SCHOOL OF OPERATIONS RESEARCH AND INDUSTRIAL ENGINEERING, CORNELL UNIVERSITY, ITHACA, NY 14853

Optimal Stocking in Reparable Parts Networks with Repair Capacity and Inventory Pooling

Technical Report No. 1310

February, 2002

John A. Muckstadt*

Peter L. Jackson*

Kathryn E. Caggiano♯

James A. Rappold♯

*School of Operations Research and Industrial Engineering, Cornell University, Ithaca, N.Y. 14853
♯School of Business, University of Wisconsin, Madison, WI 53706

This research supported in part by Xelus, Inc.

ABSTRACT

This paper addresses a tactical planning problem for a three-echelon, reparable-parts service network characterized by a capacitated repair facility and local opportunities for inventory pooling. The problem is to find the optimal total system stock. The model proposed is appropriate for high-cost, high-criticality, low-demand-rate parts for which transport times are short and for which optimization-based stock allocation is performed in the distribution system. The model includes parameters that characterize design and management decisions in the resupply system. The model can be solved in time that is $nlog(n)$ in the number of part number-location combinations, making it a practical technique for large-scale inventory problems. One implication of this approach is to emphasize the importance of optimization-based stock allocation and repair priority routines in inventory management execution systems.

1. Objective

This paper addresses a tactical planning problem for a three-echelon, reparable-parts, service network characterized by a capacitated repair facility and local opportunities for inventory pooling. The central planning problem is to determine the optimal level of total system stock for each part, particularly for high-cost, high-criticality, low-demand-rate parts. Once the stock has been acquired, its location in the system, and the resulting service performance of the system, will be managed by an inventory management execution system. The more effective this execution system is at managing repair and distribution, the less total stock will be required. Consequently, it is important to capture the possibilities for dynamic optimization when planning total stock levels, particularly for high-cost parts. Because of the large number of parts in such systems, computational efficiency in performing any inventory planning function is a critical concern. The model we propose can be solved in time that is $n \log(n)$ in the number of part number-location combinations. One implication of the approach used in this paper is to emphasize the importance of optimization-based stock allocation and repair priority routines in inventory management execution systems.

2. Multi-Echelon Reparable Parts System with Central Repair

The three-echelon distribution and repair system for reparable parts we will study is depicted in Figure 1. The system consists of a set of inventory pools, each of which contains a number of stocking locations called *cribs* ; a set of regional distribution centers, each of which supplies a set of inventory pools; a national distribution center, which supplies the regional distribution centers; a capacitated repair facility, at which defective parts are repaired and, once repaired, are sent to the national distribution center; an external supplier, which provides inventory to replace parts which have been condemned; and a third-party emergency supply source. The recovery, defect-identification, repair, and replacement processes are together referred to as the *resupply system*.

The cribs, pools, and distribution centers are referred to as the *distribution system*. The third-party emergency supply source is viewed as a separate system.

The central planning decision to be made is the number of units of each item type to have in the system. The modeling approach taken in this paper is to first determine the optimal stock level for a single part for a given repair capacity allocation while considering inventory pooling opportunities. Second, we focus on the repair capacity and propose methods to allocate this capacity in a multi-item model. Although the model can also be used to set target base stocking levels at each location, that is of secondary concern because inventory allocation is anticipated to play a greater role than simple inventory replenishment.

We begin this paper by assuming that stocking locations for each part have been pre-determined and the cost of a customer shortage is known at each location. Our focus is on items which deserve a high level of management attention: in particular, high-cost, high-criticality, low-demand-rate items. We consider various aspects of the distribution system in other papers. For example, we address the question of which parts to stock in inventory cribs when there are opportunities for pooling in [9] . In a companion paper [8] , we address the question of how to set stocking requirements across multiple part numbers to meet general customer service level requirements. The methodology developed in that paper translates general customer service requirements into part-by-part, location-by-location implicit costs of customer shortage. The general multi-item problem of satisfying service level requirements can then be disaggregated into separate single-item problems.

The system operates in the following manner. Demand for parts arises randomly at inventory cribs and is satisfied out of the closest inventory crib or distribution center which has stock available. Inventory cribs within the same pool share inventory to satisfy demand, but at some cost. The inventory pools have been formed from inventory cribs that are within close proximity of each

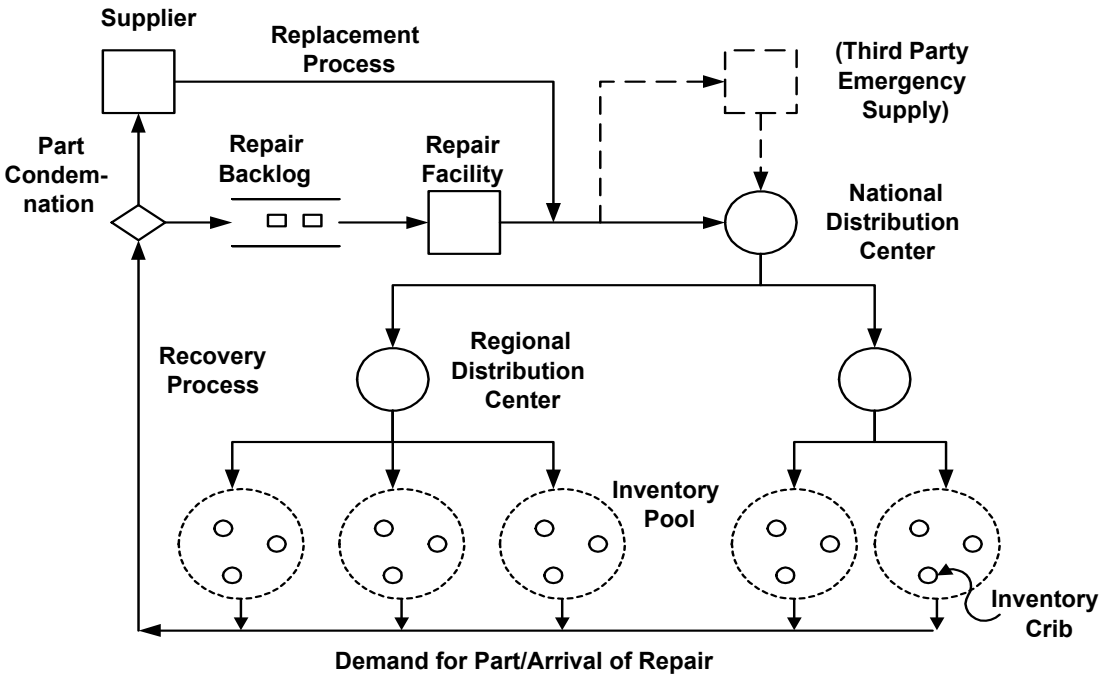

Figure 1: Multi-Echelon Distribution Network with Repair and Pooling

other (e.g., less than two hours travel time from each other). Associated with each demand for a part is a failed unit which is removed from the field. This failed unit enters a recovery process which includes defect-identification, transport, and a decision of whether to repair or replace the unit. If the decision is to repair the unit, then the unit enters the repair backlog queue where it awaits repair in the repair facility. Replacement orders for condemned units are placed with an outside supplier.

Since we assume that parts are of high criticality, backorders are very expensive in the distribution system. Specifically, we assume all demands for the parts, wherever they occur, are satisfied within one review period. As a last resort, a unit is obtained from outside this system to ensure that no backorder lasts more than one review period. In the airline industry, for example, a needed spare part may be obtained on loan from a cooperating but competitor airline. Suppliers may also hold reserve stock for such emergencies. Discussions with inventory managers of high-criticality parts reveal many creative ways in which they locate and transport units to serve the field organi-

zation. We model this process as a third-party emergency supply source. We assume that customer demands that cannot be satisfied by stock anywhere in the distribution system are satisfied with units borrowed from the third-party source and that these units are repaid one-for-one with the next physical units emerging from the resupply system. The degree to which the system is exposed to such emergency orders is controlled in the model through the cost function. We leave for further research the consideration of reciprocal emergency demands (demands on this system that might be placed by a cooperating competitor) and the optimization of industry-wide stock levels.

Once the system is operating, the units of a part that have been acquired will either be in the distribution system or they will be in the resupply system. Optimizing the total system stock level requires modeling the operating characteristics of both the distribution system and the resupply system. The model presented in this paper is a departure from previous approaches in that it treats these operational problems as dynamic optimization sub-problems. As in Cohen *et al* [12] , we assume that transport times within the distribution system are very short relative to the process and queue times experienced in the resupply system. For many high-cost, low-cubic-volume items such as electronic components, air freight is economical and distribution transport times from a central stocking facility are measured in hours. On the other hand, process and queue times through a repair or re-manufacture process are measured in weeks. Consequently, the distribution system can react to changes on a time scale that is much shorter than the resupply system. For simplicity, we assume that all emergency transport in the distribution system happens instantaneously at the end of a review period. However, the review period is not assumed to be so short as to negate any concern for positioning stock close to the customer. Reallocation of stock within a region (i.e. a regional distribution center and its associated inventory pools) is assumed to happen instantaneously at the beginning of a review period. In the application that motivated this study, for example, a repair

service provider maintains stock in the field to achieve same-day repair service but at the end of the day places replenishment orders with a regional distribution center. If the regional distribution center has the stock, the field location receives them overnight, in time to satisfy service calls on the next day. Positive resupply times are used in the model when considering the allocation of stock from the national distribution center to the regional distribution centers.

In setting total system stock levels, we further assume that the distribution system is managed dynamically to balance the distribution of available stock for best effect. That is, we assume that rather than using simple first-come, first-serve allocation policies, the real-time management system allocates available stock to optimize the tradeoff between inventory holding costs and customer shortages over the course of the review period. Similarly, we assume that the share of repair capacity allocated to an item is dynamically optimized. Such a high degree of management attention on the operation of both the distribution and resupply systems can be justified for high-cost, low-demand-rate items. This approach requires an integrated implementation of both planning and execution models, and this paper proposes an optimization-based planning model as a first step.

3. Related Literature

The management of reparable item logistics systems has received much attention over the past several decades due to the substantial economic cost of managing such systems. There are three principal bodies of research that are relevant to our work: continuous-review one-for-one replenishment models, lateral-transshipment models, and finite-production-capacity models.

In [11] , Clark and Scarf established the structure of optimal control policies and such fundamental notions as echelon reorder points and echelon holding costs. This work was later extended by Federgruen and Zipkin [21, 22] to the infinite horizon case. In another early work, Sherbrooke [51] examined one-for-one replenishment policies, or $(S-1, S)$ policies, and established the significant

practical value of developing planning models for the deployment of inventories in a multi-echelon system. This work was extended by Muckstadt [41] and Graves [28] . Muckstadt and Thomas [44] and Hausman and Erkip [30] made comparisons between managing a logistics system using multi-echelon techniques versus using a simple location-decomposition approach. The cost and service advantages of using multi-echelon techniques were shown to be substantial. See Axsäter [5] for a more detailed overview of continuous review policies, Nahmias [45] and Daniel *et al* [14] for reviews of reparable-item inventory systems research, and Federgruen [20] for a review of centralized planning models in multi-echelon systems. Extensions of this research to the case of generalized batch ordering can be found in Deuermeyer and Schwarz [16] , Lee and Moinzadeh [36, 37, 38] , Svoronos and Zipkin [52] , Axsäter [6] , and Cachon [7] .

The problem of lateral transshipments is of substantial practical importance in a variety of industrial and military applications. The problem of lateral supply is different than the problem of expediting shipments, as examined by Moinzadeh and Schmidt [39] , Aggarwal and Moinzadeh [1] , and Lawson and Porteus [34] . Das [15] and Hoadley and Heyman [31] were among the first to consider the additional complexity posed by lateral supply. Lee [35] and Axsäter [4] examined a continuous-time, two-echelon system in which lateral transshipments are permitted at some additional cost in order to avoid expected shortage costs. They tested a number of different emergency lateral transshipment rules. Cohen *at al* [12] developed a general periodic-review framework for modeling multi-echelon systems with pooling groups. While their model is quite general, it does not address the additional complexity surrounding the repair process for reparable items. We develop a different approach that requires substantially less computation. Other related lateral transshipment and inventory pooling research can be found in Tagaras [53, 54] , Tagaras and Cohen [55] , Dada [13] , Yanagi and Sasaki [58] , Tagaras and Vlachos [56] , and Grahovac and

Chakravarty [27] . The practical benefits of inventory pooling and various pooling methods are evaluated in Evers [18, 19] and Needham [46] .

More recently, Kukreja *et al* [33] developed an inventory transshipment model for a single echelon, single item, multi-location system under continuous review and Poisson demands. Using a queueing-theoretic approach, they extended the results shown in Axsäter [4] and demonstrated the benefits of a centralized approach for coordinating pooled inventory compared with a decentralized approach. Our approach differs from theirs in two ways. First, we consider a larger system with multiple echelons and a finite depot repair capacity resupplying multiple pooling groups within each echelon. Second, our model of distribution is based on a dynamic rebalancing of stock allocations.

The inventory planning models referred to thus far assume that the resupply times at the highest echelon are either deterministic or are stochastic, but independent and identically distributed. The impact of finite production capacity, or finite repair capacity, on inventory planning has been studied in Evans [17] , Tayur [57] , Güllü and Jackson [29] , Glasserman [25] , and Roundy and Muckstadt [50] . Some of these models are extensions of dam models that can be found in Prabhu [47] . The optimality of a base-stock control policy for controlling production and inventory was proven by Federgruen and Zipkin [23, 24] .

Applications of these finite capacity models to multi-echelon systems can be found in Glasserman and Tayur [26] , Aviv and Federgruen [3] , Kapuscinski and Tayur [32] , Rappold and Muckstadt [49] , and Chan *et al* [10] . In [59] , Zipkin analyses the cost impact of inventory imbalances in distribution systems. Inventory imbalance occurs when inventory investment is misallocated either across items or across locations. Rappold and Muckstadt, [49] and Chan *et al* [10] developed a lower bound approximation to the per-period expected cost. They demonstrated that inventory imbalance does not materially impact the expected cost per period when lead times are short and

allocation-balancing rules are employed. Pyke [48] also examined the impact of specific real-time allocation rules on overall system performance in the context of a reparable-item, multi-echelon system. The types of allocation rules tested include rules for allocating repair capacity and rules for allocating items to downstream locations.

Our contribution in this paper is to develop a large-scale, computationally-tractable, steady-state, tactical planning model that simultaneously considers a three-echelon system with inventory pooling, lateral transshipment, and finite repair capacity for multiple items. To our knowledge, past research has not jointly considered these three factors.

4. Linking Resupply and Distribution

Our approach is to develop a cost model based on the steady state probability distribution of the number of units in resupply for a single item. Resupply consists of three processes: recovery and transport, replacement, and repair. Assume that the repair/replace decision is made prior to the unit entering the repair queue, as illustrated in Figure 1. Let V_B denote the number of units in the repair backlog for this item in steady state, including units being repaired; let V_T denote the steady-state number of units of this item in recovery and transport within the resupply system; and let V_U denote the steady-state number of units of this item on order for replacement from the supplier. Based on the assumptions of this paper, these are independent random variables. Let V denote the total number of units in resupply, in steady state:

$$V = V_B + V_T + V_U. \tag{1}$$

The stationary probability distribution of V is thus a convolution of three stationary distributions. We assume a Poisson demand process and independently distributed replacement and recovery/transport lead times. Consequently, the steady state distribution of $V_T + V_U$ is also Poisson [**?**] . Let λ de-

note the overall demand rate for units of the part in the distribution system and let q denote the condemnation probability. Let T denote the expected recovery and transport time to the repair facility and let U denote the expected supplier lead time for replacement orders. Then the stationary distribution of $V_T + V_U$ is a Poisson distribution with rate $\lambda T + q\lambda U$. The distribution of V is the convolution of this Poisson distribution with the stationary distribution of V_B. The stationary probability distribution of V_B will depend on the repair capacity of the repair facility and the way in which this capacity is managed and allocated to units of the single item considered. It can be estimated using the techniques described in section 7.

Let $Q > 0$ denote the total planned inventory in the system for this item: both distributable units and units in resupply. The planning model focuses on determining the optimal value of this static decision variable. Let R denote the total distributable physical inventory in the system. We will be making assumptions to ensure that $R = (Q - V)^+$. Let $X = (V - Q)^+$, the number of units in resupply in excess of planned inventory. The steady state distributions of R and X can be derived from the value of Q and the stationary distribution of V. In this paper, we describe a newsvendor-style cost model based on R and X that can be optimized through the appropriate choice of Q.

The remainder of the paper is organized as follows. In section 5, we develop a model for the optimal allocation of available stock in a two-echelon distribution system with local opportunities for inventory pooling. This analysis yields a cost function which is used in section 6 to describe a model for the optimal allocation of stock in a three-echelon distribution system. This analysis, in turn, yields a cost function to describe a model for determining the optimal level of system inventory. We show how to disaggregate the optimal system inventory level into target base stock levels at all locations and we observe that the system cost function can be computed in time that is

$n \log(n)$ in the number of locations. In section 7, we review both exact and approximate techniques for determining the stationary distribution of V_B, the repair backlog, as a function of the capacity in the repair facility. We consider the priority rules that will be used in the repair facility to manage multiple items, and we develop techniques to approximate the impact of capacity management on the repair backlog of the item considered. Section 8 concludes the paper.

5. Optimal Stock Allocation for Two-Echelon Inventory Pools

Let W denote the set of two-echelon sub-systems within the overall distribution system. Each sub-system consists of a regional distribution center and a collection of inventory pools supported by this center. Let $w \in W$ index the individual sub-systems. For this section, we focus on a single sub-system and suppress the index w from all variables. In this section, we develop a single-item, single-period model for optimally allocating the physical inventory of this sub-system among its stocking locations.

5.1 Sub-System Structure, the Pooling Assumption, and Penalty Costs

Let P be the set of inventory pools served by a sub-system regional distribution center. For example, a regional distribution center in San Jose, California may serve local pools in San Jose, San Francisco, and Oakland.

Let B_p be the set of inventory cribs within pool p, $p \in P$. These cribs may be located in office buildings or institutional sites close to customer equipment installations or they may be mobile cribs, located in technician support vans, that serve fixed, non-overlapping geographical regions. Demand for service parts is tracked and forecast at the crib level.

The sub-system is supported by a real-time parts location information system from which a technician or dispatcher can identify the location of the nearest crib containing a part required to complete a customer repair. The technician/dispatcher can initiate a dedicated transfer of that part

to the customer site.

The order of events in a review period are as follows. At the beginning of the period the total stock available for distribution is reviewed and a decision is made to re-distribute this stock among the different facilities in the region. The re-distribution of stock takes place before demand is realized. Demand for stock is then realized at the various inventory cribs and transfer costs are incurred to satisfy these demands. All demands are assumed to be satisfied by the end of the review period. Four types of transfers are possible, each with different cost consequences:

1. *Nearest crib to customer site:* the cost and delay of such a transfer are unaffected by the stocking policy and are ignored in this analysis.
2. *Alternate in-pool crib to nearest crib:* let the cost of a transfer between cribs in the same pool be denoted by π^p.
3. *Alternate out-of-pool crib to nearest crib:* let the incremental cost of transfer between pools, within the same sub-system, be denoted by π^w. This is also the cost of transferring a unit from the regional distribution center to the pool. When a pool-to-pool transfer is initiated, π^w in incurred and to this cost is added the in-pool cost, π^p , of getting the part to the crib nearest to the customer site.
4. *Out-of-subsystem to nearest crib:* sometimes the required part is not in stock anywhere within the sub-system and an emergency transfer is required from another sub-system or from the national distribution center. Let the incremental cost of an emergency shipment to this subsystem be denoted by π^e. When an out-of-subsystem transfer is initiated, π^e is incurred and to this emergency shipment cost is added the cost to get the part to the pool and crib nearest the customer site: $\pi^w + \pi^p$.

Each of the three costs identified, π^p, π^w, and π^e, is assumed to include not only the incremental

transportation cost but also an imputed penalty for the incremental customer waiting time (for the use of more remote sources). Imputing shortage costs for generalized service level constraints is considered in the companion paper [8] .

5.2 Stock Allocation Decisions

Let R_b^p be the stock allocation decision for crib b, $b \in B_p$, and let R^p be the stock allocation to pool p: $R^p = \sum_{b \in B_p} R_b^p$. Let R (that is, R^w with the sub-system superscript suppressed) denote the total physical stock level available for allocation in the sub-system at the beginning of a review period. Given R, we must choose stock level allocations consistent with this total:

$$\begin{aligned} \sum_{p \in P} R^p &\leq R; \text{ and} \\ \sum_{b \in B_p} R_b^p &= R^p, \forall p \in P. \end{aligned}$$

The difference between R and $\sum_{p \in P} R^p$ is stock that is retained at the sub-system regional distribution center. Re-distribution during the review period resulting from these allocation decisions takes place before demand is realized.

5.3 The Backorder and Allocation Assumptions

Denote the net inventory in crib b at the beginning of the review period by I_b^p, for $b \in B_p$, and let $I^p = \sum_{b \in B_p} I_b^p$ be the net inventory of pool p. Lead times are so short within the sub-system that we may assume that no stock is in-transit to a crib at the beginning of a review period. We assume that there are no backorders: $I_b^p \geq 0$, for all $p \in P$ and $b \in B_p$. The cost of eliminating backorders through emergency replenishments is captured in the cost function.

A feasible allocation is one satisfying $R_b^p \geq I_b^p$, $\forall b \in B_p$; otherwise, the allocation will imply costly transhipments to redress imbalances. An execution model, one that is used for real-time allocation of system stock, cannot ignore these constraints. However, a planning model used to set

base stock policy parameters and target system inventory levels may safely ignore these constraints when review periods are short. Simulation studies have shown that when base stock policies are supplemented with intelligent allocation rules, the portion of system cost attributable to violations of this assumption is very small (Muckstadt *et al* [43, 42] and Muckstadt and Roundy [40]). Henceforth in this planning model, we ignore the current state of net inventory within and among the pools and assume that material is balanced across locations for a given amount of sub-system inventory. The state of the sub-system is therefore the total distributable inventory of the sub-system, R.

5.4 The Allocation Optimization Problem

Assume holding costs do not differ by crib within a pool. Consequently, how R^p is allocated to cribs will not affect total holding cost; however, the allocation will affect internal shortfall costs. Let D_b^p denote the random variable for demand for service parts in crib region $b \in B_p$ for one review period, a random variable. Let $C_b^p(R_b^p)$ denote the expected crib-to-crib transfer cost for crib $b \in B_p$, given an initial allocation of R_b^p to crib b:

$$C_b^p(R_b^p) \equiv \pi^p E\left[(D_b^p - R_b^p)^+\right]. \tag{2}$$

Let h^p denote the incremental holding cost of storing one unit in pool p, over the cost of holding that unit in the sub-system regional distribution center, for one review period. Holding costs are charged on inventory balances at the end of the review period. Let D^p denote the aggregate demand in pool p for one review period. Let $C^p(R^p)$ be the minimum total expected pool cost for pool p for one review period:

$$\begin{aligned} C^p(R^p) \;\equiv\; & h^p E\left[(R^p - D^p)^+\right] + \pi^w E\left[(D^p - R^p)^+\right] \\ & + \min_{\substack{s.t. \sum_{b\in B_p} R_b^p = R^p \\ R_b^p \geq 0, \text{integer}, \forall b \in B_p}} \left\{ \sum_{b \in B_p} C_b^p(R_b^p) \right\}. \end{aligned} \tag{3}$$

Two additional costs need to be considered: the cost of emergency shipments from outside the sub-system and the holding cost of reserve inventory held at the distribution center. Let D denote the random variable for aggregate demand in the sub-system for one review period. Let h^w denote the incremental holding cost of storing one unit of inventory in the sub-system for one review period, over the cost of holding it in inventory in the national distribution center for one period. It does not include the incremental cost of holding inventory in the individual pools. These costs are captured by the h^p parameters. Let $C(R)$ denote the expected cost over one review period, assuming we begin the review period with R units of distributable inventory in the sub-system and this inventory is allocated optimally. That is:

$$\begin{aligned} C(R) \quad \equiv \quad & h^w E\left[(R-D)^+\right] + \pi^e E\left[(D-R)^+\right] \\ & + \min_{\substack{s.t. \sum_{p\in P} R^p \le R \\ R^p \ge 0,\ \text{integer},\ p\in P}} \left\{ \sum_{p\in P} C^p(R^p) \right\}. \end{aligned} \tag{4}$$

We assume that the shipments required to achieve the optimal distribution of inventory within the sub-system are performed instantaneously at the beginning of the review period. The expected cost of transporting regular replenishments to any crib can be computed using Little's Law. However, this cost does not depend on the stocking policy and can be ignored. The cost of lateral transhipments to address imbalances would depend on the stocking policy but we shall ignore, for planning purposes, both the possibility and the cost of imbalance.

Observe that $C(R)$ involves a nested optimization of newsvendor-style cost functions. As shown in the appendix, this is not a difficult calculation. What we are approximating with this function is the dynamic behavior of an optimization-based inventory management execution system.

6. Optimal System Inventory

In this section, we consider the set, W, of all two-echelon sub-systems and restore the use of the

sub- or superscript $w \in W$ to index the individual sub-systems.

6.1 The Relevant Cost of Sub-System Inventory

Let R^w denote the physical inventory in sub-system w at the beginning of a review period. This include inventory at each location within the sub-system plus any inventory in transit to the regional distribution center plus any stock that is being allocated for shipment into this sub-system from the national distribution system. Let L_w denote the lead time to ship units from the national distribution system to regional distribution center w and let $D^w_{L_w}$ denote the random demand that occurs over this lead time. Thus, $\left(R^w - D^w_{L_w}\right)^+$ is the physical inventory available to sub-system w at the beginning of the review period that follows the transport lead time, including the current allocation. The relevant cost for the allocation decision is:

$$C^w_{L_w}\left(R^w\right) \equiv E\left[C_w\left(\left(R^w - D^w_{L_w}\right)^+\right)\right], \tag{5}$$

the expected one-review period cost for sub-system w based on the distributable inventory that will be in the sub-system after L_w periods. This expression is correct if we assume that any inventory that is shipped by emergency into this sub-system from some other location during the lead time L_w is withdrawn from the new allocation, R^w, and sent to the location that provided the stock in the emergency. We assume this withdrawal happens instantaneously.

6.2 The Distributable System Allocation Problem

The variable R^0 denotes the total distributable inventory of the system: it must equal the sum of the inventory levels in each sub-system, after allocation, plus inventory in transit to the regional distribution centers plus inventory that is retained at the national distribution center. As in the lower echelon models, we continue to assume that this allocation can be made without regard to the possibility of inventory imbalance in the different sub-systems, including units in transit to the sub-systems.

Let h^0 denote the base cost of having one unit of inventory in the distribution system for one review period. Let π^0 denote the cost of emergency satisfaction of a backorder in the system. We assume that backorders for the item will not be tolerated and that emergency replenishment, from an outside source, such as a competitor, will be used to satisfy backorders within one review period. Let D^0 denote the total system demand that occurs over the next review period, a random variable. Let $C^0\left(R^0\right)$ denote the expected system cost over one review period, assuming we begin the review period with R units of total distributable inventory and this inventory is allocated optimally among the sub-systems and national distribution center. That is:

$$\begin{aligned} C^0(R^0) \;\equiv\; & h^0 E\left[\left(R^0 - D^0\right)^+\right] + \pi^0 E\left[\left(D^0 - R^0\right)^+\right] \\ & + \min_{\substack{\sum_{w\in W} R^w \le R^0 \\ R^w \ge 0,\ \text{integer},\ w\in W}} \left\{ \sum_{w\in W} C^w_{L_w}\left(R^w\right) \right\}. \end{aligned} \tag{6}$$

6.3 The Single-Period System-Wide Cost Function

We have assumed that system backorders, $\left(D^0 - R^0\right)^+$, are satisfied immediately from an outside source, such as a competitor. We further assume that these units are on loan and are repaid with the first available units that complete the resupply process. This assumption ensures that total distributable physical inventory consists only of planned inventory in excess of units in resupply: $R^0 = \left(Q - V\right)^+$. It follows that $X \equiv \left(V - Q\right)^+$ represents the outstanding part-loans in the process of resupply.

Each unit-loan was charged π^0 when it was first incurred, as captured in (6). Let $\overline{\pi}$ denote the per-review-period loan cost, charged for each review-period that the unit-loan is outstanding. The single-review-period, system-wide cost is given by $C^0(R^0) + \overline{\pi}X$. We ignore the inventory holding cost of units in repair, replacement, and transit in this model. Since the inventory policy cannot affect the resupply process, this holding cost is irrelevant to determining the economically optimal

value of Q.

The steady state distribution of V can be derived as the convolution of a Poisson process and the steady state distribution of V_B, the repair backlog. Let A be a measure of the slack capacity (per-period capacity less expected per-period demand for capacity) of the repair system allocated to the item under consideration. The stationary probability distribution for V_B will be parameterized by A using techniques presented in section 7. Let $E_A[\cdot]$ denotes the expectation operation with respect to the steady-state probability distribution of V, parameterized by A , the slack capacity of the repair facility.

Let $G_A(Q)$ denote the expected steady state one-review period system-wide cost, given a slack capacity allocation, A, and total system stock level, Q:

$$G_A(Q) \equiv E_A\left[C^0\left((Q-V)^+\right) + \overline{\pi}\,(V-Q)^+\right]. \tag{7}$$

The cost function $G_A(Q)$ is the central planning model of this paper. It is a single-period, newsvendor-style objective that captures tradeoffs among inventory holding costs, shortage costs, and emergency transport costs in a dynamically-optimized, three-echelon distribution system with pooling. Furthermore, through the stationary distribution of V, this function is sensitive to design and management parameters of the resupply system.

We are now in a position to describe the optimal system inventory. Let $Q^*(A)$ denote the total system inventory level that minimizes this cost, as a function of the slack capacity of the repair system:

$$Q^*(A) \equiv \arg\min_{Q\geq 0} G_A(Q).$$

Assuming that V is stochastically decreasing in A, as will be easily seen, then $Q^*(\cdot)$ is a nonincreasing function of A. That is, as the repair facility is less highly utilized (the larger the value of A), the optimal system inventory level will not increase. The dependence of system-wide cost on

A, the repair slack capacity, is explored in section 7.3.

6.4 Disaggregating System Inventory Targets

The model developed to this point provides an approach for determining optimal system-wide inventory, Q^*, for a single part, assuming a level, A, of slack capacity allocated to that part. In practice, it will be desirable to specify target base-stock inventory levels for each location in the system. This is easily done using the allocation tools already developed. For example, denote the target base stock level for sub-system w by R^{w*} and let $(R^{w*})_{w\in W}$ solve

$$\min_{\substack{s.t.\ \sum_{w\in W} R^w \le Q^* \\ R^w \ge 0,\ \text{integer},\ w\in W}} \sum_{w\in W} E\left[C^w_{L_w}(R^w)\right].$$

That is, determine the target base stock levels assuming no units are in repair. The residual, $Q^* - \sum_{w\in W} R^{w*}$, is the target inventory to be held in reserve at the national distribution center. Similarly, for each sub-system w, set $R^*=\left(R^{w*} - E\left[D^w_{L_w}\right]\right)^+$ and let the target base stock levels for the pools, $(R^{p*})_{p\in P^w}$, solve

$$\min_{\substack{s.t.\ \sum_{p\in P} R^p \le R^* \\ R^p \ge 0,\ \text{integer},\ p\in P}} \left\{\sum_{p\in P} C^p(R^p)\right\}.$$

Finally, let the target base stock levels for the cribs, $(R^{p*}_b)_{b\in B_p}$, solve

$$\min_{\substack{s.t.\ \sum_{b\in B_p} R^p_b \le R^{p*} \\ R^p_b \ge 0,\ \text{integer},\ b\in B_p}} \sum_{b\in B_p} C^p_b(R^p_b).$$

Observe that each of these problems has a convex objective function and hence can be solved using a marginal analysis algorithm (Appendix A).

6.5 When is the System in Allocation Mode?

A system or sub-system is said to be in *allocation mode* if the corresponding distribution center has insufficient stock available for shipment to raise each sub-location to its target base stock level. In allocation mode, the distribution center should be responding to requests for stock in a way that optimally balances the available inventory. Because all target base stock levels were computed

under the optimistic assumption that no units are in repair, it will likely be the case that the system and all the sub-systems will be in allocation mode nearly all the time. It is important, therefore, to use this model only with execution systems that perform well in allocation mode.

6.6 Computational Complexity

In this section, we establish a bound on the computational complexity of finding a near-optimal value of Q, the total system stock.

The algorithmic approach is to develop piecewise linear approximations to each cost function. Let $\widetilde{C}_b^p(\cdot)$, $\widetilde{C}^p(\cdot)$, $\widetilde{C}^w(\cdot)$, $\widetilde{C}_{L_w}^w(\cdot)$, $\widetilde{C}^0(\cdot)$, and $\widetilde{G}_A(\cdot)$ denote the piecewise linear approximations to $C_b^p(\cdot)$, $C^p(\cdot)$, $C^w(\cdot)$, $C_{L_w}^w(\cdot)$, $C^0(\cdot)$, and $G_A(\cdot)$, respectively, for $b \in B_p$, $p \in P_w$, and $w \in W$. Let

$$r_b^p = \left\{ r_{b0}^p, r_{b1}^p, \ldots, r_{bn(p,b)}^p \right\}$$

denote the grid for the breakpoints of $\widetilde{C}_b^p(\cdot)$, where $n(p,b)$ denotes the number of points, less one, in the grid. We require $r_{b0}^p = 0$ and $r_{bn}^p > r_{bn-1}^p$ for $n = 1, 2, \ldots, n(p,b)$. Let

$$c_b^p = \left\{ c_{b0}^p, c_{b1}^p, \ldots, c_{bn(p,b)}^p \right\}$$

denote the breakpoints of $\widetilde{C}_b^p(\cdot)$: i.e., $c_{bn}^p = \widetilde{C}_b^p(r_{bn}^p)$ for $n = 1, 2, \ldots, n(p,b)$. Similarly, define pairs of vectors (r^p, c^p), (r^w, c^w), $\left(r_{L_w}^w, c_{L_w}^w\right)$, (r^0, c^0), and $\left(r^A, c^A\right)$ to denote the grids and breakpoints of $\widetilde{C}^p(\cdot)$, $\widetilde{C}^w(\cdot)$, $\widetilde{C}_{L_w}^w(\cdot)$, $\widetilde{C}^0(\cdot)$, and $\widetilde{G}_A(\cdot)$, respectively. Let $n_w(p)$, $n_w(0)$, $n_L(w)$, $n_L(0)$, $n_A(0)$, respectively, denote the number of points in each respective grid, less the origin. Let $\overline{n}$ denote an upper bound on the number of grid points in any of these approximations.

The piecewise linear approximations are computed by solving equations (2), (3), (4), (5), (6), and (7) using previously computed piecewise linear approximations to cost functions on the right hand side of these equations wherever appropriate. We assume constant time algorithms exist to compute the probability distributions and expectations required in each equation. Let $\overline{M}_B$ denote

an upper bound on the number of locations that must be considered in any of the pooling allocation optimizations (3):

$$\overline{M}_B = \max_{w \in W} \max_{p \in P_w} |B_p| .$$

Similarly, let $\overline{M}_P$ denote an upper bound on the number of locations that must be considered in any of the sub-system allocation optimizations (4):

$$\overline{M}_W = \max_{w \in W} |P^w| .$$

Let $\overline{M} = |W|$, the number of sub-systems that must be considered in the system-wide optimization (6). Let $\overline{M}$ denote an upper bound on the number of locations that must be considered in any of the optimizations:

$$\overline{M} \equiv \max \left\{ \overline{M}_B, \overline{M}_P, \overline{M}_W \right\} .$$

Let $\overline{N}$ denote the total number of locations to consider:

$$\overline{N} = 1 + |W| + \sum_{w \in W} |P_w| + \sum_{w \in W} \sum_{p \in P_w} |B_p| .$$

Proposition 1 *Assuming constant time algorithms exist for computing the probability distributions required, the number of calculations required to compute $\widetilde{G}_A(\cdot)$ is $O\left(\left(1 + \frac{3}{4}\log_2\left(\overline{M}\right)\right)\overline{N}\overline{n}\right)$.*

Proof. $O\left(\overline{N}\overline{n}\right)$ is a simple bound on the number of calculations to evaluate all the gridpoints, excluding optimizations. By Proposition 2 in the appendix, the number of calculations to perform the optimization in (6) is $O\left(\left(1 + \log_2\left(\overline{M}_W\right)\right)\overline{M}_W\overline{n}\right)$. Similarly, the number of calculations to perform each optimization of the form (4) is at most $O\left(\left(1 + \log_2\left(\overline{M}_P\right)\right)\overline{M}_P\overline{n}\right)$. There are $\overline{M}_W$ optimizations of that form. Likewise, the number of calculations to perform each optimization of the form (3) is at most:

$$O\left(\left(1 + \log_2\left(\overline{M}_B\right)\right)\overline{M}_B\overline{n}\right),$$

and there are at most $\overline{M}_W\ \overline{M}_P$ optimizations of that form. Assembling these facts, we have that

the number of calculations required to compute $\widetilde{G}_A(\cdot)$ is:

$$\begin{aligned} & O\left(\begin{array}{c} \overline{N}\overline{n} + \left(1+\log_2\left(\overline{M}_W\right)\right)\overline{M}_W\overline{n} + \overline{M}_W\left(1+\log_2\left(\overline{M}_P\right)\right)\overline{M}_P\overline{n} \\ +\overline{M}_W\overline{M}_P\left(1+\log_2\left(\overline{M}_B\right)\right)\overline{M}_B\overline{n} \end{array}\right) \\ \leq \;& O\left(\overline{N}\overline{n} + 3\overline{M}_W\overline{M}_P\overline{M}_B\left(1+\log_2\left(\overline{M}\right)\right)\overline{n}\right). \end{aligned}$$

Noting that $\overline{M}_W\overline{M}_P\overline{M}_B$ is of the same order of magnitude as $\overline{N}$, the result follows. ■

Remark 1 *Under the further assumption that $\overline{M} = \overline{M}_W = \overline{M}_P = \overline{M}_B$ and that $\overline{M} \simeq \overline{N}^{1/3}$, then the bound on the number of calculations is*

$$O\left(\left(1+\frac{1}{4}\log_2\left(\overline{N}\right)\right)\overline{N}\overline{n}\right).$$

Thus, the optimization of total system inventory for a single item can be performed in time that is $n\log(n)$ in the number of locations.

7. Stationary Distribution of Units in Repair

By assumption, units arrive to the repair queue according to a Poisson process. The arrival rate of units to repair is given by $\mu = (1-q)\lambda$. The number of units in repair depends only on the repair arrival process and the capacity (and capacity management) of the repair system. Let A be a measure of the slack capacity (per-period capacity less expected per-period demand for capacity) of the repair system. The stationary probability distribution for the repair backlog, V_B, will be parameterized by A. Deriving this distribution is the subject of this section. We review both exact and approximate methods for determining this distribution assuming there is only a single product item in the repair system. Under this assumption, there is no capacity management issue. In section 7.3, we consider the management of multiple product items in the repair system and explore alternative ways of disaggregating the distribution of the total number of units backlogged in repair.

7.1 Exact Analysis of Repair Backlog

Let $t = 0, 1, 2, \ldots,$ index time periods in the repair facility. Let K denote the repair capacity

available in each time period, expressed as the number of units of a single item that can be repaired in any period. Let D_t denote the number of units arriving for repair in period t, a random variable. Let $V_{B,t}$ denote the backlog of unrepaired units at the beginning of period t. The repair backlog depends on the initial backlog, $V_{B,0}$, and the history of demand:

$$V_{B,t+1} = (V_{B,t} + D_t - K)^+ , \ t = 0, 1, 2, \ldots .$$

Assume $D_t = \mu + \varepsilon_t$ where ε_t is a mean-zero noise term. Let $A = K - \mu > 0$, the per-period expected excess capacity in the repair facility. Then,

$$V_{B,t+1} = (V_{B,t} + \varepsilon_t - A)^+ , \ t = 0, 1, 2, \ldots .$$

It is clear that if the repair arrival noise process, $\{\varepsilon_t\}$, consists of independent and identically distributed random variables with a known probability distribution, then the repair backlog process, $\{V_{B,t}\}$, can be modeled as an infinite state space Markov chain. Let $p_{ij} = \mathcal{P}\{V_{B,t+1} = j \,|V_{B,t} = i\}$ describe the probability of transition from backlog state i of the chain to backlog state j, for $i, j \in \{0, 1, 2, \ldots\}$. These probabilities are easily derived:

$$p_{ij} = \begin{cases} \mathcal{P}\{\varepsilon_t = j - i + A\}, & \text{for } j > 0; \\ \mathcal{P}\{\varepsilon_t \leq A - i\}, & \text{for } j = 0, \ i \leq A; \text{and} \\ 0, & \text{for } j = 0, \ i > A. \end{cases}$$

Let $P_A = (p_{ij})$ be the matrix of transition probabilities, parameterized by A. Since we assume that $A > 0$, the chain is ergodic, and a stationary distribution for the repair backlog, V_B, exists. Denote the (exact) stationary distribution of V_B by $\nu_A = (\nu_{Ai})_{i \in \{0,1,2,\ldots\}}$. This distribution can be found by solving the (infinite) system of linear equations:

$$\begin{cases} \nu P = \nu \\ e\nu = 1 \end{cases}$$

where e is an infinite vector of 1's. For practical purposes, the matrices P and e can be truncated to yield a finite system of equations without sacrificing accuracy. For low repair arrival rates, the dimension of the truncated matrix can be kept small and the resulting computational burden of

solving the system is not high. See Chan *et al* [10] , Muckstadt *et al* [43] , and Rappold and Muckstadt [49] for details.

7.2 Approximate Analysis of Repair Backlog

When the repair arrival rates are high, however, the computational burden of computing the stationary distribution of V_B will be high when using the exact Markov chain analysis. Accordingly, we review a method to approximate the stationary distribution of repair backlog when repair arrival rates are high. A complete discussion of this analysis can be found in Glasserman [25] , and Muckstadt and Roundy [40] .

Let p_0 approximate $\mathcal{P}\{V_B = 0\}$, the probability of no backlog and let $\overline{p}_0 = 1 - p_0$. Let $F_{V_B}(\cdot)$ denote the approximate c.d.f. of V defined by its complement, $\overline{F}_{V_B}(\cdot)$, as follows:

$$\overline{F}_{V_B}(v) = \begin{cases} \overline{p}_0 e^{-\gamma v}, & v \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

This distribution has a mass of p_0 at zero ($F_{V_B}(0) = p_0$) and an exponential tail with decay rate γ. The parameters of the distribution, p_0 and γ, depend on two things: the expected slack repair capacity, A, and the probability distribution of the repair arrival noise term, $\varepsilon = \varepsilon_0$.

To derive the parameters, we begin by observing the following fundamental relation:

$$\begin{aligned} \mathcal{P}\{V_{t+1} > v\} &= \mathcal{P}\{V_t + \varepsilon_t - A > v\} \\ &= \mathcal{P}\{\varepsilon_t > v + A\} + E\left[1_{\{\varepsilon_t \leq v+A\}} \mathcal{P}\{V_t > v + A - \varepsilon_t \mid \varepsilon_t\}\right]. \end{aligned}$$

This relation must also hold in steady state:

$$\mathcal{P}\{V > v\} = \mathcal{P}\{\varepsilon > v + A\} + E\left[1_{\{\varepsilon \leq v+A\}} \mathcal{P}\{V > v + A - \varepsilon \mid \varepsilon\}\right].$$

The parameters of the approximate distribution, p_0 and γ, must therefore satisfy the following relation:

$$\overline{p}_0 e^{-\gamma v} = \mathcal{P}\{\varepsilon > v + A\} + E\left[1_{\{\varepsilon \leq v+A\}} \overline{p}_0 e^{-\gamma(v+A-\varepsilon)}\right].$$

Simplifying, we have:

$$\overline{p}_0 = e^{\gamma v}\mathcal{P}\left\{\varepsilon > v + A\right\} + \overline{p}_0 e^{-\gamma A} E\left[1_{\{\varepsilon \leq v+A\}} e^{\gamma \varepsilon}\right]. \tag{8}$$

Letting $v \to \infty$ in (8) yields:

$$\overline{p}_0 = \lim_{v\to\infty}\left\{e^{\gamma v}\mathcal{P}\left\{\varepsilon > v + A\right\}\right\} + \overline{p}_0 e^{-\gamma A} \lim_{v\to\infty}\left\{E\left[1_{\{\varepsilon \leq v+A\}} e^{\gamma \varepsilon}\right]\right\}.$$

The first term to the right of the equality sign must be zero for a stable solution to exist. The other term is simply $\overline{p}_0 e^{-\gamma A} E\left[e^{\gamma \varepsilon}\right]$. Hence, the parameter γ must solve the following:

$$e^{\gamma A} = E\left[e^{\gamma \varepsilon}\right]. \tag{9}$$

Letting $v = 0$ in (8) yields:

$$\overline{p}_0 = \mathcal{P}\left\{\varepsilon > A\right\} + \overline{p}_0 e^{-\gamma A} E\left[1_{\{\varepsilon \leq A\}} e^{\gamma \varepsilon}\right].$$

Hence, the parameter p_0 is given by:

$$p_0 = 1 - \frac{\mathcal{P}\left\{\varepsilon > A\right\}}{e^{-\gamma A} E\left[1_{\{\varepsilon \leq A\}} e^{\gamma \varepsilon}\right]}. \tag{10}$$

The parameters of the approximate stationary distribution of repair backlog are given by the solution to equations (9) and (10). For normally distributed repair arrivals, these equations can be solved in closed form. In particular, if the repair arrival noise term, ε, is normally distributed with mean zero and variance σ^2, then $p_0 = 0$ and $\gamma = \frac{2A}{\sigma^2}$. That is, for normally distributed repair arrivals, the approximate stationary distribution of repair backlog is an exponential distribution with rate $\frac{2A}{\sigma^2}$.

This concludes the description of techniques to determine the stationary distribution of the repair backlog for the single item case. The multiple item case is treated next.

7.3 Disaggregating the Repair Backlog Distribution

In this section we explore alternative ways of disaggregating the backlog distribution when there are multiple product items in the repair system. The disaggregation requires some assumption of the way in which priorities are set in the repair facility. We formulate both heuristic and optimization-

based approaches for setting those priorities.

The total system cost function, $G_A(Q)$, is now well defined given A, the slack capacity in the repair facility available for the item under consideration. In practice, there will be many items competing for repair capacity and this total capacity will be allocated dynamically. The model we have developed highlights the fact that the way in which this capacity is allocated affects the system cost for the individual items. Consequently, the capacity allocation mechanism can have a significant impact on the stocking requirements across all items. It has not been traditional to consider these issues together: determining stocking requirements is a planning activity whereas capacity allocation is a dynamic execution, or dispatching, activity.

Let N denote the set of all items repaired in this facility. Let $G_A^n(Q)$ denote the expected steady state system-wide cost for item $n \in N$, parameterized by the slack capacity available if the facility were operated as a single item repair facility. Let $Q_n^*(A)$ denote the total system inventory of item n that minimizes this cost. If we think of allocating the total slack capacity of the facility to the different items, then we could compute optimal stock levels for each part individually, assuming that each part has its own dedicated repair facility with capacity equal to the expected per period arrival rate for that part (μ_n), plus the slack capacity allocated to that part. Let A_n denote the slack capacity allocated to item $n \in N$. In a planning sense, then, we would allocate capacity to minimize total system cost:

$$\min_{\substack{A_n \geq 0 \\ n \in N}} \sum_{n \in N} G_{A_n}^n \left(Q_n^*(A_n)\right)$$
$$s.t. \sum_{n \in N} A_n = \left\lfloor K - \sum_{n \in N} \mu_n \right\rfloor.$$

We propose a simple one-pass marginal analysis algorithm to solve this problem, since it is a convex program. If repair rates differ significantly between items, this model would have to be modified to measure capacity in units of time rather than in units of product.

In other papers, we have argued that capacity-constrained facilities should be operated in a way that gives production priority to items which have unpredictable demands [43] . By keeping lead times for these items short, predictable and repeatable, the safety stock in these items can be kept to a minimum. Items with more predictable demand may face longer lead times but the investment in safety stock in these items is less than in others because of their predictability. The capacity allocation optimization concept described above should yield a result consistent with this policy: all things being equal, we would anticipate that slack capacity would be allocated to items in proportion to the variability of demand. This assumes that all items in repair have similar cost characteristics.

A simple rule based on this observation can be used in place of the optimization. Let $\widetilde{\sigma}_n$ represent the standard deviation of the repair time per period required for item n and let $\widetilde{\mu}_n$ represent the expected repair time per period required. Then, $\widetilde{\sigma}_n/\widetilde{\mu}_n$ is the coefficient of variation of repair time requirements per period for item n, a measure of the demand unpredictability for the item. One way to make the allocation, therefore, is to set

$$A_n = \frac{\widetilde{\sigma}_n/\widetilde{\mu}_n}{\sum_{n' \in N} \widetilde{\sigma}_{n'}/\widetilde{\mu}_{n'}} A.$$

This allocation could be adjusted to reflect unit costs too, by biasing the allocation of capacity to expensive items. Note that the higher the value of A_n, the lower the stock level will be for part n. Our goal is to store repair capacity in items for which the coefficient of variation of demand for repair time capacity is low. Thus, we prefer to stock items for which demand is more predictable. These are the items with low values of A_n. Items with higher coefficients of variation are less predictable; allocation of slack capacity to these items is higher in the tactical planning model.

We have proposed two ways of allocating slack repair capacity to individual items in a repair facility for the purpose of estimating optimal stocking levels of these items. One approach is optimization-based; the other is a rule, based on the anticipated form of the optimal solution.

Other approaches are possible; but, the important point is that the priority rules used for operating the repair facility should be designed in conjunction with setting the system-wide stock levels of the items that use the facility.

8. Conclusion

We have presented a model for determining optimal total system stock for a multi-echelon distribution system with reparable parts and opportunities for local pooling. This model explicitly captures the impact, in steady state, of the capacity of the repair process and the rules by which that capacity is allocated to individual products. It also explicitly considers optimized dynamic rebalancing of stock in the distribution system. Because of this, the model will likely recommend lower stock levels than models with more conservative assumptions. It will therefore be particularly valuable in environments of high-cost, low-demand-rate items, where operational management is more likely to be optimization-based. The model can be solved in time that is $n \log(n)$ in the number of part number-location combinations and is therefore a practical approach to modelling and solving large scale problems. We have also shown how to disaggregate the optimal system inventory level into target base stock levels at all locations. However, we observe that the system and its subsystems will nearly always be in allocation mode so we emphasize the importance of optimization-based stock allocation routines in execution systems.

References

[1] P. K. Aggarwal and K. Moinzadeh. Order expedition in multi-echelon production / distribution systems. *IIE Transactions*, 26:86–96, 1994.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *Data Structures and Algorithms*. Addison-Wesley, Reading, MA, 1983.

[3] Y. Aviv and A. Federgruen. Capacitated multi-item inventory systems with random and seasonally fluctuating demands: Implications for postponement strategies. *Management Science*, 47(4):512–531, 2001.

[4] S. Axsäter. Simple solution procedures for a class of two-echelon inventory problems. *Operations Research*, 38:64–69, 1990.

[5] S. Axsäter. Continuous review policies for multi-level inventory systems with stochastic demand. In S. C. Graves, A. H. G. Rinooy Kan, and P. H. Zipkin, editors, *Logistics of Production and Inventory*, volume 4, pages 175–197. North Holland, Amsterdam, 1993.

[6] S. Axsäter. Exact and approximate evaluation of batch-ordering policies for two-level inventory systems. *Operations Research*, 41:777–785, 1993.

[7] G. P. Cachon. Exact evaluation of batch-ordering inventory policies in two-echelon supply chains with periodic review. *Operations Research*, 49(1):79–98, 2001.

[8] K. E. Caggiano, P. L. Jackson, J. A. Muckstadt, and J. A. Rappold. A multi-echelon, multi-item inventory model for service parts management with generalized service level constraints. *Technical Report, School of OR&IE, Cornell University, Ithaca, NY 14853*, 2001.

[9] K. E. Caggiano, P. L. Jackson, J. A. Muckstadt, and J. A. Rappold. A simple algorithm for part stocking to satisfy pooled customer service requirements at minimum cost. *Technical Report, School of OR&IE, Cornell University, Ithaca, NY 14853*, 2001.

[10] E. W. Chan, J. A. Rappold, and J. A. Muckstadt. Determining and allocating capacity-driven safety stock in multi-item, multi-echelon systems. *School of Business, University of Wisconsin, Madison, WI*, 1999.

[11] A. J. Clark and H. E. Scarf. Optimal policies for a multi-echelon inventory problem. *Management Science*, 6(4):475–490, 1960.

[12] M. Cohen, P. Kleindorfer, and H. L. Lee. Optimal stocking policies for low usage items in multi-echelon inventory systems. *Naval Research Logistics Quarterly*, 33:17–38, 1986.

[13] M. Dada. A two-echelon inventory system with priority shipments. *Management Science*, 38(8):1140–1154, 1992.

[14] V. Daniel, R. Guide, and R. Srivastava. Repairable inventory theory: Models and applications. *European Journal of Operational Research*, 102:1–20, 1997.

[15] C. Das. Supply and redistribution rules for two-location inventory systems: One period analysis. *Management Science*, 21:765–776, 1975.

[16] B. L. Deuermeyer and L. B. Schwarz. A model for the analysis of system service level in warehouse-retailer distribution systems: The identical retailer case. In L. B. Schwarz, editor, *Multi-Level Production / Inventory Systems: Theory and Practice*, pages 163–194. North Holland, New York, 1981.

[17] R. Evans. Inventory control of a multi-product system with a limited production resource. *Naval Research Logistics Quarterly*, 14(2):173–184, 1967.

[18] P. T. Evers. Hidden benefits of emergency transshipments. *Journal of Business Logistics*, 18(2):55–77, 1997.

[19] P. T. Evers. Filling customer orders from multiple locations: A comparison of pooling methods. *Journal of Business Logistics*, 20(1):121–140, 1999.

[20] A. Federgruen. Centralized planning models for multi-echelon inventory systems under uncer-

tainty. In S. C. Graves, A. H. G. Rinooy Kan, and P. H. Zipkin, editors, *Logistics of Production and Inventory*, volume 4, pages 133–173. North Holland, Amsterdam, 1993.

[21] A. Federgruen and P. Zipkin. Approximations of dynamic multilocation production and inventory problems. *Management Science*, 30:69–84, 1984.

[22] A. Federgruen and P. Zipkin. Computational issues in an infinite horizon multi-echelon inventory model. *Operations Research*, 32:818–836, 1984.

[23] A. Federgruen and P. Zipkin. An inventory model with limited production capacity and uncertain demands i. the average-cost criterion. *Mathematics of Operations Research*, 11(2):193–207, 1986.

[24] A. Federgruen and P. Zipkin. An inventory model with limited production capacity and uncertain demands ii. the discounted-cost criterion. *Mathematics of Operations Research*, 11(2):208–215, 1986.

[25] P. Glasserman. Bounds and asymptotics for planning critical safety stocks. *Operations Research*, 45(2), 1997.

[26] P. Glasserman and S. Tayur. A simple approximation for a multistage capacitated production- inventory system. *Naval Research Logistics*, 43(1):41–58, 1996.

[27] J. Grahovac and A. Chakravarty. Sharing and lateral transshipment of inventory in a supply chain with expensive low-demand items. *Management Science*, 47(4):579–594, 2001.

[28] S. C. Graves. A multi-echelon inventory model for a repairable item with one-for-one replenishment. *Management Science*, 31(10):1247–1256, 1985.

[29] R. Güllü and P. L. Jackson. On the continuous time capacitated production / inventory problem with no setup costs. *Technical Report No. 1054, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY*, 1993.

[30] W. Hausman and N. Erkip. Multi-echelon vs. single-echelon inventory control policies for low-demand items. *Management Science*, 40(5):597–602, 1994.

[31] B. Hoadley and D. P. Heyman. A two-echelon inventory model with purchases, dispositions, shipments, returns, and transshipments. *Naval Research Logistics Quarterly*, 24:1–19, 1977.

[32] R. Kapuscinski and S. Tayur. A capacitated production-inventory model with periodic demand. *Operations Research*, 46(6):899–911, 1998.

[33] A. Kukreja, C. P. Schmidt, and D. M. Miller. Stocking decisions for low-usage items in a multilocation inventory system. *Management Science*, 47(10):1371–1383, 2001.

[34] D. G. Lawson and E. L. Porteus. Multistage inventory management with expediting. *Operations Research*, 48(6):878–893, 2000.

[35] H. L. Lee. A multi-echelon inventory model for repairable items with emergency lateral transshipments. *Management Science*, 33(10):1302–1316, 1987.

[36] H. L. Lee and K. Moinzadeh. Operating characteristics of a two-echelon inventory system for repairable and consumable items under batch ordering and shipment policy. *Naval Research Logistics*, 34:365–380, 1987.

[37] H. L. Lee and K. Moinzadeh. Two-parameter approximations for multi-echelon repairable inventory models with batch ordering policy. *IIE Transactions*, 19:140–149, 1987.

[38] K. Moinzadeh and H. L. Lee. Optimal batch size and stocking levels in multi-echelon repairable systems. *Management Science*, 32:1567–1581, 1986.

[39] K. Moinzadeh and C. P. Schmidt. An $(s-1, s)$ inventory system with emergency orders. *Operations Research*, 39:308–321, 1991.

[40] J. Muckstadt and R. Roundy. Multi-item, one-warehouse, multi-retailer distribution systems. *Management Science*, 33(12):1613–1621, 1987.

[41] J. A. Muckstadt. A model for a multi-item, multi-echelon, multi-indenture inventory system. *Management Science*, 20:472–481, 1973.

[42] J. A. Muckstadt, D. H. Murray, and J. A. Rappold. Base stock levels in capacitated multi-item, multi-echelon systems with stochastic, non-stationary, cyclic demand. *Working paper, University of Wisconsin, Madison, WI, 53706*, 2001.

[43] J. A. Muckstadt, D. H. Murray, and J. A. Rappold. Capacitated production planning and inventory control when demand is unpredictable for most items: The no b/c strategy. *Working paper, University of Wisconsin, Madison, WI, 53706*, 2001.

[44] J. A. Muckstadt and L. J. Thomas. Are multi-echelon inventory methods worth implementing in systems with low-demand rates? *Management Science*, 26(5):483–494, 1980.

[45] S. Nahmias. Managing reparable item inventory systems: A review. In L. B. Schwarz, editor, *Multi-Level Production / Inventory Systems: Theory and Practice*, pages 253–278. North Holland, New York, 1981.

[46] P. M. Needham. The influence of individual cost factors on the use of emergency transshipments. *Transportation Research. Part E, Logistics and Transportation Review*, 34:149–161, 1998.

[47] U. Prabhu. *Stochastic Storage Systems: Queues, Insurance Risk, and Dams*. Springer, New York, 1980.

[48] D. F. Pyke. Priority repair and dispatch policies for repairable-item logisitics systems. *Naval Research Logistics*, 37:1–30, 1990.

[49] J. A. Rappold and J. A. Muckstadt. A computationally efficient approach for determining safety stock levels in a capacitated multi-echelon production-distribution system. *Naval Research Logistics*, 47(5):377–398, 2000.

[50] R. O. Roundy and J. A. Muckstadt. Heuristic computation of period-review base stock policies. *Management Science*, 46(1):104–109, 2000.

[51] C. C. Sherbrooke. Metric: A multi-echelon technique for recoverable item control. *Operations Research*, 16:122–141, 1968.

[52] A. Svoronos and P. Zipkin. Evalutation of one-for-one replenishment policies for multiechelon inventory systems. *Management Science*, 37(1):68–83, 1991.

[53] G. Tagaras. Effects of pooling on the optimization and service levels of two-location inventory systems. *IIE Transactions*, 21(3):250–258, 1989.

[54] G. Tagaras. Pooling in multi-location periodic inventory distribution systems. *Omega*, 27(1), 1999.

[55] G. Tagaras and M. A. Cohen. Pooling in two-location inventory systems with non-negligible replenishment lead times. *Management Science*, 38(8):1067–1083, 1992.

[56] G. Tagaras and D. Vlachos. A periodic review inventory system with emergency replenishments. *Management Science*, 47(3):415–429, 2001.

[57] S. Tayur. Computing order-up-to levels in capacitated environments. *Stochastic Models*, 9:585–598, 1992.

[58] S. Yanagi and M. Sasaki. Reliability analysis of a two-echelon repair system considering lateral resupply, return policy, and transportation times. *Computers and Industrial Engineering*, 27(1-4):493–497, 1994.

[59] P. Zipkin. On the imbalance of inventories in multi-echelon systems. *Mathematics of Operations Research*, 9(3):402–423, 1984.

Appendix A. The Allocation Optimization

We are given a set $M = \left\{1, 2, \ldots, \overline{M}\right\}$ of locations and an augmented set $M_0 = \{0\} \cup M$ that includes one location at a higher level. For each location $m \in M_0$, we are given a set of integer gridpoints $r^m = \left\{r_0^m, r_1^m, \ldots, r_{n(m)}^m\right\}$ indexed by a set $N_m = \{0, 1, \ldots, n(m)\}$, satisfying $r_0^m = 0$ and $r_n^m > r_{n-1}^m$, for all $n > 0$. At each gridpoint, r_n^m, for $m \in M$ and $n \in N_m$,we are given a function evaluation, c_n^m, of a convex function. We define a piecewise linear approximation to each original convex function as follows. For each gridpoint, we compute a slope, $\widehat{c}_n^m$, according to the following rule:

$$\widehat{c}_n^m = \begin{cases} \frac{c_{n+1}^m - c_n^m}{r_{n+1}^m - r_n^m}, & n < n(m); \\ \frac{c_n^m - c_{n-1}^m}{r_n^m - r_{n-1}^m}, & n = n(m). \end{cases} \tag{A-1}$$

By convexity of the original function, we have $\widehat{c}_n^m \geq \widehat{c}_{n-1}^m$ for all $n > 0$. The piecewise linear approximation function for location $m \in M$ is given by $\widetilde{C}_m(r)$:

$$\begin{aligned} \widetilde{C}_m(r) \equiv\ & c_0^m + \sum_{n=0}^{n(m)-1} \left\{1_{\{r \geq r_n^m\}} \left(r \wedge r_{n+1}^m - r_n^m\right) \widehat{c}_n^m\right\} \\ & + 1_{\left\{r \geq r_{n(m)}^m\right\}} \left(r - r_{n(m)}^m\right) \widehat{c}_{n(m)}^m. \end{aligned} \tag{A-2}$$

In addition, we are given a convex function $f(\cdot)$ defined on $\mathcal{R}^+$. The allocation optimization is to find function evaluations, c_n^0, for all $n \in N_0$, satisfying

$$c_n^0 = f(r_n^0) + \min_{\substack{s.t. \sum_{m \in M} r_m = r_n^0 \\ r_m \geq 0, \\ r_m \text{ integer}, \forall m \in M;}} \sum_{m \in M} \widetilde{C}_m(r_m). \tag{A-3}$$

The following marginal analysis algorithm can be used to solve the allocation optimization:

Algorithm AllocOpt:

1. For each $m \in M$, and each $n \in N_m$, compute $\widehat{c}_n^m$ using (A-1).
2. For each $m \in M$, set $n^*(m) \leftarrow 0$ and $r^*(m) = 0$.
3. Set $m^* = \arg\min_{m \in M} \left\{\widehat{c}_0^m\right\}$.
4. Set $z \leftarrow \sum_{m \in M} c_0^m$.

5. Set $c_0^0 \leftarrow z$.

6. Set $n \leftarrow 1$.

7. While $n \leq n(0)$, do:

   (a) Set $u \leftarrow r_n^0 - r_{n-1}^0$.

   (b) While $u > 0$, do:

      (I) If $n^*(m^*) = n(m^*)$ then set $x \leftarrow u$; else set

$$x \leftarrow u \wedge \left(r^{m^*}_{n^*(m^*)+1} - r^*(m^*)\right).$$

      (II) Set $z \leftarrow z + x \cdot \widehat{c}^m_{n^*(m^*)}$.

      (III) Set $r^*(m^*) \leftarrow r^*(m^*) + x$.

      (IV) If $n^*(m^*) < n(m^*)$ and $r^*(m^*) = r^{m^*}_{n^*(m^*)+1}$, then set $n^*(m^*) \leftarrow n^*(m^*) + 1$.

      (V) Set $m^* = \arg\min_{m \in M} \left\{\widehat{c}^m_{n^*(m)}\right\}$.

      (VI) Set $u \leftarrow u - x$.

   (c) Set $c_n^0 \leftarrow z$.

   (d) Set $n \leftarrow n + 1$.

8. For $n = 0, 1, \ldots, n(0)$, set $c_n^0 \leftarrow c_n^0 + f(r_n^0)$.

Proposition 2 *Algorithm **AllocOpt** terminates with a set $c^0 = (c_n^0)_{n \in N_0}$ satisfying (A-3) for each $n \in N_0$. Assuming a constant time algorithm exists to compute $f(r)$ for any $r \in \mathcal{R}^+$, algorithm **AllocOpt** requires*

$$O\left(\left(1 + \log_2\left(\overline{M}\right)\right) \sum_{m \in M_0} n(m)\right)$$

calculations.

Proof. Observe that u, n, and $r^*(m)$, for all m, are integers throughout the algorithm. Convexity of the piecewise linear functions (A-2) ensures that a marginal analysis algorithm of the form **AllocOpt** can be used to solve (A-3). The outer loop, step 7, is performed at most $n(0)$ times. The inner loop, 7b, is performed at most $\sum_{m \in M_0} n(m)$ times. This follows because on each loop either

$n^*(m)$ is incremented by one for some m, or u is set to zero and the loop is terminated. The maximum number of times $n^*(m)$ can be incremented for any m is $n(m)$. The main optimization step, 7(b)V, requires at most $\log_2\left(\overline{M}\right)$ comparisons, provided the vector $\widehat{c} = \left(\widehat{c}^m_{n^*(m)}\right)_{m\in M}$ is maintained as a heap [2] . The number of other calculations, as in steps (1) and (8), is proportional to $\sum_{m\in M_0} n(m)$. ■

Remark 2 *Equation (A-3) requires a minimization subject to the constraint* $\sum_{m\in M} r_m = r^0_n$. *It is trivial to extend algorithm* ***AllocOpt*** *to constraints of the form* $\sum_{m\in M} r_m \leq r^0_n$. *One simply modifies the inner loop, step* 7b*, to read: "While* $u > 0$ *and* $\widehat{c}^m_{n^*(m^*)} \leq 0$*, do..."*