MULTI-ECHELON EXCHANGEABLE-ITEM REPAIR SYSTEM OPTIMIZATION

Michael Dreyfuss¹ and Yahel Giat¹

¹Department of Industrial Engineering and Management, Jerusalem College of Technology,
Jerusalem, 91160 Israel

ABSTRACT
We consider an M/G/∞ exchangeable-item repair system with spares and ample servers to which arriving customers bring failed items for repair. The system comprises a multi-item-type, multi-echelon system in which failed items may be either repaired at the first echelon or sent for repair in higher echelons. To decrease the average waiting time of a random customer, managers can invest in multiple ways: increasing the number of spares; reducing repair times; improving repair capabilities and reducing shipping times. We provide an efficient algorithm for a sub-optimal budget allocation at each location in the system and derive an upper bound for the distance of the solution from the optimal solution. We derive the waiting time distribution of a random customer and the distribution of the number of customers in the system. Finally, we provide a numerical example that is motivated by a real-life situation to demonstrate the results and intuitions learned from the model.

KEYWORDS
M/G/∞, Multi-echelon, Optimal investment, Spares allocation, Fill rate, Fill rate window, Average waiting time.
INTRODUCTION

In many repair systems a customer arriving with a faulty unit is served with a spare unit in lieu of waiting until the item he brought is repaired. In large scale repair systems, this method of replacement can be complex with a number of repair and spare storage tiers so that if at any specific location there is no spare available, a request for repair or for a spare part may be forwarded to the next echelon thus allowing improved service times.

Such situations are quite common in military units. For example, if a communication kit in a certain vehicle malfunctions, then in order to restore availability the vehicle will receive a replacement kit and the malfunctioned unit will be either repaired on site or forwarded to repair at a site with the necessary repair facilities. Once the unit returns from repair it will be placed on a shelf as a spare unit to serve future customers. This repair-exchange approach is not limited to any single type of item, but is usually replicated to many of the different components of the vehicle. While this approach is quite common in military systems (Costantino, Di Gravio and Tronci, 2013), it is also applicable in large scale civil systems. For example, in the electrical car industry, battery replacement is considered as a solution to the battery recharging problem.

The objective of these complex systems is to restore serviceability in as little time as possible. To improve times while meeting budgetary constraints, managers must consider how to optimally allocate their budget in improving repair facilities, parts transportation, and spares purchase and allocation. Meeting this objective is even more complex if the system handles multiple types of items. Managers must now decide how many spare items of each type to purchase and how to allocate them in the different service stations.

Motivated by these problems we consider a multi-echelon exchangeable item repair system into which customers bring items for repair and obtain serviceable ones in return. In the basic model
we restrict the analyses to a single type of item. If the problem can be solved onsite, it will be repaired. Otherwise, it will be sent to a higher echelon location where it will be either repaired or sent further up to the next echelon. This process may repeat itself until the item has reached the highest echelon where it is repaired or replaced. In the generalized model the system is allowed to handle multiple item types.

In our model, the criterion for optimality is minimizing customers’ average waiting time. Given a certain budget for spares, for parts shipments and for repair facilities, it is the managers’ task to allocate this budget optimally among the different repair sites. We show that for both the basic model and the generalized model the optimization problem is convex. Utilizing the convexity property we continue to describe an algorithm for finding the optimal budget allocation. The intuition gained from the basic model allows us to generalize the basic model’s algorithm to find a sub-optimal allocation for the multi-type model and provide an upper bound for the distance from the optimal solution.

We describe how to compute additional performance measures that are of interest to managers. In this respect, we are novel in applying the fill rate window (see Song, 1998) to the multi-echelon setting. We also derive the probability distribution of the number of customers in each node, a measure that can be further used to derive the distribution of the number of customers in the entire system.

To demonstrate the derivation of the optimal solution and the performance measures we conclude the paper with a numerical example. The example we use is motivated by a real-life situation and provides intuition into how budget should be shifted among the system’s components.
LITERATURE REVIEW

The research of multi-echelon systems has been developed continuously ever since it was introduced by Sherbrooke (1968), (see Basten and Houtum (2014) for a recent survey).

Sherbrooke introduced a two-level echelon system with a central depot in the higher echelon and multiple bases in the lower echelon. He then proceeds to find the optimal spares allocation in the system. Over the years extensions to the basic model include, Muckstadt (1973), Albright and Gupta (1992) and Gupta and Albright (1993) who explore multi-indenture models. Diaz and Fu (1997) consider a model in which the repair facilities are limited in number. In contrast, we assume ample repair servers so that we treat the repair capacity as infinite. More recently, Levner et al. (2011) include repair capabilities only at the highest echelon. Our model relaxes this constraint by allowing for repair to be conducted at all system levels. Similar to us van der Heijden, Alvarez and Schutten (2014) consider improvement to travel time. In their model travel time is deterministic and they also constrain each node to receive spares only from a single source.

Another strand of multi-echelon spares allocation research focuses on the operations of inventory replenishment without repair capabilities; such papers include Song (1998) and Song and Yao (2002) who consider these ideas in an assembly-to-order environment, Cagiano et al. (2007) and Chen (2000) who analyzes the complexity of multi-echelon systems with bulk arrivals or batch reordering. Our research contributes to this field in that it considers also the effects of repair capabilities as well as the possibility to improve inventory through reducing travel time and repair time.

Our research is also related to developments in single-echelon logistic systems. Examining the repair procedures only, Berg and Posner (1990) consider a one echelon exchangeable-item repair
system with ample servers with a single item class. They derive the customer delay distribution in an $M/G/\infty$ exchangeable repair system with spares. Recently, Dreyfuss and Posner (2015) derive the customer delay distribution for such a model with batch arrival. In a dynamic model, Giat (2013) examines the effects of future changes in the system operations on the investment scheme. We extend these papers by deriving this distribution for the multi-echelon setting. In our paper, we are consistent with the terminology of Dreyfuss and Posner (2015) who define the term fill rate window for a time window $w$, as the probability of a customer being satisfied within time $w$. The fill rate window for time window $w = 0$ is in fact the well-known performance measure fill rate discussed, for example, in Larsen and Thorstenson (2014), Costantino, Di Gravio and Tronci (2013). The fill rate window that we derive is a powerful tool for managers to measure the system’s performance.

**THE BASIC MODEL**

Consider a $K$-echelon repair system with $N_k$ nodes in each echelon $k = 1 \ldots K$, and $N_K = 1$, (see Figure 1). Let the pair $(n, k)$ denote the $n$th node of the $k$th echelon. Each node in the system contains a repair facility with ample repair stations. Repair time is random with a general distribution function and a mean repair time $r_{n,k}$. Each node initially holds a stock of $s_{n,k}$ spare items. At this point, we assume that the mean repair time cannot be improved through investment. We relax this assumption in the general model.
We assume that customer arrivals at each node follow a Poisson arrival process with rate $\lambda_{n,k}$. Upon arrival, if the node contains an available item in stock, the failed item is immediately replaced and the customer is released out of the system. If no such spare item exits in the node then the customer will wait until a repaired item joins the node’s stock. Customers will then obtain replacement items from this stock as they become available, based on a first-come-first-serve policy. After receiving a serviceable item, the customer leaves the system.

Every node operates as an exchangeable-item repair system. Items are considered exchangeable in the sense that customers are ready to take any serviceable item of the same kind they brought to the system. The failed item that has arrived with the customer is either taken to repair at the node or, with the exception of the $K$th echelon, an order for repair is opened and sent to a node belonging to the next (higher) echelon. The decision of whether the item is repaired on site or shipped to the next echelon typically depends on the type of the failure and whether the current site is capable of repairing it. Let $p_{n,k}$ denote the portion of failures that the current node is able to repair. In other words, $p_{n,k}$ is the probability that repair is carried out at the node itself and not forwarded to the next echelon. For now, we assume that managers cannot invest in improving the node’s repair capabilities. In the general model we relax this assumption. If the item’s failure is
such that it cannot be repaired by the node’s repair facilities, it is sent to a node belonging to a
higher echelon. The probability of being sent to the $m$th node of the next echelon is given by $p_{n,k\rightarrow m,k+1}$, $m = 1 \ldots N_{k+1}$. Thus, for all $k = 1 \ldots K - 1$, and $n = 1 \ldots N_k$ we have that $p_{n,k} + \sum_{m=1}^{N_{k+1}} p_{n,k\rightarrow m,k+1} = 1$. Since in the last echelon all items must be repaired at the node itself we have that $p_{1,k} = 1$.

Without any loss of generality, nodes are permitted to forward repair orders to the next echelon only. Upon arrival in the next echelon the order may then be forwarded to the next echelon and so forth. Thus, we assume that it is not allowed to send an order directly beyond the next echelon. This assumption, however, is not limiting as we can easily construct an equivalent system that meets this assumption but effectively allows direct multi echelon forwarding. This is done by adding dummy nodes in the intermediate echelons that are set to transfer orders to the next echelon with probability one.

Recall, $\lambda_{n,k}$ is the customer arrival rate into node $(n,k)$. Since entry of failed items into any node can be done either through the direct deposit of a customer or through the transfer from lower-echelon nodes, the total arrival rate of items into node $(n,k), k = 2 \ldots K$ is given by the sum:

$$\tilde{\lambda}_{n,k} = \lambda_{n,k} + \sum_{m=1}^{N_{k-1}} \lambda_{m,k-1} p_{m,k-1\rightarrow n,k} \text{ for all } n = 1 \ldots N_k \tag{1}$$

And for the first echelon, $\tilde{\lambda}_{n,1} = \lambda_{n,1}$ for all $n = 1 \ldots N_1$.

Let $W_{n,k}$ denote the average waiting time at node $(n,k)$, that is, the average time it takes an order arriving at the node to leave the node. Notice that an order could be either a customer arriving at the node or a request for the repair of an item being forwarded from a lower echelon.

Shipment time back and forth between any two nodes is random with a general distribution function. Let $t_{m,k-1\rightarrow n,k}$ denote the average shipment and return time between nodes $(m,k - 1)$ and $(n,k)$. As items may be shipped using different transport methods, with faster shipping time
requiring higher shipping costs, we can generally assume that managers can decide to invest in improving shipping times. For now, in the basic model, we fix shipping times. In the general model, however, we will allow for the reduction of shipping times through investment.

The mean replenishment time at node \((n, k)\), \(R_{n,k}\), is the time it takes for an item arriving with an order to the node to join the stock at that node. For the last echelon, \(R_{1,K} = r_{1,K}\) since all items are repaired on site. For earlier echelons, however, replenishment time is the weighted average of the on-site repair time and the time it takes items to return from repair in higher echelons. Thus, for \(k = 1 \ldots K - 1\), \(R_{n,k}\), is given by

\[
R_{n,k} = p_{n,k}r_{n,k} + \sum_{m=1}^{N_{k+1}} [p_{n,k \rightarrow m,k+1}(t_{n,k \rightarrow m,k+1} + W_{m,k+1})] \tag{2}
\]

Clearly, if there are no spare items in a node, waiting time and replenishment time are identical.

The existence of spares allows for waiting time to be shorter than replenishment time.

From standard Poisson formulas (I found it also in Winston), we know that for a an \(M/G/\infty\) system without spares, the probability of having \(i\) customers waiting is equal to

\[
\left(\frac{\lambda_{n,k} R_{n,k}}{(i!)\lambda_{n,k} R_{n,k}}\right)^i e^{-\lambda_{n,k} R_{n,k}}.
\]

When using spares, the probability of having \(i\) customers is equal to

\[
\left(\frac{\lambda_{n,k} R_{n,k}}{(i+s_{n,k})!(i+s_{n,k})}e^{-\lambda_{n,k} R_{n,k}}.
\]

Using Little’s formulas, the average waiting time at node \((n, k)\) is given by

\[
W_{n,k} = \frac{1}{\lambda_{n,k}} \sum_{i=1}^{\infty} i \cdot \left(\frac{\lambda_{n,k} R_{n,k}}{(i+s_{n,k})!(i+s_{n,k})} e^{-\lambda_{n,k} R_{n,k}} \right) \tag{3}
\]

Notice that \(W_{n,k}\) depends not only on the number of spares in the node, \(s_{n,k}\), but also on the spares allocation in higher echelons through the replenishment time. To emphasize this dependency, in what follows we will express the average waiting time at any node by \(W_{n,k}(\vec{s})\), where \(\vec{s}\) denotes the initial allocation of spares in the entire system.
Optimal Spares Allocation for the Basic Model

Assume now that we have a budget for $S$ spare parts. The one-item class model problem is how to optimally allocate these $S$ items among the nodes of the system so that we minimize the overall average waiting time, $W$. By overall average waiting time we mean the average waiting time of a random customer arriving to the system. Thus,

$$\min_{\bar{s}} \left[ W = \sum_{k=1}^{K} \sum_{n=1}^{N_k} \frac{\lambda_{n,k} W_{n,k}(\bar{s})}{\lambda} \right] \text{ s.t. } \sum_{k=1}^{K} \sum_{n=1}^{N_k} s_{n,k} = S \quad (4)$$

where $\lambda = \sum_{k=1}^{K} \sum_{n=1}^{N_k} \lambda_{n,k}$ is the sum of customer arrivals to the system. Notice that in (4) we use the customer arrival rate $\lambda_{n,k}$ and not the total arrival rate $\tilde{\lambda}_{n,k}$ since we are interested in minimizing only customer waiting time and exclude the waiting times of orders that have been forwarded from a lower echelon and are waiting for service. We must first verify that $W_{n,k}(\bar{s})$ is convex to meet the necessary conditions for the Karush-Kuhn-Tucker formulas, (see Boyd and Vandenberghe, 2004). In the following lemma we show this is true for the last echelon’s node.

Lemma 1: $W_{1,K}(\bar{s})$ is convex in $\bar{s}$.

Proof: We can rewrite (3) so that the average waiting time at node $(1,K)$ is given by

$$W_{1,K} = \left( \tilde{\lambda}_{1,K} R_{1,K} \right) \sum_{i=s_{1,K}}^{\infty} \left( \frac{\tilde{\lambda}_{1,K} R_{1,K}}{(i)!} \right)^{i} e^{-\tilde{\lambda}_{1,K} R_{1,K}} - s_{1,K} \sum_{i=s_{1,K}+1}^{\infty} \left( \frac{\tilde{\lambda}_{1,K} R_{1,K}}{(i)!} \right)^{i} e^{-\tilde{\lambda}_{1,K} R_{1,K}} \quad (5)$$

The first difference of the average waiting time, given by (3), (where difference is defined as $\frac{\Delta y}{\Delta x} = y(x+1) - y(x)$) is given by

$$\frac{\Delta W_{1,K}}{\Delta s_{1,K}} = -\frac{1}{\tilde{\lambda}_{1,K}} \sum_{i=1}^{\infty} \left( \frac{\tilde{\lambda}_{1,K} R_{1,K}}{(s_{1,K}+i)!} \right)^{s_{1,K}+i} e^{-\tilde{\lambda}_{1,K} R_{1,K}} < 0 \quad (6)$$

and the second difference is given by

$$\frac{\Delta^{2} W_{1,K}}{\Delta s_{1,K}^{2}} = \frac{1}{\tilde{\lambda}_{1,K}} \left( \frac{\tilde{\lambda}_{1,K} R_{1,K}}{(s_{1,K}+1)!} \right)^{s_{1,K}+1} e^{-\tilde{\lambda}_{1,K} R_{1,K}} > 0; \text{ and } \frac{\Delta^{2} W_{1,K}}{\Delta s_{n,k}^{2}} = 0 \text{ for } (n,k) \neq (1,K). \quad (7)$$

Thus, $W_{1,K}$ is convex in the number of spares. QED.
In the following lemma we show that the average waiting time in the penultimate echelon is convex.

**Lemma 2:** For all \( n = 1 \ldots N_{k-1} \), \( W_{n,K-1} \) is convex.

Proof: Similarly to (6) and (7), \( W_{n,K-1} \) is convex with the spares allocation of the \( K - 1 \) and lower echelons. It is left to show that it is also convex with the spares allocation of higher echelons. By (2), the replenishment time for node \((n,K - 1)\) is given by

\[
R_{n,K-1} = p_{n,K-1} r_{n,K-1} + p_{n,K-1-1,K} \left( t_{n,K-1-1,K} + W_{1,K} \right).
\]

Since, by Lemma 1, \( W_{1,K} \) is convex we have that \( R_{n,K-1} \) is a convex function of \( \tilde{s} \). If \( s_{n,K-1} = 0 \) then (3) reduces to \( W_{n,K-1} = R_{n,K-1} \), in which case \( W_{n,K-1} \) is convex. Assume therefore that \( s_{n,K-1} > 0 \). The first derivative of (3) is

\[
\frac{\partial W_{n,k}}{\partial R_{n,k}} = \sum_{i=0}^{\infty} \left( \frac{\tilde{\lambda}_{n,k} R_{n,k}}{(i+s_{n,k})!} \right) e^{-\tilde{\lambda}_{n,k} R_{n,k}}, \text{ which is strictly positive, and the second derivative of }
\]

\[
W_{n,k} \text{ is } \frac{\partial^2 W_{n,k}}{\partial R_{n,k}^2} = \tilde{\lambda}_{n,k} \left( \frac{\tilde{\lambda}_{n,k} R_{n,k}}{(s_{n,k}-1)!} \right) e^{-\tilde{\lambda}_{n,k} R_{n,k}}, \text{ which is strictly positive, too. Since } W_{n,K-1} \text{ is a non-decreasing convex function of } R_{n,K-1} \text{ and } R_{n,K-1} \text{ is a convex function of } \tilde{s} \text{, we have that (see Section 3.2 in Boyd and Vandenberghe, 2004) } W_{n,K-1} \text{ is a convex function of } \tilde{s}. \text{ QED.}
\]

We conclude with the following Theorem, which establishes the convexity of the overall average waiting time.

**Theorem 1:** \( W \) is a convex function of \( \tilde{s} \).

Using arguments used in the proofs of Lemma 1 and Lemma 2, we can show by induction that for any node \((n,k)\), \( W_{n,k} \) is convex in \( s_{n,k} \). By the definition of \( W \) in (4) it therefore follows that \( W \) is a convex function of \( \tilde{s} \). QED.

The convexity of \( W \) allows us to add the Lagrange multiplier \( \theta \) to the minimization problem (4), which is now given by:
\[
\text{Min}_{\vec{s}} \left[ W = \sum_{k=1}^{K} \sum_{n=1}^{N_k} \frac{\lambda_{n,k} W_{n,k}(\vec{s})}{\lambda} + \theta \left( S - \sum_{k=1}^{K} \sum_{n=1}^{N_k} s_{n,k} \right) \right] \tag{8}
\]

and thus the necessary Karush-Kuhn-Tucker conditions are:

\[
\frac{\Delta W}{\Delta s_{n,k}} = \frac{1}{\lambda} \left( \lambda_{n,k} \frac{\Delta W_{n,k}}{\Delta s_{n,k}} + \sum_{l=1}^{k-1} \sum_{m=1}^{N_l} \lambda_{m,l} \frac{\Delta W_{m,l}}{\Delta s_{n,k}} \right) - \theta = 0 \tag{9}
\]

\[
\frac{\Delta W}{\Delta \theta} = S - \sum_{k=1}^{K} \sum_{n=1}^{N_k} s_{n,k} = 0 \tag{10}
\]

In order to achieve (9), we must allocate spares in the following manner. For each node we compute the contribution of an additional item in reducing overall average waiting time. Since \( W_{n,k} \) are all decreasing and convex, only local computations are necessary and thus the location with the maximal reduction will receive the next spare item. This process is repeated until we have reached our budget limit (10), i.e. until we have allocated all the \( S \) spare units. We formalize this intuition in the following algorithm:

**Single Item Spares Allocation Algorithm**

1. Initialization: Set \( s_{n,k} = 0 \), for all nodes, denoting the number of spares allocated to location \((n,k)\). Set \( s = S \), where \( s \) is the number of unit that need to be allocated.

2. Iteration:
   a. Using (2) and (6), for each node, set
      \[
      \theta_{n,k} = \frac{1}{\lambda} \left( \lambda_{n,k} \frac{\Delta W_{n,k}}{\Delta s_{n,k}} + \sum_{l=1}^{k-1} \sum_{m=1}^{N_l} \lambda_{m,l} \frac{\Delta W_{m,l}}{\Delta s_{n,k}} \right), \]
      the improvement if node \((n,k)\) received an additional item.
   b. Choose \((n^*, k^*)\) satisfying \( \theta_{n^*, k^*} = \min_{n,k} \{ \theta_{n,k} \} \), the destination of the current spare item.
   c. \( s_{n^*, k^*} \leftarrow s_{n^*, k^*} + 1 \), \( s \leftarrow s - 1 \).

3. If \( s = 0 \), finish. Otherwise go to step 2.
We note that when repeating step 2a of the algorithm it is not necessary to compute all the \( \theta_{n,k} \).

This should be done only for nodes that are affected by the last allocation; i.e. only for the node that received the last item and the nodes belonging to lower echelons.

The implication of (9) is that the optimal solution should satisfy that \( \frac{\Delta W}{\Delta s_{n,k}} = \frac{\Delta W}{\Delta s_{m,l}} \) for any two local allocations. However, since \( W_{n,k} \) is not a continuous function of \( s_{n,k} \) the optimal solution will not always produce the same marginal savings for each node.

The complexity of the algorithm is \( O(SM^2) \) where \( M = \sum_{k=1}^{K} N_k \) is the total number of nodes in the system. The fact that the total waiting time is convex allows us to use our algorithm rather than employ non-polynomial integer or dynamic programming based algorithms.

### THE GENERALIZED MODEL

We now relax three of our previous assumptions. First, we do not assume single item types and instead, allow the system to handle the repair of several types of items. Second, managers are now able to invest in reducing travel times between the nodes. Third, managers can invest in improving each node’s repair facilities and therefore (a) the probability for a node to be repaired on site may increase and (b) the average repair time of existing facilities may decrease.

The second assumption allows reduction of the travel time. We assume that the time reduction is convex with the money invested in reducing travel time. Notice that travel time need not be a continuous function of investment. We denote this relationship by \( t_{n,k \rightarrow m,k+1,j} (b) \) where \( b \) is the sum invested and the index \( j \) denotes the item type. Convexity implies that the marginal gain in travel time is decreasing.

The third assumption affects the model in two separate ways. The first change is that managers can improve existing repair facilities so that repair time decreases. We assume that if \( b \) dollars
are invested in reducing repair time in a node then the repair time in that node is $r_{n,k,j}(b)$, where $r_{n,k,j}(b)$ is decreasing and convex. The second change is that through investment, new repair capabilities are available so that more types of failures can be repaired on site. These enhanced capabilities result with a lower probability of the item being shipped to the next echelon. If managers invest $b$ dollars improving repair capabilities in the node then the probability for repair on site is $p_{n,k,j}(b)$ where $p_{n,k,j}(b)$ is increasing and convex. As with travel time, $r_{n,k,j}(b)$ and $p_{n,k,j}(b)$ need not be continuous function.

The variables of the model are similar to the basic model but with an added index denoting the item type. For example, in the multi-item type model, $\lambda_{n,k,j}$ will denote the arrival rate of customers to node $(n,k)$ bringing item type $j$. Let $c_j$ denote the unit cost of item type $j$. Let $b_{n,k,j}^t$, $b_{n,k,j}^r$, $b_{n,k,j}^p$ denote the dollar amount invested in reducing travel time, reducing repair time and improving repair facilities in node $(n,k)$, respectively. Let $B$ denote the budget dedicated for improving waiting time in the system. The budget constraint of problem (4) is rewritten as

$$\sum_{k=1}^{K} \sum_{n=1}^{N_k} \sum_{j} \left( c_j S_{n,k,j} + b_{n,k,j}^r + b_{n,k,j}^p + \sum_{m=1}^{N_{k+1}} b_{n,k \rightarrow m,k+1,j}^t \right) \leq B$$

Note that in the basic model the budget could be expressed in terms of the number of units ($S$), whereas in this general model we express the budget in dollar values. The optimal solution dictates therefore:

1. How many items of each type must be procured and how they should be distributed throughout the system.

2. How much money should be allotted to faster shipping and the routes to receive these funds.
3. How much money should be invested to expedite repair time and the locations to receive these investments.

4. How much money should be invested to improve repair capabilities and the locations to receive these investments.

For any allocation of spares there is no interdependency between item types. Therefore, the intuition gleaned from the basic model about the convex nature of the overall average waiting time is valid in this model as well. We can therefore use an algorithm similar to that of the basic model. However, instead of examining where the next item will be allocated as in the basic algorithm, in the multi-item algorithm we examine where the next dollar should be invested and then allocated. That is, we check the return on investment (ROI) of the dollar for each item type and each node allocation and determine the type and location that will best improve the overall average waiting time. Assuming that the remaining budget suffices we purchase the item and allocate it. If there are not enough funds left, we move on to purchase the unit with the best ROI that is still feasible.

**General Model Budget Allocation Algorithm**

1. Initialization: Set \( s_{n,k,j} = 0 \), for all nodes and types, denoting the number of spares allocated to location \((n, k)\) and type \(j \). Let \( b_{n,k \rightarrow m,k+1,j}^t = 0, b_{n,k,j}^r = 0, b_{n,k,j}^p = 0 \) for all travel lines, nodes and types, denoting the sums allocated for transportation, repair time and repair capabilities, respectively. Set \( b = B \), where \( b \) is the remaining budget.

2. Iteration: Using (2), (3) and (6), set:
   
   a. \( \theta_{n,k,j}^t = \frac{\frac{1}{2} \left( \lambda_{n,k,j}^l \frac{\Delta w_{n,k,j}}{\Delta s_{n,k,j}} + \sum_{i=1}^{N_l} \sum_{m=1}^{N_l} \lambda_{m,i,j}^l \frac{\Delta w_{m,i,j}}{\Delta s_{n,k,j}} \right)}{c_j} \), the marginal improvement if node \((n,k)\) receives an additional \(j\)-type item.
b. Set $\theta^{t}_{n,k \rightarrow m,k+1,j}$ as the *marginal* improvement in $W$ if we expedite current travel time from node $(n, k)$ to node $(m, k + 1)$ for item type $j$ to the next (shorter) level of travel time.

c. Set $\theta^{r}_{n,k,j}$ as the *marginal* improvement in $W$ if we improve item type $j$ repair times in node $(n, k)$ from their current level to the next possible repair time.

d. Set $\theta^{p}_{n,k,j}$ as the *marginal* improvement in $W$ if we increase the portion of item type $j$ failures that node $(n, k)$ can repair from its current level to the next possible level.

e. Choose the type of investment $x \in \{s, t, r, p\}$ and the location $(n^*, k^*, j^*)$ satisfying $\theta^{x^*}_{n^*,k^*,j^*} = \min_{x,n,k,j} \{\theta^{x}_{n,k,j}\}$ so that the cost for this improvement is less or equal to $b$. Let $c$ denote the cost for this improvement. If there is no feasible improvement, then finish.

f. Update the investment decision, either $s_{n^*,k^*,j^*} \leftarrow s_{n^*,k^*,j^*} + 1$ or $b^{t}_{n,k \rightarrow m,k+1,j} \leftarrow b^{t}_{n,k \rightarrow m,k+1,j} + c$ or $b^{r}_{n,k,j} \leftarrow b^{r}_{n,k,j} + c$ or $b^{p}_{n,k,j} \leftarrow b^{p}_{n,k,j} + c$. Update the remaining budget, $b \leftarrow b - c_{j^*}$.

3. Go to step 2.

This algorithm does not ensure optimality. To find a bound on the distance from optimality we need to record the first time we encounter insufficient funds to purchase an item or make a next level improvement that has the highest ROI. This happens at step 2e of the algorithm. Let $\alpha$ denote the ROI of the infeasible desired change and $\beta$ denote the remaining funds. Clearly, from this point on it would be impossible to attain a waiting time reduction greater than $\alpha \beta$. Let $\gamma$ denote the decrease in waiting time achieved from the aforementioned point until the completion of the algorithm. The upper bound for the distance from optimality is $\alpha \beta - \gamma$. 

15
The Dual Problem: Minimizing investment subject to the performance target

The dual problem is finding the minimal budget that meets a target average waiting time. This problem is relevant when the budget is not predetermined and is typically used by managers as a tool to design their budget demands. The algorithm to solve the dual problem is similar to the algorithm described above with the difference that the stopping criterion is not whether we have depleted our budget but whether we have reached our average waiting time target.

SERVICE MEASURES

In our model, managerial objective is to minimize the overall average waiting time. While this service measure is commonly used, there are other service measures that are of interest. In what follow we demonstrate how to analytically derive a number of such interesting measures:

**Overall average waiting time and average queue size**

The overall average waiting time, $W$ is defined in (4). Following Little’s Law, this measure is equivalent to the average queue size in the system, $L$.

**The waiting time distribution of a random customer in the system.**

Let $F(w)$ denote the probability of a random customer to wait less than $w$ units of time in the system. This measure is also called the fill rate window, for a window size of $w$ (see Song, 1998 and Dreyfuss and Posner, 2015) A special case is the commonly used fill rate measure (see Larsen and Thorstenson, 2014), which is in our notation, equivalent to $F(0)$. To derive $F(w)$ we need to first define the probability of a customer arriving to node $(n, k)$ with type $j$ leaving the system in less than $w$ time units, $F_{n,k,j}(w)$. In what follows we describe the derivation of $F(w)$.

Recall, $R_{n,k,j}$ is the average replenishment time of item type $j$ at node $(n, k)$. Let $\hat{R}_{n,k,j}(w)$ denote the cumulative probability of an item to be replenished by time $w$. Following the same
reasoning of (2), replenishment occurs either after local repair or after repair in higher echelons. The first case happens with probability \( p_{n,k,j} \) and for replenishment to happen within time \( w \) we need repair to take no more than \( w \) units of time, which happens with probability \( Pr(r_{n,k,j} \leq w) \). The second case occurs with probability \( p_{n,k\rightarrow m,k+1,j} \) and for replenishment to occur within time \( w \) we need that waiting time in the next node will not surpass \( w - t_{n,k\rightarrow m,k+1,j} \), which is equal to \( F_{m,k+1,j}(w - t_{n,k\rightarrow m,k+1,j}) \). Obviously, for the last echelon only the first case applies. We can therefore summarize that

\[
\hat{R}_{n,k,j}(w) = \begin{cases}
Pr(r_{1,k,j} \leq w) & k = K \\
p_{n,k,j}Pr(r_{n,k,j} \leq w) + \sum_{m=1}^{N_{k+1}} [p_{n,k\rightarrow m,k+1,j}F_{m,k+1,j}(w - t_{n,k\rightarrow m,k+1,j})] & k < K
\end{cases}
\]

For expositional reasons, we assume in the above that travel time is deterministic and equal to its mean. The equation can be easily rewritten to express the case in which travel time is random. Let, \( Y_{1,n,k,j}(w) \) and \( Y_{2,n,k,j}(w) \) be independent Poisson random variables with parameters \( \lambda_{1,n,k,j}(w) \) and \( \lambda_{2,n,k,j}(w) \), respectively, given by \( \lambda_{1,n,k,j}(w) = \lambda_{n,k,j} \int_{x=0}^{w} (1 - \hat{R}_{n,k,j}(x)) \, dx \) and \( \lambda_{2,n,k,j}(w) = \lambda_{n,k,j} \int_{x=0}^{w} \hat{R}_{n,k,j}(x) \, dx \). Following arguments similar to Corollary 12 of Berg and Posner (1990), the fill rate window for time \( w \) is

\[
F_{n,k,j}(w) = P(\hat{Y}_{n,k,j}(w) \leq s_{n,k,j} - 1) + \hat{R}_{n,k,j}(w)P(\hat{Y}_{n,k,j}(w) = s_{n,k,j}),
\]

where \( \hat{Y}_{n,k,j}(w) = Y_{1,n,k,j}(w) - Y_{2,n,k,j}(w) \). The fill rate window of a random customer in the system, \( F(w) \), is the weighted average of fill rate windows in each node and is given by

\[
F(w) = \frac{\lambda_{n,k,j}F_{n,k,j}(w)}{\lambda}.
\]

The order size distribution
For any node, customers entering the node and requests for repair forwarded from lower echelons comprise the orders in the node. Let \( X_{n,k,j} \) denote the order size at each node \((n, k)\) and type \(j\). To derive the probability distribution of the queue size, we can consider the probability for the queue size to be equal \(i\) immediately after the event of an order leaving the node. Queue size is zero if either the order did not have to wait at all or that there was no arrival during the order’s waiting time. The former happens with probability \( F_{n,k,j}(0) \) and the latter happens with probability \( \int_{w=0}^{\infty} f_{n,k,j}(w) e^{-\lambda_{n,k,j}w} dw \), where \( f_{n,k,j}(w) \) is the probability to wait \(w\) units of time. Finally, queue size is \(i > 0\), if during the order’s waiting time there were \(i\) arrivals, which happens with probability \( \int_{w=0}^{\infty} f_{n,k,j}(w) \frac{(\lambda_{n,k,j}w)^i}{i!} e^{-\lambda_{n,k,j}w} dw \). Thus, the probability distribution of \( X_{n,k,j} \) is given by

\[
P(X_{n,k,j} = i) = \begin{cases} 
F_{n,k,j}(0) + \int_{w=0}^{\infty} f_{n,k,j}(w) e^{-\lambda_{n,k,j}w} dw & i = 0 \\
\int_{w=0}^{\infty} f_{n,k,j}(w) \frac{(\lambda_{n,k,j}w)^i}{i!} e^{-\lambda_{n,k,j}w} dw & i \geq 1 
\end{cases}
\]

\[= \begin{cases} 
F_{n,k,j}(0) + \lambda_{n,k,j} \int_{w=0}^{\infty} F_{n,k,j}(w) e^{-\lambda_{n,k,j}w} dw & i = 0 \\
\lambda_{n,k,j} \int_{w=0}^{\infty} F_{n,k,j}(w) \left[ \frac{(\lambda_{n,k,j}w)^i}{i!} e^{-\lambda_{n,k,j}w} - \frac{(\lambda_{n,k,j}w)^{i-1}}{(i-1)!} e^{-\lambda_{n,k,j}w} \right] dw & i \geq 1 
\end{cases}
\]

NUMERICAL EXAMPLE

The numerical example provided in the following section is motivated by a real-life problem encountered by one of the authors. Consider an armored battalion and a single item. For simplicity, we assume the battalion comprises two companies, A and B. Company A comprises Platoons A1 and A2. In Table 1 we describe the relation between the units, the arrival rates probability repair times, and the probabilities for on-site repairs. We also assume that if a unit
cannot repair the item on site it will send the failed item only to its direct commanding unit. The budget is 500, each spare unit costs 10 and time units are in days.

**Table 1: Numerical Example Input Data**

<table>
<thead>
<tr>
<th>Unit</th>
<th>Commanding unit</th>
<th>Arrival rate</th>
<th>Mean repair time</th>
<th>Probability to repair on site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Battalion</td>
<td>------</td>
<td>1</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Company A</td>
<td>Battalion</td>
<td>0.1</td>
<td>10</td>
<td>0.6</td>
</tr>
<tr>
<td>Company B</td>
<td>Battalion</td>
<td>2.5</td>
<td>10</td>
<td>0.6</td>
</tr>
<tr>
<td>Platoon A1</td>
<td>Company A</td>
<td>1</td>
<td>10</td>
<td>0.3</td>
</tr>
<tr>
<td>Platoon A2</td>
<td>Company A</td>
<td>1</td>
<td>10</td>
<td>0.3</td>
</tr>
</tbody>
</table>

In the basic scenario all the units are located in the battalion’s central base and therefore shipping times are zero. In the second scenario Platoon A1 is deployed to practice at the training site so that the back and forth travel time is one day. In the third scenario Company B is deployed to the country’s border and its back and forth travel time is four days. In the fourth scenario we allow for travel time to be decreased through investment. Platoon A1 can reduce travel time from one day to a half a day by investing 10 units of money. Company B can reduce travel time to 3 days by investing 10 or reduce it to 2 days by investing 25. We describe the results for the four scenarios in Table 2.

**Table 2: Numerical Example Outputs: Spares allocation, investment and overall measures**

<table>
<thead>
<tr>
<th>Spares Allocation</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
<th>Scenario 3</th>
<th>Scenario 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Battalion</td>
<td>21</td>
<td>20</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>Company A</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Company B</td>
<td>14</td>
<td>14</td>
<td>19</td>
<td>17*</td>
</tr>
<tr>
<td>Platoon A1</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Platoon A2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>W</strong></td>
<td>1.877</td>
<td>2.081</td>
<td>3.286</td>
<td>3.126</td>
</tr>
<tr>
<td>Fill rate**</td>
<td>14.6%</td>
<td>12.8%</td>
<td>4.0%</td>
<td>4.4%</td>
</tr>
<tr>
<td>F(1)**</td>
<td>28.7%</td>
<td>25.1%</td>
<td>9.1%</td>
<td>10.0%</td>
</tr>
</tbody>
</table>

* In Scenario 4, Company B invests 10 units in reducing travel time.

**W, fill rate and F(0) are the overall average waiting time, the probability to wait zero time in the system and the probability to wait less than a day in the system, respectively.
From Table 2 we see that during Platoon A1’s training mission it should receive an extra spare item that should be withdrawn from the battalion’s central depot. Next, Company B is deployed to the border, which is distant and requires a considerably longer travel time, and therefore Company B receives five more items. These items are taken from the central depot, its sister company and both platoons. Typically, inventory management policies do not allow for this kind of dynamic adjustment. There could be a number of reasons for this. We conjecture that the complexity of computing the optimal allocation is the most notable one. Our example demonstrates the profits attained by dynamically adjusting the spares allocation. For example, if the basic allocation is maintained overall waiting time in scenario 3 is 3.35, but by adjusting the spares allocation this time is reduced to 3.286, which equals an absolute difference of 0.076 days in average total waiting time. In our example, there is a total daily arrival of 6 units per day which are equal 2190 arrivals annually. Therefore, an absolute difference of 1 day in waiting time is equivalent to 52,560 hours of operation annually. Consequently, the absolute difference of 0.076 is equivalent to an annual gain of almost 4000(!) operation hours.

In the fourth scenario we allow travel time reduction. In this particular case it is profitable to invest 10 units in order to reduce travel time 25% by investing 10 units. This, for example, can be attained by purchasing a vehicle dedicated to item transportation. This flexibility results in a reduction of 0.16 in average total waiting time, which is equivalent to an annual gain of 8400 hours of operation. Now Company B is receiving two items less. One item is returned to the battalion’s depot and the second is lost due to the smaller budget after the spending for travel reduction.
Between the third and the fourth scenarios there is an improvement in the waiting time. This is achieved through shifting investment from an additional spare item to reducing travel time. The probability to have a spare readily available (i.e. \( F(0) \) or the fill rate) and the probability to wait less than a day (i.e. \( F(1) \) or the fill rate window for window time one) also improve when we make this change (see Table 2). Table 3 compares these measures for the third and fourth scenario and allows us to examine whether improvement in waiting time necessarily corresponds with an increase in the fill rate measures. Overall, for the entire system this is true. For Company B, however, the decrease in waiting time is not coupled by an increase in the fill rate and the fill rate window. Since Company B receives one less spare and instead has shorter travel time, the fill rate measures perform worse (due to the reduction in the number of spares) whereas the waiting time at Company B decreases (due to the shorter travel time).

**Table 3: Unit Performance Measures For the Third and Fourth Scenarios**

<table>
<thead>
<tr>
<th>Unit</th>
<th>Scenario 3</th>
<th>Scenario 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>W*</td>
<td>Fill rate*</td>
</tr>
<tr>
<td>Battalion</td>
<td>3.495</td>
<td>2.1%</td>
</tr>
<tr>
<td>Company A</td>
<td>3.215</td>
<td>3.1%</td>
</tr>
<tr>
<td>Company B</td>
<td>3.050</td>
<td>5.4%</td>
</tr>
<tr>
<td>Platoon A</td>
<td>3.700</td>
<td>3.9%</td>
</tr>
<tr>
<td>Platoon B</td>
<td>3.288</td>
<td>3.3%</td>
</tr>
</tbody>
</table>

*W, Fill rate and F(0) are the average waiting time, the probability to wait zero time and the probability to wait less than one day, for each unit, respectively.

**CONCLUSION**

In this paper, we analyzed a complex multi-echelon system with the purpose of finding an optimal investment policy for spares, repair capabilities and travel time. We describe an efficient sub-optimal algorithm and bound the distance from optimality. We give formal expressions to a number of performance measures such as the fill rate, fill rate window and the number of clients.
in the system. Finally, we provide a numerical example motivated by a real life situation to illustrate the profits attained by applying the model. In this model, the objective of managers is to reduce the average total waiting time. In other situations managers may prefer a different objective, such as to minimize the fill rate window. As we can learn from the numerical example, these problems are not equivalent and can result in different budget allocations. The derivation of an efficient algorithm for this problem is left for future research.

REFERENCES


