Proceedings of the 2011 International Conference on Industrial Engineering and Operations Management
Kuala Lumpur, Malaysia, January 22 – 24, 2011

Bottleneck Management Strategies in Semiconductor Wafer Fabrication Facilities

Shirley J. Tanjong
Department of Mechanical and Manufacturing Engineering
Universiti Malaysia Sarawak, Kota Samarahan, Sarawak 94300, MALAYSIA

Abstract

One of the manufacturing challenges in semiconductor wafer fabrication is the single machine unscheduled breakdown, which subsequently creates bottleneck. With WIP buildup due to unavailable machine, it would be a difficult task to decide on how to dispatch the WIP once the machine is available. Many methods and strategies have been presented by previous researchers, however these strategies has not been reviewed collectively. This paper aims to review and compare on the presented strategies. Based on the approach taken by these researchers, bottleneck management strategies can be categorized as TOC, LB and Optimization techniques.

Keywords
Semiconductor manufacturing, review, bottleneck management

1. Introduction
The intriguing complex nature that a semiconductor wafer fabrication facility, or wafer fab pose has motivates and brings in vast amount of researches and literatures, conducted by researchers both from the industry and academic field. The industry or the manufacturers are driven by the pressure to strive and survive in the competitive business by ensuring quality, customer satisfaction and operating cost is at par, if not ahead of the company's objective. Customers often demand for shorter cycle time and lower product cost. Failure to meet these demands will place the manufacturer at risk of losing a customer or a long term business relationship. A large number of researchers from the academic field partners with engineers from the industry to study various aspects of the wafer fab. Academicians are keen to explore this area of study as wafer manufacturing offers unique and challenging features in comparison to other manufacturing industry.

Scheduling is one aspect in semiconductor wafer fab that has gained researchers attention. Many methods and techniques were proposed but the goals are typically common: reduce cycle time, meet on time delivery (OTD) and increase throughput. Reference [1, 2] discussed on the two major production control methods in wafer fab, i.e. release strategy and dispatching strategy. Release strategy refers to the methods or practice of starting new lot into the production line. Most wafer fab employ dispatch strategies or rules such as First-In-First-Out (FIFO), earliest due date (EDD) and Critical Ratio (CR). Employing only a single dispatch rule in a wafer fab may no longer be a popular solution. It is common today that dispatching strategy is a combination of two or more rules. Reference [3] highlighted that as FIFO is superior in terms of cycle time performance, due date oriented rules like EDD and CR is less effective in terms of cycle time but provides good OTD record. Exhaustive simulation and experimental was done in [4] with the goal to compare the performance of various combination of dispatch rules in both make-to-order and make-to-stock wafer fabs. Reference [5] listed six main features of a wafer fab that contributes to the complexity of the production system: large number of processing steps, re-entrant process flows, batch equipment, random equipment breakdown, sequence dependent equipment setups and auxiliary resources requirement for some processes. In certain wafer fabs, other main challenges may includes high mix of product types (more than 100 product types) in low volume and running a mixture of products of different phases (development, prototype and production) within a single manufacturing facility.

Bottleneck sets the rate of throughput of a wafer fab. A machine or set of machines with the smallest capacity is the bottleneck of the production line. Commonly, photolithography tools are identified as bottleneck [6-9], however depending on the technology and process flow, certain wafer fabs may experience bottleneck on other areas such as implanters [10, 11]. Random unexpected breakdown of a single machine creates huge impact to a wafer fab

production. Reference [12] studied the evolution of WIP level and cycle time after a machine breakdown and highlighted that the WIP level may take weeks to return to normal after the machine is recovered. A single machine refers to the one and only resource that is available to process a particular step and/or recipe. A major breakdown may occur for as long as weeks and such occurrence creates bottleneck. The temporary shift of bottlenecks such as this creates wandering bottlenecks. With WIP buildup due to unavailable machine, it would be a difficult task to decide on how to dispatch the WIP once the machine is available. With its complex nature, a simplified strategy may not be sufficient to provide optimal solution. Thus, this paper aims to review bottleneck management strategies presented by previous researchers. The literatures included in this paper may not be extensive and does not cover all works published. However, the author considers that the amount of past researches included is sufficient to represent the general approaches proposed in the research works.

Based on the literature reviews gathered, the author suggests that bottleneck management strategies can be grouped into three main categories based on the method of approach taken by the researchers to counter the problem:

- Theory of Constraint (TOC)
- Line Balancing (LB)
- Optimization Techniques

The following part of this paper, Section 2, 3 and 4 will discuss and review on each category. Finally, Section 5 summarizes and concludes the review study.

2. Theory of Constraint (TOC)

TOC system control, which was popularized by Goldratt, is based on the theory of drum, buffer and rope [13]. Drum is the bottleneck that provides the beat or rate to be followed by all the other operations. Buffer, in terms of time, is kept before a bottleneck to make sure that the bottleneck always has material to work on. A bottleneck must never stop to wait for a material to arrive. Rope represents any kind of communication or information that coordinates all the activities within the system such that the Master Production Schedule (MPS) is met.

In the work of [14] TOC was applied with the end goal to improve the bottleneck management of a wafer fab. Based on TOC concept, the bottleneck management project was subdivided into identification, optimization and subordination. Identification of a bottleneck will be a straightforward process, because with the availability of historical data, the capacity of each machine or group of machines can be calculated. Reference [14] stated that in order to optimize the constraint, one must get the attention of almost the whole wafer fab. A bottleneck machine must run continuously, so additional and backup operators must be available to man the machine and even operator from non bottleneck areas must be trained and ready to be deployed to the bottleneck area when required. Spare parts must be permanently stocked for bottleneck machines while engineering works must be enhanced to increase the equipment's reliability and quality. Subordination is all about keeping to the beat of the bottleneck. Non bottleneck machines should be managed with the target to meet the needs of bottleneck machine. New lot starts, or lot released into the line will be assigned with a date based on when it is scheduled to be processed in the bottleneck machine. Due to the reentrant nature of the bottleneck, a lot will have multiple due dates and it is the responsibility of the non bottleneck areas to meet these due dates. In the occurrence of a wandering bottleneck due to unscheduled breakdown, a good application from [14] is to get the focus and support of the whole wafer fab. A wafer fab that operates with a common focus will results in a smooth flow of operation with minimal production conflicts. Reference [9] illustrated similar effort of implementing TOC in a wafer fab. Photolithography equipment was identified as the bottleneck in [9] and a drum-buffer-rope system was applied. The implemented system claimed to result in a twofold improvement of the wafer fab capacity.

The process of implementing TOC as described in [14] is more of a manual manner. Reference [15] proposed a method of implementing TOC within the dispatching algorithm. The algorithm proposed comprised both of CR and Hunger Ratio. CR targets to meet customer due date. Hunger Ratio aims to improve the manufacturing efficiency of the bottleneck and avoids the bottleneck from starving. It is calculated as the ratio of expected/needed time. A higher Hunger Ratio indicates WIP waiting plus WIP arriving at the bottleneck is insufficient to feed the bottleneck. Hunger Ratio is also used to schedule new lot starts into the production line based on the throughput rate of the bottleneck. Hunger Ratio is in fact another interpretation of bottleneck starvation avoidance. Earlier work of [16] presented on bottleneck starvation avoidance policy and proposed tool to detect if a bottleneck is in danger of starving. Due to computational and technology development, [15] has advantage over earlier researchers in the sense

that they are able to manipulate real time data of the wafer fab. Thus, the algorithm proposed is able to response to wandering bottleneck and product mix changes.

A more recent work of [3] proposed another bottleneck detection method with a corresponding dynamic dispatching strategy. Machine with the highest utilization is detected as bottleneck. The dispatch strategy is to prevent bottleneck from starving and avoid non bottleneck from high WIP. Bottleneck WIP level must be greater than the predefined lowest limit value and the non bottleneck WIP level must not exceed the predefined highest limit value. The detection method and the dispatch strategy suggested in [3] deem to be a proactive method in immediately identifying a bottleneck and in tackling the crisis in a proactive manner. However, certain aspects of the dispatch strategy will need to be evaluated further if a wafer fab is to employ this strategy: What is the value of lowest limit and highest limit? What would be frequency of detecting the bottleneck? Should the wafer fab identify only one bottleneck, top two or top three bottleneck machines?

An important point to note when applying TOC as the basis of control method in a wafer fab is that, the management team will need to decide on how frequent should they evaluate the capacity of the overall wafer fab. This is particularly important when a wafer fab experience a major product change. A different trend of volume mix and product types may results in changes of the bottleneck from one machine group or type to another. Note here that this change of bottleneck from one machine to another is not the ordinary wandering bottleneck. The new bottleneck will be the permanent bottleneck if the volume mix and product types remain unchanged. This will pose as a challenge to the management team to decide if the capacity of the new bottleneck should be increased by purchasing additional machine, thus the constraint of the wafer fab is shifted to a desired machine group. Reference [9] made an interesting remark in stating, although it is desirable that the bottleneck is not a reentrant process tool, photolithography aligners or steppers usually surface as the bottleneck as these are the most expensive equipment in the wafer fab.

TOC performs greatly in improving manufacturing efficiencies and equipment utilization. However, TOC has a drawback such that it does not consider the interdependence of cycle time and asset utilization [17]. Typical wafer fab performance objectives are cycle time, on time delivery, machine throughput and etc. It may be difficult for a wafer fab to decide if a control strategy based on TOC will provide the desired performance. It is impossible to be certain that a strategy based on TOC is the best method to manage bottlenecks. The best way to make a best guess is probably for the management team to identify what is the priority of each identified performance objectives such that decisions can be made in accordance to the priority.

3. Line Balancing (LB)

Bottleneck management strategies which integrate LB approach into its dispatch method or dispatch algorithm aims to distribute WIP from the bottleneck linearly throughout the wafer fab. In managing bottleneck, if the sole objective of the wafer fab is to maximize the bottleneck throughput and efficiency, then the solution to the problem is probably less complicated because the focus is only on the bottleneck area. Decision making shall be made only in the interest or benefit of the bottleneck. A number of researchers realized that taking this approach may results in negative impact on the wafer fab. Releasing of WIP from the bottleneck should also take into consideration of the downstream machines so that the WIP will not “flood” one area and left the remaining areas of the wafer fab “dry”.

Reference [18] introduced a dispatch model which allocates bottleneck resource with the purpose to reduce downstream machine lost time and improves downstream machine total moves. In this model, tool utilization and manufacturing efficiency of both the bottleneck and non bottleneck areas are taken into consideration such that the overall wafer fab efficiency is enhanced. The approach proposed by [18] in allocating WIP from the bottleneck machine group is rather simplified. Let say, there are six machines within the machine group of the bottleneck and this bottleneck area supply to two different downstream machine groups. Based on the WIP waiting before the bottleneck and the WIP waiting before the downstream machines, the model will decide how many bottleneck machines should be allocated to each downstream area; for example, three machines to one downstream area and another three to another downstream area. This model can also be applicable to single machine such that the resources of the bottleneck can be allocated by its capacity. In a similar example, if a single machine bottleneck has a capacity of running 40 lots per day and it supply to two different downstream machine groups, the capacity can be allocated so that 20 lots processed through the bottleneck will supply to one area and another 20 lots will supply to the other area.

Reference [19] integrates LB into bottleneck management and introduced a dispatch rule called Balance Work Content (BWC). The goal in BWC is to maintain throughput while balancing the work content. Work content is defined as the amount of time required to clear all the WIP waiting before the machine, or simply, WIP x Processing Time [19]. The BWC dispatch algorithm also includes the strategy of bottleneck starvation avoidance. Reference [20] presented a new WIP balancing concept called Toolset Available WIP Balancing (TAWB). The TAWB proposed supports efficient scheduling specifically for photolithography stage. Two main factors in TAWB scheduling are load levels of photolithography machines groups and production volume targets from the planning system. TAWB communicates with both the planning system and the manufacturing execution system (MES) to decide on the target production volume by product type and layer. Scheduling of other area, i.e. etching, diffusion, implant and thin film deposition should be determined by targeting to meet the photolithography schedule [20]. In the work of [21] an approach called target balance (TB) is used to model production target, due dates and WIP within each photo layer.

An aspect that should be considered when implementing LB in managing constraints is the balance between cycle time variance and WIP balancing effort. Let say, a bottleneck is supplying to five different machine groups and assuming that these five different processes have similar rate of throughput, should the WIP from the bottleneck be dispatched equally among the five machine groups? It will be a straightforward answer if the WIP mix waiting before the bottleneck is equally distributed for all the five machine groups but the solution can be complicated if, say half of the WIP waiting before the bottleneck need to be processed at one of the five machine groups. If the wafer fab tries to balance the line, then the WIP will be distributed equally among the five machine groups but what will be the impact of the WIP waiting before the bottleneck in terms of cycle time?

4. Optimization Techniques

With the advancement of computational capability, recent research works presented complex techniques that require excellent computer programming skills and higher level of operational research knowledge. Earlier researchers such as Goldratt has developed computer software called optimized production technology (OPT) [13], however it is limited by the incapability to communicate and manipulate real time data. Optimization approach is taken when the wafer fab tries to make a more accurate schedule or decision when multi-objectives is introduced into the system. Optimization technique deems by the author as another method of approach in bottleneck management strategies as the process of attempting the problem is dissimilar to TOC and LB. However, it is noted that an optimization strategy may includes TOC and/or LB as one of the objectives.

Reference [22, 23] applied the response surface methodology (RSM) to optimize the decision parameters generated by the dispatch algorithm. The proposed bottleneck dispatch model combines a set of dispatch rules, namely CR, Shortest Processing Time (SPT), Shortest Processing Time until Next Bottleneck (SPNB) and FIFO. Each lot are assigned to different priority class and the priority class is decided based on the decision parameters; hot lot, bottleneck lots, machine availability and bottleneck queue length. Subsequently, each priority class will be assigned different combination of dispatch rules. The decision parameters of the dispatch algorithm are optimized to enhance the system performance. The optimization approaches such as the one proposed in [22, 23] may be superior to other methods proposed in TOC and LB such that higher priority lots (hot lots) are taken into consideration in the scheduling. This may be an important factor especially in wafer fabs where customer has to pay higher price to have their lots running with higher priority. Reference [8] proposed an ant colony optimization based scheduling algorithm (LSA-WF) where the scheduling is made based on the following priority: 1) Schedule bottleneck machines, 2) Schedule batch processing machines with recipe constraint, 3) Schedule machines by wafer or lot, and 4) Schedule batch processing machines without recipe constraint. The optimize objectives of the algorithm are to maximize the movement of the jobs (lot moves) and minimize the total weighted tardiness of the jobs. In addition, the LSA-WF proposed is consists of two sub algorithms where LSA-WF1 works to dispatch tasks to each machine while LSA-WF2 sequences the jobs for each machine.

Incorporating optimization techniques in bottleneck management can be an appealing solution to wafer fabs as it considers multiple objectives in the decision making process. A simplified computational approach in finding optimal value may not be sufficiently robust and proactive in tackling bottleneck scenarios, thus automated and fast optimization technique will be desirable. However, complex computer programming with large set of data will not be desirable as it will results in a sluggish system. In addition, wafer fabs that are considering optimization techniques need to identify experts within their team that is familiar with these techniques. Indeed, there are

software packages that are readily integrated into existing system and easily employed by new user, however cost of purchasing, installation, training and expert advice need to be taken into consideration.

5. Concluding Remarks

This paper presented a review study on bottleneck management strategies proposed by previous researchers. Based on the approaches taken by these researchers, bottleneck management strategies can be classified into three main categories: TOC, LB and Optimization techniques. TOC supports the objectives of improving machine efficiency and utilization however it is not quite the answer in improving cycle time and on time delivery. LB promotes linear WIP distribution throughout the production line however, wafer fabs need to decide on a good balance between cycle time and machine utilization in order to employ LB. Optimization techniques may include TOC and/or LB and other decision parameters. While it does appeal as a superior solution for multi-objectives problem, cost and computational limitation may be the restraining factors. It may not be an easy job to identify which is the best approach to manage bottleneck, however a good starting point is for the wafer fab management team to identify the priority of each performance objectives. For future work, the author is interested to study and compare on the performance of each approach based on a set of the same data.

References

1. Lu, S.C.H., Ramaswamy, D., and Kumar, P.R., 1994, "Efficient Scheduling Policies to Reduce Mean and Variance of Cycle-Time in Semiconductor Manufacturing Plants," IEEE Transactions Semiconductor Manufacturing, 7(3), 374-388.
2. Pai, P.F., Lee, C.E., and Su, T.H., 2004, "A Daily Production Model for Wafer Fabrication," The International Journal of Advanced Manufacturing Technology, 23, 58-63.
3. Zhou, Z., and Rose, O., 2009, "A Bottleneck Detection and Dynamic Dispatching Strategy for Semiconductor Wafer Fabrication Facilities," Proc. of the Winter Simulation Conference, December 13-16, Austin, TX, USA, 1646-1656.
4. Bahaji, N., and Kuhl, M. E., 2008, "A Simulation Study of New Multi-Objective Composite Dispatching Rules, CONWIP and Push Lot Release in Semiconductor Fabrication," International Journal of Production Research, 46(14), 3801-3824.
5. Pfund, M.E., Mason, S.J., Fowler, J.W., 2006, "Chapter 9: Semiconductor Manufacturing Scheduling and Dispatching State of the Art and Survey of Needs," appears in Handbook of Production Scheduling, Hermann, J.W. (ed), Springer Science and Business Media, New York, 213-241.
6. Wu, M.C., Jiang, J.H., and Chang, W.J., 2008, "Scheduling a Hybrid MTO/MTS Semiconductor Fab with Machine-Dedication Features," International Journal of Production Economics, 112(1), 416-426.
7. Ding, S., Akhavan-Tabatabaei, R., and Shantikumar, J.G., 2007, "Stochastic Modeling for Serial-Batching Workstations with Heterogeneous Machines," Proc. of the IEEE International Conference on Automation Science and Engineering, September 22-25, Scottsdale, AZ, USA, 77-81.
8. Li, L., Qiao, F., Tian, X., and Wu, Q., 2009, "Ant Colony Optimization Based Scheduling for a Semiconductor Wafer Fabrication Facility with Bottleneck Stations," Proc. of the IEEE International Conference on Automation and Logistics, August 5-7, Shenyang, China, 520-525.
9. Kayton, D., 1998, "Using the Theory of Constraints' Production Application in a Semiconductor Fab with a Reentrant Bottleneck," IEEE/CPMT International Electronics Manufacturing Technology Symposium, October 19-21, Austin, TX, USA, 352-357.
10. Gupta, J.N.D., Ruiz, R., Fowlers, J.W., and Mason, S.J., 2006, "Operational Planning and Control of Semiconductor Wafer Production," Production Planning and Control, 17(7), 639-647.
11. Duwayri, Z., Mollaghasemi, M., and Nazzal, D., 2001, "Scheduling Setup Changes at Bottleneck Facilities in Semiconductor Manufacturing," Proc. of the Winter Simulation Conference, December 9-12, Arlington, VA, USA, 1208-1214.
12. Rose, O., 1998, "WIP Evolution of a Semiconductor Factory after a Bottleneck Workcenter Breakdown," Proc. of the Winter Simulation Conference, December 13-16, Washington, DC, USA, 997-1003.
13. Gaither, N., and Frazier, G., 1999, Production and Operations Management, 8th Edition, South-Western College Publishing, Ohio.
14. Villforth, R., 1994, "Applying Constraint Management Theory in a Wafer Fab," Proc. of the IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop, November 14-16, Cambridge, MA, USA, 175-178.

15. Rippenhagen, C., and Krishnaswamy, S., 1998, “Implementing the Theory of Constraints Philosophy in Highly Reentrant Systems,” Proc. of the Winter Simulation Conference, December 13-16, Washington, DC, USA, 993-996.
16. Lozinski, C., and Glassey, C.R., 1988, “Bottleneck Starvation Indicators for Shop Floor Control,” IEEE Transactions on Semiconductor Manufacturing, 1(4) 147-153.
17. Martin, D.P., 1997, “How the Law of Unanticipated Consequences Can Nullify the Theory of Constraints: the Case for Balanced Capacity in a Semiconductor Manufacturing Line,” Advanced Semiconductor Manufacturing Conference and Workshop, September 10-12, Cambridge, MA, USA, 380-385.
18. Yang, T.Y., Huang, Y.F., and Chen, W.Y., 1999, “Dynamic Dispatching Model for Bottleneck Resource Allocation,” IEEE International Symposium on Semiconductor Manufacturing Conference Proceedings, October 11-13, Santa Clara, CA, USA, 353-354.
19. Loo, H.L., Loon, C.T., and Soon, C.C., 2001, “Dispatching Heuristic for Wafer Fabrication,” Proc. of the Winter Simulation Conference, December 9-12, Arlington, VA, USA, 1215-1219.
20. Chung, J., and Jang, J., 2009, “A WIP Balancing Procedure for Throughput Maximization in Semiconductor Fabrication,” IEEE Transactions on Semiconductor Manufacturing, 22(3), 381-390.
21. Lee, B., Lee, Y. H., Yang, T., and Ignisio, J., 2008, “A Due-Date Based Production Control Policy using WIP Balance for Implementation in Semiconductor Fabrications,” International Journal of Production Research, 46(20), 5515-5529.
22. Zhang, H., Jiang, Z., Lee, Y.F., Ko, C.P., Choo, L.O.T, and Lim, L.P., 2007, “An Approach of Dynamic Bottleneck Machine Dispatching for Semiconductor Wafer Fab,” International Symposium on Semiconductor Manufacturing, October 15-17, Santa Clara, CA, USA, 1-4.
23. Zhang, H., Jiang, Z., and Guo, C., 2009, “An Optimised Dynamic Bottleneck Dispatching Policy for Semiconductor Wafer Fabrication,” International Journal of Production Research, 47(12), 3333-3343.