Modelling input–output workload control for dynamic capacity planning in production planning systems

Brian G. Kingsman*

Department of Management Science, The Management School Lancaster University, Lancaster LA1 4YX, UK

Received 21 October 1999; accepted 22 March 2000

Abstract

Workload control has been described as one of the new production planning and control concepts available for practical operations. The main principle has been defined by as to control the lengths of the queues in front of work stations on the shop floor. If these queues are to be kept short, then waiting times and hence overall manufacturing lead times will be controlled. There are four levels at which this control of queues can be attempted; priority dispatching level, job release level, job acceptance and job entry level. The first of these is a relatively weak mechanism for the control of queues if used alone. A stronger instrument, controlled job release, entails maintaining a ‘pool’ of unreleased jobs in the production planner’s office, which are only released onto the shop floor if doing so would not cause the planned queues to exceed some predetermined norms. The main aim of workload control, for those who advocate its use as a job release method, has been defined as to control the lengths of the queues in front of work stations on the shop floor. However, the true objective is to process the jobs so as to meet the promised delivery dates with the machine and workforce capacities and capabilities available. The job release stage can itself only be fully effective if the queue of jobs in the pool is also controlled. Otherwise, jobs may remain in the pool for too long so missing their promised delivery dates. Thus a comprehensive workload control system must include the customer enquiry stage, (the job entry stage), to control the input of work to the pool as well and plan the capacity to provide in future periods so the shop floor queues are also controlled. A methodology and systems to do this at both the job release and the customer enquiry stage have been presented in previous papers. The purpose of this paper is to provide a theory for workload control in a mathematical form to assist in providing procedures for implementing input and output control. It enables dynamic capacity planning to be carried out at the customer enquiry and order entry stages for versatile manufacturing make-to-order companies. The theory shows that attention should be concentrated on controlling the differences between the cumulative inputs and outputs over time, and not the period individual inputs and outputs. Although aimed at make-to-order companies, the theory and procedures give a general capacity planning method for other production planning methods; for example determining the master production schedule in MRP systems. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Make-to-order manufacturing; Production planning; Lead time management; Workload control; Input/output control; Order release

1. The input–output transformation model for production systems and workload control

The input–output model is a very useful and commonly accepted way of looking at production...
and operations. Production is regarded as a transformation process that takes inputs and transforms them to outputs that are of a higher value than the inputs. This can be viewed at the macro system level, the whole of the organisation or the whole of the production function, or at the micro level of the individual transformation operations or activities. These latter are the work centres at which the successive different transformations of the item being processed are made. The production function consists of a network of work centres, each with a capability to do a particular transformation task. The inputs to the system from a directly physical viewpoint are raw materials, components and sub-assemblies, and the outputs are the finished final goods. Another viewpoint is that the processing is transformation of information. The inputs are orders from customers for specified final products. The order is the information that specifies the particular amounts of transformation work to be carried out to produce the finished products. The output is then the transformation work completed at that work centre.

A crucial aspect for any production input–output model is that the time that the transformation process takes is a very important factor. This is because the finished product is being delivered to a customer who requires receiving that product by a particular promised delivery date. Failure to meet this promised delivery date could, and does, affect the amount of future business likely to arise from that customer and the prices that can be secured. The ability to be able to carry out the necessary transformation processes is an essential qualifier to be in that particular market, but it is the price and delivery lead time quoted are crucial order winning factors. This is particularly so in most of the produce-to-order sector of manufacturing and service companies.

Viewed from both a marketing and a production perspective, the produce-to-order sector of industry can be divided into Repeat Business Customisers (RBCs) and Versatile Manufacturing Companies (VMCs). The former produce customised products for each of its customers on a continuing basis, where the regularity of the demands from the customer may enable some production to stock to be profitable at times. RBCs tend to have a relatively small customer base and compete for the initial order of a continuing supply contract. Having won the initial order, the RBC automatically gets the repeat business on a regular basis for a certain period of time, say two or three years. The component suppliers to motor manufacturers are classic examples of RBCs.

The typical Versatile Manufacturing Company (VMC), in contrast, has to supply a variety of products, usually in small quantities, ranging from a range of standard products to all orders requiring a customised product. The arrival of customer enquiries is a stochastic process over time. Each potential order from the enquiry tends to be for a differing number of units and requires varying routings and processing times through the production facilities. VMCs include the classical job shop production environment. However, the supplier of expensive items to satisfy an intermittent customer demand also falls into this category, since the customer will often seek a quotation from other suppliers even if the previous orders had been bought from just one supplier. The manufacturing process here may be a batch or even a line production process. Many service operations, for example specialised banking services or training packages, also have the same problems. Generally, VMCs must compete for each and every order they receive, quoting either a price or a delivery lead time or both, although they may have an edge with repeat business. Furthermore, they have to make bids for many more orders than they will receive. Tobin et al. [1] found that the strike rate, the proportion of quotes that become firm orders, varied from 3% to virtually 100%.

Each order, or potential order in the form of a customer enquiry, requires transformation work on a series of work centres. It is well known that in the produce-to-order sector an order spends up to 90% of the total time in production waiting in front of or between work centres and only 10% in actual transformation work on the machines. This is due to the variability in order sizes and the number of transformation processes needed per order and the stochastic inter-arrival times between enquiries and orders. The alternative to these long queues would mean that most facilities would be idle for most of the time. The general model of the shop floor is thus
a network of work centres each with a set of orders (jobs) queuing waiting their turn to be processed.

This diverse and unpredictable nature of their production and the arrival of new orders is such that it is inappropriate for such companies to adopt production planning and control concepts such as JIT and TOC, as has been argued for example Tatsiopoulos [2], Zäpfel and Missbauer [3] and Hendry and Kingsman [4]. They do not have the type of repetitive manufacturing that enables dedicated facilities to be set up in a simplified shop floor layout. Thus, they cannot rely on the more visible 'situational management' on the shop floor as described by Johnston [5], which has led to a decrease in the importance of higher level planning in many firms. Instead, they have to continue to operate as job shop batch production systems where the higher level control of shop floor queues, as offered by Workload Input/Output Control, remains a most crucial part of their production planning and control systems.

Most of the literature on the production planning problems of VMCs has focused on the individual work centre level. The bulk of this has concentrated on priority dispatching; the order in which jobs should be scheduled through each work station. Surveys show that hundreds of such priority rules have been devised for application on the shop floor. However, experience has demonstrated that priority dispatching is a relatively weak mechanism for the control of queues. If used alone, it has little effect in reducing the long lengths and highly variability of the queues of jobs in front of machines. A stronger instrument, controlled job release, entails maintaining a 'pool' of unreleased jobs in the production planner's office, which are only released onto the shop floor if doing so would not cause the planned queues to exceed some predetermined norms. This in turn reduces the work-in-process and the task of priority dispatching is made easier.

The methods of job release are founded on the theory of Input/Output control, introduced by Wight [6], see also Plossl and Wight [7] and Belt [8]. Irastorza and Dean [9] were the first to develop a sophisticated release method, which balances and limits the queues of jobs on the shop floor. A more comprehensive class of control concepts and load-orientated release methods were developed in the eighties, see the works of Bertrand and Wortmann [10], Bechte [11], Tatsiopoulos [2], Tatsiopoulos and Kingsman [12], Kingsman et al. [13], Kurosu [14] and Fry and Smith [15]. Wiendahl [16] has reported successful implementations of these workload control concepts. Land and Gaalman [17] provides a good critical assessment of the various workload control concepts used for order release. Land and Gaalman [18] presents a more sophisticated release mechanism build on different principles, that of 'superfluous-load-avoidance-release', SLAR. Initial simulations for a balanced job shop suggest that improvements in due date performance and reductions in the direct loads at work stations are likely.

2. The basis of workload control for a single work centre

A work centre, for production planning purposes, is defined by its capacity. This is an output rate, the amount of transformation work that the work centre can perform per time period. That capacity is a rate is a crucial aspect to grasping the production situation and the problems of production planing and control. A common usage of the term capacity is the amount of some substance that a container can hold, for example the water in a bath tub. In the production situation, this water is the load of work that has to be processed. It is the input of work to be processed through the purifying process. The capacity of the system is the rate at which water can flow out of the bath through the purifying (transformation) process. Capacity is related to the output and the rate at which the transformation process can supply the output required.

The load of water in the bathtub is equivalent to the amount of transformation work on the work centre required by the queue of jobs in front of each work centre in the production system. The capacity is the output rate, the amount of processing work that can be done by the work centre per time period. Provided that the load of input work, given by the jobs in the queue, is measured in the same units of processing work as the output, then the
time taken to process all of the jobs in the queue will be

\[
\text{Lead Time} = \frac{\text{Load (measured in transformation work)}}{\text{Capacity (measured in transformation work per time period)}}.
\]

This is analogous to Little's formula in Queueing Theory relating the average time in the queue to the average number in the queue, the arrival and service rates.

If another order arrives and joins the queue, then, assuming all of the other jobs are processed first, the FIFO rule, that order will have to wait the above lead time before it can start being processed on the work centre. Thus the earliest time, using the FIFO rule, that order will have to wait the above lead time plus the time to process the job itself. If the different jobs already in the queue are independent, then priority dispatching as discussed above as a major area of research activity can only change the order in which each job emerges out of the work centre; it cannot change the time before the work centre is free and available to start processing further jobs. If it is an overloaded situation where there is too much work to do every job in time, it can help to select which jobs should be delivered on time, by giving them priority for processing at the work centre, and thus which will be allowed to be late. It cannot in these circumstances provide any assistance on ways to deliver all jobs on time.

The usual terminology is to talk in terms of orders when considering it from the customer or marketing perspective. This is then turned into the term job when the order enters production and is being processed and transformed through the various work centres.

Let us define the QUEUE as the total amount of processing work for all of the jobs currently waiting at the work centre this period and let INPUT be the total processing work of jobs that arrive in the time period, the orders that are accepted. Continuing to assume the FIFO rule, then the last order accepted will be completed and have a delivery lead time, in unit time periods, equal to

\[
\text{Delivery Lead Time} = \frac{\text{INPUT} + \text{QUEUE}}{\text{Capacity}}.
\]

If the company objective is to offer a customer service objective that the delivery lead time will be less than some pre-set value, say \(L\) time periods, then this requires that for every time period

\[
\text{Delivery Lead Time} = \frac{\text{INPUT} + \text{QUEUE}}{\text{Capacity}} \leq L.
\]

The QUEUE is a fixed amount of work that has to be processed. If the ratio value exceeds \(L\), then to ensure that the delivery lead time for all accepted orders is less than or equal to \(L\), requires either that the denominator, the capacity, is increased or that the numerator, the INPUT, is decreased.

Input control implies refusing new orders for delivery in \(L\) time periods once the INPUT exceeds

\[
\text{INPUT} \leq I_1 = L(\text{Capacity}) - \text{QUEUE}.
\]

Once the INPUT reaches \(I_1\), then either further orders must be rejected in the time period or the further orders are allocated a later lead time \(L_2\), say. This will apply to the set of orders whose INPUT amounts up to \(I_2\), where

\[
\frac{I_2}{\text{Capacity}} \leq L_2 - L = \text{output over periods } L + 1 \text{ to } L_2.
\]

The original lead time \(L\) is related to the input limit \(I_1\) and the capacity as shown above. So by substitution

\[
\frac{I_2}{\text{Capacity}} \leq L_2 - \frac{I_1}{\text{Capacity}}
\]

which can be expressed as

\[
\frac{I_1 + I_2}{\text{Capacity}} = \frac{\text{Total INPUT} + \text{QUEUE}}{\text{Capacity}} \leq L_2.
\]

So all the work to be done on the orders accepted into the company must be capable of being carried out within the longest lead time offered for any of the orders.

Alternatively, if the company wishes to accept all of the orders arriving in the time period and still
deliver them within \( L \) time periods, it has to use Output Control. It needs to increase the capacity, the output rate per time period from \( C \) to \( C_1 \), where

\[
\frac{\text{INPUT} + \text{QUEUE}}{C_1} \leq L
\]

so that

\[
C_1 \geq \frac{\text{INPUT} + \text{QUEUE}}{L}.
\]

Alternatively, one might exercise both Input and Output Control to increase the capacity by a certain amount and also turn away some orders.

The capacity is the output rate per time period. The time period is a fixed amount of calendar time, a day or a week or a month, etc. If a company only works for one shift of 8 hours per day, then the capacity can be increased by working overtime, since the work centre can be operated for a number of hours of overtime in addition to the normal shift during a calendar day. Alternatively, the work centre may have several machines, not all of which are usually in operation. So putting more operators to work on that centre increases the amount of processing work that can be done per day. Another possibility is subcontracting the processing work out to a sub-contractor. So, although time is divided into equal amounts, the capacity, the output of useful processing work that can be carried out in a time period, can vary from one time period from the next. The model and the methodology applied have to reflect this property of the situation. Thus the above analysis is better and more realistically expressed in discrete rather than continuous form.

Let the capacities in the current and future time periods be denoted by \( C(1), C(2), C(3), \ldots, C(L) \). Then if all the jobs are to be completed within \( L \) time periods, then the capacities and the INPUT of accepted orders must satisfy

\[
C(1) + C(2) + \cdots + C(L - 1) \leq \text{INPUT} + \text{QUEUE} \\
\leq C(1) + C(2) + \cdots + C(L - 1) + C(L).
\]

Output control then requires choosing the amount of capacity to provide in the current time period and all of the future \( L - 1 \) periods to follow.

3. Input–output control for the shop floor network of work centres at the order release stage

However, in the VMC sector of manufacturing and service customer orders turn into jobs that require processing on many work centres. The amount of transformation effort required on each work centre, the routing through the work centres and the sequence of work centres needed will differ from job to job. Jobs will arrive at and join the queue at any work centre not just from new orders but after having had transformation work carried out at other work centres. This means that there will be inputs of work into each work centre in future periods not just the current one. Thus the above simple model needs to be extended to this more complex situation to allow inputs of work period by period into the future from all jobs on the shop floor requiring transformation work on that work centre. In addition it has to allow outputs of work from that work centre period by period into the future, as these will be inputs to other work centres.

Any job will require a given amount of work, set-up time plus processing time, on each of a particular sequence of work centres. Note that it is possible that a job may need operations on a particular work centre more than once in the sequence. Each job will spend time queuing with other jobs in front of work centres waiting its turn to be processed. Given the varying routing sequences and varying processing times for the different jobs, some queuing is inevitable. Eliminating queues entirely is impossible without cutting significantly the total throughput of work that can be achieved per unit time. The simple analysis for a single work centre indicates that these queuing times are directly related to the amount of work that is taken on by the company. The exact relationship will be complicated by many factors, some deterministic and some stochastic, as well as the policies for dealing with customer enquiries and order acceptance and releasing work to the shop floor. These factors include the order arrival rates, the sizes of jobs and work centre processing rates, the reliability of the equipment at the work centres, the attendance records of the operators, the goodness of the management and control systems used etc.
The planned delivery date for any order will be the planned time it is released to the shop floor plus the work content of the job plus the buffer time for queuing and transportation between successive work centres summed over all the work centres required in its routing sequence. The actual values to use for the buffer times are to some extent also a management decision related to the delivery times the company generally wishes to offer customers. These values may be specific to the particular shop floor capacities and capabilities and will have to be consistent with the general mix of orders that it tends to receive and process. Ways to determine these are discussed later in Section 9.

Each job will require an amount of processing time to perform the transformation work necessary on a series of work centres, say \( L_{k,j} \) for work centre \( k \) for job \( j \) given by

\[
L_{k,j} = \text{set-up time} + \text{processing time}.
\]

The processing time will depend upon the size of the job. The buffer transfer time once the work on centre \( k - 1 \) has been completed to move the job to the next work centre \( k \) in its routing sequence and for it to join and wait in the queue of jobs in front of centre \( k \) is denoted by \( b_{k-1,k,j} \). Thus the time between leaving work centre \( k - 1 \) and centre \( k \) for job \( j \) will be \( L_{k,j} + b_{k-1,k,j} \). The time that a job leaves a work centre is denoted as the Operations Completion Date (OCD) of the job on that work centre. From the time that the job is released to the shop floor, then the job has a series of Operation Completion Dates, \( \text{OCD}_{k,j} \), at each of the work centres it will pass thorough. For the \( r \)th work centre in the routing sequence of a job, the \( \text{OCD}_{r,j} \) will be the OCD at the \( (r-1) \)th centre plus the set-up and processing times at that centre plus the buffer transfer time from centre \( r - 1 \) to \( r \), hence is the sum of the set-up, processing and buffer transfer times for all the earlier work centres in the sequence up to and including the \( r \)th one.

\[
\text{OCD}_{r,j} = \text{OCD}_{r-1,j} + L_{r-1,j} + b_{r-1,r,j} = \sum_{k=1}^{r} (L_{k,j} + b_{k-1,k,j}).
\]

Dividing time into discrete periods, for planning purposes at the customer enquiry and order acceptance stages, the job will thus impose work loads, \( L_{k,j} \), set-up time plus processing time, to be completed on each work centre in the differing future time periods in which the \( \text{OCD}_{k,j} \)'s fall. Let this \( r \)th work centre in the routing for job \( j \) be work centre \( n \) and let the period in which the OCD occurs be period \( t \). Then work centre \( n \) will have to have an output of work, \( L_{r,j} \), in period \( t \). There will be other jobs, each with their own routing sequences, currently on the shop floor being processed at other work centres, which will require processing on work centre \( n \) and thus have OCDs on centre \( n \) at different times. Summing the work loads of all such jobs with an OCD in period \( t \) at work centre \( n \) gives the total output of work which work centre \( n \) must provide in period \( t \). Let this total required output for work centre \( n \) in time period \( t \) be denoted by \( W_{n,t} \).

Now \( W_{n,t} \) is the output of work required from work centre \( n \) in time period \( t \) to ensure that the jobs conform to their planned OCDs and hence meet their planned delivery dates. Some jobs may have a total set-up and processing time on a work centre that fully occupies that work centre for several time periods. Alternatively the combined work of all jobs with an OCD on work centre \( n \) in time period \( t \) may exceed the processing capacity of the work centre within the time period. Thus the work needed may have to be carried out over several earlier time periods to meet the OCD requirement. Let \( Z_{n,t} \) be the actual work done on work centre \( n \) in period \( t \), defined in appropriate units of work, typically in machine hours. Similarly, let \( A_{n,t} \) be the sum of work already carried out before the current time on jobs done in advance of their OCDs. (This is equivalent to work in process inventories.) If all the OCDs for all the jobs involved are to be met, then the cumulative work carried out from now to any future period \( k \) plus the current work done in advance, \( A_{n,t} \), cannot be less than the cumulative outputs required over periods \( t = 1 \) to \( t = k \). Thus the \( Z_{n,t} \) values must also satisfy

\[
A_{n,1} + \sum_{t=1}^{k} Z_{n,t} \geq \sum_{t=1}^{k} W_{n,t} \quad \text{for all } k = 1, 2, 3, \ldots
\]

It is equal to or greater than inequality since some of the work over \( t = 1 \) to \( t = k \) may be advance
work for jobs with OCDs after period $k$. In addition, the work done in period $t$ cannot exceed the capacity, $C_{n,t}$ that has been provided, so that

$$Z_{n,t} \leq C_{n,t} \quad \text{for all } t.$$  

The released workload, $RB_{n,t}$, is defined as the sum of the work of all jobs currently on the shop floor at the start of period $t$ that still require processing to be carried out on work centre $n$ in future time periods:

$$RB_{n,t} = \sum_{k=t}^{\infty} W_{n,k}.$$  

In any period $t$, $Z_{n,t}$ units of work are actually carried out on work centre $n$. So the Released Workload will be reduced by $Z_{n,t}$ period by period, rather than the output required, $W_{n,t}$, and eventually to zero, starting with $RB_{n,1}$ as the released workload at the current time.

A common practice is to hold up jobs in the production planner’s office which have their material available and so can be released to the shop floor for processing work to be carried out, to allow the possibility of batch jobs into efficient production packages. This is known as the Job Pool. As with the transfer buffer times, it is assumed that there is a standard time, a pool delay, that jobs normally wait once the material has arrived, before being released to the shop floor. Let the time when the material has arrived and the job is available for transformation processing to start be defined as the Earliest Release Date (ERD). The job will then enter the shop floor and join the queue of work at the first work centre in its routing sequence at its ERD plus the pool delay, its start time for processing on the shop floor. A job is assumed to enter the released workloads of all the work centres on which the job requires processing at the same time. Given the time when a job in the pool is expected to enter the shop floor, allows its OCDs to be calculated as described earlier. The only difference is that the OCDs have the time from now to the ERD plus pool delay added to them initially rather than starting at zero. The work required at work centre $n$, calculated exactly as before for $L_{n,t}$, will then be added to the previous output needed at work centre $n$, $W_{n,t}$, in the time of its OCD. There will be a series of inputs of work to be added to the released workloads in future periods, up to the number of periods in the pool delay.

The planned workload $PB_{n,1}$ is defined as the total work of all jobs in the job pool as well as all the jobs on the shop floor that have not yet gone through work centre $n$ which require processing work on work centre $n$. Thus the planned workload is the sum of the work of all jobs currently in the pool plus jobs in the released workload. Assuming no further input of work into the pool, the planned workload will be continually reduced by the actual work done, $Z_{n,t}$, so it can be projected forward through time, period by period as

$$PB_{n,t+1} = PB_{n,t} - Z_{n,t} \quad \text{for } t = 1, 2, 3, \ldots$$  

The released workload for any work centre was defined as the sum of work required of all jobs on the shop floor that still needed to be processed on that work centre. The planned workload adds onto the released workload the work from jobs lying in the job pool in the production planner’s office awaiting release to the shop floor. This corresponds to the manufacturing lead time for a job, MLT, which is the time between the arrival of the materials that are being transformed and the completion of the transformation processing work and the delivery of the job to the customer. The MLT for a job is clearly under the direct control of the producer company because all the actions that need to be taken do not depend on the actions of any outside body.

The central basis of the methodology proposed is that the company wishes to offer a specified delivery performance to all its customers as part of its overall strategy. This will be part of the company’s ‘competitive edge’ criteria and will be what the customers in that market generally expect. This aim in general will take the form of a maximum delivery lead time being achieved for most jobs which in turn implies a maximum manufacturing lead time for all jobs. If this is to be achieved, then all the work currently in the system in the job pool and on the shop floor, which is the planned workload, must be all processed within this maximum manufacturing time, measured in periods and denoted by $T_P$. So the work performed on work centre $n$ period by period into the future, $Z_{n,t}$, for $t = 1, \ldots, T_P$, ...
must be such that the projected planned workload at the end of future period $T_p$ or more exactly at the start of period $T_p + 1$, will be zero. In addition, the planned workload can never become negative, so there are nonnegativity conditions

$$\text{PB}_{n,t+1} \geq 0 \quad \text{for } t = 1, 2, \ldots, T_p,$$

$$\text{PB}_{n,t} = 0 \quad \text{for } t = T_p + 1.$$

If the customer order is to be delivered on time, then the transformation work on every work centre needed by that job has to be completed within the delivery time promised. There may also be significant amounts of pre-production work to be done between the order being accepted and the job being released onto the shop floor. So the constraints on the inputs and outputs apply to every work centre and to every time period.

4. The hierarchy of lead times and associated workloads.

The total delivery lead time, DLT, from the customer’s point of view is the time between the receipt of the customer enquiry and its delivery to the customer. There are two other stages where jobs might be in a company that add further lead time onto the MLT just defined in addition to the time in the job pool or on the shop floor. Working backwards from the job pool, it may be a confirmed job, a bid in response to a customer enquiry that was accepted by the customer, which is awaiting the arrival of its raw materials and components and/or having the design and manufacturing configuration being specified. The producer company initially makes a bid in response to the customer enquiry. The customer may take some time to consider the bid and make a decision on whether to place a firm order with the producer company. So an order (job) may spend some time awaiting the customer decision.

Thus it can be seen that the jobs that are currently within the producer company, that it does know about, and hence can manage, fall into four different states:

1. a bid made in response to a customer enquiry and awaiting the customer decision,
2. a confirmed order awaiting the arrival of its raw material and/or having the design and manufacturing configuration being finally specified,
3. a confirmed job, with its material having been delivered, in the production planner’s office awaiting release onto the shop floor,
4. a job currently being processed at some work centre on the shop floor.

Moving forwards through time, materials will arrive so that the jobs currently in stage 2 will enter the pool in the production planner’s office ready for consideration for release in the following or subsequent periods. There will thus be inputs of extra work to increase the planned workload in future periods. All the jobs in stage 2 are confirmed, so the company knows the amounts of work that will have to be carried out on all the work centres. The date on which they were agreed and confirmed by the customer is known. When the material was ordered, the supplier gave a promised delivery date. From past data, the supplier’s delivery performance can be evaluated, so the mean and standard deviation of his lateness performance can be estimated. The mean plus say one standard deviation, as a buffer, can then be added to the promised material delivery time to give the predicted time, denoted by MAD, when that job will enter the pool in the production planner’s office. This is then its predicted ERD, at which the work for this job becomes an extra input of work to the planned workload. Thus there will be known inputs of work over future periods up to the longest MAD over all jobs awaiting material. Clearly, if a job requires materials to be ordered from several suppliers, the latest material arrival date is the one used above. Thus for all the jobs currently in Stage 2, the inputs of work for processing on work centre $n$ added onto the planned workload can be projected forward period by period. Once the ERD is known, the OCDs for all the work centres the job requires can also be calculated, so that the outputs of work required each time period, $W_{n,t}$, can also be updated. Each job in Stage 2 will need to be completed and delivered to the customer at the latest by its MAD plus the maximum manufacturing lead time, $T_p$. 


Hopefully, many of the bids made in response to the customer enquiries in Stage 1, awaiting a customer decision, will turn into confirmed jobs. A planned/predicted time for the customer to respond can be estimated from the customer history, again best taken as the mean plus one standard deviation. If it is a new customer, then the salesman will be required to provide an estimate. Thus for each enquiry in Stage 1, the job, if confirmed, can be predicted to enter the pool of jobs for release after the customer response time plus the job’s MAD, which will be its ERD. However, only a proportion of these enquiries will become confirmed jobs and work. The market segmentation and strike rate matrix analysis provide for each job an estimate of the probability of the customer accepting the bid, as described in Kingsman and Mercer [19]. The expected work from all the jobs in Stage 1 with an expected ERD in week \( t \) multiplied by the probability that the job enquiry is accepted by the customer becomes an input of work into the Planned Workload in week \( t \). This is another projected input of work to be added to the planned workloads in future periods. Similarly the outputs of work required, \( W_{n,t} \), can also be updated.

Let the projected inputs of work to work centre \( n \) in period \( t \) amount to \( I_{n,t} \). Thus the planned workload transition equation becomes

\[
P_{B_{n+1}} = P_{B_n} - Z_{n,t} + I_{n,t}
\]

where

\[
I_{n,t} = \text{Sum of work on work centre } n \text{ of all jobs awaiting material with ERD in period } t + \text{Sum over all jobs with ERD in period } t \text{ awaiting customer confirmation of the product (work on centre } n) \times (\text{probability customer accepts quotation})
\]

Thus there will be expected inputs into the planned workload from enquiries/jobs in stages (1) and (2) in future periods up to the maximum customer confirmation time, MMCT, plus the maximum of the material arrival times, MMAD. These will enter the pool by the latest at period MMCT + MMAD.

The different components of the overall delivery lead time for an order add up to give a hierarchy of lead times as illustrated in Fig. 1. The total delivery time, DLT, is thus made up of the customer confirmation time, the time awaiting materials and/or design work, the time spent in the Pool and the shop floor throughput time. The latter two sum up to the manufacturing lead time, MLT. Clearly both the MLT and the DLT must be controlled if the company is to maintain a good image in the market for reliability. The total workload of the producer company at any time is the sum of work of all jobs within the company wherever they may be. It clearly corresponds to the total delivery lead time, DLT, as defined above.

The different cumulative workloads, released, planned and total, also form a hierarchy corresponding to the various production lead times, as can also be seen in Fig. 1. If a FIFO policy is followed for processing orders through work centres in the factory then the MLT for a new order is the time it will take to process the whole of the current planned workload of work in the company. Similarly for the DLT and the total workload. Thus controlling the workloads of work at all times is one essential part of reducing significantly the stochastic variation in production and delivery times and achieving reliability in meeting promised delivery dates.

5. The need for input–output control at the customer enquiry and order entry stage

There are workloads, or backlogs of work, for every work centre on the shop floor which are the direct amount of work in the queue of jobs in front of the work centre plus the indirect work of all jobs at upstream work centres and in earlier stages of the overall process. If manufacturing lead times are to be controlled then the total work load has also to be controlled, not only overall but its occurrence over time. There are four levels at which this control of work can be attempted:

- Priority dispatching,
- Job release,
An extra stage, order acceptance, has been added to the three control levels that have been mentioned in past research. This is to deal with the situation of the customer taking a long time to consider and accept the bid made in response to the enquiry. By this time, other work may have been accepted and it may no longer be possible to produce that job to the delivery date agreed. Putting a time limit after which the bid lapses is one way of trying to deal with this problem in practice.

The main aim of workload control, for those who advocate its use as a job release method, has been defined as to control the lengths of the queues in front of work stations on the shop floor. However, the true objective is to process the jobs so as to meet the promised delivery dates with the machine and workforce capacities and capabilities available.

Long delivery lead times and a poor delivery reliability performance or poor capacity utilisation are not necessarily due to bad production scheduling. Their origin may well be wrong commitments of the sales department made at an earlier time that created major imbalances in the level of work in the company. The work in the factory at any time, which has to be processed and scheduled through production to meet attached delivery dates, is not some random sample given by an outside neutral environment. It is very much influenced by the actions of the producer company, since the lead times and income to cover costs are values to which the company has agreed. Much more is to be gained by a higher-level approach of integrating marketing and production planning effort at the customer enquiry stage so that one may try to avoid a lot of the detailed production scheduling problems in the first place. The aim has to be to use the process of bidding for orders to try to mould the order book into a shape that can be profitably manufactured.

Each job currently in the company in any of the above stages will have a promised delivery lead
time, when it will be shipped to the customer. The first aim of the workload control methodology is to ensure that these promised delivery lead times are achieved by planning forward the capacity to provide in each time period for each work centre. If the input of work is increased relative to the capacity to process work per unit time, then the MLT and DLT will inevitably increase. To maintain the same value for lead times, it is necessary to increase the rate at which work is completed per unit time, the capacity, for example by working overtime or subcontracting work out, or to reduce the input by turning work away for the moment. This typically will require adjustments to capacity at several work centres, not just one. The second aim of Workload Control is to determine what lead times are possible for the new customer and the actions, in terms of extra capacity, movement of operators and/or subcontracting etc., that are necessary to achieve particular delivery lead times. The actual manner in which input control is applied may be quite complex because the versatile manufacturing company will have a number of aims and considerations in mind when deciding how to respond to a customer enquiry.

The job release stage can itself only be fully effective if the queue of jobs in the pool is also controlled. Otherwise, jobs may remain in the pool for too long so missing their promised delivery dates. Thus a comprehensive workload control system must include the customer enquiry stage (the job entry stage) to ensure that the input of work to the pool as well as the shop floor queues are controlled. A methodology and systems to do this at both the job release and the customer enquiry stage have been devised by the author and his colleagues at Lancaster and described in several papers, see [13,20–23]. The purpose of this paper is to provide the elements of a basic theory for Input–Output modelling and workload control in a mathematical form to provide assistance with order bidding and dynamic capacity planning at the customer enquiry and order entry stages for produce-to-order companies.

6. The theory of input–output control at the order acceptance stage

All the work in the company in any of the four stages for each work centre gives the total workload of work the company has to process. The time required to process all of this work, plus the processing times for the new enquiry thus becomes the delivery time that will be required for a new enquiry made at the current time, if it is processed normally without any priority and without any change to the capacities currently planned for future periods. The planned workload can now be projected forward over the whole maximum delivery time offered, denoted by \( T_T \), rather than just the maximum manufacturing lead time, period \( T_p \). \( T_T \) is equal to \( T_p \) plus MMCT, the maximum customer confirmation time, plus MMAD, the maximum material arrival time. In order to achieve the company’s specified delivery objective for all jobs the total work involved must be completed over the time periods up to \( T_T \). Thus, for all work centres, \( n = 1 \) to \( N \), the planned workload at period \( T_T + 1 \) must be zero. Repeatedly applying the planned workload transition equation, this can be expressed as

\[
\sum_{t=1}^{T_T} Z_{n,t} = PB_{n,1} + \sum_{t=1}^{T_T} I_{n,t} \quad \text{for } n = 1 \text{ to } N. \quad (1)
\]

The left-hand side is the total work in the company at the current point in time that has to be carried out on work centre \( n \). This is the total workload of work for work centre \( n \), \( TB_n \). The actual amount of work done by work centre \( n \) in period \( t \), \( Z_{n,t} \), cannot exceed its available capacity, so

\[
Z_{n,t} \leq C_{n,t} \quad \text{for } t = 1 \text{ to } T_T, \quad n = 1 \text{ to } N. \quad (2)
\]

Thus combining the definition of \( TB_n \) and Eqs. (1) and (2) gives a simple initial constraint on the total workload of

\[
TB_n \leq \sum_{t=1}^{T_T} C_{n,t} \quad n = 1, 2, ..., N. \quad (3)
\]

This constraint provides a simple first check on whether there is capacity available to take on any new job or whether extra capacity will have to be provided. The new job will add work to the total
workload, the left-hand side of the above. If the increased total workload exceeds the cumulative capacity on any work centre over the longest allowable delivery lead time, $T_T$, then extra capacity has to be provided by some means or the job has to be rejected. This is the total workload control procedure, analogous to rough cut capacity planning in MRP systems. However, note, as will be seen, that this is only one of the constraints that the workloads in the system must satisfy.

Similarly all the output of work required over the maximum delivery lead time, $T_T$, must be achieved, so the $Z_{n,t}$ plus the initial advanced work must satisfy

$$A_{n,1} + \sum_{t=1}^{T_T} Z_{n,t} = \sum_{t=1}^{T_T} W_{n,t}$$

for all $n = 1$ to $N$. (4)

The planned workloads and advanced work for all work centres have to be nonnegative for all time periods in the range 1 to $T_T$, giving a further set of constraints on the work done each period:

$$PB_{n,t+1} = PB_{n,1} + \sum_{k=1}^{t} I_{n,k} - \sum_{k=1}^{t} Z_{n,k} \geq 0$$

for $t = 1$ to $T_T - 1$, $n = 1$ to $N$, (5)

and

$$A_{n,t+1} = A_{n,1} + \sum_{k=1}^{t} Z_{n,k} - \sum_{k=1}^{t} W_{n,k} \geq 0$$

for $t = 1$ to $T_T - 1$, $n = 1$ to $N$. (6)

with equality for $t = T_T$ as given in Eqs. (1) and (4). Eq. (6) is essential to ensure all jobs are completed by their OCDs. If in some circumstances, work could be delayed to be late by up to say $D$ time periods, then Eq. (6) would be modified to

$$A_{n,1} + \sum_{k=1}^{t} Z_{n,k} - \sum_{k=1}^{t-D} W_{n,k} \geq 0.$$ (6*)

As mentioned earlier, there might also be a constraint on how far ahead of its OCD a job can be done, both to minimise work in progress or to cover the problem of scheduling a job through the successive work centres it requires. If work cannot be done more than $d$ time periods in advance, then there is an upper bound constraint

$$A_{n,1} + \sum_{k=1}^{t} Z_{n,k} - \sum_{k=1}^{t-d} W_{n,k} \geq 0.$$ (6a)

The planned workload may initially increase over the immediate future because of the input of work from jobs in stages (1) and (2), before eventually continually falling to zero at the end of period $T_T$. This input of work may be quite variable. It is quite possible for a situation of little work currently on the shop floor or in the pool to occur but with large amounts of work in stages (1) and (2) that will enter the planned workload later. The central axiom of the methodology is that at any future time period, say $t + 1$, it must be possible to process the work in the planned workload at that time within the manufacturing lead time target set, $T_P$, from that point onwards, i.e. over the periods $t + 1$ to $t + T_P$. Otherwise, the promised delivery dates for those jobs still in the planned workload in period $k$ will not be met. This implies that for all time periods, the work processed by work centre $n$ from period $t + 1$ to period $t + T_P$ must be at least the projected planned workload on work centre $n$ at the start of period $t + 1$, provided $t + T_P$ less than the maximum planning horizon, $T_T$, i.e.

$$\sum_{k=t+1}^{t+T_P} Z_{n,k} \geq PB_{n,t+1} \text{ for all } t = 1 \text{ to } T_T - T_P.$$ (7a)

and

$$\sum_{k=t+1}^{T_T} Z_{n,k} \geq PB_{n,t+1} \text{ for all } t = T_T - T_P \text{ to } T_T.$$ (7b)

Substituting for $PB_{n,k}$ using the planned workload transition equation and simplifying gives

$$PB_{n,1} + \sum_{k=1}^{t} I_{n,k} \leq \sum_{k=1}^{t+T_T} Z_{n,k}$$

for all $t = 1 \text{ to } T_T - T_P$, (7)

or alternatively

$$PB_{n,1} + \sum_{k=1}^{t} I_{n,k} \leq \sum_{k=1}^{t+T_T} Z_{n,k}$$

for all $t = 1 \text{ to } T_T - T_P$. 

From Eqs. (5) and (6), it follows that
\[
\sum_{k=1}^{t} Z_{n,k} \geq PB_{n,1} + \sum_{k=1}^{t-T_P} I_{n,k} \text{ for all } t = T_P \text{ to } T_T
\]
for all \( t = 1 \) to \( T_T - T_P \).

This is the standard requirement given in most discussions of workload control that the cumulative input of work, including uncompleted work currently on the shop floor, must exceed the cumulative output of work, having allowed for advance work done, for all periods into the future. This is shown in Fig. 2. The introduction of the actual work done, \( Z_{n,k} \), is new and as will be seen, helps significantly to plan dynamically the capacity to provide in each period. Eq. (7) can also be represented in Fig. 2. It indicates that at any period the cumulative actual work done must exceed the cumulative input of work \( T_P \) periods earlier. The ‘horizontal distance’ in periods between the same value for the cumulative input and cumulative output is often called the planned backlog length. It is this factor that has to be controlled to be less than the manufacturing time for every period to meet agreed delivery dates. It can be seen easily from Fig. 2, that if the cumulative input of work equals the cumulative output of work required for any period \( t \), then the problem breaks down into independent sub-problems as regards feasibility and capacity planning.

7. The bidding and order acceptance planning problem

The problem may be to check if a new enquiry can be manufactured within given possible delivery lead times. It is then merely a question of increasing the outputs, \( W_{n,t} \), and the inputs, \( I_{n,t} \), for the relevant work centres in the relevant time periods and checking whether there is a solution satisfying the constraints. It is necessary to ensure that the variables \( Z_{n,t} \), the actual amounts of work to carry out at work centre \( n \) in time \( t \), satisfy the constraints (1)–(7) above with all being nonnegative, for all time periods up to the maximum delivery lead time offered as a policy and all work centres. However, in general, there will be a need to adjust the capacities at work centres over time to do the processing work for the job being considered. Thus, the important decision variables are the capacity to provide at each work centre in each future period, \( C_{n,t} \). The other decision variables, the \( Z_{n,t} \), are less important directly.

Consider a situation of 11 jobs, all to be completed over the next six weeks, requiring processing through one or both of two work centres having OCDs and requiring work, measured in set-up and processing hours, on the work centres as given in Table 1. The second column of the table shows the ERD, when processing of the job on the shop floor may commence. Clearly no work at any work centre can start until this time. Note that jobs \( H, I, J \) and \( K \) require some initial work at a subcontractor so having longer than usual material arrival times and thus later ERDs. For simplicity, it is assumed that there is currently no work on the shop floor and no advance work so \( P_{n,1} = 0 \) and \( AW_{n,1} = 0 \) for both work centres.

Using the ideas described earlier, the eleven jobs give aggregate inputs and outputs of work week by week for the two work centres as shown in Table 2. Table 2 also shows the available capacity week by week. The input times, the arrival of the material, are important since work on a job cannot be started before its input time.

There are 20 hours of spare capacity on centre 1 and 5 hours on centre 2 over the six weeks, so the total workload constraint, Eq. (3) (rough cut capacity planning) is satisfied. The cumulative inputs...
and outputs satisfy the constraints (5) and (6) for all weeks and, in addition, the cumulative outputs are lower than the cumulative capacities in each week. Thus it is possible to complete all of the jobs in time to meet the given delivery dates. There are several alternative solutions. Table 3 shows two examples, assuming processing work is carried out as early as possible and assuming that it is as late as possible.

Consider the situation if another job $L$, requiring delivery at the end of week 6 was offered, with an ERD in week 3, requiring 15 hours of work on centre 1 with an OCD in week 3 and 5 hours of work on centre 2 with an OCD in week 5 were offered. Rough cut capacity planning would suggest that the job should be accepted and could be delivered at the end of week 5. This is because the work needed is less than the 20 hours spare on centre 1 and equal to the 5 hours spare on centre 2. However, it is clear from Table 4, which results from adding the new work to that of Table 2, that it is impossible to meet the required delivery time. Including job $L$ increases the input on work centre 1 in week 3 to 15 hours and the output in week 3 to 25 hours. Accepting the job would require a cumulative output to week 3 of 45 hours of work whilst there is only 40 hours of capacity. Note that although there is a shortage of 5 hours of capacity on centre 1 in week 3, there are still 5 hours of spare capacity in both weeks 5 and 6. Thus job $L$ cannot be completed until the end of week 5 or alternatively 5 extra hours of processing capacity must be provided in week 3, unless other jobs are delayed. Currently there are 5 hours of spare capacity in week 3 or 4 on centre 2, so the work on job $L$ for centre 2 could be completed, provided it has left centre 1 in week 4. This cannot happen currently. So the job will be delivered well after week 6 if no increase in capacity is provided.

---

Table 1
Example data

<table>
<thead>
<tr>
<th>Job</th>
<th>ERD (wk)</th>
<th>Centre 1</th>
<th>Centre 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OCDs (wk)</td>
<td>Work (processing h)</td>
<td>OCDs (wk)</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>F</td>
<td>4</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>G</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>H</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>I</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>6</td>
<td>6</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2
The aggregate workloads week by week for example data

<table>
<thead>
<tr>
<th>Week</th>
<th>Aggregate input, $I_{n,t}$</th>
<th>Cumulative input</th>
<th>Aggregate output, $W_{n,t}$</th>
<th>Cumulative output</th>
<th>Capacity</th>
<th>Cumulative capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centre 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>30</td>
<td>20</td>
<td>20</td>
<td>25</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>30</td>
<td>10</td>
<td>30</td>
<td>5</td>
<td>40</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>45</td>
<td>10</td>
<td>40</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>65</td>
<td>15</td>
<td>55</td>
<td>20</td>
<td>70</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>75</td>
<td>20</td>
<td>75</td>
<td>25</td>
<td>95</td>
</tr>
<tr>
<td>Centre 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>20</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>25</td>
<td>10</td>
<td>20</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>35</td>
<td>5</td>
<td>25</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>35</td>
<td>10</td>
<td>35</td>
<td>10</td>
<td>40</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>55</td>
<td>10</td>
<td>45</td>
<td>10</td>
<td>50</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>55</td>
<td>10</td>
<td>55</td>
<td>10</td>
<td>60</td>
</tr>
</tbody>
</table>
Table 3
The actual work done each week, \( Z_{n,t} \)

<table>
<thead>
<tr>
<th>Week</th>
<th>Centre 1 Work as early as possible</th>
<th>Centre 1 Work as late as possible</th>
<th>Centre 2 Work as early as possible</th>
<th>Centre 2 Work as late as possible</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Work ( Z_{1,t} )</td>
<td>Spare capacity</td>
<td>Work ( Z_{2,t} )</td>
<td>Spare capacity</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>5</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>5</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>15</td>
<td>10</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4
Revised aggregate work loads including new job

<table>
<thead>
<tr>
<th>Week</th>
<th>Aggregate input, ( I_{n,t} )</th>
<th>Cumulative input</th>
<th>Aggregate output, ( W_{n,t} )</th>
<th>Cumulative output</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Centre 1</td>
<td>Centre 2</td>
<td>Centre 1</td>
<td>Centre 2</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>30</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>0 + 15</td>
<td>45</td>
<td>10 + 15</td>
<td>45</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>60</td>
<td>10</td>
<td>55</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>80</td>
<td>15</td>
<td>70</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>90</td>
<td>20</td>
<td>90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Week</th>
<th>Centre 2</th>
<th>Centre 2</th>
<th>Centre 1</th>
<th>Centre 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>20</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>25</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>10 + 5</td>
<td>40</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>40</td>
<td>10</td>
<td>35</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>60</td>
<td>10 + 5</td>
<td>50</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>55</td>
<td>10</td>
<td>60</td>
</tr>
</tbody>
</table>

This very simple example confirms the thesis of this paper, that if meeting the promised delivery time is crucial, as is the case for versatile manufacturing companies, then the time profile of the aggregate loads of work has to be considered directly. It is just impossible in this example, because of the timing of the arrival of orders and the OCDs of the jobs on the work centres, to meet the specified delivery times for all jobs. Using the aggregate workloads and cumulative inputs, work and outputs clearly identifies where the problem lies. It indicates where the shortfall of capacity occurs or when the earliest delivery is possible. Either 5 hours of extra processing capacity has to be provided for centre 1 in week 3 or job L will be delivered at least 2 weeks late, depending on the capacity of centre 2. Alternatively, if the operators are multi-skilled and could be moved between centres 1 and 2, it may be possible that the 5 hours on centre 1 in week 3 could be provided from centre 2 and the shortfall made up by transferring 5 hours from centre 1 to centre 2 in week 5, since there is 5 hours spare on centre 1 in both weeks 5 and 6.

Clearly the above analysis has to be done quickly, since it has to be repeated in real time for every customer enquiry. A formal planning system is required to do the analysis in most cases since there are several work centres, and deliveries and pro-
8. Dealing with a new enquiry and planning capacity decisions

The enquiry may come with a fixed delivery date and merely ask for a price. In this case the first step is to check whether this can be achieved. Taking the specified delivery date, backwards scheduling through the work centres needed gives the latest release date (LRD) the job has to leave the pool. Its ERD will be this LRD minus the pool delay. The time from now to the ERD has to be checked to ensure it is sufficient to cover the customer confirmation time plus the delivery of the necessary material. If so, then this delivery date can be accepted and it is a question of adding the work to the above model to determine what extra capacities need to be provided and where and when. If not, then the job will have to be given priority by not being delayed in the pool and by going to the heads of the queues at the work centres required, i.e. not adding the transfer buffer times to its manufacturing lead time. Again, it will be necessary to check that there is sufficient capacity at the right times to do the work required for the job to meet this new delivery lead time.

If the enquiry requests both a delivery date and a price to be quoted, then the first step is to calculate possible alternative delivery dates. Using the standard approach gives a delivery based on normal processing with standard transfer buffer times. Shorter delivery times can be achieved by giving the job priority at some of the work centres it requires, i.e. saving some transfer buffer times, and/or in the pool. Thus a series of possible alternative delivery times can be generated. Each of these alternatives will have its associated ERDs and hence sets of inputs and outputs of work required on the work centres needed in specified periods. If the existing capacity is sufficient or if the extra capacity needed can be supplied, then the job can be produced and delivered at the lead time being considered, but perhaps at a higher cost than normal.

Given an existing solution, i.e. set of values for all the parameters in the model, the analysis can be performed in terms of only the changes made by the new job for the particular alternative lead time considered. This means that only those constraints that are affected by the new job, on work centres and time periods, need to be considered in a reduced submodel to determine what changes to the $Z_{n,t}$ variables and the capacity variables need to be made to manufacture the job that is the subject of this new enquiry.

In general, there may be many alternative possible solutions that satisfy the constraints on the outputs and capacities. The interest may only be in finding a feasible solution. However, given several alternative possible solutions it is preferable to select one which best meets the company objectives. One could envisage several objectives that the versatile manufacturing company might like to try to achieve, which could all be formulated as a simple linear function of the variables of the problem. Clearly as the company wishes to keep the price as low as possible, so finding the way of providing the extra capacity needed at minimum cost might be an appropriate aim. One could attempt to minimise the amount of overtime hours worked over the planning period up to period $T_T$ or alternatively minimise the cost of overtime working. Defining $h_n$ as the overtime cost per hour on centre $n$, these objectives can be expressed as

$$\sum_{n=1}^{N} \sum_{i=1}^{T_T} OT_{n,t} \text{ and } \sum_{n=1}^{N} \sum_{i=1}^{T_T} \{h_n \cdot OT_{n,t}\}.$$

Another alternative could be to minimise the amount of work in progress. This will be all the work done in advance of the OCDs of the jobs and so could be expressed in terms of the variables $Z_{n,t}$ as

$$\text{cons} \tan t + \sum_{n=1}^{N} \sum_{i=1}^{T_T} (T_T + 1 - t)Z_{n,t}.$$

This aim thus implies that, wherever there is any choice, work should be done as late as possible.
Further alternative objectives are to smooth out the work over time, by minimising the maximum value for overtime working in any period, or minimise the number of changes of operators at a work centre from period to period, again probably best as a mini-max criterion.

The capacity is the maximum number of units of work that can be performed by a work centre in a unit time period. Capacity can be increased/decreased by changing the number of operators assigned to the work centre and also increased by overtime working. The capacity will depend upon the number of operators and the number of individual machines at the work centre. The relationship between capacity, operators and machines may be nonlinear. For example, one operator may be available to operate two machines but not necessarily produce twice the output of one machine. Furthermore, at least for normal time working, operators most likely will have to be allocated to work centres for blocks of time, say a shift or half a shift, rather than any for any general number of hours such as 2.3 hours.

The capacity available will be the sum of the normal time processing hours capacity, $\text{NT}_{n,t}$, and the overtime processing capacity $\text{OT}_{n,t}$, that is provided. Hence constraint (2) now takes the form

$$Z_{n,t} \leq C_{n,t} = \text{NT}_{n,t} + \text{OT}_{n,t}.$$  \hspace{1cm} (2*)

It is assumed that there may be circumstances where a work centre with a low load of work could only process jobs in overtime, since the usual operators have been transferred to other overloaded work centres during their normal working hours. Alternatively work centres could have one operator working during normal time and two or more in overtime. So in general there should be no constraints such that overtime working only applies if the normal time working is positive and fully utilised, but such a constraint could easily be imposed if required. For normal time working, the amount of capacity available needs to be defined in terms of the number of appropriate operator allocation time blocks provided, e.g. half shifts of 4 hours say. This can be done in a formal way through further variables and constraints, which enables the model to be defined in terms of working hours and the allocation of operators to work centres period by period.

The objective of the formal model is to find the variables that satisfy constraints (1)–(8), plus those added to define capacity, which minimise the chosen objective function. This could be done manually trying out possible solutions based on past experience checking if they satisfy the constraints. Alternatively, the problem can be presented as a mixed integer linear programming problem for a given set of ERD and OCDs associated with a particular delivery lead time. Hence it could be solved using any standard LP package. Different delivery times imply different ERD and OCDs. So, this would have to be run several times to find the 'costs' for alternative sets of delivery lead times. A mixed integer LP formulation of the example problem of Table 1, including the new job $L$, is given in the Appendix.

However, the practical problem is to deal with enquiries in real time, determining if and when they can be processed through the shop floor, what alternative lead times are feasible and what extra costs in providing extra capacity will be incurred. Solving several large integer LPs in real time may be quite time consuming.

The dynamics of the situation need to be taken into account. New orders will continue to arrive randomly. The real objective is to take on and produce as much profitable work not only now but in the future. So it could be argued, where there is a choice, that it is best to plan to do the current work in the system, the current total workload, as early as possible, as it is likely to leave more capacity later over the planning horizon, for taking on the new jobs that will arrive in the future, so increasing overall throughput and profit.

In the particular case studied in this research, a heuristic procedure was devised to allow the option of either doing work as late as possible or doing work as early as possible. These objectives were preferred by the management of the company collaborating in this research. The management preferred to make use of their experience and judgement to try out possible options for increasing capacity or extending the delivery due date rather than having this done as part of a computer based system. The system implemented is shown in Fig. 3.
It was entitled as WORKCON. The initial step in the upper part of the figure when faced with a new enquiry was to ensure that there was no overload with the existing work and enquiries in the factory. If there had been problems so that the planned outputs were not achieved, in any time period, then the advanced work and outputs were adjusted so that the system represented the actual current situation in the factory at this point in time. Once it was clear that the factory did not have an overload of work currently, the system began considering the new enquiry as outlined in the lower half of the figure. If the enquiry can be fitted into the factory schedule management can then decide the price and delivery date to quote according to the need for overtime work on that enquiry. Note that for an order already accepted the only option if extra capacity cannot be provided is to increase the pre-production lead time, so that it is known at an early stage that the order will be delivered late. In dealing

Fig. 3. The process and decision model for dealing with enquiries.
with a new enquiry by contrast, there is the option of rejecting the enquiry.

9. Setting the buffer transfer times

Deriving a good estimate of the average queuing time at work centres is necessary for the planned lead times to be a good predictor of the actuals. An initial value can be derived by management setting the maximum MLT it will offer for the largest acceptable order. Expressing this in working days, deducting the actual set-up and processing times at all work centres and dividing the remainder by the average number of work centres per job will give an initial ‘buffer norm’. This then needs to be validated by simulation. The aim of the simulation is to do a directed search of the buffer transfer times so that the average and standard deviation of the planned delivery times determined of the enquiries received are equal to the average and standard deviation of the actual delivery times coming out of the simulation model. It is necessary to work in terms of equality of average, etc., over the total enquiries processed rather than each individual enquiry.

10. Conclusions

Production planning and order acceptance are difficult management problems in produce-to-order companies because the arrival of orders into the company is a stochastic process. The arrival of enquiries cannot be predicted in advance. Whether an enquiry turns into an order depends upon the bid the company makes in response to a customer enquiry and how it compares with bids from competitors. Furthermore, each enquiry (order) tends to be different, requiring different amounts of processing work on the work centres and in a different routing sequence. Managing lead times using work load input/output control methods based on controlling a hierarchy of aggregate loads of work is a better approach than using forecast lead times. A theory for workload control that provides a mathematical model which can be used for dynamically planning capacity has been presented. The value for reducing the stochastic variability in delivery lead times, and/or obtaining more consistent use of the production capacity has been substantiated by simulation results for the situation of a sub-contracting MTO manufacturer.

Appendix A. Formulating the example data as a mixed integer programming problem

A.1. The work load constraints on Centre 1

Constraint (1) that total work done equals total input and constraint (4) that total work done equals total outputs less advance work already complete are the same for the example;

\[ Z_{11} + Z_{12} + Z_{13} + Z_{14} + Z_{15} + Z_{16} = 90. \]

Constraints (2):

\[ Z_{11} \leq C_{11}, Z_{12} \leq C_{12}, Z_{13} \leq C_{13}, \]
\[ Z_{14} \leq C_{14}, Z_{15} \leq C_{15}, Z_{16} \leq C_{16}. \]

When the capacities have to be determined, then constraint (3) that the total workload has to be less than the known cumulative capacity does not apply.

Constraints (5), cumulative work less than or equal to cumulative inputs:

\[ Z_{11} \leq 10, \]
\[ Z_{11} + Z_{12} \leq 30, \]
\[ Z_{11} + Z_{12} + Z_{13} \leq 45, \]
\[ Z_{11} + Z_{12} + Z_{13} + Z_{14} \leq 60, \]
\[ Z_{11} + Z_{12} + Z_{13} + Z_{14} + Z_{15} \leq 90. \]

Constraints (6), cumulative work greater than or equal to cumulative outputs:

\[ Z_{11} \geq 0, \]
\[ Z_{11} + Z_{12} \geq 20, \]
\[ Z_{11} + Z_{12} + Z_{13} \geq 45, \]
\[ Z_{11} + Z_{12} + Z_{13} + Z_{14} \geq 55, \]
\[ Z_{11} + Z_{12} + Z_{13} + Z_{14} + Z_{15} \geq 70. \]
A.2. The work load constraints on Centre 2

Constraints (7), assuming that the manufacturing lead time standard is set at three periods,

\[ Z_{11} + Z_{12} + Z_{13} \geq 10, \]
\[ Z_{11} + Z_{12} + Z_{13} + Z_{14} \geq 30, \]
\[ Z_{11} + Z_{12} + Z_{13} + Z_{14} + Z_{15} \geq 45, \]
\[ Z_{11} + Z_{12} + Z_{13} + Z_{14} + Z_{15} + Z_{16} \geq 60. \]

In this example, as can be clearly seen, constraints (7) are all redundant by virtue of constraints (6). This will not always be true of all practical problems.

\section*{A.2. The work load constraints on Centre 2}

Note that the same simplifications as for centre 1 can be made.

Constraint (1) is

\[ Z_{21} + Z_{22} + Z_{23} + Z_{24} + Z_{25} + Z_{26} = 60. \]

Constraints (2):

\[ Z_{21} \leq C_{21}, \quad Z_{22} \leq C_{22}, \quad Z_{23} \leq C_{23}, \]
\[ Z_{24} \leq C_{24}, \quad Z_{25} \leq C_{25}, \quad Z_{26} \leq C_{26}. \]

Constraints (5), cumulative work less than or equal to cumulative inputs:

\[ Z_{21} \leq 20, \]
\[ Z_{21} + Z_{22} \leq 25, \]
\[ Z_{21} + Z_{22} + Z_{23} \leq 40, \]
\[ Z_{21} + Z_{22} + Z_{23} + Z_{24} \leq 40, \]
\[ Z_{21} + Z_{22} + Z_{23} + Z_{24} + Z_{25} \leq 60. \]

Constraints (6), cumulative work greater than or equal to cumulative outputs:

\[ Z_{21} \geq 10, \]
\[ Z_{21} + Z_{22} \geq 20, \]
\[ Z_{21} + Z_{22} + Z_{23} \geq 25, \]
\[ Z_{21} + Z_{22} + Z_{23} + Z_{24} \geq 35, \]
\[ Z_{21} + Z_{22} + Z_{23} + Z_{24} + Z_{25} \geq 50. \]

Again in this example for Centre 2 also, constraints (7) are all redundant by virtue of constraints (6).

For the example of Table 1, including job \( L \), let us assume that the capacity is directly related to the manning of the work centres. Furthermore, let us assume for simplicity that one operator looks after both work centres, but can vary the time between the centres as required up to the limit of hours available per period. Also we assume that the operator is allocated to work centres in blocks of 4 hours. The total hours will thus be the sum of the capacities provided, see Table 2.

\begin{center}
\begin{tabular}{ccccccc}
Week & 1 & 2 & 3 & 4 & 5 & 6 \\
Times & 20 & 36 & 12 & 20 & 28 & 32 \\
\end{tabular}
\end{center}

Let \( B_{j,n,t} \) be the number of 4 hours time blocks planned on work centre \( n \) in period \( t \) with \( j \) operators assigned, the capacity variable to be determined. In addition let us assume that the maximum hours that can be worked on Centre 1 is 28 hours and on Centre 2 is 16 hours. Also it is assumed that each real time hour = one centre processing hour for all \( n \) and \( t \). Thus there are constraints on time blocks on centres of the form

\[ B_{1,1,t} \leq 7 \quad \text{for} \ t = 1 \text{ to } 6, \text{ for Centre 1} \]

and

\[ B_{1,2,t} \leq 4 \quad \text{for} \ t = 1 \text{ to } 6, \text{ for Centre 2}. \]

Constraints (2*) and (9) on the capacity in processing hours take the form

\[ C_{1,t} \leq 4B_{1,1,t} \]

and

\[ C_{2,t} \leq 4B_{1,2,t} \quad \text{for} \ t = 1 \text{ to } 6 \]

and since there is only 1 worker involved, so the total number of worker time blocks in week 1 will be \( B_{1,1,1} + B_{1,2,1} \). There are only \( 20/4 = 5 \) time blocks available in week 1, so giving the constraint

\[ B_{1,1,1} + B_{1,2,1} \leq 5 \]

and similarly for weeks 2 to 6,

\[ B_{1,1,2} + B_{1,2,2} \leq 9, \]
\[ B_{1,1,3} + B_{1,2,3} \leq 3, \]
\[ B_{1,1,4} + B_{1,2,4} \leq 5, \]
\[ B_{1,1,5} + B_{1,2,5} \leq 7, \]
\[ B_{1,1,6} + B_{1,2,6} \leq 8. \]
The problem is then to find the nonnegative values of the continuous variables $Z_{n,t}$ and $C_{n,t}$ and the integer variables $B_{1,n,t}$ for $n = 1$ to $2$ and $t = 1$ to $6$ that satisfy the above constraints and minimise or maximise the chosen objective function; for example minimise the work in progress given by

$$\sum_{n=1}^{2} \sum_{t=1}^{6} (7 - t)Z_{n,t}.$$ 

References


