Proceedings of the 2006 Winter Simulation Conference
L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, eds.

A FULL FACTORY TRANSIENT SIMULATION MODEL FOR THE ANALYSIS OF EXPECTED PERFORMANCE IN A TRANSITION PERIOD

Moti Klein
Adar Kalir

Intel Corporation
Qiriat-Gat, ISRAEL

ABSTRACT

Intel's Fab-18 is based in Israel, and has transitioned from producing 0.18-micron logic devices to producing 90nM flash products. During this transition period, the factory has de-ramped in volume of logic while ramping-up flash. AutoSched AP software was utilized for the development of a transient simulation model of the Fab's behavior during this period. It is the first attempt, at Intel, to utilize a full factory simulation in order to analyze and support decisions that pertain to a transient period of parallel de-ramp and ramp-up of technologies. Unlike typical simulation models for the analysis of factory performance and behavior in steady-state, the transient model poses several modeling challenges and requires major adjustments in dealing with these challenges. In this paper, we discuss those aspects. The benefits and contribution of such a model to decision making and the improvement of factory performance are also presented.

1 BACKGROUND

A semiconductor manufacturing process is a complex manufacturing process. It typically consists of hundreds of production stages (or process steps), performed by dozens of different tool types, on a highly reentrant process flow and with various technology and operational restrictions such as queue time and layers/tools restrictions, over a high mixture of products processed simultaneously.

Given the stochastic nature of the process, resulting from the variability in processing times and, primarily, the relatively significant portions of downtime of the tools performing the process steps, it should come as no surprise that simulation models have been used over the years to investigate the environments of semiconductor manufacturing.

However, these models have been extensively utilized to evaluate the effects and behavior of various operating environments during steady-state. Specific examples of steady-state simulations used in semiconductor manufacturing can be found in Allen et al. (1999) or DeJong and Fischbein (2000). Kalir and Avidan (2001) developed a simulation for the ramp of a new Fab in order to enhance the static capacity analysis. Their model demonstrated a method by which a full factory simulation was used to identify tools that might be regular limiters as a result of WIP flow, even if the static analysis showed there is sufficient capacity and did not elevate these tools as potential constraints throughout the ramp-up period of a new Fab.

In this paper, we describe an extension of the above efforts, by demonstrating the utilization of a full factory simulation in order to analyze and support decisions that pertain to a transient period of parallel de-ramp and ramp-up of technologies in an existing Fab. Unlike the above references to typical simulation models for the analysis of factory performance and behavior in steady-state, or even the ramp of a new Fab, the transient model for parallel de-ramp and ramp-up's of technologies in an existing Fab, poses several modeling challenges, which are discussed in this paper.

The rest of the paper uses the following abbreviations: WSA (Wafers Starts Achievable), WIP (Work In Process), and WSPW (Wafer Starts Per Week).

2 PROBLEM STATEMENT

During year 2005, Intel's Fab-18 was facing a major challenge of decreasing the volume of logic technology while ramping up volumes of chipset technology. The transition of moving from one technology to another occurred with very tight capacity due to high demand on the logic process that created some unexpected upsides in the required chipset capacity. Towards the end of 2005, Fab-18 was faced with an additional challenge of starting-up a new flash technology. The need for aggressive start-up of the new technology has been driven by customer demands. The combination of increases in the required chipset capacity and the need for a fast ramp-up of flash, under the same clean-room space restrictions, resulted in a unique situation that the new flash technology had to be ramped with many

1-4244-0501-7/06/$20.00 ©2006 IEEE

one-of-a-kind tool-sets – and this has never been done before. The risk with one-of-a-kind toolsets is that whenever each such tool goes down for a sufficiently long period, it is essentially as if the full Fab is down.

Consequently, the effects of these one-of-a-kind tools on Fab performance (output, cycle time, etc.) can be detrimental and irreversible. Thus, the full factory simulation has been developed in order to predict the Fab performance during the de-ramp of chipset technologies and the early ramp-up of flash technology, to identify: (1) opportunities for mitigating negative impact on performance owing to one-of-a-kind tools, and (2) opportunities for improvements in Fab output and cycle time.

Next, the characteristics of the model are presented.

3 THE MODEL

Commercial dedicated simulation software, named AutoSched AP, was used for modeling the Fab performance. Extensive data collection is required to populate the full factory simulation model, such as product types and their volumes, planned number of tools of each tool-type, preventive maintenance and unscheduled downtime durations and frequencies, process times and setup times, batch sizes and cascade lengths, etc. In general, the data serves for inputs to the model in five main categories as follows:

1. **Production volumes** (in wafer starts per week) – reflecting the increased loading by product type and the changing product mixtures over time.
2. **Tool inventory** – reflecting installations/demolitions of new/existing tools from each tool-type, with the changes from week to week.
   (a) Tool capacity parameters –
   (b) Availability distributions
   (c) Run-rates
   (d) Batching/Cascading rules
3. **Headcount** – reflecting the number of technicians and their certifications, thus modeling the impact of technicians on operation and maintenance support activities.
4. **Process Flow** – reflecting the order of steps that produced on the wafer and the required tool-type at each step.

The inputs to the model are depicted in Figure 1. As can be observed, two inputs are required to be modified dynamically in the model. These are the Wafer Starts by product and the changing tool inventory.

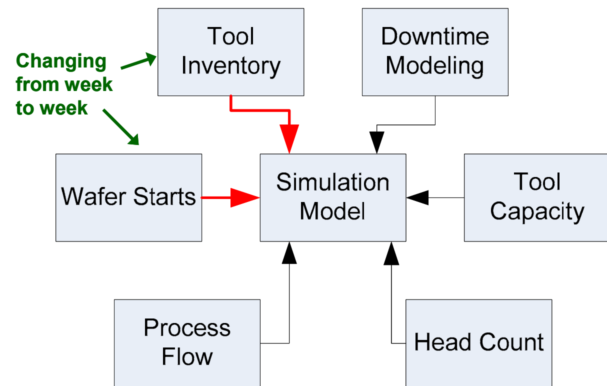

Figure 1: Main Input Categories to Simulation Model

During the construction of the model, the challenge of modeling actual availability distributions of tool-sets had to be overcome. The theoretical distributions that were utilized previously (i.e., uniform for process times and preventive maintenance tasks and exponential for unscheduled downtime) did not adequately match actual availability distributions and this caused the model to reflect faster Fab cycle time than in reality. The solution for that was to carry-out a detailed process to model all downtime activities across tool-sets more accurately. Through this process, extensive data collection, of each down time activity on each tool-set, was performed, collecting the following parameters: duration, frequency, scheduled or unscheduled, and headcount requirements for the activity (certification and quantity). By the end of the process, a sufficiently close matching of availability distributions has been accomplished.

4 METHODOLOGY

The following steps were executed in the process of developing the full factory transient simulation model:

1. **Technical modeling adjustments** to an existing steady-state full factory simulation. Technical adjustments to the *AutoSched AP* simulation model that enable changing Fab parameters over time.
2. **Data collection**, specific to the transient period that was modeled. Tools inventory WSPW & product mix that are changing from week to week. All those parameters were updated at the simulation model.
3. **Model execution** - All scenarios were run for period of two years. First year as warm up period in order to reach actual Fab status (WIP at each step, Fab WIP Turn etc.) and the second year to model the Fab status during the examined period.

4. **Verification** - through process of matching key parameters (as tools availability, utilization, Fab status etc.) and error fixing.
5. **Replications of runs** – Running the same model with different random seeds, in order to get statistically valid results. Four independent replication were run for each scenario. In this manner, the inherent variability of the performance metrics, caused for example by random machine breakdowns, can be separated from the induced variability of the metrics caused by the changes of tool inventory and WSPW (Dummler 2000). Due to the long duration of the simulation runs and the tight timelines of the project, we defined the statistical level of significance target as 80%. To satisfy this requirement, four replications per scenario were conducted.
6. **Analysis of results** – All averages were taken over a period of one week. The average values of a particular week were then averaged over the replications. The main performance characteristics under investigation are the average cycle time of lots, the average WIP and the average WIPTurn. Later this stage we identified opportunities for risk mitigation and improvements.
7. **Sensitivity analyses of what-if's**, testing the proposed opportunities (for risk mitigation and improvement) through more simulation runs in order to verify that the proposed changes will produced the expected gain.
8. **Conclusions and recommendations** for actions required for risk mitigation and improvements.
9. **Validation**: testing over time of the simulation accuracy against actual performance (more details at the next section).

5 VALIDATION

The results from the full Fab transient simulation model were compared against actual Fab performance in order to check the validity of the model.

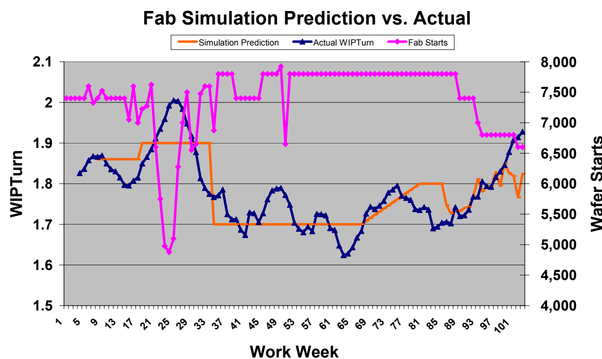

Figure 2: Validation of Simulated WIPT vs. Actual WIPT

As depicted in Figure 2, over a period of two years of changeable Fab starts (de-ramp and ramp at the amount of WSPW with periods of under-loading and full loading) the simulation model predicted the Fab performance with a high degree of precision. These results over time demonstrate that transient simulation model is valid and can accurately predict Fab performance.

6 SIMULATION RESULTS

The simulation results have shown that, during the modeled period, the Fab would not be able to meet its committed output and cycle time.

As depicted in Figure 3, the Fab WIP-Turn (a measure of velocity, approximately inversely proportional to cycle time) was expected to be below target (marked) a significant portion of the time.

In investigating the reasons for the low WIP-Turn, the following conclusions have been reached:

1. The one-of-a-kind tools **did not** accumulate cycle time, on average, or caused any major impact to Fab performance through direct downtime events, unlike the common notion before the simulation results.
2. The one-of-a-kind tools **did** generate high variability which implicitly degraded cycle time in downstream tight capacity tool-sets.

Put simply, the one-of-a-kind tools were tools with high burst capacity and, when they came back up from a down event, they were able to release any accumulated WIP very fast (due to their high capacity). The result was a generation of WIP flow variability in the form of WIP bubbles that, when hitting tight capacity tools, resulted in higher cycle times and lower WIP-Turn.

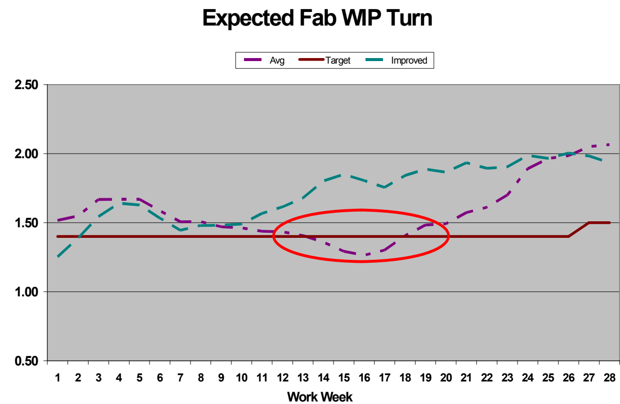

Figure 3: WIP-Turn Over Time: Expected versus Target

As depicted in Figure 4, a one-of-a-kind implant tool suffered from a long downtime of over a week that caused

WIP to accumulate. The following week (ww8), the implant tool recovered and released most of its WIP. As a result, the downstream tight capacity tools (an etch tool and another implant station of two tools) accumulated the WIP– and their cycle time increased during ww9. Next, the litho tool accumulated all the WIP owing to its tight capacity performance. The fact that there were two litho tools helped shorten the recovery period from this WIP bubble.

This was the machine sequence description per process flow that effected the case that described above:

Implant (1) → Implant (2) → Etch → Litho

This example demonstrated how the one-of-a-kind tools (in this case, implant tool) did not accumulate cycle time but generated degraded cycle time in downstream tight capacity tool-sets (in this case, litho tool).

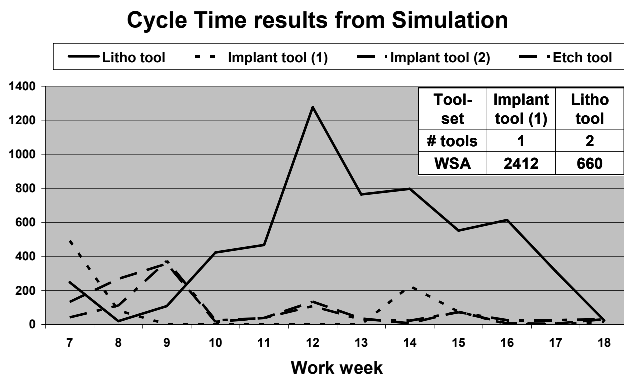

Figure 4: Formation of WIP Bubbles As a Result of One-of-a-Kind Tools

7 RECOMMENDATIONS AND ACTION PLANS

Based on the simulation results, projects to increase the capacity of the tight capacity tools were prioritized. Several options that were used to improve the capacity of those tools:

1. Pull-in's of the planned time-to-up of production tools. In most cases this option was feasible because it was a new option that was not used before.
2. Improvement in tool performance and productivity via run-rate and availability improvement projects.
3. Operational focus by preparing the area to the expected tight capacity period and taking action, such as adjusting the headcount 'on the floor' and defining 'Defcon' criteria for escalation when the area WIP increases.

Given the expected pull-in's and prioritized projects, the simulation was ran again, reflecting the improved capacity of the tight capacity tools. Results have shown, as indicated in Figure 3, that, indeed, these actions were the right actions and have resolved the WIP-Turn dip during the early ramp.

Post factum during the Fab ramp-up, the tools with the tight capacity did not accumulate cycle time, which proved that the actions taken helped prevent the case of dips in the WIP-Turn performance during the de-ramp and ramp-up period.

8 SUMMARY AND CONCLUSIONS

In this paper, a transient simulation model has been presented and discussed. The model assisted in providing useful insight that translated into effective actions for a successful start-up of a new flash technology in Fab-18, in a complex and restricted environment.

REFERENCES

Allen, D.E., A. Kalir, and K.G. Kempf, 1999, Refinement of Semiconductor Fab Capacity Planning through Dynamic Simulation, *Proceedings of the Winter Simulation Conference 1999, Dec 5-8,* Phoenix, Arizona, USA.

DeJong, C. D., and S. Fischbein, 2000, Integrating Dynamic Fab Capacity and Automation Models for 300mm Semiconductor Manufacturing, *Proceedings of the Winter Simulation Conference 2000*, Dec 10-13, pp. 1505-1509, Orlando, FL, USA.

Kalir, A., and O. Avidan, 2001, On the Uses of Simulation in the Planning and Ramp Phases of a High Volume Manufacturing Fab, *Semiconductor Manufacturing Operational Modeling and Simulation (SMOMS) Conf.*, pp. 37-41, Seattle, WA, USA.

Dummler, M. 2000. Analysis of the Instationary Behavior of a Wafer Fab During Product Mix Changes. *Proceedings of the Winter Simulation Conference, 1436-1442.*

AUTHOR BIOGRAPHIES

MOTI KLEIN received his B.Sc. in Industrial Engineering and Management from Ben-Gurion University, Israel. He is currently a Simulation and Modeling Manufacturing Eng. At Intel Corporation in Qiriat-Gat, Israel.

ADAR KALIR received his Ph.D. in Industrial and Systems Engineering from Virginia Tech, USA. He is currently Intel Israel's 300mm 45nM Manufacturing Technical Leader.