View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Sabanci University Research Database

Coping with Production Time Variability via Dynamic Lead-Time Quotation

Gökçe Kahvecioğlu,
Northwestern University, Department of Industrial Engineering and Management Sciences,
2145 Sheridan Road, Evanston, IL 60208, USA,
gokcekahvecioglu2014@u.northwestern.edu

Barış Balcıoğlu,
Sabancı University, Faculty of Engineering and Natural Sciences,
Orhanlı-Tuzla, 34956 Istanbul, Turkey,
balcioglu@sabanciuniv.edu

Abstract

In this paper, we propose two dynamic lead-time quotation policies in an $M/GI/1$ type make-to-stock queueing system serving lead-time sensitive customers with a single type of product. Incorporating non-exponential service times in an exact method for make-to-stock queues is usually deemed difficult. Our analysis of the proposed policies is exact and requires the numerical inversion of the Laplace transform of the sojourn time of an order to be placed. The first policy assures that the long-run probability of delivering the product within the quoted lead-time is the same for all backlogged customers. The second policy is a refinement of the first which improves the profitability if customers are oversensitive to even short delays in delivery. Numerical results show that both policies perform close to the optimal policy that was characterized only for exponential service times. The new insight gained is that the worsening impact of the production time variability, which is felt significantly in systems accepting all customers by quoting zero lead times, decreases when dynamic lead-time quotation policies are employed.

Keywords and Phrases: make-to-stock queues; $M_n/GI/1/K$; inventory/production policies; due date quotation; service time variability

1 Introduction

To establish a lean manufacturing system, companies aim for reducing production time variability by investing in high-technology equipment, training personnel, and conducting/improving maintenance activities to prevent unplanned production line stoppages. Attaining a minimum level of production time variability is a strategic goal because this way the adverse consequences such as the increase in overtime and work in progress levels, the decrease in output rate and work center utilizations, and the deterioration in order completion times can be avoided (see Li, 2003 and the references therein). Yet, if funding cannot be secured easily and required investment cannot be realized in the short term, can a company follow alternative policies to diminish the worsening impact of high production time variability? In this study, while exploring dynamic lead-time quotation policies as a way to increase profitability, we demonstrate that they can serve companies to this end as well.

We analyze a company that manufactures and stores a single type of product. Demand follows a Poisson process with state-dependent arrival rates (see Rajagopalan, 2002, and the references therein on the validity of using the Poisson process to model the demand process). The demand rates change due to the decision of the customers on whether or not to place an order depending on the delivery lead-time announced/quoted when there is no stock on hand. As the number of pending orders increases, to ensure reliable delivery, the company tends to quote longer lead times for newly arriving customers. Announcing longer lead times makes it more likely that customers will not place any order and will be lost. If shorter lead times are announced to secure customers to order and then the product cannot be delivered until the due date (arrival time plus the quoted lead-time), the company pays penalty costs (see Hopp and Sturgis, 2001, Slotnick and Sobel, 2005, and the references therein for examples of late delivery penalties paid in industry). In the analyzed setting, a sufficiently long lead-time is announced to effectively reject customers when the number of pending orders reaches a critical level. All these considerations lead us to model the production system as an $M_n/GI/1/K$ make-to-stock queue in Section 2. In this framework,

we propose practical dynamic lead-time quotation policies which take the number of pending orders in the production line into account. The main contribution of our study is to show the applicability of the proposed policies for general production/service time distributions. The optimal dynamic lead-time quotation policy is not tractable in this setting unless exponential service times are assumed. Via a numerical study, we show that for cases with exponential service times, the proposed policies perform close to the optimal policy. We also demonstrate that well-designed dynamic policies such as the ones developed in this study can help reduce the worsening impact of the high service time variability that the company may be unable to minimize by other means.

In the earlier literature on due date quotation, some of the studies assume that all customers accept the announced lead times and some define the due dates exogenously. In this setting, the scheduling of the orders is important. For instance, Lin (2001) considers minimizing the number of tardy jobs or maximum tardiness in a two-machine setting; Unal, Uzay, and Kiran (1997) construct a heuristic on how to insert newly arriving jobs in an existing schedule. Elhafsi (2000) explores how to assign lead times for new orders within a specified time window. Lawrence (1995) designs a method to estimate flow times to set due dates. In a make-to-order setting with all customers accepting quoted lead times, Wein (1991) addresses the sequencing of jobs in a multiclass $M/GI/1$ queue. The class with the smallest mean service time is given non-preemptive priority. For certain due date setting rules, he suggests using the distribution of the conditional sojourn time of orders but observes difficulties in doing this. Through a number of simulation studies, he concludes that a lead-time quotation policy has a more pronounced impact on meeting service levels than the priority sequencing policy. There are also studies such as the ones by Keskinocak, Ravi, and Tayur (2001), Gallien, Tallec, and Shoenmeyr (2004), and Kapuscinski and Tayur (2007), that assume accepted customers will be delivered the product within the quoted lead times. If this is not possible, customers are rejected.

In the presence of competitors, customer response may change depending on the length of the quoted lead-time. Dellaert (1991) proposes using sojourn time distribution of an order

in the $M/M/1$ queue to set due dates. Duenyas and Hopp (1995) develop dynamic lead time quotation policies in a make-to-order queue assuming both infinite and finite production capacities. They obtain partial results for the $GI/GI/1$ queue and prove that the optimal lead-time to quote in the $M/M/1$ setting increases with the number of production orders in the manufacturing system. In the presence of multiple classes of customers demanding the same type of unit, Duenyas (1995) develops a heuristic that considers the characteristics of the customer classes while setting due dates and order sequencing. The proposed policies in our study can also be considered in the multi-class setting. In addition to implementing dynamic lead-time quotation policies, holding inventory can also give an edge to companies. Via a game theoretic approach, Li (1992) shows the importance of keeping inventory in a competitive environment while also quoting due dates. Rajagopalan (2002) approximates the production facility by an $M/GI/1$ queue, and utilizes the first two moments of the sojourn time distribution and explores when finished goods inventory should be kept instead of serving on a make-to-order basis in order to meet the probability of delivery-on-time.

Our work stems from that of Savaşaneril, Griffin, and Keskinocak (2010) combining the concept of dynamic due date setting with order acceptance/rejection (i.e., rejecting customers when congestion in the production facility – the number of pending orders – reaches a critical level) in an $M/M/1$ make-to-stock queue. Using a Markov-decision process (MDP) approach, along with the maximization of profit as the objective function, they show that orders should be satisfied from stock if there is any, and otherwise, the optimal quoted lead-time is monotonically increasing in the number of pending orders. The profit under the optimal lead-time policy is unimodal in the base-stock level. Since production times can be non-exponential and high variance in service times can significantly reduce the profitability (see, Sanajian and Balcıoğlu, 2009), we relax their exponential service time assumption by allowing general service time distributions. In this setting, the characterization of the optimal due date/lead-time policy via an MDP approach is quite difficult. For instance, Çelik and Maglaras (2008) resort to diffusion approximations since employing an MDP approach is also intractable for the multiclass $M_n/GI/1$ make-to-order queue where they use dynamic pricing

to guarantee the lead times announced. Note that due to the memoryless property of the exponential service times, Savaşaneril, Griffin, and Keskinocak (2010) are able to write the optimality equation (Eq. 1 in their paper) for this problem in the $M/M/1$ queue. In doing this, they can focus only on the new customer arrival and service completion instants. When service times are general, one loses the ability to write a similar optimality equation and the analysis of the optimal policy in the $M/GI/1$ queue becomes quite difficult. In return, we develop policies that are easy to implement when production times are general while performing close to the results of the optimal policy of Savaşaneril, Griffin, and Keskinocak (2010) in the case of exponential service times.

A probable policy is simply quoting zero lead times to all customers and losing none of them at the expense of starting right away to accrue a penalty cost proportional to the length of late delivery. This reduces the problem to the setting of Sanajian and Balcıoğlu (2009) (see the references therein as well for studies with constant demand rates and constant revenues where system cost, comprising of stock holding and backlogging costs, is minimized). The only decision to make under this policy would be to determine the optimum control parameters for the finished goods inventory. In Section 4, we consider this policy as a reference point to assess the performances of the policies proposed in this study.

We design two dynamic policies as alternatives to quote lead times to newly arriving customers. Both policies consider the number of pending orders present at customer arrival instants. The first policy presented in Section 3.1 is the *Fair Quotation Policy* (FQP) under which the probability of meeting the delivery within the announced lead-time is the same for any backlogged customer. The company optimizes this probability together with the inventory control parameter. Additional service level constraints in the form of keeping the probability of delivery until the due date above a threshold can be easily included in the model. We refer the reader to Hopp and Sturgis (2001), Spearman and Zhang (1999), So and Song (1998), for designing due date quotation policies under various service levels, such as the fill rate, the fraction of tardy jobs, and the probability of meeting demand on time.

While the FQP is "fair" since it assures the same long-run probability of meeting the

demand of any awaiting customer during the quoted lead-time, an alternative policy, namely the *Preferential Quotation Policy* (PQP), designed in Section 3.2, can further increase profits when customers are oversensitive to announced lead times. Under the PQP, possible customer losses are prevented by quoting zero lead-time when the number of backlogged customers is low. In return, the PQP rejects customers sooner than the FQP by keeping a lower limit on the maximum number of customers awaiting their orders. In other words, the PQP prefers securing more customers when the number of awaiting orders is small and rejects more customers in return. Both policies require having the Laplace transform of the sojourn time distribution in the $M_n/GI/1/K$ queue (see Kerner, 2008). We employ numerical inversion techniques on the LT of the sojourn time distribution (e.g., Jagerman, 1982, Abate and Whitt, 1995) to obtain the lead times satisfying the probability of delivery until the quoted due date. Thus, the idea of using sojourn time distributions as presented here can have broader application areas for other queueing disciplines or multi-class systems as long as the sojourn time Laplace transforms are available.

While designing these policies, two important questions arose. The first one was whether they would yield profits close to the optimal results found by Savaşaneril, Griffin, and Keskinocak (2010), which turned out to be the case as demonstrated via the numerical study discussed in Section 4.1. The second question was to see the impact of the service time variability on profit when these policies were implemented. As suspected, quoting lead times dynamically can improve profitability across all service time distributions. In most of the numerical examples, the proposed FQP and PQP turn out to be more profitable than quoting zero lead times. Additional observations are also made via the numerical study in Section 4.2. The degrading impact of higher service time variability is well-known and quantified in the make-to-stock setting with constant demand rate by Sanajian and Balcıoğlu (2009). Under any policy, having deterministic service times maximizes the profits. The new finding is that the profit loss due to higher service time variance decreases significantly when a dynamic lead-time quotation policy is implemented instead of accepting all customers. Moreover, a dynamic policy better suited for the customer profile diminishes the profit loss

more. This is an important managerial insight for companies that may be unable to lower the production time variation for reasons such as not being able to invest in infrastructure. For such companies, designing the right dynamic lead-time quotation policy would provide the remedy.

The rest of the paper is organized as follows. In Section 2, we present the problem analyzed, followed by the proposed policies in Section 3. We present our numerical study in Section 4. Finally, in Section 5, we make concluding remarks and discuss how the research can be extended.

2 The $M/GI/1$ Queue with Lead-Time Quotations

In this section, we model a single item production system with a continuously reviewed inventory as a make-to-stock queue. We use a production control according to a base-stock level S. Thus, production stops when the inventory level reaches S and starts as soon as the inventory level decreases to $S-1$ (see, e.g., Savaşaneril, Griffin, and Keskinocak, 2010, who suggest that base-stock policies are applicable in repair shops or for dealers providing after-sales service). We assume that customers arrive one at a time according to a Poisson process with rate λ. Whenever there is available stock in the inventory, demand requests are satisfied right away. The system incurs a holding cost of h per unit inventory per unit time. If there is no stock, a lead-time d is announced to the arriving customer. With probability $f(d)$, the customer accepts the quoted lead-time, places an order, and waits until an item is produced and delivered. If the customer finds the quoted lead-time too long, she leaves the system immediately without placing an order (thus, there are no pending quotations). We assume that $f(d)$ is a decreasing function of d, $f(0)=1$, and there exists a maximum lead-time $d_{\max}$ such that $f(d_{\max})=0$. If the item cannot be produced and delivered during the announced lead-time, a tardiness cost l is incurred per unit time during the customer's waiting time in excess of d. Each item sold from the inventory or each order placed by a customer that accepts the quoted lead-time generates a revenue of R and

results in a production order mapped as an arrival at a single server queue which models the production stage. In the rest of the paper, we refer to customers placing orders or whose requests are directly satisfied from stock as "customers" and production orders present in the queueing system as "orders". The production/service times are assumed to be independent and identically distributed (i.i.d) random variables (r.v.s) with a Laplace transform denoted by $\widetilde{b}(\theta)$ and a mean of $1/\mu$. Furthermore, we assume that there is no further information about the production times available until the production has been completed. However, the policies developed in Section 3 make use of the sojourn time distribution of an order which depends on the distributions of the production time and the residual production time of the order in progress when the lead-time is quoted.

In this framework, $N(t)$, the number of (production) orders present in the queueing system at time t, gives the shortfall from the base-stock level S. This implies that when $N(t) \leq S$, the inventory carries $S - N(t)$ units and when $N(t) > S$, the system has $N(t) - S$ backlogged customers. Assuming that the system is stable, the steady-state probability of having n orders in the queueing system, namely $p(n|S, \mathbf{d}) = P(N = n|S, \mathbf{d})$, depends on S (unless the system quotes zero lead-time to any customer) and the vector $\mathbf{d} = [d_0, d_1, \ldots, d_S, d_{S+1}, \ldots]$ where d_n is the announced lead-time to a customer that sees n orders in the system. Since all customers are identical in terms of revenues and tardiness costs, a higher profit cannot be generated by reserving an item for a future customer instead of satisfying the demand of a customer that arrives when there is stock. Thus, requests arising when there is stock are immediately satisfied from the inventory. That is, $d_n = 0$ for $n = 0, \ldots, S-1$, and for a given S and $\mathbf{d}$, the expected profit per unit time is

$$P(S, \mathbf{d}) = \lambda R \sum_{n=0}^{\infty} p(n|S, \mathbf{d}) f(d_n) - h \sum_{n=0}^{S-1} (S-n) p(n|S, \mathbf{d}) - \lambda l \sum_{n=S}^{\infty} p(n|S, \mathbf{d}) f(d_n) L_n(d_n), \quad (1)$$

where $L_n(d_n)$ is the expected waiting time in excess of d_n of a customer that accepts the quoted lead-time d_n. Observe that the first term on the RHS of Eq. (1) is the expected revenue per unit time whereas the second and third terms are the expected inventory holding and delay penalty cost rates, respectively. Denoting the sojourn time r.v. of such a customer,

i.e., the elapsed time from the moment she places an order – when there is no stock – until she receives the finished item by T_{n+1} (the subscript referring to the $(n+1)$st order that will be sent to the make-to-stock queue due to this customer) and its probability density function (PDF) by $g_{n+1}(\cdot)$, we have

$$L_n(d_n) = \int_{d_n}^{\infty} (x - d_n) g_{n+1}(x) dx. \tag{2}$$

Various lead-time quotation policies can be considered. A possible policy is announcing zero lead-time for (and accepting) all customers. In this case, depending on the parameters such as the revenue R, the customer arrival rate λ, and costs, the system may incur loss instead of making profit. In case all customers are accepted, the only decision to make is finding the optimal S_0^* (the subscript 0 referring to announcing zero lead-time to everyone) that minimizes the holding and backlogging costs, as done in the study of Sanajian and Balcıoğlu (2009). The optimal S_0^* is also an upper bound for the optimal base-stock level if nonzero lead times are quoted in the same system which is analyzed in this study. This follows from Observation 1 by Savaşaneril, Griffin, and Keskinocak (2010).

If we have a lead-time vector $\mathbf{d}$ obtained from a policy (such as the ones proposed in Section 3) for a given base-stock level S, we compute $P(S, \mathbf{d})$ given in Eq. (1) as follows. If $\mathbf{d}$ contains nonzero lead times and $d_{\max}$ at its Kth entry (which happened to be the case whenever our proposed policies gave nonzero lead times in numerical experiments in Section 4), the underlying system is an $M_n/GI/1/K$ queue with $\lambda_n = \lambda f(d_n)$ and $p(n|S, \mathbf{d})$ in this queue can be obtained following Kerner (2008) and Abouee-Mehrizi and Baron (2015). In Section 2.1, we summarize how to obtain the required probabilities and also extend existing results when arrival rates are the same for a finitely many neighboring states where the state refers to the number of orders in the system. One can employ the analysis of the $M_n/GI/1$ queue (see, Kerner, 2008, Abouee-Mehrizi and Baron, 2015, Economou and Manou, 2015) if $d_{\max}$ is never announced while nonzero lead times for all $n \geq S$ make the customer arrival rates state-dependent (however, we have not come across such a case in our numerical experiments in Section 4). Finally, if the proposed policy yields zero lead times to be announced

for any customer (which happened to be the case for some numerical examples), with the analysis of the $M/GI/1$ queue of Sanajian and Balcıoğlu (2009) the optimal S_0^*, the optimal system (holding and backlogging) cost rate $C(S_0^*)$ can be determined and the optimal profit per unit time, $\lambda R - C(S_0^*)$, can be computed. As a special case, when service times are exponentially distributed with rate μ, the system can be modeled as a birth-and-death process from which one can derive $p(n|S, \mathbf{d})$.

2.1 The Steady-State System Size Distribution in the $M_n/GI/1/K$ Queue

In this section, we first introduce a crucial r.v. H_n for the ensuing derivations, which is the residual service time r.v. given that there are n orders in the system. Denoting its Laplace transform by $\widetilde{h}_n(\theta)$ and introducing

$$\widetilde{c}_n(\theta) = \widetilde{b}(\lambda_n)(1 - \widetilde{h}_{n-1}(\theta)) + \widetilde{b}(\theta)(\widetilde{h}_{n-1}(\lambda_n) - 1), \quad n = 1, 2, \ldots, \tag{3}$$

Kerner's (2008) recursive formula can be written as

$$\widetilde{h}_n(\theta) = \frac{\lambda_n}{1 - \widetilde{h}_{n-1}(\lambda_n)} \frac{\widetilde{c}_n(\theta)}{\theta - \lambda_n}, \quad n = 1, 2, \ldots, \tag{4}$$

with $\widetilde{h}_0(\theta) = \widetilde{b}(\theta)$. The mean of H_n can be recursively computed from Eq. (4) as follows:

$$\begin{aligned} E[H_1] &= \frac{1}{\mu(1 - \widetilde{b}(\lambda_1))} - \frac{1}{\lambda_1}, \\ E[H_n] &= \frac{\widetilde{b}(\lambda_n)}{1 - \widetilde{h}_{n-1}(\lambda_n)} E[H_{n-1}] - \frac{1}{\lambda_n} + \frac{1}{\mu}, \ n \geq 2. \end{aligned} \tag{5}$$

After defining the state of the $M_n/GI/1/$ queue as the number of customers in the system (n) and the remaining service time, Kerner (2008) obtains the flow equations relating states $n-1$, $n+1$ to state n in his Eq. (8). From here, he obtains the following recursive formulae:

$$\begin{aligned} p(n|S, \mathbf{d}) &= \frac{\lambda_0 p(0|S, \mathbf{d})}{\lambda_n} \prod_{j=0}^{n-1} \frac{1 - \widetilde{h}_j(\lambda_{j+1})}{\widetilde{b}(\lambda_{j+1})}, \quad n = 1, \ldots, K-1, \\ p(K|S, \mathbf{d}) &= 1 - \sum_{n=0}^{K-1} p(n|S, \mathbf{d}), \end{aligned}$$

for which Abouee-Mehrizi and Baron (2015) provide

$$p(0|S,\mathbf{d}) = \frac{\frac{1}{1+\lambda_{K-1}E[H_{K-1}]}}{\frac{\lambda_0}{\lambda_{K-1}}\prod_{j=0}^{K-2}\frac{1-\widetilde{h}_j(\lambda_{j+1})}{\widetilde{b}(\lambda_{j+1})} + \frac{1}{1+\lambda_{K-1}E[H_{K-1}]}\left(1+\sum_{i=1}^{K-2}\frac{\lambda_0}{\lambda_i}\prod_{j=0}^{i-1}\frac{1-\widetilde{h}_j(\lambda_{j+1})}{\widetilde{b}(\lambda_{j+1})}\right)}.$$

However, these formulae, requiring $\widetilde{h}_j(\lambda_{j+1})$, do not explicitly show what happens when $\lambda_{j+1} = \lambda_j$ because with $\widetilde{c}_j(\lambda_j) = 0$ we see that $\widetilde{h}_j(\lambda_j)$ results in a division of 0 by 0 in Eq. (4). However, in a system with $S \geq 2$, since zero lead times are announced to customers seeing 0 to $S-1$ orders in the system, we would have $\lambda_1 = \ldots = \lambda_{S-1} = \lambda_0$. In the PQP studied in Section 3.2, zero lead times may be announced to customers seeing more than S orders in the system leading to $\lambda_{j+1} = \lambda_j = \lambda_0$ for $j \geq S$. Therefore, letting $\widetilde{k}^{(m)}(\theta)$ denote the mth derivative of a Laplace transform $\widetilde{k}(\theta)$, when we apply the L'Hôpital's rule in Eq. (4), we have the following Corollary (proof is omitted since it is straightforward):

Corollary 1 *The Laplace transform of the residual service time r.v. H_j given that there are j orders in the system evaluated at $\lambda_{j+1} = \lambda_j$ is given by*

$$\widetilde{h}_j(\lambda_j) = \frac{\lambda_j}{1-\widetilde{h}_{j-1}(\lambda_j)}\widetilde{c}_j^{(1)}(\lambda_j). \tag{6}$$

There is a recursive relationship between $\widetilde{c}_j^{(m)}(\theta)$ and $\widetilde{h}_{j-1}^{(m)}(\theta)$ which is given in the following Proposition.

Proposition 1 *With $\widetilde{h}_0^{(m)}(\theta) = \widetilde{b}^{(m)}(\theta)$, there exists the following recursive relationship between $\widetilde{c}_j^{(m)}(\theta)$ and $\widetilde{h}_{j-1}^{(m)}(\theta)$:*

$$\widetilde{c}_j^{(m)}(\theta) = -\widetilde{b}(\lambda_j)\widetilde{h}_{j-1}^{(m)}(\theta) + \widetilde{b}^{(m)}(\theta)(\widetilde{h}_{j-1}(\lambda_j) - 1), \tag{7}$$

$$\widetilde{h}_j^{(m)}(\theta) = \begin{cases} \frac{\lambda_j}{1-\widetilde{h}_{j-1}(\lambda_j)} \frac{(\theta-\lambda_j)^m \widetilde{c}_j^{(m)}(\theta) - \frac{m(\theta-\lambda_j)^m \widetilde{h}_j^{(m-1)}(\theta)}{\lambda_j}}{(\theta-\lambda_j)^{m+1}}, & \text{for } \theta \neq \lambda_j, \\ \frac{\lambda_j}{1-\widetilde{h}_{j-1}(\lambda_j)} \frac{\widetilde{c}_j^{(m+1)}(\theta)}{m+1}, & \text{for } \theta = \lambda_j. \end{cases} \tag{8}$$

Proof. Eq. (7) is obtained by successively taking the derivative of Eq. (3). Starting from Eq. (4), taking successive derivatives gives the result for the case of $\theta \neq \lambda_j$ in Eq. (8). This

yields a division of 0 by 0 when $\theta = \lambda_j$. Thus, applying the L'Hôpital's rule on it yields the second line of Eq. (8). ■

Therefore, with the help of Proposition 1, Eq. (6) of Corollary 1 can be used in computing $\widetilde{h}_j(\lambda_{j+1})$ for the required probabilities when $\lambda_{j+1} = \lambda_j$. Otherwise, Eq. (4) can be employed.

2.2 Computation of the Expected Profit Rate

Now we are ready to proceed with computing the profit rate function $P(S, \mathbf{d})$ in Eq. (1). Before evaluating $L_n(d_n)$ in Eq. (2) that is required, we first note that if a customer is seeing $n \geq S$ orders in the system upon arrival accepts d_n, triggering the $(n+1)$st order to be sent to the make-to-stock queue, she will need to wait for the remaining service time of the item under production, plus, $n - S$ service completions. In the $M_n/M/1/K$ setting, T_{n+1} (the sojourn time r.v. for this customer) follows an $(n - S + 1)$-stage Erlang distribution with each exponential stage having a rate of μ, i.e., Erlang$(\mu, n - S + 1)$. Starting from Eq. (2) and using the following identity (e.g., Gross and Harris, 1998, p. 20)

$$\int_{d_n}^{\infty} \frac{\mu(\mu x)^{(n-S)}}{(n-S)!} dx = \sum_{i=0}^{n-S} \frac{(\mu d_n)^i e^{-\mu d_n}}{i!},$$

we get

$$\begin{aligned} L_n(d_n) &= \int_{d_n}^{\infty} (x - d_n) \frac{\mu(\mu x)^{(n-S)}}{(n-S)!} dx, \\ &= e^{-\mu d_n} \left(\frac{n-S+1}{\mu} \sum_{i=0}^{n-S+1} \frac{(\mu d_n)^i}{i!} - d_n \sum_{i=0}^{n-S} \frac{(\mu d_n)^i}{i!} \right). \end{aligned}$$

In the $M_n/GI/1/K$ queue with non-exponential service times, on the other hand, we have the Laplace transform of T_{n+1} as

$$\widetilde{g}_{n+1}(\theta) = \widetilde{h}_n(\theta) \widetilde{b}(\theta)^{n-S}, \tag{9}$$

where $\widetilde{h}_n(\theta)$ is given in Eq. (4).

In the remainder of the paper, for various computations, we need to numerically invert a given Laplace transform $\widetilde{k}(\theta)$ and evaluate at d which will be denoted by $\mathcal{L}^{-1}\{\widetilde{k}(\theta)\}(d)$.

Now we rewrite Eq. (2) as

$$L_n(d_n) = \int_{d_n}^{\infty} x g_{n+1}(x)dx - d_n \overline{G}_{n+1}(d_n), \tag{10}$$

where $\overline{G}_{n+1}(\cdot)$ is the complementary distribution function of T_{n+1}. Both terms on the RHS of Eq. (10) may not be available in closed-form for direct computation, however, their Laplace transforms are available and they can be numerically inverted (see, e.g., Jagerman, 1982) and evaluated at d_n. Then,

$$\begin{aligned} \overline{G}_{n+1}(d_n) &= \mathcal{L}^{-1}\{\frac{1-\widetilde{g}_{n+1}(\theta)}{\theta}\}(d_n), \\ \int_{d_n}^{\infty} x g_{n+1}(x)dx &= \mathcal{L}^{-1}\{\frac{E[T_{n+1}]+\widetilde{g}^{(1)}_{n+1}(\theta)}{\theta}\}(d_n), \\ &= E[T_{n+1}] + \mathcal{L}^{-1}\{\frac{\widetilde{g}^{(1)}_{n+1}(\theta)}{\theta}\}(d_n) \end{aligned}$$

where $E[T_{n+1}] = E[H_n] + (n-S)/\mu$ (with $E[H_n]$ given in Eq. 5) and $\widetilde{g}^{(1)}_{n+1}(\theta)$ is the derivative of $\widetilde{g}_{n+1}(\theta)$.

In summary, given a $\mathbf{d}$ vector generated via a policy and S, we are able to compute the profit. Note that in the $M_n/M/1$ queue, for a given S, the optimal lead times to announce $\mathbf{d}^* = [0, 0, \ldots, 0, d^*_S, d^*_{S+1}, \ldots]$ can be found by formulating the problem as an MDP as done by Savaşaneril, Griffin, and Keskinocak (2010). Eventually, conducting a line search from 0 to S^*_0 of the $M/M/1$ queue, the optimal base-stock level S^* and the corresponding optimal $\mathbf{d}^*$ can be found and the optimal profit P^*_E (where the subscript E refers to exponential service times) can be computed. However, the MDP approach is not practical when service times are generally distributed. Therefore, in Section 3, we design two policies that are applicable in the $M/GI/1$ system which, as demonstrated via the numerical examples in Section 4.1, also turn out to perform very close to the optimal policy in the $M/M/1$ system.

3 Lead-Time Quotation Policies

In this section, we propose two dynamic lead-time quotation policies: a) the Fair Quotation Policy (FQP) in Section 3.1, and b) the Preferential Quotation Policy (PQP) in Section 3.2.

While determining d_n, both policies consider the number of orders present in the system (n) as the only state information. That is, additional information, such as the length of time since the start of production of the current item is ignored, although this might be both available and valuable in $M/GI/1$ systems. Under the FQP, the long-run probability of producing the item within the quoted lead-time is the same for each backlogged customer. The PQP is a refinement over the FQP and when compared to the latter, it attempts to "allure" ("deter") more customers when the number of backlogged customers is small (large).

3.1 The Fair Quotation Policy

The FQP identifies lead times which assure that the long-run probability of meeting the demand within the announced lead-time is the same for any backlogged customer. We simply denote this probability by α. Eventually, the FQP finds the optimal base-stock level with the corresponding optimal α^*.

Recalling that $g_{n+1}(\cdot)$ is the PDF of the sojourn time T_{n+1} for a customer seeing n orders upon arrival,

$$\alpha_n = \int_0^{d_n} g_{n+1}(x)dx = G_{n+1}(d_n),$$

is the probability that such a customer receives the finished item within d_n. If $G_{n+1}(\cdot)$, the cumulative distribution function (CDF) of T_{n+1}, is available in closed-form and easy to invert, we can first decide on α_n for a customer seeing n orders for each $n = S, \ldots$, and then solve for

$$d_n = G_{n+1}^{-1}(\alpha_n),$$

to determine what should be announced as the lead-time. A policy giving optimal α_n^* (possibly different for different n) for all n is the optimal policy in the $M/M/1$ setting. However, identifying the optimal lead-time for each n by using $G_{n+1}^{-1}(\cdot)$ (assuming that it is available) would be difficult since this would require an ambitious search considering all possible vectors of α_n. Additionally, determining whether a finite K value (such that $\lambda_K = 0$) exists would cause additional difficulty.

In contrast, under the FQP we propose setting $\alpha_n = \alpha$ for all $n \geq S$ and then, obtain the optimal α^* for a given S. Determining this optimal α^* for all customers is not easy, either, because a closed-form $G_{n+1}(\cdot)$ usually does not exist and a direct computation of $G_{n+1}^{-1}(\alpha)$ is not possible. Yet, we can numerically invert its Laplace transform, which is $\widetilde{g}_{n+1}(s)/s$. Then, by conducting a binary search over the interval for possible d values we arrive at $L^{-1}\{\widetilde{g}_{n+1}(s)/s\}(d_n) = G_{n+1}(d_n) = \alpha$. Note that other search methods, such as the interpolation search, can be considered as well, especially when such policies are applied in more complex systems (e.g., multi-class systems) but we leave this investigation for future research. Using this idea in the following FQP algorithm, the optimal FQP parameters can be found.

In summary, the FQP performs the following: Given S and α, for a customer that sees n orders in the system, it generates a candidate lead-time d as the midpoint of a search interval whose lower and upper bounds/limits can be updated if it is necessary (initially, the search starts with the [0,$d_{\max}$] interval). If $L^{-1}\{\widetilde{g}_{n+1}(s)/s\}(d) = \alpha$ is achieved, this is the lead-time d_n to announce and the search continues with the customer seeing $n+1$ orders. Otherwise, if $L^{-1}\{\widetilde{g}_{n+1}(s)/s\}(d) > \alpha$ ($L^{-1}\{\widetilde{g}_{n+1}(s)/s\}(d) < \alpha$) , this indicates that the next candidate lead-time should be smaller (larger) than d, so the upper (lower) limit of the search interval is updated and set to d. The midpoint of the updated interval gives the next candidate lead-time. In this process, eventually the candidate lead-time hits $d_{\max}$, which gives the K value for which $\lambda_K = 0$. Then, the algorithm, having all the lead times and the state-dependent arrival rates, computes the profit. The algorithm is run for all $S \leq S_0^*$ and α values to identify the optimal policy parameters.

The Fair Quotation Policy Algorithm: This algorithm explains how the optimal FQP parameters, S_{FQ}^*, α_{FQ}^*, and $d_{FQ,n}^*$ for $n \geq S_{FQ}^*$, are found.

Initialization Step. Using λ, $\widetilde{b}(\theta)$, h, and l (as the backlogging cost) in the model of Sanajian and Balcıoğlu (2009), obtain S_0^*. $\lambda R - C(S_0^*)$ is the optimal profit for $\alpha = 0$ and can be the optimal solution if cases with nonzero lead times, generated in the Main

Step, do not yield a higher profit.

Choose ϵ_α as the acceptable error margin around α and ϵ as a measure showing that the lead-time needed exceeds $d_{\max}$ which stops the search and gives the value for K.

Main Step. This step is executed for all $S(= 0, \ldots, S_0^*)$ and α values to consider. Given S and α, set $d_0 = d_1 = \ldots = d_{S-1} = 0$, (if $S = 0$, d_0 need not be 0) and LB=0 and UB=$d_{\max}$, respectively, as the lower and upper limits for the interval for possible d values over which the binary search is conducted in Step 1. With $n = S$, go to Step 1.

Step 1 Set d_n =(LB+UB)/2.

Step 1.a If $|d_n - d_{\max}| < \epsilon$ (implying that d_n sought exceeds $d_{\max}$, hence $n = K$) then go to Step 2. Else go to Step 1.b.

Step 1.b If $L^{-1}\{\widetilde{g}_{n+1}(s)/s\}(d_n) = G_{n+1}(d_n) = \alpha \pm \epsilon_\alpha$, then store d_n as the lead-time to announce to the customer seeing n orders upon arrival. Increment n by 1, reset LB=0 and UB=$d_{\max}$, and go to Step 1. Else go to Step 1.c.

Step 1.c If $L^{-1}\{\widetilde{g}_{n+1}(s)/s\}(d_n) = G_{n+1}(d_n) < \alpha$ (implying that a longer lead-time is needed), then set LB=d_n and go to Step 1. If $L^{-1}\{\widetilde{g}_{n+1}(s)/s\}(d_n) = G_{n+1}(d_n) > \alpha$ (implying that a shorter lead-time is needed), set UB=d_n and go to Step 1.

Step 2 Since $d_K = d_{\max}$ is achieved, the vector $\mathbf{d}$ is constructed. Compute and store the profit for the current S and α values.

Final Step After the Main Step is executed for all S and α values, the case yielding the highest profit (which could be the one found in the Initialization Step for $\alpha = 0$) gives the optimal FQP profit $P_{FQ}(S_{FQ}^*, \mathbf{d}_{FQ}^*)$ along with its parameters.

Note that increasing n leads to the announcement of $d_{\max}$ eventually at Step 1 which also yields K, the maximum number of orders permitted in the system. The FQP algorithm does not assume that the optimal profit is unimodal in S or in α (for a given S).

Since $\widetilde{g}_{n+1}(s)$ in Eq. (9) via $\widetilde{h}_n(s)$ in Eq. (4) depends on S, the FQP algorithm may give different $\mathbf{d}$ vectors for different S when service times are non-exponential. In the $M/M/1$ setting, for any S value, the distribution of T_{n+1}, which is Erlang$(\mu, n-S+1)$, is independent of S for $n \geq S$. Thus, for all S values the FQP announces the same $K-S$ nonzero lead times for $d_S, \ldots, d_{K-1}, d_K = d_{\max}$ for a given α value. This implies that the Main Step, with Steps 1 and 2 in the algorithm, is run only once, let us say for $S=0$, for all α values. In the Final Step, for each $S > 0$, the profit is computed after constructing the corresponding vector $\mathbf{d}$ with 0's in the first $S-1$ entries and the last $K-S+1$ entries coming from the Main Step for the α considered. The optimal FQP parameters are consequently determined.

3.2 The Preferential Quotation Policy

The short lead times quoted under the FQP when the number of backlogged customers is small can discourage most of the customers if they are oversensitive. This prevents the system from generating more revenues which, at the expense of slight penalty cost increases, may imply higher profitability. To circumvent this problem, we propose modifying the lead-time vector of the FQP by announcing zero lead times instead of the short nonzero lead times in $\mathbf{d}^*_{FQ}$. Consequently, revenues may increase with more customers placing orders, and when d_n is replaced with 0 in Eq. (2) the penalty cost incurred may not increase significantly. With this approach, higher profit can be reaped. This revision can be supplemented by admitting fewer customers, i.e., by announcing $d_{\max}$ sooner than the FQP. This would lower the maximum number of backlogged customers permitted by the PQP below $K-S$ (K of the FQP) for a given S. In other words, the PQP tolerates a slight penalty cost increase for those customers the FQP quotes short nonzero lead times but in return declines serving customers for whom the FQP quotes lead times close to $d_{\max}$. In short, when compared to the FQP, the PQP we propose in this section "prefers" early backlogged customers over later arrivals.

The PQP determines the maximum number of customers to backlog $K' - S$ and how

many of them will be quoted zero lead times. The difference of these two numbers gives how many customers are announced nonzero lead times. For such customers, to be lined at the end of the backlog queue, – similar to the FQP – the PQP finds the optimal α^*_{PQ}, i.e., the probability of producing and delivering the item within the quoted lead-time. Since the PQP makes use of the vector $\mathbf{d}_{FQ}$ constructed by the FQP, it may only increase the profit as explained in the following PQP algorithm. Observe that if the PQP leads to a profit increase when compared to the FQP, the probability of satisfying the demand within the quoted lead times for backlogged customers to whom zero lead times are quoted is 0 and does not equal $\alpha^*_{PQ} > 0$.

In summary, the PQP performs the following: Given S and the corresponding lead-time vector $\mathbf{d}^*_{FQ}$ from the FQP, it announces zero lead times to the first, second, and so forth backlogged customers as long as the profit increases. Then, it starts rejecting the last backlogged customer, the second last backlogged customer and so forth as long as the profit increases. These two steps are iteratively repeated until no more profit increase is observed. In the lead-time vector updated this way, nonzero lead times may remain from the $\mathbf{d}^*_{FQ}$ vector to announce the last customers the PQP is to backlog. For them, similar to the FQP in Section 3.1, new lead times giving possibly a different optimal α^*_{PQ} are searched.

To demonstrate how the PQP algorithm works, consider the following example given in Table 1. In Row 0, we have the entries in $\mathbf{d}^*_{FQ}$ from 0.02 to 3.51 which are, respectively, the quoted lead times to customers seeing S to $S+8$ orders in the system (the customer seeing $S+9$ is rejected by announcing her $d_{\max} = 4$ as the lead-time). In Row 1, Step 1 of the PQP algorithm is executed which indicates that quoting 0 instead of 0.02 to the customer seeing S orders upon arrival increases the profit. In Row 2, Step 2 is executed twice which indicates that quoting $d_{\max}$ to the $S+8$th customer and then to the $S+7$th customer increases the profit. Thus, the maximum number of backlogged customers decreases by 2 (from $S+8$ to $S+6$). After Steps 1 and 2 are run again, as illustrated in Rows 3 and 4, respectively, in Row 5, we arrive at the final $\mathbf{d}^*_{PQ}$ after Step 3 is executed.

The Preferential Quotation Policy Algorithm: This algorithm explains how the opti-

Table 1: Application of the PQP algorithm

No. Orders		S	$S+1$	$S+2$	$S+3$	$S+4$	$S+5$	$S+6$	$S+7$	$S+8$
Row 0	$\mathbf{d}^*_{FQ}$:	0.02	0.15	0.44	0.83	1.28	1.79	2.34	2.91	3.51
Row 1	Step 1	0	0.15	0.44	0.83	1.28	1.79	2.34	2.91	3.51
Row 2	Step 2	0	0.15	0.44	0.83	1.28	1.79	2.34	4	4
Row 3	Step 1	0	0	0.44	0.83	1.28	1.79	2.34	4	4
Row 4	Step 2	0	0	0.44	0.83	1.28	4	4	4	4
Row 5	Step 3/$\mathbf{d}^*_{PQ}$:	0	0	0.6	1.15	2.30	4	4	4	4

mal PQP parameters, S^*_{PQ}, and $d^*_{PQ,n}$ for $n \geq S^*_{PQ}$, are found.

Main Step. This step is executed for all $S(= 0, \ldots, S^*_0)$ and corresponding $\mathbf{d}^*_{FQ}$ vectors from the FQP. Use C as a counter which is incremented by 1 when profit does not increase. Set $C = 0$. With $n = S$ and $K^{'} = K$, go to Step 1.

Step 1 Set $d_n = 0$. If the profit increases, increase n by 1, set $C = 0$ and visit Step 1 again. Otherwise, retain the nonzero d_n from $\mathbf{d}^*_{FQ}$, increment C by 1. If $C = 1$, go to Step 2, otherwise go to Step 3.

Step 2 Set $d_{K'} = d_{\max}$. If the profit increases, decrease $K^{'}$ by 1, set $C = 0$ and visit Step 2 again. Otherwise, retain the nonzero $d_{K'}$ from $\mathbf{d}^*_{FQ}$, increment C by 1. If $C = 1$, go to Step 1, otherwise go to Step 3.

Step 3 Now we have a lead-time vector of size $K^{'}$ with the first n entries, $K^{'} > n \geq S$, being zero. Implement the Main Step of the FQP algorithm to obtain the α^*_{PQ} and the corresponding nonzero lead times for the last $K^{'} - n$ entries to finalize $\mathbf{d}^*_{PQ}$. Compute and store the profit for the current S and $\mathbf{d}^*_{PQ}$.

Final Step Find the case that yields the highest profit among those stored in Step 3 which gives the optimal PQP profit $P_{PQ}(S^*_{PQ}, \mathbf{d}^*_{PQ})$ along with its parameters.

4 Numerical Experiment

In this section, we address two questions: (i) How does the FQP proposed in Section 3.1 perform with respect to the optimal policy in the $M/M/1$ setting? Can the PQP presented in Section 3.2 bring additional improvements when compared to the FQP? (ii) How does the variability in service times affect the profitability in the $M/GI/1$ setting when these dynamic lead-time quotation policies are employed instead of accepting all customers by announcing zero lead times?

For the proposed policies, whenever a Laplace transform is required to be inverted, we use the Euler technique due to Abate and Whitt (1995). The numerical inversion technique by Jagerman (1982) can be equivalently employed (see Appendix A of Jagerman and Melamed, 2003, for the algorithm of this technique). We set $\epsilon_\alpha = 0.001$ (the acceptable error margin around α in Step 1.b of the FQP algorithm), $\epsilon = 0.00001$ (as the measure showing that $d_{\max}$ has been reached in Step 1.a of the FQP algorithm). We have considered $\alpha = 0.01k$, $k = 1, \ldots, 99$.

4.1 Numerical Experiments in the $M/M/1$ Setting

In this section, we repeat the numerical study conducted by Savaşaneril, Griffin, and Keskinocak (2010) who provide us with the optimal profit P_E^* as reference values. Five values of R, $R \in \{5, 7.5, 10, 15, 25\}$, seven values of h, $h \in \{0.15, 0.3, 0.6, 0.9, 1.2, 1.5, 1.8\}$, and seven values of λ, $\lambda \in \{0.15, 0.3, 0.45, 0.6, 0.75, 0.9, 0.99\}$, are considered. The mean service time is 1. In all examples, the penalty cost rate is constant as $l = 1.5$. The six lead-time acceptance probability functions considered are presented in Table 2 and Figure 1. When compared, these six functions, in two groups with respect to $d_{\max}$, can be ordered as Convex1, Linear1, Concave1, (Convex2, Linear2, Concave2) in capturing the behaviors of customers from the most sensitive to the least to the quoted nonzero lead times.

For each $f(d)$ function, a total of $5 \times 7 \times 7 = 245$ (due to five R, seven h, and seven λ values considered) examples were used to test the performance of the FQP when compared

Table 2: The lead-time acceptance probability functions

Name	$d_{\max}$	Function
Convex1	4	$f(d) = 1 - \left(\frac{d}{4}\right)^{1/4}$
Convex2	8	$f(d) = \begin{cases} 1 - \frac{5}{8}d & \text{for } 0 \le d \le 1 \\ \frac{3}{8} - \frac{3}{8}\frac{1}{7}(d-1) & \text{for } 1 \le d \le 8 \end{cases}$
Concave1	4	$f(d) = 1 - \left(\frac{d}{4}\right)^4$
Concave2	8	$f(d) = 1 - \left(\frac{d}{8}\right)^4$
Linear1	4	$f(d) = 1 - \frac{d}{4}$
Linear2	8	$f(d) = 1 - \frac{d}{8}$

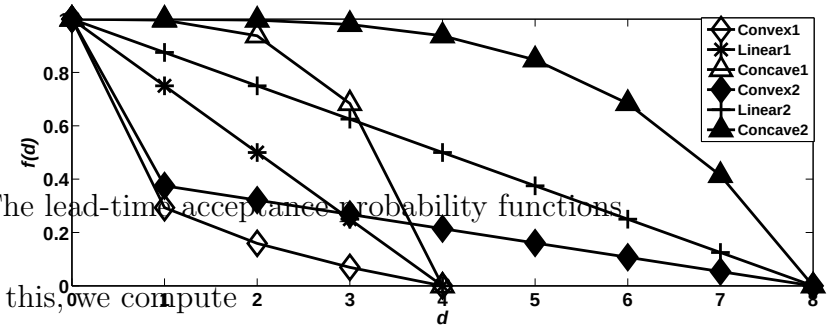

Figure 1: The lead-time acceptance probability functions

to the optimal policy. To do this, we compute

$$\Delta^*_{FQ} \equiv \frac{P(S^*_{FQ}, \mathbf{d}^*_{FQ}) - P^*_E}{P^*_E},$$

where $P(S^*_{FQ}, \mathbf{d}^*_{FQ})$ is the optimal profit under the FQP and P^*_E the profit of the optimal policy. The ratio Δ^*_{FQ} measures the profit decrease when the FQP is used instead of the optimal policy. The summary of Δ^*_{FQ} of 245 experiments for each acceptance function is presented in Table 3.

We see that the FQP performs remarkably well in most of the cases. The highest deviations from the optimal policy are observed for Convex1 acceptance probability function,

Table 3: The minimum, mean, median and maximum values of profit loss (Δ^*_{FQ}) when the FQP is used instead of the optimal policy.

	Min (%)	Mean(%)	Median(%)	Max (%)
Convex1	0	-2.67	-1.29	-19.05
Convex2	0	-0.49	-0.29	-4.32
Concave1	-0.27	-1.78	-1.66	-3.75
Concave2	-0.21	-1.04	-1.03	-1.99
Linear1	0	-0.12	-0.03	-3.87
Linear2	0	-0.07	-0.01	-3.29

but even for that group of experiments the mean profit loss due to using the FQP is 2.67%. From the detailed analysis of the numerical results, we make the following observations for the FQP:

- With higher R or λ, and lower h values, the system tends to carry more inventory.
- The system holds higher levels of stock when customers are more sensitive (declining from higher stock levels held for Convex1/2 to lower levels for Concave1/2).
- We also note the following:
  - When λ and h are fixed, increase in R decreases α.
  - When λ and R are fixed, increase in h decreases α.
  - When h and R are fixed, increase in λ increases α. However, this is not the case for Concave1 and Concave2 functions for which with increasing λ, we see that α tends to decrease, sometimes leveling off at the lowest value and sometimes increasing again from this lowest value.

  Higher/smaller α values imply longer/shorter lead times which are quoted by the supplier. Thus, via shorter lead times announced (smaller α), the supplier tries to sell

more when the revenue gets higher and holding inventory becomes costlier. Except for cases with Concave1 and Concave2 functions, higher arrival rates are reduced by announcing longer lead times.

The FQ performance deteriorates significantly if S is set to 0 when Convex1 acceptance probability function is considered. We generated 49 make-to-order examples with seven R and seven λ values, and for all cases we computed Δ^*_{FQ}, the summary of which is presented in Table 4. In 17 out of these 49 examples, Δ^*_{FQ} was below -8%. When we implemented the PQP, it turned out to give the optimal profit in 15 out of these 17 cases, and for the remaining two cases, the $P_{PQ}(S^*_{PQ}, \mathbf{d}^*_{PQ})$ was only -0.32% less than the profit of the optimal policy. The PQP appears to bring significant improvements over the FQP for Convex1 acceptance probability function.

For other acceptance probability functions, we observe that the PQP usually does not yield any improvements over the FQP in instances with the lowest Δ^*_{FQ}. This is because of higher α^* values quoted by the FQP especially for Concave acceptance probability functions. In such cases, the FQP quotes large nonzero lead times without losing many customers. Thus, replacing smallest nonzero lead times by 0 does not increase the rate of customers placing orders. Moreover, smaller penalty costs incurred under the FQP increase when 0 is substituted for a large nonzero d_n in Eq. (2). It follows that the PQP seems to provide improvement only for oversensitive customers whose behavior is best captured by Convex (and sometimes Linear) acceptance probability functions in this study.

Table 4: The minimum, mean, median and maximum values of profit loss (Δ^*_{FQ}) when the FQP is used instead of the optimal policy in make-to-order queue with Convex1 acceptance function.

Min(%)	Mean(%)	Median(%)	Max(%)
0	-10.2	-10	-25

4.2 Numerical Experiments in the $M/GI/1$ Setting

In this section, we revisit some numerical examples generated by Sanajian and Balcıoğlu (2009) who provide us with the optimal profit $P_0^* = \lambda R - C(S_0^*)$ if zero lead times are announced to all customers. In order to observe the impact of service time variability on the system profit, we consider different service time distributions with unit mean ($\mu = 1$), but different variances. In other words, the measure of variability in service time distribution is its squared-coefficient of variation (variance-to-mean ratio, which coincides with the variance in these examples) which is denoted by c_S^2. If c_S^2 is higher, the service time is deemed more variable. For our numerical examples, we consider the following three service time distributions, each presented with its density function Laplace transform:

1. The deterministic service time with the density function Laplace transform $\widetilde{b}(\theta) = e^{-\theta}$.

2. The exponential distribution with $\mu = 1$ and the density function Laplace transform

$$\widetilde{b}(\theta) = \frac{\mu}{\mu + \theta}.$$

3. The 2-stage mixed generalized Erlang (MGE2) distribution with $\mu_1 = 1.218, \mu_2 = 0.082, a_1 = 0.015$ and the density function Laplace transform

$$\widetilde{b}(\theta) = \frac{\mu_1\mu_2 + \mu_1(1-a)\theta}{\theta^2 + (\mu_1 + \mu_2)\theta + \mu_1\mu_2}.$$

Note that with probability $1 - a_1$ (a_1), an MGE2 r.v. is a an exponential r.v. with rate μ_1 (sum of two exponential r.v.s with rates μ_1 and μ_2).

In all examples, the holding cost, penalty cost rates and R are constants as $h = 1$, $l = 1$ and $R = 15$, respectively. Two values of λ, $\lambda \in \{0.7, 0.8\}$, are considered for all acceptance probability functions.

In Table 5 the first column displays the different service time distributions that are considered with their squared-coefficient of variation listed in the second column. The third column gives the Poisson arrival rate λ. The fourth column for P_0^* exhibits the profit when all

Table 5: Performance of the FQP in the $M/GI/1$ queue with Convex1, Linear1, and Concave1 lead-time acceptance probability functions

Service Time				Convex1	Linear1	Concave1
Distribution	c_S^2	λ	P_0^*	$P_{FQ}(S_{FQ}^*, \mathbf{d}_{FQ}^*)$	$P_{FQ}(S_{FQ}^*, \mathbf{d}_{FQ}^*)$	$P_{FQ}(S_{FQ}^*, \mathbf{d}_{FQ}^*)$
Deterministic	0		9.38	9.38	9.38	9.73
Exponential	1		8.57	8.57	8.73	9.11
		0.7	(-8.64%)	(-8.64%)	(-6.93%)	(-6.39%)
MGE2	5		5.34	7.34	7.94	8.24
			(-43.07%)	(-21.77%)	(-15.40%)	(-15.29%)
Deterministic	0		10.31	10.31	10.31	10.95
Exponential	1		8.9	8.96	9.71	10.09
		0.8	(-13.68%)	(-13.09%)	(-5.81%)	(-7.78%)
MGE2	5		2.67	8.21	8.75	9.14
			(-74.10%)	(-20.39%)	(-15.11%)	(-16.49%)

customers are quoted zero lead times. The terms in parenthesis capture the relative decrease in profit with respect to the system profit with deterministic production times. For instance, in the case of MGE2 service times, when $\lambda = 0.8$, the relative profit loss is 74.10% if zero lead times are quoted to all arrivals. For all policies, higher service time variability increases the profit loss compared to the base case with deterministic service times.

When the dynamic FQP is used to quote lead times, the profits tend to increase for all service time distributions. Yet, the increase in profit is more significant for the MGE2 service times. For instance, when the FQP is employed, the profits displayed in the fifth (Convex1) column show that the relative profit loss is only 21.77% and 20.39% when compared to the base case with deterministic service times when $\lambda = 0.7$ and 0.8, respectively. The last two columns list the generated profits when the FQP is employed for customers having Linear1 and Concave1 lead-time acceptance probability functions. As customers become less

sensitive to nonzero lead times from Convex1 to Linear1 and then to Concave1, and more customers tend to place orders, in each row, we see that the profits also tend to increase. This observation highlights the fact that a company should seek ways to gain confidence of its customers (such as producing a high quality product) to make them less sensitive to quoted lead times. This can help the company raise its profitability.

Table 6: Performance of the FQP in the $M/GI/1$ queue with Convex2, Linear2, and Concave2 lead-time acceptance probability functions

Service Time				Convex2	Linear2	Concave2
Distribution	c_S^2	λ	P_0^*	$P_{FQ}(S_{FQ}^*, \mathbf{d}_{FQ}^*)$	$P_{FQ}(S_{FQ}^*, \mathbf{d}_{FQ}^*)$	$P_{FQ}(S_{FQ}^*, \mathbf{d}_{FQ}^*)$
Deterministic	0		9.38	9.38	9.38	10.27
Exponential	1		8.57	8.57	8.85	9.52
		0.7	(-8.64%)	(-8.64%)	(-5.66%)	(-7.33%)
MGE2	5		5.34	7.77	8.03	8.43
			(-43.07%)	(-17.14%)	(-14.38%)	(-17.9%)
Deterministic	0		10.31	10.31	10.49	11.49
Exponential	1		8.9	9.54	9.84	10.65
		0.8	(-13.68%)	(-7.49%)	(-6.2%)	(-7.3%)
MGE2	5		2.67	8.55	8.84	9.25
			(-74.10%)	(-17.08%)	(-15.67%)	(-19.16%)

Table 6 is structurally the same as Table 5 except that in the last three columns, we list the generated profits when the FQP is employed for customers having Convex2, Linear2, and Concave2 lead-time acceptance probability functions. From Figure 1, we see that customers with Convex2, Linear2, and Concave2 lead-time acceptance functions are more likely to place orders compared to customers with Convex1, Linear1, and Concave1 functions, respectively. Thus, in Table 6, profits tend to be higher than those in Table 5. Otherwise, we see the same impact of the dynamic FQP in reducing the worsening impact of production time variability.

Table 7: Performance of the PQP in the $M/GI/1$ queue

Service Time			Convex1	Linear1	Convex2	Linear2
Distribution	c_S^2	λ	$P_{PQ}(S_{PQ}^*, \mathbf{d}_{PQ}^*)$	$P_{PQ}(S_{PQ}^*, \mathbf{d}_{PQ}^*)$	$P_{PQ}(S_{PQ}^*, \mathbf{d}_{PQ}^*)$	$P_{PQ}(S_{PQ}^*, \mathbf{d}_{PQ}^*)$
Deterministic	0		9.40	9.40	9.41	9.42
Exponential	1		8.75	8.78	8.76	8.86
		0.7	(-6.95%)	(-6.60%)	(-6.82%)	(-5.98%)
MGE2	5		8	8.01	8.01	8.06
			(-14.89%)	(-14.84%)	(-14.82%)	(-14.46%)
Deterministic	0		10.46	10.46	10.47	10.52
Exponential	1		9.67	9.72	9.69	9.85
		0.8	(-7.5%)	(-6.99%)	(-7.38%)	(-6.41%)
MGE2	5		8.81	8.82	8.81	8.84
			(-15.77%)	(-15.66%)	(-15.90%)	(-15.98%)

When the PQP is used for the cases presented in Tables 5 and 6, profits increased for Convex1, Convex2, Linear1, and Linear2 acceptance probability functions. These results are presented in Table 7. Again profit loss with respect to the cases with deterministic service times are presented in parentheses. When we compare the Convex1 and Convex2 columns in Table 7 with Convex1 column in Table 5 and Convex2 column in Table 6, respectively, we see that under the PQP profits increase more for both exponential and MGE2 distributions. This indicates that choosing a dynamic lead-time quotation policy more suitable for the customer profile can further decrease the worsening impact of service time variability.

We close this section by commenting on the computation times of running the proposed algorithms. Both algorithms have been implemented in Matlab and run on a Windows-based computer with Intel i5 CPU and 4.0 GB RAM. The computation times highly vary depending on the complexity of the service time Laplace transform and the maximum number

of orders allowed in the system. Higher S, higher arrival rates (a function of higher λ_0 and the lead-time acceptance probability function), smaller α, higher $d_{\max}$ tend to increase the maximum number of orders allowed in the system. Deterministic service times have a simpler Laplace transform when compared to that of the MGE2 distribution. For instance, for deterministic service time, when $S = 0$, $\lambda_0 = 0.7$, $\alpha = 0.99$, $d_{\max} = 4$ with Concave1 lead-time acceptance probability function, the FQ algorithm determines the lead times to announce to three backlogged customers and computes the profit in 0.75 seconds. On the other hand, for the MGE2 service time, when $S = 3$, $\lambda_0 = 0.7$, $\alpha = 0.01$, $d_{\max} = 8$ with Linear2 lead-time acceptance probability function, the FQ algorithm determines the lead times to announce to 15 backlogged customers and computes the profit in 4.41 hours. While it takes 1.3 seconds to find the lead-time for the first backlogged customer, it goes up to 46.61 minutes for the 15th backlogged customer. We can see that $\widetilde{g}_{17}(\theta)$ in Eq. (9) for this customer is quite complex and its numerical inversions needed to be done for the binary search increase the time to identify the lead-time. In this 4.41 hours, determining the entire lead-time vector $\mathbf{d}^*_{FQ}$ takes 1.67 hours. The difference of 2.73 hours is spent for computing the steady-state probabilities and the profit which involves again the Laplace transform inversions of $\widetilde{g}^{(1)}_{n+1}(\theta)$ $n = 3, \ldots, 17$ (see Eq. 10 and the ensuing discussion).

5 Conclusion and Future Work

In this paper, we propose two practical dynamic lead-time quotation policies for a company producing a single type of product. The production facility is modeled as an $M_n/GI/1/K$ queue. Both policies employ numerically inverting the Laplace transform of the sojourn time r.v. of an order to be placed. Therefore, the idea has the potential extension in other make-to-stock queues where the sojourn time Laplace transforms are available. An immediate extension we plan to pursue is the multiclass $M_n/GI/1$ queue in which different priority classes can demand the same type of product. Such an extension would incorporate multilevel rationing policy as the inventory control. A serendipitious result of this study is

that the proposed dynamic lead-time quotation policies help reduce the worsening impact of the production time variability. In the future, new dynamic lead-time quotation policies can be designed and for them and the proposed policies in this study as well, especially when they are implemented in multi-class settings, the computational performance may need be improved. To do this, other search techniques than the binary search method employed here can be considered while identifying the lead-time guaranteeing the probability of service considered. A faster search algorithm can determine the lead times whereas the profit computations can be obtained from a discrete-event simulation model using these lead times and base-stock level as input to help reduce the computation times. Another important future research effort would be the exploration of the optimal policy for systems with non-exponential service times, even if via numerical methods, which would provide reference optimal profit values against which researchers would compare the performance of the dynamic policies they would design.

Acknowledgements

This work was supported in part by TÜBİTAK, The Scientific and Technological Research Council of Turkey, under the grant number 213M428. We would like to thank Seçil Savaşaneril who provided us with their numerical results and answered our questions whenever we needed clarifications on their paper. We would also like to extend our thanks to Gabor Rudolf, Mustafa Yavaş and Şafak Yücel who have helped us at several stages of our research. The authors thank the two anonymous referees and the editors for their invaluable suggestions to improve the manuscript.

References

Abate, J., W. Whitt. 1995. "Numerical inversion of Laplace transforms of probability distributions", *ORSA Journal on Computing*, Vol. 7, 36–43.

Abouee-Mehrizi, H., O. Baron. 2015. "State-dependent $M/G/1$ queueing systems", *forthcoming*, *Queueing Systems*.

Çelik, S., C. Maglaras. 2008. "Pricing and lead-time quotation for a multiclass make-to-order queue", *Management Science*, Vol. 54, No. 6, 1132–1146.

Dellaert, N. P. 1991. "Due-date setting and production control", *International Journal of Production Economics*, Vol. 23, 59–67.

Duenyas, I. 1995. "Single facility due date setting with multiple customer classes", *Management Science*, Vol. 41, No. 4, 608–619.

Duenyas, I., W. J. Hopp. 1995. "Quoting customer lead times", *Management Science*, Vol. 41, No. 1, 43–57.

Economou, A., A. Manou. 2015. "A probabilistic approach for the analysis of the $M_n/G/1$ queue", *forthcoming*, *Annals of Operations Research*.

Elhafsi M. 2000 "An operational decision model for lead-time and price quotation in congested manufacturing systems," *European Journal of Operational Research*, Vol. 126, No. 2, 355–370.

Gallien, J., Y. L. Tallec, T. Shoenmeyr. 2004. "A model for make-to-order revenue management", *Working paper*, Massachusetts Institute of Technology, Cambridge, MA.

Gross, D., C. M. Harris. 1998. *Fundamentals of Queueing Theory*. John Wiley & Sons, New York.

Hopp, W. J., M. R. Sturgis. 2001. "A simple, robust leadtime-quoting policy", *Manufacturing & Service Operations Management*, Vol. 3, No. 4, 321–336.

Jagerman, D. L. 1982. "An inversion technique for the Laplace transform", *Bell System Technical Journal* Vol. 61, No. 8, 1995–2002.

Jagerman, D. L., B. Melamed. 2003. "Models and approximations for call center design", *Methodology and Computing in Applied Probability*, Vol. 5, No. 2, 159–181.

Kapuscinski, R., S. Tayur. 2007. "Reliable due-date setting in a capacitated MTO system with two customer classes", *Operations Research*, Vol. 55, No. 1, 56–74.

Kerner, Y. 2008. "The conditional distribution of the residual service time in the $M_n/G/1$ queue", *Stochastic Models*, Vol. 24, 364–375.

Keskinocak, P., R. Ravi, and S. Tayur. 2001. "Scheduling and reliable lead-time quotation for orders with availability intervals and lead-time sensitive revenues", *Management Science*, Vol. 47, No. 2, 264–279.

Lawrence, S. R. 1995. "Estimating flowtimes and setting due-dates in complex production systems", *IIE Transactions*, Vol. 27, 657–668.

Li, L. 1992. "The role of inventory in delivery-time competition", *Management Science*, Vol. 38, No. 2, 182–197.

Li, J-W. 2003. "Improving the performance of job shop manufacturing with demand-pull production control by reducing set-up/processing time variability", *International Journal of Production Economics*, Vol. 84, No. 3, 255–270.

Lin, B. M. T. 2001. "Scheduling in the two-machine flowshop with due date constraints", *International Journal of Production Economics*, Vol. 70, 117–123.

Rajagopalan, S. 2002. "Make to order or make to stock: Model and application", *Management Science*, Vol. 48, No. 2, 241–256.

Sanajian, N., B. Balcıoğlu. 2009. "The impact of production time variability on make-to-stock queue performance", *European Journal of Operational Research*, Vol. 194, 847–855.

Savaşaneril, S., P. M. Griffin, and P. Keskinocak. 2010. "Dynamic lead-time quotation for an M/M/1 base-stock inventory queue", *Operations Research*, Vol. 58, No. 2, 383–395.

Slotnick, S. A., M. J. Sobel. 2005. "Manufacturing lead-time rules: Customer retention versus tardiness costs". *European Journal of Operational Research*, Vol. 163, No. 3, 825–856.

So, K. C., J.-S. Song. 1998. "Price, delivery time guarantees and capacity selection", *IIE Transactions*, Vol. 111, 28–49.

Spearman, M. L., R. Q. Zhang. 1999. "Optimal lead time policies", *Management Science*, Vol. 45, No. 2, 290–295.

Unal, A. T., R. Uzsoy, and A. S. Kiran. 1997. "Rescheduling on a single machine with part-type dependent setup times and deadlines", *Annals of Operations Research*, Vol. 70, 93–113.

Wein, L. M. 1991. "Due-date setting and priority sequencing in a multiclass $M/G/1$ queue", *Management Science*, Vol. 37, No. 7, 834–850.