A short note on how to control customer demanded throughput times

Department of Operations Planning and Control--Working Paper Series

A Short Note on:
How to Control Customer Demanded Throughput Times

Henny.P.G. van Ooijen

Research Report TUE/BDK/LBS/92-03
May, 1992

Graduate School of Industrial Engineering and Management Science
Eindhoven University of Technology
P.O.Box 513, Paviljoen F1
NL-5600 MB Eindhoven
The Netherlands

DIFFERENT CUSTOMER-REQUIRED LEADTIMES.

In a number of situations it can be advantageous to have different planning leadtimes (for the same kind of products). Harrison et al. in their case study at NSC for instance notice that many high level managers at NSC endorse the concept of planning leadtimes differentiated by order priority status. With that, they mean establishing shorter planning leadtimes for production lots associated with major (OEM) customers and other urgent orders.

The major question that arises in such a situation is:

is this possible in a controlled way?

Giving priority to a (new) order for a certain customer leads to longer lead times for the orders of the customers already present on the shop floor; if many urgent orders arrive then their leadtimes will be much larger than the planned lead times. So our flexible attitude towards some customers leads to very unreliable leadtimes. This is not what we really want; we are not reliable since the lead times are not controlled.
The question if we can have different customer demanded lead times in a controlled way is the subject of this research.

Besides the already mentioned customer demanded leadtimes situation there are also a number of other situations were it can profitable to have different controlled flow rates.
Examples are:
- products with a low average demand (high demand uncertainty)
- products which use expensive materials
- products assembled from a number of different components (network structure)

The production situation we have in mind is a discrete component manufacturing department. with a functional layout and a jobshop routing structure. In a functional lay-out similar machines are grouped into work centers; a job shop routing structure implies that from each work centre the work orders can flow to a number of other work centers.
Suppose we have a production situation where all products (more or less) have the same processing time distributions for all machines in the (job-) shop (products only differ in the routing) and that we have a number of customer classes each having its own required leadtime.

Customer class	Required leadtime	Arrival rate
1	L_1	λ_1
2	L_2	λ_2
n	L_n	λ_n

Knowing the required leadtimes we can calculate the total expected waiting time per product by subtracting the total (expected) processing times. So we now have a number of customer classes, each class i having its own total expected waiting time W_i we wish to be realized. If all products are more or less the same, different total waiting times can only be achieved if the work orders for the different products have different flow rates (number of operations performed per period). Now the question is:

how can these different flow rates be realized?

One way to realize different flow rates could be by using absolute priority classes: products are divided into a number of groups giving group i priority over group j if i > j. This is a very static way leading to extremities.

We will use a more flexible method, using due dates as a means for communicating to the shop floor the different flow rates we wish to be realized (a large allowance for waiting implies a smaller flow rate than a small allowance for waiting). As has been demonstrated by Kanet and Hayya, operation due dates are very effective in this way: they realize a small variance in the work order lateness. Therefore we will use the operation due date sequencing rule.

Now the question is if such a simple heuristic rule, indeed "forces" the individual work orders to flow at rates implied by their due dates.

It will be obvious that using operation due dates for realizing certain waiting times implicitely assumes that the real waiting times are in accordance with the waiting time norms used for setting the due dates. In other words: the operation due dates must be realistic.

The actual waiting times (per work centre) obey the equation:

(for the sake of convenience we will assume that all average routinglengths are equal)

$$\lambda_1 A_{m,1}+\lambda_2 A_{m,2}+.....+\lambda_K A_{m,K}=(\lambda_1+\lambda_2+....+\lambda_K)A_m$$

with $A_{m,k}$: the actual waiting time for products of class k at work centre m

A_m : the average actual waiting time at work centre m

This means that for the waiting time norms used in the process of setting the operation due dates we must demand that:

$$\lambda_1 W_{m,1}+\lambda_2 W_{m,2}+.....+\lambda_K W_{m,K}=(\lambda_1+\lambda_2+....+\lambda_K)W_m \qquad (A)$$

(For different average routinglengths we must use the work centre arrival rates which can be found by solving a set of lineair equations, based on the external arrival rates and the transition probabilities.)

This has consequences for our production situation if we want to fulfill all the customer's leadtime wishes for the different classes.

If the waiting times are given, which happen to be so in our situation, we know what the average work centre waiting time must be (weighted average of the customer demanded waiting times) and so we can calculate the required capacity (and thus idle time) necessary to realize the customer required lead times. Knowing the required capacity we know what it costs to fulfill all the customer's lead time wishes.

The calculation of the required capacity can roughly be done by using

$$W = \frac{\varrho b}{2(1-\varrho)m} \times (1+CS^2)$$

with ϱ: utilization rate

b: average processing time

m: number of parallel machines

cs^2: squared coefficient of variation (processing time)

Example:

Suppose $cs^2 = 1$, average processing time (b) = 1 time unit, m=1 and $\varrho = 0.90$.

Then we have $W = \varrho b/(1-\varrho) = 9$ time units..

Further suppose we have 3 customer classes: customers in class 1 expecting a waiting time of 3 time units, customers in class 2 expecting a waiting time of 5 time units and customers in class 3 expecting a waiting time of 10 time units. Then, in case all customers have the same arrival rates, the average waiting time must be equal to (3+5+10)/3 = 6. This corresponds to a utilization rate of ≈85,7 %. (0,857/(1-0,857) = 5,993). So in this case we need to increase the capacity (otherwise W will be equal to 9). This can be done by buying an extra machine, working overtime, etc., and leads to an increase of costs.

Now the question is if condition (A) is not only necessary but also sufficient to have different, controllable flowrates. To investigate this we used a simulation study. The shop we simulated is the often used 5-machine job shop. The routing of a work order is determined upon arrival using equal transition probabilities, which are determined by the probability of leaving the shop (which is set according to the average routinglength). There are no setup times and the processing times are exponentially distributed with a mean value of 1 time unit. To be sure that there is enough slack to "exchange", we used a kind of input/output control: the number of orders in the shop is held constant. As soon as an order leaves the shop a new order enters the shop. The customer class a new order belongs to is determined by the ratios of the arrival rates (we do not really use the an arrival rate, we only use them for this purpose). The waiting times following from the customer demanded lead times are supposed to have a uniform distribution.

Simulation runs were performed for a utilization rate of ≈85% (number of orders on the shop floor is 23), for a utilization rate of ≈89% (number of orders on the shop floor is 34) and a utilization rate of ≈95% (number of orders on the shop floor is 70).

Since there is a lower bound > 0 for the minimum waiting time that can be achieved we used as performance measure the so-called "normalized" actual waiting time. The lower bound can be found by using the absolute priority rule or head of the line priority rule (Cobham). The normalized waiting time is calculated by dividing the actual waiting time reduction (= FCFS-waiting time - actual waiting time) by the maximum reduction that can be obtained (=FCFS-waiting time - absolute priority waiting time for the highest priority customer class):

$$W_N = \frac{W_{fcfs} - W_a}{W_{fcfs} - W_{\min}}$$

where: W_N : normalized waiting time

W_{fcfs} : waiting time under the first come first served discipline

W_a : actual waiting time

W_{min} : waiting time for the highest priority class under teh head of the line priority discipline

The results of these simulations are shown graphically in Figs. 1 to 3. From these figs. we conclude that, more or less independent of the utilization rate, up to a norm waiting time reduction of about 60% there is an approximately one-to-one relation between the norm waiting time reduction and the normalized actual waiting time reduction. We did not perform an extensive ANOVA analysis yet, but since we are only interested in relative differences and we used common

random numbers, we think that this conclusion may be generalized.

So for practical purposes we state that up to a normalized actual waiting time reduction of about 60% condition (A) is not only necessary but also sufficient. Via computer simulation we have shown that the system can be "forced" to realize the different flow rates, and, therefore, produce different throughput times, by using operation due date sequencing as priority rule.
Since we used normalized actual waiting times the general observation is that the fast categories always lag behind schedule and the slow categories always are ahead of schedule.

The question arises what this means for the actual waiting times and thus for the customer demanded leadtimes. Using a norm waiting time equal to αW for a certain class of customers we expected to get an actual waiting time for this class equal to αW. However, due to the lower bound > 0 for the minimum waiting time, we got an actual waiting time equal to $\alpha W+(1-\alpha)W_{min}$ (this can easily be found by using the one-to-one relation between norm waiting time reduction and normalized waitng time reduction). So to get a reduction of the waiting time of α% we have to set the slack equal to:

$$(1-\frac{\alpha W}{W-W_{\min}})\times W$$

where: W : waiting time under the fcfs discipline

W_{min} : waiting time for the highest priority class under the head of the line discipline

In that case we will get an actual waiting time equal to $(1-\alpha)W$.
This means that the waiting times W_i that follow from the customer demanded leadtimes and that gives the reduction parameters α_i $(=W_i/W)$ can only be achieved if the W_i obey the equation (A) and the slack used in setting the due date for class i equals:

$$(1-\frac{W-W_i}{W-W_{\min}})\times W$$

We have to investigate this further.

For more detailed information see the following related papers:

Bertrand J.W.M., and Ooijen H.P.G. van, Flow rate flexibility in complex production departments. *International Journal of Production Research,* 1991, vol.29, no.4, 713-724.

Ooijen H.P.G. van, Controlling different flow rates in job-shop like production departments. *International Journal of Production Economics,* 23 (1991) 239-249.

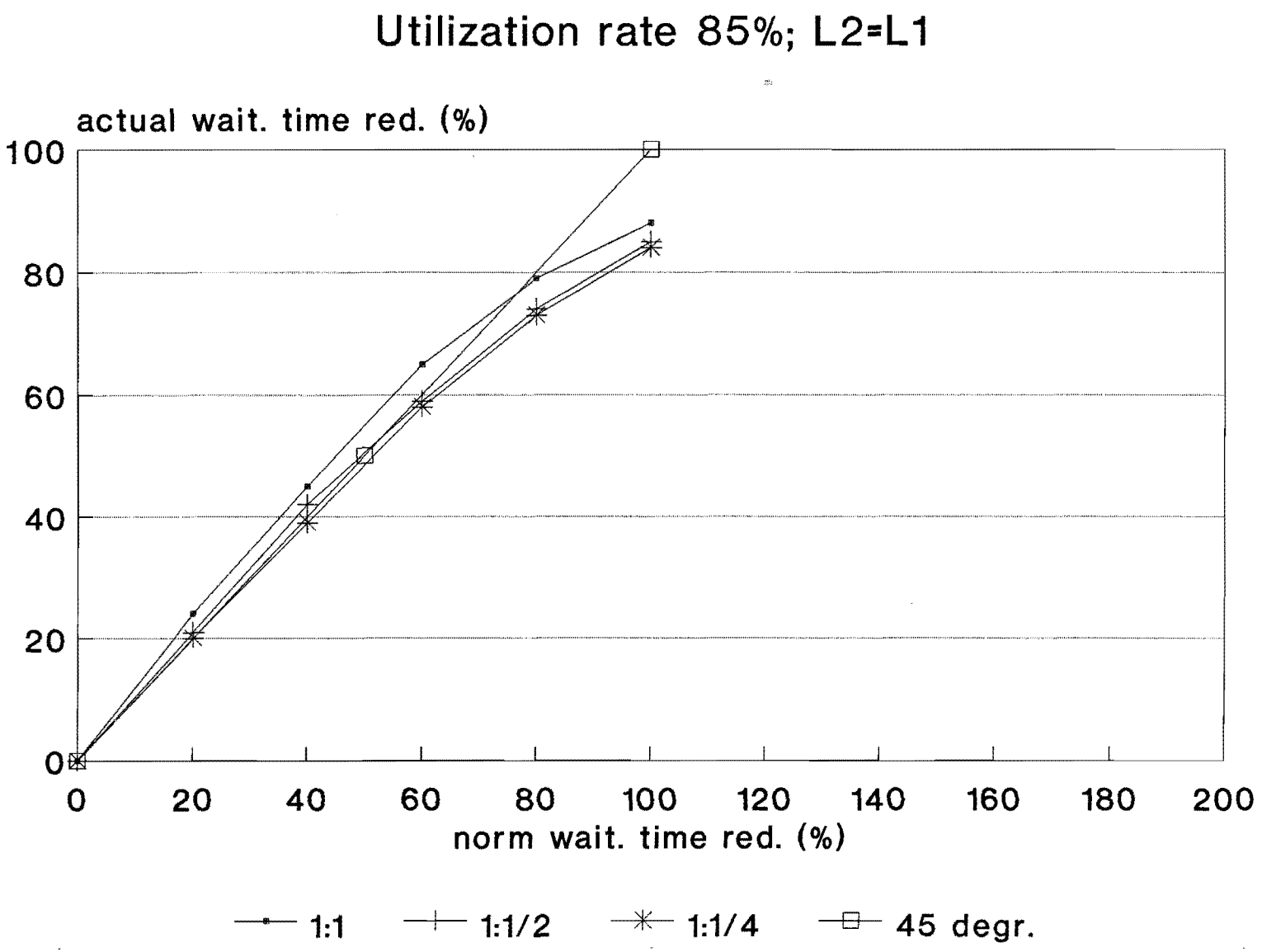
Utilization rate 85%; L2=L1
actual wait. time red. (%)
100
80
60
40
20
0
0
20
40
60
80
100
120
140
160
180
200
norm wait. time red. (%)
1:1
1:1/2
1:1/4
45 degr.

Fig. 1

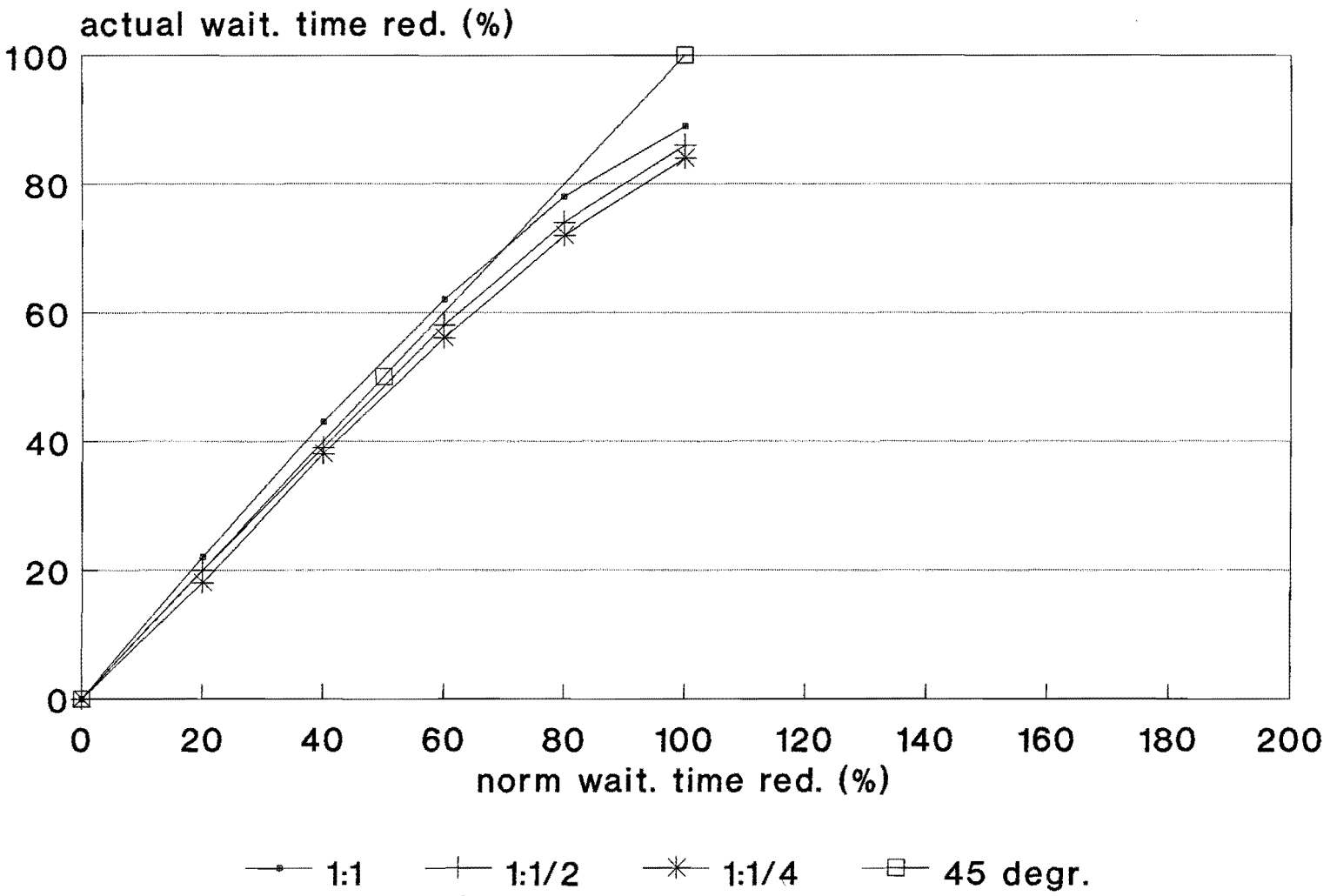
Utilization rate 90%; L2=L1
actual wait. time red. (%)
100
80
60
40
20
0
0
20
40
60
80
100
120
140
160
180
200
norm wait. time red. (%)
1:1
1:1/2
1:1/4
45 degr.

Fig. 2

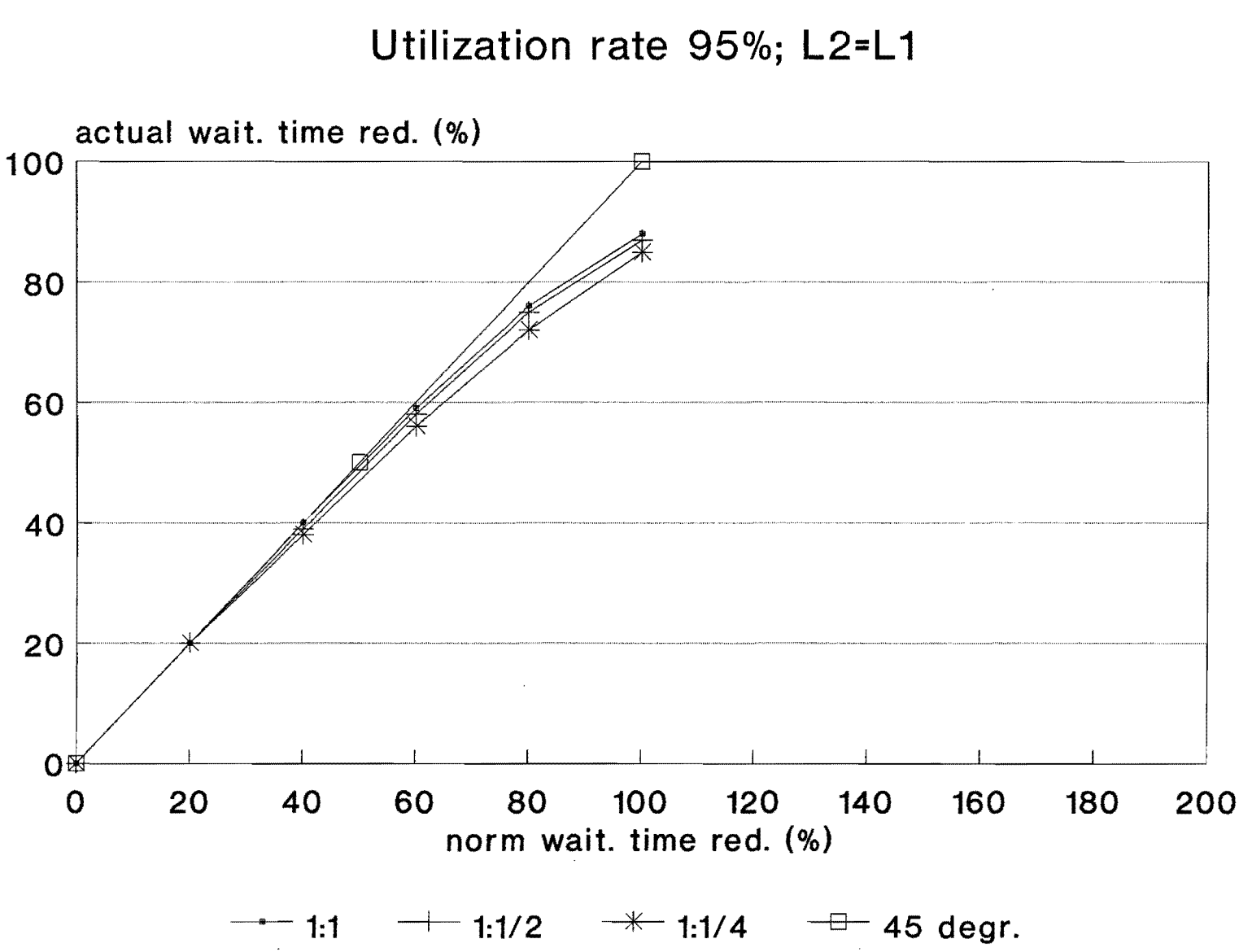
Utilization rate 95%; L2=L1
actual wait. time red. (%)
100
80
60
40
20
0
0
20
40
60
80
100
120
140
160
180
200
norm wait. time red. (%)
1:1
1:1/2
1:1/4
45 degr.

Fig. 3