Expediting as a Two-Edged Sword

Eli Schragenheim and Boaz Ronen

ABSTRACT

Expediting is a common manufacturing practice for keeping good due-date performance. The question addressed here is how much expediting is needed to get the best due-date performance with a given planning/control system. Computer simulations, under the planning scheme of drum-buffer-rope coupled with the control mechanism of buffer management, were used to investigate the impact of two expediting schemes on global due-date performance.

Eli Schragenheim is a management consultant in Israel. He combines his expertise in management information systems with the methodologies derived from the theory of constraints (TOC). He has also developed a variety of computerized management games and simulators which serve both for management education and research. Eli Schragenheim is a former partner with The A.Y. Goldratt Institute. He holds an MBA degree from Tel-Aviv University, where he is a Ph.D. student at the Faculty of Management.

Boaz Ronen is an Associate Professor at Tel Aviv University, Faculty of Management, The Leon Recanati Graduate School of Business Administration. He holds a B.Sc. in electronics engineering from Technion, Haifa and an M.Sc. and Ph.D. from Tel Aviv University, The Leon Recanati Graduate School of Business Administration. He was a visiting professor at New York University, The Stern School of Business, Information Systems Area, and in Columbia University, Graduate Business School with the Operations Management Area. Prior to his Academic Career he worked for ten years in the Hi-Tech electronics industry. His main research, teaching and consulting areas are the implementation of TOC, JIT and TQM in industrial, service and military organizations. His papers (more than 40) were published in research journals such as Management Science, Operations Research, Decision Sciences, and in practitioners' journals such as Datamation, Production and Inventory Management Journal, Industrial Engineering and Industrial Management.

INTRODUCTION

Expediting is a common manufacturing practice for keeping good due-date performance by rushing purchasing and/or production. It occurs whenever some of the orders were placed late by the client or when delays have affected the actual lead time of some orders. Expediting is done because it is assumed that by changing priorities and putting a lot of pressure on certain orders the due-date performance will be improved. Is it a valid assumption? It seems probable that the particular expedited orders will arrive earlier than without expediting. However, the question remains - will the global due-date performance improve as a result of expediting? How much expediting is needed to get the best due-date performance with a given planning/control system?

We assume that the main task of expediting is to assist in meeting all the due-dates as promised to the clients. It is not intended to reduce the average production lead-time, on the contrary, it may enlarge the average lead-time. As long as it improves the due-date performance it should be used.

We've tested the effect of expediting on the global due-date performance, using a series of simulations. The planning scheme used was the Drum-Buffer-Rope (DBR) technique (Schragenheim & Ronen, 1990). The control mechanism was the Buffer Management (BM) technique, described in Schragenheim and Ronen (1991). BM is a diagnostic tool to point out the orders for expediting. The assumption that expediting can enhance the due-date performance is embedded in the BM methodology. The DBR/BM methodologies have been successfully implemented in several types of manufacturing shop floors.

The second section deals with the research aims, assumptions and the initial expectations. The simulated environment and the decision rules are described in section 3. The results are displayed in section 4, conclusions are drawn in section 5 and suggestions for further research in section 6.

RESEARCH AIM, HYPOTHESIS, AND ASSUMPTIONS

The research aim is to set practical rules for implementing a good expediting scheme. A "good" expediting method means that the resulted due-date performance (DDP) is better than without any expediting. It is assumed that no production manager will make extensive tests in order to find the optimum expediting scheme. So, some simple and straightforward observations of the current situation should be enough to decide whether the current scheme is "good enough".

The environment chosen is planned by the DBR method (Schragenheim & Ronen, 1990). It does not have a capacity constraint. According to the DBR methodology whenever no capacity constraint exists the only schedule to implement is the release of the raw materials which are a fixed 'buffer time' prior to the due-date. The term 'buffer time' stands for an estimation of the "almost worst case lead time" which is interpreted as the average lead-time plus two or three standard deviations. We assume that the most substantial "investment" in providing the protection of the due-dates is the length of the time-buffer. The time-buffer, according to DBR methodology, is the ultimate production lead-time. It directly affects the inventory level and impacts the flexibility of the quoted lead-times determined by the marketing function.

Buffer management (Schragenheim & Ronen, 1991) serves as the control method. It has three tasks: to point out the orders which would be late unless expedited, to enable a statistical analysis of the stability of both planning and control and to point out the problematic work centers that regularly threaten the due-date performance. In this research we'll focus on the first task of BM - the identification of the orders to be expedited.

The measure of the DDP used here is the total number of late-order-days (LOD) in a given period of time. LOD means the sum of the lateness, in days, incurred by each late order. This measure is similar to throughput-dollar-days measurement suggested by Goldratt (1989) but without the dollar values. This measure was chosen because it is fairly simple and combines the number of missed orders with the number of days those orders were late. The assumption is that the damage incurred by a late order does not correspond to the money value of the order. Hence, all orders are supposed to be of equal importance.

The control on the amount of expediting is given by the length of the 'expediting zone' (Schragenheim & Ronen, 1991). The idea is to assign the term 'almost late order' to an order whose due-date is too near, that is, within a specified number of hours called the 'expediting zone'. Any order whose due-date is within this expediting zone is considered worthy of expediting efforts.

We assume that the expediting rules involve 'breaking' setups or changing a preferred sequence of operations. In other words, an expediting scheme wastes capacity from certain resources. When we enlarge the 'expediting zone' - two opposite impacts on the mean late-order-days (LOD) are created. These impacts may be shown in the form of a cause and effect tree (see Figure 1). From the basic cause - enlarging the 'expediting zone' - the first impact is drawn upwards and the other one downward.

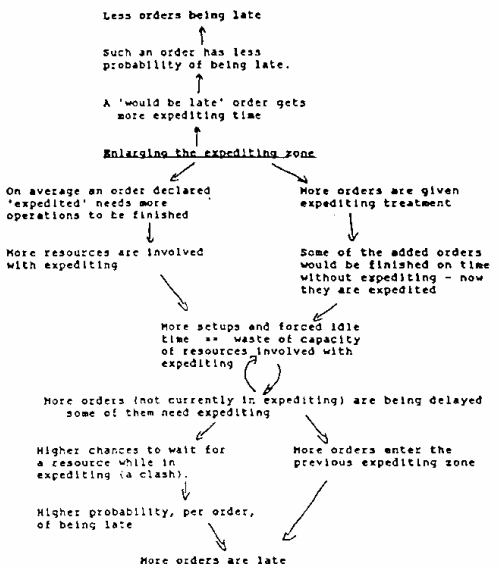

Figure 1. A Cause and Effect Tree

The upper direction outlines the impact on the orders which would have been late in the previous expediting zone. Such orders may be on time, in the enlarged expediting zone scheme, because more expediting time is provided for them. The lower direction points out that this 'extra' expediting may cause the whole system due- date performance to deteriorate.

A key term in the above argument is a 'clash'. This stands for a description of the situation where all the units of a certain resource are occupied in expediting, while there is yet another order in need of expediting for which a unit of that resource is required. There are two immediate causes for a higher number of LOD when the expediting zone is enlarged.

1. The higher number of clashes delay expedited orders thus lowering the probability of being on time.
2. There are additional delays prior to the expediting zone because certain resources are more loaded than before. More orders are pushed into the expediting zone. The additional expediting orders are exposed to the risk that the expediting time provided will not be enough.

We make two hypotheses.

a. When the expedited zone is enlarged, starting with zero, the expected number of LOD will drop at first because the

net result will be more impacted by the additional expedite time than by the increased load of some critical resources. However, at a certain size of the expediting zone, the LOD will start to rise due to lack of capacity of certain resources. In other words we expect to find an optimal size of the 'expediting zone' resulting in minimal LOD.

b. As the 'expedited zone' is enlarged, the clashes will be the most significant cause for the LOD.

If the second hypothesis is valid - then when the number of clashes becomes significant, the DDP will start to deteriorate. This will enable the production manager to assess the impact of the current 'expediting zone' on the DDP according to the amount of clashes. We should also expect, provided the second hypothesis is valid, that a significant correlation will be found between the number of clashes and the LOD. When the number of clashes is significant - it is speculated by the second hypothesis that the expediting zone is already too large. Goldratt and Fox (1986) suggested that a 'good' criterion, for the expediting zone, should be about one third of the buffer-time. This guess is checked, for the particular environment.

THE SIMULATED ENVIRONMENT

Two schemes of expediting were checked: scheme A and scheme B. Scheme A used more extreme measures to expedite than scheme B, but it demands more from the relevant resources. Scheme A expedites faster. Scheme B is more careful in using extra capacity for expediting. Our initial assumption was that scheme B would have a minimum LOD at a larger expediting zone, because of its reluctancy to waste capacity. This assumption has led us to believe that this scheme might reach, in its optimal value, a better DDP than scheme A, because more orders will be expedited with the same extra capacity for expediting.

The independent variable, within each expediting scheme, is how many hours prior to the due-hour of shipment are considered 'almost late' hence worthy of expediting. This variable directly impacts the number of orders, or rather order-hours, being expedited. The dependent variable is the total number of late-order-days (LOD) reported within the 10 weeks of simulation run.

Figure 2 displays the work centers used in the simulation. Each work center is composed of one to three identical machines that are unique to that work center. Figure 3 displays the routing of the seven different products, divided between three families of products. The simulated plant operates five days in a week, eight working hours every day. No overtime was allowed in this simulation. The market demand, shown in table 1, is fixed and it consists of regular shipments at certain days of the week.

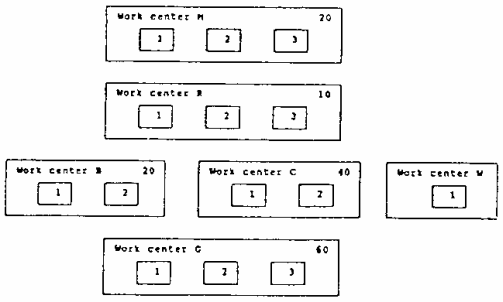

Figure 2. The layout of the shop floor
The numbers at the top right are the average setup times

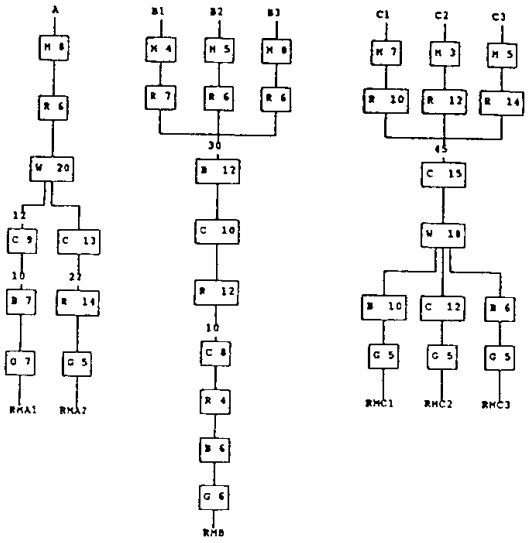

Figure 3. The routings
The letters denote the machines, the numbers are the time-per-part in minutes

Product	Weekly Shipment	Every two wk. Shipment	Day of Shipment
A	22		Friday
B1	25		Thursday
B2	24		Wednesday
B3	32		Friday
C1		50	Tuesday
C2	16		Monday
C3	20		Wednesday

Table 1. Market demand

The capacity profile for the six resources is shown in Table 2. It consists of the actual average load percentage of the production time, setup time and breakdown time. The

Machine	Downtime %	Processing %	Setup %	Total busy%
B	5.90	67.20	7.96	81.06
G	4.13	34.05	18.47	56.65
C	7.04	74.33	14.30	95.67
R	3.60	40.33	4.86	48.79
M	1.96	19.80	7.29	29.05

Table 2. Load profile

'C' resource was made to be the most loaded resource. The statistical fluctuations are of two types. The first is the fluctuations on the time-per-part (TPP) and setup times. These are uniform distributions with a coefficient of variation of 22.5% to 27.45%. The second fluctuation, a modified exponential distribution, impacts the downtime of the machines. On average it is 7.3% of the total working time (production and setup).

The planning schedule consists of the raw material release only - exactly buffer time, chosen to be 40 hours for all simulations, earlier than the due shipment. The simulation program explodes the orders to create the release schedule. The work centers work according to the following rules:

* A scan to identify new orders to be declared 'expediting' is carried out exactly 'expediting zone' hours before the end of every day. All orders are supposed to have due-date at the end of a given day.
* The expediting rules, which will be stated later, may instruct a machine to start setup immediately for that order - putting aside any other assignment. If no machine is available - a check is made every 20 minutes.
* Orders that have to be shipped within the next 20 hours, but not declared as 'expedite', have precedence over the other orders. However, no setup breakage will take place. A resource which has completed its current job will scan the waiting inventory in front of it for possible within-20-hours orders.
* If all the work in front of a work center is to be shipped later than within 20 hours - the choice criterion is the biggest load. We considered it an arbitrary choice - very easy and straightforward to implement. Choice of the smallest processing time job (SPT) is not practical, in this environment, because of the transfer mechanism employed.
* The transfer of material between work centers is automatic and immediate. Whenever a work center finishes a part, it is available to the next work center. This is in line with the concept of minimal transfer batches, mentioned by Goldratt and Fox (1986), but also used extensively in JIT and in assembly-line environments.
* The raw materials are released according to the schedule. Operations on common parts are scheduled (shipping date minus buffer-time) to prevent 'stealing'.

The expediting rules are:

* The downstream routes for every expediting order are referred to as 'the red-route'.
* The first scheme of expediting demands that every operation on the red-route will be assigned to a machine unit immediately. This means that, unless there is an idle machine, a machine will stop its current job and setup for the operation.
* The second scheme of expediting assigns a machine unit to an operation on the 'red-route' only when at least one piece of material is available for it.
* Two obstacles might prevent assigning machines to an operation on the 'red-route', no matter which expediting scheme is being used. The first when all the work center's machines are down. The other when all the work center's machines are doing other expediting jobs, meaning that it is a 'clash' situation.
* Clashes are recorded in the data base. In this paper we'll refer only to the number of clashes in a given time unit, no matter which work center is involved. When expediting can't be carried out, either because of a clash or because of a downtime period - the rest of the operations downstream will not be assigned.

The logic behind the first scheme is to push the order in a very fast way - assigning one unit of every resource needed along the way. The second scheme trades some of the speed of the expediting for the sake of less wasted capacity: assigning a unit only when at least one piece is available for the expediting job.

RESULTS

The basic experiments were performed on eight values of the 'expedited zone', the independent variable, denoted by EZ. The time-buffer was 40 hours: exactly one week of work (five days eight hours per day). Each experiment was repeated 250 times. Every simulation ran for 14 weeks. However, only data concerning the last 10 weeks was recorded and analyzed thus reducing the impact of the initial state.

The primary dependent variable is the mean number of late-order-days (LOD) - the summation of the days each order was late (per 10 week run). Another important dependent variable is the mean number of clashes.

Standard EZ	Standard Mean LOD	deviation LOD	Correlation error LOD	Mean num. of clashes	(LOD and clashes)
0	1.8	2.201	0.13919	0.016	0.192
2	1.316	2.267	0.14340	0.056	0.497
4	1.096	2.323	0.14692	0.188	0.791
6	1.572	3.911	0.24737	1.304	0.906
8	2.428	5.932	0.37520	5.900	0.962
12	4.912	7.107	0.44947	36.812	0.954
16	3.956	6.040	0.38202	64.016	0.931
20	1.74	6.071	0.38398	51.936	0.883

Table 3
Expediting Scheme A

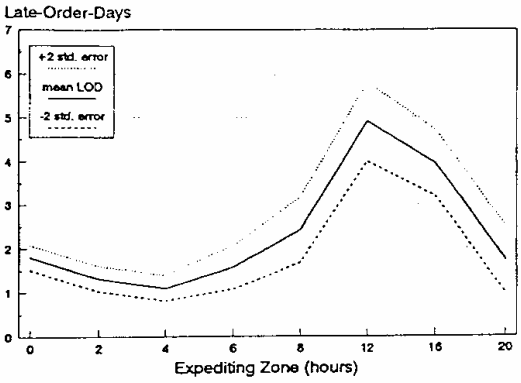

Figure 4. Late-Order-Days versus Expediting Zone: Scheme A

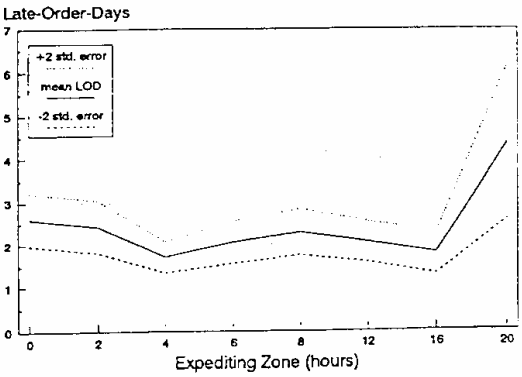

Figure 5. Late-Order-Days versus Expediting Zone: Scheme B

Expediting zone in hours	Mean 10 weeks run	LOD Std. dev.	Standard error of the mean	Mean num. of clashes per 10 wk.	Corr. LOD & clashes
0	2.604	4.901	0.3100	0.412	0.857
2	2.432	4.854	0.3070	0.368	0.779
4	1.724	2.922	0.1848	0.248	0.632
6	2.060	3.905	0.2469	0.728	0.768
8	2.288	4.273	0.2703	2.268	0.880
12	2.056	3.738	0.2364	7.188	0.943
16	1.800	3.965	0.2507	17.372	0.938
20	4.352	14.101	0.8918	41.872	0.825

Table 4
Expediting Scheme B

The results include the standard deviation of the LOD and the standard error of the mean LOD in order to assess the significance of the main results. Tables 3 and 4 outline the results for the two expediting schemes.

The LOD as a function of the EZ (the expediting zone) is shown in Figures 4 and 5. The central curve corresponds to the mean LOD per 10 weeks for the specified expediting zone. The two other curves were drawn two standard deviations from the central curve to display a significant area.

ANALYSIS AND CONCLUSIONS

The results from scheme A support the two hypothesis. Around EZ=4 hours there is a minimum which is significantly lower than with EZ=0 (expedite only late orders). The existence of inferior results to EZ=0 is also significant for EZ=8 and EZ=12.

The effect of the LOD having a minimum at the proximity of a sharp rise of the clashes level is clearly shown in the results of scheme A. From the simulation experiments we may guess a heuristic rule for an acceptable size of the expediting zone that the average number of clashes be no more than 30% of the average LOD. Notice that the number of clashes is directly related to the frequency of the scans that record the clashes. These scans were performed once in every hour in these simulations. The 30% rule is based on such a frequency. Should the scans be performed only once a day, which corresponds to the basic time units used in the LOD definition, our rule will generate no more than 4% clashes relative to the LOD. Additional experiments are needed to determine a more precise rule.

The results from scheme B are less clear. The relatively high number of clashes at EZ=0 is certainly a rare occurrence, derived from just two specific runs out of the 250. According to the stated heuristic rule, an expediting zone of four should have given the best results - which it does, but an expediting zone of six is also close enough to the stated rule. The LOD for a six-hour expediting zone, while not strictly at the minimum, is at the vicinity of the minimum and the difference is not significant. The validity of the hypotheses are also backed by the results of the generally high correlation between the LOD and the clashes.

An unexpected result is that beyond the first minimum there is at least another one. Both schemes have a maximum point and then drop again. Scheme B clearly indicates another minimum. In order to check whether scheme A has a second minimum as well, a special experiment has been carried out. It consists of just 12 simulations with the EZ covering the whole 40 hours of the buffer. The resulting LOD were very high (26.75 LOD per 10 weeks run) - sufficient to make the claim, that there is a second minimum for scheme A, significant. The existence of two minimums indicates that the three effects suggested in figure 1 are certainly not linear, and the overall effect is not monotonic.

We do not recommend use of the expediting zone of the second minimum - even if it produces lower LOD. The crucial point is that we don't know, without trial and error experiments which are not practical in real life situations, whether the system is at the minimum - or is it in a worse position relative to EZ=0. Another reason is that the stan-

dard deviation of the LOD seems to be lower at the first minimum, thus the first minimum provides a safer state.

The startling conclusion, validated by the simulations, is that **expediting is risky and may lead to inferior DDP**. This happens much earlier we expected. Notice, for instance, the results in scheme A for EZ=12 hours. The difference between EZ=0 and EZ=12 is statistically significant (less than 0.01). Expedited zone of 12 hours, relative to a buffer of 40 hours, is less than one third of the buffer suggested by Goldratt and Fox (1986).

Once an expediting zone is chosen and appropriate expediting rules are implemented, we recommend that the number of clashes relative to the level of the LOD be monitored. If the number of clashes is no more than 30% of the actual LOD - then the expediting scheme is 'good enough' in the sense that it is better than expediting orders that are already late (EZ=0). The solution we provide here may not be the optimal solution - but it provides guidance to assess the current situation.

Scheme B results are inferior to those of scheme A - contrary to our initial expectations. This is, probably, because the slower pace of the expediting was not offset by saving capacity. It seems that the superior expediting rule is to start expediting quite late but then to rush the expedited order as fast as possible. Clashes should be monitored and their data collected. If no clash appears for some time - the expediting zone should be somewhat enlarged. When the number of clashes seems to grow considerably - the expediting zone should be somewhat reduced.

FUTURE RESEARCH

The expediting mechanisms tested here didn't include the use of overtime. Overtime is a simple way to increase capacity - if needed. It is evident that BM information should be the key factor in employing overtime. Overtime should be used on top of the regular expediting measures in order to offset the additional load caused by expediting. On the other hand, overtime may waste capacity due to fatigue of the human resources. Furthermore, overtime is an operating expense driver - so research is needed to examine the economical gain of using overtime as a part of the expediting actions in order to improve the due-date performance.

The simulations used here were relatively complex considering the routing and the overall number of resources' units. However, the market demand was fairly evenly distributed. It is expected that when the market demand fluctuates - larger buffer will be needed. How will this impact the expediting scheme? We still expect the correlation between clashes and LOD to be strong enough so that the suggested determination of whether the current state is 'good enough' or 'too risky' will still hold.

REFERENCES

Schragenheim, E., & Ronen, B. (1990). Drum-buffer-rope shop floor control. *Production and Inventory Management Journal, 31*(3), 18-23.

Schragenheim, E., & Ronen, B. (1991). Buffer management: A diagnostic tool for production control. *Production and Inventory Management Journal, 32*(2), 74-79.

Goldratt, E. M., & Fox, R. (1986). *The race.* Croton-on-Hudson. NY: North River Press.

Goldratt, E. M. (1989). *The haystack syndrome.* Croton-on-Hudson. NY: North River Press.