https://doi.org/10.21872/2024IISE_6090

*Proceedings of the IISE Annual Conference & Expo 2024*
*A. Brown Greer, C. Contardo, J.-M. Frayret, eds.*

# Online Lead Time Quotation under Contingency

**Kelvin Ekun and Ana Muriel**
**Department of Mechanical and Industrial Engineering**
**University of Massachusetts, Amherst, MA**

## Abstract

The success of make-to-order (MTO) firms relies on providing competitive quotations to customers and ensuring quoted lead times align with the ability to fulfill customized orders as promised. While customers can take some time to accept or reject a quotation, the firm is liable to meet quoted lead times or otherwise face tardiness penalties. The firm must continuously provide quotations under contingent demand information, pending decisions of previous customers. Since the probability of order acceptance decreases as the quoted lead time increases, the firm benefits from optimistic quotations and hedges against losing orders by overbooking resources. We model the online lead time quotation and production scheduling problem under contingent demand and propose a primal-dual algorithm to maximize the total profit from arriving orders in a single-machine environment. Profit is defined as the net revenue penalized by the cost of tardiness. We evaluate the performance of our proposed algorithm against conventional rules and an all-knowing offline oracle. Simulation experiments show significant savings associated with the primal-dual approach, highlighting the need for future research on online lead time quotation under contingent demand.

## Keywords

Lead time quotation, contingent orders, primal-dual algorithm

## 1. Introduction

Make-to-order (MTO) firms manufacture customized products upon order request using resources with limited capacities. Each arriving customer submits a Request for Quotation (RFQ) outlining their desired product specifications. The MTO manufacturer responds by quoting the order's price, production lead time, and the time by which the customer must make their decision. While customers may accept or reject a quotation immediately, the more prevalent practice is to contact competing firms to compare quotations and negotiate terms. During this time, the MTO firm will entertain other customer requests that require limited resource capacities under the uncertainty of contingent orders. An order is considered contingent if its requesting customer is yet to accept or reject the quoted terms.

In this setting, a risk-seeking MTO manufacturer could provide an aggressive quotation to the new arriving customer without considering existing contingent orders, thereby overbooking the capacity of the shared resources. However, if customers eventually accept quotations, the MTO potentially faces production congestion and order fulfillment delays, which can negatively impact the firm's reputation and prevent repeat business. On the other extreme, a risk-averse MTO manufacturer may assume contingent orders as already confirmed and conservatively quote a new order with the resulting long lead time. If any customer eventually rejects the quotation, the MTO firm will face resource underutilization. The firm would clearly benefit from allocating some reserved capacity from contingent orders to a new arriving order, quoting a shorter lead time that the customer is more likely to accept.

In this paper, we propose an online algorithm for quoting order lead times in the presence of contingent demand. While orders are received continuously over time and quoted lead times upon arrival, production plans are updated at regular intervals and scheduling decisions are only finalized at the beginning of each period, based on the set of outstanding confirmed orders at that time. Our contribution is the proposal and performance analysis of an online algorithm that balances aggressive and conservative quotations in the presence of contingent orders, utilizing information regarding the firm's capacity as well as each order's expected profit and expected processing capacity to quote lead times.

## 2. Literature Review

Lead time quotation is a central topic in due date management and an essential aspect of managing revenue in MTO firms, such as precision manufacturers and rapid prototyping companies. With a growing literature, several studies

https://doi.org/10.21872/2024IISE_6090

have investigated the critical decision of quoting lead times to arriving customers and its benefits to an MTO firm's success and competitive advantage [1, 2]. The impact of contingent demand in the MTO bidding process, however, remains a gap in the existing literature, addressed by only a limited number of studies. Easton and Moodie [3] are referenced as the first to explicitly consider contingent demand in price and lead time quotation. They consider one order arriving into a single resource environment with confirmed and contingent workload existing in the production pipeline. The authors apply an enumerative offline approach considering all potential outcomes (acceptance or rejection) for contingent orders, resulting in $2^n$ scenarios, where $n$ is the number of contingent orders. Cakravastia and Takashi [4] expand on Easton and Moodie's model to consider multiple resources and multiple jobs per order. Watanapa and Techanitisawad [5] contribute to the literature by proposing a genetic algorithm for sequencing demand backlog in the presence of contingent demand. Piya [6] proposes a method to simultaneously quote prices and due dates to new customers, including negotiation information. Mallach et al. [7] investigate lead time quotation with contingency by providing a simulation-based approach to quoting orders based on the earliest due dates.

While the existing literature focuses on offline quotation under contingency, this study proposes an online approach to quoting lead times by adopting the primal-dual algorithm developed by Buchbinder and Naor [8]. We define the acceptance-rejection decision for contingent orders as delayed feedback. Zhalechian et al. [9] study online resource allocation problems that share some aspects of our problem. Our focus, however, is on online lead time quotation with contingent orders, and finalizing production schedules for each period at the beginning of the period.

## 3. Problem Description

Consider an MTO firm whose capacity is limited by a single bottleneck resource. Customers arrive randomly over time and place a single order characterized by the order size $v_i$, the tardiness cost per unit $c_i$, and associated processing capacity $Q_i$, where $i \in I$ and $I = \{1, 2, \ldots, |I|\}$. The firm-customer interaction is as follows. When customer order $i \in I$ arrives, the firm quotes a lead time $l$, from a menu $L = \{1, 2 \ldots, |L|\}$. The customer then has up to $n$ periods (e.g. weeks) to decide whether or not to accept the offered services with the quoted lead time, while the firm is liable to deliver on time and faces a penalty $c_i$ per unit size $v_i$ for each period of delay. The customer acceptance probability, $A(l, v_i)$, is modeled using the logistical response function [10]:

$$A(l, v_i) = \left(1 + \beta_0 \exp\left(\beta_1 \frac{l - v_i + 1}{v_i + 1} + \beta_2 l\right)\right)^{-1} \tag{1}$$

$\beta_0, \beta_1, \beta_2$ are bidding parameters capturing the convex-concave relationship between the quoted lead time $l$ and the acceptance probability, with customers having a higher probability of accepting shorter lead times. Customers provide this acceptance/ rejection response after $k$ periods, $k = 1, 2, \ldots, n$, with unknown probability $u_k$.

Production is scheduled in discrete periods. Consider orders arriving in the period set $J = \{0, 1, \ldots, |J|\}$ and being completed in the period set $T = \{1, 2, \ldots, |T|\}$ such that $|J| \leq |T|$. Order arrival and quotation are the only activities in period 0 because there are no confirmed orders yet to schedule. At the beginning of each period $j \in 1, 2, \ldots, |T|$, the manufacturer observes the current outstanding set of confirmed orders $I^f$ and determines a production schedule to complete all confirmed orders within the planning horizon. For this purpose, the manufacturer runs a schedule optimization and uses a positive tolerance $\varepsilon \ll 1$ to prioritize order scheduling in early periods whenever ties exist, so as to increase capacity utilization of earlier periods and protect future capacity for upcoming orders.

**Production scheduling model for period $j$, $j \in 1, 2, \ldots, |T|$:**

$$\max \quad \sum_{i \in I^f} \sum_{t \in T : t \geq j} \left(R_{i,t} - \varepsilon\,(t - j)\right) \cdot s_{i,t} \tag{2}$$

$$\sum_{t \in T : t \geq j} s_{i,t} = 1 \quad \forall i \in I^f \tag{3}$$

$$\sum_{i \in I^f} Q_i \cdot s_{i,t} \leq C_t \quad \forall t \in T : t \geq j \tag{4}$$

$$s_{i,t} \in \{0, 1\} \quad \forall i \in I^f, \forall t \in T : t \geq j$$

In the planning horizon under consideration, $R_{i,t} = v_i \cdot (p - c_i \cdot max(t - (j + l), 0))$ is the profit from confirmed customer $i \in I^f$, who was quoted a lead time $l$, responded as a confirmed order in period $j$, and is assigned a completion period $t$; $s_{i,t}$ is the binary decision variable of scheduling confirmed order $i \in I^f$ to a completion period $t$ in the planning horizon. The model's objective function maximizes the realized profit from orders confirmed by period $j$, prioritizing earlier completion periods. Constraint (3) ensures each confirmed customer order $i \in I^f$ must be assigned a completion period $t \in T, t \geq j$. Constraint (4) ensures the processing capacities of confirmed orders assigned to a period $t$ do not exceed its available capacity, $C_t$. The output of this model provides the firm production schedule for the current period $j$ and remaining capacity $\tilde{C}_t = C_t - \sum_{i \in I^f} Q_i \cdot s_{i,t}$ for periods $t \in T : t > j$. Table 1 summarizes the notations used.

Table 1: Notation for online quotation and offline production scheduling formulation

| **Notation** | **Description** |
|---|---|
| **Sets** | |
| $I$ | Set of incoming (initially contingent) customer orders, $I = \{1, 2, \ldots, \lvert I\rvert\}$ |
| $I^f$ | Set of confirmed customer orders at the beginning of a period, $I^f \subseteq I$ |
| $I^j$ | Set of incoming customer orders in period $j \in J$, $I^j \subseteq I$ |
| $J$ | Set of arrival periods in planning horizon, where $J = \{0, 1, \ldots, \lvert J\rvert\}$ |
| $L$ | Set of lead times, $L = \{1, 2, \ldots, \lvert L\rvert\}$ |
| $T$ | Set of completion periods in planning horizon, $T = \{1, 2, \ldots, \lvert T\rvert\}$, such that $\lvert J\rvert \leq \lvert T\rvert$ |
| **Parameters** | |
| $R_{i,t}$ | Profit of confirmed order $i \in I^f$ if completed in period $t$ |
| $r_{i,l,t}$ | Expected profit of incoming order $i \in I$ if quoted a lead time $l$ and assigned a completion period $t$ |
| $v_i, Q_i$ | Size and processing capacity of customer $i$'s order, respectively |
| $q_{i,l}$ | Expected processing capacity of customer $i$'s order if quoted a lead time $l$ |
| $A(l, v_i)$ | Probability of customer $i$ accepting lead time $l$, given order size $v_i$ |
| $p$ | price rate per unit size |
| $\lambda, c_i$ | Mean arrival of customers per period and unit tardiness of customer $i$'s order, respectively |
| $C_t, \tilde{C}_t$ | Capacity and remaining capacity of period $t$, respectively |
| $\beta_0, \beta_1, \beta_2$ | Bidding parameters to model the customer response function |
| $n, \varepsilon$ | Maximum allowed delay for customer response; i.e grace period, and a positive tolerance, $\varepsilon \ll 1$ |
| $u_k$ | Unknown probability associated with customer response delay of $k$ periods, $k = 1, 2, \ldots, n$. |
| **Variables** | |
| $y_{i,l,t}$ | 1 if customer order $i \in I$ is quoted a lead time $l$ and assigned completion period $t$, 0 otherwise |
| $s_{i,t}$ | 1 if customer $i \in I^f$ is assigned a completion period $t$, 0 otherwise |
| $x_t, z_i$ | dual capacity variable for completion period $t$ and dual assignment variable for customer $i \in I$ |

## 4. Online Lead Time Quotation for Contingent Orders in Period $j$: Primal-Dual Approach

In quoting the lead times for contingent orders that arrive during the current period $j \in J$, the manufacturer aims to maximize their expected profits for the remaining periods in the planning horizon. The expected profit from customer order $i \in I^j$ if quoted a lead time $l$ and assigned a completion period $t$ is $r_{i,l,t} = A(l, v_i) \cdot R(i,t)$ and the expected capacity used is $q_{i,l} = A(l, v_i) \cdot Q_i$. If the manufacturer knew all future customer orders arriving in period $j \in J$ and the order characteristics, the optimal lead time quotes could be derived from the offline integer program (IP):

**Offline Integer Program (IP)**

$$\max \quad \sum_{i \in I^j} \sum_{l \in L} \sum_{t \in T, t > j} r_{i,l,t} \cdot y_{i,l,t} \tag{5}$$

$$\sum_{l \in L} \sum_{t \in T, t > j} y_{i,l,t} \leq 1 \quad \forall i \in I^j \tag{6}$$

$$\sum_{l \in L} \sum_{i \in I^j} q_{i,l} \cdot y_{i,l,t} \leq \tilde{C}_t \quad \forall t \in T, t > j \tag{7}$$

$$y_{i,l,t} \in \{0, 1\}, \quad \forall i \in I^j, \forall t > j; t \in T, \forall l \in L$$

Note that these new contingent orders cannot be scheduled in the current period $j$ itself, since customer decisions are not yet known. The production schedule for period $j$ was set at the beginning of the period.

Since this offline solution requires complete information about all customers, we use a primal-dual approach [8] to develop an online algorithm for quoting lead times and assigning expected completion periods to each arriving customer. By relaxing the integrality constraint, the dual of the corresponding linear program for period $j \in J$ is:
**Corresponding Dual Linear Program (LP)**

$$\min \quad \sum_{i \in I^j} z_i + \sum_{t \in T, t > j} \tilde{C}_t \cdot x_t \tag{8}$$

$$\text{s.t.} \quad z_i + q_{i,l} \cdot x_t \geq r_{i,l,t} \quad \forall i \in I^j, \forall t \in T, t > j, \forall l \in L \tag{9}$$

$$z_i, x_t \geq 0 \quad \forall i \in I^j, \forall t \in T, t > j$$

In this dual formulation, $z_i$ and $x_t$ are the dual allocation and capacity variables, respectively. We leverage the dependence of Constraint (9) on each contingent customer $i$ order, as this information will be critical to the online quotation algorithm. The MTO manufacturer can quote contingent orders in period $j$ following Algorithm 1.

**Algorithm 1** Online Primal-Dual Lead Time Quotation in Period $j \in J$

```
 1: Initialise $C_t = \tilde{C}_t \quad \forall t \in T, t > j$
 2: Set $x_t = 0 \quad \forall t \in T, t > j$
 3: for each order $i \in I^j$ that arrives in period $j$ do
 4:     Set $z_i = 0$
 5:     Observe $v_i$, $c_i$ and calculate $q_{i,l}\ \forall l \in L$
 6:     Set $(l^\star, t^\star) \leftarrow \underset{(l \in L, t \in T: t > j)}{\arg\max} \ (r_{i,l,t} - q_{i,l} x_t)$
 7:     if $(r_{i,l^\star,t^\star} - q_{i,l^\star} x_{t^\star}) \geq 0$ then
 8:         Set $y_{i,l^\star,t^\star} \leftarrow 1$
 9:         Set $z_i \leftarrow (r_{i,l^\star,t^\star} - q_{i,l^\star} x_{t^\star})$
10:         Set $x_{t^\star} \leftarrow x_{t^\star}(1 + \frac{q_{i,l^\star}}{C_t^\star}) + \beta(\frac{r_{i,l^\star,t^\star}}{C_t^\star})$
11:     else:
12:         Set $y_{i,l,t} \leftarrow 0 \quad \forall t \in T : t > j, \forall l \in L$
13:     end if
14: end for
```

In Algorithm 1, Step 1 initializes the period capacities, $C_t$, to the remaining capacity, $\tilde{C}_t$, based on the output of the production scheduling model at the beginning of period $j \in J$. We initialize the dual variable for capacity $x_t$ to 0 in Step 2 as no quotation has been generated yet. For each new order $i$ arriving in period $j$, the dual variable for allocation $z_i$ is initialized to 0 as the manufacturer has yet to accept or reject the order for processing. The manufacturer observes the customer's order characteristics and searches for a lead time $l$ and expected completion period $t$ that maximizes the expected profitability evaluation in Step 6. The evaluation in Step 7 confirms whether the optimal combination $(l^\star, t^\star)$ results in a feasible dual solution. Intuitively, this evaluation also determines whether the expected profit from processing the order, $r_{i,l^\star,t^\star}$, is at least the expected cost of assigning available capacity, $q_{i,l^\star} x_{t^\star}$. If the evaluation is at least 0, Step 8 assigns the order $i$ a lead time $l^\star$ and expected completion period $t^\star$. Steps 9 and 10 incrementally update the dual variables to ensure the manufacturer rejects orders when the system is full. We set $\beta = \frac{\eta_{max}}{\Delta - 1}$, where $\eta_{max} = \max_{i,l,t} \left( \frac{r_{i,l,t}}{q_{i,l}} \right), \delta = \max_{i,l,t} \left( \frac{q_{i,l}}{C_t} \right), \Delta = (1 + \delta)^{\frac{1}{\delta}}$. Using these calculations for $\eta_{max}$, $\delta$, and $\Delta$ results in $\beta \to \frac{1}{(e-1)}$ as $\delta \to 0$ and $\eta_{max} \to 1$. This ensures only the online quotation phase for each period $j \in J$ converges to the competitive ratio of $(1 - \frac{1}{e})$, which is the classical performance guarantee in the primal-dual paradigm [8].

## 5. Results and Discussion

We test the performance of the proposed approach against an all-knowing offline (IP) oracle and current practice (naïve approaches) over 100 iterations of a simulated online order quotation and production scheduling environment. Each iteration is a complete planning horizon. The production scheduling models are solved using Gurobi on a laptop

https://doi.org/10.21872/2024IISE_6090

running macOS Sonoma operating system with an M1 processor, and the total execution of the online and offline phases of all algorithms for 100 simulations was recorded at 663.15 seconds. The algorithms under comparison differ solely in their online approach to quoting orders; they uniformly apply the offline production scheduling model each week with the resulting lead times. The benchmark algorithms considered are:

1. FCFS: Assigns lead times to new requests as the completion period is calculated using a first-come-first-served sequence of the confirmed orders received, ignoring contingent orders.
2. FCFS using expected capacity availability: Incorporates contingent orders by assigning lead times to new requests as the completion period calculated using a first-come-first-served sequence of both confirmed and contingent orders received, but reserving only expected capacity ($q_{i,l}$) for outstanding contingent orders.
3. All-knowing offline oracle: The oracle solution determines lead times and order scheduling to maximize the profit from all orders over all periods of the planning horizon utilizing complete information, including what lead times a particular customer will accept.

Table 2 summarizes the description and values for the parameters used in experiments. Note that they represent a congested environment, and each order's size determines the processing capacity, i.e., the units of capacity required $Q_i = v_i$; the unit price and delay penalty are set for long delays to result in negative profits (loss). Arrival periods, completion periods, and lead times are in weeks.

Table 2: Sets and Parameters for Experiments

| Description | Notation | Value |
|---|---|---|
| Number of arrival periods, completion periods and lead times | $\|J\|, \|T\|, \|L\|$ | $50, 70, 10$ |
| Capacity for each period t | $C_t$ | $\{40, 45\}$ |
| Size, unit tardiness | $v_i, c_i$ | $U[1, 10]$, respectively |
| Epsilon | $\varepsilon$ | 0.001 |
| Bidding parameters | $\beta_0, \beta_1, \beta_2$ | 0.1, 10, 1 respectively |
| Mean customer arrival per week, price rate per unit size, grace period | $\lambda, p, n$ | 12 orders, \$10, 3 periods respectively |

For each arriving customer $i$, with associated size $v_i$, we simulate acceptance-rejection decisions for each lead time $l$ based on the acceptance probability $A(l, v_i)$ given in 1. Customers have a three-week grace period post-arrival to accept or reject the quotation. Their response can occur 1, 2, or 3 weeks after the order quotation with equal probability.

Figures 1 and 2 show the total profits of each algorithm, given the weekly capacity of 40 and 45 units, where each iteration is a complete planning horizon.

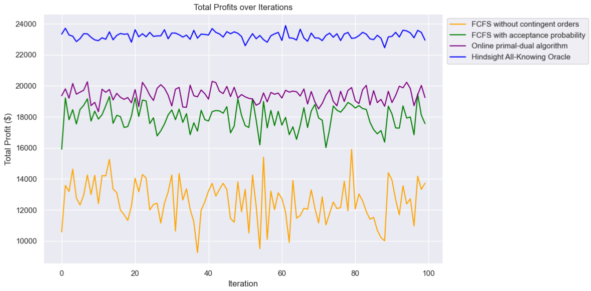

Figure 1: Total profits over 100 simulated planning horizons with 40 units of weekly capacity

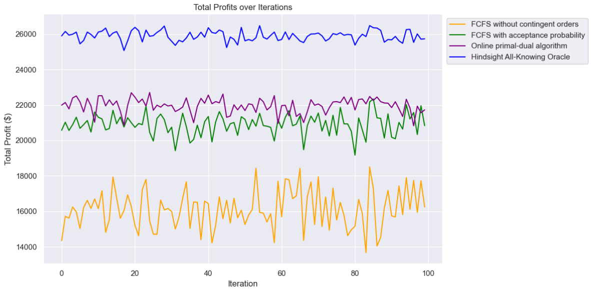

Figure 2: Total profits over 100 simulated planning horizons with 45 units of weekly capacity

To summarize the comparison results, Table 3 presents the *profit performance*, defined as the ratio of the total profit by an algorithm to the total profit of the all-knowing offline oracle over the 100 simulated 50-week environments.

The performance of all algorithms relative to the oracle improves with additional capacity (going from 40 to 45 weekly units). The proposed primal-dual algorithm achieves 6.29% (4.10%) and 29.41% (22.59 %) performance improvements relative to FCFS using expected capacity available and FCFS, respectively, with 40 (45) weekly units

Table 3: Comparison of Algorithm Performance

| | Profit performance relative to Oracle (%) | |
|---|---|---|
| Algorithm | Weekly capacity: 40 units | Weekly capacity: 45 units |
| Online Primal-Dual | 0.8359 | 0.8482 |
| FCFS with Acceptance Probability | 0.7730 | 0.8072 |
| FCFS | 0.5418 | 0.6223 |

of capacity. This performance improvement can be attributed to the fact that the primal-dual algorithm jointly considers each order's expected profit, expected processing capacity, and the firm's available capacity given contingent orders when quoting the lead time, in contrast with the naïve algorithms.

## 6. Conclusion

This study highlights the need to develop new methods for online lead time quotations under contingency in environments with delayed order acceptance. The proposed primal-dual approach results in 6.29% (4.10%) and 29.41% (22.59 %) percentage profit increase in an environment of 40 (45) capacity units compared to the naïve approaches identified in our survey of small- and medium-sized precision manufacturing firms. More research needs to be performed to develop robust methodologies for order quotations in complex scenarios, considering multiple resources, available inventory, and delay-adaptability.

## Acknowledgements

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 2223752. Any opinions, findings, and conclusions expressed are those of the authors and do not necessarily reflect the views of the NSF We are also indebted to Dr. Ted Acworth of Artaic Innovative Mosaic for introducing us to this problem and Prof. David Simchi-Levi for pointing out the effectiveness of primal-dual approaches for online decision-making.

## References

[1] Guowei Hua, Shouyang Wang, and TC Edwin Cheng. "Price and lead time decisions in dual-channel supply chains". In: *European journal of operational research* 205.1 (2010), pp. 113–126.

[2] Yue Zhai and TCE Cheng. "Lead-time quotation and hedging coordination in make-to-order supply chain". In: *European Journal of Operational Research* 300.2 (2022), pp. 449–460.

[3] Fred F Easton and Douglas R Moodie. "Pricing and lead time decisions for make-to-order firms with contingent orders". In: *European Journal of operational research* 116.2 (1999), pp. 305–318.

[4] Andi Cakravastia and Katsuhiko Takahashi. "Integrated bidding and manufacturing planning decisions with contingent orders in a make-to-order environment". In: *Journal of Japan Industrial Management Association* 54.5 (2003), pp. 291–301.

[5] Bunthit Watanapa and Anulark Techanitisawad. "A genetic algorithm for the multi-class contingent bidding model". In: *OR Spectrum* 27 (2005), pp. 525–549.

[6] Sujan Piya. "Dealing with customers enquiries simultaneously under contingent situation". In: *International Journal of Industrial Engineering Computations* 6.3 (2015), pp. 391–404.

[7] Ron Mallach, Ana Muriel, and Ted Acworth. "Lead Time Quotation with Contingent Orders". In: *IIE Annual Conference. Proceedings*. Institute of Industrial and Systems Engineers (IISE). 2019, pp. 1638–1644.

[8] Niv Buchbinder, Joseph Seffi Naor, et al. "The design of competitive online algorithms via a primal–dual approach". In: *Foundations and Trends® in Theoretical Computer Science* 3.2–3 (2009), pp. 93–263.

[9] Mohammad Zhalechian et al. "Online resource allocation with personalized learning". In: *Operations Research* 70.4 (2022), pp. 2138–2161.

[10] Joseph Berkson. "A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function". In: *Journal of the American Statistical Association* 48.263 (1953), pp. 565–599.