Proceedings of the 1998 Winter Simulation Conference
D.J. Medeiros, E.F. Watson, J.S. Carson and M.S. Manivannan, eds.

IMPLEMENTING THE THEORY OF CONSTRAINTS PHILOSOPHY IN HIGHLY REENTRANT SYSTEMS

Clay Rippenhagen
Shekar Krishnaswamy

Advanced Micro Devices
Austin, Texas 78741, U.S.A.

ABSTRACT

A significant challenge in implementing the Theory of Constraints in the semiconductor industry is the complex and reentrant nature of the manufacturing process. Managing a constraint or bottleneck in such processes is resultingly complex and is an ongoing topic of research. This paper describes a straightforward method to avoid starvation of possibly reoccurring bottleneck equipment. This method is not proximity sensitive, meaning that it does not act on material only if it is within a specified proximity of the constraint. This method is capable of reacting to changes in product mix and can act to correct line imbalances since it is based on real time data from the floor control system. The assertion is also made that for the case where a bottleneck has several occurrences in the product flow, all occurrences of the bottleneck must avoid starvation, not just the overall equipment queue. This affects the ability of the primary constraint to feed itself.

1 THE PROBLEM

In today's semiconductor industry we have a unique circumstance where the demand for semiconductor products is seemingly open ended but the cost of a new fabrication facility is in the billions of dollars. This requires the manufacturer to take advantage of every opportunity to increase the utilization of each facility. Even a small percentage increase in usable capacity translates to large additional profits and more market share. On the other hand, the semiconductor industry must face the difficult challenge of responding to a highly dynamic market characterized by rapidly changing demands and product mixes and sometimes very brief product life cycles. At the same time, failure of the manufacturer to deliver product on time can result in financial penalties or even loss of a customer. It is clear then, that semiconductor fabs have great costs associated with either under committing or over committing fab capacity and that short cycle times are of great advantage. Therefore the primary challenge in semiconductor manufacturing is to maximize the throughput of the facility while responding quickly to customer demands through low cycle times. How do we respond to this challenge?

2 ADDRESSING THE PROBLEM

Much of the industry has recognized the validity of the Theory of Constraints as explained by Eliyashu Goldratt in his book, 'The Goal' (Goldratt, 1993). This theory directly addresses our challenge. The Theory of Constraints (TOC) seeks to maximize system throughput while maintaining the minimum level of inventory. This also yields the lowest cycle time and therefore the best customer delivery performance. The basics of this theory are to identify the bottleneck, gauge the input into the system by the capacity of the bottleneck, never to let the bottleneck be idle, and then elevate the capacity of the bottleneck.

3 CHALLENGES POSED BY THE SEMICONDUCTOR INDUSTRY

The challenge has been in how to implement the theory in a complex reentrant system such as semiconductor fabrication. For example, what if the bottleneck tool is encountered several times in the process? Which occurrence of the bottleneck do you run first? Does it matter? What is a practical approach of driving inventory to the bottleneck based on the bottleneck's consumption rate? As far back as 1988, Glassey and Lozinski (1988) discussed techniques to detect starvation of the bottleneck in semiconductor manufacturing. Through the years, Glassey has described various methods to accomplish this goal. The solutions have ranged from graphical assistance for operators to queue predictions based on simulation experiments. Glassey (1989) has also attacked the problem of regulating the flow of material into the process flow based on a linear control rule called descending control. These approaches have been constrained by the unavailability of real time data.

4 FEEDING THE CONSTRAINT

The solution presented here to the problem of feeding the constraint is based on an established methodology called critical ratio. Critical ratio is a method to drive dispatching decisions based strictly on customer due date. The ratio is simply the time the lot is expected to take to complete divided by the time until it is needed to be complete. This ratio of expected/needed time gives a higher priority to product that is farther behind schedule. This concept of driving to an end point based on certain criteria is a good one, but it is obvious that there are no manufacturing considerations here. Operations people have long been aware that none of the mainstream dispatching schemes are based on or even consider the need to manage the production system in an efficient manner. Therefore, the current need is to provide a dispatching algorithm that considers not only customer delivery requirements, but also manufacturing efficiency as a means to that end.

5 EFFICIENCY AT THE CONSTRAINT

The Theory of Constraints asserts that the manufacturing system as a whole can reach optimum efficiency by identifying and feeding the constraints of the system. The results of this optimization are to minimize inventory and reduce cycle times and therefore reduce manufacturing costs and yield more aggressive and accurate customer delivery dates. This process of feeding the constraint becomes significantly more difficult when the process is highly reentrant, and the constraint tools are encountered many times in the process at differing process rates. How then, is the constraint to be fed and still maintain a linear inventory profile? This is done by recognizing that not only must the constraint tool be fed, but that all occurrences of the constraint tool be fed, and that this must done in a consistent and equitable manner. If this very important concept is not comprehended, the manufacturing system could be left with the circumstance where there are mounds of inventory in front of the constraint tool, but it is all bound for a single occurrence of the tool with all other occurrences left dry. In this case a bubble of inventory has been created in one location and a hole in the inventory profile created in other locations. A reentrant constraint must ensure that it feeds itself. Add to this the concept of maintaining a minimum buffer in front of each occurrence of the constraint to protect it from the inevitable disruptions and fluctuations that could impede a constant flow to the constraint.

By considering each occurrence of the constraint an endpoint, work in process (WIP) can be driven to the constraint based on criteria relevant to the constraint tool. If the goal is to keep each occurrence of the tool from starving, then the ratio concept of expected/needed time can be used. The expected time is the cycle time for a lot to get to the constraint, and the needed time is the time that the lot is needed at the next occurrence of the constraint in order to keep it from starving (or depleting a minimum buffer). This ratio will be referred to as the Hunger Ratio, because it describes the degree of "hunger" that the constraint experiences for each lot. The more material that is in the constraint's buffer or is likely to arrive before the current lot, the lower the Hunger Ratio for the current lot will be. A further advantage of using this 'ratio concept' is that it is not proximity sensitive. In other words, material need not be within a certain proximity to the constraint in order to be considered. Traditional approaches have only attempted to analyze WIP in a narrow window of upstream and downstream processing steps. The disadvantage of this type of approach is illustrated in Figure 1.

In the example, the proximity view favors processing lots in occurrence 2 over occurrence 1 since it appears that this occurrence has substantial WIP upstream and no WIP downstream to feed the bottleneck. A global view would clearly eliminate this problem.

Just as the "critical ratio" represents customer delivery requirements, the "hunger ratio" represents the manufacturing efficiency aspect of dispatching. These two ratios can be combined to provide a balanced and synergistic approach to scheduling. This Hunger Ratio concept can also be used to regulate the rate of starts into the manufacturing system based on the consumption rate of the bottleneck. This is another major component of the drum-buffer-rope model used in the Theory of Constraints. Other variations can be used to schedule non-production material which is driven by stock point levels (or in TOC terms a buffer level), rather than due dates.

6 THE ROLE OF SIMULATION MODELING

Simulation modeling has become an indispensable tool in the effort to accurately assess the true capacity of a facility, allowing manufacturers to more confidently commit the full capacity of the fab. Any company who fails to avail themselves of this competitive advantage in an aggressive market does so at its own peril. Simulation modeling has also provided an economical way to evaluate different dispatching philosophies, and with the marriage between simulation and real time dispatching systems, that advantage is being translated into reality.

Simulation modeling has been widely used in the semiconductor industry for a wide range of strategic objectives such as fab design, equipment selection, capacity and cycle time planning, etc. Its usage in the tactical area has been relatively limited. The pertaining areas include short-interval scheduling, dispatching of individual lots, scheduling of equipment events, short-term deployment of operators, etc. Historically, lengthy development and execution times made simulation models

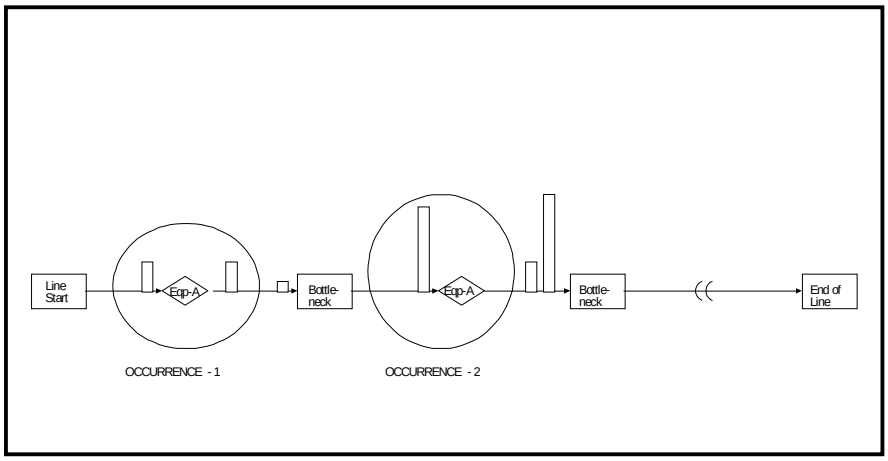


Figure 1: Global View vs. Local View: An Example

impractical to use in this area. Over the years however, commercial fab simulators have been developed that dramatically reduce the time for model development and are optimized for performance. Moreover, the cost of computer hardware has also declined steeply making it affordable to run models on high-end workstations. In spite of all these improvements, changes analyzed and deemed as beneficial by simulation models do not necessarily translate into actions in the shop floor on a day to day basis.

The primary reason for this is the lack of integration between the Manufacturing Execution Systems (MES) and the simulation and rule development systems. The majority of the information used in these two systems is common information since the simulation tool is essentially trying to mimic the physical manufacturing system. The integration and synchronization of these two systems and the use of common data is imperative in the attempt to both maintain accurate and timely simulation models as well as to translate simulation based improvements to the physical facility.

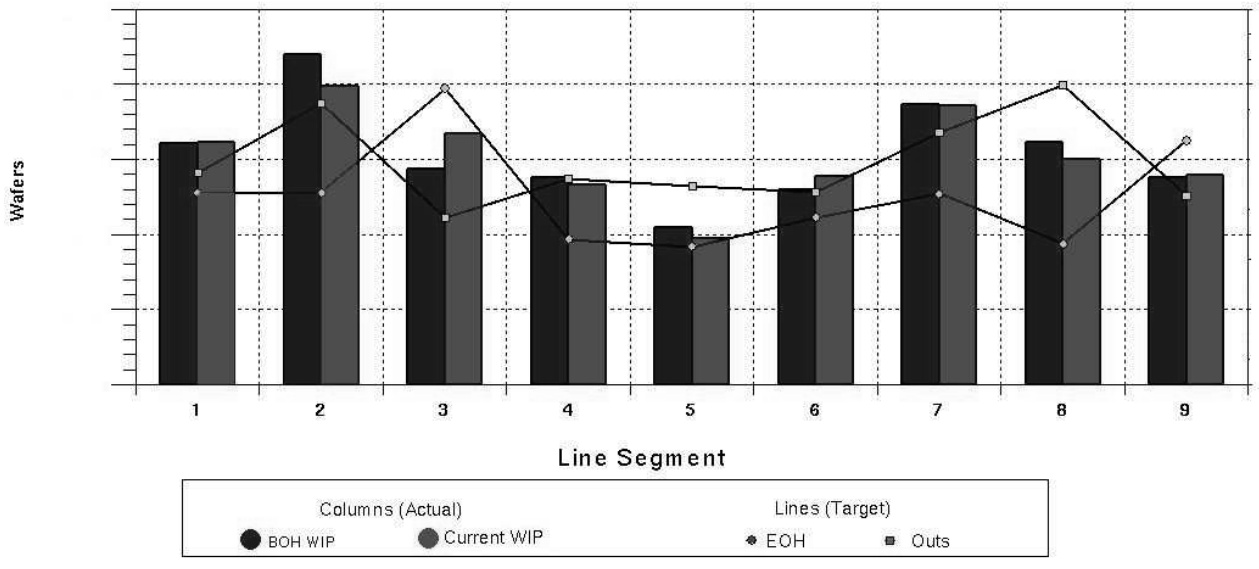
Currently data models that exist within the simulation environment encompass the basic constructs of the fab and their associated relationships, but they vary significantly from that found in the MES. Examples of these include product, route, equipment, personnel, dependent resources like masks, load boards, etc. In a tactical environment, it is essential that the virtual constructs and associated rules mesh with the real world.

7 IMPLEMENTATION- AMD'S METHOD OF INTEGRATING SIMULATED RULE LOGIC WITH DISPATCH

AMD has embarked upon a Real Time Dispatch (RTD) project with the objective to implement dispatching for equipment groups. The system used is the AutoSimulations Real Time Dispatch product called ASI-RTD. It provides the capability to implement custom rules specific to certain operations or a group of operations. An important aspect is the capability to exercise these rules using real-time information. The system's functionality will include the ability to develop and test rules or heuristics, in an off-line mode using simulation modeling. Rules deemed beneficial via simulation would then be transferred to the MES (Consilium's WorkStream) for controlling the order of processing. This system has been implemented in AMD Fab25 and is being managed by a team consisting of representatives from the production planning, manufacturing operations, modeling, and computer integrated manufacturing groups.

The combination of Hunger Ratio and Critical Ratio was implemented in this environment as dispatching rules. These rules were simulated and analyzed prior to implementation. The degree of emphasis on manufacturing efficiency and customer delivery can be empirically determined using simulation techniques. This dispatch logic has been implemented at all processing steps other than the bottleneck, with the emphasis being to move material to the bottleneck equipment at all occurrences. This approach implemented the TOC principle of making sure the bottleneck is fed. The other major principle of TOC is to elevate the capacity of the bottleneck. AMD has implemented this with a rule that is based on Hunger Ratio, but that also seeks to optimize setups of the bottleneck equipment and considers the availability of peripheral resources. This rule also relies heavily on real time data.

Apart from intelligent dispatching, this system has already provided other benefits. It has provided a user-friendly access to the MES data that enables users to develop reports, which in the past could be done only by programmers proficient in COBOL and the MES system.



	1	2	3	4	5	6	7	8	9	Total
BOH_WIP	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX
Current WIP	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX
Plan EOH	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX
Actual CT	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX
WIP Turns	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX
Start Oper	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX
Outs Oper	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX
Target CT	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX
Target Outs	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX	XX.XX

06/26/98 06:00:00 to 06/26/98 16:55:05

Figure 2: Actual and Target WIP Levels

These reports are currently providing timely and accurate information for decision making purposes.

Figure 2 is a daily chart that describes actual and target WIP levels for segments of the production line, with the targets being set via simulation analysis. It also provides information in tabular format on WIP turns and cycle times for each segment.

REFERENCES

Glassey, C. R. and Christopher Lozinski. 1988. Bottleneck starvation indicators for shop floor control. *IEEE Transactions on Semiconductor Manufacturing*, Vol. 1, No. 4.

Glassey, C. R, Jeyaveerasingam George Shanthikumar, and Sridhar Seshadri. 1989. Linear control rules for production control of semiconductor fabs. *IEEE Transactions on Semiconductor Manufacturing*, Vol. 9, No. 4.

Goldratt, Eliyashu M. 1993. *The goal*. New York: NAL/Dutton.

AUTHOR BIOGRAPHIES

CLAY RIPPENHAGEN is an Industrial Engineering Supervisor at Advanced Micro Devices (AMD) and is responsible for modeling, simulation and dispatching activities in Fab 25. Prior to AMD, Clay was a Systems Engineer in IBM's Southwest Marketing Division and an Industrial Engineer in IBM's Systems Technology Division in Austin, TX

SHEKAR KRISHNASWAMY is a Senior Manufacturing Engineer at Advanced Micro Devices (AMD) and is the project manager for the implementation of the Real Time Dispatch project. Prior to AMD, Shekar was an IBM assignee to the semiconductor consortium, SEMATECH, and an industrial engineer at IBM Microelectronics in East Fishkill, NY. His professional interests lie in the area of Scheduling & Dispatching, Simulation Modeling and their integration with planning and Manufacturing Execution Systems (MES).