Workload-dependent capacity control in production-to-order systems

Workload-dependent capacity control in production-to-order systems

G.Z. Mincsovics*, N.P. Dellaert

Department of Technology Management, Technische Universiteit Eindhoven, P.O. Box 513, NL-5600 MB, Eindhoven, The Netherlands, G.Z.Mincsovics@tue.nl, N.P.Dellaert@tue.nl

The development of job intermediation and the increasing use of the Internet allow companies to carry out ever quicker capacity changes. In many cases, capacity can be adapted rapidly to the actual workload, which is especially important in production-to-order systems, where inventory cannot be used as a buffer for demand variation. We introduce a set of Markov chain models to represent workload-dependent capacity control policies. We present two analytical approaches to evaluate the policies' due-date performance based on stationary analysis. One provides an explicit expression of throughput time distribution, the other is a fixed-point iteration method that calculates the moments of the throughput time. We compare due-date performance, capacity, capacity switching, and lost sales costs to select optimal policies. We also give insight into which situations a workload-dependent policy is beneficial to introduce. Our results can be used by manufacturing and service industries when establishing a static policy for dynamic capacity planning.

Key words: capacity planning, workload-dependency, due-date performance, Markov chain, state-dependent service, switching cost

1. Introduction

Most production-to-order companies do not have a constant flow of orders. This often leads to a varying queue of customers. Despite the uncertainty in both arrival and service of orders, customers request tight due dates, and they are also resentful of late deliveries. To compensate for the variations in the orders' frequency, they cannot use inventory buffer, but a good practical solution is to try to match the production rate to the actual workload or, when orders are statistically identical, to the length of the queue. Such a policy may bring benefits in labor cost reduction or in due-date performance. This paper investigates the value of using an optimal workload-dependent policy, where "workload" means the number of orders in the system.

Our motivation for studying workload-dependent capacity planning originates from the area of engineer-to-order (ETO). As a definition for ETO, we quote that of Gelders (1991): "In an

* Corresponding author, Tel: +31 40 2472652, Fax: +31 40 2464531

engineer-to-order environment a company *designs* and *produces* products to customer order." In the same paper, Gelders concludes that ETO companies need "fast-response capacity," which is to be considered as a general characteristic of competitive ETO production.

Similarly to ETO companies, make-to-order (MTO) companies also need to adjust their capacity to meet customer demand (see e.g. Vollmann *et al.* (2005)). Apart from the motivating example for ETO in the paper of Gelders, we give three examples of MTO companies, where workload-dependent capacity management is applicable. In our first example, we describe a general manufacturing situation with an expensive bottleneck machine. In this case, the number of shifts the bottleneck machine works determines the production capacity. The capacity can be set between zero to three shifts, depending on the workload. Our second and third examples are companies in the service industry, which can often face unforeseen variations in the number of orders. The second is a special translator service, where topic leaders share translation tasks and assign them to specialized free-lance translators. The third example is a data recording company, where audio tapes are transcribed to digital text format in order to make later searches possible. While workers listen to the tapes, they type the text into a computer.

In all examples, "fast-response capacity" is not just desirable, but also affordable. Production capacity in manufacturing does not usually require high educational labor; labor acquisition lead times are often very short. Although, in most service industry situations labor has specific knowledge, acquisition lead time substantially decreased in the last decade due to the increasing use of the Internet. In the case of the translator company, contact information of free-lance translators is carefully maintained, so distributing new tasks takes just hours. In the case of workers with high education, capacity flexibility can be gained from using overtime or from workers contracted for changing working time (see e.g. Filho & Marçola (2001)).

In our paper, we assume that we can afford instantaneous capacity changes as an abstraction of opportunity for "fast-response capacity". For our analysis, we consider stationary, homogeneous Poisson arrival process of orders, exponentially distributed service time, and FCFS-type service. The number of employees (servers) is assumed to be a decision variable that depends on the number of jobs in the system. Our objective is to minimize the total average cost per time unit, where the costs consist of labor costs per employee, hiring and firing employees, costs related to order acceptance and early or tardy order completion.

For this problem, we present our model applying two evaluation approaches. One approach aims at exact calculation of the distribution of the throughput time (time spent in the system). Another

approach, which can be used to deal with problems of a larger scale, is based on a moment-approximation of the throughput time. With respect to the effectiveness of the flexible capacity it is essential to increase and decrease capacity at the right moment. We determine the proper switching states for small scale problem instances by searching exhaustively. The optimal strategy is compared in terms of performance with the fixed capacity rules.

Previously, a number of researchers pointed out relations among service rate, work-in-process and due-date performance. One application of due-date performance measures is the evaluation of batch scheduling rules. In the scheduling rules, capacity level and work-in-process are both consistently present as parameters showing the strong relation to the due-date performance (see e.g. Philipoom *et al.* (1993)). Lead time setting is a topic where work-in-process is also recognized as an adequate parameter, which improves the performance of the rules (see e.g. Bertrand (1983)). These two examples reveal the connection between capacity level, work-in-process and due-date performance. However, there is no model in the literature that incorporates all these three closely related notions.

We think that our paper makes a valuable contribution by being the first to study a queuing system with adaptive capacity to satisfy the objective of having tight and accurate due-dates. Moreover, our examples and their analysis provide insight into when a workload-dependent strategy is effective, and also what characteristics this strategy has.

The paper is organized as follows. In the next section, we describe the related literature. Then, in section 3, we give a mathematical formulation of our problem. Results on the effectiveness and characteristics of the strategy are shown in section 4. Finally, conclusions and plans for future extensions are presented.

2. Related literature

Decisions on capacity changes were studied first in capacity expansion problems. In the case of deterministic demand with positive trend, Chenery (1952) found that gas pipelines permanently have extra capacity. This was the basis of his "excess capacity hypothesis" that says capacity is always larger than demand; optimal overcapacity is to be investigated by looking at economies of scale. Manne (1961) revised Chenery's hypothesis when extending his model. The extension of the model with a backlog option created an environment in which the "excess capacity hypothesis" no longer held. Manne also studied stochastic, stationary demand without the backlog option. This model resulted in a smaller deviation from what Chenery's model suggested. Luss (1982) gave a comprehensive review of the literature on capacity expansion.

Models with capacity expansion/reduction decisions and hiring/firing costs are usually studied by means of dynamic programming. One example is the continuous time DP model of Bentolila & Bertola (1990), where sensitivity analysis on firing costs was presented. Rocklin *et al.* (1984) studied a production-to-order environment with non-stationary demand, including both capacity expansion/reduction decisions and hiring/firing costs. Rocklin et al. showed the optimality of the (S', S'')-policy known from inventory theory by the means of discrete time DP. In their model, demand must always be met; if demand exceeds available capacity, the capacity must be increased to overcome the deficit.

Pinder (1995) deduced an approximation of optimal workload-dependent capacity control policy for stationary demand. In the model of Pinder, capacity (resources) is treated as discrete; capacity adjustments are dependent on the actual number of jobs (work), which are particular points in common with our paper. However, the workload-dependent policy class defined by Pinder seemed too broad to give an explicit formula of policy evaluation or to find an optimal solution. In addition, Pinder did not consider due-date performance in any form, which is an essential part of our model. Besides, we define a less broad policy class that entails less limitations in the analysis so that we can provide additional insights.

Queuing models have made use of servers with load-dependent service times for approximate performance analysis since the fundamental work of Avi-Itzhak & Heyman (1973). These servers are apt to represent a sub-network as they can be set to provide nearly the same characteristics. Marie (1979) introduced a similar approximation technique. A comparison of the two techniques is given in Baynat & Dallery (1993).

In some simple workload-dependent policy classes, the optimal (capacity) control policy can be analytically determined. Faddy (1974) introduced the class of P_λ^M policies for the control of water reservoirs. Namely, the policy P_λ^M sets the output rate to M if the water level in the reservoir reaches (or exceeds) λ, and the output rate is set to zero if the reservoir is empty. This policy is still a subject of research (see Kim *et al.* (2006)). Another simple policy class, the two-step service rule, was defined in Bekker & Boxma (2005). A policy of this class has a lower service rate, r_1, when the workload is not more than K, and a higher service rate, r_2, if the limit K is exceeded. This policy class has the drawback that the undesired frequent service rate changes are not excluded. Tijms & van der Duyn Schouten (1978) proposed a different class of policies for inventory control, called two switch-over level rule. A rule is characterized by two inventory levels, $y_2 \leq y_1$. An inventory decrease to y_2/increase to y_1 triggers compensating raise/reduction. This design restrains frequent service rate changes. In our paper, we specify a class of control policies for managing capacity with

a higher degree of freedom than the three simple classes discussed. As a result, finding the optimal solution in our policy class allows a better characterization of how to control capacity optimally.

3. Model formulation

In the context of this paper, a workload-dependent capacity planning policy is defined by two "switching points", one down and one up switching point, for each pair of neighboring capacity levels. Switching points are specific workload values, for which a job arrival or departure can trigger a switch in capacity. If the system is at an *up switching point*, and a job arrives, we switch from the lower capacity level to the upper. If the system is at a *down switching point*, and a job departs, we switch from the upper capacity level to the lower. Each up-switch incurs a hiring cost (c_{hiring}), and each down-switch incurs a firing cost (c_{firing}). The firm has an admissible domain of capacity levels; that is the set of integer values between $C_{\min}$ and $C_{\max}$. Naturally, the firm also needs to pay for the used capacity. The cost of having capacity level c is given by $c \cdot c_{capacity}$ per time unit.

Although this paper does not investigate the impact of efficiency aspects, one may also take into account the psychological effect of lower or higher workload Bertrand & van Ooijen (2002), and the efficiency of using different levels of capacity Schlichter (2005). Practically, the service rate can be measured for each workload and capacity level ($\mu_{w,c}$) and used in our model.

The firm works with a fixed lead time (L). It has to pay charges for each time unit when a job is late ($c_{tardiness}$). A smaller cost is due for holding if a job is ready before the due date ($c_{earliness}$). The number of jobs with which the firm can deal is constrained from above by a constant integer, $W_{\max}$. New jobs are refused if the firm already has $W_{\max}$ number of jobs. Consequently, lost sales can occur that incurs a virtual expense of $c_{lostsales}$ for each occasion. To sum up, we have a six element vector of cost coefficients ($c_{capacity}, c_{hiring}, c_{firing}, c_{lostsales}, c_{earliness}, c_{tardiness}$). Although we use linear cost components for the ease of presentation, employing general cost functions would not change our analysis.

In this section, we define a workload-dependent capacity planning policy class, and formalize the firm's capacity control problem.

3.1. Definitions, assumptions and problem formulation

We assume a stationary environment in which both arrival and service processes are homogeneous Poisson processes. Arrivals have a constant rate of λ; departures have, in general, capacity level and workload-dependent rates. We denote by $\mu_{w,c}$ the service rates for workload w and capacity level c. For the sake of better comprehension, we assume $\mu_{w,c} = c\mu$ for all w and c, which correspond

to assuming identical servers (machines, workers or shifts). Jobs are served according to FCFS discipline.

We define the set of workload-dependent capacity control policies for given values of $C_{\min}$, $C_{\max}$, and $W_{\max}$ as $\Omega(C_{\min}, C_{\max}, W_{\max})$. A feasible policy can be characterized by a triple $\pi = (\Gamma_{\min}, \Gamma_{\max}, \Theta)$. The terms $\Gamma_{\min}$, and $\Gamma_{\max}$ stand for the minimum, and maximum capacity levels which are used by the policy. Necessarily, these terms have to satisfy the inequalities, $C_{\min} \leq \Gamma_{\min} \leq \Gamma_{\max} \leq C_{\max}$. The term Θ denotes a $(\Gamma_{\max} - \Gamma_{\min})$-by-2 matrix, which describes all the switching points. For the cases when $\Gamma_{\min} = \Gamma_{\max}$, we have an empty matrix. We use indices $c \to c+1$ or $c+1 \to c$ with $c \in \{\Gamma_{\min}, \ldots, \Gamma_{\max} - 1\}$ to show that the switch occurs between the capacity levels c and $c+1$ either up or down. Particularly, $\Theta_{c \to c+1}$, and $\Theta_{c+1 \to c}$ are workloads, where the capacity is changed from c to $c+1$, upwards, or from $c+1$ to c, downwards, if a job arrival or departure occurs, respectively. The elements of Θ are constrained by $W_{\max}$, and by each other as follows,

- $1 \leq \Theta_{\Gamma_{\min}+1 \to \Gamma_{\min}}$ and $\Theta_{\Gamma_{\max}-1 \to \Gamma_{\max}} \leq W_{\max} - 1$
- $\Theta_{c+1 \to c} \leq \Theta_{c \to c+1} + 1$ for all $c \in \{\Gamma_{\min}, \ldots, (\Gamma_{\max} - 1)\}$
- $\Theta_{c+1 \to c} \leq \Theta_{c+2 \to c+1}$ and $\Theta_{c \to c+1} \leq \Theta_{c+1 \to c+2}$ for all $c \in \{\Gamma_{\min}, \ldots, (\Gamma_{\max} - 2)\}$

In Figure 1, we show two workload-dependent capacity control policies, $\pi = (1, 3, \begin{pmatrix} 3 & 1 \\ 4 & 2 \end{pmatrix})$ and $\pi = (1, 3, \begin{pmatrix} 3 & 3 \\ 4 & 5 \end{pmatrix})$ in the set $\Omega(0, 3, 6)$. Other feasible policies are e.g. $\pi = (1, 2, \begin{pmatrix} 3 & 1 \end{pmatrix})$, which uses only capacity levels one and two, or $\pi = (1, 3, \begin{pmatrix} 3 & 4 \\ 4 & 5 \end{pmatrix})$, when capacity becomes a function of the workload.

We assume that capacity adjustments can be done instantaneously, in parallel with the change in workload.

Within the defined set of workload-dependent capacity control policies, we define the subset of *fixed policies*, as having the switching point matrix, Θ, empty. Next, we define the set of *continuous fixed policies* as policies using a fixed capacity, which can be set to any real values between $C_{\min}$ and $C_{\max}$. The set of fixed policies is the intersection of the set of continuous fixed policies and the set of workload-dependent policies.

In order to simplify the exposition, we do not separate costs of hiring to firing, but use a single

cost coefficient, $c_{switching}$ that penalizes capacity changes in general. Note that as stationarity implies balance of hiring and firing, we can aggregate the cost coefficients c_{hiring} and c_{firing} by taking $c_{switching} = (c_{hiring} + c_{firing})/2$. Based on this formula, cost counting with $c_{switching}$ provides the same result as cost counting with c_{hiring} and c_{firing}.

For the cost coefficients mentioned in the model description, we have the related costs. These costs can be separated into two groups. The costs of capacity, switching, and the lost sales penalty are functions of the policy only. The costs of earliness and tardiness are functions of the lead time, additionally.

Now, we are able to formulate the capacity control problem as

$$\min_{\pi \in \Omega(C_{\min}, C_{\max}, W_{\max})} cost_{group1}(\pi) + cost_{group2}(\pi, L) \tag{1}$$

where L is the fixed lead time, $cost_{group1}(\pi)$ is the sum of $cost_{capacity}(\pi)$, $cost_{switching}(\pi)$, $cost_{lostsales}(\pi)$, and $cost_{group2}(\pi, L)$ is the sum of $cost_{earliness}(\pi, L)$, and $cost_{tardiness}(\pi, L)$. E.g. the first policy (left) in Figure 1, we can expect to outperform the second policy (right) in all the costs but $cost_{capacity}$.

3.2. Evaluation of cost functions

In this section, we aim to express the costs one by one. First, we derive the costs that are independent of the quoted lead time ($cost_{group1}$). After that, we show two approaches to evaluate the lead time dependent costs ($cost_{group2}$). We note that in real-life situations, where the cost functions need to be evaluated for general arrival and service processes, our Poisson process based approaches can help finding the optimal policy by giving a starting guess, however simulation is necessary for tuning the policy parameters afterwards.

We create a Markov chain, $MC^{\pi,\lambda,\mu}$, according to the policy, the given arrival rate, and the given service rate unit. In this Markov chain, we have the states labeled by (w, c), the workload (number of jobs in the system), and the capacity usage. A policy π can be given by the arrival matrix (A^π), and the departure matrix (D^π). These matrices are square; rows and columns are indexed by the states of the Markov chain. Elements are ones at the related arrival or departure arcs of the Markov chain and zeros elsewhere. The state space of $MC^{\pi,\lambda,\mu}$ depends on the policy. As an example, we show the transition rate diagram of $MC^{\pi,\lambda,\mu}$ in Figure 1 for the policies $\pi = (1, 3, \begin{pmatrix} 3 & 1 \\ 4 & 2 \end{pmatrix})$ and

$\pi = (1, 3, \begin{pmatrix} 3 & 3 \\ 4 & 5 \end{pmatrix})$. Note that the previously defined policy class, $\Omega(C_{\min}, C_{\max}, W_{\max})$, for $C_{\min} = 0$ contains policies, which correspond to $M|M|n|W_{\max}$ queuing systems for any $n = 0, 1, 2, \ldots, C_{\max}$.

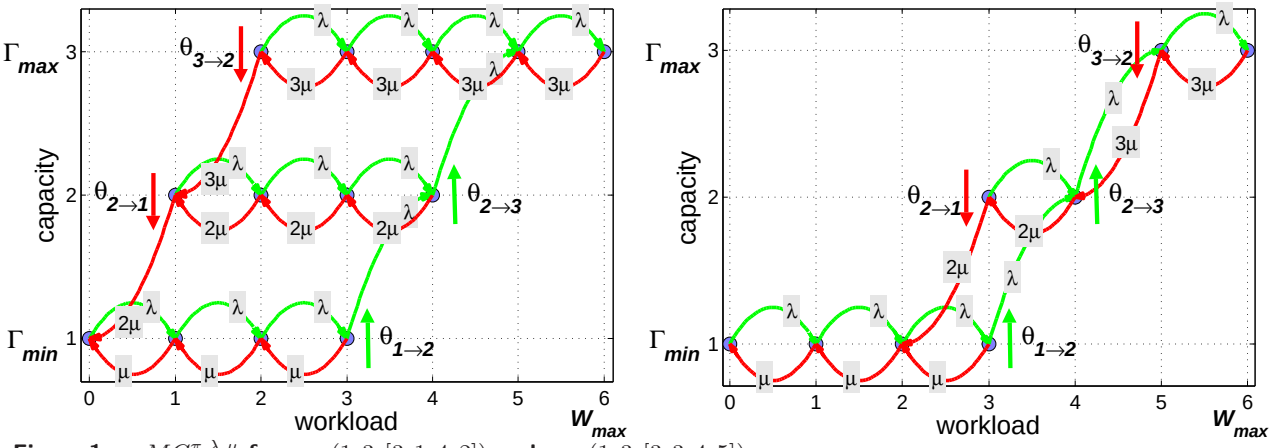

Figure 1 $MC^{\pi,\lambda,\mu}$ **for** $\pi = (1, 3, [3, 1; 4, 2])$ **and** $\pi = (1, 3, [3, 3; 4, 5])$

Via steady state analysis, the first cost group can be easily evaluated. If we add λA^{π} to μD^{π} weighted in rows (w, c) by capacity c, we can obtain Q^{π}, the transition rate matrix of the Markov chain.

$$Q^{\pi}_{(w_1,c_1),(w_2,c_2)} = \lambda A^{\pi}_{(w_1,c_1),(w_2,c_2)} + c_1 \mu D^{\pi}_{(w_1,c_1),(w_2,c_2)} \tag{2}$$

We define the diagonal elements of Q^{π} as the negative state leaving rates: $v^{\pi}_i := diag(Q^{\pi})_i = -\sum_{j \neq i} Q^{\pi}_{i,j}$. Solving the linear equation system $\begin{cases} Q^{\pi} T^{\pi} = \mathbf{0} \\ e T^{\pi} = 1 \end{cases}$, where e is the all one vector, we have $T^{\pi}_{w,c}$, the distribution of time being in a state (w, c). Eventually, we can specify all the costs of the first group.

$$\begin{aligned} cost_{capacity}(\pi) &= c_{capacity} \sum_c c \sum_w T^{\pi}_{w,c} \\ cost_{switching}(\pi) &= 2 c_{switching} \lambda \sum_{(w,c) \in U} T^{\pi}_{w,c} \\ cost_{lostsales}(\pi) &= c_{lostsales} \lambda P(lost\ sale) = c_{lostsales} \lambda T^{\pi}_{W_{\max}, \Gamma_{\max}} \end{aligned} \tag{3}$$

where $P(lost\ sale)$ is the probability that a job is rejected, and U is the set of states from which the up-switches in capacity are made. Note that it is enough to count costs for the up-switches twice because of the stationarity assumption.

In what follows, we deduce the earliness and tardiness costs in two different ways. Both approaches aim at identifying the throughput time (or sojourn time) as a random variable so that the due-date performance of the policies can be determined.

3.2.1. Derivation of the throughput time distribution

In order to derive an explicit formula for the throughput time distribution, we observe an arbitrarily selected job while it is in the system. In addition to the system's workload and the used capacity, we observe the queue position of the pointed job. Therefore, we extend our Markov chain $MC^{\pi,\lambda,\mu}$ to an extended Markov chain $EMC^{\pi,\lambda,\mu}$, which has the queue position of the pointed job (q) as an extra dimension. $EMC^{\pi,\lambda,\mu}$ has states labeled by (w,c,q), the workload, capacity, and queue position of the pointed job, respectively, with $\Gamma_{\min} \leq c \leq \Gamma_{\max}$, and $0 \leq q \leq w \leq W_{\max}$. When the pointed job arrives, it is in one of the states, for which $0 < q = w$ hold; its lifetime ends with $q = 0$. Because of the FCFS discipline, during the lifetime of the job, the queue position of the pointed job, q decreases when a job is ready, whereas it is unaffected by arrivals.

We want to obtain the probability that starting from state r at time 0 we will be in state s at time t in $EMC^{\pi,\lambda,\mu}$. This probability we denote by $\bar{P}^{\pi}_{r,s}(t)$. In the appendix, we show an example for $EMC^{\pi,\lambda,\mu}$ and derive an explicit expression for $\bar{P}^{\pi}_{r,s}(t)$ by applying uniformization.

Since the states $(w,c,0)$, where the pointed job can exit the system, are absorbing states, we can express the throughput time CDF as the sum of probabilities of getting to certain exit states as follows.

$$F^{r,\pi}(t) = \sum_{s=(w,c,0)} \bar{P}^{\pi}_{r,s}(t) \tag{4}$$

is the throughput time CDF if the pointed job arrives when the system is in state $r = (w_r, c_r, q_r)$ with $q_r = w_r$.

The stationary distribution (T^{π} for $MC^{\pi,\lambda,\mu}$) is different from the distribution right after the arrival of the pointed job, which we denote by T^{π}_{A}. We can use the arrival matrix (A^{π}) to determine the probability distribution of being in a state right after the arrival $T^{\pi}_{A} = T^{\pi}A^{\pi}$. We can express the after arrival distribution of the extended Markov chain, $EMC^{\pi,\lambda,\mu}$, which we call $\bar{T}^{\pi}_{A}$, from the after arrival distribution of $MC^{\pi,\lambda,\mu}$.

$$\bar{T}^{\pi}_{A,(w,c,q)} = \begin{cases} T^{\pi}_{A,(w,c)} \text{ , if } q = w \\ 0 \qquad\quad \text{ , if } q \neq w \end{cases} \tag{5}$$

Once we have the extended starting distribution $\bar{T}^{\pi}_{A}$, the throughput time CDF (F^{π}) can be expressed.

$$F^{\pi}(t) = \sum_{\substack{r=(w_r,c_r,q_r) \\ q_r=w_r}} \bar{T}^{\pi}_{A,r} F^{r,\pi}(t) \tag{6}$$

Finally we can give formulas for the lead time dependent cost functions.

$$\begin{aligned} cost_{earliness}(\pi, L) &= c_{earliness}\lambda(1 - P(lost\ sale)) \int_0^L (L-t)\ dF^{\pi}(t) \\ cost_{tardiness}(\pi, L) &= c_{tardiness}\lambda(1 - P(lost\ sale)) \int_L^{\infty} (t-L)\ dF^{\pi}(t) \end{aligned} \tag{7}$$

3.2.2. Moment approximation of the throughput time

Apart from the distribution function, we also derive a moment approximation for the throughput time as an alternative for the steps from applying uniformization to the step of expressing $F(t)$ in (6). This approach allows evaluation of large scale problems, as it both speeds up the calculations and needs less memory. Besides, throughput time moments can be determined with higher accuracy than extracting them from the throughput time distribution.

We evaluate the moments based on the equation that describes the relation of the conditional expected throughput times. We denote the kth moment of the throughput time if starting from state r by $E\left[(X_r)^k\right]$ and the duration of the visit to state r by Z_r. Then we have

$$E\left[(X_r)^k\right] = \sum_s \bar{P}^{\pi}_{r,s}\ E\left[(Z_r + X_s)^k\right] \tag{8}$$

where $\bar{P}^{\pi}$ is the extended transition probability matrix, containing the probabilities of going from state r to some neighboring state s. Similar to the approach in the previous subsection, a state is described by the components (w, c, q) and therefore the definition of $\bar{P}^{\pi}$ in (8) is identical to the one in the appendix, equation (15).

For small state spaces, we can solve this linear system by matrix inversion, but for larger state spaces, we need a vector iteration to determine the moments for the relevant states. This vector iteration can be established by indexing the equation by the iterator n, and declaring the starting state.

$$\begin{cases} E_{n+1}\left[(X_r)^k\right] = \sum_s \bar{P}^{\pi}_{r,s}\ E_n\left[(Z_r + X_s)^k\right] \\ E_0\left[(X_r)^k\right] = 0 \end{cases} \tag{9}$$

Using the independence of conditional throughput time and visit duration, we can write

$$E\left[(Z_r + X_s)^k\right] = \sum_{j=0}^{k} \binom{k}{j} E\left[(Z_r)^{k-j}(X_s)^j\right] = \sum_{j=0}^{k} \binom{k}{j} E\left[(Z_r)^{k-j}\right] E\left[(X_s)^j\right] \tag{10}$$

We can notice in expression (10) that for the evaluation of higher moments all the previous ones are needed in the iteration, (9). In the general form, we evaluate the first K moments.

We can use the following algorithm. We increase the moment iterator k from 1 to K. For each k we take limit in ∞ for n with the vector iteration to evaluate the moments $E\left[(X_r)^k\right]$ one after

the other, inductively. Similarly to (6), we weight the conditional moments by the after arrival distribution of the extended Markov chain, $EMC^{\pi,\lambda,\mu}$ (that is $\bar{T}_A^{\pi}$), which gives the kth moment of the throughput time, $E[X^k]$.

$$E[X^k] = \sum_{\substack{r=(w_r,c_r,q_r) \\ q_r=w_r}} \bar{T}_{A,r}^{\pi} E\left[(X_r)^k\right] \tag{11}$$

In practice, we can evaluate the equation for the first two moments, $E[X]$ and $E[X^2]$, and fit a suitable distribution, e.g. in the class of gamma distributions to approximate the throughput time CDF, $F^{\pi}(t)$. Eventually, cost of earliness and tardiness can be found using equations labeled by (7).

4. Results

We try to achieve two goals with our numerical experiments. First, we would like to show in which situations workload-dependent capacity planning is worth using. We study the effect of setting high/low switching cost coefficient as well as the different settings of lead time, and arrival rates for a fixed service rate. Second, we would like to characterize the workload-dependent policies as compared to the fixed capacity policies, the policies that can use one capacity level only. We illustrate the practical use of workload-dependent policies in the end of the section.

Our model has its limitations; finding the optimal policy using our exact, throughput time distribution evaluation approach is not possible in acceptable time for systems that can either use many capacity levels, or handle high number of orders economically. In these cases the number of workload-dependent policies and their state space increases to large values, so we need to use the approximation approach. Table 1, which we calculated via complete enumeration, gives an indication how fast the number of policies increases. Furthermore, we found that the maximum number of states in MC for $C_{min} = 0$ is $(W_{max} - C_{max} + 1)(C_{max} + 1)$, and in EMC it is $(x(x+1)/2 - 1)y + x(y(y-1)/2)$, where $y = C_{max} - C_{min} + 1$ and $x = W_{max} - (y-1) + 1$. E.g. when $C_{\min} = 0$, $C_{\max} = 6$ and $W_{\max} = 10$, the number of policies is 16844, the largest state space by the due-date performance evaluation has 203 states.

As an initial setting, we take $C_{\min} = 0$, $C_{\max} = 3$ and $W_{\max} = 6$ (number of policies is 288, maximum number of states is 57) for which we use the throughput time distribution evaluation approach. We study large scale settings in subsection 4.2, and 4.5, where we use the moment approximation approach.

We assume a service rate of $\mu = 0.04$ everywhere, but vary interarrival rate (λ) taking values from 0.01 to 0.12 by a step size of 0.01. We study lead time (L) values from the interval 0 to 180

$C_{\max} \setminus W_{\max}$	1	2	3	4	5	6	7	8	9	10	11
0	1	1	1	1	1	1	1	1	1	1	1
1	3	5	8	12	17	23	30	38	47	57	68
2	5	10	21	40	69	110	165	236	325	434	565
3	7	15	35	78	157	288	490	785	1198	1757	2493
4	9	20	49	117	260	531	1005	1783	2996	4809	7425
5	11	25	63	156	364	795	1626	3132	5719	9962	16648
6	13	30	77	195	468	1060	2275	4642	9037	16844	30162

Table 1 **Number of workload-dependent policies for** $C_{\min} = 0$

by a step size of 10. In our cost coefficient test-bed the capacity and switching cost coefficients have a pointed role. First, we fix the capacity cost coefficient to 100, to normalize the test-bed. We observe the model's behavior for different switching cost coefficients, interarrival rates, and lead time settings. For each of the remaining cost parameters we examined three values. To the coefficient, $c_{lostsales}$ 3000, 4000, and 5000 are assigned. For $c_{earliness}$ we take 1, 2, and 5. Finally, $c_{tardiness}$ has values 10, 25, and 100, respectively.

4.1. Value of workload-dependent capacity control

We investigate the value of workload-dependency in capacity control in different environments. We compare workload-dependent capacity policies with fixed capacity policies. The value of workload-dependency is defined as the expected total cost ratio of the optimal fixed capacity and the optimal workload-dependent policy. As fixed capacity policies form a subset in the set of workload-dependent capacity policies, this ratio is never less than one. Therefore, we consider only the cost excess percentage ($CE\%$) due to not using the optimal workload-dependent policy. For example, a $CE\%$ value of 5 means that if a firm uses the optimal fixed policy, it has to pay 105% compared to our reference point, the optimal workload-dependent policy as being 100%. We conduct numerical experiments to circumscribe the cases where high $CE\%$ values are to be expected.

In Figure 2, we depict contour lines of $CE\%$ value functions for $c_{switching}$ values 1000 and 3000.[1] Contour lines are drawn at the $CE\%$ values 1, 6, and 20. As an example for interpreting Figure 2, we take the $CE\%$ function for $c_{switching} = 3000$, which correspond to the solid contour lines. E.g. we look at the $CE\%$ function values above the line $L = 20$: for $\lambda = 0.08 \ldots 0.12$ the value is below 1, for $\lambda = 0.04$, 0.05 or 0.07 the value is between 1 and 6, for $\lambda = 0.02$, 0.03 or 0.06 the value is between 6 and 20, for $\lambda = 0.01$ the value is above 20. The cost coefficients, $c_{lostsales}$, $c_{earliness}$, and $c_{tardiness}$ take the values 5000, 1, and 100, respectively. In what follows, we discuss the major

[1] This type of figure is called contour map, first used in cartography. It shows level sets of a function, which has two arguments. Here, we plot two $CE\%$ functions for $c_{switching}$=1000 and 3000, respectively, with the arguments, interarrival rate and lead time

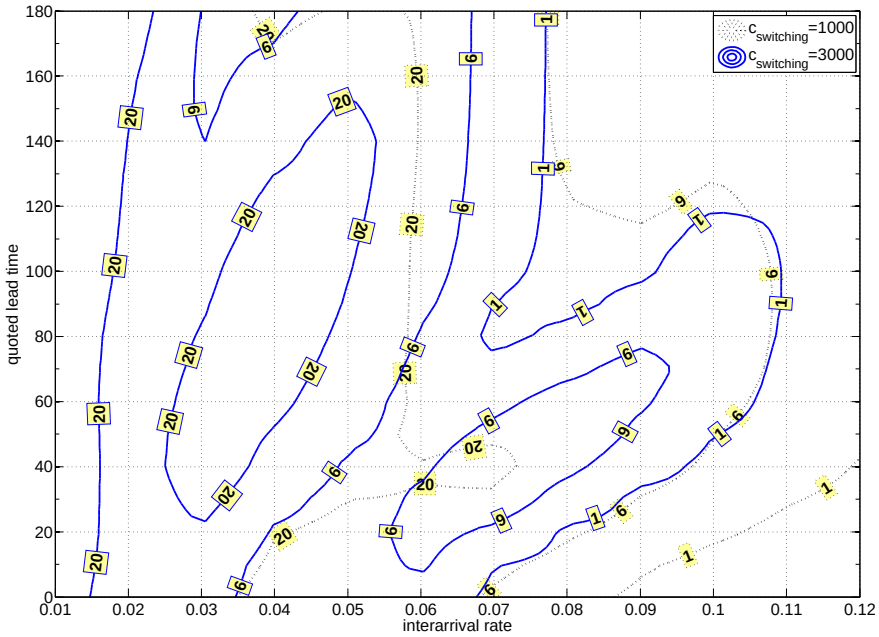

Figure 2 $CE\%$ **values for** $C_{\min} = 0$, $C_{\max} = 3$, $W_{\max} = 6$, **and** $\mu = 0.04$

properties of the $CE\%$ value that can be observed in this figure. The $3^3 = 27$ combinations of the cost coefficients $c_{lostsales}$, $c_{earliness}$, and $c_{tardiness}$ that we checked consistently support these properties and its reasonings, however they do not guarantee the properties to hold for different cost coefficient settings.

Property 1. When the switching cost coefficient decreases, $CE\%$ value increases

Since fixed capacity policies use a constant capacity level, they do not incur switching cost as the rest of the policies. Therefore, the lower the switching cost coefficient, the higher the $CE\%$ value.

Property 2. When the quoted lead time increases, $CE\%$ value decreases

Long quoted lead times, or equivalently, loose due dates penalize less the long waiting times. Long waiting times have a utilization smoothing effect similar to workload-dependent capacity flexibility. This means that when lead time (L) values are high enough, $CE\%$ values are low.

Property 3. When arrival rate increases, $CE\%$ value decreases

As interarrival rate (λ) increases, we use correspondingly more capacity. $\Gamma_{\min}$ also increases for the optimal policy. This way, there is less and less room for workload-dependent policies to use different capacity levels. This results in a decreasing trend in the $CE\%$ value.

Property 4. The effect of capacity discreteness: cuts in the $CE\%$ value

Independently of the switching cost coefficient, one can observe some regions, where the $CE\%$ value drops. The reason is the discreteness of the capacity. It would make a difference if we could set

an optimal fixed capacity level on a continuous basis. We plot the integer contours of the optimal continuous capacity level in Figure 3 keeping the contour lines of Figure 2. There are regions where the optimal fixed policies get close to the continuous fixed optimum, and regions where they are distant. Where the continuous fixed optimum is close to an integer, the (discrete) fixed capacity policies perform reasonably well, while where the continuous fixed optimum is far from an integer level, fixed capacity policies perform poor. As the workload-dependent policies are less affected by the discreteness of the capacity, the $CE\%$ value decreases around the integer contours of the optimal continuous fixed capacity.

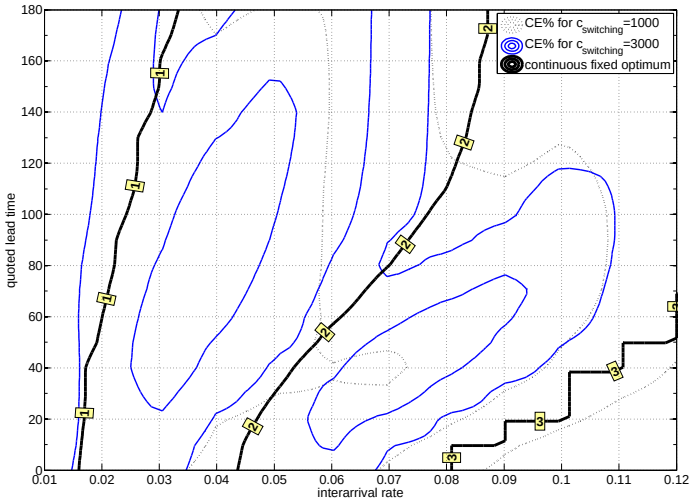

Figure 3 **Continuous fixed optimum for** $C_{\min} = 0$**,** $C_{\max} = 3$**,** $W_{\max} = 6$ **, and** $\mu = 0.04$

Property 5. Steep increase in $CE\%$ value for small interarrival rate for $C_{\min} = 0$

To the left from the continuous fixed capacity contour at the value of one in Figure 3, the $CE\%$ value steeply increases when the interarrival rate gets smaller. Fixed policies cannot adapt to low interarrival rates, as the fixed policy with a fixed zero capacity level is highly uneconomical, whereas the workload-dependent policy can make use of the zero capacity level for low workload values. As a result, the $CE\%$ value is high for low interarrival rates, when $C_{\min} = 0$.

Our further experiments showed that Property 5 does not generally hold. In particular, we need to differentiate between closable systems, which we define as $C_{\min} = 0$, and non-closable systems having $C_{\min} > 0$. We adjust Property 5 for the case of non-closable systems. Figure 4 depicts what happens if we change $C_{\min}$ in the parameter setting of Figure 2 from zero to one.

Property 5.* Low $CE\%$ value for small interarrival rates and $C_{\min} > 0$

If we consider a non-closable system with $C_{\min} = 1$, then the constant one fixed capacity is optimal among both the fixed and the workload-dependent policies as the interarrival rate tend to zero that results in $CE\% = 0$.

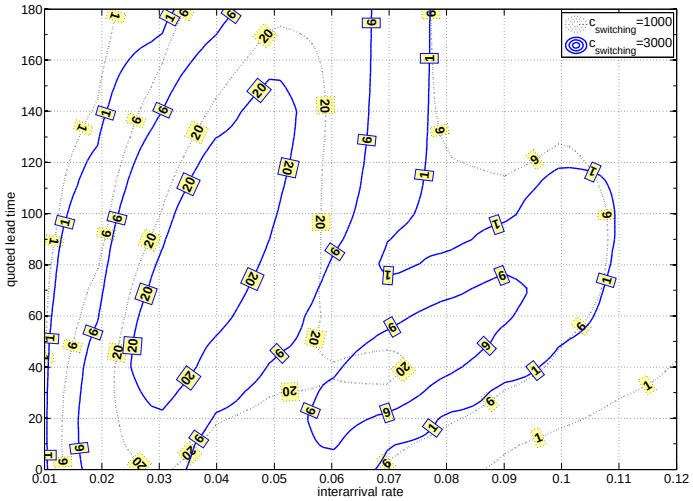

Figure 4 $CE\%$ **values for** $C_{\min}=1$, $C_{\max}=3$, $W_{\max}=6$ **, and** $\mu=0.04$

We also performed sensitivity analysis in $W_{\max}$ and $C_{\max}$. We observed an increase of the $CE\%$ value in both cases. The reason is that if we increase the $W_{\max}$ or $C_{\max}$ value by one, the number of fixed capacity policies remains the same or increases only by one, respectively, whereas the increase in the number of workload-dependent capacity policies is very large in general. The increase in $W_{\max}$ induces an overall increase in $CE\%$ values, while the increase of $C_{\max}$ from $C_{\max}^{old}=3$ to $C_{\max}^{new}=4$ affects the graph of $CE\%$ values only at λ values above around $\mu C_{\max}^{old}$.

4.2. Characterization of workload-dependent policies for uncapacitated, large scale settings

An uncapacitated situation means that one can always hire the necessary capacity. That capacity is not limited, can be expressed either by $C_{\max}=W_{\max}$ or $C_{\max}=\infty$. We study the uncapacitated setting with $C_{\min}=0$, $\lambda=0.10$, $\mu=0.04$, $L=30$, and cost coefficients, $(c_{capacity}, c_{switching}, c_{lostsales}, c_{earliness}, c_{tardiness})=(100,1000,3000,1,10)$, and observe changes of the optimal workload-dependent policy for an increasing $W_{\max}$ value. Our findings hold for all the other cost coefficient combinations from our test-bed.

We observed that optimal switching points of the first levels tend to change less and less, when $W_{\max}$ increases. In Figure 5, the first six pair of switching points can be seen, in which optimal policies did not differ for large enough $W_{\max}$ values ($14\leq W_{\max}\leq 40$). We can see that up- (to the right) and down-switching points (to the left) limit the attainable states in a closely linear manner.

We also found that the capacity/lost sales cost component is monotone increasing/decreasing in $W_{\max}$ for the optimal policy. Moreover, the simultaneous stagnation of these two costs entails the stagnation of the total costs. Figure 6 shows that

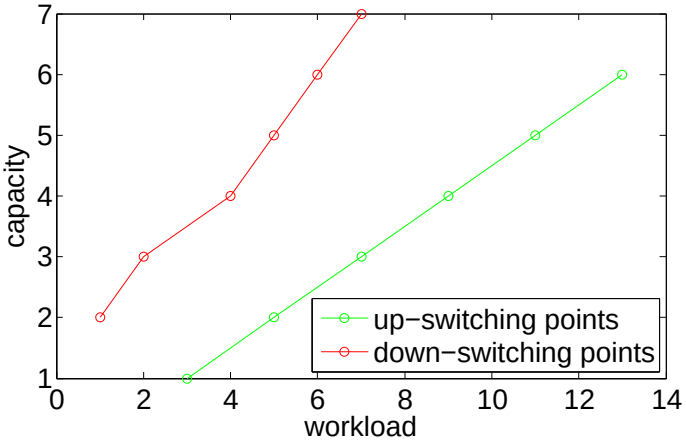

Figure 5 **First six pair switching points of optimal policies for** $14 \leq W_{\max} \leq 40$

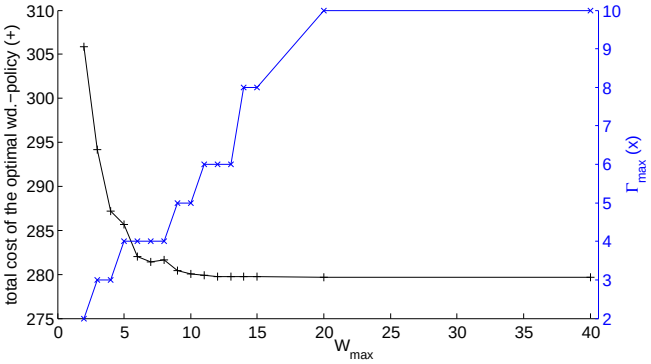

Figure 6 **Total cost and** $\Gamma_{\max}$ **of optimal workload-dependent policies for changing** $W_{\max}$

- the total cost is decreasing in $W_{\max}$ up to a certain value (about $W_{\max} = 6$); thereafter it stagnates as the visiting probabilities of the high workload levels become very small, and it converges in a not necessarily decreasing manner. As a consequence, it is safer to allow high workload values in the uncapacitated case.
- the number of used capacity levels shows a monotone increasing, concave shape. $\Gamma_{\min}$ is unaffected; here it is always one.

Under the assumption that switching points do not change for increasing W_{max} after a point, we can prove the lost sales being monotone decreasing from that point on for any reasonable arrival rate. We aggregate the MC into one state, leaving out only the state in the corner, corresponding to the highest capacity and workload level, which state we denote by s. If we increase W_{max} by one at some higher values of W_{max} (so that the switching points do not change any more), we get one new state, which we call n. We use $u := \Gamma_{max}\mu$ in order to shorten the formulas. If we compare the original Markov chain (without n) with the new one (with n), we get that the steady state probability of s in the original as $A := \frac{\lambda}{\lambda+u}$, and that of n in the new as $B := \frac{\left(1-\frac{\lambda}{u}\right)\left(1-\frac{\lambda}{u}\right)^2}{1-\left(\frac{\lambda}{u}\right)^3}$. It comes

straightforward that $A > B$ if and only if $u > \lambda$. We think that with some similar state aggregations the capacity cost component can be also handled, although it is more complicated.

4.3. Characteristics of workload-dependent policies vs. fixed policies

We try to answer the question of why workload-dependent policies outperform the fixed policies. Workload-dependent policies often compensate for switching costs by using less capacity, avoiding lost sales, and providing a throughput that gives less earliness-tardiness costs. The primary cost savings these policies achieve are normally in tardiness and lost sales cost. In an example shown in Table 2, the optimal workload-dependent policy beats the fixed optimum in all the cost components except $cost_{switching}$.

parameter	λ	μ	L	$c_{capacity}$	$c_{switching}$	$c_{lostsales}$	$c_{earliness}$	$c_{tardiness}$
value	0.07	0.04	30	100	1000	4000	2	25

policy class	optimal π	$CE\%$	$cost_{capacity}$	$cost_{switching}$	$cost_{lostsales}$	$cost_{earliness}$	$cost_{tardiness}$
workload-dep.	(1,3,[3,1;4,2])	0	182.0	18.7	12.6	0.7	18.1
fixed	(2,2,[ ])	8.9	200	0	25.9	0.9	25.9
cont. fixed	1.88	8.5	187.9	0	32.0	0.8	31.2

Table 2 **An example, where all cost components are decreased by workload-dependency, except for the switching cost.**

We emphasize that $cost_{earliness}$ and $cost_{tardiness}$ both decreased. Thus, the throughput time distribution adapts better to the lead time (L), which in turn results in a better due date performance. The reason for the better adaption is twofold. With workload-dependency the expected value of the throughput time gets closer to the actual lead time. In addition, the variance of throughput time decreases. Table 3 shows the expected value and standard deviation of throughput times ($E[X]$ and $Std[X]$) for the previous example, for which the corresponding Markov-chain, $MC^{(1,3,[3,1;4,2]),\lambda,\mu}$ was illustrated in Figure 1.

Policy	$E[X]$	$Std[X]$
$\pi = (1, 3, [3, 1; 4, 2])$	35.5	20.4
$\pi = (2, 2, [\,])$	39.0	30.3
continuous fixed 1.88	43.8	32.9

Table 3 **Expected value and standard deviation of throughput time in the same setting.**

4.4. Sensitivity analysis on the assumption of exponential service times

Although our calculation method presented in section 3 gives the optimal policy under a homogenous Poisson order arrival process and exponential service times, it is not obvious how it performs in more realistic settings. In the literature, the single item, single machine with fixed capacity to-order settings are often modeled as an $M|G|1$ queue (see Wein (1991),Baker & Bertrand (1981) and Tijms & van der Duyn Schouten (1978)) instead of an $M|M|1$ queue. While to assume Poisson order arrival process is considered reasonable, the assumption of exponential service times is found to be restrictive. This observation entails that some sensitivity analysis on the service times is necessary to ensure that the policy which is optimal under the exponential service time assumption performs well even when the assumption does not hold.

We present a simulation study to allow deterministic service times and compare our proposed policy calculation with a given plausible managerial approach. We consider a case, where the number of shifts can vary from one to three, and the maximum workload is six jobs ($C_{min} = 1$, $C_{max} = 3$, $W_{max} = 6$). The approach of the manager is to switch between one shift of capacity and two shifts at halfway to the maximum workload; the third shift is only utilized when the maximum workload is reached.

We define a new term, $MCE\%$, to measure the value of our proposed policy calculation as compared to the manager's policy, primarily under deterministic service times. Let $MCE\%$ be the ratio of expected total costs of the manager's policy and the policy, which our calculation provide and is optimal under exponential service times, minus one. In our numerical experiments, we use $3^5 = 243$ parameter settings with $\lambda = 0.04$, 0.06, and 0.08; $L = 30$, 45, and 60; $c_{capacity} = 100$; $c_{lostsales} = 3000$, 4000, and 5000; $(c_{earliness}, c_{tardiness}) = (1, 10)$, $(2, 25)$, and $(5, 100)$; $c_{switching} = 0$, 1000, and 2000.

We investigate the $MCE\%$ values under deterministic service times, but also indicate the outcomes under exponential service times in Table 4 so that we have a reference point. This table shows that the policy we propose take more advantage of the change from exponential to deterministic service times, than the manager's policy. That $MCE\%$ values in the deterministic setting can exceed 50% shows that the manager's policy can perform particularly poor in some situations.

service time distribution	minimum	maximum	mean	median	negative
exponential	+1.04%	31.01%	6.28%	4.25%	0%
deterministic	−4.41%	52.45%	7.34%	5.04%	16%

Table 4 **Statistics of the $MCE\%$ values for the 243 parameter settings tested**

The highest $MCE\%$ values occur by the combination of high $(c_{earliness}, c_{tardiness})$, high λ and short quoted lead-times. In these cases the due date performance of the manager's policy falls far behind our proposed one. The highest 9 $MCE\%$ values correspond to the 9 possible $(c_{lostsales}, c_{switching})$ settings for $(c_{earliness}, c_{tardiness}) = (5, 100)$, $\lambda = 0.08$, and $L = 30$ (the average $MCE\%$ of these 9 cases is 46%).

The smallest $MCE\%$ values occur typically for low $c_{switching}$. The lowest 40 $MCE\%$ values are negative, among which the lowest 22 $MCE\%$ values have $c_{switching} = 0$ level. Our explanation is that, because of the exponential service assumption, our policy calculation presumes higher variance in queue length than it is by deterministic service times. Therefore our approach proposes a policy that uses the extreme capacity levels too often. However, when switching costs are incurred, the extreme capacity levels become less attractive, which makes our policy calculation more robust against low service time and queue length variance.

4.5. An illustration to the use of workload-dependent policies in real-life

A firm produces colored neon light figures using a quoted production lead time of 2 weeks (L). On average, the firm receives 3.8 orders weekly (λ). One worker can manufacture one neon light figure in a week (μ). In the present situation, all the 4 regular workers work 8 hours each day. The manager would like to introduce a fast-response flexible labor arrangement in order to cope better with the varying workload. He compares two alternatives, well-known in practice: the overtime and the on-call temporary labor arrangements.

Including the present situation, we define three workload-dependent policy classes corresponding to the alternatives: the fixed, the overtime, and the temporary labor policy classes. The classes, we parameterize with the number of regular workers, an integer, which we denote by x, and the switching points, θ. The capacity levels of the classes are x, $\left(x, \frac{9x}{8}\right)$,$(x, x+1, \ldots, x+5)$, respectively. Overtime costs 1.5 times the regular capacity cost, but it does not incur switching cost, whereas temporary labor costs 1.25 times the regular capacity and it does incur switching cost. The service rate at workload w, and capacity level c is $\mu_{w,c} = c\mu$, as before. New orders are rejected if the firm already has 50 jobs (W_{max}).

We summarize the results for the cost coefficients $(c_{capacity}, c_{switching}, c_{lostsales}, c_{earliness}, c_{tardiness}) = (100, 100, 100, 2.5, 50)$ in Table 5. We conclude that the temporary labor arrangement is the manager's best alternative in this particular case.

fast-response flexible labor arrangement	optimal number of regular workers (x)	optimal switching points (θ)	total costs
fixed	4	( )	483.3
overtime	4	(12 20)	449.8
temporary labor	3	$\begin{pmatrix} 32 & 38 & 43 & 47 & 50 \\ 20 & 26 & 28 & 30 & 31 \end{pmatrix}^T$	441.0

Table 5 Expected costs of the different labor arrangements.

5. Conclusions and future research

Under the assumption that instantaneous capacity changes are possible, we defined a set of workload-dependent capacity planning policies. These policies have the ability to react to dynamic workload changes. We introduced a model that gives insight into the value of workload-dependent capacity control, via stationary analysis.

We measured the cost savings by using the workload-dependent policies as compared to the fixed policies, which can use only one capacity level. Large switching cost coefficients, high demand rates, and long quoted lead times are detrimental, while high workload limits ($W_{\max}$) are beneficial for the savings. Capacity discreteness can strongly affect the cost savings, as workload-dependent policies can counteract non-integer capacity needs, while fixed policies cannot. We showed that when the necessary capacity is between 0 and 1, we need to differentiate two cases: if the zero capacity level is feasible (closeable system) then the savings are particularly high, whereas if the zero capacity level is infeasible the savings are low.

In the uncapacitated case, we observed that using a sufficiently high order-acceptance rate, or equivalently a high workload limit ($W_{\max}$), is a safe choice when selecting the workload-dependent strategy. We found that for high workload limits, the optimal capacity up- and down-switching points tend to change less and less, and appear to form two lines. This observation may facilitate future research on the policy class comprising this linear type of policies.

Finally, we revealed that compared to the optimal fixed capacity policies, the optimal workload-dependent capacity planning policies can achieve a better due-date performance. In particular cases they can also spare capacity, and decrease lost sales probability at the same time (as shown in Table 2). We tested our proposed policy's sensitivity on our most restrictive assumption (exponential service times). The results show robustness of our policy, especially for high switching costs.

Extensions to our model are numerous. We summarize the most relevant ones only. First, more general interarrival and service time distributions can be handled by using phase-type distributions instead of exponential. Second, service time parameters ($\mu_{w,c}$) can be set so that they correspond

to situations where only one server can be assigned to a job or represent other dependencies on actual workload or capacity level. Third, switching costs for starting or suspending production are usually higher than increasing or decreasing production rate when production is already running. A simple extension can assign different costs for the different types of switchings.

A generalization of our model to positive capacity alteration lead times could reveal to what extent the assumption of instantaneous capacity changes is restrictive in workload-dependent capacity planning. However, this extension seems to be more complicated.

Appendix

In this appendix, we apply uniformization for $EMC^{\pi,\lambda,\mu}$ in order to obtain the probability that starting from state r at time 0 we will be state s in t time.

We can define both $\bar{A}^{\pi}$ extended arrival matrix, and $\bar{D}^{\pi}$ extended departure matrix for $EMC^{\pi,\lambda,\mu}$ based on A^{π}, and D^{π}, as follows.

$$\bar{A}^{\pi}_{(w_1,c_1,q_1),(w_2,c_2,q_2)} = \begin{cases} 1 \text{ , if } A^{\pi}_{(w_1,c_1),(w_2,c_2)} = 1, \text{ and } q_1 = q_2 \\ 0 \text{ , otherwise} \end{cases}$$

$$\bar{D}^{\pi}_{(w_1,c_1,q_1),(w_2,c_2,q_2)} = \begin{cases} 1 \text{ , if } D^{\pi}_{(w_1,c_1),(w_2,c_2)} = 1, \text{ and } q_1 - 1 = q_2 \\ 0 \text{ , otherwise} \end{cases} \tag{12}$$

for all (w_1, c_1, q_1), and (w_2, c_2, q_2). The transition rate diagram of $EMC^{\pi,\lambda,\mu}$ for the policy $\pi = (1, 3, [3, 1; 4, 2])$ can be seen in Figure 7.

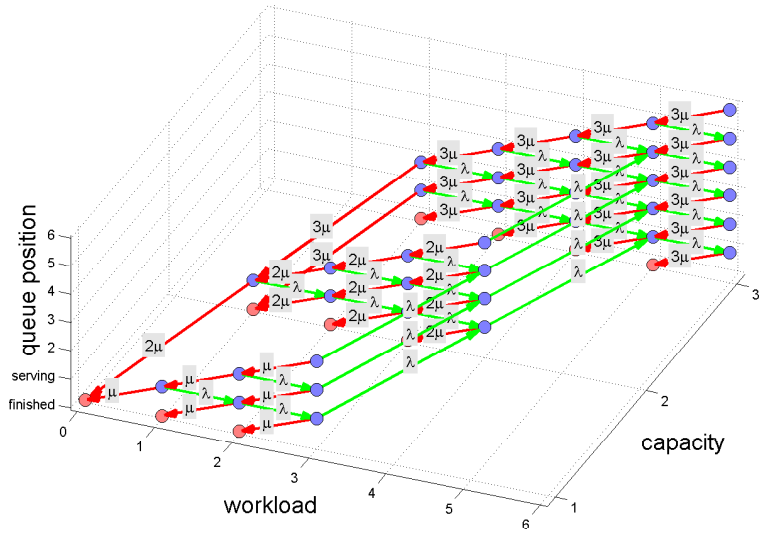

Figure 7 $EMC^{\pi,\lambda,\mu}$, **for** $\pi = (1, 3, [3, 1; 4, 2])$

The extended transition rate matrix, $\bar{Q}^{\pi}$, can be given in two steps like before.

$$\bar{Q}^{\pi}_{(w_1,c_1,q_1),(w_2,c_2,q_2)} = \lambda \bar{A}^{\pi}_{(w_1,c_1,q_1),(w_2,c_2,q_2)} + \mu_{c_1} \bar{D}^{\pi}_{(w_1,c_1,q_1),(w_2,c_2,q_2)} \tag{13}$$

if $(w_1, c_1, q_1) \neq (w_2, c_2, q_2)$.

Next, we define the vector of state leaving rates, $\bar{v}^{\pi}$.

$$\bar{v}^{\pi}_s = \sum_{s \neq r} \bar{Q}^{\pi}_{s,r} \tag{14}$$

Now, the remaining diagonal elements of $\bar{Q}^{\pi}$ can be defined as $\bar{Q}^{\pi}_{s,s} = -\bar{v}^{\pi}_s$.

From $\bar{Q}^{\pi}$ we can derive the extended transition probability matrix, $\bar{P}^{\pi}$.

$$\bar{P}^{\pi}_{r,s} = \begin{cases} \frac{\bar{Q}^{\pi}_{r,s}}{\bar{v}^{\pi}_r} \text{ , if } r \neq s \\ 0 \quad \text{ , if } r = s \end{cases} \tag{15}$$

We use uniformization; we add loop arcs to most of the states in the extended Markov chain ($EMC^{\pi,\lambda,\mu}$) so that the distribution of the time between two subsequent events (arrival or departure or loop) becomes identical. The uniformized event occurrence rate is $\bar{v}^{U\pi} = \max_s \bar{v}^{\pi}_s$. The uniformized extended transition probability matrix, elementwise, is

$$\bar{P}^{U\pi}_{r,s} = \begin{cases} \frac{\bar{v}^{\pi}_r}{\bar{v}^{U\pi}} \bar{P}^{\pi}_{r,s} \text{ , if } r \neq s \\ 1 - \frac{\bar{v}^{\pi}_r}{\bar{v}^{U\pi}} \text{ , if } r = s \end{cases} \tag{16}$$

Using the matrix, $\bar{P}^{U\pi}$, we can determine the probability of getting from state r to s in time t.

$$\bar{P}^{\pi}_{r,s}(t) = \sum_{n=0}^{\infty} \left(\left(\bar{P}^{U\pi} \right)^n \right)_{r,s} e^{-\bar{v}^{U\pi} t} \frac{\left(\bar{v}^{U\pi} t \right)^n}{n!} \tag{17}$$

References

Avi-Itzhak, B., & Heyman, D.P. 1973. Approximate queuing models for multiprogramming computer systems. *Operations Research*, **21**, 1212–1230.

Baker, K.R., & Bertrand, J.W.M. 1981. An investigation of due-date assignment rules with constrained tightness. *Journal of Operations Management*, **1**, 109–120.

Baynat, B., & Dallery, Y. 1993. A unified view of product-form approximation techniques for general closed queueing networks. *Performance Evaluation*, **18**, 205–224.

Bekker, R., & Boxma, O.J. 2005. *An M/G/1 queue with adaptable service speed.* SPOR-Report 2005-09, Eindhoven University of Technology, The Netherlands, submitted for publication.

Bentolila, S., & Bertola, G. 1990. Firing costs and labour demand: How bad is eurosclerosis? *The Review of Economic Studies*, **57**, 381–402.

Bertrand, J.W.M. 1983. The effect of workload dependent due-dates on job shop performance. *Management Science*, **29**, 799–816.

Bertrand, J.W.M., & van Ooijen, H.P.G. 2002. Workload based order release and productivity: a missing link. *Production Planning and Control*, **13**, 665–678.

Chenery, H.B. 1952. Overcapacity and the acceleration principle. *Econometrica*, **20**, 1–28.

Faddy, M.J. 1974. Optimal control of finite dams: discrete (2-stage) output procedure. *Journal of Applied Probability*, **11**, 111–121.

Filho, E.V.G.A., & Marçola, J.A. 2001. Annualized hours as a capacity planning tool in make-to-order or assemble-to-order environment: an agricultural implements company case. *Production Planning and Control*, **12**, 388–398.

Gelders, L.F. 1991. Production control in an 'engineer-to-order' environment. *Production Planning and Control*, **2**, 280–285.

Kim, J., Bae, J., & Lee, E.Y. 2006. An optimal P_{λ}^{M}-service policy for an M/G/1 queueing system. *Applied Mathematical Modelling*, **30**, 38–48.

Luss, H. 1982. Operations research and capacity expansion problems: a survey. *Operations Research*, **30**, 907–947.

Manne, A.S. 1961. Capacity expansion and probabilistic growth. *Econometrica*, **29**, 632–649.

Marie, R.A. 1979. An approximate analytical method for general queueing networks. *IEEE Transactions on Software Engineering*, **5**, 530–538.

Philipoom, P.R., Malhotra, M.K., & Jensen, J.B. 1993. An evaluation of capacity sensitive order review and release procedures in job shops. *Decision Sciences*, **24**, 1109–1133.

Pinder, J.P. 1995. An approximation of a Markov decision process for resource planning. *Journal of the Operational Research Society*, **46**, 819–830.

Rocklin, S.M., Kashper, A., & Varvaloucas, G.C. 1984. Capacity expansion/contraction of a facility with demand augmentation dynamics. *Operations Research*, **32**, 133–147.

Schlichter, M. 2005. *Stork Fokker - The production planning and control system.* Report, Logistics Management Systems, Technische Universiteit Eindhoven, The Netherlands, 2005, p55.

Tijms, H.C., & van der Duyn Schouten, F.A. 1978. Inventory control with two switch-over levels for a class of M/G/1 queueing systems with variable arrival and service rate. *Stochastic Processes and Their Applications*, **6**, 213–222.

Vollmann, T.E., Berry, W.L., Whybark, D.C., & Jacobs, F.R. 2005. *Manufacturing planning and control for supply chain management.* New York: McGraw-Hill.

Wein, L.M. 1991. Due-date setting and priority sequencing in a multiclass M/G/1 queue. *Management Science*, **37**, 834–850.