Proceedings of the 1989 Winter Simulation Conference
E.A. MacNair, K.J. Musselman, P. Heidelberger (eds.)

DISPATCHING IN AN INTEGRATED CIRCUIT WAFER FABRICATION LINE

Pravin K. Johri
AT&T Bell Laboratories
Room 3M-306
Crawfords Corner Road
Holmdel, N.J. 07733

ABSTRACT

Wafer Fabrication has been described as the most complicated manufacturing environment existing today. This paper describes a method used to dispatch lots in one of AT&T's Wafer Fabrication Clean Rooms. The objective is to minimize idle time on important facilities in the clean room. For each lot in the clean room, the method indicates the slack time the lot can incur before it is needed at the next important facility group in its route. The slack time is the amount of time the lot can be delayed in queue with the implication that a lot with a smaller slack time needs to be processed more urgently than a lot with a larger slack time.

1. INTRODUCTION

This paper qualitatively describes a method used to dispatch lots in one of AT&T's Wafer Fabrication Clean Rooms. It is a simplified version of the actual method in use - application specific details have been omitted and only the general aspects of the algorithm are discussed. The objective of the method is to minimize idle time on important facilities in the clean room.

Wafer Fabrication has been described as the most complicated manufacturing environment existing today. Dispatching is the most intricate scheduling problem. Realistic answers are required in real-time. On the other hand, for a real application data/information can be inaccurate, missing or difficult to collect. Many unpredictable events beyond the scheduler's control affect the system and need to be accounted for, generally, after the fact. Our experience with simulation indicated that even a detailed simulation model did not capture all the dynamics of the clean room.

In light of these difficulties, a detailed model approach is not used. Instead, the major aspects of the problem are identified and dealt with in the best possible manner given the practical limitations. One such aspect is facilities or machines being unavailable for work for both scheduled as well as unscheduled maintenance (breakdowns). The equipment in the clean room is highly sophisticated and needs periodic maintenance and calibration. It is also highly unreliable making the availability of facilities the single most dominant factor in dispatching/scheduling decisions. It is difficult to collect and maintain information as to when individual facilities are to undergo scheduled maintenance. Quite often, scheduled maintenance depends on the load (e.g., after a certain number of wafers have been processed) and the time it will take place cannot be specified in advance. Breakdowns are unpredictable and the cause of breakdowns difficult and, consequently, time-consuming to determine. The system does have accurate information on which facilities are up and which are down. We decided to observe this at short intervals and dispatch assuming that the state does not change till the next observation.

For the same reasons outlined previously, the dispatching method does not give specific instructions. Instead, it concentrates on providing information which can be used to make a good dispatching decision. For each lot in the clean room, the method indicates the slack time the lot can incur before it is needed at the next important facility group in its route. The slack time is the amount of time the lot can be delayed in queue with the implication that a lot with a smaller slack time needs to be processed more urgently than a lot with a larger slack time. Slack times are thus used in a relative sense.

Lozinski and Glassey (1988) develop bottleneck starvation indicators to aid in shop floor control but exactly how the control is to be applied is left to the operators. Some commercially available scheduling packages for the semiconductor industry are briefly described in this paper. Glassey and Resende (1988a,b) have a starvation avoidance rule for releasing wafers into the clean room.

This paper is organized as follows: §2 discusses the modeling complexity; §3 describes our experience with simulations written for the clean room; §4 gives a high level description of the method; §5 illustrates the calculations made by the algorithm and the conclusions are summarized in §6.

2. MODELING COMPLEXITY

There are five major process areas in wafer fabrication. These are

- chemical clean,
- photolithography,
- plasma/chemical etch,
- ion implant and
- metal deposition/oxidation.

The circuit is grown in layers, each layer essentially requiring the following sequence of operations: cleaning, metal deposition, photolithography, etching and ion implant. Consequently, a lot visits each of the process areas many times. Facilities are often dedicated to performing particular operations and different visits to the same process area may be to different facilities. The complete sequence of operations required to produce each wafer type is given in its process log which could consist of several hundred steps. Each process step has a facility group associated with it, which denotes the list of facilities on which the step can be performed. For a more technical description, the reader is referred to Burman *et al* (1986) and the references in it.

The clean room environment has certain unique characteristics which need to be considered before deciding how to dispatch. The dimensionality of such a problem is very high. Facilities can frequently be unavailable for work due to scheduled or unscheduled maintenance. A facility may need to be setup for different tasks. The time to perform a setup varies. There are many places where the product is inspected and some of the product may have to be reworked. The time duration between many events, such as facility breakdowns, setup times, some processing times etc., is variable and often unpredictable.

Dispatching decisions have a big impact on queuing times, while the queuing times determine which decisions are available at each decision point and when the decision point occurs in time. To evaluate the effect of any decision a complete specification of all future decision points, decisions and queuing times has to be made. There are countless possibilities as to how these could actually occur. An evaluation can be made by predicting one realization (or at best, a probabilistic average of a number of realizations) of these quantities. However, it is almost certain that the actual outcome will be quite different from a probabilistic average or a chosen realization.

3. EXPERIENCE WITH SIMULATION

A detailed simulation model was first developed for the clean room. This model accounted for the following features:

- product dependent routing
- shifts
- set ups
- rework
- yields
- facility breakdowns
- variance in processing times

Current data and product mix were used. Unfortunately, the simulation failed to accurately predict an average performance measure such as cycle time with errors ranging from 10-20%.

In hindsight, there can be many reasons for this disparity. We outline a few next.

First, scheduled maintenance was lumped together with unscheduled maintenance and then treated as occurring in a random manner. In reality, scheduled maintenance occurs periodically. Additionally, "control lots" are run to calibrate some facilities frequently. These lots do not follow any particular product routing. We found that control lots are a significant percentage of the total lots in the clean room.

Second, humans are not "work conserving" as servers are assumed to be in simulations. Thus, if an operator has four lots with processing time of, say, 1 hour each to finish in a shift of 8 hours, it is highly likely that the operator will space out the lots to fill the whole shift and not do them in the first four hours as the simulation assumes.

Third, it is quite likely that performance measures such as cycle time have a large variance. In this case, a long term average is not a good indicator of actual observed values from month to month.

4. DESCRIPTION OF THE ALGORITHM

4.1 The Approach Taken

Dispatching decisions are made frequently with answers required in real time. Decisions made in one area affect other areas and, thus, cannot be based purely on local considerations. On the other hand, the huge dimensionality of the problem prohibits taking the whole future into account or using detailed models. Moreover, variability and the features discussed earlier will cause the system to deviate from the best of predictions in a relatively short time. Thus, we decided to look at relatively near term (a day or so) effects of the decisions while updating frequently to take care of deviations. To account for the factors unknown to the algorithm or

ignored by it, the final dispatching decision is left to the operators and the algorithm focuses on providing the information necessary to make a good decision. For this purpose, the algorithm indicates the slack time for each lot. In the absence of other considerations, the lot with the smallest slack time should be processed first. The slack times are updated periodically.

4.2 Outline of the Algorithm

The algorithm dispatches for process steps that have been designated as important. It looks where each lot is in the clean room and determines the next important step (plus the facility group associated with it) in its route, how soon the lot can get to this step and how much work it brings for the step. Collectively, this yields how much work can arrive at the important facility groups over time. Now, based on the number of facilities that are currently up and available, the algorithm determines how much can be processed at each facility group. The difference determines the excess work at the facility group over time and indicates how long the facility group can be busy without any additional work arriving. This, in turn, determines how long each lot bound for this facility group can be delayed, i.e., its slack time.

4.3 Assumptions/Heuristics

The slack time is determined under the following assumptions:

(a) the status of facilities (up or down) does not change (till the next update),

(b) there is no contention for facilities on unimportant facility groups, i.e., queuing time on these facilities is negligible.

The main heuristic is to evaluate slack times as if there is no queuing at the unimportant facility groups. The slack times suggest the "best" queue times at these unimportant facilities and are used to control the actual queue times. Note that lots with a small slack time should be processed soon. If they are, their queuing delay is small as has been assumed. Lots with a large slack time can be delayed. If they are, they may incur large queuing delays contrary to the assumption. However, such lots are headed for a facility group which is not in danger of idling (which is why the slack times are large for such lots) and the dispatching decisions are of lesser consequence. Thus, the effect of the error is discounted. Errors, however, will accumulate with time and it is necessary to update slack times at short, regular intervals.

For simplicity, the analysis is truncated at the next important step in each lot's route. Dispatching is useful only if few steps are declared as important. We are more concerned with the short term implications and it is unlikely that a lot will visit more than one important step in a short time horizon (say, several hours). This can easily be relaxed if the conditions are not satisfied.

5. DETAILS OF THE ALGORITHM

For each lot in the clean room, the algorithm requires the next important facility group the lot will visit, the time (without any queuing delays) to get to the important facility group and the amount of work brought for the important facility group. The units of time and work are hours. This information is based on the average amount of time to perform each step in the routing log of each technology and the position of the WIP in the clean room. The time to the important facility group is the sum of the processing times of the steps between the current step and the important step. It is rounded to the nearest integer.

Finally, for each important facility group, the number of facilities currently up and available are required. The analysis for each important facility group is independent of the other facility groups, and here we will concentrate on just one facility group. Based on all the lots headed for this facility group, the algorithm tabulates the total work that could arrive each hour. This would generate an arriving work profile as follows:

Table 1 : Arriving Work Profile

Time to Facility Group (hours)	*Total Arriving Work (hours)*
0	2
1	3
2	1
3	4
4	5
.	.
.	.

This profile shows that 2 hours of work is currently at the facility group, 3 hours of work can arrive in one hour, and so on. Now, suppose there is 1 facility up and available at this facility group. It can process 1 hour of work every hour. Subtracting the amount that can be worked off from the arriving work yields the excess work at the facility at the beginning of each hour.

Table 2 : Excess Work

Time (hours)	*Excess Work (hours)*
1	1
2	3
3	3
4	6
5	10
.	.
.	.

A positive value of the excess implies that there is more work in the previous hour than the facility can handle, and some work will be carried over to this hour. A negative value implies that there is not enough work to keep the facility busy and there will be some idle time. In our example, there is 1 hour of excess work in the zero-th hour which gets carried over to hour 1. A lot which can get to this facility group in 1 hour, will find 1 hour of work already at the facility. In other words, this lot can incur 1 hour of queuing delay and arrive one hour late without the facility idling. Hence, it gets a slack time of 1 hour. All lots are assigned slack numbers based on similar reasoning.

6. CONCLUSIONS

This paper describes a practical method to dispatch lots in an integrated circuit wafer fabrication line. The method provides useful direction in a very complicated and intricate problem. There are additional uses and variations of this method. For example, the arriving work profile can be used to schedule periodic maintenance. These will be described in a future publication.

ACKNOWLEDGEMENT

The author is grateful to B. T. Doshi and W. Kahan for some very useful suggestions.

REFERENCES

D. Y. Burman, F. J. Gurrola-Gal, A. Nozari, S. Sathaye and J. P. Sitarik (1986), Performance Analysis Techniques for IC Manufacturing Lines, *AT&T Technical Journal*, 65, 46-57.

C. R. Glassey and M. G. C. Resende (1988a), Closed-Loop Job Release Control for VLSI Circuit Manufacturing, *IEEE Transactions on Semiconductor Manufacturing*, 1, 36-46.

C. R. Glassey and M. G. C. Resende (1988b), A Scheduling Rule for Job Release in Semiconductor Fabrication, *Operations Research Letters*, 7, 213-217.

C. Lozinski and C. R. Glassey (1988), Bottleneck Starvation Indicators for Shop Floor Control, *IEEE Transactions on Semiconductor Manufacturing*, 1, 147-153.

AUTHOR'S BIOGRAPHY

PRAVIN JOHRI is a Member of Technical Staff at AT&T Bell Laboratories. He received a B.Tech. in Mechanical Engineering from the Indian Institute of Technology (Delhi) in 1979, and a Ph.D. in Operations Research from the State University of New York at Stony Brook in 1983. Pravin taught for one semester before joining the Performance Analysis Department at AT&T Bell Laboratories in early 1984. His work has involved developing scheduling and dispatching algorithms for many of AT&T's Factories, as well as analyzing protocols in telecommunication networks.