + B(\tau_1) \right) \left(1 - \frac{\tau_3 - \hat{\tau}}{\tau_2 - \hat{\tau}} \right) + \frac{\tau_3 - \hat{\tau}}{\tau_2 - \hat{\tau}} B(\tau_2) \\
&\geq \left(\left(D(\tau_2) - B(\tau_1) \right) \frac{\hat{\tau} - \tau_1}{\tau_2 - \tau_1} + B(\tau_1) \right) \left(1 - \frac{\tau_3 - \hat{\tau}}{\tau_2 - \hat{\tau}} \right) + \frac{\tau_3 - \hat{\tau}}{\tau_2 - \hat{\tau}} B(\tau_2) \\
&= \left(\left(B(\tau_2) - B(\tau_1) \right) \frac{\hat{\tau} - \tau_1}{\tau_2 - \tau_1} + B(\tau_1) \right) \left(1 - \frac{\tau_3 - \hat{\tau}}{\tau_2 - \hat{\tau}} \right) + \frac{\tau_3 - \hat{\tau}}{\tau_2 - \hat{\tau}} B(\tau_2) \\
&\equiv H(\tau_3)
\end{aligned} \tag{A38}$$

The second inequality comes from noting the maximizing property of $\hat{\tau}$. On the last line we just

defined $H(\tau_3)$ as an affine function of τ_3 . Now we evaluate $H(\cdot)$ at $\hat{\tau}$ and τ_2 . This gives us

$$\begin{aligned}
H(\hat{\tau}) &= \left(\left(B(\tau_2) - B(\tau_1) \right) \frac{\hat{\tau} - \tau_1}{\tau_2 - \tau_1} + B(\tau_1) \right) \\
H(\tau_2) &= B(\tau_2)
\end{aligned} \tag{A39}$$

Exactly the same holds if we instead evaluate the function $\left(B(\tau_2) - B(\tau_1) \right) \frac{\tau_3 - \tau_1}{\tau_2 - \tau_1} + B(\tau_1)$. But

two affine functions that take the same values at two different points are identical, and so:

$$B(\tau_3) \geq H(\tau_3) = (B(\tau_2) - B(\tau_1)) \frac{\tau_3 - \tau_1}{\tau_2 - \tau_1} + B(\tau_1) \quad (\text{A40})$$

This proves that final case and the proof is complete. \square

Proof of Proposition 2. 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 (20), $d_{k+1}(t)$ can never exceed c_k , and so we must have that

$$d_{k+1}(t, t + \tau) \leq c_k \tau \leq D_k(\tau) \quad \tau \leq \tilde{\tau} \quad (\text{A41})$$

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

$$\begin{aligned} d_{k+1}(t, t + \tau) &= d_k(t, t + \tau) - BL_k(t + \tau) + BL_k(t) \\ &\leq d_k(t, t + \tau) + BL_k(t) \\ &= d_k(t, t + \tau) + (d_k(t - \tilde{n}, t) - c_k \tilde{n}) \\ &= d_k(t - \tilde{n}, t + \tau) - c_k \tilde{n} \end{aligned} \quad (\text{A42})$$

where

$$\tilde{n} \equiv \min n \geq 0 : BL_k(t - n) = 0 \quad (\text{A43})$$

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 some sufficiently low t , there must always exist such an \tilde{n} . We can replace \tilde{n} defined by (A43) with a maximizing n ; we will still have a valid (although potentially looser) bound:

$$\begin{aligned} &d_{k+1}(t, t + \tau) \\ &\leq d_k(t - \tilde{n}, t + \tau) - c_k \tilde{n} \\ &\leq \max_{n \geq 0} d_k(t - n, t + \tau) - c_k n \end{aligned} \quad (\text{A44})$$

Finally, we invoke the bound on d_k :

$$d_{k+1}(t, t + \tau) \leq \max_{n \geq 0} D_k(\tau + n) - c_k n \quad (\text{A45})$$

However, for $\tau > \tilde{\tau}$ we have, because D_k is concave and $\tilde{\tau}$ is the equality point, that (A45) is maximized for $n = 0$ and hence, for $\tau > \tilde{\tau}$, we have

$$d_{k+1}(t, t + \tau) \leq D_k(\tau) < c_k \tau \quad \tau > \tilde{\tau} \quad (\text{A46})$$

Combining (A41) and gives us the claimed relation. Finally we note that the bound (24) is tight; for example $d_{k+1}(t, t + \tau) = D_{k+1}(\tau)$ is realized if $BL_k(t - 1) = 0$ and $d_k(t, t + \tau) = D_k(\tau)$ \square

Proof of Proposition 3: We start by proving a)-c) by induction, noting that they are true by assumption for $k = 0$. 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 $k-1$, and that node k has a capacity constraint. We can then use Proposition 2 to get that $D_{k+1}(\tau) = \min(c_k \tau, D_k(\tau))$. Thus a) holds for k 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 .

Suppose now that c) holds for $k-1$, that is, any c_l ($l \geq k - 1$) 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 there is a crossing point such that

$$c_l \tilde{\tau} = D_k(\tilde{\tau}) \quad (\text{A47})$$

By the inductive assumptions a)-c), we can use Proposition 3, and so we have

$$\min(c_k \tilde{\tau}, D_k(\tilde{\tau})) = D_{k+1}(\tilde{\tau}). \quad (\text{A48})$$

Because c_l is decreasing in l , we have

$$c_l \tilde{\tau} = \min(c_k \tilde{\tau}, c_l \tilde{\tau}) \quad \forall l \geq k \quad (\text{A49})$$

Combining (A47)-(A49) gives us

$$c_l \tilde{\tau} = D_{k+1}(\tilde{\tau}) \quad \forall l \geq k \quad (\text{A50})$$

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 $c_l \tau < c_k \tau$ so $c_l \tau < \min(c_k \tau, D_k(\tau)) = D_{k+1}(\tau)$. 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 as well.

Therefore, we have shown that a)-c) for node $k-1$ imply that a)-c) hold for k as well. Since the base case $k = 1$ is true by assumption, by the induction axiom a)-c) must hold for all k . This means that the necessary assumptions for Propositions 1 and 2 are fulfilled for all k , and this proves d) and e), respectively. \square

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.
- 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.

- , 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.
- , 1995 Sensitivity analysis for base-stock levels in multiechelon production-inventory systems. *Management Science*. **41** 263-281.
- 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.
- , 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.
- 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.
- Humair, S. and S. P. Willems. 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. 2008. Optimizing strategic safety stock placement in general acyclic networks. Working Paper, August 2008, 22 pages
- Inderfurth, K. 1991. Safety stock optimization in multi-stage inventory systems. *International Journal of Production Economics*. **24** 103-113.
- Kimball, G. E. 1988. General Principles of Inventory Control. *Journal of Manufacturing and Operations Management*. **1** 119-130.

- 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. 5.
- Parker R. P., Kapuscinski R., 2004. Optimal policies for a capacitated two-echelon inventory system. *Operations Research*. **52** 739-755.
- Simpson, K. F. 1958. In-process inventories. *Operations Research*. **6** 863–873.
- Wal, J. van der, Speck, C.J. (1991). *The capacitated multi-echelon inventory system with serial structure. I. 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* . Vol. 10, No. 1, Winter, 19-23.