Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo and M. Loper, eds.

OPERATOR RESOURCE PLANNING IN A GIGA FAB DURING COVID-19 RESTRICTIONS

Ching Foong Lee
Aik Ying Tang

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

Georg Seidel

Infineon Technologies Austria AG
Siemensstraße 2
Villach 9500, AUSTRIA

Soo Leen Low
Boon Ping Gan

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

ABSTRACT

With the world facing a public health emergency due to the Coronavirus disease (COVID-19) in a global pandemic, this paper provides insight about how a simulation model was used to determine the impact of headcount variability during lockdown on fab performance. To create a robust simulation model, operator loading time was introduced as one of the input parameters. An existing and well validated Discrete Event Fab simulation model was extended with operator modelling, and was used to conduct case studies, evaluating the impact of different operator availability scenarios including work disruptions for several shifts within a week. The studies provide implications for operation to derive mitigation strategies, weighing the trade-off between cost demand and speed loss due to operator resources.

1 INTRODUCTION

Infineon's wafer fabrication facility in Kulim High Tech Park, Kulim Kedah, Malaysia is an 8 inch high volume wafer fab in Southeast Asia which presently employs more than 3,000 staff. Infineon Kulim is positioned as a manufacturing competence center focusing on megatrend technologies for the efficient use of electric energy and the electrification of automobile for improved performance and safety as well as reduced carbon emission. The wafer fab facility started to use discrete event fab simulation in 2016. A first simulation model was introduced to forecast operation KPIs with a 7 days horizon only (Seidel et. al. 2017). This model runs automatically every day, normally without any user interference, and it is called the short-term simulation model (STS). Since 2019 a second simulation model has been available as well, which can be used for scenario runs, for any chosen time period. The second model is highly interactive, and provides the user options to manipulate input data, to generate different scenarios. This model is called the long-term simulation model (LTS). The important considerations for building a good simulation model can be found in Fowler, Mönch, and Ponsignon (2015).

Both models share common basic data e.g. routes, products, equipment, dedications, etc. The model structures are very similar. But of course the LTS needs some kind of future change list, to model e.g. adding of new equipment and dedications changes over time.

978-1-6654-3311-2/21/$31.00 ©2021 IEEE

Due to the Covid-19 pandemic, governments around the world including Malaysia declared lockdowns to fight the spread of the pandemic. Semiconductor manufacturing for the essential product were allowed to continue operating though with only certain percentage of its normal workforce. This sudden declaration of lockdown imposed a significant challenge for daily plans of production activities. With the unplanned lockdown, implementation of precautionary steps to ensure stability of machines, materials and human resources became an ad-hoc discussion and decision making for operation. The focus was to minimize risks and production losses in the wafer fab.

Operating a fab production with such resource limitations created major challenges for production. There was a new requirement to judge impact of every operational decision made because there was no past data to study from. This lack of visibility through experience led to requests to the simulation team to evaluate several alternative scenarios and provide insight into the implication of decisions made. The simulation studies were conducted in collaboration with different departments, defining the input parameters to the simulation model such as choice and schedule of equipment warm down, human resource scheduling, start mix profile, etc.

In Section 2 we describe important simulation model features and the way how input data was calculated to model operator behavior. Section 3 presents scenario results and Section 4 provides conclusions and outlook.

2 SIMULATION MODEL

2.1 Modelling Elements

Discrete event simulation is commonly used in the wafer fab for the purpose of WIP flow forecasting, dynamic capacity planning, and scheduling (Mönch, Fowler, and Mason 2013). The essential modelling elements for both LTS and STS for high accuracy model forecasts are summarized in Table 1 and Table 2 below. It is the same model that has already been described in Seidel et. al. (2020). The simulation model is built on a commercial off-the-shelf simulation engine, the D-SIMCON Forecaster (D-SIMLAB Technologies 2021). The simulation model was validated by choosing a historical time period as a basis of comparison between simulation and real performance. The detailed methodology and results can be found in Seidel et. al. 2017.

As discussed in Section 1, the standard LTS and STS models do not model operators as it is not a critical limiting resource. But with the lockdown declared, with possibly less than 50% of operators returning to work, operators became the one key resource that limited the capacity moves in the fab. Operator resource constraint can be modelled as a standard feature of the D-SIMCON simulator.

An operator group is used to model an additional resource required to handle lots processing at an equipment or for any equipment maintenance events. Other activities such as searching and transportation of lots are captured as part of the time an operator resource is required to process a lot or batch. Several operator groups can be created based on the equipment characteristics or the fab layout. To model operator's shift behavior dynamically, start and end shift time can be defined for each operator group with a total number of operators available. An unique operator group name is created for each operator group and shift period in order to differentiate the behavior of each operator group at different shift period. Break times can also be defined for each operator group with a certain break time behavior. Each equipment in the simulation model is assigned to one operator group with the calculated load time. When the equipment is ready for a lot or batch processing, it will consume one operator count of the assigned operator group for the duration of the defined load time. After loading, the operator will be added back to the corresponding operator group pool. Take note that it is possible that an operator resource is still involved in a task of processing a lot at the end of shift. In this case we assume that the operator will stay until the task is done beyond the shift. If there are not enough operators to perform the task, the equipment will incur efficiency loss or idle time, and the lot or batch will be waiting for the operator. When the cycle time of lots is very high at a particular equipment with idle time, then it is likely that operator capacity is insufficient. See Figure 1 for the illustration of time elements with operator modelling. The loading time for the operator to load the lot into

the equipment will reduce the throughput of an equipment. This means that the longer the load time is, the more tool efficiency will be lost. Combining high load time and low operator quantity, the throughput of an equipment will be significantly reduced.

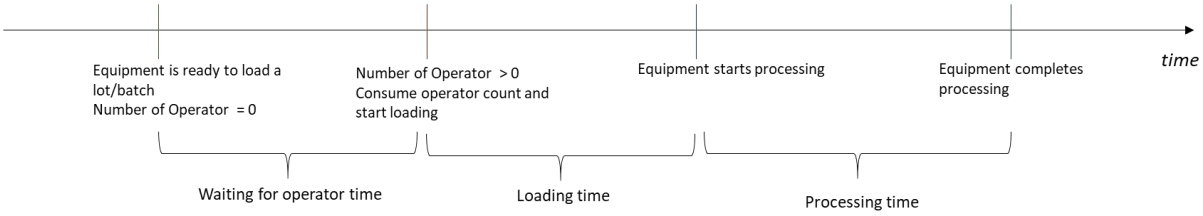

Figure 1: Operator model time element overview.

Table 1: Modelling elements and considerations for high accuracy forecast A.

Modelling Element	**Description**
Work-In-Progress (WIP)	The model is initialized with WIP in the production line, with all the lots and their associated information being captured. The essential lot information is its current step, current state: in queue, in process (current equipment and remaining processing time is required), in rework, or on-hold (estimated hold release time is required), priority, start and due date.
Initial Equipment Down	All equipment that are in down or non-productive state are initialized as down before simulation starts. An estimation of when the equipment is coming back online is required. This information is obtained from either historical data (average duration for the corresponding down type) or provided by the maintenance department.
Wafer Start Plan	A wafer start plan up to the lot level is required. Typically wafer fabs do not have lot level wafer start plan beyond a week. To address this constraint, a product level weekly volume start is obtained from the planning department, and a lot level wafer start plan is created. An algorithm of batching lots of the same product to start to enhance batching efficiency at furnaces is used for realistic wafer start plan generation.
Process Flows	All process flows required by the WIP and wafer start production lots are considered in the model. We do not choose representative process flows as we need to ensure lots are following the exact path that they will run in the reality.
Rework	Rework is modelled as a random event, where rework rate is derived from historical data for all production steps that could trigger a rework process.
Hold	Hold is modelled as a random event, where hold rate and hold duration distribution are derived from historical data for all production steps that could trigger lot hold.
Split-Merge	Some equipment type such as Chemical-Mechanical-Polishing (CMP) and Lithography require pilot runs from time to time. This is modelled with the split-merge function, where the split rate is calculated from historical data.
Dispatch Rules	Only global dispatch rules are considered in the model, such as lot priority, queue time priority, operation due date, maximum wait time and same setup. Some local dispatch rules such as prefer fast equipment were also considered.
Queue Time Constraint	Typically queue time constraints are controlled with KANBAN based dispatch rules. It is thus essential to construct a simulation model with such consideration as lots could be held back and not moving to the next step due to unavailability of KANBAN even though equipment capacity is available.

Table 2: Modelling elements and considerations for high accuracy forecast B.

Modelling Element	**Description**
Equipment	All equipment in the production line are considered. Each equipment is mapped based on its specific behavior such as: single lot, single wafer, batch, or cluster.
Dedication	Dedication is modeled at recipe and product-recipe combinations, depending on equipment type. Long term inhibits are also considered in the model to ensure constraints in production line are portrayed accurately in the simulation model.
Equipment Down	Equipment down is modelled as a random event, where the mean time to failure and mean time to repair distribution are derived from historical data.
Process Time and Throughput	Data for recipe or recipe-product based process times and throughput for each equipment are gathered. Process time is defined as the time duration that lots/batches are spending in the equipment, while throughput is defined as the rate at which lots/batches are processed by the tools. Cascading of lots/batches are thus modelled when the throughput is higher than the process time. Limping effects (losing process speed) of chamber down are also modelled.
Setup Switching	Setup switching is modelled at some of the relevant equipment such as implantation. We consider the time required to switch from one recipe class to another. This overhead is important to be modelled as it reduces the tool capacity.
Reticle	Reticles are modelled as an additional resource required before lots can start processing at lithography equipment. This is essential because lithography equipment are typically the key bottleneck of the production line, and reticle availability could alter the selection of lots for processing. discussed above

2.2 Wait for Load Estimation

There are a few papers that deal with detailed operator modeling for wafer fabs (Mosley, Teyner, and Uzsoy, 1998) (Crist, and Uzsoy 2011). However as human resources were normally not a critical limiting factor in Kulim's production line, they were not incorporated in the simulation model before the Covid-19 pandemic. As human resources limitations became the most critical bottleneck during this challenging time, it was thus essential to extend the model to include the human resources (operators) constraints. The most appropriate approach for us was to model the constraints as machine load time, which is defined as time needed for operator to find a suitable lot and load it into an equipment whenever the equipment is in idle mode. Before Covid-19 it was seldom the case that idling equipment waited a long time for loading when a suitable lot was available, because operator availability was high. The challenge now was, how to determine the average machine load time, and its relationship with the operator resources availability.

Time studies are done regularly in Kulim, where relevant data is collected. Some of this data can be used for our purpose, for example, data in how long it takes on average for an operator to search for a lot, load and unload equipment, and prepare lots for processing will be gathered in such time studies. This data will be gathered on work center level or sometimes even on equipment level and is generally used to calculate how many operator are needed to guarantee a good fab performance. We used this data to derive a load time static LT_{static} which can be interpreted as minimal loading time needed for a certain work center/equipment if operator are fully available. LT_{static} normally does not change over time.

Apart from that, timestamp data will be collected daily for the fab, for example the timestamp when a lot n is loaded to a certain equipment (*Lot movein time*$_n$) and the timestamp when equipment starts idling or is idling before processing lot n and a lot is available for processing or arrives for processing (*Equipment standby lot available*$_n$) will be stored in a database. A weekly average load time dynamic ($LT_{dynamic}$) on work center level can be calculated by using a formula (see Equation 1). $LT_{dynamic}$ can be interpreted as average load time performance for a certain work center/equipment and week. In Kulim $LT_{dynamic}$ should normally not be much higher than LT_{static}, assuming high operator availability. $LT_{dynamic}$ usually changes over time.

$$LT_{dynamic} = \frac{1}{n} * \sum_{n} (Lot\ movein\ time_n - Equipment\ standby\ lot\ available_n),$$

$$where\ we\ sum\ up\ over\ all\ lots\ n\ for\ week\ i\ and\ work\ center\ x \quad (1)$$

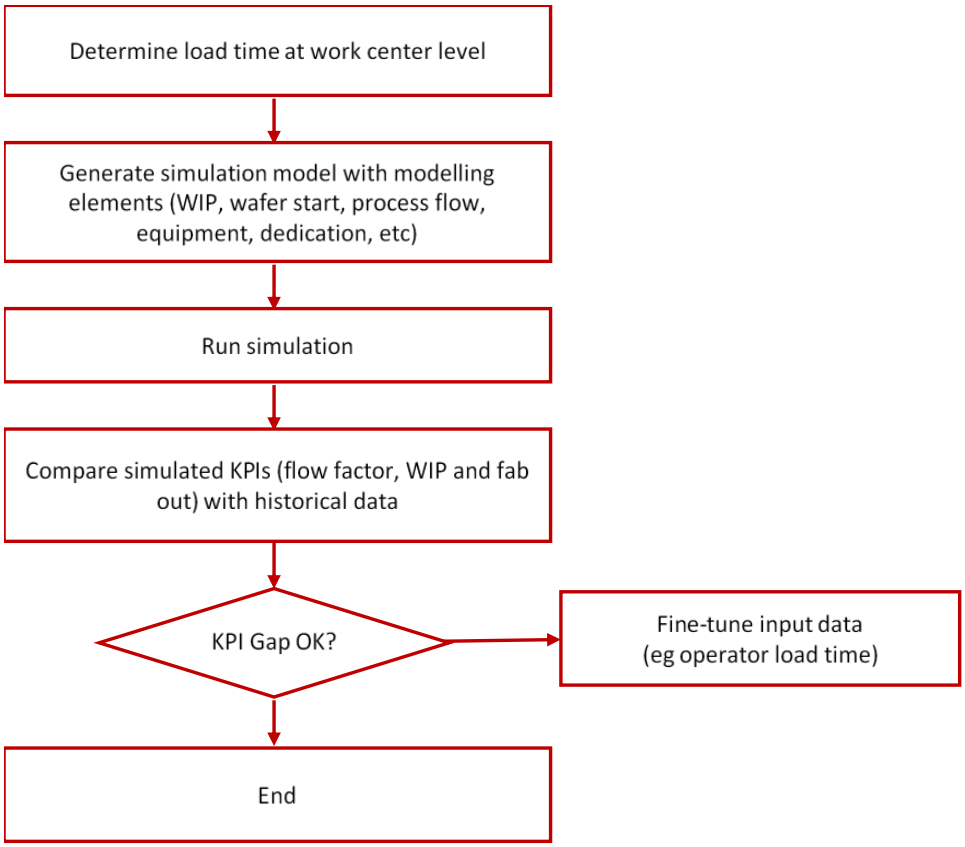

Figure 2: Methodology flow chart.

Historical $LT_{dynamic}$ data from one year and LT_{static} data were used to generate different machine load times scenarios. Simulation runs in Figure **2** were conducted for a six-week time horizon for selected historical time periods. From the simulation results and by comparing them to historical fab performance, we were able to come up with a formula that shows a correlation between load time (LT_t), $LT_{dynamic}$, LT_{static}, and operator availability (OP_t), all for the same time period t (see equation (2)). This was mainly possible because historical data was available for time periods where operator availability was limited, e.g. for a typical holiday week in a seasonal holiday period and for the beginning of the lockdown period (see Figure 2). With this data it was possible to derive weights w_i and w_j accordingly. The weights can be different for different work center. Therefore equation 2 is work center specific.

With the help of the formula, average load times for different scenarios (see Section 3) were estimated on work center level, and were used as input for simulation runs.

$$LT_t = \frac{w_i * \mathrm{OP_t}^{*}\mathrm{LT_{static}} + w_j * (1 - \mathrm{OP_t})^{*}\mathrm{LT_{dynamic}}}{2} \qquad (2)$$

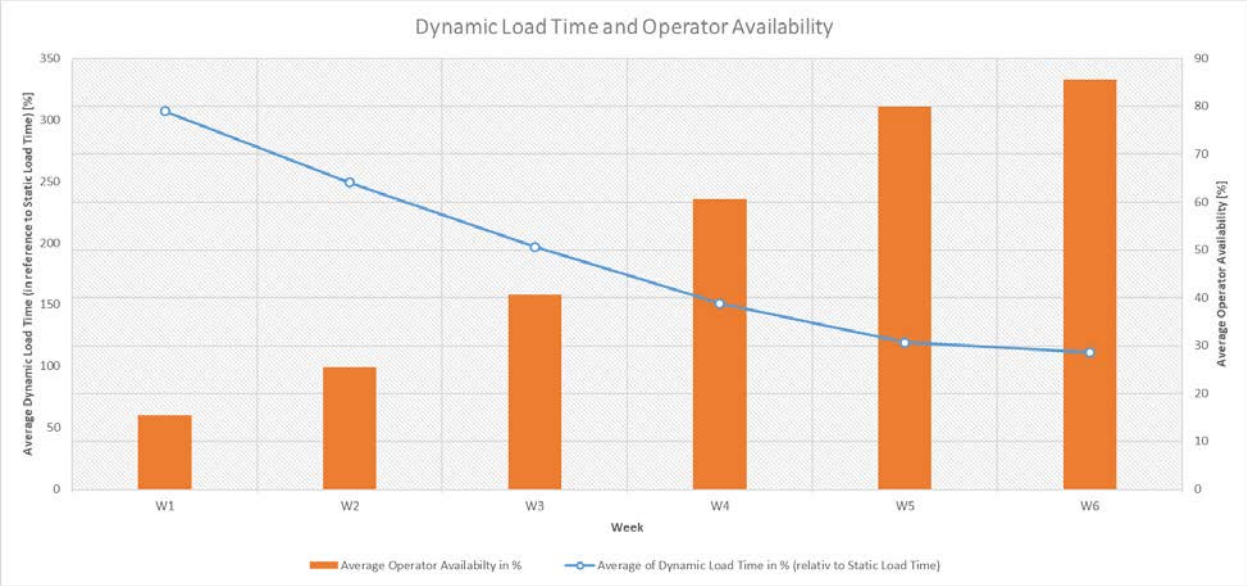

Figure 3: Correlation dynamic load time and operator availability on fab level during start of MCO period**.**

3 EXPERIMENTAL RESULTS

Before we can use the model for any meaningful study, we need to ensure that the operator model input data is a good representation of the reality. We fine-tuned the parameters with the first week forecast of the lockdown with the methodology illustrated in Figure 2 in the second week of lockdown. Based on this methodology of model calibration, we derived equation (2) as an approach to establish the relationship between load time and operator availability. The simulation runs were conducted with six weeks of forecast, and at least 10 replications for each simulation study. Figure 4 shows the overall fab performance key figures from fab simulation for the six weeks after lockdown came into effect. The values in Figure 4 were all shown relative to a baseline due to the confidentiality of the data. The baseline was a simulation with the originally planned fab loading and 100% operator availability, a situation we expected highly likely without Covid-19 pandemic.

Having established the relationship between load time and operator availability, we conducted more studies to predict the fab performance for 3 different scenarios that were likely to happen. These scenarios are summarized in Table 3. The simulation results helped the management to understand possible consequences of manpower shortage to the loss in fab level performance. One critical KPI that is of great interest for Infineon is the fab flow factor. The fab flow factor provides an indication of the production speed. The lower the fab flow factor, the better the fab performance is. It is measured as a ratio of the fab cycle time (of a product or lot) and total raw process time. The total raw process time of a product is defined

as best possible cycle time of a product, assuming no waiting times in front of equipment, optimal transport times between processes and no other disturbances.

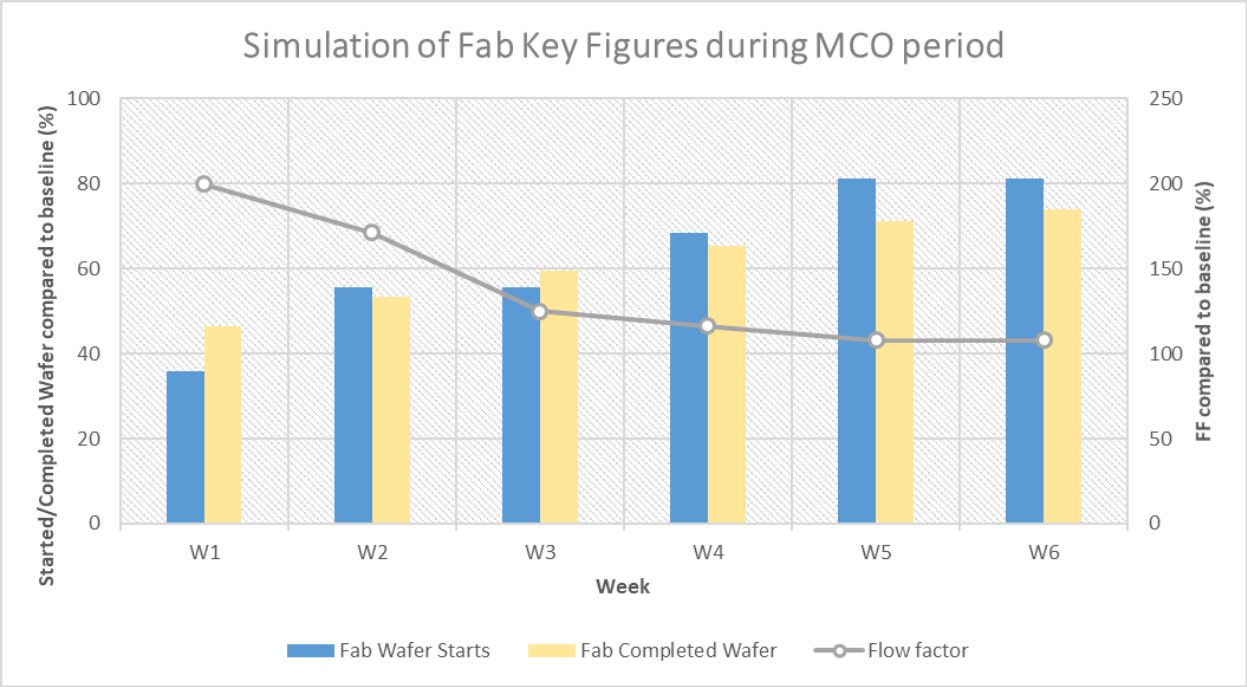

Figure 4: Simulation results for fab key figures during start of MCO period.

Table 3: Scenario study overview.

Scenario	**Impacted Area**	**Impact on Operator Availability**	**Estimated Impact**
Worst Case	Whole Fab	> 15% (Major impact by 2 shift operator absent)	Flow factor spike of 700% compared to baseline, recovery takes more than 15 weeks
Most Likely	Whole Fab	5% - 10% (Major impact by 2 shift of operator absent)	Flow factor spike of 300% compared to baseline, recovery takes approximately 10 weeks.
Best Case	Only specific areas of Fab	< 5% (Minor impact to operator attendances)	Flow factor spike of 140% compared to baseline, recover after 5 weeks.

For the worst case scenario, it was assumed that more than 15% of the operators working in the fab would be required to be quarantined for at least 2 weeks due to exposure to infected colleague, the fab would have to be closed for 2 shifts (24 hours) and a disinfection of the whole fab would be done. For the most likely scenario a similar fab shutdown like the worst case scenario was assumed but fewer operators (5% - 10%) required to quarantine. For the best case scenario we assumed that no fab shutdown would be necessary, and less than 5% of the operator would be required to be quarantined with the impact being limited to some fab areas only.

In Figure 5 the fab flow factor (FF) is shown for the three different scenarios. FF is shown relative to a baseline scenario where no operator limitations were assumed. For the best case scenario, simulation showed a recovery to baseline behaviour within 5 weeks. For the most likely scenario and the worst case scenario, FF was back to normal after approximately 10 and 15 weeks, respectively.

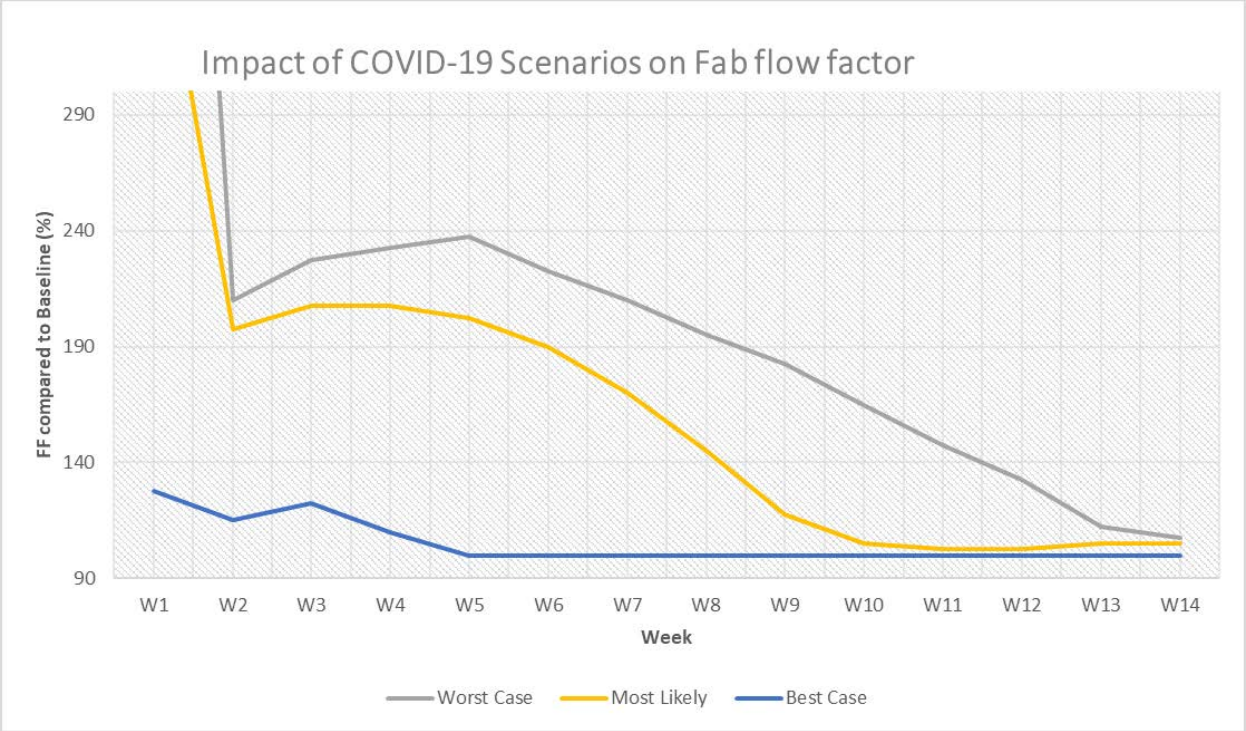

Figure 5: Flow factor forecast for scenarios.

The management at that point in time was looking for trustable forecasts to mitigate the risk of the loss to the company. The simulation results helped the management to understand possible consequences of manpower shortage to the loss in fab level performance. Different action plans were devised to reduce the impact of the operators resources constraints, such as operator cross training from different modules, increase of overtime incentives and investment in anti-microbial coating at common areas such as restaurant and cafeteria, handrails, lift areas, operator hostels, toilets and etc.

4 CONCLUSION

The Covid-19 pandemic created a totally new, unknown situation for the semiconductor industry. An existing well validated fab simulation model was adapted to determine possible impacts out of Covid-19 restrictions and to predict key fab performance indicators for different scenarios. This helped operation to quantify risks and to prepare mitigation strategies and the support was well received in these challenging times. The simulation model changes will be kept because simulation accuracy generally increased.

REFERENCES

D-SIMLAB Technologies. 2021. Forecaster and Scenario Manager. http://www.d-simlab.com/category/d-simcon/products-d-simcon/forecaster-and-scenario-manager, accessed 2nd April 2021.

Fowler, J.W., L. Mönch, and T. Ponsignon. 2015. *Discrete-event Simulation for Semiconductor Wafer Fabrication Facilities: A Tutorial*. International Journal of Industry Engineering: Theory, Applications, and Practice, Page 661-682.

Mönch, L., J.W. Fowler, and S.J. Mason. 2013. *Production Planning and Control for Semiconductor Wafer Fabrication Facilities*. Operation Research/Computer Science Interface Series. Vol. 52. Springer Science+ Business Media New York.

Seidel, G., C.F. Lee, A.M. Kam, B.P. Gan, C.W. Chan, 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, edited by Chan, W.K.V.; D'Ambrogio, A.; Zacharewicz, G.; Mustafee, N.; Wainer, G.; and Page, E., (editor(s), pages 3670-3678. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Mosley, S. A., Teyner, T., Uzsoy, R. M. (1998). Maintenance Scheduling and Staffing Policies in a Wafer Fabrication Facility. IEEE Transactions on Semiconductor Manufacturing, 11(2):316-323.

Crist, K., Uzsoy, R. (2011). Prioritising Production and Engineering Lots in Wafer Fabrication Facilities: a Simulation Study. International Journal of Production Research, 49(11):3105-3125.

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 Proceedings of the 2020 Winter Simulation Conference, edited by Bae, K; Feng, B.; Kim, S.; Lazarova-Molnar, S.; Zheng, Z.; Roeder, T.; and Thiesing , R., editor(s), pages 1777-1788. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

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.

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.

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.

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.

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.