Pricing decision and lead time quotation in supply chains with an endogenous demand sensitive to lead time and price

Abduh-Sayid Albana

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **GI : Génie Industriel**

Arrêté ministériel : 25 mai 2016

Présentée par

Abduh-Sayid ALBANA

Thèse dirigée par **Yannick FREIN, Professeur à Grenoble INP**
et codirigée par **Ramzi HAMMAMI, Professeur à Rennes School of Business**

préparée au sein du **Laboratoire des Sciences pour la Conception, l'Optimisation et la Production de Grenoble (G-SCOP)**
dans **l'École doctorale Ingénierie - Matériaux, Mécanique, Environnement, Energétique, Procédés, Production (I-MEP2)**

Choix du prix et du délai de livraison dans une chaîne logistique avec une demande endogène sensible au délai de livraison et au prix

Pricing Decision and Lead time Quotation in Supply Chains with an Endogenous Demand Sensitive to Lead time and Price

Thèse soutenue publiquement le **26/01/2018**,
devant le jury composé de :

Monsieur Lyes BENYOUCEF
Professeur à Aix-Marseille Université, Président
Monsieur Zied JEMAI
Professeur à Ecole Nationale d'Ingénieurs de Tunis, Rapporteur
Monsieur Faicel HNAIEN
Maître de conférences à l'Université de Technologie de Troyes, Rapporteur
Monsieur Jean-Philippe GAYON
Professeur à ISIMA Clermont-Ferrand, Examinateur
Monsieur Yannick FREIN
Professeur à Grenoble INP, Directeur de thèse
Monsieur Ramzi HAMMAMI
Professeur à Rennes School of Business, Co-Encadrant de thèse

Contents

List of Figures

List of Tables

Résumé en français

Chapitre 1 : Introduction

La majeure partie des travaux sur la conception d'une chaîne logistique supposent une demande exogène, c'est-à-dire connue a priori (éventuellement par une caractérisation stochastique) et donc indépendante des éventuelles décisions prises lors de cette conception. Ces travaux supposent donc en particulier que le prix, facteur influençant fortement la demande, est déjà fixé. Mais, ces travaux supposent aussi implicitement que la demande n'est pas sensible au délai de livraison ou alors que ce délai est aussi déjà fixé. Or, ces deux hypothèses sont bien sûr très discutables comme nous l'expliquons ci-après. Précisons tout d'abord que nous définissons le délai de livraison L comme étant le temps entre l'instant où le client passe sa commande et le temps où le produit est disponible pour ce client (Christopher, 2011).

Tout d'abord, il est connu depuis longtemps que le délai de livraison proposé aux clients est un facteur de compétitivité essentiel, et même une clé du succès dans de nombreuses industries (Blackburn et Stalk, 1990). La littérature fournit de nombreuses illustrations sur la façon dont les entreprises peuvent utiliser ce délai de livraison comme une arme stratégique pour obtenir un avantage concurrentiel (Blackburn et al., 1992; Hum et Sim, 1996; Suri, 1998). Geary et Zonnenberg (2000), après une enquête auprès de 110 entreprises dans cinq grands secteurs manufacturiers, relatent que nombre d'entre elles focalisent leurs efforts sur des améliorations sur les coûts, mais aussi sur les délais de livraison. Baker et al. (2001) indiquent que moins de 10% des clients finaux (BtC) et moins de 30% des clients BtB basent leurs décisions d'achat sur uniquement le prix de vente d'un article.

De plus, il est évident que le délai de livraison est largement impacté par les décisions prises lors de la conception de la chaîne. En effet, les décisions de localisation des lieux de production et d'achat des composants ont bien sûr un effet direct sur le délai de production. Les politiques de stockage (quantités et lieux de stockage) ont aussi un effet direct sur la disponibilité des produits à livrer. Il nous paraît donc intéressant de travailler sur des modèles dans lesquels on prend explicitement en compte l'impact du délai de livraison sur la demande. Au-delà de la réduction du délai, il est probablement encore plus important de satisfaire le délai annoncé. La non satisfaction du délai indiqué peut conduire à de fortes pénalités. Selon Savaşaneril et al. (2010), les exemples montrant l'importance de délais de livraison fiables sont abondants dans l'industrie. Les auteurs rapportent par exemple que le coût de livraisons tardives dans la division des équipements de FMC Wellhead pourrait atteindre $ 250 000 par jour et que les pénalités de retard dans l'industrie aéronautique vont de $ 10 000 à $ 15 000 et peuvent atteindre $ 1 000 000

par jour. En plus des conséquences directes en termes de pénalités, une livraison en retard peut affecter la réputation de l'entreprise et dissuader les clients futurs (Slotnick, 2014). Les entreprises risquent donc de perdre des marchés si elles ne sont pas capables de respecter les délais promis (Kapuscinski et Tayur, 2007). Il s'agira donc de développer des modèles dans lesquels nous choisissons le délai annoncé aux clients mais aussi garantissons un niveau de respect de ces délais annoncés.

Un temps de livraison plus court peut conduire à une augmentation de la demande, mais augmente également le risque de retard de livraison, et donc détériore le niveau de service et augmente le coût de la pénalité de retard. Inversement, la stratégie consistant à augmenter le délai annoncé conduit à une demande plus faible; cela conduira en effet les clients à commander aux concurrents qui proposent des délais de livraison plus courts (Ho et Zheng, 2004; Pekgün et al., 2016; So, 2000; Xiao et al., 2014). Et bien sûr, une diminution de la demande aura un effet négatif sur les revenus. Par contre, le côté positif de promettre un délai plus long est la possibilité d'atteindre un niveau de service plus élevé (puisque, d'une part, le délai promis est plus long et, d'autre part, la demande est plus faible). Il peut également diminuer le coût de stockage des stocks en cours. Ce dernier coût peut être significatif dans de nombreuses industries telles que l'automobile et l'électronique. La fixation du délai est donc en soi le résultat d'un compromis. Par ailleurs, même si nous avons insisté dans ce qui précède sur l'influence du délai, car encore peu abordé dans la littérature, il est bien connu que l'augmentation du prix réduit la demande mais augmente la marge et qu'une baisse du prix augmente cette demande mais au détriment de la marge. Il y a donc aussi un compromis à trouver dans cette fixation du prix. Il est donc évident que la combinaison du prix proposé et du délai promis ouvre sur de nouveaux compromis et offre des possibilités pour de nombreux travaux novateurs.

Il est intéressant de noter que très peu de recherches en gestion des opérations ont été menées dans ce cadre d'une demande sensible aux prix et délai de livraison, comme le soulignent Huang et al. (2013). La majorité de cette littérature se situe dans le contexte «Make-To-Order» (pour les articles en «Make-To-Stock» voir Panda, 2013; Savaşaneril et al., 2010; Savasaneril et Sayin, 2017; Wu et al., 2012). Le papier pionnier sur un modèle avec une demande sensible aux délais et prix dans le contexte du MTO est le papier de Palaka et al. (1998). Dans ce papier, la demande est supposée être une fonction linéaire du prix et du délai. Ils considèrent 3 variables de décision : le délai promis, la capacité de production et le prix. Ils limitent tout d'abord leur attention à un horizon court terme, et par conséquent la capacité est supposée constante. Les clients sont servis selon le principe du premier arrivé, premier servi. Ils supposent que le processus d'arrivée des clients peut être décrit par un processus de Poisson. En outre, les temps de traitement des commandes des clients sont supposés distribués de manière exponentielle. Ces hypothèses leur permettent d'utiliser un modèle M/M/1 pour représenter les opérations de l'entreprise. Dans la suite de ces travaux, différentes extensions ont été effectuées dans des cadres mono et multi-entreprise, et nous nous positionnons clairement dans ce courant de

la littérature. Avant de préciser nos contributions, nous analysons rapidement les extensions les plus importantes issues du travail fondateur de Palaka et al. (1998). Dans le cas mono-entreprise, Pekgün et al. (2008) ont étudié deux modes de prise des décisions prix et délai, en l'occurrence centralisé et décentralisé. Ils ont utilisé le cadre de Palaka et al. (1998) pour leur modèle centralisé mais sans tenir compte des coûts de stockage et de pénalité. Ray et Jewkes (2004) se focalisent sur la recherche du délai promis optimal dans une situation où le prix est sensible au délai de livraison. Zhao et al. (2012) ont également considéré cette problématique de fixation du délai et du prix dans les entreprises de services et les industries «Make-To-Order». Ils ont considéré deux stratégies : dans la première stratégie, les entreprises proposent un délai et prix uniques («uniform quotation mode») et, dans la seconde, ils proposent un menu de délais et de prix («differentiated quotation mode»). Dans le cadre multi-entreprise, Zhu (2015) considère une chaîne logistique composée d'un fournisseur et d'un détaillant face à une demande sensible aux prix et délais. Il est important de signaler que le détaillant n'a pas d'opération de production et donc pas de délai propre. Le processus de décision est modélisé comme une séquence où le fournisseur détermine la capacité et le prix de vente au détaillant, et le détaillant détermine le prix de vente et le délai de livraison. La récente recherche de Pekgün et al. (2016) est une extension de leur papier précédent (Pekgün et al., 2008). Dans ce papier récent, ils étudient deux entreprises qui se font concurrence sur les décisions de prix et de délais dans un marché commun. Une discussion détaillée de la littérature considérant une demande sensible au délai et prix est fourni dans le chapitre 2.

Dans notre étude de la littérature, nous avons trouvé quelques faiblesses. Dans tous les travaux considérés, le coût de production unitaire est supposé être constant. De plus, nous n'avons pas trouvé de travaux dans lesquels des clients peuvent être rejetés (notamment si l'entreprise a un carnet de commandes rempli). Enfin, les quelques travaux considérant plusieurs entreprises se situent dans le cadre de 2 entreprises où une seule a des opérations de production (l'autre acteur a un délai nul). Concernant la première limitation, on sait que dans de nombreuses situations le coût de production unitaire dépend du délai de livraison promis. Les entreprises peuvent en effet mieux gérer le processus de production et réduire les coûts de production lorsqu'elles disposent d'un délai plus élevé. Bien sûr, la prise en compte d'un coût de production sensible au délai pose de nouvelles difficultés surtout si on considère l'hypothèse réaliste d'une relation non linéaire entre le coût et le délai. L'hypothèse qui consiste à accepter tous les clients permet à ces travaux de considérer un modèle M/M/1 ce qui est bien sûr intéressant d'un point de vue résolution analytique. Mais, accepter les clients, même avec un nombre élevé de commandes déjà en attente, peut entraîner de longs délais pour ces clients et donc nécessite de définir un délai important si on veut satisfaire un niveau de service élevé, ce qui conduira à une réduction de la demande. Pratiquement, les entreprises peuvent choisir de rejeter les clients lorsqu'elles ont déjà trop de clients. Mais alors, le modèle obtenu sera du type M/M/1/K et la formulation du temps d'attente résiduel,

nécessaire pour calculer le coût de la pénalité pour les clients en retard, n'est pas disponible dans la littérature. Enfin, il est bien sûr intéressant de considérer une chaîne logistique comprenant plusieurs acteurs et contrairement à ce qui est présenté dans la littérature, il faudrait étudier une situation où chaque acteur a ses propres opérations de production et donc ses propres délais. Mais à nouveau, un tel modèle pose des difficultés nouvelles. En effet, le modèle global est plus complexe et si on souhaite étudier un modèle avec décentralisation des décisions, on se heurte à la difficulté de savoir imposer un taux de service global (qui intéresse le client final) à partir des contraintes de service locales.

Le plan de la thèse découle naturellement de l'analyse de ces limitations. En effet, après une étude de littérature (chapitre 2) nous proposons trois extensions: 1. Coût unitaire de production sensible au délai, 2. Politique de rejet de clients à l'aide d'un modèle M/M/1/K, et 3. Etude de la coordination d'une chaîne logistique composée de deux étages à l'aide d'un réseau tandem de type (M/M/1-M/M/1), extensions développées dans les chapitres 3, 4 et 5, respectivement. Enfin, nous concluons notre travail dans le chapitre 6.

Chapitre 2 : Revue de littérature

Comme indiqué dans l'introduction, le papier pionnier est celui de Palaka et al. (1998). Dans ce papier, les auteurs se sont intéressés au choix du délai, de la capacité et du prix pour une entreprise où les clients sont sensibles aux délais promis. Palaka et al. (1998) considèrent une entreprise qui produit avec un mode «make-to-order». Ils limitent initialement leur étude à un horizon court terme, par conséquent la capacité est supposée constante alors que le prix, le délai promis et la demande sont considérés comme des variables de décision. Comme indiqué en introduction, ils ont modélisé le système par une file d'attente de type M/M/1. Les clients sont sensibles aux délais et prix, et la demande est naturellement supposée être décroissante à la fois en fonction du prix et du délai promis. Plus précisément, la demande maximale est une fonction linéaire et modélisée comme suit:

$$\Lambda(P, L) = a - b_1 P - b_2 L \tag{1}$$

où P = prix du bien/service établi par l'entreprise, L = délai promis, $\Lambda(P, L)$ = demande maximale attendue pour le bien/service au prix P et délai promis L, a = demande maximale correspondant à un prix et un délai promis nuls, b_1 = sensibilité de la demande au prix, et b_2 = sensibilité de la demande au délai promis (b_1 et b_2 sont positifs). De plus, pour éviter des délais de livraison promis irréalistes, ils imposent que l'entreprise maintienne un niveau de service minimum (s), où ce niveau de service est défini comme la probabilité de satisfaire le délai promis. Ce niveau de service minimum peut être fixé par l'entreprise elle-même en réponse aux pressions concurrentielles.

L'entreprise étant modélisé par une M/M/1 avec un taux de service moyen, μ, et un taux d'arrivée moyen, λ, le nombre de clients moyen du système, N_s, est

donné par $N_s = \lambda/(\mu - \lambda)$ et le temps de séjour dans le système, W, est distribué de façon exponentielle avec une moyenne $1/(\mu - \lambda)$ (Hillier et Lieberman, 2001; Kleinrock, 1975). La probabilité que l'entreprise ne respecte pas le délai de livraison promis, L, est donnée par $e^{-(\mu-\lambda)L}$ et le retard moyen d'une commande en retard est de $1/(\mu - \lambda)$, identique au temps de séjour moyen en raison de la propriété sans mémoire de la distribution exponentielle.

L'objectif de l'entreprise est de maximiser le profit total attendu qui peut être exprimée par l'équation (2) ci-après. Dans la fonction objectif, $\lambda(P - m)$ représente le revenu prévu (net des coûts directs), où m est le coût unitaire. Les coûts de congestion moyens sont donnés par $F\lambda/(\mu - \lambda)$ où F est le coût de stockage unitaire et $\lambda/(\mu - \lambda)$ est le nombre moyen de clients dans le système. La pénalité de retard moyenne est donnée par $c_r(\lambda/(\mu - \lambda))e^{-(\mu-\lambda)L}$, où c_r est la pénalité par commande pour une unité de temps de retard, le nombre du client en retard étant égal à $\lambda e^{-(\mu-\lambda)L}$, et le retard moyen étant égal à $1/(\mu - \lambda)$. Enfin, en notant s le niveau de service minimum, Palaka et al. (1998) formulent le problème d'optimisation comme suit:

$$(P_{Base}) \underset{P,L,\lambda}{Maximiser} \quad \Pi(P, L, \lambda) = (P - m)\lambda - \frac{F\lambda}{\mu - \lambda} - \frac{c_r\lambda}{\mu - \lambda}e^{-(\mu-\lambda)L} \tag{2}$$

$$\text{Sous contraintes} \quad \lambda \leq a - b_1P - b_2L \tag{3}$$

$$1 - e^{-(\mu-\lambda)L} \geq s \tag{4}$$

$$0 \leq \lambda \leq \mu \tag{5}$$

$$P, L \geq 0 \tag{6}$$

La contrainte (3) exige que la demande moyenne, λ, desservie par l'entreprise ne dépasse pas la demande générée par le prix, P et le délai indiqué, L. La contrainte (4) exprime la limite inférieure du niveau de service. La contrainte (5) correspond à la restriction selon laquelle la demande moyenne, λ, est également limitée par le taux de service de l'entreprise, μ. La contrainte (6) exprime les contraintes de positivité des variables.

Dans leur papier, Palaka et al. (1998) ont montré que la contrainte (3) est serrée à l'optimalité. L'entreprise choisira le prix, P, le délai promis, L et le taux de demande, λ, de sorte que la contrainte $\lambda \leq \Lambda(P, L)$ soit en fait: $\lambda = \Lambda(P, L)$. Le problème d'optimisation est donc en fait un problème à 2 variables.

Par contre la contrainte de service (4) dans le modèle d'optimisation (P_{Base}) n'est pas obligatoirement serrée à l'optimalité. Palaka et al. (1998) ont démontré que la contrainte de service (4) n'est pas serrée si le niveau de service, s, est strictement inférieur à une valeur critique, s_c, c'est-à-dire $s < 1 - b_1/(b_2c_r)$. En outre, le niveau de service effectif sera donné par $\max(s, s_c)$ (voir la proposition 2 de Palaka et al. (1998)).

Les solutions du problème P_{Base} dans les cas serré et non serré, sont:

- La demande optimale λ^* est donnée par la racine de l'équation cubique ci-

dessous sur l'intervalle $[0, \mu]$:

$$(a - mb_1 - 2\lambda)(\mu - \lambda)^2 = G\mu$$

Où $G = b_2 \log x + Fb_1 + c_r b_1/x$ et $x = \max\{1/(1-s), b_1 c_r/b_2\}$,

- Le délai promis optimal L^* est donné par $(\log x)/(\mu - \lambda^*)$, et
- Le prix optimal, P^*, est obtenu en utilisant la relation $P^* = (a - \lambda^* - b_2 L^*)/b_1$.

Ce modèle de Palaka et al. (1998) a été à l'origine de nombreux travaux sur les modèles de demande sensible au délai dans des entreprises de type MTO.

Dans le cas mono entreprise on peut tout d'abord citer des papiers qui proposent des délais différenciés au client (Boyaci et Ray, 2006 et 2003; Çelik et Maglaras, 2008; Hafizouglu et al., 2016; Zhao et al., 2012). Pekgün et al. (2008) s'intéresse à la coordination de deux services d'une seule entreprise pour décider le prix et le délai promis. On peut aussi citer Ray et Jewkes (2004) qui modélisent un prix sensible au délai de livraison, et le papier de So et Song (1998) qui utilisent un modèle de demande log-linéaire. Des descriptions plus détaillées de ces papiers sont fournies dans la section 2.2 de la thèse.

Dans le cas multi-entreprise, avec des entreprises en compétition, nous avons les contributions de Ho et Zheng, (2004); Li, (1992); So, (2000); Xiao et al., (2014); Xiaopan et al., (2014); et Pekgün et al. (2016). Toujours pour le cas multi-entreprise mais dans des cas d'entreprises en coopération, nous avons Liu et al., (2007); Xiao et al., (2011); Xiao et Shi, (2012); Zhu, (2015); et Xiao et Qi, (2016). Des discussions détaillées sont fournies dans la section 2.3 de la thèse.

De cette revue de la littérature, nous avons pu faire les observations suivantes. Tout d'abord, tous les travaux supposent un coût unitaire de production constant. Or pratiquement, les entreprises peuvent mieux gérer leur système de production lorsqu'elles proposent un long délai. Ainsi, nous avons fait une contribution en modélisant le coût de production en fonction du délai de livraison promis. Par ailleurs, dans tous ces travaux tous les clients sont acceptés. Or, cela peut entraîner de longs délais dans le système dans certains cas, et donc nous avons étudié une politique avec possibilité de rejets de clients. Enfin, pour les travaux multi-entreprises, il est toujours supposé qu'un des acteurs agit uniquement comme un médiateur avec un délai de livraison égal à zéro. Nous avons donc étudié un système en tandem M/M/1-M/M/1. Notre travail aura pour cadre des systèmes MTO, proposant un produit unique avec un prix unique, et un modèle de demande linéaire. Nous proposerons trois contributions:

- Introduire un coût de production unitaire qui dépend du délai promis,
- Considérer une politique de rejet des clients en utilisant la file M/M/1/K,
- Introduire une chaîne logistique (multi-entreprise) où les deux acteurs ont un processus de production qui mène à réseau M/M/1-M/M/1.

Ces trois problèmes sont étudiés dans respectivement les chapitres 3, 4 et 5.

Chapitre 3 : Coût sensible aux délais de livraison

Dans le chapitre précédent nous avons présenté rapidement les travaux considérant un modèle de type M/M/1 avec une demande qui dépend du prix et du délai annoncé. Nous avons déjà souligné que tous ces travaux supposent un coût de production constant. Or, on sait que lorsque le délai de livraison est plus long, l'entreprise peut mieux gérer la production et réduire le coût.

Nous étudions donc la décision de cotation d'un délai et d'un prix en supposant que le coût de production est une fonction décroissante du délai, et ceci dans un système de production MTO. Le système est modélisé par une M/M/1. La demande suit un processus de Poisson de taux d'arrivée moyen λ, qui ne peut pas être supérieur à la valeur maximale de la demande $\Lambda(P, L)$ obtenu lorsque le prix, P et le délai, L, sont proposés aux clients. Comme très souvent supposé dans la littérature (voir chapitre 2), nous considérons que la demande diminue linéairement avec le prix et le délai promis, $\Lambda(P, L) = a - b_1 P - b_2 L$ où b_1 et b_2 sont respectivement les coefficients de sensibilité au prix et au délai de livraison. La capacité de production est constante (μ) et le temps de service est réparti exponentiellement.

Revenons sur le coût unitaire. Il est bien connu que dans de nombreuses situations le coût de production unitaire dépend des délais promis. Les entreprises qui proposent un délai court à leurs clients, et qui sont donc exposées à un risque élevé, doivent repenser différentes décisions influençant le délai pour réduire le risque autant que possible. Cela concerne l'achat d'articles auprès de fournisseurs rapides mais coûteux au lieu de fournisseurs à coût moins cher (par exemple, fournisseurs locaux au lieu de fournisseurs à l'étranger), ou bien la détention d'un stock plus élevé de matières premières en amont, ou encore l'utilisation de modes de transport plus rapides mais plus coûteux. Ces différents exemples montrent que quand des délais promis sont courts, les actions requises peuvent conduire à des coûts de production unitaire élevés. Ainsi, nous ne considérons pas un coût de production constant, mais un coût de production unitaire (m) décroissant en fonction du délai promis (L), et proposons la fonction non linéaire suivante : $m = C_1 + \frac{C_2}{L}$. Cette fonction implique que l'augmentation du coût de production unitaire résultant d'une diminution unitaire du délai n'est pas constante (comme dans les fonctions linéaires), mais cette augmentation est d'autant plus forte que les délais sont faibles. De toute évidence, le coût de production généralement utilisé dans la littérature existante est un cas particulier de notre fonction de coût avec $C_2 = 0$.

Les variables de décision et les paramètres de notre problème sont les mêmes que ceux introduits dans Palaka et al. (1998) avec les paramètres de coût de production supplémentaires: C_1 et C_2.

L'objectif de l'entreprise est de maximiser le profit total attendu, ce qui équivaut au revenu (λP) – coût de production (λm) – coût de stockage total ($F\lambda/(\mu - \lambda)$) – coût de pénalité de retard ($c_r(\lambda/(\mu - \lambda))e^{-(\mu-\lambda)L}$). Comme expliqué par Palaka et al. (1998), le coût de retard reflète la rémunération directe versée aux clients pour ne pas respecter le délai de livraison indiqué. Le coût de stockage total est donné

par $F\lambda/(\mu-\lambda)$ où F est le coût de stockage unitaire, et $\lambda/(\mu-\lambda)$ est l'inventaire moyen. La pénalité de retard peut être donnée par (pénalité par travail par unité de retard) × (taux d'arrivée des demandes) × (probabilités qu'un travail soit en retard) × (retard moyen d'une commande en retard). Ainsi, cette pénalité de retard est: $c_r(\lambda/(\mu-\lambda))e^{-(\mu-\lambda)L}$ où c_r est la pénalité par travail par unité de retard, $e^{-(\mu-\lambda)L}$ est la probabilité qu'un travail soit en retard, et $\lambda/(\mu-\lambda)$ est le débit × retard moyen (voir Palaka et al. (1998)). Enfin, la firme doit respecter son délai de livraison avec un taux de service minimum (s). La formulation de notre modèle général est donnée ci-dessous.

$$\underset{L,P,\lambda,m}{Maximiser} \quad \lambda(P-m) - \frac{F\lambda}{\mu-\lambda} - \frac{c_r\lambda}{\mu-\lambda}e^{-(\mu-\lambda)L} \tag{7}$$

Sous contraintes

$$\lambda \leq a - b_1P - b_2L \tag{8}$$

$$1 - e^{-(\mu-\lambda)L} \geq s \tag{9}$$

$$\lambda \leq \mu \tag{10}$$

$$m = C_1 + C_2/L \tag{11}$$

$$\lambda, L, P, m \geq 0 \tag{12}$$

La fonction objectif est donnée par l'équation (7). L'équation (8) garantit que le taux de demande moyen reçu par l'entreprise ne peut pas dépasser la demande générée par le prix et le délai promis. L'équation (9) garantit que la probabilité de respecter le délais promis, donnée par $1 - e^{-(\mu-\lambda)L}$ (puisque $e^{-(\mu-\lambda)L}$ est la probabilité qu'un travail soit en retard dans la file d'attente M/M/1), ne doit pas être inférieure au niveau de service requis. L'équation (10) garantit un régime stable de la M/M/1. L'équation (11) définit la fonction coût de production. Les contraintes de positivité des variables sont données dans l'équation (12).

A partir du modèle général ci-dessus, nous considérons trois cas différents : (1) le prix est fixé et les coûts de stockage et de retard sont ignorés, (2) le prix est également une variable de décision (en plus du délai), mais les coûts de stockage et de retard sont toujours ignorés et (3) le prix est une variable de décision et des coûts de stockage et de retard sont considérés, c'est à dire le modèle général ci-dessus. Pour les 2 premiers cas, nous proposons une approche pour trouver analytiquement le délai optimal et le prix optimal (s'il s'agit d'une variable); et pour le 3$^{\text{ème}}$ cas, nous développons une approche numérique pour le résoudre.

Nous résolvons analytiquement le modèle lorsque le prix est fixé (cas 1). Dans ce contexte, le problème est formulé sous la forme d'un modèle d'optimisation non linéaire sous contraintes avec une seule variable de décision L (P est fixé et la contrainte (8) sur la demande est serrée). Nous avons montré que la fonction de profit est concave en L (voir le lemme 3.2). Dans notre modèle, la contrainte de niveau de service (eq. (9)) n'est pas nécessairement serrée. En effet le compromis entre l'augmentation de la demande (λ) et la réduction du coût de production unitaire (m), en modifiant L sans violer la contrainte de niveau de service, peut conduire à des situations non serrées pour la contrainte de niveau de service (9). Nous avons

étudié ces 2 situations et obtenu la valeur optimale de L comme nous le proposons dans la proposition 3.1.

Dans le deuxième cas, nous considérons une situation plus complexe où le prix P est aussi une variable de décision (en plus du délai L). Nous transformons et formulons le problème en un problème d'optimisation à deux variables (L et λ), car la contrainte de la demande est serrée (démontrée dans lemme 3.1). Nous en déduisons plusieurs lemmes et propositions qui nous permettent de résoudre le problème analytiquement comme proposé dans la proposition 3.3. Le délai optimal est en fait la racine d'une équation cubique.

Dans le troisième cas, nous considérons un modèle avec trois composantes de coûts: le coût de production unitaire (m) mais aussi le coût de stockage unitaire (F) et le coût de pénalité (c_r). Ce modèle est très difficile à résoudre analytiquement. Ainsi, nous le résolvons numériquement avec une méta-heuristique classique telle que l'optimisation par essaims particulaires (PSO).

Nous avons conduit des expériences numériques et obtenu des résultats intéressants. Dans le cas où le prix est fixé (cas 1), nous constatons que notre modèle permet d'avoir des gains significatifs par rapport au modèle de base qui ignore la sensibilité du coût au délai. Ce gain devient plus important lorsque nous prenons en compte le prix en tant que variable de décision (cas 2). Et lorsque nous considérons le retard et le coût de stockage, nous voyons que pour la solution optimale du modèle général (cas 3) la contrainte de service est non serrée dans tous les cas testés, afin de réduire les coûts encourus.

Chapitre 4 : Politique de rejet

La plupart des articles dans la littérature, présentés dans le chapitre 2, utilisent un modèle de type M/M/1. Ce modèle a l'avantage d'être facile à résoudre, mais il implique que tous les clients sont acceptés, ce qui peut entraîner de longs temps de séjour dans le système lorsque nous acceptons des clients alors qu'il y a déjà beaucoup de clients en attente et donc nécessite de promettre un délai suffisamment élevé si on veut un taux de service satisfaisant. Une alternative consiste à rejeter les clients lorsqu'il y a déjà beaucoup de clients en attente. Cela conduit à première vue à diminuer la demande (clients rejetés) mais, en permettant de proposer un délai plus faible pour les clients acceptés, cela pourrait donner a contrario un effet positif sur la demande. Il nous est donc paru intéressant d'étudier cette politique de rejet des clients au-delà d'un certain nombre de clients déjà présents dans le système en utilisant un modèle M/M/1/K. La demande est rejetée s'il y a déjà K clients dans le système (K représente la capacité du système, c'est-à-dire le nombre maximum de clients dans le système, y compris celui en service).

Dans ce chapitre, nous formulons explicitement le problème de choix du prix et du délai annoncé pour une firme modélisée par une M/M/1/K, face à une demande linéaire basée sur le prix et le délai, en tenant compte du coût de stockage et du coût de pénalité de retard. A nouveau, la demande est supposée être une fonction

décroissante linéaire du prix et du délai de livraison annoncé $a - b_1P - b_1L$. Les variables de décision de notre modèle sont donc le prix, le délai promis et la demande.

Contrairement au comportement d'une M/M/1, pour laquelle tous les clients sont acceptés, des clients sont rejetés dans le modèle M/M/1/K et nous appellerons ($\bar{\lambda}$) la demande effective. La capacité (taille du système) K est supposée constante. La probabilité P_k d'avoir k clients dans le système ($k = 1, 2, \ldots, K$) est donnée par l'équation (13) comme dans Gross et al. (2008). P_K représente la probabilité de rejet d'un client et $(1 - P_K)$ la probabilité qu'un client soit accepté. La demande effective ($\bar{\lambda}$) est égale au taux d'arrivée moyen (λ) multiplié par la probabilité d'accepter un client $(1 - P_K)$. L'équation (14) donne le nombre moyen de clients dans le système, noté N_s (voir Gross et al. 2008). Le temps de séjour moyen W (temps total dans le système) est égal à $N_s/\bar{\lambda}$. La probabilité que l'entreprise soit en mesure de respecter le délai de livraison cité (c.-à-d. $\Pr(W \leq L)$) et la probabilité qu'un travail soit en retard (c.-à-d. $\Pr(W > L)$) sont formulées dans les équations (15) et (16) tel qu'indiqué dans Sztrik (2012).

$$P_k = \frac{1-\rho}{1-\rho^{K+1}}\rho^k \text{ si } \rho \neq 1 \text{ et } P_k = \frac{1}{K+1} \text{ si } \rho = 1 \text{ avec } \rho = \frac{\lambda}{\mu} \tag{13}$$

$$N_s = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} \tag{14}$$

$$\Pr(W \leq L) = 1 - \sum_{k=0}^{K-1} \frac{P_k}{1-P_K} \left(\sum_{i=0}^{k} \frac{(\mu L)^i}{i!} e^{-\mu L} \right) \tag{15}$$

$$\Pr(W > L) = \sum_{k=0}^{K-1} \frac{P_k}{1-P_K} \left(\sum_{i=0}^{k} \frac{(\mu L)^i}{i!} e^{-\mu L} \right) \tag{16}$$

Afin d'éviter que les entreprises citent des délais de livraison irréalistes, nous supposons que l'entreprise maintient un niveau de service minimum. Ainsi, la probabilité de respecter le délai promis doit être supérieure au niveau de service désigné par s (c'est-à-dire, nous imposons: $\Pr(W \leq L) \geq s$).

L'objectif de l'entreprise est de maximiser le profit. Étant donné que nous considérons une pénalité de retard et des coûts de stockage, le profit de l'entreprise est calculé comme suit: Profit = Revenus (net du coût direct) – Coût de stockage total – Coût de pénalité de retard.

Pour formuler le coût de retard, nous devons calculer le retard moyen d'un travail en retard (R_L) dans une file M/M/1/K. Pour ce calcul, nous avons été confrontés à un obstacle théorique car à notre connaissance, ce résultat n'est pas connu dans la littérature. Notre travail apporte donc une nouvelle contribution à la littérature de la théorie des files d'attente en calculant explicitement la valeur de R_L dans une file M/M/1/K (voir théorème 4.1 dans chapitre 4).

Ainsi, à partir du résultat annoncé dans le théorème 4.1 et des équations (13), (14), (15) et (16), nous pouvons maintenant formuler explicitement le problème de choix du délai promis et du prix pour une entreprise modélisée par une M/M/1/K,

face à une demande linéaire en fonction du prix et du délai, en tenant compte des coûts de pénalités et de stockage.

$$
\begin{aligned}
(M_K)\ \underset{P,L,\lambda}{Maximiser} \quad & \lambda(1-P_K)(P-m) - (N_s \times F) \\
& \quad - (c_r \times \lambda(1-P_K) \times \Pr(W > L) \times R_L) && (17)\\
\text{Sous contraintes} \quad & \lambda \leq a - b_1 P - b_2 L && (18)\\
& \Pr(W \leq L) \geq s && (19)\\
& \lambda, P, L \geq 0 && (20)
\end{aligned}
$$

Dans le modèle (M_K), la contrainte (18) impose que la demande moyenne (λ) ne peut pas être supérieure à la demande obtenue avec le prix (P) et le délai promis (L). La contrainte (19) exprime la contrainte de niveau de service. La contrainte (20) exprime la positivité des variables du modèle.

De toute évidence, le modèle obtenu (M_K) est très difficile à résoudre analytiquement. Donc, nous commençons par considérer le cas de $K = 1$ (M/M/1/1). Nous considérons deux situations: le cas sans coût de pénalité et de stockage, et le cas où ces coûts sont inclus.

Dans le cas sans coûts de pénalité et de stockage, nous prouvons que la contrainte de la demande (équation (18)) et la contrainte de service (eq. (19)) sont serrées à l'optimal (voir les lemmes 4.1 et 4.2). Grâce à ces deux lemmes, par méthode de substitution, nous transformons le modèle initial en un modèle avec une seule variable λ. Nous dérivons des lemmes et résolvons ce modèle analytiquement. Nous proposons notre solution dans la proposition 4.1.

Dans le cas avec coûts de pénalité et de stockage, nous prouvons que la contrainte de la demande (équation (18)) est serrée à l'optimalité, mais la contrainte de service (eq. (19)) n'est pas toujours serrée. Nous fournissons les conditions caractérisant chaque situation (contrainte de service serrée ou non) dans le lemme 4.3. Et nous résolvons le problème analytiquement comme indiqué dans la proposition 4.2.

L'expression de la solution optimale, dans le cas $K = 1$ avec coûts de pénalité et de stockage, a montré qu'une augmentation de la sensibilité au délai conduit tout d'abord à réduire le délai promis mais devient inutile si elle dépasse une certaine valeur de seuil. En effet, au-delà d'une certaine valeur de cette sensibilité, la contrainte de service devient serrée et dans ce cas le délai optimal ne dépend plus de cette sensibilité. Nous avons également observé que lorsque les clients deviennent plus sensibles au prix ou lorsque le coût de la pénalité unitaire augmente, l'entreprise peut réagir en augmentant le délai de livraison.

Ensuite, nous avons comparé le profit optimal donné par notre modèle M/M/1/1 au profit optimal obtenu lorsque l'entreprise est modélisée en tant que file M/M/1 (comme dans la littérature). Le système M/M/1/1 représente la politique de rejet alors que dans le M/M/1, tous les clients sont acceptés. Nous avons découvert qu'une politique de rejet de type M/M/1/1 peut être pour certains cas plus rentable que la politique d'acceptation de tous les clients et ceci même lorsque les coûts de stockage et de pénalité ne sont pas pris en considération. Certains de nos résultats

ont montré qu'une augmentation de la sensibilité au délai ou de la sensibilité aux prix favorise la politique de rejet. L'augmentation des coûts de stockage s'est révélée être l'un des principaux critères qui rendent la politique de rejet meilleure que la politique d'acceptation de tous les clients. Une augmentation du niveau de service ou des coûts de pénalité unitaire favorise également la politique de rejet mais a un impact beaucoup plus faible que l'augmentation des coûts de stockage unitaire.

Les modèles avec $K > 1$, n'ont pas pu être résolus analytiquement. Nous les avons résolus numériquement par une approche de type optimisation par essaims particulaires (PSO). Nous avons mené des expériences pour comparer les résultats de notre modèle, pour différentes valeurs de K, aux résultats obtenus avec une file M/M/1 sous différents paramètres. Nous avons montré, pour toutes les instances considérées, qu'il y a au moins une valeur de K pour laquelle la politique optimale obtenue avec la M/M/1/K (politique de rejet) est plus rentable que celle obtenue avec la politique d'acceptation de tous les clients (M/M/1). Dans la plupart des cas, on a également observé qu'une augmentation de la valeur de K (c'est-à-dire la taille du système) a un effet non monotone sur le profit de l'entreprise. En effet, une augmentation de K, dans un premier temps améliore le profit puis ensuite entraîne une diminution.

Chapitre 5 : Coordination de la chaîne logistique : un modèle de M/M/1-M/M/1

Dans le chapitre 2, nous avons vu que dans tous les papiers de la littérature qui considèrent une chaîne logistique composée de plusieurs étages, et en fait 2 étages, seul un des acteurs a un processus de production et un délai. L'autre acteur n'a pas de processus de production, en d'autres termes, le délai est nul. Et donc aucun papier de la littérature ne considère une chaîne de deux étages dans laquelle les deux acteurs auraient un processus de production (délai). C'est le challenge que nous avons souhaité relever dans ce chapitre 5.

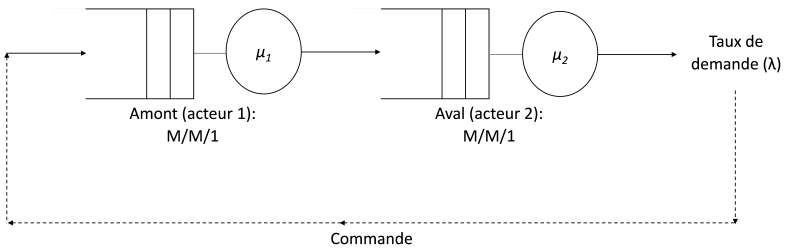


Figure 1: Modèle file d'attente

Nous considérons donc une chaîne logistique composée d'un acteur en amont (fournisseur ou fabricant) et d'un acteur en aval (fabricant ou détaillant). Les demandes arrivent à l'acteur en aval selon un processus de Poisson. Les deux acteurs

(en amont et en aval) ont une capacité fixe avec un temps de service exponentiel. Ainsi, nous modélisons le système comme un réseau en tandem de type M/M/1-M/M/1. Ce système est décrit dans la Figure 1.

Nous utilisons les mêmes notations que dans les chapitres précédents avec les compléments suivants :

Decision Variable

P_g = prix global du bien / service établi par la chaîne	L_g = délai de livraison global
P_1 = prix du bien / service mis en place par le premier acteur	L_1 = le délai promis de l'acteur 1
	L_2 = le délai promis de l'acteur 2

Parameters

m_1 = coût unitaire pour l'acteur 1	δ_2 = marge de l'acteur 2
m_2 = coût unitaire pour l'acteur 2	W_1 = temps d'attente total dans le système de l'acteur 1
μ_1 = taux de service moyen (capacité de production) de l'acteur 1	W_2 = Temps d'attente total dans le système de l'acteur 2
μ_2 = taux de service moyen (capacité de production) de l'acteur 2	

Nous avons développé différentes approches pour l'analyse de ce système avec d'une part une vision centralisée et d'autre part une vision décentralisée. Plus précisément, dans un cadre centralisé, les deux acteurs se coordonnent pour décider du prix global (P_g) et du délai promis global (L_g). Nous avons considéré 2 déclinaisons du problème selon que la contrainte de service est imposée globalement ou à chacun des acteurs. Dans un cadre décentralisé, nous considérons l'acteur aval comme leader et donc l'acteur amont comme suiveur. Nous avons considéré 2 modes de coordination. Dans le premier, l'acteur en aval décide de L_1 et L_2, et donc L_g, et l'acteur en amont décide de son propre prix (P_1), le prix global (P_g) étant fixé par $P_g = P_1 + \delta_2$. Dans le deuxième, l'acteur en aval décide du prix de l'acteur en amont (P_1) et de son délai (L_2), et l'acteur en amont décide de son propre délai (L_1).

Modèle Centralisé

Nous commençons notre analyse avec le modèle centralisé. Dans ce contexte centralisé, nous considérons les acteurs aval et amont qui décident ensemble du prix global (P_g) et du délai global (L_g). Nous modélisons le problème centralisé comme suit:

$$\underset{P_g, L_g}{Maximiser} \quad \Pi_c = (P_g - m_1 - m_2)\lambda \tag{21}$$

$$\text{Sous contraintes} \quad \lambda = a - b_1 P_g - b_2 L_g \tag{22}$$

$$\Pr(W_1 + W_2 \leq L_g) \geq s \tag{23}$$

$$\lambda < \mu_1, \mu_2 \tag{24}$$

Le temps de séjour d'un client dans le système, W_1+W_2, suit une distribution hypo-exponentielle (elle devient une loi d'Erlang dans le cas : $\mu_1 = \mu_2 = \mu$). Ce type de distribution rend très difficile une résolution analytique. Ainsi, nous résolvons le problème numériquement avec une méthode de dichotomie. La procédure détaillée se trouve section 5.2 du chapitre 5.

Modèle Centralisé Modifié

La résolution analytique du modèle centralisé est particulièrement difficile voire impossible en raison de l'équation très complexe de la distribution hypo-exponentielle du temps de séjour. Ainsi, nous avons eu l'idée de transformer la contrainte de service global en 2 contraintes de service local. Dans la section 5.3 de la thèse, nous fournissons des preuves analytiques et numériques qui montrent que quelque soient les taux de service, il existe un niveau de service minimum $s_{\min}(\mu_1, \mu_2)$, tel que pour toute valeur de s supérieure à $s_{\min}(\mu_1, \mu_2)$, si ce taux de service s est satisfait pour chacun des acteurs il est alors satisfait globalement. Cette valeur $s_{\min}$ dépend de la valeur μ_1/μ_2 et est maximale pour $\mu_1 = \mu_2$ où $s_{\min} = 0,715$. Ce résultat peut permettre d'aborder de nouveaux travaux sur une chaîne logistique comprenant 2 étages où chacun des acteurs a un délai, ce qui, rappelons-le n'a à notre connaissance jamais été abordé dans le contexte qui nous intéresse.

Dans la section 5.4, nous avons donc proposé un nouveau modèle, où nous transformons la contrainte de service global en contraintes de service pour chaque acteur. Nous appelons ce modèle «modèle centralisé modifié». Nous formulons ce modèle centralisé modifié comme suit:

$$\underset{P_g, L_g}{Maximiser} \quad \Pi_m = (P_g - m_1 - m_2)\lambda \tag{25}$$

$$\text{Sous contraintes} \quad \lambda = a - b_1 P_g - b_2 L_g \tag{26}$$

$$\Pr(W_1 \leq L_1) \geq s \tag{27}$$

$$\Pr(W_2 \leq L_2) \geq s \tag{28}$$

$$P_g = P_1 + \delta_2 \tag{29}$$

$$\lambda < \mu_1, \mu_2 \tag{30}$$

Nous avons pu transformer ce problème en un problème d'optimisation monovariable (λ). La solution optimale est racine d'une équation cubique dans le cas $\mu_1 = \mu_2$, et d'une équation du $5^{\text{ème}}$ degré sinon.

Modèle Décentralisé

Nous considérons une chaîne logistique composée de deux acteurs où chacun d'entre eux prend des décisions (prix et/ou délais) pour maximiser son propre profit en connaissance de la réaction de l'autre acteur. Le premier acteur (leader) prend sa décision sur le prix ou délai en tenant compte de la réaction du deuxième acteur,

alors le second acteur (suiveur) prendra une décision suite à la décision du premier acteur. Ce type de prise de décision est souvent appelé «Jeu de Stackelberg».

Nous avons proposé pour cette approche décentralisée deux modes de coordination. Dans les deux scénarios, l'acteur en aval agit comme le leader et l'acteur en amont agit en suiveur. Dans le premier scénario, l'acteur en amont choisit son propre délai (L_1), mais le prix P_1 et le délai de livraison L_2 sont décidés par l'acteur en aval (acteur 2). Dans le deuxième scénario, l'acteur en amont (acteur 1) décide de son propre prix (P_1) et l'acteur en aval (acteur 2) décide du délai global (L_1+L_2). La formulation détaillée et le calcul de chaque modèle décentralisé se trouvent dans la section 5.5.

À partir de nos expériences, nous voyons qu'en utilisant le premier modèle décentralisé (l'acteur en amont décide de son propre délai), le profit global est très faible par rapport à celui que l'on peut espérer avec l'approche centralisée. Nous voyons également que le profit de l'acteur en amont est nul. Cela montre que les acteurs ne se coordonnent pas naturellement de façon satisfaisante. Ainsi, nous proposons d'échanger la décision prise par chaque acteur. Dans la deuxième modèle, l'acteur amont (suiveur) décide de son propre prix (P_1), et l'acteur aval du délai de livraison L_1 et L_2, donc L_g. Il est intéressant de voir que l'équilibre obtenu est très proche de la situation où le profit global est maximum, que ce profit est proche du profit obtenu en centralisé et qu'enfin le réglage du partage des profits peut se faire avec le réglage de la marge de l'acteur aval.

Nous résolvons analytiquement les problèmes décentralisés et nous fournissons des études numériques. Nous avons comparé tous les scénarios: centralisé, centralisé modifié et les deux modèles décentralisés. Le meilleur profit est bien sûr celui obtenu avec l'approche centralisé. Toutefois, dans la plupart des cas les acteurs de la chaîne logistique ont une certaine autonomie et les scénarios décentralisés sont intéressants. Nous avons vu que le scénario où l'acteur en amont choisit son propre prix sous contrainte de délai imposé par l'acteur aval est très intéressant.

Chapitre 6 : Conclusion et perspectives

Cette thèse porte sur l'analyse et l'optimisation de systèmes de production dans le cas d'une demande sensible au prix et au délai de livraison promis aux clients. De la revue de la littérature (chapitre 2), on a identifié 3 extensions intéressantes: introduire un coût de production unitaire variable; étudier une politique de rejet de clients; étudier une chaine composée de 2 étages dans laquelle chacun des acteurs a un délai de production.

Dans la première contribution, nous avons résolu le problème du choix du délai annoncé dans une file d'attente M/M/1 lorsque coût de production est une fonction décroissante du délai. Nous avons considéré trois situations : (1) le délai est variable, mais le prix est fixé, (2) le prix et le délai sont deux variables de décision, et (3) le prix et le délai sont des variables de décision et le coût de retard et le coût de stockage sont pris en considération. Nous avons résolu analytiquement les 2

premiers cas et numériquement le troisième. Dans le cas 1, nous avons trouvé l'expression du délai (L) en fonction des paramètres du modèle. Mais, dans le cas 2, le délai optimal est racine d'une équation cubique. Et, pour le cas 3, nous avons résolu le modèle numériquement. Nous avons mené des expérimentations numériques qui montrent que nos modèles conduisent à des gains significatifs par rapport aux modèles existants où le coût est supposé être constant.

Dans la deuxième contribution, nous avons formulé le problème du choix du délai et du prix avec possibilité de rejet de clients pour une entreprise modélisée par une M/M/1/K, face à nouveau à une demande linéaire en fonction des prix et délai, en tenant compte du coût de stockage et de la pénalité de retard. Afin de déterminer la pénalité de retard, nous avons dû obtenir un nouveau résultat théorique en calculant explicitement le temps de séjour résiduel au-delà d'un temps donné d'une M/M/1/K. Ce résultat peut être utilisé à l'avenir pour différents problèmes de gestion des opérations et de théories de file d'attente. Nous avons montré que dans certaines configurations numériques, la politique de rejet (modélisée par la M/M/1/1) peut être plus rentable que la politique d'acceptation de tous les clients même lorsque les coûts de stockage et de pénalité ne sont pas pris en considération. Ceci nous a encouragés à étudier des valeurs de K plus élevées, cas que nous avons résolu numériquement. Nous avons montré sur tous les exemples traités qu'il y a au moins une valeur de K pour laquelle la M/M/1/K (politique de rejet) est plus rentable que le M/M/1 (la politique d'acceptation de tous les clients). Dans tous les cas, on a également observé qu'une augmentation de la valeur de K (c'est-à-dire la taille du système) a un effet non monotone sur le profit de l'entreprise. En effet, dans un premier temps une augmentation de K améliore le profit mais ensuite entraîne une diminution.

Dans la dernière contribution, nous avons résolu avec succès (numériquement) le modèle centralisé et le modèle dit centralisé modifié. Pour ce dernier, nous imposons des contraintes locales de service, ce qui a été possible par la démonstration d'un résultat donnant les conditions pour que la satisfaction de contraintes locales de service suffisent à la satisfaire globalement. Nous avons aussi introduit et résolu 2 modèles décentralisés et fait des expériences numériques.

Nous avons comparé les différents scénarios: centralisé, centralisé modifié et les deux modèles décentralisés. Si le meilleur profit est bien sûr celui obtenu avec l'approche centralisée, nous avons montré l'intérêt d'un scénario décentralisé où l'acteur en amont choisit son propre prix sous contrainte de délai imposé par l'acteur aval.

Notre étude peut être étendue de différentes façons. Par exemple, il serait intéressant de considérer une autre forme de demande également souvent retenue dans la littérature, en l'occurrence le modèle de demande Cobb-Douglas (demande exponentielle décroissante en fonction du délai et du prix). Il serait aussi intéressant d'approfondir le cas décentralisé en introduisant un système incitatif de partage des bénéfices notamment dans le scénario où l'acteur en amont choisit son propre délai (pour éviter un bénéfice zéro de l'acteur en amont). Une autre extension de notre

modèle de file d'attente en tandem serait d'inverser les rôles de leader-suiveur, avec donc l'acteur amont en tant que leader et l'aval comme suiveur. Enfin tous ses travaux ont considéré une seule firme ou bien 2 firmes en coopération. On pourrait également envisager une situation de concurrence entre les acteurs des chaînes logistique.

Chapter 1
Introduction

A large number of firms are using pricing and lead time quotation decisions as a strategic weapon to manage the demand and to maximize the profitability. It is well known in the business logistics literature that one of the most important customer-service elements, in addition to price, is the delivery lead time (Sterling and Lambert, 1989; Ballou, 1998; Jackson et al., 1986). Along with the price, the delivery lead time has become a key factor of competitiveness for companies and an important purchase criterion for many customers (Hammami and Frein, 2013). Since the 90's, time-based competition has been widely established as a key to success in many industries as reported by Blackburn (1991) and Stalk and Hout (1990). The academic and popular literature on time-based competition presented ample evidence on how firms can use delivery lead time as a strategic weapon to gain competitive advantage (Blackburn et al., 1992; Hum and Sim, 1996; Suri, 1998). Geary and Zonnenberg (2000) reported that the best in class performers of 110 firms in five major manufacturing sectors focus their operations on achieving breakthroughs not only in cost, but also in speed (delivery lead time). Baker et al. (2001) stated that less than 10% of end-consumers and less than 30% of corporate customers base their purchase decisions on an item's selling price only; the rest also care about other customer-service elements.

Delivery lead time is traditionally defined as the elapsed time between the receipt of customer order and the delivery of this order (Christopher, 2011). Nowadays, firms are more than ever obliged to meet their quoted lead time, that is the delivery lead time announced to the customer. The combination of pricing and lead time quotation implies new trade-offs and offers opportunities for many insights.

For instance, a shorter quoted lead time can lead to an increase in the demand but also increases the risk of late delivery, which can imply a lower service level and an increase in the lateness penalty. For many operations sectors, failure of attaining the quoted lead time might lead to a large amount of penalties. According to Savaşaneril et al. (2010), examples that show the importance of reliable lead time quotes are abundant in industry. The authors reported that the cost of late delivery in the FMC Wellhead Equipment Division may rise up to $250,000 per day and that the lateness penalties in the aircraft industry starts from $10,000-$15,000 and can go as high as $1,000,000 per day. In addition to its impact on the cost, the late delivery may affect the firm's reputation and deter future customers (Slotnick, 2014); companies risk even to lose markets if they are not capable of respecting the quoted lead time (Kapuscinski and Tayur, 2007).

A longer quoted lead time or a higher price generally yields a lower demand, which might have a negative effect on the profitability. This will drive the costumers

to order from the competitors who propose shorter quoted lead times and/or lower prices (Pekgün et al., 2016; Ho and Zheng, 2004; So, 2000; Xiao et al., 2014). The positive side of quoting longer lead time is the possibility of attaining higher service level (since, on the one hand, the quoted lead time is longer and, on the other hand, the demand is lower). It can also decrease the in-process inventory holding cost. This latter cost can be significant in many industries such as in automotive and electronics.

Despite the strategic role of joint pricing and lead time quotation decisions and their impacts on demand, the operations management literature has not paid enough attention to this problem, as reported by Huang et al. (2013). To our knowledge, most of the literature that deals with lead time quotation and pricing under endogenous demand (i.e., a demand that depends on quoted lead time and price) considered a Make-To-Order (MTO) context (for the articles that are in Make-To-Stock (MTS) context, see Savaşaneril et al., 2010; Savaşaneril and Sayin, 2017; Panda, 2013; Wu et al., 2012).

The pioneer paper on lead time quotation and pricing under lead time and price sensitive demand in MTO context is Palaka et al. (1998). Their research examined the lead time setting, pricing decisions, and capacity utilization for a firm serving customers that are sensitive to quoted lead times and price. The authors initially restricted their focus to a short time horizon, and hence capacity was assumed to be constant while price, quoted lead time, and demand were the decision variables. Customers were served on a first come-first served basis. The arrival pattern of customers was modeled by a Poisson process. Further, the processing times of the customer orders were assumed to be exponentially distributed. These assumptions led to the use of an M/M/1 queue to model the firm's operations. Demand was assumed to be a linear decreasing function in price and quoted lead time. In the last part of the paper, the authors considered a capacity expansion case where they fixed price and modeled demand, lead time, and capacity as decision variables.

Based on Palaka et al. (1998), different extensions have been studied in both single and multi-firm settings. For instance, in single firm setting, Pekgün et al. (2008) studied the centralization and decentralization of pricing and lead time decisions between production and marketing departments. They used the same framework of Palaka et al. (1998) for their centralized model but without considering the holding and penalty costs. Ray and Jewkes (2004) focused on customer lead time management where demand is a function of price and lead time, and where price itself is sensitive to lead time. Zhao et al. (2012) studied lead time and price quotation in service firms and make-to-order manufacturing industries. They considered two strategies: in the first strategy firms offer single lead time and price quotation (uniform quotation mode) and, in the second case they offer a menu of lead times and prices for customers to choose from (differentiated quotation mode). In the multi-firm setting, Zhu (2015) considered a decentralized supply chain consisting of a supplier and a retailer facing price- and lead time-sensitive demand. The decision process was modeled as a sequence where supplier determines capacity and whole-

sale price, and retailer determines sale price and lead time. Pekgün et al. (2016), which was an extension of Pekgün et al. (2008), studied two firms that compete on price and lead time decisions in a common market. A detailed discussion of the relevant literature will be developed in chapter 2.

Our review of the literature allowed to identify new perspectives for the problem of lead time quotation and pricing in a stochastic MTO context with endogenous demand. In particular,

1. The unit production cost was assumed to be constant in most published papers. In practice, the unit production cost generally depends on the quoted lead time. Indeed, the firm can manage better the production process and reduce the production cost by quoting longer lead time to the customers. However, considering a unit production cost as a function of lead time yields new analytical difficulties, especially because the relation between cost and lead time is not linear.

2. In single firm setting, only the M/M/1 queue was used. In M/M/1, all the customers are accepted, which might lead to long sojourn times (lead time) in the system. In practice, firms can choose to reject the customers when they already have too many customers. Thus, one can consider the use of the capacitated M/M/1/K queue. However, the formulation of residual waiting time, which is required to calculate the lateness penalty cost for the overdue clients, is not available in the literature for M/M/1/K.

3. In multi-firm setting, most papers considered that only one actor has production operations (the other actor has zero lead time). It is more realistic to consider a supply chain that consists of more than one stage having their own production operations. However, considering a tandem queue is challenging as it leads to a very complex service level constraint.

Based on the observations explained above, we propose three extensions in this thesis:

- Unit production cost is sensitive to lead time. In the first contribution, we use Palaka et al.'s framework and consider the production cost to be a decreasing function in quoted lead time. Indeed, a company can use different ways to reduce lead time (such as buying items from quick response but expensive subcontractors) but this generally leads to higher production cost. Moreover, a longer lead time can permit a better optimization of production process and, consequently, can lead to a decrease in production cost. The detailed analysis is provided in chapter 3.

- Firm's operations modeled by an M/M/1/K queue. In the second contribution, we still consider Palaka et al.'s framework but model the firm as an M/M/1/K queue, for which demand is rejected if there are already K customers in the system. Indeed, our idea is based on the fact that rejecting some customers

might help to quote shorter lead time for the accepted ones, which might finally lead to a higher profitability. The detailed discussion is presented in chapter 4.

- Two-stage supply chain modeled as a tandem queue (M/M/1-M/M/1). Finally, we study a new setting for the lead time quotation and pricing problem under endogenous demand as we model the supply chain by two production stages in a tandem queue. We investigate both the centralized and the decentralized settings. This will be the focus of chapter 5.

Chapter 2
Literature review

In this chapter, we will discuss the relevant articles on lead time sensitive demand models. We provide a classification of the relevant literature in figure 2.1. We classify the lead time sensitive demand models into two categories: Make-To-Order (MTO) and Make-To-Stock (MTS). In MTO context, we classify the lead time sensitive demand models into two streams: 1. single-firm models; and 2. multi-firm models.

Our research belongs to the body of literature in MTO context. Thus, we start by discussing the pioneer article Palaka et al. (1998) in section 2.1. Then, we discuss the two streams of the lead time sensitive demand models in MTO context (section 2.2 and 2.3). Next, we discuss the papers in Make-To-Stock (MTS) context (section 2.4). Finally, we conclude by pointing out our positioning and contributions in section 2.5.

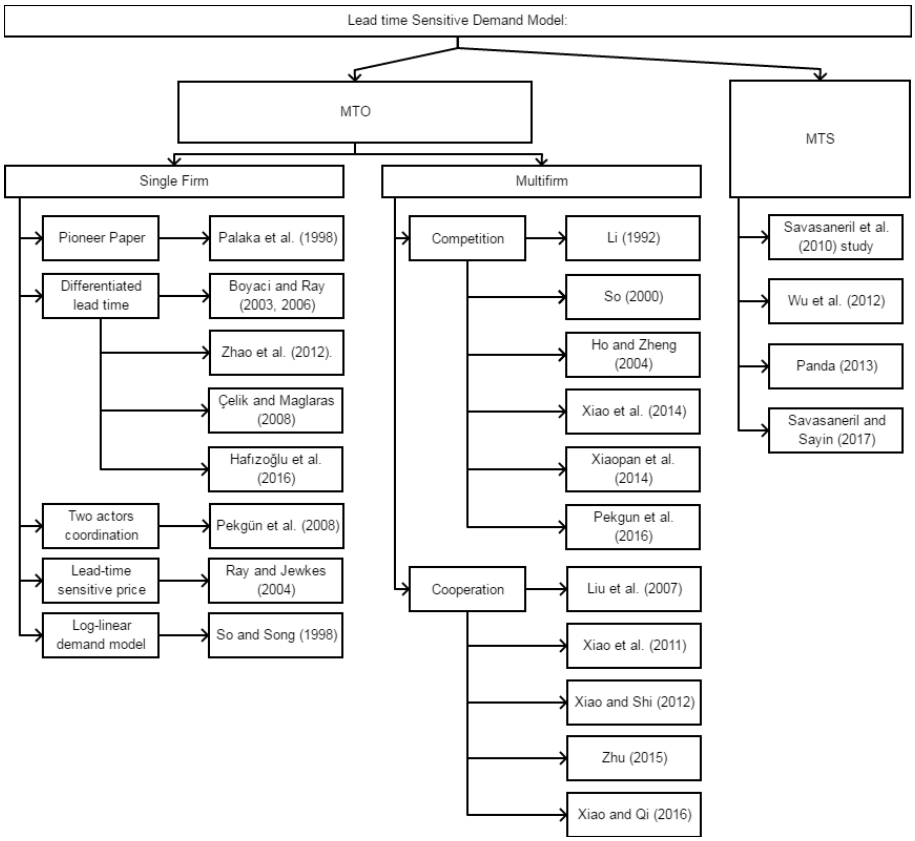


Figure 2.1: Classification of relevant studies

2.1 The pioneer paper: Palaka et al. (1998)

Palaka et al. (1998) studied lead time setting, capacity utilization, and pricing decisions of a firm where the customers are sensitive to quoted lead times. Palaka et al. (1998) considered a firm that serves customers in a make-to-order fashion. They initially restricted their model to a short time horizon, and hence capacity is assumed to be constant while price, quoted lead time, and demand are considered as decision variables. Customers are served on a first come-first served basis. They assumed that the arrival pattern of customers follows a Poisson process. The processing times of the customer orders is assumed to be exponentially distributed. These assumptions led to an M/M/1 model of the firm's operations. Customers are lead time sensitive and demand is assumed to be downward-sloping in both price and quoted lead time. The expected demand is a linear function of quoted lead time and price, which is modeled as:

$$\Lambda(P, L) = a - b_1 P - b_2 L \tag{2.1}$$

where, P is price of the good/service set by the firm, L = quoted lead time, $\Lambda(P, L)$ = expected demand for the good/service at price P and quoted lead time L, a = demand corresponding to zero price and zero quoted lead time, b_1 = price sensitivity of demand, and b_2 = lead time sensitivity of demand. Since the demand is downward sloping in both price and quoted lead time, b_1 and b_2 are restricted to be non-negative.

This linear demand function is tractable and has several desirable properties as highlighted by Palaka et al. (1998). For instance, the price elasticity of demand, given by $(-b_1 P/(a - b_1 P - b_2 L))$ is increasing in both price and quoted lead time. In other words, the percentage change in demand in response to a 1% change in price is higher for higher price and quoted lead time. Similarly, the lead time elasticity of demand, given by $(-b_2 L/(a - b_1 P - b_2 L))$, is higher for higher quoted lead time and price. Indeed, customers would be intuitively more sensitive to long lead times when they are paying more for the goods or service. Similarly, customers would be more sensitive to high prices when they also have longer waiting times.

To prevent firms from quoting unrealistically short lead times, they assumed that the firm maintains a certain minimum service level (s), where service level is defined as the probability of meeting the quoted lead time. This minimum service level may be set by the firm itself in response to competitive pressures.

Since they assumed a M/M/1 queuing system with mean service rate, μ, and mean arrival rate (demand), λ, the expected number of customers in the system, N_s, is given by $N_s = \lambda/(\mu - \lambda)$ and the actual lead time or sojourn time in the system, W, is exponentially distributed with mean $1/(\mu - \lambda)$ (Kleinrock, 1975; Hillier and Lieberman, 2001). The probability that the firm is not able to meet the quoted lead time, L, is given by $e^{-(\mu-\lambda)L}$ and the expected lateness of a late job is $1/(\mu - \lambda)$, the same as the expected lead time due to the memoryless property of the exponential distribution.

The firm's objective is to maximize the expected total profit contribution which can be expressed by equation (2.2). In the objective function, $\lambda(P-m)$ represents the expected revenue (net of direct costs), where m is the unit direct variable cost. The expected congestion costs are given by $F\lambda/(\mu-\lambda)$ where F is the unit holding cost and $\lambda/(\mu-\lambda)$ is the expected number of customers in the system. The expected lateness penalty is given by $c_r(\lambda/(\mu-\lambda))e^{-(\mu-\lambda)L}$, where c_r is the penalty per job per unit lateness, number of overdue client equaled to $\lambda e^{-(\mu-\lambda)L}$, and expected lateness given that a job is late equals to $1/(\mu-\lambda)$. Finally, Palaka et al. (1998) formulated the optimization problem as:

$$(P_{Base}) \underset{P,L,\lambda}{Maximize} \quad \Pi(P,L,\lambda) = (P-m)\lambda - \frac{F\lambda}{\mu-\lambda} - \frac{c_r\lambda}{\mu-\lambda}e^{-(\mu-\lambda)L} \tag{2.2}$$

$$\text{Subject to} \quad \lambda \leq a - b_1P - b_2L \tag{2.3}$$

$$1 - e^{-(\mu-\lambda)L} \geq s \tag{2.4}$$

$$0 \leq \lambda \leq \mu \tag{2.5}$$

$$P, L \geq 0 \tag{2.6}$$

Constraint (2.3) ensures that the mean demand, λ, served by the firm did not exceed the demand generated by price, P, and quoted lead time, L. Constraint (2.4) expresses the lower bound on the service level. The service level constraint guarantees that the probability of meeting the quoted lead time (given by $1-e^{-(\mu-\lambda)L}$ for an M/M/1 queue), must not be smaller than the minimum required service level s. It is important to note that for Poisson arrivals and exponential service times assumptions, this form of service constraint is exact. Furthermore, for high service levels, it gives a good approximation even for a G/G/s queue (refer to So and Song, 1998). Hence, the model is approximately valid for more general demand and service time characteristics. Constraint (2.5) corresponds to the restriction that mean demand served, λ, is also bounded by the firm's processing rate, μ. Constraint (2.6) restricts price and quoted lead times to non-negative values.

In their paper, Palaka et al. (1998) stated that constraint (2.3) is binding at optimality. The firm will choose price, P, quoted lead time, L, and demand rate, λ, such that $\lambda = \Lambda(P,L)$ at optimality. This can be proven by supposing that the optimal solution is given by price, P^*, quoted lead time L^*, and demand rate λ^*, and that $\lambda^* < \Lambda(P^*,L^*)$. Since the revenues are non-decreasing in P, one could increase the price to P' (while holding the demand rate and quoted lead time constant) until $\lambda^* = \Lambda(P',L^*)$. This change will increase revenues without increasing direct variable costs and lateness penalties. Therefore, (P^*,L^*,λ^*) cannot be an optimal solution.

Service level constraint (2.4) in the optimization model (P_{Base}) is not necessarily binding at optimality. Palaka et al. (1998) stated that the service level constraint (2.4) is non-binding iff the service level, s, is strictly lower than a critical value, s_c, that is, $s < 1 - b_1/(b_2c_r)$. In addition, the service level is given by $\max(s, s_c)$.

The solutions of problem P_{Base} in both binding and non-binding cases, as stated by Palaka et al. (1998), are:

(a) optimal demand λ^* is given by the root of cubic equation below on the interval $[0, \mu]$:

$$(a - mb_1 - 2\lambda)(\mu - \lambda)^2 = G\mu$$

where $G = b_2 \log x + Fb_1 + c_r b_1/x$ and $x = \max\{1/(1-s), b_1 c_r/b_2\}$,

(b) optimal quoted lead time L^* is given by $(\log x)/(\mu - \lambda^*)$, and

(c) optimal price, P^*, is obtained using the relationship $P^* = (a - \lambda^* - b_2 L^*)/b_1$.

This model of Palaka et al. (1998) has been a stepping stone for many recent studies.

In the rest of their paper, Palaka et al. (1998) considered the capacity expansion case where price is fixed. They introduced fractional increase in processing rate (Z), upper bound on capacity expansion ($\bar{Z}$), and cost of increasing the processing rate by one job/unit time (c_e). The type of capacity expansion that they consider is a short term nature, e.g., hiring part-time or temporary worker, and running overtime. The capacity expansion is given by $\mu(1 + Z)$ with total cost of expansion written as $c_e \mu Z$ (for further details see Palaka et al., 1998).

2.2 Single-firm in MTO system

Following Palaka et al. (1998)'s research, several authors dealt with lead time sensitive demand model in MTO single firm. We divide the relevant papers in this section as papers that propose differentiated lead times to customers, papers that dealt with coordination between actors in a single firm, papers that model lead time sensitive price, and papers that use log-linear demand model.

Differentiated lead times Papers that propose differentiated lead times to the customer are Boyaci and Ray (2003, 2006); Çelik and Maglaras (2008); Zhao et al. (2012); and Hafızoğlu et al. (2016).

Boyaci and Ray (2003) studied a profit-maximizing firm selling two substitutable products in a price and time sensitive market. They considered two type of products: 1. regular (slower) product with a given standard industry lead time, and 2. express (faster) product. These products are substitutable and therefore customer demand for each product depends on the price and delivery time of both products. The two products differ only in their prices and lead times. They assumed that there are dedicated capacities for each product. The firm objective is to determine the optimal price for each product, the quoted lead time for express product, and production capacity for each product. They considered the same unit constant operating cost for both regular and express products. They assumed that customers arrive to take delivery of the products at two separate facilities (one for the regular product and another for the express) according to a Poisson process. The mean arrival rate (demand) for each of the products is a linear function that depends not only on its own price and lead time but also price and lead time of the other product. The

service time for each product is exponentially distributed and the customers are served on a first-come first-served basis. The system is modeled as two M/M/1 in parallel. Boyaci and Ray (2003) found that the degree of product differentiation depends on the values of capacity costs. An increase in capacity cost differential increases price differentiation, but decreases time differentiation. They found that prices can actually decrease when the firm incurs capacity-related costs. The optimal prices depend on the market characteristic.

Boyaci and Ray (2006)'s research is an extension of Boyaci and Ray (2003). However, in this research they assumed that the capacities of both products are fixed. They added a new decision variable instead of capacities which is delivery reliability (minimum service level s). They introduced a new linear demand model as a function of price, delivery time and delivery reliabilities of both products. Their system is also modeled as a parallel M/M/1 with constant unit production cost as in Boyaci and Ray (2003). Boyaci and Ray (2006) found that customer preferences towards delivery times, reliabilities, prices, and the capacity costs have an impact on the firm's optimal product positioning policy.

Çelik and Maglaras (2008) studied the operational and demand control decisions faced by a profit-maximizing make-to-order production firm that offers multiple products to a market of price and lead time sensitive customers. They emphasized three features: 1. the joint use of dynamic pricing and lead time quotation controls to manage demand; 2. the access to a dual sourcing mode that can be used to expedite orders at a certain cost; and 3. the interaction between the demand controls and the operational decisions of sequencing and expediting. No production cost was occurred. They only considered expediting cost. Demand is assumed to be an N-dimensional non-homogeneous Poisson process with instantaneous rate vector where the arrival rate of potential customers is sensitive to price and quoted lead time. A proposed heuristics and Markov decision process formulation were used to solve the problem in their study. Unlike Boyaci and Ray (2003, 2006), Çelik and Maglaras (2008) considered one or several good(s) offered at multiple (price and lead time) combinations. Their numerical results illustrate the effectiveness of dynamic over static pricing, as well as the impact of lead time control policies.

Zhao et al. (2012) analyzed strategies where a firm offers a single lead time and price (uniform quotation mode (UQM)) or offers a menu of lead times and prices (differentiated quotation mode (DQM)). They also classified customers into two groups: lead time sensitive (LS) customers who value more the lead time reduction and price sensitive (PS) customers who value more the price reduction. They assumed that costumers for UQM arrive in Poisson process and for DQM in sub-Poisson. Unlike Boyaci and Ray (2003, 2006), Zhao et al. (2012) modeled the demand as a function of utility (willingness-to-pay) where the linear utility depend on price and lead time. The service times are exponentially distributed. Thus, they modeled the system as M/M/1 for UQM and parallel M/M/1 for DQM. They considered a constant production cost. They used their model to determine the optimal quoted lead time, price and service rate (μ). They found that DQM is dominated by UQM when LS

customers value a product or service no more than PS customers. Otherwise, which quotation mode is better depends on multiple factors, such as customer characteristics (including lead time reduction valuation and product valuation of a customer, and the proportion of LS customers) and production characteristics (including the desired service level and service or production cost).

Hafızoğlu et al. (2016) studied price and lead time quotation decisions in make-to-order system with two customer classes: (1) contract customers whose orders are always accepted and fulfilled based on a contract price and lead time agreed on at the beginning of the time horizon, and (2) spot purchasers who arrive over time and are quoted a price and lead time pair dynamically. The spot customer will place an order or not according on quoted lead time and price proposed. The objective is to maximize the long-run expected average profit per unit time, where profit from a customer is defined as revenues minus lateness penalties incurred because of lead time violations. They did not consider production cost. Hafızoğlu et al. (2016) model the dynamic quotation problem of the spot purchasers as an infinite horizon Markov decision process, given a fixed price and lead time for contract customers. They showed that the optimal price and lead time quotation policy is heavily affected by the price/lead time sensitivity of the spot purchasers and the penalty for missed due dates. When spot purchasers are highly price sensitive, it is optimal to quote a small fixed price while dynamically changing the lead time. In contrast, if spot purchasers are highly lead time sensitive, offering zero lead times and dynamic pricing is optimal.

Two actors' coordination Pekgün et al. (2008) dealt with coordination between marketing and production departments of a single firm in deciding pricing and lead time decisions. They studied a firm which serves customers that are sensitive to quoted price and lead time with pricing decisions being made by the marketing and lead time decisions by production departments. In their paper, they assumed that demand is a linear function of price and lead time, unit production costs are constant, and the firm system is modeled as M/M/1 queue. They considered two settings of decision-making: centralized and decentralized settings. In centralized setting, lead time quotation and pricing decisions are taken simultaneously. In this setting, Pekgün et al. (2008) used a simplified version of Palaka et al. (1998)'s model where they don't consider the lateness penalty and congestion cost. In decentralized setting they considered that the decisions are taken sequentially (as a Stackelberg game). Under this setting, Pekgün et al. (2008) developed two cases: 1. production acts as a leader who decides the quoted lead time and marketing acts as a follower who decides the price; and 2. marketing acts as a leader who decides the price and production acts as a follower who decides the quoted lead time. They found that inefficiencies are created by the decentralization of the price and lead time decisions. In the decentralized setting, the total demand generated is larger, lead times are longer, quoted prices are lower, and the firm's profits are lower compared to the centralized setting. They stated that coordination can be achieved using a

transfer price contract with bonus payments.

Lead time sensitive price Ray and Jewkes (2004) modeled an operating system consisting of a firm and its customers, where the mean demand rate is a function of the guaranteed delivery time (quoted lead time) offered to the customers and of market price, where price itself is determined by the length of the delivery-time. Firm's production system is modeled as M/M/1. They introduced the explicit dependence of price on lead time as an additional relationship. Ray and Jewkes (2004) stated that the new link of price and lead time captures a relationship that exists in practice. If this relationship is ignored, it could lead to a weak decision. In the latter part of their paper, they incorporated economies of scale by assuming that the unit operating cost is a decreasing function of the demand rate within a certain volume range.

Log-linear demand model So and Song (1998) studied the impact of using delivery time guarantees as a competitive strategy in service industries where demands are sensitive to both price and delivery time. They assumed that delivery reliability is crucial, and investment in capacity expansion is plausible in order to maintain a high probability of delivering the time guarantee. Thus, their objective is to find the optimal price, delivery time guarantee, and capacity selection in order to maximize profit. The production costs in So and Song (1998)'s model is assumed to be constant. So and Song (1998)'s main difference compared to Palaka et al. (1998) is their demand model. So and Song (1998) modeled the demand as log-linear model (Cobb-Douglas). As for the firm system, they modeled it as M/M/1 similar to Palaka et al. (1998).

2.3 Multi-firm in MTO system

In this subsection, we consider lead time sensitive demand models in MTO system with multi-firm. We divide the literature into two groups: competition models and cooperation models.

Competition models In papers dealing competitive firms, we have Li (1992); So (2000); Ho and Zheng (2004); Xiao et al. (2014); Xiaopan et al. (2014); and Pekgün et al. (2016).

Li (1992) studied the role of inventory in response time competition. Their objective is to determine the optimal production/inventory policy and the optimal choice between make-to-order and make-to-stock operation. He dealt first with a single firm production control in which customers are characterized by their preference of price, delivery quality and delivery-time (lead time). The demand from customers arises over time according to a Poisson process with intensity λ. Customers will buy or not buy the product depending on their utility (willingness to pay) where the

utility is a function of price, delivery quality, and delivery time. The firm incurs a constant unit production cost. In the first part of his paper, Li (1992) considered a single firm, then he extends his work to competitive multi-firm where firms compete for orders in terms of early delivery. The competition they introduced is only in timely delivery rather than price, product, or other aspect of the market. The goal is to show that competition breed a demand for make-to-stock. The competition in Li (1992) can be considered as parallel M/M/1 queue. With lead time uncertainty, they identified three important factors in firm's decision on production/inventory policies: discount, costumer characteristic, and competition. They found that the incentive for make-to-stock decreases when the nature of the market switches in the order of oligopoly racing, monopoly, and the demand sharing market. They also found that delivery-time competition increases the buyer's welfare while decreases the producer's welfare.

So (2000) developed a stylized model to analyze the impact of using time guarantees (quoted lead time) on competition. This research is an extension of So and Song (1998). Unlike So and Song (1998) who considered a single firm, So (2000) considered a competition between two firms. This competition is modeled as a parallel M/M/1 queue where each firm decides price and time guarantee. So (2000)'s demand model is similar to So and Song (1998). In their study, So (2000) found that different firm and market characteristics affect the price and delivery time (lead time) competition in the market. The equilibrium price and time guarantee decisions in an oligopolistic market with identical firms behave in a similar fashion as the optimal solution in a monopolistic situation from So and Song (1998). However, when there are heterogeneous firms in the market, these firms will exploit their distinctive characteristics to differentiate their services. Assuming all other factors being equal, the high capacity firms provide better time guarantees, while firms with lower operating costs offer lower prices, and the differentiation becomes more acute as demands become more time-sensitive. As time-attractiveness of the market increases, firms compete less on price, and the equilibrium prices of the firms increase as a result.

Ho and Zheng (2004) studied how a firm might choose a delivery time commitment to influence its customer expectation, and delivery quality in order to maximize its market share. They stated that many firms now choose to set customer expectation by announcing their maximal delivery time. Customers will be satisfied if their perceived delivery times are shorter than their expectations. They considered a firm that serves a population of homogeneous customers who are impatient and sensitive to service delivery time. The firm's objective is to maximize the demand rate, which is affected by customers' expectation for the delivery time as well as the probability that this expectation is being fulfilled. In Ho and Zheng (2004) research, no production cost is incurred. They modeled the whole service delivery process as an M/M/1 queueing system where the arrival (demand) rate depends on customer's utility for the firm's service. Unlike Zhao et al. (2012), Ho and Zheng (2004)'s customer's utility for the firm's service depends on the expected delivery

time and service quality. In the first part of their paper, they considered single firm. Then, they considered two firms in duopoly competition where firms compete for a fixed market. The competition is modeled as parallel M/M/1. These two firms compete in obtaining demand (market share). Ho and Zheng (2004) showed that the delivery time commitment game is analogous to a Prisoners' Dilemma.

Xiao et al. (2014) developed a game theory model of a one-manufacturer and one-retailer supply chain facing an outside integrated chain (a manufacturer) to study the price and lead time competition and investigate coordination within their supply chain. They considered a make-to-order production mode and consumers are sensitive to retail price and quoted lead time. In one-manufacturer and one-retailer supply chain, the manufacturer decides wholesale price and quoted lead time, then retailer decides retail price. Here, the retailer has zero lead time. In the integrated chain, the lead time quotation and pricing decision are taken simultaneously by a manufacturer. The demand in Xiao et al. (2014)'s model is a linear function in utility where the utility (willingness-to-pay) itself is sensitive to lead time and price. The unit production cost of both chains are assumed to be constant. Xiao et al. (2014) modeled their problem in Hotelling's Location Model and the objective for each chain is to maximize its own profit. Hotelling model is a model introduced by Hotelling (1929) where consumer market is conceptualized as a straight line in an ideal preference space, with exogenously specified locations for the two brands (i.e., manufacturer-retailer chain and integrated chain). Xiao et al. (2014) found that the coordination of the supply chain facing integrated chain harms the integrated chain. In addition, the existence of the outside competitor increases the lead time. They also found the Nash equilibrium of this competition.

Xiaopan et al. (2014) investigated the competition and cooperation in a duopoly setting where two dominant facilities control over a market. The demand faced by each facility is not only sensitive to its own retail price and guaranteed delivery time, but also to the differences between the two prices and guaranteed delivery times. They analyzed three different competition scenarios where 1. the two facilities compete exclusively on retail prices, 2. exclusively on guaranteed delivery times, and 3. both on retail prices and guaranteed delivery times. They also analyzed three different cooperation scenarios where 1. firms cooperate exclusively on retail prices, 2. exclusively on guaranteed delivery times, and 3. both on retail prices and guaranteed delivery times. In all scenarios, the objective of the firms is to maximize their profit. The demand in all scenarios follows Poisson process and it is a linear function in price, lead time and sensitivity of switch-over toward price and lead time difference (similar to Boyaci and Ray, 2003). The market system is modeled as parallel M/M/1 and unit production costs are assumed to be constant. In their study, Xiaopan et al. (2014) found that the competition on retail prices for given guaranteed delivery times is as same as the Bertrand game, reaching to the Nash-Bertrand equilibrium. Nash-Bertrand equilibrium occurs when both firms set price equal to unit cost (the competitive price). The equilibrium guaranteed delivery time of each facility for given retail prices only depends on its own retail price. When

price is a decision variable, the gross profits of the facilities increase significantly via coordinating retail prices.

Pekgün et al. (2016)'s research is an extension of Pekgün et al. (2008). Pekgün et al. (2016) studied two firms that compete on price and lead time decisions in a common market. They investigated the impact of decentralizing the decisions of pricing by marketing department and lead time quotation by production departments, with either marketing or production as the leader. They compared scenarios in which none, one, or both of the firms are decentralized. The market system is modeled as a parallel M/M/1. They found that under intense price competition, firms may suffer from a decentralized structure. In contrast, under intense lead time competition, a decentralized strategy with marketing as the leader can not only result in significantly higher profits, but also be the equilibrium strategy. Moreover, decentralization may no longer lead to lower prices or longer lead times if the production department chooses capacity along with lead time.

Cooperation models In papers dealing with cooperative firms, we have Liu et al. (2007); Xiao et al. (2011); Xiao and Shi (2012); Zhu (2015); and Xiao and Qi (2016).

Liu et al. (2007) studied a decentralized supply chain consisting of a supplier and a retailer facing a price- and lead time-sensitive demand. They constructed a Stackelberg game to analyze the price and lead time decisions by the supplier as leader and the retailer as follower. Upon receiving an order from the retailer, the supplier completes the finished product and delivers it to the retailer (or to the customer directly on behalf of the retailer). The product is not unique in the market and potential customers for the product are sensitive to both price and promised lead time. This requires the supply chain to offer a competitive retail price and quoted lead time. They assumed the supplier system to be modeled as a single-server queue (M/M/1) with the exponential service time and constant unit production cost. The retailer has zero lead time because no production process is occurred. They also assumed that demand process is Poisson where demand itself is a linear function in promised lead time and retailer price. They found that decentralized decisions are inefficient and lead to inferior performance due to the double marginalization effect. The decision inefficiency is strongly influenced by market and operational factors. Before pursuing a coordination strategy with retailers, a supplier should first improve his or her own internal operations.

Xiao et al. (2011) investigated coordination of a (global) supply chain consisting of one manufacturer and one retailer via a revenue-sharing contract, where a product quality assurance policy is provided and the utility of consumer is sensitive to product (physical) quality, service quality (i.e., reciprocal of delivery lead time) and retail price. The supply chain operates in a MTO environment and a defective product is returned to the manufacturer for free re-manufacturing. Consumers return the imperfect products to the manufacturer for re-manufacturing within a guaranteed period. Every reworked unit is as good as new after rework and the manufacturer requires no rework charges for consumers or the retailer. They considered centralized

and decentralized settings. In the centralized setting, all decisions are taken simultaneously via a moderator or by a higher authority. In the decentralized setting, decisions are taken as a sequence. Manufacturer determines both unit wholesale price and quoted lead time, then the retailer determines retail price following the decision of manufacturer. Consumers arrive at the retailer following a Poisson process. The order is processed in a first-come-first-served fashion and the production time is exponentially distributed. The unit production cost is assumed to be constant. The manufacturer follows an M/M/1 system. And the retailer only acts as mediator without production process happen in retailer, thus the retailer lead time is equal to zero. This demand function is a result of utility function which depends on product physical quality, lead time and return loss. In the decentralized setting, Xiao et al. (2011) found that: 1. a higher defective rate of the final product implies a higher cost for the manufacturer, 2. the optimal service quality first decreases and then increases, and 3. the optimal retail price decreases as the defective rate increases. In the coordinated supply chain, the manufacturer charges the retailer a higher unit wholesale price.

Xiao and Shi (2012) considered a supply chain consisting of one make-to-order (MTO) manufacturer and one retailer in a price and lead time sensitive market. They developed a Manufacturer-Stackelberg (MS) model with predetermined (exogenous) lead time standard and a MS model with endogenous lead time standard. Following Boyaci and Ray (2003), they assumed that the manufacturer provides two substitutable products: regular product and faster product. The difference between the two kinds of products only lies in lead time standard. When a consumer arrives at the retailer, the consumer determines to buy the regular product or the faster product. Once consumer made an order, the retailer sends the order to the manufacturer immediately and then the manufacturer sets up production. So, the retailer serves as an intermediary between consumers and the manufacturer. Both players aim to maximize their long-term average profits. The manufacturer has a dedicated facility for each product. Both facilities process orders following a first-come-first-serve service discipline. Each facility is an M/M/1 queuing service system with Poisson arrival and exponential service time. Similar to Boyaci and Ray (2003), the unit production costs (excluding capacity cost) for the two products are identical, but the faster product has a higher unit capacity cost than the regular product. Their demand model is similar to Boyaci and Ray (2003). They found that when the lead time sensitivity increases, the players decrease the unit wholesale prices and retail prices for the two products and reduce the lead time standard for the faster product. The manufacturer builds a lower capacity for the regular product while a higher capacity for the faster product.

Zhu (2015) considered a decentralized supply chain consisting of a supplier and a retailer facing price- and lead time-sensitive demand. The decision process is modeled by a Stackelberg game where the supplier, as a leader, determines the capacity and the wholesale price, and the retailer, as a follower, determines the sale price and lead time. Unlike Liu et al. (2007) who considered either pricing and capac-

ity decisions or pricing and lead time decisions, Zhu (2015) studied the integration of pricing, lead time, and capacity decisions. The supply chain framework of Zhu (2015) is similar to Liu et al. (2007) where the supplier has M/M/1 production system and retailer has zero lead time. Zhu (2015) found that the integration of pricing, lead time, and capacity decision can significantly reduce the profit loss caused by double marginalization. The revenue-sharing and two-part tariff contracts cannot coordinate the decentralized channel. Instead, a franchise contract with a contingent rebate can achieve channel coordination and a win–win outcome.

Xiao and Qi (2016) studied the equilibrium decisions in the supply chain with an all-unit quantity discount contract. They considered a two-stage supply chain with one supplier and one manufacturer. The manufacturer faces a Poisson demand process where the arrival rate depends on the selling price, the announced delivery time, and the delivery reliability defined as the probability of satisfying the announced delivery time. The supplier produces a standard product in MTS mode. The manufacturer purchases standard products from the supplier at a unit wholesale price, takes orders from end users, customizes the standard products based on order specifications, and delivers the final products to end users. The supplier and the manufacturer maximize their long-term average profit per time unit. They assumed that the supply chain adopts vendor managed inventory (VMI) mode to manage the supply chain's inventory, and the production rate of the supplier is greater than the manufacturer's service rate (capacity). Under VMI, the effect of resource constraint on the manufacturer can be ignored because the MTS supplier has a greater production rate than the manufacturer's capacity and can well control the inventory of standard products. Xiao and Qi (2016) used the M/M/1 queuing model to describe the operations of the manufacturer. They considered coordination of a decentralized supply chain where supplier decides the wholesale price, and manufacturer decides jointly the resale price and the quotation of lead time. They considered four scenarios regarding whether the lead time standard, the delivery reliability standard, and the manufacturer's capacity are endogenous, and whether the manufacturer's production cost is its private information. They found that an all-unit quantity discount scheme can coordinate the supply chain for most cases.

2.4 Other related papers: the MTS system

Other papers related on lead time sensitive demand models are in the MTS context. Examples include Savaşaneril et al. (2010); Wu et al. (2012); Panda (2013); and Savaşaneril and Sayin (2017).

Savaşaneril et al. (2010) studied a dynamic lead time quotation problem in a base-stock inventory system characterized by lead time sensitive Poisson demand and exponentially distributed service times. They show that the optimal profit is unimodal in the base-stock level. They compare the base-stock system with a make-to-order (MTO) system and show that the lead time quotes are lower in an MTO system and that increasing the base-stock level does not necessarily decrease the

expected number of customers waiting.

Wu et al. (2012) studied the news-vendor problem with endogenous demand sensitive to price and quoted lead time. The problem is observed in situations where a firm orders semi-finished product prior to the selling season and customizes the product in response to customer orders during the selling season. The total demand during the selling season and the lead time required for customization are uncertain. The demand for the product depends not only on the selling price but also on the quoted lead time. To set the quoted lead time, the firm has to carefully balance the benefit of increasing demand as the quoted lead time is reduced against the cost of increased tardiness.

Panda (2013) dealt with the coordination of a supply chain that consists of a manufacturer and a price setting retailer. The manufacturer offers a single product to the retailer, who faces time and price sensitive demand. Under explicit cost in- formation, optimal quantity–price pairs are derived for an integrated scenario and a decentralized scenario by considering the manufacturer as the Stackelberg leader. The objective of Panda (2013) is to determine price, order quantity and replenishment cycle length in order to maximize the total profit of the chain. In their model, demand is a function of time and price.

Savaşaneril and Sayin (2017) addressed the lead time quotation problem of a manufacturer serving multiple customer classes. Customers are sensitive to the quoted lead times and the manufacturer has the flexibility to keep inventory to improve responsiveness. They model the problem as a Markov decision process and characterize the optimal lead time quotation, rationing, and production policies.

2.5 Conclusion of literature review

From the literature review, we make the following observations. First, all papers assume that the unit production cost is a constant. In practice, a firm can manage better its production system when it quotes longer lead time. Thus, we will extend the existing models by considering the direct relation between the unit production cost and the quoted lead time in chapter 3. Second, in the single firm case, the vast majority of works use the M/M/1 system. Although M/M/1 is the simplest queuing model, it has a drawback since all the customers are accepted which might lead to long sojourn time in the system. Thus, we propose to test whether a new policy, which consists of rejecting customers when there are already a certain number of customers in the system, can be more profitable. We call this policy a rejection policy. The detailed discussion is provided in chapter 4. Third, in the case of multi-firm, all works assume that only one of the actors has a production process (queuing system), the other actor only acts as a mediator with a lead time that equals to zero. Thus, we will introduce a supply chain where the two actors have a production system in chapter 5.

In this thesis, we will consider MTO systems, uniform lead time quotation, single-class product, single- class customer, and linear demand model and we will propose

three contributions:

1. Introduce a unit production cost that depends on the quoted lead time,
2. Consider a policy allowing to reject the clients by using an M/M/1/K queue,
3. Introduce a supply chain (multi-firm) where the two actors have a production process, which leads to a tandem queue M/M/1-M/M/1.

These three problems are investigated in the following chapters to achieve the PhD objective.

Chapter 3

Lead time sensitive cost: a production system with demand and production cost sensitive to lead time (M/M/1 model)

Some research has been done on M/M/1 queuing systems with lead time- and price-dependent demand. However, all of the works, presented in chapter 2, assume a constant production cost. It is known that the longer the quoted lead time the better the firm can manage the production and reduce the cost. The idea of this paper stems from this observation. Indeed, we investigate the lead time quotation decision in an M/M/1 make-to-order queue while assuming the production cost to be a decreasing function in lead time. We consider three settings: (1) price is fixed and holding and lateness costs are ignored, (2) price is also a decision variable (in addition to lead time) but holding and lateness costs are ignored, and (3) price is a decision variable and holding and lateness costs are considered. For setting 1 and 2, we provide an approach to find analytically the optimal lead time and price (if it is a variable). And for setting 3, we develop a numerical approach to solve the case. We use the optimal solutions to conduct experiments and derive some insights.

3.1 The model

We model the firm as an M/M/1 Make-to-Order queue. Customers arrive and are served in a first-come-first-served fashion. The demand follows a Poisson process of mean arrival rate λ, which cannot be larger than the amount of demand $\Lambda(P, L)$ obtained when price, P, and lead time, L, are quoted to the customers. As usually assumed in the literature (see the aforementioned papers), we consider that the demand linearly decreases with price and quoted lead time, $\Lambda(P, L) = a - b_1 P - b_2 L$ where b_1 and b_2 are the sensitivity coefficients to price and lead time, respectively. The production capacity is constant (μ) and the service time is exponentially distributed.

It is well known that the unit operating cost depends on the quoted lead time in many situations. In particular, firms that quote short lead time to their customers, and that are consequently exposed to high risk, do not focus only on production capacity but rethink the different decisions along the supply chain and align them

with the lead time strategy to reduce the risk as much as possible. Examples include buying items from quick-response but expensive suppliers instead of regular low-cost suppliers (e.g. local suppliers instead of suppliers abroad), holding higher stock of raw materials in the upstream stages, using faster but more costly transportation modes (e.g., in the logistics industry, FedEx Express uses airplanes while FedEx ground uses trucks), and rethinking the production/delivery process by moving the tradeoff toward shorther lead time instead of cheaper production (e.g., less economies of scale in production, less grouped shipments, etc.). In all these examples, the shorter the quoted lead time the more difficult the required actions become and, consequently, the higher the unit production cost.

Thus, we do not assume a constant production cost but consider the unit production cost (m) to be a decreasing function in quoted lead time (L). Indeed, m is given by the following non-linear function: $m = C_1 + \frac{C_2}{L}$. This function implies that the increase in unit production cost that results from one unit decrease in lead time is not constant (as in linear functions) but is growing with smallest values of lead time. In other words, the smaller the value of the lead time the more difficult (and, consequently, the more expensive), its reduction becomes. Clearly, the production cost usually used in the existing literature is a particular case of our cost function with $C_2 = 0$.

The decision variables and parameters for this chapter are given below:

Decision variables:

λ : mean arrival rate (demand),
L : quoted lead time,
P : price of the goods/service,
m : unit production cost.

Parameters:

a : market potential,
b_1 : price sensitivity of demand,
b_2 : lead time sensitivity of demand,
μ : mean service rate (production capacity),
s : service level defined by company ($s \in [0, 1]$),
F : unit holding cost,
c_r : penalty per job per unit lateness,
C_1, C_2 : production cost parameters.

The firm's objective is to maximize the total expected profit, which equals to revenue (λP) – production cost (λm) – expected holding cost ($F\lambda/(\mu - \lambda)$) – expected lateness penalty cost ($c_r(\lambda/(\mu-\lambda))e^{-(\mu-\lambda)L}$). The lateness cost reflects direct compensation paid to customers for not meeting the quoted lead time. Firm has to make sure to meet its quoted lead time within a service rate (s). The expected congestion cost is given by $F\lambda/(\mu - \lambda)$ where F is the unit holding cost, and $\lambda/(\mu - \lambda)$ is the mean inventory. The expected lateness penalty can be given by (penalty per

job per unit lateness) × (throughput rate) × (probability that a job is late) × (expected lateness given that a job is late). Thus, the expected lateness penalty is: $c_r(\lambda/(\mu-\lambda))e^{-(\mu-\lambda)L}$ where c_r is the penalty per job per unit lateness, $e^{-(\mu-\lambda)L}$ is the probability that a job is late and $\lambda/(\mu-\lambda)$ is the (throughput rate) × (expected lateness) (see Palaka et al., 1998). The formulation of our general model is given below.

$$\underset{L,P,\lambda,m}{\text{Maximize}}\ \lambda(P-m) - \frac{F\lambda}{\mu-\lambda} - \frac{c_r\lambda}{\mu-\lambda}e^{-(\mu-\lambda)L} \tag{3.1}$$

$$\text{Subject to } \lambda \leq a - b_1P - b_2L \tag{3.2}$$

$$1 - e^{-(\mu-\lambda)L} \geq s \tag{3.3}$$

$$\lambda \leq \mu \tag{3.4}$$

$$m = C_1 + \frac{C_2}{L} \tag{3.5}$$

$$\lambda, L, P, m \geq 0 \tag{3.6}$$

The objective function is given in Equation (3.1). Equation (3.2) ensures that the mean demand rate received by the firm cannot exceed the demand generated by the quoted price and lead time. Equation (3.3) guarantees that the probability of meeting the quoted lead time, given by $1 - e^{-(\mu-\lambda)L}$ (since $e^{-(\mu-\lambda)L}$ is known to be the probability that a job is late in M/M/1 queue), must not be smaller than the required service level. Equation (3.4) guarantees a steady state of the M/M/1. Equation (3.5) defines the production cost function. The non-negativity constraints are given in Equation (3.6).

As stated before, we consider three settings: (1) Fixed price without holding and lateness penalty cost, (2) Price as decision variable without holding and lateness cost, and (3) Price as decision variable with holding and lateness cost. Before moving to the solving approach of each setting, we point out the following general result, which is verified under all settings.

Lemma 3.1. *Demand constraint (3.2) is binding and we have* $\lambda = a - b_1P - b_2L$ *at optimality.*

Proof. Suppose that the optimal solution is given by quoted lead time L^*, price P^*, and demand rate λ^* such that $\lambda^* < a - b_1P^* - b_2L^*$. The profit increases when L increases as an increase in L leads to decrease in production cost ($m = C_1 + \frac{C_2}{L}$) and in lateness penalty cost (for setting 3). An increase in L also leads to decreasing the probability of having late delivery ($e^{-(\mu-\lambda)L}$). Thus, by keeping price P^* and demand rate λ^* constant, one could increase the lead time from L^* to L' until $\lambda^* = a - b_1P^* - b_2L'$. This change will increase profit as production cost decreases, lateness penalty cost decreases, and holding cost stays the same (because λ doesn't change), while the service level constraint remains satisfied. Therefore, we would get a new solution: P^*, λ^*, and L' with a better profit than the supposed optimal solution, which is impossible. ■

Given the result of lemma 3.1, we will use $\lambda = a - b_1P - b_2L$ for the rest of this chapter. Note also that constraint (3.4) is immediately satisfied if service constraint is satisfied and the lead time is positive. Therefore, constraint (3.4) can be removed.

3.2 Setting 1: Model with variable lead time and fixed price

In this section, we consider the model when price is fixed (not a decision variable).

3.2.1 Optimal policy with variable lead time and fixed price

We solve analytically the model when price is fixed. By considering the result of Lemma 3.1 and integrating the expression of production cost (m) into the objective function, the problem can be formulated as a constrained non-linear optimization model with a single decision variable L. Indeed, we respectively replace λ and m by $a - b_1P - b_2L$ and $C_1 + \frac{C_2}{L}$. Thus, the model with fixed price becomes equivalent to the following model.

$$\underset{L\geq 0}{\text{Maximize}}\ f(L) = (a - b_1P - b_2L)\left(P - C_1 - \frac{C_2}{L}\right) \tag{3.7}$$

$$\text{Subject to } (\mu - a + b_1P)L + b_2L^2 \geq \ln\left(\frac{1}{1-s}\right) \tag{3.8}$$

$$L \leq \frac{a - b_1P}{b_2} \tag{3.9}$$

Equation (3.8) represents the service constraint after rewriting $1 - e^{-(\mu-\lambda)L} \geq s$ as $(\mu - \lambda)L \geq \ln\left(\frac{1}{1-s}\right)$. In the rest of the paper, we let γ denote $\ln\left(\frac{1}{1-s}\right)$. Equation (3.9) guarantees that $\lambda \geq 0$, which is equivalent to $L \leq \frac{a-b_1P}{b_2}$. In order to solve the model, we firstly consider the following lemma.

Lemma 3.2. *The feasible region of profit $f(L)$ is defined by:*

$$\left[\max\left\{\frac{C_2}{P-C_1}, \frac{a-b_1P-\mu+\sqrt{(a-b_1P-\mu)^2+4b_2\gamma}}{2b_2}\right\}, \frac{a-b_1P}{b_2}\right]$$

Proof. Service level constraint is satisfied iff $(\mu - a + b_1P)L + b_2L^2 \geq \gamma$ (see eq. (3.8)). The equation $(\mu - a + b_1P)L + b_2L^2 - \gamma = 0$ has a discriminant $\Delta = (\mu - a + b_1P)^2 + 4b_2\gamma$ and two roots: $\frac{a-b_1P-\mu\pm\sqrt{(a-b_1P-\mu)^2+4b_2\gamma}}{2b_2}$. Thus, the service level constraint (3.8) is satisfied iff: $L \geq \frac{a-b_1P-\mu+\sqrt{(a-b_1P-\mu)^2+4b_2\gamma}}{2b_2}$, which is positive, or $L \leq \frac{a-b_1P-\mu-\sqrt{(a-b_1P-\mu)^2+4b_2\gamma}}{2b_2}$, which is negative. Furthermore, we have $L \geq \frac{C_2}{P-C_1}$ (indeed to have a positive profit, we must have cost lower than price: $P - C_1 - \frac{C_2}{L} \geq$

$0 \Leftrightarrow L \geq \frac{C_2}{P-C_1}$) and $L \leq \frac{a-b_1P}{b_2}$ (eq.(3.9)). Consequently, we have a feasible domain of L in the interval: $\left[\max\left\{\frac{C_2}{P-C_1}, \frac{a-b_1P-\mu+\sqrt{(a-b_1P-\mu)^2+4b_2\gamma}}{2b_2}\right\}, \frac{a-b_1P}{b_2}\right]$. ■

Given the result of lemma 3.2, we must consider the following conditions to make the problem feasible.

- We must assume that $\frac{C_2}{P-C_1} \leq \frac{a-b_1P}{b_2}$ and $\frac{a-b_1P-\mu+\sqrt{(a-b_1P-\mu)^2+4b_2\gamma}}{2b_2} \leq \frac{a-b_1P}{b_2}$ since otherwise the feasible region is empty and the model is infeasible.
- We are going to prove that the peak of the curve $f(L)$ is in the interval $\left[\frac{C_2}{P-C_1}, \frac{a-b_1P}{b_2}\right]$:
  - First, over the interval of $[0, +\infty]$, the profit $f(L)$ is concave (proven by negative second derivative function $\frac{\partial^2}{\partial L^2}f(L) = -\frac{2C_2(a-b_1P)}{L^3} < 0$ for $L > 0$) and reaches its maximum at $\sqrt{\frac{(a-b_1P)C_2}{(P-C_1)b_2}}$ (proven by $\frac{\partial}{\partial L}f(L) = 0 \Leftrightarrow L^2 = \frac{(a-b_1P)C_2}{(P-C_1)b_2}$ with the only positive root $L = \sqrt{\frac{(a-b_1P)C_2}{(P-C_1)b_2}}$).
  - Second, $\frac{C_2}{P-C_1} < \sqrt{\frac{(a-b_1P)C_2}{(P-C_1)b_2}} \Leftrightarrow \left(\frac{C_2}{P-C_1}\right)^2 < \frac{(a-b_1P)C_2}{(P-C_1)b_2} \Leftrightarrow \frac{C_2}{P-C_1} < \frac{a-b_1P}{b_2}$.
  - Third, $\sqrt{\frac{(a-b_1P)C_2}{(P-C_1)b_2}} < \frac{a-b_1P}{b_2} \Leftrightarrow \frac{(a-b_1P)C_2}{(P-C_1)b_2} < \left(\frac{a-b_1P}{b_2}\right)^2 \Leftrightarrow \frac{C_2}{P-C_1} < \frac{a-b_1P}{b_2}$.

In the published papers, in which the production cost is assumed to be constant, the service constraint is binding, which simplifies the solving approach. In our model, the service level constraint is not necessarily binding. Indeed, if we have a given lead time L for which the service constraint is not tight, we could reduce L without violating this constraint. By decreasing L, we increase the demand and, consequently, the revenue λP. But, we also increase unit production cost. Thus, this does not necessarily improve the overall profit. This trade-off may lead to non-binding situations for service level constraint (3.3) or (3.8).

In the following lemma, we identify the candidates for optimality in each of binding and non-binding situations.

Lemma 3.3. *The service level constraint is not necessarily binding, and we have:*

- *In the binding situation, the candidate for optimality of the base model is* $L^B = \frac{a-b_1P-\mu+\sqrt{(a-b_1P-\mu)^2+4b_2\gamma}}{2b_2}$ *with* $\gamma = \ln\left(\frac{1}{1-s}\right)$,
- *In the non-binding situation, the candidate for optimality of the base model, if satisfying the service constraint (i.e.* $L^{NB} \geq L^B$*), is* $L^{NB} = \sqrt{\frac{(a-b_1P)C_2}{(P-C_1)b_2}}$.

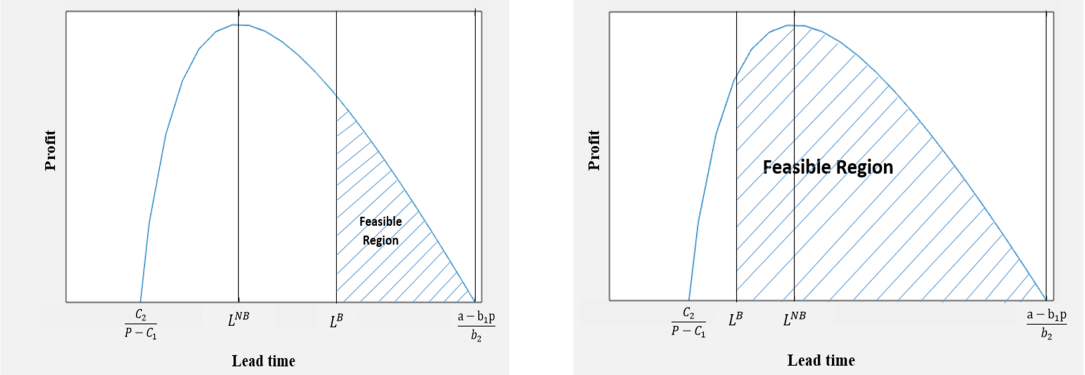
Proof. **Binding situation (figure 3.1a):** In this case, the service level constraint (3.8) is binding, implying that $(\mu - a + b_1P)L + b_2L^2 = \gamma$ at optimality. Thus, we can obtain the lead time L directly from service level equation. Quadratic

equation $b_2L^2 + (\mu - a + b_1P)L - \gamma = 0$ has positive discriminant $\Delta = (\mu - a + b_1P)^2 + 4b\gamma$, implying that we have two roots, with only one positive root $L_1 = \frac{a-b_1P-\mu+\sqrt{(\mu-a+b_1P)^2+4b\gamma}}{2b_2}$. Thus, in binding situation, the candidate for optimality of base model is $L^B = L_1$.

Non-binding situation (figure 3.1b): If we ignore the service level constraint (3.8), the problem becomes:

$$\underset{0\leq L\leq \frac{a-b_1P}{b_2}}{\text{Maximize}} \qquad f(L) = (a - b_1P - b_2L)\left(P - C_1 - \frac{C_2}{L}\right)$$

We have $\frac{\partial}{\partial L}f(L) = \frac{(a-b_1P)C_2}{L^2} + (C_1 - P)b_2$. Thus, $\frac{\partial}{\partial L}f(L) = 0 \Leftrightarrow L^2 = \frac{(a-b_1P)C_2}{(P-C_1)b_2}$. We only consider the positive root, which is $L = \sqrt{\frac{(a-b_1P)C_2}{(P-C_1)b_2}}$. Thus the candidate for optimality in the non-binding situation, if satisfying the service constraint (i.e. $L^{NB} \geq L^B$), is $L^{NB} = \sqrt{\frac{(a-b_1P)C_2}{(P-C_1)b_2}}$ which is smaller than $\frac{a-b_1P}{b_2}$. ■



(a) Case: $L^B > L^{NB}$ (b) Case: $L^B \leq L^{NB}$

Figure 3.1: Illustration of situations: $L^B \leq L^{NB}$ and $L^B > L^{NB}$

Based on the results of the previous lemma, we can now announce the optimal lead time of the base model with variable lead time and fixed price.

Proposition 3.1. *The optimal solution of the problem with variable lead time and fixed price is* $L^* = \max(L^B, L^{NB})$*, (where* L^B *and* L^{NB} *are defined as in Lemma 3.3)*

Proof. Based on the binding and non-binding situations, we have two cases: $L^B \leq L^{NB}$ or $L^B > L^{NB}$.

Case of $L^B > L^{NB}$ (figure 3.1a): In this case, L^{NB} doesn't satisfy the service constraint. Given the concavity of $f(L)$ and its maximum in $\sqrt{\frac{(a-b_1P)C_2}{(P-C_1)b_2}} = L^{NB}$, which is lower than L^B, we deduce that the best feasible value of lead time is L^B.

Case of $L^B \leq L^{NB}$ (figure 3.1b): In this case, L^{NB} satisfies the service constraint as it is greater than L^B. Given in addition that $f(L)$ is concave in the range of $[0, +\infty]$ and reaches its maximum in $\sqrt{\frac{(a-b_1P)C_2}{(P-C_1)b_2}} = L^{NB}$, which is equal to or greater than L^B, we deduce that $f(L^B) \leq f(L^{NB})$. Consequently, the optimal solution is L^{NB}. ■

Before moving to the experiments and insights, one can verify that when $C_2 = 0$ (i.e. when the unit production cost is constant), we obtain $L^{NB} = 0$, and the optimal $L^* = L^B$ according to proposition 3.1. Thus, for $C_2 = 0$ we find the optimal solution given in the literature for model with constant unit production cost (see Pekgün et al., 2008).

3.2.2 Experiments and insights with fixed price

To illustrate our analytical results and understand better the behavior of the model, we conduct experiments with the following base example with parameters: $a = 50$, $b_1 = 4$, $b_2 = 6$, $\mu = 10$, $s = 0.95$, $C_1 = 2$, and $C_2 = 3$. This parameters, except C_1 and C_2, are taken from Pekgün et al. (2008). In particular, we study the effect of customers' sensitivity to lead time (b_2) and cost sensitivity to lead time (C_2). We do experiments by varying one parameter and keep the other parameters constant. In each case, we provide the optimal values of quoted lead time, demand, revenue, total cost, profit ($\Pi 1$) (i.e., revenue – total cost), and observed service level (which is not necessarily the service level s, initially defined by the company).

Effect of customers' sensitivity to lead time (b_2) with fixed price

We fix the price to 8.81 (this price is obtained from the optimization with price as decision variable and $b_2 = 2$, see table 3.4) and vary the value of b_2 from 2 to 20. The results are reported in Table 3.1. Observing Table 3.1, we derive the following insights:

- As expected, an increase in customers' sensitivity to lead time (b_2) leads to a decrease in quoted lead time. Note also that, despite the decrease of quoted lead time, which favors attracting more customers, the demand and profit always go down when b_2 goes up.

- Unlike the demand and profit, the total cost is not a monotonous function in lead time sensitivity (b_2). Indeed, for values of b_2 ranging from 2 to 10, total production cost increases with an increase in b_2. This region corresponds to the cases where service constraint is binding (observed service level = s). Then, an increase in b_2 from 12 to 20 leads to decreasing the total cost. This region corresponds to non-binding situations (observed service level > s). In this case, the effect of demand decrease on reducing the total cost becomes

Table 3.1: Effect of demand sensitivity to lead time (b_2) with fixed price

b_2	Lead time	Demand	Total cost	Profit ($\Pi 1$)	$\Pi 1'$	Gains	Serv. lev. realized
2	2.89	8.96	27.23	51.77	51.77	0.0%	95%
4	1.64	8.18	31.28	40.78	40.78	0.0%	95%
6	1.21	7.51	33.73	32.50	32.50	0.0%	95%
8	0.98	6.93	35.16	25.94	25.94	0.0%	95%
10	0.83	6.41	35.87	20.61	20.61	0.0%	95%
12	0.74	5.92	35.98	16.19	16.19	0.0%	95.03%
14	0.68	5.21	33.39	12.55	12.50	0.4%	96.16%
16	0.64	4.55	30.56	9.58	9.41	1.9%	96.89%
18	0.60	3.94	27.53	7.16	6.81	5.1%	97.38%
20	0.57	3.35	24.35	5.19	4.63	12.1%	97.74%

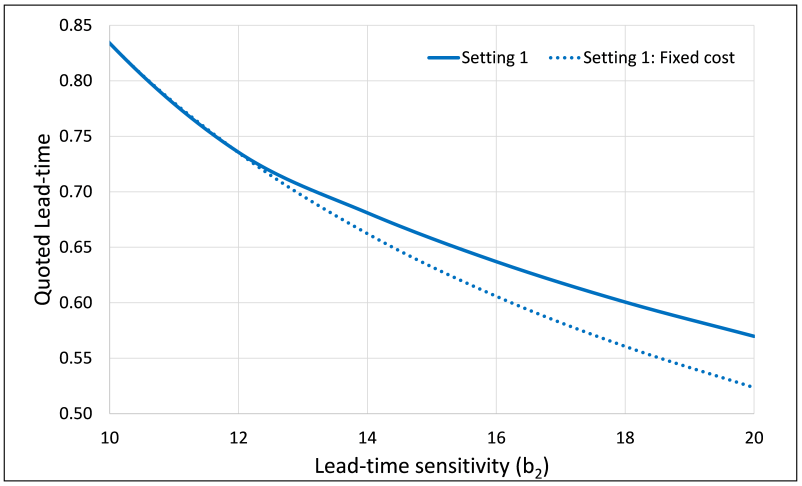


Figure 3.2: Variable cost vs Fixed cost: quoted lead time

more important than the effect of lead time decrease on increasing the total cost.

- It is important to note that when b_2 becomes significantly large (here, $b_2 \geq 12$), the model chooses to do better than the required service level (non-binding situation). More generally, we observe that the higher the demand sensitivity to lead time the higher the service level. This results is not intuitive since an increase in lead time sensitivity leads to shorter lead time that is more difficult to guarantee and, consequently, one can expect that this will lead to tight service constraint. In our problem, the model has to quote a short lead time in order to maintain a certain amount of demand. However, this implies a high unit production cost, which may annihilate the gain obtained by reducing

the lead time. Consequently, the model reacts by reducing quoted lead time but not as much as allowed by service constraint, which implies a non-binding situation. Clearly, this behavior cannot be captured by the existing models where unit production cost is constant.

- In order to show the impact of taking into account the sensitivity of cost to lead time (as we do in our model), we re-solve the model while considering a constant unit production cost. Indeed, we calculate the average unit cost $(C_1 + \frac{C_2}{L})$ over all instances of Table 3.1 and take the obtained value as a constant cost in these new experiments. In other words, as the average unit cost in Table 3.1 is 5.55, we consider in these experiments $C_1 = 5.55$, and $C_2 = 0$ (cost does not depend here on lead time). We firstly compare the results of our base model to the results of the model with constant cost in terms of quoted lead time. The results, reported in Fig. 3.2, show that both models leads to the same optimal lead time for low values of b_2, which corresponds to the binding situation. In this case, we can check analytically that the quoted lead time does not depend neither on C_1 nor on C_2 (lemma 3.3). However, when the customers are very sensitive to lead time (non-binding-situation), we see that the lead time quoted by our base model is greater than the lead time quoted by the model with constant cost.

- To evaluate the impact of not quoting the right lead time when the cost is assumed constant, we take the optimal lead time in this case and inject it in the objective function of our base model (with variable cost). We denote the obtained profit by $\Pi 1'$ and let $\Pi 1$ denote the optimal profit of our base model (given in Table 3.1). Then, we calculate the percentage of gain $= \frac{100\times(\Pi 1-\Pi 1')}{\Pi 1'}$ for different values of b_2. We obtain an average gain 4.81% in non-binding situations. Over all instances, the average gain is 1.93%.

Effect of cost sensitivity to lead time (C_2) with fixed price

We fix the price to 8.35 (this price is obtained from the optimization with price as decision variable and $C_2 = 2$, see table 3.5) and vary the value of C_2 from 2 to 10. The results are reported in Table 3.2. Observing Table 3.2, we deduce the following:

- At first, an increase in C_2 does not impact the quoted lead time (this corresponds to the binding situation). Then, when we are in the non-binding situation, if the cost is more sensitive to lead time (C_2 increases) then the model quotes longer lead time.

- For relatively high values of C_2, it is important to note that the realized service level is very high (close to 1). In this case, the model chooses to quote a relatively long lead time (with comparison to what the existing capacity allows to do) in order to reduce the cost. Although this implies less demand, the loss

Table 3.2: Effect of cost sensitivity to lead time (C_2) with fixed price

C_2	Lead time	Demand	Total cost	Profit(Π1)	Serv. lev. realized
2	1.44	7.93	26.83	39.37	95%
3	1.44	7.93	32.32	33.88	95%
4	1.44	7.93	37.80	28.39	95%
5	1.48	7.74	41.71	22.93	96.44%
6	1.62	6.89	39.38	18.20	99.34%
7	1.75	6.12	36.77	14.33	99.89%
8	1.87	5.39	33.91	11.14	99.98%
9	1.98	4.72	30.87	8.51	99.99%
10	2.09	4.07	27.67	6.35	99.99%

of revenue is here offset by the saving in cost. Thus, when the operating cost is very sensitive to lead time, it is optimal for the system to work like an almost guaranteed service model (i.e., 99.99% of demands are satisfied on time, no lateness) despite the uncertainties in both demand and processing time.

3.3 Setting 2: Model with both lead time and price as decision variables

In this section, we consider a more complex setting with price P as decision variable in addition to lead time.

3.3.1 Optimal policy with both lead time and price as decision variables

Given that demand constraint is binding (as demonstrated in Lemma 3.1), we deduce that $P = \frac{a-b_2L-\lambda}{b_1}$. Thus, we can formulate the problem with only two variables (L and λ), as follows. Objective function is given in Equation (3.10). Service constraint, expressed as a function of L and λ, is given in (3.11). As the price must be positive, we must have $a - b_2L - \lambda \geq 0$, which is guaranteed by constraint (3.12).

$$\underset{L,\lambda>0}{\text{Maximize}}\ g(L,\lambda) = \lambda\left(\frac{a-b_2L-\lambda}{b_1} - C_1 - \frac{C_2}{L}\right) \tag{3.10}$$

$$\text{Subject to } \lambda \leq \mu - \frac{\gamma}{L} \tag{3.11}$$

$$\lambda \leq a - b_2L \tag{3.12}$$

The model above is still hard to solve. We start by finding an alternative formulation of the model above that will be easier to solve. With this scope in mind, we present intermediate results in Lemma 3.4, 3.5 and 3.6, and then provide an equivalent formulation of model above in Proposition 3.2. The optimal policy will be described in Proposition 3.3.

Lemma 3.4. *The feasible region of quoted lead time L is $[L^{\min}, L^{\max}]$, where* $L^{\min} = \max\left\{\frac{a-C_1b_1-\sqrt{(a-C_1b_1)^2-4b_1b_2C_2}}{2b_2}, \frac{\gamma}{\mu}\right\}$ *and* $L^{\max} = \frac{a-C_1b_1+\sqrt{(a-C_1b_1)^2-4b_1b_2C_2}}{2b_2}$.

Proof. Given that $0 < \lambda \leq \mu - \frac{\gamma}{L}$ (equation 3.11) and $0 < \lambda \leq a - b_2L$ (equation 3.12), the lead time L must belong to $\left[\frac{\gamma}{\mu}, \frac{a}{b_2}\right]$. For any given L, profit function $g(L,\lambda)$ is concave in λ (proven by $\frac{\partial^2}{\partial\lambda^2}g(L,\lambda) = -\frac{2}{b_1} < 0$). Therefore, for given L, $g(L,\lambda)$ reaches its maximum in $\lambda = \frac{1}{2}(a - C_1b_1 - b_2L - \frac{C_2b_1}{L})$ (obtained form the first derivative of $g(L,\lambda)$). For the problem to be feasible, this demand must be positive and therefore $\lambda = \frac{1}{2}(a - C_1b_1 - b_2L - \frac{C_2b_1}{L}) \geq 0 \Leftrightarrow (a - C_1b_1)L - b_2L^2 - C_2b_1 \geq 0$. This quadratic equation has a discriminant $(a - C_1b_1)^2 - 4b_1b_2C_2$. Then, we must assume $(a - C_1b_1)^2 - 4b_1b_2C_2 \geq 0$ since, otherwise, $a - C_1b_1 - b_2L - \frac{C_2b_1}{L}$ will be negative for any positive value of L, which implies that demand will be negative. We assume this condition holds, thus the two roots of quadratic function $a - C_1b_1 - b_2L - \frac{C_2b_1}{L}$ are $L_1^{\min}$ and $L_1^{\max}$. Then we must consider the values of L such as $L_1^{\min} \leq L \leq L_1^{\max}$ where $L_1^{\min} = \frac{a-C_1b_1-\sqrt{(a-C_1b_1)^2-4b_1b_2C_2}}{2b_2}$ and $L_1^{\max} = \frac{a-C_1b_1+\sqrt{(a-C_1b_1)^2-4b_1b_2C_2}}{2b_2}$.
We can deduce that $L^{\min} = \max\left\{\frac{a-C_1b_1-\sqrt{(a-C_1b_1)^2-4b_1b_2C_2}}{2b_2}, \frac{\gamma}{\mu}\right\}$.
Next, we check that:

$$\begin{aligned}\frac{a - C_1b_1 + \sqrt{(a-C_1b_1)^2 - 4b_1b_2C_2}}{2b_2} \leq \frac{a}{b_2} &\Leftrightarrow \sqrt{(a-C_1b_1)^2 - 4b_1b_2C_2} \leq a + C_1b_1 \\ &\Leftrightarrow (a-C_1b_1)^2 - 4b_1b_2C_2 \leq (a+C_1b_1)^2 \\ &\Leftrightarrow -b_2C_2 \leq aC_1\end{aligned}$$

Given that $a, b_2, C_1, C_2 \geq 0$, this condition is always true.
Thus $L^{\max} = \frac{a-C_1b_1+\sqrt{(a-C_1b_1)^2-4b_1b_2C_2}}{2b_2}$. ■

Lemma 3.5. *Given lead time L in $\left[L^{\min}, L^{\max}\right]$, optimal demand* $\lambda^* = \min\left\{\frac{1}{2}(a - C_1b_1 - b_2L - \frac{C_2b_1}{L}),\ \mu - \frac{\gamma}{L}\right\}$.

Proof. Profit function $g(L,\lambda)$ is concave in λ for any given L. Indeed, $\frac{\partial^2}{\partial\lambda^2}g(L,\lambda) = -\frac{2}{b_1} < 0$. Therefore, for given L, $g(L,\lambda)$ reaches its maximum in $\lambda = \frac{1}{2}(a - C_1b_1 - b_2L - \frac{C_2b_1}{L})$. This value of λ is obtained from the $\frac{\partial}{\partial\lambda}g(L,\lambda) = 0 \Leftrightarrow \frac{aL - b_2L^2 - 2\lambda L - C_1b_1L - C_2b_1}{b_1L} = 0 \Leftrightarrow \lambda = \frac{1}{2}\left(a - C_1b_1 - b_2L - \frac{C_2b_1}{L}\right)$ (knowing that $b_1L \neq 0$).
Given the constraints on the value of λ from (3.11) and(3.12), we deduce that $\lambda^* = \min\left\{\frac{1}{2}(a - C_1b_1\ -b_2L - \frac{C_2b_1}{L}),\ a - b_2L,\ \mu - \frac{\gamma}{L}\right\}$.

However, as we are interested in the values of L such as $a - b_2L \geq 0$ (otherwise, we obtain a negative demand), we always have $\frac{1}{2}(a - C_1b_1 - b_2L - \frac{C_2b_1}{L}) \leq a - b_2L$. Hence, $\lambda^* = \min\left\{\frac{1}{2}(a - C_1b_1 - b_2L - \frac{C_2b_1}{L}),\ \mu - \frac{\gamma}{L}\right\}$. ■

In the previous lemma, we found out that, for given L, optimal demand λ^* is the minimum between two functions of L. Now, we determine the conditions under which each of this function is the minimum.

Lemma 3.6. *Denote* $\Delta = (a - C_1b_1 - 2\mu)^2 + 4b_2(2\ln\left(\frac{1}{1-s}\right) - C_2b_1)$ *and* $M = \frac{a-C_1b_1-2\mu+\sqrt{\Delta}}{2b_2}$. *For given* L,

- *If* $(\Delta \leq 0)$ *then* $\lambda^* = \frac{1}{2}(a - C_1b_1 - b_2L - \frac{C_2b_1}{L})$,
- *If* $(\Delta > 0)$, *then* $\begin{cases} \text{if} \quad L \geq M \text{ then } \lambda^* = \frac{1}{2}(a - C_1b_1 - b_2L - \frac{C_2b_1}{L}) \\ \text{if} \quad L < M \text{ then } \lambda^* = \mu - \frac{\ln\left(\frac{1}{1-s}\right)}{L} \end{cases}$

Proof. We let $h_1(L) = \frac{1}{2}(a - C_1b_1 - b_2L - \frac{C_2b_1}{L})$ and $h_2(L) = \mu - \frac{\ln\left(\frac{1}{1-s}\right)}{L}$. Let us recall that to simplify the presentation we denote $\ln\left(\frac{1}{1-s}\right)$ by γ. We need to determine the conditions under which $h_1(L) \leq h_2(L)$ and vice versa. Clearly, we are interested only in the positive values of L.
For feasible values of L, we have:

$$h_1(L) \leq h_2(L) \Leftrightarrow b_2L^2 - (a - C_1b_1 - 2\mu)L - (2\gamma - C_2b_1) \geq 0$$

The discriminant of this quadratic function is $\Delta = (a - C_1b_1 - 2\mu)^2 + 4b_2(2\gamma - C_2b_1)$. Thus, if $(\Delta \leq 0)$ then we always have $h_1(L) \leq h_2(L)$ for any feasible value of L. If $(\Delta > 0)$ then the equation has only one positive root, M. Consequently, $h_1(L) \leq h_2(L)$ if $L \geq M$ and $h_1(L) \geq h_2(L)$ if $L \leq M$. ■

Based on the result of the previous Lemma, we can now determine an equivalent formulation of the base model with only one variable, which is easier to solve.

Proposition 3.2. *We let* $g_1(L) = \frac{\left(a - C_1b_1 - b_2L - \frac{C_2b_1}{L}\right)^2}{4b_1}$ *and* $g_2(L) = \left(\mu - \frac{\gamma}{L}\right)\left(\frac{a - b_2L - \mu + \frac{\gamma}{L}}{b_1} - C_1 - \frac{C_2}{L}\right)$. *Recall that* $\Delta = (a - C_1b_1 - 2\mu)^2 + 4b_2(2\gamma - C_2b_1)$ *and* $M = \frac{a-C_1b_1-2\mu+\sqrt{\Delta}}{2b_2}$.

- *If* $\Delta \leq 0$ *then the setting 2 model is equivalent to:* $\underset{L^{\min} \leq L \leq L^{\max}}{Max}\ g_1(L)$,
- *If* $\Delta > 0$ *then*

$$\begin{cases} \text{if } M < L^{\min} \text{ then model is equivalent to } \underset{L^{\min} \leq L \leq L^{\max}}{Max} g_1(L) \\ \text{if } L^{\min} \leq M \leq L^{\max} \text{ then model is equivalent to} \\ \quad Max\left\{\underset{L^{\min} \leq L \leq M}{Max} g_2(L),\ \underset{M \leq L \leq L^{\max}}{Max} g_1(L)\right\} \\ \text{if } M > L^{\max} \text{ then model is equivalent to } \underset{L^{\min} \leq L \leq L^{\max}}{Max} g_2(L) \end{cases}$$

Proof. According to the previous lemma:

if $\Delta \leq 0$ then $g(L,\lambda)$ reaches its maximum in $\lambda = \frac{1}{2}(a - C_1 b_1 - b_2 L - \frac{C_2 b_1}{L})$. The feasible region of quoted lead time L is $[L^{\min}, L^{\max}]$. Replacing λ by $\frac{1}{2}(a - C_1 b_1 - b_2 L - \frac{C_2 b_1}{L})$ in the objective function, we immediately conclude that the base model becomes equivalent to $\underset{L^{\min}<L<L^{\max}}{Max}\ g_1(L)$.

If $\Delta > 0$ then we have to consider three situation:

- if $M < L_1^{\min}$. Here, $\lambda^* = \frac{1}{2}(a - C_1 b_1 - b_2 L - \frac{C_2 b_1}{L})$ over the relevant interval $\left[L^{\min}, L^{\max}\right]$. Hence, the model is equivalent to $\underset{L^{\min}\leq L\leq L^{\max}}{Max} g_1(L)$.

- If $L^{\min} \leq M \leq L^{\max}$ then we have to consider 2 regions: $[L^{\min}, M]$ and $[M, L^{\max}]$. For $L \in [L^{\min}, M]$, the best feasible value of λ is $\mu - \frac{\ln\left(\frac{1}{1-s}\right)}{L}$, which is positive only for $L \geq L^{\min}$ (otherwise, the problem is not relevant). Replacing λ by this value in the objective function, we deduce that, over interval $[L^{\min}, M]$, the base model is equivalent to $\underset{L^{\min}\leq L<M}{Max}\ g_2(L)$. Over interval $[M, L^{\max}]$, we obtain $\lambda = \frac{1}{2}(a - C_1 b_1 - b_2 L - \frac{C_2 b_1}{L})$ and the base model is equivalent to $\underset{M\leq L\leq L^{\max}}{Max}\ g_1(L)$. Consequently, when $\Delta > 0$, the base model is equivalent to $Max\left\{\underset{L^{\min}\leq L<M}{Max}\ g_2(L),\ \underset{M\leq L\leq L^{\max}}{Max}\ g_1(L)\right\}$.

- Case of $M > L^{\max}$. Here, $\lambda^* = \mu - \frac{\ln\left(\frac{1}{1-s}\right)}{L}$ over the relevant interval $\left[L^{\min}, L^{\max}\right]$. Hence, the model is equivalent to $\underset{L^{\min}\leq L\leq L^{\max}}{Max} g_2(L)$.

■

At this stage in proposition 3.2, we have transformed our complex problem into a set of subproblems with only one variable L. We can now provide the optimal policy of lead time quotation and pricing for our setting 2 model.

Proposition 3.3. *The optimal result of setting 2 model are:*

- *If* ($\Delta \leq 0$) *or* ($\Delta > 0$ *and* $M < L^{\min}$) *then the optimal lead time* $L^* = max\left(\sqrt{(C_2 b_1)/b_2}, \frac{\gamma}{\mu}\right)$.

- *If* ($\Delta > 0$ *and* $M > L^{\max}$) *then the optimal lead time* $L^* = \underset{x\in\{R_1,R_2,L^{\min}\}\cap[L^{\min},L^{\max}]}{\arg\max} g_2(x)$, *where* R_1 *and* R_2 *are the two potential positive roots of cubic equation* $\mu b_2 L^3 - (a\gamma - 2\mu\gamma + \mu C_2 b_1 - C_1 b_1 \gamma)\, L - 2\gamma(\gamma - C_2 b_1) = 0$ *(this equation has at maximum two positive roots).*

- *If* ($\Delta > 0$ *and* $L^{\min} \leq M \leq L^{\max}$) *then we let* $L_1^* = max\left(\sqrt{(C_2 b_1)/b_2}, M\right)$ *and* $L_2^* = \underset{x \in \{R_1, R_2, L^{\min}\} \cap [L^{\min}, M]}{\arg\max} g_2(x)$. *The optimal lead time* $L^* = \begin{cases} L_1^*, \text{ if } g_1(L_1^*) \geq g_2(L_2^*) \\ L_2^*, \text{ otherwise} \end{cases}$.

In all cases, the optimal price can be easily deduced from this expression $P^* = \frac{a - b_2 L^* - \lambda^*}{b_1}$.

Proof. Here, we consider following situations:

- Case of ($\Delta \leq 0$) or ($\Delta > 0$ and $M < L^{\min}$). The model here is equivalent to $\underset{L^{\min} \leq L \leq L^{\max}}{Max} g_1(L)$ according to Proposition 3.2. The first derivative of $g_1(L)$: $g_1'(L) = \frac{\left((a - C_1 b_1)L - b_2 L^2 - C_2 b_1\right)\left(-b_2 + \frac{C_2 b_1}{L^2}\right)}{2 b_1 L}$, implying that there are three extrema. Two of them are the roots of quadratic equation $(a - C_1 b_1) L - b_2 L^2 - C_2 b_1 = 0$ (i.e., $L_1^{\min}$ and $L_1^{\max}$), leading to a null profit. The third extrema, solution of $g_1'(L) = 0$ is $L = \sqrt{(C_2 b_1)/b_2}$, leading to positive profit $\frac{(a - C_1 b_1)^2}{4 b_1}$. Hence, $g_1(L)$ is concave over $\left[L_1^{\min}, L^{\max}\right]$, which includes $\left[L^{\min}, L^{\max}\right]$, and reaches its maximum in $\sqrt{(C_2 b_1)/b_2} \in \left[L_1^{\min}, L^{\max}\right]$. However, $\sqrt{(C_2 b_1)/b_2}$ is not necessarily feasible (i.e., does not necessarily belong to $\left[L^{\min}, L^{\max}\right]$ since $L^{\min} = \max\left\{L_1^{\min}, \frac{\gamma}{\mu}\right\}$).Consequently, $L^* = max\left(\sqrt{(C_2 b_1)/b_2}, \frac{\gamma}{\mu}\right)$.

- Case of ($\Delta > 0$ and $M > L^{\max}$). The model is equivalent to $\underset{L^{\min} \leq L \leq L^{\max}}{Max} g_2(L)$ according to Proposition 3.2. $g_2'(L) = 0 \Leftrightarrow \mu b_2 L^3 - (a\gamma - 2\mu\gamma + \mu C_2 b_1 - C_1 b_1 \gamma) L - 2\gamma(\gamma - C_2 b_1) = 0$. Using the properties of cubic equations, if the three roots are reals, it can be proven that there are at maximum two positive roots (see Appendix A). We denote the two positive roots by R_1 and R_2. Note that $g_2(L^{\min}) = g_2(L^{\max}) = 0$ and that R_1 and R_2 are not necessarily in $\left[L^{\min}, L^{\max}\right]$. Consequently, $L^* = \underset{x \in \{R_1, R_2, L^{\min}\} \cap [L^{\min}, L^{\max}]}{\arg\max} g_2(x)$.

- Case of ($\Delta > 0$ and $L^{\min} \leq M \leq L^{\max}$). Model is equivalent to $Max\left\{\underset{L^{\min} \leq L \leq M}{Max} g_2(L), \underset{M \leq L \leq L^{\max}}{Max} g_1(L)\right\}$ according to Proposition 3.2. On the one hand, we just demonstrated that $g_1(L)$ is concave over $\left[L_1^{\min}, L^{\max}\right]$ and reaches its maximum in $\sqrt{(C_2 b_1)/b_2}$. As we have $L^{\min} \leq M \leq L^{\max}$, we deduce that the optimal solution of $\underset{M \leq L \leq L^{\max}}{Max} g_1(L)$ is $L_1^* = max\left(\sqrt{(C_2 b_1)/b_2}, M\right)$. On the other hand, as $g_2(L)$ has at maximum two positive extrema R_1 and R_2, and that $g_2(L^{\min}) = g_2(M) = 0$, we deduce that the optimal solution of $\underset{L^{\min} \leq L \leq M}{Max} g_2(L)$ is $L_2^* = \underset{x \in \{R_1, R_2, L^{\min}\} \cap [L^{\min}, M]}{\arg\max} g_2(x)$. Consequently, $L^* = L_1^*$ if $g_1(L_1^*) \geq g_2(L_2^*)$ and $L^* = L_2^*$ otherwises.

■

3.3.2 Experiments and insights with variable price

We recall that we use a base example with parameters: $a = 50$, $b_1 = 4$, $b_2 = 6$, $\mu = 10$, $s = 0.95$, $C_1 = 2$, and $C_2 = 3$. The objective of this section is to illustrate the model behavior when both price and lead time are variables. In particular, we will focus on the comparison between the case of fixed price and this case of variable price.

Effect of demand sensitivity to price (b_1) with variable price

We vary the demand sensitivity to price (b_1) and report in Table 3.3 the optimal lead time, price, demand, total cost, profit, and realized service level. The main observations are summarized hereafter:

Table 3.3: Effect of demand sensitivity to price (b_1)

b_1	Lead time	Price	Demand	Total cost	Profit ($\Pi 2$)	$\Pi 2'$	Gains	Serv. lev. realized
2	1.45	16.70	7.93	32.31	100.08	98.33	1.78%	95%
3	1.52	10.96	8.03	31.92	56.02	53.25	5.19%	95%
4	1.59	8.08	8.12	31.54	34.08	30.19	12.88%	95%
5	1.67	6.36	8.20	31.17	20.99	15.86	32.37%	95%
6	1.74	5.21	8.28	30.81	12.33	5.83	111.43%	95%
7	1.87	4.57	6.78	24.41	6.56	-1.75	-	99.76%
8	2.00	4.13	5.00	17.50	3.13	-7.70	-	99.99%
9	2.12	3.78	3.27	11.17	1.19	Negative L	-	99.99%
10	2.24	3.50	1.58	5.29	0.25	Negative L	-	99.99%

- An increase in demand sensitivity to price leads, as expected, to a decrease in offered price. The model also reacts by quoting a longer lead time in order to reduce the cost and offset the effect of offering a small price. For $b_1 \geq 7$, the quoted lead time is so long that the observed service level becomes much higher than the minimum required level. Thus, if customers are very sensitive to price then they will be offered longer lead time but with better delivery reliability.

- We also observe that demand is concave in b_1. At the beginning, the increase in b_1 leads to significant decrease in price implying an increase in demand in spite of longer quoted lead time. Then, the price decreases much slower and the lead time continues to be longer, leading to a decreasing demand.

- In order to show the impact of taking into account the sensitivity of cost to lead time, we re-solve the model while considering a constant unit production cost. Indeed, we calculate the average unit cost ($C_1 + \frac{C_2}{L}$) over all instances of Table 3.3 and take the obtained value (namely, 3.70) as a constant cost in these new

experiments (i.e., $C_1 = 3.70$, and $C_2 = 0$). Note that the cases of $b_1 = 9$ and 10 are not considered here as they lead to non-feasible problems. We firstly compare the results of our base model to the results of the model with constant cost in terms of quoted lead time. The results, reported in Fig. 3.3, show that an increase in the value of b_1 leads to an increase in quoted lead time in our model while it has the opposite effect when the cost is assumed to be constant. Indeed, with an increase in b_1 in case of constant cost, the model doesn't have any interest in quoting longer lead time as this will not imply a lower unit production cost. Therefore, the fixed cost model involves a different trade-off by shortening the lead time, as much as allowed by the service constraint, in order to obtain more demand. This will also offset: 1. the effect of the price increase (triggered by the increase in b_1) on losing the amount of demand; and 2. the effect offering smaller price on decreasing revenue.

- To evaluate the impact of not quoting the right price and lead time when the cost is assumed constant, we take the optimal price and lead time in this case and inject them in the objective function of our base model (with variable cost). We denote the obtained profit by $\Pi 2'$ and let $\Pi 2$ denotes the optimal profit of our base model (given in Table 3.3). Then, we calculate the percentage of gain $= \frac{100\times(\Pi 2-\Pi 2')}{\Pi 2'}$ for the different values of b_1. In some cases ($b_1 = 7$ and 8), we obtain for $\Pi 2'$ a negative profit when we consider the optimal solution given by the model with constant cost. For the other cases which give positive profits ($\Pi 2'$), our model leads to an average gain 32.73%.

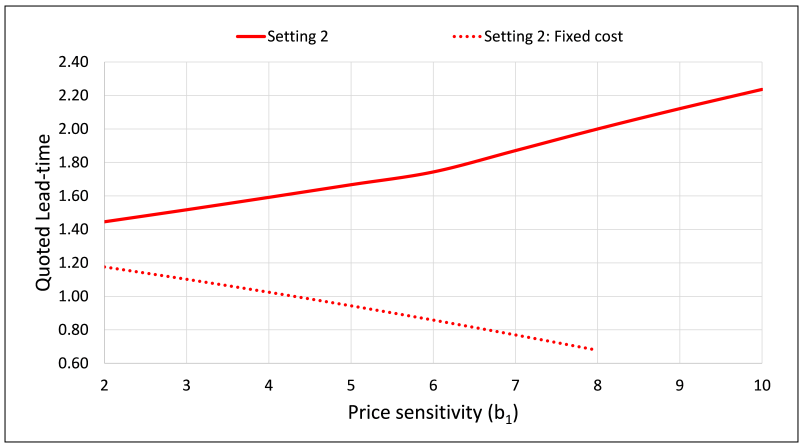


Figure 3.3: Effect of b_1 with variable and constant cost

Effect of demand sensitivity to lead time (b_2) with variable price

We have studied the effect of b_2 in case of fixed price. Now, we conduct the experiments in case of variable price and report the results in Table 3.4.

Table 3.4: Effect of demand sensitivity to lead time (b_2) with variable price

b_2	Lead time	Price	Demand	Total cost	Profit ($\Pi 2$)	$\Pi 2'$	Gains	Serv. lev. realized
2	2.89	8.81	8.96	27.23	51.77	46.28	11.86%	95%
4	1.99	8.38	8.50	29.78	41.44	35.76	15.89%	95%
6	1.59	8.08	8.12	31.54	34.08	28.70	18.74%	95%
8	1.35	7.85	7.78	32.85	28.26	23.36	21.00%	95%
10	1.18	7.67	7.47	33.86	23.45	19.08	22.87%	95%
12	1.06	7.52	7.18	34.65	19.34	15.54	24.47%	95%
14	0.96	7.41	6.89	35.24	15.79	12.54	25.88%	95%
16	0.88	7.31	6.61	35.66	12.68	9.97	27.17%	95%
18	0.82	7.25	6.30	35.77	9.93	7.73	28.41%	95.11%
20	0.77	7.25	5.51	32.35	7.58	5.78	31.23%	96.92%

- It is firstly important to note that an increase in b_2 leads not only to a decrease in lead time but also to a decrease in price. We expected that the model will react by quoting a shorter lead time (in order to guarantee a profitable amount of demand) but the price decrease was not expected since shortening the lead time implies a higher cost, so a price increase was more likely. Here, the price decreases to offset the demand loss due to the increase in $b_2 L$. As customers become more sensitive to lead time, the model prefers not to quote a very short lead time (with comparison to the case of fixed price), but to react by decreasing lead time and price simultaneously in order to find the best trade-off between cost and demand. If we compare the demand in table 3.1 and table 3.4, we see that the demand is always greater in case of variable price although a longer lead time is quoted in this case.

- In addition, for high values of b_2 (two last rows of Table 3.4), the observed service level is higher than the minimum required level. In this case, the firm can offer shorter lead time but does not do it in order to limit the increase in unit operating cost. Thus, when demand is very sensitive to lead time, the customers can benefit of smaller price, shorter lead time, and also more reliable deliveries.

- In Fig 3.4, we report the variation of lead time for increasing values of b_2 in three situations: variable price with variable unit cost ($C_1 + \frac{C_2}{L}$) (setting 2), variable price with fixed unit cost (taken as the average value of unit costs obtained in case of variable cost) (setting 2 fixed cost), and fixed price with variable cost (setting 1, discussed previously in section 3.2). We can see that the longest lead time is quoted when the price is variable and the cost depends on lead time. The shortest lead time is quoted when the price is variable and the cost depends on lead time.

- In order to evaluate the impact of not quoting the right price and lead time when the cost is assumed constant, we take the optimal price and lead time in this case and inject them in the objective function of our base model (with variable cost). We denote the obtained profit by $\Pi 2'$ and let $\Pi 2$ denotes the optimal profit of our base model (given in Table 3.4). Then, we calculate the percentage of gain $= \frac{100\times(\Pi 2-\Pi 2')}{\Pi 2'}$ for the different values of b_2. We found that our model leads to an average gain 22.75%. This shows once again the interest of our model with comparison to existing models where the cost is assumed to be constant.

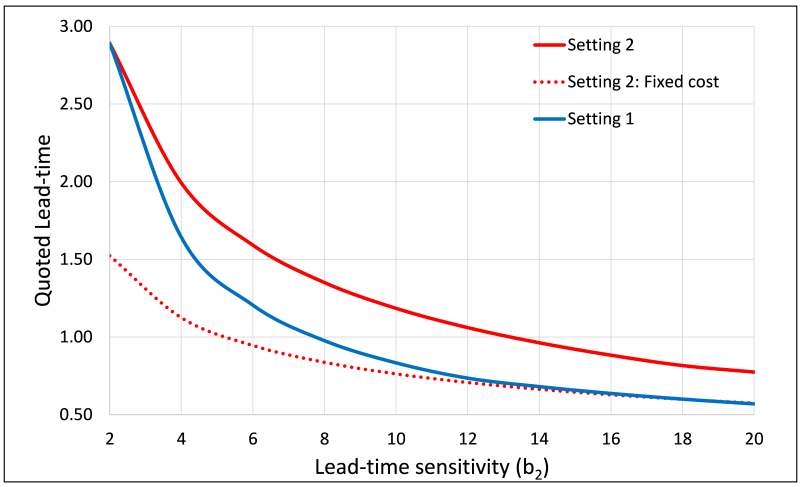


Figure 3.4: Variation of lead time for increasing values of b_2

Effect of cost sensitivity to lead time (C_2) with variable price

We vary the value of C_2 from 2 to 10. The results are reported in Table 3.5. Observing Table 3.5, we deduce the following:

- Unlike the case of fixed price, for which a first increase in C_2 (from 2 to 4, see Table 3.2) does not impact the quoted lead time, we see that the lead time is sensitive to C_2 in case of variable price even for small values of C_2.

- An increase in C_2 implies a longer lead time which favors a decrease in demand. In order to offset the effect of lead time on demand, the firm reacts by decreasing the price as it favors a greater amount of demand. Thus, surprisingly, when the production cost is more sensitive to shorter lead times, the firm reacts by offering a smaller price.

- In addition, an increase in C_2 leads to higher service level. In Table 3.5, we can see that the firm operates in almost guaranteed service level for $C_2 \geq 8$. In fact, due to cost increase, the lead time quoted in this case is longer than necessary, implying a higher service level.

Table 3.5: Effect of cost sensitivity to lead time (C_2) with variable price

C_2	lead time	Price	Demand	Total cost	Profit (Π2)	Serv. lev. realized
2	1.44	8.35	7.93	26.83	39.37	95%
3	1.59	8.08	8.12	31.54	34.08	95%
4	1.74	7.83	8.28	35.61	29.15	95%
5	1.88	7.58	8.41	39.18	24.54	95%
6	2.01	7.35	8.51	42.38	20.19	95%
7	2.16	7.25	8.04	42.12	16.15	98.55%
8	2.31	7.25	7.14	39.03	12.76	99.86%
9	2.45	7.25	6.30	35.77	9.93	99.98%
10	2.58	7.25	5.51	32.35	7.58	99.99%

- As we observe in the figure 3.5, the firm quotes a longer lead time when the price is variable. Indeed, in case of fixed price (see section 3.2.2), the firm is obliged to keep a relatively short lead time even with an increase in C_2 in order to guarantee a sufficient amount of demand. However, when the price is variable, the firm can quote a longer lead time (as this implies a smaller cost) and offset its impact on demand by decreasing the price as we discussed in the previous point.

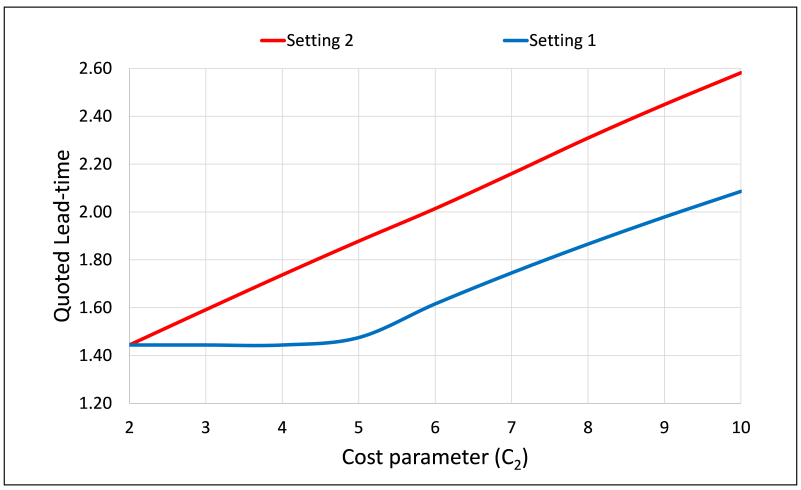


Figure 3.5: Effect of C_2 with variable and constant price

3.4 General model (Setting 3): price is a decision variable; congestion & lateness costs are considered

In this section, we consider the general model (setting 3) that has been explained in section 3.1. We consider three cost components: unit production cost ($m = C_1 + \frac{C_2}{L}$), unit holding cost (F), and lateness penalty cost (c_r).

3.4.1 Optimal policy for general model

As stated in lemma 3.1, the demand constraint (3.2) is tight at optimality. Thus, we have $P = (a - b_2 L - \lambda)/b_1$. The service level constraint (3.3) can also be rewritten as: $\lambda \leq \mu - \frac{\ln(1/(1-s))}{L}$. Thus, the problem can be rewritten as:

$$\underset{0 \leq L \leq \frac{a-\lambda}{b_2}}{\text{Maximize}} \; f(L,\lambda) = \lambda \left(\frac{a - b_2 L - \lambda}{b_1} - C_1 - \frac{C_2}{L} \right) - \frac{F\lambda}{\mu - \lambda} - \frac{c_r \lambda}{\mu - \lambda} e^{-(\mu-\lambda)L} \tag{3.13}$$

$$\text{Subject to } \lambda \leq \mu - \frac{\ln(1/(1-s))}{L} \tag{3.14}$$

Lemma 3.7. *If service level constraint (3.14) is non-binding then optimal demand is* $\lambda^* = \frac{\ln(b_2/(b_1 c_r) - C_2/(L^2 c_r))}{L} + \mu$. *If service level (3.14) is binding, optimal demand is equal to* $\lambda^* = \mu - \frac{\ln(1/(1-s))}{L}$.

Proof. In non-binding situations, we have service level constraint as $\lambda < \mu - \frac{\ln(1/(1-s))}{L}$. This service level constraint (3.14) can be ignored. The optimal lead time in non-binding situation should respect: $\frac{\partial}{\partial L} f(L,\lambda) = \lambda \left(\frac{C_2}{L^2} - \frac{b_2}{b_1} \right) + c_r \lambda e^{-(\mu-\lambda)L} = 0$, which leads to the following demand: $\lambda^* = \frac{\ln(b_2/(b_1 c_r) - C_2/(L^2 c_r))}{L} + \mu$. In binding situation, from (3.14)we have $\lambda^* = \mu - \frac{\ln(1/(1-s))}{L}$. ■

Then from $\lambda = a - b_1 P - b_2 L$, we derive the maximum value for L is $\frac{a}{b_2}$ ($\lambda = P = 0$). Thus, we have the range of L as $0 \leq L \leq a/b_2$. And adding lemma 3.7, we get a single variable optimization problem which is:

$$\underset{0 \leq L \leq \frac{a}{b_2}}{\text{Maximize}} \; f(L) = \lambda \left(\frac{a - b_2 L - \lambda}{b_1} - C_1 - \frac{C_2}{L} \right) - \frac{F\lambda}{\mu - \lambda} - \frac{c_r \lambda}{\mu - \lambda} e^{-(\mu-\lambda)L} \tag{3.15}$$

$$\text{where } \lambda = \frac{\ln(b_2/(b_1 c_r) - C_2/(L^2 c_r))}{L} + \mu \quad \text{(non-binding situation)} \quad \text{or}$$

$$\lambda = \mu - \frac{\ln(1/(1-s))}{L} \quad \text{(binding situation)}$$

To find the optimal profit, we have to compare the profit of both binding and non-binding case. Then we should take the maximum one. But, the problem above

is very difficult to solve analytically. Thus, we solve it numerically with classical meta-heuristic. Given that our decision variables are real numbers and we want to do some explorations, hence we choose to use a population based algorithm such as particle swarm optimization (PSO).

3.4.2 Experiments and insights for general model

In these experiments, we use our base case with parameters: $a = 50$, $b_1 = 4$, $b_2 = 6$, $\mu = 10$, $s = 0.95$, $C_1 = 2$, $C_2 = 3$, $F = 2$, and $c_r = 10$. We do experiments by varying one parameter and keep other parameters constant. We do the experiment in varying b_2, b_1, C_2, F and c_r. As an example, in the first experiment we vary the price sensitivity parameters (b_1) and keep the others parameters constant, then, in second experiment, we vary the lead time sensitivity (b_2) and keep other parameters constant, and so on.

Effect of demand sensitivity to price (b_1) for general model

To see the effect of demand sensitivity to price, we vary (b_1) and report the result in table 3.6. We discovered several findings:

Table 3.6: Effect of demand sensitivity to price (b_1) for general model

b_1	lead time	Price	Demand	Total cost	Profit (Π3)	Π3′	Gains	Serv. lev. realized
2	1.27	17.37	7.64	41.43	91.29	89.62	2%	95%
3	1.34	11.51	7.44	38.29	47.32	44.53	6%	96.76%
4	1.47	8.54	6.99	33.15	26.58	21.61	23%	98.81%
5	1.60	6.80	6.39	28.37	15.11	7.63	98%	99.67%
6	1.74	5.68	5.53	23.11	8.29	-1.72	-	99.96%
7	1.87	4.91	4.42	17.50	4.18	-8.12	-	99.99%
8	2.00	4.34	3.25	12.32	1.78	-12.19	-	99.99%
9	2.12	3.94	1.85	6.76	0.51	-13.93	-	99.99%
10	2.24	3.60	0.58	2.08	0.03	-12.89	-	99.99%

- We are always in non-binding situation (except for $b_1 = 2$). This result is not surprising as the production cost and lateness penalty cost are lower with long lead time. The optimal lead time is longer than the lead time needed to satisfy the service level constraint.
- As b_1 increases, intuitively price (P) will decrease. The production cost (m) decreases due to a decline in price to keep the profit ($P - m$) positive. This will cause quoted lead time L to increase.
- In order to show the impact of taking into account the sensitivity of cost to lead time, we re-solve the model while considering a constant unit production cost.

We calculate the average unit cost ($C_1 + \frac{C_2}{L}$) over all instances of Table 3.6 and take a value of 3.79 as a constant cost. These new experiments use $C_1 = 3.79$ and $C_2 = 0$. Note that the cases of $b_1 > 5$ are not considered here as they lead to non-feasible problems. We compare the results of our general model to the results of the model with constant cost in terms of quoted lead time. The results, reported in Fig. 3.6, show that this general model has the same behavior as setting 2 (variable price without holding cost and lateness penalty cost in figure 3.3). An increase in the value of b_1 leads to an increase observed in figure 3.6 for quoted lead time in our model while it has the opposite effect when the cost is assumed to be constant.

- In order to evaluate the impact of not quoting the right price and lead time when the unit production cost is assumed constant, we take the optimal price and lead time in this case and inject them in the objective function of our base model (the general model). We denote the obtained profit by $\Pi 3'$ and let $\Pi 3$ denote the optimal profit of our base model (given in Table 3.6). Then, we calculate the percentage of gain $= \frac{100 \times (\Pi 3 - \Pi 3')}{\Pi 3'}$ for the different values of b_1. In some cases ($b_1 > 5$), we obtain a negative profit when we consider the optimal solution given by the model with constant cost. For the other cases which give positive profits, our model leads to an average gain 32.32%.

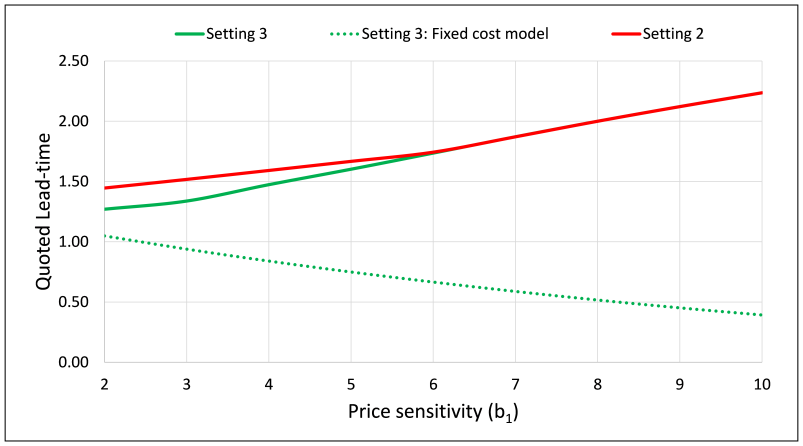


Figure 3.6: Effect of b_1 with variable and constant cost for general model

Effect of demand sensitivity to lead time (b_2) for general model

In table 3.7, we vary lead time sensitivity (b_2). The summary of our observations is presented below:

- In our experiment of b_2 from 2 to 20, we are always in non-binding situation. The model prefers to quote longer lead time than the necessary lead time to satisfy the service level constraint, in order to reduce the production cost and lateness cost. These costs favor long lead times.

Table 3.7: Effect of demand sensitivity to lead time (b_2) for general model

b_2	lead time	Price	Demand	Total cost	Profit(Π3)	Π3′	Gains	Serv. lev. realized
2	2.53	9.30	7.73	31.51	40.36	30.81	30.99%	99.68%
4	1.81	8.86	7.34	32.63	32.41	24.98	29.74%	99.18%
6	1.47	8.54	7.00	33.16	26.58	20.39	30.33%	98.81%
8	1.27	8.29	6.65	33.28	21.92	16.65	31.66%	98.58%
10	1.13	8.10	6.30	33.01	18.04	13.48	33.81%	98.47%
12	1.03	7.94	5.93	32.34	14.74	10.78	36.77%	98.47%
14	0.95	7.80	5.55	31.35	11.92	8.44	41.24%	98.52%
16	0.88	7.70	5.11	29.84	9.49	6.40	48.30%	98.66%
18	0.83	7.61	4.66	28.05	7.40	4.61	60.76%	98.80%
20	0.78	7.56	4.13	25.56	5.63	3.02	86.26%	98.99%

- As noted in the experiment of b_2 in setting 2, an increase in b_2 leads not only to a decrease in lead time but also to a decrease in price. In setting 3, the model behaves the same ways as setting 2. As costumer become more sensitive to lead time, the model chooses to decrease lead time and price simultaneously in order to find the trade-off between cost and demand.

- In Fig 3.7, we report the variation of lead time for increasing values of b_2 in the following situations: variable unit cost ($C_1 + \frac{C_2}{L}$), and fixed unit cost (taken as the average value of unit costs obtained in case of variable cost which equal to 4.61). We see the quoted lead time in lead time sensitive model is higher than in the fixed cost model. This happens as the lead time sensitive model tries to reduce the cost by having longer quoted lead time.

- In order to evaluate the impact of not quoting the right price and lead time when the cost is assumed constant, we take the optimal price and lead time in this case and inject them in the objective function of our general model. We denote the obtained profit by Π3′ and let Π3 denotes the optimal profit of our base model (given in Table 3.7). Then, we calculate the percentage of gain $= \frac{100\times(\Pi 3-\Pi 3')}{\Pi 3'}$ for the different values of b_2. We found that our model leads to an average gain 42.99%. This value is twice bigger compared to the value obtained by comparing setting 1 and setting 2 (without holding and lateness cost, see table 3.3).

Effect of cost sensitivity to lead time (C_2) for general model

In table 3.8, we vary C_2. Our main findings is:

- We are always in non-binding situations unlike setting 2 (table 3.5).

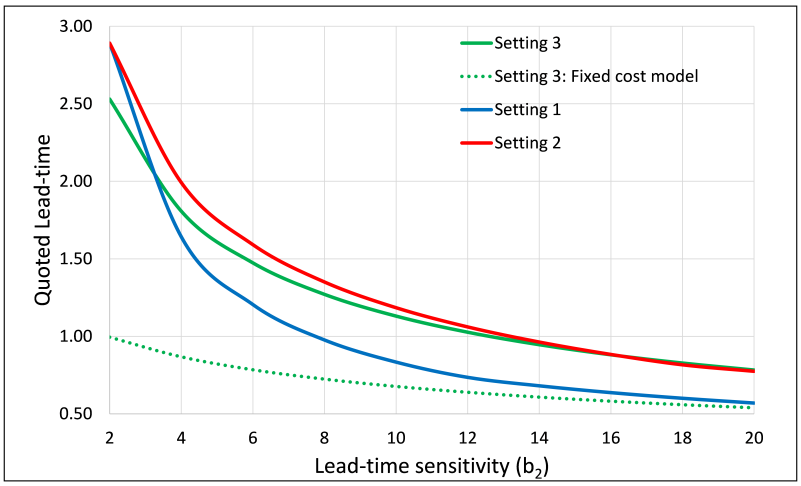


Figure 3.7: Variation of lead time for increasing values of b_2 in general model

Table 3.8: Effect of cost sensitivity to lead time (C_2) for general model

C_2	lead time	Price	Demand	Total cost	Profit(Π3)	Serv. lev. realized
2	1.27	8.82	7.10	30.94	31.73	97.46%
3	1.47	8.54	6.99	33.15	26.58	98.81%
4	1.66	8.31	6.81	34.42	22.17	99.50%
5	1.84	8.11	6.52	34.60	18.34	99.83%
6	2.00	7.94	6.20	34.24	15.02	99.95%
7	2.16	7.80	5.81	33.24	12.14	99.99%
8	2.31	7.71	5.31	31.29	9.65	99.99%
9	2.45	7.37	5.81	35.73	7.10	99.99%
10	2.58	7.56	4.28	26.66	5.71	99.99%

- In this setting, the lead time L increases to reduce $m = C_1 + (C_2/L)$ as a compensation for an increase in C_2. Price decreases because the lead time increases (to capture the maximum demand). And the demand in non-binding situation is sensitive to both lead time (L) and C_2 (see lemma 3.7).

- Figure 3.8 shows the comparison in term of quoted lead time in function of C_2 for our three settings: (1) case of fixed price, (2) variable price, and (3) our general model. The general model behaves in the same ways as the model with variable price. However, at first the general model has a lower lead time compared to the variable price model, then the lead times obtained in setting 2 & 3 converge.

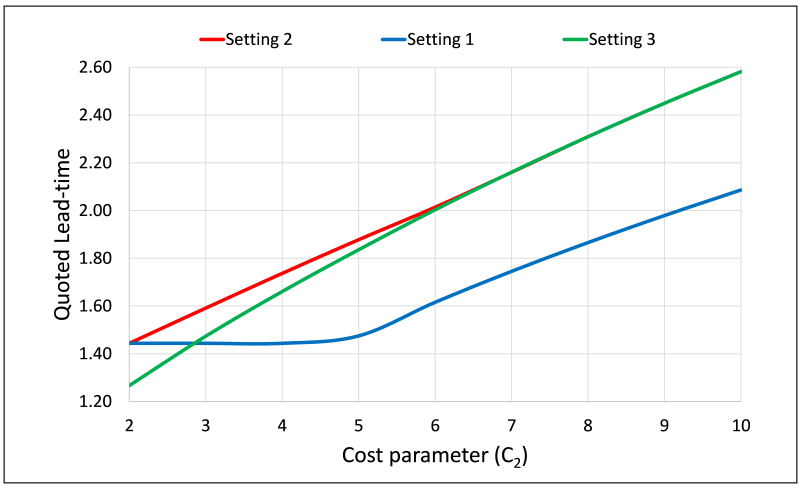


Figure 3.8: Effect of C_2 in general model

Effect of holding cost (F) for general model

In table 3.9, we vary holding cost (F). Here are our observations:

Table 3.9: Effect of holding cost (F) for general model

F	lead time	Price	Demand	Total cost	Profit ($\Pi 3$)	$\Pi 3'$	Gains	Serv. lev. realized
0	1.58	8.17	7.80	31.43	32.30	27.69	16.64%	96.96%
1	1.51	8.38	7.38	32.75	29.14	24.23	20.25%	98.10%
2	1.47	8.54	6.99	33.16	26.58	21.23	25.19%	98.81%
3	1.45	8.66	6.64	33.10	24.44	18.60	31.41%	99.23%
4	1.44	8.76	6.31	32.72	22.60	16.26	39.04%	99.50%
5	1.43	8.85	6.03	32.33	20.99	14.15	48.37%	99.66%
6	1.43	8.92	5.75	31.76	19.56	12.24	59.83%	99.77%
7	1.42	8.99	5.50	31.20	18.28	10.52	73.71%	99.83%
8	1.42	9.05	5.28	30.68	17.11	8.95	91.18%	99.88%
9	1.42	9.10	5.09	30.30	16.03	7.50	113.72%	99.91%
10	1.42	9.16	4.87	29.58	15.04	6.18	143.52%	99.93%

- Again results in table 3.9 show that we are always in binding situation.
- As F increases, the firm chooses to quote short lead times to reduce holding cost. Due to decrease in L, cost m increases thus price P increases. In Fig 3.9, we report the variation of lead time for increasing values of F in the situations where unit cost depends on quoted lead time ($C_1 + \frac{C_2}{L}$), and unit cost is fixed (taken as the average value of unit costs obtained in case of variable cost which is 4.07).

- To evaluate the impact of not quoting the right price and lead time when the cost is assumed constant, we take the optimal price and lead time in this case and inject them in the objective function of our general model. $\Pi 3'$ denote the obtained profit of this new case and $\Pi 3$ denotes the optimal profit of our general model (given in Table 3.9). The percentage of gain is $\frac{100\times(\Pi 3-\Pi 3')}{\Pi 3'}$ for the different values of F. We found that our model leads to an average gain 60.26%.

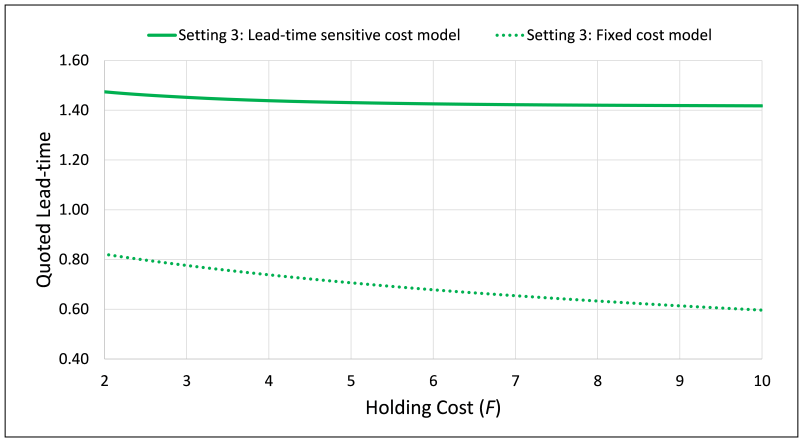


Figure 3.9: Effect of F in general model

Effect of lateness cost (c_r) for general model

In table 3.10, we vary lateness cost (c_r) for general model. Here are our findings:

Table 3.10: Effect of lateness cost (c_r) for general model

c_r	lead time	Price	Demand	Total cost	Profit ($\Pi 3$)	$\Pi 3'$	Gains	Serv. lev. realized
0	1.41	8.57	7.24	35.12	26.95	22.86	17.93%	97.97%
10	1.47	8.54	7.00	33.16	26.58	21.40	24.19%	98.81%
20	1.51	8.52	6.87	32.20	26.35	20.04	31.46%	99.11%
30	1.53	8.51	6.78	31.55	26.17	18.75	39.59%	99.27%
40	1.55	8.50	6.72	31.07	26.03	19.80	31.48%	99.38%
50	1.56	8.49	6.67	30.68	25.92	20.42	26.89%	99.45%
60	1.58	8.48	6.62	30.36	25.81	20.84	23.89%	99.51%
70	1.59	8.47	6.59	30.08	25.72	21.14	21.71%	99.56%
80	1.60	8.46	6.55	29.84	25.64	21.36	20.07%	99.59%
90	1.61	8.46	6.53	29.63	25.57	21.52	18.79%	99.62%
100	1.62	8.45	6.50	29.43	25.50	21.66	17.75%	99.65%

- Even when $c_r = 0$, the model chooses to be non-binding. In our analysis, the model decides to reduce the production cost by increasing L in order to get more profit $(P - m)$, and that even in the situation where there is no penalty $(c_r = 0)$. When the lateness cost is not zero, it is obvious that the model will be non-binding.

- Obviously, as lateness penalty cost (c_r) increases, firms will quote longer lead time to avoid paying high lateness penalty cost. As L increases, m decreases. Demand decreases because $\mu - \frac{\ln(1/(1-s))}{L}$ decreases (see constraint 3.14). Price P decreases to attract more demand as L increases (recall that $\lambda = a - b_1 P - b_2 L$). In Fig 3.10, we report the variation of lead time for increasing values of c_r. We consider following situations: variable unit cost ($C_1 + \frac{C_2}{L}$), and fixed unit cost (taken as the average value of unit costs obtained in case of variable cost which is 3.94). In fixed cost model (see figure 3.10), we can see that in the beginning the quoted lead time L is decreasing (the service level constraint is binding) then after certain point L increases (the service level constraint is non-binding).

- In order to evaluate the impact of not quoting the right price and lead time when the cost is assumed constant, we take the optimal price and lead time in this case and inject them in the objective function of our general model. We denote the obtained profit by $\Pi 3'$ and let $\Pi 3$ denotes the optimal profit of our base model (given in Table 3.10). Then, we calculate the percentage of gain $= \frac{100 \times (\Pi 3 - \Pi 3')}{\Pi 3'}$ for the different values of c_r. We found that our model leads to an average gain 24.89%.

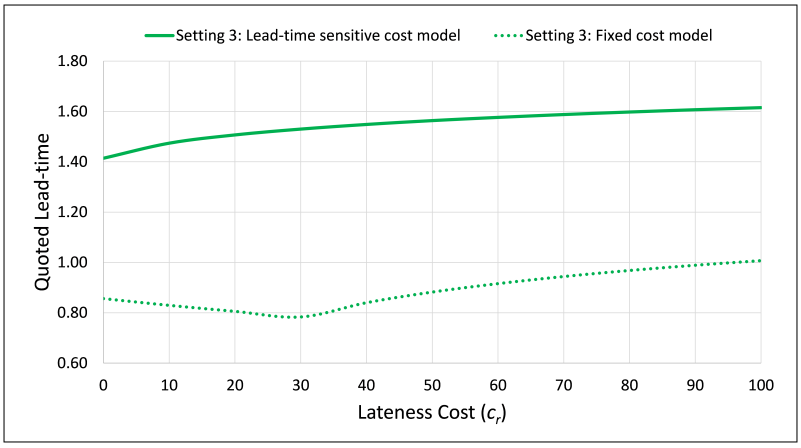


Figure 3.10: Effect of c_r in general model

3.5 Conclusion

We solved the problem of lead time quotation in an M/M/1 make-to-order queue

while assuming the production cost to be a decreasing function in lead time. We considered three settings: (1) lead time is variable but price is fixed, (2) price and lead time are both decision variables, and (3) price and lead time are decision variables where the lateness and holding costs are considered. We conducted experiments and derived interesting insights.

In case of variable lead time and fixed cost (setting 1), some of our insights indicated that the higher the demand sensitivity to lead time the higher the service level. This results is not intuitive since an increase in lead time sensitivity leads to shorter lead time that is supposed to be more difficult to guarantee. In addition, this behavior cannot be captured by the existing models where the unit operating cost is assumed constant since the service constraint is always binding in this case. We also found that when customers are very sensitive to lead time, our model quotes longer lead time than the benchmark model with constant cost. Furthermore, we observed that when the operating cost is very sensitive to lead time, it can be optimal for the system to work like an almost guaranteed service model (i.e., 99.99% of demands are satisfied on time) despite the uncertainties in both demand and processing time.

When both lead time and price are endogenous variables (Setting 2), our results showed that an increase in the demand sensitivity to price leads to an increase in quoted lead time in our model while it has the opposite effect for models with constant cost. We also found that when demand is very sensitive to lead time, the customers can benefit of smaller price, shorter lead time, and also more reliable deliveries. Some of our findings showed that, surprisingly, when the operating cost is more sensitive to lead time, the firm reacts by offering a smaller price. In addition, we saw that the firm quotes a longer lead time when the price is also a variable (in addition to lead time) with comparison to the case of fixed price.

When we consider the lateness penalty and holding costs (Setting 3), the model quotes a longer lead time higher than the service constraint. The model prefers to be in non-binding situation. Indeed, quoting a longer lead time is favorable as it will reduce the incurred cost (i.e. production and lateness cost).

From our numerical experiments, we quantified the gain brought by using the solution of our model versus the solution of the benchmark model where the cost is constant. In the case where the price is fixed (setting 1), we found that our model leads to small gains. This prove that there is an impact of not quoting the right lead time when the cost is assumed to be constant. This impact becomes more significant (as the gains become bigger) when we take into account the price as a decision variable (setting 2). And when the lateness and holding costs are included, this gains increase compared to the settings 1 and 2.

Chapter 4

Rejection policy: a lead time quotation and pricing in an M/M/1/K make-to-order queue

All of papers in the literature, presented in chapter 2, use the M/M/1 system. Although M/M/1 is the simplest queuing model, it has a drawback. In M/M/1, all the customers are accepted, which might lead to long sojourn times (lead time) in the system when we accept customer while there are already a lot of customers waiting. Realistically, firms can choose to reject customers when they already have too many customers. Thus, we propose a customer's rejection policy using an M/M/1/K model. Demand is rejected if there are already K customers in the system (K represents the system capacity, that is the maximum number of customers in the system including the one under service). Our idea, that we want to check in this chapter, is that rejecting some customers might help to quote shorter lead time for the accepted ones, which might finally lead to a higher profitability.

In this chapter, we explicitly formulate the problem of lead time quotation and pricing for a profit-maximizing firm, modeled as an M/M/1/K system, facing a linear price- and lead time-dependent demand with the consideration of inventory holding and lateness penalty costs. In order to formulate the lateness cost, we overcome a theoretical obstacle by calculating the expected lateness given that a job is late in an M/M/1/K queue when a delivery lead time, L, is quoted to the customers (i.e., the expected sojourn time in the system after L). Then, we bring a second analytical contribution by determining the optimal firm's policy (optimal price and quoted lead time) in case of M/M/1/1. We solve the problem numerically for $K > 1$. We conduct experiments and use the expression of the optimal solution obtained for M/M/1/1 to derive some insights. We also compare the results of our model, firstly for $K = 1$ and then for different values of K, to the results obtained in the literature with an M/M/1 queue. We show in different situations that a customer rejection policy, represented by the M/M/1/K, can be better than the all-customers' acceptance policy, represented by the M/M/1, even when the inventory holding and lateness penalty costs are ignored.

4.1 General model: M/M/1/K

We consider a firm operating under a make-to-order M/M/1/K setting. Customers are served in first-come, first-served basis. The demand arrival is assumed to

be a Poisson process. The processing time of customers in the system is assumed to be exponentially distributed. Similarly to many works in the literature such as Liu et al. (2007); Palaka et al. (1998); and Pekgün et al. (2008), the demand is assumed to be a linear decreasing function in price and quoted lead time.

The decision variables of our model are the price, the quoted lead time and the demand as in previous chapter. We let P refers to the price of the good/service set by the firm, and L is the quoted lead time. Thus, the expected demand for the good/service with price P and quoted lead time L is given by $a - b_1P - b_1L$, where a is the market potential, b_1 is the price sensitivity of demand, and b_2 is the lead time sensitivity of demand. Since the demand is downward sloping in both price and quoted lead time, b_1 and b_2 are restricted to be non-negative.

The considered M/M/1/K queueing system has a mean service rate denoted by μ, a mean arrival rate (demand), λ, and a throughput rate (effective demand), $\bar{\lambda}$. Unlike the assumptions of M/M/1, for which all customers are accepted, the customers are rejected in the M/M/1/K model when there are already K clients in the system. The capacity (system size) K is assumed to be constant. The probability P_k of having k customers in the system ($k = 1, 2, \ldots, K$) is given by equation (4.1) as in Gross et al. (2008). Therefore, P_K represents the probability of rejecting a customer and $(1-P_K)$ is the probability that a customer is accepted. Consequently, the effective demand ($\bar{\lambda}$) is equal to the mean arrival rate (λ) multiplied by the probability of accepting a customer $(1 - P_K)$. Equation (4.2) gives the expected number of customers in the system, denoted by N_s (see Gross et al., 2008). The sojourn time W is total time in the system with mean $N_s/\bar{\lambda}$. The probability that the firm is able to meet the quoted lead time (i.e., $\Pr(W \leq L)$) and the probability that a job is late (i.e., $\Pr(W > L)$) are formulated in eq. (4.3) as given in Sztrik (2016).

$$P_k = \frac{1-\rho}{1-\rho^{K+1}}\rho^k \text{ if } \rho \neq 1 \text{ and } P_k = \frac{1}{K+1} \text{ if } \rho = 1 \text{ with } \rho = \frac{\lambda}{\mu} \tag{4.1}$$

$$N_s = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} \tag{4.2}$$

$$\begin{aligned} \Pr(W \leq L) &= 1 - \sum_{k=0}^{K-1} \frac{P_k}{1-P_K} \left(\sum_{i=0}^{k} \frac{(\mu L)^i}{i!} e^{-\mu L} \right) \text{ and} \\ \Pr(W > L) &= \sum_{k=0}^{K-1} \frac{P_k}{1-P_K} \left(\sum_{i=0}^{k} \frac{(\mu L)^i}{i!} e^{-\mu L} \right) \end{aligned} \tag{4.3}$$

In order to prevent the firms from quoting unrealistically short lead times, we assume that the firm maintains a minimum service level. Thus, the probability of meeting the quoted lead time must be greater than the service level denoted by s (i.e., we impose $\Pr(W \leq L) \geq s$).

The objective of the firm is to maximize the profit. Since we consider the lateness penalty and the inventory holding costs, the firm's profit is calculated as follows:

Profit = Expected revenue (net of direct cost) − Total in-process inventory holding cost − Total Lateness penalty cost. In what follows, we explain how we calculate each part of this profit function.
Expected revenue (net of direct cost) is given by $\lambda(1-P_K)(P-m) = \bar{\lambda}(P-m)$, where m denotes the unit direct variable cost.
Total in-process inventory holding cost, as in chapter 3, is given by $N_s \times F$, where F denotes the unit holding cost. Recall that N_s is the mean number of customers in the system (given in eq. (4.2)).
Total lateness penalty cost is expressed, as in chapter 3, as (penalty per job per unit lateness) × (number of overdue clients) × (expected lateness given that a job is late). The penalty cost per job per unit lateness (denoted by c_r) reflects the direct compensation paid to a customer for not meeting the quoted lead time. The number of overdue clients is equal to (throughput rate) × (probability that a job is late); it is then given by $\bar{\lambda} \times \Pr(W > L)$. We let R_L denote the expected lateness given that a job is late. Therefore, the total lateness penalty cost is given by $(c_r \times \bar{\lambda} \times \Pr(W > L) \times R_L)$.

We need to calculate R_L in order to determine the total lateness penalty cost and, consequently, to formulate the objective function of our model. The calculation of R_L in an M/M/1/K queue is challenging. To the best of our knowledge, this result is not known in the literature. Our work brings a new contribution to the queuing theory literature by explicitly calculating the value of R_L in an M/M/1/K queue.

Theorem 4.1. *Consider an M/M/1/K queueing system with mean service rate, μ, and mean arrival rate λ. We let W denote the sojourn time in the system and $f_W(.)$ its probability density function. Given a quoted lead time, L, the expected lateness given that a job is late in an M/M/1/K queue, denoted by R_L, is given by:*

$$R_L = \int_L^{\infty} (t-L) f_{W|W\geq L}(t)dt = \frac{\sum_{k=0}^{K-1} \frac{P_k}{k!} \left[\frac{(\mu L)^{k+1}}{\mu} + \left(\frac{k+1}{\mu} - L \right) \sum_{i=0}^{k} \frac{k!}{i!} (\mu L)^i \right]}{\sum_{k=0}^{K-1} P_k \sum_{i=0}^{k} \frac{(\mu L)^i}{i!}}$$

where, $P_k = \dfrac{1-\rho}{1-\rho^{K+1}}\rho^k$ *if* $\rho \neq 1$, $P_k = \dfrac{1}{K+1}$ *if* $\rho = 1$, *and* $\rho = \dfrac{\lambda}{\mu}$.

Proof. See Appendix B. ■

In order to help understanding the result of theorem 4.1, we can consider the case $K = 1$. One can check that the result of theorem 4.1, for $K = 1$, becomes $\frac{1}{\mu}$ which corresponds to effectively the residual waiting time since there is only one client.

Next, from theorem 4.1, we deduce by standard calculus that the total lateness

penalty cost is given by:

$$c_r \times \lambda(1-P_K) \times \sum_{k=0}^{K-1} \frac{P_k}{1-P_K} \left(\sum_{i=0}^{k} \frac{(\mu L)^i}{i!} e^{-\mu L} \right) \times \frac{\sum_{k=0}^{K-1} \frac{P_k}{k!} \left[\frac{(\mu L)^{k+1}}{\mu} + \left(\frac{k+1}{\mu} - L \right) \sum_{i=0}^{k} \frac{k!}{i!} (\mu L)^i \right]}{\sum_{k=0}^{K-1} P_k \sum_{i=0}^{k} \frac{(\mu L)^i}{i!}}$$

$$\Leftrightarrow c_r \times \lambda(1-P_K) \times \frac{e^{-\mu L}}{1-P_K} \sum_{k=0}^{K-1} P_k \left(\sum_{i=0}^{k} \frac{(\mu L)^i}{i!} \right) \times \frac{\sum_{k=0}^{K-1} \frac{P_k}{k!} \left[\frac{(\mu L)^{k+1}}{\mu} + \left(\frac{k+1}{\mu} - L \right) \sum_{i=0}^{k} \frac{k!}{i!} (\mu L)^i \right]}{\sum_{k=0}^{K-1} P_k \sum_{i=0}^{k} \frac{(\mu L)^i}{i!}}$$

$$\Leftrightarrow c_r \times \lambda \times e^{-\mu L} \times \sum_{k=0}^{K-1} \frac{P_k}{k!} \left[\frac{(\mu L)^{k+1}}{\mu} + \left(\frac{k+1}{\mu} - L \right) \sum_{i=0}^{k} \frac{k!}{i!} (\mu L)^i \right]$$

Thus, based on the result announced in Theorem 4.1 and on equations (4.1), (4.2), and (4.3), we can now explicitly formulate the problem of lead time quotation and pricing for a profit-maximizing firm, modeled as an M/M/1/K system, facing a linear price- and lead time-dependent demand with the consideration of inventory holding and lateness penalty costs. We denote this model by (M_K).

$$\begin{aligned}
(M_K)\ \underset{P,L,\lambda}{Maximize}\ & \lambda(1-P_K)(P-m) - \left(\left(\frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} \right) \times F \right) \\
& - \Big(c_r \times \lambda \times e^{-\mu L} \\
& \times \sum_{k=0}^{K-1} \frac{P_k}{k!} \left[\frac{(\mu L)^{k+1}}{\mu} + \left(\frac{k+1}{\mu} - L \right) \sum_{i=0}^{k} \frac{k!}{i!} (\mu L)^i \right] \Big) \qquad (4.4) \\
\text{Subject to } & \lambda \leq a - b_1 P - b_2 L \qquad (4.5) \\
& 1 - \sum_{k=0}^{K-1} \frac{P_k}{1-P_K} \left(\sum_{i=0}^{k} \frac{(\mu L)^i}{i!} e^{-\mu L} \right) \geq s \qquad (4.6) \\
& \rho = \frac{\lambda}{\mu} \qquad (4.7) \\
& P_k = \frac{1-\rho}{1-\rho^{K+1}} \rho^k \text{ if } \rho \neq 1 \text{ and } P_k = \frac{1}{K+1} \text{ if } \rho = 1 \qquad (4.8) \\
& \lambda, P, L \geq 0 \qquad (4.9)
\end{aligned}$$

where, as in chapter 3,

Decision Variables

P = price of the good/service set by the firm	λ = mean arrival rate (demand)
L = quoted lead time	

Parameters

a = market potential	s = service level set by the company
b_1 = price sensitivity of demand	F = unit holding cost
b_2 = lead time sensitivity of demand	c_r = penalty cost per job per unit lateness
μ = mean service rate (production capacity)	K = system size (notation added for this chapter)
m = unit direct variable cost	

In model (M_K), constraint (4.5) imposes that the mean demand (λ) cannot be greater than the demand obtained with price (P) and quoted lead time (L). Constraint (4.6) expresses the service level constraint. Equality (4.7) gives the value of ρ. Equality (4.8) calculates the probability of rejecting customers. The non-negativity constraint (4.9) gives the domain of model variables. Clearly, the model obtained is very hard to solve analytically in the general case. In the following section, we show how an analytical solution can be found in case of $K = 1$ with and without lateness penalty and holding costs.

4.2 The M/M/1/1 model: Analytical solution

Solving analytically the general case (M_K) seems to be very difficult. So, in this section, we consider the case of K = 1. We will consider two situations: the case without penalty and holding cost; and the case where these costs are included. For both cases, we will compare the obtained optimal solution with the optimal solution of the M/M/1 approach and derive insights in section 4.3.

4.2.1 The M/M/1/1 model: congestion & lateness costs are ignored

For $K = 1$ without lateness penalty and holding costs ($c_r = 0$ and $F = 0$), it can be shown by standard calculus that model (M_K) is equivalent to model $(M_1^{'})$ given below.

$$(M_1^{'}) \underset{P,L,\lambda}{Maximize} \left(\frac{\lambda\mu}{\mu+\lambda}\right)(P-m) \tag{4.10}$$

$$\text{Subject to } \lambda \leq a - b_1 P - b_2 L \tag{4.11}$$

$$1 - e^{-\mu L} \geq s \tag{4.12}$$

$$\lambda, P, L \geq 0 \tag{4.13}$$

Under its present form, model $(M_1^{'})$ is a three-variables constrained non-linear optimization model, which is still hard to solve analytically. In order to solve model $(M_1^{'})$, we firstly reduce the number of variables by using the following lemma.

Lemma 4.1. *The demand constraint is binding (eq. (4.11)) and we have* $P = \frac{a-b_2L-\lambda}{b_1}$ *at optimality.*

Proof. We let price P^*, quoted lead time L^*, and demand rate λ^* denote the optimal solution and suppose that $\lambda^* < a - b_1P^* - b_2L^*$. Since the objective function is increasing in P, one could increase the price from P^* to P' (while keeping L^* and λ^* constant) until $\lambda^* = a - b_1P' - b_2L^*$. This change will increase the profit, which is impossible since (L^*, P^*, λ^*) was assumed to be the optimal solution. Consequently, the demand constraint is binding and $P = (a - b_2L - \lambda)/b_1$ at optimality. ■

Lemma 4.2. *Service constraint (eq. (4.12)) is binding and we have* $L = \frac{\ln(1/(1-s))}{\mu}$ *at optimality.*

Proof. We let price P^*, quoted lead time L^*, and demand rate λ^* denote the optimal solution and suppose that $1 - e^{-\mu L} > s \Leftrightarrow \mu L^* > \ln(1/(1-s))$. Since constraint (4.11) is binding (according to lemma 4.1) and demand rate is decreasing in quoted lead time, one could increase λ^* to λ' by decreasing L^* to L' (while keeping the price constant) until $\mu L' = \ln(1/(1-s))$. Given that the objective function is increasing in demand rate λ (assuming the profit is positive), then solution P^*, L', and λ' will increase the profit. This is impossible since P^*, L^*, and λ^* is the optimal solution. Hence, service constraint is binding and implies that $L = (\ln(1/(1-s)))/\mu$ at optimality. ■

Now, we will use the results of Lemma 4.1 and 4.2 in order to transform model (M_1) into a single variable model. We substitute L by $(\ln(1/(1-s)))/\mu$ into equation (4.11). Given that constraints (4.11) and (4.12) are tight at optimality, we obtain:

$$P = \frac{a\mu - b_2 \ln(1/(1-s)) - \lambda\mu}{\mu b_1}$$

Price P must be greater than m in order to obtain a positive profit, which implies that λ must satisfy $\lambda \leq a - \frac{b_2 \ln(1/(1-s))}{\mu} - b_1 m$. Thus, substituting L and P by their values, we get the following equivalent formulation of model (M_1) with a single variable (λ):

$$\underset{0\leq\lambda\leq a-\frac{b_2\ln(1/(1-s))}{\mu}-b_1m}{Maximize} \quad \Pi(\lambda) = \frac{\lambda a\mu - \lambda b_2 \ln(1/(1-s)) - \lambda^2\mu - \lambda m\mu b_1}{\mu b_1 + \lambda b_1} \tag{4.14}$$

Clearly, this problem is relevant only when $a - \frac{b_2 \ln(1/(1-s))}{\mu} - b_1 m > 0$. We assume this condition holds.

Proposition 4.1. *Assuming* $a - \frac{b_2 \ln(1/(1-s))}{\mu} - b_1 m > 0$ *(otherwise the problem is not relevant), the optimal solution of the (M/M/1/1) model without penalty and holding costs is:*

- *Optimal lead time:* $L^* = \frac{\ln(1/(1-s))}{\mu}$,

- *Optimal price:* $P^* = \frac{a - b_2 \frac{\ln(1/(1-s))}{\mu} + \mu - \sqrt{\mu^2 + a\mu - b_2 \ln(1/(1-s)) - m\mu b_1}}{b_1}$,
- *Optimal demand:* $\lambda^* = -\mu + \sqrt{\mu^2 + a\mu - b_2 \ln(1/(1-s)) - m\mu b_1}$

Proof. Firstly, we identify the stationary points of function $\Pi(\lambda)$. Let us calculate $\frac{d\Pi(\lambda)}{d\lambda}$:

$$\frac{d\Pi(\lambda)}{d\lambda} = 0 \Leftrightarrow a\mu - b_2 \ln(1/(1-s)) - m\mu b_1 - 2\lambda\mu - \lambda^2 = 0$$

The discriminant of this quadratic equation is:

$$\Delta = 4\mu^2 + 4a\mu - 4b_2 \ln(1/(1-s)) - 4m\mu b_1$$

As it was assumed that $a - \frac{b_2 \ln(1/(1-s))}{\mu} - b_1 m > 0$ (since, otherwise, the problem is not relevant), we deduce that $\Delta \geq 0$. Hence, we obtain two real roots (two stationary points):

$$\lambda_1 = -\mu - \sqrt{\mu^2 + a\mu - b_2 \ln(1/(1-s)) - m\mu b_1}$$

and

$$\lambda_2 = -\mu + \sqrt{\mu^2 + a\mu - b_2 \ln(1/(1-s)) - m\mu b_1}$$

The first root λ_1 is negative and, consequently, non-feasible.
We are going to prove that λ_2 is feasible (i.e., $\lambda_2 \geq 0$ and $\lambda_2 \leq a - \frac{b_2 \ln(1/(1-s))}{\mu} - b_1 m$). First, we search the condition that allows $\lambda_2 \geq 0$ which equals to:

$$\begin{aligned} &-\mu + \sqrt{\mu^2 + a\mu - b_2 \ln(1/(1-s)) - m\mu b_1} \geq 0 \\ &\Leftrightarrow a\mu - b_2 \ln(1/(1-s)) - m\mu b_1 \geq 0 \\ &\Leftrightarrow a - \frac{b_2 \ln(1/(1-s))}{\mu} - b_1 m \geq 0 \end{aligned}$$

Then, we have our second condition $\lambda_2 \leq a - \frac{b_2 \ln(1/(1-s))}{\mu} - b_1 m$ which is equivalent to:

$$\begin{aligned} &-\mu + \sqrt{\mu^2 + a\mu - b_2 \ln(1/(1-s)) - m\mu b_1} \leq a - \frac{b_2 \ln(1/(1-s))}{\mu} - b_1 m \\ &\Leftrightarrow a\mu - b_2 \ln(1/(1-s)) - m\mu b_1 \leq \left(a - \frac{b_2 \ln(1/(1-s))}{\mu} - b_1 m + \mu\right)^2 - \mu^2 \\ &\Leftrightarrow a\mu - b_2 \ln(1/(1-s)) - m\mu b_1 \leq \left(a - \tfrac{b_2 \ln(1/(1-s))}{\mu} - b_1 m + \right)^2 + 2\mu \left(a - \tfrac{b_2 \ln(1/(1-s))}{\mu} - b_1 m\right) \\ &\Leftrightarrow a - \tfrac{b_2 \ln(1/(1-s))}{\mu} - b_1 m \leq 2\left(a - \tfrac{b_2 \ln(1/(1-s))}{\mu} - b_1 m\right) + \tfrac{1}{\mu}\left(a - \tfrac{b_2 \ln(1/(1-s))}{\mu} - b_1 m\right)^2 \end{aligned}$$

which is also equivalent to:

$$-\mu \leq a - \frac{b_2 \ln(1/(1-s))}{\mu} - b_1 m$$

This condition must already be satisfied to verify $\lambda_2 \geq 0$. Thus, λ_2 is the unique feasible stationary point of our problem.
We have $\Pi(\lambda_2) = \frac{\lambda_2(a\mu - b_2 \ln(1/(1-s)) - \lambda_2\mu - m\mu b_1)}{\mu b_1 + \lambda_2 b_1}$. Given that $a - \frac{b_2 \ln(1/(1-s))}{\mu} - b_1 m > 0$, one can easily check that $\Pi(\lambda_2) > 0$.
In addition, $\lim_{\lambda \to 0} \Pi(\lambda) = 0$ and $\lim_{\lambda \to a - \frac{b_2 \ln(1/(1-s))}{\mu} - b_1 m} \Pi(\lambda) = 0$. Hence, λ_2 is the optimal demand as given in proposition 4.1. The optimal price and profit follow immediately. ■

It has been shown when the penalty and holding cost are removed the problem can be solved analytically. Therefore, we increase the difficulty by considering lateness penalty and holding cost. We investigate the problem with lateness penalty and holding cost in the next subsection.

4.2.2 The M/M/1/1 model: congestion & lateness costs are considered

With the consideration of penalty and holding costs, the objective function will be composed by three terms: expected revenue, total congestion costs, and total lateness penalty costs. The formulation of this objective function has been presented in section 4.1. With $K = 1$, it can be shown by standard calculus that model (M_K) is equivalent to model (M_1) given below.

$$(M_1)\ \underset{P,L,\lambda}{Maximize}\ \frac{\lambda}{\mu + \lambda}\left(\mu(P - m) - F - c_r e^{-\mu L}\right) \tag{4.15}$$

$$\text{Subject to } \lambda \leq a - b_1 P - b_2 L \tag{4.16}$$

$$1 - e^{-\mu L} \geq s \tag{4.17}$$

$$\lambda, P, L \geq 0 \tag{4.18}$$

As was demonstrated for the case without penalty and holding costs, demand constraint (eq. (4.16)) is binding and we have $P = \frac{a - b_2 L - \lambda}{b_1}$ at optimality (proof similar to the proof of Lemma 4.1).

Substituting price P by its value and rewriting $1 - e^{-\mu L} \geq s$ as $L \geq \frac{1}{\mu} \ln\left(\frac{1}{1-s}\right)$, we obtain the following equivalent formulation of (M_1).

$$\underset{L,\lambda}{Maximize}\ \Pi(L, \lambda) = \frac{\lambda\left[\frac{\mu(a - b_2 L - \lambda)}{b_1} - m\mu - F - c_r e^{-\mu L}\right]}{\mu + \lambda} \tag{4.19}$$

$$\text{Subject to } L \geq \frac{1}{\mu} \ln\left(\frac{1}{1-s}\right) \tag{4.20}$$

$$L \leq \frac{a - \lambda}{b_2} \tag{4.21}$$

$$\lambda, L \geq 0 \tag{4.22}$$

Note that, in constraint (4.21), we forced lead time L to be smaller than $(a-\lambda)/b_2$ in order to guarantee that the price is positive according to the expression of P. In case where the inventory holding and lateness penalty costs are ignored, it is proven that the service level constraint is binding, which simplifies the solving approach. However, under our setting with holding and penalty costs, the service constraint (constraint (4.20)) is not necessarily binding. Indeed, for large values of unit penalty cost (c_r), the achieved service level has to be very high (close to 1) to avoid a high penalty cost. This means that the achieved service level can be greater than the imposed service level (s). In order to solve the problem, we will now determine in Lemma 4.3 when service constraint is binding.

Lemma 4.3. *There exists a critical value of service level $s_c = 1 - \frac{b_2}{c_r b_1}$ such as the service constraint is binding if and only if $s \geq s_c$.*

Proof. We have $\frac{\partial}{\partial L}\Pi(\lambda, L) = -\frac{\lambda\mu[b_2 - b_1 c_r e^{-\mu L}]}{b_1(\mu+\lambda)}$.
Case of $s \geq s_c$. It can be verified by standard calculus that if $s \geq s_c$, which is equivalent to $1 - e^{-\mu L} \geq 1 - \frac{b_2}{c_r b_1} \Leftrightarrow b_2 - b_1 c_r e^{-\mu L} \geq 0$, then we have $\frac{\partial}{\partial L}\Pi(\lambda, L) \leq 0$ for the feasible values of L (eq. 4.20-4.21), that is for $L \in \left[\frac{1}{\mu}\ln\left(\frac{1}{1-s}\right), \frac{a-\lambda}{b_2}\right]$. In this case, the profit is therefore a decreasing function in L. Hence, the profit maximum can be obtained from the smallest feasible L. Consequently, if $s \geq s_c$ then the service level constraint is binding.
Case of $s < s_c$. The optimal lead time L^* must verify $\frac{\partial}{\partial L}\Pi(\lambda, L) = 0$. This equation has a unique solution, given by $\frac{1}{\mu}\ln\left(\frac{b_1 c_r}{b_2}\right) = \frac{1}{\mu}\ln\left(\frac{1}{1-s_c}\right)$. This solution is therefore the only candidate for optimality. If $s < s_c$ then we have $\frac{1}{\mu}\ln\left(\frac{1}{1-s_c}\right) > \frac{1}{\mu}\ln\left(\frac{1}{1-s}\right)$, implying that the service constraint is satisfied for the candidate solution $L = \frac{1}{\mu}\ln\left(\frac{1}{1-s_c}\right)$ and, consequently, the service constraint is not binding. ■

Based on the results of lemma 4.3 and expression of P, we can now announce the optimal solution of lead time quotation and pricing problem in M/M/1/1 with penalty and holding costs.

Proposition 4.2. *Assuming $a \geq \frac{b_2 \ln\left(\max\left\{\frac{1}{1-s}, \frac{b_1 c_r}{b_2}\right\}\right) + b_1\left(\mu m + F + c_r \max\left\{1-s, \frac{b_2}{b_1 c_r}\right\}\right)}{\mu}$ (since otherwise the problem is not relevant), the optimal solution of lead time quotation and pricing problem in M/M/1/1 with penalty and holding costs is the following.*

- *Optimal lead time:*
  $L^* = \frac{\ln\left(\max\left\{\frac{1}{1-s}, \frac{b_1 c_r}{b_2}\right\}\right)}{\mu}$,
- *Optimal price:*
  $P^* = \frac{a - \frac{b_2}{\mu}\ln\left(\max\left\{\frac{1}{1-s}, \frac{b_1 c_r}{b_2}\right\}\right) + \mu - \sqrt{\mu^2 + a\mu - b_2 \ln\left(\max\left\{\frac{1}{1-s}, \frac{b_1 c_r}{b_2}\right\}\right) - b_1\left(\mu m - F - c_r \max\left\{1-s, \frac{b_2}{b_1 c_r}\right\}\right)}}{b_1}$,

- *Optimal demand:*
$\lambda^* = -\mu + \sqrt{\mu^2 + a\mu - b_2 \ln\left(\max\left\{\frac{1}{1-s}, \frac{b_1 c_r}{b_2}\right\}\right) - b_1\left(\mu m - F - c_r \max\left\{1-s, \frac{b_2}{b_1 c_r}\right\}\right)}$

Proof. We firstly consider the case of $s \geq s_c$ (i.e., when service constraint is binding) and then turn to the case of $s < s_c$ (non-binding situation).
Case 1: $s \geq s_c$. In this case, $L^* = \frac{1}{\mu}\ln\left(\frac{1}{1-s}\right)$ according to the service constraint (4.20). Substituting L^* by its value, we transform the problem into a one variable optimization model in λ. We use the first derivative condition to obtain the candidates for optimality.

$$\frac{d}{d\lambda}\Pi(\lambda) = 0 \Leftrightarrow -\lambda^2 - 2\mu\lambda + a\mu - b_2 \ln\left(\frac{1}{1-s}\right) - \mu m b_1 - F b_1 - b_1 c_r (1-s) = 0$$

The discriminant of this quadratic equation in λ is:

$$\Delta = 4\mu^2 + 4\left(a\mu - b_2 \ln\left(\frac{1}{1-s}\right) - \mu m b_1 - F b_1 - b_1 c_r (1-s)\right)$$

The case of $\Delta < 0$ is not relevant to our study (since, in this case, the profit is decreasing in λ, implying that the optimal demand and profit are equal to zero or the problem is infeasible).
We focus on the case of $\Delta \geq 0$. If $\Delta \geq 0$ then we have two roots:

$$\lambda_1 = -\mu + \sqrt{\mu^2 + a\mu - b_2 \ln\left(\frac{1}{1-s}\right) - \mu m b_1 - F b_1 - b_1 c_r (1-s)}$$

and

$$\lambda_2 = -\mu - \sqrt{\mu^2 + a\mu - b_2 \ln\left(\frac{1}{1-s}\right) - \mu m b_1 - F b_1 - b_1 c_r (1-s)}$$

Root λ_2 is negative, so infeasible. The first root λ_1 is positive if and only if $a \geq \frac{b_2 \ln\left(\frac{1}{1-s}\right) + b_1(\mu m + F + c_r(1-s))}{\mu}$. We assume this condition holds (since otherwise the optimal demand and profit are equal to zero or the problem is infeasible).
Under this condition, one can check that λ_1 satisfies constraint (4.21). Hence, the only candidate for optimality is λ_1. Given that the profit obtained with λ_1 is positive and that the limits of the objective function in the endpoints do not improve this profit, we deduce that the optimal solution of the problem is: $L^* = \frac{1}{\mu}\ln\left(\frac{1}{1-s}\right), \lambda^* = \lambda_1$.
Case 2: $s < s_c$. In this case, service constraint (4.20) is not binding, so can be ignored.

$$\frac{\partial}{\partial L}\Pi(\lambda, L) = 0 \Leftrightarrow \lambda\mu\left[b_2 - b_1 c_r e^{-\mu L}\right] = 0 \qquad \text{(the case of } \lambda = 0 \text{ is not relevant)}$$

$$\Leftrightarrow L = \frac{1}{\mu}\ln\left(\frac{b_1 c_r}{b_2}\right)$$

$$\frac{\partial}{\partial \lambda}\Pi(\lambda, L) = 0 \Leftrightarrow a\mu - \mu b_2 L - \mu m b_1 - F b_1 - b_1 c_r e^{-\mu L} - 2\mu\lambda - \lambda^2 = 0$$

Substituting L by $\frac{1}{\mu}\ln\left(\frac{b_1 c_r}{b_2}\right)$ in $\frac{\partial}{\partial\lambda}\Pi(\lambda, L)$ equation, we obtain:

$$-\lambda^2 - 2\mu\lambda + a\mu - b_2 \ln\left(\frac{b_1 c_r}{b_2}\right) - \mu m b_1 - F b_1 - b_2 = 0$$

We focus on the case where the discriminant is positive (since otherwise the problem is not relevant as explained earlier).
Under this condition, the equation has two roots:

$$\lambda_1 = -\mu + \sqrt{\mu^2 + a\mu - b_2 \ln\left(\frac{b_1 c_r}{b_2}\right) - b_1\left(\mu m + F + \frac{b_2}{b_1}\right)}$$

and

$$\lambda_2 = -\mu - \sqrt{\mu^2 + a\mu - b_2 \ln\left(\frac{b_1 c_r}{b_2}\right) - b_1\left(\mu m + F + \frac{b_2}{b_1}\right)}$$

The second root is always negative, so infeasible. The first root λ_1 is positive if and only if $a \geq \frac{b_2 \ln\left(\frac{b_1 c_r}{b_2}\right) + b_1\left(\mu m + F + \frac{b_2}{b_1}\right)}{\mu}$. Assuming this condition holds, there is a unique candidate for optimality $\left(L = \frac{1}{\mu}\ln\left(\frac{b_1 c_r}{b_2}\right),\ \lambda = \lambda_1\right)$. One can check that this solution is feasible (satisfies constraint (4.21)). In addition, it leads to a positive profit while the limits of the objective function in the endpoints do not give a better profit. Thus, $\left(\frac{1}{\mu}\ln\left(\frac{b_1 c_r}{b_2}\right), \lambda_1\right)$ is the optimal solution.
Note finally that in both cases (binding or non-binding situation), the optimal price can be directly obtained by using the expression of $P = \frac{a - b_2 L - \lambda}{b_1}$. ■

The expression of the optimal quoted lead time in Proposition 4.2 shows that an increase in lead time-sensitivity b_2 can firstly lead to reducing the quoted lead time, but when b_2 becomes greater than $(1-s)\, b_1 c_r$, the lead time sensitivity has no more effect on the quoted lead time. Furthermore, we can deduce that an increase in price sensitivity (beyond $\frac{b_2}{(1-s)c_r}$) or in the unit lateness penalty cost (beyond $\frac{b_2}{(1-s)c_r}$) always favors quoting a longer lead time. Note also that the optimal quoted lead time does not depend on the unit holding cost F. Finally, the quoted lead time is decreasing in production capacity (i.e., in mean service rate μ), as expected.

In the next section, we use the analytical result of Proposition 4.1 and 4.2 to compare the rejection policy, modeled as an M/M/1/1, with the all customers' acceptance policy, M/M/1.

4.3 Performance of the rejection policy (M/M/1/1) with comparison to the all-customers' acceptance policy (M/M/1)

The fact that rejecting some customers might help to quote shorter lead time for

the accepted ones raise the question about the interest of the all-customers' acceptance policy with comparison to the rejection policy when customers are sensitive to lead time. The consideration of holding and lateness costs is also expected to impact on this trade-off. Under different parameters setting, we investigate in this section whether a customer rejection policy, represented by an M/M/1/1 model, can be more profitable for the firm than an all-customers' acceptance policy, represented by an M/M/1 model.

Thus, we compare the performance of our model, where the firm is modeled as an M/M/1/1 queue to two relevant models of the literature where the firm is represented by an M/M/1 queue: Palaka et al. (1998) and Pekgün et al. (2008). Both of these models consider a similar framework than the one used in this chapter (but all customers are accepted in their model unlike us). Nevertheless, recall that Palaka et al. (1998) include holding and lateness penalty costs while Pekgün et al. (2008) ignore such costs. We compare the results of Palaka et al. (1998) to our results obtained from model M_1 (section 4.2.2) and the results of Pekgün et al. (2008) to our result of model $M_1^{'}$ (section 4.2.1).

To conduct our experiments, we consider again the base scenario used by Pekgün et al. (2008) with the following setting: market potential $(a) = 50$, lead time sensitivity $(b_2) = 6$, price sensitivity $(b_1) = 4$, production capacity $(\mu) = 10$, service level $(s) = 0.95$, unit direct variable cost $(m) = 5$. In the comparison to Palaka et al. (1998), we consider the unit holding $(F) = 2$ and the cost per job per unit lateness $(c_r) = 10$. In each experiment, we vary one parameter while keeping the others constant, and deduce the relative gain resulting from using the M/M/1/1 model instead of M/M/1. This gain is given by $\frac{Profit^{M/M/1/1} - Profit^{M/M/1}}{Profit^{M/M/1}} \times 100$. Clearly, a positive gain means that the rejection policy (M/M/1/1) is better and vice versa.

Note that binding and non-binding situation can be seen from the ^{nb} symbol in the profit (i.e., 36.22^{nb}). The ^{nb} means that it is a case in non-binding situation. Otherwise, it is a case in binding situation.

4.3.1 Effect of lead time-sensitivity

First, we study the impact of lead time sensitivity (b_2) and report the results in Table 4.1. As expected, the M/M/1 is better for small values of b_2 whether the holding and penalty costs are considered or not. It is then interesting to note that an increase in lead time-sensitivity favors the rejection policy, even in the absence of holding and penalty costs. In M/M/1, all customers are accepted, which leads to a long time in the system with comparison to M/M/1/1. We can observe in Table 4.1 that the quoted lead time is always longer for M/M/1. Therefore, even when the holding and penalty costs are not considered, the firm cannot quote a very short lead time since, otherwise, the service level cannot be satisfied. When the customers are highly sensitive to lead time, the impossibility of quoting short lead time in M/M/1 leads to a much smaller demand with comparison to M/M/1/1, which explains why the M/M/1/1 can be better for high values of b_2 even without

holding and penalty costs. The consideration of these costs favors the M/M/1/1. For this reason, the M/M/1/1 becomes more rapidly better than M/M/1 when we increase b_2 under holding and penalty costs (with comparison to the case where these costs are ignored). Indeed, the holding cost is higher in M/M/1 because there is more congestion compared to M/M/1/1. The lateness penalty cost also favors the rejection policy as it obliges the M/M/1 to quote longer lead time in order to reduce the expected lateness, which decreases the demand and the firm's profit.

Intuitively, one could expect that when customers are more sensitive to lead time, the firm reacts by quoting a higher price in order to capitalize on the existing demand. However, in both models (M/M/1 and M/M/1/1), an increase in lead time sensitivity implies a smaller quoted price. Indeed, since the firm cannot always reduce the lead time as desired (because of service constraint and penalty cost), the increase in lead time sensitivity will finally lead to a significant decrease in demand. In order to offset this decrease in demand, the firm reacts by setting a smaller price.

Table 4.1: M/M/1/1 vs M/M/1 for different values of b_2

b_2	**with** holding and penalty costs							**without** holding and penalty costs						
	M/M/1			M/M/1/1			Gains	M/M/1			M/M/1/1			Gains
	L^*	P^*	Profit	L^*	P^*	Profit		L^*	P^*	Profit	L^*	P^*	Profit	
0.001	4.53	10.58	36.22^{nb}	1.06	10.05	24.00^{nb}	-33.7%	55.03	10.00	49.73	0.30	10.00	25.00	-49.7%
2	0.96	10.31	30.95	0.30	9.95	23.02	-25.6%	1.45	9.79	38.02	0.30	9.89	24.25	-36.2%
4	0.84	10.06	27.99	0.30	9.84	22.29	-20.4%	1.08	9.61	33.34	0.30	9.78	23.51	-29.5%
6	0.76	9.85	25.51	0.30	9.73	21.57	-15.5%	0.91	9.46	29.90	0.30	9.66	22.78	-23.8%
8	0.70	9.67	23.37	0.30	9.62	20.85	-10.8%	0.81	9.31	27.13	0.30	9.55	22.05	-18.7%
10	0.66	9.50	21.48	0.30	9.51	20.14	-6.3%	0.74	9.17	24.77	0.30	9.44	21.33	-13.9%
12	0.62	9.35	19.80	0.30	9.40	19.43	-1.8%	0.68	9.04	22.72	0.30	9.33	20.61	-9.3%
14	0.59	9.20	18.27	0.30	9.29	18.73	2.5%	0.64	8.91	20.91	0.30	9.22	19.90	-4.8%
16	0.56	9.07	16.88	0.30	9.18	18.04	6.9%	0.61	8.79	19.28	0.30	9.11	19.20	-0.4%
18	0.54	8.93	15.61	0.30	9.07	17.36	11.2%	0.58	8.67	17.80	0.30	9.00	18.50	4.0%
20	0.52	8.81	14.44	0.30	8.96	16.68	15.5%	0.56	8.56	16.45	0.30	8.89	17.81	8.3%

4.3.2 Effect of price-sensitivity

Second, we study the impact of price sensitivity (b_1) and report the results in Table 4.2. We can observe that an increase in price-sensitivity favors the rejection policy whether the holding and penalty costs are considered or not. When b_1 goes up, both models react by decreasing the price as expected. Regarding the quoted lead time, it remains constant in M/M/1/1 as we are in the binding situation for all values of b_1 between 0.001 until 9 (see Lemma 4.3 and Proposition 4.2) while it decreases in M/M/1. Let us recall that lead time doesn't depend on b_1 in binding situation of M/M/1/1. The motivation of decreasing the quoted lead time is to maintain an interesting amount of demand given that the increase in price-sensitivity leads to a significant decrease in demand. However, with comparison to M/M/1/1, the firm

cannot set the lead time as short as desired in M/M/1 because it is more difficult to satisfy the service level constraint in this case. This explains why the M/M/1/1 performs better for high values of b_1 even when the holding and penalty costs are not considered. As expected, the integration of these costs favors the rejection policy, which makes the M/M/1/1 better than M/M/1 for smaller values of price-sensitivity. Note finally that the price in the first row of Table 4.2 is unrealistically high because b_1 tends to 0.

Table 4.2: M/M/1/1 vs M/M/1 for different values of b_1

b_1	**with** holding and penalty cost							**without** holding and penalty cost						
	M/M/1			M/M/1/1			Gains	M/M/1			M/M/1/1			Gains
	L^*	P^*	Profit	L^*	P^*	Profit		L^*	P^*	Profit	L^*	P^*	Profit	
0.001	1.31	34400.16	265493.88	0.30	34078.44	199490.73	-24.9%	1.31	34399.50	265502.3	0.30	34078.39	199492.19	-24.9%
2	1.01	18.47	88.70	0.30	18.18	70.10	-21.0%	1.12	17.96	95.06	0.30	18.12	71.46	-24.8%
4	0.76	9.85	25.51	0.30	9.73	21.57	-15.5%	0.91	9.46	29.90	0.30	9.66	22.78	-23.8%
6	0.55	7.03	7.13	0.30	6.98	6.70	-6.0%	0.67	6.74	9.63	0.30	6.90	7.69	-20.1%
8	0.38	5.72	0.81	0.30	5.68	0.93	14.4%	0.43	5.54	1.66	0.30	5.59	1.52	-8.4%
9	0.31	5.31	0.018	0.30	5.30	0.024	35.5%	0.34	5.18	0.24	0.30	5.19	0.25	4.5%

4.3.3 Effect of service level

Third, we vary the service level (s) and report the results in Table 4.3. For all models, a higher service level leads, as expected, to a longer quoted lead time since, otherwise, the service constraint cannot be satisfied. Consequently, an increase in service level could lead to a significant decrease in demand. In M/M/1/1, the firm is able to quote relatively short lead times even for high values of service level. This favors the M/M/1/1, when the holding and penalty costs are also considered. In the absence of these costs, we can see that the impact of increasing the service level is not significant enough to make the M/M/1/1 more profitable than the M/M/1.

Table 4.3: M/M/1/1 vs M/M/1 for different values of s

s	**with** holding and penalty cost							**without** holding and penalty cost						
	M/M/1			M/M/1/1			Gains	M/M/1			M/M/1/1			Gains
	L^*	P^*	Profit	L^*	P^*	Profit		L^*	P^*	Profit	L^*	P^*	Profit	
0.5	0.50	10.20	26.54^{nb}	0.19	9.88	21.88^{nb}	-17.6%	0.40	9.85	39.96	0.07	9.92	24.48	-38.7%
0.6	0.50	10.20	26.54^{nb}	0.19	9.88	21.88^{nb}	-17.6%	0.46	9.80	38.50	0.09	9.90	24.32	-36.8%
0.7	0.50	10.20	26.54^{nb}	0.19	9.88	21.88^{nb}	-17.6%	0.54	9.75	36.89	0.12	9.86	24.10	-34.7%
0.8	0.50	10.20	26.54^{nb}	0.19	9.88	21.88^{nb}	-17.6%	0.64	9.68	34.95	0.16	9.82	23.80	-31.9%
0.85	0.50	10.20	26.54	0.19	9.88	21.88	-17.6%	0.70	9.63	33.73	0.19	9.79	23.59	-30.1%
0.9	0.60	10.05	26.36	0.23	9.82	21.83	-17.2%	0.78	9.57	32.20	0.23	9.74	23.29	-27.7%
0.95	0.76	9.85	25.51	0.30	9.73	21.57	-15.5%	0.91	9.46	29.90	0.30	9.66	22.78	-23.8%
0.99	1.05	9.52	22.70	0.46	9.54	20.60	-9.2%	0.99	9.67	25.04	0.46	9.49	21.61	-13.7%
0.999	1.39	9.16	18.88	0.69	9.28	19.02	0.7%	1.49	8.92	21.05	0.69	9.23	19.96	-5.2%

4.3.4 Effect of holding cost

Fourth, we vary the holding cost (F) and report the results in Table 4.4. As expected, an increase in unit holding cost, F, favors the rejection policy. Indeed, in M/M/1, the probability that a client spends a long time in the system is high (with comparison to M/M/1/1), which leads to a large holding cost when the value of F goes up.

As shown in Proposition 4.2, the optimal quoted lead time does not depend on the unit holding cost when the firm is modeled as an M/M/1/1 queue. This explains why we have the same optimal quoted lead time for M/M/1/1 in Table 4.4 (note that we have $L^* = 0.3$ because we are in the binding situation. If we had a non-binding situation then L^* would be different from 0.3 but would remain constant when we increase F). When the holding cost increases the firm reacts by increasing the price in order to ensure the profitability. For $F \neq 0$, since the holding cost is higher in M/M/1, we observe that the price with M/M/1 is always greater than the price associated with M/M/1/1. Given in addition that the quoted lead time is higher in M/M/1, the demand for M/M/1 is significantly smaller than the demand associated with the M/M/1/1 system.

Table 4.4: M/M/1/1 vs M/M/1 for different values of F

F	M/M/1			M/M/1/1			Gains
	L^*	P^*	Profit	L^*	P^*	Profit	
0	0.87	9.55	28.92	0.30	9.68	22.54	-22.1%
2	0.76	9.85	25.51	0.30	9.73	21.57	-15.5%
4	0.68	10.07	22.71	0.30	9.78	20.61	-9.3%
6	0.63	10.24	20.33	0.30	9.83	19.66	-3.3%
8	0.59	10.38	18.26	0.30	9.89	18.73	2.5%
10	0.56	10.51	16.44	0.30	9.94	17.81	8.3%
12	0.53	10.62	14.81	0.30	10.00	16.90	14.1%
14	0.51	10.71	13.35	0.30	10.05	16.00	19.9%
16	0.49	10.80	12.03	0.30	10.11	15.12	25.7%
18	0.47	10.88	10.83	0.30	10.16	14.25	31.6%
20	0.46	10.96	9.74	0.30	10.22	13.40	37.6%

4.3.5 Effect of lateness penalty cost

Finally, we vary the unit penalty cost (c_r) and report the results in Table 4.5. In these experiments, the M/M/1/1 is never better than M/M/1 even with high values

of penalty cost. Nevertheless, an increase in penalty cost closes the gap between the two models. Note also that we have a relatively stable profit with M/M/1/1 for increasing values of penalty cost while the profit decreases significantly under M/M/1. This is explained by the fact that we have much less overdue clients in M/M/1/1.

It is also interesting to note that when the unit penalty cost goes up, the M/M/1 model first reacts by decreasing the quoted lead time and increasing the price while, for M/M/1/1, the quoted lead time remains constant (as we are in the binding situation according to Lemma 4.3) and the price slightly increases. If we continue increasing the unit penalty cost, both models react by quoting longer lead time and smaller price.

Table 4.5: M/M/1/1 vs M/M/1 for different values of c_r

c_r	M/M/1			M/M/1/1			Gains
	L^*	P^*	Profit	L^*	P^*	Profit	
0	0.78	9.79	26.30	0.30	9.72	21.81	-17.1%
5	0.77	9.82	25.90	0.30	9.72	21.69	-16.3%
10	0.76	9.85	25.51	0.30	9.73	21.57	-15.5%
15	0.75	9.88	25.13	0.30	9.74	21.45	-14.7%
20	0.74	9.91	24.77	0.30	9.74	21.33	-13.9%
25	0.73	9.94	24.40	0.30	9.75	21.21	-13.1%
30	0.72	9.97	24.05	0.30	9.75	21.09	-12.3%
35	0.75	9.94	23.73^{nb}	0.31	9.74	20.98^{nb}	-11.6%
40	0.77	9.91	23.46^{nb}	0.33	9.72	20.88^{nb}	-11.0%
45	0.79	9.89	23.23^{nb}	0.34	9.71	20.80^{nb}	-10.5%
50	0.81	9.87	23.02^{nb}	0.35	9.70	20.72^{nb}	-10.0%

In this section, it has been shown that the client rejection policy (M/M/1/1) can be better than the all-customers' acceptance policy (M/M/1) in different situations even when the penalty and holding costs are not considered. In M/M/1/1, customers are rejected if there is already one customer in the system, which is very restrictive. It is therefore interesting to study the performance of the rejection policy when we increase the system size (namely, increase K). We investigate this problem in the next section.

4.4 The M/M/1/K model: numerical solution and experiments

The formulation of lead time quotation and pricing problem in an M/M/1/K queue with penalty and holding costs was presented at the beginning of the chapter (model M_K). Given that the demand constraint is binding, M_K can be formulated as follows.

$$(M_K)\ \underset{L,\lambda}{Maximize}\ \lambda(1-P_K)\left(\frac{a-b_2L-\lambda}{b_1}-m\right)-\left(\left(\frac{\rho}{1-\rho}-\frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}\right)\times F\right) - \left(c_r\times\lambda\times e^{-\mu L}\times\sum_{k=0}^{K-1}\frac{P_k}{k!}\left[\frac{(\mu L)^{k+1}}{\mu}+\left(\frac{k+1}{\mu}-L\right)\sum_{i=0}^{k}\frac{k!}{i!}(\mu L)^i\right]\right) \tag{4.23}$$

$$\text{Subject to } 1-\textstyle\sum_{k=0}^{K-1}\frac{P_k}{1-P_K}\left(\sum_{i=0}^{k}\frac{(\mu L)^i}{i!}e^{-\mu L}\right)\geq s \tag{4.24}$$

Recall that the service constraint is not necessarily binding. Since it is not possible to find an analytical solution of model (M_K), we are going to use a numerical solving approach. We firstly transform (M_K) into an unconstrained optimization model by adding a penalty $(\eta G(\lambda, L))$ for violating the service constraint, where η is a very large number and $G(\lambda, L) = \max\left(0, \left[s-1+\sum_{k=0}^{K-1}\frac{P_k}{1-P_K}\left(\sum_{i=0}^{k}\frac{(\mu L)^i}{i!}e^{-\mu L}\right)\right]\right)$. Each time the constraint is violated, we will have a negative profit which is infeasible. And by the iterative process of the meta-heuristic, negative profit will disappear (replaced by a better solution). Thus, we consider the following model denoted by (M_K^*).

$$(M_K^*)\ \underset{L,\lambda}{Maximize}\ \lambda(1-P_K)\left(\frac{a-b_2L-\lambda}{b_1}-m\right)-\left(\left(\frac{\rho}{1-\rho}-\frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}\right)\times F\right) - \left(c_r\times\lambda\times e^{-\mu L}\times\sum_{k=0}^{K-1}\frac{P_k}{k!}\left[\frac{(\mu L)^{k+1}}{\mu}+\left(\frac{k+1}{\mu}-L\right)\sum_{i=0}^{k}\frac{k!}{i!}(\mu L)^i\right]\right) - \eta[G(\lambda, L)] \tag{4.25}$$

To solve model (M_K^*), we use a numerical solving approach based on the Particle Swarm Optimization (PSO) method. Given that our decision variables are real numbers and we want to do some explorations, hence we choose to use a population based algorithm such as particle swarm optimization (PSO). The pseudo code of PSO is given in Appendix C. In our experiments, we consider $K = 2$, $K = 3$, $K = 5$, and $K = 10$. Recall that our main objective is to investigate the performance of the rejection policy for different values of K. We use the same parameters setting of the previous section: market potential $(a) = 50$, lead time sensitivity $(b_2) = 6$, price sensitivity $(b_1) = 4$; Production capacity $(\mu) = 10$; service level $(s) = 0.95$;

unit direct variable cost (m) = 5, holding cost (F) = 2 and penalty cost (c_r) = 10. For each instance, we run the PSO heuristic 31 times and take the solution that gives the maximum profit.

The following figures show the profit obtained with M/M/1, M/M/1/1, M/M/1/2, M/M/1/3, M/M/1/5, and M/M/1/10 while varying one parameter at each time. The analysis of these figures leads to the following observations:

- Obviously as K increases the observed behavior of M/M/1/K becomes similar to M/M/1.

- In all our experiments (for the different values of model parameters), there is always at least one value of K for which the M/M/1/K is more profitable than the M/M/1.

- For high values of lead time-sensitivity b_2 and, mainly, for high values of holding cost F, a tough rejection policy (i.e., M/M/1/1) can be better than a more flexible rejection policy (M/M/1/2, M/M/1/3, or M/M/1/5). In most of the other cases, the M/M/1/1 is less profitable than M/M/1/K (for $K > 1$).

- In most cases, an increase in K has a non-monotonous effect on the firm's profit. For instance, we can see in Fig. 4.5 and 4.3 (for all considered values of penalty cost c_r and service level s, respectively) that an increase in K first improves the profit and then leads to decreasing it (see how the profit of $K = 5$ is smaller than the profit of $K = 3$ while the profit with $K = 3$ is greater than the profit of $K = 2$, which is greater than the profit of $K = 1$). We observe the same situation in Fig 4.1, 4.2 and 4.4 for certain ranges of b_2, b_1, and F, respectively. It might be expected that an increase in K improves the profit as this gives more flexibility to the rejection policy. However, beyond a certain value of K, the waiting queue becomes relatively long (too many clients are accepted) and we lose the interest of the rejection policy.

- Although this would be a very hard problem, the previous points show the relevance of studying the optimal system size K beyond which customers must be rejected in order to get the maximum profit. Thus, we have performed some experiments in finding numerically the K^{opt}. We put the detailed discussion of these experiments in appendix D. Our experiments show that an increase in lead time-sensitivity or price-sensitivity leads to a decrease in K_{opt}.

- When K becomes relatively large, we expect that the M/M/1/K will have the same behavior than the M/M/1. In Fig. 4.1 and 4.5, we can see that the curve associated with M/M/1/10 is closed to the M/M/1 curve. We would need to test larger values of K in order to better illustrate this observation. However, when K goes up it becomes very tough to solve the model with the numerical approach.

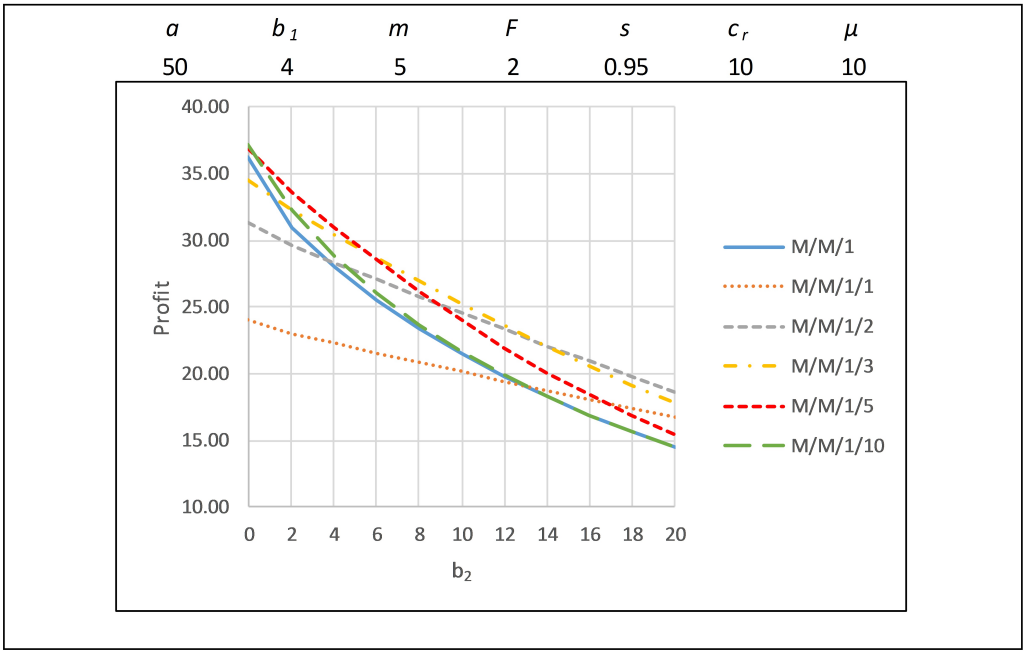


Figure 4.1: M/M/1/K performance as a function of b_2

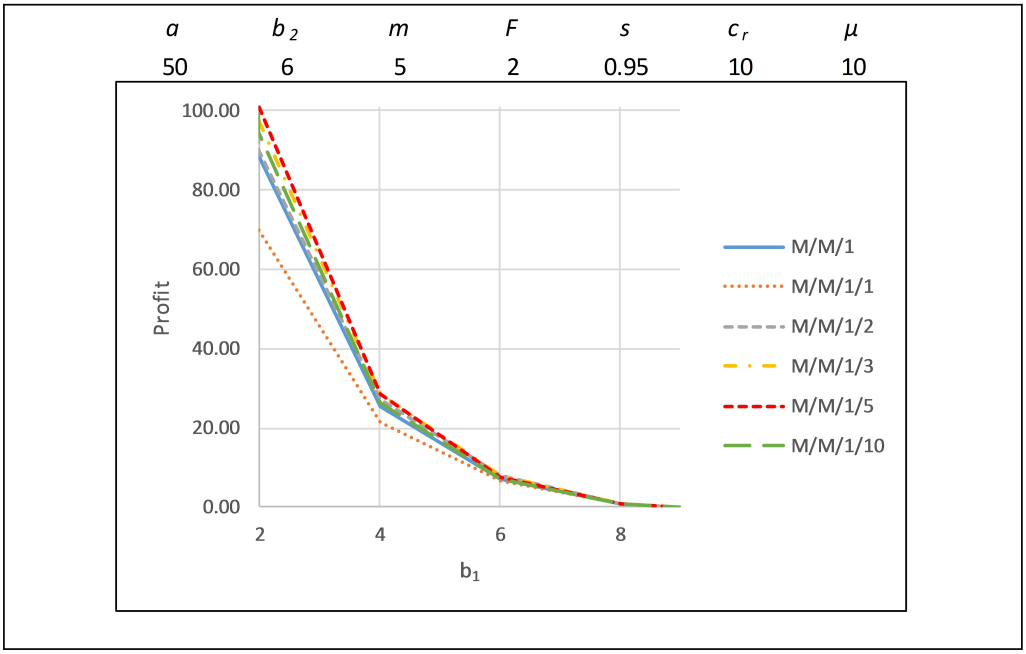


Figure 4.2: M/M/1/K performance as a function of b_1

4.5 Conclusion

In this chapter, we formulated the problem of lead time quotation and pricing for a profit-maximizing firm, modeled as an M/M/1/K system, facing a linear price-

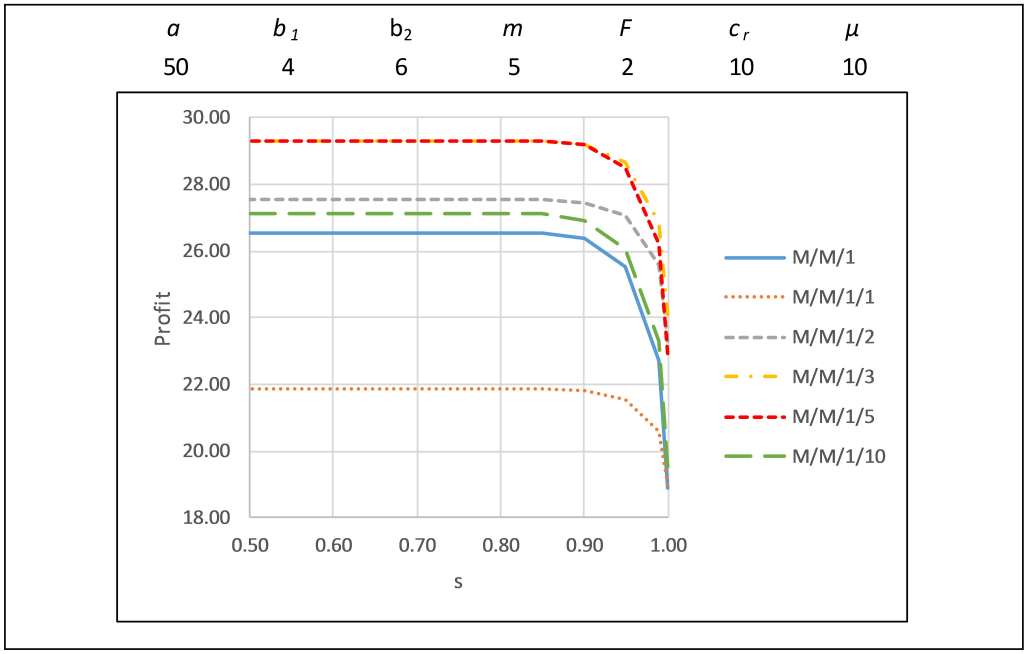


Figure 4.3: M/M/1/K performance as a function of s

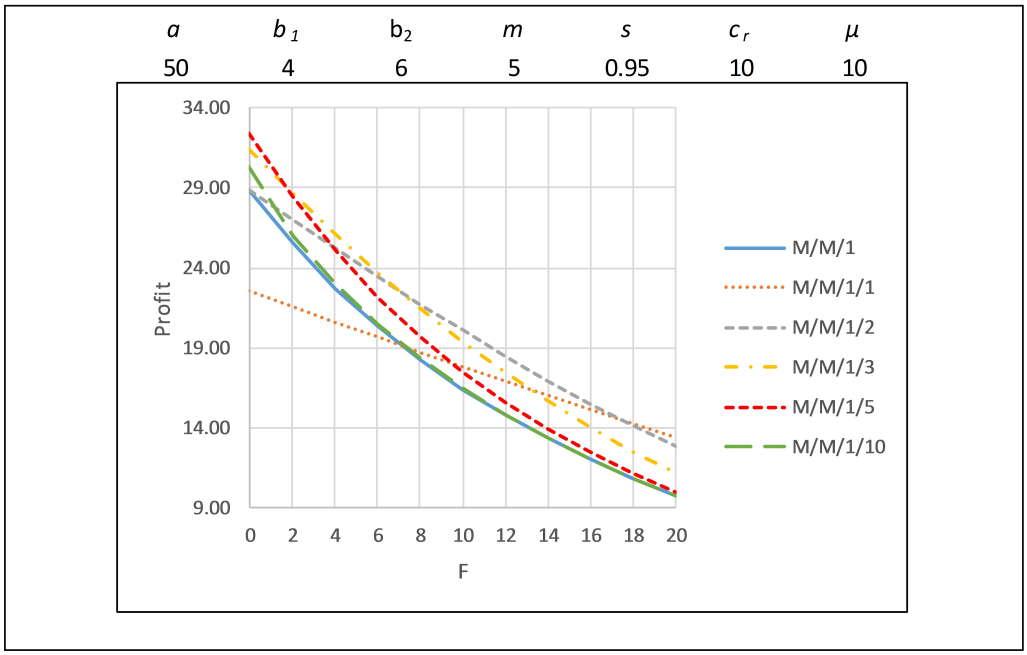


Figure 4.4: M/M/1/K performance as a function of F

and lead time-dependent demand with the consideration of inventory holding and lateness penalty costs. In order to determine the lateness penalty cost, we knocked out a theoretical barrier by explicitly calculating the expected lateness given that a job is late in an M/M/1/K queue when a certain delivery lead time is quoted to the

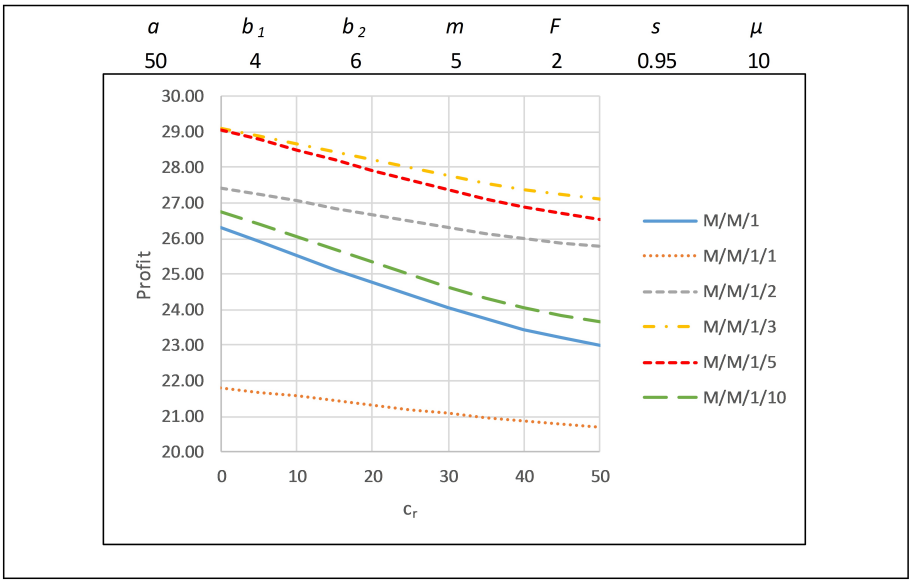


Figure 4.5: M/M/1/K performance as a function of c_r

customers. This result can be used in the future for different operations management and queuing theory problems.

We first focused on the case of $K = 1$ and analytically determined the optimal firm's policy (optimal price and quoted lead time) when the firm is modeled as an M/M/1/1 queue. The expression of the optimal solution showed that an increase in lead time-sensitivity first leads to reducing the quoted lead time but can rapidly become useless if it exceeds a certain threshold value (since the service constraint becomes binding). We also deduced that when the customers become more sensitive to price or when the unit lateness penalty cost increases, the firm can react by increasing the quoted lead time.

Then, we compared the optimal profit given by our M/M/1/1 model to the optimal profit obtained when the firm is modeled as an M/M/1 queue (as given in the literature). The M/M/1/1 system represents the rejection policy while in the M/M/1 all customers are accepted. We found out that a rejection policy can be more profitable than an all-customers' acceptance policy even when the holding and penalty costs are not considered. Some of our results showed that an increase in lead time sensitivity or in price-sensitivity favors the rejection policy. The increase in unit holding cost has been proven to be one of the main criteria that make the rejection policy better than the all-customers' acceptance policy. An increase in service level or in unit penalty cost also favors the rejection policy but has a much smaller impact than an increase in unit holding cost.

Finally, we solved the problem numerically for an M/M/1/K queue ($K > 1$). We conducted experiments to compare the results of our model, for different values of K, to the results obtained with an M/M/1 queue under different parameters settings. We showed, for all considered instances, that there is at least one value

of K for which the M/M/1/K (rejection policy) is more profitable than the M/M/1 (all customers' acceptance policy). In most cases, it has also been observed that an increase in the value of K (i.e., the system size) has a non-monotonous effect on the firm's profit. Indeed, an increase in K first improves the profit and then leads to decreasing it.

Chapter 5

Coordination of upstream-downstream supply chain under price and lead time sensitive demand: a tandem queue model

In the industrial world nowadays, a competition between enterprises is a competition between supply chains. And different stages of the supply chain are used to response the expectation of the clients. Let us take an example of airplane industry, such as Airbus. When an order arrives, Airbus has to quote the right price and lead time that fit the customer needs. Airbus will contact its suppliers, which will quote their price and lead time taking into account some constraints imposed by Airbus. Then, Airbus announces the final price and lead time to the costumer. The costumer will then decide on the basis of the lead time and price offered. If we add the feature of a lead time dependent demand in this supply chain, the problem will become complex. Thus it is interesting to study the behavior of such chain and how to quote the right lead time and price under the endogenous demand (demand sensitive to price and lead time). In this chapter, we only consider a two-stage supply chain which we model as a tandem queue network (M/M/1-M/M/1).

In chapter 2, we saw several papers in the literature who model different stages of a supply chain in cooperative firms, such as Liu et al. (2007); Xiao et al. (2011); Xiao and Shi (2012); Zhu (2015); and Xiao and Qi (2016). However, in this literature only one of its actor has a production process and lead time. The other actor doesn't have production process, in other words, the lead time is zero. Based on our knowledge, no paper considers two stages with both actors having production process (lead time).

5.1 System description (M/M/1–M/M/1)

We consider a two-stage supply chain consisting of an upstream actor (supplier or manufacturer) and a downstream actor (manufacturer or retailer). The demands arrive at the downstream actor according to a Poisson process. Both actors (upstream and downstream) have a fixed capacity with exponential service time. Thus,

we model the system as a tandem queuing network (M/M/1-M/M/1). This system is described in Figure 5.1.

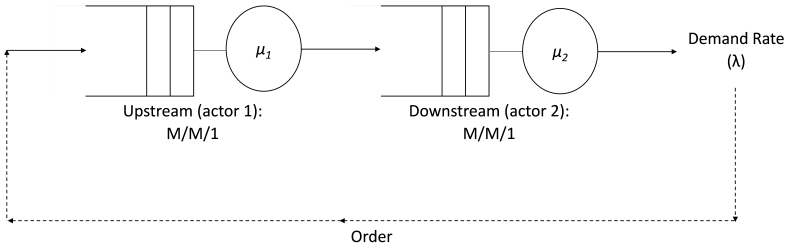


Figure 5.1: Queuing model

We use the following notations throughout this chapter:

Decision Variable

P_g = Global price of the good/service set by the chain	λ = mean arrival rate (demand)
P_1 = price of the good/service set by the first actor	L_g = Global quoted lead time
	L_1 = quoted lead time of actor 1
	L_2 = quoted lead time of actor 2

Parameters

a = market potential	s = service level set by the chain
b_1 = price sensitivity of demand	μ_1 = mean service rate (production capacity) of actor 1
b_2 = lead time sensitivity of demand	μ_2 = mean service rate (production capacity) of actor 2
m_1 = unit direct variable cost for actor 1	W_1 = Total waiting time in actor 1 system
m_2 = unit direct variable cost for actor 2	W_2 = Total waiting time in actor 2 system
δ_2 = margin price set by the actor 2	

We do analyses for centralized, and two decentralized settings decisions. In centralized setting, both actors coordinate to decide the global price (P_g) and the global lead time (L_g). In decentralized setting, we consider the downstream actor as a leader and the upstream actor as a follower. Both actors are assumed to know the reaction of the other actor when decision are taken. In the first decentralized setting, the downstream actor decides the global lead time ($L_g = L_1 + L_2$) which takes into account the reaction of the upstream actor, and upstream actor decides his own price (P_1). The global price (P_g) is fixed by $P_g = P_1 + \delta_2$. In the second decentralized setting the downstream actor decides the upstream actor's price (P_1) and his lead time (L_2) (taking into account the reaction of upstream actor), and the upstream actor acts as a follower who decides his own lead time (L_1).

5.2 Centralized Setting: Model and Experiments

We start our analysis with the centralized setting. In this centralized setting, we consider downstream and upstream actors who work together to decide the global price (P_g) and global lead time (L_g). We model the centralized problem as follows:

$$\underset{P_g, L_g \geq 0}{\text{Maximize}} \ (P_g - m_1 - m_2)\lambda \tag{5.1}$$

$$\text{Subject to } \lambda = a - b_1 P_g - b_2 L_g \tag{5.2}$$

$$\Pr(W_1 + W_2 \leq L_g) \geq s \tag{5.3}$$

$$0 \leq \lambda < \mu_1, \mu_2 \tag{5.4}$$

When $\mu_1 \neq \mu_2$, the actual lead time $(W_1 + W_2)$ will follow a hypo-exponential distribution (Bolch et al., 2006). Thus, the service constraint (5.3) will be expressed as below:

$$1 - \frac{\mu_2 - \lambda}{(\mu_2 - \lambda) - (\mu_1 - \lambda)} e^{-(\mu_1 - \lambda)L_g} + \frac{\mu_1 - \lambda}{(\mu_2 - \lambda) - (\mu_1 - \lambda)} e^{-(\mu_2 - \lambda)L_g} \geq s$$

$$\Leftrightarrow 1 - \frac{\mu_2 - \lambda}{\mu_2 - \mu_1} e^{-(\mu_1 - \lambda)L_g} + \frac{(\mu_1 - \lambda)}{(\mu_2 - \mu_1)} e^{-(\mu_2 - \lambda)L_g} \geq s$$

However, if the upstream actor has exactly same capacity as the downstream actor $(\mu_1 = \mu_2 = \mu)$(Bolch et al., 2006), the service constraint (5.3) becomes:

$$1 - e^{-(\mu - \lambda)L_g} - e^{-(\mu - \lambda)L_g}(\mu - \lambda)L_g \geq s$$

The overall formulation for both problem with $\mu_1 \neq \mu_2$ and $\mu_1 = \mu_2$ can be rewritten as:

$$\underset{P_g, L_g \geq 0}{\text{Maximize}} \ \Pi_c(P_g, L_g) = (P_g - m_1 - m_2)(a - b_1 P_g - b_2 L_g) \tag{5.5}$$

$$\text{Subject to } \Pr(W_1 + W_2 \leq L_g) \geq s \Leftrightarrow \begin{cases} 1 - \frac{\mu_2 - a + b_1 P_g + b_2 L_g}{\mu_2 - \mu_1} e^{-(\mu_1 - a + b_1 P_g + b_2 L_g)L_g} \\ \quad + \frac{\mu_1 - a + b_1 P_g + b_2 L_g}{\mu_2 - \mu_1} e^{-(\mu_2 - a + b_1 P_g + b_2 L_g)L_g} \geq s, \\ \quad \text{for } \mu_1 \neq \mu_2 \\ 1 - e^{-(\mu - a + b_1 P_g + b_2 L_g)L_g} - e^{-(\mu - a + b_1 P_g + b_2 L_g)L_g} \\ \quad (\mu - a + b_1 P_g + b_2 L_g)L_g \geq s, \\ \quad \text{for } \mu_1 = \mu_2 \end{cases} \tag{5.6}$$

$$0 \leq a - b_1 P_g - b_2 L_g < \mu_1, \mu_2 \tag{5.7}$$

Then, we deduce several lemmas.

Lemma 5.1. *The profit $\Pi_c(P_g, L_g)$ is concave for any given L_g and P_g.*

Proof. The second derivative of $\Pi_c(P_g, L_g)$ is $\frac{\partial^2}{\partial {P_g}^2}\Pi_c(P_g, L_g) = -2b_1$ which proves that the function $\Pi_c(P_g, L_g)$ is concave in P_g for given L_g. The second derivative of $\Pi_c(P_g, L_g)$ in function of L_g is null ($\frac{\partial^2}{\partial {L_g}^2}\Pi_c(P_g, L_g) = 0$) and $\frac{\partial^2}{\partial P_g \partial L_g}\Pi_c(P_g, L_g) = -b_2$. Hence, the function $\Pi_c(P_g, L_g)$ is concave given that $\frac{\partial^2 \Pi_c(P_g,L_g)}{\partial {P_g}^2}\frac{\partial^2 \Pi_c(P_g,L_g)}{\partial {L_g}^2} - \left[\frac{\partial^2 \Pi_c(P_g,L_g)}{\partial P_g \partial L_g}\right]^2 = {b_2}^2 \geq 0$, $\frac{\partial^2 \Pi_c(P_g,L_g)}{\partial {P_g}^2} \leq 0$, and $\frac{\partial^2 \Pi_c(P_g,L_g)}{\partial {L_g}^2} \leq 0$ (see Hillier and Lieberman, 2001 on how to prove convexity or concavity of functions with several variables) ■

Lemma 5.2. *service constraint (5.6), for both $\mu_1 \neq \mu_2$ and $\mu_1 = \mu_2$, is binding at optimality.*

Proof. Let's assume that at optimality we have ${P_g}^*$ and ${L_g}^*$ which give a service level superior to s ($\Pr(W_1 + W_2 \leq L_g) > s$). We have profit of the firm as $\Pi_c({P_g}^*, {L_g}^*)$. If we decrease the ${L_g}^*$ to ${L_g}'$ (while keeping P_g constant) until $\Pr(W_1 + W_2 \leq L_g) = s$, we will get $\Pi_c({P_g}^*, {L_g}^*) < \Pi_c({P_g}^*, {L_g}')$ since the demand has increased. We have a better solution which is ${P_g}^*$ and ${L_g}'$. Thus, service level is binding at optimality. This condition is true with $\mu_1 = \mu_2$ and $\mu_1 \neq \mu_2$. ■

From lemma 5.2 and constraint (5.6), we can deduce a link between P_g and L_g. For each value of P_g, we can obtain a unique value of L_g. Thus the initial problem become a single optimization problem in function of P_g. Then, we can find the solution for profit Π_c numerically as proposed in proposition below.

Proposition 5.1. *The following single optimization problem (centralized setting):*

$$\underset{P_g \geq 0}{Maximize}\ \Pi_c(P_g) = (P_g - m_1 - m_2)(a - b_1 P_g - b_2 L_g(P_g))$$

can be solved with bisection method with the following steps:

1. *We fixed a value P_g within the range of $(m_1 + m_2) \leq P_g \leq \frac{a}{b_1}$,*
2. *We obtain a unique solution of L_g in function of P_g from:*

$$g(P_g, L_g) = \begin{cases} 1 - s - \frac{\mu_2 - a + b_1 P_g + b_2 L_g}{\mu_2 - \mu_1} e^{-(\mu_1 - a + b_1 P_g + b_2 L_g)L_g} \\ \quad + \frac{\mu_1 - a + b_1 P_g + b_2 L_g}{\mu_2 - \mu_1} e^{-(\mu_2 - a + b_1 P_g + b_2 L_g)L_g} = 0 & \text{for } \mu_1 \neq \mu_2 \\ 1 - s - e^{-(\mu - a + b_1 P_g + b_2 L_g)L_g} - e^{-(\mu - a + b_1 P_g + b_2 L_g)L_g} \\ \quad (\mu - a + b_1 P_g + b_2 L_g)L_g = 0 & \text{for } \mu_1 = \mu_2 \end{cases}$$

   with the range of $\frac{a - P_g b_1 - \min[\mu_1, \mu_2]}{b_2} < L_g \leq \frac{a - b_1 P_g}{b_2}$.

3. *With the obtained L_g, we can calculate $\Pi_c(P_g)$ and use the bisection method procedures.*

Proof. To obtain a positive profit, we must have $P_g - m_1 - m_2 \geq 0 \Leftrightarrow P_g \geq (m_1 + m_2)$ and $a - b_1 P_g - b_2 L_g \geq 0 \Leftrightarrow P_g \leq \frac{a}{b_1}$ for $L_g = 0$.

From lemma 5.2, the constraint (5.6) becomes equation $g(P_g, L_g)$ which links P_g and L_g. The $g(P_g, L_g)$ increases when L_g increases. Given that $\lim_{L_g \to 0} g(P_g, L_g) = -s$ and $\lim_{L_g \to +\infty} g(P_g, L_g) = 1 - s \geq 0$, we can deduce that for a given value of P_g there is a unique value of L_g that give $g(P_g, L_g) = 0$. The obtained L_g must respect constraint (5.7), thus $0 \leq a - b_1 P_g - b_2 L_g \Leftrightarrow L_g \leq \frac{a - b_1 P_g}{b_2}$ and $a - b_1 P_g - b_2 L_g < \min[\mu_1, \mu_2] \Leftrightarrow L_g > \frac{a - P_g b_1 - \min[\mu_1, \mu_2]}{b_2}$.

After we obtain the L_g for a given P_g, we can calculate the profit as:

$$\underset{P_g \geq 0}{\text{Maximize}}\ \Pi_c(P_g) = (P_g - m_1 - m_2)(a - b_1 P_g - b_2 L_g(P_g))$$

The initial problem $\Pi_c(P_g, L_g)$ is concave in P_g, thus we can use bisection method to solve the problem. ■

Experiments

Using the bisection method, we do sensitivity analysis on parameters b_1 and b_2 while considering $\mu_1 = \mu_2$ and $\mu_1 \neq \mu_2$. The result are presented in tables 5.1 - 5.4.

We start with the case $\mu_1 = \mu_2$, we use $a = 50$, $b_1 = 4$, $b_2 = 6$, $m_1 = 3$, $m_2 = 2$, $s = 0.95$, $\mu_2 = 20$, and $\mu_1 = 20$. Table 5.1 reports the result of experimentation on b_1. As expected, when b_1 increases, the price will decrease as the customers are more sensitive to price.

Table 5.1: Centralized setting experiment on b_1 with $\mu_1 = \mu_2$

b_1	Centralized Setting			
	L_g	P_g	λ	Profit (Π_c)
1	0.8121	30.9688	14.1586	367.6809
2	0.7212	16.1254	13.4221	149.3263
3	0.6304	11.2478	12.4743	77.9376
4	0.5432	8.8683	11.2675	43.5856
5	0.4640	7.4878	9.7767	24.3228
6	0.3959	6.6012	8.0176	12.8375
7	0.3398	5.9888	6.0395	5.9719
8	0.2947	5.5413	3.9016	2.1119
9	0.2586	5.1994	1.6543	0.3298

Table 5.2 reports the result of varying b_2 with $\mu_1 = \mu_2$. As b_2 increases, customer are more sensitive to quoted lead time. Thus, the chain has no other option than reducing the quoted lead time.

Table 5.2: Centralized setting experiment on b_2 with $\mu_1 = \mu_2$

b_2	Centralized Setting			
	L_g	P_g	λ	Profit (Π_c)
2	0.6817	8.8989	13.0409	50.8456
4	0.5945	8.9005	12.0201	46.8844
6	0.5432	8.8683	11.2675	43.5856
8	0.5076	8.8211	10.6547	40.7125
10	0.4806	8.7658	10.1302	38.1487
12	0.4591	8.7057	9.6676	35.8255
14	0.4413	8.6424	9.2514	33.6976
16	0.4263	8.5771	8.8712	31.7334
18	0.4132	8.5103	8.5204	29.9092
20	0.4018	8.4426	8.1935	28.2071

For case $\mu_1 \neq \mu_2$: we use $a = 50$, $b_1 = 4$, $b_2 = 6$, $m_1 = 3$, $m_2 = 2$, and $s = 0.95$. We use $\mu_1 = 30$ & $\mu_2 = 15$ for case where $\mu_1 > \mu_2$ and $\mu_1 = 15$ & $\mu_2 = 30$ for $\mu_1 < \mu_2$. We vary b_1 in table 5.3 and b_2 in 5.4. We found that even though both cases $\mu_1 = \mu_2$ and $\mu_1 \neq \mu_2$ have $\frac{1}{\mu_1} + \frac{1}{\mu_2} = \frac{1}{10}$, the lead times obtained aren't the same. We found that changing the capacity, from $\mu_1 = 15$ & $\mu_2 = 30$ to $\mu_1 = 30$ & $\mu_2 = 15$, leads to the same value of L_g and P_g.

From table 5.3, case $\mu_1 \neq \mu_2$ behaves in the same way as case $\mu_1 = \mu_2$. When b_1 increases, customers are more sensitive to price, and naturally firm has to decrease price. We also observed that lead time decreases to compensate at least partially the loss of demand due to the increase of b_1.

Table 5.3: Centralized setting experiment on b_1 with $\mu_1 \neq \mu_2$

b_1	Centralized Setting			
	L_g	P_g	λ	Profit (Π_c)
1	0.9077	33.0877	11.4664	322.0649
2	0.8204	17.0065	11.0645	132.8462
3	0.7285	11.6998	10.5297	70.5467
4	0.6333	9.1001	9.7997	40.1800
5	0.5385	7.5948	8.7946	22.8205
6	0.4505	6.6418	7.4462	12.2250
7	0.3756	5.9997	5.7484	5.7468
8	0.3162	5.5416	3.7697	2.0419
9	0.2706	5.1971	1.6026	0.3158

Table 5.4 reports the result of varying b_2 with $\mu_1 \neq \mu_2$. As b_2 increases, obviously, clients are more sensitive to quoted lead time, and therefore the firm has to decrease quoted lead time to capture more demand.

Table 5.4: Centralized setting experiment on b_2 with $\mu_1 \neq \mu_2$

b_2	Centralized Setting			
	L_g	P_g	λ	Profit (Π_c)
2	0.8962	9.1973	11.4183	47.9264
4	0.7196	9.1628	10.4705	43.5859
6	0.6333	9.1001	9.7997	40.1800
8	0.5785	9.0275	9.2619	37.3022
10	0.5394	8.9504	8.8049	34.7828
12	0.5094	8.8711	8.4034	32.5300
14	0.4853	8.7908	8.0425	30.4873
16	0.4654	8.7101	7.7131	28.6163
18	0.4485	8.6294	7.4089	26.8898
20	0.4340	8.5489	7.1254	25.2873

5.3 Local & Global Service Level

Solving analytically the centralized setting with hypo-exponential distribution (global service constraint) is very difficult, because of the very complex equation of the global service constraint (eq. 5.6). Thus, we come out with an idea to decouple the global service constraint into local service constraints. Furthermore in numerous situations, the two actors (upstream and downstream) correspond to two different companies which want to take their own decision. However, this cause a problem: how to guarantee that respecting the local service constraints will also allow to respect the global service constraint?

In this section, we are going to demonstrate, analytically or numerically according to cases, that the global service constraint is respected when each actor satisfy their service constraint for all value of $s \geq 0.715$. We start by proposing lemma 5.3.

Lemma 5.3. *If we have s such that $f(s) \geq 0$ where:*

$$f(s) = \begin{cases} s - 2(1-s)\ln\left(\frac{1}{1-s}\right) \geq 0 & \text{for } \mu_1 = \mu_2 \\ \frac{V_2}{V_1}\left(1-(1-s)^{\frac{V_1}{V_2}}\right) - \left(1-(1-s)^{\frac{V_2}{V_1}}\right) \geq 0 & \text{for } \mu_1 < \mu_2 \\ \frac{V_1}{V_2}\left(1-(1-s)^{\frac{V_2}{V_1}}\right) - \left(1-(1-s)^{\frac{V_1}{V_2}}\right) \geq 0 & \text{for } \mu_1 > \mu_2 \end{cases}$$

where $V_2 = \mu_2 - \lambda$ and $V_1 = \mu_1 - \lambda$, thus $[\Pr(W_1 \leq L_1) \geq s]$ *and* $[\Pr(W_2 \leq L_2) \geq s]$ *implies* $[\Pr(W_1 + W_2 \leq L_g) \geq s]$.

Proof. **Case 1:** $\mu_1 = \mu_2$. We start by recalling the formulation of the global service level. Given that $L_g = L_1 + L_2$, the global service constraint ($\Pr(W_1 + W_2 \leq L_g) \geq s$) where $\mu_1 = \mu_2$ is:

$$1 - e^{-(\mu-\lambda)(L_1+L_2)} - e^{-(\mu-\lambda)(L_1+L_2)}(\mu-\lambda)(L_1+L_2) \geq s$$

We have the local service constraints which are:
$\Pr(W_1 \leq L_1) \geq s \Leftrightarrow e^{-(\mu_1-\lambda)L_1} \leq 1-s$ and $\Pr(W_2 \leq L_2) \geq s \Leftrightarrow e^{-(\mu_2-\lambda)L_2} \leq 1-s$. We consider calculations with $1-s = e^{-(\mu_1-\lambda)L_1}$ and $1-s = e^{-(\mu_1-\lambda)L_2}$. And obviously if we prove that in this case the global service constraint is satisfied, it will be also verified when $1-s > e^{-(\mu_1-\lambda)L_1}$ and/or $1-s > e^{-(\mu_1-\lambda)L_2}$. The global service level equation is then equivalent to:

$$1 - e^{-(\mu-\lambda)(L_1+L_2)} - e^{-(\mu-\lambda)(L_1+L_2)}(\mu-\lambda)(L_1+L_2) = \\ 1-(1-s)^2-(\mu-\lambda)(L_1+L_2)(1-s)^2 \geq s$$

With $s \neq 1$, we can simplify the equation above into:

$$s-(\mu-\lambda)(L_1+L_2)(1-s) \geq 0$$

From $e^{-(\mu_1-\lambda)L_1} = 1-s$ and $e^{-(\mu_2-\lambda)L_2} = 1-s$, we can deduce:

$$(\mu-\lambda)(L_1+L_2) = 2\ln\left(\frac{1}{1-s}\right)$$

and therefore:

$$s-(\mu-\lambda)(L_1+L_2)(1-s) = s-2(1-s)\ln\left(\frac{1}{1-s}\right) \geq 0$$

Therefore we have proven that if $f(s) = s-2(1-s)\ln\left(\frac{1}{1-s}\right) \geq 0$ the global service level is respected.

Case 2: $\mu_1 \neq \mu_2$. The global service constraint ($\Pr(W_1+W_2 \leq L_g) \geq s$) where $\mu_2-\mu_1 \neq 0$ is:

$$1 - \frac{\mu_2-\lambda}{\mu_2-\mu_1}e^{-(\mu_1-\lambda)(L_1+L_2)} + \frac{\mu_1-\lambda}{\mu_2-\mu_1}e^{-(\mu_2-\lambda)(L_1+L_2)} \geq s \\ \Leftrightarrow 1 - \frac{\mu_2-\lambda}{\mu_2-\mu_1}e^{-(\mu_1-\lambda)L_1}e^{-(\mu_1-\lambda)L_2} + \frac{\mu_1-\lambda}{\mu_2-\mu_1}e^{-(\mu_2-\lambda)L_1}e^{-(\mu_2-\lambda)L_2} \geq s$$

Given that the local service constraints are $e^{-(\mu_1-\lambda)L_1} \leq 1-s$ and $e^{-(\mu_2-\lambda)L_2} \leq 1-s$, we use the same logic as for case $\mu_1 = \mu_2$: we consider the case in which local service constraints are binding ($e^{-(\mu_1-\lambda)L_1} = 1-s$ and $e^{-(\mu_2-\lambda)L_2} = 1-s$). The obtained condition in this case will be obviously a sufficient condition when we will have $e^{-(\mu_1-\lambda)L_1} < 1-s$ and/or $e^{-(\mu_2-\lambda)L_2} < 1-s$. Thus the global service constraint is equivalent to:

$$1 - \frac{\mu_2-\lambda}{\mu_2-\mu_1}e^{-(\mu_1-\lambda)L_1}e^{-(\mu_1-\lambda)L_2} + \frac{\mu_1-\lambda}{\mu_2-\mu_1}e^{-(\mu_2-\lambda)L_1}e^{-(\mu_2-\lambda)L_2} = \\ 1 - \frac{1-s}{\mu_2-\mu_1}\left[(\mu_2-\lambda)e^{-(\mu_1-\lambda)L_2} + (\mu_1-\lambda)e^{-(\mu_2-\lambda)L_1}\right] \geq s$$

Here, we have two sub-cases: $\mu_2 - \mu_1 > 0$ and $\mu_2 - \mu_1 < 0$.
Sub-case 2.1: $\mu_2 - \mu_1 > 0$. We derive from previous inequality:

$$1 - \frac{1-s}{\mu_2 - \mu_1}\left[(\mu_2 - \lambda)e^{-(\mu_1-\lambda)L_2} + (\mu_1 - \lambda)e^{-(\mu_2-\lambda)L_1}\right] \geq s$$
$$\Leftrightarrow \mu_2 - \mu_1 \geq (\mu_2 - \lambda)e^{-(\mu_1-\lambda)L_2} + (\mu_1 - \lambda)e^{-(\mu_2-\lambda)L_1}$$

Recall that we consider $e^{-(\mu_1-\lambda)L_1} = 1-s$ which is equivalent to $L_1 = \frac{\ln(1/(1-s))}{\mu_1-\lambda}$ and $e^{-(\mu_2-\lambda)L_2} = 1-s$ which is equivalent to $L_2 = \frac{\ln(1/(1-s))}{\mu_2-\lambda}$. Thus we obtain:

$$\mu_2 - \mu_1 - (\mu_2 - \lambda)e^{-(\mu_1-\lambda)\frac{\ln(1/(1-s))}{\mu_2-\lambda}} - (\mu_1 - \lambda)e^{-(\mu_2-\lambda)\frac{\ln(1/(1-s))}{\mu_1-\lambda}} =$$
$$\mu_2 - \mu_1 - (\mu_2 - \lambda)(1-s)^{\frac{\mu_1-\lambda}{\mu_2-\lambda}} - (\mu_1 - \lambda)(1-s)^{\frac{\mu_2-\lambda}{\mu_1-\lambda}} \geq 0$$

Replace $(\mu_2 - \lambda) = V_2$ and $(\mu_1 - \lambda) = V_1$, we obtain:

$$V_2 - V_1 - V_2(1-s)^{\frac{V_1}{V_2}} - V_1(1-s)^{\frac{V_2}{V_1}} \geq 0$$
$$\Leftrightarrow f(s) = \frac{V_2}{V_1}\left(1 - (1-s)^{\frac{V_1}{V_2}}\right) - \left(1 - (1-s)^{\frac{V_2}{V_1}}\right) \geq 0$$

Thus, if $f(s) \geq 0$ implies $\Pr(W_1 + W_2 \leq L_g) \geq s$ for $\mu_2 > \mu_1$.
Sub-case 2.2: $\mu_2 - \mu_1 < 0$. Using the same procedure as sub-case 2.1, we will have:

$$f(s) = \frac{V_1}{V_2}\left(1 - (1-s)^{\frac{V_2}{V_1}}\right) - \left(1 - (1-s)^{\frac{V_1}{V_2}}\right) \geq 0$$

Thus, if $f(s) \geq 0$ implies $\Pr(W_1 + W_2 \leq L_g) \geq s$ for $\mu_1 > \mu_2$. ■

Note that the expressions of $f(s)$ in sub-case 2.1 and 2.2 (in the proof above) are symmetric. Indeed, suppose that we have (μ, μ') with $\mu < \mu'$. If $\mu_1 = \mu$ and $\mu_2 = \mu'$, we use the expression of $f(s)$ in sub-case 2.1 and obtain a value of s such that $f(s) = 0$ called s_1. If $\mu_1 = \mu'$ and $\mu_2 = \mu$, we use the expression of $f(s)$ in sub-case 2.2 and obtain a value of s such that $f(s) = 0$ called s_2. It is obvious, from the expression of $f(s)$ in both sub-cases, that the obtained value of s is equal $(s_1 = s_2)$.

From lemma 5.3, we can deduce following corollary.

Corollary 5.1. *For $s \geq 0.715$, satisfying the local service constraints allows to satisfy the global service constraint.*

Proof. For $\mu_1 = \mu_2$, the condition to verify is: $f(s) = s - 2(1-s)\ln\left(\frac{1}{1-s}\right)$; we have $\frac{d}{ds^2}f(s) = \frac{2}{1-s} > 0$. This proves that $f(s)$ is convex for $\mu_1 = \mu_2$.
Furthermore $\frac{d}{ds}f(s) = 0$ will result in $s_0 = 1 - e^{-\frac{1}{2}}$ and $f(s_0) = 1 - 2e^{-\frac{1}{2}} < 0$. In addition, $\lim_{s\to 0} s - 2(1-s)\ln\left(\frac{1}{1-s}\right) = 0$ and $\lim_{s\to 1} s - 2(1-s)\ln\left(\frac{1}{1-s}\right) = 1$. Therefore,

we are sure that there is only one solution to $f(s) = s - 2(1-s)\ln\left(\frac{1}{1-s}\right) = 0$. We can find numerically the exact value of $f(s) = 0$; we draw $f(s)$ in figure 5.2. We found numerically that $f(s) \geq 0$ for $s \geq 0.715$.

For $\mu_1 < \mu_2$, $f(s) = \frac{V_2}{V_1}\left(1-(1-s)^{\frac{V_1}{V_2}}\right) - \left(1-(1-s)^{\frac{V_2}{V_1}}\right)$, we will have $\frac{d}{ds^2}f(s) = \frac{(V_2-V_1)\left({V_1}^2(1-s)^{\frac{V_1}{V_2}} + {V_2}^2(1-s)^{\frac{V_2}{V_1}}\right)}{{V_1}^2V_2(s-1)^2} > 0$ for $0 \leq s < 1$ given that $V_1, V_2 > 0$ which means that the curve is convex. Furthermore $\frac{d}{ds}f(s) = \frac{V_1(1-s)^{\frac{V_1}{V_2}} - V_2(1-s)^{\frac{V_2}{V_1}}}{V_1(1-s)} = 0 \Leftrightarrow s = 1 - \left(\frac{V_2}{V_1}\right)^{\frac{V_1V_2}{{V_1}^2-{V_2}^2}}$. In addition $\lim_{s\to 0}\frac{d}{ds}f(s) = \frac{V_1-V_2}{V_1} < 0$, $\lim_{s\to 0}\frac{V_2}{V_1}\left(1-(1-s)^{\frac{V_1}{V_2}}\right) - \left(1-(1-s)^{\frac{V_2}{V_1}}\right) = 0$ and $\lim_{s\to 1}\frac{V_2}{V_1}\left(1-(1-s)^{\frac{V_1}{V_2}}\right) - \left(1-(1-s)^{\frac{V_2}{V_1}}\right) = \frac{V_2}{V_1} - 1 > 0$, we are sure that there is only single solution of $f(s) = \frac{V_2}{V_1}\left(1-(1-s)^{\frac{V_1}{V_2}}\right) - \left(1-(1-s)^{\frac{V_2}{V_1}}\right) = 0$ for $0 \leq s < 1$.

When $\mu_1 > \mu_2$ we have the same demonstration procedure.

Then we can find numerically the solution of $f(s) = 0$. Let us give the two following cases: (1) with $V_1 = 1$ and $V_2 = 2$, and (2) with $V_1 = 1$ and $V_2 = 10$ (see Figure 5.2). We can observe that for these two curves $f(s) \geq 0$ for $s \geq 0.715$. We also solve numerically $f(s) = 0$ for different value of $\frac{V_1}{V_2}$ (reported in figure 5.3).

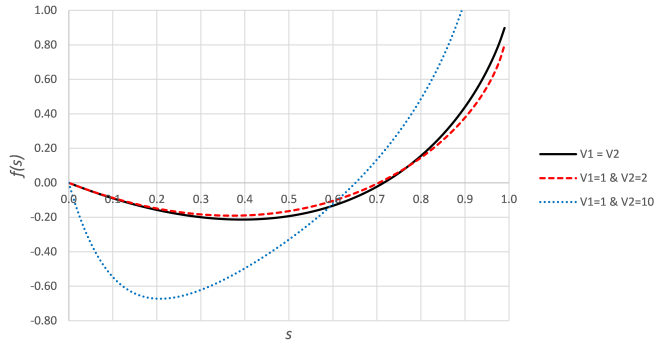


Figure 5.2: $f(s)$ in function of s

■

The result of corollary 5.1 allows us to simplify the global service constraint into local service constraints as seen as conditions given in this corollary are verified. This also allows us to model the problem as a modified centralized setting where we replace the global service constraint with local service constraints.

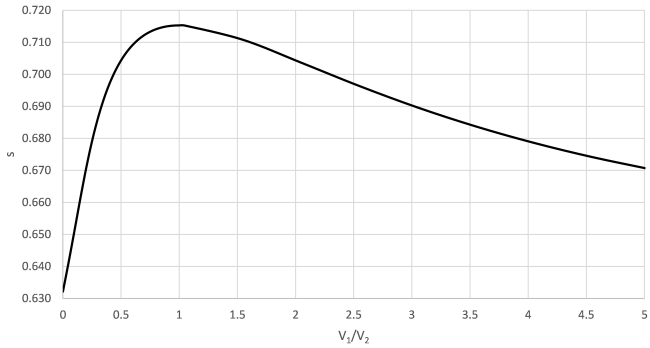


Figure 5.3: s in function of V_1/V_2

5.4 Modified Centralized: Model and Experiments

As shown in section 5.2, the global service constraint with hypo-exponential distribution is hard to solve. In this section, we proposed a new model, where we transform the global service constraint into local service constraints for each actor. We call this model "Modified Centralized". We formulate the modified centralized model as follows.

$$\underset{P_g, L_g \geq 0}{\text{Maximize}} \ (P_g - m_1 - m_2)\lambda \tag{5.8}$$

$$\text{Subject to } \lambda = a - b_1 P_g - b_2 L_g \tag{5.9}$$

$$\Pr(W_1 \leq L_1) \geq s \tag{5.10}$$

$$\Pr(W_2 \leq L_2) \geq s \tag{5.11}$$

$$0 \leq \lambda < \mu_1, \mu_2 \tag{5.12}$$

$$\text{where } L_g = L_1 + L_2 \tag{5.13}$$

Given that $\lambda = a - b_1 P_g - b_2 L_g \Leftrightarrow P_g = \frac{a-(L_1+L_2)b_2-\lambda}{b_1}$, the problem can be rewritten as:

$$\underset{\lambda, L_1, L_2 \geq 0}{\text{Maximize}} \ \Pi_m(\lambda, L_1, L_2) = \left(\frac{a - (L_1 + L_2)b_2 - \lambda}{b_1} - m_1 - m_2 \right) \lambda \tag{5.14}$$

$$\text{Subject to } 1 - e^{-(\mu_1 - \lambda)L_1} \geq s \tag{5.15}$$

$$1 - e^{-(\mu_2 - \lambda)L_2} \geq s \tag{5.16}$$

$$\lambda < \mu_1, \mu_2 \tag{5.17}$$

$$\text{where } P_g = P_1 + \delta_2 \tag{5.18}$$

To solve the problem above, we first derive the following lemma.

Lemma 5.4. *Both local service constraints (eq.5.15 & eq.5.16) are binding.*

Proof. Let's assume that at optimality we have λ^*, ${L_1}^*$, and ${L_2}^*$ which give a service level superior to s ($\Pr(W_1 \leq L_1) > s$. We have profit of the firm as $\Pi_m(\lambda^*, {L_1}^*, {L_2}^*)$.

If we decrease the ${L_1}^*$ to ${L_1}'$ (by keeping the L_2 and λ constant) until $\Pr(W_1 \leq L_1) = s$, we will get $\Pi_m(\lambda^*, {L_1}^*, {L_2}^*) < \Pi_m(\lambda^*, {L_1}', {L_2}^*)$ because price has increased $\left(P_g = \frac{a - b_2(L_1+L_2) - \lambda}{b_1}\right)$. We have a better solution which is λ^*, ${L_1}'$, and ${L_2}^*$. Thus, service level $\Pr(W_1 \leq L_1)$ is binding at optimality. Using the same reasoning, we can prove $\Pr(W_2 \leq L_2) = s$ at optimality by decreasing ${L_2}^*$ to ${L_2}'$ and keeping the L_1 and λ constant. ■

From lemma 5.4, we get $1 - e^{-(\mu_1 - \lambda)L_1} = s \Leftrightarrow L_1 = \frac{\ln(1-s)}{\lambda - \mu_1}$ and $1 - e^{-(\mu_2 - \lambda)L_2} = s \Leftrightarrow L_2 = \frac{\ln(1-s)}{\lambda - \mu_2}$. We deduce $P_g = \frac{a - (L_1+L_2)b_2 - \lambda}{b_1}$ from the expression of demand (eq.(5.9)). Then, we can transform our model into single variable optimization.

$$\underset{0 \leq \lambda \leq \min[\mu_1, \mu_2]}{\text{Maximize}} \ \Pi_m(\lambda) = \left(\frac{a - \left(\frac{\ln(1-s)}{\lambda - \mu_1} + \frac{\ln(1-s)}{\lambda - \mu_2} \right) b_2 - \lambda}{b_1} - m_1 - m_2 \right) \lambda \qquad (5.19)$$

The problem above can solved using the following proposition.

Proposition 5.2. *The solution of the modified centralized model is:*

- *For $\mu_1 \neq \mu_2$, candidates for the optimum demand are the roots of the quintic equation below:*

$$\lambda b_2 \left(\frac{\ln(1-s)}{(\lambda - \mu_1)^2} + \frac{\ln(1-s)}{(\lambda - \mu_2)^2} \right) - b_2 \left(\frac{\ln(1-s)}{\lambda - \mu_1} + \frac{\ln(1-s)}{\lambda - \mu_2} \right) - 2\lambda + a - (m_1 + m_2)b_1 = 0$$

- *For $\mu_1 = \mu_2 = \mu$, candidates for the optimum demand are the roots of the cubic equation below:*

$$2\mu b_2 \ln(1-s) - (2\lambda - a + (m_1 + m_2)b_1)(\lambda - \mu)^2 = 0$$

The optimum demand is the root which gives a maximum $\Pi_m(\lambda)$ in regard to $0 \leq \lambda \leq \min[\mu_1, \mu_2]$; lead time for each actor can be deduced from $L_1 = \frac{\ln(1-s)}{\lambda - \mu_1}$ and $L_2 = \frac{\ln(1-s)}{\lambda - \mu_2}$; and price can be deduced from $P_g = \frac{a - (L_1+L_2)b_2 - \lambda}{b_1}$.

Proof. The optimal solution must satisfy $\frac{d}{d\lambda}\Pi_m(\lambda)$ equals zero:

$$\begin{aligned}
&\frac{d}{d\lambda}\Pi_m(\lambda) = 0 \\
&\quad \Leftrightarrow \frac{\lambda\left(b_2\left(\frac{\ln(1-s)}{(\lambda-\mu_1)^2} + \frac{\ln(1-s)}{(\lambda-\mu_2)^2}\right) - 1\right)}{b_1} - \frac{\lambda - a + b_2\left(\frac{\ln(1-s)}{\lambda-\mu_1} + \frac{\ln(1-s)}{\lambda-\mu_2}\right)}{b_1} - m_1 - m_2 = 0 \\
&\quad \Leftrightarrow \lambda b_2 \left(\frac{\ln(1-s)}{(\lambda - \mu_1)^2} + \frac{\ln(1-s)}{(\lambda - \mu_2)^2} \right) - b_2 \left(\frac{\ln(1-s)}{\lambda - \mu_1} + \frac{\ln(1-s)}{\lambda - \mu_2} \right) - 2\lambda + a - (m_1 + m_2)b_1 = 0
\end{aligned}$$

For the case $\mu_1 = \mu_2 = \mu$, the quintic equation above becomes:

$$\begin{aligned}
&\lambda b_2 \left(\tfrac{\ln(1-s)}{(\lambda-\mu_1)^2} + \tfrac{\ln(1-s)}{(\lambda-\mu_2)^2}\right) - b_2 \left(\tfrac{\ln(1-s)}{\lambda-\mu_1} + \tfrac{\ln(1-s)}{\lambda-\mu_2}\right) - 2\lambda + a - (m_1+m_2)b_1 = 0 \\
&\quad \Leftrightarrow \lambda b_2 \left(\tfrac{2\ln(1-s)}{(\lambda-\mu)^2}\right) - b_2 \left(\tfrac{2\ln(1-s)}{\lambda-\mu}\right) - 2\lambda + a - (m_1+m_2)b_1 = 0 \\
&\quad \Leftrightarrow \left[\lambda b_2 \left(\tfrac{2\ln(1-s)}{(\lambda-\mu)^2}\right) - b_2 \left(\tfrac{2\ln(1-s)}{\lambda-\mu}\right) - 2\lambda + a - (m_1+m_2)b_1\right](\lambda-\mu)^2 = 0 \\
&\quad \Leftrightarrow 2\mu b_2 \ln(1-s) - (2\lambda - a + (m_1+m_2)b_1)(\lambda-\mu)^2 = 0
\end{aligned}$$

The optimum demand can be found by comparing the profit $\Pi_m(\lambda)$ (see eq.(5.19)) of each root which respect $0 \leq \lambda \leq \min[\mu_1, \mu_2]$. Let us note that we have not encountered the case where there are several feasible roots. But we still not yet prove that it can not happen.
The expression of L_1 and L_2 can be deduced using lemma 5.4. And the expression of P_g can be found from the expression of demand (eq.(5.9)). ■

Experiments

We do some experiments using the bisection. The parameters explored are b_1, and b_2. We vary one parameter and fix other parameters. For μ_1 and μ_2, we define cases where $\mu_1 = \mu_2$ and $\mu_1 \neq \mu_2$.

Our first experiment in this section uses $\mu_1 = \mu_2 = 20$ and base parameters are: $a = 50$, $b_1 = 4$, $b_2 = 6$, $m_1 = 3$, $m_2 = 2$, and $s = 0.95$. It is obvious that when both capacity of the actors are the same, it will result the same waiting time. This leads to both actors have the same quoted lead time. This can be seen in our result (see Table 5.5 and 5.6). In table 5.5, as b_1 increases, the price decreases. In table 5.6, as b_2 increases the lead time decreases. These behaviors are similar to those observed in section 5.2.

Let us point the following interesting finding: the profit, that we get from the modified centralized setting, isn't too far from the profit of the centralized setting. We use $\frac{\Pi_c - \Pi_m}{\Pi_c} \times 100\%$ to calculate the loss of using this modified centralized setting compared to the centralized setting. For the experiment with b_1, we get an average loss of 6.66% and for experiment with b_2 we get an average loss of 8.41%.

Table 5.5: Experiment on b_1 with $\mu_1 = \mu_2$ for modified centralized setting

b_1	Modified Centralized Setting									
	L_1	L_2	P_g	λ	Π_m	Π_c	Loss	Serv. Lvl. Real.		
								Actor 1	Actor 2	Global
2	0.4213	0.4213	16.0273	12.8896	142.1375	149.3263	4.81%	95.00%	95.00%	98.25%
4	0.3248	0.3248	8.8319	10.7754	41.2901	43.5856	5.27%	95.00%	95.00%	98.25%
6	0.2425	0.2425	6.5742	7.6450	12.0349	12.8375	6.25%	95.00%	95.00%	98.25%
8	0.1833	0.1833	5.5183	3.6545	1.8940	2.1119	10.32%	95.00%	95.00%	98.25%

In the second experiment, we consider $\mu_1 \neq \mu_2$. We use $\mu_1 = 30$ & $\mu_2 = 15$ for $\mu_1 > \mu_2$ and vice-versa for $\mu_1 < \mu_2$. We only display the results where $\mu_1 > \mu_2$ (see

Table 5.6: Experiment on b_2 with $\mu_1 = \mu_2$ for modified centralized setting

b_2	Modified Centralized Setting									
	L_1	L_2	P_g	λ	Π_m	Π_c	Loss	Serv. Lvl. Real.		
								Actor 1	Actor 2	Global
2	0.4122	0.4122	8.9049	12.7317	49.7162	50.8456	2.22%	95.00%	95.00%	98.25%
4	0.3567	0.3567	8.8861	11.6019	45.0861	46.8844	3.84%	95.00%	95.00%	98.25%
6	0.3248	0.3248	8.8319	10.7754	41.2901	43.5856	5.27%	95.00%	95.00%	98.25%
8	0.3027	0.3027	8.7628	10.1048	38.0225	40.7125	6.61%	95.00%	95.00%	98.25%
10	0.2862	0.2862	8.6860	9.5321	35.1359	38.1487	7.90%	95.00%	95.00%	98.25%
12	0.2730	0.2730	8.6049	9.0278	32.5441	35.8255	9.16%	95.00%	95.00%	98.25%
14	0.2622	0.2622	8.5210	8.5745	30.1908	33.6976	10.41%	95.00%	95.00%	98.25%
16	0.2530	0.2530	8.4355	8.1608	28.0366	31.7334	11.65%	95.00%	95.00%	98.25%
18	0.2451	0.2451	8.3491	7.7790	26.0525	29.9092	12.89%	95.00%	95.00%	98.25%
20	0.2382	0.2382	8.2620	7.4237	24.2163	28.2071	14.15%	95.00%	95.00%	98.25%

table 5.7 and 5.8). If readers are interested in the case where $\mu_1 < \mu_2$, the readers can just inverse the value in column L_1 and L_2 (the price, demand and profit stay the same). Theoretically, the actor who has a higher capacity will have lower lead time. It is proven in our experiments as can be seen in tables 5.7 and 5.8. Using modified centralized setting, we can see the portion of lead time given for each actor. However, the $L_1 + L_2$, that we got here, is higher than the L_g in centralized setting. For the increase on price in table 5.7 as well as the increase on lead time in table 5.8, we observe the same behavior as for the experiments in section 5.2. We use $\frac{\Pi_c - \Pi_m}{\Pi_c} \times 100\%$ to calculate the loss. For the experiment with b_1, we get an average loss of 4.79% and for experiment with b_2 we get an average loss of 5.83%.

Table 5.7: Experiment on b_1 with $\mu_1 > \mu_2$ for modified centralized setting

b_1	Modified Centralized Setting									
	L_1	L_2	P_g	λ	Π_m	Π_c	Loss	Serv. Lvl. Real.		
								Actor 1	Actor 2	Global
2	0.1574	0.7428	16.8158	10.9671	129.5849	132.8462	2.45%	95.00%	95.00%	96.64%
4	0.1472	0.5595	9.0284	9.6461	38.8582	40.1800	3.29%	95.00%	95.00%	96.91%
6	0.1316	0.3861	6.6089	7.2405	11.6491	12.2250	4.71%	95.00%	95.00%	97.27%
8	0.1134	0.2623	5.5207	3.5801	1.8642	2.0419	8.70%	95.00%	95.00%	97.59%

In all instances, as expected, satisfying the local service constraints will also satisfy the global service constraint. We can see that in all instances, local service level of 95% will result in global service level greater than 95%. To support this argument, let us consider results obtained for centralized setting (table 5.1); then we calculate the local service constraints. Given that $\mu_1 = \mu_2$, thus $L_1 = L_2$ where $L_1 = L_g/2$. We can see in table 5.9 that satisfying the global service constraint doesn't lead to satisfy the local service constraints. This reinforces our argument that local service constraints are more compelling than global service constraint.

Table 5.8: Experiment on b_2 with $\mu_1 > \mu_2$ for modified centralized setting

b_2	Modified Centralized Setting									
	L_1	L_2	P_g	λ	Π_m	Π_c	Loss	Serv. Lvl. Real.		
								Actor 1	Actor 2	Global
2	0.1607	0.8222	9.1694	11.3565	47.3501	47.9264	1.20%	95.00%	95.00%	96.54%
4	0.1525	0.6456	9.1119	10.3599	42.5986	43.5859	2.27%	95.00%	95.00%	96.77%
6	0.1472	0.5595	9.0284	9.6461	38.8582	40.1800	3.29%	95.00%	95.00%	96.91%
8	0.1431	0.5051	8.9362	9.0693	35.6984	37.3022	4.30%	95.00%	95.00%	97.01%
10	0.1398	0.4663	8.8405	8.5762	32.9369	34.7828	5.31%	95.00%	95.00%	97.09%
12	0.1370	0.4367	8.7434	8.1408	30.4744	32.5300	6.32%	95.00%	95.00%	97.15%
14	0.1346	0.4131	8.6459	7.7482	28.2491	30.4873	7.34%	95.00%	95.00%	97.21%
16	0.1325	0.3936	8.5486	7.3886	26.2189	28.6163	8.38%	95.00%	95.00%	97.25%
18	0.1306	0.3771	8.4517	7.0556	24.3536	26.8898	9.43%	95.00%	95.00%	97.29%
20	0.1288	0.3629	8.3553	6.7446	22.6305	25.2873	10.51%	95.00%	95.00%	97.32%

Table 5.9: Verification of local service constraint for centralized setting

b_1	Centralized Setting with $\mu_1 = \mu_2$								
	L_g	L_1	L_2	P_g	λ	Π_c	Serv. Lvl. Real.		
							Global	Actor 1	Actor 2
2	0.7212	0.3606	0.3606	16.1254	13.4221	149.3263	95.00%	90.67%	90.67%
4	0.5432	0.2716	0.2716	8.8683	11.2675	43.5856	95.00%	90.67%	90.67%
6	0.3959	0.1980	0.1980	6.6012	8.0176	12.8375	95.00%	90.67%	90.67%
8	0.2947	0.1473	0.1473	5.5413	3.9016	2.1119	95.00%	90.67%	90.67%

5.5 Decentralized setting (Downstream Leader – Upstream Follower)

In lemma 5.3, we have provided the conditions such that respecting the local service constraints leads to respect the global service constraint. This will help us to address the decentralized setting. In this section, we consider a supply chain consisting of two actors where each of them undertakes decisions (price or/and lead time) to maximize its own profit. The sequence of decision is the following. The first actor (leader) will take his decision of price or lead time knowing the reaction of the second actor (follower), then the second actor (follower) will take a decision following the decision of the first actor. This type of decision making is often called as Stackelberg Game.

We divide this decentralized setting into two scenarios. In both scenarios, the downstream actor acts as a leader and the upstream actor acts as a follower. In the first scenario, the upstream actor has the right to choose his own lead time (L_1) but the price P_1 and lead time L_2 are decided by the downstream actor (actor 2). In the second scenario, the upstream actor (actor 1) decide his own price (P_1) and downstream actor (actor 2) decides the global lead time ($L_1 + L_2$) and L_1.

5.5.1 Upstream decides his own lead time

In this setting, the upstream actor decides his own lead time (L_1). The upstream price (P_1) and downstream lead time (L_2) are decided by the downstream actor. In this setting the global price (P_g) is obtained from P_1 plus a fixed margin taken by the downstream actor (δ_2): $P_g = P_1 + \delta_2$. The formulation for both upstream and downstream problems are:

Upstream Problem:

$$\underset{L_1 \geq 0}{\text{Maximize}}\ \Pi_1(L_1) = (P_1 - m_1)\lambda \tag{5.20}$$

$$\text{Subject to } \lambda = a - b_1(P_1 + \delta_2) - b_2(L_1 + L_2) \tag{5.21}$$

$$\Pr(W_1 \leq L_1) \geq s \tag{5.22}$$

$$0 \leq \lambda < \mu_1 \tag{5.23}$$

Downstream Problem:

$$\underset{P_1, L_2 \geq 0}{\text{Maximize}}\ \Pi_2(P_1, L_2) = (P_g - P_1 - m_2)\lambda \tag{5.24}$$

$$\text{Subject to } \lambda = a - b_1(P_1 + \delta_2) - b_2(L_1 + L_2) \tag{5.25}$$

$$\Pr(W_2 \leq L_2) \geq s \tag{5.26}$$

$$0 \leq \lambda < \mu_2 \tag{5.27}$$

$$\text{knowing that } P_g = P_1 + \delta_2 \tag{5.28}$$

$$L_g = L_1 + L_2 \tag{5.29}$$

To solve the problem, we do a backward induction. Thus, we start the analysis by solving the upstream problem.

Upstream Problem

In the upstream problem, L_2 and P_1 are given by the downstream actor thus the only option of upstream actor to maximize its profit is to choose his lead time (L_1). From here we can deduce lemma 5.5.

Lemma 5.5. *The upstream's service constraint is binding, implying that:*

$$L_1(L_2, P_1) = \frac{a - b_1(P_1+\delta_2) - b_2 L_2 - \mu_1 + \sqrt{(\mu_1 - a + b_1(P_1+\delta_2) + b_2 L_2)^2 + 4b_2 \ln\left(\frac{1}{1-s}\right)}}{2b_2}$$

Proof. L_2 and P_1 are given by the downstream actor, the only option of upstream actor to maximize his profit is to increase demand by reducing his lead time (L_1) but in respecting to service constraint. Thus, the upstream's service level will be binding. From here we can deduce that:

$$\Pr(W_1 \leq L_1) = s \Leftrightarrow 1 - e^{-(\mu_1 - \lambda)L_1} = s$$

$$\Leftrightarrow (\mu_1 - a + b_1(P_1 + \delta_2) + b_2 L_2)L_1 + b_2 {L_1}^2 = \ln\left(\frac{1}{1-s}\right)$$

The discriminant of this equation is:

$$\Delta = (\mu_1 - a + b_1(P_1 + \delta_2) + b_2 L_2)^2 + 4b_2 \ln\left(\frac{1}{1-s}\right) \geq 0$$

Thus, we will obtain two roots, knowing that one of the roots is negative. We are only interested in the positive root which is:

$$L_1(L_2, P_1) = \frac{a - b_1(P_1+\delta_2) - b_2 L_2 - \mu_1 + \sqrt{(\mu_1 - a + b_1(P_1+\delta_2) + b_2 L_2)^2 + 4b_2 \ln\left(\frac{1}{1-s}\right)}}{2b_2}$$

■

Having obtained L_1 from the upstream, we can move to the downstream problem.

Downstream Problem

The service constraint ($\Pr(W_2 \leq L_2) \geq s$) can be rewritten as: $1 - e^{-(\mu_2-\lambda)L_2} \geq s \Leftrightarrow (\mu_2 - a + b_1(P_1 + \delta_2) + b_2(L_1 + L_2))L_2 \geq \ln\left(\frac{1}{1-s}\right)$. Then, the problem can be rewritten as:

$$\underset{P_1, L_2 \geq 0}{\text{Maximize}}\ \Pi_2(P_1, L_2) = (\delta_2 - m_2)(a - b_1(P_1 + \delta_2) - b_2(L_1 + L_2)) \tag{5.30}$$

$$\text{Subject to } (\mu_2 - a + b_1(P_1 + \delta_2) + b_2(L_1 + L_2))L_2 \geq \ln\left(\frac{1}{1-s}\right) \tag{5.31}$$

$$0 \leq a - b_1(P_1 + \delta_2) - b_2(L_1 + L_2) < \mu_2 \tag{5.32}$$

From the problem above, we can deduce several lemmas.

Lemma 5.6. *The downstream's service level is binding.*

Proof. Assume that at optimality we have ${L_2}^*$ and ${P_1}^*$ which give $(\mu_2 - a + b_1(P_1 + \delta_2) + b_2(L_1 + L_2))L_2 > \ln(1/(1-s))$. We have profit of the downstream as $\Pi_2({L_2}^*, {P_1}^*)$. If we decrease the ${L_2}^*$ to ${L_2}'$ (by keeping the P_1 constant) until $(\mu_2 - a + b_1(P_1 + \delta_2) + b_2(L_1 + L_2))L_2 = \ln(1/(1-s))$, we will get $\Pi_2({L_2}^*, {P_1}^*) < \Pi_2({L_2}', {P_1}^*)$ because demand has increased. Thus, we will have a better solution which is ${P_1}^*$ and ${L_2}'$. ■

Before continuing, let us consider the following remarks:

- The downstream actor's profit (Π_2) is increasing in λ which is proven by $\Pi_2 = (\delta_2 - m_2)\lambda$ (from eq.(5.24) and eq.(5.28)).
- Demand λ is increasing in function of L_2 which is proven by $\Pr(W_2 \leq L_2) = s \Leftrightarrow (\mu_2 - \lambda)L_2 = \ln\left(\frac{1}{1-s}\right) \Leftrightarrow \lambda = \mu_2 - \frac{\ln\left(\frac{1}{1-s}\right)}{L_2}$ (from eq.(5.26) and lemma 5.6).
- L_1 is an increasing function in λ which is proven by $\Pr(W_1 \leq L_1) = s \Leftrightarrow (\mu_1 - \lambda)L_1 = \ln\left(\frac{1}{1-s}\right) \Leftrightarrow L_1 = \frac{\ln\left(\frac{1}{1-s}\right)}{\mu_1 - \lambda}$ (from eq.(5.22) and lemma 5.5).

- And price P_1 decreases when L_1, L_2, and λ increase which is proven by $P_1 = \frac{1}{b_1}\left[a - b_1\delta_2 - b_2(L_1 + L_2) - \lambda\right]$ (from eq.(5.25)).

Based on those remarks, we derive the following lemma.

Lemma 5.7. *The optimum price P_1 proposed by the downstream actor is $P_1 = m_1$.*

Proof. Suppose that at optimality we have ${P_1}^*$, λ^*, ${L_1}^*$, and ${L_2}^*$ with ${P_1}^* > m_1$; we have downstream actor's profit as ${\Pi_2}^*$. If we increase ${L_2}^*$ to ${L_2}'$, keeping the service constraint binding for actor 2, the demand λ will increase thus we would obtain profit ${\Pi_2}' > {\Pi_2}^*$. Because demand increases, L_1 will increase if we keep binding the service constraint for actor 1. As L_2, λ, and L_1 increase, consequently price P_1 will decrease. But this is possible only until a value of P_1 such that $P_1 = m_1$ (otherwise the profit of actor 1 would be negative). Thus, downstream actor can increase L_2 to get a better profit until L_2 such that it leads to $P_1 = m_1$.

■

To illustrate this lemma 5.7, we provide the profit curve in function of L_2 for each actor (Π_1 and Π_2) and the global profit which is the sum of each actor profit ($\Pi_g = \Pi_1 + \Pi_2$) (see figure 5.4). The curves are drawn with price:

$$P_1(L_2) = \frac{\left(2\frac{\ln\left(\frac{1}{1-s}\right)}{L_2} - 2\mu_2 + a - b_1\delta_2 - b_2L_2 + \mu_1\right)^2 - (\mu_1 - a + b_1\delta_2 + b_2L_2)^2 - 4b_2\ln\left(\frac{1}{1-s}\right)}{2b_1\left[2\frac{\ln\left(\frac{1}{1-s}\right)}{L_2} - 2\mu_2 + 2\mu_1\right]}$$

which can be obtained from the service constraint (5.31) using lemma 5.6. We observe the same behavior for all experiments that we have done. We see that Π_2 is increasing in function of L_2 given that $\Pi_1 \geq 0$. In consequence, the downstream actor will choose the longer L_2 to maximize his own profit. This will lead to zero profit for upstream actor ($\Pi_1 = 0$). And since $\Pi_1 = \lambda(P_1 - m_1)$, it means that we effectively have $P_1 = m_1$.

Lemma 5.8. *Given that the downstream will propose $P_1 = m_1$, for the case $\mu_1 \neq \mu_2$ the candidates for optimum L_2 will be the solution of the following cubic equation:*

$$(\mu_2 - \mu_1)b_2{L_2}^3 + (\mu_2(\mu_2 - \mu_1) - \alpha(\mu_2 - \mu_1) + b_1m_1(\mu_2 - \mu_1) - 2b_2\gamma){L_2}^2 + \gamma(\mu_1 - 2\mu_2 + \alpha - b_1m_1)L_2 + \gamma^2 = 0$$

where $\alpha = a - b_1\delta_2$ and $\gamma = \ln(1/(1-s))$.
The optimum L_2 is the root which gives a maximum Π_2 in regard to the problems constraint.

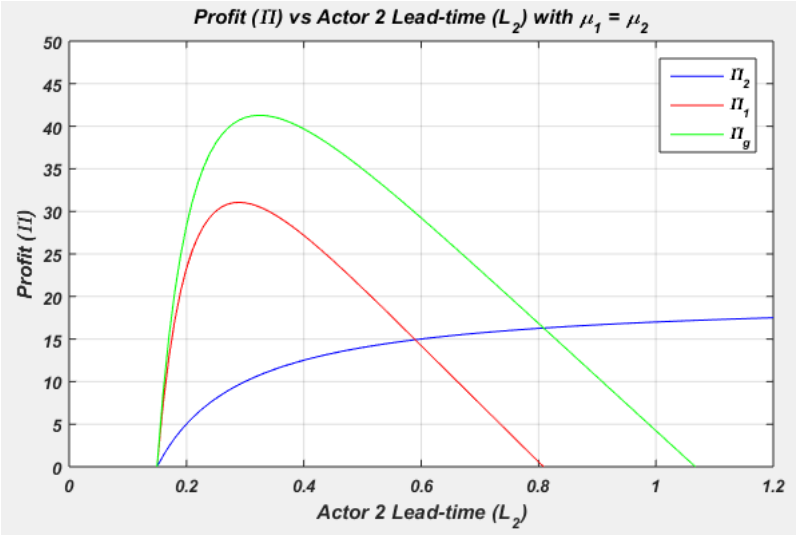


Figure 5.4: Π_1, Π_2 and Π_g with $\mu_1 = \mu_2$

Proof. From lemma 5.6 and 5.7, we get:

$$(\mu_2 - a + b_1(P_1 + \delta_2) + b_2(L_1 + L_2))L_2 = \ln\left(\frac{1}{1-s}\right)$$
$$\Leftrightarrow (\mu_2 - a + b_1(m_1 + \delta_2) + b_2(L_1 + L_2))L_2 = \ln\left(\frac{1}{1-s}\right)$$
$$\Leftrightarrow b_1 m_1 = \frac{\ln\left(\frac{1}{1-s}\right)}{L_2} - \mu_2 + a - b_1\delta_2 - b_2(L_1 + L_2)$$

Given the expression of L_1 in upstream problem (see lemma 5.5), we obtain:

$$b_1 m_1 = \frac{\ln\left(\frac{1}{1-s}\right)}{L_2} - \mu_2 + a - b_1\delta_2 - b_2 L_2 - \frac{a - b_1(m_1+\delta_2) - b_2 L_2 - \mu_1 + \sqrt{(\mu_1 - a + b_1(m_1+\delta_2) + b_2 L_2)^2 + 4b_2 \ln\left(\frac{1}{1-s}\right)}}{2}$$
$$\Leftrightarrow b_1 m_1 = 2\frac{\ln\left(\frac{1}{1-s}\right)}{L_2} - 2\mu_2 + a - b_1\delta_2 - b_2 L_2 + \mu_1 - \sqrt{(\mu_1 - a + b_1(m_1+\delta_2) + b_2 L_2)^2 + 4b_2 \ln\left(\frac{1}{1-s}\right)}$$
$$\Leftrightarrow \sqrt{(\mu_1 - a + b_1(m_1+\delta_2) + b_2 L_2)^2 + 4b_2 \ln\left(\frac{1}{1-s}\right)} = 2\frac{\ln\left(\frac{1}{1-s}\right)}{L_2} - 2\mu_2 + a - b_1\delta_2 - b_2 L_2 + \mu_1 - b_1 m_1$$
$$\Leftrightarrow (\mu_1 - a + b_1\delta_2 + b_1 m_1 + b_2 L_2)^2 + 4b_2 \ln\left(\frac{1}{1-s}\right) = \left(2\frac{\ln\left(\frac{1}{1-s}\right)}{L_2} - 2\mu_2 + \mu_1 + a - b_1\delta_2 - b_2 L_2 - b_1 m_1\right)^2$$
$$\Leftrightarrow 2b_1 m_1 (2\frac{\ln\left(\frac{1}{1-s}\right)}{L_2} - 2\mu_2 + 2\mu_1) = (2\frac{\ln\left(\frac{1}{1-s}\right)}{L_2} - 2\mu_2 + a - b_1\delta_2 - b_2 L_2 + \mu_1)^2 - (\mu_1 - a + b_1\delta_2 + b_2 L_2)^2 - 4b_2 \ln\left(\frac{1}{1-s}\right)$$
$$\Leftrightarrow m_1 = \frac{\left(2\frac{\ln\left(\frac{1}{1-s}\right)}{L_2} - 2\mu_2 + a - b_1\delta_2 - b_2 L_2 + \mu_1\right)^2 - (\mu_1 - a + b_1\delta_2 + b_2 L_2)^2 - 4b_2 \ln\left(\frac{1}{1-s}\right)}{2b_1\left[2\frac{\ln\left(\frac{1}{1-s}\right)}{L_2} - 2\mu_2 + 2\mu_1\right]}$$
$$\Leftrightarrow \frac{4(\mu_2-\mu_1)b_2 {L_2}^3 + 4(\mu_2(\mu_2-\mu_1) - \alpha(\mu_2-\mu_1) + b_1 m_1(\mu_2-\mu_1) - 2b_2\gamma){L_2}^2 + 4\gamma(\mu_1 - 2\mu_2 + \alpha - b_1 m_1)L_2 + 4\gamma^2}{{L_2}^2} = 0$$

where $\alpha = a - b_1\delta_2$ and $\gamma = \ln(1/(1-s))$. Knowing that $\frac{4}{{L_2}^2} \neq 0$, thus we will get

a cubic expression as follows:

$$(\mu_2-\mu_1)b_2{L_2}^3+(\mu_2(\mu_2-\mu_1)-\alpha(\mu_2-\mu_1)+b_1m_1(\mu_2-\mu_1)-2b_2\gamma){L_2}^2 \\ +\gamma(\mu_1-2\mu_2+\alpha-b_1m_1)L_2+\gamma^2=0$$

The cubic expression above will have three roots namely L_{21}, L_{22}, and L_{23} (see Appendix E for the detailed demonstration). We will have either one or three real roots. Recall that we are only interested in positive lead time. The best lead time is the one which gives a maximum Π_2 and satisfies the problem's constraints. ■

Lemma 5.9. *For the case $\mu_1=\mu_2$, the optimum quoted lead time of downstream actor is:*

$$L_2=\frac{\alpha-m_1b_1-\mu_1+\sqrt{(\alpha-\mu_1-b_1m_1)^2+8b_2\gamma}}{4b_2}$$

where $\alpha=a-b_1\delta_2$ and $\gamma=\ln(1/(1-s))$

Proof. For $\mu_1=\mu_2$, the equation of lemma 5.8 is equivalent to:

$$-2b_2{L_2}^2+(\alpha-\mu_1-b_1m_1)L_2+\gamma=0$$

with discriminant $\Delta=(\alpha-\mu_1-b_1m_1)^2+8b_2\gamma\geq 0$.
We will have two roots where one of the roots is negative. We only use the positive root which equals to: $L_2=\frac{\alpha-m_1b_1-\mu_1+\sqrt{(\alpha-\mu_1-b_1m_1)^2+8b_2\gamma}}{4b_2}$. ■

As a summary, we put our result in the proposition 5.3 below.

Proposition 5.3. *The solution of the decentralized setting problem when upstream actor decides his own lead time is:*

1. *The optimum L_2 is:*
   - *For the case where $\mu_1\neq\mu_2$, L_2 is one of the roots of the cubic equation of lemma 5.8 which gives a maximum Π_2 and satisfies the problem's constraints,*
   - *For the case where $\mu_1=\mu_2$, $L_2=\frac{\alpha-m_1b_1-\mu_1+\sqrt{(\alpha-\mu_1-b_1m_1)^2+8b_2\gamma}}{4b_2}$, with $\alpha=a-b_1\delta_2$ and $\gamma=\ln(1/(1-s))$,*
2. *The optimum price is $P_1=m_1$,*
3. *The optimum $L_1(L_2)=\frac{a-b_1(P_1+\delta_2)-b_2L_2-\mu_1+\sqrt{(\mu_1-a+b_1(P_1+\delta_2)+b_2L_2)^2+4b_2\ln\left(\frac{1}{1-s}\right)}}{2b_2}$,*
4. *The optimum demand $\lambda(L_1,L_2)=a-b_1(m_1+\delta_2)-b_2(L_1(L_2)+L_2)$.*

Profit of each actor can be calculated as:
$\Pi_1(L_2)=0$,
$\Pi_2(L_2)=(\delta_2-m_2)(a-b_1(m_1+\delta_2)-b_2(L_1(L_2)+L_2))$,
$\Pi_g(L_2)=\Pi_2(L_2)$.

Numerical experiment

We do some experiments on effect of b_1 and b_2. We vary one parameter and fix other parameters. For the capacity parameters (μ), we define cases where $\mu_1 = \mu_2$ and $\mu_1 \neq \mu_2$.

For the first experiments, we choose $\mu_1 = \mu_2 = 20$. We set the other base parameters: $a = 50$, $b_1 = 4$, $b_2 = 6$, $m_1 = 3$, $m_2 = 2$, $\delta_2 = 3$ and $s = 0.95$. Here, logically we will have $L_1 = L_2$ as both capacities of the actors are the same. As we can see in table 5.10 and 5.11, both lead time are effectively the same. The explanation of price decreases as b_1 decreases and lead time decreases as b_2 decreases are similar to explanation in section 5.2. To give a better understanding whether both actors can cooperate effectively or not, we show the value of Π_g^{Max} (which is the maximum of the Π_g's curve) and the value of loss (with respect to the centralized setting). We use $\frac{\Pi_c - \Pi_g}{\Pi_c} \times 100\%$ to calculate the loss. And to find Π_g^{Max}, we use the MatLab Optimization function (*fminsearch*). For the experiment with b_1, we get an average loss of 65.36% and for experiment with b_2 we get an average loss of 61.21%. The obtained global profit (Π_g) is very low compared to the centralized setting profit (Π_c). However, we can see that Π_g^{Max} is close to Π_c. This signifies that the decentralized setting approach could provide a favorable global profit provided that the proposed coordination system could lead the actors into this situation. Note that Π_g^{Max} is obtained numerically with MatLab function.

Table 5.10: b_1 for $\mu_1 = \mu_2$

b_1	Decentralized Setting - Upstream decide L_1									
	L_1	L_2	P_1	λ	Π_1	Π_2	Π_g	Π_g^{Max}	Π_c	Loss
2	1.6512	1.6512	3.0000	18.1857	0.0000	18.1857	18.1857	142.1375	149.3263	87.82%
4	0.8087	0.8087	3.0000	16.2956	0.0000	16.2956	16.2956	41.2901	43.5856	62.61%
6	0.3087	0.3087	3.0000	10.2956	0.0000	10.2956	10.2956	12.0349	12.8375	19.80%
8	0.1512	0.1512	3.0000	0.1857	0.0000	0.1857	0.1857	1.8940	2.1119	91.21%

Table 5.11: b_2 for $\mu_1 = \mu_2$

b_2	Decentralized Setting - Upstream decide L_1									
	L_1	L_2	P_1	λ	Π_1	Π_2	Π_g	Π_g^{Max}	Π_c	Loss
2	1.8952	1.8952	3.0000	18.4193	0.0000	18.4193	18.4193	49.7162	50.8456	63.77%
4	1.0927	1.0927	3.0000	17.2584	0.0000	17.2584	17.2584	45.0861	46.8844	63.19%
6	0.8087	0.8087	3.0000	16.2956	0.0000	16.2956	16.2956	41.2901	43.5856	62.61%
8	0.6591	0.6591	3.0000	15.4547	0.0000	15.4547	15.4547	38.0225	40.7125	62.04%
10	0.5651	0.5651	3.0000	14.6985	0.0000	14.6985	14.6985	35.1359	38.1487	61.47%
12	0.4998	0.4998	3.0000	14.0057	0.0000	14.0057	14.0057	32.5441	35.8255	60.91%
14	0.4513	0.4513	3.0000	13.3625	0.0000	13.3625	13.3625	30.1908	33.6976	60.35%
16	0.4138	0.4138	3.0000	12.7597	0.0000	12.7597	12.7597	28.0366	31.7334	59.79%
18	0.3836	0.3836	3.0000	12.1905	0.0000	12.1905	12.1905	26.0525	29.9092	59.24%
20	0.3588	0.3588	3.0000	11.6497	0.0000	11.6497	11.6497	24.2163	28.2071	58.70%

In our second experiment, we consider $\mu_1 \neq \mu_2$. We use the base parameters as: $a = 50$, $b_1 = 4$, $b_2 = 6$, $m_1 = 3$, $m_2 = 2$, $\delta_2 = 3$ and $s = 0.95$. For $\mu_1 > \mu_2$ we use $\mu_1 = 30$ & $\mu_2 = 15$ and for $\mu_1 < \mu_2$ we use $\mu_1 = 15$ & $\mu_2 = 30$. We only display the result with $\mu_1 > \mu_2$. If readers are interested with the case $\mu_1 < \mu_2$, the readers can swap the value of L_1 and L_2 for each instances. The other values such as price, demand and profit remain the same. As expected, the higher capacity of the actor, the lower the lead time will be (see table 5.12 and 5.13). The explanation of price decreases as b_1 decreases and lead time decreases as b_2 decreases are similar to those given in experiment of section 5.2. For the experiments with b_1, we get an average loss of 66.85% between Π_g and Π_c; and for the experiments with b_2 we get an average loss of 63.98%; but again Π_g^{Max} is closed to Π_c.

Table 5.12: b_1 for $\mu_1 > \mu_2$

b_1	Decentralized Setting - Upstream decide L_1									
	L_1	L_2	P_1	λ	Π_1	Π_2	Π_g	Π_g^{Max}	Π_c	Loss
2	0.1897	3.7759	3.0000	14.2066	0.0000	14.2066	14.2066	129.5849	132.8462	89.31%
4	0.1808	1.9134	3.0000	13.4344	0.0000	13.4344	13.4344	38.8582	40.1800	66.56%
6	0.1477	0.5667	3.0000	9.7137	0.0000	9.7137	9.7137	11.6491	12.2250	20.54%
8	0.1005	0.2022	3.0000	0.1840	0.0000	0.1840	0.1840	1.8642	2.0419	90.99%

Table 5.13: b_2 for $\mu_1 > \mu_2$

b_2	Decentralized Setting - Upstream decide L_1									
	L_1	L_2	P_1	λ	Π_1	Π_2	Π_g	Π_g^{Max}	Π_c	Loss
2	0.1928	5.5758	3.0000	14.4627	0.0000	14.4627	14.4627	47.3501	47.9264	69.82%
4	0.1865	2.8283	3.0000	13.9408	0.0000	13.9408	13.9408	42.5986	43.5859	68.02%
6	0.1808	1.9134	3.0000	13.4344	0.0000	13.4344	13.4344	38.8582	40.1800	66.56%
8	0.1756	1.4565	3.0000	12.9432	0.0000	12.9432	12.9432	35.6984	37.3022	65.30%
10	0.1709	1.1825	3.0000	12.4666	0.0000	12.4666	12.4666	32.9369	34.7828	64.16%
12	0.1665	0.9999	3.0000	12.0039	0.0000	12.0039	12.0039	30.4744	32.5300	63.10%
14	0.1624	0.8694	3.0000	11.5543	0.0000	11.5543	11.5543	28.2491	30.4873	62.10%
16	0.1586	0.7715	3.0000	11.1172	0.0000	11.1172	11.1172	26.2189	28.6163	61.15%
18	0.1552	0.6953	3.0000	10.6916	0.0000	10.6916	10.6916	24.3536	26.8898	60.24%
20	0.1519	0.6343	3.0000	10.2769	0.0000	10.2769	10.2769	22.6305	25.2873	59.36%

From these experiments, we see that using this decentralized setting, the global profit is very low with comparison to the centralized setting. We also see that the profit of upstream actor is null. It shows that the actors cannot cooperate naturally. However, if we observe the comparison between Π_g^{Max} and Π_c, we see that the difference is small. Thus, it would be interesting to introduce some innovation mechanisms which could lead to a better coordination. In the next section we consider another way; indeed we propose to swap the decision taken by each actor. The upstream (follower) decides his own price (P_1), and downstream decides the global lead time ($L_g = L_1 + L_2$).

5.5.2 Upstream decides his own price

In this model we consider a decentralized setting where again the leader is the downstream actor and follower is the upstream actor. The upstream decides his own price (P_1), and downstream decides the global lead time (L_g) by deciding his own lead time (L_2) and upstream's lead time (L_1). The global price P_g is obtained from P_1 by $P_g = P_1 + \delta_2$.

Upstream Problem:

$$\underset{P_1 \geq 0}{\text{Maximize}}\ \Pi_1(P_1) = (P_1 - m_1)\lambda \tag{5.33}$$

$$\text{Subject to } \lambda = a - b_1(P_1 + \delta_2) - b_2(L_1 + L_2) \tag{5.34}$$

$$\Pr(W_1 \leq L_1) \geq s \tag{5.35}$$

$$0 \leq \lambda < \mu_1 \tag{5.36}$$

Downstream Problem:

$$\underset{L_1, L_2 \geq 0}{\text{Maximize}}\ \Pi_2(L_1, L_2) = (P_g - P_1 - m_2)\lambda \tag{5.37}$$

$$\text{Subject to } \lambda = a - b_1(P_1 + \delta_2) - b_2(L_1 + L_2) \tag{5.38}$$

$$\Pr(W_2 \leq l_2) \geq s \tag{5.39}$$

$$0 \leq \lambda < \mu_2 \tag{5.40}$$

$$\text{knowing that } P_g = P_1 + \delta_2 \tag{5.41}$$

To solve the problems above we use a backward induction, we start by solving the upstream problem.

Upstream Problem

The service constraint ($\Pr(W_1 \leq L_1) \geq s$) can be rewritten using the expression of λ from eq. (5.38) as:

$$\begin{aligned} 1 - e^{-(\mu_1 - \lambda)L_1} \geq s \Leftrightarrow 1 - e^{-(\mu_1 - a + b_1(P_1 + \delta_2) + b_2(L_1 + L_2))L_1} \geq s \\ \Leftrightarrow P_1 \geq \frac{\frac{\ln\left(\frac{1}{1-s}\right)}{L_1} - \mu_s + a - b_1\delta_2 - b_2(L_1 + L_2)}{b_1} \end{aligned}$$

Then, we have a new formulation of the problem as:

$$\underset{P_1 \geq 0}{\text{Maximize}}\ \Pi_1(P_1) = (P_1 - m_1)(a - b_1(P_1 + \delta_2) - b_2(L_1 + L_2)) \tag{5.42}$$

$$\text{Subject to } P_1 \geq \frac{\frac{\ln\left(\frac{1}{1-s}\right)}{L_1} - \mu_1 + a - b_1\delta_2 - b_2(L_1 + L_2)}{b_1} \tag{5.43}$$

$$0 \leq a - b_1(P_1 + \delta_2) - b_2(L_1 + L_2) < \mu_1 \tag{5.44}$$

Different from the previous case (section 5.5.1), in this new case (section 5.5.2) the service constraint (5.43) isn't necessarily binding. Indeed, if we have a given P_1 for which the service constraint is not tight, we could reduce P_1 without violating this constraint. By decreasing P_1, we increase $(a - b_1(P_1 + \delta_2) - b_2(L_1 + L_2))$ but we also decrease $(P_1 - m_1)$. Thus, this does not necessarily improve the profit. This trade off complicates the solving approach. Thus, in order to solve the problem, we consider the following lemma.

Lemma 5.10. *If service level is binding, we have:* $P_1 = \frac{\frac{\ln\left(\frac{1}{1-s}\right)}{L_1} - \mu_1 + a - b_1\delta_2 - b_2(L_1+L_2)}{b_1}$. *Otherwise, service level is non-binding and we have:* $P_1 = \frac{a - b_1(\delta_2 - m_1) - b_2(L_1+L_2)}{2b_1}$.

Proof. Let us consider the two possible cases: (1) service constraint is binding, and (2) service constraint is non-binding.

Case 1: service constraint is binding. From eq.(5.43), we have immediately: $P_1 = \frac{\frac{\ln\left(\frac{1}{1-s}\right)}{L_1} - \mu_1 + a - b_1\delta_2 - b_2(L_1+L_2)}{b_1}$.

Case 2: service constraint is non-binding. We can find P_1 directly from the objective function: $\Pi_1(P_1) = (P_1 - m_1)(a - b_1(P_1 + \delta_2) - b_2(L_1 + L_2))$.
Thus, the first derivative in P_1 is:

$$\frac{d}{dP_1}\Pi_1(P_1) = 0 \Leftrightarrow a - b_1(2P_1 + \delta_2) - b_2(L_1 + L_2) + b_1 m_1 = 0$$
$$\Leftrightarrow P_1 = \frac{a - b_1(\delta_2 - m_1) - b_2(L_1 + L_2)}{2b_1}$$

■

Having obtained the expression of P_1 in both binding and non-binding cases, we move to the downstream problem.

Downstream Problem

The service constraint $(\Pr(W_2 \leq l_2) \geq s)$ can be rewritten as: $1 - e^{-(\mu_2 - \lambda)L_2} \geq s \Leftrightarrow (\mu_2 - a + b_1(P_1 + \delta_2) + b_2(L_1 + L_2))L_2 \geq \ln\left(\frac{1}{1-s}\right)$. Then the problem can be rewritten as:

$$\underset{L_1, L_2 \geq 0}{\text{Maximize}}\ \Pi_2(L_1, L_2) = (\delta_2 - m_2)(a - b_1(P_1 + \delta_2) - b_2(L_1 + L_2)) \tag{5.45}$$

$$\text{Subject to } (\mu_2 - a + b_1(P_1 + \delta_2) + b_2(L_1 + L_2))L_2 \geq \ln\left(\frac{1}{1-s}\right) \tag{5.46}$$

$$0 \leq a - b_1(P_1 + \delta_2) - b_2(L_1 + L_2) < \mu_2 \tag{5.47}$$

From the problem above, we deduce several lemmas.

Lemma 5.11. *Service constraint (5.46) is binding at optimality.*

Proof. Let's assume that at optimality we have ${L_2}^*$ and ${L_1}^*$ which give $(\mu_2 - a + b_1(P_1+\delta_2)+b_2(L_1+L_2))L_2 > \ln(1/(1-s))$. We have profit of the downstream actor as $\Pi_2({L_2}^*, {L_1}^*)$. If we decrease the ${L_2}^*$ to ${L_2}'$ (by keeping the L_1 constant) until $(\mu_2 - a + b_1(P_1 + \delta_2) + b_2(L_1 + L_2))L_2 = \ln(1/(1-s))$, we will get $\Pi_2({L_2}^*, {L_1}^*) < \Pi_2({L_2}', {L_1}^*)$ because the demand has increased. We have a better solution which are ${L_2}'$ and ${L_1}^*$. Thus, service level is binding at optimality. ■

And taking into account the two expressions of P_1 in lemma 5.10, we have the following lemma.

Lemma 5.12. *We have two situations based on the uptream's service level.*

- *If upstream service level is binding, then*
$$L_2(L_1) = \frac{\ln\left(\frac{1}{1-s}\right)}{\frac{\ln\left(\frac{1}{1-s}\right)}{L_1} - \mu_1 + \mu_2},$$
$$P_1(L_1) = \frac{\frac{\ln\left(\frac{1}{1-s}\right)}{L_1} - \mu_1 + a - b_1\delta_2 - b_2\left(L_1 + \frac{\ln\left(\frac{1}{1-s}\right)}{\frac{\ln\left(\frac{1}{1-s}\right)}{L_1} - \mu_1 + \mu_2}\right)}{b_1},$$
$$\lambda(L_1) = \mu_1 - \frac{\ln\left(\frac{1}{1-s}\right)}{L_1}.$$

- *If upstream service level is non-binding, then*
$$L_2(L_1) = \frac{a - b_1(\delta_2+m_1) - b_2L_1 - 2\mu_2 + \sqrt{(2\mu_2 - a + b_1(\delta_2+m_1) + b_2L_1)^2 + 8b_2\ln\left(\frac{1}{1-s}\right)}}{2b_2},$$
$$P_1(L_1) = \frac{a - b_1(\delta_2 - 3m_1) - b_2L_1 + 2\mu_2 - \sqrt{(2\mu_2 - a + b_1(\delta_2+m_1) + b_2L_1)^2 + 8b_2\ln\left(\frac{1}{1-s}\right)}}{4b_1},$$
$$\lambda(L_1) = \frac{a - b_1(\delta_2+m_1) - b_2L_1 + 2\mu_2 - \sqrt{(2\mu_2 - a + b_1(\delta_2+m_1) + b_2L_1)^2 + 8b_2\ln\left(\frac{1}{1-s}\right)}}{4}$$

Proof. We have two possible situations: (1) upstream service constraint is binding and (2) upstream service constraint is non-binding.

Case 1: If upstream service constraint is binding.
We have $P_1 = \frac{\frac{\ln\left(\frac{1}{1-s}\right)}{L_1} - \mu_1 + a - b_1\delta_2 - b_2(L_1+L_2)}{b_1}$ from lemma 5.10. The service level of the downstream is always binding (lemma 5.11), thus the service constraint:

$$(\mu_2 - a + b_1(P_1 + \delta_2) + b_2(L_1 + L_2))L_2 = \ln\left(\frac{1}{1-s}\right) \Leftrightarrow L_2 = \frac{\ln\left(\frac{1}{1-s}\right)}{\frac{\ln\left(\frac{1}{1-s}\right)}{L_1} - \mu_1 + \mu_2}$$

Then, substitute the expression of L_2 into $P_1(L_1, L_2)$, we get:

$$P_1(L_1, L_2) = \frac{\frac{\ln\left(\frac{1}{1-s}\right)}{L_1} - \mu_1 + a - b_1\delta_2 - b_2(L_1 + L_2)}{b_1}$$

$$\Leftrightarrow P_1(L_1) = \frac{\frac{\ln\left(\frac{1}{1-s}\right)}{L_1} - \mu_1 + a - b_1\delta_2 - b_2\left(L_1 + \frac{\ln\left(\frac{1}{1-s}\right)}{\frac{\ln\left(\frac{1}{1-s}\right)}{L_1} - \mu_1 + \mu_2}\right)}{b_1}$$

Next, substitute the expression of $P_1(L_1)$ and L_2 into $\lambda(P_1, L_1, L_2)$, we get:

$$\lambda(P_1, L_1, L_2) = a - b_1(P_1 + \delta_2) - b_2(L_1 + L_2) \Leftrightarrow \lambda(L_1) = \mu_1 - \frac{\ln\left(\frac{1}{1-s}\right)}{L_1}$$

Case 2: If upstream service constraint is non-binding.
We have $P_1 = \frac{a - b_1(\delta_2 - m_1) - b_2(L_1+L_2)}{2b_1}$. The service level of the downstream is always binding (lemma 5.11), thus the service constraint:

$$(\mu_2 - a + b_1(P_1 + \delta_2) + b_2(L_1 + L_2))L_2 = \ln\left(\frac{1}{1-s}\right)$$

$$\Leftrightarrow b_2{L_2}^2 + (2\mu_2 - a + b_1(\delta_2 + m_1) + b_2L_1)L_2 - 2\ln\left(\frac{1}{1-s}\right) = 0$$

Given that $\Delta = (2\mu_2 - a + b_1(\delta_2 + m_1) + b_2L_1)^2 + 8b_2\ln\left(\frac{1}{1-s}\right) \geq 0$, we get two roots and we only consider the positive root. Thus we have:

$$L_2 = \frac{a - b_1(\delta_2 + m_1) - b_2L_1 - 2\mu_2 + \sqrt{(2\mu_2 - a + b_1(\delta_2 + m_1) + b_2L_1)^2 + 8b_2\ln\left(\frac{1}{1-s}\right)}}{2b_2}$$

Then, substitute the expression of L_2 into $P_1(L_1, L_2)$, we get:

$$P_1(L_1, L_2) = \frac{(a - b_1(\delta_2 - m_1) - b_2(L_1 + L_2))}{2b_1}$$

$$\Leftrightarrow P_1(L_1) = \frac{a - b_1(\delta_2 - 3m_1) - b_2L_1 + 2\mu_2 - \sqrt{(2\mu_2 - a + b_1(\delta_2 + m_1) + b_2L_1)^2 + 8b_2\ln\left(\frac{1}{1-s}\right)}}{4b_1}$$

Demand of the downstream actor becomes:

$$\lambda(P_1, L_1, L_2) = a - b_1(P_1 + \delta_2) - b_2(L_1 + L_2)$$

$$\Leftrightarrow \lambda(P_1, L_1) = \frac{2(a - b_1\delta_2 - b_2L_1) + 2b_1m_1 + 4\mu_2 - 2\sqrt{(2\mu_2 - a + b_1(\delta_2 + m_1) + b_2L_1)^2 + 8b_2\ln\left(\frac{1}{1-s}\right)}}{4} - b_1P_1$$

$$\Leftrightarrow \lambda(L_1) = \frac{a - b_1(\delta_2 + m_1) - b_2L_1 + 2\mu_2 - \sqrt{(2\mu_2 - a + b_1(\delta_2 - m_1) + b_2L_1)^2 + 8b_2\ln\left(\frac{1}{1-s}\right)}}{4}$$

■

From here, we see that we now only need to find L_1. We define the objective function of the downstream actor in L_1 for both binding and non-binding situations as below:

$$\Pi_2{}^{Binding}(L_1) = (\delta_2 - m_2)\left(\mu_1 - \frac{\ln\left(\frac{1}{1-s}\right)}{L_1}\right)$$

$$\Pi_2{}^{NonBinding}(L_1) = (\delta_2 - m_2)\left(\frac{a - b_1(\delta_2 + m_1) - b_2 L_1 + 2\mu_2 - \sqrt{(2\mu_2 - a + b_1(\delta_2 - m_1) + b_2 L_1)^2 + 8b_2 \ln\left(\frac{1}{1-s}\right)}}{4}\right)$$

We can easily see from the expression of $\Pi_2{}^{Binding}$ that it is increasing in function of L_1. And from the expression of $\Pi_2{}^{NonBinding}$, we see that it is decreasing in function of L_1. Thus, the curve of $\Pi_2(L_1)$ increases at the beginning (as it is in binding situation) then after a certain point the curve will decrease (as it is in non-binding situation). Thus, the maximum profit for downstream actor is when $\Pi_2{}^{Binding}(L_1) = \Pi_2{}^{NonBinding}(L_1)$.

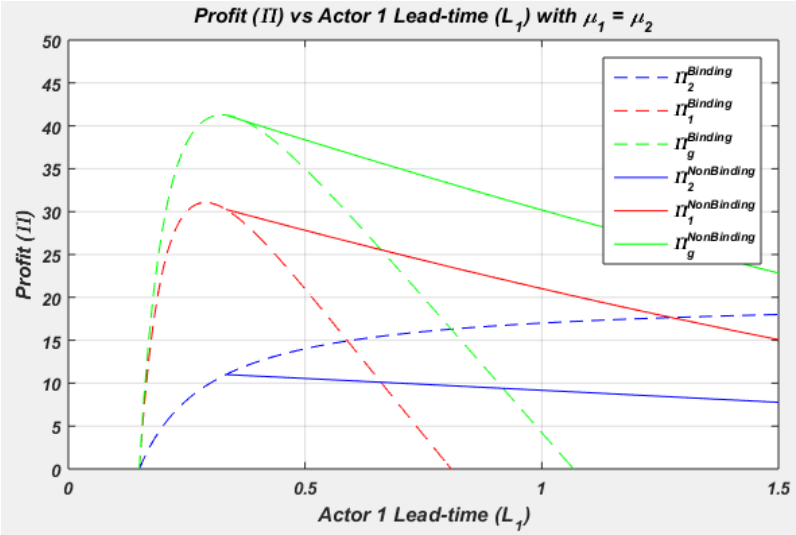


Figure 5.5: Theoritical profit of each actor with $\mu_1 = \mu_2$

To illustrate this result, we draw the profit curve for each actor (Π_1 and Π_2) and profit global (Π_g) in both binding and non-binding situations. The curves can be seen in figure 5.5. We found the same behavior for all tests done. In this figure 5.5, we draw $\Pi_1(L_1)$, $\Pi_2(L_1)$, and $\Pi_g(L_1)$ obtained for the two situations: binding and non-binding cases. But of course, we cannot have both situations simultaneously thus we draw figure 5.6 with only the significant parts of the curves.

From figure 5.6, we can see effectively that $\Pi_2{}^{Binding}$ is increasing in function of L_1 and $\Pi_2{}^{NonBinding}$ is decreasing in function of L_1 which confirm our result.

Lemma 5.13. *Given that the downstream will propose the L_1 such that $\Pi_2{}^{Binding}(L_1) = \Pi_2{}^{NonBinding}(L_1)$, for the case where $\mu_1 \neq \mu_2$, the candidates for the optimum L_1*

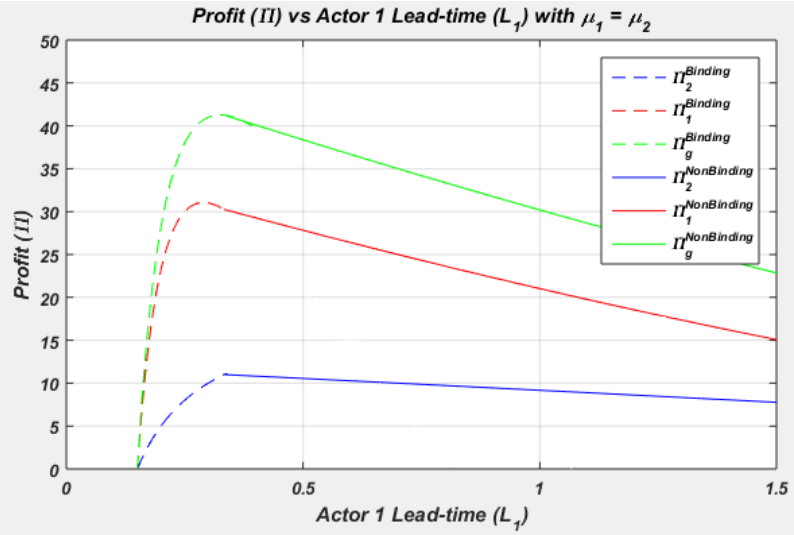


Figure 5.6: Real profit of each actor with $\mu_1 = \mu_2$

can be obtained from the following cubic equation:

$$(\mu_2-\mu_1)b_2{L_1}^3+((\alpha+2\mu_2-2\mu_1)\mu_1-\alpha\mu_2+2b_2\gamma){L_1}^2+(4\mu_1-2\mu_2-\alpha)\gamma L_1-2\gamma^2)=0$$

where: $\alpha = a - b_1\delta_2 - b_1 m_1$ *and* $\gamma = \ln(1/(1-s))$.
The optimum L_1 *is the smallest positive root.*

Proof. Given that the downstream will propose the L_1 such that:

$$\begin{aligned}
&{\Pi_2}^{Binding}(L_1) = {\Pi_2}^{NonBinding}(L_1)\\
&\Leftrightarrow (\delta_2 - m_2) \times \lambda^{Binding}(L_1) = (\delta_2 - m_2) \times \lambda^{NonBinding}(L_1)\\
&\Leftrightarrow \lambda^{Binding}(L_1) = \lambda^{NonBinding}(L_1)\\
&\Leftrightarrow \mu_1 - \frac{\gamma}{L_1} = \frac{\alpha - b_2 L_1 + 2\mu_2 - \sqrt{(2\mu_2 - \alpha + b_2 L_1)^2 + 8b_2\gamma}}{4}\\
&\Leftrightarrow \sqrt{(2\mu_2 - \alpha + b_2 L_1)^2 + 8b_2\gamma} = \alpha - b_2 L_1 + 2\mu_2 - 4\mu_1 + 4\frac{\gamma}{L_1}\\
&\Leftrightarrow \frac{8((\mu_2-\mu_1)b_2{L_1}^3+((\alpha+2\mu_2-2\mu_1)\mu_1-\alpha\mu_2+2b_2\gamma){L_1}^2+(4\mu_1-2\mu_2-\alpha)\gamma L_1-2\gamma^2)}{{L_1}^2} = 0
\end{aligned}$$

where $\alpha = a - b_1\delta_2 - b_1 m_1$ and $\gamma = \ln(1/(1-s))$. Knowing that $\frac{8}{{L_1}^2} > 0$, we will obtain a cubic equation:

$$(\mu_2-\mu_1)b_2{L_1}^3+((\alpha+2\mu_2-2\mu_1)\mu_1-\alpha\mu_2+2b_2\gamma){L_1}^2+(4\mu_1-2\mu_2-\alpha)\gamma L_1-2\gamma^2=0$$

The cubic expression above will have three roots namely L_{11}, L_{12}, and L_{13} (see Appendix F for the detailed demonstration). We will have either one or three real roots. Recall that we are only interested in positive lead time and we consider L_1 such that $\lambda^{NonBinding}(L_1) = \lambda^{Binding}(L_1) = \mu_1 - \frac{\gamma}{L_1}$. Let us consider that a case

where there would have several positive roots, that satisfy the problem's constraints, demand being a decreasing function in L_1, the profit (Π_2) is a decreasing function in L_1. Thus, the best lead time is the smallest positive one which satisfy the problem's constraints. Let us note that we have not encountered this several feasible roots case, but we still have not yet prove that it can not happen. ■

Lemma 5.14. *For the case where $\mu_1 = \mu_2$, the optimum quoted lead time of actor 1 is:*

$$L_1 = \frac{\alpha - 2\mu_2 + \sqrt{(2\mu_2 - \alpha)^2 + 16b_2\gamma}}{4b_2}$$

where $\alpha = a - b_1\delta_2 - b_1 m_1$ and $\gamma = \ln(1/(1-s))$

Proof. For the case where $\mu_1 = \mu_2$, the equation of lemma 5.13 becomes:

$$2b_2{L_1}^2 + (2\mu_2 - \alpha)L_1 - 2\gamma = 0$$

with $\Delta = (2\mu_2 - \alpha)^2 + 16b_2\gamma \geq 0$. We have two roots where one of them is negative. We only use the positive root which is $L_1 = \frac{\alpha - 2\mu_2 + \sqrt{(2\mu_2-\alpha)^2+16b_2\gamma}}{4b_2}$. ■

As a summary, we put our results in the proposition below.

Proposition 5.4. *The solution of the decentralized setting problem where upstream actor decides his own price is:*

1. *The optimum L_1 is:*
   - *For case where $\mu_1 \neq \mu_2$, L_1 is the smallest positive root of cubic equation in lemma 5.13,*
   - *For case where $\mu_1 = \mu_2$, $L_1 = \frac{\alpha - 2\mu_2 + \sqrt{(2\mu_2-\alpha)^2+16b_2\gamma}}{4b_2}$, with $\alpha = a - b_1\delta_2 - b_1 m_1$ and $\gamma = \ln(1/(1-s))$.*
2. *The optimum demand: $\lambda(L_1) = \mu_1 - \frac{\ln\left(\frac{1}{1-s}\right)}{L_1}$,*
3. *The optimum price is: $P_1(L_1) = \dfrac{\frac{\ln\left(\frac{1}{1-s}\right)}{L_1} - \mu_1 + a - b_1\delta_2 - b_2\left(L_1 + \frac{\ln\left(\frac{1}{1-s}\right)}{\frac{\ln\left(\frac{1}{1-s}\right)}{L_1} - \mu_1 + \mu_2}\right)}{b_1}$,*
4. *The optimum L_2 is: $L_2(L_1) = \dfrac{\ln\left(\frac{1}{1-s}\right)}{\frac{\ln\left(\frac{1}{1-s}\right)}{L_1} - \mu_1 + \mu_2}$.*

Profit of each actor can be calculated as:
$\Pi_2(L_1) = (\delta_2 - m_2) \times \lambda(L_1)$,
$\Pi_1(L_1) = (P_1(L_1) - m_1) \times \lambda(L_1)$,
$\Pi_g(L_1) = (P_1(L_1) + \delta_2 - m_2 - m_1) \times \lambda(L_1)$.

Proof. Recall that the optimum profit of downstream actor is obtained for ${\Pi_2}^{Binding}(L_1) = {\Pi_2}^{NonBinding}(L_1)$, thus the optimal point is the limit between binding and non-binding situations. It means that the value obtained, for L_1, $\lambda(L_1)$, $P_1(L_1)$, and $L_2(L_1)$ from the binding and non-binding expressions, are equal. For the simplification of presentation, we choose to use the expressions in binding situation. ■

Numerical experiment

We do the experiments on the impact of b_1, b_2, and δ_2. We vary one parameter and fix other parameters. For the capacity parameters (μ), we define cases with $\mu_1 = \mu_2$ and $\mu_1 \neq \mu_2$.

In our first experiment $\mu_1 = \mu_2 = 20$. For the other parameters, we use: $a = 50$, $b_1 = 4$, $b_2 = 6$, $m_1 = 3$, $m_2 = 2$, $\delta_2 = 3$, and $s = 0.95$. We can see in tables 5.14 and 5.15 that both lead times are the same. When b_1 increases, the price will decrease as the customers are more sensitive to price. As b_2 increases, customer are more sensitive to quoted lead time. Thus, the chain has no other option than reducing the quoted lead time.

We use $\frac{\Pi_c - \Pi_g}{\Pi_c} \times 100\%$ to calculate the loss with respect to the centralized setting. The Π_g^{Max} is the maximum Π_g where we calculate with MatLab optimization function. For the experiment with b_1, we get an average loss of 31.33% and for experiment with b_2 we get an average loss of 8.77%.

Table 5.14: b_1 for $\mu_1 = \mu_2$

b_1	Decentralized Setting - Upstream decide P_1									
	L_1	L_2	P_1	λ	Π_1	Π_2	Π_g	Π_g^{Max}	Π_c	Loss
2	0.6282	0.6282	10.6155	15.2310	115.9916	15.2310	131.2226	142.1375	149.3263	12.12%
4	0.3329	0.3329	5.7506	11.0023	30.2628	11.0023	41.2651	41.2901	43.5856	5.32%
6	0.2101	0.2101	3.9566	5.7396	5.4904	5.7396	11.2300	12.0349	12.8375	12.52%
8	0.1505	0.1505	3.0121	0.0969	0.0012	0.0969	0.0981	1.8940	2.1119	95.36%

Table 5.15: b_2 for $\mu_1 = \mu_2$

b_2	Decentralized Setting - Upstream decide P_1									
	L_1	L_2	P_1	λ	Π_1	Π_2	Π_g	Π_g^{Max}	Π_c	Loss
2	0.3855	0.3855	6.0572	12.2290	37.3871	12.2290	49.6161	49.7162	50.8456	2.42%
4	0.3557	0.3557	5.8943	11.5773	33.5085	11.5773	45.0858	45.0861	46.8844	3.84%
6	0.3329	0.3329	5.7506	11.0023	30.2628	11.0023	41.2651	41.2901	43.5856	5.32%
8	0.3147	0.3147	5.6205	10.4820	27.4683	10.4820	37.9503	38.0225	40.7125	6.78%
10	0.2997	0.2997	5.5008	10.0033	25.0164	10.0033	35.0197	35.1359	38.1487	8.20%
12	0.2869	0.2869	5.3894	9.5575	22.8363	9.5575	32.3938	32.5441	35.8255	9.58%
14	0.2758	0.2758	5.2846	9.1386	20.8785	9.1386	30.0171	30.1908	33.6976	10.92%
16	0.2661	0.2661	5.1856	8.7423	19.1070	8.7423	27.8493	28.0366	31.7334	12.24%
18	0.2575	0.2575	5.0913	8.3653	17.4946	8.3653	25.8599	26.0525	29.9092	13.54%
20	0.2497	0.2497	5.0013	8.0050	16.0201	8.0050	24.0251	24.2163	28.2071	14.83%

We want to see how the actors behave when we modify the margin δ_2. Results are given in table 5.16. As expected, the change in δ_2 doesn't change the Π_g^{Max} due to Π_g doesn't depend on δ_2. We see that when we increase δ_2, the downstream actor will receive more profit, and therefore the profit sharing between actors can be balanced with the right δ_2 ($\delta_2 = 4.2297$ for our example). However, the $\Pi_1 = \Pi_2$ will result in a lower Π_g compared to Π_g^{Max}. In our experiment, the Π_g^{Max} can be obtained using $\delta_2 = 3.138$.

Table 5.16: δ_2 for $\mu_1 = \mu_2$

δ_2	Decentralized Setting - Upstream decide P_1							
	L_1	L_2	P_1	λ	Π_1	Π_2	Π_g	Π_g^{Max}
2.0	0.4036	0.4036	6.1445	12.5782	39.5526	0.0000	39.5526	41.2901
2.5	0.3656	0.3656	5.9516	11.8063	34.8473	5.9032	40.7504	41.2901
3.0	0.3329	0.3329	5.7506	11.0023	30.2628	11.0023	41.2651	41.2901
3.138	0.3248	0.3248	5.6939	10.7755	29.0276	12.2625	41.2901	41.2901
3.5	0.3048	0.3048	5.5428	10.1712	25.8636	15.2569	41.1204	41.2901
4.0	0.2804	0.2804	5.3294	9.3174	21.7035	18.6348	40.3384	41.2901
4.2297	0.2703	0.2703	5.2297	8.9186	19.8855	19.8855	39.7711	41.2901
5.0	0.2407	0.2407	4.8889	7.5556	14.2719	22.6669	36.9387	41.2901

Our second experiment in this section is with $\mu_1 \neq \mu_2$. We use the base parameters as: $a = 50$, $b_1 = 4$, $b_2 = 6$, $m_1 = 3$, $m_2 = 2$, $\delta_2 = 3$, and $s = 0.95$. For $\mu_1 > \mu_2$ we use $\mu_1 = 30$ & $\mu_2 = 15$ and for $\mu_1 < \mu_2$ we use $\mu_1 = 15$ & $\mu_2 = 30$. We only display the result with $\mu_1 > \mu_2$. If readers are interested with the case $\mu_1 < \mu_2$, the readers can swap the value of L_1 and L_2 for each instances. The other values such as price, demand and profit stay the same. Logically, the higher capacity of the actor the lower the lead time. In the experiment with b_1, we get an average loss of 33.41% and in the experiments with b_2 we get an average loss of 7.62%.

Table 5.17: b_1 for $\mu_1 > \mu_2$

b_1	Decentralized Setting - Upstream decide P_1									
	L_1	L_2	P_1	λ	Π_1	Π_2	Π_g	Π_g^{Max}	Π_c	Loss
2	0.1791	1.7311	9.6347	13.2695	88.0395	13.2695	101.3090	129.5849	132.8462	23.74%
4	0.1539	0.6700	5.6321	10.5285	27.7125	10.5285	38.2411	38.8582	40.1800	4.83%
6	0.1231	0.3210	3.9446	5.6676	5.3537	5.6676	11.0213	11.6491	12.2250	9.85%
8	0.1002	0.2010	3.0121	0.0964	0.0012	0.0964	0.0976	1.8642	2.0419	95.22%

In the last part of the experiments, we want to see how the actors react to the increase of δ_2. Again, changing the δ_2 doesn't change the Π_g^{Max}. As we explained earlier in case $\mu_1 = \mu_2$, the curve of Π_g doesn't depend on δ_2. Again, we also see that the increase of δ_2 favors the leader (downstream actor). Changing the δ_2 also indicates that we change the proportion between Π_1 and Π_2. The Π_g^{Max} can be obtained using $\delta_2 = 3.6170$

Table 5.18: b_2 for $\mu_1 > \mu_2$

b_2	Decentralized Setting - Upstream decide P_1									
	L_1	L_2	P_1	λ	Π_1	Π_2	Π_g	Π_g^{Max}	Π_c	Loss
2	0.1653	0.9589	5.9690	11.8758	35.2588	11.8758	47.1346	47.3501	47.9264	1.65%
4	0.1588	0.7747	5.7833	11.1330	30.9861	11.1330	42.1192	42.5986	43.5859	3.37%
6	0.1539	0.6700	5.6321	10.5285	27.7125	10.5285	38.2411	38.8582	40.1800	4.83%
8	0.1498	0.5995	5.5007	10.0028	25.0141	10.0028	35.0170	35.6984	37.3022	6.13%
10	0.1463	0.5477	5.3825	9.5300	22.7050	9.5300	32.2350	32.9369	34.7828	7.32%
12	0.1433	0.5074	5.2740	9.0958	20.6834	9.0958	29.7792	30.4744	32.5300	8.46%
14	0.1406	0.4749	5.1729	8.6917	18.8863	8.6917	27.5780	28.2491	30.4873	9.54%
16	0.1381	0.4479	5.0779	8.3117	17.2712	8.3117	25.5829	26.2189	28.6163	10.60%
18	0.1359	0.4250	4.9880	7.9518	15.8079	7.9518	23.7597	24.3536	26.8898	11.64%
20	0.1338	0.4053	4.9022	7.6089	14.4739	7.6089	22.0828	22.6305	25.2873	12.67%

Table 5.19: δ_2 for $\mu_1 > \mu_2$

δ_2	Decentralized Setting - Upstream decide P_1							
	L_1	L_2	P_1	λ	Π_1	Π_2	Π_g	Π_g^{Max}
2.0	0.1641	0.9206	5.9365	11.7459	34.4914	0.0000	34.4914	38.8582
3.0	0.1539	0.6700	5.6321	10.5285	27.7125	10.5285	38.2411	38.8582
3.5	0.1484	0.5783	5.4549	9.8198	24.1069	14.7296	38.8365	38.8582
3.6170	0.1472	0.5595	5.4115	9.6459	23.2608	15.5974	38.8582	38.8582
4.0	0.1431	0.5042	5.2646	9.0583	20.5131	18.1166	38.6297	38.8582
4.189516	0.1410	0.4799	5.1895	8.7581	19.1759	19.1759	38.3519	38.8582
5.0	0.1327	0.3951	4.8542	7.4169	13.7525	22.2506	36.0032	38.8582

In general, if we compare the Π_g, Π_g^{Max}, and Π_c in all instances, we see that the differences are small. If we see the difference between Π_g and Π_g^{Max}, the decision taken (Π_g) is near the optimum profit (Π_g^{Max}). Furthermore, the peak of global profit Π_g^{Max} is close to Π_c. Thus, we can conclude that the natural coordination between actors, using this decision sequence, is much better compared to the previous decision sequence (Section 5.5.1).

5.6 Conclusion

We deal with a two-stage supply chain consisting of an upstream actor (supplier or manufacturer) and a downstream actor (manufacturer or retailer). We consider that both actors have a production process (lead time). We do analyses for centralized, and two decentralized settings decisions.

In the centralized setting, we consider downstream and upstream actors who work together to decide the global price (P_g) and global lead time (L_g). We successfully solve the centralized problem numerically.

With hypo-exponential distribution of sojourn time, the service constraint is complex, and solving analytically the centralized setting is very difficult. Thus, we

come out with an idea to decouple the global service constraint into local service constraints. We have proved that for $s \geq 0.715$, satisfying the local service constraints allows to satisfy the global service constraint. Then, we proposed a new model, where we transform the global service constraint into local service constraints for each actor. We call this model "Modified Centralized". Based on our experimentations, decouple the global service constraints leads to lower profit (in our experiments we get approximately 15% of loss compared to the centralized setting).

Thanks to this result, we have been able to model our problem into the decentralized setting. We divide the decentralized setting into two scenarios. In both scenarios, the downstream actor acts as a leader and the upstream actor acts as a follower. In the first scenario, the upstream actor chooses his own lead time (L_1) but the price P_1 and lead time L_2 are decided by the downstream actor (actor 2). In the second scenario, the upstream actor (actor 1) decides his own price (P_1) and downstream actor (actor 2) decides the global lead time ($L_1 + L_2$) and L_1. The global price (P_g) is obtained from P_1 plus a fixed margin taken by the downstream actor (δ_2). We solve the decentralized problems analytically and we provide some experiments.

In the first scenario (upstream actor decides his own lead time), from the experiments we see that the global profit is very low compared to the centralized setting. We also see that the profit of upstream actor is null. It shows that the actors cannot cooperate naturally. However, if we observe the comparison between maximum global profit and profit in centralized setting, we see that the difference is small. This signifies that the decentralized setting approach could provide a favorable global profit in condition that the proposed coordination system could lead the actors into this situation.

In second scenario (upstream actor decides his own price), the global profit curve doesn't depend on the downstream actor's margin (δ_2). We see that the increase in δ_2 favors the leader (downstream actor). Changing the δ_2 also indicates that we change the proportion of profit for each actor. The maximum global price can be obtained numerically.

If we compare all scenarios: centralized, modified centralized, and two decentralized settings, we can naturally see that the best setting for maximizing the global profit is centralized setting. However, in most supply chain actors have a given autonomy. So in decentralized schemes, we suggest the upstream actor chooses his own price. This will result in a better profit for both actors and for the profit global of the chain compared to the other scheme (upstream actors chooses the lead-time).

Chapter 6

General conclusion

6.1 Conclusion

From the literature review, we see at least three weaknesses. First, all works assume that the unit production cost is a constant. Second, in single firm case all papers use the M/M/1 system. Although M/M/1 is the simplest queuing model, the fact that all customers are accepted might lead to long sojourn times (lead time) in the system. Third, in multi-firm all contributions assume that only one of the actor has production process, the other actor only acts as a mediator with a lead-time equals to zero.

Thus, in this study, we propose three extensions: 1. Lead-time sensitive production cost, 2. Customer's rejection policy using an M/M/1/K model, and 3. Two-stage supply chain coordination using a tandem queue (M/M/1-M/M/1) model. In the first contribution, we consider the production cost to be a decreasing function in lead time. Indeed, a company can use different ways to reduce lead time (such as buying items from quick response but expensive subcontractors) but this generally leads to higher production cost. The second contribution, we model the firm as an M/M/1/K queue, for which demand is rejected if there are already K customers in the system. The last contribution is modeling the system in a tandem queue. It is more realistic when a supply chain consists of more than one stage and in each stage there is a production process.

In the first contribution, we solved the problem of lead-time quotation in an M/M/1 make-to-order queue while assuming the production cost to be a decreasing function in lead-time. We considered three settings: (1) lead-time is variable but price is fixed, (2) price and lead-time are both decision variables, and (3) price and lead-time are decision variables where the lateness and holding cost are considered. We solve settings 1 and 2 analytically; and setting 3 numerically. We conducted experiments and derived interesting insights.

In case of variable lead time and fixed cost (setting 1), some of our insights indicated that the higher the demand sensitivity to lead time the higher the service level. This results is not intuitive since an increase in lead time sensitivity leads to shorter lead time that is supposed to be more difficult to guarantee. This behavior cannot be captured by the existing models where the unit operating cost is assumed constant since the service constraint is always binding in this case.

When both lead time and price are endogenous variables (Setting 2), our results showed that an increase in the demand sensitivity to price leads to an increase in quoted lead time in our model while it has the opposite effect for models with

constant cost. We also found that when demand is very sensitive to lead time, the customers can benefit of smaller price, shorter lead time, and also more reliable deliveries.

When we consider the lateness penalty and holding costs (Setting 3), the model quotes a longer lead time higher than the one imposed by the service constraint. The model prefers to be in non-binding situation. Indeed, quoting a longer lead time is favorable as it will reduce the incurred costs (i.e. production and lateness costs).

From our numerical experiments, we quantified the gain brought by using the solution of our model versus the solution of the benchmark model where the cost is constant. In the case where the price is fixed (setting 1), we found that our model leads to small gains. This proves that there is an impact of not quoting the right lead time when the cost is assumed to be constant. This impact becomes more significant (as the gains become bigger) when we take into account the price as a decision variable (setting 2). And when the lateness and holding costs are included, this gains increase compared to the settings 1 and 2.

In the second contribution, we formulated the problem of lead time quotation and pricing for a profit-maximizing firm, modeled as an M/M/1/K system, facing a linear price- and lead time-dependent demand with the consideration of inventory holding and lateness penalty costs. In order to determine the lateness penalty cost, we knocked out a theoretical barrier by explicitly calculating the expected lateness given that a job is late in an M/M/1/K queue when a certain delivery lead time is quoted to the customers. This result can be used in the future for different operations management and queuing theory problems.

In M/M/1/1 queue, the expression of the optimal solution showed that an increase in lead time-sensitivity first leads to reducing the quoted lead time but can rapidly become useless if it exceeds a certain threshold value (since the service constraint becomes binding). We also deduced that when the customers become more sensitive to price or when the unit lateness penalty cost increases, the firm can react by increasing the quoted lead time. Based on our comparison of the optimal profit given by our M/M/1/1 model to the optimal profit obtained when the firm is modeled as an M/M/1 queue (as given in the literature), we found out that a rejection policy can be more profitable than an all-customers' acceptance policy even when the holding and penalty costs are not considered. Some of our results showed that an increase in lead-time sensitivity or in price-sensitivity favors the rejection policy. The increase in unit holding cost has been proven to be one of the main criteria that make the rejection policy better than the all-customers' acceptance policy.

Then, we solved numerically the M/M/1/K queue ($K > 1$). We showed that there is at least one value of K for which the M/M/1/K (rejection policy) is more profitable than the M/M/1 (all customers' acceptance policy). In most cases, it has also been observed that an increase in the value of K (i.e., the system size) has a non-monotonous effect on the firm's profit. Indeed, an increase in K first improves the profit and then leads to decreasing it.

In the last contribution, we deal with a two-stage supply chain consisting of an upstream actor (supplier or manufacturer) and a downstream actor (manufacturer or retailer). We consider that both actors have production process (lead time). We do analyses for centralized, and two decentralized settings decisions.

In the centralized setting, we consider downstream and upstream actors who work together to decide the global price (P_g) and global lead time (L_g). We successfully solve the centralized problem numerically.

With hypo-exponential distribution of sojourn time, the service constraint is complex, and solving analytically the centralized setting is very difficult. Thus, we come out with an idea to decouple the global service constraint into local service constraints. We have proved, from a given value of service level, that satisfying the local service constraints allows to satisfy the global service constraint. We proposed a new model with these new constraints, called "Modified Centralized". Based on our experimentations, decouple the global service constraints naturally leads to lower profit, but the difference is not very high.

Thanks to the decoupling result, we have been able to model our problem into the decentralized setting. We divide the decentralized setting into two scenarios. In both scenarios, the downstream actor acts as a leader and the upstream actor acts as a follower. In the first scenario, the upstream actor chooses his own lead time (L_1) but the price P_1 and lead time L_2 are decided by the downstream actor (actor 2). In the second scenario, the upstream actor (actor 1) decides his own price (P_1) and downstream actor (actor 2) decides the global lead time ($L_1 + L_2$) and L_1. We solve the decentralized problems analytically and we provide some experiments.

If we compare all scenarios: centralized, modified centralized, and two decentralized settings, we can naturally see that the best setting for maximizing the global profit is centralized setting. However, most supply chains nowadays are decentralized. So in decentralized schemes, we suggest the upstream actor chooses his own price. This will result in a better profit for both actors and for the profit global of the chain compared to the other scheme (upstream actors chooses the lead-time).

6.2 Future works and perspectives

Our study can be extended in different ways. For instance, it would be interesting to investigate the log-linear model of demand. Another extension of our tandem queue model would be to inverse the role of leader-follower. We could consider the upstream acts as a leader and downstream acts as a follower. Other extensions could be: to introduce sharing bonus scheme for the decentralized setting where upstream actor chooses his own lead-time (this will avoid the zero profit of upstream actor). One could also consider a competition between actors in the multi stage supply chains. We could also consider a new stream of research which considers a dynamic lead time quotation.

Bibliography

Albana, A. S., Y. Frein, and R. Hammami
2016. Optimal firm's policy under lead time- and price-dependent demand: interest of customers rejection policy. In *POMS 27th Annual Conference*, Orlando, Florida.

Albana, A. S., R. Hammami, and Y. Frein
2017a. Expected lateness in an M/M/1/K queue. *Working paper*. `http://hal.univ-grenoble-alpes.fr/hal-01626006`.

Albana, A. S., R. Hammami, and Y. Frein
2017b. Impact of lead-time sensitive cost on lead-time quotation and pricing problems. In *7th IESM Conference*, Pp. 474–479, Saarbrücken, Germany.

Albana, A. S., R. Hammami, and Y. Frein
2017c. Lead time quotation and pricing in a finite buffer queue with endogenous demand: Effect of customers' rejection and buffer capacity sizing. *Working paper*.

Albana, A. S., R. Hammami, and Y. Frein
2017d. Lead time quotation and pricing in a stochastic make-to-order system with lead time-dependent cost and endogenous demand. Submitted to *International Journal of Production Economics*.

Baker, W., M. Marn, and C. Zawada
2001. Price smarter on the net. *Harvard business review*, 79(2):122–7.

Ballou, R. H.
1998. *Business Logistics Management*, 4th edition. Upper Saddle River, New Jersey: Prentice-Hall.

Blackburn, J. D.
1991. *Time-based competition: the next battleground in American manufacturing.* Homewood, Illinois: Irwin Professional Pub.

Blackburn, J. D., T. Elrod, W. B. Lindsley, and A. J. Zahorik
1992. The strategic value of response time and product variety. *Manufacturing Strategy – Process and Content. Chapman and Hall, London.*

Bolch, G., S. Greiner, H. de Meer, and K. S. Trivedi
2006. *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*, 2nd edition. Hoboken, New Jersey: John Wiley & Sons.

Boyaci, T. and S. Ray
2003. Product differentiation and capacity cost interaction in time and price sensitive markets. *Manufacturing & Service operations management*, 5(1):18–36.

Boyaci, T. and S. Ray
2006. The impact of capacity costs on product differentiation in delivery time, delivery reliability, and price. *Production and Operations Management*, 15(2):179–197.

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

Christopher, M.
2011. *Logistics & supply chain management*, 4th edition. Harlow: FT Prentice Hall.

Geary, S. and J. P. Zonnenberg
2000. What it means to be best in class. *Supply Chain Management Review, V.4, NO.3 (July/Aug. 2000), P. 43-48: ILL.*

Gross, D., J. F. Shortle, J. M. Thompson, and C. M. Harris
2008. *Fundamentals of Queueing Theory*, 4th edition. Hoboken, New Jersey: John Wiley & Sons.

Hafızoğlu, A. B., E. S. Gel, and P. Keskinocak
2016. Price and lead time quotation for contract and spot customers. *Operations Research*, 64(2):406–415.

Hammami, R. and Y. Frein
2013. An optimisation model for the design of global multi-echelon supply chains under lead time constraints. *International Journal of Production Reseach*, 51(9):2760–2775.

Hillier, F. and G. Lieberman
2001. *Introduction To Operations Research*, 7th edition. New York: McGraw Hill.

Ho, T. H. and Y.-S. Zheng
2004. Setting customer expectation in service delivery: An integrated marketing-operations perspective. *Management Science*, 50(4):479–488.

Hotelling, H.
1929. Stability in competition. *The economic journal*, 39(153):41–57.

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

Hum, S.-H. and H.-H. Sim
1996. Time-based competition: literature review and implications for modelling. *International Journal of Operations & Production Management*, 16(1):75–90.

Irving, R.
2013. *Beyond the quadratic formula*. Washington: Mathematical Association of America.

Jackson, D. W., J. E. Keith, and R. K. Burdick
1986. Examining the relative importance of physical distribution service elements. *Journal of Business Logistics*, 7(2).

Kapuscinski, R. and S. Tayur
2007. Reliable due-date setting in a capacitated mto system with two customer classes. *Operations research*, 55(1):56–74.

Kleinrock, L.
1975. *Queueing systems*, volume 1. New York: Wiley.

Li, L.
1992. The role of inventory in delivery-time competition. *Management Science*, 38(2):182–197.

Liu, L., M. Parlar, and S. X. Zhu
2007. Pricing and lead time decisions in decentralized supply chains. *Management Science*, 53(5):713–725.

Palaka, K., S. Erlebacher, and D. H. Kropp
1998. Lead-time setting, capacity utilization, and pricing decisions under lead-time dependent demand. *IIE transactions*, 30(2):151–163.

Panda, S.
2013. Coordinating a manufacturer–retailer chain under time and price dependent demand rate. *International Journal of Management Science and Engineering Management*, 8(2):84–92.

Pekgün, P., P. M. Griffin, and P. Keskinocak
2008. Coordination of marketing and production for price and leadtime decisions. *IIE transactions*, 40(1):12–30.

Pekgün, P., P. M. Griffin, and P. Keskinocak
2016. Centralized versus decentralized competition for price and lead-time sensitive demand. *Decision Sciences*.

Ray, S. and E. M. Jewkes
2004. Customer lead time management when both demand and price are lead time sensitive. *European Journal of operational research*, 153(3):769–781.

Savaşaneril, S., P. M. Griffin, and P. Keskinocak
2010. Dynamic lead-time quotation for an M/M/1 base-stock inventory queue. *Operations research*, 58(2):383–395.

Savaşaneril, S. and E. Sayin
2017. Dynamic lead time quotation under responsive inventory and multiple customer classes. *OR Spectrum*, 39(1):95–135.

Schechter, E.
2013. The Cubic Formula - Solve Any 3rd Degree Polynomial Equation. `https://math.vanderbilt.edu/schectex/courses/cubic/`.

Slotnick, S. A.
2014. Lead-time quotation when customers are sensitive to reputation. *International Journal of Production Research*, 52(3):713–726.

So, K. C.
2000. Price and time competition for service delivery. *Manufacturing & Service Operations Management*, 2(4):392–409.

So, K. C. and J.-S. Song
1998. Price, delivery time guarantees and capacity selection. *European Journal of operational research*, 111(1):28–49.

Stalk, G. and T. M. Hout
1990. *Competing against time: How time-based competition is reshaping global mar*. New York: Free Press.

Sterling, J. U. and D. M. Lambert
1989. Customer service research: past, present and future. *International journal of physical distribution & materials management*, 19(2):2–23.

Suri, R.
1998. *Quick response manufacturing: a companywide approach to reducing lead times*. Portland, Oregon: Productivity Press.

Sztrik, J.
2016. *Basic Queueing Theory: Foundations of System Performance Modeling*. Saarbrücken, Germany: GlobeEdit.

Thomopoulos, N. T.
2012. *Fundamentals of Queuing Systems: Statistical Methods for Analyzing Queuing Models*. New York: Springer Science & Business Media.

Wu, Z., B. Kazaz, S. Webster, and K.-K. Yang
2012. Ordering, pricing, and lead-time quotation under lead-time and demand uncertainty. *Production and Operations Management*, 21(3):576–589.

Xiao, T. and X. Qi
2016. A two-stage supply chain with demand sensitive to price, delivery time, and reliability of delivery. *Annals of Operations Research*, 241(1-2):475–496.

Xiao, T. and J. Shi
2012. Price, capacity, and lead-time decisions for a make-to-order supply chain with two production modes. *International Journal of Applied Management Science*, 4(2):107–129.

Xiao, T., J. Shi, and G. Chen
2014. Price and leadtime competition, and coordination for make-to-order supply chains. *Computers & Industrial Engineering*, 68:23–34.

Xiao, T., D. Yang, and H. Shen
2011. Coordinating a supply chain with a quality assurance policy via a revenue-sharing contract. *International Journal of Production Research*, 49(1):99–120.

Xiaopan, L., W. Jianjun, Z. Binghang, and Z. Zongbao
2014. Price and guaranteed delivery time competition/cooperation in a duopoly. In *Service Systems and Service Management (ICSSSM), 2014 11th International Conference on*, Pp. 1–6. IEEE.

Zhao, X., K. E. Stecke, and A. Prasad
2012. Lead time and price quotation mode selection: uniform or differentiated? *Production and Operations Management*, 21(1):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.

Appendix A

Root of cubic equation Q(L) in Proposition 3.3

This appendix is intended to find the roots of the cubic equation $Q(L)$ in proposition 3.3 (optimal solution for case with both lead-time and price as decision variables). Recall that $\Delta = (a - C_1b_1 - 2\mu)^2 + 4b_2(2\gamma - C_2b_1)$.

In the case of $\Delta > 0$, we have: $g_2(L) = \frac{(\gamma-\mu L)(b_2L^2+(\mu-a+C_1b_1)l+C_2b_1-\gamma)}{b_1L^2}$. The first derivative $\frac{d}{dL}g_2(L) = g_2'(L) = 0$ is:

$$\frac{2\gamma(\gamma - C_2b_1) + (a\gamma - 2\mu\gamma + \mu C_2b_1 - \gamma C_1b_1)L - \mu b_2L^3}{b_1L^3} = 0$$

Knowing that $b_1L^3 \neq 0$, it is equivalent to :

$$Q(L) = 2\gamma(\gamma - C_2b_1) + (a\gamma - 2\mu\gamma + \mu C_2b_1 - \gamma C_1b_1)l - \mu b_2l^3 = 0$$

The cubic equation $Q(L)$ has at minimum one real root. Its first real roots can be found using Cardano's formula (see Irving, 2013) which is:

$$R_1 = \sqrt[3]{\frac{\gamma(\gamma - C_2b_1)}{b_2\mu} + \sqrt{\left(\frac{\gamma(\gamma - C_2b_1)}{b_2\mu}\right)^2 - \frac{(a\gamma - 2\mu\gamma + \mu C_2b_1 - \gamma C_1b_1)^3}{27(\mu b_2)^3}}} + \sqrt[3]{\frac{\gamma(\gamma - C_2b_1)}{b_2\mu} - \sqrt{\left(\frac{\gamma(\gamma - C_2b_1)}{b_2\mu}\right)^2 - \frac{(a\gamma - 2\mu\gamma + \mu C_2b_1 - \gamma C_1b_1)^3}{27(\mu b_2)^3}}}$$

The discriminant of the cubic equation $Q(L)$ is:

$$4(\mu b_2)(a\gamma - 2\mu\gamma + \mu C_2b_1 - \gamma C_1b_1) - 3(\mu b_2)^2R_1^2$$

The next step of the calculation depends on whether the discriminant is positive or not.

If the discriminant is positive, the other two real roots can be found by factorization:

$$\begin{aligned} &2\gamma(\gamma - C_2b_1) + (a\gamma - 2\mu\gamma + \mu C_2b_1 - \gamma C_1b_1)L - \mu b_2L^3 = 0 \\ &\Leftrightarrow (L - R_1)(-\mu b_2L^2 - (\mu b_2R_1)L + a\gamma - 2\mu\gamma + \mu C_2b_1 - \gamma C_1b_1 - (\mu b_2){R_1}^2 = 0 \end{aligned}$$

The other two real roots are:

$$R_2 = \frac{-\mu b_2 R_1 + \sqrt{4(\mu b_2)(a\gamma - 2\mu\gamma + \mu C_2 b_1 - \gamma C_1 b_1) - 3(\mu b_2)^2 {R_1}^2}}{2\mu b_2}$$

$$R_3 = \frac{-\mu b_2 R_1 - \sqrt{4(\mu b_2)(a\gamma - 2\mu\gamma + \mu C_2 b_1 - \gamma C_1 b_1) - 3(\mu b_2)^2 {R_1}^2}}{2\mu b_2}$$

R_3 is always negative, thus it is infeasible.
Given the product of the roots: $R_1 R_2 R_3 = \frac{2\gamma(\gamma - C_2 b_1)}{\mu b_2}$ with $\frac{2\gamma}{\mu b_2} > 0$, thus:
If $\gamma > C_2 b_1$, we have one positive root (R_1 or R_2).
Otherwise, we have two positive roots (R_1 and R_2).

If the discriminant is negative, then there is only one real root: R_1.
If $R_1 \geq 0$, then it is the solution. Otherwise problem is infeasible.

Appendix B

Expected lateness in a M/M/1/K

Given a quoted lead time, L. We denote by R_L the expected lateness given that a job is late in an M/M/1/K queueing system with mean service rate, μ, and mean arrival rate, λ. We let W denote the sojourn time (waiting time in the system) with probability density function $f_W(.)$ and cumulative distribution function $F_W(.)$. Note that W is exponentially distributed with mean $N_s/\bar{\lambda}$ where $N_s = \dfrac{\rho}{1-\rho} - \dfrac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}$ and $\rho = \dfrac{\lambda}{\mu}$. Our objective is to calculate R_L.

R_L can be given by the following integral function: $\int_L^{\infty}(t-L)f_{W|W\geq L}(t)dt$, where $f_{W|W\geq L}(t)$ is the probability density function of having a sojourn time W given that W is greater than the quoted lead time L. For clarity of presentation, we first calculate $f_{W|W\geq L}(t)$ and then calculate the integral function R_L.

Calculation of $f_{W|W\geq L}(t)$

We have $f_{W|W\geq L}(t) = \dfrac{d}{dt}F_{W|W\geq L}(t)$ and $F_{W|W\geq L}(t) = \dfrac{F_W(t)-F_W(L)}{1-F_W(L)}$. It is known that in M/M/1/K, $F_W(x) = 1 - \sum_{k=0}^{K-1}\dfrac{P_k}{1-P_K}\left(\sum_{i=0}^{k}\dfrac{(\mu x)^i}{i!}e^{-\mu x}\right)$ where $P_k = \dfrac{1-\rho}{1-\rho^{K+1}}\rho^k$ if $\rho \neq 1$, and $P_k = \dfrac{1}{K+1}$ if $\rho = 1$ (see Gross et al., 2008; Kleinrock, 1975; Thomopoulos, 2012). Thus, we obtain by standard calculus:

$$
\begin{aligned}
F_{W|W\geq L}(t) &= \frac{\sum_{k=0}^{K-1}\frac{P_k}{1-P_K}\left(\sum_{i=0}^{k}\frac{(\mu L)^i}{i!}e^{-\mu L} - \frac{(\mu t)^i}{i!}e^{-\mu t}\right)}{\sum_{k=0}^{K-1}\frac{P_k}{1-P_K}\left(\sum_{i=0}^{k}\frac{(\mu L)^i}{i!}e^{-\mu L}\right)} \\
&= 1 - e^{-\mu(t-L)}\frac{\sum_{k=0}^{K-1}P_k\left(\sum_{i=0}^{k}\frac{(\mu t)^i}{i!}\right)}{\sum_{k=0}^{K-1}P_k\left(\sum_{i=0}^{k}\frac{(\mu L)^i}{i!}\right)}
\end{aligned}
$$

$$
\begin{aligned}
f_{W|W\geq L}(t) &= \frac{d}{dt}F_W(t|t\geq L) \\
&= \frac{1}{\sum_{k=0}^{K-1} P_k\left(\sum_{i=0}^{k}\frac{(\mu L)^i}{i!}\right)}\left[\mu e^{-\mu(t-L)}\sum_{k=0}^{K-1}P_k\left(\sum_{i=0}^{k}\frac{(\mu t)^i}{i!}\right)\right. \\
&\quad \left. -e^{-\mu(t-L)}\sum_{k=0}^{K-1}P_k\left(\sum_{i=0}^{k}\frac{\mu(\mu t)^{i-1}}{(i-1)!}\right)\right] \\
&= \frac{\mu e^{-\mu(t-L)}\sum_{k=0}^{K-1}P_k\left(\left(\sum_{i=1}^{k}\left(\frac{(\mu t)^i}{i!}-\frac{(\mu t)^{i-1}}{(i-1)!}\right)\right)+1\right)}{\sum_{k=0}^{K-1}P_k\left(\sum_{i=0}^{k}\frac{(\mu L)^i}{i!}\right)}
\end{aligned}
$$

Given that $\sum_{i=1}^{k}\left(\frac{(\mu t)^i}{i!}-\frac{(\mu t)^{i-1}}{(i-1)!}\right)=\frac{(\mu t)^k}{k!}-1$, we deduce:

$$
f_{W|W\geq L}(t)=\frac{\mu e^{-\mu(t-L)}}{\sum_{k=0}^{K-1}P_k\left(\sum_{i=0}^{k}\frac{(\mu L)^i}{i!}\right)}\sum_{k=0}^{K-1}P_k\frac{(\mu t)^k}{k!}
$$

Back to the expression of R_L

Consequently, we have:

$$
\begin{aligned}
R_L &= \int_L^\infty (t-L)f_{W|W\geq L}(t)dt = \int_L^\infty \frac{(t-L)\mu e^{-\mu(t-L)}}{\sum_{k=0}^{K-1}P_k\left(\sum_{i=0}^{k}\frac{(\mu L)^i}{i!}\right)}\sum_{k=0}^{K-1}P_k\frac{(\mu t)^k}{k!}dt \\
&= \frac{1}{\sum_{k=0}^{K-1}P_k\left(\sum_{i=0}^{k}\frac{(\mu L)^i}{i!}\right)}\sum_{k=0}^{K-1}\frac{P_k}{k!}\int_L^\infty (t-L)\mu e^{-\mu(t-L)}(\mu t)^k dt
\end{aligned}
$$

We let $I=\int_L^\infty (t-L)\mu e^{-\mu(t-L)}(\mu t)^k dt$. We need to calculate I in order to find the

expression of R_L.

$$
\begin{aligned}
I &= \int_L^\infty t\mu e^{-\mu(t-L)}(\mu t)^k dt - \int_L^\infty L\mu e^{-\mu(t-L)}(\mu t)^k dt \\
&= \frac{e^{\mu L}}{\mu}\int_L^\infty \mu e^{-\mu t}(\mu t)^{k+1} dt - Le^{\mu L}\int_L^\infty \mu e^{-\mu t}(\mu t)^k dt \\
&= \frac{e^{\mu L}}{\mu}N_{k+1} - Le^{\mu L}N_k \text{ with } N_k = \int_L^\infty \mu e^{-\mu t}(\mu t)^k dt
\end{aligned}
$$

Thus, $R_L = \dfrac{1}{\sum_{k=0}^{K-1} P_k\left(\sum_{i=0}^{k}\frac{(\mu L)^i}{i!}\right)}\sum_{k=0}^{K-1}\dfrac{P_k}{k!}\left(\dfrac{e^{\mu L}}{\mu}N_{k+1} - Le^{\mu L}N_k\right)$. In order to calculate R_L, we need now to calculate N_k.

Calculation of N_k

Lemma B.1. $N_k = \int_L^\infty \mu e^{-\mu t}(\mu t)^k dt = \left(\sum_{i=0}^{k}\frac{k!}{i!}(\mu L)^i\right)e^{-\mu L}$

Proof. We demonstrate the lemma by recursive induction. For $k = 0$, we verify that $N_0 = \int_L^\infty \mu e^{-\mu t}dt = \left[-e^{-\mu t}\right]_L^\infty = e^{-\mu L} = \left(\sum_{i=0}^{0}\frac{0!}{i!}(\mu l)^i\right)e^{-\mu L}$. Assume that $N_k = \left(\sum_{i=0}^{k}\frac{k!}{i!}(\mu l)^i\right)e^{-\mu L}$ and let's demonstrate that $N_{k+1} = \left[\sum_{i=0}^{k+1}\frac{(k+1)!}{i!}(\mu l)^i\right]e^{-\mu L}$.

We have $N_{k+1} = \int_L^\infty \mu e^{-\mu t}(\mu t)^{k+1}dt$. We use partial integration with $f'(t) = \mu e^{-\mu t}$ and $g(t) = (\mu t)^{k+1}$ to transform the expression of N_{k+1}.

Thus, $N_{k+1} = \left[-e^{-\mu t}(\mu t)^{k+1}\right]_L^\infty + (k+1)\int_L^\infty \mu e^{-\mu t}(\mu t)^k dt = \left[-e^{-\mu t}(\mu t)^{k+1}\right]_L^\infty + (k+1)N_k$. Given that $\lim_{t\to\infty} e^{-\mu t}(\mu t)^{k+1} = 0$, we deduce that $N_{k+1} = e^{-\mu L}(\mu L)^{k+1} + (k+1)N_k$.

Since we assumed $N_k = \left(\sum_{i=0}^{k} \frac{k!}{i!}(\mu L)^i\right) e^{-\mu L}$, we deduce that:

$$\begin{aligned}
N_{k+1} &= e^{-\mu L}(\mu L)^{k+1} + (k+1)\left(\sum_{i=0}^{k} \frac{k!}{i!}(\mu L)^i\right) e^{-\mu L} \\
&= e^{-\mu L}\left[(\mu L)^{k+1} + (k+1)\left(\sum_{i=0}^{k} \frac{k!}{i!}(\mu L)^i\right)\right] \\
&= e^{-\mu L}\left[(\mu L)^{k+1} + \left(\sum_{i=0}^{k} \frac{(k+1)!}{i!}(\mu L)^i\right)\right] \\
&= e^{-\mu L}\left[\sum_{i=0}^{k+1} \frac{(k+1)!}{i!}(\mu L)^i\right],
\end{aligned}$$

which demonstrate the lemma. ■

Back to the expression of R_L

We have established that $R_L = \frac{1}{\sum_{k=0}^{K-1} P_k \left(\sum_{i=0}^{k} \frac{(\mu L)^i}{i!}\right)} \sum_{k=0}^{K-1} \frac{P_k}{k!}\left(\frac{e^{\mu L}}{\mu} N_{k+1} - L e^{\mu L} N_k\right)$.

Given the result of the previous Lemma, we deduce that $\frac{e^{\mu L}}{\mu} N_{k+1} - L e^{\mu L} N_k = \sum_{i=0}^{k+1} \frac{(k+1)!}{i!}\frac{(\mu L)^i}{\mu} - \sum_{i=0}^{k} \frac{k!}{i!}(\mu L)^i L = \frac{(\mu L)^{k+1}}{\mu} + \left(\frac{k+1}{\mu} - L\right)\sum_{i=0}^{k} \frac{k!}{i!}(\mu L)^i$. Consequently, we finally conclude that:

$$R_L = \int_L^{\infty} (t-L) f_{W|W\geq L}(t)dt = \frac{\sum_{k=0}^{K-1} \frac{P_k}{k!}\left[\frac{(\mu L)^{k+1}}{\mu} + \left(\frac{k+1}{\mu} - L\right)\sum_{i=0}^{k} \frac{k!}{i!}(\mu L)^i\right]}{\sum_{k=0}^{K-1} P_k \left(\sum_{i=0}^{k} \frac{(\mu L)^i}{i!}\right)}$$

Appendix C

Particle Swarm Optimization for M/M/1/K

In this appendix, we present the Particle Swarm Optimization (PSO) heuristic which has been used to solve the general problem with M/M/1/K (chapter 4). The algorithm is the following.

1. Initial iteration ($i = 1$)

   (a) *Decide the lower and upper bound for lead time (L) and demand (λ) for this initial iteration*
   For lead time (L), $L_{min} = 0$ and $L_{max} = (a/b_2)$ and for demand (λ), $\lambda_{min} = 0$ and upper bound $\lambda_{max} = a$.

   (b) *Generate n individuals, with a uniform random process*
   Each individual represents a lead time and a demand. The lead time and demand have to satisfy the conditions of step 1.(a): $x_{i,j} = [L_{j,1}, \lambda_{j,1}]$ with $L_{min} \leq L_{j,i} \leq L_{max}$ and $\lambda_{min} \leq \lambda_{j,i} \leq \lambda_{max}$ and $j = 1, \ldots, n$. Note that $x_{j,i}$ refers to individual j for iteration 1.

   (c) *Calculate the objective function for each individual*
   The objective function $f(x_{j,i})$ is calculated based on equation (4.25), for $j = 1, \ldots, n$.

   (d) *Generate a velocity matrix*
   For each individual, the velocity is a random number between 0 and 1: $v_{j,1} = [v_{L_j,1} \;\; v_{\lambda_j,1}]$ for$j = 1, \ldots, n$ where $v_{L_j,1}$, represents the velocity for lead time of j^{th} individual on iteration i and $v_{\lambda_j,1}$, represents the velocity for demand of j^{th} individual on iteration i.

   (e) *Find the personal best position for each individual ($Pbest_{j,i}$)*
   $Pbest_{j,i}$ refers to the best L and λ that give us the currently known maximum profit for j^{th} individual in iteration i. In the first iteration ($i = 1$), the best position for each individual is its initial position. Thus, $Pbest_{j,i} = x_{j,i} = [L_{j,i} \;\; \lambda_{j,1}]$ for $j = 1, \ldots, n$.

   (f) *Find the global best position for all individuals ($Gbest_i$)*
   $Gbest_i$ is the best position among all individuals in iteration i. Thus, $\max\{f(Pbest_{1,i}), f(Pbest_{2,i}), \ldots, f(Pbest_{n,i})\}$ will give us ${L_i}^*$ and ${\lambda_i}^*$ in iteration i, and then $Gbest_i = [{L_i}^* \;\; {\lambda_i}^*]$. For the first iteration, $Gbest_1 = [{L_1}^* \;\; {\lambda_1}^*]$.

2. Iterations i

   (a) *While stopping conditions are not satisfied do*
   The stopping condition for our algorithm is the number of iterations. We use 10,000 iterations because no significant improvement is obtained for 100,000 iterations.

      i. *Update the velocity for each individual*
      $v_{j,i+1} = v_{j,i} + r_1(Pbest_{j,i} - x_{j,i}) + r_2(Gbest_i - x_{j,i})$ for $j = 1, 2, \ldots, n$, where r_1 and r_2 are random numbers on interval [0,1].

      ii. *Update the position for each individual*
      $x_{j,i+1} = x_{j,i} + v_{j,i+1}$ for $j = 1, 2, \ldots, n$.

      iii. *Calculate the objective function for each individual*
      We use equation (4.25) and obtain $f(x_{j,i+1})$ for $j = 1, 2, \ldots, n$.

      iv. *Update the personal best position for each individual*
      If $f(x_{j,i+1}) > f(x_{j,i})$ then $Pbest_{j,i+1} = x_{j,i+1}$
      Else $Pbest_{j,i+1} = x_{j,i}$
      for $j = 1, 2, \ldots, n$.

      v. *Update the global best position for all individuals*
      We calculate $\max\{f(Pbest_{1,i+1}), f(Pbest_{2,i+1}), \ldots, f(Pbest_{n,i+1})\}$ and deduce ${L_{i+1}}^*$ and ${\lambda_{i+1}}^*$. $Gbest_{i+1} = [{L_{i+1}}^* \ {\lambda_{i+1}}^*]$.

      vi. *Back to step 2.(a).i*

Appendix D
Experiment with K^{opt}

We did some experiments with K equals 1 to 10 (note that $K = +\infty$ leads to M/M/1 model) with the base case parameters of section 4.4 in chapter 4. As expected, the M/M/1/1 result is worse than M/M/1. When K increase, the profit increases (up to certain point) then it will decrease. When K goes to $+\infty$, the profit of M/M/1/K will be closer to the M/M/1 profit (see figure D.1).

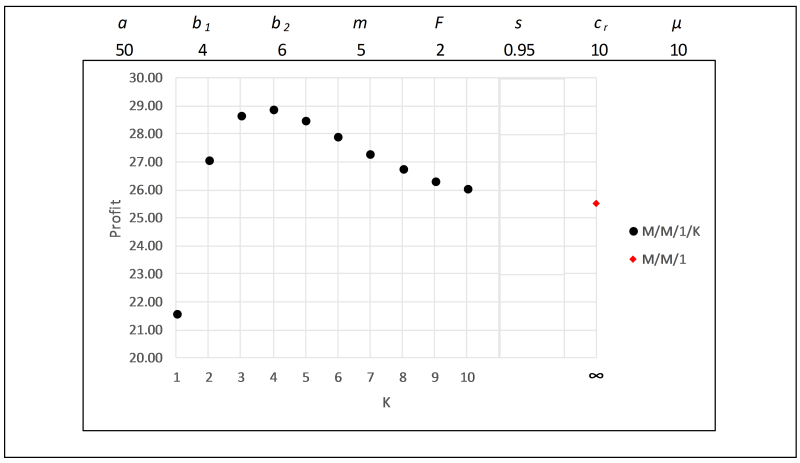


Figure D.1: Profit in function of K

In figure D.2, we did some experiments with the b_2 and K^{opt}. As expected, when the clients are very sensitive to lead time, the firm reacts by accepting less clients. This happen as an attempt to reduce the waiting time, and avoid having a high lateness penalty cost.

In table D.1, we provide the detailed results of the optimum profit for each value of b_2 with $K = 1 \ldots 10$. For each value of b_2, we observe the same behavior of the profit in function of K as what we observed in figure D.1. The profit increases then decreases after a certain value of K. When the client isn't sensitive to lead time (b_2 goes to zero), firm reacts by accepting many clients. Although the demand isn't affected by the quoted lead time when b_2 is null, the lateness penalty cost is still affected by the number of client accepted. Thus, to reduce the penalty cost, firms will only accept certain number of client which they see profitable.

In figure D.3, we did some experiment to find K^{opt} for each value of price sensitivity (b_1). As b_1 increases, clients are more sensitive to price. When b_1 goes to 0, firms can quote any price which is profitable. In addition, the objective function is formulated as *throughput* $\times$ (*price–unit production costs*) $-$ *total holding cost* $-$

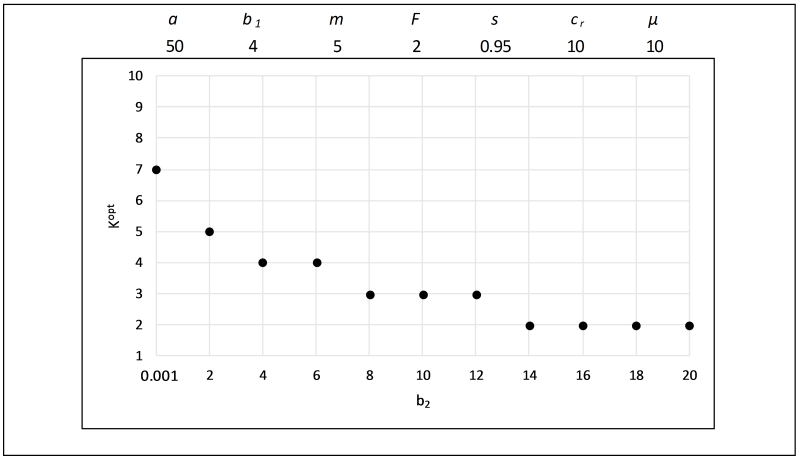


Figure D.2: K^{opt} in function of b_2

Table D.1: Detailed result of K^{opt} in function of b_2

b_2	Profit with $K =$										K^{opt}	Max Profit
	1	2	3	4	5	6	7	8	9	10		
1.00E-27	24.01	31.36	34.56	36.14	36.92	37.27	**37.39**	37.37	37.28	37.16	7	37.39
0.0001	24	31.36	34.56	36.14	36.92	37.27	**37.39**	37.37	37.28	37.16	7	37.39
0.001	24	31.35	34.56	36.13	36.91	37.27	**37.38**	37.37	37.28	37.16	7	37.38
2	23.02	29.65	32.24	33.27	**33.57**	33.5	33.25	32.93	32.59	32.26	5	33.57
4	22.29	28.34	30.43	**31.03**	30.96	30.59	30.12	29.64	29.21	28.85	4	31.03
6	21.57	27.05	28.67	**28.87**	28.49	27.89	27.28	26.74	26.33	26.03	4	28.87
8	20.85	25.78	**26.95**	26.8	26.16	25.41	24.75	24.25	23.9	23.69	3	26.95
10	20.14	24.53	**25.28**	24.82	23.98	23.16	22.52	22.09	21.83	21.68	3	25.28
12	19.43	23.31	**23.66**	22.93	21.97	21.14	20.56	20.21	20.02	19.92	3	23.66
14	18.73	**22.11**	22.1	21.15	20.11	19.32	18.83	18.56	18.42	18.34	2	22.11
16	18.04	**20.93**	20.59	19.47	18.42	17.7	17.29	17.08	16.98	16.93	2	20.93
18	17.36	**19.78**	19.14	17.9	16.87	16.24	15.91	15.75	15.68	15.64	2	19.78
20	16.68	**18.65**	17.76	16.44	15.46	14.92	14.66	14.54	14.48	14.44	2	18.65

total penalty cost. The throughput and costs (in this case holding and penalty cost) are affected by the number of accepted client. Thus, in the case where the client isn't sensitive to price, firms will set a very high price at the same time accept many clients. When the client is very sensitive to price (b_1 is high), firm has to set the price as low as possible (minimum price is unit production cost m) in the same time minimizing the holding and penalty cost. Thus, the firm reacts by reducing the number of accepted clients.

In table D.2, we provide more detailed result of the figure D.3. We see that the behaviors of the profit for each b_1 with K from 1 to 10 are similar to what we see in figure D.1. The profit will increase up to a certain value of K then decreases. For our case where $b_1 = 1E - 27$, the better K we obtained is 10. But the profit is still

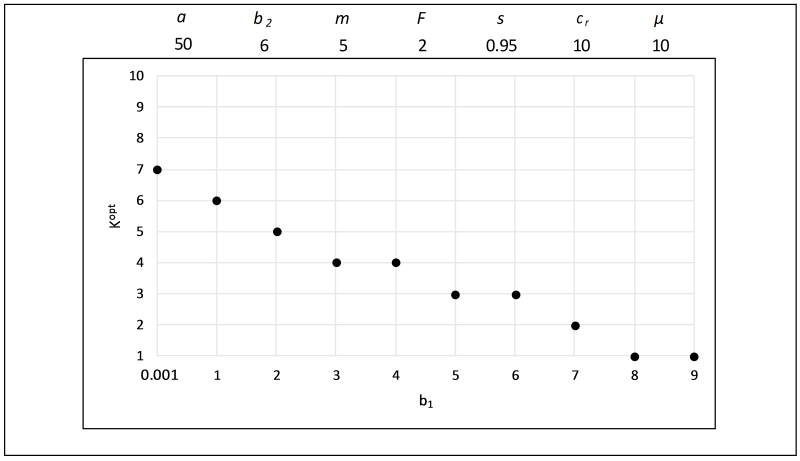


Figure D.3: K^{opt} in function of b_1

increasing, thus the optimum K is probably bigger than 10.

Table D.2: Detailed result of K^{opt} in function of b_1

b1	Profit with K =										Kopt	Max Profit
	1	2	3	4	5	6	7	8	9	10		
1.00E-27	2.10E+29	2.78E+29	3.10E+29	3.29E+29	3.42E+29	3.51E+29	3.57E+29	3.62E+29	3.66E+29	**3.69E+29**	10	3.69E+29
0.0001	1994972.8	2580128.4	2832251.5	2949976	3006745.1	3028204.5	**3029339.4**	3016529.5	2994080.8	2967796.5	7	3029339.4
0.001	199487.48	258279.02	283207.39	295073.14	300682.08	302822.11	**302882.57**	301641.06	299506.67	296690	7	302882.57
1	169.3	218.5	238.55	247.33	250.69	**251.09**	249.78	247.4	244.41	241.08	6	251.09
2	70.1	89.99	97.63	100.53	**101.19**	100.67	99.5	98	96.35	94.77	5	101.19
3	37.48	47.69	51.25	**52.28**	52.16	51.47	50.55	49.57	48.66	47.9	4	52.28
4	21.57	27.05	28.67	**28.87**	28.49	27.89	27.28	26.74	26.32	26.04	4	28.87
5	12.4	15.2	**15.77**	15.62	15.25	14.86	14.56	14.35	14.21	14.13	3	15.77
6	6.7	7.89	**7.95**	7.73	7.5	7.34	7.24	7.19	7.16	7.14	3	7.95
7	3.0845	**3.3889**	3.2797	3.1578	3.0854	3.0501	3.0346	3.0281	3.0243	3.0242	2	3.39
8	**0.9309**	0.8989	0.8444	0.8228	0.816	0.814	0.8135	0.8134	0.8132	0.8132	1	0.93
9	**0.0241**	0.0182	0.0178	0.0178	0.0178	0.0178	0.0178	0.0178	0.0177	0.0177	1	0.02

Appendix E

Root of cubic equation in Lemma 5.8

This appendix is intended to find the roots of the cubic equation in lemma 5.8 which is:

$$(\mu_2-\mu_1)b_2{L_2}^3+(\mu_2(\mu_2-\mu_1)-\alpha(\mu_2-\mu_1)+b_1m_1(\mu_2-\mu_1)-2b_2\gamma){L_2}^2 + \gamma(\mu_1-2\mu_2+\alpha-b_1m_1)L_2+\gamma^2=0$$

where $\alpha = a - b_1\delta_2$ and $\gamma = \ln(1/(1-s))$.

The first real root of cubic equation above can be found using Cardano formula (see Schechter, 2013) which is:

$$L_{21}=\sqrt[3]{Q+\sqrt{Q^2+R^3}}+\sqrt[3]{Q-\sqrt{Q^2+R^3}} - \frac{\mu_2(\mu_2-\mu_1)-\alpha(\mu_2-\mu_1)+b_1m_1(\mu_2-\mu_1)-2b_2\gamma}{3(\mu_2-\mu_1)b_2}$$

where,

$$Q=\frac{-(\mu_2(\mu_2-\mu_1)-\alpha(\mu_2-\mu_1)+b_1m_1(\mu_2-\mu_1)-2b_2\gamma)^3}{27(\mu_2-\mu_1)^3{b_2}^3} + \frac{(\mu_2(\mu_2-\mu_1)-\alpha(\mu_2-\mu_1)+b_1m_1(\mu_2-\mu_1)-2b_2\gamma)(\mu_1-2\mu_2+\alpha-b_1m_1)\gamma}{6(\mu_2-\mu_1)^2{b_2}^2} - \frac{\gamma^2}{2(\mu_2-\mu_1)b_2}$$

$$R=\frac{(\mu_1-2\mu_2+\alpha-b_1m_1)\gamma}{3(\mu_2-\mu_1)b_2}-\frac{(\mu_2(\mu_2-\mu_1)-\alpha(\mu_2-\mu_1)+b_1m_1(\mu_2-\mu_1)-2b_2\gamma)^2}{9(\mu_2-\mu_1)^2{b_2}^2}$$

The discriminant of the cubic equation of lemma 5.8 is:

$$\Delta=(\mu_2(\mu_2-\mu_1)-\alpha(\mu_2-\mu_1)+b_1m_1(\mu_2-\mu_1)-2b_2\gamma)^2 - 4b_2\gamma(\mu_2-\mu_1)(\mu_1-2\mu_2+\alpha-b_1m_1)-2b_2L_{21}(\mu_2-\mu_1)(\mu_2(\mu_2-\mu_1) - \alpha(\mu_2-\mu_1)+b_1m_1(\mu_2-\mu_1)-2b_2\gamma)-3{b_2}^2{L_{21}}^2(\mu_2-\mu_1)^2$$

If $\Delta \geq 0$, we have two other real roots, which can be found by factorization. These two roots are:

$$L_{22} = \frac{-(\mu_2(\mu_2-\mu_1) - \alpha(\mu_2-\mu_1) + b_1 m_1(\mu_2-\mu_1) - 2b_2\gamma) - (\mu_2-\mu_1)b_2 L_{21} - \sqrt{\Delta}}{2(\mu_2-\mu_1)b_2}$$

$$L_{23} = \frac{-(\mu_2(\mu_2-\mu_1) - \alpha(\mu_2-\mu_1) + b_1 m_1(\mu_2-\mu_1) - 2b_2\gamma) - (\mu_2-\mu_1)b_2 L_{21} + \sqrt{\Delta}}{2(\mu_2-\mu_1)b_2}$$

If $\Delta < 0$, we only have one candidate which is L_{21}. If L_{21} is negative then the problem is infeasible.

Appendix F

Root of cubic equation in Lemma 5.13

This appendix is intended to find the roots of the cubic equation in lemma 5.13 which is:

$$(\mu_2-\mu_1)b_2{L_1}^3+((\alpha+2\mu_2-2\mu_1)\mu_1-\alpha\mu_2+2b_2\gamma){L_1}^2+(4\mu_1-2\mu_2-\alpha)\gamma L_1-2\gamma^2)=0$$

where: $\alpha = a - b_1\delta_2 - b_1 m_1$ and $\gamma = \ln(1/(1-s))$.

The first real root of cubic equation of above can be found using Cardano formula (see Schechter, 2013):

$$L_{11} = \sqrt[3]{Q+\sqrt{Q^2+R^3}} + \sqrt[3]{Q-\sqrt{Q^2+R^3}} - \frac{(\alpha+2\mu_2-2\mu_1)\mu_1-\alpha\mu_2+2b_2\gamma}{3(\mu_2-\mu_1)b_2}$$

where,

$$\begin{aligned} Q &= \frac{-((\alpha+2\mu_2-2\mu_1)\mu_1-\alpha\mu_2+2b_2\gamma)^3}{27(\mu_2-\mu_1)^3{b_2}^3} \\ &\quad + \frac{((\alpha+2\mu_2-2\mu_1)\mu_1-\alpha\mu_2+2b_2\gamma)(4\mu_1-2\mu_2-\alpha)\gamma}{6(\mu_2-\mu_1)^2{b_2}^2} + \frac{2\gamma^2}{2(\mu_2-\mu_1)b_2} \\ R &= \frac{(4\mu_1-2\mu_2-\alpha)\gamma}{3(\mu_2-\mu_1)b_2} - \frac{((\alpha+2\mu_2-2\mu_1)\mu_1-\alpha\mu_2+2b_2\gamma)^2}{9(\mu_2-\mu_1)^2{b_2}^2} \end{aligned}$$

The discriminant of the cubic equation from lemma 5.13 is:

$$\begin{aligned} \Delta = &((\alpha+2\mu_2-2\mu_1)\mu_1-\alpha\mu_2+2b_2\gamma)^2 - 4b_2\gamma(\mu_2-\mu_1)(4\mu_1-2\mu_2-\alpha) \\ &- 2b_2L_{11}(\mu_2-\mu_1)((\alpha+2\mu_2-2\mu_1)\mu_1-\alpha\mu_2+2b_2\gamma) - 3{b_2}^2{L_{11}}^2(\mu_2-\mu_1)^2 \end{aligned}$$

If $\Delta \geq 0$, then the other two roots can be found by factorization. These two roots are:

$$\begin{aligned} L_{12} &= \frac{-((\alpha+2\mu_2-2\mu_1)\mu_1-\alpha\mu_2+2b_2\gamma)-(\mu_2-\mu_1)b_2L_{11}-\sqrt{\Delta}}{2(\mu_2-\mu_1)b_2} \\ L_{13} &= \frac{-((\alpha+2\mu_2-2\mu_1)\mu_1-\alpha\mu_2+2b_2\gamma)-(\mu_2-\mu_1)b_2L_{11}+\sqrt{\Delta}}{2(\mu_2-\mu_1)b_2} \end{aligned}$$

If $\Delta < 0$, we only have one candidate which is L_{11}. If L_{11} is negative then the problem is infeasible.

Pricing decision and lead time quotation in supply chains with an endogenous demand sensitive to lead time and price

Abstract – Along with the price, the delivery lead time has become a key factor of competitiveness for companies and an important purchase criterion for many customers. Nowadays, firms are more than ever obliged to meet their quoted lead time, which is the delivery lead time announced to the customers. The combination of pricing and lead time quotation implies new trade-offs and offers opportunities for many insights. For instance, on the one hand, a shorter quoted lead time can lead to an increase in the demand but also increases the risk of late delivery and thus may affect the firm's reputation and deter future customers. On the other hand, a longer quoted lead time or a higher price generally yields a lower demand. Despite the strategic role of joint pricing and lead time quotation decisions and their impacts on demand, in the operations management literature an exogenous demand (a priory a known demand) is generally used in supply chain models, even if the design of the supply chain has a strong impact on lead times (i.e., sites location, inventory position, etc.) and thus affects the demand. Therefore, we are interested in the lead time quotation and pricing decisions in a context of endogenous demand (i.e., demand sensitive to price and quoted lead time).

The literature dealing with pricing and lead time quotation under an endogenous demand mainly considered a make to order (MTO) context. A pioneer paper, Palaka et al. (1998), investigated this issue by modeling the company as an M/M/1 queue, and our work follows their footsteps. Our review of the literature allowed to identify new perspectives for this problem, which led to three main contributions in this thesis.

In our first contribution, using Palaka et al.'s framework, we consider the unit production cost to be a decreasing function in quoted lead time. In most published papers, the unit production cost was assumed to be constant. In practice, the unit production cost generally depends on the quoted lead time. Indeed, the firm can manage better the production process and reduce the production cost by quoting longer lead time to the customers.

In the second contribution, we still consider Palaka et al.'s framework but model the firm as an M/M/1/K queue, for which demand is rejected if there are already K customers in the system. In the literature on single firm setting following Palaka et al.'s research, only the M/M/1 queue was used, i.e., where all customers are accepted, which might lead to long sojourn times in the system. Our idea is based on the fact that rejecting some customers, might help to quote shorter lead time for the accepted ones, which might finally lead to a higher profitability, even if in the first glance we lose some demand.

In the third contribution, we study a new framework for the lead time quotation and pricing problem under endogenous demand as we model the supply chain by two production stages in a tandem queue (M/M/1-M/M/1). In the literature with multi-firm setting, all papers considered that only one actor has production operations and the other actor has zero lead time. We investigated both the centralized and decentralized decision settings.

For each problem studied, we formulated a profit-maximization model, where the profit consists of a revenue minus the production, storage and lateness penalty costs, and provides the optimum result (analytically or numerically). These resolutions led us to demonstrate new theoretical results (such as the expected lateness in an M/M/1/K, and the sufficient condition required to satisfy the global service constraint in a tandem queue by only satisfying the local service constraints). We also conducted numerical experiments and derived managerial insights.

Keywords: supply chains, endogenous demand, lead time quotation, pricing, queuing theory

Choix du prix et du délai de livraison dans une chaîne logistique avec une demande endogène sensible au délai de livraison et au prix

Résumé – Parallèlement au prix, le délai de livraison est un facteur clé de compétitivité pour les entreprises. De plus les entreprises sont plus que jamais obligées de respecter ce délai promis. La combinaison du choix du prix et du délai promis implique de nouveaux compromis et offre de nombreuses perspectives. Un délai plus court peut entraîner une augmentation de la demande, mais augmente également le risque de livraison tardive et donc décourager les clients. A contrario un délai plus long ou un prix plus élevé entraîne généralement une baisse de la demande. Or malgré le rôle stratégique conjoint du prix et des délais et leurs impacts sur la demande, dans la littérature en gestion des opérations on suppose très généralement une demande exogène (fixée a priori) même si la conception de la chaîne impacte fortement les délais (localisation des sites, positionnement des stocks,..) et donc la demande. Nous nous sommes donc intéressés à ces choix de fixation des délais promis et du prix dans un contexte de demande endogène.
La littérature traitant du choix du délai et du prix sous demande endogène a principalement considéré un contexte de fabrication à la commande (Make to Order). Un papier fondateur de Palaka et al en 1998 a présenté cette problématique avec une modélisation de l'entreprise par une file d'attente M/M/1 et nos travaux se placent dans la suite de ce travail. Notre revue de la littérature a permis d'identifier de nouvelles perspectives et nous proposons trois extensions dans cette thèse.
Dans notre première contribution, en utilisant le cadre de Palaka et al, nous considérons que le coût de production est une fonction décroissante du délai. Dans tous les articles publiés dans ce contexte, le coût de production unitaire a été supposé constant. Pourtant en pratique, le coût de production unitaire dépend du délai promis, l'entreprise pouvant mieux gérer le processus de production et réduire les coûts de production en proposant des délais plus longs aux clients.
Dans la deuxième contribution, nous considérons toujours le cadre de Palaka et al, mais modélisons l'entreprise comme une file d'attente M/M/1/K, pour laquelle la demande est donc rejetée s'il y a déjà K clients dans le système. Dans la littérature issue du travail de Palaka seule la file d'attente M/M/1 a été utilisée, ce qui signifie que tous les clients sont acceptés, ce qui peut entraîner de longues durées de séjour dans le système. Notre idée est basée sur le fait que rejeter certains clients, même si cela peut apparaitre dans un premier temps comme une perte de demande, pourrait aider à proposer un délai plus court pour les clients acceptés, et finalement conduire à une demande et donc un profit plus élevé.
Dans la troisième contribution nous étudions un nouveau cadre pour le problème du délai et du prix en fonction de la demande endogène, en modélisant une chaîne logistique composée de deux étapes de production, modélisée par un réseau de files d'attente tandem (M/M/1-M/M/1). Dans la littérature avec ce cadre multi-entreprise, tous les articles ont considéré qu'un seul acteur avait des opérations de production, l'autre acteur ayant un délai nul. Nous avons étudié les scénarios centralisés et décentralisés.
Pour chacun des nouveaux problèmes nous avons proposé des formulations maximisant le profit composé du revenu diminué des coûts de production, de stockage et pénalité de retard, et fourni des résolutions optimales, analytiques ou numériques. Ces résolutions nous ont amenés à démontrer de nouveaux résultats (retard moyen dans une M/M/1/K ; condition pour que des contraintes de service locales permettent d'assurer une contrainte de service globale dans un système en tandem). Nous avons mené des expériences numériques pour voir l'influence des différents paramètres.

Mots clés: chaînes logistiques, demande endogène, délai de livraison, prix, files d'attente.