*Proceedings of the 2022 Winter Simulation Conference*
*B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.*

# APPLICATION OF A SIMULATION MODEL TO FORECAST CYCLE TIME BASED ON STATIC MODEL INPUT

Syahril Ridzuan Ab Rahim
Gwendolene Haw
Wei Jin Lee

GlobalFoundries Singapore
60 Woodlands Industrial Park D
Street 2
Singapore, 738406, SINGAPORE

Oliver Diehl

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

## ABSTRACT

The semiconductor wafer fabrication industry is having the challenge to address increasing demand during the global chip shortage. The semiconductor industry is increasing its capacity utilization while maintaining its cycle time to address this issue. Currently, static capacity modelling is used to calculate capacity utilization. Due to the complexity of the cycle time calculation, the existing static capacity model is unable to estimate cycle time. One of the pain points for cycle time modelling is the preparation and maintenance of the real-time data as inputs into a simulation model. Therefore, this paper aims to demonstrate how the dynamic capacity engine (DCE) utilizes most of the ready input data from static capacity modelling to forecast fab cycle time. The high accuracy of the DCE is achieved and validated with the actual input (wafer start, actual WIP) and output (dynamic cycle time) which will be discussed in this paper.

## 1 INTRODUCTION

The semiconductor industry has demonstrated a significant rise in demand in recent years. In the year 2021, Semiconductor Industry Association reported global semiconductor industry sales reached a record $556 billion, which is a 26.2% increase compared to the year 2020 (Kharpal 2022). According to the forecast from the industry expert, the demand for semiconductors is expected to rise in the next few years as chips become more heavily embedded in the technologies today and future as well. Due to smartphone usage and the computing power boom, the demand for microcontrollers, memory chips, and processors has grown fast. On top of this, we are seeing mechanical machines like cars requiring more chips to support the innovative smart system inside the vehicle.

Data analytics is used in GlobalFoundries (GF) to alleviate some of the challenges arising from the increasing demand for chips which typically involve hundreds of steps in the manufacturing process (Shikalgar et al. 2002). The wafer fabrication process consists of several basic processing units such as photolithography, etching, and ion implantation. Wafer lots go through each processing unit multiple times during production. In an industry where production equipment is expensive, static production data of considerable size is collected in a database and analyzed to identify the opportunities to maximize long-term profits and reduce costs and losses (Seidel et al. 2020).

Among the myriad of performance metrics to focus on is the cycle time, defined as the total time taken to complete a recurring task which in the case of a wafer fabrication plant (fab) is also known as the time taken to produce a single wafer lot. As a critical factor to measure a business's efficiency, shorter cycle times mean an optimized wafer production process and faster time to market. Since the manufacturing of wafers continues until a defect is detected and eliminated, another benefit from the reduction of cycle time is the opportunity cost savings when defects in defective products can be detected earlier and eliminated faster as suggested by (Nemoto et al. 1996). Lengthening of cycle times would mean that there are inefficiencies in the process, and delays are to be expected for customers. Cycle times can be estimated based on historical data collected in the database to calculate the expected delivery dates. However, these estimated cycle times are not always accurate when the fab conditions change, especially when the fab loading is increasing to its historical high and cycle time is expected to be affected. Thus, the historical cycle time data has its limitation to forecast future cycle time.

With the high demand for chips from various industries, the fab is running at its maximum utilization and even above its available capacity by improving tools' planned uptime and throughput. GF is taking a calculated risk to run the production line above 100% planned utilization to support the customers' orders. When the fab is running at high utilization, there will be a trade-off with the extended cycle time to maximize fab output.

In GF, static capacity modelling is used to calculate equipment utilization based on the equipment planned moves divided by equipment capable moves. The ultimate objective of this modelling is to identify overutilized equipment groups and focus on productivity improvement initiatives to meet the fab capacity support. However, the static capacity modelling is unable to reflect the cycle time impact with the increase in fab utilization.

Historical fab performance data is the key input for the simulation model to analyze relevant performance metrics to evaluate fab effectiveness. In other sectors, the existing practice of conducting simulations incorporates a substantial amount of real-time data and resources effort to maintain the database. With the growth in product mix complexity, the effort of managing this large database has grown significantly more onerous.

As an alternative solution, GF has enhanced the internally developed static capacity model with D-SIMLAB's dynamic simulation model known as DCE which can provide important information which is not available in the static capacity model, such as dynamic cycle time entitlement, projected fabout, and WIP movement. Fabout is defined as the number of wafers that completed the last step of fab processing. This dynamic simulation model leverages most of the available input data in the static capacity modelling to ensure sustainability and minimize additional resources to maintain the database.

This paper presents sections discussing the details of the DCE model including the methodology, assumptions, important key elements, and validation results. Results of the simulations conducted with the model will be discussed.

## 2 BACKGROUND

Static models can enable faster implementation, are easier to use and maintain, and can immediately provide preliminary estimations with broad-based assumptions. In contrast, simulation requires fewer simplifying assumptions but data requirements need to be defined systematically. The challenge of a simulation model is that it takes longer to provide results than static models and requires significant resources and efforts to maintain, and the unpredictability inherent in simulation models does not ensure consistent outcomes between runs.

The benefits of simulation include the ability to properly represent random events and to incorporate rules that are influenced by the current state of the system during the simulation run. Variables in the system can also be tracked and analyzed statistically (Foster et al. 1998).

Before the development of DCE, the fab was using an empirical approach based on actual performance as a quick method to understand the fab cycle time characteristic and performance forecasting.

In the semiconductor industry, static cycle times and dynamic cycle times are frequently used. Static cycle time is the actual measurement of a single wafer from start to finish. Dynamic cycle time is the overall average cycle time. While static cycle time can fluctuate depending on equipment availability, resource availability, and equipment utilization, dynamic cycle time reflects more accurately the steady-state of production.

## 2.1 Cycle-time Characteristic Curve

Before the DCE was developed to forecast cycle time, the fab used the turn ratio-cycle time curve for cycle time projection.

First, we can calculate the actual turn ratio, which is one of the important indices used by fab operation to measure the dynamic speed of WIP movement using equation (1). Then we can plot the corresponding dynamic cycle time based on the historical data in Figure 1. With the turn ratio-cycle time curve, the fab can know what the required turn ratio is to meet certain desired dynamic cycle times.

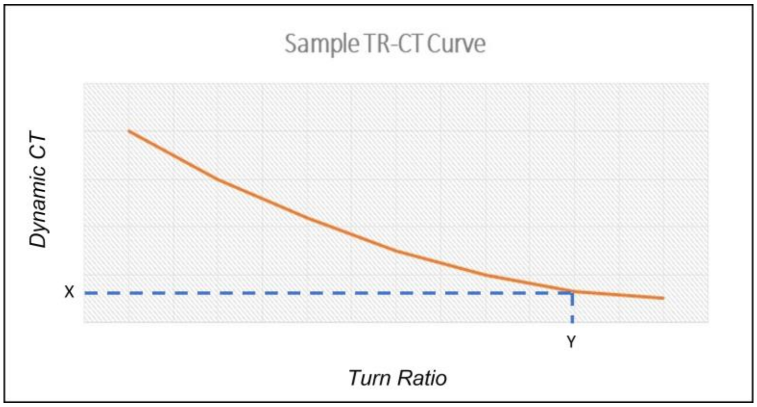

Figure 1: Turn ratio - cycle time curve.

For example, if the fab is targeting dynamic cycle time of X value, then the fab needs to perform at Y value of turn ratio.

This method is used as a quick reference for the fab to understand the cycle-time capability of the fab before the simulation model is developed. The validation result showed that the actual cycle-time performance was higher by 6.9% than the cycle-time characteristic curve. Due to the relatively big gap between the actual and forecast, there is a motivation for the team to look for better solutions to forecast the cycle-time performance. On top of that, the output produced by the cycle-time characteristic curve is limited. For example, there is no other output such as the fabout projection, and WIP projection data available. Hence, DCE is developed to overcome the limitations that the team is facing.

## 2.2 Dynamic Cycle Time Formula in DCE Applications

For DCE applications, dynamic cycle time is one of the most important outputs. We will discuss the formula of dynamic cycle time that DCE used in this section.

First, we need to find out one of the key indices in daily fab operation which is the Turn Ratio. It is being calculated as:

$$Turn\ Ratio(TR) = \frac{Process\ Moves}{Active\ WIP} \tag{1}$$

Turn ratio is a measurement of the average number of steps processed each day in the fab. Process moves are defined as the number of wafer passes at process steps, and active WIP is the number of WIP currently in our production line at the time when the data is extracted. Hence, we can calculate the dynamic cycle time in days and days per mask layer (DPML) using the following formulas:

$$\text{Dynamic Cycle Time (in Days)} = \frac{Total\ Weighted\ Steps}{Turn\ Ratio}$$

$$\text{Dynamic Cycle Time (in DPML)} = \frac{Dynamic\ Cycle\ Time\ (in\ Days)}{Total\ Weighted\ Mask\ Layers}$$

By rearranging these 2 formulas, we can conclude the dynamic cycle time formula as below:

$$\text{Dynamic Cycle Time (in DPML)} = \frac{Total\ Weighted\ Steps}{Total\ Weighted\ Mask\ Layers} \times \frac{1}{Turn\ Ratio}$$

## 3 DYNAMIC CAPACITY ENGINE (DCE) APPROACH

### 3.1 Static Capacity Planning Simplifications

Certain simplifications are usually made for the static capacity model to keep the model simple and the data maintenance manageable.

In a wafer fab often a few hundred or even thousands of different products are manufactured concurrently, however, most of the time there are certain products with only small differences in terms of the process steps they are passing through. For static capacity planning, they can be grouped to a specific representative product process flow, which is used for the modelling. The representative routes used in the model are normally high runner products with significant volume running in the fab.

Another simplification is to use a recipe dedication by equipment group, which means that all equipment belonging to one equipment group has the same capability to process the recipes dedicated to the equipment group. This simplification is applied as it is usually difficult to obtain detailed dedication by equipment for the long-term future, as the dedication is adjusted dynamically by inhibiting certain recipes for certain equipment for a shorter period of time or releasing the recipe on additional equipment. For lithography equipment, the dedication has been provided equipment specific, as for these equipment typically the dedication is assigned not only by recipes but also by products and layers.

### 3.2 Dynamic Capacity Engine versus Static Capacity Planning

The DCE is a discrete event simulation model. It simulates the lots flowing through the virtual equipment in the virtual fab. This allows the measurement of the cycle time of the lots, which is not easily possible with a mathematical model, due to the high complexity of the manufacturing process in a wafer fab.

Table 1 shows a comparison of some KPIs between the static capacity model and the dynamic model, which are important in the capacity planning process.

Table 1: KPI Comparison – DCE vs Static Capacity Planning.

| KPI | Static Model | Dynamic Capacity Engine |
|---|---|---|
| Dynamic Cycle Time | Not considered | Output of model |
| Equipment Batch Factor | Input to model | Output of model |
| On-time Delivery | Not considered | Output of model |
| Queue Time | Not considered | Output of model |
| WIP | Not considered | Starting WIP as Input and WIP forecast as Output of the model |
| Fabout | Input to model | Output of model |

GF's static capacity model is built by incorporating various inputs such as demand forecast, ramp plan, recipes UPH, equipment dedication, and equipment quantity. These parameters are used to compute equipment utilization, generate capacity reports, and conduct bottleneck analysis. The static capacity engine allocation algorithm allocates the demand to the available equipment with the objective to balance the equipment utilization in the fab. If the fab is loaded with a specific amount of product mix, a bottleneck analysis will then be performed based on the calculated equipment utilization. Mitigation plans, such as productivity improvement, are then required to support the specified loading.

## 3.3 DCE Approach

Compared to the static model, some additional input data and modelling elements are required to generate a model which better represents reality and improves the simulation results. The approach taken is to build a hybrid model, which relies mainly on the data of the static capacity model, enhanced with some additional data. For the additional data, it was emphasized to use existing data sources and automatically feed them into the model to reduce the user effort as much as possible. As a result of the approach, the time to generate the model is significantly reduced and results can be obtained quickly. The underlying simulation engine from D-SIMLAB is specifically designed to represent all the features required to model the semiconductor fab operation with very high fidelity and accuracy.

### 3.3.1 Recipe Process Time (Full Lot Cycle Time)

In convention, for a static capacity model, only the throughput (UPH, defined as units per hour) of a recipe at an equipment is considered, which describes the production rate of lots being processed and completed by the equipment. For the simulation model, it is important to portray how long a lot is staying within the equipment for processing, as this impacts the lot cycle time. Therefore, the Recipe Process Time is an additional input and describes the time from the lot track in time into the equipment to the lot track out time.

### 3.3.2 Equipment Type

The simulation model considers the different equipment types to more accurately portray the equipment behavior in the simulation. The information is directly included from an online database, which was already available. It describes if the equipment is a batch equipment and what is the maximum number of lots per batch or if the equipment is a cluster tool with a set of chambers and the operating mode of the chambers (parallel or serial). During the simulation, the lots at these equipment are then also processed in batches if lots with the same recipes are available.

### 3.3.3 Metrology Equipment

For the lot cycle time, it is important that the metrology steps in the process flow are considered, as they significantly contribute to the lot cycle time. In DCE, two different approaches for this have been implemented. A big challenge for the metrology steps is the dynamic behavior of the sampling rate and sampling rules, which often can change over the time of the scenario horizon and are difficult to predict.

The metrology steps are automatically added to the process flow based on historical data. So only steps which are processed and their respective sampling rates are added to the respective process flow of the product. For future products, this is done manually.

In the first modelling approach, these actual sampling rates and process times are applied as per the data from the system.

A second approach is that the user defines the total number of days a product on average accumulates on the metrology steps and this duration is then distributed across the metrology steps of the process flow. This is called Fixed Metro Cycle Time Modelling.

### 3.3.4 Dispatch Rules

In the actual operation of the fab with a Real-Time-Dispatch System (RTD), a lot of different dispatch rules are implemented at the different equipment types, often designed in a way to optimize some local objective. Not all of these detailed dispatch rules are required for a long-term simulation used for cycle time and capacity planning. The current model implements a few of the Global Dispatch Rules, which are typically defined in most of the equipment.

1. Lot Priority: Lots are given a lot priority at the wafer starts. For example, it can be defined as the percentage of high-priority lots per product. Lots with higher priority are dispatched first.
2. Batch Forming: The waiting time to form a full batch at a batch equipment is a maximum of 50% of the recipe processing time.
3. Step Goal: For every lot at the wafer start it is defined how many process steps the lot requires to complete per day to complete on time, this is the step goal. Lots which have already completed the number of steps per day required are less urgent as compared to lots that have not reached their step goal.
4. Critical Ratio: The critical ratio is the remaining total process time divided by the remaining time to planned fabout including the waiting time. Lots with a higher critical ratio are more urgent.

Under the Batch Forming dispatch rule, if the batch factor is 0.98, that signifies that 2% of the total lots are not part of a full batch. As an example, if a recipe's total processing time is 4 hours then the Batch Forming time is at a maximum of 2 hours, lots will proceed even if a batch is not full. A combined score of the step goal and the critical ratio is calculated and used for the lot dispatch for lots of the same lot priority.

### 3.3.5 Fab WIP

For a longer-term simulation horizon of 1-2 years, it is important to consider the current state of the fab and specifically the current WIP, which includes all lots and their current process step. Considering WIP will take several months to complete, the current WIP must be taken into account in the simulation run for the first five months. While the existing WIP is no longer necessary after the sixth month of simulation because it has already been completed and this WIP has been replaced with new wafer starts. Active WIP must be included in simulation runs since it competes for equipment capacity with the new wafer start and has a substantial impact on simulation results.

The simulation model gets initialized with the actual WIP at the simulation start time. The WIP is available every day from the online database. As per section 3.1, the products of the WIP lots have to be mapped to the available high runner products, which are mapped to the respective high runner process flow so that the lots can be completed according to this flow. This is done automatically based on a mapping table. Due to the mapping sometimes the exact same process step at which the WIP lot is currently processing is not available. For this case, a logic model is implemented to map the lot to the next nearest process step according to step number and stage information.

## 4 DCE APPLICATION AND RESULTS

While constructing this simulation model, several fine tunings have been done before achieving the intended accuracy of more than 95% as set by fab management. Before running any simulation, simulation input data such as equipment quantity, equipment capable moves, product dedication, wafers loading quantity, WIP quantity, and product low runner grouping are validated against the fab's planned figures to ensure that the model uses the correct information. An Excel comparison template is created for validation. The simulation input data and the fab planned data are loaded into the template and compared to see if there is any mismatch. Any mismatch data discovered during validation is corrected, and the validation process is repeated until all of the simulation input data matches the planned figures.

The simulation model is configured to produce the dynamic cycle time forecast and other important output like fabout quantity and WIP levels used by fab management. Each of the results is validated with the actual data to understand its deviation. Detailed validation of the results is further discussed below.

### 4.1 Dynamic Cycle Time

Figure 2 indicates the monthly overall dynamic cycle time forecast numbers from simulation compared with the actual data from the fab. The cycle time shown is the overall fab dynamic cycle time inclusive of various products ran in the fab. We are seeing there is a delta ranging from 1.2% to 3.8% from month to month when we compare the simulation result and actual data. On average, the simulation shows a delta dynamic cycle time of 2.1% compared to the actual data. This difference between forecast and actual is significantly lower than the empirical approach, which has a delta of 6.9%. The delta between simulation and actual is because some real-time issues are not captured in the simulation, for example, equipment recipe inhibition, product on-hold by customers, time-link constraint, etc. By understanding the delta between actual and simulation, the team can predict the future dynamic cycle time by adding in a 2% offset from the simulation result.

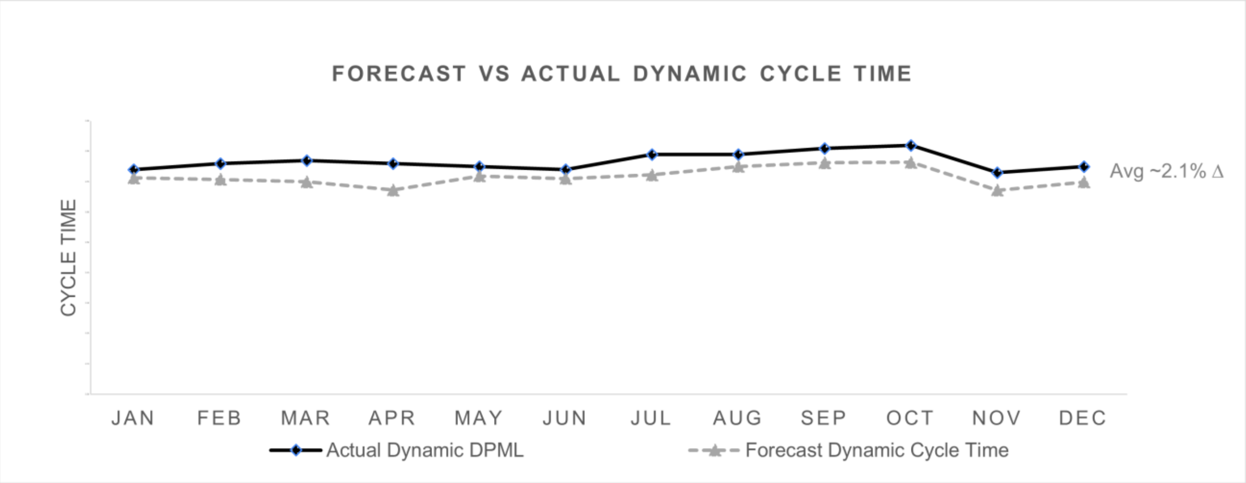

Figure 2: Forecast vs actual dynamic cycle time.

### 4.2 Fabout Quantity

Besides cycle time, this simulation model is also able to provide a forecast for fabout quantity. Figure 3 shows the comparison between quarterly forecast fabout and actual fabout numbers. The delta between forecast and actual are ranging between 0% to 4%. Forecast numbers are higher due to fewer variables interacting with the lots in the simulation as compared to the actual situation in the fab. An example of a variable that is not configured inside the simulation model is the dynamic equipment inhibition that varies day to day that is causing less equipment to be available at the point of time. From this simulation model, the differences between forecast and actual numbers are within our target of 5% and we have validated the data over a 12-month period. Thus, this model is capable of being used to forecast the fabout quantity for better planning.

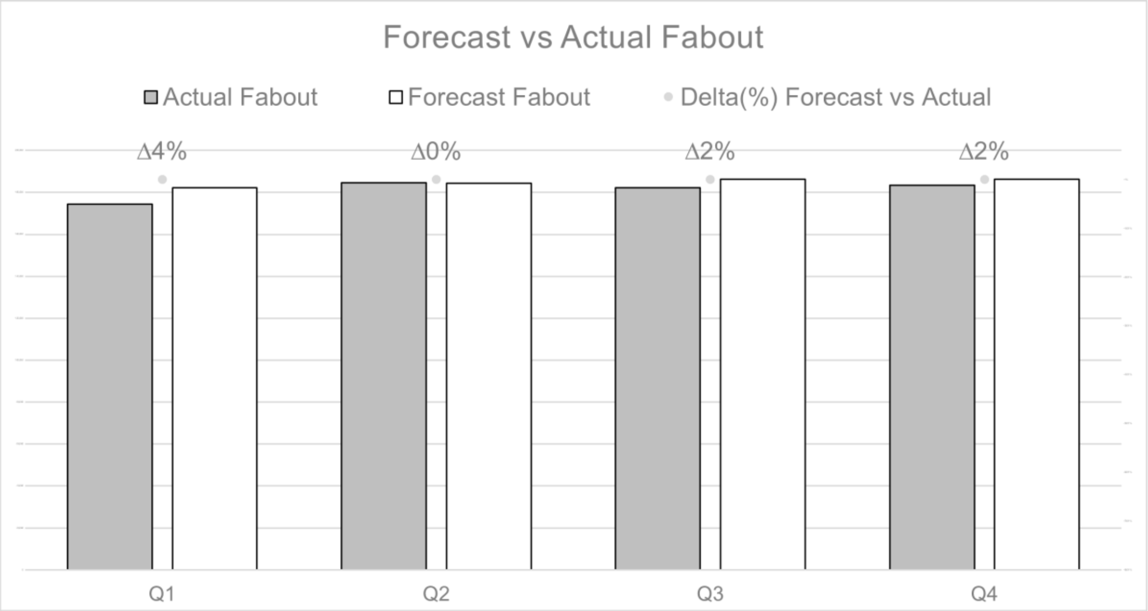

Figure 3: Quarterly forecast versus actual fabout.

### 4.3 WIP

Figure 4 shows the comparison between the simulated WIP level and the actual WIP level in the fab. The deltas range between 0.5% to 1.5%. From the same figure, we can observe that the trend in the simulation is aligned with the trend in the actual data. Thus, we can conclude that the simulation model is capable to forecast the WIP level in the fab. WIP forecasting is critical for the fab to ensure that there is enough WIP to achieve the overall fabout target by the end of the year, as well as to allow capacity planning for carriers and storage space.

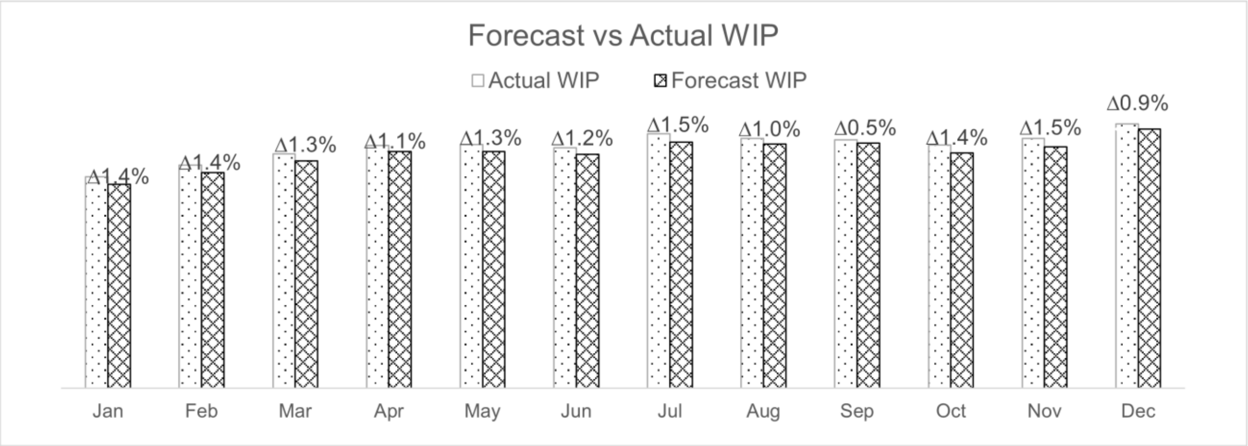

Figure 4: Forecast vs actual WIP level.

## 5 CONCLUSION

Although both the cycle-time characteristic curve and the DCE simulation results can project the expected cycle time, the cycle-time characteristic curve has a large variance between the projected and actual cycle times. DCE, on the other hand, can use the existing static capacity model to produce not only the entitled cycle time but also projections on the monthly fabout and WIP inventory with higher accuracy. Contrary to conventional wisdom, the simulation findings revealed DCE's capability to provide high-quality results without the need for a substantial amount of real-time data by using static data and minimal amount of user input.

Moving forward, this simulation model may be expanded in numerous ways. With the new fab expansion planned by GF, the simulation model can be expanded to simulate cross-fab moves. The purpose of this expanded feature is to forecast the capacity requirement of the automated material handling system that moves between two fabs. At the same time, the team is expecting the model to identify the equipment groups that have high cross-fab moves for the team to explore improvements to reduce the cross-fab moves.

## ACKNOWLEDGMENTS

We would like to express our gratitude to GF IE management, Mr. Eng Teck Yeo, Mr. Jackson Tan, Mr. Akmal Syamsul, and the CEO of D-SIMLAB Technologies (Singapore), Mr. Boon Ping Gan, for their contributions to the successful completion of this paper.

## REFERENCES

Foster, B., D. Meyersdorf, J. M. Padillo and R. Brenner. 1998. "Simulation Of Test Wafer Consumption In A Semiconductor Facility". In *IEEE/SEMI 1998 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop (Cat. No.98CH36168)*, edited by Semiconductor Equipment and Materials International (SEMI), 298-302, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Kharpal, A. 2022. Global Semiconductor Sales Top Half A Trillion Dollars for First Time as Chip Production Gets Boost. https://www.cnbc.com/2022/02/15/global-chip-sales-in-2021-top-half-a-trillion-dollars-for-first-time.html, accessed 17th March 2022.

Nemoto, K., E. Akcali and R. Uzsoy. 1996. "Quantifying The Benefits Of Cycle Time Reduction In Semiconductor Wafer Fabrication". In *IEEE Transactions on Electronics Packaging Manufacturing*, edited by R. W. Johnson, 39-47, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Seidel, G., C. F. Lee, A. Y. Tang, S. L. Low, B. P. Gan and W. Scholl. 2020. "Challenges Associated With Realization Of Lot Level Fab Out Forecast In A Giga Wafer Fabrication Plant". In *2020 Winter Simulation Conference (WSC)*, edited by K. G. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder and R. Thiesing, 1777-1788, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Shikalgar, S. T., D. Fronckowiak and E. A. MacNair. 2002. "300 Mm Wafer Fabrication Line Simulation Model". In *Proceedings of the Winter Simulation Conference*, edited by E. Yucesan, C. H. Chen, J. L. Snowdon and J. M. Charnes, 1365-1368, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

## AUTHOR BIOGRAPHIES

**SYAHRIL RIDZUAN BIN AB RAHIM** is a Senior Engineer in the Line Analytics Department of Industrial Engineering at GlobalFoundries Singapore. His scope of work is in the discrete event simulation model and throughput improvement analysis. He has been serving in the wafer fabrication industry since 2011. His email address is syahrilridzuan.abrahim@globalfoundries.com.

**GWENDOLENE QIN LING HAW** is an Engineer in the Line Analytics Department of Industrial Engineering at GlobalFoundries Singapore. She is currently responsible for the discrete event simulation model for Fab7 & Fab7H in GF. She earned her Bachelor’s degree in Mechanical Design and Manufacturing Engineering at Singapore Institute of Technology – Newcastle University in 2021. Her email address is gwendoleneqinling.haw@globalfoundries.com.

**WEI JIN LEE** is a Member of Technical Staff Engineer in the Line Analytics Department of Industrial Engineering at Globalfoundries Singapore. He has been involved in Industrial Engineering topics like capacity planning, productivity improvement, capacity optimization, and simulation since 2007. He is now responsible to lead the Line Analytics team in GF Singapore. His email address is weijin.lee@globalfoundries.com.

**OLIVER DIEHL** is VP Projects Semicon at D-SIMLAB Technologies (Singapore). He is responsible for projects of static and dynamic capacity planning in various wafer fabs worldwide for the past 8 years. Prior to this he has gained experience in different areas of the semiconductor and electronics manufacturing, such as failure analysis, automatic optical inspection and circuit simulation. His email address is oliver@d-simlab.com.