Safety stock placement with market selection under load-dependent lead times

Foad Ghadimi [a,d], Tarik Aouam [a,b] and Reha Uzsoy [c]

[a] Faculty of Economics and Business Administration, Ghent University, Ghent, Belgium

[b] Africa Business School, Mohammed VI Polytechnic University, Rabat, Morocco

[c] Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh 27695-7906, NC, USA

[d] OM PARTNERS NV, Koralenhoeve 23, 2160 Wommelgem (Antwerp), Belgium

Abstract

We study the problem of safety stock placement in a supply chain with market selection decisions. A manufacturer with deterministic, load-dependent lead time supplies multiple warehouses, each serving multiple retailers. Each retailer has access to a set of potential markets with different characteristics. Serving more markets increases revenues, but also increases the manufacturer's lead time, resulting in higher inventory costs. Adopting the Guaranteed Service Approach, we present a nonlinear mixed integer programming model and reformulate it to eliminate integer variables related to service times at warehouses. We then propose a successive piecewise linearization algorithm and a mixed-integer conic quadratic formulation to solve the resulting nonlinear binary formulation. Computational experiments show that the successive piecewise linearization algorithm outperforms two state-of-art solvers, BARON and CPLEX, which are used to solve instances of the original formulation and the mixed-integer conic quadratic reformulation, respectively. The value of incorporating load-dependent lead times is greatest when capacity is limited relative to available demand. The benefit of integrating market selection and safety stock decisions is greatest when capacity is limited and marginal revenue is relatively low.

Keywords: Supply chain management, load-dependent lead time, safety stocks, guaranteed service, market selection.

1 Introduction

The safety stock placement problem determines which nodes in a supply chain network should hold safety stocks and in what quantity to achieve a desired service level under stochastic demand. We adopt the Guaranteed Service Approach (GSA) (Eruguz et al., 2016) under which each inventory location quotes a guaranteed outgoing service time within which all customer orders will be delivered with certainty. Most studies assume exogenous

demand over which the firm has no control. In practice, however, the firm faces the demands of many distinct customers or markets, allowing it to choose which to serve (Bakal et al., 2008; Geunes et al., 2009). However, the higher revenue achieved by serving more markets may be offset by increased production, inventory and backordering costs. Hence aligning marketing, production and inventory decisions is essential to profitability (Jalali et al., 2019).

We consider a supply chain consisting of a manufacturer with load-dependent lead time supplying multiple warehouses each of which, in turn, supplies multiple retailers. Each retailer has access to a set of potential markets with different demand distributions, revenues, and outgoing service time requirements. Serving a market commits the firm to satisfying its entire demand; rejecting it foregoes that market's entire revenue. All warehouses and retailers are potential safety stock holding locations, following a periodic-review base-stock policy with a common review period. Queueing theory (Buzacott and Shanthikumar, 1993; Curry and Feldman, 2000; Hopp and Spearman, 2011) has shown that the average cycle time at the manufacturer is a convex non-decreasing function of its average utilization, which is determined by the total demand of the accepted markets. Accepting more markets increases revenue, but also the manufacturer's utilization and hence its lead time and inventory costs, eventually reducing profit.

The problem can be formulated as a mixed-integer nonlinear program (MINLP) with a non-convex objective function. We first reformulate the problem to eliminate the integer variables related to the service times at warehouses, and linearize all bilinear terms. We then propose a successive piecewise linearization algorithm and a mixed-integer conic quadratic formulation. Computational experiments show that the successive piecewise linearization algorithm outperforms BARON, a state-of-the-art non-convex solver, and CPLEX, which is used to solve instances of the mixed-integer conic quadratic reformulation. Numerical experiments on a serial system with a single market show that the amount of demand accepted is increasing in manufacturing capacity, while the manufacturer's lead time is decreasing. Experiments on a distribution network show that markets with higher marginal revenue and longer outgoing service time are accepted before those with shorter outgoing service times and lower marginal revenue. Markets with shorter lead time and higher inventory holding costs are accepted before those with lower inventory holding cost and longer lead time.

Section 2 reviews relevant literature and highlights our contributions. Section 3 presents

our assumptions and problem formulation. Section 4 describes the successive piecewise linearization algorithm and a mixed-integer conic quadratic reformulation of the original problem. Section 5 presents computational experiments, and Section 6 concludes the paper.

2 Literature review

This paper draws on three related streams of research: inventory-location models, guaranteed service models for safety stock placement with load-dependent lead times, and inventory models with market selection decisions.

Facility Location Models with Inventory Considerations. These models consider facility location/allocation problems in the face of stochastic demand, explicitly considering the savings in inventory costs obtained by pooling safety stocks at facilities serving multiple customers (Farahani et al., 2015; Fathi et al., 2021). Daskin et al. (2002) and Shen et al. (2003) study a location-inventory problem with risk pooling effect, which is formulated and solved as a nonlinear integer programming model. These works are extended by Ozsen et al. (2008) to consider capacitated facilities and by Sourirajan et al. (2007, 2009) to include congestion effects. Lee and Ozsen (2020) propose an efficient tabu search procedure. Atamtürk et al. (2012) formulate several location–inventory problems in a supply chain comprising distribution centers and retailers under a stochastic service approach as conic quadratic mixed-integer programs that can be solved using commercial solvers.

The above-mentioned studies all adopt the stochastic service approach to safety stock planning, in which the amount of material delivered to meet an order during the replenishment lead time is a random variable due to the possibility of stockouts. Under the Guaranteed Service Approach, in contrast, each inventory location specifies a service time within which all orders will be filled with certainty (Eruguz et al., 2014; Graves and Willems, 2000; Simpson Jr, 1958). You and Grossmann (2010) formulate a joint inventory-location problem for uncapacitated warehouses and retailers with fixed endogenous replenishment lead times, and propose a spatial decomposition algorithm based on a Lagrangian heuristic and piecewise linearization. Puga et al. (2019) consider an inventory location problem in a two-stage supply chain with two customer classes with different delivery time requirements and formulate the problem as a conic quadratic mixed integer program. In this paper the lo-

cations of the facilities are fixed, but service level guarantees apply only to markets selected to be served, while no revenue is received from declined markets. The inventory location literature, in contrast, assumes all customers must be served by some facility.

Guaranteed Service (GS) models for safety stock placement. In contrast to inventory location problems, the safety stock placement problem assumes fixed facility locations and seeks to determine the cost-minimizing safety stock levels at these facilities that will ensure the desired customer service level. Stochastic service approaches assume safety stock will be held at all facilities, while GSA approaches allow each inventory location to quote an outgoing service time within which all orders will be delivered in full. Simpson Jr (1958) and Minner (2000) show that in a serial network the optimal solution to the safety stock placement problem under the GSA will be such that a facility either holds no safety stock at all, or enough to decouple it from its downstream stage. Inderfurth (1991) and Inderfurth and Minner (1998) prove this property for distribution and assembly networks, respectively. This *all or nothing* property forms the basis for several dynamic programming algorithms that can solve quite large instances in reasonable solution times (Graves and Willems, 2000), as well as mathematical programming models (Magnanti et al., 2006).

Most early studies of GSA models for safety stock placement (Graves and Willems, 2000; Simpson Jr, 1958) assumed constant, exogenous replenishment lead times. Recent work on GSA models has considered capacitated production nodes with endogenous lead times dependent on production decisions. Kumar and Aouam (2018a,b) use queuing theory to formulate the problem of jointly optimizing lot-sizing decisions and multi-echelon inventory policies in supply networks. Aouam and Kumar (2019) model production facilities as G/M/1 queues to study the impact of subcontracting and overtime on safety stock placement. Kumar and Aouam (2019) use the Tactical Planning Model of Graves (1986) to analyze the impact of production smoothing on safety stock placement decisions, finding significant benefit in coordinating production and inventory decisions. Ghadimi et al. (2020) and Aouam et al. (2021) study the problem of jointly optimizing capacity allocation to production stages and safety stock placement in a general acyclic network. They consider general lead time functions that are decreasing and convex in capacity, and develop a novel Lagrangian decomposition method. Ghadimi and Aouam (2021) optimize processing capacity and safety stocks under a manufacturer budget and warehouse storage capacity.

They consider a serial supply chain consisting of a manufacturer with multiple workcenters supplying multiple products to one warehouse and one retailer, and propose a nested Lagrangian heuristic. The current paper also considers a manufacturer with load-dependent lead time; however, capacity expansion is not an option. The load-dependent lead time at the manufacturer is endogenous to the model and driven by the total accepted demand. Thus even with sufficient capacity to serve all markets, accepting demand up to its nominal capacity may extend the manufacturer's lead time to a point that both pipeline and safety stocks increase to an unacceptable level.

Several papers have used nonlinear clearing functions (Missbauer and Uzsoy, 2020) to integrate safety stock considerations into production planning models. The clearing function captures the workload-dependent nature of the production lead times, while chance constraints seek to ensure service levels are met in the face of stochastic demand. Aouam and Uzsoy (2015, 2012) compare chance constrained models with stochastic programming and robust optimization for single-stage production-inventory systems. Albey et al. (2015) propose a chance-constrained formulation capturing forecast evolution (Heath and Jackson, 1994), which is extended to multistage production systems by Ziarnetzky et al. (2018, 2020). The focus of this work is the release of work into production systems to meet exogenous stochastic demand, as opposed to the endogenous demand addressed in this work.

Inventory models with market selection decisions. Geunes et al. (2004) generalize the classical Economic Order Quantity and Economic Production Quantity models to the case where a producer can choose which markets to serve to maximize their average net profit. Geunes et al. (2005) review demand selection and assignment problems and discuss an optimization model for integrated production and demand planning. Levi et al. (2005) study inventory/facility location models with market selection using a two stage decision model. Bakal et al. (2008) study simultaneous market selection, pricing and order quantity decisions when (i) the firm must offer the same selling price in all markets selected, and (ii) the firm can offer market-specific prices. Taaffe et al. (2008) formulate the problem of jointly determining market selection and ordering decisions as a selective newsvendor problem where demand in each market is normally distributed and dependent on the marketing effort exerted, and show that it can be solved efficiently by ranking the markets according to the ratio of net expected revenue to demand variance.

Shu et al. (2011) extend the model of Geunes et al. (2004) to consider demand uncertainty in a (Q, r) inventory system, where the mean and variance of the demand are known, and propose a polynomial-time algorithm. Geunes et al. (2011) propose a general framework for integrated supply chain planning and logistics problems with market choice. They derive conditions under which a polynomial-time constant-factor approximation algorithm exists for a cost-minimization version of the problem. Van den Heuvel et al. (2012) show that the integrated market selection and production planning problem is NP-hard, and no constant-factor polynomial-time approximation algorithm exists unless P=NP. They identify several special cases that can be solved in polynomial time, and propose a heuristic for large instances. Shu et al. (2013) study an integrated demand selection and multi-echelon inventory problem where inventory is held at both a distribution center and at retailers. The problem is to simultaneously determine the set of demands to fulfill and the multi-echelon inventory control policy to maximize the net profit.

Demand flexibility can also be introduced in the form of order acceptance, where individual customer orders are accepted or rejected rather than all demand from a particular customer or location. Aouam and Brahimi (2013) formulate the problem of integrated order acceptance and production planning with load-dependent lead times while allowing partial order acceptance under uncertain demand. They adopt a robust optimization approach resulting in a linear program. Brahimi et al. (2015) study the combined effect of load-dependent lead times, order acceptance and flexible customer due dates under deterministic demand, proposing two relax-and-fix heuristics for the integrated problem. Aouam et al. (2018) use a robust optimization approach to model demand uncertainty and clearing functions (Missbauer and Uzsoy, 2020) to capture queueing behavior in an integrated order acceptance and production planning problem. Ghadimi et al. (2022) present centralized and decentralized models for coordinating order acceptance and release planning under load-dependent lead times. Their centralized models use detailed information at the item level, while the decentralized models decompose the decision process into order acceptance and order release subproblems that are solved sequentially.

3 Problem statement

We present our formulation of the integrated safety stock placement and market selection problem using the following notation:

Sets

$\mathbb{I}$	set of warehouses
$\mathbb{J}$	set of retailers
$\mathbb{J}_i$	set of retailers supplied by warehouse i
$\mathbb{N}$	set of all potential markets
$\mathbb{N}_j$	set of potential markets available to retailer j
$\mathbb{A}$	set of arcs (i, j) representing material flow from warehouse i to retailer j
$\mathbb{Q}$	set of feasible lead time - utilization pairs (l_q, u_q) at the manufacturer

Parameters

λ_n	demand rate of market n (units per period)
π_n	marginal revenue of market n (€ per unit per period)
c	unit production cost at the manufacturer (€ per unit)
w	unit work-in-process holding cost at the manufacturer (€ per unit per period)
c^o	unit expediting cost of delayed items at the manufacturer (€ per unit per period)
c^{exp}	unit expediting cost at the manufacturer (€ per unit per period)
w_i^{whs}	unit marginal pipeline inventory cost between the manufacturer and warehouse i (€ per unit).
w_j^{ret}	unit marginal pipeline inventory cost between retailer j and the warehouse supplying it (€ per unit).
r	capacity at the manufacturer (units per period)
τ_i^{whs}	logistics delay at warehouse i (periods)
τ_j^{ret}	logistics delay at retailer j (periods)
z_i^{whs}	safety factor at warehouse i
z_j^{ret}	safety factor at retailer j
h_i^{whs}	inventory holding cost at warehouse i (€ per unit per period)
h_j^{ret}	inventory holding cost at retailer j (€ per unit per period)
h_j^{exp}	unit expediting cost at retailer j (€ per unit per period)
$H^{whs}(z_i)$	(lead time demand) uncertainty cost at warehouse i (€ per unit per period)
$H^{ret}(z_j)$	(lead time demand) uncertainty cost at retailer j (€ per unit per period)
s_n	maximum external outgoing service time for the demand at market n (periods)
l_q	qth possible lead time (integer value) at the manufacturer (periods)
u_q	manufacturer utilization level corresponding to lead time l_q

Decision variables

S_i	outgoing service time at warehouse i (integer valued, in periods)
A_n	binary variable taking the value of 1 if market n is selected, and 0 otherwise

X_q binary variable taking value of 1 if lead time - utilization pair (l_q, u_q) is selected at the manufacturer, and 0 otherwise

P_i binary variable taking value of 1 if inventory is coupled, i.e., held only at downstream retailers of warehouse i; and 0 otherwise. Let $P_i' = 1 - P_i$.

3.1 *Model description and assumptions*

Supply chain network. We consider a single manufacturer supplying warehouses $i \in \mathbb{I}$, each of which then supplies retailers $j \in \mathbb{J}_i$ as shown in Figure 1. Each retailer j is supplied by a specific warehouse i. The set of arcs defining material flow between warehouses and retailers is denoted by $\mathbb{A}$. Each retailer j has access to potential markets $n \in \mathbb{N}_j$ with marginal revenue π_n and external outgoing service time s_n, denoting the maximum time it is prepared to wait for an order to be delivered. Demand for each market n arrives following a stationary Poisson process with rate λ_n. Inventory can be held at warehouses and retailers, but not at the manufacturer. Each warehouse and retailer has a fixed lead time (logistics delay) denoted by τ_i^{whs} and τ_j^{ret}, respectively. The manufacturer is modelled as a workstation with service rate (capacity) r and staging areas for raw materials and finished goods. The load-dependent lead time l_q at the manufacturer is deterministic and depends on the utilization level u_q, which is determined by the total accepted demand. Set $\mathbb{Q}$ collects all the lead time and utilization level pairs (l_q, u_q). The (l_q, u_q) pair adopted at the manufacturer is determined in our model by the binary decision variable X_q.

Demand process. Demand for market n arrives at the retailer following a Poisson process, implying exponentially distributed inter-arrival times with mean $\frac{1}{\lambda_n}$. The demand $D_n(t)$ of market n in period t at the retailer follows a Poisson process with average rate λ_n that is observed at the beginning of period t (Axsäter, 2015). Demand at each market is assumed to be independent and identically distributed (i.i.d.) across periods and markets.

The total external demand that must be served by the network is defined by the market selection decisions represented by the binary decision variables A_n. The firm commits to satisfying the entire demand of all accepted markets over the horizon, and receives no revenue from rejected markets. Once a subset of markets is selected, specifying the values of the binary decision variables $A(n)$, retailer j's external demand in period t is given by the random variable $D_j(t) = \sum_{n \in \mathbb{N}_j} D_n(t) A_n$. The demand at warehouse i in period t is

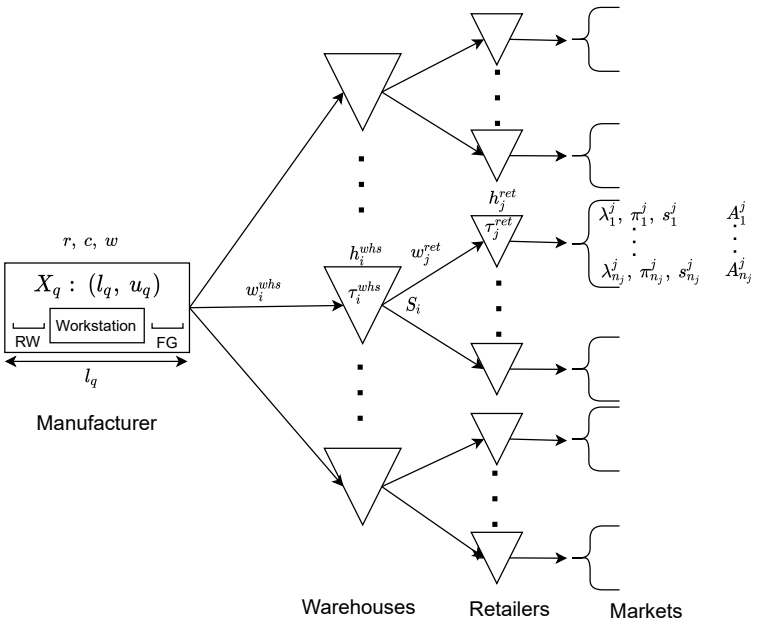

Figure 1: Schematic model of the considered supply chain network

then $D_i(t) = \sum_{j\in\mathbb{J}_i} \sum_{n\in\mathbb{N}_j} D_n(t)A_n$, and that at the manufacturer $D(t) = \sum_{n\in\mathbb{N}} D_n(t)A_n$.

Guaranteed service times. Each warehouse or retailer quotes a guaranteed outgoing service time to all its downstream nodes within which it can satisfy all orders with certainty. Since we assume single sourcing, each node's incoming service time is equal to the outgoing service time quoted by its upstream supplier. The outgoing service time at retailers cannot exceed the lead time, i.e. $s_n \leq \tau_j^{ret}$ for all $j \in \mathbb{J}$ and all $n \in \mathbb{N}_j$.

Replenishment process. Inventory at each stage is replenished following a periodic review base stock policy with a common review period shared by all stages. A base stock level is set at each warehouse and retailer to cover demand over the net replenishment lead time T with a target service level α_i^{whs} at warehouse i and α_j^{ret} at retailer j. Demand over the net replenishment time T is normally distributed with mean $\lambda_n T$ and standard deviation $\sqrt{\lambda_n T}$. Following standard GSA calculations (Graves and Willems, 2000), the net replenishment time of retailer j for serving market n is $T_{jn} = S_i + \tau_j^{ret} - s_n$, implying a base stock level of $\mathcal{B}_j^{ret} = \sum_{n\in\mathbb{N}_j} \lambda_n T_{jn} A_n + z_j^{ret} \sqrt{\sum_{n\in\mathbb{N}_j} \lambda_n T_{jn} A_n}$ at this retailer, where z_j^{ret} is the safety factor corresponding to the service level α_j^{ret}. Similarly, the net replenishment time at warehouse i is $T_i = \sum_{q\in\mathbb{Q}} l_q X_q + \tau_i^{whs} - S_i$, yielding a base stock level $\mathcal{B}_i^{whs} = \left(\sum_{j\in\mathbb{J}_i} \sum_{n\in\mathbb{N}_j} \lambda_n A_n\right) T_i + z_i^{whs} \sqrt{\left(\sum_{j\in\mathbb{J}_i} \sum_{n\in\mathbb{N}_j} \lambda_n A_n\right) T_i}$. Following Klosterhalfen et al. (2013) and Aouam and Kumar (2019), any demand exceeding the base stock levels is met

by expediting items from the pipeline inventory. We assess a common unit cost for each expedited item, irrespective of the duration, and assume that the pipeline inventory always exceeds the amount to be expedited.

Manufacturer lead time. The manufacturer specifies a guaranteed, deterministic lead time l_q based on its planned utilization level $u_q = \frac{\sum_{n\in\mathbb{N}} \lambda_n A_n}{r}$. The time l_q between an order being received by the manufacturer and its shipment to a warehouse can be represented as the sum of two components such that $l_q = t_1 + t_2$. t_1 represents the average time an item spends in the raw material staging area waiting to be processed, and t_2 the deterministic time it will spend in process and waiting in the finished goods staging area. The value of t_2 is determined such that a target fraction α of processed items is completed and ready to be shipped within this fixed time. Units that are not available in the finished good staging area, and thus prevent shipment of a complete order to the warehouses, are expedited from the workstation queue, ensuring that the entire order placed by a warehouse is satisfied within the manufacturer's guaranteed service time, given by its lead time l_q. The average fraction of time the manufacturer must resort to expediting is $(1-\alpha)$, and the expected number of delayed items per period is $(1-\alpha)\sum_{n\in\mathbb{N}} \lambda_n A_n$. The manufacturer receives orders in its raw material staging area at the start of a period. Because demand at each market follows a Poisson process, the total demand also follows a Poisson process with average rate $\sum_{n\in\mathbb{N}} \lambda_n A_n$. Hence the manufacturer releases items to the workstation queue with exponentially distributed interarrival times with mean $\frac{1}{\sum_{n\in\mathbb{N}} \lambda_n A_n}$, after which they are processed in First-Come First-Served (FCFS) order. The service time at the workstation can follow any distribution with mean $\frac{1}{r}$, so we model the manufacturer as a M/G/1 queue.

Uncertainty costs $H^{whs}(z_i)$ and $H^{ret}(z_j)$. The uncertainty costs associated with lead time demand at warehouses and retailers, consisting of the unit inventory holding cost and expediting costs, are given by $H^{whs}(z_i) = h_i^{whs} z_i^{whs} + \left(c^{exp} + h_i^{whs} - w\right) G(z_i^{whs})$ and $H^{ret}(z_j) = h_j^{ret} z_j^{ret} + \left(h_j^{exp} + h_j^{ret} - \frac{w_j^{ret}}{\tau_j^{ret}}\right) G(z_j^{ret})$, respectively. The derivation of these costs is given in Appendix A of the Electronic Supplement.

3.2 *Problem Formulation*

The integrated safety stock placement and market selection problem **SSPM** determines the set of markets to be served by each retailer j, the utilization u_q and lead time l_q at the

manufacturer and the service times S_i at the warehouses to maximize expected profit. The formulation is as follows:

$$\textbf{SSPM} \quad \max \quad \sum_{n\in\mathbb{N}}(\pi_n - c)\lambda_n A_n - w\sum_{q\in\mathbb{Q}} l_q X_q \left(\sum_{n\in\mathbb{N}}\lambda_n A_n\right) - c^o(1-\alpha)\left(\sum_{n\in\mathbb{N}}\lambda_n A_n\right)$$
$$- \sum_{i\in\mathbb{I}} w_i^{whs} \sum_{j\in\mathbb{J}_i}\sum_{n\in\mathbb{N}_j}\lambda_n A_n - \sum_{j\in\mathbb{J}} w_j^{ret} \sum_{n\in\mathbb{N}_j}\lambda_n A_n - \sum_{i\in\mathbb{I}} H^{whs}(z_i)\sqrt{\sum_{j\in\mathbb{J}_i}\sum_{n\in\mathbb{N}_j}\lambda_n A_n\left(\sum_{q\in\mathbb{Q}} l_q X_q + \tau_i^{whs} - S_i\right)}$$
$$- \sum_{i\in\mathbb{I}}\sum_{j\in\mathbb{J}_i} H^{ret}(z_j)\sqrt{\sum_{n\in\mathbb{N}_j}\lambda_n A_n(S_i + \tau_j^{ret} - s_n)} \tag{1}$$

subject to:

$$\sum_{n\in\mathbb{N}}\lambda_n A_n \le r\sum_{q\in\mathbb{Q}} u_q X_q \tag{2}$$

$$\sum_{q\in\mathbb{Q}} X_q \le 1 \tag{3}$$

$$\sum_{q\in\mathbb{Q}} l_q X_q + \tau_i^{whs} - S_i \ge 0 \qquad \forall i \in \mathbb{I} \tag{4}$$

$$S_i \in \mathbb{Z}^+ \qquad \forall i \in \mathbb{I} \tag{5}$$

$$A_n,\ X_q \in \{0,1\} \qquad \forall n \in \mathbb{N}, \forall q \in \mathbb{Q} \tag{6}$$

The objective function (1) maximizes the total profit, where the first term is the total revenue obtained from serving the accepted markets minus total production costs, the second the total expected WIP holding cost at the manufacturer, and the third the cost incurred at the manufacturer to guarantee lead time l_q by expediting delayed units. The fourth and fifth terms represent the total pipeline inventory cost between the manufacturer and warehouses and the warehouses and retailers, respectively, while the sixth and seventh capture the inventory costs at the warehouses and retailers. Constraints (2) define the utilization level at the manufacturer, while (3) ensure that only one lead time - utilization level pair is selected at the manufacturer. Constraints (4) restrict the net replenishment lead time at the warehouses to nonnegative values, while (5) and (6) define all service times as integer variables and market selection variables X_q as binary.

4 Reformulations and solution methods

Problem **SSPM** is a mixed-integer nonlinear program (MINLP) with a nonconvex objective function and nonlinear constraints. The nonlinearity of **SSPM** arises from several bilinear terms and the square root terms that compute the safety stock costs. We first present an

alternative formulation **A-SSPM** eliminating the integer variables related to service times S_i at the warehouses and linearizing all bilinear terms. We then explore two different ways of dealing with the square root terms remaining in **A-SSPM**: 1) a conic quadratic formulation **CQ** that can be solved using standard solvers while adding valid inequalities to improve computational efficiency, and 2) a successive piecewise linearization algorithm **SPLA** that iteratively improves lower and upper bounds.

4.1 An alternative formulation of SSPM

Simpson Jr (1958) and Minner (2000) show that in a serial network the optimal solution to the safety stock placement problem under the GSA has the all-or-nothing property such that a stage either holds no safety stock or enough to decouple it from its downstream stage. Inderfurth (1991) and Inderfurth and Minner (1998) prove this property for distribution and assembly networks, respectively. This property implies two different policies for safety stock placement at a retailer j served by warehouse i: *inventory coupling*, holding inventory only at the retailer, and *inventory decoupling* when it is held at both warehouse and retailer.

Under the coupling policy, indicated by $P_i = 1$, safety stock is held only at retailers served by warehouse i and none at warehouse i. In this case $S_i = \sum_{q\in\mathbb{Q}} l_q X_q + \tau_i^{whs}$, and the net replenishment lead time at warehouse i is zero. The net replenishment lead time at retailer j served by warehouse i will be $NLT_{ijn} = \sum_{q\in\mathbb{Q}} l_q X_q + \tau_i^{whs} + \tau_j^{ret} - s_n$. Under the second policy, denoted by $P_i' = 1$, inventory is held at warehouse i and all retailers it serves. In this case $S_i = 0$, and the net replenishment lead time at warehouse i is $NLT_i = \sum_{q\in\mathbb{Q}} l_q X_q + \tau_i^{whs}$. Retailer j served by warehouse i will have net replenishment lead time $NLT_{ijn} = \tau_j^{ret} - s_n$. The decision variables P_i and P_i' allow the service time variables S_i to be eliminated. The bilinear terms arising from the product of binary variables can be linearized as presented in Appendix B in the Electronic Supplement, yielding the following

formulation:

$$\textbf{A-SSPM} \quad \max \quad \sum_{n\in\mathbb{N}} \pi_n \lambda_n A_n - w \sum_{q\in\mathbb{Q}} \sum_{n\in\mathbb{N}} \lambda_n l_q X A_{qn} - c^o(1-\alpha)\left(\sum_{n\in\mathbb{N}} \lambda_n A_n\right) - \sum_{i\in\mathbb{I}} w_i^{whs} \sum_{j\in\mathbb{J}_i} \sum_{n\in\mathbb{N}_j} \lambda_n A_n$$
$$- \sum_{j\in\mathbb{J}} w_j^{ret} \sum_{n\in\mathbb{N}_j} \lambda_n A_n - \sum_{i\in\mathbb{I}} \sum_{j\in\mathbb{J}_i} H^{ret}(z_j) \sqrt{\sum_{n\in\mathbb{N}_j} \lambda_n (\tau_i^{whs} + \tau_j^{ret} - s_n) PA_{in} + \sum_{n\in\mathbb{N}_j} \sum_{q\in\mathbb{Q}} \lambda_n l_q PAX_{inq}}$$
$$- \sum_{i\in\mathbb{I}} H^{whs}(z_i) \sqrt{\sum_{j\in\mathbb{J}_i} \sum_{n\in\mathbb{N}_j} \lambda_n \tau_i^{whs} PA'_{in} + \sum_{j\in\mathbb{J}_i} \sum_{n\in\mathbb{N}_j} \sum_{q\in\mathbb{Q}} \lambda_n l_q PAX'_{inq}}$$
$$- \sum_{i\in\mathbb{I}} \sum_{j\in\mathbb{J}_i} H^{ret}(z_j) \sqrt{\sum_{n\in\mathbb{N}_j} \lambda_n (\tau_j^{ret} - s_n) PA'_{in}} \tag{7}$$

subject to: constraints (2), (3), (6), (B.11)-(B.21), and (B.23)-(B.29)

$$P_i + P'_i = 1 \qquad \forall i \in \mathbb{I} \tag{8}$$

$$P_i, P'_i \in \{0,1\} \qquad \forall i \in \mathbb{I} \tag{9}$$

Constraints (8) ensure that only one of the two inventory policies (coupling or decoupling) is selected for each retailer j supplied by warehouse i, while (9) define binary variables. However, **A-SSPM** remains difficult to solve due to the square root terms in the objective function. We now propose two alternative approaches to solving **A-SSPM**.

4.2 *A mixed-integer conic quadratic reformulation*

Noting that for any binary variable X we have $X = X^2$, we can reformulate **A-SSPM** as the following mixed-integer conic quadratic program (MICQP) by introducing three new auxiliary variables Y_{ij}, Y'_i and Y''_{ij}.

$$\textbf{CQ} \quad \max \quad \sum_{n\in\mathbb{N}} \pi_n \lambda_n A_n - w \sum_{q\in\mathbb{Q}} \sum_{n\in\mathbb{N}} \lambda_n l_q X A_{qn} - c^o(1-\alpha)\left(\sum_{n\in\mathbb{N}} \lambda_n A_n\right) - \sum_{i\in\mathbb{I}} w_i^{whs} \sum_{j\in\mathbb{J}_i} \sum_{n\in\mathbb{N}_j} \lambda_n A_n$$
$$- \sum_{j\in\mathbb{J}} w_j^{ret} \sum_{n\in\mathbb{N}_j} \lambda_n A_n - \sum_{i\in\mathbb{I}} \sum_{j\in\mathbb{J}_i} H^{ret}(z_j) Y_{ij} - \sum_{i\in\mathbb{I}} H^{whs}(z_i) Y'_i - \sum_{i\in\mathbb{I}} \sum_{j\in\mathbb{J}_i} H^{ret}(z_j) Y''_{ij} \tag{10}$$

subject to: constraints (2), (3), (6), (8), (9), (B.11)-(B.21), and (B.23)-(B.29).

$$Y_{ij}^2 \geq \sum_{n\in\mathbb{N}_j} \lambda_n (\tau_i^{whs} + \tau_j^{ret} - s_n) PA_{in}{}^2 + \sum_{n\in\mathbb{N}_j} \sum_{q\in\mathbb{Q}} \lambda_n l_q PAX_{inq}{}^2 \qquad \forall (i,j) \in \mathbb{A} \tag{11}$$

$$Y_i'^2 \geq \sum_{j\in\mathbb{J}_i} \sum_{n\in\mathbb{N}_j} \lambda_n \tau_i^{whs} PA'_{in}{}^2 + \sum_{j\in\mathbb{J}_i} \sum_{n\in\mathbb{N}_j} \sum_{q\in\mathbb{Q}} \lambda_n l_q PAX'_{inq}{}^2 \qquad \forall i \in \mathbb{I} \tag{12}$$

$$Y_{ij}''^2 \geq \sum_{n\in\mathbb{N}_j} \lambda_n (\tau_j^{ret} - s_n) PA'_{in}{}^2 \qquad \forall (i,j) \in \mathbb{A} \tag{13}$$

$$Y_{ij},\ Y'_i,\ Y''_{ij} \in \mathbb{R}^+ \qquad \forall i \in \mathbb{I},\ \forall (i,j) \in \mathbb{A} \tag{14}$$

Model **CQ** has a linear objective function and conic quadratic constraints, and can be solved directly by standard solvers such as CPLEX (Atamtürk et al., 2012; Shahabi et al., 2013,

2014). However, the size of the formulation increases rapidly in network size, especially the number of constraints and variables for the linearization, and the computation of constraints (11)-(13) becomes time-consuming.

To improve the solution time of **CQ**, we introduce a set of valid inequalities. Recall that $XA_{qn} = A_n X_q$. Since $\sum_{q\in\mathbb{Q}} X_q \leq 1$, we can write the valid inequalities

$$\sum_{q\in\mathbb{Q}} XA_{qn} \leq 1 \qquad \forall n \in \mathbb{N} \tag{15}$$

Similarly for binary variables $PAX_{inq} = P_i A_n X_q$ and $PAX'_{inq} = P'_i A_n X_q$, we get

$$\sum_{q\in\mathbb{Q}} PAX_{inq} \leq 1 \qquad \forall (i,j) \in \mathbb{A},\ \forall n \in \mathbb{N}_j \tag{16}$$

$$\sum_{q\in\mathbb{Q}} PAX'_{inq} \leq 1 \qquad \forall (i,j) \in \mathbb{A},\ \forall n \in \mathbb{N}_j \tag{17}$$

PA_{in}, PA'_{in}, PAX_{inq}, and PAX'_{inq} are products of binary variables $P_i A_n$, $P'_i A_n$, $P_i A_n X_q$, and $P'_i A_n X_q$, respectively. Since $P_i + P'_i = 1$ and $\sum_{q\in\mathbb{Q}} X_q \leq 1$, we have the valid inequalities

$$\sum_{q\in\mathbb{Q}} PAX_{inq} + PAX'_{inq} \leq 1 \qquad \forall (i,j) \in \mathbb{A},\ \forall n \in \mathbb{N}_j \tag{18}$$

$$PA_{in} + PA'_{in} \leq 1 \qquad \forall (i,j) \in \mathbb{A},\ \forall n \in \mathbb{N}_j \tag{19}$$

$$PAX_{inq} + PAX'_{inq} \leq 1 \qquad \forall (i,j) \in \mathbb{A},\ \forall n \in \mathbb{N}_j,\ \forall q \in \mathbb{Q} \tag{20}$$

Our computational experiments show that the addition of these valid inequalities significantly improves the optimality gap and solution time.

4.3 *The successive piecewise linearization algorithm (SPLA)*

SPLA iteratively refines lower and upper bounds on the objective function by successively refining piecewise linear approximations that yield integer linear programs. We use the "multiple-choice" formulation of Magnanti et al. (2006) to approximate the square root term $\sqrt{\Theta_{ij}}$, using a piecewise linear function defined over a set of intervals $K_{ij} = \{1,\ 2,\ 3,\ ...,\ k\}$ and the lower and upper bounds on Θ_{ij} in each interval given by ν_{ij0}, ν_{ij1}, ν_{ij2}, ..., $\nu_{ij,k-1}$, ν_{ijk}. Applying the multiple-choice approach to **A-SSPM**, with $\Theta_{ij} = \sum_{n\in\mathbb{N}_j} \lambda_n(\tau_i^{whs} + \tau_j^{ret} - s_n)PA_{in} + \sum_{n\in\mathbb{N}_j}\sum_{q\in\mathbb{Q}} \lambda_n l_q PAX_{inq}$, $\Theta'_i = \sum_{j\in\mathbb{J}_i}\sum_{n\in\mathbb{N}_j} \lambda_n \tau_i^{whs} PA'_{in}$ $+ \sum_{j\in\mathbb{J}_i}\sum_{n\in\mathbb{N}_j}\sum_{q\in\mathbb{Q}} \lambda_n l_q PAX'_{inq}$, and $\Theta''_{ij} = \sum_{n\in\mathbb{N}_j} \lambda_n(\tau_j^{ret} - s_n)PA'_{in}$ gives the following

piecewise linear approximation problem:

$$\textbf{PLA-SSPM} \quad \max \quad \sum_{n\in\mathbb{N}} \pi_n\lambda_n A_n - w\sum_{q\in\mathbb{Q}}\sum_{n\in\mathbb{N}} \lambda_n l_q X A_{qn} - c^o(1-\alpha)\left(\sum_{n\in\mathbb{N}}\lambda_n A_n\right)$$
$$- \sum_{i\in\mathbb{I}} w_i^{whs} \sum_{j\in\mathbb{J}_i}\sum_{n\in\mathbb{N}_j} \lambda_n A_n - \sum_{j\in\mathbb{J}} w_j^{ret} \sum_{n\in\mathbb{N}_j} \lambda_n A_n - \sum_{i\in\mathbb{I}}\sum_{j\in\mathbb{J}_i} H^{ret}(z_j) \sum_{k\in\mathbb{K}} \left(f_{ijk}\Psi_{ijk} + a_{ijk}O_{ijk}\right)$$
$$- \sum_{i\in\mathbb{I}} H^{whs}(z_i) \sum_{k\in\mathbb{K}} \left(f'_{ik}\Psi'_{ik} + a'_{ik}O'_{ik}\right) - \sum_{i\in\mathbb{I}}\sum_{j\in\mathbb{J}_i} H^{ret}(z_j) \sum_{k\in\mathbb{K}} \left(f''_{ijk}\Psi''_{ijk} + a''_{ijk}O''_{ijk}\right) \tag{21}$$

subject to: constraints (2), (3), (6), (8), (9), (B.11)-(B.21), and (B.23)-(B.29).

$$\sum_{k\in\mathbb{K}} O_{ijk} \geq \sum_{n\in\mathbb{N}_j} \lambda_n(\tau_i^{whs} + \tau_j^{ret} - s_n) P A_{in} + \sum_{n\in\mathbb{N}_j}\sum_{q\in\mathbb{Q}} \lambda_n l_q P A X_{inq} \qquad \forall i\in\mathbb{I},\ \forall j\in\mathbb{J} \tag{22}$$

$$\sum_{k\in\mathbb{K}} \Psi_{ijk} = 1 \qquad \forall i\in\mathbb{I},\ \forall j\in\mathbb{J} \tag{23}$$

$$\nu_{ijk-1}\Psi_{ijk} \leq O_{ijk} \leq \nu_{ijk}\Psi_{ijk} \qquad \forall i\in\mathbb{I},\ \forall j\in\mathbb{J},\ \forall k\in\mathbb{K} \tag{24}$$

$$\sum_{k\in\mathbb{K}} O'_{ik} \geq \sum_{j\in\mathbb{J}_i}\sum_{n\in\mathbb{N}_j} \lambda_n \tau_i^{whs} P A'_{in} + \sum_{j\in\mathbb{J}_i}\sum_{n\in\mathbb{N}_j}\sum_{q\in\mathbb{Q}} \lambda_n l_q P A X'_{inq} \qquad \forall i\in\mathbb{I} \tag{25}$$

$$\sum_{k\in\mathbb{K}} \Psi'_{ik} = 1 \qquad \forall i\in\mathbb{I} \tag{26}$$

$$\nu'_{ik-1}\Psi'_{ik} \leq O'_{ik} \leq \nu'_{ik}\Psi'_{ik} \qquad \forall i\in\mathbb{I},\ \forall k\in\mathbb{K} \tag{27}$$

$$\sum_{k\in\mathbb{K}} O''_{ijk} \geq \sum_{n\in\mathbb{N}_j} \lambda_n(\tau_j^{ret} - s_n) P A'_{in} \qquad \forall i\in\mathbb{I},\ \forall j\in\mathbb{J} \tag{28}$$

$$\sum_{k\in\mathbb{K}} \Psi''_{ijk} = 1 \qquad \forall i\in\mathbb{I},\ \forall j\in\mathbb{J} \tag{29}$$

$$\nu''_{ijk-1}\Psi''_{ijk} \leq O''_{ijk} \leq \nu''_{ijk}\Psi''_{ijk} \qquad \forall i\in\mathbb{I},\ \forall j\in\mathbb{J},\ \forall k\in\mathbb{K} \tag{30}$$

$$O_{ijk}, O'_{ik}, O''_{ijk} \in \mathbb{R}^+,\ \Psi_{ijk}, \Psi'_{ik}, \Psi''_{ijk} \in \{0,1\} \qquad \forall i\in\mathbb{I},\ \forall j\in\mathbb{J},\ \forall k\in\mathbb{K} \tag{31}$$

where $a_{ijk} = \frac{\sqrt{\nu_{ijk}}-\sqrt{\nu_{ijk-1}}}{\nu_{ijk}-\nu_{ijk-1}}$ and $f_{ijk} = \sqrt{\nu_{ijk}} - a_{ijk}\nu_{ijk}$ are the slope and intercept, respectively, of the approximating line segments. Similarly, we have $a'_{ik} = \frac{\sqrt{\nu'_{ik}}-\sqrt{\nu'_{ik-1}}}{\nu'_{ik}-\nu'_{ik-1}}$, $f'_{ik} = \sqrt{\nu'_{ik}} - a'_{ik}\nu'_{ik}$, $a''_{ijk} = \frac{\sqrt{\nu''_{ijk}}-\sqrt{\nu''_{ijk-1}}}{\nu''_{ijk}-\nu''_{ijk-1}}$ and $f''_{ijk} = \sqrt{\nu''_{ijk}} - a''_{ijk}\nu''_{ijk}$. Variables O_{ijk}, O'_{ik}, and O''_{ijk} define the level of variables Θ_{ij}, Θ'_i, and Θ''_{ij}, respectively, in each interval k.

To manage the model size, **SPLA** adds linear segments iteratively to improve the approximation. In the first step, a one-piece linear approximation is considered, i.e. all square root terms are replaced with their secants. In this case, the optimal objective value of **PLA-SSPM** provides an upper bound on the optimal value of **A-SSPM** and the optimal solution of **PLA-SSPM** is feasible for **A-SSPM**, providing a lower bound. In each subsequent iteration, new intervals are introduced, based on the optimal values of variables Θ_{ij}, Θ'_i, and Θ''_{ij} obtained in the previous iteration, and the resulting instance of **PLA-SSPM** is solved. Each iteration provides lower and upper bounds on the optimal value of **A-SSPM**.

Iterations continue until the best lower and upper bounds are within a specified tolerance *tol*. Appendix C of the Electronic Supplement summarizes the **SPLA** procedure.

5 Computational experiments

Our computational study consists of three experiments. We first examine a simple serial system with a single market, where the principal decision is what fraction of the available demand to serve. The parameters for this system are presented in Appendix D of the Electronic Supplement. Our second experiment, reported in Appendix E of the Electronic Supplement for brevity, examines a distribution system. Our final computational experiments examine the computational performance of the solution procedures.

5.1 *Analysis of a serial system with one market*

We first consider a serial supply chain network with one manufacturer, one warehouse and one retailer as depicted in Figure 2 to obtain insight into the structure of an optimal solution. The retailer has access to only one market, so the firm's decision is what fraction of this market's demand should be accepted to maximize profit, allowing all decision variables to take fractional values. The manufacturer is modelled as an $M/M/1$ queue.

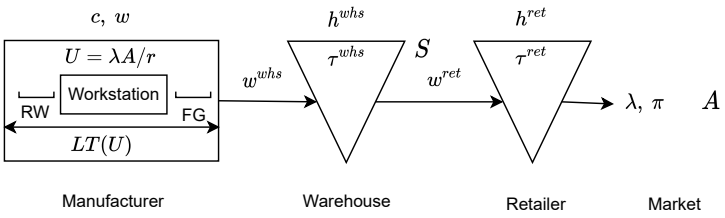

Figure 2: Schematic model of the serial system

The problem, denoted by **P**, can be formulated as follows:

$$\mathbf{P} \max A\Bigg((\pi - c)\lambda - wLT(U)\lambda - c^o(1-\alpha)\lambda - w^{whs}\lambda - w^{ret}\lambda - H^{whs}(z)\sqrt{\lambda(LT(U) + \tau^{whs} - S)} - H^{ret}(z)\sqrt{\lambda(S + \tau^{ret})}\Bigg) \tag{32}$$

subject to:

$$\lambda A \leq rU \tag{33}$$

$$LT(U) + \tau^{whs} - S \geq 0 \tag{34}$$

$$0 \leq A,\ U \leq 1 \tag{35}$$

$$U,\ LT(U),\ S,\ A \in \mathbb{R}^+ \tag{36}$$

Effect of capacity on profit, market selection and inventory placement. Figure 3 shows the effect of the capacity r on the system profit for low ($\pi = 2c = 20$) and high ($\pi = 5c = 50$) levels of marginal revenue. As r increases from 1 to 35 units per period, profit increases from 0.00106 to 53.21 € when $\pi = 20$, and from 13.678 to 803.21 € when $\pi = 50$. As r increases, more demand is accepted, increasing revenue and profit which, as expected, is increasing in the marginal revenue π.

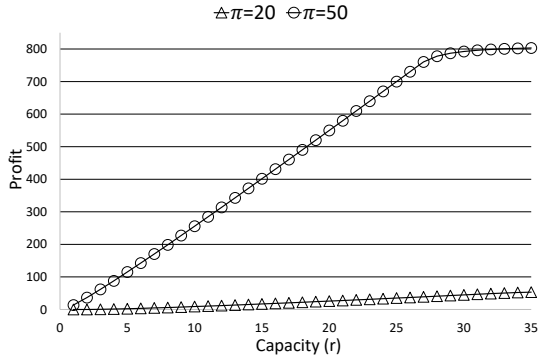

Figure 3: Effect of capacity level r on the profit for $\pi = 20$ and $\pi = 50$.

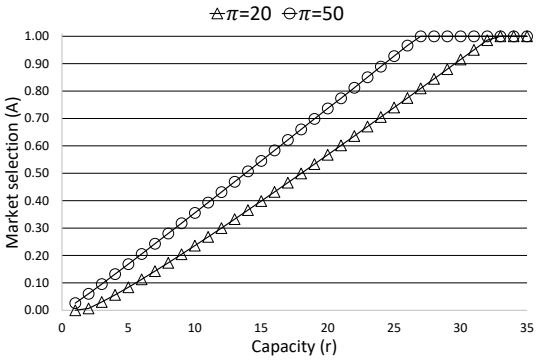

Figure 4: Effect of capacity level r on market selection decision A for $\pi = 20$ and $\pi = 50$.

Figure 4 shows the effect of r on the fraction A of demand served (i.e.,market selection decisions) for $\pi = 20$ and $\pi = 50$. As r increases from 1 to 35 units per period, the fraction of demand served increases from 0.0001 and 0.0261 to 1, for $\pi = 20$ and $\pi = 50$, respectively. A higher value of r reduces the manufacturer's lead time, allowing the firm to accept more demand and generate more revenue. There is a capacity level ($r = 33$, when $\pi = 20$; $r = 27$ when $\pi = 50$) above which the entire demand is accepted, i.e., $A = 1$. This threshold is

decreasing in the marginal revenue π; all available demand can be accepted for lower values of r when π is high enough.

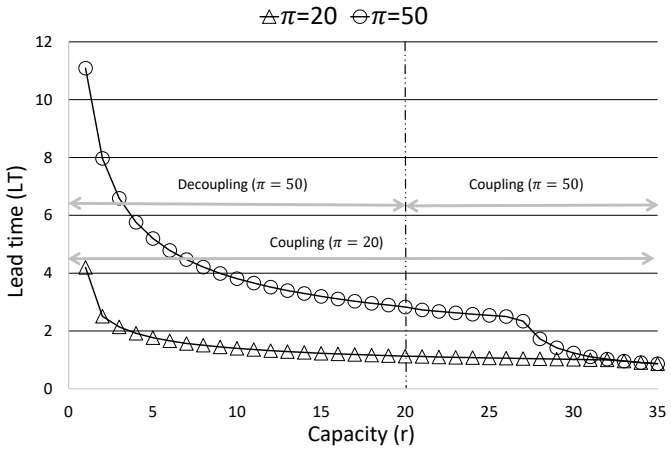

Figure 5: Effect of capacity r on the manufacturer's lead time and inventory placement for $\pi = 20$ and $\pi = 50$.

Figure 5 shows the manufacturer's lead time as a function of capacity r. As r increases from 1 to 35, lead time decreases from 4.2 to 0.87 periods when $\pi = 20$, and from 11.09 to 0.87 periods, when $\pi = 50$. Lead time is decreasing in the capacity r and increasing in the accepted demand λA. As r increases more demand is accepted, but only as long as lead time decreases. Figure 5 also shows that when $r < 33$, the lead time is always higher for $\pi = 50$ than for $\pi = 20$, because under the former the high revenue offsets the extra inventory costs due to the higher lead time. When $r > 33$, the lead time is equal for both $\pi = 20$ and $\pi = 50$ since the entire demand is accepted.

Figure 5 shows that there is a lead time threshold above which inventory is decoupled, i.e., held at both the warehouse and the retailer. When lead time is below this threshold no inventory is held at the warehouse, and inventory is only held at the retailer. This threshold is at $r = 20$ and $LT = 2.83$ when $\pi = 50$. When $\pi = 25$, this threshold happens when $r < 1$. Therefore, when marginal revenue is low, inventory coupling is optimal for lower capacity.

Effect of considering nonlinear relationship between capacity and lead time. We now examine the benefit of considering load-dependent lead times in the SSPM problem. For comparison purposes we consider a Two-Step approach that sequentially optimizes the market selection and safety stock decisions assuming a fixed exogenous lead time at the manufacturer. In this approach, we first replace $LT(U)$ in problem **P** with a fixed lead

time LT based on a target utilization level U to obtain the optimal market selection A^* for this utilization level. We then input the resulting market selection A^* into the problem **P** to obtain the optimal utilization, lead time and service time decisions. Comparing the objective function value of the Two-Step approach $obj^{Two-Step}$ and that of problem **P**, $obj^{\mathbf{P}}$ yields the value of considering the nonlinear relationship between capacity and lead time, computed as $VOCN = \dfrac{obj^{\mathbf{P}} - obj^{Two-Step}}{obj^{Two-Step}} \times 100$.

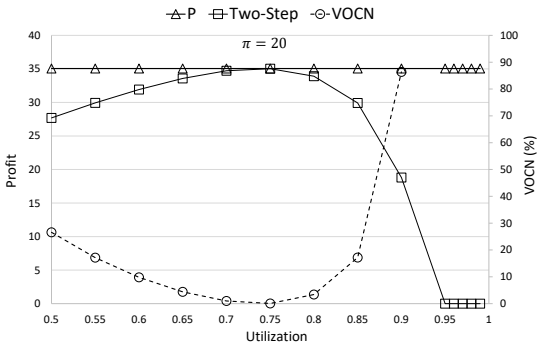

Figure 6: Value of considering the nonlinear relationship between capacity and lead time when $r = 25$ and $\pi = 20$.

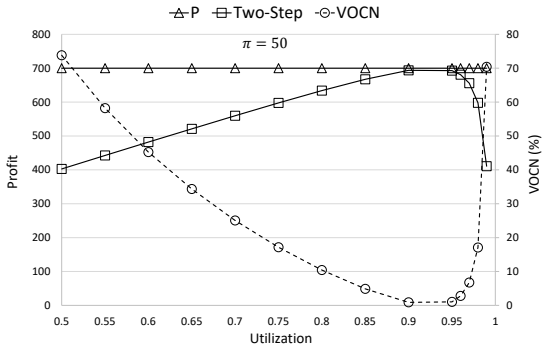

Figure 7: Value of considering the nonlinear relationship between capacity and lead time when $r = 25$ and $\pi = 50$.

In Figure 6, when $\pi = 20$, the optimal profit from problem **P** is 35.06. The profit of the Two-Step approach, which never exceeds that of problem **P**, depends on the target utilization level U chosen in the first step. It increases from 27.69 for $U = 50\%$ to 35.06 for the $U = 74\%$, the optimal utilization level in problem **P**, then decreases to 18.81 for $U = 90\%$. **VOCN** decreases from 26.6% for target utilization of 50% to 0% for target utilization of 74%, increasing to 86.38% for target utilization of 90%.When $U > 90\%$, the Two-Step approach yields negative profit so we do not calculate **VOCN**. Figure 7 shows similar results for $\pi = 50$, with **VOCN** decreasing from 73.83% for $U = 50\%$ to 0% for $U = 92.8\%$, then increasing to 70.41% for $U = 99\%$.

Figure 6 shows the danger of loading the manufacturer beyond the optimal utilization level suggested by our approach. The **VOCN** values reported here may well underestimate those that would be observed in practice, since the fixed lead time and the corresponding utilization are computed using the queueing model. In practice the lead time incurred at high utilization is often underestimated. Even this conservative approach, however, suggests that considering load-dependent lead times can be beneficial, preventing the adverse

consequences of extremely high utilization levels under unfavorable cost structures as in Figure 6. When $\pi = 50$, as in Figure 7, the unit revenue is high enough to offset the additional costs incurred due to the very high lead times at the manufacturer.

Value of integrating market selection and safety stock decisions. We now study the value of integrating market selection and safety stock decisions. We compare the profit obtained from solving the integrated problem **P** with a sequential approach in which we first make the market selection decision by optimizing the profit of accepting demand, i.e., $A\left((\pi - c)\lambda - wLT(U)\lambda - c^o(1-\alpha)\lambda - w^{whs}\lambda - w^{ret}\lambda\right)$, subject to capacity constraints, i.e., $\lambda A \leq rU$ and $U < 1$. We then input the market selection decisions as parameters into Problem **P** and obtain the optimal profit $obj^{Sequential}$ for these market selection decisions. In this way, the sequential approach does not take into account safety stock decisions and costs when selecting the markets to serve. The value of integrating market selection and safety stock decisions is given by $VOI = 100 \times \frac{Profit^P - Max(Profit^{Sequential}, 0)}{Profit^P}$. Because safety stock costs are only considered after market selection decisions, the sequential procedure may yield negative profit. Thus we only consider positive values of $obj^{Sequential}$ when calculating VOI. Figures 8 and 9 plot the profits (left y-axis), of problem **P** and the sequential approach, and VOI (right y-axis) as a function of capacity for $\pi = 20$ and $\pi = 50$, respectively.

When $\pi = 20$, as capacity increases from 1 to 35, the profits of **P** and the sequential approach increase from 0.0018 and 0 to 53.21, respectively. The profit of **P** is always greater than that of the sequential approach, giving VOI$\geq$ 0. When $r \leq 3$, the marginal profit of the sequential approach is negative, i.e., the cost of accepting demand exceeds its revenue. Hence, under the sequential approach it is optimal not to serve any market and the profit is zero, mainly due to the safety stock costs. In contrast, the integrated approach is able to serve a (small) fraction of the market and sets safety stocks to yield a positive profit. The VOI in this case 100%. As capacity increases, VOI decreases to 35.16% for $r = 5$, 12.78% for $r = 10$, 5.18% for $r = 25$, and to 0% when $r \geq 35$. Figure 9 exhibits qualitatively similar results when $\pi = 50$, but the VOI is very small. In fact, when $\pi = 50$ then VOI = 0.096% for $r = 5$, VOI = 0.068% for $r = 10$ and VOI = 0.03% for $r = 25$. Comparing Figures 8 and 9, we observe that when π is very high the VOI is very small relative to the case when marginal revenue is low. This is to be expected when the marginal revenue is sufficiently high to offset both pipeline and safety stock costs.

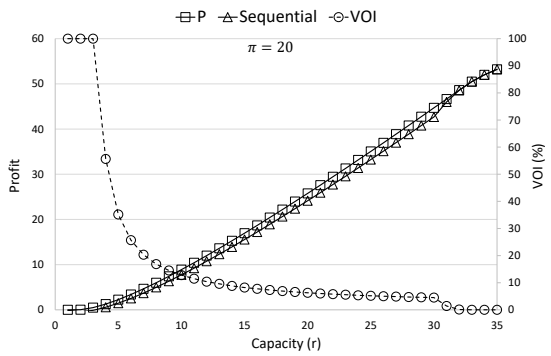

Figure 8: Value of integration as a function of capacity for $\pi = 20$.

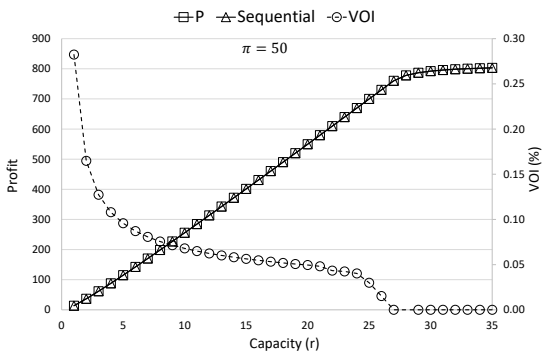

Figure 9: Value of integration as a function of capacity for $\pi = 50$.

5.2 *Comparison of solution methods*

We now compare the computational performance of the **SPLA** algorithm presented in Section 4.3 and the mixed-integer conic quadratic formulation **CQ** of Section 4.2. Both methods are coded in GAMS 31.2.0 and solved using the CPLEX solver with standard settings. Model **SSPM** is coded in GAMS and solved using BARON, a well-established non-convex MINLP solver (Tawarmalani and Sahinidis, 2005; Sahinidis, 2017), with default settings. All methods are terminated when the CPU time exceeds 3600 seconds. In each iteration of **SPLA**, we set a time limit of 1500 seconds to solve the linear approximation problem **PLA-SSPM**. At the end of each iteration, we calculate the total time used, and terminate the algorithm if this exceeds 3600 seconds. All experiments were run on a 64-bit computer with a 2.7 GHz Intel Core i5 processor and 8 GB of RAM under OSX 10.15.6.

We conduct our experiments on three sets A, B, and C of small, medium and large instances, respectively. Each set consists of 10 randomly generated instances. The number of warehouses $|\mathbb{I}|$, number of retailers $|\mathbb{J}|$ and number of markets $|\mathbb{N}|$ for each instance are given in Table 4 in Appendix F of the Electronic Supplement.

The mean demand λ_n for each market n is uniformly distributed between 10 and 100 units per period. The manufacturer's capacity level is then randomly generated as $r = unif(0.5, 1) \times \sum_{n \in \mathbb{N}} \lambda_n$. The production cost c at the manufacturer is generated randomly between 1€ and 10€ per unit per period. The work-in-process holding cost at the manufacturer is set as $w = c \times unif(0.1, 0.3)$. The lead time at the warehouses τ_i^{whs} and retailers τ_j^{ret} are generated randomly between 1 and 10 periods. The maximum external

outgoing service times for markets s_n follow a discrete uniform distribution between 1 and τ_j^{ret} for all $n \in \mathbb{N}_j$. The pipeline inventory costs between the manufacturer and warehouses are generated based on $w_i^{whs} = w\tau_i^{whs} \times unif(1.05, 1.10)$. The inventory holding cost at warehouses are generated as $h_i^{whs} = \frac{w_i^{whs}}{\tau_i^{whs}} \times unif(1.05, 1.10)$, and the pipeline inventory cost between warehouses and retailers as $w_j^{ret} = h_i^{whs}\tau_j^{ret} \times unif(1.1, 1.2)$. The inventory holding cost at retailers are generated randomly as $h_j^{ret} = \frac{w_j^{ret}}{\tau_j^{ret}} \times unif(1.1, 1.2)$. The safety factor for all warehouses and retailers is set as $z_i^{whs} = z_j^{ret} = 1.96$, corresponding to a 97.5% service level. Based on Aouam and Kumar (2019), the unit expediting cost at warehouses is set as $h_i^{exp,\ whs} = \frac{h_i^{whs}}{1-\Phi(z_i^{whs})} - (h_i^{whs} - w)$ and at the retailers as $h_j^{exp,\ ret} = \frac{h_j^{ret}}{1-\Phi(z_j^{ret})}$. The overtime cost at the manufacturer is set as $c^o = \frac{\sum_{i\in\mathbb{I}} h_i^{exp,\ whs}}{|\mathbb{I}|}$. Following Ghadimi and Aouam (2021), the uncertainty cost at the warehouse is set as $H^{whs}(z_i) = h_i^{whs} z_i^{whs} + (h_i^{exp,\ whs} + h_i^{whs} - w)G(z_i^{whs})$ and at the retailer as $H^{ret}(z_j) = h_i^{ret} j z_j^{ret} + h_j^{exp,\ ret} G(z_j^{ret})$, where $G(z) = \phi(z) - z(1-\Phi(z))$. $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal probability density function and cumulative distribution function, respectively.

The time an item spends in the workstation (queuing plus processing) is the production cycle time of the manufacturer, PCT which is a random variable. For the $M/G/1$ queue, the mean and variance of the PCT are given by the Pollaczek-Khintchine formulas (Allen, 2014) as $\mathrm{E}(PCT) = \frac{1}{r}\left[\left(\frac{1+C_s^2}{2}\right)\left(\frac{u_q}{1-u_q}\right) + 1\right]$ and $\mathrm{Var}(PCT) = \frac{\lambda}{3(1-u_q)}\Gamma + \frac{1}{2}\left(\frac{1+C_s^2}{r}\right)^2\left(\frac{u_q}{1-u_q}\right)^2 + \frac{1+C_s^2}{r^2(1-u_q)} - \mathrm{E}(PCT)^2$. Where C_s is the coefficient of variation of the processing times and Γ the third moment about zero of the processing time distribution. The fixed portion t_2 of the manufacturer's lead time is defined such that $\mathrm{P}(PCT \leq t_2) \geq \alpha$. The probability distribution of PCT can be approximated by a hyper-exponential or gamma distribution. When α =95%, Martin (1972) estimates the production lead times as $t_2(u_q) = \mathrm{E}(PCT) + 2\sqrt{\mathrm{Var}(PCT)}$, which we use to set the value of t_2. We assume that processing time at the manufacturer follows a lognormal distribution with mean $\frac{1}{r}$, and the coefficient of variation C_s of the service time is generated uniformly between 0.5 and 1.5. On average, orders spend 0.5 periods in the raw material staging area, i.e., $t_1 = 0.5$. We then discretize these to obtain integer values $l_q = t_1 + t_2$ and their corresponding utilization levels u_q.

The marginal revenue of accepting demand from market n is generated as $\pi_n = \Big(c + wl_q + c^o(1-\alpha) + \overline{w}^{whs} + \overline{w}^{ret} + \overline{H}^{whs}(z)\sqrt{l_q + \overline{\tau}^{whs}} + \overline{H}^{ret}(z)\sqrt{\overline{\tau}^{ret} - \overline{s}}\Big) \times unif(1.1, 1.5)$, where l_q is the lead time values related to a utilization level of 80%, and $\overline{w}^{whs} = \frac{\sum_{i\in\mathbb{I}} w_i^{whs}}{|\mathbb{I}|}$,

$\overline{H}^{whs}(z) = \frac{\sum_{i\in\mathbb{I}} H^{whs}(z_i)}{|\mathbb{I}|}$, $\overline{\tau}^{whs} = \frac{\sum_{i\in\mathbb{I}} \tau_i^{whs}}{|\mathbb{I}|}$, $\overline{w}^{ret} = \frac{\sum_{j\in\mathbb{J}} w_j^{ret}}{|\mathbb{J}|}$, $\overline{H}^{ret}(z) = \frac{\sum_{j\in\mathbb{J}} H^{ret}(z_j)}{|\mathbb{J}|}$, $\overline{\tau}^{ret} = \frac{\sum_{j\in\mathbb{J}} \tau_j^{ret}}{|\mathbb{J}|}$, and $\overline{s} = \frac{\sum_{n\in\mathbb{N}} s_n}{|\mathbb{N}|}$.

We compare the solution procedures using three performance measures: 1) *Optimality gap* (Gap%) which is the relative difference between the best upper bound (GUB) and the best lower bound (GLB), i.e. Gap% $= \frac{GUB-GLB}{GLB} \times 100$. 2) *Relative deviation* Dev$_I$% which shows the relative deviation between the best lower bound of **SPLA** and the best lower bound obtained by BARON, i.e. Dev$_I$%$= \frac{GLB_{\mathbf{SPLA}}-GLB_{\mathbf{BARON}}}{GLB_{\mathbf{BARON}}} \times 100$. 3) *Relative deviation* Dev$_{II}$% which is the relative deviation between the best lower bound of **CQ** and the best lower bound obtained by BARON, i.e. Dev$_{II}$% $= \frac{GLB_{\mathbf{CQ}}-GLB_{\mathbf{BARON}}}{GLB_{\mathbf{BARON}}} \times 100$. Positive values of Dev$_I$% and Dev$_{II}$% indicate that **SPLA** and **CQ** yield higher profit than BARON.

Tables 2-4 report the lower and upper bounds, optimality gap Gap%, CPU time, Dev$_I$% and Dev$_{II}$% for all instances in sets A, B, and C. In Table 2, BARON, **SPLA**, and **CQ** find optimal solutions for all the small instances in set A in an average CPU time of 87, 96, and 130 seconds, respectively. As all methods find an optimal solution Dev$_I$=Dev$_{II}$ =0%.

Table 2: Performance of BARON, **SPLA** and **CQ** for Instance Set A

Ins. #	BARON				SPLA					CQ				
	LB	UB	Gap%	Time (s)	LB	UB	Gap%	Time (s)	Dev$_I$%	LB	UB	Gap%	Time (s)	Dev$_{II}$%
1	19866	19866	0.00	224	19866	19866	0.00	152	0.00	19866	19866	0.00	149	0.00
2	8681	8681	0.00	177	8681	8681	0.00	70	0.00	8681	8681	0.00	105	0.00
3	43649	43649	0.00	9	43649	43650	0.00	74	0.00	43649	43649	0.00	102	0.00
4	10313	10313	0.00	66	10313	10313	0.00	66	0.00	10313	10313	0.00	190	0.00
5	7279	7279	0.00	81	7279	7279	0.00	37	0.00	7279	7279	0.00	126	0.00
6	33830	33830	0.00	62	33830	33830	0.00	31	0.00	33830	33830	0.00	58	0.00
7	99537	99537	0.00	57	99537	99538	0.00	97	0.00	99537	99537	0.00	136	0.00
8	37505	37505	0.00	36	37505	37505	0.00	86	0.00	37505	37505	0.00	87	0.00
9	74066	74066	0.00	117	74066	74067	0.00	255	0.00	74066	74066	0.00	270	0.00
10	6892	6892	0.00	37	6892	6892	0.00	90	0.00	6892	6892	0.00	73	0.00
	Average		**0.00**	**87**			**0.00**	**96**	**0.00**			**0.00**	**130**	**0.00**

Table 3 shows that BARON obtains feasible solutions for all medium-sized instances in set B within the one hour time limit with an average gap of 38.95%. **SPLA** finds the optimal solution for all instances in this set in an average of 1129 seconds, while **CQ** obtains optimal solutions for all instances except instance 1 (whose optimality gap is 0.07%) in an average CPU time of 2063 seconds. The average relative deviation for **SPLA** and **CQ** are Dev$_I$=0.29% and Dev$_{II}$=0.29%, meaning that both **SPLA** and **CQ** obtain higher profit than BARON. The small gaps suggest that BARON is actually close to an optimal solution,

but unable to confirm it in the available time.

Table 3: Performance of BARON, **SPLA** and **CQ** for Instance Set B

# Ins.	BARON				SPLA					CQ				
	LB	UB	Gap%	Time (s)	LB	UB	Gap%	Time (s)	Dev_I%	LB	UB	Gap%	Time (s)	Dev_{II}%
1	132444	182374	37.70	3600	132974	132977	0.00	1388	0.40	132974	133067	0.07	3600	0.40
2	60791	84430	38.89	3600	61109	61110	0.00	1306	0.52	61109	61109	0.00	1929	0.52
3	43518	59476	36.67	3600	43543	43543	0.00	1664	0.06	43543	43543	0.00	2482	0.06
4	152086	229769	51.08	3600	152224	152226	0.00	1299	0.09	152224	152224	0.00	1635	0.09
5	18653	23971	28.51	3600	18752	18752	0.00	1192	0.53	18752	18752	0.00	1340	0.53
6	53737	76213	41.83	3600	53916	53917	0.00	757	0.33	53916	53916	0.00	3100	0.33
7	53428	69807	30.66	3600	53625	53626	0.00	1196	0.37	53625	53625	0.00	1295	0.37
8	171308	237163	38.44	3600	171340	171342	0.00	888	0.02	171340	171341	0.00	1857	0.02
9	93885	132354	40.97	3600	94309	94310	0.00	1032	0.45	94309	94309	0.00	2240	0.45
10	46190	66842	44.71	3600	46258	46258	0.00	567	0.15	46258	46258	0.00	1153	0.15
	Average		**38.95**	**3600**			**0.00**	**1129**	**0.29**			**0.01**	**2063**	**0.29**

Table 4 shows that BARON obtains feasible solutions for all the large instances in set C within the one hour time limit with an average Gap of 50.33%. **SPLA** finds optimal or near optimal solutions for all instances in set C with an average Gap of 0.01% within an average CPU time of 3522 seconds. **CQ** is able to provide feasible solutions for all instances in set C with an average Gap of 11.32% within the one hour time limit. **SPLA** obtains solutions with higher profit than BARON and **CQ** with Dev_I=0.48%. **CQ** is able to provide better solutions than BARON only for instances 1 and 2 with Dev_I of 0.04% and 0.24%, respectively. The average relative gap for **CQ** is Dev_I=-1.14%, implying that on average **CQ** cannot provide better solutions than BARON. As the size of the instances increases the performance of **CQ** deteriorates.

Table 4: Performance of BARON, **SPLA** and **CQ** for set C

# Ins.	BARON				SPLA					CQ				
	LB	UB	Gap%	Time (s)	LB	UB	Gap%	Time (s)	Dev_I%	LB	UB	Gap%	Time (s)	Dev_{II}%
1	316886	454494	43.42	3600	318657	318679	0.01	2886	0.56	317017	324229	2.28	3600	0.04
2	253898	371804	46.44	3600	255050	255052	0.00	3456	0.45	254497	258668	1.64	3600	0.24
3	527948	722859	36.92	3600	529518	529526	0.00	3465	0.30	526323	541040	2.80	3600	-0.31
4	84449	125056	48.08	3600	84776	84776	0.00	3235	0.39	83686	88512	5.77	3600	-0.90
5	387902	559365	44.20	3600	388554	388575	0.01	2015	0.17	384075	412247	7.33	3600	-0.99
6	203813	318153	56.10	3600	205047	205171	0.06	4809	0.61	198191	220863	11.44	3600	-2.76
7	99639	162862	63.45	3600	100120	100127	0.01	3477	0.48	97328	111209	14.26	3600	-2.32
8	117121	178946	52.79	3600	117460	117468	0.01	3549	0.29	115592	123081	6.48	3600	-1.31
9	97948	152823	56.02	3600	98738	98748	0.01	4436	0.81	97816	145608	48.86	3600	-0.13
10	117286	182862	55.91	3600	118174	118175	0.00	3887	0.76	113838	127937	12.39	3600	-2.94
	Average		**50.33**	**3600**			**0.01**	**3522**	**0.48**			**11.32**	**3600**	**-1.14**

Overall, **CQ** provides better upper bounds, BARON better feasible solutions, and **SPLA**

better feasible solutions and better upper bounds. The performance of BARON and **CQ** deteriorates as the instance size increases, while that of **SPLA** remains consistent.

Finally Table 5 shows the optimality gap Gap% and CPU time for all instances in sets A, B, and C for the **QC** problem with and without the valid inequalities developed in Section 4.2. For instance sets A and B, both $\mathbf{QC}^0$ and **QC** are able to find optimal solutions significantly more rapidly when the valid inequalities are incorporated. In set C, the inclusion of the valid inequalities results in smaller optimality gaps at the end of the one hour time limit, although none of the gaps is reduced below 2.28%. Hence adding valid inequalities improves the performance of the MICQ formulation **QC**.

Table 5: Effect of valid inequalities on the performance of **CQ**

Ins. #	Set A				Set B				Set C			
	$\mathbf{QC}^0$		**QC**		$\mathbf{QC}^0$		**QC**		$\mathbf{QC}^0$		**QC**	
	Gap%	Time (s)	Gap%	Time (s)	Gap%	Time (s)	Gap%	Time (s)	Gap%	Time (s)	Gap%	Time (s)
1	0.00	441	0.00	149	27.35	3600	0.07	3600	55.42	3600	2.28	3600
2	0.00	127	0.00	105	34.81	3600	0.00	1929	60.60	3600	1.64	3600
3	0.00	170	0.00	102	21.29	3600	0.00	2482	53.60	3600	2.80	3600
4	0.00	411	0.00	190	12.74	3600	0.00	1635	56.75	3600	5.77	3600
5	0.00	146	0.00	126	29.16	3600	0.00	1340	46.89	3600	7.33	3600
6	0.00	98	0.00	58	28.45	3600	0.00	3100	51.61	3600	11.44	3600
7	0.00	337	0.00	136	23.42	3600	0.00	1295	56.63	3600	14.26	3600
8	0.00	175	0.00	87	29.25	3600	0.00	1857	48.80	3600	6.48	3600
9	0.00	1379	0.00	270	29.32	3600	0.00	2240	53.05	3600	48.86	3600
10	0.00	76	0.00	73	28.53	3600	0.00	1153	49.40	3600	12.39	3600
Average	**0.00**	**336**	**0.00**	**130**	**26.43**	**3600**	**0.01**	**2063**	**53.28**	**3600**	**11.32**	**3600**

6 Conclusion

This paper addresses the safety stock placement problem with market selection decisions in a production-distribution system with load-dependent lead time at the manufacturer. We formulate the problem as a mixed-integer nonlinear programming (MINLP) model with a non-convex objective function. After reformulating the problem to eliminate some integer variables and all bilinear terms, we propose a successive piecewise linearization algorithm and a mixed-integer conic quadratic formulation. Computational experiments show that the successive piecewise linearization algorithm outperforms BARON, a state-of-art solver, and the mixed-integer conic quadratic formulation.

Sequential approaches that first determine the amount of demand to be served and then

determine safety stock levels based on this decision perform well only under ample capacity, which allows short manufacturing lead times. When the available demand exceeds the available capacity, accepting demand levels close to the system capacity results in high lead times that induce high pipeline inventory and safety stock costs that can more than offset any additional revenue. Sequential approaches that first determine the demand to be served and then compute safety stocks fail to consider the impact of producing safety stocks on lead times; those that assume a fixed lead time implicitly assume that the system can always maintain the target lead time with no additional costs. Ifinventory holding costs are low relative to the marginal revenue and customers are willing to tolerate long lead times and occasional delivery delays, the additional complexity of models such as those proposed is unlikely to be justified. However, for capital-intensive industries where additional capacity is expensive, lead times to bring it online are long, inventory holding costs are substantial or long lead times lead to reduced demand, integrated consideration of queueing behavior, safety stocks and market selection takes on considerable importance.

The approach proposed in this paper can be extended to the case where each market involves a set of different products, and to include multiple manufacturers. Modifications to include finite inventory budgets together with production decisions, such as lot sizing, production smoothing, or lot scheduling are also interesting to explore in terms of their interaction with safety stock placements and market selection. The present study assumes that logistics nodes follow a base stock policy with a common review period. Considering different review periods at logistics nodes may further decrease supply chain costs and take into account setup or ordering costs. The emerging optimization problems would be interesting to study but would present computational challenges. Developing efficient solution procedures for these new problems will also be an interesting future research direction.

Acknowledgements

We acknowledge the support provided by the Research Foundation - Flanders (FWO) for the project with contract number FWO.OPR.2016.0019.01.

References

Albey, E., A. Norouzi, K. G. Kempf, and R. Uzsoy (2015). Demand modeling with forecast evolution: An application to production planning. *IEEE Transactions on Semiconductor Manufacturing 28*, 374–384.

Allen, A. O. (2014). *Probability, Statistics, and Queueing Theory*. Academic Press.

Aouam, T. and N. Brahimi (2013). Integrated production planning and order acceptance under uncertainty: A robust optimization approach. *European Journal of Operational Research 228*(3), 504–515.

Aouam, T., K. Geryl, K. Kumar, and N. Brahimi (2018). Production planning with order acceptance and demand uncertainty. *Computers & Operations Research 91*, 145–159.

Aouam, T., F. Ghadimi, and M. Vanhoucke (2021). Finite inventory budgets in production capacity and safety stock placement under the guaranteed service approach. *Computers & Operations Research 131*, 105266.

Aouam, T. and K. Kumar (2019). On the effect of overtime and subcontracting on supply chain safety stocks. *Omega 89*, 1–20.

Aouam, T. and R. Uzsoy (2012). Chance-constraint-based heuristics for production planning in the face of stochastic demand and workload-dependent lead times. In *Decision Policies for Production Networks*, pp. 173–208. Springer.

Aouam, T. and R. Uzsoy (2015). Zero-order production planning models with stochastic demand and workload-dependent lead times. *International Journal of Production Research 53*, 1–19.

Atamtürk, A., G. Berenguer, and Z.-J. Shen (2012). A conic integer programming approach to stochastic joint location-inventory problems. *Operations Research 60*(2), 366–381.

Axsäter, S. (2015). *Inventory control*. Springer.

Bakal, I. S., J. Geunes, and H. E. Romeijn (2008). Market selection decisions for inventory models with price-sensitive demand. *Journal of Global Optimization 41*(4), 633–657.

Brahimi, N., T. Aouam, and E.-H. Aghezzaf (2015). Integrating order acceptance decisions with flexible due dates in a production planning model with load-dependent lead times. *International Journal of Production Research 53*(12), 3810–3822.

Buzacott, J. A. and J. G. Shanthikumar (1993). *Stochastic models of manufacturing systems*. Prentice Hall Englewood Cliffs, NJ.

Curry, G. L. and R. M. Feldman (2000). *Manufacturing Systems Modelling and Analysis*. Springer Science & Business Media.

Daskin, M. S., C. R. Coullard, and Z.-J. M. Shen (2002). An inventory-location model: Formulation, solution algorithm and computational results. *Annals of Operations Research 110*(1-4), 83–106.

Eruguz, A. S., Z. Jemai, E. Sahin, and Y. Dallery (2014). Optimising reorder intervals and order-up-to levels in guaranteed service supply chains. *International Journal of Production Research 52*(1), 149–164.

Eruguz, A. S., E. Sahin, Z. Jemai, and Y. Dallery (2016). A comprehensive survey of guaranteed-service models for multi-echelon inventory optimization. *International Journal of Production Economics 172*, 110–125.

Farahani, R. Z., H. R. Bajgan, B. Fahimnia, and M. Kaviani (2015). Location-inventory problem in supply chains: a modelling review. *International Journal of Production Research 53*(12), 3769–3788.

Fathi, M., M. Khakifirooz, A. Diabat, and H. Chen (2021). An integrated queuing-stochastic optimization hybrid genetic algorithm for a location-inventory supply chain network. *International Journal of Production Economics 237*, 108139.

Geunes, J., R. Levi, H. E. Romeijn, and D. B. Shmoys (2011). Approximation algorithms for supply chain planning and logistics problems with market choice. *Mathematical programming 130*(1), 85–106.

Geunes, J., Y. Merzifonluoğlu, and H. E. Romeijn (2009). Capacitated procurement planning with price-sensitive demand and general concave-revenue functions. *European Journal of Operational Research 194*(2), 390–405.

Geunes, J., Y. Merzifonluoğlu, H. E. Romeijn, and K. Taaffe (2005). Demand selection and assignment problems in supply chain planning. In *Emerging Theory, Methods, and Applications*, pp. 124–141. INFORMS.

Geunes, J., Z.-J. Shen, and H. E. Romeijn (2004). Economic ordering decisions with market choice flexibility. *Naval Research Logistics (NRL) 51*(1), 117–136.

Ghadimi, F. and T. Aouam (2021). Planning capacity and safety stocks in a serial production–distribution system with multiple products. *European Journal of Operational Research 289*(2), 533–552.

Ghadimi, F., T. Aouam, S. Haeussler, and R. Uzsoy (2022). Centralized and decentralized systems for coordinating order acceptance and release planning. *European Journal of Operational Research*.

Ghadimi, F., T. Aouam, and M. Vanhoucke (2020). Optimizing production capacity and safety stocks in general acyclic supply chains. *Computers & Operations Research 120*, 104938.

Graves, S. C. (1986). A tactical planning model for a job shop. *Operations Research 34*(4), 522–533.

Graves, S. C. and S. P. Willems (2000). Optimizing strategic safety stock placement in supply chains. *Manufacturing & Service Operations Management 2*(1), 68–83.

Heath, D. C. and P. L. Jackson (1994). Modeling the evolution of demand forecasts with application to safety stock analysis in production/distribution systems. *IIE Transactions 26*, 17–30.

Hopp, W. J. and M. L. Spearman (2011). *Factory physics*. Waveland Press.

Inderfurth, K. (1991). Safety stock optimization in multi-stage inventory systems. *International Journal of Production Economics 24*(1-2), 103–113.

Inderfurth, K. and S. Minner (1998). Safety stocks in multi-stage inventory systems under different service measures. *European Journal of Operational Research 106*(1), 57–73.

Jalali, H., R. Carmen, I. Van Nieuwenhuyse, and R. Boute (2019). Quality and pricing decisions in production/inventory systems. *European Journal of Operational Research 272*(1), 195–206.

Klosterhalfen, S. T., D. Dittmar, and S. Minner (2013). An integrated guaranteed-and stochastic-service approach to inventory optimization in supply chains. *European Journal of Operational Research 231*(1), 109–119.

Kumar, K. and T. Aouam (2018a). Effect of setup time reduction on supply chain safety stocks. *Journal of Manufacturing Systems 49*, 1–15.

Kumar, K. and T. Aouam (2018b). Integrated lot sizing and safety stock placement in a network of production facilities. *International Journal of Production Economics 195*, 74–95.

Kumar, K. and T. Aouam (2019). Extending the strategic safety stock placement model to consider tactical production smoothing. *European Journal of Operational Research 279*(2), 429–448.

Lee, K. and L. Ozsen (2020). Tabu search heuristic for the network design model with lead time and safety stock considerations. *Computers & Industrial Engineering 148*, 106717.

Levi, R., J. Geunes, H. E. Romeijn, and D. B. Shmoys (2005). Inventory and facility location models with market selection. In *International Conference on Integer Programming and Combinatorial Optimization*, pp. 111–124. Springer.

Magnanti, T. L., Z.-J. M. Shen, J. Shu, D. Simchi-Levi, and C.-P. Teo (2006). Inventory placement in acyclic supply chain networks. *Operations Research Letters 34*(2), 228–238.

Martin, J. (1972). *Systems analysis for data transmission*. Prentice Hall, Inc., Englewood Cliffs, New Jersey.

Minner, S. (2000). *Strategic Safety Stocks in Supply Chains*, Volume 490. Springer Science & Business Media.

Missbauer, H. and R. Uzsoy (2020). *Production Planning with Capacitated Resources and Congestion*. Springer.

Ozsen, L., C. R. Coullard, and M. S. Daskin (2008). Capacitated warehouse location model with risk pooling. *Naval Research Logistics (NRL) 55*(4), 295–312.

Puga, M. S., S. Minner, and J.-S. Tancrez (2019). Two-stage supply chain design with safety stock placement decisions. *International Journal of Production Economics 209*, 183–193.

Sahinidis, N. V. (2017). BARON 17.8.9: Global Optimization of Mixed-Integer Nonlinear Programs, *User's Manual*.

Shahabi, M., S. Akbarinasaji, A. Unnikrishnan, and R. James (2013). Integrated inventory control and facility location decisions in a multi-echelon supply chain network with hubs. *Networks and Spatial Economics 13*(4), 497–514.

Shahabi, M., A. Unnikrishnan, E. Jafari-Shirazi, and S. D. Boyles (2014). A three level location-inventory problem with correlated demand. *Transportation Research Part B: Methodological 69*, 1–18.

Shen, Z.-J. M., C. Coullard, and M. S. Daskin (2003). A joint location-inventory model. *Transportation Science 37*(1), 40–55.

Shu, J., Z. Li, and L. Huang (2013). Demand selection decisions for a multi-echelon inventory distribution system. *Journal of the Operational Research Society 64*(9), 1307–1313.

Shu, J., Z. Li, and W. Zhong (2011). A market selection and inventory ordering problem under demand uncertainty. *Journal of Industrial & Management Optimization 7*(2), 425.

Simpson Jr, K. F. (1958). In-process inventories. *Operations Research 6*(6), 863–873.

Sourirajan, K., L. Ozsen, and R. Uzsoy (2007). A single-product network design model with lead time and safety stock considerations. *IIE Transactions 39*(5), 411–424.

Sourirajan, K., L. Ozsen, and R. Uzsoy (2009). A genetic algorithm for a single product network design model with lead time and safety stock considerations. *European Journal of Operational Research 197*(2), 599–608.

Taaffe, K., J. Geunes, and H. E. Romeijn (2008). Target market selection and marketing effort under uncertainty: The selective newsvendor. *European Journal of Operational Research 189*(3), 987–1003.

Tawarmalani, M. and N. V. Sahinidis (2005). A polyhedral branch-and-cut approach to global optimization. *Mathematical Programming 103*(2), 225–249.

Van den Heuvel, W., O. E. Kundakcioglu, J. Geunes, H. E. Romeijn, T. C. Sharkey, and A. P. Wagelmans (2012). Integrated market selection and production planning: complexity and solution approaches. *Mathematical programming 134*(2), 395–424.

You, F. and I. E. Grossmann (2010). Integrated multi-echelon supply chain design with inventories under uncertainty: MINLP models, computational strategies. *AIChE Journal 56*(2), 419–440.

Ziarnetzky, T., L. M onch, and R. Uzsoy (2018). Rolling horizon production planning for wafer fabs with chance constraints and forecast evolution. *International Journal of Production Research 56*, 6112–6134.

Ziarnetzky, T., L. M onch, and R. Uzsoy (2020). Simulation-based performance assessment of production planning formulations with safety stock and forecast evolution using a large-scale wafer fab model. *IEEE Transactions on Semiconductor Manufacturing 33*, 1–12.