Two-Echelon Inventory Model With Service Consideration and Lateral Transshipment

SAMUEL CHIABOM ZELIBE
Department of Mathematics, Federal University of Petroleum Resources, Effurun
NIGERIA

UNANAOWO NYONG BASSEY
Department of Mathematics, University of Ibadan, Ibadan
NIGERIA

Abstract: This paper considers a two-echelon inventory system with service consideration and lateral transshipment. So far, researchers have not extensively considered the use of lateral transshipment for such systems. Demand arrivals at both echelons follow the Poisson process. We introduce a continuous review base stock policy for the system in steady state, which determined the expected level for on-hand inventory, expected lateral transshipment level and expected backorder level. We showed that the model satisfied convexity with respect to base stock level. Computational experiments showed that the model with lateral transshipment performed better that the model without lateral transshipment.

Key-Words: - Inventory, Lateral Transshipment, Base Stock Policy, Service Constraint, Replenishment Systems, Multi-Echelon, Supply Chain.

Received: March 17, 2021. Revised: May 4, 2021. Accepted: May 14, 2021. Published: May 24, 2021.

1 Introduction

Lateral Transshipment (LT) is the transfer of items between service centers on same echelon. Research on multi-echelon inventory systems with Lateral Transshipment (LT) is mainly motivated by firms in the manufacturing industries where the desire exists to simultaneously minimise cost and improve service. Cost minimisation and service improvement are two desirable but contrasting objectives in service parts supply systems. The importance of having an appropriate inventory policy cannot be overemphasized [1]. The objective of cost minimisation could lead to reduction in customer service level, while the objective of service improvement could lead to increase in total system costs. Thus, decision makers need to always strike the right balance between these two contrasting objectives. This will help to prevent instances where minimising cost results to deteriorating customer service, or instances where improving customer service results to shrinking profit margins due to rising system costs. Consequently, in order for the system to remain sustainable, the decision maker is required to always make accurate or near accurate decisions. This highlights the need for continuous research on more efficient means of balancing the conflicting objectives of service improvement and cost minimisation. This study explores the use of LT as a means to achieving a balance between cost minimisation and service improvement for an inventory system with service consideration in a two-echelon arena.

This work is applicable in spare parts supply chain, especially for firms that deal with parts which are expensive and slow moving. Here, an order for a spare part from a customer also implies that the customer is desperate and needs to get his machine to full functionality as soon as possible. The sensitivity of customers to service time imply that in order to retain customers, spare parts dealers would continually seek means to lower inventory costs without violating their customers' waiting time preferences.

The consequences of using LT has been studied for many multi echelon environments. However, the use of LT in an inventory system with uniform service constraint across all lower echelon facilities in a two-echelon environment has not been considered extensively. [2], [3], [4], [5] and [6] are some two-echelon inventory systems with service considerations for which the effect of LT is yet to be studied. Thus, this study was designed to integrate LT into a two-echelon inventory system with uniform service consideration across all service centers. This work can be treated as a natural extension of the work by [2], the integration of LT leads to a change in expressions for steady state levels of inventory on hand and backorder. In this study, the service constraint or service consideration is a common threshold for response times across all service centers. A measure of the system's service quality is the waiting time of customers, thus, our service constraint demands that waiting time does not exceed a given threshold. This follows because generally, customers desire to have short service times for service parts. The desire to minimise cost is taken care of by the system's objective function. Thus, we aim to determine an inventory policy that minimises system wide costs for a two-echelon inventory system with LT, subject to a constraint on response time.

Our two-echelon inventory system is made up of a set of Service Centers (SVCs) at the lower echelon and the top echelon consists of a single plant. A Service Center (SVC) satisfies demand from geographically spaced customers, while the plant produces and stores items to replenish SVC stock. The SVCs are divided into pools such that LT is allowed only between SVCs in the same pool. Customer demand at each SVC is met from on hand inventory if the SVC has available stock. If the SVC has a stockout occurrence, the demand is satisfied by LT from a SVC in same pool. Demand is backordered if all SVCs in the pool are out of stock. The sensitivity of customers to response times makes LT a more preferable option than direct shipment from plant. This is because a SVC which runs out of stock may be closer to other nearby SVCs than the plant, thus, satisfying incoming demands via LT might be quicker than satisfying them via direct shipment from plant.

This study is a fusion of two major areas: (i) two-echelon inventory problems with service consideration and (ii) lateral transshipment. The use of time-based service constraints have been considered in the literature of inventory systems with two echelons. [2] considered a two-echelon model with service constraint, which controlled inventory at both echelons using (S-1,S) policies. They imposed an upper bound on the expected response time and created efficient algorithms to minimise stock related costs at both echelons. [7] studied a two-echelon problem on inventory location that incorporated service consideration. They modelled the manufacturing process as a queue and formulated a nonlinear mixed integer problem. They solved the problem using a Lagrange heuristic. [8] considered an inventory model with a service constraint which gave a threshold for backorder. Their model resulted to efficient convex curve for backorder costs on application of marginal analysis and greedy algorithm. [9] considered the network for service parts logistics of an aerospace firm with service requirements. Their service requirements resulted to service constraints which were highly stochastic and non-linear. They presented a solution procedure which was exact. They also presented a new decomposition scheme. Their procedure could handle cases having 60 items in reasonable time. [10] investigated scenarios in the utility industry where service measures go beyond availability of parts to consider the effect of utility downtime on customers. [11] considered inventory models with stochastic service constraints. All the papers mentioned so far considered inventory systems with service constraints; the non-inclusion of lateral transshipment is a noticeable gap for the systems considered in these papers. This makes it necessary to consider the incorporation of lateral transshipment into such systems. In this study, we focus on system wide service constraints.

There have been lots of literature that consider inventory control with lateral transshipment. [12] gives a well detailed review on lateral transshipments. They classified transshipments into proactive and reactive based on the time the lateral transshipments occur. They further classified reactive transshipment into two categories based on whether they are in centralised systems or decentralised systems. Most models with centralised systems [13], [14], [15], [16] and [17] assumed negligible transshipment times and found that lateral transshipments improved the system performance. [13] analysed a two-echelon inventory system controlled with continuous review base stock policy. The system considered had identical bases and assumed negligible transshipment. Demand occurred due to part failure, and the demand arrival process was assumed to be a Poisson process. Demand is satisfied by on-hand inventory or lateral transshipments in a stock out situation. The demand fraction satisfied from on-hand inventory and the demand fraction satisfied by LT were evaluated following three rules for the selection of LT source: random selection, maximum on-hand inventory, and smallest number of outstanding orders. The study found no significant difference in the performance of the three transshipment rules and that LT led to substantial cost savings because less base stocks were needed at the bases. [14] relaxed the assumption of identical facilities and presented better methods for determining approximate service levels. The model by [14] was extended by [15] who allowed emergency shipments from a central warehouse and a manufacturing facility such that no demand was backordered. They found that the use of LT and the flexibility of direct shipment resulted in significant reduction of cost compared with using no supply flexibility at all. [16] and [17] conducted simulation studies for negligible transshipment times and showed that a policy which permitted LT performed better than one without lateral transshipments if the benefits of avoiding facility stock out dominated the additional costs resulting from transshipments. On the other hand, some studies ([18], [19], [20], [21]) considered nonnegligible LT times in their models.

Another practical feature which is considered within a spare parts model is time based service levels, where targeted proportions of demand are to be satisfied within a certain time period. [22] considered a system with time based service level. They pointed out that in service parts system, time-based service requirements are more appropriate than fill rates, and that the performance of response times improved with LT. [23] considered a service parts location inventory problem with flexible replenishment stock and LT. They proposed a service measure which was customer oriented. They also provided an approximation for optimising inventory allocation subject to the this measure. However their response time measure did not cater for demand not met by inventory on hand or LT. The service measures considered by [22] and [23] only considered a fraction of total system demand. Thus, it was possible that a certain fraction of demand might not be satisfied within the time threshold for service. This makes it necessary to examine service measures which guarantee the satisfaction of all customers. The service constraints in this study consider all customers, thus our constraints differ greatly from those of [23] and [22]. [24] extensively reviewed literature on system-oriented inventory systems. They reviewed systems with service level constraints and systems with LT. From [12], [24], and other studies mentioned so far, there is scarcity of literature on the incorporation of LT into a two echelon inventory system with service constraint (response time requirement).

[2] used the (S-1, S) policy to control inventory and presented the steady state relationship between on-hand inventory and backorder for a two-echelon system with service constraint. This study extends their work by developing an inventory policy that incorporates LT. This was done via derivation of the steady state relationship between on-hand inventory, LT, and backorder. We also derive approximations for on-hand inventory, LT, and backorder using METRIC approximation [25]. The service constraint considered is a single response time threshold across all facilities in the lower echelon. Thus, this study established a significant contribution by enhancing the literature on two-echelon systems via the incorporation of lateral transshipment into a centralised two-echelon location-inventory system with finite number of facilities at the lower echelon and response time requirement across all facilities. We introduce a new model and also introduce steady state relationships and optimal policies for the new system.

The structure of this paper is as follows. In Section 1, the introduction is presented. In Section 2, we present the model description and formulation, and also determine the steady state expected levels for inventory, LT and backorder. In Section 3, some properties of the model are presented. In Section 4, computational experiments are presented. In Section 5 we present our conclusion.

2 Model description and Formulation

The two-echelon inventory system considered in this study is made up of a plant at the top echelon and a set of Service Centers (SVCs) at the lower echelon. A Service Center (SVC) satisfies demand from geographically spaced customers, while the plant produces and stores items to replenish SVC stock. In this section we give a description of the system, introduce the basic notations, and present the basic model formulation.

2.1 System description and notations

1. The item (spare part) is manufactured and stored at the plant to fulfil resupply requests from SVCs within a SVC specific response time.

2. Arrival of orders at a SVC are independent and follow a Poisson process. Customer orders are satisfied at the SVCs.

3. We use a (S-1, 1) policy to control inventory at both echelons. This follows because [26] showed that (S-1,S) policies are appropriate for slow moving items. This study deals with a slow moving and expensive single item inventory.

4. If a SVC has positive stock level, it immediately satisfies its arriving customer orders and instantly sends replenishment requests to the plant.

5. If stock level at a SVC (SVC A) is non-positive and there exists one or more pooled SVCs with positive stock level, then a demand arriving at SVC A is satisfied instantaneously via LT from any of its pooled SVCs with positive stock level.

6. If all SVCs in a pool have non-positive stock levels, then demand arrival at any SVC in that pool will be backordered.

7. Customer demand at a SVC are satisfied via any of the following: on hand stock, LT or backorder.

8. We assume negligible LT times. This implies that all demands satisfied from on-hand stock at a SVC or via LT were satisfied instantaneously. Thus, demand is backordered if and only if the entire pool is out of stock. Hence, the customer's waiting time can be constrained by putting a bound on the backorder time. The waiting time includes the deterministic plant to SVC transportation time and any delay at the plant.

9. If the plant has positive stock level, it fulfills replenishment request from a SVC on arrival, and instantly sends a production order to its production line.

10. If the plant has non-positive stock level, replenishment requests from the SVCs are backordered.

11. The finished products from the plant's production line are stored as inventory or used to satisfy backorders as soon as the production process ends. We assume finite servers at the production line, where each server has an exponential service rate and processes one unit at a time. Controlling each SVC with the (S-1,S) policy implies that replenishment requests at the plant arrive one at a time. Thus, the plant's demand arrival process is Poisson and the plant possesses the properties of a Markovian queue.

12. All demand and replenishment requests are handled in a First-Come,First-Served $(FCFS)$ manner.

The following are the costs considered in the system:

1. cost of storing inventories at the plant and SVCs (holding costs),

2. cost of backordering customer demand at a SVC (backorder cost)

3. cost of satisfying demand via LT (LT cost)

Below, we present notations used in this study
Sets
Y = Set of SVCs
Z = Set of Pools

Parameters
h_{yz} $=$ Per unit holding cost at SVC y in pool z per unit time
q_{yz}= Per unit lateral transshipment at SVC y in pool z
p_{yz} = Per unit backorder cost at SVC y in pool z per unit time
λ_{yz} $=$ Demand rate at SVC y in pool z
λ_z $=$ Demand rate at pool z $= \sum_{y\in Y} \lambda_{yz}$
λ_0 $=$ Demand rate at plant $= \sum_{z\in Z} \lambda_z$
τ $=$ response time threshold
ρ $=$ Plant utilisation rate$(= \frac{\lambda_0}{\mu})$
μ $=$ Plant order processing rate
α_w $=$ exact lead time from the plant to pool z
C_{yz} $=$ Capacity available at SVC y in pool z, this is uniform for all SVCs in pool z
C_z $=$ $|z|C_z$ = Pool z's total capacity, where $|z|$ represents number of SVCs in pool z
C_0 = Plant's total storage capacity

Other Decision Variables
S_{yz} is the required stock level at SVC y in pool z
$S_z = |z|S_{yz}$ is the required stock level at pool z
S_0 is the plant's required level of stock

Service Variables
I_{yz} $=$ Expected inventory level in steady state for SVC y in pool z
I_z $=$ $\sum_{y\in Y} I_{yz}$ = Pool z's expected steady state inventory level
B_{yz} $=$ Expected backorder level in steady state for SVC y in pool z
B_z $=$ Pool z's expected backorder level in steady state
T_{yz} $=$ Expected LT level in steady state for SVC y in pool z
Wt_{yz} $=$ Expected response time in steady state for SVC y in pool z
I_0 $=$ Expected inventory level in steady state at the plant
B_0 $=$ Expected backorder level in steady state at the plant
$N_{yz}(t)$ $=$ Total replenishment orders by SVC y in pool z yet to arrive by time t
$N_z(t)$ $=$ Total pool z replenishment orders yet to arrive by time t.
$N_0(t)$ $=$ Total plant replenishment orders yet to arrive by time t.

2.2 Model formulation

The basic model formulation is given below.

$$\min \sum_{z\in Z}\sum_{y\in Y}(h_{yz}I_{yz}+p_{yz}B_{yz}+q_{yz}T_{yz})+h_0I_0 \tag{1}$$

Subject to:

$$S_{yz} \le C_{yz}, \text{ for each, } y\in Y \tag{2}$$
$$S_z \le |z|C_{yz}, \text{ for each, } z\in Z \tag{3}$$
$$S_0 \le C_0 \tag{4}$$
$$Wt_{yz} \le \tau, \text{ for each, } y\in Y \tag{5}$$
$$S_{yz} \ge 0, \text{ integer, for each, } y\in Y \tag{6}$$
$$S_0 \ge 0 \tag{7}$$

The objective (1) is to determine the minimum sum of SVC inventory holding costs, backorder costs at SVCs, LT costs and plant inventory holding costs. Our system is centralised hence plant backorders are considered internal to the system and not incur a monetary cost. Constraints (2), (3) and (4) state that SVCs, pools and plant stock levels cannot be greater than available storing capacity. From our service constraints (5), the expected service time cannot exceed the specified threshold. Finally, (6) and (7) are nonnegativity constraints.

The total waiting time of a customer is a measure of the system's service quality, thus, our service constraint demands that waiting time does not exceed a given threshold. This is similar to the service constraints found in [2], [7], and [27]. Generally, customers desire to have short service times for service parts. This is because they often desire to have their failed equipments fixed in the shortest possible time. We treat each SVC as queue system such that the objects passing through the systems are customer demands [28]. The quantity of objects in line gives the level of backorder, while, the entire time spent by object till fulfillment of order gives the response time. An application of Little's law [29] results to the following:

$$Wt_{yz} = \frac{B_{yz}}{\lambda_{yz}} \tag{8}$$

Hence, the service constraint can be rewritten as

$$B_{yz} \leq \tau\lambda_{yz} \tag{9}$$

The basic form of the model does not give readily give any information on the structure of the model. In order to determine the problem structure, we need to first determine the on-hand inventory level, LT level, and backorder level for the model given S_0 and S_{yz} respectively. We present this in the next subsection.

We present the expected levels in steady state for on-hand inventory, LT and backorder. We utilise the METRIC method [25] to approximate these levels. [30] presented an exact procedure for deriving steady state levels for on hand inventory and backorder at the second echelon. However this procedure is computationally burdensome and not ideal for large optimisation problems. To ease computation, researchers have proposed various approximations. The METRIC method uses Palm's theorem [31] to find an approximate distribution for orders in replenishment at each facility in the lower echelon. This is done by means of a Poisson distribution with corresponding mean. METRIC approximation ignores the dependence of successive lead times from the top echelon to facilities in the lower echelon. The lead times are actually dependent on the inventory situation at the top echelon. Another important approximation was proposed by [30]. He approximated the number of outstanding orders in the lower echelon facilities by a negative binomial distribution comprising of two parameters, which are the corresponding mean and variance. METRIC approximation, in general, is appropriate as long as each lower echelon facility demand is low compared to the total system demand. The METRIC approximation will work well for a system with many facilities in the lower echelon, this reduces the dependence between successive lead times to the top echelon [2]. Thus the METRIC approximation is appropriate for the system considered in this study.

In this system each customer's demand follow a Poisson process, hence the demand process at each service center is also Poisson because it is an aggregation of independent Poisson processes. We assume that the SVCs and plant are controlled with an order-up-to policy. This implies that arrival process for the plant's demand is also Poisson. [2] derived expected inventory and backorder levels in steady state for the plant, we state their result and give a new proof.

Proposition 2.2.1

1. In steady state the plant's expected backorder level is given by

$$B_0 = I_0 - S_0 + E[N_0], \tag{10}$$

2. In steady state the plant's expected inventory level is given by

$$I_0 = S_0 - E[N_0] + B_0. \tag{11}$$

proof

Individual customer arrival process at each SVC is Poisson. The aggregation of Poisson processes at each SVC imply that the demand process at each SVC is also Poisson. Controlling each SVC with the (S-1,S) policy implies that replenishment requests at the plant arrive one at a time. Thus, the plant's demand arrival process is Poisson and possesses the properties of a Markovian queue. The balance equation for a Markovian queue imply that steady state inflow is equal to steady state outflow. The steady state inflow to the plant is denoted by $E[N_0]$. $S_0 \geq I_0$ by definition and the steady state expected number of plant demand fulfilled from on-hand inventory is $S_0 - I_0$. The steady state expected number of plant demand satisfied from backorder is represented by B_0. Therefore steady state expected outflow is $S_0 - I_0 + B_0$. Hence the balance equation of this system is

$$E[N_0] = S_0 - I_0 + B_0 \tag{12}$$

Thus

$$I_0 = S_0 - E[N_0] + B_0. \tag{13}$$

and

$$B_0 = I_0 - S_0 + E[N_0]\square \tag{14}$$

Furthermore, [2] showed that the plant backorder and inventory levels in steady state can also be given respectively as

$$B_0 = E[N_0] - \sum_{s=0}^{S_0-1} [1 - F_0(s)], \quad (15)$$

and

$$\bar{I}_0 = \sum_{s=0}^{S_0-1} F_0(s) \quad (16)$$

where

$$F_0(s) = \sum_{m=0}^{s} P\{N_0 = m\}$$

.

For variety of manufacturing queuing systems, plugging in the long-run probabilities into the formulas above readily give the expected levels for the plant's backorder and on-hand inventory [7].

For an M/M/1 queue, [32] showed the expected levels for plant backorder and on-hand inventory in steady state to be

$$B_0 = \frac{\rho^{S_0+1}}{1-\rho} \quad (17)$$

$$I_0 = S_0 - \frac{\rho}{1-\rho}(1-\rho^{S_0}) \quad (18)$$

By Little's law, the expected plant response time is given by

$$W_0 = \frac{\rho^{S_0+1}}{\lambda_0(1-\rho)} \quad (19)$$

In this subsection, we treat each pool as a single facility. This makes our problem have a structure similar to the problems considered by [2] and [7], with the plant at the top echelon and the pools at the lower echelon. The demand at each pool is Poisson, by the aggregation of the demand processes of all SVCs in the pool. Pool demand is fulfilled either through on-hand inventory or through backorders.

We assume identical stock levels S_{yz} for all SVCs in pool z. Hence the pool stock level is given by $S_z = |z|S_{yz}$, where $|z|$ represents the number of SVCs in pool z. The pool expected inventory level in steady state is

$$I_z = \sum_{s=0}^{|z|S_{yz}-1} (|z|S_{yz} - s)P\{N_z = s\}$$

or

$$I_z = \sum_{s=0}^{|z|S_{yz}-1} F_z(s) \quad (20)$$

where

$$F_w(z) = \sum_{m=0}^{s} P\{N_z = m\}$$

The following proposition establishes the steady state expected pool backorder level.

Proposition 2.2.2

The expected pool backorder level in steady state is

$$B_z = \frac{\lambda_z}{\lambda_0}\frac{\rho^{S_0+1}}{1-\rho} + \lambda_z\alpha_z - |z|S_{yz} + \sum_{s=0}^{|z|S_{yz}-1} F_z(s) \quad (21)$$

proof

Backorders can only take place in $Pool_z$ if every SVC in that pool experience a stockout situation at the same time, this follows from the assumption of instantaneous transshipment times. Then the expected pool backorder level in steady state is

$$B_z = E[N_z] - \sum_{s=0}^{|z|S_{yz}-1} [1 - F_z(s)] \quad (22)$$

$$= E[N_z] - |z|S_{yz} + \sum_{s=0}^{|z|S_{yz}-1} F_z(s)$$

Thus, the distribution of pool z outstanding orders ,N_z, in steady state, is needed before inventory and backorder expected levels can be determined. At any given time t, total number of pool z's outstanding orders comprises of:

(a) backorders at the plant emanating from pool z at $t - \alpha_z$ (this amount was in backorder status at time $t - \alpha_z$, implying that they were not instantly shipped out to pool z. Thus, they will not get to pool z prior to t) and

(b) quantity of fresh arrivals in $(t - \alpha_z, t)$.

The queue discipline for order processing at the plant is First Come First Served, thus, we can randomly split plant backorders [30]. The implication of this is that the probability a backorder at the plant emanated from pool z is proportional to the demand rate at pool z. Pool z's expected backorder value is $(\frac{\lambda_z}{\lambda_0})B_0$. Given a time interval length of α_z, the long term average of fresh arrivals in α_z is $\lambda_z\alpha_z$. Therefore, the expected value of N_z in steady state is given by:

$$E[N_z] = \frac{\lambda_z}{\lambda_0}B_0 + \lambda_z\alpha_z = \frac{\lambda_z}{\lambda_0}\frac{\rho^{S_0+1}}{1-\rho} + \lambda_z\alpha_z \quad (23)$$

Therefore

$$B_z = \frac{\lambda_z}{\lambda_0}\frac{\rho^{S_0+1}}{1-\rho} + \lambda_z\alpha_z - |z|S_{yz} + \sum_{s=0}^{|z|S_{yz}-1} F_z(s) \tag{24}$$

⊡

Demand faced at each SVC is fulfilled instantly through on hand inventory if the SVC has a positive inventory level. If the inventory level is zero, the demand arrival at the SVC is also satisfied instantly via LT from any other available SVC in same pool with positive inventory level. In the event that all SVCs in the pool have non-positive inventory levels, then the demand is backordered. So SVC demand is satisfied from any one of inventory on hand, LT, and backorder.

The next result determines the steady state relationship between inventory level, LT level and backorder level at a SVC. This result builds on work done by [2] and [32].

Proposition 2.2.3

1. The steady state expected inventory level at SVC y in pool z is

$$I_{yz} = S_{yz} - E[N_{yz}] + T_{yz} + B_{yz} \tag{25}$$

2. The steady state expected backorder level at SVC y in pool z is

$$B_{yz} = I_{yz} - S_{yz} + E[N_{yz}] + T_{yz} \tag{26}$$

3. The steady state expected LT level at SVC y in pool z is is

$$T_{yz} = I_{yz} - S_{yz} + [N_{yz}] - T_{yz} \tag{27}$$

proof

$$\begin{aligned} I_{yz} &= \sum_{s=0}^{S_{yz}-1}(S_{yz}-s)P\{N_{yz}=s\} \\ &= S_{yz}\sum_{s=0}^{S_{yz}-1}P\{N_{yz}=s\} - \sum_{s=0}^{S_{yz}-1}sP\{N_{yz}=s\} \\ &= S_{yz}(1-\sum_{s=S_{yz}}^{\infty}P\{N_{yz}=s\}) - (\sum_{s=0}^{\infty}sP\{N_{yz}=s\}) \\ &- \sum_{s=S_{yz}}^{\infty}sP\{N_{yz}=s\}) \\ &= S_{yz} - S_{yz}\sum_{s=S_{yz}}^{\infty}P\{N_{yz}=s\} \\ &- (E[N_{yz}] - \sum_{s=S_{yz}}^{\infty}sP\{N_{yz}=s\}) \\ &= S_{yz} - E[N_{yz}] + \sum_{s=S_{yz}}^{|z|S_{yz}}(s-S_{yz})P\{N_{yz}=s\} \\ &+ \sum_{s=|z|S_{yz}+1}^{\infty}(s-|z|S_{yz})P\{N_{yz}=s\} \\ &= S_{yz} - E[N_{yz}] + T_{yz} + B_{yz} \end{aligned}$$

This proves (25). Making B_{yz} and T_{yz} the subject of the formula yield (26) and (27) respectively.
The next proposition gives the SVC steady state expressions for on hand inventory, backorder and LT.

Proposition 2.2.4

1. The steady state expected inventory level at each SVC is

$$I_{yz} = \sum_{s=0}^{S_{yz}-1}(S_{yz}-s)P\{N_{yz}=s\} \tag{28}$$

2. The steady state expected backorder level at each SVC is

$$\begin{aligned} B_{yz} =& \frac{\lambda_{yz}}{\lambda_0}\frac{\rho^{S_0+1}}{1-\rho} + \lambda_{yz}\alpha_w \\ &+ \frac{\lambda_{yz}}{\lambda_w}\left(\sum_{s=0}^{|z|S_{yz}-1} F_z(s) - |z|S_{yz}\right) \end{aligned} \tag{29}$$

3. The steady state expected LT level at each SVC is

$$T_{yz} = \sum_{s=0}^{S_{yz}-1} F_{yz}(s) - S_{yz} - \frac{\lambda_{yz}}{\lambda_w}\left(\sum_{s=0}^{|z|S_{yz}-1} F_z(s) - |z|S_{yz}\right) \quad (30)$$

where

$$F_{yz}(s) = \sum_{m=0}^{s} P\{N_{yz} = m\}$$

and

$$F_z(s) = \sum_{m=0}^{s} P\{N_z = m\}$$

proof The proof for SVC inventory steady state level is similar to that of the steady state expected pool inventory level

$$I_{yz} = \sum_{s=0}^{S_{yz}-1} (S_{yz} - s)P\{N_{yz} = s\}$$

$$I_{yz} = \sum_{s=0}^{S_{yz}-1} F_{yz}(s) \quad (31)$$

By the model assumptions, backorders can only occur if all SVCs in a pool are out of stock. Let $S_z = |z|S_{yz}$ be the total base stock level of the pool z. Then the steady state expected backorder level at each SVC in $Pool_z$ is

$$B_{yz=}\left(\frac{\lambda_{yz}}{\lambda_z}\right) B_w \quad (32)$$

$$B_{yz} = \left(\frac{\lambda_{yz}}{\lambda_z}\right)\left(\frac{\lambda_z}{\lambda_0}\frac{\rho^{S_0+1}}{1-\rho} + \lambda_z\alpha_z + \left(\sum_{s=0}^{|z|S_{yz}-1} F_z(s) - |z|S_{yz}\right)\right)$$

hence

$$B_{yz} = \frac{\lambda_{yz}}{\lambda_0}\frac{\rho^{S_0+1}}{1-\rho} + \lambda_{yz}\alpha_z + \frac{\lambda_{yz}}{\lambda_z}\left(\sum_{s=0}^{|z|S_{yz}-1} F_z(s) - |z|S_{yz}\right) \quad (33)$$

The behaviour of a Markovian queue in steady state along with the implications of using the (S-1,S) policy help to determine LT level at SVCs. The balance equation for the SVC is

$$T_{yz} + B_{yz} + S_{yz} - I_{yz} = E[N_{yz}] \quad (34)$$

The expected number of requests filled from on-hand inventory is given by $S_{yz} - I_{yz}$

Hence, the distribution in steady state of orders outstanding ,N_{yz} , defined as the quantity of orders in line and in service at the SVC, is needed in order to ascertain the inventory, backorder and transshipment levels.

The quantity of unfulfilled orders from a SVC at a given time t, comprises of:
(a) plant backorders at time $t - \alpha_z$ which emanated from the SVC (these orders won't get to the SVC before t)
(b) the quantity of fresh order arrivals in $(t - \alpha_z, t)$.

Since First Come First Served is the plant's queue discipline, we can randomly disaggregate plant backorders [30]. Consequently, the probability a backorder at the plant emananted from a particular SVC is proportional to the demand rate at that SVC. The expected backorder value in steady state for a SVC is given by

$$\frac{\lambda_{yz}}{\lambda_0}B_0.$$

The expected fresh order arrivals during a time interval of length α_z is $\lambda_z\alpha_z$. Consequently, the expected value of N_{yz} in steady state is given by:

$$E[N_{yz}] = \frac{\lambda_{yz}}{\lambda_0}\frac{\rho^{S_0+1}}{1-\rho} + \lambda_{yz}\alpha_w \quad (35)$$

Hence

$$T_{yz} = E[N_{yz}] + I_{yz} - S_{yz} - B_{yz} \quad (36)$$

$$= \frac{\lambda_{yz}}{\lambda_0}\frac{\rho^{S_0+1}}{1-\rho} + \lambda_{yz}\alpha_z + \sum_{s=0}^{S_{yz}-1} F_{yz}(s) - S_{yz} \quad (37)$$

$$- \frac{\lambda_{yz}}{\lambda_0}\frac{\rho^{S_0+1}}{1-\rho} - \lambda_{yz}\alpha_w - (\frac{\lambda_{yz}}{\lambda_z})\sum_{s=0}^{|z|S_{yz}-1}[1 - F_z(s)]$$

Therefore

$$T_{yz} = \frac{\lambda_{yz}}{\lambda_z} \sum_{s=0}^{|z|S_{yz}-1} [1 - F_z(s)] - \sum_{s=0}^{S_{yz}-1} [1 - F_{yz}(s)] \tag{38}$$

$$= \sum_{s=0}^{S_{yz}-1} F_{yz}(s) - S_{yz} - \left(\frac{\lambda_{yz}}{\lambda_z} (\sum_{s=0}^{|z|S_{yz}-1} F_z(s) - |z|S_{yz}) \right)$$

Note

$$\sum_{s=0}^{|z|S_{yz}-1} F_z(s) - |z|S_{yz} = - \sum_{s=0}^{|z|S_{yz}-1} [1 - F_z(s)] \tag{39}$$

⊡

Following METRIC method, we derive approximations for outstanding orders at the pool and SVCs to be

$$P[N_z = m] = \frac{e^{\lambda_z L_z} (\lambda_z L_z)^m}{m!} \tag{40}$$

and

$$F_z(s) = \sum_{m=0}^{s} \frac{e^{\lambda_z L_z} (\lambda_z L_z)^m}{m!} \tag{41}$$

In the above, L_z is the expected replenishment lead time which consists of expected response time of the plant and the delivery lead time:

$$L_z = W_0 + \alpha_z = \frac{\rho^{S_0+1}}{\lambda_0(1-\rho)} + \alpha_z \tag{42}$$

Similarly for SVC

$$P[N_{yz} = m] = \frac{e^{-\lambda_{yz} L_z} (\lambda_{yz} L_z)^m}{m!} \tag{43}$$

and

$$F_{yz}(s) = \sum_{m=0}^{s} \frac{e^{-\lambda_{yz} L_z} (\lambda_{yz} L_z)^m}{m!} \tag{44}$$

In the above, L_z is the expected replenishment lead time which consists of expected response time of the plant and the delivery lead time. Note that L_z here is the same as that for the pool; this is because it is assumed that lateral transshipment between SVCs in a pool is instantaneous.

Substituting the expressions for I_{yz}, B_{yz} and T_{yz} into our model gives the following reformulation, which shows the true structure of the problem.

$$\begin{aligned} \min \sum_{z \in Z} \sum_{y \in Y} \Bigg\{ & (h_{yz} + q_{yz}) \sum_{s=0}^{S_{yz}-1} F_{yz}(s) - q_{yz} S_{yz} \\ & + \lambda_{yz} \left(\frac{p_{yz} \rho^{S_0+1}}{\lambda_0(1-\rho)} + p_{yz} \alpha_z \right) \\ & + (p_{yz} - q_{yz}) \frac{\lambda_{yz}}{\lambda_z} \left(\sum_{s=0}^{|z|S_{yz}-1} F_z(s) - |z| S_{yz} \right) \Bigg\} \\ & + h_0 [S_0 - \frac{\rho}{1-\rho}(1 - \rho^{S_0})] \end{aligned} \tag{45}$$

Subject to

$$S_{yz} \leq C_{yz}, \text{ for each, } y \in Y \tag{46}$$
$$S_z \leq C_z = |z| C_{yz}, \text{ for each, } z \in Z \tag{47}$$
$$S_0 \leq C_0 \tag{48}$$

$$\left[\frac{\rho^{S_0+1}}{\lambda_0(1-\rho)} + \alpha_z - \tau \right] \lambda_{yz} \leq \frac{\lambda_{yz}}{\lambda_z} \sum_{s=0}^{|z|S_{yz}-1} [1 - F_z(s)] \tag{49}$$

$$S_{yz}, S_z, S_0 \geq 0 \text{ integer , for each } y \in Y \tag{50}$$

(49) can also be written as

$$\left[\frac{\rho^{S_0+1}}{\lambda_0(1-\rho)} + \alpha_z - \tau \right] \leq \frac{\sum_{s=0}^{S_z-1} [1 - F_z(s)]}{\lambda_z}$$

A close inspection shows that the model is a mixed integer programming problem.

3 Model Properties

In this section, we exploit the structure of the model to highlight some of it's properties. The model reformulation obtained in the previous section is solvable using GAMs. Thus, there is no urgency on our part to find heuristic solutions for our model. However, the properties highlighted in this section are steps that can lead to future development of heuristic solutions.

3.1 Lagrange relaxed solution

The existence of capacity constraint, in addition to our assumption of low system demand, indicate that the stock levels required to ensure the satisfaction of a desired service level lies within a small range, which has the capacity as its upper bound. [33] and [7] exploited similar properties to develop solution algorithms which enumerated over all feasible points. For this model, the capacity constraint on the plant, implies that a number of problems can be solved where the base stock level at the plant is fixed to each feasible value. The solution with the least cost is the original problem's optimal solution. When S_0 is fixed, terms dependent on just S_0 are treated as constants. This property, makes it easy to break down the model into smaller problems.

The reformulation of the model does not give any obvious clue on the properties or structure of the model. Thus, there is need to utilise a decomposition technique to decompose the model. Lagrange relation has been shown to be efficient for problems with complicating constraints. Hence, we use Lagrange relaxation to decompose our model. With S_0 and S_{yz} are fixed, we observe that the complicating constraint is the response time constraint. Thus, we relax the service constraints (49) in the restricted problem so as to decompose the model and further exploit it's structure. Using γ_{yz} to denote the corresponding dual multiplier for (49), the following Lagrangian Dual problem is obtained:

$$\begin{aligned}
&\max_{\gamma\geq 0} \min \sum_{z\in Z}\sum_{y\in Y}\Bigg\{(h_{yz}+q_{yz})\sum_{s=0}^{S_{yz}-1} F_{yz}(s) - q_{yz}S_{yz} \\
&+(p_{yz}-q_{yz}+\gamma_{yz})\frac{\lambda_{yz}}{\lambda_z}\sum_{s=0}^{|z|S_{yz}-1} F_z(s) \qquad (51)\\
&-\frac{\lambda_{yz}}{\lambda_z}(p_{yz}-q_{yz}+\gamma_{yz})|z|S_{yz} \\
&+\frac{(p_{yz}+\gamma_{yz})\rho^{S_0+1}}{\lambda_0(1-\rho)}\lambda_{yz} + (p_{yz}\alpha_z+\gamma_{yz}\alpha_z-\gamma_{yz}\tau)\lambda_{yz}\Bigg\}
\end{aligned}$$

Subject to

$$S_{yz} \leq C_{yz}, \text{ for each,} y \in Y, z \in Z \qquad (52)$$

$$S_0 \leq C_0 \qquad (53)$$

$$S_{yz}, S_z, S_0 \geq 0 \text{ integer , for each,} y \in Y, z \in Z(54)$$

The objective function (55) can be rewritten as

$$\begin{aligned}
&\max_{\gamma\geq 0} \min_{S} \sum_{z\in Z}\sum_{y\in Y}\Bigg\{\sum_{s=0}^{S_{yz}-1} [(h_{yz}+q_{yz})F_{yz}(s) - q_{yz}] \\
&\qquad\qquad (55)\\
&+(p_{yz}-q_{yz}+\gamma_{yz})\frac{\lambda_{yz}}{\lambda_z}\sum_{s=0}^{|z|S_{yz}-1} (F_z(s)-1) \\
&+\left(\frac{(p_{yz}+\gamma_{yz})\rho^{S_0+1}}{\lambda_0(1-\rho)} + (p_{yz}\alpha_z+\gamma_{yz}\alpha_z-\gamma_{yz}\tau)\right)\lambda_{yz}\Bigg\}
\end{aligned}$$

Lagrange relaxation decomposes the problem by SVCs and associated pools. The decomposed problem is

$$\begin{aligned}
&\max_{\gamma\geq 0} \min_{S} \sum_{s=0}^{S_{yz}-1} [(h_{yz}+q_{yz})F_{yz}(s) - q_{yz}] \qquad (56)\\
&+(p_{yz}-q_{yz}+\gamma_{yz})\frac{\lambda_{yz}}{\lambda_z}\sum_{s=0}^{|z|S_{yz}-1} (F_z(s)-1) \\
&+\left(\frac{(p_{yz}+\gamma_{yz})\rho^{S_0+1}}{\lambda_0(1-\rho)} + (p_{yz}\alpha_z+\gamma_{yz}\alpha_z-\gamma_{yz}\tau)\right)\lambda_{yz}
\end{aligned}$$

Subject to

$$S_{yz} \leq C_{yz} \text{ for each,} y \in Y, z \in Z \qquad (57)$$

$$S_{yz} \text{ integer , for each,} y \in Y, z \in Z \qquad (58)$$

The constraints of the decomposed model depends on only S_{yz}. We proceed to show that the objective function is convex with respect to S_{yz} and S_z. Convexity is desirable property for optimisation models because it guarantees the existence of a global minimum solution. Recall that $|z|S_{yz} = S_z$. Let $K(S_{yz}, S_z)$ represent the terms depending entirely on S_{yz} and S_z, and let $\triangle_{S_{yz}}(K(S_{yz}, S_z))$ denote the first difference of $K(S_{yz}, S_z)$ with respect to S_{yz}.

$$\begin{aligned}
K(S_{yz}, S_z) =&(h_{yz}+q_{yz})\sum_{s=0}^{S_{yz}-1} F_{yz}(s) - q_{yz}S_{yz} \\
&+(p_{yz}-q_{yz}+\gamma_{yz})\frac{\lambda_{yz}}{\lambda_z}(\sum_{s=0}^{S_z-1}(F_z(s)-1))
\end{aligned} \qquad (59)$$

$$\begin{aligned}
\triangle_{S_{yz}}(K(S_{yz}, S_z)) &= K(S_{yz}+1, S_z) - K(S_{yz}, S_z) \\
&= (h_{yz}+q_{yz})F_{yz}(S_{yz}) - q_{yz}
\end{aligned} \qquad (60)$$

$$\begin{aligned}
&\triangle_{S_{yz}}(\triangle_{S_{yz}}(K(S_{yz}, S_z))) = \triangle_{S_{yz}}(K(S_{yz}+1, S_z)) \\
&-\triangle_{S_{yz}}(K(S_{yz}, S_z)) \\
&= (h_{yz}+q_{yz})(F_{yz}(S_{yz}+1) - F_{yz}(S_{yz})) \\
&= (h_{yz}+q_{yz})(\sum_{m=0}^{S_{yz}+1} P\{N_{yz}=s\} - \sum_{m=0}^{S_{yz}} P\{N_{yz}=s\}) \\
&= (h_{yz}+q_{yz})P\{N_{yz} = S_{yz}+1\} \qquad (61)
\end{aligned}$$

$$\begin{aligned}
\triangle_{S_z}K(S_{yz}, S_z) &= K(S_{yz}, S_z+1) - K(S_z) \\
&= (p_{yz}-q_{yz}+\gamma_{yz})\frac{\lambda_{yz}}{yz}(F_z(S_z)-1) \\
&= -(p_{yz}-q_{yz}+\gamma_{yz})\frac{\lambda_{yz}}{\lambda_z}[1-F_z(S_z)] < 0
\end{aligned} \qquad (62)$$

$$\triangle_{S_{yz}}(\triangle_{S_z}K(S_{yz},S_z)) = 0 \tag{63}$$

Similarly,

$$\triangle_{S_z}(\triangle_{S_{yz}}K(S_{yz},S_z)) = 0 \tag{64}$$

$$\begin{aligned}
&\triangle_{S_z}(\triangle_{S_z}K(S_{yz},S_z)) = \triangle_{S_z}K(S_{yz},S_z+1)\\
&-\triangle_{S_z}K(S_{yz},S_z)\\
&= (p_{yz}-q_{yz}+\gamma_{yz})\frac{\lambda_{yz}}{\lambda_z}(F_z(S_z+1)-F_z(S_z))\\
&+(p_{yz}-q_{yz}+\gamma_{yz})\frac{\lambda_{yz}}{\lambda_z}(-1+1)\\
&= (p_{yz}-q_{yz}+\gamma_{yz})\frac{\lambda_{yz}}{\lambda_z}(\sum_{m}^{S_z+1}P[N_z=m]\\
&-\sum_{m}^{S_z}P[N_z=m])\\
&= (p_{yz}-q_{yz}+\gamma_{yz})\frac{\lambda_{yz}}{\lambda_z}(P[N_z=S_z+1])>0
\end{aligned} \tag{65}$$

Let $H_1 = \triangle_{S_{yz}}(\triangle_{S_{yz}}K(S_{yz},S_z))$, $H_2 = \triangle_{S_{yz}}(\triangle_{S_z}K(S_{yz},S_z))$, $H_3 = \triangle_{S_z}(\triangle_{S_{yz}}K(S_{yz},S_z))$ and $H_4 = \triangle_{S_z}(\triangle_{S_z}K(S_{yz},S_z))$ The Hessian Matrix of the problem $Hess(S_{yz},S_z)$ is

$$Hess(S_{yz},S_z) = \begin{pmatrix} H_1 & H_2 \\ H_3 & H_4 \end{pmatrix} \tag{66}$$

$\triangle_{S_{yz}}(\triangle_{S_{yz}}K(S_{yz},S_z)) > 0$, $\triangle_{S_z}(\triangle_{S_z}K(S_{yz},S_z))$ therefore $\det(Hess(S_{yz},S_z)) > 0$. This implies that the Hessian of the problem is strictly positive definite with respect to S_{yz} and S_z, thus the problem is convex with respect to S_{yz} and S_z.

Remark: We utilised Lagrange relaxation to decompose the model and have showed that the decomposed problem is convex. The solution of the relaxed problem is only a lower bound to the solution of the initial problem. We proceed to show that the primal problem is convex for fixed values of S_0.

Proposition 3.1.1

For fixed values of S_0, the primal problem is a convex optimisation problem.

proof

We showed convexity of the relaxed problem for fixed multiplier values. Thus, the objective function of the relaxed problem is convex for $\gamma_{vw} = 0$. Also, the objective function of the relaxed problem is equal to the objective function of the model with LT when $\gamma_{yz} = 0$. Hence the objective function of the model with LT (primal problem) is convex. With S_0 fixed, the response time constraint depends only on the variable $S_z = |z|S_{yz}$ and can be written as

$$L_z - \tau + \frac{1}{\lambda_z}\sum_{s=0}^{S_z-1}(F_z(s)-1) \le 0$$

Let

$$\bar{J}(S_z) = L_z - \tau + \frac{1}{\lambda_z}\sum_{s=0}^{S_z-1}(F_z(s)-1)$$

$$\begin{aligned}
\triangle\bar{J}(S_z) &= \bar{J}(S_z+1) - \bar{J}(S_z)\\
&= \frac{1}{\lambda_z}(F_z(S_z)-1)\\
&= -\frac{1}{\lambda_z}(1-F_z(S_z-1)) < 0
\end{aligned}$$

where, $\triangle\bar{J}(S_z)$ is the first difference of $\bar{J}(S_z)$.

$$\begin{aligned}
\triangle^2\bar{J}(S_z) &= \triangle\bar{J}(S_z+1) - \triangle\bar{J}(S_z)\\
&= \frac{1}{\lambda_z}(F_z(S_z+1)-F_z(S_z))\\
&= \frac{1}{\lambda_z}\left(\sum_{m}^{S_z+1}P[N_z=m] - \sum_{m}^{S_z}P[N_z=m]\right)\\
&= \frac{1}{\lambda_z}(P[N_z=S_z+1]) > 0
\end{aligned}$$

where, $\triangle^2\bar{J}(S_z)$ is the second difference of $\bar{J}(S_z)$. Since, $\triangle^2\bar{J}(S_z) > 0$, we say that our service constraint is convex. We haved shown that the objective function of our model and the inequality constraint (service constraint) are convex when S_0 is fixed. Thus, for fixed values of S_0, the primal problem is convex with respect to S_z.$\boxdot$

Convexity implies that our model can be solved using convex optimisation solvers

3.2 Optimal base stock level for service centers

Having exploited the properties of Model I by means of Lagrange relaxation, we proceed to determine the nature of the optimal solution. Here, the optimal solution is the basestock level that gives minimum cost and also satisfies the service or response time constraint.

Let the terms of the objective function of the primal problem (45) that depends on S_{yz} be given as:

$$H(S_{yz}) = (h_{yz} + q_{yz}) \sum_{s=0}^{S_{yz}-1} F_{yz}(s) - q_{yz}S_{yz}$$
$$+(p_{yz} - q_{yz} + \gamma_{yz})\frac{\lambda_{yz}}{\lambda_z}(\sum_{s=0}^{|z|S_{yz}-1} (F_z(s) - 1)) \quad (67)$$

To determine optimal S_{yz} without response time constraint, let $\triangle H(S_{yz})$ be the change in the objective value due to an increase in base stock level from S_{yz} to $S_{yz} + 1$. Then

$$\begin{aligned}&\triangle H(S_{yz}) = \bar{H}(S_{yz} + 1) - H(S_{yz})\\ &= (h_{yz} + q_{yz})F_{yz}(S_{yz}) - q_{yz}\\ &+ (p_{yz} - q_{yz} + \gamma_{yz})\frac{\lambda_{yz}}{\lambda_z}\left(\sum_{s=|z|S_{yz}}^{|z|S_{yz}+|z|-1} (F_z(s) - 1)\right)\end{aligned} \quad (68)$$

$$\triangle H(S_{yz}) = F_{yz}(S_{yz}) - \frac{(q_{yz} + p_{yz} - \gamma_{yz})\lambda_{yz} \sum_{s=|z|S_{yz}}^{|z|S_{yz}+|w|-1} [1 - F_z(s)] + \lambda_z q_{yz}}{\lambda_z(h_{yz} + q_{yz})} \quad (69)$$

When $\triangle(S_{yz}) < 0$, increasing S_{yz} by 1 will cause a decrease in cost. Also, by definition $F_{yz}(S_{yz})$ is monotone increasing in S_{yz} and lies between 0 and 1. Hence the unconstrained optimal S_{yz} can be found as follows:
Fix $S_0 = C_0$ and follow the following steps to determine a local minimum cost for $S_0 = C_0$.

1. If $\triangle H(C_{yz}) \leq 0$, then $S_{yz} = C_{yz}$ remains.
2. If $\triangle H(C_{yz}) > 0$, select S_{yz} as the largest integer such that $\triangle H(S_{yz}) \leq 0$.

Decrease the value of S_0 by one and follow the steps above. To get all local minimum solutions, the process is repeated until S_0 reaches zero. We pick the minimum of all local solutions, this becomes the global minimum solution.

The optimal S_{yz} with service constraint can be found as follows:
Fix $S_0 = C_0$ and follow the following steps to determine a local minimum cost for $S_0 = C_0$.

1. If $\triangle H(C_{yz}) \leq 0$ and $[L_w - \tau]\,\lambda_{yz} \leq \frac{\lambda_{yz}}{\lambda_z}\sum_{s=0}^{|z|C_{yz}-1}[1 - F_z(s)]$ then $S_{yz} = C_{yz}$ remains.
2. If $\triangle H(C_{yz}) > 0$ select S_{yz} as the largest integer such that $\triangle H(S_{yz}) \leq 0$ and $[L_z - \tau]\,\lambda_{yz} \leq \frac{\lambda_{yz}}{\lambda_z}\sum_{s=0}^{|z|S_{yz}-1}[1 - F_z(s)]$

Decrease the value of S_0 by one and follow the steps above. To get all local minimum solutions, the process is repeated until S_0 reaches zero. We pick the minimum of all local solutions, this becomes the global minimum solution.

The optimal solution to the model can be obtained using GAMS software, thus there is no urgency to immediately develop any specialised heuristics for solving it.

4 Computational Experiments

In this section, computational experiments which investigate properties of the model are designed. We utilise two data sets made up of of 37 nodes and 109 nodes. The 37 nodes represent the 36 state capitals and capital city in Nigeria. The 109 nodes represent the 3 most populous cities in each of the 36 states and the capital city. The population of each city was obtained from the 2006 census. Nodes in same geopolitical zone form a pool and LT is permitted only among SVCs in the same pool. Demand rates are obtained by multiplying the population at a node by 10^{-5}. The demand rate for each node is constrained to be no more than 10 for nodes with very large population. The pure lead time for transportation from the plant to pool z is set to be the maximum of the transportation lead times from the plant to SVCs in pool z.

4.1 Model performance

Here we compare our model with the model without LT. This test is conducted with the 37 node and 109 node data sets for UR= (0.9, 0.5) and RTR =(0.4, 0.5, 0.6), where, UR and RTR are abbreviations for utility rate (ρ) and response time requirement (ρ), respectively. We take note of the objective function value of our model (OBJ LT) and the objective function value of the model without LT (OBJ WLT). The model without LT was obtained from the model by [2] by imposing it with our pooling criterion. The pooling criterion implies that the model with LT can be partitioned into sub problems by geographical region. Hence, a fair comparison will be to compare with a model which is also a collection of sub problems by geographical region. This idea follows from the paper by [13] who partitioned the facilities into disjoint pools, and considered one of such pools. Thus, the total expected cost of the system comprises of the sum of the costs from all pools and the cost at the plant.

He divides the stock locations into pooling groups, and focuses on one such group.

Table 1: Model performance

S/N	NODES	UR	RTR	OBJ LT	OBJ WLT
1	37	0.9	0.6	50848.58	533559.76
2	37	0.5	0.6	51343.16	55568.75
3	37	0.9	0.5	50848.58	53559.76
4	37	0.5	0.5	51343.16	55568.75
5	37	0.9	0.4	50848.58	53559.76
6	37	0.5	0.4	51343.16	55568.75
7	109	0.9	0.6	50483.05	50531.07
8	109	0.5	0.6	49730.65	55462.37
9	109	0.9	0.5	50483.05	50531.07
10	109	0.5	0.5	49730.65	55462.37
11	109	0.9	0.4	504483.05	50531.07
12	109	0.5	0.4	49730.65	55462.37

We summarise our results in Table 1. For all instances tested, the total system cost of our LT model was lower than that of the model without LT. This illustrates the cost savings that can be achieved via the incorporation of LT.

4.2 Effect of response time requirement and base stock level

In this experiment we check the behaviour of the model as the response time is varied. We make use of the 37-node dataset and set ρ to 0.9. We vary response time requirement values between 0.272 and 0.668. Figure 1 shows that expected cost remains stable with varying response time requirement values. This occurs as a result of LT and pooling which ensure uniform response time constraint for all SVCs in a pool. The implication of this is that, within feasible values, the decision maker can slacken or tighten the response time requirement to fit into the contract signed with a customer, and this will have negligible effect on the expected cost. This will help decision makers in two-echelon systems to know the bounds to response times during negotiation with customers.

In order to test the effect of basestock level, we utilise the 37 node dataset and set (ρ) to 0.9. In the first case, we consider the effect of plant base stock level on response time, S_{yz} is fixed at 5, while S_0 is varied between the feasible range. In the second case, we consider the effect of SVC base stock level on response time, S_0 is fixed at 3 and the value of S_{yz} is varied within the feasible range. In all cases, the minimum feasible response time requirement and the corresponding expected costs are recorded. The expected costs and response time requirement are plotted against the stock level.

In the first case, the difference between the largest expected cost (attained at the maximum plant capacity) and the lowest expected cost (attained when $S_0 = 1$) is 1% of the lowest expected cost. Also, for all instances tested, the difference between the expected cost for $S_0 = n$ and for $S_0 = n - 1$ was no more than 0.11% of the value of the expected cost when $S_0 = n - 1$. Thus, for this experiment, increase in plant base stock level results to negligible increase in expected cost and causes a decrease in the minimum response time requirement. This is obvious from Figure 2. Figure 3, shows that increasing SVC stock level results to an increase in expected cost and causes no change in minimum response time requirement. If the decision maker intends to reduce response times to customers with minimum increase in cost, she has to increase the plant base stock level. In real life, the value of S_0 is usually constrained by capacity. Comparing Figure 2 and Figure 3, it is obvious that increase in plant base stock level is preferable for achieving stable costs and minimum response time. This follows because the decent in response time is steeper when plant base stock is increased compared to when SVC base stock is increased. Also, the ascent in expected cost is steeper when SVC base stock is increased compared to when plant base stock is increased.

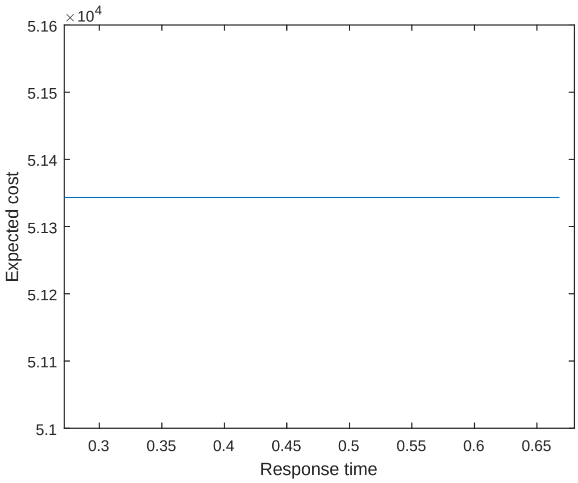

Figure 1: Effect of response time

5 Conclusion

In this study, we considered the integration of lateral transshipment into an inventory system with two echelons and a service time constraint across all faciities. The service centers and plants use a continuous review (S-1,S) policy for inventory management. We formulated a two echelon inventory model for our system and established the relationship between inventory on-hand, lateral transshipment and backorder, in steady state. We also determined steady state expected levels for inventory on-hand, lateral transshipment and backorder. The model was decomposed

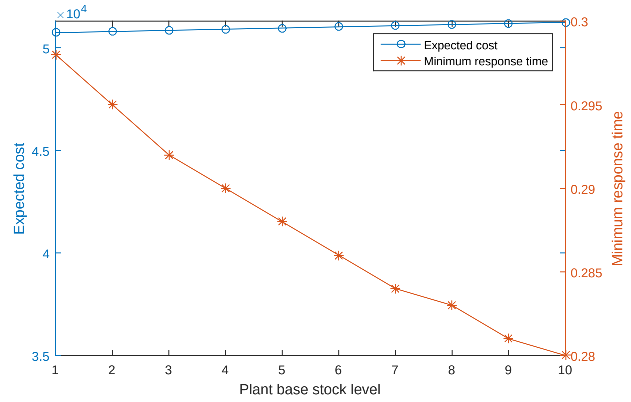

Figure 2: Effect of plant base stock level

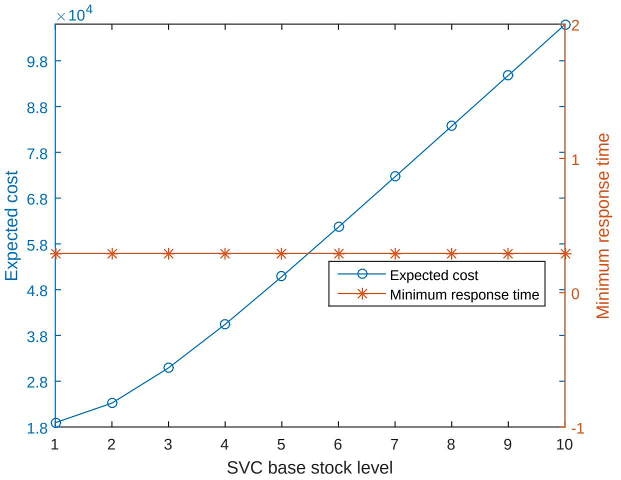

Figure 3: Effect of SVC base stock level

using Lagrange relaxation method. The model was shown to be convex. The solution to the model was found using GAMS. In all instances our model returned lower costs when compared with the model without lateral transshipment. From our results, increasing the plant base stock level will cause a reduction in minimum response time to with slight increase in cost. Our results show that a major effect of using lateral transshippment is that the expected cost remains consistent when the response time is varied between its feasible points. Conclusively, results from this study show that lateral transshipment is efficient for achieving a balance between the desire to simultaneously minimize cost and maintain acceptable response times in an inventory system with two echelons and service constraints.

There are several possible extensions to this work. Our pooling criterion is geographical, this just gave us the opportunity to analyse the problem structure. However, this might not be the optimal pooling criterion for optimising cost and service; it is possible that the closest facility which can fulfill a transshipment request might be a facility in a different geographical region (this scenario could arise for facilities in geographical regions which share same boundary). Thus, a possible future research direction will be to explore the problem using other pooling criteria, e.g. increasing order of distance as in [23]. Also, our model assumes that a transshipment source with positive on-hand stock always releases inventory to satisfy transshipment requests. [34] has considered cases where not all inventory available in a facility are available for transshipment. So, this work can be extended by looking at a pooling rule that holds back inventory for transshipment in order to hedge against future stock out. Another limitation of this study is the assumption of negligible lateral transshipment times. Thus, investigating the effects of non-negligible lateral transshipment on the system will be an interesting direction for future research.

References:

[1] Mpwanya, M. F., The Relationship Between Inventory-Management Polices and Customer Service in Manufacturing Industries Logistics in Gauteng Province, South Africa, *WSEAS Transactions on Business and Economics* ,Vol. 13, 556-572, 2016.

[2] Caglar, D., Li, C.L. and Simchi-Levi, D.,Two Echelon Spare Parts Inventory System Subject to a Service Constraint, *IIE Transactions*, Vol.36, 2004, pp.665-666.

[3] Shen, H., Tian, T. and Zhu, H., A Two-Echelon Inventory System with a Minimum Order Quantity Requirement, *Sustainability*, 2019.

[4] Lai, X., Chen, Z., Giri, B. C. and Chiu, C., Two-Echelon Inventory Optimization for Imperfect Production System under Quality Competition Environment, *Mathematical Problems in Engineering*, 2015.

[5] Marije, N., Wout, D., David, S.W.L. and de Leeuw, S., A Simulation–Optimization Approach for a Service-Constrained Multi-Echelon Distribution Network, *Transportation Research Part E: Logistics and Transportation Review*, Vol.114, 2018.

[6] Johansson, L., Multi-Echelon Inventory Control with Consideration of Emissions and Service Differentiation, *Lunds universitet, Lunds Tekniska Högskola, Lund University*, 2020.

[7] Mak, H. Y. and Shen, Z. J., A Two Echelon-Inventory Location Problem With Service Considerations.,*Naval Research Logistics*,2009.

[8] Basten, R, J. I., Van der Heijden, M. C. and Schutten, J. M. J., Joint optimization of level of repair analysis and spare parts stocks, *European J. Oper. Res.*, Vol.222, No.3, 2012, pp.474-443.

[9] Wheatley, D., Gzara, F. and Jewkes, E., Logic-based Benders decomposition for an inventory location problem with service constraints, *Omega*, Vol.55, 2015, pp.10-23.

[10] van den Berg, D., van der Heijden M. C., and Schuur, P. C., Allocating service parts in two echelon networks at a utility company, *International Journal of Production Economics*, Vol.181, No. Part A, 2016.

[11] Jiang, Y., Shi, C. and Shen, S., Service Level Constrained Inventory Systems, *Production and Operations Management*, Vol. 28, Issue 9, 2365-2389, 2019.

[12] Paterson, G., Kiesmullery, G., Teunterz, R. and Glazebrook, K., Inventory Models with Lateral Transshipments: A Review, *European Journal of Operational Research*, Vol.210, No.2, 2011, pp.125-136.

[13] Lee, H., A multi-echelon inventory model for repairable items with emergency lateral transshipments, *Management Science*,Vol.33, No.10, pp.1302-1316.

[14] Axsater, S., Modelling emergency lateral transshipments in inventory systems, *Management Science*, Vol.36, No.11, 1990, pp. 1329-1338.

[15] Alfredsson, P. and Verrijdt, J., Modeling emergency supply flexibility in a two-echelon inventory system, *Management Science*, Vol.45, No.10, 1999, pp.1416-1431.

[16] Banerjee, A. Burton, J. and Banerjee, S., A simulation study of lateral shipments in single supplier, multiple buyers supply chain networks, *International Journal of Production Economics*, Vol.81-82, 2003, pp.103-114.

[17] Burton, J. and Banerjee, A., Cost-parametric analysis of lateral transshipment policies in two-echelon supply chains, *International Journal of Production Economics*, Vol.93-94, 2005, pp.169-178.

[18] Grahovac, J. and Chakravarty, A., Sharing and lateral transshipment of inventory in a supply chain with expensive low demand items, *Management Science*, Vol.47, No.4, 2001, pp.579-594.

[19] Tagaras, G. and Vlachos, D., Effectiveness of stock transshipment under various demand distributions and non negligible transshipment times, *Production and Operations Management*, Vol.11, No.2, 2002, pp.183-198.

[20] Wong, H., van Houtum, G. J., Cattrysse, D. and van Oudheusden, D., Stocking decisions for repairable spare parts pooling in a multi-hub system, *International Journal of Production Economics*, Vol.93-94, 2005, pp.309-317.

[21] Kutanoglu, E.and Mahajan, M., An Inventory Sharing and Allocation Method for a Multi-Location Service Parts Logistics Network with Time-Based Service Levels, *European Journal of Operational Research*, Vol.194, 2009, pp.728-742.

[22] Kutanoglu, E., Insights Into Inventory Sharing in Service Parts Logistics Systems with Time-based Service Levels, *Computers & Industrial Engineering*, Vol.54, 2008, pp.341-358.

[23] Yang, G., Dekker, R., Gabor, F. A., and Axsater, S., Service parts inventory control with lateral transshipment and pipeline stock flexibility, *Int. J.Production Economics*, Vol.142, 2013, pp.278-289.

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

[25] Shebrooke, C., A multi-echelon technique for recoverable item control, *Operations Research*, Vol.16, 1968, pp.122-141.

[26] Moinzadeh, K. and Lee, H. L., Batch Size and Stocking Levels in Multi-Echelon Reparable Systems, *Management Science*,Vol.32, No.12, 1986, pp.1567-1581.

[27] Riaz, M. W., Two-Echelon Supply Chain Design for Spare Parts with Time Constraints, *UWSpace. http://hdl.handle.net/10012/7914*, 2013.

[28] Kruse, K., Waiting Time in Continuous Review (s,S) Inventory Systems with Constant Lead times,*Operations Research*, Vol.29, 1981, pp.202-207.

[29] Little, J. D., A Proof for the Queiung Formula: $L = \lambda W$., *Operations Research* , Vol.9, No.3, 1961, pp.383-387.

[30] Graves, S., A Multi-Echelon Inventory Model for a Repairable Item with One-for-one Replenishment, *Management Science*, Vol.40, 1985, pp.567-602.

[31] Palm, C., Analysis of the Erlang traffic formulae for busy- signal arrangements, *Ericsson Technics*, Vol.4, 1938, pp.39-58.

[32] Buzacott, J. A. and Shanthikumar, J. G., Stochastic Models of Manufacturing Systems, *Prentice Hall, New Jersey*, 1993.

[33] Candas, M.F. and Kutanoglu, E., Benefits of Considering Inventory in Service Parts Logistic Network Design Problems with Time-based Service Constraints, *IIE Transactions*, Vol.39, 2007, pp.159-176.

[34] van Wijk, A. A. C., Adan, I. and van Houtum G. J., Approximate evaluation of multilocation inventory models with lateral transshipments and hold back levels, *European J. Oper. Res.*, Vol.218, No.3, 2012, pp.624-635.

Creative Commons Attribution License 4.0 (Attribution 4.0 International , CC BY 4.0)