Proceedings of the 2004 Winter Simulation Conference
R .G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

ANALYSIS OF A BORDERLESS FAB SCENARIO IN A DISTRIBUTED SIMULATION TESTBED

Peter Lendermann
Boon Ping Gan
Yoon Loong Loh

Production and Logistics Planning Group
Singapore Institute of Manufacturing Technology
71 Nanyang Drive, Singapore 638075, SINGAPORE

Hiap Keong Tan
Sip Khean Lieu

Chartered Semiconductor Manufacturing Ltd.
60 Woodlands Industrial Park D
Street 2, Singapore 738406, SINGAPORE

Leon F. McGinnis

School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205, U.S.A.

John W. Fowler

Department of Industrial Engineering
Arizona State University
Tempe, AZ 85287-5906, U.S.A.

ABSTRACT

Distributed simulation based on the High Level Architecture standard is adopted to realize the simulation of a borderless fab that involves two wafer fabs located in close proximity. The two fabs pool together their resources for capacity sharing. To demonstrate the benefits of this concept, experiments were conducted to measure the cycle time changes resulting from introduction of an additional product into either one of the fabs. In the case without cross fab material flow, the capacity of each fab alone is not sufficient to handle the increasing release rate of the new product as bottleneck machines surface. However, for the cross fab case where the front-end of the new product's process is done in the first fab, while the back-end in the second, it is possible to avoid the bottleneck situation. As a result, the two fabs are able to increase their aggregated capacity without investing in new equipment.

1 INTRODUCTION

Semiconductor manufacturing operations face three fundamental drivers of change: cost per function, speed of delivery, and profitability. Cost per function improvements (the result of "Moore's law") are fundamentally technology driven, achieved by making smaller features on bigger wafers with better yields. Speed of delivery is the result of both technology and market forces, where the ability to design products faster enables more product innovation and greater access to the market "sweet spot" (see, e.g., (Leachman et al. 1999) on the value of speed to market). Shareholders demand acceptable returns, requiring competitive products delivered to customers at competitive costs.

Together, these three drivers are dramatically altering the landscape for semiconductor manufacturing operations. Complexity is increasing rapidly—greater product complexity (more masks, more materials, more process steps), and more products with shorter lifecycles and faster ramps—at the same time capital cost of factories is stretching toward $3.5 billion. The cost of factories, the optimum scale of operation, and the search for high asset utilization creates opportunities for foundries and contract manufacturers, transforming integrated command-and-control of supply chains into distributed negotiate-and-compromise in complex supply networks.

As features become smaller and wafers become larger, a single bottleneck process represents a larger increment of device capacity. Fine-tuning fab capacity becomes more difficult, and the impacts of bottleneck process failure or sudden changes in product mix are magnified. Foundries provide access to capacity, perhaps with cost and speed penalties. The basic research question is "How can we establish a strategic plan for in-sourcing and out-sourcing, and how can we manage the transition of products between sources?" Especially in an environment of product proliferation, extremely competitive markets, and volatile global economies, these are quite difficult questions to answer.

For wafer fabs in particular, as mentioned above, a major difficulty is the high capital cost. Also, because of the upstream position in the supply network for electronics

products the effects of the "bullwhip" effect are more apparent and the ability to react to demand changes is more important. Efficient capacity management in between wafer fabs is most critical in addressing these difficulties.

There also is the more prosaic problem of coordinating operations to achieve capacity sharing, as in (Lee 2002) and (Weng 1998). For the situation in which several wafer fabs are located within reasonable proximity, and are willing to pool some fraction of capacity, there is the basic question of deciding when and how to realize a "borderless fab" through exploitation of capacity pooling. There already are examples of products being moved between facilities for non-lithographic processes. This problem is especially interesting when the fabs do not want to share detailed information about plans and schedules of products not involved in capacity pooling.

Many semiconductor manufacturers have more than one fab qualified to produce a given semiconductor device (IC). In addition to these options, outsourcing to one or more foundries is increasingly a feasible option. This leads to the need for two types of decisions concerning these facilities.

The first decision includes how to allocate work to production segments or resource groups. This decision is labelled in the literature as mid-term or aggregate Production Planning (PP) (Stadler and Kilger 2002). It is essentially very similar to the long-term Master Planning. However, rather than the whole Supply Network, only the production processes of a single echelon are considered. The planning tasks tackled are, e.g., the allocation of production quantities (of product groups) to the production segments (fabs), production smoothing (by means of subcontracting, seasonal inventory, back-logging or external purchasing), and aggregate lot-sizing for groups of final items.

The second decision arises when there is an unexpected disturbance that puts one of these facilities significantly behind in its commitments. In this case, the company may decide to "cross site" wafers between fabs, by moving partially completed wafers from one fab to another. Both of these decisions require sharing data and information between the facilities.

2 DISTRIBUTED SIMULATION TESTBED

A "borderless fab" is a large and complex dynamic system, with many sources of uncertainty, and aspects that simply cannot be modeled analytically. Currently, the most promising approach for developing understanding, theory and methods for analyzing such systems is to study simulations of them. In this kind of research, achieving high fidelity simulation is more important than achieving fast run times.

For systems as large and complex as a "borderless fab", distributed (or federated) simulation is the only realistic approach to model development and model maintenance (and also is a more accurate representation of reality) for the following reasons:

1. Maintaining a representation of several fabs in one single model would be complex and difficult to handle.
2. Since the individual simulation models could also be used for online-scheduling purposes in the respective fabs, they need to be maintained locally in the fabs.
3. Running the simulation in a distributed "Plug & Play" environment will enable (a) subsequent incorporation of planning procedures into the simulation that will further enhance the quality of the model representation and the range of experiments that can be conducted, and (b) subsequent extension of the simulation to downstream nodes of the supply chain (assembly & test) and optimization of the coordination mechanisms between wafer fab and assembly & test, with the flexibility of integrating an assembly & test model based on any simulation software that is compliant with the synchronization middleware, but without having to share sensitive information in one single simulation model.

Consequently, an important component of research in this area is the development of a distributed simulation testbed necessary to test and evaluate the theories and methodologies that result from work on the identified integration problems. We are using the High Level Architecture (HLA) as the mechanism for integrating federates representing supply network operations and decision-making (DMSO, 2004), (Lutz, 1998). Currently, the HLA (IEEE standard 1516) is emerging as a standard for Plug & Play of simulation-based decision support components for manufacturing and logistics systems. It is driven by the HLA Commercial Simulation Package Integration Forum (HLA-CSPIF, `http://www.cspif.com/`) which is co-founded by Singapore Institute of Manufacturing Technology (SIMTech) and endorsed by the Simulation Interoperation Standards Organization (SISO).

Earlier work by Georgia Tech and SIMTech has demonstrated the technical feasibility of distributed supply chain simulation in the context of semiconductor manufacturing (Lendermann et al. 2003a). For a totally integrated testbed, the best way to achieve an optimal representation of the customer order management and scheduling processes in a simulation testbed is to incorporate and reuse the corresponding software applications within the simulation by wrapping them as HLA federates (Lendermann et al. 2003b).

Although we have already identified a way to make a commercial simulation package such as AutoSched AP

compliant with the HLA through a middleware that is able to transfer and manage the synchronization messages between the simulation package and the Runtime Infrastructure of the HLA and does not require any modifications within the simulation packages, we used SIMTech's existing C++ waferfab simulator for the purposes of this project. Two instances of this simulator, based on the industry datasets were connected through the HLA-RTI.

3 BORDERLESS FAB ANALYSIS

To demonstrate the benefits of the borderless fab and the use of the HLA-RTI to enable seamless integration of simulation models, a case study comprising two wafer fabs was conducted.

3.1 Simulation Model

The two wafer fabs modeled for have similar capabilities of producing ten wafer product types (0.35 micro technology logic devices and antifuse gate devices) but their capacities are different. The process flows of the wafer products considered range from 200 to 300 steps. There are a total of 73 machine sets modeled, including wet benches, furnaces, steppers, implanters and metrology tools. The downtime behavior of each machine set is also modeled.

The need to route wafer lots from one fab to another can be triggered by two conditions: 1) unavailability of resources due to a resource breakdown, 2) insufficient capacity in one fab while appropriate resources are available in the other fab. Lots are accumulated before routing. Rerouted lots can either continue in the destination fab or be moved back to the original fab once processing by certain resources is completed, depending on which option is more cost effective. Due to the close proximity of the two fabs considered in this study, time to transfer lots from one fab to another has been capped at an hour, including handling time and transportation delays.

3.2 Scenarios

For the following scenarios whereby the two fabs are manufacturing 9 types of products out of 10, the simulation testbed is used to measure the impact of introducing the tenth product (known as Product X hereafter) on the fabs' current production levels. Three particular scenarios were studied: 1) First Fab - introducing Product X to the first fab only, 2) Second Fab - introducing Product X to the second fab only, and 3) Cross Fab - introducing Product X to the first fab, and reroute Product X to the second fab when Product X reaches an inspection step in the middle of its process flow (Step 117). In these three scenarios, loading of Product X is increased until the fabs' capacity threshold has been reached. In Scenario 3, four lots are accumulated before moving to the second fab. This helps to fill up furnaces in the second fab for better utilization.

3.3 Experimental Results

The experimental results obtained here are for a simulation run length of 2 years, having each fab running on separate computers, communicating through the HLA-RTI. Product X was first released to the fabs based on a release interval of 50 hours per lot. The release interval is then reduced by a factor of 1.5, 2.0, 2.5, 3.0, 3.5, and 4.0. As observed in Figure 1, the capacity of the first fab and the second fab on its own is insufficient to handle the loading of Product X. The cycle time of Product X increases from 12 to 21 days for the first fab, and from 10 to 14 days for the second fab as the release interval is reduced. However, if we allow Product X to cross fab, the cycle time decreases from 13 to 10 days. This is due to the fact that the reduced release interval has also reduced the time that is required to accumulate four lots need prior to routing to the second fab.

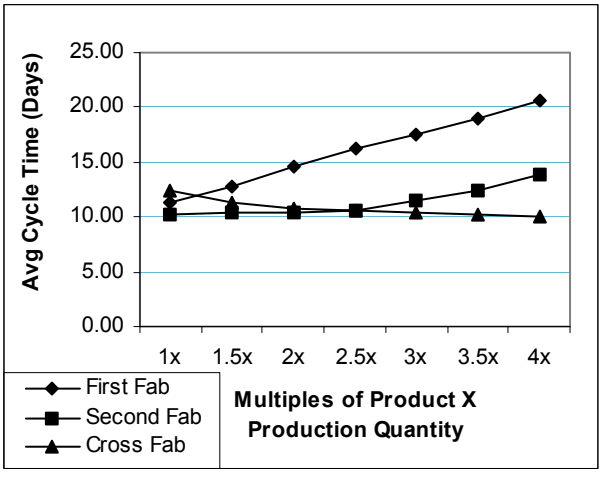

Figure 1: Effects of Cross Fab Production for Product X

Figures 2 and 3 show the cycle time of the other 9 products for the First Fab and Second Fab scenarios, respectively. It is apparent from these figures that the introduction of Product X to the first fab and second fab (standalone basis) has breached the capacity threshold of both fabs. As the release interval for Product X is reduced, the cycle time of all 9 products increase significantly. However, the cycle time of the 9 products remain unchanged in the Cross Fab scenario.

This can be explained by referring to Figure 4 and 5. When Product X is introduced to the first fab alone, it creates a bottleneck at Furnace A (used at Step 192). The average waiting time at the furnace increases from 1 day to 8 days. On the other hand, when Product X is introduced to the second fab alone, it creates a bottleneck at Furnace B (used at Steps 5, 17, 24, 42, and 56). The average waiting time at the furnace increases from 4 minutes to 16 minutes. In the cross fab scenario, we reroute the lots of Product X at Step 117 (after inspection) in the first fab thereby avoiding the creation of bottlenecks at Furnace A. Similarly,

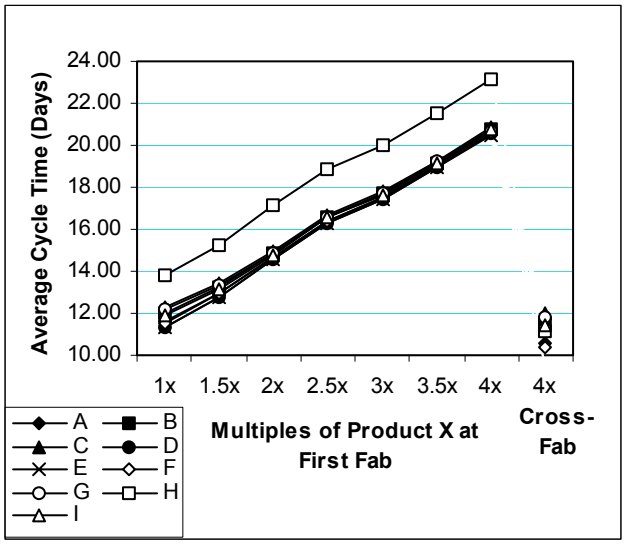

Figure 2: Effects of Product X Introduction on Other Products in the First Fab Scenario

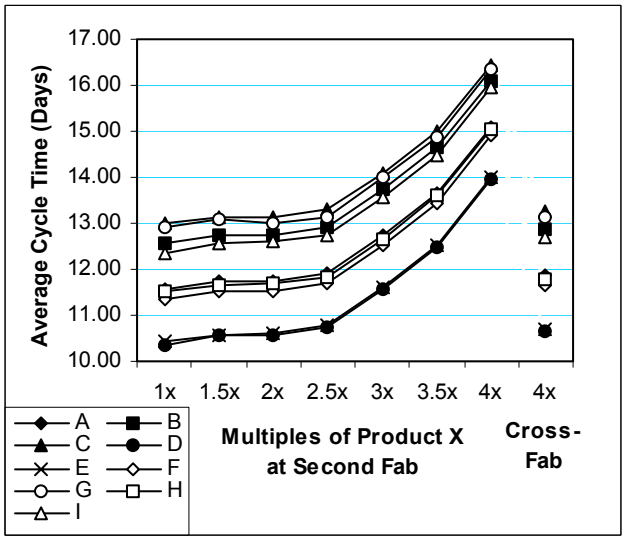

Figure 3: Effects of Product X Introduction on Other Products in the Second Fab

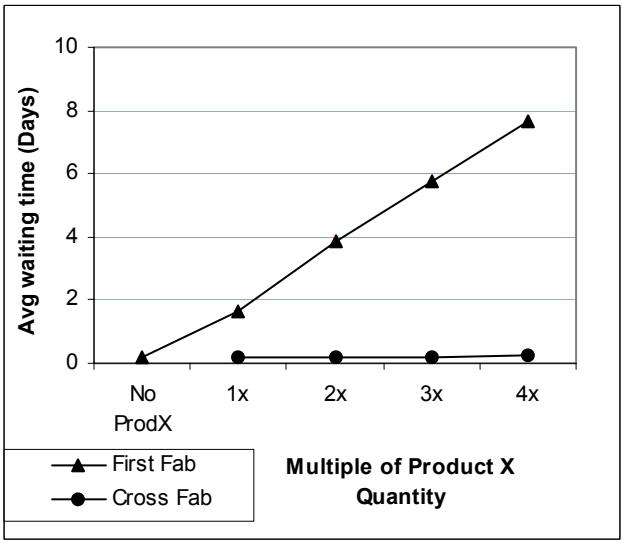

Figure 4: Effects of Cross Fab Production on Waiting Time (Bottleneck Furnace A at First Fab

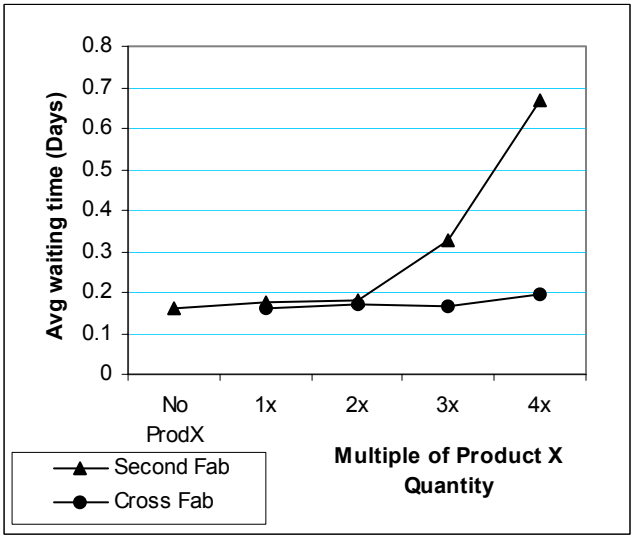

Figure 5: Effects of Cross Fab Production on Waiting Time (Bottleneck Furnace B at the Second Fab

Product X is only introduced to the second fab after Step 117 which again avoids creation of bottlenecks at Furnace B. The findings clearly demonstrate that the cross fab scenario is more suitable to handle the introduction of Product X into the manufacturing process.

4 TIME MANAGEMENT ISSUES

Realization of this Plug & Play simulation also involves investigation of new time management mechanisms to account for the need for aperiodic synchronization of simulation models to portray the material flow scenarios. Material flows from one fab to another are represented by a time stamp-order interaction, which incorporates information such as the product type, the process flow the product follows, the step the lot will continue on in the destination fab, the name of the destination fab and the quantity of the lot. The material flow interaction is sent only when four lots are accumulated, based on the scenario presented earlier. This means that the time at which the interaction is sent cannot be determined in advance. This introduces a constraint to which the fabs can advance their time as we cannot be sure which event of the simulation can trigger an interaction. Consequently, the safest way to advance the time of the fabs would be to request for time advance permission from the RTI for every event that the fab is simu-

lating. This restricts the simulation progress unnecessarily and in turn worsens the execution time of the simulation.

To overcome this problem, we first need to narrow down the events that potentially can trigger external events. We called them potential events hereafter. With this knowledge, we can attempt to advance the time of the fab up to the time that these potential events are triggered. This helps to reduce the number of time advance requests to the RTI. Based on the simulation scenario of this study, an interaction will only be sent when a lot arrives at a step that will reroute the lot to the second fab. We define the lot arrival to this step as the potential event type. Whenever such an event exists in the event list, we will attempt to advance the time of the fab up to the timestamp of this event. When such an vent does not exist, we revert to a time-step approach. The time of the fab is advanced with a time step of t_s where t_s is the smallest time interval in which a potential event can be triggered. With this approach, we manage to relax the synchronization constraint of simulation and improves the execution time of the simulation by approximately 5 times, compared to the case in which time advance is requested for each event of the fab. The execution time achieved using this mechanism is approximately in the range of 10 to 15 minutes for two years of simulation time, depending on the scenario being simulated.

5 CONCLUSIONS AND FUTURE WORK

Through distributed simulation, the case study illustrated the benefit of capacity pooling between two wafer fabs with similar manufacturing operations. The simulation models of each fab were executed on a separate computer and communicated through the RTI. This simulation can be used to evaluate various policy of triggering lot rerouting or to study trade-off of sharing capacity with different fabs. Questions associated with the performance of the borderless fab can be answered using this simulation testbed. The simulation is not restricted to only two fabs as in the case study. It can be extended easily to include other simulation models that are HLA-compliant, through modification of the fab's simulation object model (SOM).

Some work still needs to be done to further explore the synchronization mechanism discussed in this paper. The mechanism is customized for the scenario that we were investigating in this case study. Generalizing the mechanism to handle different kind of simulation scenarios is crucial as different scenarios may introduce different synchronization constraints to the simulation.

REFERENCES

DMSO. 2004. High Level Architecture. <https://www.dmso.mil/public/transition/hla/> [accessed April 13, 2004].

Leachman, Robert C., J. Plummer and N. Sato-Misawa. 1999. Understanding Fab Economics. Report CSM-47, Engineering Systems Research Center, University of California at Berkeley. Berkeley, CA 94720.

Lee, Young Hae. 2002. Advanced Planning and Scheduling with Outsourcing in Manufacturing Supply Chain. *Computers and Industrial Engineering* 43 (1-2): 351-357.

Lendermann, P., N. Julka, B.P. Gan, D. Chen, L.F. McGinnis, and J.P. McGinnis. 2003a. Distributed Supply Chain Simulation as a Decision-Support Tool for the Semiconductor Industry. *Simulation* 79 (3): 126-138.

Lendermann, P., N. Julka, L.P. Chan, and B.P. Gan. 2003b. Integration of Discrete Event Simulation Models with Framework-Based Business Applications. In *Proceedings of the 2003 Winter Simulation Conference*, ed. S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 1797-1804. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Lutz, R. 1998. High Level Architecture Object Model Development and Supporting Tools. *Simulation* 71 (6): 401-409.

Stadtler, H., and C. Kilger (ed.). 2002. *Supply Chain Management and Advanced Planning*. 2nd edition. Springer Verlag, Berlin.

Weng, Z.K. 1998. Managing Production with Flexible Capacity Deployment for Serial Multi-Stage Manufacturing Systems. *European Journal of Operational Research* 109 (3): 587-598.

AUTHOR BIOGRAPHIES

PETER LENDERMANN is a Senior Scientist at the Singapore Institute of Manufacturing Technology (SIMTech). Previously he was a Managing Consultant with agiConsult in Germany where his focus was on the areas of supply chain management and production planning. He also worked as a Research Associate at the European Laboratory for Particle Physics CERN in Geneva (Switzerland) and Nagoya University (Japan). He obtained a Diploma in Physics from the University of Munich (Germany), a Doctorate in Applied Physics from Humboldt-University in Berlin (Germany) and a Master in International Economics and Management from Bocconi-University in Milan (Italy). His research interests include modeling and analysis of high-tech manufacturing networks and distributed simulation. His email address is <peterl@SIMTech.a-star.edu.sg>.

BOON PING GAN is a Research Engineer with the Production and Logistics Planning Group at the Singapore Institute of Manufacturing Technology. He is currently leading a research project that aims to apply distributed simulation technology for supply chain simulation. He received a Bachelor of Applied Science in Computer Engineering and a Master of Applied Science from Nanyang

Technological University of Singapore in 1995 and 1998, respectively. His research interests are parallel and distributed simulation, parallel programs scheduling, and application of genetic algorithms. His email address is <bpgan@SIMTech.a-star.edu.sg>.

YOON LOONG LOH is a Research Scientist with the Production and Logistics Planning Group at Singapore Institute of Manufacturing Technology. He graduated from Nanyang Technological University (Singapore) with a second class upper honor in Mechanical and Production Engineering in 1989. He later continued to pursue his M. Eng degree in the area of Optimization and Simulation in 1994. Mr. Loh has since been a consultant for the manufacturing industry and has been a Certified Industrial Automation Consultant since 1996. His research interest is in the systems optimization and facilities layout planning using simulation. His email address is <ylloh@SIMTech.a-star.edu.sg>

HIAP KEONG TAN is a Industrial Engineering Manager with Chartered Semiconductor. He is responsible for cross fab productivity activities and manages the 300mm plant layout and capacity. He received a Bachelor in Mechanical Engineering and a Master in Industrial and System Engineering from National University of Singapore. His email address is <hktan@charteredsemi.com>.

SIP KHEAN LIEU is a Senior Industrial Engineer with Chartered Semiconductor. He is currently leading simulation projects in the company which include the 300mm Fab. He held various positions in manufacturing with Leica, Maxtor and Delphi Automotive. He has a BEng in Manufacturing & Manuafacturing Management from Middlesex University (UK). His email address is <lieusk@charteredsemi.com>.

LEON F. McGINNIS is Eugene C. Gwaltney Professor of Manufacturing Systems at Georgia Tech, where he also serves as Associate Director of the Manufacturing Research Center and founding Director of the Keck Virtual Factory Lab. His research focuses on the application of operations research and computer science to solve decision problems arising in the design and operation of industrial logistics systems. His email address is <leon.mcginnis@isye.gatech.edu>.

JOHN W. FOWLER is a Professor in the Industrial Engineering Department at Arizona State University. Prior to his current position, he was a Senior Member of Technical Staff in the Modeling, CAD,, and Statistical Methods Division of SEMATECH. His research interests include modeling, analysis, and control of semiconductor manufacturing systems. Dr. Fowler is the co-director of the Modeling and Analysis of Semiconductor Manufacturing Laboratory at ASU. The lab has had research contracts with NSF, SRC, International SEMATECH, Infineon Technologies, Intel, Motorola, ST Microelectronics, and Tefen Ltd. His email address is <john.fowler@asu.edu>