A two-echelon spare parts network with lateral and emergency shipments: a product-form approximation

A Two-Echelon Spare Parts Network with Lateral and Emergency Shipments: A Product-Form Approximation

Richard J. Boucherie* Geert-Jan van Houtum† Judith Timmer*

Jan-Kees van Ommeren*

* Stochastic Operations Research, Department of Applied Mathematics, University of Twente,

P.O. Box 217, 7500 AE, Enschede, The Netherlands

{r.j.boucherie, j.b.timmer, j.c.w.vanommeren}@utwente.nl

† School of Industrial Engineering, Technische Universiteit Eindhoven,

P.O. Box 513, 5600 MB, Eindhoven, The Netherlands, g.j.v.houtum@tue.nl

May 10, 2017

Abstract

We consider a single-item, two-echelon spare parts inventory model for repairable parts for capital goods with high downtime costs. The inventory system consists of multiple local warehouses, a central warehouse, and a central repair facility. When a part at a customer fails, if possible his request for a ready-for-use part is fulfilled by his local warehouse. Also, the failed part is sent to the central repair facility for repair. If the local warehouse is out of stock, then, via an emergency shipment, a ready-for-use part is sent from the central warehouse if it has a part in stock. Otherwise, it is sent via a lateral transshipment from another local warehouse, or via an emergency shipment from the external supplier. We assume Poisson demand processes, generally distributed leadtimes for replenishments, repairs, and emergency shipments, and a basestock policy for the inventory control.

Our inventory system is too complex to solve for a steady-state distribution in closed form. We approximate it by a network of Erlang loss queues with hierarchical jump-over blocking. We show that this network has a product-form steady-state distribution. This enables an efficient heuristic for the optimization of basestock levels, resulting in good approximations of the optimal costs.

Keywords: Spare parts inventory control, multi-echelon, emergency shipments, product-form solution.

1 Introduction

Operational and manufacturing processes commonly depend on the availability of expensive capital goods such as large-scale computers, medical equipment, material handling systems and production equipment. For many of such systems, full service contracts are offered by Original Equipment Manufacturers (OEMs). Those contracts cover a whole range of maintenance activities, and one of these activities concerns the on-time deliveries of spare parts to replace failed parts in systems installed at the customers. Typically, spare parts have to be delivered within a limited number of hours, and therefore OEMs have networks consisting of one (or a few) central warehouse(s) and multiple local warehouses in the vicinity of their customers. Spare parts may be defined at different levels of the bill of material of a system. Relatively cheap spare parts are disposed of when they fail, while more expensive spare parts are repaired at repair shops. In many cases, expensive spare parts are modules that are produced by external suppliers and those external suppliers also take care of the repairs of failed modules.

In this paper, we consider a two-echelon inventory model for spare parts consisting of one central warehouse, one central repair facility and multiple local warehouses. We assume that a basestock policy is used for inventory control. The central repair facility is owned by an external supplier and we refer to it accordingly. We consider inventory control for a single repairable item that is part of an expensive capital good. Each local warehouse serves a group of customers where the capital goods are installed, and demands for ready-for-use parts occur according to Poisson processes with constant rates. When a demand occurs at a local warehouse, the failed part is sent to the external supplier, and the demand itself is fulfilled by the local warehouse if it has a part in stock. If the local warehouse is out of stock, then a ready-for-use part is sent from the central warehouse via an emergency shipment.

In case the central warehouse is also out of stock the part is supplied by another local warehouse at considerable costs. This is referred to as a *lateral transshipment*, which is often applied in practice. If the other local warehouses have no parts, then the part is supplied by the external supplier as soon as possible (the external supplier can always deliver). The latter types of shipments are also known as *emergency shipments*, and they are applied to avoid long and expensive downtimes of the capital goods at the customers. The external supplier is assumed to have ample repair capacity, and thus repair leadtimes for different failed parts are mutually independent. Lateral transshipments are typically used when they are (significantly) faster than emergency shipments from the external supplier.

Our inventory system is too complex to solve for a steady-state distribution in closed form. We approximate it by a network of Erlang loss queues with so-called hierarchical or nested jump-over blocking. We show that for a given basestock policy this network has a product-form steady-state distribution. Through this product-form solution, we find that the steady-state behavior is insensitive to the distribution of the order leadtimes. (It depends on that distribution only through the mean leadtimes.) Therefore, our results also hold for *deterministic leadtimes*, which are commonly assumed in practice. Due to the explicit results for the steady-state distribution, an efficient procedure is obtained for the approximation of the optimal basestock levels. This is illustrated through numerical results for a setting that includes inventory holding costs, transportation costs (regular costs for transportation among the warehouses and higher costs for sending the part from a warehouse to a customer), costs for repair and penalty costs for delayed fulfillments of demands (lateral and emergency shipments).

This paper contributes to a rich literature on multi-echelon inventory models for spare parts. This literature started with the seminal paper of Sherbrooke [25] in 1968. Sherbrooke formulated the so-called METRIC model, which is a two-echelon model *without* lateral and emergency shipments (instead unfilled demand is backordered). He derived an approximate procedure for the performance evaluation under a given basestock policy. A first exact analysis of the same model was derived by Simon [29], and Axsäter [4] developed an alternative exact analysis. The analysis of Simon was++= extended by Kruse [17] to multi-echelon systems. Slay [30] and Graves [14] independently developed similar approximation methods that were more accurate than the approximation method of Sherbrooke. Graves' approximation is based on exact recursions for pipeline inventories, which also allow for exact evaluations. Muckstadt [19] extended the METRIC approximation method to so-called two-indenture structures, under which spare parts are partitioned into assemblies and subassemblies. This approximation was improved by Sherbrooke [26] by extending Slay's approach. Rustenburg et al. [24] generalized the exact and approximate evaluation method of Graves to multi-echelon, multi-indenture systems. More recent, Caggiano et al. [11, 12] derived exact and practical methods for the evaluation of time-based fill rates. Selçuk [28] studies adaptive basestock levels for repairable item inventory control in a two-dimensional Markov model, that is solved analytically by matrix geometric methods.

An approximate evaluation procedure for a two-echelon system with emergency shipments was derived by Muckstadt and Thomas [20]. For the same system, Hausman and Erkip [15] developed a procedure based on single-echelon models. Recently, Özkan, Van Houtum and Serin [22] developed a new approximate evaluation method for key performance measures, that enables fast calculations. For different versions of

two-echelon systems with both emergency and lateral transshipments, approximate evaluation procedures were developed by Alfredsson and Verrijdt [2] and Grahovac and Chakravarty [13]. Approximate procedures for two-echelon models with lateral transshipments, but without emergency shipments, were developed by Lee [18], Axsäter [3], and Sherbrooke [27]. In many of the above papers, also heuristic optimization procedures (e.g. greedy procedures and Lagrangian heuristics) were developed for both single-item and multi-item settings; see also Wong et al. [32, 33], and the references therein. Recent reviews on inventory models may be found in [6, 7, 23].

Our work is most closely related to Alfredsson and Verrijdt [2] and Grahovac and Chakravarty [13]. Our work differs from the work of Alfredsson and Verrijdt because they assume a different order for the alternative options to satisfy a demand in case of a stockout. They first look at a lateral transshipment from another local warehouse, next the possibility of an emergency shipments from the central warehouse is considered, and an emergency shipment from the external supplier is applied when the first two options are not possible. In our paper, the first two options are interchanged. In real-life systems, one generally follows the rule that the fastest options are tried first. It then depends on geographical positions of warehouses and logistics procedures which order is preferred. Grahovac and Chakravarty assume emergency shipments from the central warehouse and lateral transshipments from other local warehouses, as we do, but in their model, no emergency shipments from the external supplier are possible (hence, demands are backordered at a local warehouse when the whole network is out of stock). The absence of this last option is realistic for certain types of networks in practice. This is typically so for networks managed by users of systems or by third parties. Our model includes the option of emergency shipments from the external supplier and this is generally realistic for networks managed by OEMs; see [6] for examples of the different types of networks in practice. The above differences are the main ones, and lead to significant differences in the steady-state behavior of the inventories (under a given choice of all basestock levels). At a more detailed level, our work differs from Alfredsson and Verrijdt [2] and Grahovac and Chakravarty [13] because we assume a slightly different procedure for replenishment orders of the local warehouse (see Section 2). We assume this slightly different procedure to facilitate our analysis. We believe that this does not lead to a significant change in the steady-state behavior of the inventory. Regarding the cost factors, we assume the same cost factors as Alfredsson and Verrijdt. In comparison to Grahovac and Chakravarty, we do not need a cost factor for backordered demands but we do have a cost factor for an emergency shipment from the external supplier.

The contribution of this paper is as follows. Our two-echelon model with emergency shipments and lateral transshipments is approximated by a network of Erlang loss queues with hierarchical jump-over blocking.

We show that the steady-state behavior of the network is described by a product-form solution. The underlying assumptions are in line with practice, as we argue when the precise model is described, and differ from those of e.g. Muckstadt and Thomas [20] and Alfredsson and Verrijdt [2]. The closed-form solution has several important implications. First, it enables an efficient heuristic for the approximation of single-item optimization that gives good results (as we demonstrate in Section 5). Second, it implies that the steady-state distribution and several relevant approximating performance measures only depend on the distributions of the repair and replenishment leadtimes via their means (i.e., they are insensitive to the underlying distributions except for their means).

This paper is organized as follows. In Section 2 we describe our model and the optimization problem. In Section 3 we formulate and in Section 4 we prove our main result. In Section 5 we consider the optimization problem and present numerical results. Finally, we conclude in Section 6.

2 Mathematical formulation of the problem

Consider a two-echelon spare parts inventory system that stocks spare parts of a single stockkeeping unit (SKU). The system consists of *multiple local warehouses* and *one central warehouse*. The local warehouses are numbered $1, \ldots, J$ ($J \in \{1, 2, \ldots\}$), and the index 0 is used for the central warehouse. The local warehouses are located in different continents or different parts of a region, and each local warehouse supports many technical systems installed in its neighborhood. The central warehouse is located at a central place of the total area that is serviced. The SKU is assumed to be relatively expensive and to have such a low demand, that the part is only stocked at the central warehouse and the local warehouses, i.e., there are no spare parts in stock at individual customers. Hence, when one of the installed technical systems has a failure of the SKU that we consider, then a spare part has to be delivered as soon as possible in order to minimize the downtime.

Time is continuous and failures within the installed base supported by local warehouse i occur according to a Poisson process with constant rate λ_i. Often parts have exponential life times, and thus it is natural to assume a Poisson demand process. Alternatively, parts may have non-exponential lifetimes, but the merged demand processes of all technical systems together may be close to Poisson, see [1]. Further, downtimes of technical systems are short in general, leading to a constant total failure rate.

Below, we first describe how demands are fulfilled. Next, we describe the inventory control, which includes

the procedures used for returns and repairs of failed parts and for replenishments of the stocks in all warehouses. Finally, we discuss the relevant performance measures and we formulate our optimization problem.

2.1 Fulfillment of demands

If a part fails at a technical system, then immediately a demand for a spare part is placed at its supporting local warehouse i, say. Four cases may be distinguished:

(i) the part is available at local warehouse i;

(ii) the part is not available at local warehouse i, but is available at the central warehouse;

(iii) the part is neither available at local warehouse i nor at the central warehouse, but is available at another local warehouse;

(iv) all local warehouses and the central warehouse are out of stock.

In case (i), the demand is immediately fulfilled from the local warehouse. On average, it takes a leadtime $T_{l,i}$ to pick the part from the local warehouse i and to bring the part from the local warehouse to the technical system that needs the part. In this case, the failed part is initially sent to local warehouse i. From there, the failed part is sent to the central warehouse and next it is sent to the external supplier; this is further explained in the next subsection.

In case (ii), the demand is fulfilled from the central warehouse via an emergency shipment. This means that a fast procedure is used for picking up the part and that a fast transport mode is used to bring the part to the technical system that needs the part. On average, it takes a leadtime $T_{c,i} \geq T_{l,i}$ to have the part available at the technical system that requires the part. In this case, we assume that the failed part is directly sent to the central warehouse.

In case (iii), the demand is fulfilled from another local warehouse, local warehouse j, $j \in \{1, \ldots, J\} \setminus \{i\}$, say. The part is picked up and sent to the technical system that requires the part via a fast transport mode. The part may go directly to the technical system or it may go via local warehouse i. In both cases, we say that the demand is fulfilled by a lateral transshipment. The leadtime for this lateral transshipment involves transportation from local warehouse j to i, for example transportation to an airport, air freight shipping, and transportation from the airport to warehouse i. We assume differences in leadtimes for

different warehouses j may be neglected. The leadtime for lateral transshipment is on average $T_{a,i} \geq T_{c,i}$. We assume that the failed part is sent to local warehouse j. When multiple local warehouses have a part available, one can choose from where the part is sent via a lateral transshipment. We use the rule that the part is sent from the local warehouse which has the highest physical stock level; in case of a tie, we simply choose the warehouse with the smallest index. (Notice that at this point also another rule may be chosen. In the approximate evaluation that we develop, this rule plays no role. Hence, the approximate evaluation would remain the same if one would use another rule.)

In case (iv), the demand is fulfilled by the external supplier via an emergency shipment. The external supplier keeps no ready-for-use parts in stock, but may finish the repair of one of the failed parts in the repair shop via a fast procedure. Next, the repaired part is sent to the technical system that needs the part via a fast transport mode. The average leadtime for the whole procedure is denoted by $T_{s,i} \geq T_{a,i}$. We assume that in this case the failed part is directly sent to the external supplier.

Because downtime of the technical system is expensive, we assume that demand will be fulfilled in a relatively short time in all cases. Therefore, the mean leadtimes $T_{l,i}$, $T_{c,i}$, $T_{a,i}$, and $T_{s,i}$ are assumed to range from a few hours till a few days.

2.2 Inventory control

In the description above, we see that each time a demand is fulfilled, the failed part is sent to the location (a local warehouse, the central warehouse, or the external supplier) that fulfilled the demand. When a failed part is returned to a local warehouse, it will be sent to the central warehouse and at the same time a replenishment order is placed at the central warehouse (below, we describe the precise procedure in more detail). All failed parts that arrive at the central warehouse, either from local warehouses or directly from technical systems at customers, are sent to the external supplier and after repair they are returned to the central warehouse as ready-for-use parts. At the external supplier, failed parts arrive from the central warehouse and directly from technical systems at customers. We assume that all failed parts can be repaired.

It is easily seen that the total stock of spare parts in the system (of the local warehouses, the central warehouse, the external supplier, and the various transport lines) remains constant. In addition, under the above assumptions, the inventory position (= physical stock minus the backlog if applicable plus the

amount on order) of the central warehouse and of each local warehouse remains at a constant level. Hence, we may also say that we follow a basestock policy $(S_0, S_1, \ldots, S_J)$ for the inventory control, where S_i is the basestock level (constant inventory position) of warehouse i.

The detailed order and replenishment procedure at a local warehouse i, $i \in \{1, \ldots, J\}$, is as follows. At the moment that local warehouse i fulfills a demand, either in case (i) or case (iii) as described in the previous subsection, the failed part is sent to this warehouse. Local warehouse i places a replenishment order immediately upon sending a part to the customer, and it sends the failed part immediately to the central warehouse. However, the central warehouse accepts the replenishment order only after the failed part arrives at the external supplier. With this rule, the local warehouse knows that it will not receive a replenishment of a ready-for-use part if the failed part is not sent back. This avoids that a local warehouse forgets to send the failed part back or that it is slow in doing so. The time until the failed part arrives at the central warehouse is typically one or two weeks. Once the materials coordinator at the central warehouse sees that the failed part has arrived, the spare part is shipped to the local warehouse. This is assumed to take a short time only. To facilitate the analysis, we assume that this second part is instantaneous. Thus, the leadtime for a replenishment order consists of the time until the failed part arrives at the central warehouse. This leadtime is generally distributed with mean $1/\mu_i$. The corresponding random variable is denote by X_i. The leadtimes X_i of multiple replenishment orders are assumed to be mutually independent.

If there is no part available at the central warehouse, then the replenishment order is backordered and fulfilled as soon as possible. In this case, the replenishment leadtime consists of the time X_i and the backorder time at the central warehouse.

The central warehouse fulfills demands that come directly from failed machines at customers (see case (ii) in the previous subsection) and replenishment orders placed by the local warehouses. In both cases, a failed part is returned and the central warehouse sends this part to the external supplier for repair. This part is sent from the external supplier after a generally distributed repair leadtime with mean $1/\mu_0$. The corresponding random variable is denoted by X_0. The repair leadtimes of multiple repair orders are assumed to be mutually independent. When a repaired part returns to the central warehouse as a ready-for-use part, it is added to the physical stock unless there are backordered replenishment orders.

Looking in more detail at the repair leadtime X_0, we see two different situations. When the central warehouse receives a replenishment order, a failed part arrives at the same moment at the central warehouse and the repair leadtime consists of sending the failed part to the external supplier, the time that the part

is at the external supplier, and the time to send the ready-for-use part to the central warehouse. When the central warehouse fulfills a demand that comes directly from a customer, the repair leadtime consists of the same three components plus the time needed to send the failed part from the customer to the central warehouse. We assume that the latter component is relatively small and can be neglected. In practice, the repair leadtime is typically in the order of months, while the time to return a failed part to the central warehouse is in the order of 1-2 weeks. This return time is relatively small, and thus we assume that it is not necessary to distinguish between the repair leadtimes in the above two situations.

2.3 Performance measures and optimization problem

We are interested in the following performance measures: the average delay W_i per request at local warehouse i to place a spare part at a failed technical system, and the total average costs g. Let $\beta_{l,i}$, $\beta_{c,i}$, $\beta_{a,i}$, and $\beta_{s,i}$ denote the fraction of demand at local warehouse i that is fulfilled by the local warehouse, the fraction that is fulfilled by the central warehouse, the fraction that is fulfilled by lateral transshipments, and the fraction fulfilled by the external supplier, respectively. Then the average delay per request equals

$$W_i = \beta_{l,i}T_{l,i} + \beta_{c,i}T_{c,i} + \beta_{a,i}T_{a,i} + \beta_{s,i}T_{s,i}. \tag{1}$$

The costs consist of inventory holding costs for the spare parts and costs for transport and emergency deliveries. Furthermore, we introduce costs for the delay. We distinguish the following cost parameters:

h_i the inventory holding cost per part per time unit in warehouse i $(i = 0, \ldots, J)$;

$c_{l,i}$ the average cost when a demand at local warehouse i is fulfilled by the local warehouse itself, which includes the cost of fast transport from the local warehouse to the customer and the cost to return the failed part to local warehouse i;

$c_{c,i}$ the average cost when a demand at local warehouse i is fulfilled by the central warehouse, which includes the cost of fast transport from the central warehouse to the customer and the cost to return the failed part to the central warehouse;

$c_{a,i}$ the average cost when a demand at local warehouse i is fulfilled by a lateral transshipment from another local warehouse, which includes the cost of fast transport from the other local warehouse to the customer and the cost to return the failed part to the other local warehouse;

$c_{s,i}$ the average cost when a demand at local warehouse i is fulfilled by the external supplier, which

includes the cost to finish the repair of one of the failed parts at the supplier via a fast procedure, the cost of fast transport from the external supplier to a customer and the cost to return the failed part to the external supplier;

c_i^{repl} the average cost of a replenishment order placed by local warehouse i, which includes the cost of regular transport from the central warehouse to local warehouse i and the cost to send a failed part from the local warehouse i to the central warehouse;

c_0^{rep} the average cost of a repair order placed by the central warehouse, which includes the regular cost of repair of a failed part, the cost of regular transport from the external supplier to the central warehouse, and the cost to send a failed part from the central warehouse to the external supplier;

p_i the penalty for delay per part per time unit for demand at local warehouse i.

Let $\beta_{a,i,j}$ denote the fraction of lateral transshipments to local warehouse i from warehouse j; then $\beta_{a,i} = \sum_{j:j\neq i} \beta_{a,i,j}$. The average costs, as a function of the basestock levels, consist of holding costs, costs for fulfilling demands, costs of replenishment orders, costs of repair orders and delay costs:

$$g(S_0,\ldots,S_J) = \sum_{i=0}^{J} h_i S_i + \sum_{i=1}^{J} \lambda_i \left[\beta_{l,i} c_{l,i} + \beta_{c,i} c_{c,i} + \beta_{a,i} c_{a,i} + \beta_{s,i} c_{s,i} \right. \tag{2}$$
$$\left. + \beta_{l,i} c_i^{repl} + \sum_{j\in\{1,\ldots,J\}\setminus\{i\}} \beta_{a,i,j} c_j^{repl} + (\beta_{l,i} + \beta_{c,i} + \beta_{a,i}) c_0^{rep} + W_i p_i \right].$$

As we see from (1) and (2), to evaluate all delays W_i and costs $g(S_0,\ldots,S_J)$, we need the fractions $\beta_{l,i}$, $\beta_{c,i}$, $\beta_{a,i}$, and $\beta_{s,i}$. These are obtained from the steady-state probabilities; see the end of Section 3.

The goal is to minimize the average costs by selecting appropriate basestock levels. This leads to the following optimization problem.

$$\begin{aligned} &\min\ g(S_0,\ldots,S_J) \\ &\text{s.t. } S_0,\ldots,S_J \text{ nonnegative and integer} \end{aligned} \tag{3}$$

Sections 3 and 4 below are devoted to determining the steady-state distribution of our model. This distribution is used in section 5 to obtain the parameters $\beta_{l,i}$, $\beta_{c,i}$, $\beta_{a,i}$, and $\beta_{s,i}$, and optimize the costs $g(S_0,\ldots,S_J)$.

3 Main Result

In this section, we formulate our main result. First, we model the two-echelon spare part inventory network with multiple local warehouses and exponential replenishment and repair leadtimes. The resulting inventory system is too complex to solve for a steady-state distribution in closed form. Fortunately, we may arrange the lateral transshipments such that the resulting approximating model has a product-form solution for the steady-state probabilities.

Let n_i, $i = 1, \ldots, J$, be the number of failed parts that are send from local warehouse i to the central warehouse; we also refer to this as the number of outstanding repair orders of local warehouse i. Let n_0 be the number of outstanding repair orders at the central warehouse, and $n_{tot} = n_0 + \cdots + n_J$ the total number of outstanding repair orders. Then $n_i \in \{0, 1, \ldots, S_i\}$, $i = 1, \ldots, J$, and $n_0 \in \{0, 1, \ldots, S_0 + \cdots + S_J\}$ such that $0 \leq n_{tot} \leq S_{tot}$, with $S_{tot} = S_0 + \cdots + S_J$ the total inventory. Also, let n_{0i} be the number of outstanding repair orders at the central warehouse due to local warehouse i. It must be that $n_0 = n_{01} + \cdots + n_{0J}$. Let $N = (N_{01}, \ldots, N_{0J}, N_1, \ldots, N_J)$ denote the Markov chain recording the numbers of outstanding repair orders of the warehouses in the inventory system. It has states $n = (n_{01}, \ldots n_{0J}, n_1, \ldots, n_J)$ and state space

$$S = \{n \ : \ 0 \leq n_i \leq S_i, \ n_{0i} \geq 0, \ i = 1, \ldots, J; \ 0 \leq n_{tot} \leq S_{tot}\}.$$

To describe the transition rates, let e_i and e_{0i} denote the unit vectors with element 1 at position i and $0i$, respectively, and zero elsewhere. These vectors have the same dimension as the state vector. Recall cases (i) – (iv) described in Subsection 2.1. If local warehouse i has stock available (case (i)), then the transition rate is straightforward, see (4). If it does not have stock, then the central warehouse is considered. If the central warehouse has stock available (case (ii)), then the part is supplied from there (5). If not, then the other local warehouses are considered. If one of these has stock available, then the part is supplied by another local warehouse (case (iii)); this is a lateral transshipment (6). Case (iv) occurs if $n_{tot} = S_{tot}$, $n_i = S_i$. In that case the state is unaffected by the repair request as this request is directly fulfilled by the external supplier. Furthermore, the central warehouse replenishes the inventories of the local warehouses (7), and the external supplier does so for the central warehouse (8). Implementing these system dynamics

in a transition scheme gives the following equations.

For $i = 1, \ldots, J$

$$q(n, n + e_i) = \lambda_i, \qquad n_i < S_i,\ n_{tot} < S_{tot}, \tag{4}$$

$$q(n, n + e_{0i}) = \lambda_i, \qquad n_i = S_i,\ n_0 = \sum_{j=1}^{J} n_{0j} < S_0,\ n_{tot} < S_{tot}, \tag{5}$$

$$q(n, n + e_k) = \lambda_i, \qquad n_i = S_i,\ n_0 = S_0,\ n_k < S_k,\ n_{tot} < S_{tot}, \tag{6}$$

$$q(n, n + e_{0i} - e_i) = n_i \mu_i, \tag{7}$$

$$q(n, n - e_{0i}) = n_{0i} \mu_0. \tag{8}$$

This model is numerically hard to solve. One drawback of the lateral transshipment scheme is that it requires specification of the local warehouse k that is selected to replenish the request in the transition rates (6). To circumvent this specification, in our mathematical model we consider the following *approximation* for handling lateral transshipments. If local warehouse i is out of stock, but either the central warehouse or another local warehouse has inventory, then we assume that the demand at warehouse i is fulfilled by an emergency shipment from the central warehouse. We assume that the central warehouse always sends an item, even if the central warehouse is out of stock. An item sent from the central warehouse when it is empty is called a *virtual item*; this looks like a backlog, only it occurs under the restriction that somewhere in the system there is an item available. Note that in the original system this item would be supplied from another local warehouse. Under this approximation, the inventory system with multiple local warehouses has the following transition rates.

For $i = 1, \ldots, J$

$$q(n, n + e_i) = \lambda_i, \qquad n_i < S_i,\ n_{tot} < S_{tot}, \tag{9}$$

$$q(n, n + e_{0i}) = \lambda_i, \qquad n_i = S_i,\ n_{tot} < S_{tot}, \tag{10}$$

$$q(n, n + e_{0i} - e_i) = n_i \mu_i, \tag{11}$$

$$q(n, n - e_{0i}) = n_{0i} \mu_0. \tag{12}$$

The equations (9), (11) and (12) coincide with the equations (4), (7) and (8), respectively. The former equations (5) and (6) are replaced by (10), representing the emergency shipments by the central warehouse of normal and virtual items.

Four cases of transition rates are distinguished in (9)–(12). The rates (9), (11), and (12) show a regular pattern inside the state space. At the boundary of the state space the rates (9) are modified into (10). This structure coincides with that resulting from a network of Erlang loss queues where a customer (demand for a spare part) arrives to queue i, routes from queue i to queue 0, and leaves the network from queue 0. The queues have common capacity restrictions, namely the number of customers in queue i does not exceed S_i, $i = 1, \ldots, J$, and the total number of customers does not exceed S_{tot}. When the total number of customers in the system equals its upper limit S_{tot} a customer arriving to the system is blocked and discarded. When the system is not full, but a customer arriving to queue i finds this queue full, it jumps over queue i to receive service at queue 0. Thus, $N = (N_{01}, \ldots N_{0J}, N_1, \ldots, N_J)$ records the number of customers in the queues of the network of Erlang loss queues with capacity restrictions.

Notice that the Erlang loss queue is a queue with Poisson arrivals, and capacity C, say, where each arriving customer obtains its own server. A customer arriving when all servers are occupied is blocked and cleared. The steady-state distribution $\pi_C(m)$ of m customers in the Erlang loss queue with arrival rate λ and service rate μ is (see [31])

$$\pi_C(m) = \left(\frac{\lambda}{\mu}\right)^m \frac{1}{m!} G_C^{-1}, \quad m = 0, \ldots, C, \qquad G_C = \sum_{m=0}^{C} \left(\frac{\lambda}{\mu}\right)^m \frac{1}{m!},$$

where G_C is the normalizing constant.

The Erlang loss queue behaves as an infinite server queue to which customers that arrive when C customers are present are blocked and cleared. It is well-known that the infinite server queue is a so-called BCMP queue (see [5]). A network of infinite server queues in which customers arrive to queue i, route from queue i to queue 0, and leave the network from queue 0, with arrival rate λ_i to queue i, service rate μ_i at queue i, and service rate μ_0 at queue 0, has a product-form steady-state distribution. Namely, for state n with $0 \leq n_i$, $0 \leq n_{0i}$, $i = 1, \ldots, J$, the equilibrium probability $\pi(n)$ is

$$\pi(n) = \left(\prod_{i=1}^{J} \left(\frac{\lambda_i}{\mu_0}\right)^{n_{0i}} \frac{1}{n_{0i}!}\right) \left(\prod_{i=1}^{J} \left(\frac{\lambda_i}{\mu_i}\right)^{n_i} \frac{1}{n_i!}\right) \exp\left[-\left(\frac{\sum_{i=1}^{J} \lambda_i}{\mu_0} + \sum_{i=1}^{J} \frac{\lambda_i}{\mu_i}\right)\right].$$

The jump-over blocking protocol is a product-form preserving blocking protocol, see [8, 31]. Below we show that it may also be used in a nested fashion with capacity restrictions at both queues and groups of queues. Under jump-over blocking, the steady-state distribution is un-altered except for normalization. From these observations, the steady-state distribution of the process recording the number of customers in the queues is of product-form. We obtain the following result. The proof is given in Section 4.

Theorem 1 *The Markov chain* $N = (N_{01}, \dots N_{0J}, N_1, \dots, N_J)$ *with state space*

$$S = \{n \ : \ 0 \le n_i \le S_i, \ 0 \le n_{0i}, \ i = 1, \dots, J; \ n_{tot} \le S_{tot}\}$$

and transition rates

$$\begin{aligned} q(n, n + e_i) &= \lambda_i, \quad n_i < S_i, \ n_{tot} < S_{tot}, \\ q(n, n + e_{0i}) &= \lambda_i, \quad n_i = S_i, \ n_{tot} < S_{tot}, \\ q(n, n + e_{0i} - e_i) &= n_i \mu_i, \\ q(n, n - e_{0i}) &= n_{0i} \mu_0, \ i = 1, \dots, J, \end{aligned}$$

has steady-state distribution

$$\pi(n) = \frac{1}{G} \left(\prod_{i=1}^{J} \left(\frac{\lambda_i}{\mu_0} \right)^{n_{0i}} \frac{1}{n_{0i}!} \right) \left(\prod_{i=1}^{J} \left(\frac{\lambda_i}{\mu_i} \right)^{n_i} \frac{1}{n_i!} \right), \tag{13}$$

where G *is the normalizing constant,*

$$G = \sum_{n \in S} \left(\prod_{i=1}^{J} \left(\frac{\lambda_i}{\mu_0} \right)^{n_{0i}} \frac{1}{n_{0i}!} \right) \left(\prod_{i=1}^{J} \left(\frac{\lambda_i}{\mu_i} \right)^{n_i} \frac{1}{n_i!} \right).$$

Moreover, the steady-state distribution is insensitive to the distributions of the leadtimes of the replenishment orders except for their means $1/\mu_i$, $i = 0, \dots, J$.

A common method to evaluate the normalising constant G is via Monte-Carlo summation. The key observation is that the equilibrium distribution is a multi-dimensional Poisson distribution with rate $\nu = (\lambda_i/\mu_0, \ i = 1, \dots, J, \lambda_i/\mu_i, \ i = 1, \dots, J)$ truncated to the state space S. Let X_i, $i = 1, \dots, I$, be iid samples from this Poisson(ν) distribution. An unbiased estimator of the normalising constant is obtained from the fraction of samples X_i in S. For additional results on efficient estimation of performance measures from Poisson distributions, see [9].

Now the steady-state distribution is known, we can approximate the fractions $\beta_{l,i}$, $\beta_{c,i}$, $\beta_{a,i}$ and $\beta_{s,i}$. Let V_i denote the number of virtual items due to demand at local warehouse i. This number may be interpreted as the amount of service from other local warehouses to local warehouse i. We assume that a virtual item is replaced by a real item as soon as possible, that is when an item arrives from the supplier to the central warehouse.

Let $\mathsf{P}(V_i = v | n_0, n_{0i})$ denote the conditional probability of v virtual items due to demand at local warehouse

i given that $N_0 = n_0$ and $N_{0i} = n_{0i}$. When $n_0 \leq S_0$ there are no virtual items. Now let $n_0 > S_0$ and assume that all n_0 requests are numbered in order of arrival at the central warehouse. The first S_0 requests represent items that have been sent; the remaining $n_0 - S_0$ requests represent virtual items. Since requests at the local warehouses arrive according to a Poisson process, the order of requests is random. So, the conditional probability is hypergeometric:

$$\mathsf{P}(V_i = v | n_0, n_{0i}) = \frac{\binom{S_0}{n_{0i}-v}\binom{n_0-S_0}{v}}{\binom{n_0}{n_{0i}}},$$

for $v = \max\{0, S_0 - n_{0i}\}, \ldots, \min\{n_0 - S_0, n_{0i}\}$, and 0 otherwise.

Recall that an emergency shipment occurs whenever a request arrives at an empty local warehouse, say local warehouse i. This is not only the case when $N_i = S_i$, but also when the total number of requests $N_i + V_i$ due to requests at local warehouse i is at least S_i. Using the conditional probability and the steady-state distribution, the fractions $\beta_{l,i}$, $\beta_{c,i}$, $\beta_{a,i}$ and $\beta_{s,i}$, $i = 1, \ldots, J$, may be calculated as follows.

$$\begin{aligned}
\beta_{l,i} &= \sum_{n \in S: n_i < S_i} \pi(n) \mathsf{P}(V_i < S_i - n_i | n_0, n_{0i}), \\
\beta_{c,i} &= \sum_{n \in S: n_i = S_i, n_0 < S_0} \pi(n), \\
\beta_{a,i} &= \sum_{\substack{n \in S: n_i = S_i, n_0 \geq S_0, \\ \sum_{j=1}^J n_j < \sum_{j=1}^J S_j}} \pi(n) \mathsf{P}(V_i \geq S_i - n_i | n_0, n_{0i}), \\
\beta_{s,i} &= \sum_{n \in S: n_{tot} = S_{tot}} \pi(n).
\end{aligned}$$

4 Proof of Theorem 1

For a system of one central warehouse and two local warehouses, we show that the steady-state distribution of the number of items in stock in that network has a product-form solution, and is insensitive to the distributions of the replenishment times except for their means. The generalization to the network of multiple Erlang loss queues is immediate, and is left to the reader.

First consider exponential leadtimes. The Markov chain N is irreducible at finite state space S. Therefore,

the steady-state distribution π is the unique solution of the global balance equations, for all $n \in S$,

$$\sum_{n' \in S} (\pi(n)q(n,n') - \pi(n')q(n',n)) = 0.$$

It is readily shown that the partial balance equations are satisfied. These equations read, for $n \in S$, $i = 1, 2$,

$$\pi(n)q(n, n - e_{0i}) - \pi(n - e_{0i} + e_i)q(n - e_{0i} + e_i, n) - \pi(n - e_{0i})q(n - e_{0i}, n) = 0, \tag{14}$$

$$\pi(n)q(n, n - e_i + e_{0i}) - \pi(n - e_i)q(n - e_i, n) = 0, \tag{15}$$

and

$$\sum_{i=1}^{2} (\pi(n) (q(n, n + e_i) + q(n, n + e_{0i})) - \pi(n + e_{0i})q(n + e_{0i}, n)) = 0. \tag{16}$$

Now consider the network of queues 0, 1, 2 with phase-type service requirements. For a general introduction to phase-type distributions and insensitivity results, see e.g. [10, 31]. The service request S_i at queue i has the following phase-type distribution

$$P(S_i \leq x) = \sum_{k=1}^{\infty} p_{i,k} Erl(k, \nu_i)(x), \quad x \geq 0,$$

$$Erl(k, \nu)(x) = 1 - \sum_{t=0}^{k-1} \frac{(\nu x)^t}{t!} e^{-\nu x}, \quad x \geq 0,$$

$$\sum_{k=1}^{\infty} p_{i,k} = 1,$$

that is, with probability $p_{i,k}$ the service request at queue i has an Erlang distribution with k exponential phases with rate ν_i, $k = 1, 2, \ldots$. The phase-type distribution has mean service time

$$\frac{1}{\mu_i} = \sum_{k=1}^{\infty} \frac{k p_{i,k}}{\nu_i}.$$

Thus, a customer arriving to queue i selects with probability $p_{i,k}$ a service requirement containing k exponential phases with rate ν_i. Phase-type distributions are dense in the class of distributions with non-negative support [16]. Due to the exponential phases, we can model the system as a Markov chain. To this end, we add for each customer its remaining number of phases in the state description.

We represent a state with n_{01}, n_{02}, n_1, n_2, customers in queues 0, 1, 2, as $r = (r_{(01)1}, \ldots, r_{(01)n_{01}}, r_{(02)1}, \ldots, r_{(02)n_{02}}, r_{11}, \ldots, r_{1n_1}, r_{21}, \ldots, r_{2n_2})$, where $r_{(0i)j}$, r_{ij} denote the remain-

ing number of phases of the customer at position j in queues $0i$ and i, respectively, where queue $0i$ contains the customers in queue 0 allocated to queue i. Upon completion of an exponential phase, the customer moves to the next phase, so that the remaining number of phases, $r_{(0i)j}$ or r_{ij}, for that customer decreases by one unit. When $r_{(0i)j} = 1$ the customer leaves the network and for $r_{ij} = 1$ the customer moves from queue i to queue 0 as part of customers allocated to queue i, i.e., n_{0i} increases by 1 customer. As servers are indistinguishable, customers may be placed in arbitrary order in the queues. We will model this by selecting an arbitrary location for a customer arriving to a queue.

Before introducing the transition rates, we need additional notation. Let $r + r_{(0i)j}$, or $r + r_{ij}$ denote the state obtained from state r due to an arrival at queue $0i$ or i that is placed in position j with $r_{(0i)j}$ or r_{ij} phases of service, $i = 1, 2$. The customers in positions $j, \ldots, n_{0i}$, or $j, \ldots, n_i$ now receive the labels $j+1, \ldots, n_{0i}+1$, or $j+1, \ldots, n_i+1$. Note that an arrival to $0i$ is possible only if $n_i = S_i$. Let $r - r_{(0i)j}$ denote the state obtained from state r due to a departure from queue $0i$ from position j (note that this requires that $r_{(0i)j} = 1$). The customers in positions $j+1, \ldots, n_{0i}$ now receive the labels $j, \ldots, n_{0i}-1$. Let $(r; r_{(0i)j} - 1)$, or $(r; r_{ij} - 1)$ denote the state obtained from state r due to the completion of a phase of the customer in position j at queue $0i$ or i, i.e., due to $r_{(0i)j} \to r_{(0i)j} - 1$, or $r_{ij} \to r_{ij} - 1$, and $(r; r_{(0i)j} + 1)$, or $(r; r_{ij} + 1)$ the state that yields state r due to completion of a phase of the customer in position j at queue $0i$, or i. Let $r - r_{ik} + r_{(0i)j}$ denote the state obtained from state r due to the departure of a customer from queue i in position k ($r_{ik} = 1$) that moves to position j in queue $(0i)$ with $r_{(0i)j}$ phases. The transition rates are, for $i = 1, 2$,

$$
\begin{aligned}
q(r, r + r_{ij}) &= \lambda_i p_{i,r_{ij}} \frac{1}{n_i + 1} \mathbb{1}(n_i < S_i,\ n_{01} + n_{02} + n_1 + n_2 < S_0 + S_1 + S_2), \\
q(r, r + r_{(0i)j}) &= \lambda_i p_{0,r_{(0i)j}} \frac{1}{n_{0i} + 1} \mathbb{1}(n_i = S_i,\ n_{01} + n_{02} + n_1 + n_2 < S_0 + S_1 + S_2), \\
q(r, r - r_{ik} + r_{(0i)j}) &= \nu_i p_{0,r_{(0i)j}} \frac{1}{n_{0i} + 1} \mathbb{1}(r_{ik} = 1), \\
q(r, r - r_{(0i)j}) &= \nu_0 \mathbb{1}(r_{(0i)j} = 1), \\
q(r, (r; r_{ij} - 1)) &= \nu_i \mathbb{1}(r_{ij} > 1), \\
q(r, (r; r_{(0i)j} - 1)) &= \nu_0 \mathbb{1}(r_{(0i)j} > 1).
\end{aligned}
$$

Theorem 2 *The steady-state distribution is*

$$\pi(r) = \frac{1}{G}\left(\prod_{i=1}^{2}\left(\frac{\lambda_i}{\mu_0}\right)^{n_{0i}}\frac{1}{n_{0i}!}\prod_{j=1}^{n_{0i}}H_0(r_{(0i)j})\right)\left(\prod_{i=1}^{2}\left(\frac{\lambda_i}{\mu_i}\right)^{n_i}\frac{1}{n_i!}\prod_{j=1}^{n_i}H_i(r_{ij})\right),$$

$$G = \sum_{n\in S}\left(\prod_{i=1}^{2}\left(\frac{\lambda_i}{\mu_0}\right)^{n_{0i}}\frac{1}{n_{0i}!}\right)\left(\prod_{i=1}^{2}\left(\frac{\lambda_i}{\mu_i}\right)^{n_i}\frac{1}{n_i!}\right),$$

$$H_i(k) = \frac{\mu_i}{\nu_i}\sum_{t=k}^{\infty}p_{i,t}, \quad k = 1,2,\ldots,\ i = 0,1,2.$$

Moreover,

$$\pi(n) = \frac{1}{G}\left(\prod_{i=1}^{2}\left(\frac{\lambda_i}{\mu_0}\right)^{n_{0i}}\frac{1}{n_{0i}!}\right)\left(\prod_{i=1}^{2}\left(\frac{\lambda_i}{\mu_i}\right)^{n_i}\frac{1}{n_i!}\right), \quad n \in S.$$

Proof: Observe that $H_i(k)$ may be interpreted as the steady-state distribution of the renewal process obtained from the phase-type distribution when transitions from phase 1 to phase t are introduced with probability $p_{i,t}$, see e.g. [31]. Then, for $i = 1, 2$, $H_i(k)$ is the unique solution of the equations

$$H_i(k) = H_i(1)p_{i,k} + H_i(k+1), \quad k = 1,2,\ldots, \tag{17}$$

$$H_i(1) = \frac{\mu_i}{\nu_i}. \tag{18}$$

This can readily be verified by insertion of $H_i(k)$.

The Markov chain is irreducible and regular. Therefore, the steady-state distribution is the unique solution of the global balance equations

$$\sum_{r'}\{\pi(r)q(r,r') - \pi(r')q(r',r)\} = 0.$$

Inserting the transition rates and the proposed steady-state distribution into the global balance equations

gives the following.

$$
\begin{aligned}
\pi(r)\sum_{i=1}^{2}\Bigg(&\sum_{j=1}^{n_i+1}\sum_{r_{ij}=1}^{\infty} q(r,r+r_{ij}) + \sum_{j=1}^{n_{0i}+1}\sum_{r_{(0i)j}=1}^{\infty} q(r,r+r_{(0i)j}) + \sum_{k=1}^{n_i}\sum_{j=1}^{n_{0i}+1}\sum_{r_{(0i)j}=1}^{\infty} q(r,r-r_{ik}+r_{(0i)j}) \\
&+\sum_{j=1}^{n_i} q(r,r-r_{(0i)j}) + \sum_{j=1}^{n_i} q(r,(r;r_{ij}-1)) + \sum_{j=1}^{n_{0i}} q(r,(r;r_{(0i)j}-1))\Bigg) \\
-\sum_{i=1}^{2}\Bigg(&\sum_{j=1}^{n_i}\pi(r-r_{ij})q(r-r_{ij},r) + \sum_{j=1}^{n_{0i}}\pi(r-r_{(0i)j})q(r-r_{(0i)j},r) \\
&+\sum_{k=1}^{n_i+1}\sum_{j=1}^{n_{0i}}\pi(r-r_{(0i)j}+r_{ik})q(r-r_{(0i)j}+r_{ik},r) + \sum_{j=1}^{n_{0i}+1}\pi(r+r_{(0i)j})q(r+r_{(0i)j},r) \\
&+\sum_{j=1}^{n_{0i}}\pi((r;r_{(0i)j}+1))q((r;r_{(0i)j}+1),r) + \sum_{j=1}^{n_i}\pi((r;r_{ij}+1))q((r;r_{ij}+1),r)\Bigg) \\
=\pi(r)\sum_{i=1}^{2}\Bigg(&(\lambda_i\mathbb{1}(n_i<S_i)+\lambda_i\mathbb{1}(n_i=S_i))\,\mathbb{1}(n_{01}+n_{02}+n_1+n_2<S_0+S_1+S_2) \\
&+\sum_{j=1}^{n_i}\pi(r)\nu_i\mathbb{1}(r_{ij}=1) + \sum_{j=1}^{n_{0i}}\pi(r)\nu_0\mathbb{1}(r_{((0i)j}=1) \\
&+\sum_{j=1}^{n_i}\pi(r)\nu_i\mathbb{1}(r_{ij}>1) + \sum_{j=1}^{n_{0i}}\pi(r)\nu_0\mathbb{1}(r_{(0i)j}>1)\Bigg) \\
-\sum_{i=1}^{2}\Bigg(&\sum_{j=1}^{n_i}\pi(r-r_{ij})\lambda p_{i,r_{ij}}\frac{1}{n_i} - \sum_{j=1}^{n_{0i}}\pi(r-r_{(0i)j})\lambda p_{0,r_{(0i)j}}\frac{1}{n_{0i}}\mathbb{1}(n_i=S_i) \\
&-\sum_{k=1}^{n_i+1}\sum_{j=1}^{n_{0i}}\pi(r-r_{(0i)j}+r_{ik})\nu_i p_{0,r_{(0i)j}}\frac{1}{n_{0i}}\mathbb{1}(r_{ik}=1)\mathbb{1}(n_i<S_i) \\
&-\sum_{j=1}^{n_{0i}+1}\pi(r+r_{(0i)j})\nu_0\mathbb{1}(r_{(0i)j}=1)\mathbb{1}(n_{01}+n_{02}+n_1+n_2<S_0+S_1+S_2) \\
&-\sum_{j=1}^{n_i}\pi((r;r_{ij}+1))\nu_i - \sum_{j=1}^{n_{0i}}\pi((r;r_{(0i)j}+1))\nu_0\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
= \sum_{i=1}^{2} \Bigg(& \pi(r)\lambda_i \mathbb{1}(n_{01}+n_{02}+n_1+n_2 < S_0+S_1+S_2) + \sum_{j=1}^{n_i} \pi(r)\nu_i + \sum_{j=1}^{n_{0i}} \pi(r)\nu_0 \\
& - \pi(r) \sum_{j=1}^{n_i} \frac{\mu_i}{\lambda_i} \frac{1}{H_i(r_{ij})} \lambda_i p_{i,r_{ij}} - \pi(r) \sum_{j=1}^{n_{0i}} \frac{\mu_0}{\lambda_i} \frac{1}{H_0(r_{(0i)j})} \lambda_i p_{0,r_{(0i)j}} \mathbb{1}(n_i = S_i) \\
& - \pi(r) \sum_{k=1}^{n_i+1} \sum_{j=1}^{n_{0i}} \frac{\lambda_i}{\mu_i} \frac{\mu_0}{\lambda_i} \frac{1}{n_i+1} \frac{H_i(r_{ik})}{H_0(r_{(0i)j})} \nu_i p_{0,r_{(0i)j}} \mathbb{1}(r_{ik}=1) \mathbb{1}(n_i < S_i) \\
& - \pi(r) \sum_{j=1}^{n_{0i}+1} \frac{\lambda_i}{\mu_0} \frac{1}{n_{0i}+1} H_0(r_{(0i)j}) \nu_0 \mathbb{1}(r_{(0i)j}=1) \mathbb{1}(n_{01}+n_{02}+n_1+n_2 < S_0+S_1+S_2) \\
& - \pi(r) \sum_{j=1}^{n_i} \frac{H_i(r_{ij}+1)}{H_i(r_{ij})} \nu_i - \pi(r) \sum_{j=1}^{n_{0i}} \frac{H_i(r_{(0i)j}+1)}{H_0(r_{(0i)j})} \nu_0 \Bigg) \\
= \sum_{i=1}^{2} \Bigg(& \pi(r)\lambda_i \mathbb{1}(n_{01}+n_{02}+n_1+n_2 < S_0+S_1+S_2) + \sum_{j=1}^{n_i} \pi(r)\nu_i + \sum_{j=1}^{n_{0i}} \pi(r)\nu_0 \\
& - \pi(r) \sum_{j=1}^{n_i} \mu_i \frac{1}{H_i(r_{ij})} p_{i,r_{ij}} - \pi(r) \sum_{j=1}^{n_{0i}} \mu_0 \frac{1}{H_0(r_{(0i)j})} p_{0,r_{(0i)j}} \\
& - \pi(r)\lambda_i \mathbb{1}(n_{01}+n_{02}+n_1+n_2 < S_0+S_1+S_2) \\
& - \pi(r) \sum_{j=1}^{n_i} \frac{H_i(r_{ij}+1)}{H_i(r_{ij})} \nu_i - \pi(r) \sum_{j=1}^{n_{0i}} \frac{H_i(r_{(0i)j}+1)}{H_0(r_{(0i)j})} \nu_0 \Bigg)
\end{aligned}
$$

We have used $H_i(1) = \frac{\mu_i}{\nu_i}$ in the $\mathbb{1}(r_{ij}=1)$ terms and we have simplified and combined a number of terms. Now observe that the term $\sum_{i=1}^{2} \pi(r)\lambda_i \mathbb{1}(n_{01}+n_{02}+n_1+n_2 < S_0+S_1+S_2)$, representing arrivals to the network and departures from the network, cancels. Invoking (17) and (18) shows that the above equation equals zero, so that global balance is indeed satisfied. The second assertion of the theorem follows by aggregation over all r such that $n = (n_{01}, n_{02}, n_1, n_2)$. ■

Remark 1 *The results of this section readily generalise to a multi-echelon system in which each warehouse serves a number of subwarehouses. The resulting tree-like multi-echelon system with nested jump-over blocking has a product-form steady-state distribution that is insensitive to the distributions of the leadtimes except for their means.*

5 Optimization and numerical results

In this section we return to the optimization problem of the multi-echelon spare part inventory system as formulated in Section 2. In the previous sections, we approximated the steady-state distribution of our system. From this we readily derive all parameters required for a complete specification of the optimization problem. Our aim is to minimize the average costs, which depend on the basestock levels of the central warehouse and the local warehouses.

The optimization problem can readily be solved by complete enumeration in the total basestock level $S_{tot} = \sum_{j=0}^{J} S_j$. To this end, let $h_{\min} = \min(h_0, h_1, ..., h_J)$ be the minimum inventory costs. An obvious lower bound on the total costs with total basestock S_{tot} is $h(S_{tot}) = S_{tot} h_{\min} + \sum_{i=1}^{J} \lambda_i T_{l,i} p_i$, which ignores all costs except for the minimum holding costs and delay costs. Notice that $h(S_{tot})$ is linearly increasing in the total basestock level. Furthermore, if $h(S'_{tot}) > g(S_0, ..., S_J)$ for some level $S'_{tot} > S_{tot}$, then $g(s_0, ..., s_J) \geq h(S'_{tot}) > g(S_0, ..., S_J)$ for all $s_0, ..., s_J$ such that $s_0 + \cdots + s_J \geq S'_{tot}$. That is, the minimum cannot be attained for basestock levels resulting in a total basestock exceeding S'_{tot}. As a consequence, enumeration in S_{tot} yields the optimum basestock levels.

We now proceed as follows. Let $g_{S_{tot}}$ denote the minimum costs for the optimization problem with total basestock S_{tot}. First, evaluate g_0, for which $S_0 = ... = S_J = 0$. Let $g^* = g_0$. Now consider $S_{tot} = 1$, with minimal costs g_1. If $g_1 < g^*$ then set $g^* = g_1$, with corresponding basestock levels $(S_0^*, ..., S_J^*)$. Proceed with $S_{tot} = 2, 3, ...$ until $h(S_{tot}) > g^*$. Then g^* is the minimum cost level.

Using this heuristic, we ran a numerical experiment for a two-echelon system with one central warehouse and three local warehouses, i.e., $J = 3$. The values of the parameters are displayed in Table 1 and correspond to a typical setting in the high-tech industry. We assume the replenishment costs for the local warehouses are identical. This implies that the costs for local warehouse i may be calculated from the fractions $\beta_{a,i} = \sum_{k \neq i} \beta_{a,i,k}$ of lateral transshipments since $\sum_{j \neq i} \beta_{a,i,j} c_j^{repl} = c_1^{repl} \beta_{a,i}$. In our experiments we varied the values of the demand rates (λ_1, λ_2, and λ_3), the penalty costs p_i of the delay, and the inventory holding costs h_i. For multiple combinations we obtained the basestock levels that minimize the average costs per time unit, using simulation for optimal results and our heuristic for approximations. In the simulation a lateral transshipment is performed by the local warehouse with the largest inventory level; if there are multiple such warehouses, the one with the smallest index is selected.

Our results are displayed in Tables 2 and 3. In the column 'optimal by simulation' the optimal basestock

par.	default			par.	default		
$T_{l,i}$	4	hours	$i = 1, 2, 3$	λ_1	0.1	demands/week	
$T_{c,i}$	24	hours	$i = 1, 2, 3$	λ_2	0.2	demands/week	
$T_{a,i}$	36	hours	$i, j = 1, 2, 3,\ i \neq j$	λ_3	0.3	demands/week	
$T_{s,i}$	48	hours	$i = 1, 2, 3$	h_i	200	Euros/part/week	$i = 1, 2, 3$
$c_{l,i}$	400	Euros	$i = 1, 2, 3$	p_i	1000	Euros/hour	$i = 1, 2, 3$
$c_{c,i}$	1000	Euros	$i = 1, 2, 3$	$1/\mu_i$	1	weeks/order	$i = 1, 2, 3$
$c_{a,i}$	2500	Euros	$i, j = 1, 2, 3,\ i \neq j$	$1/\mu_0$	10	weeks/order	
$c_{s,i}$	4000	Euros	$i = 1, 2, 3$				
c_i^{repl}	100	Euros	$i = 1, 2, 3$				
c_0^{rep}	1000	Euros					

Table 1: Default values of the parameters in the experiments.

levels (S_{opt}) are stated with the corresponding 95% confidence interval for the costs (g_{opt}) as found by simulation. In the column 'optimal approximated' our heuristic is used to approximate the optimal basestock levels (S_{app}) and corresponding costs (g_{app}). Further, given these basestock levels S_{app} confidence intervals for its true costs g_{sim} as found by simulation are stated in the final column. The costs are rounded to two decimals, and the delivery fractions to three decimals.

Table 2 displays the results for varying demand rates for the local warehouses. The costs for delay and the inventory holding costs are equal to the default values in Table 1. The results of the heuristic are close to those of the simulation. The true costs g_{sim} of the heuristic solutions differ at most 4.5% from the optimal true costs g_{opt}.

Higher demand rates result in larger basestock levels for the warehouses. This increase is not linear. For instance, if the demand rate increases from 0.05 to 0.30 by a factor of 6, then the basestock levels at the warehouses increase by a factor of approximately 3. Further, in case of higher demand rates, the warehouses keep more stock and are more often able to respond to requests from their customers; the delivery fractions $\beta_{l,i}$ are larger. Also, emergency shipments β_s by the external supplier are required less often. For small demand rates, emergency shipments are fulfilled more often by the central warehouse than by lateral transshipments. On the other hand, for larger demand rates, the central warehouse is more often out of stock. Then emergency shipments are fulfilled more often by lateral transshipments than by the central warehouse. Notice that the fractions of delivery by the external supplier are symmetric among the local warehouses, even in case of nonsymmetric demand rates.

Table 3 shows the results for varying delay costs, and varying holding costs. The demand rates are equal to the default values in Table 1. Also here, the approximation gives good results. The approximated optimal basestock levels S_{app} result in true costs g_{sim} that differ at most 1.3% from the optimal true costs g_{opt}.

	optimal by simulation		optimal approximated		by simulation
$(\lambda_1, \lambda_2, \lambda_3)$	S_{opt}	g_{opt}	S_{app}	g_{app}	g_{sim}
$(0.05, 0.05, 0.05)$	$(4, 1, 1, 1)$	2480.24 ± 3.76	$(3, 1, 1, 1)$	2331.97	2495.70 ± 4.78
$(0.07, 0.07, 0.07)$	$(3, 2, 2, 2)$	3160.63 ± 4.31	$(4, 1, 1, 1)$	3060.19	3304.50 ± 6.16
$(0.10, 0.10, 0.10)$	$(5, 2, 2, 2)$	4075.16 ± 3.52	$(4, 2, 2, 2)$	3925.79	4127.60 ± 7.24
$(0.15, 0.15, 0.15)$	$(8, 2, 2, 2)$	5580.24 ± 8.34	$(7, 2, 2, 2)$	5373.39	5612.56 ± 11.11
$(0.20, 0.20, 0.20)$	$(8, 3, 3, 3)$	7073.40 ± 7.46	$(9, 2, 2, 2)$	6833.53	7260.12 ± 11.76
$(0.30, 0.30, 0.30)$	$(13, 3, 3, 3)$	9869.60 ± 17.67	$(12, 3, 3, 3)$	9549.06	9984.16 ± 16.40
$(0.40, 0.40, 0.40)$	$(16, 4, 4, 4)$	12638.44 ± 19.92	$(17, 3, 3, 3)$	12257.02	12762.32 ± 21.85
$(0.05, 0.20, 0.30)$	$(9, 1, 2, 3)$	6523.84 ± 9.39	$(8, 1, 2, 3)$	6296.41	6653.56 ± 9.16
$(0.10, 0.20, 0.30)$	$(9, 2, 3, 3)$	7043.52 ± 9.50	$(9, 2, 2, 3)$	6826.52	7096.52 ± 11.41

$(\lambda_1, \lambda_2, \lambda_3)$	β_l	β_c	β_a	β_s
$(0.05, 0.05, 0.05)$	$(0.927, 0.927, 0.927)$	$(0.039, 0.039, 0.039)$	$(0.030, 0.030, 0.030)$	$(0.005, 0.005, 0.005)$
$(0.07, 0.07, 0.07)$	$(0.910, 0.910, 0.910)$	$(0.055, 0.055, 0.055)$	$(0.029, 0.029, 0.029)$	$(0.007, 0.007, 0.007)$
$(0.10, 0.10, 0.10)$	$(0.972, 0.972, 0.972)$	$(0.003, 0.003, 0.003)$	$(0.023, 0.023, 0.023)$	$(0.002, 0.002, 0.002)$
$(0.15, 0.15, 0.15)$	$(0.978, 0.978, 0.978)$	$(0.008, 0.008, 0.008)$	$(0.013, 0.013, 0.013)$	$(0.001, 0.001, 0.001)$
$(0.20, 0.20, 0.20)$	$(0.969, 0.969, 0.969)$	$(0.014, 0.014, 0.014)$	$(0.015, 0.015, 0.015)$	$(0.002, 0.002, 0.002)$
$(0.30, 0.30, 0.30)$	$(0.986, 0.986, 0.986)$	$(0.003, 0.003, 0.003)$	$(0.010, 0.010, 0.010)$	$(0.001, 0.001, 0.001)$
$(0.40, 0.40, 0.40)$	$(0.987, 0.987, 0.987)$	$(0.006, 0.006, 0.006)$	$(0.007, 0.007, 0.007)$	$(0.001, 0.001, 0.001)$
$(0.05, 0.20, 0.30)$	$(0.935, 0.963, 0.983)$	$(0.039, 0.013, 0.003)$	$(0.024, 0.021, 0.012)$	$(0.002, 0.002, 0.002)$
$(0.10, 0.20, 0.30)$	$(0.990, 0.969, 0.987)$	$(0.004, 0.014, 0.003)$	$(0.005, 0.016, 0.009)$	$(0.001, 0.001, 0.001)$

Table 2: The upper table shows optimal and approximated basestock levels and minimal costs for varying demand rates. The lower table shows the approximated delivery fractions.

Increasing the delay cost provides an incentive to have faster deliveries, and results in larger basestock levels for the local warehouses. This implies that fractions of delivery $\beta_{l,i}$ by these warehouses increase, and emergency shipments by the external supplier occur less often. Lateral transshipments remain. Since the basestock level of the central warehouse remains the same while the basestock levels of the local warehouses are larger, emergency shipments by the central warehouse also occur less often.

If the holding costs h_i are low compared to the delay costs p_i, then it is relatively cheap to keep the goods in stock, while it is more expensive to have a delayed delivery to the customer. Hence, the basestock levels are larger. Consequently, the local warehouses are more often able to respond to requests from their customers; the delivery fractions $\beta_{l,i}$ are higher. Emergency requests are needed less often.

If the holding costs increase, it is more expensive to keep stock. The basestock levels decrease, resulting in lower fractions of delivery for the local warehouses. Emergency shipments by the central warehouse, by the external supplier, and lateral transshipments occur more often. Also here, the fractions of delivery by the external supplier are symmetric.

The tables show that our heuristic based on the product-form approximation performs well. Furthermore,

		optimal by simulation		optimal approximated		by simulation
h_i	p_i	S_{opt}	g_{opt}	S_{app}	g_{app}	g_{sim}
200	500	$(9,1,2,3)$	5592.48 ± 9.13	$(8,1,2,3)$	5364.71	5665.12 ± 9.31
200	1000	$(9,2,3,3)$	7043.52 ± 9.50	$(9,2,2,3)$	6826.52	7096.52 ± 11.41
200	2000	$(9,2,3,4)$	9666.28 ± 12.09	$(8,2,3,4)$	9444.44	9795.96 ± 13.48
50	1000	$(10,2,3,4)$	4346.72 ± 5.54	$(9,2,3,4)$	4290.58	4382.52 ± 5.55
500	1000	$(9,1,2,3)$	11780.28 ± 16.70	$(8,1,2,3)$	11183.61	11823.88 ± 18.88
1000	1000	$(7,1,2,3)$	18658.36 ± 26.99	$(7,1,2,2)$	17463.57	18764.00 ± 28.47

h_i	p_i	β_l	β_c	β_a	β_s
200	500	$(0.869, 0.955, 0.978)$	$(0.068, 0.012, 0.003)$	$(0.059, 0.028, 0.015)$	$(0.005, 0.005, 0.005)$
200	1000	$(0.990, 0.969, 0.987)$	$(0.004, 0.014, 0.003)$	$(0.005, 0.016, 0.009)$	$(0.001, 0.001, 0.001)$
200	2000	$(0.986, 0.991, 0.993)$	$(0.003, 0.001, 0.000)$	$(0.011, 0.008, 0.006)$	$(0.000, 0.000, 0.000)$
50	1000	$(0.991, 0.995, 0.997)$	$(0.004, 0.001, 0.000)$	$(0.005, 0.004, 0.003)$	$(0.000, 0.000, 0.000)$
500	1000	$(0.869, 0.955, 0.978)$	$(0.068, 0.012, 0.003)$	$(0.059, 0.028, 0.015)$	$(0.005, 0.005, 0.005)$
1000	1000	$(0.838, 0.931, 0.884)$	$(0.056, 0.010, 0.021)$	$(0.087, 0.040, 0.076)$	$(0.019, 0.019, 0.019)$

Table 3: The upper table shows optimal and approximated basestock levels and minimal costs for varying holding and delay costs with $(\lambda_1, \lambda_2, \lambda_3) = (0.10, 0.20, 0.30)$. The lower table shows the approximated delivery fractions.

the approximated basestock levels S_{app} result in costs g_{sim} that are close to the optimal costs g_{opt}.

Our heuristic is based on enumeration, which appears to be fast enough for the instances considered in this paper. For larger instances, e.g., instances with more local warehouses or higher basestock levels (needed when demand rates are higher), the computation time will become too high, and then the enumeration could be replaced by a local search procedure. We expect that then still a good heuristic is obtained. (In several spare parts inventory optimization problems with service-level constraints, greedy heuristics, which may be seen as a particular type of local search algorithms, have been shown to perform well; see [6].)

6 Conclusion

In this paper, we have considered a two-echelon inventory model for spare parts consisting of one central warehouse, one central repair facility, and multiple local warehouses. Because our inventory system is too complex to solve for a steady-state distribution in closed form, we approximate it by a network of Erlang loss queues with so-called hierarchical or nested jump-over blocking. We show that under a given basestock policy this network has a steady-state distribution in product-form.

The closed-form solution enables an efficient heuristic for the approximation of the basestock levels and costs of the inventory model. This heuristic gives good results. Also, the steady-state distribution and

several relevant approximating performance measures only depend on the distributions of the repair and replenishment leadtimes via their means, i.e., they are insensitive to the underlying distributions.

Similar to the approach demonstrated in this paper, one could solve single-item optimization problems with service level constraints. In future research, we plan to exploit the results of this paper in multi-item optimization problems and in networks with both emergency shipments and lateral transshipments. For these problems, it is interesting to develop efficient and effective heuristics for the approximation of multi-item optimization procedures (such as greedy approaches and Lagrangian heuristics).

References

[1] Albin, L. 1982. On Poisson approximations for superposition arrival processes in queues. *Management Science 28*, 126-137.

[2] Alfredsson, P., and Verrijdt, J. 1999. Modeling emergency supply flexibility in a two-echelon inventory system. *Management Science 45*, 1416-1431.

[3] Axsäter, S. 1990. Modelling emergency lateral transshipments in inventory systems. *Management Science 36*, 1329-1338.

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

[5] Baskett, F., Chandy, K.M., Muntz, R.R., and Palacios, F.G. 1975. Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM 22*, 248-260.

[6] Basten, R.J.I., Van Houtum, G.J. 2014. System-oriented inventory models for spare parts. *Surveys in Operations Research and Management Science 19*, 34-55.

[7] Bijvank, M., and Vis, I.F.A. 2011. Lost-sales inventory theory: A review. *European Journal of Operational Research 215* (1), 1-13.

[8] Boucherie, R.J. 1996. Batch routing queueing networks with jump-over blocking. *Probability in the Engineering and Informational Sciences 10*, 287-297.

[9] Boucherie, R.J., Mandjes, M. 1998. Estimation of performance measures for product form cellular mobile communications networks. *Telecommunication Systems 10*, 321-354.

[10] Boucherie, R.J., and van Dijk, N.M. (Eds) 2011. *Queueuing Networks: A Fundamental Approach*, International Series in Operations Research & Management Science 154, Springer, New York.

[11] Caggiano, K.E., Jackson, P.L., Muckstadt, J.A., and Rappold, J.A. 2007. Optimizing service parts inventory in a multi-echelon, multi-item supply chain with time-based customer service level agreements. *Operations Research 55* (2), 303-318.

[12] Caggiano, K.E., Jackson, P.L., Muckstadt, J.A., and Rappold, J.A. 2009. Efficient computation of time-based customer service levels in a multi-item, multi-echelon supply chain: A practical approach for inventory optimization. *European Journal of Operational Research, 199* (3), 744-749.

[13] Grahovac, J., and Chakravarty, A. 2001. Sharing and lateral transshipments of inventory in a supply chain with expensive low-demand items. *Management Science 47*, 579-594.

[14] Graves, S.C. 1985. A multi-echelon inventory model for a repairable item with one-for-one replenishment. *Management Science 31*, 1247-1256.

[15] Hausman, W.H., and Erkip, N.K. 1994. Multi-echelon vs. single-echelon inventory control policies for low demand items. *Management Science 40*, 597-602.

[16] Hordijk, A., and Schassberger, R. 1982. Weak convergence for generalized semi-Markov processes. *Stochastic Processes and Their Applications 12*, 271-291.

[17] Kruse, K.C. 1984. An exact N echelon inventory model: The simple Simon method, U.S. Army Research Office, Technical Report TR 79-2.

[18] Lee, H.L. 1987. A multi-echelon inventory model for repairable items with emergency lateral transshipments. *Management Science 33*, 1302-1316.

[19] Muckstadt, J.A. 1973. A Model for a multi-item, multi-echelon, multi-indenture inventory system. *Management Science 20*, 472-481.

[20] Muckstadt, J.A., and Thomas, L.J. 1980. Are multi-echelon inventory methods worth implementing in systems with low-demand-rate items? *Management Science 26*, 483-494.

[21] Muckstadt, J.A., and Sapra, A. 2009. *Principles of Inventory Management: When You are Down to Four, Order More.* Springer, New York.

[22] Özkan, E., Van Houtum, G.J., and Serin Y. 2015. A New Approximate Evaluation Method for Two-Echelon Inventory Systems with Emergency Shipments. *Annals of Operations Research 224*, 147-169.

[23] Paterson, C., Kiesmüller, G., Teunter, R., and Glazebrook, K. 2011. Inventory models with lateral transshipments: A review. *European Journal of Operational Research, 210* (2), 125-136.

[24] Rustenburg, J.W., Van Houtum, G.J., and Zijm, W.H.M. 2003. Exact and approximate analysis of multi-echelon, multi-indenture spare parts systems with commonality. Chapter 7 in Shantikumar, J.G., Yao, D.D. and Zijm, W.H.M. (Eds.), *Stochastic Modeling and Optimization of Manufacturing Systems and Supply Chains*, Kluwer, Boston.

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

[26] Sherbrooke, C.C. 1986. VARI-METRIC: Improved approximations for multi-indenture, multi-echelon availability models. *Operations Research 34*, 311-319.

[27] Sherbrooke, C.C. 1992. Multi-echelon inventory systems with lateral supply. *Naval Research Logistics 39*, 29-40.

[28] Selçuk, B. 2013. An adaptive base stock policy for repairable item inventory control. *International Journal of Production Economics 143*, 304-315.

[29] Simon, R.M. 1971. Stationary properties of a two-echelon inventory model for low demand items. *Operations Research 19*, 761-773.

[30] Slay, F.M. 1984. VARI-METRIC: An approach to modeling multi-echelon resupply when the demand process is Poisson with a Gamma prior. Report AF301-3, Logistics Management Institute, Washington D.C.

[31] Van Dijk, N.M. 1993. *Queueing Networks and Product Forms: A System's Approach*. Wiley, New York.

[32] Wong, H., Kranenburg, A.A., Van Houtum, G.J., and Cattrysse, D. 2007. Efficient heuristics for two-echelon spare parts inventory systems with an aggregate mean waiting time constraint per local warehouse. *OR Spectrum 29*, 699-722.

[33] Wong, H., Van Oudheusden, D., and Cattrysse, D. 2007. Two-echelon multi-item spare parts systems with emergency supply flexibility and waiting time constraints. *IIE Transactions 39* (11), 1045-1057.