Journal Pre-proof

Planning capacity and safety stocks in a serial production-distribution system with multiple products

Foad Ghadimi, Tarik Aouam

PII: S0377-2217(20)30635-4
DOI: https://doi.org/10.1016/j.ejor.2020.07.024
Reference: EOR 16647

To appear in: *European Journal of Operational Research*

Received date: 2 July 2018
Accepted date: 14 July 2020

Please cite this article as: Foad Ghadimi, Tarik Aouam, Planning capacity and safety stocks in a serial production-distribution system with multiple products, *European Journal of Operational Research* (2020), doi: https://doi.org/10.1016/j.ejor.2020.07.024

Planning capacity and safety stocks in a serial production-distribution system with multiple products

Foad Ghadimi [1] Tarik Aouam[1,2,*]

July 11, 2020

Abstract

This study jointly optimizes the production capacity and safety stocks in a serial production-distribution system supplying multiple products under a guaranteed service approach (GSA). The network comprises one manufacturer operating multiple workcenters, one warehouse with limited storage capacity, and one retailer. The manufacturer must efficiently allocate capacity to the workcenters under a limited budget, while the warehouse and retailer need to maintain safety stocks to achieve a target service level. For a single workcenter processing a single product, the interaction between the production lead time, storage capacity, inventory costs, and safety stock placement is characterized. When the manufacturer has multiple workcenters, the integrated problem is formulated as a non-convex program and is solved using a nested Lagrangian relaxation heuristic. The algorithm dualizes the storage constraint in the first phase and the budget constraint in the second phase. A simulation study is conducted to assess the value of the integration, and computational experiments demonstrate that the nested Lagrangian relaxation heuristic can identify optimal or near-optimal solutions in reasonable CPU times.

Keywords— Safety stocks, guaranteed service, multi-echelon model, Lagrangian method, supply chain management

1 Introduction

A production-distribution network comprises manufacturers, wholesalers, and retailers, each of which is a separate unit that typically acts independently (Kotler and Armstrong, 2010). In such a system, production focuses on the allocation of resources to satisfy customer demand, while distribution manages the inventory (Simchi-Levi et al., 2008). From a supply chain management perspective, there are various reasons for which production and distribution decisions should be coordinated (Pyke and Cohen, 1993). In particular, capacity planning at the manufacturer and safety stock placement in

*Corresponding author

[1] Faculty of Economics and Business Administration, Ghent University, Tweekerkenstraat 2, 9000 Gent, Belgium

[2] BearLab - Rabat Business School, Université Internationale de Rabat, Morocco

Email addresses: foad.ghadimi@ugent.be (F. Ghadimi) tarik.aouam@ugent.be (T. Aouam)

the distribution part of the network must be coordinated. When the manufacturer operates multiple workcenters under a limited budget, the capacity allocation affects utilization and production cycle times (PCTs) owing to congestion effects. The assignment of low capacity to a workcenter results in elongated PCTs and requires the distribution side to maintain a high inventory. The cost of holding inventory and warehouse storage capacity should therefore be considered when setting the capacity at the manufacturer. The present paper addresses the integrated capacity planning and safety stock placement in a multi-product production-distribution network comprising one manufacturer, one warehouse with storage capacity, and one retailer in series.

This work integrates capacity planning and safety stock placement. Capacity planning determines the service rates or capacities (in units per time period) at each production stage (facility or workstation) in the supply chain. The capacity can be controlled at the production stages via the number of machines, modernizing or updating equipment, additional maintenance, the number of workers, the number of shifts, the use of overtime, providing additional worker training, etc. (Bretthauer, 1995). Capacity planning is considered at the strategic level when it pertains to assets, i.e. changes in the facilities, over a period of several years. It is considered at the tactical level, dealing with the medium term, which is typically one year, to modify the size of the staff and the amount of working time, but not that of the equipment (Martínez-Costa et al., 2014). In addition, (Martínez-Costa et al., 2014) report that capacity investment is typically not financed by equity without debt, and hence, capacity planning decisions are typically constrained by a finite budget (Bitran and Tirupati, 1989a; Bretthauer and Côté, 1997; Wang et al., 2007; Thomas and Bollapragada, 2010; Woerner et al., 2018). Safety stock placement, which is a tactical supply chain decision, determines the optimal locations and quantities of safety stock in the network to meet a target cycle service level (Graves and Willems, 2003). Safety stock placement should consider storage capacity, which is a scarce resource owing to the increase in land acquisition costs (Hariga, 2010).

Although the problems of capacity and safety stock planning are intimately related, they have been largely treated independent of each other in the literature. Bitran and Tirupati (1989b,c) formulate the manufacturer's capacity planning problem, which determines the processing rate (or capacity) of workcenters to minimize the total work in progress (WIP) holding cost subject to a budget constraint. Based on the resulting capacity allocation, the production lead time (LT) for each workcenter is set. Given these lead times, the strategic safety stock placement problem determines inventory levels at potential inventory locations in order to cope with demand uncertainty (Graves and Willems, 2003). In such a sequential approach, capacity planning does not consider safety stock-related costs or storage capacity at the warehouse. When capacity levels and safety stocks are jointly optimized, which is the approach taken in this paper, lead times become endogenous variables depending on capacity and utilization (Hopp and Spearman, 2011). Setting a high capacity at a workcenter decreases its utilization level as well as the mean and variability of its PCT (Hopp and Spearman, 2011). In turn, this results in a lower average WIP inventory, shorter replenishment lead time (i.e. the time that elapses from the moment an order is placed until it is received), and lower inventories on the distribution side.

This work addresses the problem of jointly optimizing capacity planning and safety stock placement for a production-distribution system supplying multiple products and consisting of one manufacturer, one warehouse with limited storage capacity, and one retailer. The system operates according to the guaranteed service approach (GSA) (Graves and Willems, 2000), where each facility quotes a guaranteed outgoing service time to its downstream customer within which all orders must be satisfied. The warehouse and retailer are considered to be potential safety stock holding locations, and they follow a periodic review base stock policy with a common review period to replenish their inventory. Base stock levels are set to guarantee a target service level and demands exceeding the base stock level are expedited using countermeasures, such as overtime. The manufacturer sets the capacity (or processing rate) of multiple workcenters under a limited budget, and each workcenter sets a deterministic lead time, i.e. the time from when an order arrives at the workcenter until it is shipped to the warehouse. Delayed items at workcenters are expedited using overtime.

For a production-distribution system with a single-workcenter and a single-product, the interaction between the manufacturer's lead time, storage capacity at the warehouse, inventory holding costs, and safety stock placement is analytically characterized. For systems with multiple workcenters and products, the integrated capacity planning and safety stock placement problem is formulated and then solved using a nested Lagrangian relaxation heuristic. The algorithm dualizes the storage constraint in the first phase and the budget constraint in the second phase. This approach decomposes the integrated problem into subproblems, each corresponding to a single-workcenter, which are easy to solve. In both phases, lower bounds are computed by iteratively solving relaxed problems, and efficient greedy heuristics find tight upper bounds. Subgradient procedures update the Lagrangian multipliers until an acceptable optimality gap is reached. Computational experiments show that the nested Lagrangian relaxation heuristic outperforms BARON, a commercial nonlinear optimization solver, in terms of solution quality and run time. In addition, a simulation study is conducted to evaluate the accuracy of the proposed mathematical model and to compare the solution of the integrated approach with a sequential approach for setting capacity and safety stocks. Experiments show that there is a great value in integrating capacity planning and safety stock placement for the considered production-distribution system.

The following section briefly reviews related literature. Section 3 presents the integrated problem formulation and Section 4 analyzes a supply chain with a single workcenter and a single product. Section 5 details the nested Lagrangian relaxation heuristic for solving the integrated problem. Section 6 discusses a simulation study to assess the value of the integration. Computational experiments are presented in Section 7, and Section 8 concludes the paper.

2 Literature review

There are two alternative approaches for optimizing supply chain safety stocks: the GSA and the stochastic service approach (Graves and Willems, 2003). While replenishment lead times are guaranteed or deterministic in GSA, they are stochastic in the stochastic service approach. Simpson Jr (1958)

was the first to present GSA for a serial supply chain, and his work was then extended by Graves and Willems (2000) to supply chains with a spanning tree structure. They assume that each stage operates according to a base stock policy, and quotes a guaranteed outgoing service time to satisfy a bounded demand. When the demand exceeds the demand bound, managers resort to special measures such as expediting, overtime, or subcontracting in order to satisfy excess demand. In this way, replenishment lead times are guaranteed. In contrast, under the stochastic service approach, when the demand exceeds the base stock level, a stockout occurs and replenishment lead times are stochastic (Clark and Scarf, 1960). We refer the readers to Graves and Willems (2003) for a comparison between GSA and the stochastic service approach. The GSA can model large supply chains with general network topologies, and corresponding safety stock placement problems can be efficiently solved using either dynamic programming (Graves et al., 1988; Graves and Willems, 2000) or mixed-integer programming (MIP) techniques (Magnanti et al., 2006). For this reason, GSA has been applied to many industrial cases such as Eastman Kodak (Graves and Willems, 2000), Hewlett-Packard (Billington et al., 2004), Intel (Hsieh, 2011; Wieland et al., 2012), and Nike (Polak, 2014).

There are few studies that propose GSA models with capacitated stages. Graves and Schoenmeyr (2016) considered stages with a fixed capacity, and modified the base stock policy in a way that a stage never places an order to its upstream stage, which is greater than the available capacity. They show that the dynamic programming algorithm proposed by Graves and Willems (2000) can be modified to solve the capacitated case. Graves and Schoenmeyr (2016) made the assumption that lead times are fixed, independent of utilization, which is restrictive in the presence of congestion effects due to variability (Hopp and Spearman, 2011). Lemmens et al. (2016) and Kumar and Aouam (2018b) present models that capture the relationship between the capacity, batch size, and lead times using queuing theory. While the former expresses the mean and variability of the PCT as a function of the batch size in a G/G/m system, the latter considers a G/G/1 queue and optimizes batch sizes. Kumar and Aouam (2018a) considered a distribution network with a capacitated manufacturer and studied the effect of setup reduction on the production lead time and safety stock placement. Kumar and Aouam (2019) studied the effect of production capacity and production smoothing on safety stock placement. Aouam and Kumar (2019) also modeled production lead times based on queuing theory, while considering endogenous safety times and optimized overtime and subcontracting. Ghadimi et al. (2020) extended the model of Aouam and Kumar (2019) by considering limited budgets for allocating capacity to production stages in general acyclic supply chains.

The present paper also reflects the dependency between utilization and production lead times, but differs from the above studies in three ways. First, the above papers consider a single workcenter at a production stage with a known processing rate (capacity), while the present work considers multiple workcenters where processing rates are decision variables that the manufacturer has to optimize under a budget constraint. Second, this work does not assume any queuing model; rather, it assumes that the production lead time of a workcenter is convex, decreasing in its processing rate. A simulation procedure is then presented to illustrate how such a production lead time function can be estimated.

Third, this paper considers a limited storage capacity at the warehouse, which is an important realistic restriction (Akinc and Khumawala, 1977; Ozsen et al., 2008; Liu et al., 2010), and considers the effect of storage capacity on safety stock placement.

Capacity planning subject to congestion is studied for systems that can be modeled as a network of queues. Bitran and Tirupati (1989b,c) considered the processing rate setting, or equivalently the capacity expansion and contraction, in a multi-product manufacturing system with a number of discrete capacity options. They modeled the interaction between the capacity and system performance measures, such as the WIP, throughput, and lead time. Bretthauer (1995) formulated the problem of selecting among discrete capacity options to minimize the cost subject to a given target WIP. The author also presented the dual problem that optimizes WIP while satisfying a budget constraint based on the amount of money available for additional capacity. Bretthauer and Côté (1997) studied the multi-period capacity planning problem, where the capacity at each work station may be varied in discrete time periods. Rajagopalan and Yu (2001) addressed a capacity expansion problem subject to a target cycle time in a single workcenter, multiple-product, and a multiple-server system. They used an M/G/1 queuing model to derive the mean and variance of the cycle time at each workstation. Kim and Uzsoy (2008, 2009) investigated capacity planning subject to congestion effects using concave clearing functions to capture the relationship between the throughput and WIP. While this stream of research recognizes the effect of capacity and utilization on WIP and lead times, it does not consider the effect on the distribution part of the network.

In fact, lead times affect safety stock placement decisions. Hua and Willems (2016) analyzed a two-stage supply chain under GSA, and analytically showed that lead times and inventory costs jointly affect the optimal safety stock placement policy. They found that when the upstream holding cost is low or the upstream lead time is long, maintaining the inventory at both upstream and downstream leads to lower cost, i.e. a better solution. This means that failing to jointly optimize capacity allocation and safety stocks at downstream locations would certainly erode the performance of production-distribution systems. The present paper builds on the above fact and contributes to the GSA literature in four important ways. First, it proposes a model for integrating capacity and safety stock planning for a multi-product supply chain with a limited capacity budget and storage capacity. Second, the interaction between the production lead time, storage capacity, inventory costs, and safety stock placement is characterized in a single-workcenter, single-product supply chain. Third, a nested Lagrangian relaxation heuristic exploits the structure of the formulated problem to decompose it into several subproblems that are easy to solve. Fourth, a simulation study is conducted to evaluate the quality of the solutions and validate the proposed model.

3 Model formulation

This section presents notations, assumptions, and a mathematical formulation for the integrated capacity planning and safety stock placement problem under budget and storage constraints.

3.1 Notations

Sets

$\mathbb{M}$	set of products
$\mathbb{W}$	set of workcenters
$\mathbb{M}_w$	set of products produced at workcenter w

Parameters

λ_j	mean demand for product j at the retailer (units per period)
σ_j	standard deviation of demand for product j at the retailer (units per period)
h_j^{wip}	work-in-process holding cost at the manufacturer for product j (Euro per unit per period)
b_w	cost of one unit of capacity at workcenter w (Euro per unit per period)
α_w^{LT}	fraction of on-time completion at workcenter w
B	manufacturer's budget for allocating capacity (Euro)
K	storage capacity at the warehouse (units)
$h_j^{0,whs}$	inventory holding cost for product j at the warehouse (Euro per unit per period)
$H_j(z_j^{whs})$	(augmented) inventory cost for product j at the warehouse (Euro per unit per period)
$h_j^{0,ret}$	inventory holding cost for product j at the retailer (Euro per unit per period)
$H_j(z_j^{ret})$	(augmented) inventory cost for product j at the retailer (Euro per unit per period)
c_j^m	overtime (expediting) cost for product j at the manufacturer (Euro per unit per period)
c_j^{ds}	overtime (expediting) cost for product j at the distribution side (Euro per unit per period)
$\alpha_j^{whs}, z_j^{whs}$	target service level and corresponding safety factor for product j at the warehouse
$\alpha_j^{ret}, z_j^{ret}$	target service level and corresponding safety factor for product j at the retailer
τ_j^{whs}	delay at the warehouse, which includes review period in addition to transportation and material handling time for product j (periods)
τ_j^{ret}	delay at the retailer, which includes review period in addition to transportation and material handling time for product j (periods)
si_j	external incoming service time for product j at the manufacturer (periods)
s_j	external outgoing service time for product j at the retailer (periods)

Decision variables

R_w	capacity of workcenter w (units per period)
SI_j	incoming service time for product j at the manufacturer (periods)
S_j^{whs}	outgoing service time for product j at the warehouse (periods)
S_j^{ret}	outgoing service time for product j at the retailer (periods)

3.2 Model description and assumptions

Supply chain network. We consider a serial network consisting of three facilities in tandem: a manufacturer, a warehouse, and a retailer, as depicted in Figure 1. Customer requests for multiple products $j \in \mathbb{M}$ arrive continuously at the retailer. The warehouse and retailer are considered to be potential safety stock holding locations and follow a periodic review base stock policy with a common review period to replenish their inventory. The fixed lead times at the warehouse and retailer for product j are τ_j^{whs} and τ_j^{ret}, respectively. The warehouse has a storage capacity K, which limits the amount of inventory that it can hold, i.e. the sum of base stock levels for all products (we assume a unit volume for all items). The manufacturer is composed of multiple workcenters $w \in \mathbb{W}$, each comprising a workstation and staging areas for raw materials and finished goods. The manufacturer has a limited budget B to set production rates of workcenters. The capacity unit cost at workcenter w is b_w. Each item $j \in \mathbb{M}_w$ is processed by a single workcenter w, which sets a deterministic lead time LT_w that is dependent on its capacity. This lead time is the elapsed time between the moment an order of a period is received until it is completed and shipped to the warehouse.

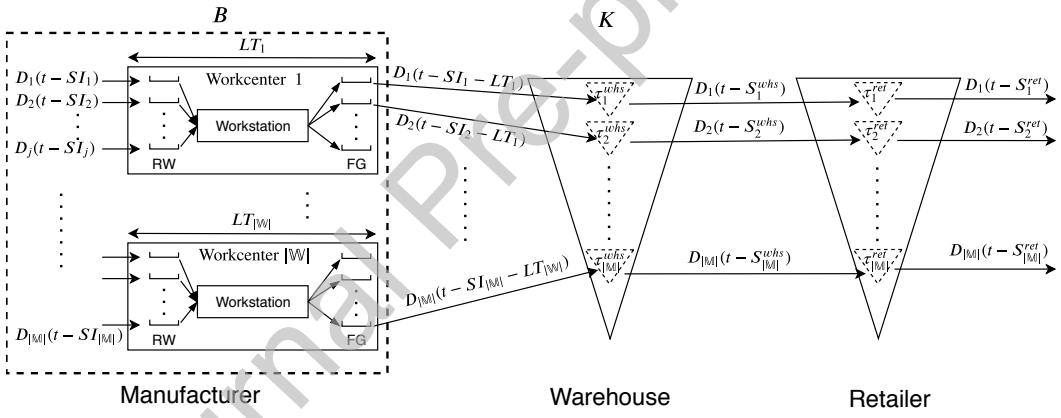

Figure 1: Schematic model of the production-distribution system.

Demand and order processes. The demand $D_j(t)$ for item j at the retailer is a random variable that is observed at the beginning of period t. Although demands may arrive at the retailer continuously, one unit at a time, during period t, it is commonly assumed in periodic review inventory models that the demand is observed at the beginning of t (Axsäter, 2015). The demand is assumed to be independent and identically distributed (i.i.d.) from period to period and independent across products. Demand during net replenishment lead time T (i.e. the difference between the replenishment time and the outgoing service time) is assumed to be normally distributed with mean $\lambda_j T$ and standard deviation $\sigma_j\sqrt{T}$. The demand for product j at the warehouse is $D_j^{whs}(t) = D_j(t)$, and the demand at workcenter w is $D_w^{wc}(t) = \sum_{j\in\mathbb{M}_w} D_j(t)$.

The ordering process at each facility follows a periodic review base stock policy with a common review period among all facilities. A base stock level is set at the warehouse and retailer to cover

demand during net replenishment lead time T with a target service level α_j^{whs} at the warehouse and α_j^{ret} at the retailer. As an example, at the retailer, the base stock level is $\mathcal{B}_j(T) = \lambda_j T + z_j^{ret}\sigma_j\sqrt{T}$, where z_j^{ret} is the safety factor corresponding to α_j^{ret}. Similar to Klosterhalfen et al. (2013) and Aouam and Kumar (2019), we assume that demand exceeding base stock levels is satisfied by expediting items from the pipeline inventory using extraordinary measures, such as overtime. We follow their work and consider a single cost value for each expedited item, irrespective of the duration, and we assume that the pipeline inventory always exceeds the expediting amount.

Guaranteed service times. Each facility promises a guaranteed outgoing service time for product j, denoted by $S_j^{facility}$, to its downstream customer within which it can satisfy all orders. Accordingly, incoming service times to a facility $SI_j^{facility}$ are equal to the outgoing service time guaranteed by the facility's upstream supplier, i.e. $SI_j^{whs} = S_j^{wc}$ and $SI_j^{ret} = S_j^{whs}$. The incoming service time promised by an external supplier to the manufacturer is denoted by SI_j. A workcenter w that processes product j quotes an outgoing service time of $S_j^{wc} = SI_j + LT_w$ to the warehouse because no safety stocks are held at the workcenters. Moreover, for each product, the incoming service time at the manufacturer SI_j must be greater than or equal to the external incoming service time si_j and the outgoing service time at the retailer S_j^{ret} cannot exceed the external outgoing service time s_j.

Production lead times. Each workcenter w sets a guaranteed, deterministic lead time $LT_w(R_w)$ that is dependent on its capacity R_w. At the beginning of a period, orders are received in the raw material staging areas of workcenters. Then, items are continuously released (i.e. one-by-one) to the workstation queue in the same period. After processing, units are placed in the finished goods staging area, and once the lead time is elapsed, completed orders are shipped to the warehouse. When items of an order are delayed during processing, they are expedited using overtime to guarantee a complete shipment. A similar assumption has been made in Çelik and Maglaras (2008), Plambeck and Ward (2007) and Plambeck and Ward (2008) to model overtime as a measure to expedite delayed items at production nodes. The cost related to overtime due to the delay of items in the workcenter is not optimized in our model. Further, we assume that LT_w is a decreasing and convex function of capacity R_w. This assumption is supported by the literature (Yang et al., 2007, 2008; Hopp and Spearman, 2011) and is verified by our simulation study in Section 6.

Description of the production-distribution system. This paper studies a production-distribution system with one manufacturer, consisting of multiple workcenters, supplying one warehouse and one retailer under the GSA.

The demand $D_j(t)$ is observed at the beginning of period t. At the end of period t, the retailer first receives $D_j(t - S_j^{whs} - \tau_j^{ret})$, which is the order placed to the warehouse in period $t - S_j^{whs} - \tau_j^{ret}$, and then satisfies $D_j(t - S_j^{ret})$. For example, when $S_j^{ret} = 0$, the observed demand $D_j(t)$ at the beginning of period t is fulfilled at the end of the same period. If $D_j(t - S_j^{ret})$ is greater than the retailer's

on-hand inventory, the excess demand is satisfied by expediting items from the pipeline inventory. This means that the retailer always satisfies the whole customer demand within its guaranteed service time S_j^{ret}. Finally, the retailer orders $D_j(t)$ to raise its inventory position to the base stock level $\mathcal{B}_j^{ret} = \lambda_j(S_j^{whs} + \tau_j^{ret} - S_j^{ret}) + z_j^{ret}\sigma_j\sqrt{S_j^{whs} + \tau_j^{ret} - S_j^{ret}}$.

At the warehouse, the order placed in period $t - SI_j - LT_w - \tau_j^{whs}$, i.e. $D_j(t - SI_j - LT_w - \tau_j^{whs})$, is received from workcenter w which produces item j. The warehouse then fulfills $D_j(t - S_j^{whs})$, that is, the retailer's order from period $t - S_j^{whs}$. In case of a shortage, i.e. $D_j(t - S_j^{whs})$ is greater than the warehouse's on-hand inventory, the excess demand is satisfied by expediting items from the pipeline inventory so that the retailer's order is satisfied in full within S_j^{whs}. Afterwards, the warehouse orders $D_j(t)$ to the manufacturer in order to raise its inventory position to the base stock level $\mathcal{B}_j^{whs} = \lambda_j(SI_j + LT_w + \tau_j^{whs} - S_j^{whs}) + z_j^{whs}\sigma_j\sqrt{SI_j + LT_w + \tau_j^{whs} - S_j^{whs}}$.

Each workcenter w at the manufacturer processes items $j \in \mathbb{M}_w$. We assume that the quantity $D_j(t - SI_j)$ of each processed item is available at the beginning of period t in the workcenter's RM staging area. These quantities are released continuously, i.e. one unit at a time, to the workstation queue in period t. Units of the various items are released and processed on a first-come first-served (FCFS) basis, according to the sequence of arrival at the retailer. Once a unit is processed at a workstation, it is placed at the corresponding finished goods staging area. At the end of period t, workcenter w fulfills $D_j(t - SI_j - LT_w)$, which is the order of the warehouse in period $t - SI_j - LT_w$ for product j. When $D_j(t - SI_j - LT_w)$ is greater than the on-hand inventory for product j at the finished goods area, the excess order is satisfied by expediting items from the WIP inventory at the workcenter. In this manner, the complete order of the warehouse is satisfied within the guaranteed service time $S_j^{wc} = SI_j + LT_w$.

3.3 Formulation of the integrated capacity planning and safety stock placement problem

The objective of the integrated capacity planning and safety stock placement problem is to set capacity (processing rates of workcenters) and service times to minimize WIP holding costs of the manufacturer and inventory cost at the warehouse and retailer, in addition to overtime costs for expediting shortages. The derivation of the expected total cost is based on the work of Klosterhalfen and Minner (2010) and Aouam and Kumar (2019) and provided in Appendix A. The integrated capacity planning and safety stock placement problem **P** can be formulated as follows,

$$\mathbf{P}\ \min \sum_{w\in\mathbb{W}}\sum_{j\in\mathbb{M}_w}\Bigg(h_j^{wip}\lambda_j LT_w(R_w) + H_j(z_j^{whs})\sigma_j\sqrt{SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}} + H_j(z_j^{ret})\sigma_j\sqrt{S_j^{whs} + \tau_j^{ret} - S_j^{ret}}\Bigg) \tag{1}$$

subject to:

$$\sum_{w\in\mathbb{W}} b_w R_w \le B \tag{2}$$

$$\sum_{j\in\mathbb{M}_w} \lambda_j < R_w \qquad \forall w \in \mathbb{W} \tag{3}$$

$$\sum_{w\in\mathbb{W}}\sum_{j\in\mathbb{M}_w} \lambda_j\left(SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}\right) + z_j^{whs}\sigma_j\sqrt{SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}} \le K \tag{4}$$

$$\max(0, S_j^{ret} - \tau_j^{ret}) \le S_j^{whs} \le SI_j + LT_w(R_w) + \tau_j^{whs} \qquad \forall w \in \mathbb{W}, \forall j \in \mathbb{M}_w \tag{5}$$

$$si_j \le SI_j \qquad \forall j \in \mathbb{M} \tag{6}$$

$$S_j^{ret} \le s_j \qquad \forall j \in \mathbb{M} \tag{7}$$

$$SI_j, S_j^{whs}, S_j^{ret} \in \mathbb{Z}^+ \qquad \forall j \in \mathbb{M} \tag{8}$$

$$LT_w(R_w) \in \mathbb{Z}^+, R_w \in \mathbb{R}^+ \qquad \forall w \in \mathbb{W} \tag{9}$$

The objective function minimizes the expected total cost of the production-distribution system. The augmented inventory costs at the warehouse and retailer, $H_j(z_j^{whs})$ and $H_j(z_j^{ret})$ respectively, include the unit cost of holding inventory and expediting excess demand. Constraint (2) ensures that the cost for allocating capacities to workcenters does not exceed the manufacturer's budget. Constraints (3) state that the capacity assigned to a workcenter should be greater than its total arrival rate (demand). Constraint (4) ensures that the sum of base stock levels does not exceed the storage capacity of the warehouse. Constraints (5) restrict the net replenishment lead time of the warehouse and retailer to be positive. Constraints (6) put a limit on the lowest possible amount of incoming service time at the manufacturer, and constraints (7) ensure that the external demand will be satisfied within its maximum acceptable outgoing service time. Constraints (8) and (9) define all service times and lead times as integer variables.

4 Impact of the storage capacity on safety stock placement

In this section, we study the safety stock placement problem in a supply chain that consists of a single workcenter processing a single product and supplying a warehouse with limited storage capacity and a retailer, as depicted in Figure 2. We characterize the optimal safety stock placement and the effect of the lead time and inventory costs on safety stocks. The derived results extend the work of Minner (2000) and Hua and Willems (2016), which studied safety stock placement in two-stage serial systems with infinite storage capacity under the GSA.

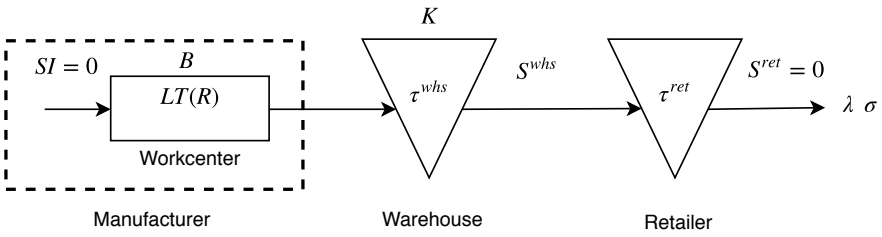

Figure 2: Schematic model of a single-product production-distribution system.

In the rest of the paper, *inventory coupling* refers to the situation when inventory is held only at the retailer, and no inventory is kept at the warehouse, while *inventory decoupling* refers to the case when inventory is kept at both the warehouse and the retailer, and the warehouse capacity is not binding. In addition, *inventory partial decoupling* refers to the case when keeping inventory at the warehouse results in a lower cost; however, because of the storage capacity, part of the inventory is shifted to the retailer. Below, we first derive the optimal outgoing service time of the warehouse. Then, the impact of the production lead time and the ratio of the upstream to downstream inventory costs $\left(\text{i.e. } \nu = \frac{H(z^{whs})}{H(z^{ret})}\right)$ on optimal safety stock placement is characterized.

4.1 Optimal safety stock placement

In the above supply chain, it is optimal to set $R = \frac{B}{b}$ because there is only one workcenter and lead time is decreasing in capacity. The integrated problem with storage capacity K reduces to optimizing the outgoing service time at the warehouse S^{whs}, and can be re-written as follows:

$$\mathbf{P}^K \quad \min \quad H(z^{whs})\sigma\sqrt{RLT - S^{whs}} + H(z^{ret})\sigma\sqrt{S^{whs} + \tau^{ret}} \tag{10}$$

$$\lambda\left(RLT - S^{whs}\right) + z^{whs}\sigma\sqrt{RLT - S^{whs}} \leq K \tag{11}$$

$$0 \leq S^{whs} \leq RLT \tag{12}$$

where $RLT = LT(R) + \tau^{whs}$ is the effective replenishment lead time that the warehouse is facing. The objective function (10) minimizes the total inventory cost, and constraint (11) ensures that the base stock level does not exceed the storage capacity at the warehouse. Constraint (12) restricts the net replenishment lead time at the warehouse and retailer to be non-negative.

Based on Minner (2000) and Hua and Willems (2016), Lemma 1 provides the optimal service time at the warehouse when there is infinite storage capacity. This Lemma states that the optimal service time lies at one of two extreme points.

Lemma 1. *The optimal outgoing service time at the warehouse in problem* $\mathbf{P}^K$ *with infinite storage capacity is given by* $S^{whs,*} \in \{0, RLT\}$.

Proposition 1 extends the result of Lemma 1 to the case of limited storage capacity at the warehouse. The proposition states that the optimal service time at the warehouse lies at one of three extreme points.

Proposition 1. *The optimal outgoing service time at the warehouse in problem* $\mathbf{P}^K$ *with storage capacity* K *is given by* $S^{whs,*} \in \{0, S^K, RLT\}$*, where* $S^K = RLT - \frac{(z^{whs})^2\sigma^2+2\lambda K-z^{whs}\sigma\sqrt{(z^{whs})^2\sigma^2+4\lambda K}}{2\lambda^2}$ *is the outgoing service time at the warehouse when the storage constraint* (11) *is binding.*

Proof. See Appendix B.1. □

This Proposition extends the two extreme points property of the optimal safety stock placement under GSA in serial systems to reflect the storage capacity. The optimal solution now lies at three extreme points: (i) $S^{whs,*} = RLT$ for *inventory coupling* at the retailer, (ii) $S^{whs,*} = 0$ corresponds to *inventory decoupling*, and (iii) $S^{whs,*} = S^K$ when the storage capacity is binding and we have *inventory partial decoupling*, i.e. some of the inventory is shifted to the retailer.

4.2 Impact of the replenishment lead time

When there is infinite storage capacity at the warehouse, Lemma 2 shows that the optimal safety stock placement is determined by a threshold on replenishment lead time RLT_1^0 (see Minner (2000) and Hua and Willems (2016)). This threshold is a function of the ratio of inventory costs ν. When the replenishment lead time is less than RLT_1^0, the inventory is coupled at the retailer; otherwise, the inventory is decoupled, i.e. the inventory is kept at both the warehouse and retailer.

Lemma 2. *Fixing all parameters and changing only the replenishment lead time* RLT*, the optimal solution of problem* $\mathbf{P}^K$ *when there is infinite storage capacity is given by:*

a) when $RLT_1^0 \leq RLT$*, then* $S^{whs,*} = 0$*, and both the warehouse and retailer keep a safety stock (decoupling),*

b) when $RLT \leq RLT_1^0$*, then* $S^{whs,*} = RLT$*, and the safety stock is only kept at the retailer (coupling).*

where $RLT_1^0 = \max\{\tau^{ret}\frac{4\nu^2}{(1-\nu^2)^2}, 0\}$.

Proof. See Appendix B.2. □

Proposition 2 extends the results of Lemma 2 by considering a finite storage capacity, and defines a storage capacity threshold K^0 and a threshold on the replenishment lead time RLT_2^0 that together define the optimal safety stock placement. The case of $K \geq K^0$ is similar to the infinite storage capacity case, and the optimal safety stock placement can be obtained based on Lemma 2. When $K \leq K^0$, i.e. there is a finite storage capacity, the solution in Lemma 2 does not satisfy the storage constraint (11). In this case, the optimal safety stock placement is determined by Proposition 2.

Proposition 2. *By fixing all parameters and changing only the replenishment lead time* RLT*, the optimal solution of problem* $\mathbf{P}^K$ *when there is a finite storage capacity (i.e.* $K \leq K^0$*) is given by:*

a) when $RLT \leq RLT_2^0$*, then* $S^{whs,*} = S^K$*, the storage capacity is binding, and both the warehouse and retailer maintain a safety stock (partial decoupling),*

b) when $RLT_2^0 \leq RLT$*, then* $S^{whs,*} = RLT$*, and the safety stock is only kept at the retailer (coupling),*

where $K^0 = \lambda RLT + z^{whs}\sigma\sqrt{RLT}$ *and* $RLT_2^0 = \max\left\{\frac{(z^{whs})^2\sigma^2+2\lambda K-z^{whs}\sigma\sqrt{(z^{whs})^2\sigma^2+4\lambda K}}{2\lambda^2}\left(\frac{(1+\nu^2)^2}{4\nu^2}\right) - \tau^{ret}, 0\right\}$.

Proof. See Appendix B.3. □

Proposition 2 shows that when the storage capacity is less than K^0, the safety stock placement depends on RLT_2^0, which is a function of both ν and K. In this case, the inventory is coupled at the retailer if the replenishment lead time is greater than RLT_2^0; otherwise, the inventory is partially decoupled. Table 2 defines the optimal safety stock placement for both infinite and finite storage capacity cases depending on the replenishment lead time RLT.

Table 2: Optimal service time at the warehouse based on the replenishment lead time threshold.

Storage capacity	Replenishment lead time	Optimal service time
$K \geq K^0$ (Infinite)	$RLT_1^0 \leq RLT$	$S^{whe,*} = 0$
	$RLT \leq RLT_1^0$	$S^{whe,*} = RLT$
$K \leq K^0$ (Finite)	$RLT \leq RLT_2^0$	$S^{whe,*} = S^K$
	$RLT_2^0 \leq RLT$	$S^{whe,*} = RLT$

The following corollaries provide properties of thresholds K^0, RLT_1^0, and RLT_2^0.

Corollary 2.1. *The threshold for storage capacity* (K^0) *is an increasing function of the replenishment lead time* RLT.

Corollary 2.2. *For infinite storage capacity, i.e.* $K \geq K^0$*, the replenishment lead time threshold* RLT_1^0 *is not a function of* K*, and is an increasing function of* ν.

Proof. See Appendix B.4. □

Corollary 2.3. *For a finite storage capacity, i.e.* $K \leq K^0$*, the replenishment lead time threshold* RLT_2^0 *is an increasing function of* K *and a decreasing function of* ν.

Proof. See Appendix B.5. □

To illustrate our analytical results in this section, we use the following instance as a base case: $\lambda = 10$ and $\sigma = 3.16$ units per period, $b = 1$ Euro, $\nu = 0.5$, $\tau^{whs} = 2$ and $\tau^{ret} = 2$ periods, $\alpha_j^{whs} = 0.95$. The budget at the manufacturer is $B = 10.53$ Euro, which is obtained based on 95% utilization, i.e. $B = \frac{b\lambda}{0.95}$. The storage capacity at the warehouse is $K = 0.75 \times K^0 = 68.51$ units, where K^0 is the storage capacity threshold derived in Proposition 2. We also assume that the production lead time has the following form $LT(R) = \frac{\ln(1-0.95)}{R-\lambda}$, which is obtained based on the M/M/1 queuing model, with 95% as the fraction of on-time completion (Kleinrock, 1975).

Figure 3 plots K^0 as a function of the replenishment lead time RLT and illustrates Corollary 2.1. Figure 4 shows the replenishment lead time threshold RLT_1^0 as an increasing function of ν, which is in line with Corollary 2.2. In fact, for infinite storage capacity, i.e. $K \geq K^0$, as the difference between the upstream and downstream inventory costs becomes insignificant, i.e. ν increases, the inventory is coupled for longer replenishment.

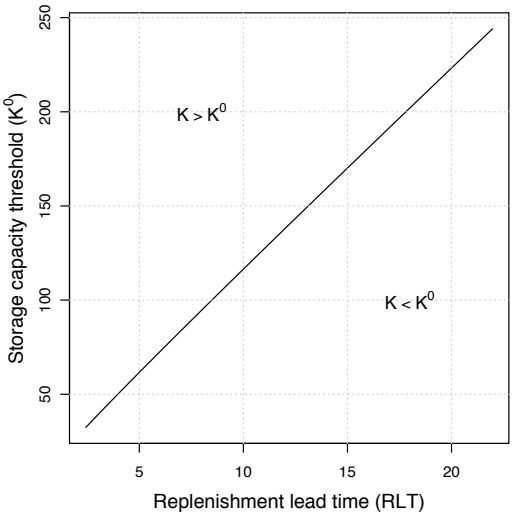

Figure 3: Storage capacity threshold K^0 as a function of replenishment lead time RLT.

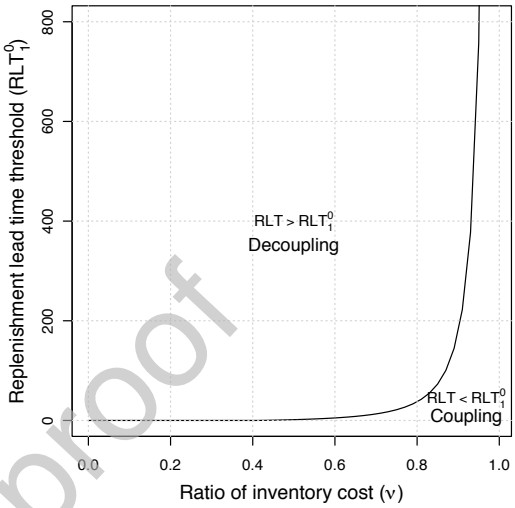

Figure 4: Replenishment lead time threshold RLT_1^0 as a function of ν.

Corollary 2.3 states that for finite storage capacity, i.e. $K \leq K^0$, as the storage capacity increases, the inventory is partially decoupled for longer replenishment lead times. Figure 5 shows the replenishment lead time threshold RLT_2^0 as an increasing function of K for three different values of ν. For a given K, the inventory is partially decoupled when as ν decreases. Based on Figure 5, we can also see that there is a minimum value of storage capacity below which the inventory is always coupled at the retailer. Corollary 2.3 also states that as ν increases, the inventory is partially decoupled for shorter replenishment LTs. This is illustrated by Figure 6, which shows the replenishment lead time threshold RLT_2^0 as a decreasing function of ν for three different values of storage capacity.

4.3 Impact of the ratio of inventory costs

When there is infinite storage capacity at the warehouse, Lemma 3, which is obtained based on the work of Minner (2000) and Hua and Willems (2016), shows that the optimal safety stock placement depends on a threshold based on the ratio of inventory costs ν_1^0. This threshold is a function of the replenishment lead time RLT. When $\nu \geq \nu_1^0$, the inventory is coupled at the retailer; otherwise, the inventory is decoupled, i.e. the inventory is kept at both the warehouse and retailer. In fact, for all values of $\nu \leq \nu_1^0$, representing situations where the inventory cost at the warehouse is much cheaper than the one at the retailer, the inventory is decoupled. As the difference between the inventory costs becomes insignificant, i.e. $\nu \geq \nu_1^0$, inventory coupling at the retailer leads to a lower cost.

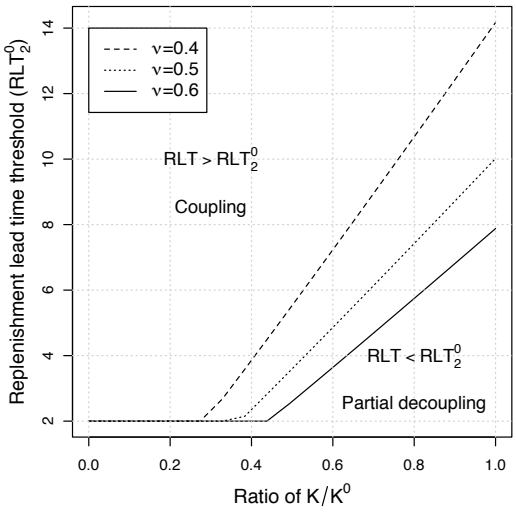

Figure 5: Replenishment lead time threshold RLT_2^0 as a function of K for different values of ν.

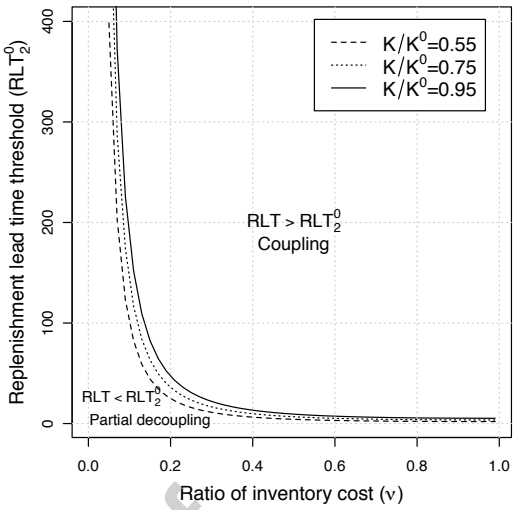

Figure 6: Replenishment lead time threshold RLT_2^0 as a function of ν for different values of K.

Lemma 3. *By fixing all parameters and changing only the ratio of inventory costs ν, the optimal solution of problem $\mathbf{P}^K$ when there is infinite storage capacity is given by:*

a) when $\nu \leq \nu_1^0$, then $S^{whs,*} = 0$ and both the warehouse and retailer maintain a safety stock (decoupling),

b) when $\nu \geq \nu_1^0$, then $S^{whs,*} = RLT$ and a safety stock is only kept at the retailer (coupling)

where $\nu_1^0 = \sqrt{1 + \frac{\tau^{ret}}{RLT}} - \sqrt{\frac{\tau^{ret}}{RLT}}$.

Proof. See Appendix B.6. □

Proposition 3 extends the results of Lemma 3 by considering a finite storage capacity, and defines a threshold based on the ratio of inventory costs ν_2^0 that defines the optimal safety stock placement. The case of $K \geq K^0$ is similar to the infinite storage capacity case, and the optimal safety stock placement can be obtained based on Lemma 3. When $K \leq K^0$, i.e. there is a finite storage capacity, the solution in Lemma 3 does not satisfy the storage constraint (11). In this case, the optimal safety stock placement is given by Proposition 3.

Proposition 3. *By fixing all parameters and changing only the ratio of inventory costs ν, the optimal solution of problem $\mathbf{P}^K$ when there is finite storage capacity (i.e. $K \leq K^0$) is given by:*

a) when $\nu \leq \nu_2^0$, then $S^{whs,*} = S^K$, the storage capacity is binding, and both the warehouse and retailer maintain a safety stock (partial decoupling),

b) when $\nu \geq \nu_2^0$, then $S^{whs,*} = RLT$, and a safety stock is only kept at the retailer (coupling).

where $\nu_2^0 = \sqrt{\frac{2\lambda^2(RLT+\tau^{ret})}{(z^{whs})^2\sigma^2+2\lambda K-z^{whs}\sigma\sqrt{(z^{whs})^2\sigma^2+4\lambda K}}} - \sqrt{\frac{2\lambda^2(RLT+\tau^{ret})}{(z^{whs})^2\sigma^2+2\lambda K-z^{whs}\sigma\sqrt{(z^{whs})^2\sigma^2+4\lambda K}} - 1}$.

Proof. See Appendix B.7. □

Proposition 3 shows that when the storage capacity is less than K^0, the safety stock placement depends on ν_2^0, which is a function of both RLT and K. In this case, the inventory is coupled at the retailer when the difference between the inventory costs is not significant, i.e. $\nu \geq \nu_2^0$. When $\nu \leq \nu_2^0$, the inventory is partially decoupled. Table 3 shows the optimal safety stock placement for both infinite and finite storage capacity based on the threshold for the ratio of inventory costs.

Table 3: Optimal service time at the warehouse based on the threshold on the ratio of inventory costs.

Storage capacity	Ratio of inventory cost	Optimal service time
$K \geq K^0$ (Infinite)	$\nu \leq \nu_1^0$	$S^{whe,*} = 0$
	$\nu \geq \nu_1^0$	$S^{whe,*} = RLT$
$K \leq K^0$ (finite)	$\nu \leq \nu_2^0$	$S^{whe,*} = S^K$
	$\nu \geq \nu_2^0$	$S^{whe,*} = RLT$

The properties of ν_1^0 and ν_2^0 are given by the following corollaries.

Corollary 3.1. *For an infinite storage capacity, i.e. $K \geq K^0$, the threshold ν_1^0 is an increasing function of RLT.*

Proof. See Appendix B.8. □

Corollary 3.2. *For a finite storage capacity, i.e. $K \leq K^0$, the threshold ν_2^0 is a decreasing function of RLT, and increases in K.*

Proof. See Appendix B.9. □

Corollary 3.1 characterizes the effect of the replenishment lead time RLT on the threshold ν_1^0. This effect is illustrated in Figure 7, which shows that ν_1^0 is an increasing function of the replenishment lead time RLT. This means that when there is infinite storage capacity, as the replenishment lead time increases inventory decoupling happens only when the difference between inventory costs at the warehouse and retailer is very high, i.e. ν is very low.

Corollary 3.2 states that when there is a finite storage capacity, ν_2^0 is a decreasing function of RLT and an increasing function of K. Figure 8 plots the threshold ν_2^0 as a decreasing function of the replenishment lead time RLT for three different values of K. This figure shows that when the replenishment lead time increases, the inventory is coupled at the retailer only when the difference between the inventory cost at the warehouse and retailer is not significant, i.e. ν is high. Figure 9 shows the threshold ν_2^0 as an increasing function of K for three different values of utilization. This figure also shows that there is a maximum value of storage capacity at the warehouse above which the inventory is always partially decoupled. Similarly, based on Figure 9, one can observe that there is a minimum value of utilization below which the inventory is always partially decoupled.

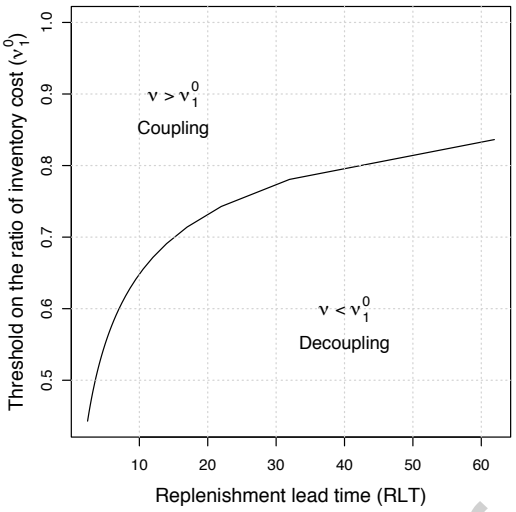

Figure 7: ν_1^0 as a function of the replenishment lead time RLT.

From the above analysis, it is clear that the storage capacity at the warehouse, replenishment lead time, and inventory costs affect the safety stock placement in a supply chain. The replenishment lead time is a function of the manufacturer's budget, and hence the manufacturer's budget also affects the safety stock placement. For finite storage capacity at the warehouse, inventory coupling results in a lower cost when the replenishment lead time is very long or when the inventory cost downstream is very expensive.

5 Solving the integrated problem using a nested Lagrangian relaxation heuristic

When the manufacturer operates multiple workcenters, each of which processes multiple products, the production lead times are not fixed, and the objective function is no longer concave, as was the case in the previous section. In fact, production lead times become variables that are dependent on capacity decisions, and the objective function of problem **P** is non-convex. Therefore, problem **P** is a non-convex problem and falls into the difficult class of global optimization problems (Horst et al., 2000). In this section, we propose a nested Lagrangian relaxation heuristic as an integrated solution approach of problem **P**.

The proposed nested Lagrangian relaxation heuristic dualizes constraints (2) and (4), which link workcenters and products, over two phases. In the first phase, the storage constraint (4) is relaxed with a Lagrangian multiplier κ. The objective value of the relaxed problem provides a lower bound for problem **P**. The solution of this relaxed problem satisfies the budget constraint, but not necessarily the storage constraint. By repairing this solution, i.e. providing feasible solutions, and updating the Lagrangian multiplier κ, an optimal or near-optimal solution of **P** can be obtained. To solve the relaxed problem of the first phase, the budget constraint (2) is relaxed with a Lagrangian multiplier $\gamma \geq 0$

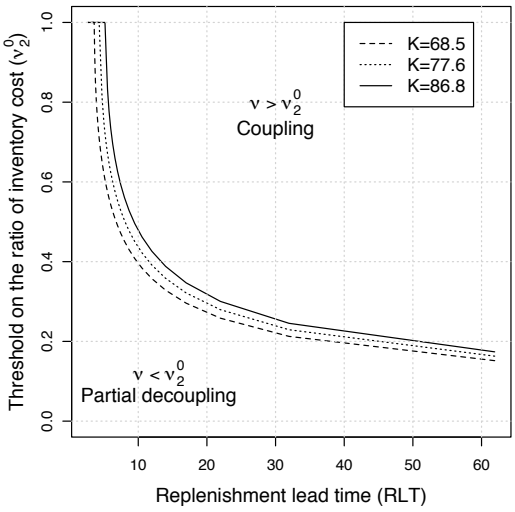

Figure 8: ν_2^0 as a function of the replenishment lead time RLT for different values of K.

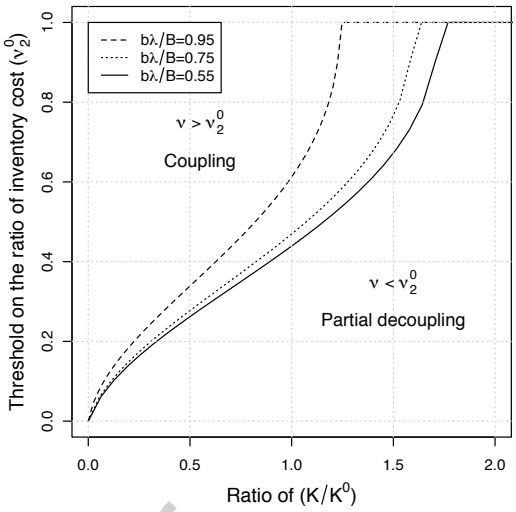

Figure 9: ν_2^0 as a function of K for different values of B.

in the second phase. The objective value of the resulting relaxed problem provides a lower bound for the relaxed problem of the first phase. The lower bound solution for the relaxed problem of the first phase does not necessarily satisfy the budget constraint. Using an efficient greedy heuristic to repair the lower bound solutions, upper bounds can be generated for the relaxed problem of the first phase. By updating the Lagrangian multiplier γ in each iteration, an optimal or near-optimal solution for the relaxed problem of the first phase can be obtained. The nested Lagrangian relaxation algorithm is summarized in Figure 10.

5.1 First phase: relaxing the storage constraint

In the first phase of the nested Lagrangian relaxation, the storage constraint (4) is relaxed with a Lagrangian multiplier κ. The relaxed problem of the first phase $\mathbf{NLR}_\kappa^1$ for a given κ is formulated as follows:

$$\begin{aligned}\mathbf{NLR}_\kappa^1 \quad NL^1(\kappa) = \min \sum_{w\in\mathbb{W}}\sum_{j\in\mathbb{M}_w} \Bigg(& h_j^{wip}\lambda_j LT_w(R_w) + H_j(z_j^{whs})\sigma_j\sqrt{SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}} \\ & + H_j(z_j^{ret})\sigma_j\sqrt{S_j^{whs} + \tau_j^{ret} - S_j^{ret}}\Bigg) + \kappa\Bigg(\sum_{w\in\mathbb{W}}\sum_{j\in\mathbb{M}_w}\lambda_j\left(SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}\right) \\ & + z_j^{whs}\sigma_j\sqrt{SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}} - K\Bigg) \qquad (13)\end{aligned}$$

subject to: constraints (2), (3), (5), (6), (7), (8) and (9).

By solving $\mathbf{NLR}_\kappa^1$ for a given κ, a lower bound on the objective value of problem $\mathbf{P}$ can be obtained. An optimal or near-optimal solution of $\mathbf{NLR}_\kappa^1$ is obtained in the second phase of the nested Lagrangian relaxation heuristic. This solution satisfies the budget constraint (2), but does not necessarily satisfy the storage constraint (4). Next, we present an upper bound method that generates a feasible solution

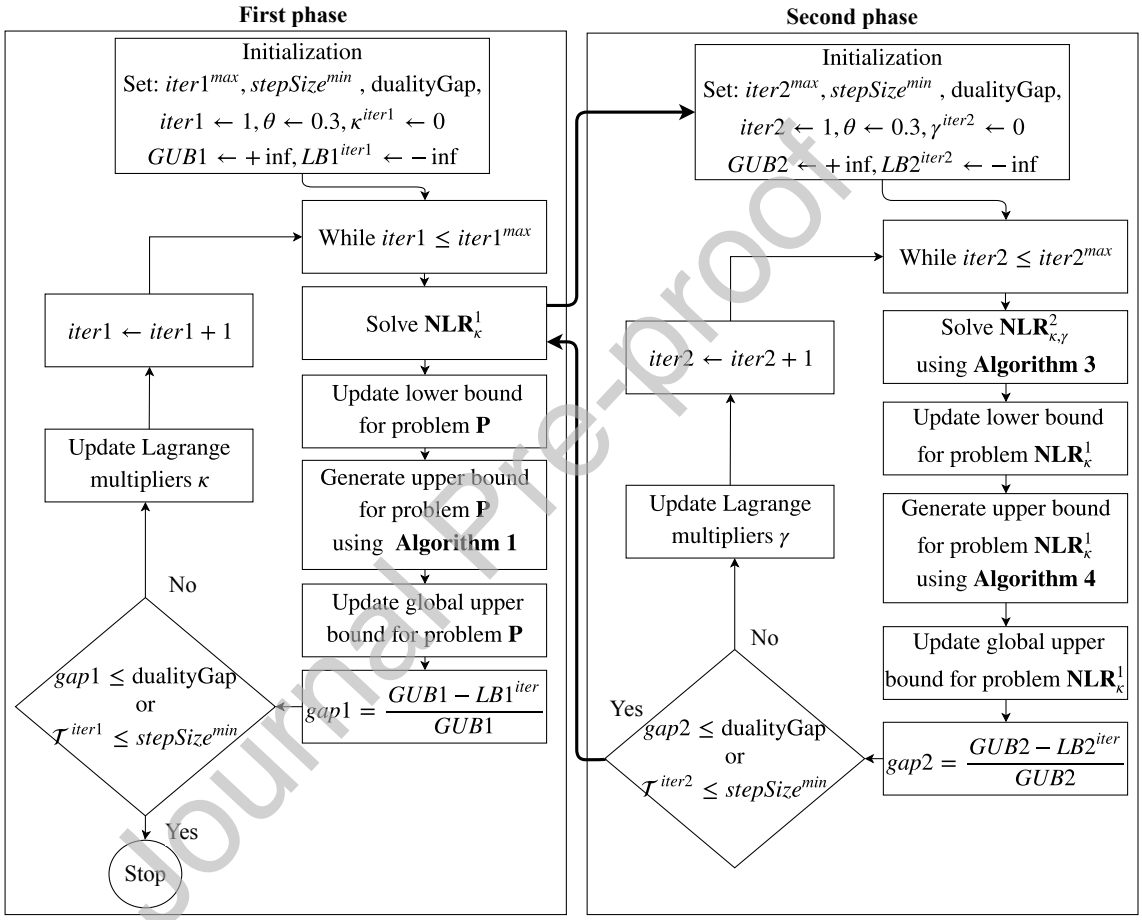

Figure 10: Flowchart diagram for the nested Lagrangian relaxation heuristic.

for $\mathbf{P}$ by repairing the solution of $\mathbf{NLR}^1_\kappa$.

5.1.1 Generating upper bounds for problem P

When the solution of $\mathbf{NLR}^1_\kappa$ satisfies the storage constraint (4), it is also a feasible solution for problem $\mathbf{P}$. As a result, the corresponding objective value becomes an upper bound on the objective value of $\mathbf{P}$. When this is not the case, we use a greedy heuristic to eliminate the surplus of storage capacity, denoted by $Surplus1$. The proposed greedy heuristic does not change the capacity level R_w because it satisfies the budget constraint, but it modifies the outgoing service times at the warehouse. Algorithm 1 lists the steps of the greedy heuristic, where $f_{\mathbf{P}}$ is the objective value of problem $\mathbf{P}$.

The algorithm identifies all products that are stored at the warehouse in the lower bound solution, i.e. products with positive net replenishment lead time at the warehouse or $NRT_j^{whs} > 0$ (Line 7). For each product j stored at the warehouse, the algorithm calculates $Loss1[j]$ which corresponds to the marginal increment in the objective value when the inventory of j is shifted from the warehouse to the retailer, i.e. when $S_j^{whs} = SI_j + LT_w(R_w) + \tau_j^{whs}$ (Line 8). If a product j is only stored at the retailer in the solution then $Loss1[j] = +\infty$ (Line 9). For all products with negative $Loss1$, we set the outgoing service time at the warehouse to be equal to $S_j^{whs,UB} = SI_j + LT_w(R_w) + \tau_j^{whs}$ because it results in a lower objective value (Line 11). These products are then shifted from the warehouse to the retailer and we set $Loss1[j] = +\infty$.

Afterwards, we start from the product with the lowest value of $Loss1$ (Line 13). The corresponding product j results in the largest free space at the warehouse with the smallest increment in the objective function. For this product, we set $S_j^{whs,UB} = SI_j + LT_w(R_w) + \tau_j^{whs}$ and $Loss1[j] = +\infty$ since this product is now moved to the retailer (Line 15). $Surplus1$ is also updated by subtracting the base stock level of j. The same process is repeated until the storage constraint (4) is satisfied, i.e. $Surplus1 \leq 0$. In the last iteration, if the base stock level of the last product is greater than the surplus, i.e. $\mathcal{B}_j > Surplus1$, then we change the base stock level to $\mathcal{B}_j - Surplus1$, which means that a quantity $Surplus1$ of product j is moved to the retailer. This corresponds to setting $S_j^{whs,UB} = \left\lceil \frac{(z_j^{whs})^2\sigma_j^2 + 2\lambda_j(\mathcal{B}_j - Surpluss1) - z_j^{whs}\sigma\sqrt{(z_j^{whs})^2\sigma_j^2 + 4\lambda_j(\mathcal{B}_j - Surpluss1)}}{2\lambda_j^2} - \tau_j^{whs} - SI_j \right\rceil$ (Line 17). In this case, the storage capacity constraint (4) would be satisfied.

Algorithm 1 Greedy heuristic to generate upper bounds for problem **P**.

function UpperBound1(κ)
  $UB1 \leftarrow 0$
  for all $w \in \mathbb{W},\ j \in \mathbb{M}_w$ **do** $R_w^{\text{UB}} \leftarrow R_w\ S_j^{whs,UB} \leftarrow S_j^{whs},\ NRT_j^{whs} \leftarrow SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}$ **end for**
  if $\sum_{w\in\mathbb{W}}\sum_{j\in\mathbb{M}_w} \lambda_j(NRT_j^{whs}) + z_j^{whs}\sigma_j\sqrt{NRT_j^{whs}} > K$ **then**
  $Surplus1 \leftarrow \sum_{w\in\mathbb{W}}\sum_{j\in\mathbb{M}_w} \lambda_j(NRT_j^{whs}) + z_j^{whs}\sigma_j\sqrt{NRT_j^{whs}} - K$
  for all $j \in \mathbb{M}$ **do**
  if $NRT_j^{whs} > 0$ **then**
  $Loss1[j] = \frac{f_{\mathbf{P}}\left(SI_j + LT_w(R_w) + \tau_j^{whs}\right) - f_{\mathbf{P}}\left(S_j^{whs}\right)}{\mathcal{B}_j}$
  else $Loss1[j] = +\infty$ **end if**
  if $Loss1[j] < 0$ **then**
  $S_j^{whs,UB} \leftarrow SI_j + LT_w(R_w) + \tau_j^{whs},\ Surplus1 \leftarrow Surplus1 - \mathcal{B}_j,\ Loss1[j] = +\infty$ **end if end for**
  while $Surplus1 > 0$ **do**
  FindProductWithLowestValueIn($Loss1$)
  if $Surplus1 - \mathcal{B}_j > 0$ **then**
  $S_j^{whs,UB} \leftarrow SI_j + LT_w(R_w) + \tau_j^{whs},\ Surplus1 \leftarrow Surplus1 - \mathcal{B}_j,\ Loss1[j] = +\infty$
  else
  $S_j^{whs,UB} \leftarrow \left\lceil \frac{(z_j^{whs})^2\sigma_j^2 + 2\lambda_j(\mathcal{B}_j - Surpluss1) - z_j^{whs}\sigma\sqrt{(z_j^{whs})^2\sigma_j^2 + 4\lambda_j(\mathcal{B}_j - Surpluss1)}}{2\lambda_j^2} - \tau_j^{whs} - SI_j \right\rceil,$
  $Surplus1 \leftarrow 0$ **end if end while end if**
  for all $w \in \mathbb{W},\ j \in \mathbb{M}_w$ **do** $UB1 \leftarrow UB1 + f_{\mathbf{P}}\left(R_w^{\text{UB}},\ S_j^{whs,UB}\right)$ **end for**
  return $UB1$
end function

5.1.2 Updating the Lagrangian multiplier κ

The subgradient method is the common method used to solve the Lagrangian dual problem (see Fisher (1981, 1985) and Shor (2012)). This method converges for smooth and non-differentiable functionals at the rate of a geometric progression (Polyak, 1969; Held et al., 1974; Allen et al., 1987). We implemented the subgradient-based heavy ball method proposed by Polyak (1964) to accelerate the convergence by using the history of the last search direction. The direction of the search is defined based on the current gradient and the last search direction in equation (14), where $0 \leq \theta \leq 1$ is a constant parameter that defines how much memory the algorithm uses. When $\theta = 0$, the method reduces to the standard subgradient method. The step size of the heavy ball method is calculated using equation (15), and the Lagrangian multiplier is updated based on equation (16).

$$d^{iter1} = (1-\theta)G^{iter1} + \theta d^{iter1-1} \tag{14}$$

$$\mathcal{T}^{iter1} = \frac{UB - f_{\mathbf{NLR}^1_{\kappa^{iter1}}}}{(d^{iter1})^2} \tag{15}$$

$$\kappa^{iter1+1} = \kappa^{iter1} + \mathcal{T}^{iter1} d^{iter1} \tag{16}$$

5.2 Second phase: relaxing the budget constraint

To obtain a lower bound of the first phase relaxation problem $\mathbf{NLR}^1_\kappa$, the budget constraint is dualized in the second phase with a Lagrangian multiplier $\gamma \geq 0$. The formulation of the resulting relaxed problem $\mathbf{NLR}^2_{\kappa,\gamma}$ for a given γ is defined as follows:

$$\begin{aligned}\mathbf{NLR}^2_{\kappa,\gamma}\ NL^2(\kappa,\gamma) = \min \sum_{w\in\mathbb{W}}\sum_{j\in\mathbb{M}_w}\Bigg(& h_j^{wip}\lambda_j LT_w(R_w) + H_j(z_j^{whs})\sigma_j\sqrt{SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}} \\ &+ H_j(z_j^{ret})\sigma_j\sqrt{S_j^{whs} + \tau_j^{ret} - S_j^{ret}}\Bigg) + \kappa\Bigg(\sum_{w\in\mathbb{W}}\sum_{j\in\mathbb{M}_w}\lambda_j\left(SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}\right) \\ &+ z_j^{whs}\sigma_j\sqrt{SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}} - K\Bigg) + \gamma\left(\sum_{w\in\mathbb{W}} b_w R_w - B\right) \qquad (17)\end{aligned}$$

subject to: constraints (3), (5), (6), (7), (8) and (9).

Problem $\mathbf{NLR}^2_{\kappa,\gamma}$ can be decomposed into $|\mathbb{W}|$ subproblems $\mathbf{D}-\mathbf{NLR}^2_{\kappa,\gamma,w}$, each corresponding to a single workcenter w. Note that the terms γB and κK in the objective function of $\mathbf{NLR}^2_{\kappa,\gamma}$ for a given γ and κ are constant, and are dropped from the objective function of subproblems $\mathbf{D}-\mathbf{NLR}^2_{\kappa,\gamma,w}$, which is expressed as

$$\begin{aligned}\mathbf{D}-\mathbf{NLR}^2_{\kappa,\gamma,w}\ \min \sum_{j\in\mathbb{M}_w}\Bigg(& h_j^{wip}\lambda_j LT_w(R_w) + H_j(z_j^{whs})\sigma_j\sqrt{SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}} \\ &+ H_j(z_j^{ret})\sigma_j\sqrt{S_j^{whs} + \tau_j^{ret} - S_j^{ret}}\Bigg) + \kappa\Bigg(\sum_{j\in\mathbb{M}_w}\lambda_j\left(SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}\right) \\ &+ z_j^{whs}\sigma_j\sqrt{SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}}\Bigg) + \gamma b_w R_w \qquad (18)\end{aligned}$$

$$\text{subject to:} \quad \sum_{j\in\mathbb{M}_w}\lambda_j < R_w \qquad (19)$$

$$\max(0, S_j^{ret} - \tau_j^{ret}) \leq S_j^{whs} \leq SI_j + LT_w(R_w) + \tau_j^{whs} \qquad \forall j \in \mathbb{M}_w \qquad (20)$$

$$si_j \leq SI_j \qquad \forall j \in \mathbb{M}_w \qquad (21)$$

$$S_j^{ret} \leq s_j \qquad \forall j \in \mathbb{M}_w \qquad (22)$$

$$S_j^{whs}, S_j^{ret} \in \mathbb{Z}^+ \qquad \forall j \in \mathbb{M}_w \qquad (23)$$

$$LT_w(R_w) \in \mathbb{Z}^+ \qquad (24)$$

For a fixed level of capacity R_w at the workcenter, problem $\mathbf{D}-\mathbf{NLR}^2_{\kappa,\gamma,w}$ can be decomposed into $|\mathbb{M}_w|$ subproblems $\mathbf{D}-\mathbf{Prod}^2_{\kappa,\gamma,w,j}$, each corresponding to a single-workcenter and a single-product. The term $\gamma b_w R_w$ in the objective of $\mathbf{D}-\mathbf{NLR}^2_{\kappa,\gamma,w}$ for a given R_w is constant and is dropped from the objective of subproblems $\mathbf{D}-\mathbf{Prod}^2_{\kappa,\gamma,w,j}$ defined as

$$\mathbf{D-Prod}^2_{\kappa,\gamma,w,j} \quad \min \quad h_j^{wip}\lambda_j LT_w(R_w) + H_j(z_j^{whs})\sigma_j\sqrt{SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}}$$
$$+ H_j(z_j^{ret})\sigma_j\sqrt{S_j^{whs} + \tau_j^{ret} - S_j^{ret}} + \kappa\lambda_j\left(SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}\right)$$
$$+ \kappa z_j^{whs}\sigma_j\sqrt{SI_j + LT_w(R_w) + \tau_j^{whs} - S_j^{whs}} \tag{25}$$

$$\text{subject to:} \quad \max(0, S_j^{ret} - \tau_j^{ret}) \leq S_j^{whs} \leq SI_j + LT_w(R_w) + \tau_j^{whs} \tag{26}$$

$$si_j \leq SI_j \tag{27}$$

$$S_j^{ret} \leq s_j \tag{28}$$

$$S_j^{whs}, S_j^{ret} \in \mathbb{Z}^+ \tag{29}$$

The objective function of problem $\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}$ is concave in S_j^{whs} when $H_j(z_j^{whs}) + \kappa z_j^{whs} \geq 0$. Therefore, for a given capacity level R_w, when $\kappa \geq \frac{-H_j(z_j^{whs})}{z_j^{whs}}$, the optimal outgoing service time at the warehouse lies at one of two extreme points, i.e. $S_j^{whs,*} = \max(0, S_j^{ret} - \tau_j^{ret})$ or $S_j^{whs,*} = SI_j + LT_w(R_w) + \tau_j^{whs}$ (Lines 3 to 6). When $\kappa < \frac{-H_j(z_j^{whs})}{z_j^{whs}}$, we search all possible integer values of the outgoing service time at the warehouse and compute the objective value $f_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}}$. This full search determines the best outgoing service time at the warehouse with the lowest objective value $f^*_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}}$ (Lines 9 to 12). Algorithm 2 shows the procedure for solving problem $\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}$ for a given capacity level R_w.

Algorithm 2 Procedure for solving $\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}$.

1: **function** SOLVE $\mathbf{D-Prod}^2(\kappa, \gamma, w, j, R_w)$
2: $S_j^{1,whs} \leftarrow \max(0, S_j^{ret} - \tau_j^{ret})$, $S_j^{2,whs} \leftarrow SI_j + LT_w(R_w) + \tau_j^{whs}$, $f^*_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}} \leftarrow +\infty$
3: **if** $f_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}}\left(S_j^{1,whs}\right) < f_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}}\left(S_j^{2,whs}\right)$ **then**
4: $S_j^{whs,*} \leftarrow S_j^{1,whs}$, $f^*_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}} \leftarrow f_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}}\left(S_j^{1,whs}\right)$
5: **else**
6: $S_j^{whs,*} \leftarrow S_j^{2,whs}$, $f^*_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}} \leftarrow f_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}}\left(S_j^{2,whs}\right)$ **end if**
7: **if** $\kappa < \frac{-H_j(z_j^{whs})}{z_j^{whs}}$ **then**
8: $S_j^{whs} \leftarrow S_j^{1,whs}$
9: **while** $S_j^{whs} \leq S_j^{2,whs}$ **do**
10: **if** $f_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}}\left(S_j^{whs}\right) < f^*_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}}$ **then**
11: $S_j^{whs,*} \leftarrow S_j^{whs}$, $f^*_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}} \leftarrow f_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}}\left(S_j^{whs}\right)$ **end if**
12: $S_j^{whs} \leftarrow S_j^{whs} + 1$ **end while end if**
13: **return** $f^*_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}}$
14: **end function**

To solve $\mathbf{D-NLR}^2_{\kappa,\gamma,w}$, we search all possible values of R_w that result in integer lead times, and we compute the corresponding objective value $f_{\mathbf{D-NLR}^2_{\kappa,\gamma,w}}$. This full search determines the best capacity and lowest cost $f^*_{\mathbf{D-NLR}^2_{\kappa,\gamma,w}}$. The steps employed to solve $\mathbf{NLR}^2_{\kappa,\gamma}$ are summarized in Algorithm 3.

Algorithm 3 Procedure for solving $\mathbf{NLR}^2_{\kappa,\gamma}$.

1: **function** SOLVE $\mathbf{NLR}^2(\kappa,\gamma)$
2: $f^*_{\mathbf{NLR}^2_{\kappa,\gamma}} \leftarrow 0$
3: **for all** $w \in \mathbb{W}$ **do**
4: $f^*_{\mathbf{D-NLR}^2_{\kappa,\gamma,w}} \leftarrow +\infty,\ R_w \leftarrow \sum_{j\in\mathbb{M}_w} \lambda_j$
5: **while** $R_w \leq \sum_{j\in\mathbb{M}_w} \lambda_j + \frac{B-\sum_{i\in\mathbb{W}}\sum_{j\in\mathbb{M}_w} b_i\lambda_j}{b_w}$ **do**
6: $f_{\mathbf{D-NLR}^2_{\kappa,\gamma,w}} \leftarrow 0$
7: **for all** $j \in \mathbb{M}_w$ **do**
8: $f_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}} \leftarrow$ SOLVE $\mathbf{D-Prod}^2(\kappa,\gamma,w,j,R_w)$ ▷ Using Algorithm 2
9: $f_{\mathbf{D-NLR}^2_{\kappa,\gamma,w}} \leftarrow f_{\mathbf{D-NLR}^2_{\kappa,\gamma,w}} + f_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}}$ **end for**
10: $f_{\mathbf{D-NLR}^2_{\kappa,\gamma,w}} \leftarrow f_{\mathbf{D-NLR}^2_{\kappa,\gamma,w}} + \gamma\left(b_w R_w - \frac{B}{|\mathbb{W}|}\right) - \kappa K$
11: **if** $f_{\mathbf{D-NLR}^2_{\kappa,\gamma,w}} \leq f^*_{\mathbf{D-NLR}^2_{\kappa,\gamma,w}}$ **then** $f^*_{\mathbf{D-NLR}^2_{\kappa,\gamma,w}} \leftarrow f_{\mathbf{D-NLR}^2_{\kappa,\gamma,w}}$ **end if**
12: $R_w \leftarrow$ FINDNEXTR **end while** ▷ Find next R_w that results in integer lead time
13: $f^*_{\mathbf{NLR}^2_{\kappa,\gamma}} \leftarrow f^*_{\mathbf{NLR}^2_{\kappa,\gamma}} + f^*_{\mathbf{D-NLR}^2_{\kappa,\gamma,w}}$ **end for**
14: **return** $f^*_{\mathbf{NLR}^2_{\kappa,\gamma}}$
15: **end function**

5.2.1 Generating upper bounds for problem $\mathbf{NLR}^1_\kappa$

To generate upper bounds on the objective value of the first phase relaxation problem $\mathbf{NLR}^1_\kappa$, a greedy heuristic is used in each iteration of the second phase. After solving $\mathbf{NLR}^2_{\kappa,\gamma}$, we check whether the obtained lower bound solution satisfies the budget constraint (2). If this is the case, the lower bound solution is feasible for $\mathbf{NLR}^1_\kappa$, and the corresponding objective value is an upper bound. When $\sum_{w\in\mathbb{W}} b_w R_w > B$, i.e. the lower bound solution is not feasible, a greedy heuristic is used to repair the solution by eliminating the surplus of the capacity budget, denoted by $Surplus2$. The output of this heuristic is a repaired, feasible solution for which the objective value is an upper bound on problem $\mathbf{NLR}^1_\kappa$.

In each iteration of the greedy heuristic, $Surplus2$ is reduced by a predefined capacity budget reduction, which we refer to as *Step*. Subtracting *Step* from the capacity budget of a workcenter reduces its capacity level by $\frac{Step}{b_w}$ and increases the objective value of $\mathbf{NLR}^1_\kappa$ by $Loss2$ (Line 9). We start from the workcenter with the lowest value of $Loss2$ (Line 10), reduce its capacity by $\frac{Step}{b_w}$, and update $Surplus2$ (Line 11). This process is repeated until the budget constraint (2) is satisfied. Once the obtained capacities become feasible, we need to determine the corresponding optimal service times. When $\kappa \geq \frac{-H_j(z_j^{whs})}{z_j^{whs}}$, problem $\mathbf{NLR}^1_\kappa$ is concave in S_j^{whs}, and the optimal outgoing service time at the warehouse lies at one of two extreme points: $S_j^{whs,*} = \max(0, S_j^{ret} - \tau_j^{ret})$ or $S_j^{whs,*} = SI_j + LT_w(R_w) + \tau_j^{whs}$. When $\kappa < \frac{-H_j(z_j^{whs})}{z_j^{whs}}$, we search over all possible values of S_j^{whs} and compute the objective value of $\mathbf{NLR}^1_\kappa$. Algorithm 4 lists the steps of the greedy heuristic to generate upper bounds for $\mathbf{NLR}^1_\kappa$.

Algorithm 4 Greedy heuristic to generate upper bounds for $\mathbf{NLR}^1_\kappa$.

function UpperBound2(γ)
  $UB2 \leftarrow 0$
  if $\sum_{w\in\mathbb{W}} b_w R_w \leq B$ **then**
  for all $w \in \mathbb{W}$ **do** $R_w^{\text{UB}} \leftarrow R_w$ **end for** ▷ R_w is the solution of problem $\mathbf{NLR}^2_{\kappa,\gamma}$
  else
  $Surplus2 \leftarrow \sum_{w\in\mathbb{W}} b_w R_w - B$, $Step \leftarrow \frac{Surplus2}{\sum_{w\in\mathbb{W}} b_w}$
  while $0 < Surplus2$ **do**
  if $Surplus2 - Step < 0$ **then** $Step \leftarrow Surplus2$ **end if**
  for all $w \in \mathbb{W}$ **do** $Loss2[w] = f_{\mathbf{P}}\left(\max\left(\sum_{j\in\mathbb{M}_w} \lambda_j, R_w - \frac{Step}{b_w}\right)\right) - f_{\mathbf{P}}(R_w)$ **end for**
  FindWorkcenterWithLowestValueIn($Loss2$)
  $Surplus \leftarrow Surplus - b_w\left(R_w - \max(\sum_{j\in\mathbb{M}_w} \lambda_j, R_w - \frac{Step}{b_w})\right)$
  $R_w^{\text{UB}} \leftarrow R_w$, $R_w \leftarrow \max\left(\sum_{j\in\mathbb{M}_w} \lambda_j, R_w - \frac{Step}{b_w}\right)$ **end while end if**
  for all $w \in \mathbb{W}$, $j \in \mathbb{M}_w$ **do**
  $f_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}} \leftarrow$ Solve $\mathbf{D-Prod}^2(\kappa, \gamma, w, j, R_w^{\text{UB}})$ ▷ Using Algorithm 2
  $UB2 \leftarrow UB2 + f_{\mathbf{D-Prod}^2_{\kappa,\gamma,w,j}}$ **end for**
  return $UB2$
end function

5.2.2 Updating the Lagrangian multiplier γ

The subgradient-based heavy ball method, presented in Section 5.1.2, is used to update the Lagrangian multiplier γ. The current gradient is $\sum_{w\in\mathbb{W}} b_w R_w - B$ and the direction of the search is defined based on equation (30), where $0 \leq \theta \leq 1$. The step size of the heavy ball method is calculated using equation (31), and the Lagrangian multiplier is updated based on equation (32).

$$d^{iter2} = (1-\theta)G^{iter2} + \theta d^{iter2-1} \tag{30}$$

$$\mathcal{T}^{iter2} = \frac{UB - f_{\mathbf{NLR}^2_{\kappa^{iter1}\gamma^{iter2}}}}{(d^{iter2})^2} \tag{31}$$

$$\gamma^{iter2+1} = \max\left(\gamma^{iter2} + \mathcal{T}^{iter2} d^{iter2}, 0\right) \tag{32}$$

6 A simulation study

This section presents a simulation study to evaluate the accuracy of the mathematical model and to compare two approaches for setting capacity and safety stocks in order to assess the value of the integration. The simulation study includes three steps. In the first step, production lead times are estimated using simulation-based cycle time-throughput curves. Based on the obtained approximate lead time functions $LT_w(R_w)$, the second step sets capacities and service times. In the third step, the obtained solutions are taken as inputs of a discrete-event simulation model for the production-distribution system described in Section 3.2. This last step evaluates the quality of the solutions and the accuracy of the model with respect to several performance criteria.

6.1 Simulation steps

6.1.1 Production lead time estimation

Each workcenter w sets a guaranteed, deterministic lead time LT_w that is dependent on its capacity R_w. This lead time is the elapsed time between the moment an order of a period is received by the workcenter until it is completed and shipped to the warehouse.

We set LT_w to include two parts, the first part is the average time an item spends in the raw material staging area plus a second part, which is a fixed time to be spent by an item at the workstation and the finished good staging area. The second part of the lead time, i.e. the fixed time, is determined such that a target fraction α_w^{LT} of processed items are completed and ready to be shipped within this fixed time. Units that are delayed, i.e. units that are not available at the finished good staging area to be able to ship a complete order to the warehouse, are expedited from the queue of the workstation. Expediting can be implemented using special measures such as overtime (Çelik and Maglaras, 2008; Plambeck and Ward, 2007, 2008). Inderfurth (1993) and Minner (2000) also used a similar approach to define lead times in GSA models. In this manner, the complete order of the warehouse is satisfied within the guaranteed service time of the workcenter, i.e. $SI_j + LT_w$ for product $j \in \mathbb{M}_w$. In this case, $(1 - \alpha_w^{LT})$ represents the percentage of time the manufacturer resorts to expediting measures at workcenter w, and the expected number of delayed items per period is $(1 - \alpha_w^{LT}) \sum_{j \in \mathbb{M}_w} \lambda_j$.

Specifically, at the beginning of each period, orders are placed in the raw material staging area. Subsequently, units of the various items are released and processed on a FCFS basis, according to the sequence of arrival at the retailer. On average, orders spend 0.5 periods in the raw material staging area. The time that an item spends in the workstation (queuing plus processing) is the production cycle time of workcenter w, PCT_w which is a random variable. The fixed time t_w, which is the second part of lead time, is defined such that $\mathbb{P}(PCT_w \leq t_w) \geq \alpha_w^{LT}$.

To estimate the fixed time t_w and hence production lead times as a function of the allocated capacity R_w, we use simulation-based cycle time-throughput curves based on the work reported by Yang et al. (2008). We start by estimating the α_w^{LT}-percentile of PCT_w by fitting curves to the first three moments of PCT_w, and subsequently matching a generalized gamma distribution function to these three moments. Afterwards, a nonlinear regression model can be fitted on the obtained percentiles to estimate the fixed time as a function of R_w at a given α_w^{LT}. This procedure is described in Appendix C.

6.1.2 Two approaches for setting capacity and safety stocks

Two methods that are employed to set capacity and safety stocks in the production-distribution system are considered, namely a sequential approach and an integrated approach. The sequential approach considers the relationship between the capacity, cycle time, and WIP, while setting the capacity, but it ignores the effect of the capacity on safety stocks. The integrated approach jointly optimizes capacity and safety stocks.

Sequential approach (Seq). In this approach, the capacity planning problem is solved first to set capacities of workcenters R_w, and subsequently, the safety stock placement problem is solved to obtain optimal service times and safety stock levels. The capacity planning problem determines the processing rate for each workcenter while considering the manufacturer's budget and WIP inventory cost. Based on Bretthauer (1995) and Bretthauer and Côté (1997), the capacity planning problem (**CAP**) can be formulated as follows:

$$\textbf{CAP} \quad \min_{R} \sum_{w \in \mathbb{W}} \sum_{j \in \mathbb{M}_w} h_j^{wip} \lambda_j LT_w(R_w) \tag{33}$$

$$\text{subject to:} \quad \text{constraints (2), (3) and (9).}$$

Once problem **CAP** is solved, production lead times $LT_w(R_w)$ are taken as inputs to the safety stock placement problem (**SSP**), which is solved using Algorithm 1 to determine safety stocks. The safety stock placement problem **SSP** is given by

$$\textbf{SSP} \quad \min_{SI,\, S} \; H_j(z_j^{whs})\sigma_j \sqrt{SI_j + LT_w + \tau_j^{whs} - S_j^{whs}} + H_j(z_j^{ret})\sigma_j \sqrt{S_j^{whs} + \tau_j^{ret} - S_j^{ret}} \tag{34}$$

$$\text{subject to:} \quad \text{constraints (4), (5), (6), (7) and (8).}$$

Integrated approach (Int) The integrated approach jointly optimizes capacity and safety stocks. The integrated problem is formulated in Section 3 and solved using the nested Lagrangian relaxation heuristic presented in Section 5. The improvement obtained through integration is measured using the value of integration (VOI)%, which is defined as the savings obtained using the integrated approach relative to the other approaches or

$$\text{VOI}\% = \frac{f^{\textbf{Seq}} - f^{\textbf{Int}}}{f^{\textbf{Seq}}} \times 100.$$

Where $f^{\textbf{Seq}}$ is the total cost of the sequential approach, and $f^{\textbf{Int}}$ is the total cost of the integrated approach.

6.1.3 Evaluation

Once capacity and safety stocks (or base stock levels) are set using one of the above two approaches, they are taken as inputs for a discrete-event simulation model. To evaluate the quality of the solutions, simulation experiments were conducted, and statistics on various performance measures are calculated. Performance measures include: the fraction of on-time completion at workcenters (α_w^{LT}), service levels at the warehouse and retailer $\left(\alpha_j^{whs}, \alpha_j^{ret}\right)$, the WIP level at workcenters, expediting (backorders) at warehouse and retailers (BO_j^{whs}, BO_j^{ret}), on-hand inventory at the warehouse and retailers (OH_j^{whs}, OH_j^{ret}), and the total cost of the system.

6.2 Simulation experiments

The simulation steps described in the previous section were applied to a production-distribution system example. The simulation model was implemented using FlexSim simulation software version 7.0.

Production-distribution system example. We consider an example of a production-distribution system, where the production part is motivated by the semiconductor manufacturing process at IBM, presented in Woerner et al. (2018). The semiconductor manufacturing system consists of two assembly lines. Each assembly line processes a single product, and includes five production steps. In this paper, we assume that all five production steps are done in one workcenter. Therefore, the production-distribution system consists of one manufacturer with two workcenters, one warehouse, and one retailer, as depicted in Figure 11. Workcenter 1 processes product 1, while workcenter 2 produces product 2. Demand for products 1 and 2 follow a normal distribution with means 100 and 1000, and standard deviation values of 200 and 10 units per period, respectively. Product 1 has a lower demand rate and higher coefficient of variation (COV=2) in comparison to product 2 (COV=0.01). The duration of each review period is 1440 min or one day.

Product 1 is assumed to be more expensive to hold and expedite through the supply chain compared with product 2. Based on Woerner et al. (2018), to set the WIP cost, we take the average WIP cost of the five production steps in each assembly line. Therefore, the WIP costs are $h_j^{wip} = \{26, 2.6\}$ Euro per unit per period. We define the value addition at the warehouse and retailer as $\beta_j^{whs} = \frac{h_j^{0,whs}}{h_j^{wip}}$ and $\beta_j^{ret} = \frac{h_j^{0,ret}}{h_j^{0,whs}}$, respectively, and set $\beta_j^{whs} = \beta_j^{ret} = 1.2$. Based on Woerner et al. (2018), the capacity unit cost at workcenters is set to $b_w = \{0.1, 1\}$ in Euro per unit per period. The total available budget at the manufacturer is $B = 1030.6$ Euro, which is obtained based on an average utilization of 98% at the manufacturer. Processing times at both workcenters follow a log-normal distribution function with a squared coefficient of variation $C_s^2 = 2$.

The service factor for both products at the warehouse and retailer are $z_j^{whs} = 1.281$ and $z_j^{ret} = 1.96$, respectively, which corresponds to a 90% and 97.5% service level. Unit expediting costs are computed based on Aouam and Kumar (2019). The logistics delays τ_j^{whs} and τ_j^{ret} are set equal to two periods. We also assume that costumers are willing to wait for their items. Accordingly, the maximum outgoing service times at the retailer are $s_1 = s_2 = 1$ period. In addition, we set the minimum incoming service times at the manufacturer to zero, i.e. $si_1 = si_2 = 0$. In the base case instance, we consider four different values for the storage capacity, $K \in \{\infty,\ 23000,\ 20000,\ 1000\}$.

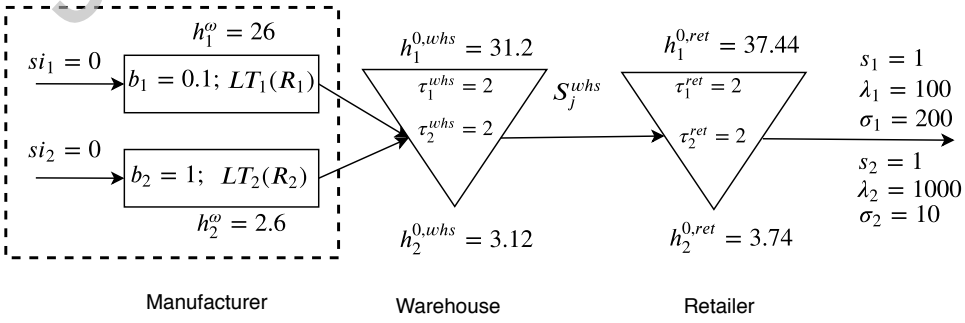

Figure 11: Supply chain network of the base case instance.

Lead time estimation. Production lead times of workcenters 1 and 2 are estimated based on Yang

et al. (2008). The average time that an item spends in the raw material staging area is calculated and added to the fixed time, as explained in section 6.1.1. For each workcenter, we consider 500 capacity levels and 100 simulation runs for each capacity level, where each run consists of 20,000 periods. Inter-release times follow log-normal distributions with mean $\frac{1}{\lambda_j}$ and coefficient of variation equal to one. Once an order (a period demand) arrives as a batch at the staging area, the average time an item waits in batch is 0.5 periods. Based on Yang et al. (2008) and for α_w^{LT} =95%, the lead time functions for workcenters 1 and 2 are estimated as follows:

$$LT_1(R_1) = 0.5 + \frac{4.685 - 23.763\left(\frac{100}{R_1}\right) + 47.683\left(\frac{100}{R_1}\right)^2 - 42.25\left(\frac{100}{R_1}\right)^3 + 13.645\left(\frac{100}{R_1}\right)^4}{\left(1 - \frac{100}{R_1}\right)^{2.017}} \tag{35}$$

$$LT_2(R_2) = 0.5 + \frac{2.459 - 9.676\left(\frac{1000}{R_2}\right) + 15.894\left(\frac{1000}{R_2}\right)^2 - 10.306\left(\frac{1000}{R_2}\right)^3 + 1.928\left(\frac{1000}{R_2}\right)^4}{\left(1 - \frac{1000}{R_2}\right)^{1.007}} \tag{36}$$

Solutions of the two approaches and effect of storage capacity

Given the above estimated lead times, the integrated and sequential approaches presented in Section 6.1.2 determine the capacity levels and service times. Table 4 shows the solutions of these two approaches for the base case with $K \in \{\infty,\ 23000,\ 20000,\ 1000\}$. From Table 4, one can observe that the two approaches result in different replenishment lead times and safety stocks.

Table 4: Integrated and sequential solutions for the base case instance.

Storage capacity		Integrated approach				Sequential approach			
K (in units)		RLT	SS^{whs}	SS^{ret}	Type	RLT	SS^{whs}	SS^{ret}	Type
∞	Product 1	5	572.88	392.00	Decoupling	7	677.84	392.00	Decoupling
	Product 2	23	61.43	19.60	Decoupling	21	58.70	19.60	Decoupling
	TC	104736.76 (VOI=4.10%)				109210.40			
23000	Product 1	5	572.88	392.00	Decoupling	7	677.84	392.00	Decoupling
	Product 2	23	58.70	33.95	Partial decoupling	21	58.70	19.60	Decoupling
	TC	104789.17 (VOI=4.05%)				109210.40			
20000	Product 1	5	572.88	392.00	Decoupling	7	677.84	392.00	Decoupling
	Product 2	23	0.00	96.02	Coupling	21	54.35	39.20	Partial decoupling
	TC	104815.51 (VOI=4.08%)				109279.32			
1000	Product 1	5	0.00	960.20	Coupling	7	0.00	1108.74	Coupling
	Product 2	23	0.00	96.02	Coupling	21	0.00	91.93	Coupling
	TC	105708.85 (VOI=5.89%)				112324.16			

From Table 4, the replenishment lead time at the warehouse $RLT = (7,\ 21)$ periods in the sequential approach. When safety stock placement is taken into consideration in the integrated approach, capacity allocation changes and $RLT = (5,\ 23)$ periods. The different capacity allocation results in different safety stock levels. The sequential approach first sets the capacity and subsequently optimizes safety stocks. The approach considers the differences in the WIP holding costs, capacity unit costs, and demand rates between products, and allocates more capacity to product 1. This results in $RLT_1 = 7$ periods for product 1 that is much shorter than $RLT_1 = 21$ periods for product 2. When safety stock placement is taken into consideration in the integrated approach, RLT for product 1 is further reduced

to $RLT_1 = 5$ periods because product 1 is expensive to hold and has more demand variability. The integrated approach minimizes production and distribution costs simultaneously, and this explains the cost savings that are realized. Therefore, the integrated approach allocates capacity more efficiently, which leads to a cost reduction (value of integration) relative to the sequential approach of $VOI = 4.1\%$.

Table 4 shows that RLT is not affected by the warehouse storage capacity K in both approaches. This means that for any value of K, it is more cost-efficient to adjust safety stock placement than to change capacity allocation. When storage capacity is reduced from infinity to $K = 23000$ units, the optimal safety stock placement remains unchanged and inventory is decoupled for both products in the sequential approach. In the integrated approach, inventory is decoupled for product 1, while inventory is partially decoupled for product 2. When storage capacity is further reduced to $K = 20000$ units, the integrated approach couples inventory of product 2 at the retailer while the sequential approach partially decouples inventory of product 2. When storage capacity is very small $K = 1000$ units, both approaches place inventory at the retailer for both products. Therefore, storage capacity at the warehouse seems to impact safety stock placement and not capacity allocation in both approaches. Storage capacity has more effect on safety stock placement in the integrated approach. As storage capacity becomes smaller, inventory is pushed towards the retailer.

Validation of the mathematical model and evaluation

To validate the mathematical model and evaluate the solutions of the integrated approach, performance measures are computed using simulation experiments, and the results are compared with those of the mathematical model. Based on 100 simulation runs, the mean and 95% confidence interval (CI) of the performance measures are estimated. The simulation and mathematical model results are reported in Tables 5 and 6, respectively.

Table 5: Simulation results for the integrated approach.

Storage capacity (units)		Performance measures based on the simulation							
		α^{LT}	WIP	α^{whs}	α^{ret}	BO^{whs}	BO^{ret}	OH^{whs}	OH^{ret}
∞	Product 1	95.11%±0.15	294.76±0.44	90.01%±0.09	97.51%±0.05	21.14±0.07	1.89±0.007	595.98±2.73	395.44±1.67
	Product 2	95.09%±0.19	20913.34±53.42	90.02%±0.05	97.50%±0.08	2.27±0.01	0.09±0.002	63.91±0.27	19.77±0.08
	TC	104432.67±326.08							
23000	Product 1	95.14%±0.22	294.68±0.47	90.00%±0.09	97.51%±0.07	21.15±0.08	1.89±0.006	596.16±2.67	395.35±1.61
	Product 2	95.15%±0.21	20910.20±56.56	90.05%±0.06	97.53%±0.04	2.17±0.01	0.16±0.003	60.98±0.19	34.21±0.11
	TC	104471.35±336.87							
20000	Product 1	95.16%±0.27	294.34±0.84	90.02%±0.10	97.53%±0.11	21.13±0.10	1.88±0.006	595.83±2.58	395.33±1.83
	Product 2	95.11%±0.23	20907.05±59.71	-	97.52%±0.08	-	0.46±0.004	-	96.81±0.37
	TC	104455.28±366.22							
1000	Product 1	95.16%±0.27	294.72±0.46	-	97.51%±0.03	-	4.63±0.008	-	968.25±3.91
	Product 2	95.14%±0.19	20932.19±26.19	-	97.52%±0.09	-	0.46±0.002	-	96.76±0.32
	TC	105468.07±246.81							

Table 5 indicates that the fraction of on-time completion α^{LT} in all three cases is always higher than the target $\alpha_w^{LT} = 95\%$ in the mathematical model. This means that the production lead time estimation is conservative, leading to fewer delayed items at workcenters. In addition, the average cycle

service levels α^{whs} and α^{ret} are very close to but slightly higher than their target values. The reason is that to set the base stock level in the simulation model, we round up the base stock level obtained from the mathematical model. Therefore, in the simulation model, we have a higher base stock level, which results in higher service level. When a stage does not carry inventory, i.e. the net replenishment lead time is equal to zero, we do not report the service level.

Table 6: Results of the mathematical model for the integrated approach.

Storage capacity		Performance measures based on the mathematical models							
(units)		α^{LT}	WIP	α^{whs}	α^{ret}	BO^{whs}	BO^{ret}	OH^{whs}	OH^{ret}
∞	Product 1	95%	295	90%	97.5%	21.20	1.89	594.08	393.89
	Product 2	95%	20950	90%	97.5%	2.27	0.09	63.71	19.69
	TC	104736.76							
23000	Product 1	95%	295	90%	97.5%	21.20	1.89	594.08	393.89
	Product 2	95%	20950	90%	97.5%	2.17	0.16	60.87	34.11
	TC	104789.17							
20000	Product 1	95%	295	90%	97.5%	21.20	1.89	594.08	393.89
	Product 2	95%	20950	-	97.5%	-	0.46	-	96.48
	TC	104815.51							
1000	Product 1	95%	295	-	97.5%	-	4.63	-	964.83
	Product 2	95%	20950	-	97.5%	-	0.46	-	96.48
	TC	105708.85							

In addition, when comparing both Tables 5 and 6 we observe that backorders (BO^{whs} and BO^{ret}) as well as on-hand inventories (OH^{whs} and OH^{ret}) of the mathematical model are close to those of the simulation and within the 95% CI. Average WIP levels in the simulation are slightly lower compared with those of the mathematical model owing to the conservative production lead time estimation. Furthermore, for all three cases, the total cost of the system in the mathematical model falls within the 95% CI estimated through simulation. The above results show that the approximations used in the mathematical model are accurate.

Effect of different parameters on the VOI%

In the following, we study the effect of different parameters on the VOI%. In each experiment, the parameter of product 1 is changed while the parameter of product 2 remains constant. The warehouse storage capacity is set to $K = 20000$ units in the base case.

Coefficient of variation of the demand. Figure 12 plots the total cost of the system as a function of the demand coefficient of variation for product 1, i.e. COV_1. This figure shows that as COV_1 increases, the total cost of the system for both integrated and sequential approaches increases, while the total cost of the integrated approach is always lower. In fact, as the demand variability increases, it is necessary for more safety stocks to be kept on the distribution side in order to hedge against uncertainty. This figure also displays that the total cost of the system is very close to the total cost of the simulation. Figure 13 shows that as COV_1 increases, the VOI% increases. In fact, the sequential approach does not consider the effect of demand variability in the capacity allocation phase, while the integrated approach is able to jointly optimize capacity and safety stocks, which results in savings up to 5.06%.

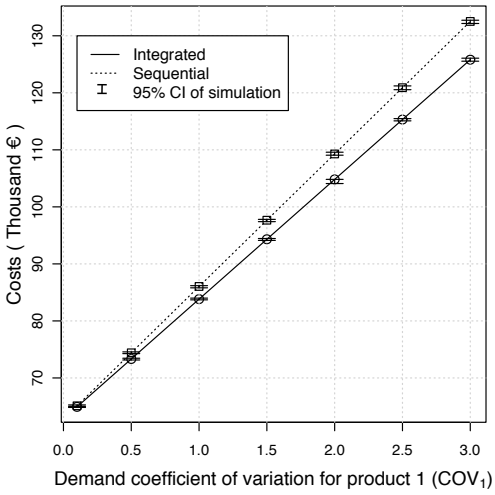

Figure 12: Impact of demand variability (COV_1) on total cost of the system.

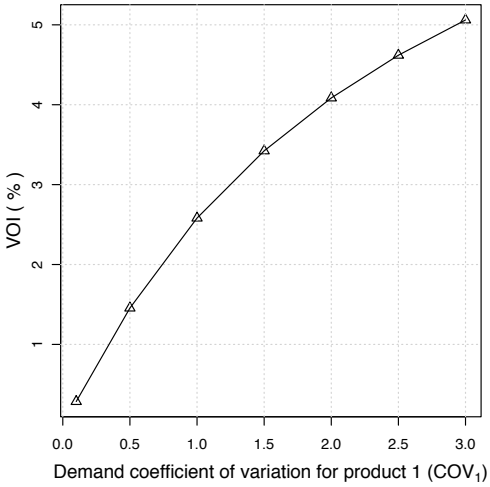

Figure 13: Impact of demand variability (COV_1) on VOI%.

External outgoing service time s. Figure 14 illustrates the effect of the external outgoing service time s_1 of product 1 on the total cost of the system for both integrated and sequential approaches. The total cost in both approaches decreases with s_1 because constraints (7) are less restrictive. However, the total cost in the integrated approach is always lower, and hence, there is a positive VOI% up to 19.73%, as shown in Figure 15. This figure also shows that there is a threshold on s_1, $s_1^0 = 9$ periods, above which the total cost of the integrated and sequential approaches is the same. This threshold is reached when the lead time of product 1 takes its maximum value $LT_1^{max} = 5$ periods in our example, and the threshold can be computed as $s_1^0 = LT_1^{max} + \tau_1^{whs} + \tau_1^{ret} = 9$ periods. Furthermore, we observe that the VOI% is particularly high for $s_1 = 6, 7$ and 8 periods and reaches its maximum value at $s_1 = 7$ periods. In fact, the integrated approach reallocates capacity to increase RLT_1 when s_1 increases, while the capacity allocation remains unchanged in the sequential approach when s_1 is increased. In the integrated approach, the retailer and warehouse do not carry inventory of product 1 when $s_1 \geq 6$, i.e. product 1 is made to order. This makes sense since product 1 is expensive to hold in inventory and the customer is willing to wait. The same solution is optimal in the case of the sequential approach, but only when $s_1 \geq 9$. The difference in the type of safety stock placement solutions between the two approaches explains the high VOI%.

Value addition at the warehouse (β^{whs}) and at the retailer (β^{ret}). The total cost of the system for both integrated and sequential approaches increases with β_1^{whs} and β_1^{ret}. This can be explained by the increase in holding costs at the warehouse $h_1^{0,whs}$ and the retailer $h_1^{0,ret}$, which increase with β_1^{whs} and β_1^{ret}. This increase is captured by the integrated approach while allocating capacity to workcenters. As a result, the total cost of the integrated approach is always lower than the cost of the sequential approach. As displayed in Figure 16, the VOI% ranges from 4.08% to 5.01%. We also notice a major change in the VOI% at $\beta_1^{whs} = \beta_1^{ret} = 1.2$, which can be explained by the change in the type of safety stock placement solution in the integrated approach. While inventory of product 1 is decoupled for

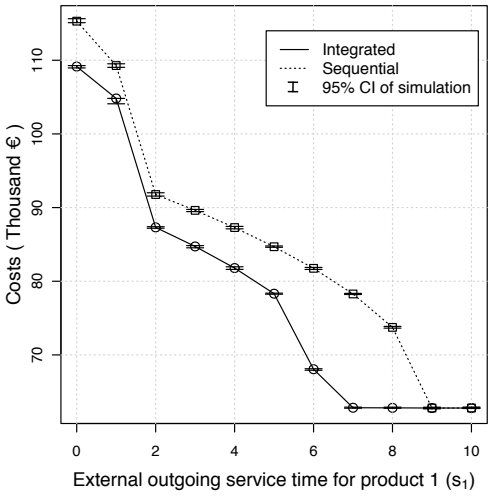

Figure 14: Impact of s_1 on total cost of the system.

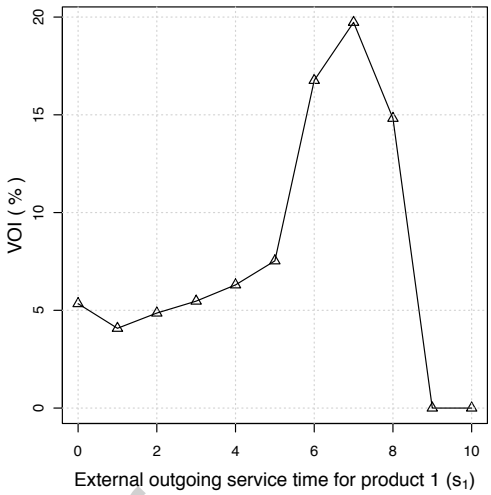

Figure 15: Impact of s_1 on VOI%.

all considered values of β_1^{whs} and β_1^{ret} in the sequential approach, it is coupled at the retailer when $\beta_1^{whs} = \beta_1^{ret} < 1.2$ and then decoupled for $\beta_1^{whs} = \beta_1^{ret} \geq 1.2$ in the integrated approach.

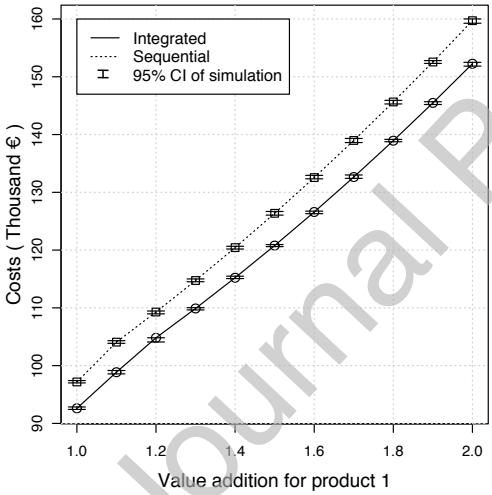

Figure 16: Impact of β_1^{whs} and β_1^{ret} on total cost of the system.

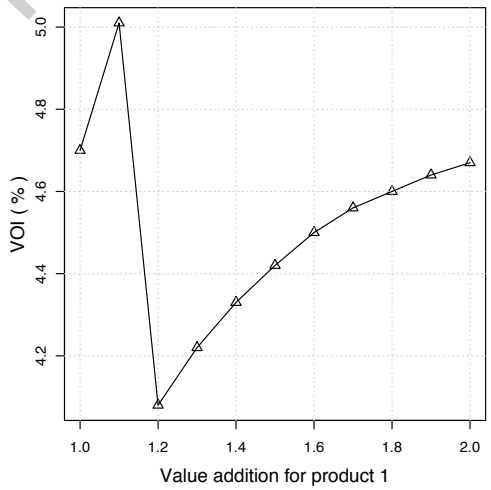

Figure 17: Impact of β_1^{whs} and β_1^{ret} on VOI%.

7 Performance of the nested Lagrangian relaxation heuristic

The performance of the proposed solution procedure, i.e. the nested Lagrangian relaxation heuristic presented in section 5, is assessed in this section. Solutions and CPU times of the proposed algorithm are compared to those of BARON, a standard solver for solving non-convex mathematical programming. The nested Lagrangian relaxation heuristic is coded in JAVA 8 and run on a 64-bit computer with 2.7 GHz Intel Core i5 processor and 8 GB of RAM under OSX 10.11.6. The integrated problem **P** was

coded in GAMS 24.7.4 and solved using BARON solver.

7.1 Instance generation

We consider three instance sets A (5 workcenters), B (10 workcenters), and C (30 workcenters). Each instance set consists of 10 randomly generated instances. Next, we describe how these instances are constructed.

The number of products processed by workcenter is uniformly generated between 1 and 3 for set A and between 10 and 20 for sets B and between 20 and 40 for set C. For all products, demand arrival is modeled as a Poisson process and for each instance mean demand is randomly generated between 100 and 1000 units per period. Lead times are estimated through simulation as describe in section 6.1.1. We assume log-normal processing times for workcenters in the simulation and the fraction of on-time completion is $\alpha_w^{LT} = 0.95$. The unit cost of capacity b_w for each workcenter is generated uniformly between 100 and 1000 Euro per unit per period and the capacity budget B is set randomly in a way that the average utilization to be between 0.85 and 0.95, i.e $\mathrm{B}=\frac{\sum_{w\in\mathbb{W}}\sum_{j\in\mathbb{M}_w}\lambda_j b_w}{85\%\text{ to }95\%}$. The manufacturer's WIP holding cost for each product (h_j^{wip}) is generated uniformly between 10 and 100 Euro.

The inventory holding costs at the warehouse ($h_j^{0,whs}$) are generated randomly to be on average between 1.05 and 2 times of the WIP holding cost at the manufacturers. Similarly, the inventory holding costs at the retailer ($h_j^{0,ret}$) are generated randomly to be on average between 1.1 and 1.4 times of the inventory holding cost at the warehouse. Safety factor at both warehouse and retailer is equal to 1.96 which corresponds to a 97.5% service level. The overtime costs and augmented inventory costs are set based on Aouam and Kumar (2019). The external incoming service times for products at the manufacturer (si_j) and the external outgoing service time at the retailer (s_j) are generated uniformly between 0 and 10 periods. The delays τ_j^{whs} and τ_j^{ret} are drawn uniformly between 1 and 20 periods. The storage capacity is set randomly as a function of the sum of base stock levels such that $K = (50\%\text{ to }150\%) \times \sum_{j\in\mathbb{M}} \left(\lambda_j \tau_j^{whs} + z_j^{whs}\sigma_j\sqrt{\tau_j^{whs}}\right)$.

7.2 Computational results

In the nested Lagrangian relaxation heuristic, the memory parameter of the heavy ball method is $\theta = 0.3$ and the initial values for all Lagrangian multipliers are taken to be zero. The algorithm stops when the optimality gap is less than 0.01%, the step size is less than 10^{-10}, or the number of iterations exceeds 1000. For large instances in set C, we add an additional stopping criterion, where the algorithm stops when the lower bound does not improve for five consecutive iterations.

To compare the quality of BARON's upper bound with the one obtained using the nested Lagrangian relaxation heuristic, we define relative gap GAP′% as follows:

$$\text{GAP}'\% = \frac{UB_{\mathbf{NLR}} - UB_{\mathbf{BARON}}}{UB_{\mathbf{BARON}}} \times 100.$$

Where $UB_{\mathbf{NLR}}$ and $UB_{\mathbf{BARON}}$ are the obtained upper bound of the nested Lagrangian relaxation heuristic and BARON solver, respectively. Negative values of GAP′% indicate that the nested La-

grangian relaxation heuristic finds a lower cost solution in comparison to BARON. Tables 7-9 report lower bound, upper bound, optimality gap GAP %, CPU time, number of iterations and relative gap GAP′% for all instances in sets A, B, and C. The second column in these tables refers to the total number of products in each instance.

From Table 7, we see that BARON solves all instances from set A to optimality within an average CPU time of 261 seconds. The average optimality gap of BARON increases from 0.0% in set A to 68.5% in set B in Table 8. In Table 9, we can see that except for instance number 6 for which the optimality gap is 98.4%, BARON does not find feasible solutions for instances in set C. We can conclude that as the number of workcenters and the number of products in the network increases, the performance of BARON greatly deteriorates.

Table 7: Results for set A with 5 workcenters

Ins.#	Prd.#	BARON				Nested Lagrangian relaxation					Gap′ %
		LB	UB	Gap %	Time (s)	LB	UB	Gap %	Time(s)	Iter	
1	9	11776874.0	11776874.0	0.00	288	11769152.2	11776874.0	0.07	1.2	38	0.00
2	9	7285517.6	7285517.6	0.00	62	7283156.9	7285517.6	0.03	0.7	40	0.00
3	10	7731110.4	7731110.4	0.00	264	7729345.7	7731110.4	0.02	1.1	39	0.00
4	9	18860806.0	18860806.0	0.00	53	18849170.5	18860806.0	0.06	1.6	31	0.00
5	12	9530299.4	9530299.4	0.00	1179	9527438.2	9530299.4	0.03	1.0	41	0.00
6	9	7522070.0	7522070.0	0.00	108	7519698.3	7522070.0	0.03	1.0	35	0.00
7	10	9395256.7	9395256.8	0.00	214	9390639.6	9395256.8	0.05	0.6	39	0.00
8	9	6421194.3	6421194.3	0.00	140	6418537.1	6421194.3	0.04	0.7	37	0.00
9	11	20007963.0	20007963.0	0.00	288	20006168.7	20007963.0	0.01	0.6	30	0.00
10	6	6254402.2	6254402.2	0.00	16	6244826.6	6254402.2	0.15	0.6	43	0.00
Average				***0.00***	***261***			***0.05***	***0.9***	***37***	***0.00***

The nested Lagrangian relaxation heuristic finds optimal or near optimal solutions for all instances in set A, B, and C in reasonable CPU times. The average optimality gap is 0.05%, 0.05%, and 0.06% and the average solution time is 0.9, 9.0 and 98.7 seconds for instance set A, B, and C, respectively. We also notice that the average number of iterations in set C is 26, which is lower than 37 and 32 iterations for sets A and B, respectively. This is because the nested Lagrangian relaxation algorithm is terminated if the lower bound does not improve after five consecutive iterations for instances in set C.

When comparing the solutions of BARON with those of the nested Lagrangian relaxation heuristic, we can see that for small size instances in set A, the average GAP′%= 0.0. Which means that the nested Lagrangian relaxation heuristic is able to find the optimal solutions in all instances. From Table 8, one can observe that the nested Lagrangian relaxation heuristic provides better solutions for all instances in set B since GAP′%< 0 for all instances, with an average of -46.8%. Further, Table 9 shows that the optimality gap is very small with an average GAP = 0.06% and the lower bounds of the nested Lagrangian are much higher than those provided by BARON. In addition, GAP′ = -68.5% for instance 6 for which BARON finds a feasible solution.

Table 8: Results for set B with 10 workcenters

Ins.#	Prd.#	BARON				Nested Lagrangian relaxation					Gap′ %
		LB	UB	Gap %	Time(s)	LB	UB	Gap %	Time(s)	Iter	
1	157	82811643	401060250	79.4	3600	223051491	223120120	0.03	7.6	26	-44.4
2	156	94744226	330564870	71.3	3600	168272393	168307780	0.02	9.4	36	-49.1
3	138	76940600	282948000	72.8	3600	153586905	153655015	0.04	18.6	43	-45.7
4	143	76383256	439425000	82.6	3600	142019460	142119768	0.07	8.1	35	-67.7
5	136	123519000	302441000	59.2	3600	173603949	173743092	0.08	6.7	30	-42.6
6	137	170635560	336925460	49.4	3600	236931787	236981006	0.02	6.9	24	-29.7
7	144	56245926	310623800	81.9	3600	139680261	139783514	0.07	10.7	33	-55.0
8	126	133239540	264551820	49.6	3600	175354154	175384711	0.02	6.9	30	-33.7
9	159	62575199	321347260	80.5	3600	144426656	144503319	0.05	7.9	29	-55.0
10	138	123389790	298290740	58.6	3600	162868925	162945945	0.05	7.7	33	-45.4
Average				***68.5***	***3600***			***0.05***	***9.0***	***32***	***-46.8***

Table 9: Results for set C with 30 workcenters

Ins.#	Prd.#	BARON				Nested Lagrangian relaxation					Gap′ %
		LB	UB	Gap %	Time(s)	LB	UB	Gap %	Time(s)	Iter	
1	934	45016209	-	-	3600	1091682660	1092637780	0.09	77.6	20	-
2	919	40973842	-	-	3600	1003378835	1003840743	0.05	113.3	32	-
3	900	41000300	-	-	3600	859477698	860049665	0.07	229.8	23	-
4	831	39596200	-	-	3600	889559131	890113578	0.06	77.7	28	-
5	844	37224343	-	-	3600	913484638	913903133	0.05	67.7	22	-
6	920	43079500	2677180000	98.4	3600	844110355	844379870	0.03	70.6	24	-68.5
7	865	41400800	-	-	3600	933763974	934339898	0.06	79.3	26	-
8	895	40651627	-	-	3600	933587117	934316778	0.08	74.8	30	-
9	940	42127887	-	-	3600	937714247	938334400	0.07	95.3	26	-
10	915	43539152	-	-	3600	983672233	984140914	0.05	100.6	28	-
Average				***-***	***3600***			***0.06***	***98.7***	***26***	***-***

8 Conclusion

This paper addresses the problem of jointly optimizing capacity planning and safety stock placement under the GSA in a serial production-distribution system with multiple products. The production-distribution network consists of a capacitated manufacturer that supplies a warehouse with limited storage capacity and a retailer. The manufacturer must efficiently allocate capacity to multiple workcenters, which makes the lead times of these workcenters endogenous. We formulate the integrated problem of planning capacity and safety stocks with the objective of minimizing the WIP, inventory and overtime costs subject to budget, and storage constraints.

For a single workcenter processing a single-product, the interaction between the manufacturer's lead time, the storage capacity at the warehouse, inventory costs, and the safety stock placement is analytically characterized. For a given budget at the manufacturer, there is a storage threshold above which the problem is similar to the problem with infinite storage capacity. When the storage capacity is greater than the storage threshold, there is a replenishment lead time threshold that determines the safety stock placement. This replenishment lead time threshold is increasing in the ratio of inventory costs. Similarly, there is also a threshold for the ratio of the inventory cost, which is an increasing function of the replenishment lead time. When the storage capacity is less than the storage threshold, the safety stock placement depends on another replenishment lead time threshold, which is increasing

in the storage capacity and decreasing in the ratio of inventory costs. Similarly, there is also a threshold for the ratio of the inventory cost, which is a decreasing function of the replenishment lead time and increasing in the storage capacity.

When the manufacturer has multiple workcenters, the integrated problem is formulated as a non-convex program and solved using a nested Lagrangian relaxation heuristic. The algorithm dualizes the storage constraint in the first phase and the budget constraint in the second phase. This decomposes the problem into subproblems, each corresponding to a single workcenter, that are easy to solve. To solve the integrated problem, lower bounds are computed by iteratively solving the relaxed problems and employing efficient greedy heuristics to find tight upper bounds. Subgradient procedures update the Lagrangian multipliers in both phases until an acceptable optimality gap is reached. Our computational experiments show that the nested Lagrangian relaxation heuristic is able to find optimal or near-optimal solutions in reasonable CPU times, and outperforms BARON, which is a commercial mixed-integer nonlinear optimization solver, in terms of the average optimality gap and run time.

In addition, a simulation study was conducted to evaluate the accuracy of the mathematical model and to compare the solution of the integrated approach with that of the sequential approach employed to set capacity and safety stocks. These experiments illustrate that for multiple products with highly different demand variability and inventory costs, the integrated approach results in a high value of integration relative to the sequential approach. The reason is that the integrated approach considers the effect of demand variability, inventory costs, and the cost trade-offs between different products competing for the shared manufacturer's budget and warehouse storage and allocates capacity more efficiently.

This paper considers a serial supply chain and can be extended to analyze more complex supply chain structures. Manufacturers may belong to different firms each with its own capacity budget. Another interesting future research direction would be to model other production decisions, such as lot sizing and scheduling at the manufacturer, and to study the impact on capacity allocation and safety stocks.

Acknowledgment

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

References

Akinc, U. and Khumawala, B. M. (1977). An efficient branch and bound algorithm for the capacitated warehouse location problem. *Management Science*, 23(6):585–594.

Allen, E., Helgason, R., Kennington, J., and Shetty, B. (1987). A generalization of Polyak's convergence result for subgradient optimization. *Mathematical Programming*, 37(3):309–317.

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

Axsäter, S. (2015). *Inventory control*, volume 225. Springer.

Billington, C., Callioni, G., Crane, B., Ruark, J. D., Rapp, J. U., White, T., and Willems, S. P. (2004). Accelerating the profitability of Hewlett-Packard's supply chains. *Interfaces*, 34(1):59–72.

Bitran, G. R. and Tirupati, D. (1989a). Capacity planning in manufacturing networks with discrete options. *Annals of Operations Research*, 17(1):119–135.

Bitran, G. R. and Tirupati, D. (1989b). Capacity planning in manufacturing networks with discrete options. *Annals of Operations Research*, 17(1):119–135.

Bitran, G. R. and Tirupati, D. (1989c). Tradeoff curves, targeting and balancing in manufacturing queueing networks. *Operations Research*, 37(4):547–564.

Bretthauer, K. (1995). Capacity planning in networks of queues with manufacturing applications. *Mathematical and Computer Modelling*, 21(12):35–46.

Bretthauer, K. M. and Côté, M. J. (1997). Nonlinear programming for multiperiod capacity planning in a manufacturing system. *European Journal of Operational Research*, 96(1):167–179.

Çelik, S. and Maglaras, C. (2008). Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. *Management Science*, 54(6):1132–1146.

Clark, A. J. and Scarf, H. (1960). Optimal policies for a multi-echelon inventory problem. *Management Science*, 6(4):475–490.

Fisher, M. L. (1981). The Lagrangian relaxation method for solving integer programming problems. *Management Science*, 27(1):1–18.

Fisher, M. L. (1985). An applications oriented guide to Lagrangian relaxation. *Interfaces*, 15(2):10–21.

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

Graves, S. C. et al. (1988). Safety stocks in manufacturing systems. *Journal of Manufacturing and Operations Management*, 1(1):67–101.

Graves, S. C. and Schoenmeyr, T. (2016). Strategic safety-stock placement in supply chains with capacity constraints. *Manufacturing & Service Operations Management*, 18(3):445–460.

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

Graves, S. C. and Willems, S. P. (2003). Supply chain design: Safety stock placement and supply chain configuration. In de Kok, A. G. and Graves, S. C., editors, *Handbook of Operations Research and Management Science, Vol. 11*, chapter 3, pages 95–131. Elsevier.

Hariga, M. A. (2010). A single-item continuous review inventory problem with space restriction. *International Journal of Production Economics*, 128(1):153–158.

Held, M., Wolfe, P., and Crowder, H. P. (1974). Validation of subgradient optimization. *Mathematical Programming*, 6(1):62–88.

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

Horst, R., Pardalos, P. M., and Van Thoai, N. (2000). *Introduction to global optimization*. Springer Science & Business Media.

Hsieh, M. F. (2011). *Applying a MEIO approach to manage Intel's VMI Hub Supply*. PhD thesis, Massachusetts Institute of Technology.

Hua, N. G. and Willems, S. P. (2016). Analytical insights into two-stage serial line supply chain safety stock. *International Journal of Production Economics*, 181:107–112.

Inderfurth, K. (1993). Valuation of leadtime reduction in multi-stage production systems. In *Operations Research in Production Planning and Control*, pages 413–427. Springer.

Kim, S. and Uzsoy, R. (2008). Exact and heuristic procedures for capacity expansion problems with congestion. *IIE Transactions*, 40(12):1185–1197.

Kim, S. and Uzsoy, R. (2009). Heuristics for capacity planning problems with congestion. *Computers & Operations Research*, 36(6):1924–1934.

Kleinrock, L. (1975). Queueing systems, vol. 1: Theory, vol. 2: Computer applications. *J. Wiley and Sons*.

Klosterhalfen, S. and Minner, S. (2010). Safety stock optimisation in distribution systems: a comparison of two competing approaches. *International Journal of Logistics: Research and Applications*, 13(2):99–120.

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

Kotler, P. and Armstrong, G. (2010). *Principles of marketing*. Pearson Education.

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

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

Lemmens, S., Decouttere, C., Vandaele, N., Bernuzzi, M., and Reichman, A. (2016). Performance measurement of a rotavirus vaccine supply chain design by the integration of production capacity into the guaranteed service approach. *Available at SSRN 2841176*.

Liu, K., Zhou, Y., and Zhang, Z. (2010). Capacitated location model with online demand pooling in a multi-channel supply chain. *European Journal of Operational Research*, 207(1):218–231.

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

Martínez-Costa, C., Mas-Machuca, M., Benedito, E., and Corominas, A. (2014). A review of mathematical programming models for strategic capacity planning in manufacturing. *International Journal of Production Economics*, 153:66–85.

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

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

Plambeck, E. L. and Ward, A. R. (2007). Note: A separation principle for a class of assemble-to-order systems with expediting. *Operations Research*, 55(3):603–609.

Plambeck, E. L. and Ward, A. R. (2008). Optimal control of a high-volume assemble-to-order system with maximum leadtime quotation and expediting. *Queueing Systems*, 60(1-2):1.

Polak, B. M. (2014). *Multi-echelon inventory strategies for a retail replenishment business model*. PhD thesis, Massachusetts Institute of Technology.

Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17.

Polyak, B. T. (1969). Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29.

Pyke, D. F. and Cohen, M. A. (1993). Performance characteristics of stochastic integrated production-distribution systems. *European Journal of Operational Research*, 68(1):23–48.

Rajagopalan, S. and Yu, H.-L. (2001). Capacity planning with congestion effects. *European Journal of Operational Research*, 134(2):365–377.

Shor, N. Z. (2012). *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media.

Simchi-Levi, D., Kaminsky, P., Simchi-Levi, E., and Shankar, R. (2008). *Designing and managing the supply chain: concepts, strategies and case studies*. Tata McGraw-Hill Education.

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

Thomas, B. G. and Bollapragada, S. (2010). General electric uses an integrated framework for product costing, demand forecasting, and capacity planning of new photovoltaic technology products. *Interfaces*, 40(5):353–367.

Wang, K.-J., Wang, S.-M., and Yang, S.-J. (2007). A resource portfolio model for equipment investment and allocation of semiconductor testing industry. *European Journal of Operational Research*, 179(2):390–403.

Wieland, B., Mastrantonio, P., Willems, S. P., and Kempf, K. G. (2012). Optimizing inventory levels within Intel's channel supply demand operations. *Interfaces*, 42(6):517–527.

Woerner, S., Laumanns, M., and Wagner, S. M. (2018). Joint optimisation of capacity and safety stock allocation. *International Journal of Production Research*, 56(13):4612–4628.

Yang, F., Ankenman, B., and Nelson, B. L. (2007). Efficient generation of cycle time-throughput curves through simulation and metamodeling. *Naval Research Logistics*, 54(1):78–93.

Yang, F., Ankenman, B. E., and Nelson, B. L. (2008). Estimating cycle time percentile curves for manufacturing systems via simulation. *INFORMS Journal on Computing*, 20(4):628–643.

Highlights:

i) Capacity and safety stocks are optimized in a serial production-distribution system.
ii) The effect of warehouse storage capacity on safety stocks is characterized.
iii) A nested Lagrangian relaxation heuristic is proposed to solve the problem.
iv) A simulation study demonstrates the value of integration.