Lead Time Quotation under Time-Varying Demand and Capacity

Thanh-Ha Nguyen
Lancaster University
Management School
Lancaster, LA1 4YX, UK
e-mail: n.thanh-ha@lancaster.ac.uk

Mike Wright
Lancaster University
Management School
Lancaster, LA1 4YX, UK
e-mail: n.thanh-ha@lancaster.ac.uk

***Abstract* — In this paper, we consider a telecommunication service company with time-varying demand and capacity. The firm applies a uniform lead-time policy, i.e. a lead time which corresponds to the maximum time span a customer has to wait before receiving the required service is uniformly quoted to all customers. The objective is to schedule service jobs such that the shortest possible lead-time can be quoted to customers. We present an easy-to-compute approach to tackle this problem.**

Keywords: service capacity management; service level agreement; standard lead time; telecommunications strategic review

I. PROBLEM DESCRIPTION

We consider a major telecommunication service company. The company uses an integrated planning system, based on hierarchical planning concepts that allow to decompose the entire planning problem into partial planning tasks while still considering their interdependencies and coordinating their solutions. This planning system consists of different modules, such as demand forecasting, resource planning and work scheduling, which are interlinked. It makes use of solution approaches known as mathematical programming and meta-heuristics and provides support at different levels for planning tasks along the company's service chain, from long-term strategic decision making to short-term operational decisions. The levels of planning may overlap or may be distinct. Either way, there is a flow of information from strategic to tactical planning and then to operational planning and vice versa.

Demand The firm faces seasonal demand for a particular service, e.g. broadband installation. The estimated demand data are provided by the responsible module for forecasting demand. Thus, the start and the end point of the seasonal cycle are known. Further, the seasonal demand pattern, which repeats itself every cycle, is also given. By dividing the seasonal cycle into time periods 1,…,M, the demand pattern can be expressed by a vector $[\lambda_1,\ldots,\lambda_M]^T$. Each element of this vector represents the demand (measured by the number of jobs) that occurs throughout a particular time period.

Capacity The firm has a fixed number of permanent employees and a number of seasonal technicians with repeated fixed term contracts. The latter are retained in order to meet peaks in demand (e.g. surge of demand for broadband installations at the beginning of school terms). The information concerning availability of the workforce per time period is provided by means of the medium-term, anticipatory deployment plan. As it is possible to estimate the average time a technician needs to complete a job, we represent capacity during a time period in terms of the number of jobs to better match it with customer demand. Capacity levels are assumed to follow a cycle of N time periods with the pattern $[c_1,\ldots, c_N]^T$.

Planning horizon The planning horizon τ is the minimum time interval after which both demand and capacity pattern will repeat themselves. Thus, τ is determined as the least common multiple of M and N, τ = lcm(M,N). Throughout this paper, we use the notations i, j $\in \{1,\ldots,\tau\}$ to denote, respectively, arrival periods and completion periods. Note that all computations with i and j are performed modulo τ. For ease of notation, we suppress the notation mod τ when referring to time periods. For example, we write i + 1 instead of (i + 1) mod τ. The demand and the available capacity during the planning horizon are represented by the vectors $\lambda = [\lambda_i]$ and $c = [c_j]$. The vector λ is obtained by τ/M-times concatenation of $[\lambda_1,\ldots,\lambda_M]^T$, and the vector c by τ/N-times concatenation of $[c_1, \ldots, c_N]T$.

Lead-time A uniform lead-time L, which corresponds to the maximum time span a customer has to wait before receiving the required service, is quoted to all customers. In practice, to prevent the firm from breaking promises to customers, lead-time is usually computed based on empirical data as the longest period of time needed for completing a service and is offered as part of a service level agreement to customer [4]. A manually and inaccurately calculated lead-time may lead to inefficient resource utilization due to the mismatch between customer demand and the firm's capacity. Let ℓ denote the minimum waiting time before which a job cannot be processed; thus L is bounded by $1 + \ell$ and $\tau + \ell$. This means $\ell < L \leq \tau + \ell$

The central questions addressed in this paper are:

1. Given demand and capacity, what is the shortest possible lead-time to quote, which is not too long so as not to frustrate customers, and not too short so that the firm does not end up handling customer demand inefficiently?
2. How to find a job scheduling scheme which ensures that all demands are met within the quoted lead-time?

Such a scheduling scheme is referred to as an optimal solution in this paper.

The remainder of this paper is structured as follows: a literature review is presented in Section 2. In Section 3, structural properties of a class of optimal solutions are derived, based on which an easy-to-compute approach for finding the shortest possible lead-time is presented in Section 4. Finally, conclusions are given in Section 5.

II. LITERATURE REVIEW

A number of studies are related to ours in some ways, but with different emphases. For lead-time quotation, research has developed in multiple directions. Most of the research deals with quoting lead-time in real-time for each customer. See [16, 2] for examples in manufacturing environments, where problems about lead-time quotation and production planning decisions are investigated. See also [15] and references therein for studies on service systems (e.g. telephone call centers) that make lead-time predictions in order to improve customer waiting experience.

There have been several research efforts studying the use of uniform quoted lead-times in service firms, such as [14, 8, 13], and [12]. These papers are concerned with the optimal selection of capacity level, lead-time and/or price to maximize overall profit. Here, the decision problem is modeled as a M/M/1 queue and a linear or log-linear relationship between demand, lead-time or price is assumed. However, time variation of neither demand nor capacity is considered. Our decision problem falls within the context of an integrated planning system consisting of modules for supporting diverse planning tasks. While such systems for the service industry are still in a nascent stage, they are widely used in manufacturing environments under the term Advanced Planning Systems (APS). The class of planning problem that is supported by APS, and is somewhat related to our topic, is called Demand Fulfillment (DF). The prime task of DF is order promising, i.e. determining whether to accept a given customer request and quoting the delivery quantity and lead-time. The basis for these decisions are the so-called available-to-promise (ATP) quantities. The quantity of ATP at a given point in time is given by the forecast-driven medium-term master planning, and is computed as inventory on hand plus scheduled replenishments which have not yet been committed to customers. Models and solution methods have been developed for DF problems in different manufacturing environments, such as make-to-order [1, 9], assembly-to-order [17, 5], make-to-stock [6, 10, 11]. However, none of these papers are concerned with determining an optimal uniform lead-time.

The problem under consideration was first investigated by [3], who solved it in two steps:

1. First, in order to improve capacity utilization, the workload must be balanced across time periods. For a given lead-time, they proposed an approach to solve a combinatorial problem of scheduling jobs to match the firm's capacity to customer demand in such a way that: a) the quoted uniform lead-time is satisfied and b) the capacity utilization rate per period is distributed as equally as possible along the time line. We refer to this problem as the Workload Balancing Problem (WBP).
2. Second, they find the shortest possible lead-time L^* in an iterative manner. The search is limited to the range between $1+ \ell$ and $\tau + \ell$. In each iteration the WBP is solved, given the value of L. The search stops when a schedule is obtained where the optimized capacity utilization rate in each time period is less than or equal 100%.

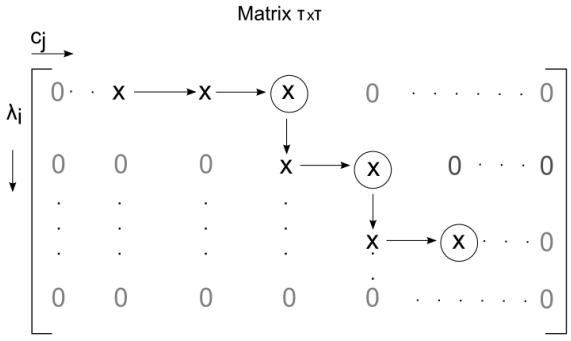
In this paper, we show that the lead-time quotation problem can be solved independently from the WBP and we propose an easy-to-compute approach to tackle this problem. New efficient solution algorithms for the WBP can be found in [7].

III. STRUCTURAL PROPERTIES

Let $X = [x_{ij}]_{\tau\times\tau}$ be a matrix that represents a job assignment scheme, where x_{ij} denotes the number of jobs that arrive in period i and are scheduled to be completed in period j. We now analyze some structural properties of a class of optimal solutions that are useful for solution finding.

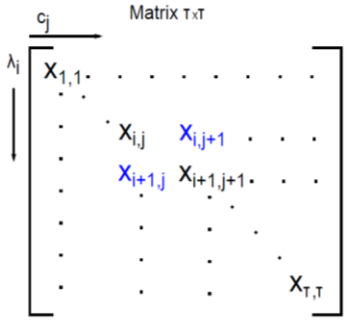
Property 1 There always exists an optimal solution which has "first come first served" (FCFS) structure.

The FCFS property means that the allocation scheme has a pattern like the one depicted below:

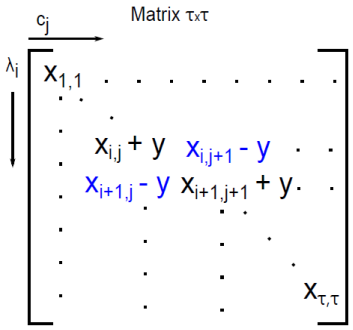


with each x indicating a node on the assignment route, which at every step moves either to the right or down. Further, at the circled position in each row i of the matrix, the value of x is positive and the demand λ_i is completely fulfilled.

Insight: If this property does not hold, then there exist problems for which all optimal solutions have the following structure



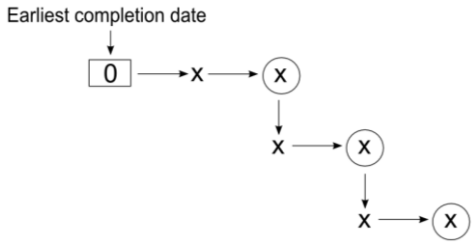
with both $x_{i,j+1}$ and $x_{i+1,j}$ being positive, meaning that the processing of λ_{i+1} begins before the processing of λ_i has finished. Write $y = \min(x_{i+1,j}, x_{i,j+1}) > 0$ and consider the new solution



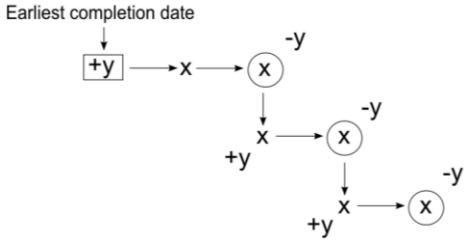
One of the elements, $x_{i,j+1} - y$ or $x_{i+1,j} - y$, is now zero, while the remainder of the solution is untouched. This solution is still feasible with respect to the quoted lead-time, but may have moved the completion of demand λ_i forward in time.

Property 2 In at least one period it must be true that some of its demand is met at the earliest possible point of time.

Insight: Suppose there exists an optimal solution where demand in every period is delayed beyond its earliest point. Fix one such period. We have a solution in the following form



Note that the allocated quantities of jobs at the circled positions are positive. Call the smallest of them y. We can construct a new solution as illustrated below:



The resulting solution still satisfies all demands within the quoted lead-time, but moves all completions forward and now occupies the "earliest completion date".

IV. Shortest Possible Lead-Time

In order to determine the shortest possible lead-time L^*, we distinguish between the following two cases:

- If the total demand during the planning horizon exceeds the total capacity, the firm will not be able to accommodate the total demand within $\tau + \ell$ periods without exceeding the total capacity. In this case, the firm is better off increasing capacity by hiring more labor or outsourcing part of its service activities, but modeling these aspects is out of scope of this paper.
- If the total capacity of the entire planning horizon is sufficient to accommodate the total demand, the shortest possible uniform lead-time L^* to quote is the minimum time by which all jobs have been completed.

In the second case, L^* can be trivially determined by taking advantage of the structural properties of some optimal solutions. It is obvious that there exists a job scheduling scheme with shortest possible lead-time which possesses Properties 1 and 2 and which satisfies demand as much as possible at every step to the right or down. There are τ such schemes; each starts to satisfy a different demand λ_i as much as possible at its earliest completion date. Denote the set of these scheduling schemes as $\Theta = \{\theta_t \mid t \in \{1,\ldots,\tau\}$. The shortest possible lead-time is determined as

$$L^* = \min_{\theta_\tau \in \Theta} \{L(\theta_\tau)\}. \qquad (1)$$

where $L(\theta_t)$ is the standard lead-time to be quoted if the schedule θ_t is applied. Under each schedule θ_t, $L(\theta_t)$ corresponds to the longest time between the arrival of a demand and its completion. We illustrate the solution approach in the following example.

Example 1 We consider a real-life problem scenario presented in [3]. The planning horizon is one week, i.e. $\tau = 7$. The demand pattern is forecasted as

$$\lambda = [\lambda_i] = [178, 191, 106, 136, 55, 2, 2]^T$$

The field force of the company is organized into multiple patches across the country. A patch covered by 8 technicians is considered. Only 6 of them work on Saturday and Sunday. Thus the number of available technicians during a week is presented as

$$n = [n_j] = [8, 8, 8, 8, 8, 6, 6]^T$$

Each technician can complete a job within 30 minutes, i.e. the service rate μ corresponds to 16 jobs per technician per day. Assuming that a technician works 8 hours per day, the weekly available capacity (measured in number of jobs) is given by

$$c = [c_j] = [128, 128, 128, 128, 128, 96, 96]^T$$

Let the minimum time ℓ a job has to wait before being processed be 1 day. The set $\Theta = \{\theta_1,\ldots,\theta_7\}$ then consists of 7 scheduling schemes that need to be examined.

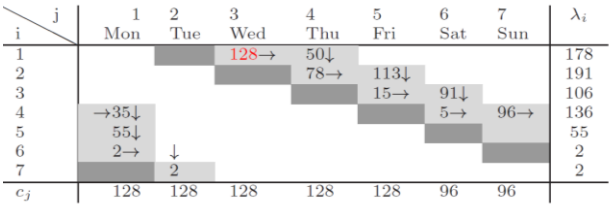


Figure 1. Scheduling scheme θ_1, $L(\theta_1) = 4$.

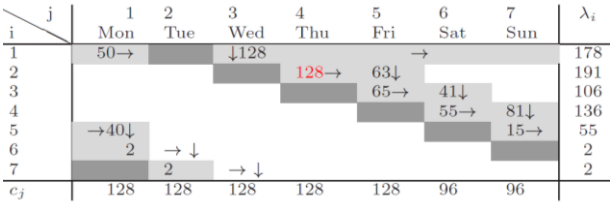


Figure 2. Scheduling scheme θ_2, $L(\theta_2) = 7$

Consider, for instance, scheduling scheme θ_1 (Figure 1). Here, demands are satisfied as much as possible at every step to the right or down, starting with demand λ_1 arriving Monday (day 1) which cannot be completed earlier than Wednesday (day 3) of the same week. Since the capacity available on Wednesday is only sufficient for the completion of 128 jobs, demand λ_1 of 178 jobs can only be completely satisfied on the following Thursday (day 4). For jobs that arrive on Thursday (day 5), the earliest possible completion day is Saturday of the same week. However, a part of the capacity available in this day has already been allocated to jobs that arrived earlier, so that only 5 jobs of demand λ_3 can be fulfilled on Saturday. Thus, the remainder of demand λ_5 is scheduled for the following Sunday (96 jobs) and for Monday of the subsequent week (55 jobs). As a result, demand λ_7 can only be completely fulfilled 4 days after its arrival. This is also the longest waiting time under this scheme. In other words, if scheme θ_1 is applied, the standard lead-time to be quoted will be $L(\theta_1) = 4$ days.

Analogously, the standard lead-time can be obtained for every scheduling scheme $\theta_t \in \Theta$. Figure 2 shows scheduling scheme θ_2, which begins with the starting point from the cell [2,4] where demand λ_2 is satisfied as much as possible at its earliest possible completion date. Under this scheme, the maximum waiting time for a job to be completed is $L(\theta_2) = 7$ days, because demand λ_1 that arises on Monday can only be completed on Monday of the subsequent week. Table 1 shows the shortest possible lead-time that can be quoted to all customers is 4 days, which can be achieved under schemes θ_1, θ_5, θ_6 or θ_7.

TABLE I. COMPARISON OF DIFFERENT SCHEDULING SCHEMES

i	1	2	3	4	5	6	7
$L(\theta_i)$	4	7	6	5	4	4	4

V. CONCLUSION

In this paper, we considered a telecommunication service that faces time-varying demand and capacity and quotes a uniform lead-time to all customers. This study is motivated by previous research by [3] who proposed an iterative approach for determining the shortest possible lead-time based on a combinatorial job scheduling problem. Thus, their solution method is time-consuming. We show that the optimal job schedule with respect to quoted lead-time has a simple, intuitive structure. Based on this result, we are able to derive a simple approach to determine the shortest possible lead-time. It would be interesting for future research to investigate the problem of lead-time quotation under stochastic demand.

REFERENCES

[1] Chen CY, Zhao ZY, Ball MO (2002) A model for batch advanced available-to-promise. Production & Operations Management 11(4):424–440

[2] Duenyas I, Hopp WJ (1995) Quoting customer lead times. Management Science 41(1):43–57

[3] Li Y, He B (2008) Optimizing lead time and resource utilization for service enterprises. Service Oriented Computing and Applications 2(2-3):65–78

[4] Li Y, Voudouris C, Thompson SG, Owusu G, Anim-Ansah G, Liret A, Lee H, Kern M (2006) Self-service reservation in the fieldforce. BT Technology Journal 24(1):40–47

[5] Lin JT, Hong I, Chen T, Wang K (2009) A multi-site available-to-promise model for TFT-LCD manufacturing. In: APIEMS, pp 2790–2798

[6] Meyr H (2009) Customer segmentation, allocation planning and order promising in make-to-stock production. OR Spectrum 31(1):229–256

[7] Nguyen TH, Wright M (20013) Variable neighbor-hood search for the workload balancing problem in service enterprises. Computers & Operations Research DOI 10.1016/j.cor.2013.07.027

[8] Palaka K, Erlebacher S, Kropp DH (1998) Lead-time setting, capacity utilization, and pricing decisions under lead-time dependent demand. IIETransactions 30(2):151–168

[9] Pibernik R, Yadav P (2008) Dynamic capacity reservation and due date quoting in a make-to-order system. Naval Research Logistics 55(7):593–611

[10] Pibernik R, Yadav P (2009) Inventory reservation and real-time order promising in a make-to-stock system. OR Spectrum 31(1):281–307

[11] Quante R, Fleischmann M, Meyr H (2009) A stochastic dynamic programming approach to revenue management in a make-to-stock production System. (No. ERS-2009-015-LIS). Erasmus Research Institute of Management (ERIM)

[12] Ray S, Jewkes E (2004) Customer lead time management when both demand and price are lead time sensitive. European Journal of Operational Research 154(3):769–781

[13] So KC (2000) Price and time competition for service delivery. Manufacturing and Service Operations Management 2(4):392–409

[14] So KC, Song JS (1998) Price, delivery time guarantees and capacity selection. European Journal of Operational Research 111(1):28–49

[15] Stanford DA, Pagurek B,Woodside CM (1983) Optimal prediction of times and queue lengths in the gi/m/1 queue. Operations Research 31:322–337

[16] Weng ZK (1996) Manufacturing lead times, system utilization rates and lead-time-related demand. European Journal of Operational Research 89(2):259–268

[17] Zhao ZY, BallMO, KotakeM (2005) Optimization-based available-to-promise with multi-stage resource availability. Annals of Operations Research135(1):65–85