Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

CAPACITY PLANNING ACCURACY AND THE EFFECT OF DYANMIC DEDICATION CHANGES FOR A SINGLE WAFER LOT SEMICONDUCTOR FACTORY

Richard Surman[1], Matt Nehl[2], Cole Evanson[2], Soo Leen Low[3], Kern Chern Chan[3], Hui Sian Liu[3], and Boon Ping Gan[3]

[1]Seagate Technology, Londonderry, UNITED KINGDOM
[2]Seagate Technology, Bloomington, MN, USA
[3]D-SIMLAB Technologies Pte Ltd, Singapore, SINGAPORE

ABSTRACT

In the context of a single wafer lot semiconductor factory characterized by high levels of Research and Development (RD) work-in-progress (WIP), low levels of product-based lots and lengthy cycle times, this paper investigates the accuracy of capacity planning given the impact of dynamic changes in dedication. We delve into several critical aspects related to dedication planning, drawing insights from historical data and dispatch logic used. The experimental results show the improvement in model accuracy with the incorporation of dedication changes as distribution functions.

1 INTRODUCTION

Semiconductor manufacturing facilities play a pivotal role in the production of electronic devices. The Seagate facility (NRM), exemplifies this significance by focusing on the fabrication of hard disk storage recording head components. The Seagate facility comprises multiple clean rooms that house the intricate machinery responsible for wafer processing. Notably, equipment is strategically grouped by type, ensuring streamlined workflows. However, the challenge lies in maintaining efficient colocation practices within limited space constraints.

Effective WIP management is essential for optimizing production. At the Seagate facility, single wafer lots are evenly distributed between research and development (RD) and production wafers. Four distinct priority levels govern product lots, with one specific product accounting for 60-70% of the overall production volume. The dynamic nature of RD lot shipments, which vary weekly, necessitates agile management strategies.

The typical production process comprises a substantial number of steps, resulting in a significant line yield with only minimal rework. The planned cycle time for the product extends across several months, leveraging a diverse array of equipment types or configurations. As the facility readies itself for the upcoming product iteration, simulation-based route planning will extend the process to a considerably larger number of steps.

Complex dispatch rules are dictated by lot priorities. The system employs auto priority adjustments to balance the number of lots in the queue. KANBAN systems (Ohno 1988), among other rule-based logic, facilitating efficient dispatching. However, challenges persist in managing hold processes, such as future staging, unavailable recipes, and process capability limitations.

The Seagate facility operates various single and dual-path equipment. Rigorous monitoring ensures optimal performance. Key equipment metrics, including Overall Equipment Effectiveness (OEE), availability, Mean Time Between Failures (MTBF), Mean Time To Repair (MTTR), and Green2Green assessments (G2G = Assuming WIP is at the tool it is the duration from last wafer finished to next wafer started when a tool unavailable event occurs), guide maintenance decisions. Scheduled qualifications and ad hoc assessments further enhance equipment reliability.

979-8-3315-3420-2/24/$31.00 ©2024

Discrete event simulation (DES) has widely been used in the semiconductor industry for short-term and long-term planning. Some examples are Bagchi et. al. (2008), Biwer et. al. (2018), Scholl et. al. (2011), Seidel et. al. (2017). In Seagate a DES model was built with a commercial off-the-shelf simulation engine, the *D-SIMCON DCMF Planner* (D-SIMLAB 2024). Accurate simulation models allow NRM to assess production capacity and improve velocity based on running different scenarios without impacting real world operations. Using different key parameters inputs (denoted as KPIVs hereafter) for the NRM model validation versus the KPIVs for scenarios helps NRM to determine the accuracy of the simulation model and provides confidence when simulating scenarios.

With the high level of complexity and variability at NRM, capacity planning proves to be a significant challenge. Currently the Industrial Engineering team (IE) is responsible for the factory capacity planning using a static capacity model (SCM) which relies on historical data of average equipment available time to calculate an equipment utilization and an average achievable factory output. One of the main limitations of this approach is the assumption that all equipment of equivalent hardware are all potential paths for a given product-stage-step (PSS) if the equipment is up. However, in reality there is a frequent adjustment of equipment dedication for a PSS even within the same equipment group.

There are several different ways that Process Engineering (PE) manages equipment dedication for a PSS to meet product quality and performance objectives and reduce the scrap risk: IPP (intended process path), recipe level, product level, and lot level.

At the highest-level, PE can enable additional equipment by PSS that share the same set of process modules (chambers) known as an IPP. This allows PE to group similar processes across product lines and then enable or disable larger groupings of like processes on a equipment if there is an issue with one of the process modules that is part of an IPP definition. A more granular level of enablement would be a PSS that has a similar recipe on a given set of equipment which would be enabled on any equipment capable of performing those set of instructions whether it is depositing a certain thickness of metal or polishing for a certain duration. With the high level of RD in the factory a more granular level of dedication is often required even within the same stage-step when new products are developed. Product level dedication allows for PE and RD to enable certain products within a stage step to specific equipment within a equipment group to have more control or experiment with different recipe parameters. Finally, the most granular level, PE can control equipment dedication at the lot level which would only allow for a single lot of a PSS to go through a particular equipment at once. This reduces the risk when equipment comes back from a maintenance event or if there has been a process shift that could put more wafers at risk.

In the short term, all changes in daily enablement influence variability and may degrade capacity with fewer available paths to process the same PSS beyond what is visible in the SCM capacity plan. The impact seen in the factory is an increase in cycle time that is difficult to model due to the high level of randomness and event-driven process control.

On a longer time horizon, there are additional planning challenges with regard to how dedication is controlled for new products. Often as a new product is being transferred from RD to PE to begin mass production it has fewer paths compared to existing volume products as PE needs to qualify additional paths and validate that other equipment in an equipment group are sufficiently matched to the existing paths. As the product matures there is often an increase in paths, but the time duration can be over months or even a year to reach a full level of enablement that would be expected by the SCM plan. This poses a challenge when using historical data to predict future performance as there is the expectation that additional paths would become available over time but with product lifecycles shortening and cycle time growing the impact of trying to ramp and build a product in parallel drives a wider gap between current and future dedication for a given PSS or a brand-new PSS with no history.

In this paper, an overview of the wafer fab simulation model is provided in Section 2. In Section 3, the dedication approaches that were adopted to address specific modelling challenges in Seagate to improve the accuracy of the simulation model are described. This is then followed by experimental results where a standard dedication approach is compared with a dedication that changes based on statistical distribution.

The paper is concluded with an outline of future work to further enhance the model quality, especially with regard to the uniqueness of Seagate's manufacturing line.

2 WAFER FAB SIMULATION MODEL

All the essential semiconductor wafer fab modelling elements are required to be captured in the simulation model to achieve high accuracy. Based on our previous work, fab level KPIs such as wafer out, work-in-progress, wafer moves, and cycle time could reach 95% accuracy (Seidel et. al. 2017; Mosinski et. al. 2017). Below is the summary of Seagate simulation modelling elements.

Table 1: Modeling elements overview.

Modelling Element	**Description**
Work-In-Progress (WIP)	Initialize the simulation model with the current snapshot of WIP lots in the production line, with information such as current step in queue, on-hold, in buffer zone and in process, in rework, priority and fab start date.
Initial Equipment Down	Initialize the simulation model with the current snapshot of all equipment are in down or non-productive state with an estimate of when the equipment is expected to get back online by using historical data (average down duration of the corresponding down type) or provided by the maintenance department.
Process Flows	All process flows required by the WIP and wafer start lots are considered in the model. Rework is modelled as a random event, where rework rates are derived from historical data for all process steps that could trigger a rework process. Hold is modelled as a random event, where hold rate and hold duration distributions are derived from historical data for all process steps that could trigger lot hold.
Equipment	All equipment that are currently active in the wafer fab are considered. Each equipment is modelled based on its specific behavior such as single wafer or batch processing.
Dedication	Dedication is modeled on recipe, stage, and product combinations. Current active dedication is obtained at the snapshot time together with the inactive dedication. Current active and inactive dedication could be disabled and enabled correspondingly, which is obtained from the historical data.
Chamber Process Path	For each equipment, the specific chamber process path is captured for a recipe, stage, and product. The sequence of chambers required for each chamber process path is modeled in the simulation model. This is essential due to different chemical or target can be used by each chamber.
Process Time and Throughput	Equipment, Recipe, Stage and Product level of process time and throughput is gathered to model the lot processing time at the equipment and the speed of the equipment correspondingly. Limping effect (losing process speed) of chamber down is also modelled.
Equipment Down	The variability of the equipment downs is modeled as a random event. Mean time to repair (MTTR) and mean time to failure (MTTF) are determined using historical data.
Reticle	Reticles are modelled as a resource required before the lots can start processing at the lithography equipment. This is essential due to different product mix and multiple layers for each product could result in significant effect on the forecast accuracy.
Dispatch Rules	Specific Seagate dispatch rules are implemented, such as queue time constraints, batching rules, priority lots and reticle switch interval times.
Qual Monitor Lots	Qual monitor lots are required in some process equipment in the Seagate wafer fab to obtain a better yield. Thus, the availability of qual monitor lots and the consumption of the qual monitor lots at the equipment are captured in the simulation model.

3 DEDICATION MODELLING APPROACHES

This section discusses the dedication constraints in the wafer fab and the importance of modelling them accurately. In addition, some unique characteristics of Seagate production process relevant for dedication modelling, in particular which dedication disabling/enabling to be used as the control mechanism for the production line, is described. The discussion extends to how a discrete event simulator models the behavior correctly based on the bin-based distribution that is captured from the historical data.

3.1 Dedication As a Constraint

Dedication is the permission for certain processes to be run on certain equipment. In the context of Seagate, dedication is defined at product-stage-step (PSS) level. An equipment is said to be dedicated to certain PSS when the PSS can be processed by the equipment. The number of dedications of certain PSS is the number of equipment that can run the PSS. Dedication pair basically means the combination of PSS + Dedicated Equipment.

Dedication imposes a significant constraint to both the simulation (virtual fab) and the actual production line (real fab). With more flexible dedication where the PSS is dedicated to more equipment, the lots arriving to the step will have more available paths to run. This means that such lots do not have to wait for a specific set of equipment to be available before it can be processed. This will increase the velocity (average moves per day) of the lots that run on the step, with a shorter Cycle Time and Queue Time for the step.

On the contrary, if there is none or limited dedication for certain PSS, lots will get stuck and WIP will pile up at the step with limited available paths before the lots can get processed. The waiting time or queue time of the equipment group that is dedicated to run the PSS will also increase over time. Furthermore, this will affect all the downstream processes as there will be fewer arrivals to the downstream steps/ equipment groups.

Hence, proper modelling is required to capture the constraints while modeling the capacity correctly so that what is happening in the actual production line can be replicated in simulation.

3.2 Statistical Distribution for Dedication Enabling/Disabling

The following part discusses how statistical distribution of dedication disabling and enabling is being used and calculated/generated from the historical data.

For the simulator to decide whether to enable or disable a particular dedication pair, it will read from the respective disabling/enabling bin-based distribution that defines the probability of the disabling/enabling the dedication pair for certain duration. If the bin-based distribution can capture the behavior of the production line, the long-term behavior of the dedication disabling/enabling will be like the actual production line.

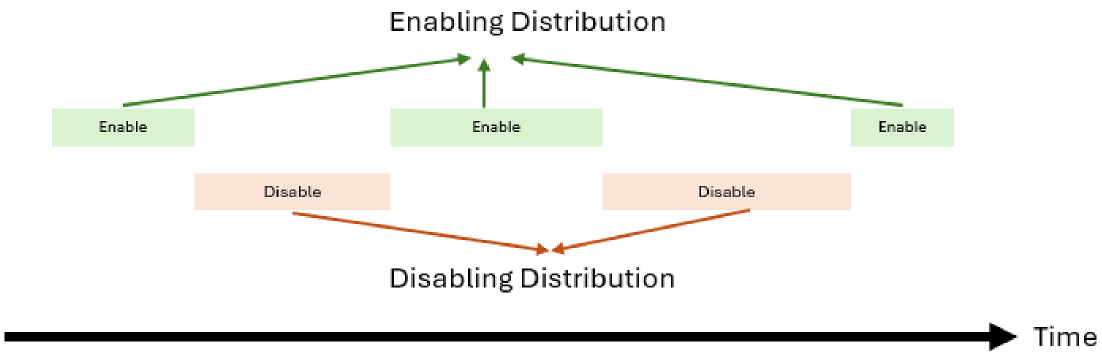

Figure 1: Generating dedication enabling/disabling statistical distribution from historical data.

To calculate the distribution, the valid disabling entries and period for the dedication pair will need to be identified from the dedication enabling/disabling history. The dedication enabling/disabling trace will cover the time horizon spanning at least one fab level cycle time. The data points will then be grouped into a bin-based distribution with pre-defined time intervals for the durations.

3.3 Weightage Considerations

Due to the nature of an RD fab where process improvements are always happening, the disabling/enabling events that are closer to the current date are always more relevant and a better representation of what will happen in the future. This means that these data points are more important and will need to have a higher weightage when the bin-based distribution is being calculated. On top of that, there are always new products in the line. For such products, considerable qualification work is to be done, and many of the PSS that are used by the product will be having a longer disabling duration. As the volume ramps up, the processes used by the product will be more mature, and hence there will be fewer and shorter disabling events for these products. Similarly, for a product that is ramping down, there will be more disabling events as fewer equipment are required to run the PSS. To reduce the impact of the ramp up/ ramp down product – which will skew the distributions – the disabling/enabling events by the moves made within the same period are also weighted.

Another particular characteristic of Seagate dedication enabling/disabling events is that some of the PSS are enabled/disabled on certain equipment on a lot level basis. This means that the PSS can be disabled for all other lots, and only enabled for certain lots, and vice versa. This is mainly due to the process control concern when the production line wants to control the productive moves for certain PSS. To take this into consideration, a separate distribution which includes the consideration of the lot level enabling/disabling events is generated. The two distributions (lot level distribution and original distribution) will be combined by applying a weightage on each distribution, based on the number of lots being disabled/enabled on a lot level basis vs the number of lots being disabled/enabled not on a lot level basis.

3.4 Valid Data Points

Finally, the question is how to classify a dedication disable event as valid. For some new products, the PSS could be disabled for a very long time because there have not been any moves or sampling for the newly introduced PSS. As the volume of the product slowly increases, the disabling/enabling of the dedication of the PSS will converge to the behavior that will be replicating itself in the future. Hence, the initial long disabling events will be considered as invalid events, and if these data points were included into the distribution it would be skewed. Besides that, the PSS could be disabled on certain equipment when the equipment is down. However, since the down behavior is captured in the down modelling of the equipment, these events are considered as invalid and discarded in the calculation, to avoid over-constraining the production line. By disregarding the invalid datapoints, the bin-based distribution can be calculated. To be more robust, an outlier filtering for the long tail is also be applied based on the pre-defined cutoff percentile.

In short, a bin-based distribution that can capture the valid historical behavior of the dedication pair needs to be calculated so that the dedication constraints can be captured. Unique behavior patterns in the Seagate fab are captured in the distribution by applying filtering and weightage on the historical events based on different condition checking.

4 EXPERIMENTAL RESULTS AND ANALYSIS

To determine the impact of adding a distribution for dedication, two scenarios were analyzed to assess the impact. To establish a baseline that can be used to compare the benefit of the path distribution, the first scenario was created using a static number of paths based on any equipment that was enabled at least once for a PSS. This provides the simplest to model and most flexible case which will allow an estimate of the impact of Seagate's path enablement/disablement impact on factory velocity. To reduce the noise from other sources of variation in the factory during the analysis period the following modeling elements were input into the model from historical data:

- Real Wafer Start
- Real WIP (historical WIP release time from hold + termination)
- Real Equipment Down
- Buffer Release Based on Logic (Not actual lot level historical release schedule)
- Hold Modelling Based on Historical probability + Average Duration (Not actual lot level historical release schedule)
- Snapshot dispatching parameters

The second scenario created uses the same historic deterministic inputs as the baseline while adding in the time and moves weighted dedication distributions for each PSS. Both scenarios were run for 8 weeks, each with 5 replications, with the results shown below:

The major KPI to assess how the new method worked is by comparing factory velocity (average moves/wafer) to determine how well the simulation reflects the factory in this period with and without the dedication distribution:

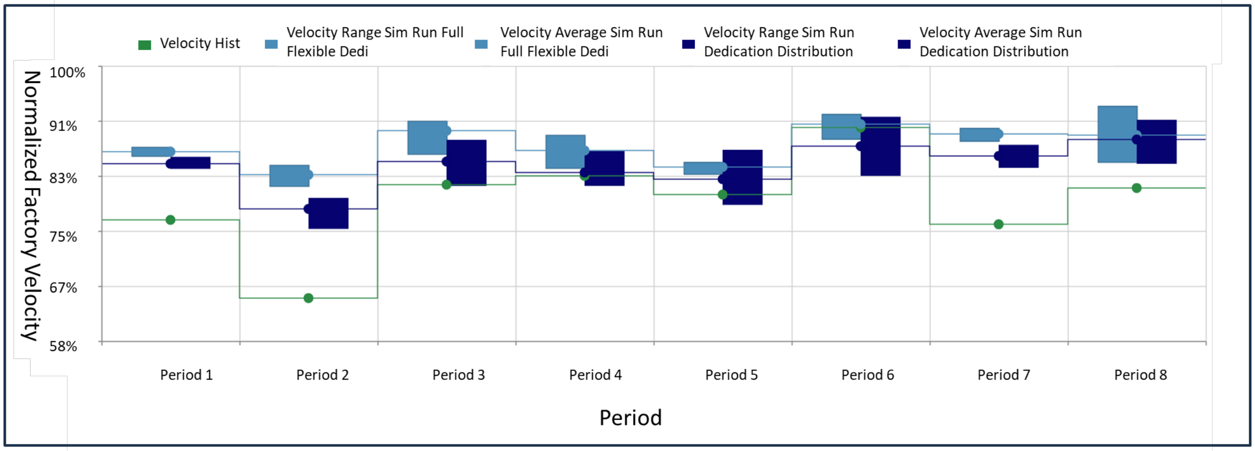

Figure 2: Velocity range across replications.

In Figure 2 the blue range represents the min and max factory level velocity range with a static dedication based on the history vs the green bar which represents the range with the time moves weighted dedication distribution included. As Figure 2 shows in every period the scenario that includes the dedication distribution is closer to the actual historical performance indicated by the dark blue line. To compare the change in performance versus actuals an overall average forecast quality (FQ) metric is calculated as shown in Figure 3.

$$Forecast\ Quality\ (FAB\_KPOV) = 1 - \frac{ABS(FAB\_KPOV(SIM) - FAB\_KPOV(HIST))}{MAX(FAB\_KPOV(SIM),\ FAB\_KPOV(HIST))}$$

Figure 3: Forecast quality measurement.

The FQ vs hist is 90.4% versus an improved FQ of 92.8% using the weighted distribution. Another key factory metric measured is average factory moves per week with results shown below in Figure 4.

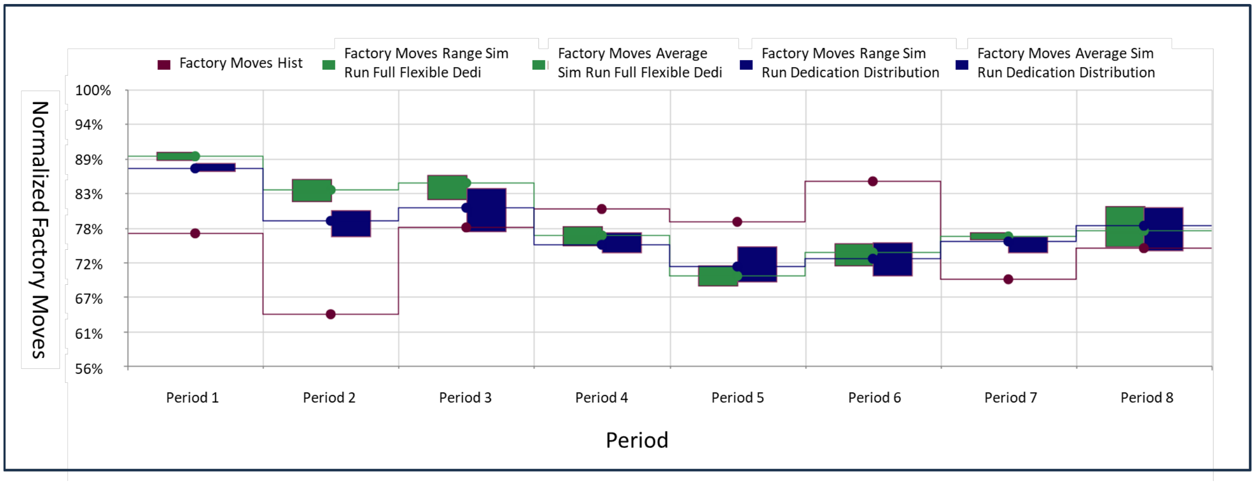

Figure 4: Factory moves comparison.

Similarly to the factory velocity comparison, the moves with the dedication distribution applied are closer to the historical performance in almost every period with the factory moves FQ improving from 88.9% to 90.2%.

To assess the impact at the equipment level a high moves equipment was analyzed to see the impact at a lower level as shown in Figure 5.

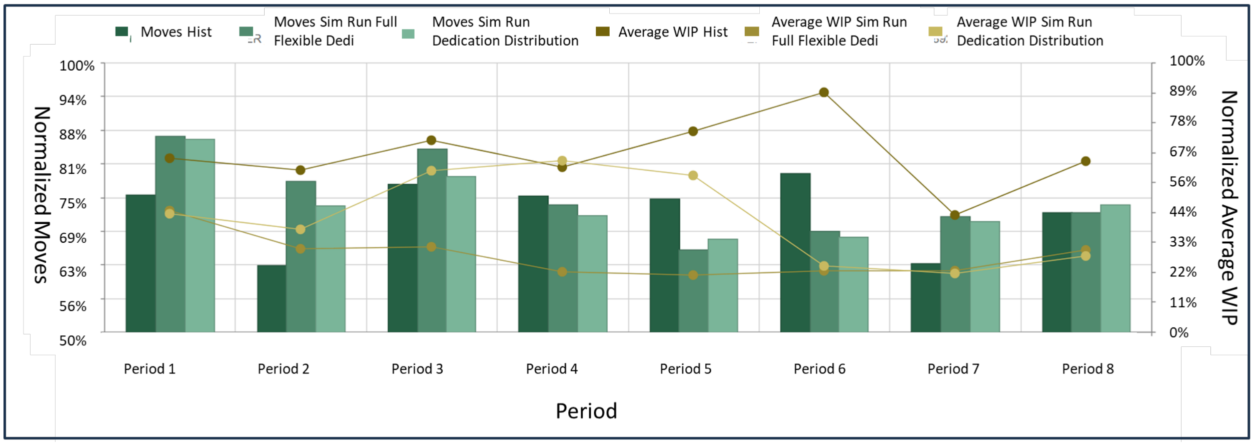

Figure 5: Equipment moves vs WIP levels.

The left most dark bar represents historical average moves while the middle bar represents the scenario without the dedication distribution moves and the lightest rightmost bar represents moves with the dedication distribution included. In addition, the brown line represents average historical WIP, the middle yellow line with the dedication distribution and the brown line without the dedication distribution. With the addition of the dedication distribution the overall average FQ improved from 89.8 to 91%. The average WIP also improved with the middle 3 weeks showing a far closer behavior in WIP even at similar moves rates when comparing the addition of dedication distribution versus history.

5 CONCLUSIONS AND FUTURE WORK

As the scenario outputs show at several different levels from the overall factory performance accuracy down to the weekly moves and WIP accumulation at the equipment group level the addition of a distribution for PSS path dedication improves FQ. By accounting for the engineering practices of controlling processes at multiple levels from overall step to as specific as product stage step or even a single lot at a stage step, the slowing down of the factory can be observed and the accuracy of the simulation improved.

While this experiment captures one aspect of the variation in PSS paths seen at Seagate, there are additional considerations that can also be studied to improve long term forecast accuracy. As mentioned in Section 3.3, as new products are beginning to ramp, they have a longer disabled duration and often less potential paths which often is not the case as the product reaches full maturity. One limitation of using a distribution by PSS is that the full number of future paths may not be included at all in a historical data set which would have a larger impact the farther into the future the simulation projects as the volume of the path-limited product grows. One future scenario study to address this issue would be to estimate future paths for new products based on a hierarchy of path enablement. For longer time horizon simulations, the number of paths enabled for a new product could be used as a stage step distribution instead of the more constraining PSS distribution. If the stage step is completely new, then the future paths estimate could fall back to a more general equipment group and step level distribution. Since the same stage step across different products often exhibits the same process behavior or physical properties in the case of consumable constrained equipment, a more general case is often a good estimate for future products. This would ensure that low volume products with a small volume of data and a low number of historic enabled paths would not be artificially slowed down by lack of future potential paths. This method would also capture the historical impact of the dedication distribution experienced for other products which was shown to be an improvement in FQ over a static distribution.

REFERENCES

Bagchi, S., C.H. Chen, S.T. Shikalgar, and M. Toner. 2008. “A Full-Factory Simulator as a Daily Decision-Support Tool for 300mm Wafer Fabrication Productivity”. In *2008 Winter Simulation Conference (WSC)*, 2021-2029 https://doi.org/10.1109/WSC.2008.4736297.

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 *2018 Winter Simulation Conference (WSC)*, 3639 – 3650 https://doi.org/10.1109/WSC.2018.8632286.

D-SIMLAB Technologies. 2024. Dynamic Capacity & Material Flow Performance Planner. https://d-simlab.com/category/d-simcon/products-d-simcon/dynamic-capacity-planner/, accessed April 24, 2024.

Ohno, Taiichi. 1988. *Toyota Production System - Beyond Large-Scale Production.* Productivity Press. pp. 25–28.

Scholl, W., D. Noack, O. Rose, B. P. Gan, P. Lendermann, P. Preuss *et al*. 2011. “Implementation of a Simulation-based Short-term Lot Arrival Forecast in a Mature 200mm Semiconductor Fab”. In *2011 Winter Simulation Conference (WSC)*, 1932-1943 https://doi.org/10.1109/WSC.2011.6147907.

Seidel, G., B. P. Gan, C. W. Chan, C. F. Lee, A. M. Kam, A. Naumann *et al*. 2017. “Harmonizing Operations Management of Key Stakeholders in Wafer Fab using Discrete Event Simulation”. In *2017 Winter Simulation Conference (WSC)*, 3670-3678 https://doi.org/10.1109/WSC.2017.8248079.

AUTHOR BIOGRAPHIES

RICHARD SURMAN is an accomplished Staff Manufacturing Engineer at Seagate Technology, based in Springtown, Ireland. He started work in semiconductor manufacturing in 1994. His expertise extends across various domains, including manufacturing analytics, capacity planning, equipment installations, production control, and labor modelling. In recent years, Richard has focused on simulation projects, collaborating with multiple simulation packages. His work has involved investigating complex dispatch rules, factory automations, stocker implementations, factory velocity, bottleneck analysis, and the impact of capital investments. His dedication to advancing manufacturing processes and optimizing efficiency underscores his role within Seagate Technology. His email address is richard.b.surman@seagate.com.

MATT NEHL is an Industrial Engineer at Seagate Technology, based out of the Minnesota Normandale Campus. He has a bachelor's degree in industrial engineering from the University of Wisconsin-Madison and a master's in industrial engineering from the University of Minnesota. He started working for Seagate Technology in 2017 and had several different roles, primarily planning capacity for vacuum metal and dielectric tools. In later roles he has shifted to factory modelling primarily with mixed integer linear programming strategic decision models as well as labour planning and simulation forecasting. His email address is matthew.nehl@seagate.com.

COLE EVANSON is an Industrial Engineer at Seagate Technology, based at the Minnesota Normandale Campus. He graduated from the University of Minnesota with a degree in Industrial and Systems Engineering in 2019, and he began working at Seagate in 2021. He has experience in manufacturing capacity and capital planning, cycle time data analytics, and data visualization. He currently works as a wafer owner with a focus on inspection and vacuum dielectric toolsets. His email address is Cole.evanson@seagate.com.

SOO LEEN LOW is a VP Operation at D-SIMLAB Technologies (Singapore). She is responsible for managing simulation project delivery, building simulation models and conducting model validation analysis for 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.

KERN CHERN CHAN is a Senior Data Analyst at D-SIMLAB Technologies (Singapore). He is responsible for building and validating simulation models of wafer fabrication plants. He graduated with a Bachelor of Science in Physics from National University of Singapore (NUS) in 2020. He also holds a Master of Science degree from NUS, specialising in Data Science and Machine Learning. His email address is kern.chern@d-simlab.com.

HUI SIAN LIAU is a Software Engineer at D-SIMLAB Technologies (Singapore). She is responsible for data transformation and validating simulation models of wafer fabrication plants. She graduated with a Bachelor of Applied Science in Applied Statistics from Universiti Sains Malaysia (USM) in 2020. She also holds a Master of Science degree from USM in 2022, specialising in Data Science and Analytics. Her email address is chris.liau@d-simlab.com.

BOON PING GAN is the CEO 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 led a team of researchers and developers in building a suite of products in solving wafer fabrication operational problems. 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.