Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds

A SIMULATION STUDY ON LINE MANAGEMENT POLICIES WITH SPECIAL FOCUS ON BOTTLENECK MACHINES

Lixin Wang
Vinoth Chandrasekaran

Industrial Engineering Department
Micron Technology Inc.
9600 Godwin Drive, Manassas, VA 20136, USA

ABSTRACT

A 300mm wafer fab is one of the most complex systems in the world. How to optimize this system in terms of planning and scheduling is critical for profitability of the semiconductor companies considering billions of dollars of initial investment involved. Line management or fab wide scheduling is more important than area level scheduling although the latter has higher resolution and is considered as a harder problem. Traditional line management policies focus on pre-determined bottlenecks and has proven to be successful. However, for a dynamic fab with changing bottlenecks, some potential issues have been discovered. This paper used simulation as a tool to study the issues involved and propose an improved line management policy.

1 INTRODUCTION

A 300mm wafer fab is one of the most complex manufacturing systems in the world. It has the following characteristics. The system is huge involving thousands of machines and dozens of part types, and each part type needs to go through hundreds of steps before it leaves the fab. The wafer fabrication process is a re-entrant flow, where a wafer goes through same machine multiple times, and in some cases the wafer is required to go through the same component inside the machine for each step. Also, machine may break down at unexpected times. When a machine breaks down, it disrupts the normal movement of wafers in the system, and the impact may cascade to downstream steps. More details regarding a wafer fab system can be found in Wang (2008).

Typically, a 300mm fab has an initial investment of several billion dollars. Manufacturing department is asked to meet customer demand and maximize revenue. Planning and scheduling play a critical role for this mission. Planning makes sure the fab has right machines, both in quantity and type, to produce parts. Some types of machine, like lithography machines, are very expensive, thus quantity is limited and they are normally bottlenecks of the fab. Scheduling assigns individual wafers to individual machine by time. Optimized scheduling methods help achieve the manufacturing goal, especially when unexpected event happens, such as machine down or changed customer demands.

978-1-4799-2076-1/13/$31.00 ©2013 IEEE

2 LITERATURE REVIEW

2.1 Wafer Fab Scheduling

Wafer fab scheduling is a well-known NP hard problem. Thus it is impossible to find a optimal scheduling sequence within meaningful time period. As a result, very limited studies focused on mathematical optimization methods. Bixby, Burda and Miller (2006) developed a short-interval scheduling approach using integer and constraint programming. It was implemented in an IBM fab for Diffusion area. The approach provided benefits in throughput, cycle time and hot lot performance. SmartSched advance scheduling package (Hanny D. and S. Marteney 2011) is a constraint programming based scheduler designed for Photo area. The software has been proven to improve photo lithography machine utilization significantly at a 300mm fab. Note that there are extensive research studies on optimization for similar scheduling problems, but they are either for a much simplified problem, e.g., one machine or two machine problems, or their solutions have not been proven to be implemented in an actual size fab successfully. Recent studies can be found in Monch et al. (2011).

On the other hand, heuristic rules based scheduling has been used commonly in wafer fabs, though it does not provide optimal solutions. It does have its advantages, such as easier to implement in fab, simple to understand, and so on. A complete survey for this type of approaches can be found in (Subhash, Varadarajan, and Wang 2011).

2.2 Line Management Policies

Due to the complexity of wafer fab scheduling, one favorite method by fabs is a hierarchical method. It includes two layers: fab level (or line level) scheduling and area level scheduling (Figure 1). Line level scheduling is also referred as line management policies at fabs. Line management policies coordinate needs of different areas, and make sure decisions at one area do not have negative impact on other areas. Area scheduling is responsible for scheduling within the area only. This proposed framework can be found at Hanny D. and S. Marteney (2011), and Klemmt et al. (2010), as illustrated in Figure 1.

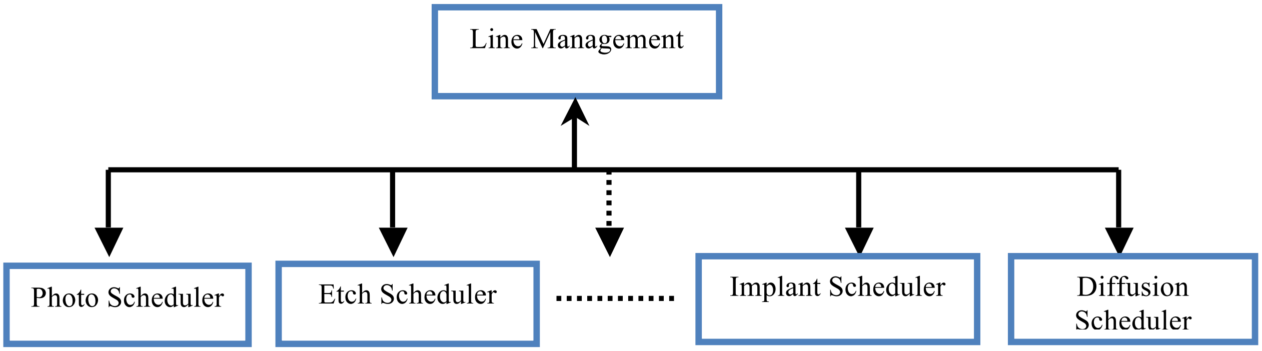

Figure 1: Two layer scheduling framework for a 300mm wafer fab

Main focus of this paper is on line management policy, as it is at higher level, its decision has fabwide impact, and thus plays more important role in achieving manufacturing goals. This also aligns with observations we have made over years working in the industry.

3 SLIM METHOD

SLIM, the acronym for Short cycle time and Low Inventory in Manufacturing, is a set of methodologies and scheduling applications for managing cycle time in semiconductor manufacturing (Leachman, Kang and Lin 2002). TCT (target cycle time), IPQ (ideal production quantity), and SS (schedule score) are the foundation of SLIM. TCT is the target time from lots entering the fab to leaving the fab and is given. Let PT_j denote processing time at step j, DCT_j denote the difference between sum of historical average step cycle time and step processing time for the portion of the line that ends at the preceding step of bottleneck step j, and starts immediately after the closest bottleneck upstream. Then, the target cycle time for step j, TCT_j, is

$$TCT_j = PT_j + BT_j,$$

where BT_j is buffer time at step j. For a non-bottleneck step, it is zero; and for a bottleneck step, it can be written as

$$BT_j = \frac{DCT_j}{\sum_{j=1}^{N}(DCT_j)}\left(TCT - \sum_{j=1}^{J}(PT_j)\right),$$

where J is total number of steps, and N is total number of bottleneck steps. The above TCT calculation method can be illustrated by Table 1. In this example, step 3 and step 7 are bottleneck steps.

Table 1: An Example to Illustrate TCT Calculation

Step sequence	Step processing time (PT_j)	Historical average step cycle time	Step TCT (TCT_j)	TCT
1	2	4	2	42
2	1	2	1	42
3*	5	10	9.875	42
4	3	5	3	42
5	2	3	2	42
6	2	4	2	42
7*	8	12	16.125	42
8	3	4	3	42
9	2	4	2	42
10	1	4	1	42

Once TCT for each step is determined, with help of Little's Law, target WIP (work in progress) can be computed as TCT times target throughput rate. Target throughput rate is determined by assuming fab achieves exactly TCT and fab outs schedule. More details and proof can be found in Leachman, Kang, and Lin 2002. Short term target IPQ for each step is expressed as delta of target WIP and actual WIP from its immediate downstream step to the last step. And SS is then computed with IPQ divided by target fab throughput rate from immediate downstream step to the last step. Note that SS is used to prioritize all lots at different part steps in the fab, thus equivalent to a line management policy.

The philosophy underlying SLIM is to allocate buffer time to bottleneck steps only. Doing this, target WIP at a bottleneck step will be much higher than actually needed. This inflates IPQ and SS for steps in front of bottleneck step and help to move as much as possible WIP to bottleneck steps. Thus it helps improve fab performance, e.g. throughput, which is determined by bottleneck steps. This is reasonable from perspective of long term results. SS based line management policy has proven to be successful at Samsung fabs (Leachman, Kang, and Lin 2002).

However, there are some disadvantages of SS policy from short term perspective. A 300mm fab is a dynamic environment. Once a machine is down or not qualified for a part step anymore, capacity loss incurs for the part step. When the capacity loss is severe, it becomes a dynamic bottleneck step. Dynamic, in this context, means time horizon of several days, up to a week. One direct solution for this is to allocate some buffer time to the dynamic bottleneck steps when calculating TCT. But how much buffer time to allocate remains a question. Another one is to use historical cycle time, which hopefully captures this capacity loss when it happened in the past. The goal of this paper is to develop new line management policies based on these ideas in order to handle dynamic bottlenecks better.

4 SIMULATION STUDIES

4.1 Line Management Policies

In this paper, we propose two line management policies in order to handle dynamic bottlenecks effectively. Together with SLIM, three policies will be studied, and the detailed descriptions are listed in Table 2. HCT_SS is very similar to TCT_SS. The only difference is to use historical cycle time(HCT) to replace TCT. HCT is computed with exponentially weighted moving average (EWMA) method. Let *N* denote number of weeks considered, WCT_{jk} is weekly cycle time for week *k* step *j*, k=1, 2, …., *N*, HCT for step j can be written as

$$HCT_j = \alpha \times (WCT_{jK} + (1-\alpha) \times WCT_{jK-1} + (1-\alpha)^2 WCT_{jK-2} + \cdots + (1-\alpha)^{K-1} WCT_{j1}$$

Where α is a constant smoothing factor between 0 and 1, and week *K* is the latest week. Similarly, the only difference between Pred_HCT_SS and TCT_SS is to use predicted cycle time to replace TCT. There are various methods to predict future cycle time, such as queueing theory or simulation (Zisgen et al. 2008, Fronckowiak, Peikert and Nishinohara 1996). The primary focus of this paper is on dynamic bottlenecks thus predicted cycle time for near future (usually several shifts or days) is used.

Table 2: Three Line Management Policies

Line Management Policy	Descriptions
TCT_SS	SS based on target cycle time proposed by Leachman, Kang and Lin 2002.
HCT_SS	SS based on historical cycle time
Pred_HCT_SS	SS based on predicted cycle time in future

4.2 A Simulation Model

Due to the complexity of a 300mm fab, a discrete event simulation model is built to evaluate the above three line management policies. The model is written with AutoMod scripts from AMAT (Applied Material). Initially, a real size fab model was run with these different policies, and output data showed some pattern. However, it was not intuitive to analyze and find out what caused the output difference, mainly due to the amount of noise introduced by interactions of reentrant flows and multiple part types. Therefore, a simplified line that has 11 steps, 10 machines, and a reentrant loop was constructed. Details are included in Figure 2 and Table 3.

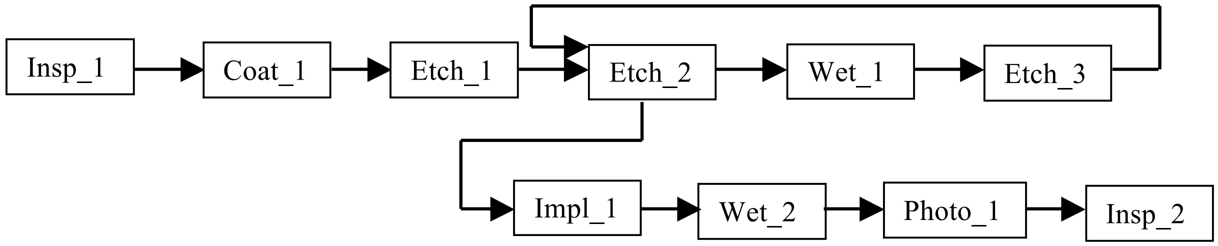

Figure 2: The process flow of a simple example with re-entrant flow

Table 3: The Process Flow

Step sequence	Machine name	Processing time (mins/lot)
1	Insp_1	5
2	Coat_1	15
3	Etch_1	20
4	Etch_2	10.2
5	Wet_1	20
6	Etch_3	20
7	Etch_2	10.2
8	Impl_1	20
9	Wet_2	18
10	Photo_1	24.1
11	Insp_2	12

Looking at Table 3, it is clear that Step 10 Photo step is the bottleneck step, and machine Photo_1 is the bottleneck machine. The model ran for 80 days and average cycle time in the last 70 days was collected to construct HCT_SS policy. The sum of the average step cycle time is used as TCT. Machine Wet_1 will be down from day 3 to day 4, thus it limits the throughput of this simple line from days 3 afterwards. The model ran with machine Wet_1 down scenario, and average cycle time for the first 7 days was collected to construct policy Pred_HCT_SS. The step cycle time used to construct each policy is listed in Table 4. For both cases, TCT_SS was used to compute priorities of all the lots, and if lots have equal priority, first in first out (FIFO) was used to break the tie.

The only stochastic input data for this model is machine MTTR (mean time to repair) and MTBF (mean time between failures). Actual MTTR and MTBF from a fab in the past 90 days are collected to construct the distributions.

4.3 Simulation Results

The model did not ran with zero WIP. Instead, a modified real fab WIP snapshot for certain steps are used as initial WIP profile. The model ran with 300 replications and with simulation length of 7 days as our primary interest is the short term performance of these three line management policies. The performance metric is wafer outs (number of wafers leaving the fab in the chosen time period). Data was collected every 12 hours and the results are summarized in Table 5 and Figure 3.

Table 4: Step Cycle Time (seconds) Used for All Three Policies

Step sequence	TCT_SS	HCT_SS	Pred_HCT_SS
1	300	360	2672
2	900	3306	3453
3	1200	3742	10653
4	612	1609	6958
5	1200	2641	71088
6	1200	2401	7112
7	612	1362	3897
8	1200	2532	8479
9	1080	1995	6710
10	59652	48008	15087
11	720	720	764

Table 5: Average Wafer Outs by 12 Hour Time Period for All Three Line Management Policies

Time period	TCT_SS	HCT_SS	Pred_HCT_SS
12	625	625	625
24	593	593	567
36	633	633	629
48	644	643	638
60	633	634	643
72	337	341	412
84	50	51	74
96	2	2	3
108	378	379	380
120	551	550	552
132	566	566	566
144	581	581	579
156	605	605	605
168	620	620	620
7 days	6818	6823	6893

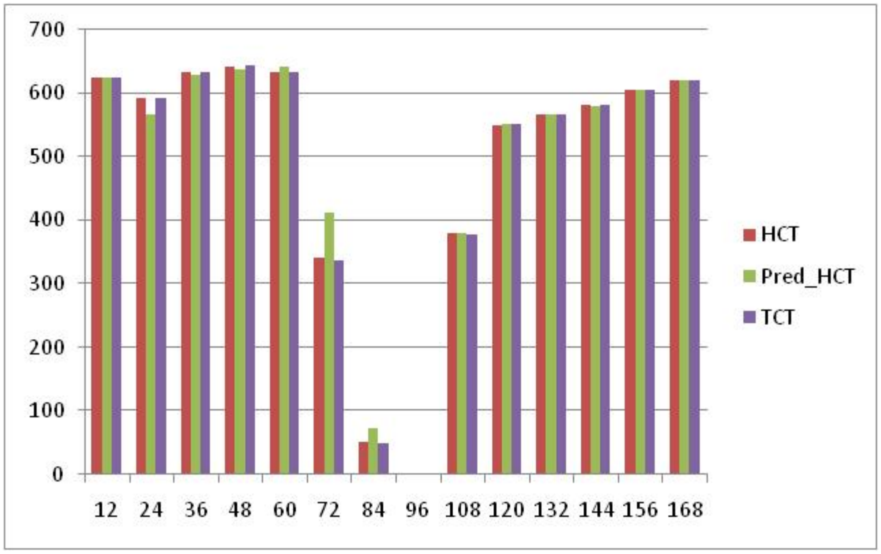

Figure 3: Average wafer outs by 12hour time period for all three policies

ANOVA analysis in Figure 4 showed that Pred_HCT_SS performs significantly better than the other two policies with p value of 0.05 and wafer outs improvement of 75 wafers. For Pred_HCT_SS, Step 5 cycle time was much greater than that for the other two policies, as shown in Table 4. Thus, SS was high at Step 4 and Step 4 was prioritized. Step 7 was de-prioritized because Step 7 and 4 shared the same machine. For the first 48 hours, more wafers went through Step 4 rather than Step 7. This caused the first 48 hour wafer outs reduced. However, more wafers went through Step 5 and accumulated in the backend of the flow during the first 48 hours period. Once Wet_1 machine was down for 48 hours starting from hour 48, no more new wafers could go through Step 5 to arrive at the backend of the line. For Pred_HCT_SS, at hour 48, the existing WIP at the back portion was greater than that for the other two policies, and resulted in improved wafer outs from hour 60 to 84. Note that for the 12 hour time period ending at hour 96, almost there was no wafer outs difference for all policies, as WIP has been depleted at the steps after Step 5.

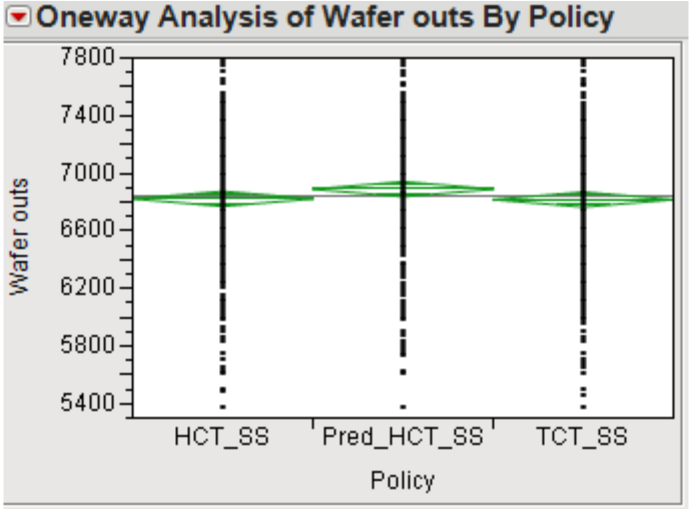

Figure 4: ANOVA analysis for wafer outs in 7 days for three policies

ANOVA analysis in Figure 4 also showed there is no significant difference between HCT_SS and TCT_SS. Step cycle time used for HCT_SS reflects step performance at steady state in the longer term. For this small example, it is not close to what will happen in the next 7 days. So it would not have benefit for the dynamic bottleneck scenario. It is also worth mentioning that HCT_SS may give higher priority to some non-bottleneck steps, compared to TCT_SS, as non-bottleneck step cycle time used in HCT_SS is greater than that used in TCT_SS. However, the difference was not big enough to impact the wafer outs in this example.

5 CONCLUSIONS AND FUTURE STUDIES

In this paper, two new line management policies, HCT_SS and Pred_HCT_SS, based on TCT_SS from SLIM method, are proposed in order to consider dynamic bottleneck better. The simulation model was used to evaluate the performance of these line management policies. Simulation results indicated that when a dynamic bottleneck is present, Pred_HCT_SS performs better over TCT_SS and HCT_SS in terms of wafer outs for the short term. However, based on the simulation experiment, there are several factors for Pred_HCT_SS to be successful. The first one is that we have enough information and knowledge on what will be the dynamic bottleneck in future. The second one is that assuming there is a good way to predict future step cycle time very well. Simulation can be one of the methods. Other factors include values of actual downstream WIP at the beginning, variations of MTTR and MTBF for non-bottleneck machines, target throughput rate, among others. The reason is that Pred_HCT_SS determines priorities based on these values. One simple case is that Pred_HCT_SS has little improvement when initial WIP in the line is low.

The next step will evaluate the performance of these line management policies for a full size fab model. Since Pred_HCT_SS does not guarantee to work successfully due to several factors mentioned above, another direction for future work is to consider fundamentally different line management approaches, such as optimization models or simulation models.

REFERENCES

Bixby, R., R. Burda, and D. Miller. 2006. "Short-Interval Detailed Production Scheduling in 300mm Semiconductor Manufacturing using Mixed Integer and Constraint Programming." In *Proceedings of the 2006 17th Annual SEMI/IEEE Advanced Semiconductor Manufacturing Conference*, 148-154.

David, F., A. Peikert, and K. Nishinohara. 1996. "Using Discrete Event Simulation to Analyze the Impact of Job Priorities on Cycle Time in Semiconductor Manufacturing." In Proceedings of 1996 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop, 151-155.

Hanny D. and S. Marteney 2011. "Advanced Predictive Scheduling Technology Can Improve Litho Cell Productivity." *Nanochip Fab Solutions* 6:1-2.

Horst, Z., I. Meents, B. R. Wheeler, and T. Hanschke. 2008. "A Queueing Network Based System to Model Capacity and Cycle Time for Semiconductor Fabrication." In *Proceedings of the 2008 Winter Simulation Conference*, Edited by S. J. Mason, R. R. Hill, L. Monch, O. Rose, T. Jefferson, J. W. Fowler, 2067-2074. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Klemmt, A., J. Lange, G. Weigert, F. Lehmann, and J. Seyfert. 2010. "A Multistage Mathematical Programming Based Scheduling Approach for the Photolithography Area in Semiconductor Manufacturing ." In *Proceedings of the 2010 Winter Simulation Conference*, Edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yucesan, 2474-2485. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Leachman, R. C., J. Kang, and V. Lin. 2002. "SLIM: Short Cycle Time and Low Inventory in Manufacturing at Samsung Electronics." *Interfaces* 32:61-77.

Monch, L., J. W. Fowler, S. Dauzere-Peres, S. J. Mason, and O. Rose. 2011. "A Survey of Problems, Solution Techniques, and Future Challenges in Scheduling Semiconductor Manufacturing Operations." *Journal of Scheduling* 14:583-599.

Subhash, C. S., A. Varadarajan, and L. Wang. 2011. "A Survey of Dispatching Rules for Operations Control in Wafer Fabrication." *Production Planning and Control* 22:4-24.

Wang, L. 2008. "Optimal and Approximate Algorithms for the Multiple-Lots-per-Carrier Scheduling and Integrated Automated Material Handling and Lots Scheduling Problems in 300mm Wafer Fabs." Ph.D thesis, Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia.

AUTHOR BIOGRAPHIES

LIXIN WANG is a manufacturing science engineer at the Industrial Engineering department at Micron Technology Inc., Virginia. He received B.S. in Mechanical Engineering from Tsinghua University, Beijing, China and Ph.D. in Industrial and Systems Engineering from Virginia Tech. His interest is mathematical modeling and simulation of semiconductor manufacturing systems. His e-mail is stanley-wang@micron.com.

Vinoth Chandrasekaran is a Manufacturing Science Engineer at the Industrial Engineering Department in Micron Technology Inc., Virginia. He received B.S. in Production Engineering from University of Madras, India and M.S. in Industrial Engineering from Arizona State University. His interest is in Statistics, Operations Research and Scheduling. His e-mail is vinothchandr@micron.com.