Periodic capacity management under a lead-time performance constraint

OR Spectrum (2013) 35:221–249
DOI 10.1007/s00291-011-0261-4

REGULAR ARTICLE

Periodic capacity management under a lead-time performance constraint

Nasuh C. Buyukkaramikli · J. Will M. Bertrand · Henny P. G. van Ooijen

Published online: 30 June 2011

Abstract In this paper, we study a production system that operates under a lead-time performance constraint which guarantees the completion of an order before a pre-determined lead-time with a certain probability. The demand arrival times and the service requirements for the orders are random. To reduce the capacity-related operational costs, the production system under study has the option to use flexible capacity. We focus on periodic capacity policies and model the production system as a queuing system that can change its capacity periodically and choose to operate in one of the two levels: a permanent capacity level and a permanent plus contingent capacity level. Contingent capacity is supplied if needed at the start of a period, and is available during that period, at a cost rate that is decreasing in period length in different functional forms. Next, we propose a search algorithm that finds the capacity levels and the switching point that minimizes the capacity-related costs for a given period length. The behaviour of the capacity-related costs changes drastically under different period lengths and cost structures. In our computational study, we observe that the periodic capacity flexibility can reduce the capacity-related operational costs

N. C. Buyukkaramikli (✉) · J. W. M. Bertrand · H. P. G. van Ooijen
School of Industrial Engineering,
Eindhoven University of Technology, P.O. Box 513,
5600 MB Eindhoven, The Netherlands
e-mail: n.c.buyukkaramikli@tue.nl

J. W. M. Bertrand
e-mail: j.w.m.bertrand@tue.nl

H. P. G. van Ooijen
e-mail: h.p.g.v.ooijen@tue.nl

N. C. Buyukkaramikli
EURANDOM, P.O. Box 513,
5600 MB Eindhoven, The Netherlands

significantly (up to 35%). However, in order to achieve these savings, the period length must be chosen carefully depending on ambition level and capacity-related costs. We also observe that the percentage savings are higher for more ambitious lead-time performance constraints. Moreover, we observe that the use of contingent capacity makes the total system costs more insensitive to the lead-time performance requirements.

Keywords Production control · Capacity management · Lead-time management · Queuing theory · Transient analysis

1 Introduction

Customer service, one of the most important keys for success in today's business world, is a broad concept that entails many dimensions of customer satisfaction. Numerous studies in the business literature highlight the importance of fast and reliable deliveries in customer satisfaction (Ballou 1998). Therefore, speed has emerged as a determinant factor of competitive advantage next to the price.

For highly customized products, production is mostly achieved through MTO (make-to-order) or ATO (assemble-to-order) systems with little or no finished goods inventory, where the lead-time arises as a medium of coordination to accomplish fast and reliable deliveries. Lead time is often exploited as a marketing tool to signal a firm's commitment to its customers, resulting in a uniform lead time communicated in the market. Examples of guaranteed uniform lead times in furniture, construction equipment manufacturing and telecommunication field service industries are mentioned by Rao et al. (2005).

In line with these trends, companies are inclined to set targets for uniform, short and reliable lead times. Usually, these targets are considered beforehand as higher level tactical issues which may further affect pricing decisions, market demand, capacity/production planning and cash flows. These targets are then communicated to the shop floor manager, who is expected to meet the targets at the lowest possible operating costs.

In this context, we focus on a specific form of lead-time performance, which guarantees the completion of an order within a pre-determined lead time (e.g. 1 week) with a certain (e.g. 95%) probability. Compared with other targets (e.g. average delivery lead time), this performance target provides more certainty to the customers about the completion time of an order. Therefore, before giving his/her order, the customer can schedule other activities (like the preparation for the use of the product) during the pre-determined lead time more efficiently and with more certainty. In a different setting, the use of this form of performance targets is quite common in call centers, in the shape of service level agreements (see, e.g. Gans et al. 2003).

Demanding markets mostly require ambitious lead-time performance targets. In our study, a more demanding market either dictates a shorter lead time or a better on-time delivery performance. As performance targets get more ambitious, more capacity is needed, which leads to higher operating costs. Under these circumstances, flexible capacity management can play a soothing role for the shop floor manager, who is

stuck between the conflicting objectives of attaining ambitious lead-time performance targets and reducing the operating costs.

Actually, in the presence of demand uncertainty, flexible capacity management can be of high value to hedge against the under-utilization of deployed capacity. Empirical studies show that flexible capacity management policies (e.g. flexible staffing, under/over working hours, outsourcing) are commonly used in the manufacturing as well as service industries (Houseman 2001; Kalleberg et al. 2003).

For various reasons, flexible capacity control practices in real life are often periodic. First, a company's reach to the external capacity pool may be restricted to certain specific times like the start of a day or the start of a week. Second, decisions about working times (e.g. working over/under time) are often taken on a periodic basis, in order to abide to labour regulations and to accomplish the timely communication of these working time decisions to the relevant employees. In addition, periodic flexible capacity policies are compatible with the modus operandi of resource planning software systems, most of which also operate on a periodic basis due to decision-information synchronization issues. For instance, a ground-handling company in Turkey uses a software system that creates weekly work schedules for its employees. Within the software, weekly workforce size (in terms of total working hours) can be adjusted based on the flight traffic. Similarly, a German car manufacturer and a Dutch lighting equipment manufacturer make use of periodic capacity control techniques such as hiring of temporary workers, implementing variable working hours, employing multifunctional employees and shifting work internally in order to deliver the customer orders on time. These capacity control actions are taken periodically, in the most ambitious case, on a daily basis.[1]

Motivated by these observations, in this paper we analyse the periodic flexible capacity control problem for a single production system in a MTO environment, which operates under a fixed lead time and a delivery performance target. Customer orders arrive according to a stationary Poisson process and each order requires a random processing time. We assume that the processing time is inversely proportional to the total capacity of the system. For the sake of convenience and practicality, we assume two levels of capacity: permanent and permanent plus contingent capacity levels.

At the start of each period, the shop floor manager decides whether to deploy the contingent capacity for that period or not. The contingent capacity is available and ready to be deployed at the start of each period before the manager's decision. This decision is based on the workload in the system at the decision instant. Uncertainty on the use of the reserved contingent capacity in a period creates an opportunity cost, which makes per time unit cost for contingent capacity always larger than that of the permanent capacity. This opportunity cost decreases with period length, because a longer period length implies an improved job security for contingent capacity resources. The reflection of improved job security/working conditions on wages is a well-studied topic in labour economics (see, e.g. Ehrenberg and Smith 1994).

For a given permanent/contingent capacity cost structure, the shop floor manager tries to minimize capacity-related operating costs while satisfying the communicated

[1] The company names are not mentioned for confidentiality.

lead-time performance constraint, which is an indicator of how demanding the market is. Decision variables are the period length, the size of the permanent capacity, the size of the contingent capacity and the workloads at which the contingent capacity is deployed.

In order to analyse this problem, we develop a queuing model, where the periodic capacity policy is reflected in the change of the service rate. At the start of each period, the service rate of the queue is set to either high or low level based on the number of orders in the system. The low service rate corresponds to the permanent capacity level whereas the difference between the high and low service rates corresponds to the contingent capacity level. Per time unit cost for permanent capacity is fixed and per time unit cost for contingent capacity is the sum of permanent capacity cost and the opportunity costs per time. We provide several functional forms for the opportunity cost per time; all are decreasing with the period length and depending on two additional parameters: the maximum value of the opportunity and the decrease of the opportunity cost with period length. Subsequently, we use a computational approach to assess the performance of two-level, threshold type, periodic capacity policies under various decision frequencies (period lengths) and permanent/contingent capacity cost structures for three markets with different demands on lead-time performance. The results show under which conditions substantial savings can be obtained and highlight the importance of the decision on the period lengths for the cost performance of the shop floor. Finally, we assume the case in which setting up and running a flexible capacity system comes at a cost. In such a case, we investigate the cost circumstances, where the deployment of a flexible capacity policy is more beneficial compared to the best fixed capacity system.

The remainder of the paper is organized as follows: In Sect. 2, we present an overview of the relevant literature that paved our way to this study. Section 3 discusses the specifications of the model, the cost structure of the permanent and contingent capacity costs and the formulation of the problem. In Sect. 4, we provide the analysis of the production system under a two-level periodic capacity policy for a given period length. Subsequently, we present and discuss some computational results in Sect. 5. Finally in Sect. 6, concluding remarks and a discussion on future research are presented.

2 Literature review

Decisions on capacity investment are first studied in Economics/Econometrics literature as capacity investment problems. (see, e.g. Chenery 1952; Eberly and van Mieghem 1997).

Holt et al. (1960) were the first to address the problem of the coordination of production and capacity decisions, and they develop the aggregate planning model, in which the production, inventory and workforce decisions (such as hiring/firing and over time/under time working hours) are taken for a finite horizon based on forecasted demand over that horizon.

Pinker (1996), Milner and Pinker (2001) and Pinker and Larson (2003) develop models with different types of flexible capacity arrangements (such as contingent labour contracting or overtime working hours) in the presence of demand/supply

uncertainty over a finite discrete-time horizon. In these studies, different stochastic dynamic programming models are presented in order to obtain the optimal decisions on the capacity levels.

Later studies extend the problem to integrated capacity and inventory control. Bradley and Glynn (2002) provided a Brownian motion approximation to study the joint optimal control of the inventory and the capacity in a make-to-stock system with a subcontracting option. Similarly, Tan and Alp (2009) use stochastic dynamic programming formulations for the integrated capacity and inventory management problem of a make-to-stock system.

Tan (2004) and Tan and Gershwin (2004) provide a modelling framework for the production and subcontracting control problem with limited capacity and volatile demand environment. They analyse their models as stochastic flow rate control problems. Different factors such as the availability guarantee of the subcontractor or the backlog-dependent demand structures are incorporated to their models, as well.

If a production system is modelled as a queuing system, the service rate of the queue can be interpreted as the capacity level. Mostly, stochastic dynamic programming formulations are utilized in order to determine the optimal service rates of the queuing systems with the help of the uniformization technique (Lippman 1975). Sennott (1999) provides a comprehensive overview of the usage of stochastic dynamic programming in queuing systems for different control aspects.

Due date management is a very broad research area with sheer number of studies and many dimensions (see, e.g. Keskinocak and Tayur 2004). Up to now, due-date performance metrics are largely neglected in dynamic queuing control literature. There are only a few studies that incorporate average lead-time performance metrics into the capacity control problem (see, e.g. Mincsovics and Dellaert 2009). In order to calculate other due-date performance metrics than the average performance, the sojourn time distribution of an order is needed, and it is quite difficult to reflect the effects of the capacity control mechanism on the actual sojourn time distribution.

It is still an open problem to determine the structure of an optimal periodic capacity control policy with due-date performance metrics. In this paper, we do not tackle this problem but rather follow a prescriptive approach for the periodic capacity policy structure. We are not aware of any previous work on the use of periodic capacity management in a MTO system in the presence of congestion effects and lead-time performance targets. The main contributions of this paper can be listed as follows:

a. Different from many other studies; we focus on periodic capacity control, which requires the modelling of a queuing system that changes its service rate periodically according to a workload threshold policy. Therefore, period length arises both as a decision variable and as a dimension of a system's flexibility measure.
b. We reflect the effects of the period length on per time unit contingent capacity costs due to the improved job security in line with the theory of compensating differentials (Rosen 1986).
c. Rather than a performance target on the average lead time, we study a MTO system that operates with a fixed lead time and a delivery performance target, which gives more certainty about the order delivery time to the customers before they give their order.

In the next section, we present the modelling approach of our system and formulate the problem that the shop floor manager is facing.

3 A periodic two-level service rate control policy

The production system under study is modelled as a single-server queue which faces a stationary Poisson demand. The arrival rate at the system is equal to λ, and each order admitted requires an exponentially distributed amount of service time. The production system operates under a lead time L and an on-time delivery target γ. This requires that each order should be completed within L units of time with a probability of γ. At the start of each period of length T, the number of orders in the system is observed and according to this number, the service rate is adjusted based on a certain policy.

Simple capacity policies, such as two-level policies, are highly valued by the practitioners because of their operational simplicity. In a two-level capacity policy, permanent capacity is the capacity that is always employed in the system and contingent capacity is either temporarily outsourced from an external supplier or achieved by the efforts of in-house capacity (e.g. working over-hours). Both of these measures are modelled in this paper as a change in the service rate of a single server queue.

Motivated by the optimality results in queuing control literature (i.e. threshold-type policies are optimal in many problems where there is no capacity switching costs, Crabill 1972), we focus on two-level threshold type periodic capacity policies in this paper. However, the analysis provided in this section can be extended to any type of workload-dependent periodic capacity policy.

A two-level, threshold type capacity policy $\pi(k, \mu_l, \mu_h)$ consists of a switching point k, which is a positive integer, and a low and high service rate pair (μ_l, μ_h). In such a policy, μ_l can be interpreted as the permanent capacity level and the $\mu_h - \mu_l$ can be interpreted as the contingent capacity level. For stability, we assume that $\mu_h > \lambda$.

At the start of each period, the number of orders in the system is observed. The contingent capacity is deployed for a period if the number of orders in the system is greater than or equal to the switching point k at the start of that period. Figure 1 illustrates the system under study.

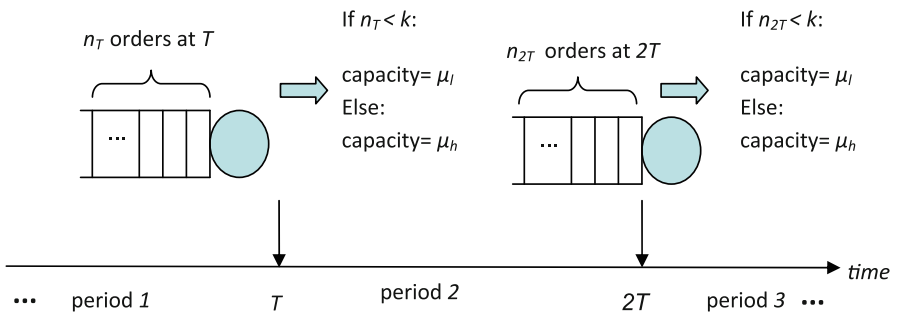

Fig. 1 Illustration of the system under study

3.1 Modelling the capacity-related costs

In order to supply the required amount of capacity for each period, the contingent capacity supplier has to be prepared at the start of each period before the decision is taken. Although contingent capacity is ready to be deployed at the start of each period, it is not guaranteed whether it will be used, since its use is dependent on the number of orders in the system at the start of each period, which cannot be known in advance with certainty. The uncertainty on the use of the contingent capacity creates an economic factor that causes an opportunity cost, because that capacity could be used somewhere else if it was not reserved for that period.

A longer period length T mitigates the severity of the lost opportunity effects due to the contingent capacity availability at the start of each period because it gives more room to the capacity supplier to benefit from the possibility of re-assigning the contingent capacity for other tasks until the start of the next period. Also, a longer T implies an improved job security for the contingent capacity, which would decrease per time unit contingent capacity costs. These effects are in line with the wage differential theory, a research area in Labour Economics that analyses the relations between the wage rate and the unpleasantness, risk or other undesirable attributes of a particular job (Rosen 1986).

Suppose c_p denotes the usage cost per unit time for the permanent capacity, c_c denotes the usage cost per unit time for the contingent capacity and o_c denotes the opportunity cost per unit time due to the reservation of the contingent capacity in each period. We assume that c_p is fixed and c_c is the sum of c_p and o_c. The opportunity cost o_c is always greater than or equal to zero and it decreases with the period length. We propose three different functional forms for o_c. Note that other functions (which can be constructed after an empirical investigation) can be used to model the opportunity costs per unit time, as well.

All proposed functions depend on two more additional parameters: Δ and α. $\Delta > 0$ represents the maximum opportunity cost per time unit due to the availability of the contingent capacity at the start of each period, and $\alpha > 0$ reflects the decreasing rate of the opportunity cost with period length. The proposed functions can be seen in Table 1. For these suggested functional forms of o_c, the effects of Δ and α on c_c are illustrated in Fig. 2.

Suppose the system operates under a stable policy $\pi(k, \mu_\mathrm{l}, \mu_\mathrm{h})$ for infinite horizon. Let $\mathrm{ACU}(\pi(k, \mu_\mathrm{l}, \mu_\mathrm{h}), T)$ denote the average capacity usage and $\mathrm{ACC}(\pi(k, \mu_\mathrm{l}, \mu_\mathrm{h}), T)$ denote the average capacity-related cost resulting from capacity policy $\pi(k, \mu_\mathrm{l}, \mu_\mathrm{h})$ with period length T.

For given c_c and c_p values, $\mathrm{ACC}(\pi(k, \mu_\mathrm{l}, \mu_\mathrm{h}), T)$, can be directly derived from $\mathrm{ACU}(\pi(k, \mu_\mathrm{l}, \mu_\mathrm{h}), T)$:

Table 1 Opportunity cost functions

Name of the function	$o_\mathrm{c}(\Delta, \alpha)$
1. Linear	$(\Delta - \alpha T)^+$
2. Inverse proportional	$\Delta/(1 + \alpha T)$
3. Exponential	$\Delta e^{-\alpha T}$

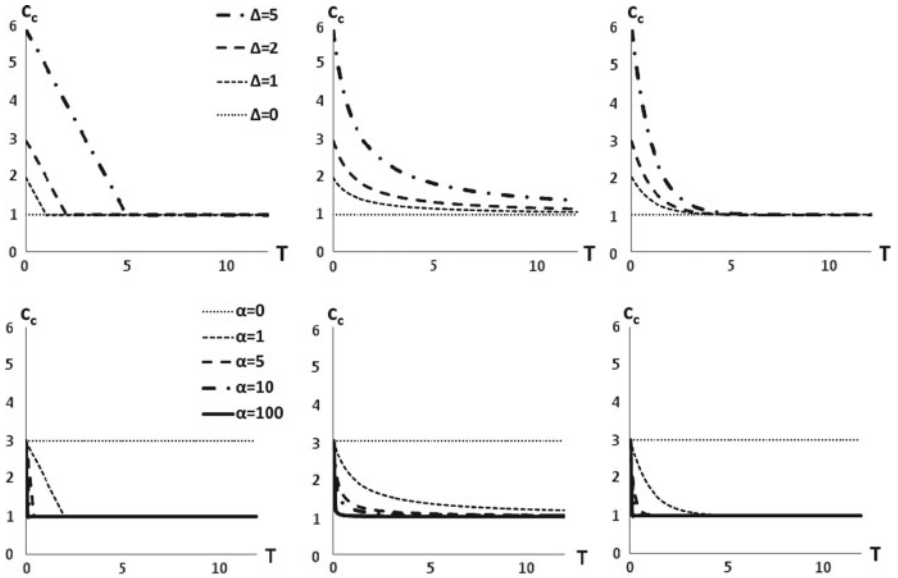

Fig. 2 The figures on the *top* depict the behaviour of c_c for $\alpha = 1$ and $\Delta = 0, 1, 2, 5$. The figures on the *bottom* depict the behaviour of c_c for $\Delta = 2$ and $\alpha = 0, 1, 5, 10, 100$. *Right* to the *left*: o_c is of the linear, inverse proportional and exponential forms

$$\mathrm{ACC}(\pi(k, \mu_l, \mu_h), T) = \mu_l \times c_p + (\mathrm{ACU}(\pi(k, \mu_l, \mu_h), T) - \mu_l) \times c_c \qquad (1)$$

3.2 Problem formulation

For a given permanent labour cost per unit time c_p, and given Δ, α coefficients, the shop floor manager tries to minimize the average capacity costs while satisfying the lead-time constraint with probability γ. In order to achieve the minimum $\mathrm{ACC}(\pi(k, \mu_l, \mu_h), T)$, the shop floor manager has to decide on the following decision variables:

1. Length of the period: T
2. The size of the permanent and the contingent capacity levels: μ_l & μ_h
3. The switching point k to decide on the use of the contingent capacity.

Let the random variable $S(\pi(k, \mu_l, \mu_h), T)$ denote the throughput time of an order in a MTO system under policy $\pi(k, \mu_l, \mu_h)$ with a period length T. The optimization problem can be formulated as follows:

$$\begin{aligned} &\min_{T,\mu_h,\mu_l,k} \quad \mathrm{ACC}(\pi(k, \mu_1, \mu_h), T) \\ &\text{s.t.} \\ &P(S(\pi(k, \mu_1, \mu_h)T) > L) \leq 1 - \gamma \end{aligned} \qquad (2)$$

Calculation of the $\mathrm{ACC}(\pi(k, \mu_l, \mu_h), T)$ using Eq. 1 necessitates the steady state probability vector of the number of the orders at the start of a period in the system.

Each capacity policy results in a different throughput time distribution. In order to check whether a capacity policy satisfies the lead-time performance constraint, the distribution of the throughput time of an order is needed under that capacity policy. The structure of the lead-time performance target requires the percentile information of the throughput time distribution. This detailed information cannot be extracted from Little's Law (Little 1961), which gives information about the average throughput time. Bounds from average throughput time (e.g. by using Markov Inequality) can be used; however, we prefer to follow a constraint satisfaction approach, since satisfying the lead-time performance constraint with the minimum capacity lies in the core of the MTO's responsibilities. In the next section, we analyse the model under study and provide the steps needed to calculate $\mathrm{ACC}(\pi(k, \mu_\mathrm{l}, \mu_\mathrm{h}), T)$ as well as the distribution of $S(\pi(k, \mu_\mathrm{l}, \mu_\mathrm{h}), T)$.

4 Analysis

4.1 Steady-state probabilities of the number of orders at the start of a period

In this subsection, we derive the steady-state probability vector of the number of orders at the start of a period for the production system under study.

Note that in the remainder of the paper, throughput time distribution is denoted as the *sojourn time distribution* and $S(\pi, T)$, $\mathrm{ACU}(\pi, T)$ and $\mathrm{ACC}(\pi, T)$ are used as the shortened versions for the expressions: $S(\pi(k, \mu_\mathrm{l}, \mu_\mathrm{h}), T)$, $\mathrm{ACU}(\pi(k, \mu_\mathrm{l}, \mu_\mathrm{h}), T)$ and $\mathrm{ACC}(\pi(k, \mu_\mathrm{l}, \mu_\mathrm{h}), T)$, respectively.

The production system that operates under periodic capacity policy $\pi(k, \mu_\mathrm{l}, \mu_\mathrm{h})$ with a period length T, has a switching point k, low (permanent) service rate μ_l and the high (permanent + contingent) service rate μ_h.

Without any constraints on the waiting room capacity, the formulas needed to analyse the system would contain infinite sums of Bessel functions, which would make the numerical computations more time-consuming and difficult. However, it is known from the literature (see, e.g. Stern 1979) that the transient behaviour of a Markovian queue with an infinite waiting room can be approximated with that of the same queue but with a finite waiting room. Hence, we model the system as an $M/M/1/K$ queue with periodically adjustable service rates. The accuracy of the approximation is of course dependent on the size of the waiting room K.

Let $X(t)$ denote the number of orders present at time t. The service rate is set to $\mu_\mathrm{l}(\mu_\mathrm{h})$ at the start of period number $n = 1, 2, \ldots$ if $X((n-1)T) < k$ (if $X((n-1)T) \geq k$). When the service rate is set to $\mu_\mathrm{l}(\mu_\mathrm{h})$, the behaviour of $X(t)$ in this dynamic system is identical to the behaviour of the number of orders at time $t\,(t < T)$ in a system with a constant service rate with $\mu_\mathrm{l}\,(\mu_\mathrm{h})$. Therefore, we first analyse the transient behaviour of the $X(t)$ with a constant service rate.

Let $P_{ij}(t, \mu)$ be the probability that there will be j orders at time t given that there are i orders at time "0", under constant service rate of μ. We use numerical methods for the computation of the $P_{ij}(t, \mu)$ values, which are available in the literature (see, Kulkarni 1999; Ledermann and Reuter 1954).

The derivation of the $P_{ij}(t, \mu)$ holds for an arbitrary t for a system with a constant service rate μ. Now, consider our system, which operates under the periodic capacity policy $\pi(k, \mu_{\mathrm{l}}, \mu_{\mathrm{h}})$ with a period length T. Recall that we are interested in the number of orders at the capacity adjustment points, which is denoted as $X(nT)$.

Remark 1 For any period length T, the number of orders at the beginning of each period, $X(nT)$, satisfies the Markov property.

Let $P(T, \pi)$ denote the transition probability matrix of the $X(nT)$ process for $n = 0, 1, \ldots$ under the capacity policy: $\pi(k, \mu_{\mathrm{l}}, \mu_{\mathrm{h}})$. $P_{ij}(T, \pi)$ is the probability that there will be j orders at the end of the period, given that there are i orders at the start of that period.

If there are $i < k$ orders at the start of a period, the service rate is updated to μ_{l} and it is updated to μ_{h} otherwise. The service rate that is updated at the start of a period remains the same until the end of that period.

If we define $\vec{P}_i(T, \pi) = (P_{i0}(T, \pi), P_{i1}(T, \pi), \ldots, P_{iK}(T, \pi))$, as the ith row of $P(T, \pi)$, then we can say $\vec{P}_i(T, \pi) = \vec{P}_i(T, \mu_{\mathrm{l}})$ for $i < k$ and $\vec{P}_i(T, \pi) = \vec{P}_i(T, \mu_{\mathrm{h}})$ for $i \geq k$.

If $\vec{P}_i(T, \pi)$ is obtained when $\mu = \mu_{\mathrm{l}} (\mu = \mu_{\mathrm{h}})$ and $t = T$ for $i < k$ (for $i \geq k$), then we can obtain $P(T, \pi)$, which is the transition probability matrix of $X(nT)$ under $\pi(k, \mu_{\mathrm{l}}, \mu_{\mathrm{h}})$. Let $v(T, \pi)$ be the steady-state vector of the probabilities of the number of orders in the system at the start of a period under policy $\pi(k, \mu_{\mathrm{l}}, \mu_{\mathrm{h}})$ with period length T. After deriving $P(T, \pi)$, the $v(T, \pi)$ vector can now easily be obtained from the following equalities:

$$v(T, \pi) = v(T, \pi) P(T, \pi), \quad \sum_{i=0}^{K} v_i(T, \pi) = 1 \tag{3}$$

Having the steady-state probability vector $v(T, \pi)$, we can calculate the average capacity usage of the policy $\pi(k, \mu_{\mathrm{l}}, \mu_{\mathrm{h}})$ with period length T:

$$\mathrm{ACU}(\pi, T) = \sum_{i=0}^{k-1} \mu_{\mathrm{l}} \times v_i(T, \pi) + \sum_{i=k}^{K} \mu_{\mathrm{h}} \times v_i(T, \pi) \tag{4}$$

4.2 Sojourn time distribution of an arriving job

In order to satisfy the lead-time performance constraint, we need the sojourn time distribution of an arriving order. To derive an explicit formula for the sojourn time distribution of an order, we first define an extended Markov Process, $(X(t), Y(t))$, where $X(t)$ denotes the number of orders in the system at time t, just like in the previous section, and $Y(t)$ denotes the position of a tagged order in the queue (including the order that is being processed) at time t.

Let $\mathbf{Z}$ denote the set of all states of the stochastic process $(X(t), Y(t))$. Then $|Z| = \frac{(K+3)K}{2}$ is the cardinality of $\mathbf{Z}$ because the total number of states is equal to

$$|Z| = 1 + 2 + \cdots + K + (K+1) - 1 = \frac{(K+1)(K+2)}{2} - 1 = \frac{K(K+3)}{2}.$$

Further analysis of the $(X(t), Y(t))$ process under a constant service rate is given in Appendix.

Since the transient probability matrices for the $(X(t), Y(t))$ process under low $(U^l(t))$ and high $(U^h(t))$ service rates are obtained in Appendix, the transient probabilistic behaviour of the $(X(t), Y(t))$ process under the periodic capacity policy $\pi(k, \mu_l, \mu_h)$ can be characterized for $t \leq T$, as well.

Now, we focus on the $(X(t), Y(t))$ process under policy $\pi(k, \mu_l, \mu_h)$ at the start of each period $(t : t = nT$ for $n = 0, 1, 2, \ldots)$. Note that the $(X(nT), Y(nT))$ process has the following property for $n = 1, 2, \ldots$

Remark 2 For any period length T, $(X(nT), Y(nT))$ also satisfies the Markovian property, similar to $X(nT)$ that is defined in Remark 1.

Now, let $A(T, \pi)$ be the $|Z| \times |Z|$ transition probability matrix of $(X(nT), Y(nT))$ process under capacity policy $\pi(k, \mu_l, \mu_h)$ with period length T and for positive integer n values.

$$A_{r,s}(T, \pi) = P((X((n+1)T), Y((n+1)T)) = (s_1, s_2) | (X(nT), Y(nT)) = (r_1, r_2))$$

for $r = (r_1, r_2)$ and $s = (s_1, s_2)$ where $r, s \in \mathbf{Z}$.

We can describe $A_{r,s}(T, \pi)$ as the probability that the system will be in state $s = (s_1, s_2)$ at the end of a period (which means there will be s_1 orders in the system and the tagged order's position will be s_2), given that the system is at state $r = (r_1, r_2)$ at the start of that period (which means there are r_1 orders in the system and the tagged order's position is r_2).

From the definition of the $U^l(t)$ and $U^h(t)$ matrices, we have $A_{r,s}(T, \pi) = U^l_{r,s}(T)$ if $r_1 < k$ and $A_{r,s}(T, \pi) = U^h_{r,s}(T)$ if $r_1 \geq k$ for all $r = (r_1, r_2) \in \mathbf{Z}$. Hence, the $A(T, \pi)$ matrix can be easily constructed from $U^l(T)$ and $U^h(T)$ matrices.

Let $S(\pi, T)$ denote the sojourn time of an arriving order under policy π and period length T. After analysing the transient behaviour of $(X(t), Y(t))$ process under policy π and period length T, we can start deriving $P(S(\pi, T) > x)$ for an arbitrary x.

The sketch of our method is as follows: Suppose it is known that there has been an arrival in a period. For the sake of the convenience, let *origin* denote the start of the period that the order arrives. If the number of orders in the system at the origin is less than k, the initial service rate is equal to μ_l; otherwise, it is equal to μ_h.

Let t denote the time between the arrival of the order and the end of the first period after the arrival. From the conditional distribution of the arrival times (Ross 1996), it is known that the arrival time of an order $(T - t)$ is uniformly distributed over $(0, T)$. If an arrival occurs between $(0, \lceil x/T \rceil^* T - x)$ in a period, then the duration x spreads over $\lceil x/T \rceil$ consecutive periods, whereas if the arrival occurs between $(\lceil x/T \rceil^* T - x, T)$, then x spreads over $\lceil x/T \rceil + 1$ periods.

Hence, for any x, the initial state belongs to one of these events: (1) x spreads over $\lceil x/T \rceil$ periods and initial service rate is μ_l; (2) x spreads over $\lceil x/T \rceil$ periods and

initial service rate is μ_{h}; (3) x spreads over $\lceil x/T \rceil + 1$ periods and initial service rate is μ_{l}; and (4) x spreads over $\lceil x/T \rceil + 1$ periods and initial service rate is μ_{h}.

For each of these event spaces, the probability vector for the number of orders that an arriving order finds in the system upon its arrival is generated. If an arriving order finds j orders ($j < K$) in the system, the extended state of the system, $(X(T-t), Y(T-t))$, upon that arrival will be $(j+1, j+1)$ and the transient behaviour is traced throughout x from that arrival point.

Note that if an order finds K other orders in the system, that arriving order is not accepted. Let R denote the probability of such an event. R can be calculated from

$$R = \sum_{i=0}^{k-1} \frac{1}{T} v_i(T,\pi) \int_{t=x}^{T} P_{iK}(T-t,\mu_{\mathrm{l}})\mathrm{d}t + \sum_{i=k}^{K} \frac{1}{T} v_i(T,\pi) \int_{t=x}^{T} P_{iK}(T-t,\mu_{\mathrm{h}})\mathrm{d}t. \tag{5}$$

While deriving $P(S(\pi,T) > x)$ we condition the probability that an arriving order is accepted. Arbitrarily small values of R can be obtained by taking large enough K. The mathematical formulation of the above sketch is given in Theorem 1.

Theorem 1 *$P(S(\pi,T) > x)$ for $0 \le x < T$ can be written as*

$$\frac{1}{T(1-R)} \sum_{i=h,l} \sum_{j=1,2} \mathrm{P}\left(\mathrm{S}\left(\pi,\mathrm{T}\right) > \mathrm{x}|\mathrm{i},\mathrm{j}\right) \tag{6}$$

where

$$\mathrm{P}\left(\mathrm{S}\left(\pi,\mathrm{T}\right) > \mathrm{x}|l,1\right) = \sum_{i=0}^{k-1} v_i\left(T,\pi\right) \sum_{j=0}^{K-1} \int_{t=x}^{T} P_{ij}\left(T-t,\mu_{\mathrm{l}}\right) \bar{F}_{\mathrm{l}}^{(j+1)}(x)\mathrm{d}t$$

$$\mathrm{P}\left(\mathrm{S}\left(\pi,\mathrm{T}\right) > \mathrm{x}|l,2\right) = \sum_{i=0}^{k-1} v_i\left(T,\pi\right) \sum_{j=0}^{K-1} \int_{t=0}^{x} P_{ij}\left(T-t,\mu_{\mathrm{l}}\right) \vec{U}^{l}_{(j+1,j+1)}(t) \vec{\bar{F}}(x-t)^{\mathrm{Tr}}\mathrm{d}t$$

$$\mathrm{P}\left(\mathrm{S}\left(\pi,\mathrm{T}\right) > \mathrm{x}|h,1\right) = \sum_{i=k}^{k} v_i\left(T,\pi\right) \sum_{j=0}^{K-1} \int_{t=x}^{T} P_{ij}\left(T-t,\mu_{\mathrm{h}}\right) \bar{F}_{\mathrm{h}}^{(j+1)}(x)\mathrm{d}t$$

$$\mathrm{P}\left(\mathrm{S}\left(\pi,\mathrm{T}\right) > \mathrm{x}|h,2\right) = \sum_{i=k}^{k} v_i\left(T,\pi\right) \sum_{j=0}^{K-1} \int_{t=0}^{x} P_{ij}\left(T-t,\mu_{\mathrm{h}}\right) \vec{U}^{h}_{(j+1,j+1)}(t) \vec{\bar{F}}(x-t)^{\mathrm{Tr}}\mathrm{d}t$$

In a similar manner, $P(S(\pi,T) > x)$ for $(n-1)T \le x < nT$ if $n \ge 2$ can be written as follows:

$$\frac{1}{T(1-R)} \sum_{i=h,l} \sum_{j=n,n+1} P\left(S(\pi,T) > x|i,j\right) \tag{7}$$

$$P\left(S\left(\pi,T\right)>x|l,n\right)=\sum_{i=0}^{k-1}\sum_{j=0}^{K-1}v_i\left(T,\pi\right)$$
$$\times\int_{t=x-(n-1)T}^{T}P_{i,j}\left(T-t,\mu_1\right)\vec{U}^{l}_{(j+1,j+1)}(t)A\left(T,\pi\right)^{n-2}\vec{\bar{F}}\left(x-(n-2)T-t\right)^{\mathrm{Tr}}\mathrm{d}t$$

$$P\left(S\left(\pi,T\right)>x|l,n+1\right)=\sum_{i=0}^{k-1}\sum_{j=0}^{K-1}v_i\left(T,\pi\right)$$
$$\times\int_{t=0}^{x-(n-1)T}P_{i,j}\left(T-t,\mu_1\right)\vec{U}^{l}_{(j+1,j+1)}(t)A\left(T,\pi\right)^{n-1}\vec{\bar{F}}\left(x-(n-1)T-t\right)^{\mathrm{Tr}}\mathrm{d}t$$

$$P\left(S\left(\pi,T\right)>x|h,n\right)=\sum_{i=k}^{K}\sum_{j=0}^{K-1}v_i\left(T,\pi\right)$$
$$\times\int_{t=x-(n-1)T}^{T}P_{i,j}\left(T-t,\mu_{\mathrm{h}}\right)\vec{U}^{h}_{(j+1,j+1)}(t)A\left(T,\pi\right)^{n-2}\vec{\bar{F}}\left(x-(n-2)T-t\right)^{\mathrm{Tr}}\mathrm{d}t$$

$$P\left(S\left(\pi,T\right)>x|h,n+1\right)=\sum_{i=k}^{K}\sum_{j=0}^{K-1}v_i\left(T,\pi\right)$$
$$\times\int_{t=0}^{x-(n-1)T}P_{i,j}\left(T-t,\mu_{\mathrm{h}}\right)\vec{U}^{h}_{(j+1,j+1)}(t)A\left(T,\pi\right)^{n-1}\vec{\bar{F}}\left(x-(n-1)T-t\right)^{\mathrm{Tr}}\mathrm{d}t$$

Where a^{Tr} is the transpose of vector a,

$$\vec{U}^{l}_{(j+1,j+1)}(t)=\left(U^{l}_{(j+1,j+1),s^1}(t),U^{l}_{(j+1,j+1),s^2}(t),\ldots,U^{l}_{(j+1,j+1),s^{|Z|}}(t)\right),$$

$$\vec{U}^{h}_{(j+1,j+1)}(t)=\left(U^{h}_{(j+1,j+1),s^1}(t),U^{h}_{(j+1,j+1),s^2}(t),\ldots,U^{h}_{(j+1,j+1),s^{|Z|}}(t)\right),$$

$$\vec{\bar{F}}\left(x\right)=\left(I\left(s_1^1,s_2^1,k,x\right),I\left(s_1^2,s_2^2,k,x\right),\ldots,I\left(s_1^{|Z|},s_2^{|Z|},k,x\right)\right)$$

for $I\left(i,j,k,x\right)=\bar{F}_{\mathrm{l}}^{j}(x)=P(B>x)$ where $B\sim\mathrm{Erlang}(j,\mu_{\mathrm{l}})$ for $i<k$ and $\mathrm{I}\left(i,j,k,x\right)=\bar{F}_{\mathrm{h}}^{j}(x)=P(B>x)$ where $B\sim\mathrm{Erlang}(j,\mu_{\mathrm{h}})$ for $i\geq k$.

Note that the sth element of $\vec{U}^{l}_{(j+1,j+1)}(t)$ $\left(\vec{U}^{h}_{(j+1,j+1)}(t)\right)$ above denotes the probability that after t time units from the order arrival, the number of total orders will be s_1, $(X(T)=s_1)$, and the position of the order which has arrived at time $T-t$ will be s_2, $(Y(T)=s_2)$, for $\mu=\mu_{\mathrm{l}}$ $(\mu=\mu_{\mathrm{h}})$. The formula in Eq. 7 generalizes the

formula given in Eq. 6 for $S > T$ by keeping the track of the extended state of an order throughout a period with the help of the $A(T, \pi)$ matrix.

As the sojourn time distribution of an order is derived, we can incorporate this result with the lead-time performance constraint. As explained before, in a lead-time performance constraint with a lead time L, the on-time delivery target $0 < \gamma < 1$ guarantees that the proportion of jobs whose throughput time is more than L will not exceed $1 - \gamma$. Therefore, the lead-time performance constraint under a periodic capacity policy $\pi(k, \mu_l, \mu_h)$ can be expressed as

$$P(S(\pi, T) \leq L) \leq \gamma.$$

4.3 Randomized switching option

From Stochastic Optimization Theory, we know that for unconstrained MDP problems, there exists a nonrandomized optimal policy. However when the problem is constrained, randomized action taking in some states can yield a better result (Puterman 1994). Since the problem under study is a constrained problem (due to the imposed lead-time performance constraint); the randomized action taking should be included to the capacity control policy.

Therefore, in this subsection we introduce the randomized switching point notion. A randomized switching point can be introduced to an existing non-randomized policy $\pi(k, \mu_l, \mu_h)$ via a probability factor p. Due to the randomized switching option, the switching point is no longer an integer point, but can be any positive real number.

In the two-level capacity policy $\pi(k, \mu_l, \mu_h)$ that was analysed so far, the service rate is set to μ_l if there are less than k orders in the system at the start of a period; otherwise it is set to μ_h. Now in the randomized policy $\pi(k, \mu_l, \mu_h)$ with probability p, if there are exactly k orders in the system at the start of a period, the service rate of the system is μ_h with probability p, and μ_l with probability $1 - p$. On the other hand, if there are less (more) than k orders, the service rate is set to μ_l (μ_h). Such a randomized switching in a two-level threshold policy can be interpreted such that, instead of k, the new switching point is $p(k-1) + (1-p)k$.

Incorporating the randomized switching at state k to the analysis can be achieved as follows: Under the non-randomized policy, when there are k orders at the start of a period, a service rate of μ_h is used during the derivation of the expressions from Sects. 4.1 and 4.2. Let $g(k, \mu_h)$ denote one of these expression in question. Under the randomized policy, when there are k orders, each $g(k, \mu_h)$ should be replaced with the weighted average of the same expression with service rates μ_l and μ_h. Mutatis mutandis, when there are k orders, $(1-p)g(k, \mu_h) + pg(k, \mu_l)$ should be used rather than $g(k, \mu_h)$ during the calculations under the randomized policy.

After the necessary expression updates are completed, $\mathrm{ACU}(\pi, T)$ with switching probability p can be re-obtained as follows:

$$\mathrm{ACU}(\pi, T) = \mu_l + (\mu_h - \mu_l)\left(\sum_{i=k+1}^{K} v_i(T, \pi) + (1-p) \times v_k(T, \pi)\right) \tag{8}$$

In the next section, we use the results from this section to determine the benefits that can be obtained from employing a periodic, two-level capacity policy as a function of lead time and lead-time performance requirement on the one hand, and the additional costs of deploying flexible capacity policy as a function of period length due to the opportunity cost factors Δ and $\alpha.$, on the other hand.

5 Computational study

This section consists of three subsections. In Sect. 5.1, we describe the computational study and explain how we decide on the best policy parameters under certain lead-time performance requirement (L, γ) and opportunity cost function $o_c(\Delta, \alpha)$. Before giving cost results/comparisons, in Sect. 5.2, we discuss the interrelations among the system and policy parameters, like the interrelations between L and the period length T and the impact of the randomized policies. Finally, in Sect. 5.3, we present the savings in operating costs when the best flexible periodic capacity is employed compared with the best fixed capacity. These savings infer the (Δ, α) regions where the use of flexible capacity policies would still be preferable if there was an additional fixed cost rate for the set up and operating of the periodic capacity policy.

5.1 Design of the computational study and search for the best policy parameters

In our computational study, we normalize the arrival rate, $\lambda = 1$ (customers per time unit) and also the permanent capacity cost per time unit, $c_p = 1$. We investigate three markets with different demands on lead-time performance ($L = 10$ and $\gamma = 0.90$, $L = 5$ and $\gamma = 0.90$, $L = 5$ and $\gamma = 0.95$), which represent increasing levels of ambition.

As mentioned before, we approximate the real system with a system having a finite waiting room of size K. Naturally, the quality of approximation highly depends on the choice of K. In our experiments we take K between 50 and 60. We have several means to validate the accuracy of K. One of them is the R value that is presented just before Theorem 1 in Sect. 4.2. In our experiments, all parameter combinations yield an R value that is practically equal to zero. Another alternative to assess the quality of the choice of K can be to compare the steady-state solutions of the number of orders at the start of a period under increasing K values. The change in the steady-state solutions should become negligible after some K value. We observe this behaviour in our studies, as well.

For a given lead time L and a performance level γ, we first determine the optimal capacity level for the reference case with constant capacity. Let $\mu_{L,\gamma}$ denote this capacity level. From the sojourn time properties of the $M/M/1$ queue, $\mu_{L,\gamma}$ can be found analytically from $\mu_{L,\gamma} = \lambda - \frac{\ln(1-\gamma)}{L}$. From $\mu_{L,\gamma}$ and $\lambda = c_p = 1$, we can find the minimum cost rate for the constant capacity as a reference point: $c_p \mu_{L,\gamma} = \mu_{L,\gamma} = 1 - \frac{1}{L}\ln(1-\gamma)$. In addition, $\mu_{L,\gamma}$ level plays an important role in determiningμ_l and μ_h levels for the periodic capacity policy. These levels should satisfy $\mu_l \leq \mu_{L,\gamma} \leq \mu_h$. Otherwise, either the constraint $P(S(\pi, T) \leq L) \leq \gamma$ is not satisfied or ACU(π, T) becomes unnecessarily high.

For a given lead-time performance constraint (L, γ) and a given opportunity cost function $o_c(\Delta, \alpha)$, the shop floor manager has to decide on the period length T, permanent and contingent capacity levels μ_l and $\mu_h - \mu_l$, and the switching point k to switch from high to low service rate (or vice versa). The corresponding optimization problem cannot be solved with standard optimization techniques since the objective function and the sojourn time distribution are to be derived numerically for every new policy. Therefore, we propose a search method to find the best policy parameters.

Before starting the search, for a given L and γ, we first create the $\Upsilon_l = \{\mu_{l1}, \mu_{l2}, \mu_{l3}, \mu_{l4}, \mu_{l5}\}$ and $\Upsilon_h = \{\mu_{h1}, \mu_{h2}, \mu_{h3}, \mu_{h4}, \mu_{h5}\}$ sets for candidate μ_l and μ_h levels, where $\mu_{li} = \frac{i \times \mu_{L,\gamma}}{6}$ and $\mu_{hi} = \mu_{L,\gamma} + \mu_{li}$ for all $i = 1, 2, \ldots, 5$.

Similarly, for every problem instance, we have a set of candidate period lengths, θ, which are the integer multiples of 0.5 up to L. After an μ_l level from Υ_l and an μ_h level from Υ_h, and a candidate period length T from θ are chosen, the corresponding switching point $k^*(\mu_l, \mu_h, T)$ is found from the following sub-problem:

$$\begin{aligned} &\min_k \text{ACC}(\pi(k, \mu_l, \mu_h), T) \\ &\text{s.t.} \\ &P(S(\pi(k, \mu_l, \mu_h), T) > L) \leq 1 - \gamma \end{aligned} \tag{9}$$

From Eq. 1, it can be seen that $\text{ACC}(\pi(k, \mu_l, \mu_h), T)$ is increasing with $\text{ACU}(\pi(k, \mu_l, \mu_h), T)$. In all of our computational results, we observe the following:

- For given μ_l, μ_h and T, $\text{ACU}(\pi(k, \mu_l, \mu_h), T)$ is non-increasing in k.
- For given μ_l, μ_h and T, $P(S(\pi(k, \mu_l, \mu_h), T) > L)$ is non-decreasing in k.

These aforementioned behaviours of $\text{ACU}(\pi(k, \mu_l, \mu_h), T)$ and $P(S(\pi(k, \mu_l, \mu_h), T) > L)$ can be seen in Fig. 3 for $\lambda = 1$, $\mu_l = 0.24342$, $\mu_h = 1.7039$, $\gamma = 0.9$, $L = 5$ and $T = 2$ for increasing levels of real k values.

Assuming the aforementioned monotonicity of $\text{ACU}(\pi(k, \mu_l, \mu_h), T)$ and $P(S(\pi(k, \mu_l, \mu_h), T) > L)$, from the KKT conditions, we can state that the optimal switching point $k^*(\mu_l, \mu_h, T)$ is the largest possible switching point that satisfies $P(S(\pi(k, \mu_l, \mu_h), T) > L) \leq \gamma$. After finding $k^*(\mu_l, \mu_h, T)$ for all $\mu_l \in \Upsilon_l$, $\mu_h \in \Upsilon_h$ the best policy $\pi^*(T)$, for a given period length $T \in \theta$ can be found via brute force search:

$$\text{ACU}(\pi^*(T), T) = \min_{\mu_l \in \Upsilon_l, \mu_h \in \Upsilon_h} \{\text{ACU}(\pi(k^*(\mu_l, \mu_h, T), \mu_l, \mu_h), T)\}$$

Finally, the best period length and the minimum achievable costs can be found:

$$T^* = \arg\min_{T \in \theta} \{\text{ACC}(\pi^*(T), T)\}$$

Due to randomized policies, we can have non-integer switching point values which enables the system to meet the lead-time performance constraints more tightly (in the ideal case with equality) with less average capacity usage. In our numerical study, we restrict switching points to be integer multiples of $p = 0.1$. Suppose k^*_{NR} is the

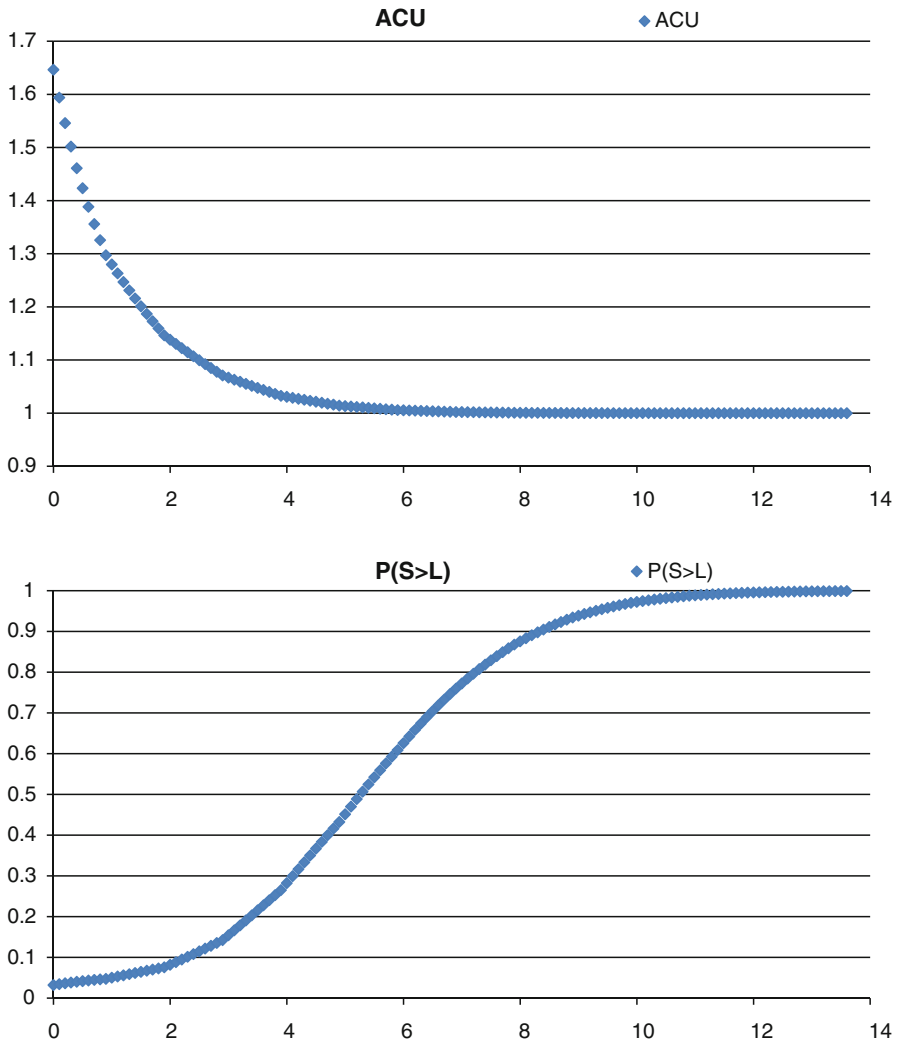

Fig. 3 The figures *above* depict how $\mathrm{ACU}(\pi(k, \mu_{\mathrm{l}}, \mu_{\mathrm{h}}), T)$ and $P(S(\pi(k, \mu_{\mathrm{l}}, \mu_{\mathrm{h}}), T) > L)$ behave for increasing levels of switching points (k is not necessarily an integer due to the randomization) when $\lambda = 1,\ \mu_{\mathrm{l}} = 0.24342,\ \mu_{\mathrm{h}} = 1.7039,\ \gamma = 0.9,\ L = 5$ and $T = 2$

optimal switching point to (8) when k can only be an integer and k_R^* is the optimal switching point when k is an integer multiple of 0.1. In Table 2, we present the average percentage increase in average capacity usage (ACU) due to using non-randomized policies for every (L, γ) and T, which can be derived from

$$100 \times \frac{\sum_{\mu_{\mathrm{l}} \in \Upsilon_{\mathrm{l}}, \mu_{\mathrm{h}} \in \Upsilon_{\mathrm{h}}} \left[\mathrm{ACU}\left(\pi\left(k_{NR}^*\left(\mu_{\mathrm{l}}, \mu_{\mathrm{h}}, T\right), \mu_{\mathrm{l}}, \mu_{\mathrm{h}}\right), T\right) - \mathrm{ACU}\left(\pi\left(k_R^*\left(\mu_{\mathrm{l}}, \mu_{\mathrm{h}}, T\right), \mu_{\mathrm{l}}, \mu_{\mathrm{h}}\right), T\right)\right]}{25}.$$

Table 2 Mean percentage increase in ACU when non-randomized policies are used under different TL values and lead-time performance constraints

T/L	1/10	2/10	3/10	4/10	5/10	6/10	7/10	8/10	9/10	1
$L = 10,\ \gamma = 0.9$	0.21	0.25	0.38	0.43	0.53	0.60	0.93	0.84	0.85	1.05
$L = 5,\ \gamma = 0.9$	1.12	1.38	1.78	1.89	2.15	3.05	2.50	2.58	3.77	2.85
$L = 5,\ \gamma = 0.95$	1.41	1.94	2.56	2.37	2.90	2.74	2.62	4.04	4.06	5.44

From Table 2, it can be observed that the use of randomized policies bring more savings in ACU for more ambitious lead-time performance constraints and larger period lengths. These values in Table 2 constitute a lower bound for the increase in ACC, since the savings of randomized policies are the savings from contingent capacity usage and the c_c is at least same as or more expensive than c_p.

5.2 Interrelations between the system and policy parameters

In this subsection we discuss the interrelations between the system parameters and the optimal policy parameters. From the numerical studies, we observe that, when the other parameters are the same, the optimal switching point $k^*(\mu_l, \mu_h, T)$ increases with μ_l and μ_h but decreases with the ambition level of the lead-time performance constraint and the period length T. Since $k^*(\mu_l, \mu_h, T)$ is determined only from the policy parameters and the lead-time performance constraint, cost parameters such as Δ and α do not affect the choice of the switching point.

Next, we investigate the effects of period length and capacity costs on the choice of permanent capacity levels. From our computational study, we observe that the production system tends to employ less permanent capacity (μ_l) for smaller T, smaller Δ and larger α values. This behaviour can be explained as follows: as the opportunities to update the capacity become more frequent and less costly, hiring contingent capacity temporarily becomes more attractive than deploying permanent capacity.

Subsequently, we discuss the interrelations between L and T and their impacts on $\text{ACC}(\pi^*(T), T)$ under different opportunity cost parameters (Δ, α) when $o_c(\Delta, \alpha) = \Delta/(1 + \alpha T)$. In our periodic capacity control framework, (L/T) value arises as a flexibility metric that shows the number of capacity update opportunities during the lead-time L. In the presence of opportunity costs, more update opportunities come at a higher price.

We first investigate the effects of Δ, the maximum value of the opportunity, on the capacity-related costs. Figure 4 shows the minimum capacity costs for each of the three lead-time performance constraints with a constant α, ($\alpha = 1$) and different Δ values ($\Delta = 0, 0.5$ and 1) as a function of period length: $\text{ACC}(\pi^*(T), T)$.

We can see in Fig. 4 that minimum capacity costs with positive Δ are higher than minimum capacity costs with $\Delta = 0$, but lower than $c_{p*}\mu_{L,\gamma}$ for every period length T under all three lead-time performance constraints.

In each of the three lead-time performance constraints, when period length T gets close to L, we observe that $\text{ACC}(\pi^*(T), T)$ values with $\Delta = 0.5$ and $\Delta = 1$ almost

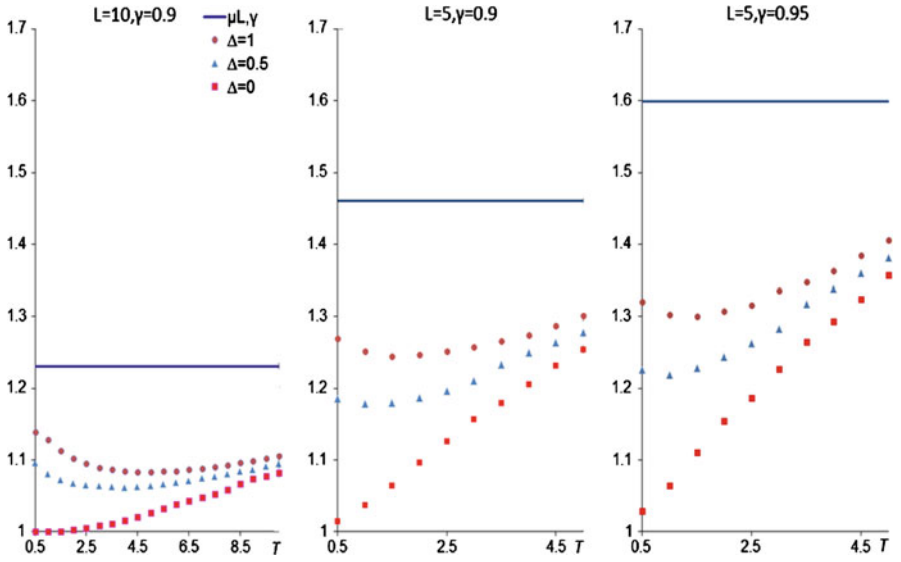

Fig. 4 ACC($\pi^*(T)$, T) for increasing T under three lead-time performance constraints when $c_p = 1$, $\alpha = 1$ and $\Delta = 0, 0.5, 1$. (from *left* to the *right* ambition level of the lead-time constraint increases)

Table 3 T^* for different Δ when $c_p = 1$ and $\alpha = 1$ under three lead-time performance constraints

	$\Delta = 0$	$\Delta = 0.5$	$\Delta = 1$
$L = 10,\ \gamma = 0.9$	0.5	4	5
$L = 5,\ \gamma = 0.9$	0.5	1.5	1.5
$L = 5,\ \gamma = 0.95$	0.5	1	1.5

overlap with that of ACU($\pi^*(T)$, T). The gaps between ACC($\pi^*(T)$, T) with different Δ values are biggest for the smallest possible period length ($T = 0.5$), due to the structure of the contingent capacity cost function.

Note that for the low-ambition lead-time performance constraint ($L = 10$, $\gamma = 0.9$), when $\Delta > 0$, the minimum capacity costs with the shortest period length are higher than the minimum capacity costs with the longest period length. On the other hand, for the higher-ambition settings, we observe the opposite. As the lead-time performance constraint becomes more ambitious, minimum capacity costs for large period lengths (around L) increase and get more expensive compared to the minimum capacity costs short period lengths (around 0.5).

From the figure, for positive Δ, it can be seen that the minimum capacity costs first decrease and then increase with T. The period length that yields the minimum capacity costs, T^*, is affected by both the specifications of lead-time performance constraint and the specifications of contingent capacity cost structure. In Table 3, we present the best period lengths T^* for all three lead-time performance constraints when $\alpha = 1$ and $\Delta = 0, 0.5, 1$.

From Table 3, we can see that T^* increases with Δ and decreases with the ambition level of the lead-time performance constraint (when $\Delta > 0$). If the lead-time perfor-

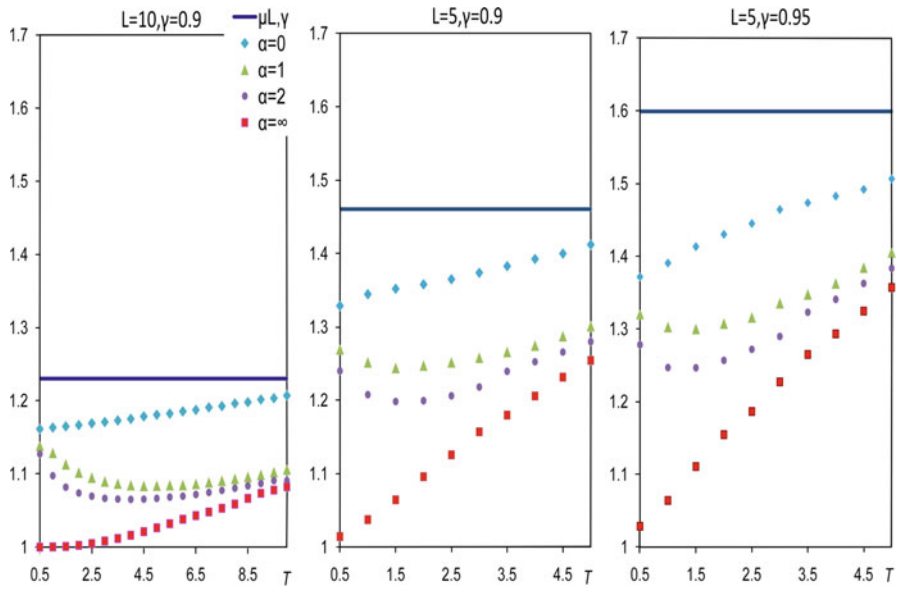

Fig. 5 ACC($\pi^*(T)$, T) for increasing T under three lead-time performance constraints for $c_p = 1$, $\Delta = 1$ and $\alpha = 0, 1, 2, \infty$ (from *left* to *right* the ambition level of the lead-time constraint increases)

mance constraints are more ambitious, the system appreciates the flexibility option more and would prefer to tailor its capacity more frequently at the expense of higher capacity costs. If the contingent capacity cost per unit time is same as the permanent capacity cost per unit time, then operating with the smallest possible period length is the right thing to do. The more expensive the contingent capacity costs become, closer T^* value get to the L.

The decreasing rate of the opportunity cost, α, is also a very important factor that determines the behaviour of the minimum capacity costs in response to the period length. Two extreme values that α can take are 0 and ∞, respectively. When $\alpha = \infty$, the contingent capacity can be immediately assigned to another task if it is not deployed by the system at the start of a period. Therefore, the cost burden of the lost opportunities disappears and ACC($\pi^*(T)$, T) behaves as if $\Delta = 0$. On the other hand, when $\alpha = 0$, the assignment of the contingent capacity to another task is not possible; hence, the lost opportunity cost during a non-used period of the contingent capacity is not affected by the period length and it is a constant. In Fig. 5, minimum capacity costs as a function of T, for $\Delta = 1$, with different α, under three different lead-time performance constraints are given.

From Fig. 5, it can be observed that ACC($\pi^*(T)$, T) decreases with increasing α, since the contingent capacity gets cheaper for larger α values when Δ is positive. It can also be observed that as T approaches to zero, ACC($\pi^*(T)$, T) values with $0 < \alpha < \infty$ are closer to the ACC($\pi^*(T)$, T) values with $\alpha = 0$, and as T approaches to L, the ACC($\pi^*(T)$, T) values are closer to the ACC($\pi^*(T)$, T) values with $\alpha = \infty$.

When $\alpha = 0$ or $\alpha = \infty$, ACC($\pi^*(T)$, T) behaves as a monotone increasing function of T; however, for other mid-values of α, ACC($\pi^*(T)$, T) has a more

Table 4 T^* for different α with $c_p = 1$, $\Delta = 1$ and under three lead-time performance constraints

	$\alpha = 0$	$\alpha = 1$	$\alpha = 2$	$\alpha = \infty$
$L = 10,\ \gamma = 0.9$	0.5	5	4.5	0.5
$L = 5,\ \gamma = 0.9$	0.5	1.5	1.5	0.5
$L = 5,\ \gamma = 0.95$	0.5	1.5	1.5	0.5

U-shaped structure. Therefore, for these values of α, the best period length T^* is not necessarily the smallest period length ($T = 0.5$), and is affected by the choice of α. In Table 4, we present the best period lengths T^* for all three lead-time performance configurations when $\Delta = 1$ and $\alpha = 0, 1, 2$ and ∞.

From the table, it can be seen that T^* first increases and then decreases with α. When $\alpha = 1$ or $\alpha = 2$, we can see that T^* decreases as the lead-time performance gets more ambitious; however, when $\alpha = 0$ and $\alpha = \infty$, the contingent capacity cost becomes independent of the period length and therefore we have $T^* = 0.5$, the smallest period length in our test bed.

5.3 Possible savings in operating costs for different environments

Under a single-level capacity policy, the effects of a change in the lead-time performance constraint on the minimum capacity costs can be seen from the formula provided in Sect. 5.1. As can be seen from the $\mu_{L,\gamma}$ formula, an increase in the ambition level of the lead-time performance constraint requires an additional capacity to be deployed.

Let $\text{ACC}_i^* = \text{ACC}(\pi^*(T^*), T^*)$ be the best operating costs that can be achieved under a two-level periodic capacity policy with the best period length T^* and a lead-time performance constraint i for $i = 1, 2, 3$.

A two level periodic capacity policy not only reduces the average capacity costs, but also may balance the increase in the ACC* when a more ambitious lead-time performance constraint is used. In Fig. 6, we can see the percentage increases in ACC* when a more ambitious lead-time performance constraint is adopted.

On the left, the percentage increases in $\text{ACC}^*(100 \times \frac{\text{ACC}_2^* - \text{ACC}_1^*}{\text{ACC}_1^*})$ are shown when the low-ambition lead-time performance constraint ($L = 10,\ \gamma = 0.9$) is changed to a more ambitious (medium level) lead-time performance constraint ($L = 5,\ \gamma = 0.9$) and similarly on the right, the percentage increases in $\text{ACC}^*(100 \times \frac{\text{ACC}_3^* - \text{ACC}_2^*}{\text{ACC}_2^*})$ are shown when the medium level ambition lead-time performance constraint ($L = 5, \gamma = 0.9$) is changed to an even more ambitious (high level) lead-time performance constraint ($L = 5, \gamma = 0.95$) for the single level capacity policy, for two-level capacity policy with $c_p = \Delta = \alpha = 1$ and the two-level capacity policy with $c_p = 1$, $\Delta = 0$. We can see that the usage of a two-level capacity policy can soothe the drastic changes in ACC* upon the adoption of a different lead-time performance constraint. Especially, when $\Delta = 0$, i.e. when $c_c = c_p$, adopting a more ambitious lead-time performance constraint would barely increase the average capacity-related costs under a two-level capacity policy.

From this, we get the following conclusion: if the contingent capacity is not that expensive compared with the permanent capacity, two-level capacity policies are quite

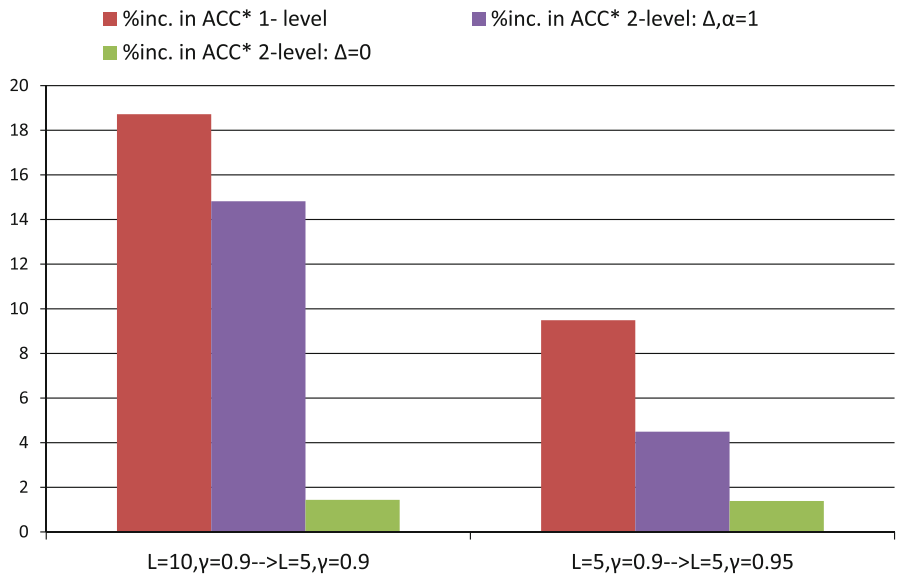

Fig. 6 % increase in ACC* when the lead-time performance constraint is changed under single and two-level periodic capacity policies

robust to the changes in the lead time or delivery performance target agreements in terms of capacity costs.

Finally, we investigate how much savings that a two-level capacity policy can bring compared to a single-level policy under different (Δ, α) settings. We are interested in the percentage savings in the minimum capacity costs that can be achieved from a two-level periodic capacity policy compared with $c_p \times \mu_{L,\gamma}$, which is the ACC* under the best single level capacity policy. Under lead-time performance constraint i for $i = 1, 2, 3.$, the percentage cost savings that a two-level capacity policy brings can be found from $100 \times \frac{c_p \times \mu_{L_i,\gamma_i} - \text{ACC}_i^*}{c_p \times \mu_{L_i,\gamma_i}}$. In Table 5, the percentage savings are given for different functional forms (linear, inverse proportional and exponential) of opportunity cost, $o_c(\Delta, \gamma)$ with Δ and α varying from 1 to 5.

From Table 5, it can be seen that the percentage savings that a two-level policy can bring in ACC* increases with the ambition level of the lead-time performance constraint, increases with α and decreases with Δ for all functional forms of $o_c(\Delta, \gamma)$. Recall that Δ denotes the maximum value that the per unit time opportunity cost can get, and α is a factor that affects the decreasing rate of o_c with regard to T. The benefits of periodic capacity flexibility are more tangible when the lead-time delivery performance target settings are ambitious or when the contingent capacity is not that expensive compared to the permanent capacity. Note that for all the three functional forms, we can find instances where the cost performance of the best two-level flexible capacity policy is worse than the cost performance of the single-level capacity policy. These cases are typically the (Δ, α) combinations, where Δ is quite high in comparison with c_p and where α is quite small; thus, the o_c is rather insensitive to the period

Table 5 Percentage savings that a two-level capacity policy can bring in ACC* compared with the single-level capacity policy under different lead-time performance constraints and different functional forms (linear–inverse proportional–exponential) of opportunity costs with varying $\Delta(1–5)$ and $\alpha(1–5)$ values

$c_p\mu_{L,\gamma} = 1.599$	Linear $1-L=5,\ \gamma=95\%$						Inverse proportional $1-L=5,\ \gamma=95\%$						Exponential $1-L=5,\ \gamma=95\%$					
$\alpha\backslash\Delta$	0 (%)	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	0 (%)	1(%)	2 (%)	3 (%)	4 (%)	5 (%)	0 (%)	1 (%)	2 (%)	3(%)	4 (%)	5 (%)
0	−35.7	−14.2	−8.1	−5.2	−2.3	0.6	−35.7	−14.2	−8.1	−5.2	−2.3	0.6	−35.7	−14.2	−8.1	−5.2	−2.3	0.6
1	−35.7	−33.5	−27.8	−23.3	−19.1	−15.1	−35.7	−18.7	−14.3	−10.8	−8.5	−7.4	−35.7	−23.0	−20.5	−19.2	−17.8	−16.8
2	−35.7	−35.7	−33.5	−30.5	−27.8	−25.8	−35.7	−22.1	−17.4	−15.1	−13.1	−11.0	−35.7	−28.2	−26.4	−25.6	−24.8	−24.4
3	−35.7	−35.7	−33.5	−33.5	−30.5	−27.8	−35.7	−23.8	−19.5	−17.0	−15.5	−14.0	−35.7	−31.0	−29.4	−28.8	−28.4	−28.0
4	−35.7	−35.7	−35.7	−33.5	−33.5	−30.5	−35.7	−25.2	−21.3	−18.2	−16.9	−15.7	−35.7	−32.6	−31.6	−30.7	−30.0	−29.9
5	−35.7	−35.7	−35.7	−33.5	−33.5	−33.5	−35.7	−26.3	−22.4	−19.8	−17.8	−16.8	−35.7	−33.1	−32.8	−32.5	−32.1	−31.8
$c_p\mu_{L,\gamma} = 1.460$	Linear $2-L=5,\ \gamma=90\%$						Inverse proportional $2-L=5,\ \gamma=90\%$						Exponential $2-L=5,\ \gamma=90\%$					
$\alpha\backslash\Delta$	0 (%)	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	0 (%)	1(%)	2 (%)	3 (%)	4 (%)	5 (%)	0 (%)	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)
0	−30.5	−9.0	−4.0	0.3	4.5	8.7	−30.5	−9.0	−4.0	0.3	4.5	8.7	−30.5	−9.0	−4.0	−0.3	−4.5	8.7
1	−30.5	−29.0	−25.0	−20.8	−17.4	−14.1	−30.5	−14.8	−10.1	−7.3	−6.0	−4.7	−30.5	−19.8	−17.9	−16.4	−15.3	−14.6
2	−30.5	−30.5	−29.0	−27.1	−25.0	−22.9	−30.5	−18.0	−13.7	−11.1	−8.8	−7.5	−30.5	−24.7	−23.4	−22.7	−21.9	−21.4
3	−30.5	−30.5	−29.0	−29.0	−27.1	−25.0	−30.5	−19.7	−15.6	−13.3	−11.6	−10.0	−30.5	−26.5	−25.9	−25.4	−24.9	−24.5
4	−30.5	−30.5	−30.5	−29.0	−29.0	−27.1	−30.5	−21.0	−17.3	−14.6	−13.1	−11.8	−30.5	−28.0	−27.0	−26.7	−26.6	−26.4
5	−30.5	−30.5	−30.5	−29.0	−29.0	−29.0	−30.5	−22.0	−18.4	−15.9	−14.2	−13.0	−30.5	−28.6	−28.2	−27.9	−27.5	−27.1

Table 5 continued

$c_p\mu_{L,\gamma}=1.230$	Linear $3-L=10,\ \gamma=90\%$						Inverse proportional $3-L=10,\ \gamma=90\%$						Exponential $3-L=10,\ \gamma=90\%$					
$\alpha\backslash\Delta$	0 (%)	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	0 (%)	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	0 (%)	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)
0	−18.7	−5.6	−0.1	5.5	11.0	16.5	−18.7	−5.6	−0.1	5.5	11.0	16.5	−18.7	−5.6	−0.1	−5.5	−11.0	16.5
1	−18.7	−18.7	−18.5	−18.0	−17.4	−16.6	−18.7	−11.9	−9.1	−7.5	−6.7	−6.0	−18.7	−16.5	−16.0	−15.7	−15.5	−15.3
2	−18.7	−18.7	−18.7	−18.6	−18.5	−18.3	−18.7	−13.4	−11.6	−10.2	−8.8	−7.7	−18.7	−17.9	−17.8	−17.7	−17.6	−17.5
3	−18.7	−18.7	−18.7	−18.7	−18.6	−18.5	−18.7	−14.5	−12.7	−11.5	−10.6	−9.6	−18.7	−18.4	−18.3	−18.2	−18.2	−18.1
4	−18.7	−18.7	−18.7	−18.7	−18.7	−18.6	−18.7	−15.1	−13.3	−12.4	−11.5	−10.8	−18.7	−18.5	−18.5	−18.5	−18.4	−18.4
5	−18.7	−18.7	−18.7	−18.7	−18.7	−18.7	−18.7	−15.6	−13.9	−12.9	−12.1	−11.5	−18.7	−18.6	−18.6	−18.5	−18.5	−18.5

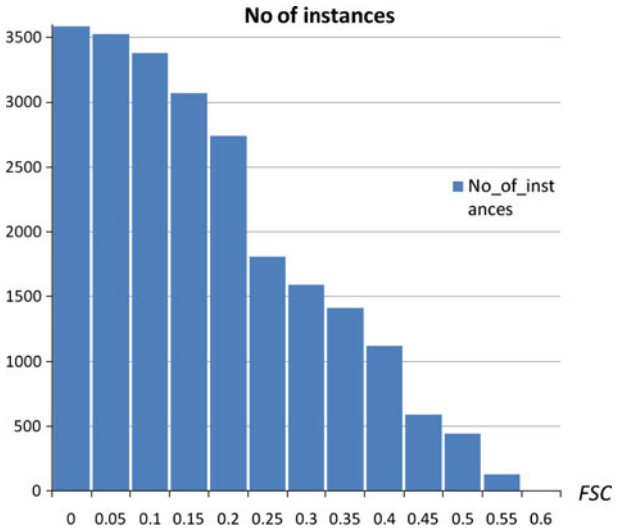

Fig. 7 Number of instances (out of different 3780 scenario instances), where using capacity flexibility is preferable even if there is a fixed flexible system cost rate FSC

length T. For these instances it is better to use the single-level capacity policies, or in other words, set $\mu_{\mathrm{l}} = \mu_{\mathrm{h}} = \mu_{L,\gamma}$.

There can be an additional cost factor for operating a flexible capacity policy. This cost is due to the system that needs to be in place in order to be able to deploy a flexible capacity policy; such a system requires additional human capacity, information and communication systems, as well as training and maintenance. We call this the flexibility system cost (FSC in short). The flexibility system cost can be at most $c_{\mathrm{p}}\mu_{L,\gamma} - \mathrm{ACC}^*$, because otherwise the MTO system would prefer to operate with a single capacity. In the following figure, we show the number of instances (out of 3780 instances, which extends the experiment in Table 5), where using capacity flexibility is still preferred even in the presence of a flexible system cost rate.

From Fig. 7, it can be seen that the capacity flexibility can be affordable up to $0.6 \times c_{\mathrm{p}}$. There are around 200 instances where capacity flexibility is not preferable even if FSC $= 0$. The number of instances that can afford flexibility decreases as FSC increases and finally diminishes when FSC $= 0.6$. Especially, a dramatic decrease occurs in the number of instances when FSC is between 0.2 and 0.4, which is due the fact that most of the possible savings lie in that interval.

6 Conclusion

In this paper, we have studied a production system that operates under a lead-time performance constraint, which guarantees the completion of a job order before a given lead-time with a certain delivery performance target. We assumed that the demand follows a unit Poisson distribution with rate λ. The service time requirement of a job is assumed to follow an exponential distribution. This system is modelled as an $M/M/1$ queuing system and the capacity level of this system corresponds to the service

rate. In a fixed capacity system, the minimum service rate that satisfies a given lead-time performance constraint can be found analytically. We studied flexible capacity systems which adapt their capacity periodically in view of the number of job orders in the system. For the sake of practicality, we focussed on a two-level capacity policy with a permanent capacity which is always deployed, and a contingent capacity, which can be supplied on demand. As the decision on the use of the contingent capacity is taken at the start of each period, the contingent capacity provider does not know in advance whether the reserved capacity will be actually demanded or not. This uncertainty on the use of the contingent capacity creates an opportunity cost and it is reflected on the contingent capacity costs per unit time. The period length T has a smoothing effect on this opportunity cost, since the contingent capacity provider will have much more flexibility to schedule the contingent capacity to other tasks for longer period lengths. Therefore, we modelled the contingent capacity cost per time as a function of the period length, of the cost effect of lost opportunities and of the time-leniency factor of the contingent capacity to switch among different jobs/tasks.

For a given lead-time performance constraint and a permanent/contingent capacity cost structure, the shop floor manager has to decide on the period length T, permanent and contingent capacity levels and the workload level where the contingent capacity should be deployed. This resulting minimization problem cannot be solved with standard optimization techniques. Therefore, we have developed a procedure to create problem-specific sets of values for the decision variables and next developed a layered search method to find the best decision variables in these sets.

We finally conducted a computational study to investigate the behaviour of the optimal period length and the minimum capacity costs under different lead-time performance constraints and different permanent/contingent capacity cost structures.

From the computational study, we observed that, under a periodic two-level capacity policy, the capacity-related costs can be substantially reduced compared with the reference system which uses a fixed capacity to satisfy the lead-time performance constraint. These savings can be particularly high for ambitious lead-time performance constraints. Also, in flexible systems, the optimal period lengths are smaller for more ambitious lead-time performance constraints. Moreover, we observed that under the flexible capacity system, capacity-related costs are quite insensitive to variations in the lead-time performance constraints. However, data from the computational study show that the flexible capacity policy is not always to be preferred over the reference fixed capacity system. First, under prohibitively expensive contingent capacity costs, the capacity-related costs for the flexible system may exceed those in the reference system. Second, even if the total capacity-related cost under the flexible capacity policy is lower than the total cost under fixed capacity, the cost difference may not be sufficient enough to compensate for the additional cost incurred for running the flexible capacity system itself.

In this paper, we have investigated how a production system that faces a lead-time performance constraint imposed by her customers can decrease her costs by employing a periodic flexible capacity policy. An interesting future research question would be how to coordinate the consequences of the production system's capacity decisions with the decision making of the customers on the lead-time performance constraint.

This integration would yield to a decision feedback loop that hopefully iterates to a better, more economical design of the system as a whole.

Acknowledgments The authors would like to thank the anonymous review team for their valuable feedback and suggestions, Dr. Turgut Aykin from ac^2 solutions for insightful discussions and Daniel Hamermesh from UT of Austin for his guidance in Labor Economics.

Appendix A: Analysis of (X(t),Y(t)) process under constant service rate

If we assume that the production system employs a FCFS priority rule, we always have: $0 \leq Y(t) \leq X(t) \leq K$, since the number indicating the position of a tagged order cannot exceed the total number of orders in the system. Also, because of the FCFS policy, $Y(t)$ is non-increasing in t, since the position of the tagged order decreases one by one as the services of the orders before the tagged order are completed. When $Y(t) = 0$, the tagged order's service is literally finished. So, any $(j, 0)$ is an absorbing state of the $(X(t), Y(t))$ process for $0 \leq j \leq K - 1$.

Note that if a tagged order finds $n - 1$ orders in the queue upon its arrival at time t, then $(X(t), Y(t)) = (n, n)$ for all $n > 0$. As discussed earlier, because of the service rate policy π, the service rate can change at the start of each period. Therefore, to analyse the $(X(t), Y(t))$ process under policy π, we first need to characterize the transient behaviour of the same process under a constant service rate of μ.

The state diagram of $(X(t), Y(t))$ under a constant service rate policy with μ can be seen in Fig. 8.

Let Q be the transition rate matrix of the $(X(t), Y(t))$ process under constant service rate μ. Since there are $|Z|$ states in total, Q is a $|Z| \times |Z|$ matrix. Note that state $r = (r_1, r_2)$ means that there are r_1 orders in the system and the tagged order's

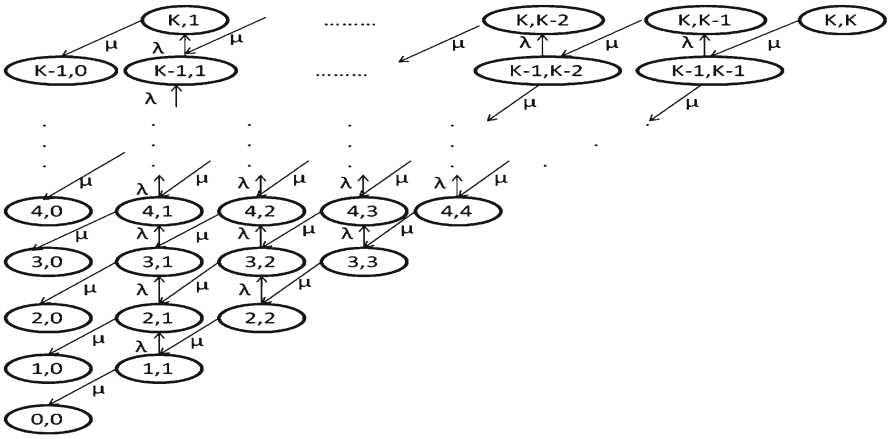

Fig. 8 State diagram of $(X(t),\ Y(t))$

position is r_2. Let $Q_{r,s}$ denote the transition rate from state $r = (r_1, r_2)$ to state $s = (s_1, s_2)$. The $Q_{r,s}$ for the $(X(t), Y(t))$ process under service rate μ is as follows:

$$\begin{aligned} Q_{r,s} &= \lambda \quad \text{if } s_1 = r_1 + 1 \quad \text{for } r_1 = 1, 2, \ldots, K-1 \text{ and } r_2 = s_2 \quad \text{for } 0 < r_2 \le r_1; \\ &= \mu \quad \text{if } s_1 = \quad r_1 - 1 \text{ and } s_2 = r_2 - 1 \quad \text{for } 0 < r_1 \le K, r_2 \le r_1; \\ &= -\sum_{r \neq m} Q_{r,m} \quad \text{if } s = r \quad \text{when } r \text{ is a non-absorbing state;} \\ &= 0 \quad \text{for all other } (r, s) \text{ pairs.} \end{aligned} \tag{A.1}$$

After constructing Q for an arbitrary μ, the transient probability behaviour of the $(X(t), Y(t))$ process can be analysed when $\mu = \mu_1$ and $\mu = \mu_h$. Let $U^l_{r,s}(t)(U^h_{r,s}(t))$ denote the probability that the system will be in state s, $(X(t) = s_1, Y(t) = s_2)$, given that it was in state r in the beginning: $(X(0) = r_1, Y(0) = r_2)$, when the service rate is $\mu_1(\mu_h)$ throughout time t.

We can find $U^l_{r,s}(t)(U^h_{r,s}(t))$ from Q with the help of the uniformization technique (see, e.g. Kulkarni 1999 for details).

References

Ballou RH (1998) Business logistics management. Prentice Hall, Upper Saddle River

Bradley J, Glynn P (2002) Managing capacity and inventory jointly in manufacturing systems. Manag Sci 48:273–288

Chenery HB (1952) Overcapacity and the acceleration principle. Econometrica 20:1–28

Crabill T (1972) Optimal control of a service facility with variable exponential service times and constant arrival rate. Manag Sci 18:560–566

Eberly JC, van Mieghem JA (1997) Multi-factor dynamic investment under uncertainty. J Econ Theory 75:345–387

Ehrenberg RG, Smith RS (1994) Modern labor economics: theory and public policy. HarperCollins, New York

Gans N, Koole G, Mandelbaum A (2003) Telephone Call Centers: tutorial, review and research prospects. Manuf Serv Oper Manag 5(2):79–141

Holt CC, Modigliani F, Muth JF, Simon HA (1960) Planning production, inventories and work force. Prentice-Hall, Englewood Cliffs

Houseman SN (2001) Why employers use flexible staffing arrangements: evidence from an establishment survey. Ind Labor Relat Rev 55:149–170

Kalleberg AL, Reynolds J, Marsden PV (2003) Externalizing employment: flexible staffing arrangements in U.S. organizations. Soc Sci Res 32:525–552

Keskinocak P, Tayur S (2004) Due date management policies. In Simchi-Levi D, Wu SD, Chen ZM (eds) Handbook of quantitative supply chain analysis: modelling in the eBusiness era. Kluwer, Norwell, pp 311–328

Kulkarni VG (1999) Modeling, analysis, design, and control of stochastic systems. Springer, New York

Ledermann W, Reuter GEH (1954) Spectral theory for the differential equations of simple birth and death processes. Phil Trans R Soc Lond A 246:321–369

Lippman SA (1975) Applying a new device in the optimization of exponential queuing systems. Oper Res 23:687–710

Little J (1961) A proof of the theorem $L = \lambda W$. Oper Res 9:383–387

Milner J, Pinker E (2001) Contingent labor contracting under demand and supply uncertainty. Manag Sci 47(8):1046–1062

Mincsovics G, Dellaert NP (2009) Workload-dependent capacity control in production-to-order systems. IIE Trans 41(10):853–865

Pinker EJ (1996) Models of flexible workforce management in uncertain environments. PhD thesis, Massachusetts Institute of Technology

Pinker EJ, Larson RC (2003) Optimizing the use of contingent labor when demand is uncertain. Eur J Oper Res 144(1):39–55

Puterman ML (1994) Markov decision processes-discrete stochastic dynamic programming. Wiley, New York

Rao US, Swaminathan JM, Zhang J (2005) Demand and production management with uniform guaranteed lead time. Prod Oper Manag 14(4):400–412

Rosen S (1986) The theory of equalizing differentials. In: Ashenfelter O, Layard R (eds) Handbook of labor economics, vol 1. North-Holland, pp 641–692

Ross SM (1996) Stochastic processes, 2nd edn. Wiley, New York

Sennott LI (1999) Stochastic dynamic programming and the control of queueing systems. Wiley, New York

Stern TE (1979) Approximations of queue dynamics and their application to adaptive routing in computer communication networks. IEEE Trans Commun 27(9):1331–1335

Tan B (2004) Subcontracting with availability guarantees: production control and capacity decisions. IIE Trans 36:711–724

Tan B, Gershwin S (2004) Production and subcontracting strategies for manufacturers with limited capacity and volatile demand. Ann Oper Res 125:205–232

Tan T, Alp O (2009) An integrated approach to inventory and flexible capacity management subject to fixed costs and non-stationary stochastic demand. OR Spectrum 31:337–360