EFFICIENT MULTI-ECHELON INVENTORY OPTIMIZATION MODELS FOR INDUSTRIAL PRACTICE

V. G. Achkar[1,2*], B. B. Brunaud[3] and I. E. Grossmann[1]
[1]Carnegie Mellon University, [2]Universidad Nacional del Litoral, [3]Johnson & Johnson
[1]Pittsburgh, PA 15213, [2]Santa Fe, Argentina, [3]New Jersey, NJ 08901

Abstract

In this work, we propose a generalized Multi-Echelon Inventory Optimization (MEIO) model based on the Guaranteed-Service approach for allocating safety stocks across the network, seeking to meet customer service levels at minimum holding costs. This is especially challenging in multi-echelon supply chains, which face supply and demand uncertainty, and the decisions on one echelon impact the other echelons. The proposed model offers a more realistic representation of real-world supply chains, accounting for several features such as non-normal demand distributions, specified time periods between reviews, Minimum Order Quantities (MOQ), and different service level performance indicators (fill rate and cycle service levels). To improve the computational efficiency and make the model effective to use in industrial practice, we propose to reformulate the nonlinear programming model as a Mixed-Integer Quadratically Constrained Program (MIQCP) reformulation, a stepwise function and a piecewise linear approximation. This solution strategy leads to order of magnitude reductions in computational time. The performance of the model is evaluated by solving several instances of a real-world case study.

Keywords

Multi-echelon inventory optimization, safety stock, non-normal demand, QCP, piecewise approximation.

Introduction

Multi-echelon inventory optimization (MEIO) seeks to optimize safety stock setting for the entire supply chain with centralized coordination (You & Grossmann, 2009). MEIO eliminates over-buffering of inventory, and builds confidence in product availability. De Kok et al. (2019) present a general typology and review of stochastic MEIO models in which they summarize the extensive research on multi-echelon inventory management. The authors state that multi-echelon inventory systems are still a very active area of research because of their complexity and practical relevance. Recently, Gonçalves et al. (2020) present a systematic literature review, and they also highlight that the number of contributions to MEIO has seen a significant increase from the year 2005 onwards, and they list many potential directions and trends for future research.

The present paper relies on the Guaranteed-Service Model (GSM) to optimize safety stocks (Graves & Willems, 2000; Simpson, 1958). The objective is to develop an MEIO model to address the problem of safety stocks in real-world complex systems. Managing a real-world supply chain gives rise to two important challenges: (i) the model must represent industrial practice features and dynamics, and (ii) the optimal solution must be found with modest computational time. Over the years, many authors have worked on extending the original GSM assumptions to enable real-world supply chain characteristics to be captured, as presented in the survey by Eruguz et al. (2016). However, the GSM for multi-echelon supply chains proposed in the literature does not fully account for all the issues and characteristics arising in industrial practice. We

[*] vachkar@fiq.unl.edu.ar

extend the work of Achkar et al. (2022) to account for a novel approach to include stochastic lead times, and we specially focus on new solution strategies. In this work, the following characteristics are included: (i) non-normally distributed demands, (ii) manufacturing plants placed at any location in the network, (iii) fill rate as an alternative key customer service performance indicator apart from the Cycle Service Level (CSL), and (iv) Minimum Order Quantities (MOQ) for replenishment orders and the period between reviews. Furthermore, this work presents solution strategies to increase computational efficiency. From an optimization perspective, inventory decisions in multi-echelon systems are challenging because the problems are nonlinear and nonconvex. We introduce a solution approach that allows solving real-world size problems in modest computational time. The NLP model is reformulated as an MIQCP by exploiting the structure of the constraints of the base model. In addition, the problem involves piecewise and stepwise functions, reducing additional complexities. Computational examples for a real industrial system are presented to illustrate the application of the proposed model and its resulting improved computational performance.

The outline of the paper is as follows. The problem statement is described first, followed by the model formulation. Next, an extension to account for nonnormal distributions for the demand is described. The paper ends with the application of the model to illustrate its application to real-world case studies. Conclusions are drawn in the final section.

Problem Statement

Given is a supply chain with a fixed design for locations $j \in J$ and a set of materials $p \in P$ that can be either raw materials or finished goods. Holding costs are incurred at all the nodes. If the customer places an order of size $d_{jp}(t)$ on node j at time t, this order will be fulfilled by time t + S_{jp}, with S_{jp} being the guaranteed-service time of node j (Graves & Willems, 2000). Moreover, each node j receives a service commitment from its upstream node $i \in J$, called inbound service time SI_{jp}, and has an order processing time or lead time of LT_j. The Net Lead Time (NLT), as shown in Figure 1, represents the time period that is not covered within the guaranteed service time, and must be covered with safety stock, defined as $NLT_{jp} = SI_{jp} + LT_{jp}$ - S_{jp}. If $NLT_{jp} = 0$, the node works under a make-to-order policy without storing any inventory. On the other hand, if $NLT_{jp} > 0$, there is a time period that should be covered with safety stock. The objective is to determine the guaranteed-service times (SI_{jp} and S_{jp}) for each material at each location, and consequently how much safety stock to maintain to minimize the total holding costs and satisfy a specified customer service level. The service times at the initial and the final nodes are given.

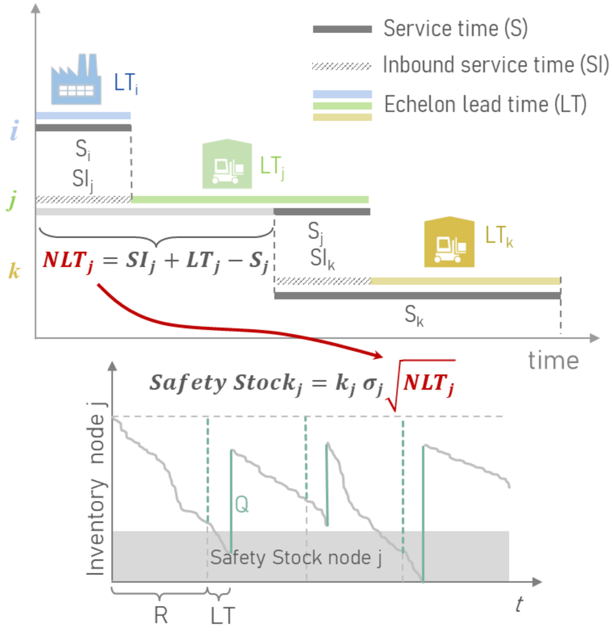

Figure 1. Guaranteed-service model elements

Demand is assumed to be bounded. If the demand in a certain time period exceeds the bound, it is assumed that other extraordinary measures are used to satisfy the excess demand. In addition, each stage of the supply chain operates under a periodic review inventory policy with a basestock level. The demand is independent and identically distributed at each demand node. The statistical distributions of the demand are not necessarily normal. Similarly, lead times are variable but assumed to be normally distributed.

Demand is propagated upstream considering the risk pooling assumptions described in You and Grossmann (2009). At plant locations, there is a coefficient ϕ_{pq} that represents the bill of materials for product transformation and depends on material-finished good relationships. We assume divergent networks: a node that holds a material p can only receive this material from a single node and can distribute it to one or more locations, as is usual in finished goods supply chains. The same node can be supplied with another material $q \in P$ from another location, but this last one should be the only supplier of q for that node. Furthermore, a Minimum Order Quantity moq_{jp} may be required for the replenishment orders. Each node makes its own replenishment decisions and has no delay in ordering. For each node, there is a safety stock factor k_{jp} related to the Cycle Service Level (CSL), which reflects the percentage of replenishment cycles with non-stocking out. Alternatively, the modeler can also ask for a fill rate to be considered as a target service measure, which equals the proportion of the demand met over the total demand placed. This option is assumed to be active only if demand is normally distributed.

Model Formulation

The following formulation involves four positive continuous variables: the guaranteed-service time S_{jp} of a product on a given node, the inbound service SI_{jp}, the safety factor $\widehat{K}_{jp}$ related to CSL, and the Net Lead Time NLT_{jp}. The objective function is to minimize safety stock holding cost as defined by Eq. (1), where h_{jp} is the coefficient that represents holding cost for each material p at each location j, and the square root represents the "sigma-combination" formula, known from the stochastic-service approach (Clark & Scarf, 1960). We introduce stochastic lead times into the GSM by extending the approach proposed by Minner (1998) because it faces all the variability sources using safety stock. The mean demand and the standard deviation are represented by μ_{jp} and σ_{jp}, respectively. We introduce a new parameter σ_{NLTjp} that refers to the NLT variability instead of using the lead time variability.

$$min \sum_{p} \sum_{j \in B_{jp}} h_{jp}\, \widehat{K}_{jp} \sqrt{NLT_{jp}\, \sigma_{jp}^{2} + \mu_{jp}^{2}\, \sigma_{NLT_{jp}}^{2}} \tag{1}$$

The first set of constraints comes from the classic GSM. Equation (2) defines the first inbound service time for the starting nodes in the network J^0, where si^0 is a given input. Equation (3) links the inbound guaranteed-service time SI_{jp} and the guaranteed-service time of upstream nodes S_{iq}. A is a set with elements (i,j,p) indicating that there is an enabled route for material p from i to j. Note that $q = p$ and $i \neq j$ if it is a distribution link (i to j) of the same product p, and $q \neq p$ and $i = j$ if node j is a plant location that produces p from q. The service time S_{jp} is bounded by Eq. (4), and Eq. (5) becomes active if there is a maximum accepted service time. The NLT is defined in Eq. (6), with lt_{jp} being the lead time and r_{jp} the time period between reviews.

$$SI_{jp} = si_{jp}^{0} \qquad \forall\, j\ \in J^{0}, p \in P_{j} \tag{2}$$

$$SI_{jp} \geq S_{iq} \qquad \forall\, (i,j,p) \in A, (q,p) \in \Phi, p \in P_{j} \tag{3}$$

$$S_{jp} \leq SI_{jp} + lt_{jp} + r_{jp} \qquad \forall\, j\ \in J, p \in P_{j} \tag{4}$$

$$S_{jp} \leq maxS_{jp} \qquad \forall\, j\ \in J, p \in P_{j} \tag{5}$$

$$NLT_{jp} \geq SI_{jp} - S_{jp} + lt_{jp} + r_{jp} \qquad \forall\, j\ \in J, p \in P_{j} \tag{6}$$

As mentioned above, the GSM uses the CSL as the customer service performance indicator when setting safety stocks. As mentioned before, the fill rate is another measure that represents the fraction of demand that was met on-time from the inventory. We extend the GSM to allow specifying fill rates if desired ($j,p \in F$). Chopra and Meindl (2013) propose a formula to obtain the corresponding fill rates given a safety stock level for continuous review policies. From this formula, we can derive Eq. (7) that links the fill rate (fr_{jp}) to the safety factor k_{jp}, and consequently to the CSL. The safety factor, in this case, becomes a positive continuous variable, and the aim is to find the lower required CSL that is able to meet the desired fill rate. We assume that Q_{jp} is the average replenishment quantity of product p on location j, with Q_{jp} = max{$\mu_{jp}\, r_{jp}$, moq_{jp}}for periodic-review policies. $F_s(k_{jp})$ and $f_s(k_{jp})$ correspond to the cumulative and normal density distribution functions, respectively.

$$fr_{jp} \leq 1 + \frac{\sqrt{NLT_{jp}\, \sigma_{jp}^{2} + \mu_{jp}^{2}\, \sigma_{NLT_{jp}}^{2}}}{Q_{jp}} \left(\widehat{K}_{jp}[1 - F_{s}(\widehat{K}_{jp})] - f_{s}(\widehat{K}_{jp})\right) \tag{7}$$
$$\forall\, j \in J,\ p \in P_{j},\ (j,p) \in\ F$$

On the other hand, if the service level target is CSL, the safety factor is given as an input k_{jp}, as in Eq. (8):

$$\widehat{K}_{jp} = k_{jp} \qquad \forall\, j \in J,\ p \in P_{j},\ (j,p) \notin\ F \tag{8}$$

Solution Approach

The model (1)-(8) corresponds to a nonconvex nonlinear problem (NLP). These problems can be in principle be solved with nonlinear global optimization solvers. However, the computational time required to find a global solution with these solvers may be very expensive. Eq. (7) includes the normal distribution density and cumulative functions. Moreover, Eqs. (1) and (7) include the value σ_{NLTjp}, whose calculation significantly affect the computational burden. To overcome this difficulty, we propose three solution strategies: (i) an exact quadratically constrained reformulation, (ii) a stepwise function to obtain the value of σ_{NLTjp}, and (iii) a piecewise linear approximation of Eq. (7) to replace normal density and cumulative functions.

First, we reformulate the NLP as a MIQCP, that is guaranteed to obtain the global optimum. Solvers like CPLEX and Gurobi can solve MIQCPs quite effectively in reasonable computational times and we avoid to use general nonlinear solvers such as BARON. To account for this reformulation, a new positive continuous variable, U_{jp}, represents the square root, i.e. $U_{jp} \geq \sqrt{\tau_{jp}}$, as shown in Figure 2, being τ the argument of the square root. We include Eq. (9) to define U_{jp}^{2}:

$$U_{jp}^{2} \geq NLT_{jp}\, \sigma_{jp}^{2} + \mu_{jp}^{2}\, \sigma_{NLT_{jp}}^{2} \qquad \forall\, j\ \in J, p \in P_{j} \tag{9}$$

Hence, the objective function becomes a linear function given by Eq. (10):

$$min \sum_{p} \sum_{j \in B_{jp}} h_{jp}\, \widehat{K}_{jp}\, U_{jp} \tag{10}$$

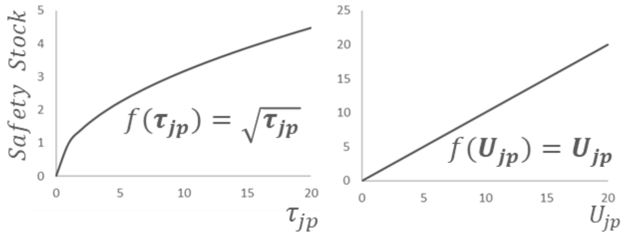

Figure 2. Reformulation from NLP to QCP

In second place, we introduce a stepwise function to represent the NLT standard deviation, σ_{NLTjp}. We assume that the variability of the lead time is pushed downstream if the NLT of node *j* for product *p* is not enough to cover the total replenishment time inherent to node *j*. In other words, if $NLT_{jp} \le SI_{jp} + lt_{jp} + r_{jp}$, then, σ_{LTjp} should be propagated downstream. We introduce a stepwise function as depicted in Figure 3 to represent the value of σ_{NLTjp}, dependent of the value of NLT_{jp}.

Equation (9) is replaced by Eq. (11), with X_{jp} being a positive continuous variable that represents σ_{NLTjp}.

$$U_{jp}^2 \ge NLT_{jp}\, \sigma_{jp}^2 + \mu_{jp}^2\, X_{jp} \qquad \forall\, j \in J, p \in P_j \qquad (11)$$

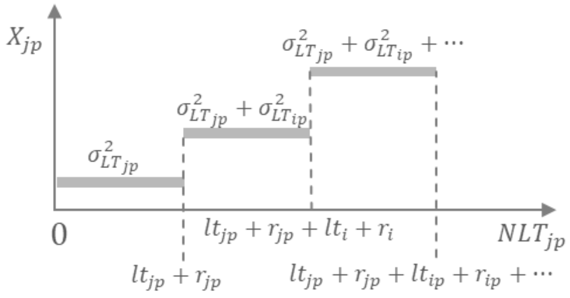
X_{jp}
$\sigma_{LT_{jp}}^2 + \sigma_{LT_{ip}}^2 + \cdots$
$\sigma_{LT_{jp}}^2 + \sigma_{LT_{ip}}^2$
$\sigma_{LT_{jp}}^2$
0
$lt_{jp} + r_{jp} + lt_i + r_i$
NLT_{jp}
$lt_{jp} + r_{jp}$
$lt_{jp} + r_{jp} + lt_{ip} + r_{ip} + \cdots$

Figure 3. Stepwise function to define σ_{NLTjp}

The lower and upper bounds for each step $s \in S_1$ are defined as lb_{jps} and ub_{jps}, respectively. For example, for the first step (s_1) of product *p* at node *j*, $lb_{jps1} = 0$ and $ub_{jps1} = lt_{jp} + r_{jp}$. As NLT_{jp} increases, the bounds increase by adding the lead times and the review periods of upstream nodes. The new binary variable V_{jps} defines what the active step in Eqs. (12) and (13) is according to the value of NLT_{jp}. Equation (14) ensures that only one step is active for each product on each location. Eq. (15) assigns the value of σ_{NLTjp} associated with the step *s*, given by the parameter c_{jps}, which represents the sum of lead time variances according to Figure 3. In the example, $c_{jps1} = 0$, $c_{jps2} = \sigma_{LT_{jp}}^2$, and $c_{jps3} = \sigma_{LT_{jp}}^2 + \sigma_{LT_{ip}}^2$. It is worth to mention that this coefficient also includes the lead time variance of raw materials if they are involved in the production of product *p*.

$$\sum_{s \in S_1} lb_{jps} V_{jps} \le NLT_{jp} \qquad \forall\, j \in J, p \in P_j \qquad (12)$$

$$NLT_{jp} \le \sum_{s \in S_1} ub_{jps} V_{jps} \qquad \forall\, j \in J, p \in P_j \qquad (13)$$

$$\sum_{s \in S_1} V_{jps} = 1 \qquad \forall\, j \in J, p \in P_j \qquad (14)$$

$$X_{jp} = \sum_{s \in S_1} V_{jps}\, c_{jps} \qquad \forall\, j \in J, p \in P_j \qquad (15)$$

Finally, the last solution strategy involves the fill rate constraint. We aim to improve the model tractability by replacing $h(\hat{K}) = \hat{K}[1 - F_s(\hat{K})] - f_s(\hat{K})$ with a piecewise linear approximation, as shown in Figure 4. This function will be the same for all nodes on all locations, while demand is normally distributed. In the example there are four breakpoints and three segments. If the number of breakpoints increases, the precision of the estimation also does, but the computational efficiency decrease. For the current formulation we propose 7 breakpoints.

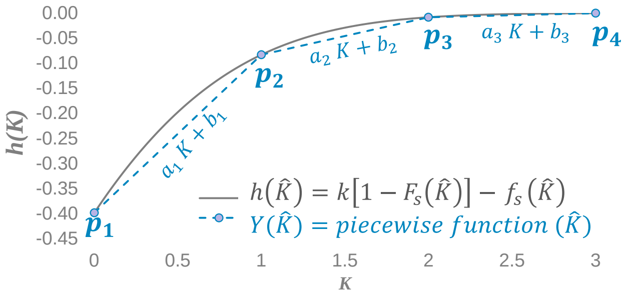

Figure 4. Piecewise linear approximation

Variable Y_{jp} is added to the model to represent the piecewise linear function as in Eq. (16). The value of $\hat{K}_{jp}$ defines which segment $s \in S_2$ is active, and it forces the activation of the binary variable W_{jps} and the continuous variable $\tilde{K}_{jps}$ as in Eqs. (17), (18). Equations (19) and (20) force only one segment to be active.

$$Y_{jp} = \sum_{s \in S_2} \left(a_{jps}\tilde{K}_{jps} + b_{jps}\, W_{jps}\right) \qquad \forall\, j \in J, p \in P_j \qquad (16)$$

$$lb_{jps} W_{jps} \le \tilde{K}_{jps} \qquad \forall\, j \in J, p \in P_j, s \in S_2 \qquad (17)$$

$$\tilde{K}_{jps} \le ub_{jps} W_{jps} \qquad \forall\, j \in J, p \in P_j, s \in S_2 \qquad (18)$$

$$\sum_{s \in S_2} W_{jps} = 1 \qquad \forall\, j \in J, p \in P_j \qquad (19)$$

$$\sum_{s \in S_2} \widetilde{K}_{jps} = \widehat{K}_{jp} \qquad \forall\, j \in J, p \in P_j \qquad (20)$$

Finally, Eq. (7) is redefined including the new variables related to the piecewise and stepwise functions:

$$fr_{jp} \leq 1 + \frac{U_{jp}}{Q_{jp}} Y_{jp} \qquad \forall\, j \in J,\ p \in P_j,\ (j, p) \in F \qquad (21)$$

The new mathematical reformulation is a MIQCP given by Equations (2)-(6), (8), (10)-(20).

Extension for non-normal demand

In industrial practice, it is frequently found that demand data histograms do not fit the shape of a normal distribution. This is generally detected when the coefficient of variation (CV) increases. For distributions with large CV, the model predicts a slightly lower CSL than expected when targets are large, as shown in Figure 5. The plot presents the results of several simulations in Excel for 10,000 periods, assuming a deterministic lead time equal to 1, period between reviews equal to 1, and normally distributed demand datasets with different CVs.

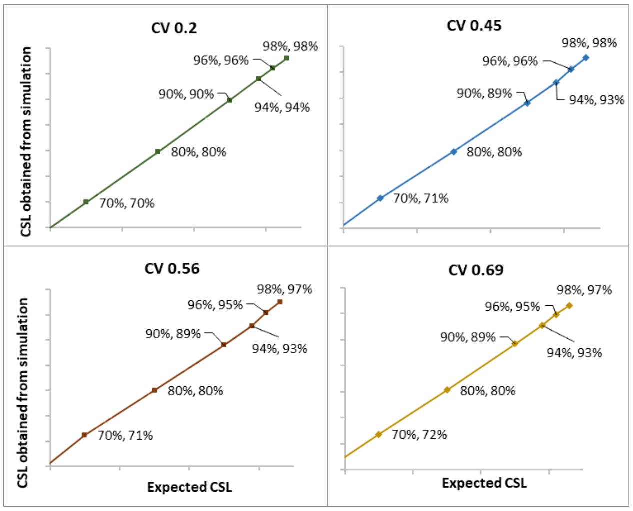

Figure 5. Expected vs Effective CSL

In this work, we aim to find an alternative to adapt the model to set safety stocks for cases that are able to achieve at least the desired customer service level. Mirzaee (2017) proposes an alternative way to set safety stocks by adjusting the service level, using the equivalent *k* safety factor that is equal to the empirical cumulative distribution function (*h(x)*) value. We simulate several datasets of gamma-distributed demands. In Table 1 we present the results obtained for CSL using the *k* safety factors obtained from the normal distribution, and the *k* equivalent from the empirical distribution function *h(x)*. Note that the empirical correlation fits better than the normal for large CSL but it does poorly for low values. We propose a new approach that selects the maximum value between the classic safety factor and the one obtained using the empirical distribution, as stated in Eq. (22):

$$\hat{k}_{jp} = max\left\{\Phi^{-1}(CSL_{jp}),\ \frac{h_{jp}^{-1}(CSL_{jp}) - \mu_{jp}}{\sigma_{jp}}\right\} \qquad (22)$$

Table 2 presents the results of using Eq. (22). All CSLs are achieved or surpassed. Note that this extension only applies when the target service level is the CSL, and the main disadvantage is that it can over-buffer (as in Table 2, row 80%). Future work will extend it for fill rate targets.

Table 1. Effective CSL for large CVs with original and adapted safety factors

Expected CSL	CV = 0.56		CV = 2		CV = 5	
	Normal (N)	Empirical (E)	N	E	N	E
50%	55%	55%	68%	68%	68%	68%
60%	64%	58%	75%	68%	75%	68%
70%	73%	68%	81%	68%	81%	68%
80%	81%	79%	86%	75%	86%	75%
90%	89%	90%	91%	88%	91%	88%
96%	95%	96%	94%	96%	94%	96%
98%	97%	98%	95%	98%	95%	98%

Table 2. Effective CSL for proposed safety factor

Expected CSL	CV=0.56	CV=2	CV=5
50%	55%	68%	68%
60%	64%	75%	75%
70%	73%	81%	81%
80%	81%	86%	86%
90%	90%	91%	91%
96%	96%	96%	96%
98%	96%	98%	98%

Application

In order to illustrate the application of the proposed solution strategy, we carry out a computational experiment based on a real-world case study with 800 products and 18 locations, as shown in Figure 6. The numbers next to each node refer to the number of products that can be stored in that location, because not all products follow the same route. Nodes with people icons mean that those nodes receive external demand. The demand is different for each product, all of them are all stochastic, independent and identically distributed. Lead times are normally distributed.

Figure 7 displays the computational time required to solve different instances ranging from 100 to 800 products. All the instances are modeled with Pyomo and solved with Gurobi 9.5.0, on an Intel Core i7 16 MB RAM. The same instances were tested using both targets: the orange line corresponds to the problem with CSL targets of 97% for all products, and the blue line represents the results fixing a fill rate (FR) target of 98%. Note that real-world cases can be solved to optimality within few seconds of CPU time. The computational burden increase when the fill rate is the target measure, because more constraints and variables become active. Table 3 shows the detailed model sizes and the objective values for some instances. It is worth to mention that using BARON to solve the smallest instance for both the NLP and the QCP formulations could not yield a feasible solution within 1000 seconds. This demonstrates that the proposed MIQCP reformulation solving with Gurobi yields order of magnitude increases in efficiency.

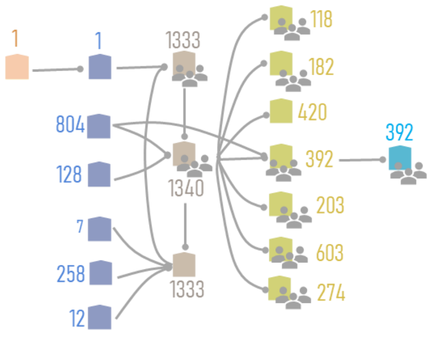

Figure 6. Case-study network

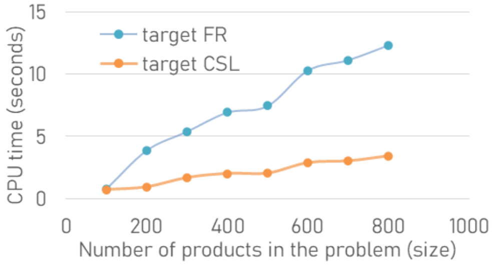

Figure 7. Instance size vs. computational time

Table 3. Model sizes and optimal solutions

Items	Constraints	Continuous variables	Integer variables	Total cost ($)
200	40 502	22 581	30 171	2 652 585
400	78 284	43 499	58 422	5 655 908
600	111 471	63 139	84 507	6 814 902
800	153 123	86 819	115 877	13 738 148

Conclusion

The proposed model integrates several features commonly found in industrial practice that have a strong impact on inventory levels. Moreover, the model with the proposed solution strategy outperforms the original NLP formulation, by finding optimal solutions at minimal computational expense for large scale problems. Therefore, the proposed tool is able to accurately represent real systems, and to set tight safety stock levels in order to achieve target customer service levels with minimum inventory cost. To the best of our knowledge, this is the first model that brings together an MIQCP reformulation with piecewise linear functions to improve the computational efficiency of this optimization model. Future work will address an extension by including responsive characteristics to account for supply chain disruptions and by including storage capacity limitations.

Acknowledgments

The authors gratefully acknowledge the financial support from Johnson and Johnson and the Center for Advanced Process Decision-making (CAPD) from Carnegie Mellon University.

References

Achkar, V. G., Brunaud, B. B., Pérez, H. D., Musa, R., Méndez, C. A., & Grossmann, I. E. (2022). Extensions to the Guaranteed Service Model for Industrial Applications of Multi-Echelon Inventory Optimization. *Submitted for Publication.*

Chopra, S., & Meindl, P. (2013). Supply Chain Management. In *Pearson*. http://www.doiserbia.nb.rs/Article.aspx?ID=0013-32640670067A

Clark, A. J., & Scarf, H. (1960). Optimal Policies for a Multi-Echelon Inventory Problem. *Management Science*, *50*(12_supplement), 1782–1790. https://doi.org/10.1287/mnsc.1040.0265

de Kok, T. (2019). Inventory Management: Modelling Real-life Supply Chains and Empirical Validity. *Foundations and Trends in Technology, Information and Operations Management*, *12*(4), 349–433. http://www.nowpublishers.com/article/Details/TOM-057

Eruguz, A. S., Sahin, E., Jemai, Z., & Dallery, Y. (2016). A comprehensive survey of guaranteed-service models for multi-echelon inventory optimization. *International Journal of Production Economics*, *172*, 110–125. https://doi.org/10.1016/j.ijpe.2015.11.017

Gonçalves, J. N. C., Sameiro Carvalho, M., & Cortez, P. (2020). Operations research models and methods for safety stock determination: A review. *Operations Research Perspectives*, *7*(April), 100164. https://doi.org/10.1016/j.orp.2020.100164

Graves, S. C., & Willems, S. P. (2000). Optimizing Strategic Safety Stock Placement in Supply Chains. *Manufacturing & Service Operations Management*, *2*(1), 68–83. https://doi.org/10.1287/msom.2.1.68.23267

Minner, S. (1998). *Strategic Safety Stocks in Supply Chains*.

Mirzaee, A. (2017). *Alternative Methods for Calculating*. University of Missouri-Columbia.

Simpson, K. F. . (1958). In-Process Inventories. *INFORMS*, *6*(6), 863–873.

You, F., & Grossmann, I. E. (2009). Integrated multi-echelon supply chain design with inventories under uncertainty: MINLP models, computational strategies. *AIChE Journal*, *59*(4), 420–440. https://doi.org/10.1002/aic.12010