Delivery time quotation and pricing in two-stage supply chains: Centralized decision-making with global and local managerial approaches

Ramzi Hammami, Yannick Frein, Abduh Albana

Version of Record: https://www.sciencedirect.com/science/article/pii/S0377221720302204
Manuscript_535b16a39b977772bcf9b39b4959873e

Delivery Time Quotation and Pricing in Two-Stage Supply Chains: Centralized Decision-making with Global and Local Managerial Approaches

Ramzi Hammami*[a], Yannick Frein[b], Abduh S. Albana[c]

[a]Rennes School of Business, 2 Rue Robert d'Arbrissel - 35065 Rennes - France.

hammami.ramzi@gmail.com, ramzi.hammami@rennes-sb.com.

[b]Univ. Grenoble Alpes, CNRS, G-SCOP, 38 000 Grenoble- France. yannick.frein@gscop.inpg.fr.

[c]Univ. Grenoble Alpes, CNRS, G-SCOP, 38 000 Grenoble- France.

abduh-sayid.albana@grenoble-inp.fr.

* Corresponding author

Abstract

This study investigates the delivery time quotation and pricing in a two-stage make-to-order supply chain facing a time- and price-sensitive demand. We consider different managerial approaches which results in different models. First, we study a global model where a pair of price and delivery time are quoted to customers to maximize the expected overall profit while satisfying a global service level on the whole system. Second, we study a local model where each stage is required to quote a local delivery time while satisfying a local service level, and the delivery time quoted to customers consists of both local delivery times and must satisfy the global service level. The objective is similar to that of the global model. When both stages target the same service level than the one imposed to the whole system, we demonstrate under realistic conditions that satisfying the local service constraints enables to satisfy the global service constraint. This allows to remove the global constraint from the local model and solve it analytically. With comparison to the global model, the local model presents several managerial advantages with a limited profit loss. The mean gap is only 1.68% for a service level of 95%. We perform sensitivity analyses to derive insights into the impact of market characteristics and capacities on the performance of each stage and the overall performance. Finally, we extend the local model by allowing each stage to targeting a different service level. This leads to closing the profit gap with the global model.

Keywords: Supply chain management; Lead time quotation; Pricing; Two-stage tandem queue; Endogenous demand.

1 Introduction

The delivery time (DT) quotation and pricing are strategic decisions that determine the demand of a substantial majority of customers, and significantly impact the firm's profitability (Celik and Maglaras, 2008, Huang et al., 2013, Hammami and Frein, 2014, Xiao and Qi, 2016). While the selling price has a well-understood impact on demand and profitability, the DT factor may be more complicated. Indeed, quoting a shorter DT can yield a higher demand but increases the risk of late delivery. As reliability in meeting agreed DTs is of value to customers (Boyaci and Ray, 2006, Kingsman et al. 1998), firms generally target a given service level. However, to satisfy the service level constraint (i.e. to guarantee that the probability of respecting the quoted DT is greater than the service level), it may be required to quote a longer DT, which may deter some customers and then yields a lower demand. Combining DT quotation and pricing implies more complex trade-offs since pricing impacts the demand rate and, consequently, has an effect on the DT quotation (a smaller price generates a higher demand but increases the risk of late delivery).

To offer a short DT, some companies adopt a make-to-stock (MTS) production policy. In this case, the work is released based on demand forecast, and demand is satisfied from the available stock. However, when there are diversified and uncertain customer requirements or when holding inventory is costly, the downstream operations may prefer to adopt a make-to-order (MTO) approach, i.e., release the work only in response to customers' orders. Many practical systems combine both MTS and MTO operations. Indeed, there is generally a Customer Order Decoupling Point (CODP) that divides the operations into forecast-driven operations (upstream of the CODP), and customer order-driven operations (downstream of the CODP). The DT of a customer order mostly depends on the lead times associated with the MTO operations (if one ignores the stockout in the CODP) (Hammami et al. 2017). This study focuses on the MTO operations occurring after the CODP. According to Haskose et al. (2004), the production process associated with the MTO operations can be viewed as a network of queues, of different types. The authors explained that one possible configuration is the tandem network with at least two stages. In practice, this corresponds to MTO manufacturing systems where each

customer order requires processing (transformation work) on a series of workstations through the production facilities Haskose et al. (2004).

This research investigates the DT quotation and pricing in a MTO supply chain (SC) facing a DT- and price-sensitive demand. The extant works typically aggregate the SC into a single operation stage and model the system as a $M/M/1$ queue. One remark is in order here. A few works investigated a two-stage SC (e.g., Liu et al. 2007, Zhu, 2015, Xiao and Qi, 2016). However, these works also consider only one operation stage, whereas the other stage just plays a pricing role and does not affect the sojourn time of customers' orders in the system, so does not impact the DT quotation. Consequently, the system is also modeled as a single queue (similar to single-stage models). A more detailed analysis of processing and waiting lead times suggests to model the system as a network of queues (Haskose et al. 2004). This is often more realistic than considering a single aggregated operation.

This is the first paper to investigate the DT quotation and pricing in a MTO tandem queue model consisting of two operations stages (each stage has a finite capacity) and facing a random DT- and price-sensitive demand. This corresponds, for instance, to a system where the upstream stage manufactures a semi-finished product and the downstream stage manufactures and/or assembles the final product. Our approach is particularly interesting when the whole system can be decomposed into two independent sub-systems, such as in the following situations.

- When the processing operations are located in two different facilities, it is generally more suitable to model each facility as a separate stage instead of aggregating the system into a single operation stage, especially when each facility has a different role. Indeed, this enables to set specific objectives to each facility and to assess its local performance. Considering two stages is also more representative of real SCs. There are many examples of SCs with MTO operations located in two different facilities, and where each facility has a specific role (see e.g., Zhu, 2015, Liu et al. 2007).

- The operations stages are usually of different natures. For instance, they may correspond to a machining process followed by an assembly process, or to a design/configuration process followed by a manufacturing process, which is a common situation for the manufacturers of capital goods equipment, mainly producers of machines for making things. In such cases, it is more realistic to consider a distinct stage for each type of operations instead of a single aggregated stage.

We model the system as a tandem queue $M/M/1-\ M/M/1$. Both stages have a finite capacity and impact the sojourn time of customers' orders in the system. The demand arrives at the downstream stage according to a Poisson process and the mean demand rate linearly decreases in the offered price and the quoted DT. The service time in each stage is exponentially distributed. The assumption of exponential service times is suitable for the cases where there is a high service time variability. Kingsman et al. (1998) argued that there is often a high level of variability with respect to the processing times. Haskose et al. (2004) also reported that service times are often unreliable due to the large proportion of time spent in the queues.

We address this problem with two different managerial approaches.

- A global approach. The decision maker chooses a price and a DT to maximize the overall expected profit while guaranteeing a minimum service level, denoted by s, to customers. Thus, the probability that the sojourn time in the system is smaller than the quoted DT, must be greater than s. This constraint is referred to by the global service constraint as it applies to the whole system. The global service constraint leads to a challenging model since the sojourn time in a $M/M/1-\ M/M/1$ queue follows a hypo-exponential distribution (whereas it is exponentially distributed for a single $M/M/1$ queue).

- A local approach. While considering the same objective function of the global model, each stage is here required (by the decision maker) to quote a local DT while satisfying locally the service level s. The DT quoted to customers consists of both local DTs. Since the customers are only interested in the global service level, the global service constraint must also be satisfied. This strategy can be interesting in practice as it gives a clear objective (in terms of DT quotation) to each stage. Hence, on the one hand, it is easier to implement and, on the other hand, it enables to control and assess the local performance of each stage (facility).

The global model is very hard to solve. We simplify this model and propose a numerical solving approach. For the local model, we demonstrate that satisfying the local service constraints enables to satisfy the global service constraint when (i) the DT quoted to customers is the sum of local DTs, and (ii) both stages target the same service level than the one imposed to the whole system. This enables to obtain a simpler formulation of the local model. We then solve this model and provide the optimal solution analytically. We quantify the profit loss resulting from using the local model instead of the global model and show that this loss is

relatively small, especially in the general case where the stages do not have the same capacity. Thus, the local model can also be used as an approximation of the global model, which is an interesting result since we solve analytically the local model but the global model cannot be solved. Then, we conduct sensitivity analyses and derive insights into the impact of market characteristics and capacities.

Finally, we extend the local model to consider the case where each stage may target a different service level (s_1 and s_2) and where these service levels are also decision variables to be optimized. In case of balanced capacity, we solve the model analytically and show that the optimal profit is very close to the profit obtained with the global model. In case of unbalanced capacity, we show numerically that considering variable service levels can improve the profit with comparison to the basic local model (it is recalled that $s_1 = s_2 = s$ in this latter model). We also study the robustness of our results to the assumption of exponential service times. We simulate different service time distributions and show numerically that the exponential assumption can be a good approximation.

In Section 2, we review the relevant literature. We dedicate Section 3 to the study of the global model. We develop and solve the local model in Section 4. In Section 5, we compare both formulations and conduct experiments to derive insights. In Section 6, we study the robustness of our models and investigate some extensions. We finally conclude and give future work directions.

2 Literature review

The present work is related to the stream of research on DT quotation and pricing in MTO environments with endogenous demand. In Table 1, we classify the relevant papers according to three dimensions: (1) The number of stages in the SC, single-stage or two-stage models; (2) The decision process, centralized (i.e. only one decision maker undertakes all decisions simultaneously) or decentralized (i.e. different decision makers, each one of them undertakes a subset of decisions); and (3) The number of stages impacting the sojourn time of customers' orders in the system and, consequently, affecting the DT quotation.

Table 1. Classification of relevant papers

	Single-stage SC	Two-stage SC	Centralized decision	Decentralized decision	Number of stages impacting the DT quotation
Palaka et al. (1998)	X		X		1
So and Song (1998)	X		X		1
Boyaci and Ray (2003)	X		X		1
Ray and Jewkes (2004)	X		X		1
Boyaci and Ray (2006)	X		X		1
Zhao et al. (2012)	X		X		1
Albana et al. (2018)	X		X		1
Pekgün et al. (2008)		X		X	1
Pekgün et al. (2017)		X		X	1
Liu et al. (2007)		X		X	1
Zhu (2015)		X		X	1
Xiao and Qi (2016)		X		X	1
Our paper		X	X		2

In a pioneer paper, Palaka et al. (1998) studied the problem of DT quotation, pricing, and capacity utilization of a profit-maximizing firm modeled as a single-stage $M/M/1$ queue and facing a linear price- and DT-sensitive demand. Basically, they investigated two situations with either a fixed capacity or a fixed price. In both cases, they characterized the optimal solution with cubic equations. So and Song (1998) investigated a quite similar problem while using a log-linear demand model (Cobb-Douglas) instead of the linear demand. The framework developed by Palaka et al. (1998) has been extended by many authors. Boyaci and Ray (2003) considered a firm selling two substitutable products (a regular product with a given standard DT, and an express product that is supposed to have a faster DT) in a price- and DT-sensitive market. Each type of demand is served from a different facility, each facility is modeled as a $M/M/1$ queue. This work has been extended by Boyaci and Ray (2006) to incorporate the delivery reliability (i.e. the service level) as a new decision variable. Ray and Jewkes (2004) incorporated the economies of scale by assuming that the unit operating cost is decreasing convex with respect to the mean demand rate and studied a simplified version of Palaka et al.'s model where it is assumed that the price is not an independent variable but is a linear decreasing function in the

quoted DT. Another interesting extension was proposed by Zhao et al. (2012). The authors compared the strategy where a firm offers a single DT and price quotation to the strategy when a firm offers a menu of DT and prices for customers to choose from. Similar to the previous works, the system is modeled as a $M/M/1$ queue, and the demand is linear in price and DT. Recently, Albana et al. (2018) extended the existing works by modeling the unit operating cost, not as a constant, but as a convex decreasing function in the quoted DT. This assumes that the operational cost decreases if a longer DT is quoted to customers. The authors studied three settings: (i) DT is the unique variable, (ii) DT and price are variables with a fixed capacity, and (iii) DT, price, and capacity are variables.

The papers described above studied a single-stage SC. In what follows, we focus on the papers that considered a two-stage SC, which is more relevant to our study. Pekgün et al. (2008) studied the centralization and decentralization of pricing and DT decisions for a MTO firm modeled as a $M/M/1$ queue and facing a linear price- and DT-sensitive demand. Their centralized model is similar to Palaka et al.'s model but with a constant capacity and without holding and penalty costs. For the decentralized model, they studied two settings with either marketing or production as a leader. The production department quotes a DT and the marketing department quotes a price. The authors observed, for instance, that a higher capacity results in a greater flexibility and a higher profit for a centralized firm. However, a higher capacity does not necessarily result in a higher profit for a decentralized firm. The authors also studied a coordination mechanism where marketing pays a given amount to production for each unit produced, and both departments receive a bonus payment as the fraction of the total revenue generated. Pekgün et al. (2017) extended the previous research by considering two firms that compete on price and DT decisions in a common market. For each firm, both centralization and decentralization (with either marketing or production as the leader) were considered. The authors found that under intense price competition, the firms may suffer from a decentralized structure. In contrast, under intense DT competition, a decentralized strategy with marketing as the leader can not only result in significantly higher profits, but also be the equilibrium strategy.

Liu et al. (2007) studied a decentralized SC with a supplier and a retailer facing a price- and DT-dependent demand. The supplier's decisions are the quoted DT (to customers) and the wholesale price (to the retailer). The retailer's decision is the final price quoted to customers. The Stackelberg game was used to analyze the problem of the supplier (as a leader) and the

retailer (as a follower). The authors illustrated some of their results by considering the supplier as a $M/M/1$ queue, the retailer plays only a pricing role and does not impact the sojourn time in the system. Using the performance of the corresponding centralized system as a benchmark, the authors showed that the decentralized decisions are inefficient and lead to inferior performance due to the double marginalization effect. Zhu (2015) considered a decentralized SC consisting of a supplier and a retailer where the supplier (as a leader) determines the capacity and wholesale price, and the retailer (as a follower) decides the final price and the DT. Similar to Liu et al. (2007), the demand is linear in the retailer's price and the quoted DT. The supplier's facility was modeled as a $M/M/1$ queue. The retailer plays an intermediate role and does not impact the quoted DT. By using the decentralized chain without capacity decision as a benchmark, the authors demonstrated that the integration of capacity decisions can significantly reduce the profit loss caused by the double marginalization. Finally, Xiao and Qi (2016) considered a two-stage SC with one supplier, operating in MTS where it was assumed that stock-out cannot happen, and one MTO manufacturer, modeled as a $M/M/1$ queue. Thus, the quoted DT depends only on the DT of the manufacturer. In the basic model, the supplier chooses the wholesale price and the manufacturer determines the resale price and the quoted DT. The authors also studied the case where the manufacturer determines the resale price, the quoted DT, and the delivery reliability or the capacity, and where demand is a linear function in these three variables. The authors investigated the coordination of the channel via an all-unit quantity discount contract under different scenarios and derived some managerial insights, for example, how the delivery reliability may affect the demand rate and the channel profit, and whether the all-unit quantity discount scheme can still coordinate the SC.

The literature overview shows that the extant works (either single-stage or two-stage models) calculate the sojourn time of customers' orders in the system based on only one operation stage, so typically model the system as a single $M/M/1$ queue. This research extends the extant literature by considering a two-stage centralized SC where both stages perform operations and impact the DT quoted to customers.

3 Model with global service constraint

We first describe the general modeling framework adopted in this paper and then formulate the global model.

3.1 Modeling framework

We consider a SC consisting of two MTO operations stages. The demand arrives at the downstream stage according to a Poisson process with a mean arrival rate λ. The mean demand rate linearly decreases in the quoted DT and price. We respectively denote by p and l the selling price and the DT quoted to customers. Thus, $\lambda = a - \alpha p - \beta l$, where a is the market potential, and α and β are respectively the price-sensitivity and the DT-sensitivity of demand. Both upstream and downstream stages have a finite capacity, and their service (processing) times are exponentially distributed with mean service rates μ_1 and μ_2, respectively. Thus, we model the system as a tandem queuing network ($M/M/1 - M/M/1$) as illustrated in Figure 1.

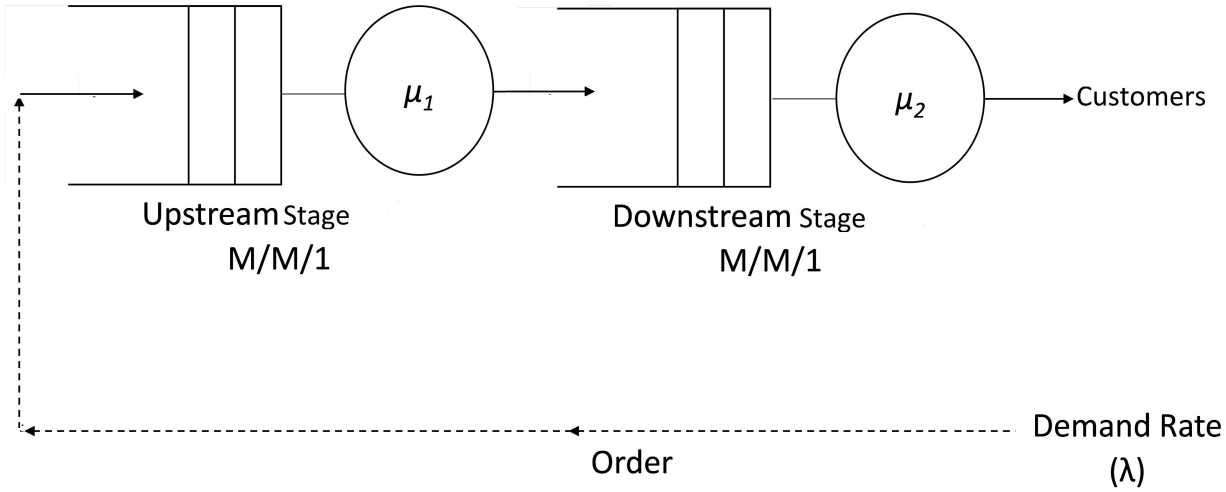

Figure 1. The studied system modeled as a Tandem queue

To prevent the SC from quoting unrealistic and unreliable DT, a service level s (defined by the firm) must be satisfied. Thus, the probability of serving customers' demand on time must not be smaller than s. We respectively denote by w_1 and w_2 the sojourn times (waiting time + processing time) in the upstream and downstream stages. Hence, the total sojourn time in the system is $w = w_1 + w_2$, and the service constraint is thus given by $\Pr(w_1 + w_2 \leq l) \geq s$. We refer to this constraint by the global service constraint (as it applies to the whole SC).

3.2 Model formulation and analysis

In the global model, the firm decides the price and the DT (p and l) to maximize the overall expected profit while satisfying a global service constraint. We let (GSM) denote the global service model. The formulation of this model is described below. The profit is calculated in equation (1) as the difference between the revenue and the cost, where m_1 and m_2 respectively

denote the unit direct variable cost for stage 1 and stage 2. Equation (2) gives the mean demand rate as a function of price and DT. In a tandem queue $M/M/1-M/M/1$, it is known that the sojourn time in each stage is exponentially distributed with mean $\frac{1}{\mu_1-\lambda}$ and $\frac{1}{\mu_2-\lambda}$ for upstream and downstream stage respectively, and that the total sojourn time in the system (i.e. $w = w_1 + w_2$) follows the hypo-exponential distribution if $\mu_1 \neq \mu_2$, and follows the Erlang $(2, \mu)$ distribution if $\mu_1 = \mu_2 = \mu$. Therefore, the global service constraint can be formulated as given in constraint (3). Constraint (4) guarantees a steady state at each stage and imposes a positive demand.

$$(GSM) \ \underset{l,p\geq 0}{\text{Maximize}} \ \ \Pi(l,p) = (p - m_1 - m_2)\lambda \tag{1}$$

$$\text{Subject to} \ \ \lambda = a - \alpha p - \beta l \tag{2}$$

$$\begin{cases} 1 - \frac{\mu_2-\lambda}{\mu_2-\mu_1}e^{-(\mu_1-\lambda)l} + \frac{\mu_1-\lambda}{\mu_2-\mu_1}e^{-(\mu_2-\lambda)l} \geq s \text{ if } \mu_1 \neq \mu_2 \\ 1 - e^{-(\mu-\lambda)l} - (\mu-\lambda)le^{-(\mu-\lambda)l} \geq s \text{ if } \mu_1 = \mu_2 = \mu \end{cases} \tag{3}$$

$$0 \leq \lambda < \min\{\mu_1, \mu_2\} \tag{4}$$

Obviously, given the complexity of the service constraint, model (GSM) cannot be solved analytically. We shall try to simplify the model in order to solve it numerically with an optimization software. We firstly derive the following Lemma.

Lemma 1 *The service constraint is binding (for both cases $\mu_1 \neq \mu_2$ and $\mu_1 = \mu_2$) and, consequently, we have $\Pr(w_1 + w_2 \leq l) = s$ at optimality.*

Proof. We prove this result by contradiction. Suppose that we have an optimal solution p^* and l^* such that $\Pr(w_1 + w_2 \leq l^*) > s$. The optimal profit in this case is $\Pi^*(l^*, p^*)$. If we decrease the DT from l^* to $l^{'}$ while keeping the price constant until we have $\Pr(w_1 + w_2 \leq l^{'}) = s$, then we will get $\Pi^{'}(l^{'}, p^*) > \Pi^*(l^*, p^*)$ because demand has increased. Thus, $\left(l^{'}, p^*\right)$ is feasible and gives a higher profit than (l^*, p^*), which is impossible. The service constraint is therefore binding. Of course, this result holds for both cases $\mu_1 \neq \mu_2$ and $\mu_1 = \mu_2$. ■

We let $g(p,l) = \Pr(w_1+w_2 \leq l) - s = \begin{cases} 1 - s - \frac{\mu_2-\lambda}{\mu_2-\mu_1}e^{-(\mu_1-\lambda)l} + \frac{\mu_1-\lambda}{\mu_2-\mu_1}e^{-(\mu_2-\lambda)l} \text{ if } \mu_1 \neq \mu_2 \\ 1 - s - e^{-(\mu-\lambda)l} - (\mu-\lambda)le^{-(\mu-\lambda)l} \text{ if } \mu_1 = \mu_2 = \mu \end{cases}$.
Since constraint (3) is binding, we must have $g(p,l) = 0$ at optimality. Clearly, equation $g(p,l) = 0$ cannot be solved analytically. Nevertheless, for each fixed price p, we can solve

equation $g_p(l) = 0$ numerically and get the corresponding values of l. To simplify the problem, we consider the result of Lemma 2.

Lemma 2 *For a given price p, equation $g_p(l) = 0$ has only one positive root.*

Proof. On the one hand, it is known that the longer the quoted DT is, the higher the probability of satisfying the service constraint becomes. Therefore, $\Pr(w_1 + w_2 \leq l)$ is increasing in l and, consequently, $g_p(l)$ is increasing in l. On the other hand, $g_p(l)$ is continuous, and one can verify that $\lim\limits_{l\to 0} g_p(l) = -s < 0$, and $\lim\limits_{l\to +\infty} g_p(l) = 1 - s > 0$. Hence, equation $g_p(l) = 0$ has only one positive root. ■

We let $l_0(p)$ denote the positive root of equation $g_p(l) = 0$ (obviously, this root depends on the value of p). Therefore, model (GSM) becomes equivalent to the following single-variable model.

$$(GSM) \ \underset{p\geq 0}{\text{Maximize}} \ \ \Pi(p) = (p - m_1 - m_2)\,(a - \alpha p - \beta l_0(p)) \tag{5}$$

$$\text{Subject to} \ \ 0 \leq a - \alpha p - \beta l_0(p) < \min\{\mu_1, \mu_2\} \tag{6}$$

As we cannot get the closed-form expression of $l_0(p)$, no more analytical development can be made. Nevertheless, given that the problem was simplified and reduced to a single-variable optimization model and that $l_0(p)$ can be obtained numerically by any optimization software, we can solve model (GSM) with a numerical approach.

Indeed, we first highlight that we are interested only in the values of p ranging from $m_1 + m_2$ to $\frac{a}{\alpha}$ since, otherwise, the profit cannot be positive (price is smaller than cost) or the model cannot be feasible (negative demand). To solve model (GSM), we proceed as follows. For each given price $p_i \in \left[m_1 + m_2, \frac{a}{\alpha}\right]$, we solve the equation $g_{p_i}(l) = 0$ and obtain the unique positive root $l_0(p_i)$. Then, we calculate its associated profit $\Pi(p_i)$ according to equation (5). This procedure enables to draw numerically the curve $\Pi(p)$ as a function of p and to deduce the optimal solution.

Remark 1 *We performed extensive numerical tests and found that the curve $\Pi(p)$ is concave for all tested instances. However, we were not able to prove the concavity as we do not have the explicit expression of $l_0(p)$.*

4 Alternative model with local service constraints

In many practical cases, the company targets a global service level, and then each stage (or facility) is asked to satisfy this same service level. In addition, for many companies, the service levels have always been viewed as input parameters (sometimes imposed by the market) and not as decision variables to be optimized. When optimizing the service levels is not a priority, a natural approach consists in imposing the same service level everywhere. For instance, a global leader in clinical diagnostics and industrial microbiology operates in MTO the last production phases of some products that have a limited shelf life; this company targets a service level of 97% at all stages in order to achieve a competitive advantage in a market that is highly sensitive to lead time issues.

In this section, we develop an alternative formulation of the problem by considering that each stage is asked to quote a local DT (l_1 and l_2 for stage 1 and stage 2, respectively) while satisfying locally the service level s. The whole SC quotes the DT $l = l_1 + l_2$ to customers. Given that the customers are interested only in the global service level, it is important to also satisfy the global service constraint. The local model, denoted by (LSM), is provided below. The objective function is given in Eq. (7). The mean demand rate λ is given by Eq. (8). The local service constraints are $\Pr(w_1 \leq l_1) \geq s$ and $\Pr(w_2 \leq l_2) \geq s$ for stage 1 and stage 2, respectively. Given that w_1 and w_2 follow the exponential distribution, the local service constraints are given by constraints (9) and (10). Constraint (11) represents the global service constraint (i.e. $\Pr(w_1 + w_2 \leq l) \geq s$ with $l = l_1 + l_2$). It is highlighted that we now have 3 independent decision variables l_1, l_2, and p (instead of 2 independent variables, l and p, for model (GSM)).

$$(LSM)\ \underset{l_1,l_2,p\geq 0}{\text{Maximize}}\ \ \Pi(l_1,l_2,p)=(p-m_1-m_2)\lambda \tag{7}$$

$$\text{Subject to}\ \ \lambda=a-\alpha p-\beta l \tag{8}$$

$$1-e^{-(\mu_1-\lambda)l_1}\geq s \tag{9}$$

$$1-e^{-(\mu_2-\lambda)l_2}\geq s \tag{10}$$

$$\begin{cases} 1-\frac{\mu_2-\lambda}{\mu_2-\mu_1}e^{-(\mu_1-\lambda)l}+\frac{\mu_1-\lambda}{\mu_2-\mu_1}e^{-(\mu_2-\lambda)l}\geq s \text{ if } \mu_1\neq\mu_2 \\ 1-e^{-(\mu-\lambda)l}-(\mu-\lambda)le^{-(\mu-\lambda)l}\geq s \text{ if } \mu_1=\mu_2=\mu \end{cases} \tag{11}$$

$$l=l_1+l_2,\ 0\leq\lambda<\min\{\mu_1,\mu_2\} \tag{12}$$

The motivations and challenges of this alternative formulation are discussed below.

- Managerial perspective. In practical SCs, imposing local service constraints helps to give a clear objective to each stage and, consequently, to efficiently manage and evaluate its performance. If we just impose a global service constraint, then it is known that the whole system must satisfy $\Pr(w_1+w_2\leq l)\geq s$, but it is not clear which DT must be quoted by each facility and with which service level. It is therefore more practical to impose a service constraint to each facility, and then to deduce the total DT that can be quoted to customers. However, imposing local service constraints will necessarily lead to a smaller profit for the firm (since we have more constraints). If the amount of profit loss is significant, then the local model might be useless despite its advantages. It is then important to evaluate the profit loss resulting from using the local model instead of the global model. This will be discussed in Section 5.

- Analytical perspective. In a tandem queue $M/M/1-M/M/1$, satisfying the local service constraints does not necessarily guarantee the satisfaction of the global service constraint. In fact, if each stage i quotes DT l_i while satisfying the local service level s then this does not necessarily imply that the SC can quote the DT $l=l_1+l_2$ with the same service level s. This is why we added the global service constraint to model (LSM) (in addition to local constraints). This leads to a very hard model. However, if we demonstrate under realistic conditions that satisfying the local service constraints can lead to satisfying the global service constraint then we can simplify the model. This result, not known in the

extant literature, will be the focus of the next subsection.

4.1 From global to local service constraints

The objective of this technical section is to demonstrate how satisfying the local service constraints can lead to satisfying the global service constraint when (i) the DT quoted to customers is the sum of local DTs, and (ii) both stages target the same service level than the one imposed to the whole system. We verified that this result is not always guaranteed, i.e. it is possible that the local constraints are satisfied but the global constraint is not. However, our analysis shows that this result holds when the service level s verifies one condition. In this section, we characterize this condition using function $f(s)$ defined below. This function is obtained analytically from the characteristics of the sojourn time distributions in a tandem queue $M/M/1-M/M/1$ (see the proof of Proposition 1). It is recalled that each of w_1 and w_2 follows the exponential distribution, whereas $w = w_1 + w_2$ follows the Erlang $(2, \mu)$ distribution for $\mu_1 = \mu_2 = \mu$, and the Hypo-exponential distribution for $\mu_1 \neq \mu_2$. We thus have different expressions of $f(s)$ based on the values of μ_1 and μ_2.

$$f(s) = \begin{cases} s - 2(1-s)\ln\left(\frac{1}{1-s}\right) & \text{if } \mu_1 = \mu_2 \\ \frac{V_2}{V_1}\left(1-(1-s)^{\frac{V_1}{V_2}}\right) - \left(1-(1-s)^{\frac{V_2}{V_1}}\right) & \text{if } \mu_1 < \mu_2 \\ \frac{V_1}{V_2}\left(1-(1-s)^{\frac{V_2}{V_1}}\right) - \left(1-(1-s)^{\frac{V_1}{V_2}}\right) & \text{if } \mu_1 > \mu_2 \end{cases}$$

, where $V_1 = \mu_1 - \lambda$ and $V_2 = \mu_2 - \lambda$.

Proposition 1 *When $f(s) \geq 0$, the following result holds:*

If $[\Pr(w_1 \leq l_1) \geq s$ *and* $\Pr(w_2 \leq l_2) \geq s]$, *then* $\Pr(w_1 + w_2 \leq l_1 + l_2) \geq s$.

Proof. It is firstly noted that the problem is symmetric in μ_1 and μ_2. Therefore, when $\mu_1 \neq \mu_2$ we assume without loss of generality that $\mu_1 < \mu_2$ (i.e. $V_2 > V_1$). $f(s)$ has different expressions, so we start with the case of $\mu_1 = \mu_2 = \mu$ and then focus on the case of $\mu_1 < \mu_2$.

- Case of $\mu_1 = \mu_2 = \mu$. Suppose that the local constraints are satisfied (i.e. $\Pr(w_1 \leq l_1) \geq s$ and $\Pr(w_2 \leq l_2) \geq s$). If we demonstrate the result when the local constraints are binding (i.e. for $\Pr(w_1 \leq l_1) = s$ and $\Pr(w_2 \leq l_2) = s$), then the result also holds when $\Pr(w_1 \leq l_1) > s$ and $\Pr(w_2 \leq l_2) > s$. In what follows, we consider that the local service constraints are binding (i.e. we have at optimality $\Pr(w_1 \leq l_1) = s$ and $\Pr(w_2 \leq l_2) = s$). $\Pr(w_1 \leq l_1) = s \Leftrightarrow e^{-(\mu-\lambda)l_1} = 1 - s \Leftrightarrow l_1 = \frac{\ln\left(\frac{1}{1-s}\right)}{(\mu-\lambda)}$ and similarly, $\Pr(w_2 \leq l_2) = s \Leftrightarrow l_2 = \frac{\ln\left(\frac{1}{1-s}\right)}{(\mu-\lambda)}$. Hence, $l_1 = l_2 = \frac{\ln\left(\frac{1}{1-s}\right)}{(\mu-\lambda)} = \frac{l}{2}$. For the global service constraint, $\Pr(w_1 + w_2 \leq l) \geq s \Leftrightarrow$

$1-e^{-(\mu-\lambda)l}-(\mu-\lambda)le^{-(\mu-\lambda)l}\geq s$. Since $l=\frac{2\ln\left(\frac{1}{1-s}\right)}{(\mu-\lambda)}$ and $e^{-(\mu-\lambda)l}=(1-s)^2$, it comes that $\Pr(w_1+w_2\leq l)\geq s \Leftrightarrow 1-(1-s)^2-2\ln\left(\frac{1}{1-s}\right)(1-s)^2-s\geq 0$. Thus, the global service constraint is satisfied when $1-(1-s)^2-2\ln\left(\frac{1}{1-s}\right)(1-s)^2-s\geq 0$. Since $1-s>0$, this condition is equivalent to $f(s)\geq 0$.

- Case of $\mu_1<\mu_2$ (i.e. $V_2>V_1$). Similar to the previous case, we suppose that the local constraints are binding (i.e. $e^{-(\mu_1-\lambda)l_1}=1-s$ and $e^{-(\mu_2-\lambda)l_2}=1-s$). It is recalled that $\Pr(w_1\leq l_1)=s\Leftrightarrow l_1=\frac{\ln\left(\frac{1}{1-s}\right)}{\mu_1-\lambda}$ and $\Pr(w_2\leq l_2)=s\Leftrightarrow l_2=\frac{\ln\left(\frac{1}{1-s}\right)}{\mu_2-\lambda}$. For $\mu_1<\mu_2$, $\Pr(w_1+w_2\leq l)=1-\frac{\mu_2-\lambda}{\mu_2-\mu_1}e^{-(\mu_1-\lambda)l}+\frac{\mu_1-\lambda}{\mu_2-\mu_1}e^{-(\mu_2-\lambda)l}$. Given that $e^{-(\mu_1-\lambda)l}=e^{-(\mu_1-\lambda)l_1}e^{-(\mu_1-\lambda)l_2}=(1-s)(1-s)^{\frac{\mu_1-\lambda}{\mu_2-\lambda}}$ and $e^{-(\mu_2-\lambda)l}=(1-s)(1-s)^{\frac{\mu_2-\lambda}{\mu_1-\lambda}}$, it comes that $\Pr(w_1+w_2\leq l)=1-\frac{\mu_2-\lambda}{\mu_2-\mu_1}(1-s)(1-s)^{\frac{\mu_1-\lambda}{\mu_2-\lambda}}+\frac{\mu_1-\lambda}{\mu_2-\mu_1}(1-s)(1-s)^{\frac{\mu_2-\lambda}{\mu_1-\lambda}}$. Therefore, the global service constraint $\Pr(w_1+w_2\leq l)\geq s\Leftrightarrow 1-s-\frac{\mu_2-\lambda}{\mu_2-\mu_1}(1-s)(1-s)^{\frac{\mu_1-\lambda}{\mu_2-\lambda}}+\frac{\mu_1-\lambda}{\mu_2-\mu_1}(1-s)(1-s)^{\frac{\mu_2-\lambda}{\mu_1-\lambda}}\geq 0$, which is equivalent to $1-\frac{\mu_2-\lambda}{\mu_2-\mu_1}(1-s)^{\frac{\mu_1-\lambda}{\mu_2-\lambda}}+\frac{\mu_1-\lambda}{\mu_2-\mu_1}(1-s)^{\frac{\mu_2-\lambda}{\mu_1-\lambda}}\geq 0$. Using the notation $V_1=\mu_1-\lambda$ and $V_2=\mu_2-\lambda$, it comes that $\Pr(w_1+w_2\leq l)\geq s\Leftrightarrow 1-\frac{V_2}{V_2-V_1}(1-s)^{\frac{V_1}{V_2}}+\frac{V_1}{V_2-V_1}(1-s)^{\frac{V_2}{V_1}}\geq 0$. Given that $V_2>V_1$ (since $\mu_1<\mu_2$), this becomes equivalent to $V_2-V_1-V_2(1-s)^{\frac{V_1}{V_2}}+V_1(1-s)^{\frac{V_2}{V_1}}\geq 0$, which can be written as $f(s)=\frac{V_2}{V_1}\left(1-(1-s)^{\frac{V_1}{V_2}}\right)-\left(1-(1-s)^{\frac{V_2}{V_1}}\right)\geq 0$. ■

Based on the result of Proposition 1, we derive the following Corollary.

Corollary 1 *Consider a two-stage supply chain, modeled as a tandem queue $M/M/1-M/M/1$. The upstream stage quotes a delivery time l_1 with a service level s (i.e. $\Pr(w_1\leq l_1)\geq s$), and the downstream stage quotes a delivery time l_2 with the same service level s (i.e. $\Pr(w_2\leq l_2)\geq s$).*

There exists s_0, unique solution of $f(s)=0$ over $[0,1]$, such that:

If $s\geq s_0$, then the whole system can quote the delivery time $l=l_1+l_2$ with the service level s (i.e. $\Pr(w_1+w_2\leq l)\geq s$).

Proof. - Case of $\mu_1=\mu_2=\mu$.

In this case, according to Proposition 1, we just need to prove that $\exists\ s_0\in[0,1]$ such that, if $s\geq s_0$, then $f(s)=s-2(1-s)\ln\left(\frac{1}{1-s}\right)\geq 0$.

We have $\frac{\partial^2 f(s)}{\partial s^2}=\frac{2}{1-s}>0$, implying that function $f(s)$ is convex. According to the first derivative condition, $f(s)$ reaches its minimum in $s_{\min}=1-e^{-1/2}$. We have $f(s_{\min})=1-2e^{-1/2}<0$. In addition, $\lim\limits_{s\to 0}f(s)=0$, and $\lim\limits_{s\to 1}f(s)=1$. Therefore, over $[0,1]$, $f(s)$ decreases from 0 to $1-2e^{-1/2}$, and then increases to reach 1. Consequently, equation $f(s)=0$ has a unique solution s_0 over $[0,1]$, and we have $f(s)\geq 0$ for $s\geq s_0$. To illustrate, we draw $f(s)$ as a function of s in Figure 2 below (please see the case $\mu_1=\mu_2$, which corresponds to $V_1=V_2$).

- Case of $\mu_1 < \mu_2$ (i.e. $V_2 > V_1$).

According to Proposition 1, we need to demonstrate that $\exists\ s_0 \in [0,1]$ such that, if $s \geq s_0$, then $f(s) = \frac{V_2}{V_1}\left(1-(1-s)^{\frac{V_1}{V_2}}\right) - \left(1-(1-s)^{\frac{V_2}{V_1}}\right) \geq 0$.

We have $\frac{\partial^2 f(s)}{\partial s^2} = \frac{(V_2-V_1)\left(V_1^2(1-s)^{\frac{V_1}{V_2}}+V_2^2(1-s)^{\frac{V_2}{V_1}}\right)}{V_1^2V_2(1-s)^2} > 0$ for $s \in [0,1[$ (since $V_2 > V_1$). Thus, $f(s)$ is convex over $[0,1[$. The first derivative function $\frac{\partial f(s)}{\partial s} = \frac{V_1(1-s)^{\frac{V_1}{V_2}} - V_2(1-s)^{\frac{V_2}{V_1}}}{V_1(1-s)}$. Therefore, $\lim\limits_{s\to 0^+} \frac{\partial f(s)}{\partial s} = \frac{V_1-V_2}{V_1} < 0$, which implies that $f(s)$ is decreasing at the neighborhood of 0. Furthermore, one can verify that $\lim\limits_{s\to 0} f(s) = 0$, and $\lim\limits_{s\to 1} f(s) = \frac{V_2}{V_1} - 1 > 0$. Consequently, over $[0,1]$, $f(s)$ decreases from 0 to reach a negative value, and then increases to reach the positive value $\frac{V_2}{V_1} - 1$. Consequently, equation $f(s) = 0$ has a unique solution s_0 over $[0,1]$, and we have $f(s) \geq 0$ for $s \geq s_0$. Illustrations are given in Figure 2.

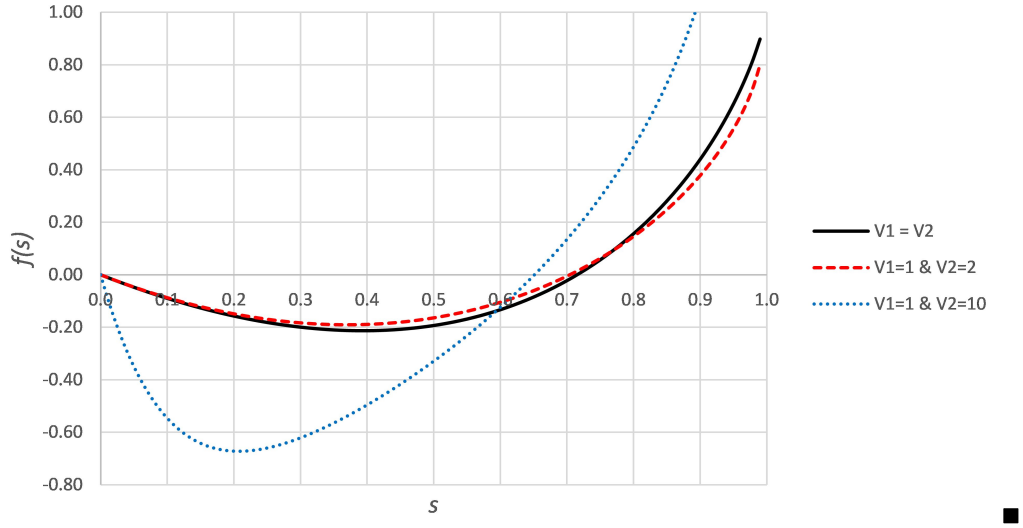

Figure 2. $f(s)$ as a function of s in different situations

If $\mu_1 = \mu_2$ (i.e. $\frac{V_1}{V_2} = 1$) then s_0 is independent of model's parameters since, in this case, s_0 is the unique solution of equation $s - 2(1-s)\ln\left(\frac{1}{1-s}\right) = 0$ over $[0,1]$. Therefore, we solve this equation and obtain $s_0 = s_0(1) = 0.715$. However, when $\mu_1 \neq \mu_2$, s_0 is not constant but depends on the ratio $\frac{V_1}{V_2}$. Given that V_1 and V_2 are variables (as they depend on λ), s_0 cannot be determined beforehand but depends on the model's solution. To overcome this obstacle, we calculate the highest value of s_0 over all possible values of $\frac{V_1}{V_2}$. We denote it by $s_0^{\max}$. It is recalled that $\frac{V_1}{V_2} \in]0,+\infty[$ and that three cases should be distinguished: (i) When $V_1 < V_2$ (i.e. $\frac{V_1}{V_2} \in]0,1[$), s_0 is the unique solution of $\frac{V_2}{V_1}\left(1-(1-s)^{\frac{V_1}{V_2}}\right) - \left(1-(1-s)^{\frac{V_2}{V_1}}\right) = 0$ over $[0,1]$, (ii) When $V_1 > V_2$ (i.e. $\frac{V_1}{V_2} \in]1,+\infty[$, s_0 is the unique solution of $\frac{V_1}{V_2}\left(1-(1-s)^{\frac{V_2}{V_1}}\right) -$

$\left(1-(1-s)^{\frac{V_1}{V_2}}\right) = 0$ over $[0,1]$, and (iii) when $V_1 = V_2$ (i.e. $\frac{V_1}{V_2} = 1$), $s_0 = s_0(1) = 0.715$. Since $s_0^{\max}$ is the highest value of s_0 over all these possible situations, it is calculated as $s_0^{\max} = \max_{\frac{V_1}{V_2} \in]0,+\infty[} s_0(\frac{V_1}{V_2})$. We draw numerically $s_0(\frac{V_1}{V_2})$. Indeed, we vary the value of $\frac{V_1}{V_2}$ with a step of 0.01 and calculate its associated $s_0(\frac{V_1}{V_2})$ with the MATLAB function "fzero". We find that $s_0^{\max} = s_0(1) = 0.715$, as illustrated in Figure 3. Clearly, if $s \geq s_0^{\max}$, then s is greater than any potential value of s_0.

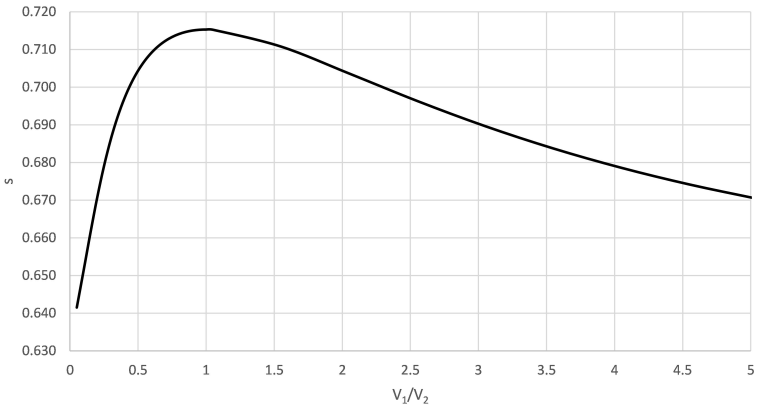

Figure 3. s_0 as a function of $\frac{V_1}{V_2}$

The above analysis leads to the following general result.

Corollary 2 *If $s \geq 0.715$, then satisfying the local service constraints enables to satisfy the global service constraint, i.e., if we have $\Pr(w_1 \leq l_1) \geq s$ and $\Pr(w_2 \leq l_2) \geq s$, then $\Pr(w_1+w_2 \leq l_1+l_2) \geq s$.*

It is important to highlight that this result has a wide applicability since the minimum service level is greater than 0.715 in most practical cases.

4.2 Analytical solving approach

The purpose of this section is to solve model (LSM). We consider $s \geq 0.715$. Thanks to the result obtained in the previous section, it is then possible to simplify model (LSM) by removing the global service constraint (as it is automatically satisfied). We first derive the following Lemma.

Lemma 3 *Both local service constraints are binding and, consequently, we have $1-e^{-(\mu_1-\lambda)l_1} = s$ and $1-e^{-(\mu_2-\lambda)l_2} = s$ at optimality.*

Proof. We prove this result by contradiction. Suppose that we have an optimal solution p^*, l_1^* and l_2^* such that $1-e^{-(\mu_1-\lambda^*)l_1^*} > s$ or $1-e^{-(\mu_2-\lambda^*)l_2^*} > s$. We denote the optimal profit in

this case by Π^*. We can decrease the DTs from l_1^* to l_1', or from l_2^* to l_2' while keeping the price constant until we have a tight service constraint. We denote the new profit by $\Pi^{'}$. This decrease in DTs will lead to a smaller quoted DT $l^{'} = l_1' + l_2'$ and, consequently a higher demand. Since price remains constant, it comes that $\Pi^{'} > \Pi^*$, which is impossible. Hence, both local service constraints are binding. ■

Based on the result of Lemma 3, we deduce that $l_1 = \frac{\ln(1-s)}{\lambda-\mu_1}$ and $l_2 = \frac{\ln(1-s)}{\lambda-\mu_2}$ at optimality. Given in addition that $\lambda = a - \alpha p - \beta\left(l_1 + l_2\right)$, it comes that $p = \frac{a-\beta\left(\frac{\ln(1-s)}{\lambda-\mu_1}+\frac{\ln(1-s)}{\lambda-\mu_2}\right)-\lambda}{\alpha}$. Consequently, model (LSM) can be reformulated as a single-variable optimization model (with λ as the unique variable). We write l_1, l_2, and p as a function of λ and obtain the objective function given in Eq. (13). Constraint (14) imposes a positive demand an ensures the stability condition. Furthermore, to obtain a positive price, we add constraint (15).

$$(LSM)\ \underset{\lambda}{\text{Maximize}}\ \ \Pi(\lambda) = \left(\frac{a - \beta\left(\frac{\ln(1-s)}{\lambda-\mu_1} + \frac{\ln(1-s)}{\lambda-\mu_2}\right) - \lambda}{\alpha} - m_1 - m_2\right)\lambda \tag{13}$$

$$\text{Subject to } 0 \leq \lambda < \min\{\mu_1, \mu_2\} \tag{14}$$

$$a - \beta\left(\frac{\ln\left(1-s\right)}{\lambda-\mu_1} + \frac{\ln\left(1-s\right)}{\lambda-\mu_2}\right) - \lambda \geq 0 \tag{15}$$

Thus, we need to maximize $\Pi(\lambda)$ while satisfying constraints (14) and (15). According to constraint (14), we know that $\lambda \in [0, \min\{\mu_1, \mu_2\}[$. However, constraint (15) adds more restrictions on the value of λ. To identify the feasible domain of λ, we analyze constraint (15) and provide the result in Lemma 4.

Lemma 4 *Equation* $\Psi(\lambda) = a-\beta\left(\frac{\ln(1-s)}{\lambda-\mu_1} + \frac{\ln(1-s)}{\lambda-\mu_2}\right)-\lambda = 0$ *has only one root in* $[0, \min\{\mu_1, \mu_2\}[$. *We let* $\lambda_{\max}$ *denote this root. The feasible domain of model* (LSM) *is given by* $[0, \lambda_{\max}]$.

Particular case: if $\mu_1 = \mu_2 = \mu$, *then* $\lambda_{\max} = \frac{a+\mu-\sqrt{(a+\mu)^2-4(a\mu+2\beta\ln(1-s))}}{2}$.

Proof. We have $\frac{\partial\Psi(\lambda)}{\partial\lambda} = \beta\ln\left(1-s\right)\left(\frac{1}{(\lambda-\mu_1)^2} + \frac{1}{(\lambda-\mu_2)^2}\right) - 1 < 0$ (recall that $\ln\left(1-s\right) < 0$). Hence, $\Psi(\lambda)$ is decreasing in λ. Furthermore, $\Psi(0) = a + \beta\ln\left(1-s\right)\left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right)$. We assume that $a + \beta\ln\left(1-s\right)\left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right) > 0$ since, otherwise, $\Psi(\lambda)$ cannot be positive and, consequently, constraint (15) cannot be satisfied. In real situations, the market potential is generally large enough to have this condition satisfied.

Furthermore, $\lim_{\lambda \to \min\{\mu_1,\mu_2\}} \Psi(\lambda) = -\infty$. Thus, over $[0, \min\{\mu_1, \mu_2\}[$, $\Psi(\lambda)$ decreases monotonously from a positive value to $-\infty$. Therefore, $\Psi(\lambda)$ has only one root $\lambda_{\max}$ in $[0, \min\{\mu_1, \mu_2\}[$, and we have $\Psi(\lambda) \geq 0$ for $\lambda \in [0, \lambda_{\max}]$.

If $\mu_1 = \mu_2 = \mu$, then $\Psi(\lambda) = a - \frac{2\beta \ln(1-s)}{\lambda - \mu} - \lambda$. Thus, $\Psi(\lambda) = 0 \Leftrightarrow \lambda^2 - (a+\mu)\lambda + a\mu + 2\beta \ln(1-s) = 0$. This equation has two roots: $\lambda_{0,1} = \frac{a+\mu-\sqrt{(a+\mu)^2 - 4(a\mu + 2\beta \ln(1-s))}}{2}$ and $\lambda_{0,2} = \frac{a+\mu+\sqrt{(a+\mu)^2 - 4(a\mu + 2\beta \ln(1-s))}}{2}$, and we have $\Psi(\lambda) \geq 0$ before $\lambda_{0,1}$ and after $\lambda_{0,2}$. Given that $a + \beta \ln(1-s)\left(\frac{1}{\mu} + \frac{1}{\mu}\right) > 0$, it comes by standard calculus that $a+\mu > \sqrt{(a+\mu)^2 - 4(a\mu + 2\beta \ln(1-s))}$. Consequently, both of $\lambda_{0,1}$ and $\lambda_{0,2}$ are positive. We know that there is only one root $\lambda_{\max}$ in $[0, \min\{\mu_1, \mu_2\}[$. Hence, $\lambda_{\max} = \min\{\lambda_{0,1}, \lambda_{0,2}\} = \lambda_{0,1} = \frac{a+\mu-\sqrt{(a+\mu)^2 - 4(a\mu + 2\beta \ln(1-s))}}{2}$. ■

Based on the result of Lemma 4, we obtain the following simplified equivalent formulation of model (LSM).

$$(LSM) \quad \underset{\lambda}{\text{Maximize}} \quad \Pi(\lambda) = \left(\frac{a - \beta\left(\frac{\ln(1-s)}{\lambda-\mu_1} + \frac{\ln(1-s)}{\lambda-\mu_2}\right) - \lambda}{\alpha} - m_1 - m_2\right)\lambda \tag{16}$$

$$\text{Subject to } 0 \leq \lambda \leq \lambda_{\max} \tag{17}$$

Note that in the general case (i.e. for $\mu_1 \neq \mu_2$.), we cannot get the closed-form expression of $\lambda_{\max}$. We let $\Omega(\lambda) = \frac{\partial \Pi(\lambda)}{\partial \lambda} = \frac{1}{\alpha}\left(\lambda\beta\left(\frac{\ln(1-s)}{(\lambda-\mu_1)^2} + \frac{\ln(1-s)}{(\lambda-\mu_2)^2}\right) - \beta\left(\frac{\ln(1-s)}{\lambda-\mu_1} + \frac{\ln(1-s)}{\lambda-\mu_2}\right) - 2\lambda + a - (m_1+m_2)\alpha\right)$. If $\mu_1 = \mu_2 = \mu$, then $\Omega(\lambda) = \frac{1}{\alpha}\left(2\mu\beta \ln(1-s) - (2\lambda - a + (m_1+m_2)\alpha)(\lambda-\mu)^2\right)$. We denote by λ_0 a root of equation $\Omega(\lambda) = 0$ in $[0, \lambda_{\max}]$ (if this root exists). We provide in the following Proposition an analytical approach to solve model (LSM) to optimality.

Proposition 2 *If equation $\Omega(\lambda) = 0$ has a root in $[0, \lambda_{\max}]$, then this root is unique.*

- *If this root exists, then the optimal demand rate $\lambda^* = \lambda_0$; otherwise the problem is not relevant since the profit cannot be positive.*
- *Optimal lead times: $l_1^* = \frac{\ln(1-s)}{\lambda_0 - \mu_1}$ and $l_2^* = \frac{\ln(1-s)}{\lambda_0 - \mu_2}$.*
- *Optimal price: $p^* = \frac{a - \beta\ln(1-s)\left(\frac{1}{\lambda_0-\mu_1} + \frac{1}{\lambda_0-\mu_2}\right) - \lambda_0}{\alpha}$.*

Proof. After simplification, we have $\frac{\partial^2 \Pi(\lambda)}{\partial \lambda^2} = \frac{2\beta \ln(1-s)}{\alpha}\left(-\frac{\mu_1}{(\lambda-\mu_1)^3} - \frac{\mu_2}{(\lambda-\mu_2)^3}\right) - \frac{2}{\alpha}$. We remind the reader that we consider the values of λ such that $0 \leq \lambda \leq \lambda_{\max} < \min\{\mu_1, \mu_2\}$, which implies that $(\lambda-\mu_1)^3 \leq 0$ and $(\lambda-\mu_2)^3 \leq 0$. Given, in addition, that $\ln(1-s) < 0$, we conclude that $\frac{\partial^2 \Pi(\lambda)}{\partial \lambda^2} < 0$. Thus, $\Pi(\lambda)$ is strictly concave over $[0, \lambda_{\max}]$.

Given this strict concavity, we deduce that equation $\Omega(\lambda) = 0$ has at maximum only one root in $[0, \lambda_{\max}]$. If $\Omega(\lambda) = 0$ has a root in $[0, \lambda_{\max}]$, then this root represents the optimal demand rate λ^*. Otherwise, the optimal demand is equal to either 0 or $\lambda_{\max}$ (since $\Pi(\lambda)$ is concave), which cannot yield a strictly positive profit. ■

5 Experiments and insights

Model (LSM) gives the DT to be quoted at each stage and enables to efficiently manage and evaluate the performance of each facility. From this managerial perspective, it might be preferred to model (GSM) in practice. However, the interest of model (LSM) could be questionable if this model leads to a significant profit loss with comparison to (GSM). In this section, we first evaluate the profit gap between (LSM) and (GSM). Then, we perform sensitivity analyses to derive insights from the models.

To solve model (LSM), we use the analytical approach of Proposition 2, which provides the optimal solution. As for model (GSM), we use the numerical approach presented at the end of Section 3.

5.1 Profit gap between models (LSM) and (GSM)

We conduct extensive numerical experiments to quantify the profit gap between models (LSM) and (GSM). We consider three different service levels: $s = 95\%$, 97% and 99%. For each s, we generate a large number of test cases, 30720 cases for $\mu_1 = \mu_2$ and 6912 cases for $\mu_1 \neq \mu_2$ as described in Table 2. Therefore, the total number of cases is equal to 92160 and 20736 for $\mu_1 = \mu_2$ and $\mu_1 \neq \mu_2$, respectively.

Table 2. Test cases

Parameter	Test values for $\mu_1 = \mu_2$	Test values for $\mu_1 \neq \mu_2$
a	50, 60, 70, 80, 90, 100	50, 60, 70
α	1, 2, 3, 4, 5, 6, 7, 8	1, 2, 3, 4
β	1, 2, 3, 4, 5, 6, 7, 8	1, 2, 3, 4
m_1	1, 2, 3, 4	1, 2, 3, 4
m_2	1, 2, 3, 4	1, 2, 3, 4
μ_1	10, 20, 30, 40, 50	10, 20, 30
μ_2	$\mu_2 = \mu_1$	10, 20, 30

For each instance, we calculate the gap between the optimal profit of model (LSM) and the one resulting from model (GSM), $Gap_{(LSM)/(GSM)} = \frac{100\times\left(\Pi^*_{(GSM)}-\Pi^*_{(LSM)}\right)}{\Pi^*_{(GSM)}}$. The results are presented in Table 3.

Table 3. Comparison between models (LSM) and (GSM)

s		Profit gap for $\mu_1 = \mu_2$	Profit gap for $\mu_1 \neq \mu_2$
95%	Mean (%)	2.58	1.68
	Standard deviation	4.25	1.02
	Confidence interval (95%)	(2.52, 2.65)	$(1.64, 1.72)$
97%	Mean (%)	3.43	2.13
	Standard deviation	5.12	1.36
	Confidence interval (95%)	$(3.36, 3.50)$	$(2.08, 2.18)$
99%	Mean (%)	4.92	3.06
	Standard deviation	6.57	1.91
	Confidence interval (95%)	$(4.84, 5.01)$	$(2.99, 3.12)$

It is firstly reminded that the profit given by model (GSM) is the highest possible profit. Observing Table 3, it can be concluded from a mathematical point of view that model (LSM) is a good approximation of model (GSM) with a relatively acceptable mismatch, especially when $\mu_1 \neq \mu_2$. We observe that the gap relative to the case of $\mu_1 = \mu_2$ is higher than the one obtained for $\mu_1 \neq \mu_2$, but is still acceptable. Furthermore, Table 3 shows that the higher the service level is, the greater the profit gap becomes. Nevertheless, this gap is still acceptable even for $s = 99\%$ (around 3% for $\mu_1 \neq \mu_2$ and less than 5% for $\mu_1 = \mu_2$).

To better understand the impact of the service level on the profit gap, we vary the value of s and study the profit gap between models (LSM) and (GSM). We consider the following numerical example: $a = 50$, $\alpha = 4$, $\beta = 4$, $m_1 = 2$, and $m_2 = 3$.We respectively illustrate the results in Figures 4 and 5 for the cases ($\mu_1 = \mu_2 = 20$) and ($\mu_1 = 30$ and $\mu_2 = 15$). We respectively restrict our analysis to the values of s greater than 0.715 and 0.667. Indeed, for the cases $\mu_1 = \mu_2$ and $\mu_1 > \mu_2$, the condition of Proposition 1 (i.e. $f(s) \geq 0$) is verified only for $s \geq 0.715$ and $s \geq 0.667$, respectively.

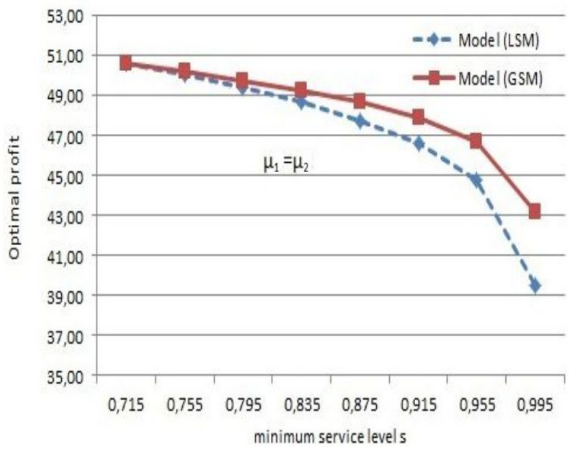

Figure 4. Effect of s ($\mu_1 = \mu_2$)

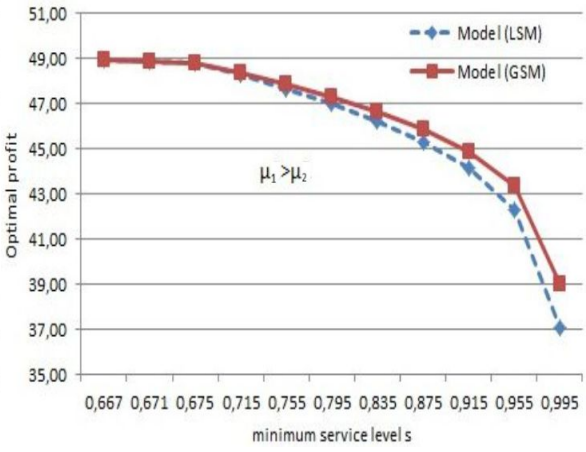

Figure 5. Effect of s ($\mu_1 > \mu_2$)

These experiments illustrate our analytical results. In fact, for the value of s verifying $f(s) = 0$ (i.e. for $s = 0.715$ and $s = 0.667$ in Figures 4 and 5, respectively), models (LSM) and (GSM) are equivalent. Then, an increase in the value of s leads to an increase in $f(s)$ (we refer the reader to Section 4.1 and to Figure 2 for illustration). This implies an increase in the gap between models (LSM) and (GSM) as we can see in Figures 4 and 5.

5.2 Insights from the models

We now study the effect of market characteristics and capacities on optimal decisions and profits. We consider the following basic numerical example, inspired from that used by Pekgün et al. (2008): $a = 50$, $\alpha = 4$, $\beta = 4$, $s = 0.95$, $m_1 = 2$, and $m_2 = 3$. In case of $\mu_1 = \mu_2 = \mu$, we consider $\mu = 20$. In case of $\mu_1 \neq \mu_2$, we focus without lost of generality on the case of $\mu_1 > \mu_2$ (since the problem is symmetric) and consider $\mu_1 = 30$ and $\mu_2 = 15$.

5.2.1 Effect of DT-sensitivity

We vary the DT-sensitivity β from 1 to 8 and report the results in Tables 4 and 5 for $\mu_1 = \mu_2$ and $\mu_1 > \mu_2$, respectively. This range of variation covers most situations since we start with the case where customers are almost insensitive to DT ($\beta = 1$ means that price-sensitivity is 4 times higher than DT-sensitivity) and finally consider the case where customers are very sensitive to DT ($\beta = 8$ means that DT-sensitivity is 2 times higher than price-sensitivity). It is noted that Π^*_{GSM} refers to the best profit obtained for model (GSM), whereas Π^*_{LSM} represents the optimal profit of model (LSM). The real service level realized by the SC is referred to by $s.real$. Obviously, $s.real$ is greater than or equal to s.

Table 4. Effect of DT-sensitivity (μ_1= μ_2= 20)

β	Model (GSM)					Model (LSM)						
	l^*	p^*	λ^*	Π^*_{GSM}	$s.real$	l^*_1	l^*_2	l^*	p^*	λ^*	Π^*_{LSM}	$s.real$
1	0.76	8.88	13.78	53.25	95%	0.46	0.46	0.93	8.88	13.56	52.58	98.25%
2	0.68	8.90	13.04	50.85	95%	0.41	0.41	0.82	8.90	12.73	49.72	98.25%
3	0.63	8.91	12.48	48.76	95%	0.38	0.38	0.76	8.90	12.11	47.27	98.25%
4	0.59	8.90	12.02	46.88	95%	0.36	0.36	0.71	8.89	11.60	45.09	98.25%
5	0.57	8.89	11.62	45.17	95%	0.34	0.34	0.68	8.86	11.16	43.11	98.25%
6	0.54	8.87	11.27	43.59	95%	0.32	0.32	0.65	8.83	10.78	41.29	98.25%
7	0.52	8.85	10.95	42.11	95%	0.31	0.31	0.63	8.80	10.43	39.60	98.25%
8	0.51	8.82	10.66	40.71	95%	0.30	0.30	0.61	8.76	10.11	38.02	98.25%

Table 5. Effect of DT-sensitivity (μ_1= 30 and μ_2= 15)

β	Model (GSM)					Model (LSM)						
	l^*	p^*	λ^*	Π^*_{GSM}	$s.real$	l^*_1	l^*_2	l^*	p^*	λ^*	Π^*_{LSM}	$s.real$
1	1.12	9.18	12.16	50.84	95%	0.17	1.04	1.21	9.16	12.13	50.52	96.32%
2	0.90	9.20	11.42	47.93	95%	0.16	0.82	0.98	9.17	11.36	47.35	96.54%
3	0.79	9.19	10.89	45.59	95%	0.16	0.71	0.87	9.15	10.81	44.80	96.68%
4	0.72	9.16	10.47	43.59	95%	0.15	0.65	0.80	9.11	10.36	42.60	96.77%
5	0.67	9.13	10.11	41.80	95%	0.15	0.60	0.75	9.07	9.98	40.64	96.85%
6	0.63	9.10	9.80	40.18	95%	0.15	0.56	0.71	9.03	9.65	38.86	96.92%
7	0.60	9.06	9.52	38.69	95%	0.15	0.53	0.67	8.98	9.34	37.22	96.97%
8	0.58	9.03	9.26	37.30	95%	0.14	0.51	0.65	8.94	9.07	35.70	97.02%

As expected, we observe in both models that the higher the sensitivity of customers to DT is, the shorter the quoted DT (l^*) becomes. It is also interesting to note that the optimal price is a non-monotonous concave function in DT-sensitivity. Indeed, below a given threshold value, an increase in β leads to an increase in price in both models. Then, above this value, when the customers become more sensitive to DT, both models react by decreasing the price. Indeed, decreasing the price generates more demand and thus offsets the decrease of demand caused by the increase of DT-sensitivity.

We observe that the total DT (i.e. l^*) quoted by model (GSM) is always slightly shorter than that of model (LSM), and that the price of model (GSM) is always slightly higher than that of

model (LSM). Indeed, in model (GSM), the system may satisfy the global service constraint without satisfying the local constraints, which means that model (GSM) has a relatively higher flexibility. This enables model (GSM) to quote a shorter DT and consequently, to offer a higher price.

For model (LSM), we have $l_1^* = \frac{\ln(1-s)}{\lambda_0-\mu_1}$ and $l_2^* = \frac{\ln(1-s)}{\lambda_0-\mu_2}$ (see Proposition 2). Therefore, for $\mu_1 = \mu_2$, we obtain $l_1^* = l_2^*$, and for $\mu_1 \neq \mu_2$, the stage with a higher capacity always quotes a shorter DT than the stage with a lower capacity. This is confirmed by the results of Tables 4 and 5.

We observe that the realized service level obtained with model (LSM) is higher than s. We provide hereafter a qualitative explanation. Indeed, it has been demonstrated that when stage 1 and stage 2 respectively guarantee the DTs l_1 and l_2 with the service level s, then the whole SC can guarantee the DT $l = l_1 + l_2$ with the same s. However, $l = l_1 + l_2$ is not the shortest DT that can be guaranteed by the SC in this case. Thus, the DT quoted by model (LSM) is longer than the shortest possible DT. This explains why $s.real$ is greater than s for model (LSM). It is easy to figure out that the smaller the gap between s and $s.real$ for model (LSM) is, the smaller the gap between the solution of model (LSM) and that of model (GSM) becomes. Our results illustrate this remark.

When $\mu_1 = \mu_2$, it is also important to note that $s.real$ for model (LSM) is always equal to 98.25% whatever the value of β is. This result can be proven analytically. In fact, in case of $\mu_1 = \mu_2 = \mu$, the realized service level is given by $1 - e^{-(\mu-\lambda)l} - (\mu - \lambda)le^{-(\mu-\lambda)l}$ where $l = l_1 + l_2$ and we have at optimality $l_1 = l_2 = \frac{l}{2}$. Given that we considered local service constraints in model (LSM) and that these constraints are tight at optimality (see Lemma 3), we have $e^{-(\mu-\lambda)\frac{l}{2}} = 1 - s$, which implies that $e^{-(\mu-\lambda)l} = (1-s)^2$ and $(\mu-\lambda)l = -\ln\left((1-s)^2\right)$. Thus, it comes that $s.real$ is equal to $1-(1-s)^2+(1-s)^2\ln\left((1-s)^2\right)$. Consequently, in case of $\mu_1 = \mu_2$, the realized service level for model (LSM) depends only on s. For $s = 95\%$, one can verify that $s.real = 98.25\%$ as we have obtained in Table 4. In model (GSM), the global service constraint is always binding (see Lemma 1), which explains why we get $s.real = s = 95\%$ in both tables.

5.2.2 Effect of price-sensitivity

We vary the price-sensitivity α and report the results in Tables 6 and 7 for $\mu_1 = \mu_2$ and $\mu_1 > \mu_2$, respectively.

Table 6. Effect of price-sensitivity ($\mu_1 = \mu_2 = 20$)

α	Model (GSM)					Model (LSM)						
	l^*	p^*	λ^*	Π^*_{GSM}	$s.real$	l^*_1	l^*_2	l^*	p^*	λ^*	Π^*_{LSM}	$s.real$
1	0.94	31.24	14.98	393.08	95%	0.55	0.55	1.09	31.11	14.52	379.07	98.25%
2	0.83	16.22	14.25	159.96	95%	0.48	0.48	0.96	16.18	13.79	154.11	98.25%
3	0.71	11.30	13.29	83.65	95%	0.42	0.42	0.84	11.27	12.83	80.52	98.25%
4	0.59	8.90	12.02	46.88	95%	0.36	0.36	0.71	8.89	11.60	45.09	98.25%
5	0.50	7.52	10.43	26.26	95%	0.30	0.30	0.60	7.50	10.07	25.21	98.25%
6	0.41	6.63	8.55	13.96	95%	0.26	0.26	0.51	6.62	8.26	13.35	98.25%
7	0.35	6.02	6.46	6.59	95%	0.22	0.22	0.44	6.00	6.24	6.26	98.25%
8	0.30	5.57	4.24	2.41	95%	0.19	0.19	0.38	5.56	4.06	2.25	98.25%

Table 7. Effect of price-sensitivity ($\mu_1 = 30$ and $\mu_2 = 15$)

α	Model (GSM)					Model (LSM)						
	l^*	p^*	λ^*	Π^*_{GSM}	$s.real$	l^*_1	l^*_2	l^*	p^*	λ^*	Π^*_{LSM}	$s.real$
1	1.07	33.66	12.05	345.17	95%	0.17	1.00	1.16	33.35	12.00	340.08	96.36%
2	0.96	17.23	11.68	142.90	95%	0.16	0.89	1.05	17.10	11.62	140.50	96.47%
3	0.84	11.81	11.18	76.19	95%	0.16	0.77	0.93	11.73	11.10	74.72	96.61%
4	0.72	9.16	10.47	43.59	95%	0.15	0.65	0.80	9.11	10.36	42.60	96.77%
5	0.60	7.63	9.45	24.88	95%	0.14	0.53	0.67	7.60	9.31	24.22	96.98%
6	0.48	6.67	8.02	13.43	95%	0.14	0.42	0.56	6.65	7.87	13.00	97.19%
7	0.39	6.03	6.22	6.40	95%	0.13	0.34	0.46	6.01	6.06	6.15	97.39%
8	0.33	5.57	4.13	2.36	95%	0.12	0.27	0.39	5.56	3.99	2.22	97.56%

As expected, an increase in price-sensitivity leads to a decrease in the offered price for both models. It is also interesting to see that both models react to an increase in α by quoting a shorter DT. Indeed, this aims to offset the decrease of demand (caused by the increase of α) and consequently, to keep a profitable amount of demand. Similar to the previous observations (see Tables 4 and 5), we see that the price quoted by model (GSM) is always slightly higher than that of model (LSM), and that the total DT quoted by model (GSM) is always slightly shorter than that of model (LSM).

5.2.3 Effect of capacity

Finally, we vary the capacity μ_1 of upstream stage (with μ_2 fixed to 20) and report the results in Table 8. It is recalled that the problem is symmetric in μ_1 and μ_2, so the same results are obtained if we fix μ_1 and vary μ_2.

Table 8. Effect of capacity ($\mu_2 = 20$)

μ_1	Model (GSM)					Model (LSM)						
	l^*	p^*	λ^*	Π^*_{GSM}	$s.real$	l^*_1	l^*_2	l^*	p^*	λ^*	Π^*_{LSM}	$s.real$
10	1.18	9.50	7.27	32.69	95%	1.04	0.23	1.27	9.45	7.11	31.66	96.71
20	0.59	8.90	12.02	46.88	95%	0.36	0.36	0.71	8.89	11.60	45.09	98.25
30	0.47	8.90	12.50	48.77	95%	0.17	0.39	0.56	8.85	12.37	47.56	97.59
40	0.45	8.90	12.61	49.17	95%	0.11	0.40	0.51	8.86	12.53	48.33	96.96
50	0.44	8.90	12.65	49.34	95%	0.08	0.40	0.48	8.87	12.60	48.70	96.55
60	0.43	8.90	12.67	49.42	95%	0.06	0.41	0.47	8.87	12.63	48.91	96.28
70	0.43	8.90	12.69	49.48	95%	0.05	0.41	0.46	8.88	12.65	49.05	96.09
80	0.43	8.90	12.70	49.52	95%	0.04	0.41	0.45	8.88	12.67	49.15	95.95

As expected, an increase in the capacity of one stage enables the SC to quote a shorter DT and consequently, to generate more demand and to increase profit. However, it is important to note that after a threshold value, increasing the capacity has no longer a significant effect neither on DT quotation nor on profit. In Figure 6, we illustrate the variation of optimal profits as a function of μ_1. We observe that the profits increase but with an asymptotic behavior. Furthermore, it is noted that the highest profit gap between models (LSM) and (GSM) is obtained when the capacities of upstream and downstream stages are more or less the same, which is in line with our previous numerical results.

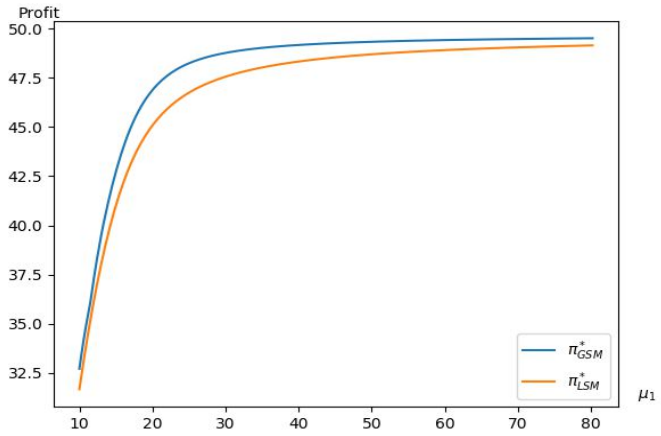

Figure 6. Effect of capacity on optimal profits

For model (LSM), it is important to note that an increase in the capacity of one stage leads to a shorter quoted DT for this stage but to a longer DT for the other stage. In Figure 7, we draw the optimal DTs as a function of μ_1. If we increase μ_1, we see that l_1^* decreases whereas l_2^* increases. Indeed, as the first stage quotes a shorter DT, the customers' demand increases and consequently the second stage, which has a fixed capacity, needs to quote a longer DT to satisfy its local service constraint.

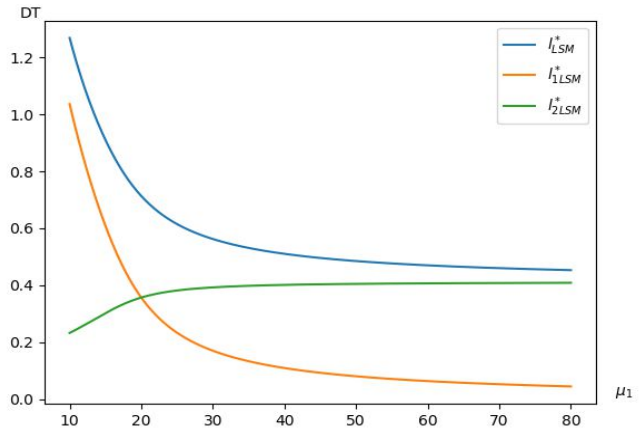

Figure 7. Effect of capacity on DT quotation

6 Model extensions and robustness

In this section, we first extend the local model to consider the case where each stage may target a different service level (s_1 and s_2) and where these service levels are also decision variables to be optimized. Then, we investigate whether the assumption of an exponential service time is a reliable approximation. Indeed, we simulate different service time distributions and compare the obtained profits.

6.1 Variable service levels

In model (LSM), we have considered that both stages target the same service level than the one imposed to the whole SC. We now generalize model (LSM) by considering that each stage may target a different service level and that these service levels are also decision variables. Thus, the firm's problem is to determine the DT and the service level at each stage as well as the price of the product in order to maximize the overall expected profit. We let s_1 and s_2 denote the service levels in upstream and downstream stages, respectively. The new model with variable service levels is denoted by (VSM). We have demonstrated that when the local service constraints are satisfied then, under realistic conditions, the global service constraint is also satisfied if the same service level s is targeted by each stage and by the whole system. This allowed to remove the global service constraint from model (LSM). For model (VSM), however, each stage may target a different service level (s_1 and s_2). In this case, we cannot guarantee that the whole system can satisfy the global service constraint with service level s. Hence, the global service constraint cannot be removed from model (VSM). The formulation of model (VSM) is given below.

$$(VSM) \text{ Maximize } \Pi(l_1, l_2, s_1, s_2, p) = (p - m_1 - m_2)\lambda \tag{18}$$

$$\text{Subject to } \lambda = a - \alpha p - \beta\,(l_1 + l_2) \tag{19}$$

$$1 - e^{-(\mu_1 - \lambda)l_1} \geq s_1 \tag{20}$$

$$1 - e^{-(\mu_2 - \lambda)l_2} \geq s_2 \tag{21}$$

$$\begin{cases} 1 - \frac{\mu_2 - \lambda}{\mu_2 - \mu_1} e^{-(\mu_1 - \lambda)l} + \frac{\mu_1 - \lambda}{\mu_2 - \mu_1} e^{-(\mu_2 - \lambda)l} \geq s \text{ if } \mu_1 \neq \mu_2 \\ 1 - e^{-(\mu - \lambda)l} - (\mu - \lambda) l e^{-(\mu - \lambda)l} \geq s \text{ if } \mu_1 = \mu_2 = \mu \end{cases} \tag{22}$$

$$0 \leq \lambda < \min\{\mu_1, \mu_2\} \tag{23}$$

$$l = l_1 + l_2, l_1 \geq 0, l_2 \geq 0,\ s_1 \in [0, 1[, s_2 \in [0, 1[, p \geq 0 \tag{24}$$

In case of $\mu_1 = \mu_2$, we determine analytically the optimal solution and then compare model (VSM) to models (LSM) and (GSM). When $\mu_1 \neq \mu_2$, we solve model (VSM) numerically. In this case, our main goal is to assess the quality of the solution that has been obtained with model (LSM).

6.1.1 Case of $\mu_1 = \mu_2$

We first consider the case $\mu_1 = \mu_2 = \mu$ and derive the following result.

Proposition 3 *For model (VSM) with $\mu_1 = \mu_2 = \mu$, we have at optimality $s_1^* = s_2^* = r$, where r is the unique root over $[0,1]$ of equation $1-(1-x)^2+2(1-x)^2\ln(1-x)-s=0$. In addition, for $s_1^* = s_2^* = r$, the global service constraint is automatically satisfied.*

Proof. It is firstly noted that the local service constraints are binding and consequently, we have at optimality $1-e^{-(\mu-\lambda)l_1} = s_1$ and $1-e^{-(\mu-\lambda)l_2} = s_2$. In addition, the problem is symmetric, so it can be demonstrated that there is an optimal solution such that $s_1 = s_2$. We let $x = s_1 = s_2$. It comes that $l_1 = l_2 = \frac{-\ln(1-x)}{\mu-\lambda}$, implying that $l = \frac{-2\ln(1-x)}{\mu-\lambda}$. The global service constraint is therefore equivalent to $1-(1-x)^2+2(1-x)^2\ln(1-x)-s \geq 0$. We let $h(x) = 1-(1-x)^2+2(1-x)^2\ln(1-x)-s$. $h^{'}(x) = -4(1-x)\ln(1-x)$, implying that $h^{'}(x) > 0$ over $[0,1]$. Given that $\underset{x\to 0}{lim}\ h(x) = -s < 0$ and $\underset{x\to 1}{lim} h(x) = 1-s > 0$, the equation $h(x) = 0$ admits a unique solution over $[0,1]$. We denote it by r. Thus, the global service constraint is equivalent to $x \geq r$. Since there is not any interest in increasing the local service levels, we have at optimality $s_1^* = s_2^* = r$. ■

Using the result of the previous Proposition, we can replace s_1 and s_2 with their optimal value r and remove the global service constraint from model (VSM). Consequently, in case of $\mu_1 = \mu_2$, model (VSM) becomes equivalent to model (LSM) if we replace s with r in this latter model. Therefore, we use the analytical approach developed earlier for model (LSM) to solve model (VSM) to optimality.

We then conduct extensive numerical experiments to assess the quality of the solution of model (VSM) with comparison to the ones obtained for models (LSM) and (GSM). To perform the experiments, we use the 30720 test cases generated according to Table 2 for each given service level. For each instance, we calculate the following profit gaps:

- Gap between model (LSM) and model (VSM), given by $Gap_{(VSM)/(LSM)} = \frac{100\times\left(\Pi^*_{(VSM)}-\Pi^*_{(LSM)}\right)}{\Pi^*_{(VSM)}}$.
- Gap between model (VSM) and model (GSM), given by $Gap_{(GSM)/(VSM)} = \frac{100\times\left(\Pi^*_{(GSM)}-\Pi^*_{(VSM)}\right)}{\Pi^*_{(GSM)}}$.

We consider three values of s and provide the results in Table 9.

Table 9. Quality of the solution obtained for model (VSM)

Profit gaps		$s = 0.95$	$s = 0.97$	$s = 0.99$
$Gap_{(VSM)/(LSM)}$	Mean	2.45	2.94	4.31
	Std. Deviation	4.11	4.65	6.10
	Conf. Interval (95%)	[2.39, 2.51]	[2.88, 3.00]	[4.23, 4.39]
$Gap_{(GSM)/(VSM)}$	Mean	0.15	0.55	0.71
	Std. Deviation	0.29	0.99	1.85
	Conf. Interval (95%)	[0.145, 0.154]	[0.54, 0.57]	[0.69, 0.72]

We roughly deduce that most values of $Gap_{(VSM)/(LSM)}$ are between 2.39% and 4.39%. In case of $s = 95\%$, the mean gap is 2.45% and most values are between 2.39% and 2.51%. Therefore, model (VSM) slightly outperforms model (LSM) when $\mu_1 = \mu_2$. This gain is due to imposing a lower local service level (r instead of s), which enables model (VSM) to quote a shorter DT and consequently to generate a higher demand and to increase profit.

As for $Gap_{(GSM)/(VSM)}$, most values are between 0.14% and 0.72% and they are even comprised between 0.145% and 0.154% for $s = 95\%$. Hence, the optimal profit obtained with model (VSM) is very close to that of model (GSM), which is the highest profit that can be achieved by the firm. It is important to note that model (VSM) cannot achieve the optimal profit of model (GSM) in the general case. The reason is that model (VSM) has local service constraints and assumes that the overall DT quoted to customers is equal to the sum of local DTs while these constraints are not considered in model (GSM). Thus, the optimal solution of model (GSM) may be infeasible for model (VSM) which implies that model (GSM) generally leads to a higher profit than that of model (VSM).

In case of $\mu_1 = \mu_2$, model (VSM) presents many advantages. First, it can be solved analytically to optimality. Second, it can be easily implemented in practice and has managerial advantages (as explained earlier). Third, it yields a high profit that is very close to the one obtained with model (GSM). Consequently, model (VSM) can be considered as the best alternative in case of $\mu_1 = \mu_2$.

6.1.2 Case of $\mu_1 \neq \mu_2$

Now, we focus on the case of $\mu_1 \neq \mu_2$. Since it is not possible to obtain an analytical solution, our main objective here is to assess the quality of the solution obtained with model (LSM). In

other words, we aim to quantify the loss resulting from imposing $s_1 = s_2 = s$. To solve model (VSM), we use a numerical approach based on an iterative procedure. Indeed, we firstly ignore the global constraint and test, for a given s, different combinations of s_1 and s_2 starting from 0.90 until s with a step of 0.01 (e.g., $s_1 = 0.90$ and $s_2 = 0.91$, $s_1 = 0.90$ and $s_2 = 0.92$,..., $s_1 = 0.94$ and $s_2 = 0.92$, etc.). More precisely, for each combination we solve the resulting model with Matlab and obtain in particular l_1 and l_2. Then, we verify for each combination whether the global constraint (with $l = l_1 + l_2$) is satisfied. The combinations that do not satisfy the global constraint are rejected, and we select the combination that yields the highest profit while satisfying the global constraint. It is recalled that the profit gap between models (VSM) and (LSM) is given by $Gap_{(VSM)/(LSM)} = \frac{100\times\left(\Pi^*_{(VSM)} - \Pi^*_{(LSM)}\right)}{\Pi^*_{(VSM)}}$. For each given service level s ($s = 0.95$, 0.97 and 0.99), we generate 6912 test cases according to the procedure described in Table 2. The comparison results are reported in Table 10.

Table 10. Comparison between models (VSM) and (LSM) when $\mu_1 \neq \mu_2$

Profit gap		$s = 0.95$	$s = 0.97$	$s = 0.99$
$Gap_{(VSM)/(LSM)}$	Mean	1.57	2.01	2.76
	Std. Deviation	1.00	1.36	1.93
	Conf. Interval (95%)	[1.53, 1.61]	[1.96, 2.06]	[2.69, 2.83]

As expected, model (VSM) dominates model (LSM) in terms of profit. However, in most cases, the gap is between 1.53% and 2.83%. In addition, in the industrial sectors where s is not very high, the gap can even be smaller than 2% as indicated in Table 10. This confirms the interest of model (LSM) since this model is tractable, can be easily implemented, presents managerial advantages, and does not lead to a significant loss with comparison to (GSM) and (VSM).

6.2 Other service times distributions

Our models assume an exponential service time at each stage which is of course an approximation. As this assumption enables to develop tractable models, it has been widely adopted in the literature, in general, and in the vast majority of papers on DT quotation, in particular. The random aspect of service times is of course realistic in several situations. In addition, many authors argued that there is often a high level of variability with respect to service times (e.g., Kingsman et al. 1998, Haskose et al. 2004). In these cases, the exponential distribution

might be relevant because it has a high coefficient of variation. Nevertheless, when there is a small service time variability, the assumption of an exponential service time is often pessimistic. With comparison to the exponential distribution, the consideration of another service time distribution that has a smaller variance is expected to yield a shorter quoted DT (for the same service level) which consequently generates a higher demand and more profit. In this section, we assess the amount of demand loss (and, consequently, profit loss) caused by the exponential assumption.

Thus, we compare the profit of model (GSM) with exponential distribution (coefficient of variation equal to 1) to the profits obtained when we change the variability of the service time distribution but without changing the mean service time. We test the Erlang-2 distribution (coefficient of variation equal to $\frac{1}{\sqrt{2}}$) and the extreme case of a deterministic distribution. We proceed as follows:

- We solve model (GSM) for the basic numerical example ($a = 50$, $\alpha = 4$, $\beta = 4$, $m_1 = 2$, $m_2 = 3$, and $s = 95\%$) with $\mu_1 = \mu_2 = 20$ and $\mu_1 = 60$ and $\mu_2 = 30$, and get the associated profits.

- We then simulate the system, first with the Erlang-2 and then with the deterministic distribution, in order to find the new mean demand rate λ that satisfies the service level $s = 95\%$ (the price being fixed to its value obtained with model (GSM)).

- The new mean demand rate, of course higher than the one obtained with the exponential distribution (i.e. for model (GSM)), is used to calculate the new profit.

We report the results in Table 11.

Table 11. Robustness to exponential assumption

	$\mu_1 = \mu_2 = 20$		$\mu_1 = 60$ and $\mu_2 = 30$	
	λ	Profit	λ	Profit
Exponential	12.02	46.88	10.47	43.59
Erlang-2	12.52	48.83	10.93	45.50
Deterministic	13.28	51.80	11.47	47.75

We observe in the balanced case that the exponential assumption leads to a loss of demand of 4.0% compared to the Erlang-2 distribution and 9.5% compared to the deterministic case, and therefore a profit loss of respectively 4.0% and 9.5%. For the unbalanced case, the values

of the different gaps (for both demand and profit) become 4.2% with the Erlang-2 distribution and 8.7% with the deterministic distribution.

The results obtained with the exponential distribution are of course pessimistic but remain interesting with errors of less than 10% on the demands and profits even in the deterministic case which is obviously very far from the exponential case. With comparison to the Erlang-2, the gaps are even less than 5%. These results demonstrate in our opinion the interest of our models even if it would be interesting to investigate other cases with other types of variabilities that may be very different from the exponential distribution.

7 Conclusion

We considered the problem of DT quotation and pricing in a two-stage MTO supply chain modeled as a tandem queue $M/M/1 - M/M/1$ and facing a DT- and price-sensitive demand. We addressed this problem with two different managerial approaches. In the first approach (global model), a pair of price and DT are quoted to customers to maximize the expected overall profit while satisfying a global service level. In the second approach (local model), a DT is quoted at each stage while satisfying a local service level, and the DT quoted to customers, calculated as the sum of local DTs, must satisfy the global service level. When both stages target the same service level than the one imposed to the whole system, we demonstrated under realistic conditions that satisfying the local service constraints enables to satisfy the global service constraint. Based on this result, we simplified the local model and solved it to optimality with an analytical approach.

The local model presents several managerial and mathematical advantages but leads to a smaller profit with comparison to the global model. We quantified this profit loss and showed that it is relatively small, especially when the stages do not have the same capacity or when the service level is not too high. Thus, we deduced that the local model can also be used as an approximation of the global model, which is interesting since the local model was solved analytically. Then, we conducted sensitivity analyses with both models and derived insights. As expected, we found that an increase in DT-sensitivity (respectively, price-sensitivity) leads to quoting a shorter DT (respectively, offering a lower price). Less expected are the facts that the optimal price is a non-monotonous concave function in DT-sensitivity and that both models react to an increase in price-sensitivity by quoting a shorter DT. We also found that the total

DT quoted by the global model is always slightly shorter than that of the local model, and that the price of the global model is always slightly higher than that of the local model. It was also interesting to investigate the effect of capacities since our work is the first to consider two operations stages impacting the DT quotation with finite capacities in both stages. We showed that an increase in the capacity of one stage leads to a shorter quoted DT for this stage but a longer DT for the other stage and, consequently, does not necessarily lead to a shorter DT quoted to the customers. Beyond a threshold value, increasing the capacity has no longer a significant effect neither on DT quotation nor on the overall profit.

Finally, we extended the local model to consider the case where each stage may target a different local service level and where these service levels are also decision variables. In case of balanced capacity, we solved the model analytically and showed that its resulting optimal profit is very close to the one obtained with the global model. In case of unbalanced capacity, we showed numerically that considering different service levels slightly improves the profit of the local model. We also studied the robustness of our results to the assumption of exponential service times and showed numerically that the exponential assumption can be a good approximation.

This research studied a centralized setting with only one decision maker for the two stages. Thus, a natural extension would be to investigate the decentralized setting where each actor optimizes his own profit. Extending the model to more than two stages can also be the focus of future works but is expected to be a very hard problem, especially in case of different service levels between the stages. In this case, a solving approach based on metaheuristics could be used.

References

Albana, A.S., Frein, Y., Hammami, R., 2018. Effect of a lead time-dependent cost on lead time quotation, pricing, and capacity decisions in a stochastic make-to-order system with endogenous demand. International Journal of Production Economics, 203, 83 - 95.

Boyaci, T., Ray, S., 2003. Product differentiation and capacity cost interaction in time and price sensitive markets. Manufacturing & Service operations management, 5, 18 - 36.

Boyaci, T., Ray, S., 2006. The impact of capacity costs on product differentiation in delivery time, delivery reliability, and price. Production and Operations Management, 15, 179 - 197.

Çelik, S., Maglaras, C., 2008. Dynamic pricing and lead-time quotation for a multiclass

make-to-order queue. Management Science, 54, 1132 - 1146.

Hammami, R., Frein, Y., 2014. A Capacitated Multi-echelon Inventory Placement Model under Lead Time Constraints. Production and Operations Management, 23, 446 - 462.

Hammami, R., Frein, Y., Bahli, B., 2017. Supply chain design to guarantee quoted lead time and inventory replenishment: model and insights. International Journal of Production Research, 55(12), 3431-3450.

Haskose, A., Kingsman, B. G., Worthington, D., 2004. Performance analysis of make-to-order manufacturing systems under different workload control regimes. International journal of production economics, 90(2), 169-186.

Huang, J., M. Leng, M. Parlar, 2013. Demand functions in decision modeling: A comprehensive survey and research directions. Decision Sciences, 44(3), 557 - 609

Kingsman, B.G., Tatsiopoulos, I.P., Hendry, L.C., 1998. A structural methodology for managing manufacturing lead times in make-to-order companies. European Journal of Operational Research, 40, 196 - 209.

Liu, L., Parlar, M., Zhu, S.X., 2007. Pricing and lead time decisions in decentralized supply chains. Management Science, 53, 713 - 725.

Palaka, K., Erlebacher, S., Kropp, D.H., 1998. Lead-time setting, capacity utilization, and pricing decisions under lead-time dependent demand. IIE Transactions, 30, 151 - 163.

Pekgün, P., Griffin, P.M., Keskinocak, P., 2008. Coordination of marketing and production for price and leadtime decisions. IIE Transactions, 40, 12 - 30.

Pekgün, P., Griffin, P.M., Keskinocak, P., 2017. Centralized versus Decentralized Competition for Price and Lead-Time Sensitive Demand, Decision Sciences, 48 (6), 1198-1227

Ray, S., Jewkes, E.M., 2004. Customer lead time management when both demand and price are lead time sensitive. European Journal of Operational Research, 153, 769 - 781.

So, K.C., Song, J.-S., 1998. Price, delivery time guarantees and capacity selection. European Journal of Operational Research, 111, 28 - 49.

Xiao, T., Qi, X., 2016. A two-stage supply chain with demand sensitive to price, delivery time, and reliability of delivery. Annals of Operations Research, 241, 475 - 496.

Zhao, X., Stecke, K.E., Prasad, A., 2012. Lead time and price quotation mode selection: uniform or differentiated? Production and Operations Management, 21, 177 - 193.

Zhu, S.X., 2015. Integration of capacity, pricing, and lead-time decisions in a decentralized supply chain. International Journal of Production Economics, 164, 14 - 23.