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.

INFLUENCE OF SPARE PARTS SERVICE MEASURES ON THE PERFORMANCE OF FRONT-END WAFER PRODUCTION PROCESS

Douniel Lamghari-Idrissi
Daniel Soellaart
Rob Basten
Nico Dellaert

School of Industrial Engineering
Eindhoven University of Technology
P.O. Box 513
Eindhoven, 5600MB, THE NETHERLANDS

ABSTRACT

We are interested in the influence of spare part service measures on the performance of front-end wafer fabrication process. This process is characterized by re-entrant flows exacerbating variability differences. We focus on the bottleneck resource. First, we simulate the spare part supply chain to show the impact of the spare part service measures on the time to repair distribution. Second, we use this distribution to assess the performance of the front-end wafer fabrication process. We conclude that the choice of the spare parts service measure has a high impact on the front-end wafer fabrication process performance. Our methodology could help practitioners making improved decisions regarding spare parts service measure.

1 INTRODUCTION

We study the impact of spare parts service measures on the performance of remote front-end wafer fabrication process. The front-end fabrication of semiconductors is an example of a high-volume industry with a continuous production process (Hopp and Spearman 2011). In order to produce a semiconductor, Mönch et al. (2018) identify four different phases in the production of wafers, i.e., building layers, electrical probing, assembly and testing. We focus on the first phase, since, within the semiconductor manufacturing process, building layers is the most costly and time-consuming process step according to Gupta et al. (2006). They characterize this process as having a large number of production steps, between 600 and 800, re-entrance of work in process components for the multiple layers (recirculating flows), random equipment failures, sequence-dependent tool set ups, the need for batch processing tools and expensive capacity expansion.

The capital goods that are used in these factories are characterized as high-tech, complex and expensive. A wafer production facility costs several billions of dollars. Therefore, manufacturers expect to use their assets for a long period of time and require high system availability. System availability, defined as the percentage of time a system operates (uptime) in a certain time interval, is of crucial importance to produce cost-effectively (Smets et al. 2012). When a system failure occurs, the production stagnates and high costs are incurred due to, for example, reduced production output. The time a system is not producing is called downtime, i.e., the time to repair. The time to repair can be decomposed in four steps namely, diagnosing the problem, waiting for spare parts, swapping the defective component by a functioning one, and the recovery sequence. Each of these steps can take a different amount of time, creating variation in the time to repair. Therefore, production managers use work in progress inventory (WIP) to limit the risk of starvation (Hopp and Spearman 2011). However, WIP is costly and results in deterioration of the

978-1-7281-3283-9/19/$31.00 ©2019 IEEE

production performance, higher cycle time for example. To find the optimal WIP level, fab managers require an accurate prediction of the machines' downtime and, therefore, the time to repair. This is even more relevant when considering the bottleneck resource. For example, in a front-end fab, the lithography equipment is the bottleneck. The costs associated with the downtime of the lithography machine are in the range of 20 euros per second.

To reach the goal of high availability, original equipment manufacturers (OEMs) provide the option to maintain the capital good in exchange of a certain fee through a service contract. Service contracts generally contain a section on spare parts availability, which is the focus of this paper. Not meeting the commitment set in the service contract can result in high penalty costs for the OEMs. To ensure high availability and avoid penalty costs, the OEMs stock spare parts close to the fab and optimize their stock in order to provide an affordable service. Two types of spare parts service measures are commonly used: *aggregate fill rate* and *aggregate mean waiting time* (Basten and van Houtum 2014). The aggregate fill rate is defined as the probability that an arbitrary demand for the total group of stock keeping units (SKUs) is fulfilled immediately. Under an *aggregate fill rate* commitment, there is no commitment from the service provider on how long the customer has to wait in case the component is not delivered within the agreed time. The *aggregate mean waiting time* commitment is a step in that direction. It is defined as the expected waiting time until an arbitrary spare part demand is fulfilled. Under an *aggregate mean waiting time* commitment, the service provider (the OEM or a third party) can compensate a very slow delivery by multiple fast ones. This is possible since there is a commitment on the *mean waiting time*. Beyond the *mean*, a spare part service measure influences the *time to repair distribution*. Current service measures do not take this distribution into account, at least not directly.

The situation described in the previous paragraphs constitutes a double buffer. On one hand, the OEM stocks spare parts to decrease the impact of a machine failure. On the other hand, the customer carries WIP to limit the risk of starvation. To the best of our knowledge, only Kiesmüller and Zimmermann (2018) investigate the relationship between spare parts provisioning and the performance of a production system. They refine the definition of starvation as follows: "when a WIP buffer before a machine is exhausted, and there are no spare parts available, the machine cannot produce and consequently stands still". To gain analytical insights, they focus on a model with two workstations and one spare part. We extend their work, using simulation, to study a more realistic production system and inventory model whereby multiple spare part service measures are taken into account. Furthermore, next to the WIP level, we also investigate other performance measures, namely cycle time and throughput. Another performance measure in the semiconductor industry is the yield. In this paper, we do not take it into account since spare part service measures do not directly influence it.

We are interested in understanding the influence of spare part service measures on the performance of a front-end semiconductor production system. Since the photo-lithography is the bottleneck in the front-end process, it is the focus of this study. Our contribution consists of the definition of a methodology to study this question, setting corresponding values between the service measures, assessing the impact on the *time to repair distribution* and quantifying the impact on the performance measures of a production system. Our results show that the choice of the service measure has high impact on the front-end wafer fabrication process performance. In the studied case, an *aggregate mean waiting time* commitment delivers, on average, 15% more availability, a quarter of the cycle time, a third of the WIP and 20% more throughput than an *aggregate fill rate* commitment.

Our paper is organized as follows. In Section 2, we introduce the various models including the assumptions. Section 3 presents the results of the simulations and provides the key managerial insights. We conclude in Section 4.

2 MODELS

To study the impact of different service measures on the performance of a front-end wafer fab, multiple links need to be established. In Section 2.1, we optimize the base stock levels for each SKU, at each warehouse and for each service measure. A translation between *aggregate fill rate* and *aggregate mean waiting time* is established to ensure the viability of the comparison. Using the base stock levels corresponding to each service measure and using simulation we derive, in Section 2.2, the *time to repair distribution* associated with each service measure. Section 2.3 covers the repair process of the bottleneck machine, linking the spare parts side with the front-end wafer fab side. Section 2.4 focuses on the front-end wafer fab.

2.1 Base Stock Levels Optimization Model

We focus on a single type of machine for the bottleneck resource of a front-end wafer fab, the photo-lithography equipment. We consider only critical components for which a failure brings the machines to a halt. Upon failure, the defective component is swapped and sent to the upstream supply chain for repair. These different SKUs are used in a mathematical model that consists of four parts, the supply network, the costs, the service measures and the stock levels. To stay close to practice, we consider a multi-location multi-item model with lateral transshipments representing the network of a photo-lithography equipment supplier, illustrated in Figure 1.

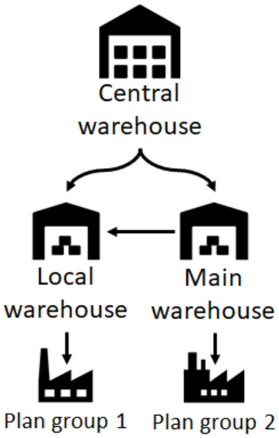

Figure 1: Spare pare network.

We set the number of machines supported by each warehouse at fifteen. Demand for a component arrives according to a Poisson process. Once demand from plan group 1 occurs, it is first satisfied by the local warehouse. In case of stock out, a lateral transshipment is triggered from the main warehouse. If no spare part is available, the central warehouse supplies the spare part directly to the local warehouse via an emergency shipment. For plan group 2, the demand can only be fulfilled by either the main or the central warehouse. We assume that both replenishment and emergency shipments are supplied by the central warehouse where we assume ample stock. The shipment times from the central warehouse are the same for both the local and the main warehouses. Each warehouse uses an $(S-1,S)$ base stock inventory control policy and fulfils demand on a first come first served basis. We assume that the warehouses are either very close to or even inside the front-end wafer fab and therefore, we do not consider the shipment time from the warehouse to the customer. Replenishment, lateral and emergency shipment times are deterministic following the general literature (Basten and van Houtum 2014). As highlighted in Section 1, under an aggregate fill rate commitment, the OEM has no obligation to deliver as fast as possible a component that is not locally available. Therefore, when under an *aggregate fill rate commitment*, we assume the emergency shipment time to be equal to the replenishment shipment time. Since we use emergency shipments if demand cannot be fulfilled from stock, we have an Erlang loss system, i.e., an $M/G/c/c$ queue. We consider two scenarios: either both warehouses are under an *aggregate fill rate*, FR scenario, or under an *aggregate mean waiting time*, WT scenario. For the *aggregate fill rate* and the *aggregate mean waiting time*, we follow Van Houtum and Kranenburg (2015). The costs are composed of the holding costs, lateral transshipment costs and emergency shipment costs. Being always incurred, we do not consider the ordering costs and the replenishment shipment costs. The objective of the optimization model is to set the base stock levels at the level that minimizes the total costs while meeting the service measure target. We use the optimization procedure described in Van Aspert (2015), who gives a mathematical model for a complete spare part network.

2.2 Spare Parts Simulation Model

In the simulation model, we use the averages commonly observed in practice. For all warehouses and all components, we set the replenishment shipment time at fourteen days, the lateral transshipment time at 12 hours and the emergency shipment time at 72 hours for the aggregate mean waiting time commitment and equal to the replenishment lead time for the aggregate fill rate commitment. Setting the emergency lead time equal to the replenishment lead time is in line with the aggregate fill rate not committing to a delivery time in case of a lost sale. Additionally, both warehouses support fifteen machines. If the main warehouse is under an *aggregate mean waiting time* commitment, we set the target at 1%. In case it is under an *aggregate fill rate* commitment, we set the target at 95%. We normalize the ASML costs by setting the emergency costs at 100. The lateral transshipment costs are set at 42. The holding cost rate is set at 17% and is composed of the average cost of capital and warehousing costs.

The simulation model is a discrete-event based simulation. In order to provide reliable results, we first define the warm-up time to avoid the start-up problem described by Law (2007) and follow his approach. We start by making r replications, where $r \geq 5$, with a relatively large length l. This leads to the output Y_{pq} where p is the number of observation and q the number of replication. We then create an average process, $\bar{Y}_p = \sum_{q=1}^{r} Y_{pq}/r$. Let u denote the number of time points used to calculate an average. u is calculated following Law (2007). We set $r = 5$ and $l = 10$ years to investigate when the SKU stock at the local warehouse reaches steady state. As can be seen in Figure 2, after 30 weeks the line is stabilized, representing the end of the warm-up time.

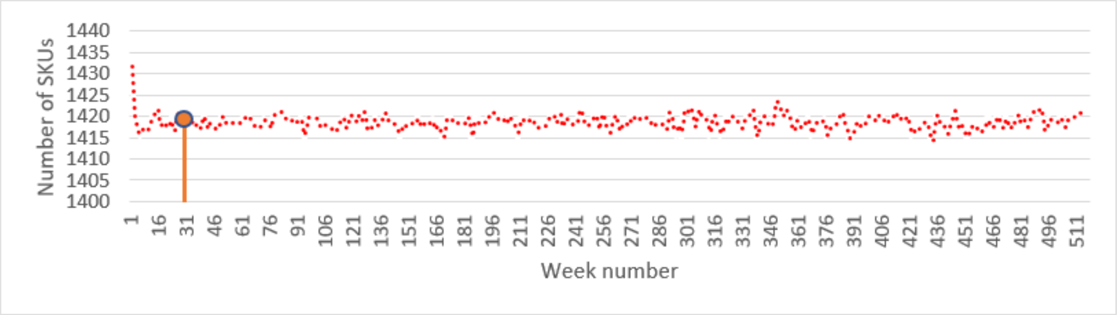

Figure 2: Warm up time for the spare part network simulation.

Using the methodology of Byrne (2013) which is based on confidence intervals, we set at 101 the number of replications, or sub-runs, needed to have 95% of the outcomes within plus minus one percent of the *mean* value. Finally, we validate our model using the approach proposed by Sargent (2011). Table 1 gives an overview of the first validation step, the comparison to other models; in our case this means comparing the result of the simulation with the result of the optimization model. The maximum deviation is of one percent. This is only due to a specific case where, when the demand for a specific SKU arrives, the warehouse has no part on stock, and an emergency shipment is triggered while a replenishment shipment is on its way. In this special case, it can happen that the replenishment order arrives before the emergency order. The demand is then fulfilled by the replenishment order and the stock is replenished by the emergency one. This leads to a lower *mean waiting time* than planned for in the optimization model.

Table 1: Comparison table used for validation.

Scenario	**Service measure**	**Optimization**	**Simulation**	**Average relative difference**
WT	Agg. *mean* waiting time local	1.00%	1.00%	0.00%
WT	Agg. *mean* waiting time main	1.00%	1.01%	1.00%
FR	Agg. fill rate local	93.13%	92.30%	0.89%
FR	Agg. fill rate main	96.04%	96.09%	0.05%

Next, we perform an extreme condition test by using high base stock levels of ten units for each components at each warehouse. As can be seen in Table 2, the simulation performs as expected since the *aggregate fill rate* and *aggregate mean waiting time* reached are extremely high and low, respectively. The above results validate our simulation model.

Table 2: Results of the extreme condition test.

Aggregate fill rate	100%
Aggregate *mean* waiting time	0.37%
Number of emergency shipments	0
Number of lateral transshipment	0

2.3 Repair Process

For the spare part simulation model and the front-end fab simulation model to be linked, we analytically model the repair process of the bottleneck resource since this is the focus of the paper. As described earlier, due to the high utilization of the bottleneck resource, we can assume that the machine is idle only when it is broken. This is also what can be observed in practice where preventive maintenance is performed in the shadow of corrective maintenance. This is a common practice done to increase the overall system availability. As described in Section 2.1, we consider fifteen machines. Additionally, we set at four the number machines that can be repaired at the same time. We model the repair process as a continuous Markov chain with the number of machines down $b \in \{0,...,15\}$ as the state space. Let λ denote the *mean time to failure* and μ the *mean time to repair*. Figure 3 illustrates the flow model of this process and Q is the transition matrix.

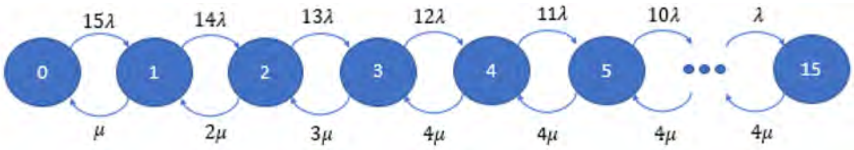

Figure 3: Flow diagram for the repair process.

$$Q = \begin{pmatrix} -15\lambda & 15\lambda & \dots & & & \\ \mu & -(\mu+14\lambda) & 14\lambda & \dots & & \\ \vdots & 2\mu & -(2\mu+13\lambda) & \dots & & \\ & \vdots & \vdots & \ddots & & \\ & & & & 4\mu & -(4\mu+\lambda) \end{pmatrix}$$

We assume that λ and μ are exponentially distributed. These assumptions are in line with general literature and were validated through interviews with practitioners. We can model the repair process as an $M/M/c$ queueing model. For the model analysis of this queue, we follow Adan and Resing (2015). Furthermore, we analyse this system with a local balance equation to calculate the probabilities p_b to be in state b, i.e., the probability that b machines are down. Once this probability is known, we are able to calculate the average number of machines down.

2.4 Front-end Wafer Fabrication Simulation Model

To produce the multiple layers of a chip on a wafer, a multistep process is repeated multiple times, i.e., a re-entrant process. To produce a layer, seven key process steps are required (Gkorou et al. 2017), deposition, photoresist coating, exposure, developing, etching, ion implantation and stripping. Due to the complexity, simulation is the most common modelling technique in the semiconductor industry (Mönch et al. 2018). We use a discrete-event simulation model to analyse the front-end wafer fabrication process as a whole. In a paper on cycle-time improvement, Akcalt et al. (2001) describe a scaled down representation of the production line including production and set up times. Our simulation model is based on their paper and use the same assumptions and values. The goal of our model is to assess how various performance measures are impacted by different time to repair *distribution*s. The performance measures we are interested in, are the average WIP level, the average cycle time, the throughput, the utilization of the bottleneck resource and the availability of the machines. Often, multiple wafer types move through the wafer fabrication process in a batch. Wafer types differ in the number of arrivals per hour and the number of production rounds. We assume that each wafer type has the same fixed route. The route is predetermined and consists of multiple workstations as described earlier. We assume that the exposure step is the bottleneck as is commonly the case in practice. According to Lin and Lee (2001), the number of operations before the bottleneck resource is relatively low. As a result, we assume that the bottleneck resource is the start of the process loop. Figure 4 gives a visual overview of the wafer production process as we model it.

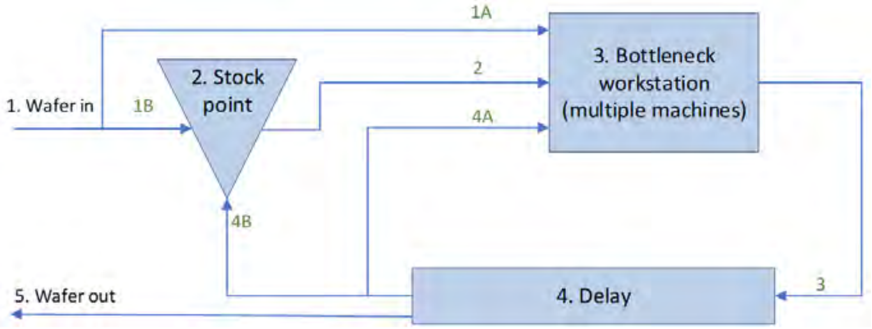

Figure 4: Overview of the modelled wafer fabrication process.

The inflow of wafers being planned, we assume the batch arrival process to be deterministic. Following Akcalt et al. (2001) and practice, we assume a dedicated assignment policy, balancing set up time and processed wafer stock levels. Therefore, a new batch can either be directly assigned to a workstation (1A) or be put on stock (1B). We assume that the stock point follows a First Come First Serve policy with preemption, due to the dedicated assignment policy, and has infinite capacity. We assume the production process to be balanced, i.e., the amount of wafers arriving in the system (sum of 1A, 2 and 4A) is strictly less than the maximum output. Similarly to Section 2.2, we set the number of bottleneck machines at fifteen and assume the processing times as deterministic. The input values of our simulation model, including the processing times, are based on the test bed of Akcalt et al. (2001). Each wafer type has its own characteristics and input values given in Table 3. The *mean processing time* for a bottleneck machine is set at 1.1 hours per batch and 13.4 hours per batch for the delay process, giving us a maximum capacity of 13.5 batches per hour. We assume a target utilization of 80%. We consider only critical component failures. Upon failure, we assume the diagnostic, swapping and recovery to be deterministic with only the spare parts supply being stochastic. The failure rate and time to repair are from Section 2.2. We assume that the number of machines that can be repaired at the same time is fixed, at four. When a wafer is finished at the bottleneck machine, it is sent to the delay process. The delay process consists of all non-bottleneck processes, i.e., developing, etching, ion-implantation and stripping. We assume, for these stations, deterministic processing times and parallel identical servers with an infinite size buffer. The transportation times between workstations are not modelled. Upon completion of the delay process, the wafer can be in three different states. If the wafer is not yet complete and a bottleneck machine, set up for this wafer type, is available, the wafer will re-enter the bottleneck workstation (4A). If the wafer is not complete but there is no bottleneck machine, set up for this wafer type, available, the incomplete wafer is transferred to the stock point (4B). Last, if the wafer has completed all production loops then all layers are complete and it leaves the system (5).

Table 3: Characteristics of the different wafer types.

Wafer - Type	**A**	**B**	**C**	**D**	**E**
Batch size (units)	25	25	25	25	25
Number of required production rounds (loops)	21	19	20	19	21
Assigned number of machines (number of machines)	4	2	3	2	4
Set up time (hrs)	0.02	0.01	0.02	0.01	0.01

Following the methodologies described in Section 2.2, we assess the warm up time for the front-end fab simulation model, the number of replications and then proceed with the validation of our simulation model. 500 hours of the simulated production are needed for the warm up in order to eliminate the initial bias. This number is set using the results of the analysis described by Law (2007). The results are shown in Figure 5.

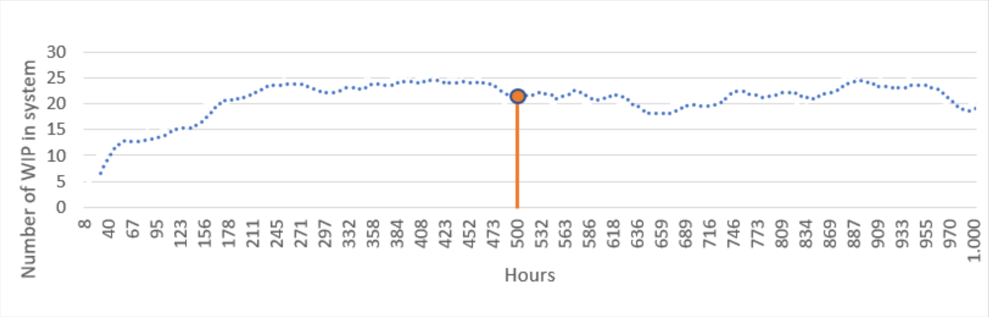

Figure 5: Warm up time for the front-end wafer fab simulation.

Following the methodology of Byrne (2013), we conclude that the minimum number of replications required is 46. We validate our simulation model using the extreme condition test and face validity. For the former, we input the arrival of all wafer types at ten batches per hour. The results are summarized in Table 4. The utilization is almost equal to the availability. This is in line with what one would expect since, the process being overloaded with wafers, the utilization is limited only by the time to repair of machines.

Table 4: Results of the extreme condition test.

Average utilization	81%
Average WIP (units)	$9.18e+04$
Average throughput (units)	0.53
Average cycle time (hours)	$1.69e+05$
Average availability	81%

Moreover, we note that the average WIP and cycle time are extremely high. Figure 6 shows that the WIP level in the system is increasing over time since the line is unbalanced. This was what we expected, so we conclude that the simulation model works well.

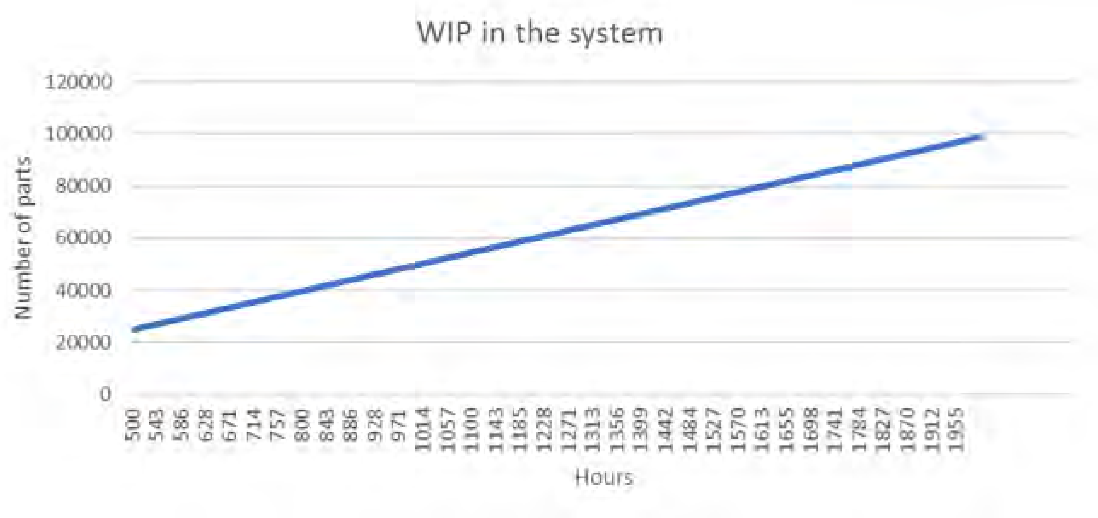

Figure 6: WIP level evolution during the extreme condition test.

3 RESULTS

Section 3.1 provides the different results needed to quantify the impact of the service measures on the spare part *waiting time distribution*. Section 3.2 uses these results to assess the impact on the fab performance.

3.1 Spare Parts Simulation

We, first, use the output of the spare part optimization model to set the base stock levels for each scenario. To find a translation between aggregate *mean* waiting time and aggregate fill rate, we start with WT and set the *aggregate mean waiting time* target at 1%. Using the optimization model of Section 2.1, we set the base stock levels. As discussed in Section 1, the *aggregate fill rate* varies from the aggregate *mean* waiting time by the way stock outs from the local warehouse are fulfilled. In the former, a replenishment shipment takes place as opposed to a lateral transshipment or an emergency shipment for the latter. Building on this difference and using the base stock levels set for the *aggregate mean waiting time* target of 1%, we can calculate the corresponding aggregate fill rate level. The stock outs being fulfilled via a replenishment order, we are then able to calculate the corresponding *aggregate mean waiting time* for scenario FR. This being done using an optimization model, we provide in Table 5 the translation between service measures as well as the costs linked to each scenario. The costs are normalized on the most expensive scenario. In line with intuition, the FR scenario has a higher *aggregate mean waiting time* due to how stock outs are fulfilled while being less expensive.

Table 5: Corresponding service measure levels and associated costs.

Scenario	**Local aggregate *mean* waiting time**	**Local aggregate fill rate**	**Total costs**
WT	1%	94.0%	100.0
FR	1%	95.0%	96.7

The different scenarios lead to different spare part waiting time *distributions*. Table 6 shows the normalized results of the simulation where the mean is set at 1%, in line with the *aggregate mean waiting time* target. The results show that, when the local warehouse is under an *aggregate mean waiting time* commitment, the coefficient of variation is much lower than when under an *aggregate fill rate* commitment.

Table 6: *Time to repair distribution* parameters per scenario.

Scenario	**Mean**	**Standard deviation**	**Coefficient of variation**
WT	**1%**	**5%**	**0.23**
FR	14%	79%	2.24

We conclude that an *aggregate mean waiting time* for both the local and the main warehouse, scenario WT, leads to the lowest coefficient of variation. The higher costs is a direct consequence of the additional commitment on the *mean*.

3.2 Front-end Wafer Fab Simulation

We start by analysing the impact of the different scenarios on the repair process of the bottleneck resource. The result is summarized in Table 7. As a result of the lower coefficient of variation of the time to repair, the average number of machines being down is lower when the local and the main warehouses are under an *aggregate mean waiting time* commitment. Consequently, the average availability is higher. Using these results and keeping the line balanced, we are able to calculate the number of arrivals per hour, based on the production rate of the bottleneck workstation. This is done using the Excel solver. The results are summarized in Table 8. Following the results from the repair process analysis, the production rate of the bottleneck station is the highest when under an *aggregate mean waiting time* commitment.

Table 7: Impact on the repair process.

Scenario	**WT**	**FR**
Probability of zero machines down	**5%**	2%
Mean number of machine down	**3.0**	5.1
Probability of four machines down	**34%**	71%
Mean availability	**80%**	67%

Table 8: Throughput of the system per hour for each scenario.

Wafer type	Scenario WT	Scenario FR
A	0.095	0.079
B	0.116	0.094
C	0.073	0.064
D	0.116	0.095
E	0.132	0.101
Total	**0.532**	0.435

It is important to note that the throughput of the system is not the same as the one of the bottleneck workstation. The process by re-entrant, batches visit all workstations multiple times, twenty on average as described in Section 2.4. For example, for scenario WT, the throughput of the system is 0.532 batches per hour which means that the bottleneck resource will process $20 * 0.532 = 10.6$ batches per hour.

Using our front-end fab simulation model, we can now evaluate the impact of the different service measures on the performance of the wafer fabrication process. The results are summarized in Table 9.

Table 9: Performance of the fab for each scenario.

Scenario	Throughput	WIP	Cycle time	Availability	Utilization
	(batches/hr)	(batches)	(hours)		
WT	**0.532**	**27.83**	**52.18**	**81%**	**77%**
FR	0.435	92.06	215.19	67%	63%

Scenario WT performs better than the FR scenario on all performance measures. It is clear that the FR scenario is degrading considerably the performance of the front-end wafer fabrication process across all performance measures, for example, 20% less throughput, over three times more WIP and over four times higher cycle times.

3.3 Performance Under a High *Aggregate Fill Rate* Target

In this section, we are interested at higher *aggregate fill rate* commitment to study the impact on the previous results. We focus on a commitment of 98% as it is the highest seen in practice. Table 10 shows the normalized results of the simulation where the mean is set at 1%, in line with the *aggregate mean waiting time* target. The performance of the front-end fab improves significantly compared to a 95% *aggregate fill rate* commitment. Nevertheless, the impact of the 2% of orders that are not fulfilled directly from stock still results in a high coefficient of variation, i.e., over 1.33. As a consequence, the performance of the front-end wafer fab remains worse than when an *aggregate mean waiting time* commitment is used. These results confirm that an *aggregate mean waiting time* for both the local and the main warehouse is more beneficial for the studied setting.

Table 10: Performance of the fab for each scenario.

Scenario	Mean	Coefficient	Throughput	WIP	Cycle time	Availability	Utilization
		of variation	(batches/hr)	(batches)	(hours)		
WT	**1%**	**0.23**	**0.532**	**27.83**	**52.18**	**81%**	**77%**
FR 95%	11%	2.24	0.435	92.06	215.19	67%	63%
FR 98%	4%	1.46	0.522	52.39	100.36	77%	75%

4 CONCLUSION

Front-end fab managers strive for the best performance of their production processes due to the highly complex and expensive nature of the equipments used. In this paper, we focused on the front-end fab and its double buffering of spare parts at the OEM and WIP at the fab. We aimed at studying the impact of different spare parts service measures on the performance of the production process of a front-end wafer fab. To reach this goal, we used a combination of optimization and simulation models. For a multi-item, multi-location model, we optimized the base stock levels for each scenario. Using these base stock levels and a simulation model, we assessed the impact of each scenario on the *time to repair distribution*. We found that an *aggregate mean waiting time* commitment across the network gives the lowest coefficient of variation. Focusing on the bottleneck resource, the photo-lithography, the repair process was modelled analytically and the outcome used in a front-end fab simulation model to generate the sought after insights. In the studied settings, an *aggregate mean waiting time* commitment delivers, up to 15% more availability, 20% more throughput and a reduction of up to a quarter of the cycle time, a third of the WIP than an *aggregate fill rate* commitment. This shows that important fab performance gain can be reached when using our methodology as a decision tool for spare parts service measures.

Further work includes the addition of the extreme long down service constraint introduced by Lamghari-Idrissi et al. (2020). To the best of our knowledge, there exists no optimization method for this service constraint yet. Adding this service measure to our analysis would be relevant since this service measure aims at limiting the number of infrequent long downs, notorious for disturbing the performance of production processes and particularly relevant for remote front-end fabs.

REFERENCES

Adan, I., and J. Resing. 2015. *Queueing Systems*. Eindhoven: Department of Mathematics and Computing Science, Eindhoven University of Technology. https://www.win.tue.nl/~iadan/queueing.pdf.

Akcalt, E., K. Nemoto, and R. Uzsoy. 2001. "Cycle-Time Improvements For Photolithography Process In Semiconductor Manufacturing". *IEEE Transactions on Semiconductor Manufacturing* 14(1):48–56.

Basten, R. J. I., and G. J. van Houtum. 2014. "System-Oriented Inventory Models For Spare Parts". *Surveys in operations research and management science* 19(1):34–55.

Byrne, M. D. 2013. "How Many Times Should A Stochastic Model Be Run? An Approach Based On Confidence Intervals". In *Proceedings of the 12th International conference on cognitive modeling*. July 11th-14th, Ottawa, Canada, 445-450.

Gkorou, D., A. Ypma, G. Tsirogiannis, M. Giollo, D. Sonntag, G. Vinken, R. van Haren, R. J. van Wijk, J. Nije, and T. Hoogenboom. 2017. "Towards Big Data Visualization For Monitoring And Diagnostics Of High Volume Semiconductor Manufacturing". In *Proceedings of the Computing Frontiers Conference*. May 15th-17th, Siena, Italy, 338-342.

Gupta, J., R. Ruiz, J. Fowler, and S. Mason. 2006. "Operational Planning And Control Of Semiconductor Wafer Production". *Production Planning and Control* 17(7):639–647.

Hopp, W., and M. Spearman. 2011. *Factory Physics: Third Edition*. Boston: Waveland Press.

Kiesmüller, G. P., and J. Zimmermann. 2018. "The Influence Of Spare Parts Provisioning On Buffer Size In A Production System". *IISE Transactions* 50(5):367–380.

Lamghari-Idrissi, D., R. Basten, and G. J. van Houtum. 2020. "Spare Parts Inventory Control Under A Fixed-Term Contract With A Long-Down Constraint". *International Journal of Production Economics* 219:123 – 137.

Law, A. 2007. *Simulation Modeling and Analysis*. Chicago: McGraw-Hill Inc.

Lin, Y. H., and C. E. Lee. 2001. "A Total Standard WIP Estimation Method For Wafer Fabrication". *European Journal of Operational Research* 131(1):78–94.

Mönch, L., R. Uzsoy, and J. W. Fowler. 2018. "A Survey Of Semiconductor Supply Chain Models Part I: Semiconductor Supply Chains, Strategic Network Design, And Supply Chain Simulation". *International Journal of Production Research* 56(13):4524–4545.

Sargent, R. 2011. "Verification And Validation Of Simulation Models". *Engineering Management Review, IEEE* 37:166 – 183.

Smets, L. P. M., G. J. van Houtum, and F. Langerak. 2012. "Design for availability: A holistic approach to create value for manufacturers and customers of capital goods". *Journal of Systems Science and Systems Engineering* 21(4):403–421.

Van Aspert, Martijn 2015. "Design Of An Integrated Global Warehouse And Field Stock Planning Concept For Spare Parts". PDEng. thesis - Logistics design project at ASML - Confidential until 01-01-2020, School of Industrial Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands.

Van Houtum, G. J., and B. Kranenburg. 2015. *Spare Parts Inventory Control Under System Availability Constraints*. New York: Springer.

AUTHOR BIOGRAPHIES

DOUNIEL LAMGHARI-IDRISSI is a part-time Ph.D. candidate at the school of Industrial Engineering at Eindhoven University of Technology. His research interest is on service contracts, outcome economy and new technologies. Beside his research, he is program manager at ASML in customer supply chain management. His email address is d.p.t.lamghari-idrissi@tue.nl.

DANIEL SOELLAART was a Master's student at the school of Industrial Engineering at Eindhoven University of Technology. His Master thesis project, conducted at ASML, formed the basis for this paper. He graduated in January 2019 and started his career in the industry. His email address is d-soellaart@live.nl.

ROB BASTEN is an Associate Professor at the school of Industrial Engineering at Eindhoven University of Technology where he is primarily occupied with maintenance and service logistics and its interfaces. He is especially interested in using new technologies to improve after sales services. His email address is r.j.i.basten@tue.nl.

NICO DELLAERT is an Associate Professor at the school of Industrial Engineering at Eindhoven University of Technology. His research has always been related to quantitative modelling of business processes. Currently, his prime research interests are on the integration of capacity and production decisions. His email address is n.p.dellaert@tue.nl.