A SHORT TERM CAPACITY ADJUSTMENT POLICY FOR MINIMIZING LATENESS IN JOB SHOP PRODUCTION SYSTEMS

Henny P.G. van Ooijen
J. Will M. Bertrand
Technische Universiteit Eindhoven
Department of Technology Management
phone:+31402472230; e-mail: h.p.g.v.ooijen@tm.tue.nl

Abstract

In stochastic, dynamic job shop production systems, order flow times can be estimated with considerable accuracy by taking into account order processing information, shop workload information and work center workload information. This can be used to quote highly reliable due dates. However, customer order due dates based on this information are highly variable and therefore the customers cannot anticipate to that. Customers prefer constant lead times. In this study we assume that the work center capacity at certain time intervals can be varied (to some extent) around a given long term level. We develop a capacity adjustment policy that periodically selects capacity levels per work center such that for each order the estimated lateness is minimized.

1. Introduction

It is well known that in stochastic, dynamic job shop production systems, the throughput times may vary heavily over time. Depending on the production situation and the required delivery reliability, this might lead to rather long lead times which might not be acceptable for the customers. Moreover, the lead times for the same products might be different over at different points in time. In many production situations customers prefer fixed lead times. For instance if the customer needs the products for its own operation, it is important that the lead-time of these products is constant and, moreover, reliable. If lead-times vary over time and/or not reliable, then it is difficult to control their own production. A way to work with constant lead times that lead to a certain delivery reliability is to take the variance of the throughput times into account in setting the lead-times. However, if the variance of the throughput times is rather high, this will lead to long lead times. Shorter lead times can be obtained by accepting less customer orders (which leads to lower utilization rates) but this influences also the income and thus the profit.
In situations where the capacity can be varied on the short term, one might use this to control the throughput times of the orders, which might reduce the variance and thus the lead times. This control leads to the following: if there are many orders on the shop floor that tend to be delivered late, one might extend the capacity at some work centers. On he opposite, if there are many orders that tend to be delivered too early, capacity might be decreased at some work centers. The question is how to determine whether orders will be late and how much capacity expansion is needed at a certain instance. To answer the first part of this question we will use a method developed in Van Ooijen and Bertrand (2001). To determine the required change in capacity we develop a capacity adjustment policy that, given a certain required constant lead-time, minimizes the estimated lateness of the current orders.

The remainder of this paper is organized as follows. In Section 2 we will discuss some related literature. In Section 3 the method that will be used for determining whether orders will be late. The capacity adjustment policy will be discussed in Section 4 and in Section 5 we will investigate the performance of this policy. Finally, in Section 6 the conclusions will be given.

2. Related literature

Weng (1996) studies the problem of maximizing the expected profit through optimizing the trade-off between short manufacturing lead times and high system utilization rates. His model shows how the expected profit can be maximized by taking advantage of the existence of lead time-sensitive demand and lead time-insensitive customer orders.
Palaka et al. (1997) present a model to study the lead-time setting, capacity utilization, and pricing decisions of a profit-maximizing firm serving customers that are sensitive to quoted lead times. Their analysis shows that the capacity utilization should be lower when (1) customers are more sensitive to lead times and/or (2) the firm incurs higher congestion related costs and/or (3) the penalty for lateness is higher.
So and Song (1997) studies the impact of using delivery time guarantees as a competitive strategy in service industries where demands are sensitive to both price and delivery time. They assume that delivery reliability is crucial, and investment in capacity expansion is plausible in order to maintain a high probability of delivering the time guarantee. A mathematical framework is proposed to understand the interrelations among pricing, delivery guarantee and capacity expansion decisions.
Ray and Jewkes (2004) extend previous research by explicitly modeling a relationship between price and delivery time. The firm they investigate can invest in increasing capacity to guarantee a shorter delivery time but must be able to satisfy the guarantee according to a pre-specified reliability level. The model accounts for whether customers are "price sensitive" or "lead time sensitive" by capturing the dependence of *both* price and demand on delivery time.
Barut and Sridharan (2004) develop a heuristic for short term constrained capacity allocation to multiple product classes in make-to-order manufacturing, deploying a decision theory based approach.

All studies thus far do not investigate the adjustment of capacity on the (very) short term using the expected throughput times. They set capacity levels on the medium or long term, based on the expected profit, or they investigate measures that can be taken before orders are released to the shop floor (outsourcing or allocating capacity to the most profitable classes). The effect of possible measures with regard to the performance of orders that have been released already is not investigated. In this research we want to investigate whether adjustment of capacities based on the expected lateness of the orders that are being processed can improve the (economical) performance. In general there are costs related to changing the capacities. However, there are often also costs related to late deliveries. So, there is a trade off between costs of changing the capacities and late deliveries. The expected lateness will be determined using the method developed in Van Ooijen and Bertrand (2001), where it is used to set cost optimal due dates. This method will be discussed in the next section.

3. The expected lateness

In this study we look at symmetric job-shops, where orders arrive according to a Poisson process, order routings are determined upon arrival, and work center processing times are generated from a negative exponential distribution function. Upon arrival orders get a due date that is based on a fixed lead-time. For orders that have a completion time that deviates from this due data a penalty has to be paid. We assume that in this job shop the work center capacities at certain time intervals can be varied, to a certain extent, around a given long term level. This varying of the capacities goes at a certain cost. Varying of the capacity only makes sense if this leads to a better delivery performance and thus less lateness costs. The change of the capacities must thus be based on the expected lateness. To determine the required capacity adjustment at a certain instance, we therefore first need to have estimates of the difference between the delivery date, as communicated to the customer, and the expected delivery date of the orders that have been released at the time the adjustment of capacities is being considered. Therefore we need to have an estimate of the completion date of the orders that are released at that time. To get this estimate we will use the method that is developed in Van Ooijen and Bertrand (2001) for setting economic due dates. In that study for each category of orders, where a category is determined by the number of operations they use so-called normalized waiting times distribution

functions. These are obtained by normalizing the observed waiting times with regard to the work-in-process.
In this study we will also start with determining normalized waiting time distribution functions for each category of orders. However, instead of using aggregate information, as in Van Ooijen and Bertrand (2001), we will use routing related workload information (see Bertrand and Van Ooijen (2003)) since this policy leads to the best performance. In this way we get empirically constructed routing normalized waiting time distribution functions per order category using:

$$w_j^N = \frac{q}{q_j + 1} w_j$$

where w_j^N = routing normalized waiting time of order j
w_j = waiting time of order j
q = the long term average number of orders at the work centers on the routing of an order
q_j = the actual number of orders in the work centers on the routing of order j at the arrival time of order j

In determining the expected lateness of a released order at a certain time *t,* we use these empirically constructed normalized waiting time distribution functions. First we determine the remaining number of operations. This remaining number of operations determines the category of the routing normalized waiting time distribution functions that we will use in estimating the completion date of the order. Using this routing normalized waiting time distribution function we determine the routing normalized remaining waiting time that corresponds to a certain required delivery reliability. This routing normalized remaining waiting time is transformed into an "actual" remaining waiting time by multiplying it by the actual number of orders at the work centers on the (remaining) routing of that order at time *t,* and dividing it by the long term average number of orders at the work centers on the (remaining) routing of that order. For instance, if at time *t,* a certain order *x*, has *g* remaining operations, then the remaining waiting time for this order *x* that will be realized with a reliability of α% is:

$$\frac{q_x}{q} F_g^{\leftarrow}(\alpha) \qquad (1)$$

with q_x :actual number of orders at the work centers of the remaining operations of order *x* at time *t*
q :the long-term average number of orders at the work centers on the routing of order *x*
$F_g^{\leftarrow}(\alpha)$:the routing-workload normalized waiting time for orders of category *g*, that with a probability of α is not exceeded.

If we add to this expected remaining waiting time the expected processing times of the remaining operations of that order, we get an estimate of the completion time that has the required delivery reliability.
Comparing this completion time with the due date gives the expected lateness of this order.

4. The capacity adjustment procedure

Let us start with the situation where the adjustment of the capacities can be done without any restriction. Suppose that given the situation at the shop floor at a certain time *t* where we can vary the capacity, a certain order *x,* with *g* remaining operations, will get an expected α% reliable completion date of *C*. If this is equal to the due date that is given to the order at arrival, the available capacity matches the required capacity. If not, we have too much or too less capacity.

Conjecture: The capacity required to deliver order x to its due date, can be determined by interpreting the load q as the load in relation to the installed capacity (relative load).

For instance, the situation with a routing load of 40 and 50 hours installed capacity on that routing is equal to the situation with a load of 60 and 75 hours installed capacity. With this interpretation a load of L with M installed hours is transformed into a load L^* if the number of installed hours is changed to $(M*L)/L^*$.

Using (1) it can be determined which value q_x^* is required to deliver order x to its due date DD with reliability α:

$$q_x^* = \frac{DD - t - \sum_{\substack{remaining \\ operations\ ;\ j}} p_{xj}}{F_g^{\leftarrow}(\alpha)} q \qquad (2)$$

with q equal to the long-term average number of work orders at the work centers on the remaining routing of order x at time t.
To get the right value for q_x^*, the capacity must be changed in such a way that the load, in relation to the changed capacity, equals q_x^*. This means that the changed total capacity at the work centers on the remaining routing of order x, CC, must be equal to:

$$\frac{q * AC_t}{q_x^*}$$

with AC_t equal to the actual total capacity at the work centers on the remaining routing of order x at time t.

This can be done for all orders on the shop floor at time t. However, in general, this will lead to the situation where for a certain order the capacity at certain resources must be increased, whereas, at the same time, for another order these same capacities should be decreased. Therefore, varying the capacities at a certain time, such that all orders present at the shop floor will be delivered exactly in time, will seldom be possible. There will always be some orders that will be delivered too late or too early.

No restrictions to the variations in capacity.
Suppose that the capacities can be varied without any (cost) restriction. Then the goal of our procedure is to vary the capacities in such a way that the sum of the deviations of the expected completion times from the corresponding due dates is minimized. Now suppose that at a certain time t the number of orders at resource i is equal to n_{ti}, that we use a delivery reliability of α% and that the costs of lateness is an exponential function of the lateness. Then we will get the following objective function:

$$\min_{\beta(.)} \sum_{\substack{released \\ orders\ x}} \left(DD_x - \left(t + \sum_{\substack{remaining \\ operations\ j \\ of\ order\ x}} p_{xj} + \frac{\sum_{\substack{workcenters \\ remaining\ operations\ j \\ of\ order\ x}} \beta_j . n_{tj}}{q} F_{g_x}^{\leftarrow}(\alpha)\right)\right)^2 \qquad (3)$$

where $\boldsymbol{\beta}$ is a vector with as entries β_i, such that $\beta_i.n_{ti}$ is the relative workload that should be present at work center i to get a delivery reliability of α%.

Now suppose that in first instance we can only vary the capacity of the shop as *a whole*. That is, if we increase the capacity of a certain resource, using a certain ratio, we also (have to) increase the capacity

of the other resources with the same ratio. In that case all β_j in (3) are equal and the problem is then to determine this value for the β_j such that it minimizes the objective function (3). Some straightforward calculations then give that β_j, j=1,….,m (m is number of work centers) has to be equal to:

$$q\frac{\sum_{\substack{\text{released}\\ \text{orders } x}}(DD_x - t - \sum_{\substack{\text{remaining}\\ \text{operations } j}} p_{xj})}{\sum_{\substack{\text{workcenters}\\ \text{remaining}\\ \text{operations } j}} n_{tj} F_{g_x}^{\leftarrow}(\alpha)}$$

If the capacities of the resources can be varied *independently* of each other, we have a vector $\boldsymbol{\beta}$ with as many elements β_j as the number of work centers. Now we have to determine the vector $\boldsymbol{\beta}$, such that it satisfies (3). Rewriting (3) gives that we have to minimize:

$$\min_{\beta(.)} \sum_{\substack{\text{released}\\ \text{orders } x}} ((DD_x - t - \sum_{\substack{\text{remaining}\\ \text{operations } j}} p_{xj}) - \frac{F_{g_x}^{\leftarrow}(\alpha)}{q} \sum_{\substack{\text{workcenters}\\ \text{remaining}\\ \text{operations } j}} \beta_j . n_{tj})^2$$

If we define

N_{xjt} :the number of the orders in the queue at work center *j* at time *t*, multiplied by the remaining number of visits of order *x* to work center *j*

$a(x,t)$: $(DD_x\text{-}t\text{-}\Sigma p_{xj})*q/F_{gx}^{\leftarrow}(\alpha)$

we get as objective function:

$$\min_{\beta(.)} \sum_{\substack{\text{released}\\ \text{orders } x}} (a(x,t) - \beta_1 N_{x1t} - \beta_2 N_{x2t} - .....\beta_m N_{xmt})^2$$

This is of the form:

$$\min_{\beta(.)} \| a(t) - N_t \beta \|$$

with N_t :m x n matrix with n equal to the number of orders at the shop floor at time t; elements are N_{xjt} ; j=1…m, x=1….n

a(t) :vector with elements a(x,t); x=1,....n

<u>*Restrictions to the variations in capacity.*</u>

Thus far, we have considered the situation where variation of the capacities can be done without any restriction. We derived an expression for determining the ratios that determine which number of orders in the queue we should have in relation to the available capacity. Since we cannot change the number of orders in the queue, but we can change the capacity, the derived ratios can be used to determine the capacity that is needed to get the required relative loads. In general, there will be a limit to the variations, mainly due to the fact that there are costs associated with the variations. Let us assume that these costs can be interpreted as costs associated with changing the relative load, that these costs are linear to the size of adjustment and that the costs of changing the relative load with one unit equals c_1. If costs of lateness are an exponential function of the lateness and the cost parameters c_2, the objective function becomes:

$$\min_{\beta(.)}\left(\sum_{i=1}^{m} c_1((1-\beta_i)n_{ti}) \;+ c_2 \sum_{\substack{released\\ orders;x}} (DD_x - (t + \sum_{\substack{remaining\\ operations;j}} p_{xj} + \frac{\sum_{\substack{workcenters\\ remaining\\ operations;j}} \beta_j . n_{tj}}{q} F_{g_x}^{\leftarrow}(\alpha)))^2\right)$$

which is quivalent to

$$\min_{\beta(.)}\left(-c_1 \sum_{i=1}^{m} \beta_t n_{ti} + c_2 \sum_{\substack{released\\ orders;x}} (DD_x - (t + \sum_{\substack{remaining\\ operations;j}} p_{xj} + \frac{\sum_{\substack{workcenters\\ remaining\\ operations;j}} \beta_j . n_{tj}}{q} F_{g_x}^{\leftarrow}(\alpha)))^2\right) \qquad (4)$$

If again we define:

N_t :m x n matrix with n equal to the number of orders at the shop floor at time t;
elements are N_{xjt} ; j=1…m, x=1….n

a(.,t) :vector with elements a(x,t); x=1,....n

and

y(.,t) :vector with elements $y(i)=-c_1.n_{ti}/2.c_2$ i=1,…m

(4) can be rewritten as:

$$\min_{\beta}\left(y^T.\beta + \frac{1}{2}\left\| a - N_t\beta \right\|_2^2\right)$$

Since negative relative load must be excluded and, in general, there will be a limit B for the adjustment of the capacity, we have the constraint:

$$\beta \geq B$$

Now we have a so-called constrained least squares problem and that can be used by using the software package LSSOL.

5. The simulation study

For the investigation of the performance of our capacity adjustment procedure, we performed a simulation study of a job-shop with the following characteristics:

- The job-shop model consists of five single machine work centers (as in many research of this type, see Conway et al. [1967]).
- Orders arrive according to a Poisson process. We assume that the delivery performance has no influence on the arrival rate. So, customers do not have a memory with regard to the past performance of the shop.
  Order routings are determined upon arrival. The routings are generated in such a way that each work center has an equal probability of being selected as the first work center. After the first operation the probabilities of going to any of the other work centers are equal and depend on the probability of leaving the shop, which in turn depends on the average routing length. We have used an average routing length of 5, so the probability of leaving the shop after each completed operation is 0.2, and the work center transition probabilities all equal 0.8/4=0.2.
- At each work center processing times are generated from a negative exponential probability density function with a mean value of one time unit. Set-up times and transportation times are considered to be zero.

- The mean value of the order inter-arrival time is equal to 10/9, which implies a machine utilization rate of 90%.
- First Come First Served sequencing at the work centers.
- Capacities can be varied weekly (this is roughly about the average throughput time) or monthly. This is done by determining the required variations in capacity at the beginning of the week or month

Like in Sabuncuoglu and Comlekci [2002], we carried out two sets of simulation experiments. With the first set we determined the non-normalized, routing workload-normalized order waiting time distribution functions per category. Next we performed a set of experiments to investigate the performance of the system using the distribution functions constructed in the first set of experiments for determining capacity adjustments that minimize the lateness. Each measurement results from 10 simulation runs, with order streams that differ from the order streams that were used to construct the (normalized) waiting time distribution functions. The common random number technique was used to reduce the variance between experiments with different settings.

6. Conclusions

References

Barut, M. and V. Sridharan, (2004), *"Design and evaluation of a dynamic capacity apportionment procedure",* European Journal of Operational Research, 155, pp. 112-133.

Bertrand, J.W.M. and H.P.G. van Ooijen, (2003), "*Using order routing specific flow time and workload information to improve lead-time and due date performance in job shops",* Working paper, Technische Universiteit Eindhoven, Dept. Technology Management.

Conway, R.W., Maxwell, W.L. and Miller, W.M. (1967), 'Theory of Scheduling', Addison-Wesley, Reading Massachusetts.

Palaka, K., Erlebacher, S. and D.H. Kropp, (1998), "*Lead-time setting, capacity utilization, and pricing decisions under lead-time dependent demand",* IIE Transactions, 30, 2, pp. 151-163.

Ray, S. and E.M. Jewkes, (2004), *"Customer lead time management when both demand and price are lead time sensitive",* European Journal of Operational Research, 153, pp. 769-781.

Sabuncuoglu, I. and Comlekci, A. (2002), Operation-based flow time estimation in a dynamic job shop, Omega, 30, pp.423-442.

So, K.C. and Song, J.S., (1998), "*Price, delivery time guarantee and capacity selection",* European Journal of Operational Research, 111, pp. 28-49.

Van Ooijen, H.P.G. and J.W.M. Bertrand, (2001), "*Economic due date setting in job shops based on routing and workload dependent flow time distribution functions",* Int. J. Production Economics, 74, pp. 261-268.