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.

CLOSING THE GAP: A DIGITAL TWIN AS A MECHANISM TO IMPROVE SPARE PARTS PLANNING PERFORMANCE

Joan Stip
Lois Aerts

ASML
De Run 6501
Veldhoven, 5504DR, THE NETHERLANDS

Geert-Jan van Houtum

Department of Industrial Engineering and
Innovation Sciences
Eindhoven University of Technology
P.O. Box 513
Eindhoven, 5600MB, THE NETHERLANDS

ABSTRACT

In order to meet service level agreements at minimal cost, Original Equipment Manufacturers (OEMs) use spare parts planning models to determine the optimal base stock levels at the warehouses in their service network. In practice, however, these optimized base stock levels result in a *realized performance* that deviates from the *expected performance*. Therefore, it is beneficial for these companies to evaluate the base stock levels in terms of service performance, inventory value, and costs. In order to measure this *planning performance*, we developed a *digital twin* that is able to measure the planning performance and identify root causes for the performance gap. Our digital twin helped ASML, an OEM in the semiconductor industry, to create a feedback loop between the spare parts planning model and its realized performance in practice, providing a mechanism to learn from past results and determine actions to close the gap between the expected and realized performance.

1 INTRODUCTION

Nowadays a lot of companies rely on equipment from original equipment manufacturers (OEMs) for the production of goods and services. In the semiconductor industry, ASML is the world's leading OEM of lithography systems that are an essential component in chip manufacturing. The lithography machines produced by ASML form the bottleneck workstation in the multi-stage production-inventory system of the chip manufacturing process of their customers. This means that downtime of ASML's lithography systems has a negative effect on the performance of downstream processes, leading to an overall reduced performance for the customers of ASML, potentially leading to high downtime costs. Accordingly, it can be stated that the service provided by ASML is critical to its customers. As a consequence, ASML has strict service contracts with its customers which include strict service level agreements (SLAs). These agreements guarantee a certain availability of an installed base. When ASML fails to meet the agreed upon target, a considerable amount of penalty costs must be paid to the customer. In order to be able to meet the SLAs, ASML manages a network of service engineers, service tools, and spare parts. This paper focuses on the SLAs for spare parts.

In order to meet the SLAs for spare parts, ASML keeps inventory of spare parts in warehouses across the globe, located close to its customers. The key challenge for ASML in managing its spare parts inventory is balancing the total costs and the value of service for the customer. While on the one hand holding inventory in the warehouses leads to costs (e.g. purchasing, handling, and warehousing costs), on the other hand, inventory facilitates service by enabling the provision of parts demand without the delay of production and transportation. To set the base stock levels for the service materials at the local warehouses,

ASML executes a planning model which is based on the work of van Houtum and Kranenburg (2015) and Lamghari-Idrissi et al. (2022).

The planning model which is used for evaluating and optimizing the base stock levels of all stock keeping units (SKUs), makes use of a set of assumptions as well as an estimate for the demand rate of the specific spare parts. This estimated demand rate can be different from the actual demand rate. Consequently, the realized planning will not always be optimal. Additionally, the algorithm assumes, among other things, that over an infinite time horizon, the SLAs will be met, however, the SLAs are reviewed monthly. Lugtigheid et al. (2007) note that a finite time horizon can have a considerable effect on the optimal policy. Infinite time horizon models can only be considered as an approximation for short-term operational decisions and their accuracy needs to be tested (Topan et al. 2020). Due to these modeling assumptions, uncertainties with regard to the demand of the SKUs, and other disturbances on operational level, the planning method used will not always lead to the desired performance for ASML. A gap occurs between the expected- and realized performance. Figure 1 provides an overview of the relevant performance gaps.

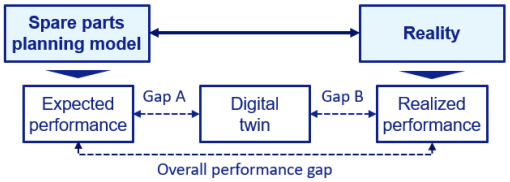

Figure 1: Overview of the relevant performance gaps.

As stated by Axsäter (2015), continuously evaluating the performance of an inventory control system is needed to achieve a cost reduction while still maintaining satisfactory customer service, which is in general the purpose of such a system. Currently, ASML faces challenges with regard to the evaluation of the base stock levels in terms of costs, service performance and inventory value, which should serve as a basis to identify areas of improvement. Due to the fact that the spare parts planning model is a simplification of reality and not all events occurring on operational level can be predicted beforehand, the expected performance in terms of service performance, costs, and inventory value deviates from the realized performance (the overall performance gap). In order to determine areas of improvement for the spare parts planning, we need to gain insights into performance gap A, as well as the impact of individual factors on this gap. Consequently, performance gap B consists of the factors occurring on operational level that cannot be predicted beforehand. Examples of such factors are the ordering behavior of engineers, and operational disturbances such as supplier capacity issues that will influence the current stock levels but cannot be resolved by changing the base stock levels.

We are interested in developing a simulation tool that is able to capture the dynamics of the service supply chain of ASML and is thereby a digital representation of the physical environment. In recent literature such a digital representation of a physical product or process is referred to as a "Digital Twin". Since the introduction of the concept in 2003, as later described in Grieves (2014), there has been a growing interest in the concept. This growth is largely driven by advances in related technologies and initiatives such as data management and processing, big data, and the Internet-of-Things. Originating from the field of product life-cycle management, where the digital twin is a virtual representation of a physical product, use cases nowadays have extended towards representations of processes in e.g. manufacturing (Kritzinger et al. 2018) and shop floor design (Guo et al. 2019). For a complete literature review, we would like to refer the reader to Jones et al. (2020).

The key concept of the Digital Twin as described by Grieves (2014) is that it consists of three components: A physical product or process, a virtual representation of the physical entity, and the bi-directional data

connections that feed data from the physical to the virtual representation, and vice-versa. In our research, the physical process is the spare parts supply chain of ASML, which is virtually represented in the simulation model. Data is flowing from the physical to the virtual representation (i.e. the state of the physical supply chain). The intended use of the digital twin is that data will also flow vice-versa, where the results of the simulation model will be used to reduce the overall performance gap. However, this is not yet implemented.

Our main contribution is in the fact that we developed a method to investigate performance gap A, using the concept of a digital twin. Furthermore, we show that using this digital twin, we are not only able to quantify the performance gap, but more importantly we are able to attribute this gap to certain root causes. The remainder of this paper is organized as follows. In Section 2 we introduce the digital twin. Section 3 presents a number of results obtained by using the digital twin for various use cases. Finally, we conclude in Section 4.

2 MODEL

2.1 Service Network

ASML operates a network of global and local warehouses which together form a multi-echelon network that allows for lateral transshipments (i.e. shipments between two nearby local warehouses) and emergency shipments. The service network of ASML is graphically presented in Figure 2, where the red arrows represent emergency shipments. The service network currently consists of two global replenishment warehouses and a set of local warehouses. The global warehouses serve as an inventory buffer between the suppliers and the local warehouses. Each local warehouse is associated with one or multiple plan groups, and each plan group is assigned to a single warehouse. Each plan group consists of multiple machines at customer sites which have the same SLA. Demand arrives according to a Poisson process to these plan groups for each SKU. The demand is satisfied directly from the local stock in case it is available at the dedicated warehouse of this plan group.

Depending on the type of contracts and the set SLAs, the tactical planners position the different SKUs at the different warehouses. For customers with low utilization fabs and less strict SLAs, more downtime of the system is acceptable. Therefore, in that case, the dedicated warehouse of ASML can be located further away from the customer's fab. However, for customers with a relatively high system utilization and strict SLAs, the transportation times of a spare part have a big impact on the up-time of the system. Consequently, ASML desires to shorten the delivery times effectively by operating local warehouses close to the fabs of these customers. Within this service network design the emphasis is on transportation times, where for every plan group a set of supporting local warehouses is defined that is capable of supplying within a pre-agreed time. These local warehouses together form a region, and support each other through lateral transshipments. A local warehouse can only be part of one region. In case the requested material is out of stock at the dedicated local warehouse, a lateral transshipment from another close-by local warehouse can be performed. In case the requested material is not available at one of the local warehouses within the region, it is checked whether the material can be delivered from another local warehouse or from one of the global warehouses through an emergency shipment. As a final option, the part can be delivered from one of the factories; this, however, would cause a disturbance to the production process of ASML and is therefore the least preferred option.

Since spare parts can only be used to fulfill demand once they are delivered to one of the global warehouses, we consider it relevant to track the location of an SKU in the digital twin from that moment onward. Accordingly, for SKUs that need to be converted into a service part in an ASML factory before they can be put on stock, we consider a supplier lead time. On top of the original supplier lead time (i.e. the time between the placement of the purchase order and the receipt of the part at the factory), this supplier lead time covers the time needed for conversion as well as the time needed to transport the converted part from the factory to the global warehouse.

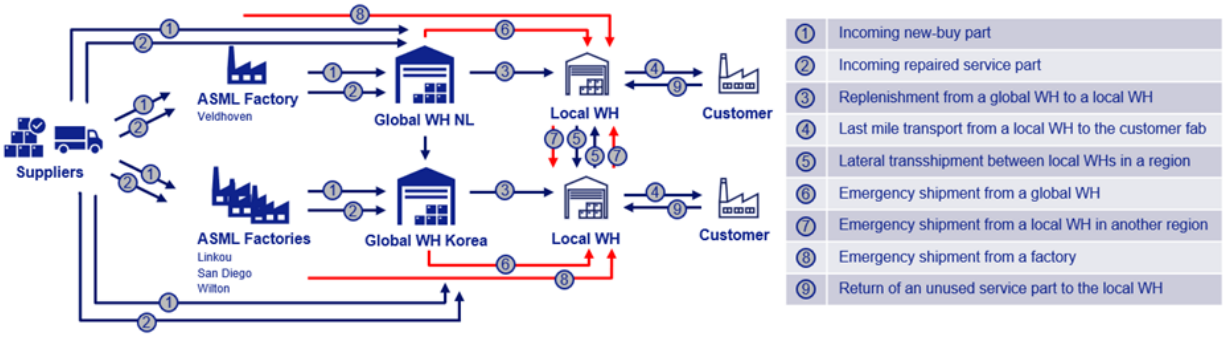

Figure 2: ASML's service network and its relevant flows.

Moreover, when the supplier of a part is able to repair a defect part, the part enters the repair flow. Due to the complexity of the repair flow within ASML, we excluded the repair flow from the digital twin. Nevertheless, the possibility that a part can be repaired does influence the stock levels, because a successfully repaired part will re-enter the service network. Consequently, the return of repaired parts is identified as one of the factors causing the expected- versus planning performance gap. Therefore, it is decided to separate the inflow from the supplier into a "regular" new-buy flow and a repair flow with a difference in lead time.

2.2 Base Model

In the base model of the digital twin, we only consider all local warehouses. The demand is satisfied directly from the local warehouse stock in case it is available at the dedicated warehouse of this plan group. In that case a last-mile shipment from the dedicated warehouse to the plan group is performed and the demand is fulfilled. In case it cannot be satisfied directly, we can use a lateral transshipment from another local warehouse in the network which is located in the same region as the dedicated warehouse. Here, the other local warehouses within the region are checked for stock in a pre-specified order. In case the local warehouses in the region do not have stock of the required SKU, an emergency shipment is used. Additionally, in this model, all local warehouses are continuously replenished from the global warehouse for all SKU positions. This means that once a demand occurs, a replenishment order is immediately placed at the global warehouse, and the part is sent to the local warehouse.

The spare parts planning model used by ASML to determine the base stock levels at the local warehouse assumes an infinite stock point from which demand can always be fulfilled in case of an emergency shipment. In reality however, an emergency shipment can only be performed from one of the global warehouses when stock is available. In case no stock is available in practice, and a part is needed as soon as possible, the part is pulled out of one of the ASML factories. Accordingly, we decide to not include back-orders in the digital twin and to consider the factories as an infinite stock point.

In line with the costs components which are also included in the spare parts planning model, the following costs are considered in the digital twin:

- Holding costs c_i^h for each SKU $i \in I$. The holding costs drive the base stock levels down, as higher levels would correspond to higher costs. The holding costs cover the warehousing component of the service cost and are computed using a holding cost rate for each local warehouse. Here the holding rate multiplied by the unit cost of the SKU expresses the costs incurred to hold one unit for one year.
- Lateral transshipment costs $c_{n,j}^{lat}$ incurred for a shipment from local warehouse n to j, whereby warehouses n and j are located in the same region. This cost factor includes both a freight component and a duty component (in case the transport crosses country borders). These costs drive the local

stock levels up, as high local stock levels cause more demand being satisfied through local stock, thus requiring fewer lateral transshipments.

- Emergency transportation costs $c^{em}_{n,j}$ incurred for an emergency shipment between warehouses n and j, whereby warehouse n is either a global warehouse or a local warehouse that is located in a different region than local warehouse j. This factor corresponds to the costs of shipping a part in case no stock is available in the dedicated warehouse nor in another warehouses within the region. Similarly as for the lateral transshipment costs, this cost factor, which includes both a freight component and a duty component, drives local stock levels up.

2.3 Switches

The main modeling challenge is that in order to quantify and explain the impact of a single cause on performance gap A, its effect has to be isolated. The digital twin should be able to make a statement about what happens with the performance when a specific cause does, or does not, occur, without changing anything else. A more complex simulation structure is chosen which follows an "on-demand" approach whereby the single causes can be included or excluded in the digital twin before the simulation runs are started. The conceptual idea is as follows. The base model exists of a simulation of the spare parts planning model, following all assumptions made in the planning model. With this simulation, the expected performance is measured. Next, the rules of the digital twin can be changed. Alternative realities are created in which a specific performance gap cause does, or does not, occur, while keeping all other processes unchanged. With this set-up, the impact of the changes on the performance can be tested. Using this idea as a building block, Figure 3 illustrates the key concept whereby the identified causes are depicted as on-off switches.

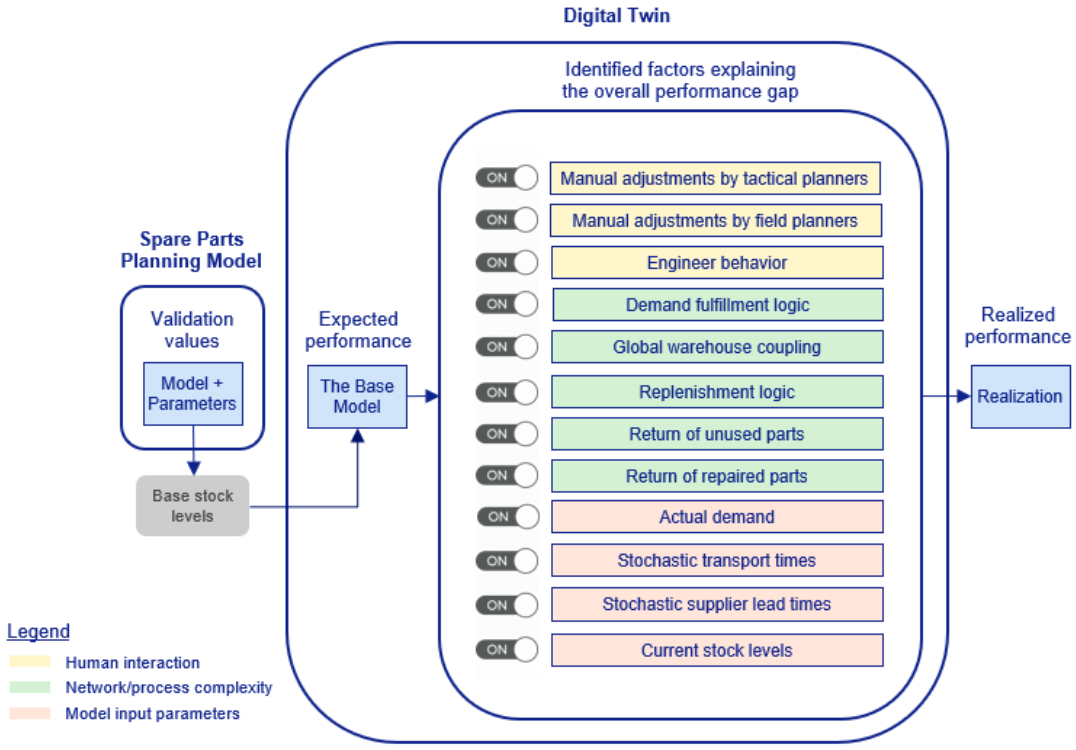

Figure 3: Conceptual model including switches.

By switching one or multiple of the switches included in the simulation tool on, the base model, as described in the previous section, is altered. In the following subsections we briefly describe the changes for each individual switch.

2.3.1 Human Interaction

Once the optimal base stock levels are determined by the planning model, the results are first reviewed by the tactical planners at the central office of ASML. In case the switch related to the manual adjustments made by the tactical planners is switched on, these reviewed base stock levels are provided as input to the model and used in the simulation. After the review performed by the tactical planners, the base stock levels are also reviewed by the field planners who work in the regions and can suggest changes for the base stock levels in their region. Accordingly, in case the switch related to the manual adjustments made by the field planners is switched on, these reviewed base stock levels are used. In case the engineer behavior switch is switched on, the transportation time of a part is determined based on the priority of the order as registered by the field engineer who places the order.

2.3.2 Network and Process Complexity

The planning model used to determine the base stock level at the local warehouses assumes an infinite stock at the global warehouses. In case the global warehouse coupling switch is switched on, the two global warehouses are added to the scope of the simulated service supply network. In practice, depending on the region the requesting plan group is located in, local warehouses of other close-by regions are checked for stock as well. Therefore, the demand fulfillment logic switch influences the order and number of warehouses checked for stock. In case the replenishment logic switch is switched on, the moment at which purchase and replenishment orders are placed is replaced by the moment at which the requested part arrives at the customer. Additionally, a new approach to perform replenishment actions is introduced. While the base model uses a First-In-First-Out (FIFO) approach, when this switch is on replenishment actions are based around the concept of non-availability (NAV) risk - which is the probability of being out of stock during the lead time for replenishment. In contrast to the base model which assumes that every requested part is used, the return to unused parts switch relates to the fact that a part that is not built into the system returns to the warehouse. In case the return of repaired parts switch is switched on and a defective service part is built out of the system, it is checked whether the defective part is successfully repaired. If so, the repaired part re-enters the network after the repair lead time.

2.3.3 Model Input Parameters

The base model uses simulated demand based on a demand forecast. In case the actual demand switch is switched on, actual historic demand is used in the simulation. In case the current stock levels switch is turned on, the actual stock present at the warehouses at the start of the to-be simulated period is used as starting state for the simulation. To incorporate uncertainty in the transportation times in the model, in case the stochastic transportation times switch is on, we use Gamma distributed transportation times. For the stochastic supplier lead times switch we include normally distributed errors for the supplier lead times.

2.4 Discrete Event Simulation

In essence, a digital twin is often a discrete event simulation, which seems to be suitable for being used in a broad spectrum of applications due to its efficiency and its flexibility stemming from its stochastic nature (Agalianos et al. 2020).

In our digital twin, we also apply discrete-event simulation to analyze performance gap A. By using such a simulation, it is possible to simulate the inventory control process and by changing the rules used to schedule events, this system can easily be adjusted to reflect what would have happened if a specific cause

occurred or was absent. Events are related to the time epochs at which the state of the system changes. In the digital twin, we simulate a period of 6 months, which is equivalent to the time period between two consecutive planning model runs whereby we measure the costs, service performance, and inventory turnover over that period.

At the moment that the base stock levels are updated, stock is already present in the warehouses, which may either be lower or higher than the newly set base stock levels. Accordingly, the inventory levels may not be equal to the base stock levels. When neglecting this characteristic in the digital twin, the twin might overestimate the inventory levels. Therefore, the twin might return higher service performance than realistic over the 6-month period. As a consequence, the digital twin might not give a valid representation of real-life. Therefore, it is not desired to start the simulation with the assumption that stock levels are equal to the base stock levels at the start of the planning horizon. Consequently, a warm-up period is defined. The warm-up period is determined using the approach described by Law and Kelton (2000). This approach uses a minimum of five replications (r) of the simulation to retrieve the total stock level in the service supply network (Y), over a relatively large period of time (l). We then compute the average total stock level in the supply network $\overline{Y} = \sum_{x=1}^{r} Y_x/r$. We set $l = 10$ years and $r = 10$ to investigate when the stock present in the network reaches steady state for both the base model and the model whereby all switches related to the planning performance are switched on.

As can be seen in Figure 4, the stock level at the initialization of the simulation declines rapidly. This effect occurs due to the fixed replenishment lead time (i.e. the set target waiting time) which differs between 6 and 29 days depending on the SKU. Accordingly, in the first few weeks, demand for spare parts arises while no replenishment are received yet, explaining the steep decline. Thereafter, an increase in replenishment occurs which results in a stabilization of stock levels. Based on these findings it was decided to use a warm-up interval of 1 year for measuring the expected performance (i.e. 365 days). For measuring the planning performance, Figure 5 shows that the stock level in the network declines less rapidly due to the fact that stock can now be immediately replenished from the global warehouse. Additionally, a small increase in stock is visible in the first few weeks which can be explained by unused parts returning to stock. Based on these findings it was decided to use a warm-up period of 5 years. Please note that the current stock levels are modeled as a switch. In case this switch is turned on, no warm-up period is required since we will take the current stock levels as the starting state of the simulation.

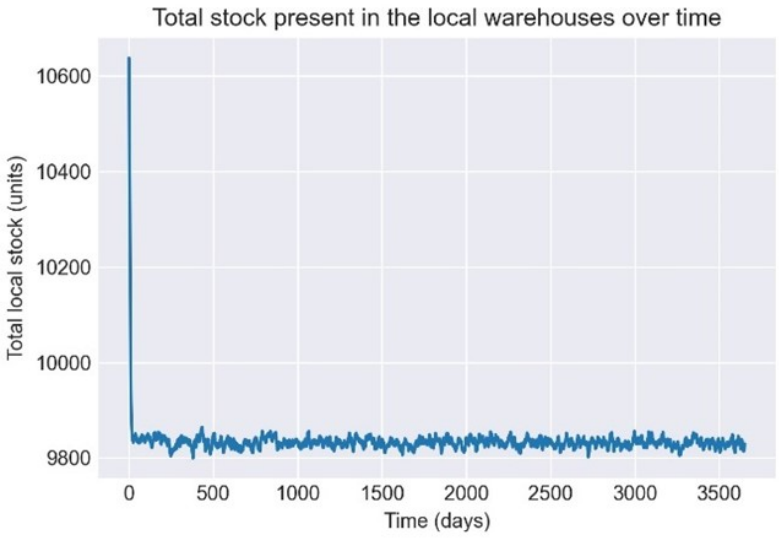

Figure 4: Total stock present in the local warehouses over time.

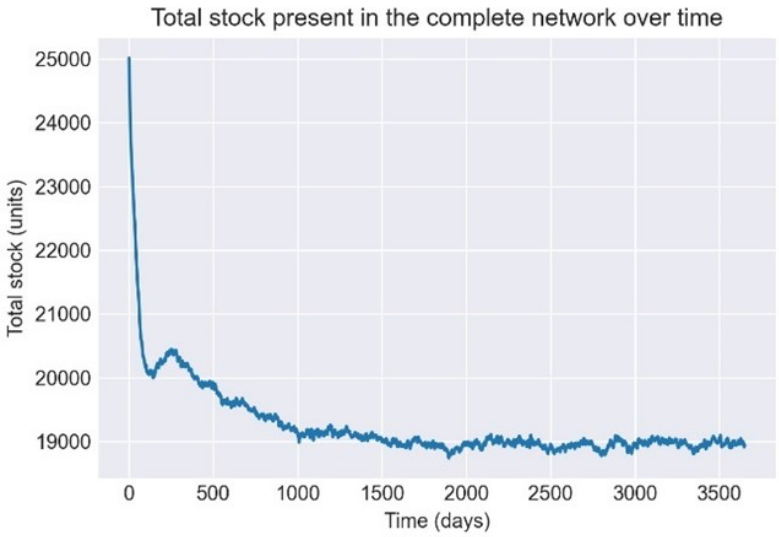

Figure 5: Total stock present in the complete network over time.

Using the approach as described by Byrne (2013), the minimum required number of replications needed to have 95 % of the outcomes within plus minus 0.5 percent of the mean value are obtained. Finally, we validate our simulation by comparing the simulation output of the base model with the output of the spare parts planning model. We measure the direct fill rate, lateral transshipment rate and the emergency shipment rate per combination of local warehouse and mover type. The mover type classification we use classifies the SKUs in low, medium and high demanded SKUs. Here based on historic demand data over the past 12 months, SKUs that were requested less than 5 times are classified as low demanded, SKUs that have between 5 and 24 requests are classifies as medium demanded, and SKUs with 24 or more request are classified as high demanded. The results with a simulation period of 10 years and the number of replications equal to 200 are presented in Table 1. In this table the local warehouses that start with the same letter are located within the same region. The results show that for most cases the model outcome falls within the confidence interval resulting from the simulation tool. Since the spare parts planning model makes use of an approximate evaluation procedure, the model may not be equal to the results obtained when using evaluation by simulation. van Houtum and Kranenburg (2015) have tested the accuracy of the model by comparing approximate results with exact results. A maximum absolute error of 1 % was found for the item fill rate. Taking this into account, we conclude that the base model of the digital twin is valid.

For the validation of the on-off switches several techniques are used. Firstly, for the switches that influence the input that is being used in the simulation simple input checks are performed. This technique deals with the switches related to the (reviewed) base stock levels for the warehouses. Accordingly, the functionality of those switches was easily validated by checking the set base stock levels for the warehouses at the start of the simulation period and comparing them with the corresponding input data sets. The same holds for the switch related to the current stock levels. The functionality of this switch was validated by checking the starting state of the simulation (i.e. the stock levels at the warehouses and the stock in-transit to the warehouses) and comparing it to the input data set. Secondly, output checks are performed to check the results for reasonability based on expert opinions. Moreover, the results are compared to the results of the base model. For example, when switching on the cause related to the return of unused parts, we expect the direct fill rates to increase as compared to the base model, this effect was indeed observed. Thirdly, using a small subset of the data, component tracing is performed to determine whether the simulation works as expected. This subset includes three different SKUs for which we trace the stock levels at the warehouses as well as the order of events occurring. Following the simulation with hand calculations, it

Table 1: Validation results for each combination of local warehouse and mover type.

Local WH	**Mover Type**	**Direct Fill Rate (%)**		**Lateral Transship. Rate (%)**		**Emergency Shipment Rate (%)**	
		Planning model	Digital twin	Planning model	Digital twin	Planning model	Digital twin
A.1	Low	86.08	(86.0, 86.14)	1.87	(1.84, 1.89)	12.05	(12.0, 12.13)
A.1	Medium	90.04	(90.04, 90.13)	1.32	(1.29, 1.32)	8.65	(8.56, 8.65)
A.1	High	99.09	(99.08, 99.10)	0.61	(0.60, 0.62)	0.30	(0.29, 0.30)
A.2	Low	45.82	(45.41, 46.02)	43.94	(43.67, 44.25)	10.23	(10.17, 10.49)
A.2	Medium	62.85	(62.55, 62.89)	29.40	(29.37, 29.69)	7.74	(7.66, 7.83)
A.2	High	94.76	(94.66, 94.76)	4.70	(4.7, 4.79)	0.54	(0.53, 0.56)
B.1	Low	89.65	(89.59, 89.77)	1.34	(1.32, 1.39)	9.01	(8.88, 9.06)
B.1	Medium	94.29	(94.23, 94.32)	2.43	(2.41, 2.47)	3.28	(3.26, 3.32)
B.1	High	99.36	(99.3, 99.32)	0.61	(0.63, 0.66)	0.04	(0.04, 0.04)
B.2	Low	89.46	(89.4, 89.51)	3.04	(3.01, 3.07)	7.49	(7.46, 7.56)
B.2	Medium	93.96	(93.94, 94.0)	2.86	(2.84, 2.88)	3.18	(3.15, 3.2)
B.2	High	99.39	(99.38, 99.4)	0.59	(0.58, 0.6)	0.02	(0.02, 0.02)
B.3	Low	88.15	(88.02, 88.22)	4.68	(4.63, 4.77)	7.17	(7.1, 7.27)
B.3	Medium	92.67	(92.62, 92.74)	4.88	(4.82, 4.91)	2.45	(2.42, 2.49)
B.3	High	99.39	(99.28, 99.31)	0.59	(0.66, 0.68)	0.02	(0.03, 0.03)
C.1	Low	33.81	(33.88, 34.81)	0.0	(0.0, 0.0)	66.19	(65.19, 66.12)
C.1	Medium	45.36	(45.08, 45.68)	0.0	(0.0, 0.0)	54.64	(54.32, 54.92)
C.1	High	90.77	(90.65, 90.86)	0.0	(0.0, 0.0)	9.23	(9.14, 9.35)
D.1	Low	90.38	(90.29, 90.45)	0.0	(0.0, 0.0)	9.62	(9.55, 9.71)
D.1	Medium	93.83	(93.76, 93.84)	0.0	(0.0, 0.0)	6.17	(6.16, 6.24)
D.1	High	99.17	(99.16, 99.18)	0.0	(0.0, 0.0)	0.83	(0.82, 0.84)
E.1	Low	28.35	(27.77, 28.76)	0.0	(0.0, 0.0)	71.65	(71.24, 72.23)
E.1	Medium	49.07	(48.68, 49.32)	0.0	(0.0, 0.0)	50.93	(50.68, 51.32)
E.1	High	91.77	(91.72, 91.9)	0.0	(0.0, 0.0)	8.23	(8.1, 8.28)

can be determined whether the simulation tool works as expected. Specifically, this approach is performed for all five switches related to the network and process complexity. Using these three techniques, the on-off switched included in the simulation tool have been validated.

3 DEMONSTRATION OF TOOL FUNCTIONALITY

In the first use case, the digital twin is used to quantify the performance gap between the expected and digital twin performance. We will also quantify how this performance gap is relevant to the overall performance gap. To further demonstrate the functionality of the digital twin, a second use case is formulated. This use case investigates the influence of the actual demand switch on the performance gap. Many more interesting use cases can be formulated by various configurations of the switches, and the two cases presented here merely serve as a demonstration of how the tool can be used. Also note that the simulation results and model output presented are retrieved by considering a subset of the complete set of warehouses, plan groups, and service parts. Therefore, these results do not represent the overall (expected) performance of ASML.

3.1 Quantifying the Performance Gap

To quantify the performance gap between the expected and digital twin performance, almost all switches of the digital twin are on, except for two: the *engineer behavior* and the *current stock levels*. Using the warm-up period of 5 years and setting the number of replications equal to 46 such that accurate results are obtained on local warehouse level, the normalized digital twin performance per local warehouse is retrieved and presented in Table 2. In addition, Table 2 quantifies the gap with the expected performance. For the service level measures, the gap presents the percentage difference between the expected performance and the performance obtained by the digital twin. The service level measures we report are: direct, local, and regional fill rate. In contrast to the direct fill rate which takes into account both used and unused demand, the local fill rate measure only considers used demand. For the regional fill rate, the part may also be delivered from one of the other local warehouses within the region. The results show that overall the

identified factors causing the performance gap, have a negative effect. In the table, negative effects are presented in red (i.e. a decrease in service level, an increase in costs, an increase in the inventory turnover). It can be seen that for most local warehouses, the service levels as well as the holding costs are negatively impacted, which indicates that the parts that were stocked locally were not requested as much as expected, leading to higher holding costs, and the part that were requested often were not stocked locally, leading to a lower direct fill rate.

Table 2: Normalized digital twin performance per local warehouse and the gap with expected performance.

	Service Level Measures						Service OPEX				Inventory	
Local WH	**Direct Fill Rate**		**Local Fill Rate**		**Regional Fill Rate**		**Holding Costs**		**Transport Costs**		**Inventory Turnover**	
		Gap		Gap		Gap		Gap		Gap		Gap
A.1	86.48	-13.52	87.84	-12.16	91.00	-9.00	129.80	+42.81	15.79	+13.17	22.95	-20.56
A.2	76.10	-23.90	72.93	-27.07	93.22	-6.78	7.37	+0.20	3.20	+2.67	12.98	-26.51
B.1	91.51	-8.49	92.87	-7.13	99.24	-0.76	74.08	+11.49	2.37	-0.08	35.12	-49.30
B.2	89.25	-10.75	88.95	-11.05	96.94	-3.06	132.82	+32.82	14.81	+9.47	22.40	-25.90
B.3	81.14	-18.86	80.94	-19.06	97.48	-2.52	52.66	-14.25	3.49	+1.85	27.37	-72.63
C.1	77.78	-22.22	83.62	-16.38	83.62	-16.38	4.93	+3.82	3.00	-0.90	34.57	+12.12
D.1	90.60	-9.40	96.40	-3.60	96.40	-3.60	131.60	+71.62	15.54	+10.24	51.94	-13.22
E.1	102.76	+2.76	106.59	+6.59	106.59	+6.59	5.83	+4.27	0.73	-2.54	44.45	+10.36

Next to quantifying the gap between the expected and digital twin performance, we would like to indicate how this gap is relevant to the overall performance gap. Therefore, we denote the realized local fill rate performance over the same period as the simulation data. We show the figures, normalized towards expected local fill rate performance, in Table 3. First of all, it can be observed that compared to the expected performance, the digital twin performance is usually lower, and the realized performance is usually higher. This can be explained by the fact that most switches represent factors which are not explicitly considered by the planning model, and can therefore be expected to have an overall negative effect on the local fill rate performance. On the contrary, factors not included in the digital twin, which are factors occurring on operational level that cannot be predicted beforehand appear to have an overall positive effect on the service level performance, which even leads to over-performance. It is advisable for future model iterations, that the performance gap between digital twin and realized performance is investigated further to identify factors which can be added to the twin, in order to narrow the gap between digital twin and realized performance.

3.2 Impact of Actual Demand

To quantify the impact of actual demand, we will turn this switch off, contrary to the previous scenario where actual demand was switched on. This means that we will use simulated demand instead of the actual demand. Table 4 presents the normalized results for the performance obtained. Again, we provide the gap with the expected performance. Looking at the service measures, it can be seen that the gap with the expected performance is much smaller for all warehouses as compared to the performance gap that was measured in the previous experiment. Hence, also for most warehouses, the transportation costs are smaller compared to the previous scenario, resulting in a smaller gap with the expected performance. For

Table 3: Expected, digital twin, and realized local fill rate performance.

Local WH	**Expected performance**	**Digital twin performance**	**Realized performance**
A.1	100	91.00	102.31
A.2	100	93.22	114.47
B.1	100	99.24	101.83
B.2	100	96.94	101.39
B.3	100	97.48	101.81
C.1	100	83.62	130.40
D.1	100	96.40	101.30
E.1	100	106.59	120.92
Average	100	95.56	109.30

the holding costs and the inventory turnover rate, less deviations from the previous scenario are obtained. This can be explained by the fact that the base stock levels are not changed.

This use case shows that by separately considering the impact of the switches, the gap between the expected and realized performance can be attributed to specific root causes. By addressing these root causes in the physical environment, the spare parts planning can be improved.

Table 4: Normalized performance per local warehouse using simulated demand and its gap with the expected performance.

	Service Level Measures						Service OPEX				Inventory	
Local WH	**Direct Fill Rate**		**Local CSD**		**Regional CSD**		**Holding Costs**		**Transport Costs**		**Inventory Turnover Rate**	
		Gap		Gap		Gap		Gap		Gap		Gap
A.1	99.51	-0.49	99.57	-0.43	99.37	-0.63	117.92	+30.93	5.66	+3.04	19.66	-23.85
A.2	81.29	-18.71	85.20	-14.80	92.55	-7.45	6.09	-1.08	1.65	+1.11	11.10	-28.39
B.1	95.90	-4.10	96.27	-3.73	99.18	-0.82	69.73	+7.14	3.07	+0.61	31.13	-53.29
B.2	98.58	-1.42	98.55	-1.44	99.33	-0.67	121.13	+21.13	6.67	+1.33	19.51	-28.79
B.3	93.76	-6.24	94.44	-5.56	98.68	-1.32	48.19	-18.72	3.51	+1.86	24.02	-75.98
C.1	102.00	+2.00	101.75	+1.75	101.75	+1.75	4.78	+3.68	2.94	-0.96	31.60	+9.15
D.1	99.59	-0.41	99.36	-0.64	99.36	-0.64	121.09	+61.11	8.75	+3.45	43.89	-21.27
E.1	100.17	+0.17	99.18	-0.82	99.18	-0.82	5.48	+3.92	1.72	-1.55	38.02	+3.94

4 CONCLUSION

From theory, it is known that a mismatch between the model output and its realization in practice is to some extent inevitable, due to the fact the a model will always be a simplification of reality. Nevertheless, investigating and understanding what causes this mismatch and acknowledging the limitations of the model is vital in managing its consequences and proposing improvements to narrow the performance gap in the future. However, very little research has focused on understanding the factors that drive the performance gap and the impact of those factors. We developed a digital twin that can be used to quantify causes of a gap between the expected performance and the realized performance at ASML. For the test case data set, a relatively large negative gap between the expected- and digital twin performance was found for most local warehouses. It was found that the actual demand negatively influences the service performance at most warehouses and is a large contributor to the expected- versus digital twin performance gap.

It can be concluded that the developed simulation tool provides ASML with the ability to investigate the planning performance and generate insights on the impact of different causes on the gap. Using the understanding of the reasons for the performance gap and the impact of every cause in the future, a better control on the process is gained, potentially resulting in more effective inventory management and an increased trust in the planning. Therefore, ASML aims to set up a feedback loop from the digital twin towards the planning model for the spare parts network, so a digital twin with two-way interaction is realized.

REFERENCES

Agalianos, K., S. Ponis, E. Aretoulaki, G. Plakas, and O. Efthymiou. 2020. "Discrete Event Simulation and Digital Twins: Review and Challenges for Logistics". *Procedia Manufacturing* 51:1636–1641.

Axsäter, S. 2015. *Inventory Control*. International Series in Operations Research and Management Science. Cham: Springer.

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.*, 445–450.

Grieves, M. 2014. "Digital Twin: Manufacturing Excellence through Virtual Factory Replication". Technical report, Auburn Hills, Michigan: Delmia. https://www.3ds.com/fileadmin/PRODUCTS-SERVICES/DELMIA/PDF/Whitepaper/DELMIA-APRISO-Digital-Twin-Whitepaper.pdf, accessed 7th July 2022.

Guo, J., N. Zhao, L. Sun, and S. Zhang. 2019. "Modular Based Flexible Digital Twin for Factory Design". *Journal of Ambient Intelligence and Humanized Computing* 10(3):1189–1200.

Jones, D., C. Snider, A. Nassehi, J. Yon, and B. Hicks. 2020. "Characterising the Digital Twin: A Systematic Literature Review". *CIRP Journal of Manufacturing Science and Technology* 29:36–52.

Kritzinger, W., M. Karner, G. Traar, J. Henjes, and W. Sihn. 2018. "Digital Twin in Manufacturing: A Categorical Literature Review and Classification". *IFAC-PapersOnLine* 51(11):1016–1022.

Lamghari-Idrissi, D., R. van Hugten, G. J. van Houtum, and R. Basten. 2022. "Increasing Chip Availability Through a New After-Sales Service Supply Concept at ASML". *INFORMS Journal on Applied Analytics* 52(5):460–470.

Law, A. M., and W. D. Kelton. 2000. *Simulation Modeling & Analysis*. 3rd ed. New York: McGraw-Hill, Inc.

Lugtigheid, D., A. K. S. Jardine, and X. Jiang. 2007. "Optimizing the Performance of a Repairable System under a Maintenance and Repair Contract". *Quality and Reliability Engineering International* 23(8):943–960.

Topan, E., A. Eruguz, W. Ma, M. van der Heijden, and R. Dekker. 2020. "A Review of Operational Spare Parts Service Logistics in Service Control Towers". *European Journal of Operational Research* 282(2):401–414.

van Houtum, G. J., and A. Kranenburg. 2015. *Spare Parts Inventory Control under System Availability Constraints*. International Series in Operations Research & Management Science. New York: Springer.

AUTHOR BIOGRAPHIES

JOAN STIP is a part-time Ph.D. candidate at the department of Industrial Engineering at Eindhoven University of Technology. He earned his M.Sc. and P.D.Eng degree at the same school. His research interests are on service logistics, human algorithm interaction, and simulation and optimization methodologies. Next to his research, he is working at ASML as a supply chain engineer in the customer supply chain management department. His e-mail address is j.stip@tilburguniversity.edu.

LOIS AERTS recently completed her master Operations Management and Logistics at the Eindhoven University of Technology. In her master thesis project, conducted at ASML, she developed the digital twin which is described in this paper. She graduated in March 2022 and started her career at ASML. Her email address is lois-aerts@hotmail.com.

GEERT JAN VAN HOUTUM is Full Professor and chair of Maintenance and Reliability at Eindhoven University of Technology (TU/e) since 2008. Since September 2017, he is also vice-dean IE of the Department IE & IS. His areas of expertise include maintenance optimization, spare parts inventory control, inventory theory in general, and operations research. Geert-Jan carries out research on the maintenance and reliability of capital goods. In particular, his focus is on design and control of spare parts networks, predictive maintenance concepts, and product design choices that have a strong effect on system availability and TCO. In this research, also the value of remote monitoring data and other degradation data is investigated. Much of this research is carried out in cooperation with the industry, working with companies such as ASML, Dutch Railways, Lely, Philips, Marel, the Royal Netherlands Navy, and Vanderlande. His e-mail address is g.j.v.houtum@tue.nl