Proceedings of the 2001 Winter Simulation Conference
B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, eds.

DISPATCHING HEURISTIC FOR WAFER FABRICATION

Loo Hay Lee
Loon Ching Tang
Soon Chee Chan

Department of Industrial & Systems Engineering
National University of Singapore
Kent Ridge, Singapore 119260 SINGAPORE

ABSTRACT

As the semiconductor industry moves into the next millennium, companies increasingly will be faced with production obstacles that impede their ability to remain competitive. Effective equipment and line management planning will increasingly be required to maximize profitability while maintaining the flexibility to keep pace with rapidly changing manufacturing environment. In this paper, the authors present a two-bottleneck machines center model for wafer operations analysis. A new dispatching rule Balance Work Content, BWC, is introduced. This is a selective dispatching rule whereby it attempts to maximize the utilization of bottleneck machine. A systematic approach to assessing the impact of BWC is presented. Extensive simulation runs on both the deterministic and stochastic models developed shows its supremacy over conventional approaches of FIFO and SPT.

1 INTRODUCTION

The wafer fabrication industry is among the world's frontiers of ingenuity, entrepreneurial activity, and amazing accomplishments in the development of technology. It is also the industry which faces a lot of competition because both price and shelf life for new electronic products are decreasing exponentially over time due to rapid technological advancements. As a result, effective equipment and line management planning will increasingly be required to maximize profitability while maintaining the flexibility necessary to keep pace with rapidly changing manufacturing environment.

To-date, the manufacturing process of an integrated chip can broadly be classified into three major phases, namely wafer preparation, wafer fabrication and packaging. Wafer fabrication is the most complex of the three. Research interest has been focused on controlling wafer fabrication with its unique characteristics, such as reentrant product flows, diverse types of equipment, complex production processes and unpredictable yield and equipment downtime. All of these factors make production planning and scheduling complicated.

Scheduling policies attempt to get the right products done at the right time (Shayan and Fallah 1999). Some aspects of wafer fabrication scheduling that have attracted the interest of researchers are dispatching rule, reticle management, operator cross training, and management of ion implantation processes. This paper will focus mainly on the aspect of dispatching rules.

Many production scheduling and control techniques have been developed with the objectives of increasing key machine utilization, increasing throughput rates and decreasing WIP inventory level. Scheduling techniques range from simple dispatching rules such as FIFO and SPT to more advanced techniques such as PAC/ Kanban (Huang and Sha 1998), which generally requires more complicated information systems.

Due to the complexity in a wafer fabrication process, usually dispatching rules will be used to decide what job to schedule next when the machine center becomes free. Surveys of such rules can be found in Chandrasekharan and Holthaus 1999, Jain and Meeran 1999, and Johari 1993.

Apart from dispatching rules, there has also been considerable research on input regulation policies. Input regulation policies attempt to achieve shorter and more reliable flow times by releasing work to the fab in a controlled manner. General results indicate that the mean and variance of time spent in the fab can be reduced by input regulation rules.

In 1988, Lozinski and Glassey presented a concept called bottleneck starvation indicators and it is used in developing input regulation policies and dispatching rules. (Glassey and Resende 1988, Glassey and Petrakian 1989, Glassey and Weng 1991, Glassey and Resende 1988, and Lozinski and Glassey 1988). The starvation avoidance indicator is an indicator to assess if the bottleneck is in the risk of starving. For each process, WIP may visit the bot-

tleneck machine center several times in the whole manufacturing process. For each visit to the bottleneck machine, there is a different sequence of upstream machines from the bottleneck where the wafers need to visit. Each of the sequence is denoted as a flow. The risk of starvation can be evaluated by computing the desired amount of WIP required in the flow for the bottleneck machine center. Since the bottleneck will on average be the biggest queue in the factory, the lost time represents an irretrievable loss of final output. Hence by keeping the workload for this single bottleneck at or above a preset limit, the regulation of work in the shop can be achieved. A drawback of the starvation avoidance approach is the assumption of a fixed bottleneck whose location is known a priori. Besides, the lead time calculations did not include either transportation times or queue times, thus leading to possible uncertainties in the estimates.

Another prominent researcher, Wein, (Wein 1988, 1990, 1991, 1992, and Wein and Chevalier 1992), also has focused his work on wafer fab scheduling. He has conducted comprehensive studies on the effect of various dispatching rules with four different input regulation rules in 3 different fab setups, namely, 1 bottleneck, 2 bottleneck and 4 bottleneck fab setup, The four input regulation rules he considered are no input control, where lots arrive according to a poisson distribution; constant rate input, where lots are released into the fab at a uniform rate; constant WIP, where the amount of WIP in the fab is kept constant and new lots are released only as others have completed; and workload regulation, where new lots of wafers are released into the fab when the total work content for the single heavily loaded station falls below the preset limit. . He used cycle time as the performance measure. He observed that the effect of specific dispatching rules is highly dependent on both the type of input control used and the number of bottleneck work centers in the fab.

In this paper, we will propose a new dispatching rule, Balancing Work Content (BWC), and the details of this rule will be discussed section 2. Then in section 3, we will compare the rule developed with other dispatching rules at different experiment scenarios. Finally in section 4, conclusions are made.

2 BALANCING WORK CONTENT (BWC) DISPATCHING RULE

A dispatching rule, targeting to maintaining throughput while balancing the work content, will be introduced. Wafer fabrication technology is progressing rapidly. Changes in fabrication process are being made regularly in order to keep up with this progress. This may proved to be an enormous challenge to manage the fabrication line as machines and their operating parameters have to be reviewed and adjusted constantly. Since wafer fabrication is such a delicate operation, product yield is very sensitive to the adjustments of machine parameters. Special qualification period may be required to ensure that product meets the specifications. These entire disturbances, if not given proper attention, may result in lumpy WIP distribution and this will have adverse impact on the overall cycle time.

In order to minimize the impact of randomness caused by unplanned events such as process change or unscheduled machine down, a dynamic production schedule is required. The proposed heuristic is built with intention to overcome these problems through acquiring the capability to automatically adjust WIP level in a fab according to the prevailing operating conditions. In this heuristic, production lots are dynamically sequenced according to work content balancing algorithm which targets to balance the work content between different bottlenecks and simultaneously minimize the risk of bottleneck machine centers from running out of WIP to process while the others are choked with long queues of WIP.

The selection criteria is based on attempts to generate a more balance line and at the same time prevent the bottleneck machine centers from running out of WIP to process. Starvation at the bottleneck simply implies that capacity not utilized is lost capacity forever.

Assume that there is only one product, and the dispatching rule is to decide which layer to be processed at the station when it becomes idle.

Let

$t_{i,1}$ = processing time for the WIP of layer i at bottleneck 1.

$t_{i,2}$ = processing time for the WIP of layer i at bottleneck 2.

($t_{i,j} = 0$ if layer i doesn't pass through work station j.)

$P_{i,1}$ = priority for WIP of layer i at bottleneck 1.

$P_{i,2}$ = priority for WIP of layer i at bottleneck 2.

WIP_{i1} = WIP of layer i at bottleneck machine center 1.

WIP_{i2} = WIP of layer i at bottleneck machine center 2.

Work content is the amount of time required to clear all the WIP parked at the machine center and is defined as the product of the WIP to the processing time. The priority can be defined as

$$P_{i,1} = WIP_{i,1}\left(t_{i,1} - t_{i,2}\right) \quad (1)$$

$$P_{i,2} = WIP_{i,2}\left(t_{i,2} - t_{i+1,1}\right). \quad (2)$$

Priority will be given to the layer which has the lowest value of $P_{i,1}$ or $P_{i,2}$, or which has the most negative difference in the work content. The rational is that the layer that has more work ,i.e. product of WIP and processing time is large, to be done at the next bottleneck machine center should be processed first. In other words, the algorithm will drive towards loading the other bottleneck machine center with more work so as not to starve that machine.

Notice the layer which will not pass through the other bottleneck machine will always be given the lowest priority.

The strength in this heuristic is that it not only provides a global view of the WIP distribution at the bottleneck machine, but also considers the distinct processing time at each layer and each bottleneck machine. As such, the work content is computed to provide a global view of the relationship between bottleneck machine centers. All these indices provide excellent tools for production planners to create effective schedule for the line.

3 SIMULATION ANALYSIS OF A TWO-STATION MODEL

We have considered a two-station model to mimic the wafer fabrication process, where the two stations are referred to the two bottleneck machines in the wafer fab. The model is shown in Figure 1.

We consider 14 layers of move at the production flow, and among them, 8 of the flow will pass through both the bottleneck machines, while the rest will only pass through either bottleneck 1 or 2. Moreover, whenever a lot leave the bottleneck machine, it will experience a delay before reaching the next bottleneck machine. This delay is assumed to be at least 20 times of the process time of the bottleneck machine.

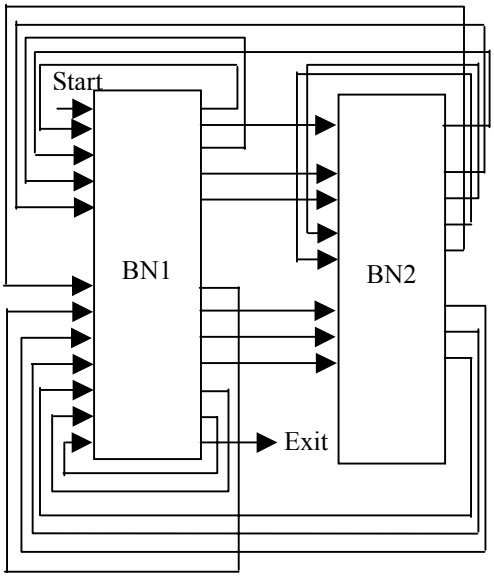

Figure 1: Two-station Model

This two-station model was built using Extend 4.0, and 3 different dispatching rules, i.e., First in First Out (FIFO), Shortest Processing Time (SPT) and Balancing Work Content (BWC), are tested using the simulation model.

3.1 Deterministic Model

For the deterministic model, every parameter is assumed to be constant, except for the time between failures and machine repair time, is exponentially distributed. We are adopting the constant release rule, and the input rate is maintained at the level so as to keep the target utilization for the bottleneck machine around 95%. For each rule, we repeat the simulation for 10 times, and the mean cycle time and standard deviation of the cycle time is recorded at Table 1.

Table 1: Summary of Simulation Results of Different Sequencing Policies

	Mean Cycle Time (CT)	Standard Deviation (SD)
FIFO	1151	131.94
SPT	1125	92.72
BWC	1110	42.61

The results show that BWC outperforms the other dispatching rule. BWC not only has a smaller mean but also a smaller standard deviation for the cycle time, which means it can provide a smoother output than the other dispatching rules. This is not surprising because BWC tends to balance the work contents between the two bottlenecks which will in turn give the benefit of smoothening the lines as well as the output. Since it is believed that most of the wafer fabs are adopting the constant release rule, BWC is believed to be a good coupling sequencing rule.

3.2 Stochastic Model

In section 3.1, we assume that all the parameter values are known with certainty. This is almost certainly untrue in practice. The best that one can honestly claim is that the probability of various outcomes can be determined and quantified with high degree of certainty. Here the processing time at the bottleneck machine centers are considered to be random. The objective of this analysis is to find out the impact, if any, on the performance of different sequencing rules. Three degree of variability for the processing time will be considered here namely, Coefficient of Variation (CV) = 0.5, 1.0 and 1.5. Constant rate input rule will be employed while maintaining constant throughput rate at about 0.0415, i.e., at a target utilization of 95% for the bottleneck machine. Similarly, we have repeated the simulation for 10 times, and the mean cycle time and standard deviation of the cycle time is recorded at Table 2.

Table 2: Summary of Simulation Results of Dynamic Processing Time

	CV=0.5		CV=1.0		CV=1.5	
	CT	SD	CT	SD	CT	SD
FIFO	1240	79.14	1237	75.70	1240	74.00
SPT	1185	115.48	1184	117.24	1185	118.30
BWC	1127	20.62	1124	21.33	1125	18.71

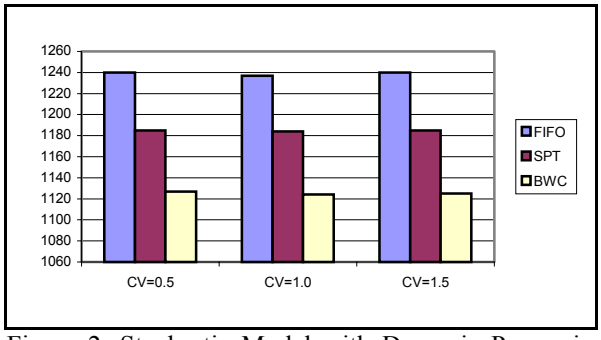

Figure 2: Stochastic Model with Dynamic Processing Time

"Deliberate" sequencing control policies such as SPT and BWC are observed to perform better under dynamic processing time conditions. In fact, improvement in throughput time when using BWC in a stochastic model is observed to be approximately double of that when the same rule is applied in a deterministic model. Hence, one can draw from these results that "deliberate" control is required especially when processing time is stochastic which is a true observation in any real fab. Also, we can observe that the standard deviation of BWC rule is also observed to be the smallest, which means it can give a more stable output.

4 CONCLUSIONS

In this paper, the authors have developed a new sequencing rule, Balance Work Content, BWC, and this rule was tested on a two-station machine center model which mimics the operations of a wafer fab. For the deterministic model with constant rate input control, BWC exhibits qualities of an excellent sequencing strategy, as compared to FIFO and SPT.

For stochastic analysis, we have analyzed the impact of variable processing time. Simulation results collected also show that BWC performs well and even better than with the deterministic model.

BWC is a dynamic dispatching rule because it makes its decision based on the current WIP distribution at the bottleneck machine, and also the distinct processing time of each layer at each bottleneck machine, and so it can handle the situation where the system is very dynamic.

Although BWC shows promising results, however, in order to enhance the value of this dispatching rule, we need to compare its performance under different simulation scenario, e.g. under different input policies. Moreover, using a real fab model would also help to determine its feasibility in practice.

REFERENCES

Brah, S. and G. E. Wheeler. 1998. Comparison of scheduling rules in a flow shop with multiple processors: a simulation. *Simulation* 71 (5): 302-311.

Chandrasekharan, R. and O. Holthaus. 1999. A comparative study of dispatching rules in dynamic flowshops and jobshops. *European Journal of Operational Research* 116: 156-170.

Glassey, C. R. and M. G. C. Resende. 1988. Closed-loop job release control for VLSI circuit manufacturing. *IEEE Transactions on Semiconductor Manufacturing* 1 (1): 36-46.

Glassey, C. R. and R. G. Petrakian. 1989. The use of bottleneck starvation avoidance with queue predictions in shop floor control. In *Proceedings of the 1989 Winter Simulation Conference*, 908-917.

Glassey, C. R. and W. W. Weng. 1991. Dynamic batching heuristic for simultaneous processing. *IEEE Transactions on Semiconductor Manufacturing* 4 (2): 77-82.

Glassey, C. R. and M. G. C. Resende. 1988. A Scheduling Rule for Job Release in Semiconductor Fabrication. *Operations Research Letters* 7 (5): 213-217.

Glassey, C. R. and R. G. Petrakian. 1989. The use of bottleneck starvation avoidance with queue predictions in shop floor control. In *Proceedings of the 1989 Winter Simulation Conference* :908-917.

Huang, J. Y. and D. Y. Sha. 1998. Constructing procedures of an effective production activity control technique for a wafer fabrication environment. *International Journal of Industrial Engineering* 5 (3): 235-243.

Jain, A. S. and S. Meeran. 1999. Deterministic job-shop scheduling: Past, present and future. *European Journal of Operational Research* 113: 390-434.

Johri, P. K. 1993. Practical Issues in Scheduling and Dispatching in Semiconductor Wafer Fabrication. *Journal of Manufacturing Systems* 12 (6): 474 – 485.

Lou, S. X. C. and Kager. 1989. A robust production control policy for VLSI wafer fabrication. *IEEE Transactions on Semiconductor Manufacturing* 2 (4): 159-164.

Lozinski, C. and C. R. Glassey. 1988. Bottleneck starvation indicators for shop floor control. *IEEE Transactions on Semiconductor Manufacturing* 1 (4): 147-153.

Shayan, E., and H. Fallah. 1999. A new approach to finite scheduling. *International Journal of Production Research* 37 (8): 1903-1915.

Wein, L. M. and P. B. Chevalier. 1992. A broader view of the job-shop scheduling problem. *Management Science* 38 (7): 1018-1033.

Wein, L. M. 1990. Scheduling networks of queues : heavy traffic analysis of a two-station network with controllable inputs. *Operations Research*, 38 (6): 1065-1077.

Wein, L. M. 1988. Scheduling semiconductor wafer fabrication. *IEEE Transaction on Semiconductor Manufacturing* 1 (3): 115-130.

Wein, L. M. 1988. Brownian networks with discretionary routing. *Operations Research* 39 (2): 322-340.

Wein, L.M. 1992. Scheduling networks of queues: heavy traffic analysis of a multistation network with controllable inputs. *Operations Research* 40 (2): 312-334.

AUTHOR BIOGRAPHIES

LOO HAY LEE is an Assistant Professor in the Department of Industrial and Systems Engineering, National University of Singapore, since 1997. He received his B.S (Electrical Engineering) degree from the National Taiwan University in 1992 and his S.M and PhD degrees in 1994 and 1997 From Harvard University. He is currently a member of IEEE, ORSS, and INFORMs. His research interests include simulation-based optimization, production scheduling and sequencing, logistics and supply chain planning, and vehicle routing. His email and web addresses are <iseleelh@nus.edu.sg> and <www.ise.nus.edu.sg/staff/leelh/>.

LOON CHING TANG is an Associate Professor and Deputy Head (Research) in the Department of Industrial and Systems Engineering, National University of Singapore. He received his PhD degrees in operations research from Cornell University. His research interests include quality and reliability engineering, simulation and operation planning, He has consulted for semiconductor, telecommunication, defense and health care industries in the above areas and has been on the examination board of CRW, CQE and CQS examination since 1993. His email and web addresses are <isetlc@nus.edu.sg> and <www.ise.nus.edu.sg/staff/tanglc/>.

SOON CHEE CHAN is a Systems Engineer in the TECH Semiconductor, Singapore. He received his B.ENG (Mechanical Engineering) degree from the National University of Singapore in 1998 and his M.ENG degrees in 2000. His research interests include simulation and semiconductor scheduling. His email address is <JoeChan@techsemi.com.sg>.