Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.

AN INTEGRATION OF STATIC AND DYNAMIC CAPACITY PLANNING FOR A RAMPING FAB

Georg Seidel

Infineon Technologies Austria AG
Siemensstraße 2
Villach 9500, AUSTRIA

Patrick Preuss

D-SIMLAB Technologies GmbH
Wiener Platz 6
Dresden, 01069, GERMANY

Cevahir Canbolat

Infineon Technologies Regensburg GmbH
Wernerwerkstraße 2
Regensburg, 93049, GERMANY

Soo Leen Low
Chew Wye Chan
Boon Ping Gan

D-SIMLAB Technologies Pte Ltd
8 Jurong Town Hall Road #23-05 JTC Summit
Singapore, 609434, SINGAPORE

Ching Foong Lee
Prakash Manogaran
Aik Ying Tang

Infineon Technologies (Kulim) Sdn Bhd
Jalan Hi-Tech 7
Industrial Zone Phase II, Hi-Tech Park
Kulim, Kedah 09000, MALAYSIA

ABSTRACT

In the semiconductor industry, production planning is often complicated due to constantly changing product mixes, reentrant process flows, and high variations of capacity uptime. In this paper we discuss the combination of static capacity and dynamic simulation approaches for production planning, highlighting how these approaches complement each other in our daily business process. The typical static capacity planning is based on a fixed product lead time, allocating the production volume of each product to available capacity with an objective of capacity minimization and a constraint of utilization limit (plan load limit) to absorb production variability. Whereas the dynamic simulation models the process of lots flowing through the production line, and consume the capacity at each process steps, with additional consideration of fab WIP at the beginning of simulation. With simulation we can additionally provide forecasts for important production key figures, for example product cycle times and fab flow factor.

1 INTRODUCTION

Semiconductor wafer fabrication is one of the most complex manufacturing processes today. A diverse product mix that changes over time (Hoop et al. 2002; Robinson 2003; Geng and Jiang 2009; Pappert et al. 2017), reentering process flows (Zhang et al. 2006), different machinery uptimes and shifting bottlenecks (Nyhuis and Filho 2002) are only some typical challenges that have to be considered in the planning process.

Production planning experts try to optimize fab loading, often based on personal knowledge and historical performance. They must provide a loading plan for production that incorporates e.g. machine performance and uptime, availability of raw material and operator resources, and customer orders (Nyhuis and Filho 2002). Product fine lead times are used to determine when a product has to be started in the fab to fulfill customer demands on time. Lead times are normally derived through historical data analysis. Underestimation or overestimation of lead times could cause problems. Overestimation will most likely reduce the customer satisfaction because delivery dates cannot be met. Underestimation will increase the WIP level and can lead to additional congestion in the line (Geng and Jiang 2002; Nyhuis and Filho 2002; Wang et al. 2016; Biwer et al. 2018).

Infineon's frontend factory in Kulim, Malaysia, is undergoing rapid ramping for the last months. In a ramping fab it is very tough to estimate the lead times with high accuracy out of historical data, because too many influencing parameters are changing constantly. We therefore provide a long-term simulation model to forecast fab key figures like product cycle time, fab flow factor, fab out, WIP, work center and machine utilization. Our long-term simulation model can be used to forecast the next weeks and even months using strategic loading plans, actual WIP in the line, product routes, expected machine down times, expected process times, future change lists (list of new and obsolete equipment) and dispatching rules. By using the long-term simulation model we can complement the static capacity planning process, with additional forecasts and an increased insight in our production which improves our planning process.

In this paper, we provide an overview of the static capacity planning and dynamic capacity planning approaches in Section 2 and Section 3 respectively. In Section 4, we describe some details about the discrete event simulation model we used. Section 5 is about the interaction between static planning and simulation results. Lastly, we provide an outlook of how the planning business process could be improved even further in the future with a combination of these approaches.

2 STATIC CAPACITY PLANNING

The input to the static capacity planning are plan uptime, raw tool time, plan load limit, recipe dedication (see Table 1 for definition), process flows, and fab loading. These values are defined for a specific tool, work center or tool/recipe combination. Work center is a collection of tools that are capable of processing similar recipes.

Table 1: Definitions of static capacity planning input parameters.

Parameter	Description
Raw Tool Time (RTT)	Raw tool time is the planned tool time for each product/recipe running on the tool. It does not include downtime, quality sampling, production test, or rework, which would be considered process inefficiencies
Plan Uptime (PUT)	Uptime is the percent of time the equipment is in a condition to perform its intended function during the period of operations time (SEMI E10-0304, operational uptime).
Plan Load Limit (PLL)	The load limit is the time an equipment should be idle to avoid overloading and congestion due to high queuing times (see Figure 1).
Recipe Dedication	The list of tools that are capable of processing a recipe, where each recipe is associated with each process step (operation).
Process Flows	The process steps (operations) where each product requires to run through.
Fab Loading	The number of wafers of product to be produced for each time period.

Based on the provided fab loading, a weekly going rate for each product on any operation of the process flow can be calculated. Each operation is tied to a tool, or group of tools. This yields us to a mathematical optimization problem of minimizing the maximum utilization for all tools in the fab, and derives a lower bound for the expected utilization of the tools or tool groups. This will be compared with the expected uptime and plan load limit. If the expected utilization exceeds the expected uptime minus the PLL (time

available for production), a bottleneck is detected. With this method it is possible to forecast which tools or tool groups are potential bottlenecks for a given loading scenario.

If the input parameters of RTT and PUT are reliable, this is a fast method to evaluate different loading scenarios. However, it is not possible to predict product cycle times or fab flow factor. Experience is needed to know how high work center can be utilized to keep cycle times and flow factor stable. In a ramping fab experience can be misleading due to changing basic conditions.

The setting of different PLL for equipment is also based on experience, by using factory physics knowledge and queuing theory (Hopp et al. Factory physics). Different values for the PLL can lead to overestimation or underestimation of work center capacity. This can result in reduced or increased fab loading if the planner reacts accordingly. Again the impact on cycle time and flow factor cannot be predicted.

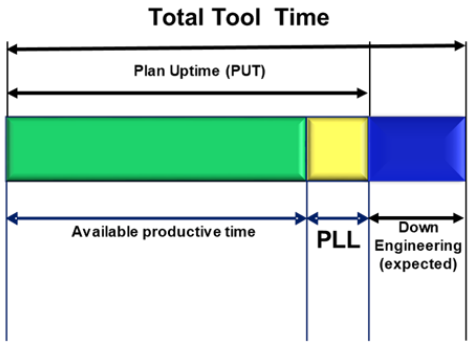

Figure 1: Plan load limit (PLL) and plan uptime (PUT).

Static capacity planning can work well if you have a stable factory and experienced planner. Additional capacity will be installed if a tool overload is predicted with a committed loading forecast. The definition of tool overload is critical in this process. If the predicted productive time exceeds the total tool time the decision is easy. New capacity must be installed in this case. If the predicted productive time is less than total tool time minus the expected down and engineering time (within the PLL) there is room for discussion. Maybe capacity increase can be delayed and investment can be avoided. However cycle time for products using this tool will increase in this case. This also means that the affected product has to be re-validated as the lead time would have changed in this case.

In a stable fab the necessity for capacity increase (or decrease) will happen from time to time due to product mix changes and slight loading changes. Validation of required lead time changes and product cycle time changes are not easily done but are normally manageable and within certain ranges.

In a ramping fab this is trickier. A lot of work center will see capacity increase at the same time. Analyzing historical data will not help that much because basic conditions changed. It will be impossible to predict product cycle times and therefore also necessary changes of product lead times. Without simulation the approach for changing lead times can only be done retrospectively. Only after a cycle time increase for products can be observed in reality, lead times will be changed accordingly, which means that could already introduce problems to the production line.

3 DYNAMIC CAPACITY PLANNING

As opposed to the static capacity planning, the dynamic capacity planning approach (with the application of discrete event simulation technique) provides additional insight into the implication of ramp plan towards fab KPIs. This approach does not only answer the question of whether the planned capacity is sufficient for the production ramp, but also provides visibility into the expected flow factor, the WIP profile development and the evolving tool utilization over time. This visibility is very important because the dynamic behavior of the production line could spell problems that cannot be observed with the static capacity approach.

Situation such as a short term capacity issue that caused by unfavorable random events distribution, such as ad-hoc tool down, hold lots, sampling, and rework, could trigger a WIP buildup scenario that is unrecoverable. In addition, batching, dispatching and changing delivery targets are factors that could create variability within production line. For example, an incorrectly configured batching rule might result in a low batching efficiency that consumes tool capacity higher than expected, which in turn causes capacity shortage. This situation cannot be observed with static capacity planning because batching efficiency is an assumed input, which usually is based upon past history. Table 2 provides an overview of the distinct differences between the static and dynamic models.

Table 2: Static vs Dynamic Capacity Model.

Model Aspect	**Description**	**Static**	**Dynamic**
Tool Uptime	Time available to process production lots	Uniformly distributed uptime over the time bucket	A result of random down events, which occurs following statistical distributions derived from historical data
Tool Efficiency Loss	Time loss due to non-optimal lots cascading, meaning lots are not immediately available to be loaded when the tool is ready to cascade the next lot	A fixed percentage of time loss	Cascading loss is a result of non-optimal lot arrival pattern
Tool Setup Time Loss	Time loss due to overhead of setup changes	A fixed percentage of setup time loss	Setup time loss is a result of lot arrival pattern
Planned Load Limit	The tool utilization limit, usually provided to absorb line variability	A fixed percentage that limits the time available for capacity allocation	Not required to provide as the dynamics of the line is portrayed
Capacity Allocation	The allocation of lot (work) to tools	Capacity consumption is allocated by assuming that all lots are available at a defined time bucket	Capacity consumption is a result of lot arrival at a specific point in time with applied dispatch rules to decide which lot is going to be allocated, which in turn dictates the lot arrival pattern at the subsequent tools
Process Flow	All steps each production lot needs to go through	Determines the number of passes for each recipe that is required by the product	Used to generate the lot flow through the production line, which in turn generates the lot arrival pattern
Recipe Dedication	Define which recipe can run on which tool or chamber	Used to calculate the total capacity available for each recipe	Use to decide which lot can be processed by which tool at time of dispatching

The dynamic capacity model of our wafer fab is built with a commercial discrete event simulation package (D-SIMLAB Technologies 2019). To ensure easy usage, the model is generated automatically from the fab execution and planning data sources (Seidel et al 2017). Whenever a capacity study is required,

a simulation model is first generated and verified. The objective of the verification is to ensure that stable WIP is observed with the model over the study time period. Thus, the verification is done by feeding the simulation model with a feasible wafer start plan (specific product mix) that maintains the fab WIP at a defined level. If a rising or falling fab WIP is observed, that could mean that there are some issues with the data input to the model. This would trigger an investigation and find solutions for the data issue, such that the expected model output is observed.

4 THE DISCRETE EVENT SIMULATION MODEL

As already mentioned in Section 3, our simulation model is generated automatically. On demand the model will be validated and can be used for dynamic capacity planning. The running time for capacity scenarios is normally fixed to 8 weeks or 6 months. This is due to the restrictions of quality data availability for the specified time horizon. The simulation user is encouraged to conduct confidence runs of at least 10 replications to achieve required statistical confidence of the results.

The main source of variability within the simulation runs is coming from the tool downs. MTTR (mean time to repair) and MTBF (mean time between failure) values and distributions are derived from historical data up-to 6 months. The same is true for sampling rates, yield, percentage of hold lots and split lots. At the initialization of the model tool status are considered. A tool will be down at simulation start when it is down in reality. MTTR value and distribution will be used to bring the tool up again. Lots in process will be considered too at initialization and will be released from tool considering the real process start and the planned tool time for the recipe (RTT).

5 EXPERIMENTAL RESULTS

The general planning approach uses both methods, static and dynamic capacity planning. First static capacity planning is used to determine a set of reasonable fab loading scenarios. Each of these scenarios is then analyzed further by using the dynamic approach.

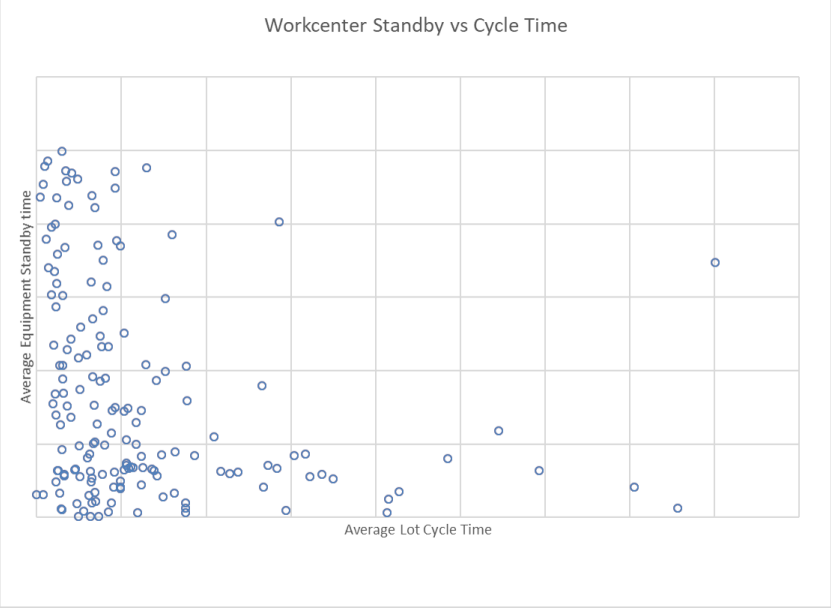

Figure 2: Work center average equipment standby time (idle time) vs lot cycle time.

By using fab simulation it is possible to forecast work center cycle time. Therefore it is possible to detect hot spots in the line that can cause cycle time problems even if the work center utilization is not close

to the plan load limit yet. Figure 2 shows the correlation between the expected average standby time and lot cycle time at the work center. With this information, planner can fine-tune the loading by reducing lot starts of products which will positively impact high cycle time work center and maybe even increase loading for work center with low cycle time and already high utilized. After fine-tuning, simulation can be rerun to determine the impact of the loading changes. Furthermore it is possible to determine potential cycle time risks by conducting confidence runs with different seeds. Max and min results of the confidence runs for cycle times, flow factor, and other important key figures are used to detect high results variability. A high product cycle time variability between different confidence runs is often a sign for potential bottlenecks provided the product loading volume is not very low.

6 THE CHALLENGES

Simulation runs and analysis are time consuming. Therefore it is not possible to conduct too many scenario runs. It requires certain knowledge to determine which scenarios should be analyzed in detail. Minor loading changes often leads to neglectable changes of fab key figures but could have a big impact on work center and product key figures.

It is also challenging to detect areas where small changes possess high risk. Small uptime changes in critical work center can change predictions drastically. There is a huge number of potential different scenarios if you change uptime distributions on work center level and loading scenarios on product level. Experience is needed to choose appropriate scenarios to highlight risk areas. Until now there is no guarantee that all critical areas can be detected.

A reliable future change list is required for a good prediction. The future change list must contain data about new incoming equipment, and recipe dedications for these new tools. Expected changes of RTT times, changes of dispatch rules, and expected fab shut down dates should be in too. Incorrect future change list may lead to wrong forecast results.

Another very important topic is the prediction of future loading. Impact of changes in the loading structure will impact the prediction quality. Changing fab loading structures, compared to the fab loading of the conducted scenario can change the fab key figures dramatically. However it is possible to use simulation to show afterwards the impact of different loadings or incorrect future change lists by rerunning simulation and comparing the results. Therefore insight is gained how the production reacts to certain loading changes.

7 CONCLUSIONS

Dynamic capacity planning complements static capacity planning and should be used in combination. It helps to identify hot spots in the line and supports the planner to fine-tune the fab loading. Information about expected cycle time and flow factor can be derived additionally. Impact on fab key figures induced by changes in the fab loading structure can be validated upfront. It's also possible to determine how much certain fab key figures have been influenced in hindsight by rerunning simulation. Static capacity planning will be not obsolete but is used to provide a preselection of suitable scenarios.

REFERENCES

Biwer. S., M. Filipek, E. Arikan, and W. Jammernegg. 2018. "Capacity planning challenges in a global production network with an example from the semiconductor industry". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A.A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 3639 – 3650. Piscataway, New Jersey: Institute of Electtrical and Electronics Engineers, Inc.

D-SIMLAB Technologies. 2019. Forecaster and Scenario Manager. http://www.d-simlab.com/category/d-simcon/products-d-simcon/forecaster-and-scenario-manager, accessed 01st July 2019.

Geng. N. and Z. Jiang. 2009. "A review on strategic capacity planning for the semiconductor manufacturing industry". *International Journal of Production Research* 47(13): 3639-3655.

Hopp, W. J. and M. L. Spearman. 2001. *Factory physics: foundations of manufacturing management*. Boston: Irwin/McGraw-Hill.

Hopp, W.J., M.L. Spearman, S. Chayet, K.L. Donohue, and E.S. Gel. 2002. "Using an optimized queueing network model to support wafer fab design". *IIE Transactions (Institute of Industrial Engineers)* 34: 119-130.

Nyhuis. F. and N.A.P. FilhoP. 2002. "Methods and tools for dynamic capacity planning and control". *Gestão & Produção* 9(3):245-260.

Robinson. J., J. Fowler, and E. Neacy. 2003. Capacity Loss Factors in Semiconductor Manufacturing. http://www.fabtime.com/files/CapPlan.pdf, accessed 01st July 2019.

Seidel, G., B.P. Gan, C.W. Chan, C.F. Lee, A.M. Kam, A. Naumann, and P. Preuss. 2017. "Harmonizing Operations Management of Key Stakeholders in Wafer Fab Using Discrete Event Simulation". In *Proceedings of the 2017 Winter Simulation Conference*, edit by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

SEMI E10-0814E. 2014 . *SEMI E10 – Specification for definition and measurement of equipment reliability, availability, and maintainability (RAM)*. Semiconductor Equipment and Material International (SEMI).

Tanjoing, S. 2011. "Bottleneck Management Strategies in Semiconductor Wafer Fabrication Facilities", In *Proceedings of the International Conference on Industrial Engineering and Operations Management*, Jan 22nd – 24th, Singapore, 3-8.

Wang, L.C., A. Wang, and C.Y. Chueh. 2018. "Development of a capacity analysis and planning simulation model for semiconductor fabrication". *The International Journal of Advanced Manufacturing Technology* 99(1-4): 37-52.

Zhang, M.T., J. Fu, and E. Zhu. 2008. "Dynamic capacity modeling in semiconductor assembly manufacturing". *International Journal of Production Research* 46(3): 739-752.

AUTHOR BIOGRAPHIES

GEORG SEIDEL is Senior Staff Engineer of Infineon Technologies Austria AG (Villach, Austria). He has been involved in simulation, WIP flow management and Industrial Engineering topics since 2000. He was responsible for WIP flow management, especially for Lot dispatching at Infineon's site in Kulim (Malaysia) from 2012 until 2015. He is now responsible to rollout Fab Simulation in Kulim and Villach. He holds a Master degree of Technical Mathematics. His email address is georg.seidel@infineon.com.

PATRICK PREUSS is a Project Manager and the Deputy Manager Germany Operations of D-SIMLAB Technologies (Germany). He has been working in the development of simulation-based applications for Airbus, German Aerospace Centre and Infineon with focus on data analysis and heuristic optimization methods since 2005. Patrick holds a M.S. degree in computer science from Dresden University of Technology. His email address is patrick@d-simlab.com.

SOO LEEN LOW is a Project Manager at D-SIMLAB Technologies (Singapore). She is responsible for simulation modelling and analysis of Wafer Fabrication plants. She earned a Bachelor of Engineering in Computer Engineering from National University of Singapore (NUS) in 2014. Her email address is soo.leen@d-simlab.com.

CHEW WYE CHAN is a Senior Software Engineer of D-SIMLAB Technologies (Singapore). He has been working in the development of simulation-based applications for logistics and semiconductor industry with focus on simulation-based optimization since 2006. Chew Wye holds a Master of Computing degree from National University of Singapore. His research interest is in the area of the application of Artificial Intelligence techniques for production operation execution. His email address is chew.wye@d-simlab.com.

BOON PING GAN is the CTO of D-SIMLAB Technologies (Singapore). He has been involved in simulation technology application and development since 1995, with primary focus on developing parallel and distributed simulation technology for complex systems such as semiconductor manufacturing and aviation spare inventory management. He was also responsible for several operations improvement projects with wafer fabrication clients which concluded with multi-million dollar savings. He holds a Master of Applied Science degree, specializing in Computer Engineering. His email address is boonping@d-simlab.com.

CHING FOONG LEE is Senior Specialist Engineer of Infineon Technologies (Kulim) Sdn. Bhd. She has been involved in Semiconductor System Development and Datamining since 2004. She joined Infineon Technologies Kulim in 2010 driving various projects in Production System Setup, Reporting and System Improvement under Factory Integration department. Currently she is responsible in Kulim for Simulation and WIP flow management topics under Operation Research and Engineering department. She holds Master of Business Administrative(MBA) and Bachelor Degree of Information Technology, majoring in Software Engineering. Her email address is chingfoong.lee@infineon.com.

PRAKASH MANOGARAN is the Capacity Planner of Infineon Technologies (Kulim). He has been involved functions of engineering, manufacturing, training & capacity planning since 1992 in backend and front-end semiconductor. He was also responsible for several operations improvement projects with wafer fabrication engineering solution in multi-million dollar savings. He also has been in the two green field manufacturing startup. He holds a Master in Mechanical Engineering, specializing in

Manufacturing Engineering. His email address is Prakash.Manogaran@infineon.com

AIK YING TANG is an engineer of Infineon Technologies (Kulim) Sdn. Bhd. She is currently involved in simulation topics under WIP Flow Management department. She holds a Doctor of Philosophy Degree specializing in Mathematics. Her email address is aikying.tang@infineon.com.