Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

OPTIMIZATION OF TIMELINKS IN SEMICONDUCTOR MANUFACTURING

Nina Dybowski
Maria Sander
Ralf Sprenger

Infineon Dresden GmbH & Co KG
Königsbrücker Str. 180
01099 Dresden
GERMANY

ABSTRACT

Impact of timelinks to semiconductor manufacturing has risen due to shrinking technology sizes. Their operational control defines on the one hand how good the time restrictions are met and on the other the impact to fab capacity. This paper discusses both aspects and the influencing factors like uptime stability, length of the timelink etc. A control approach is proposed, evaluated, and discussed. Furthermore, a monitoring system is introduced that enables for fast decision making and optimization of the control parameters. Finally, a simulation study is done for evaluating different parameters and impact of influencing factors.

1 INTRODUCTION

Semiconductor Manufacturing is characterized by complex re-entrance flows with a lot of process steps, multiple machine sets, that can process the material, and tight constraints that lead to restrictions where material can be processed (e. g. dedications). A limitation that is typical for smaller structural size on wafers are time critical areas within the production flow where material must pass consecutive steps within a given time to avoid yield loss, rework or potential scrap due to e. g. oxidation. In addition to technology-related restrictions there are also requirements resulting from highly competitive sales market environment for semiconductor products. Cost pressure on the one hand but also fast production cycle times on the other need to be under consideration. Production restrictions like time constraints have a negative influence on both factors.

New technologies with smaller structures lead to increasing amount of time constraint areas within the wafer production. Beside the technology driven challenges, this limitation affects capacity planning and controlling strategies in the production. Although, the wafer fabs are facing this problem more and more often, there is only very little number of publications concerning this issue. Robinson and Giglio (1999) showed a planning approach for time constraints with a M/M/c approximation to define the maximum planned load for machines which are linked by a time constraint. This is shown for a simplified system with two machine groups and reduced variability. This approach considers not all sources of variability, e. g. for machine availability. Also, Tu et al. (2010) used a queuing model for showing the capacity impact of time constraints but with focus to batching processes. By using a GI/G/m approximation they are modelling the influences of the systems' capacity in order to the probability of exceeding the time constraint. They also recommend including additional factors like failure behavior of the machines. These approximations show that queuing models can be used for capacity planning models but for further usage to improve the production controlling strategy these models cannot represent the entire complexity and variability. The main issue we are facing due to variability are influences that lead to an exceedance of time constraint and a resulting scrap decision. The approach from Scholl and Domaschke (1999) focuses more on a controlling

strategy using a Kanban-like system and evaluating the usage between wet etch and furnace processes. An important finding was, that the controlling strategy depends on the available capacity of both machine groups. Klemmt and Mönch (2012) suggest a solution using a scheduling algorithm as controlling strategy for timelinks, but are also using a simplified system to prove the concept.

These theoretical approaches could not cover all the influences and complexities we are dealing with in the daily production. We are also not able to use an adapted scheduling solution for every new time constraint that is necessary for the development of new products. We need a simple and transparent but also changeable controlling strategy that executes some WIP limitations within the time constraint area by only adapting some less parameters for increasing or decreasing the capacity with strong focus to yield results.

Literature sometimes discusses the impact of time constraints to capacity planning models. The capacity plans are directly linked to the costs and the cycle time. The issue we are facing in the daily production control is linked to the planning (available capacity) but it is even more complex due to the variability of production area influences. Time constraints are limited to theoretical or simplified scenarios. Practical solutions on how to handle this problem are not available. Beside on how to control the material flow into such an area, also the influencing factors like uptime stability and mean time to repair need to be under further consideration. Furthermore, uncertainty needs to be considered hence available capacity within these areas might change over time. It is also worth to discuss when a control mechanism has influence on the fab capacity hence these areas sometimes require backup capacity to handle material waves generated by the control approach. For optimizing the control approach, setting the parameters and work on the important influencing factors, a Decision Support System (DSS) is needed to monitor the current situation in the line and to evaluate the root cause (very often laying in the past) that let to it.

This paper gives a wide overview about these different aspects and discusses the problem using an existing timelink chain. Control approaches from the literature are discussed and a simple KANBAN-like approach will be proposed. Pros and cons and the influencing factors on the control parameters are discussed. Monitoring charts and a Decision Support System for area visualization and fast decision making are shown. Finally, this timelink chain is simulated and enables us to optimize parameter settings.

2 TIMELINKS

2.1 Definition of Timelinks

In semiconductor manufacturing, the wafers are combined as so-called lots that are being processed according to a process flow $p_0 \dots p_k$, $k \in \mathbb{N}$. Timelinks are defined as a maximum time $tl > 0$, $tl \in \mathbb{N}$ between process end at operation p_m and process start at p_n where $m < n$ must apply. The time of a lot within a timelink is defined by the process time pt and waiting time wt at all operations within the timelink and the waiting time at the operation p_n so that if

$$wt(p_n) + \sum_{x=m+1}^{n-1} wt(p_x) + pt(p_x) \leq tl$$

is valid, the timelink has been passed without violating the constraint.

Timelinks can be connected to others as a timelink chain (Figure 1). This means at operation p_n an additional timelink might start. Detailed description of different timelink variants can be found in Klemmt and Mönch (2012). This paper focuses on timelinks that are shown in Figure 1. Overlapping timelinks and other variants are not discussed.

There are two ways to prevent material from violating the constraints within a chain:

- timelink material can be handled with higher priority against non-timelink material (e.g. Scholl and Domaschke (1999)) and

- material can be stopped at the first operation (timelink gate) of a chain in order to limit the amount of material within the timelink chain. The material flow can be controlled by a KANBAN like approach.

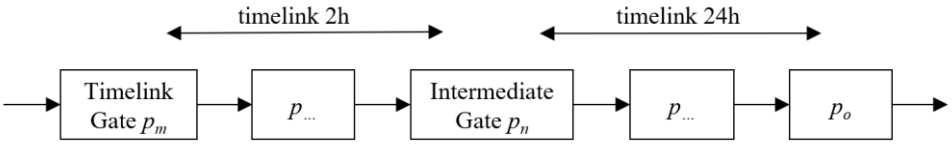

Figure 1: Example of a timelink chain.

3 INFLUENCING FACTORS

Influences on capacity within semiconductor productions are described in publications like Ignizio (2009) (e. g. pp. 152 ff.). During our analysis, we found additional factors that have an impact on the fab performance due to timelink chains. The machine groups within the timelink chain with the highest planned utilization are defined as potential bottlenecks. By using this definition we can identify the backup capacity to catch up process moves that have been missed in the past due to low level of material (WIP) in front of the machines. In reality the backup capacity is usually low due to high utilization of the machines and the cost depends very much on utilization. Therefore, high utilization is an important indicator to avoid costs. Especially, these bottleneck machine groups are crucial for the performance of the whole timelink chain. The timelink itself has an influence on the capacity of these bottlenecks. Due to the described interconnections, the following factors become very important to the available capacity within this timelink chains.

3.1.1 Uptime Stability

Stability of machine uptime is of the essence within timelink chains; this is especially valid for the bottleneck groups. A huge drop in uptime leads to a risk for violating the timelink restriction. This has two consequences. First, the buffer for the WIP limit must be set according to the WIP that can be absorbed with the remaining capacity. Secondly, the WIP feed must be stopped immediately hence the amount of WIP within the chain is much too high after a machine breakdown. This leads to machines going to standby (and therefore capacity loss), starting at the gate up to the step where the uptime deteriorated. In addition, the number of tools within a machine group must be taken into consideration. The more tools are available for a specific process, the less is the influence of one machine down on the relative throughput decrease.

3.1.2 Ability to Forecast and Align Machine Downs

The ratio between scheduled and unscheduled down events can be leveraged in two directions. On the one hand, if the time is known when machines will go down, the WIP feed from the gate can be reduced beforehand. This avoids WIP accumulation in front of the bottleneck, reduces stand by times due to fully stopped WIP feed and reduces the risk of timelink violations. On the other hand, downs of other critical machine groups within the chain can be aligned so that at the same time the capacity of the whole chain goes down to a certain degree with the result that it will come back with well-maintained tools and constant uptime for the whole chain afterwards.

3.1.3 Position of Bottlenecks within the Timelink Chain

The more operations are upstream the bottleneck machine group within the timelink chain, the longer is the reaction time of the material feed from the gate. E. g. the bottleneck uptime goes down and the control

mechanism reduces the amount of material to start into the chain, the already started material will continue to arrive at the bottleneck, possibly for additional days, if the chain is long.

3.1.4 Length of the Timelink Restriction

The length of the timelink restrictions linearly affects the amount of WIP, we can hold within the chain at the same time. So, if the restriction is relaxed by factor two, we can feed double of the material into the chain without increased risk for timelink violations while reducing potential capacity loss at the same time.

3.1.5 Ratio of Processing non Timelink Material

Machine groups within the timelink chain that also process non timelink material hold the chance to prioritize timelink material when needed, e. g. the bottleneck within the chain must be filled fast hence WIP level is too low. If the bottleneck itself processes non timelink material to a high share, timelink critical lots should be processed first in case of a high uptime decrease.

3.1.6 Implementation of Intermediate Gate Blocks

If two or more timelink restrictions are following on each other, there is a problem in a later timelink with too high WIP level and lots getting critical, the material can be prevented to being processed at the last step of one former timelink if there is still time left. This gives some additional buffer. Especially, in addition to higher variability in these systems this holds some potential for improvement.

4 CONTROL APPROACH

WIP feed into this area must be controlled hence the risk of timelink violations must be minimized. There are several approaches in the literature. Some depend on MIP and CP models (e. g. Kalir and Tirkel 2016) that decide which material is allowed to be fed into the area. Hence these approaches are unable to handle real world problem sizes, we suggest the following simple KANBAN like approach. More complex approaches for rule-based control are to be found e.g. in Arima et al. (2015), Sadeghi et al. (2015), Lima et al. (2017) and Wang et al. (2018). A simulation based approaches are discussed e.g. in Anthouard et al. (2022) and Winkler et al. (2016).

- Each part of the timelink is controlled by itself. The first operation of timelink that directly follows onto a previous one is controlled in the same way as the first one.
- Decision if a lot is allowed to be dispatched by the first operation is made depending on the WIP level within the area. This WIP level is determined for each single machine group within the timelink.
- Hence the throughput of a machine and the timelink restriction are known, we can set a safety value $sv = \{0,1\},\ sv \in \mathbb{R}$ where the allowed WIP level is calculated as follows:

$$\mathrm{WIP}_{allowed} = \mathrm{tl} * \mathrm{sv} * \sum_{\mathrm{m} \in \mathrm{M}} m_{up} m_{THP}$$

where *m* defines the machines within one group, $m_{up} \in \{0,1\}$, $m_{up} \in \mathbb{N}$ is 0 if a machine is down and 1 if up. $m_{THP} \in \mathbb{N}$ is the throughput of a single tool.
If *sv* is set to 1, $\mathrm{WIP}_{allowed}$ gives the amount of material, that the timelink can process within its timelink restriction *tl*. Hence, machines can break down, the material flow might be resorted due to prioritization of lots, some lots will have longer cycle time within the timelink and others shorter.

The expectation is that 50% of the lots will meet the timelink restriction whereas the other 50% will take longer. Hence this is not acceptable, the amount of material must be lower than the theoretic processable amount and is set using *sv* (Figure 2).

- $WIP_{allowed}$ is compared to the whole WIP that is running within this timelink and feeding to this machine group including the WIP that is waiting there.
- If all WIP levels of the machine groups are below their respective $WIP_{allowed}$, the lot is dispatched and processed at the first step of the timelink. If $WIP_{allowed}$ is exceeded, the lot will be blocked and reevaluated again at a later point in time, e. g. trigger based after a move has been done within the timelink or after a given period.
- If the decision within a timelink chain is made to not dispatch a lot at the first step of the 2nd or 3rd timelink, material will persist within the previous timelink until the WIP level reach an allowed value within the following one. It is clear that more complex decision strategies could be implemented that depend on the criticality (e. g. based on expected yield impact) of a timelink.

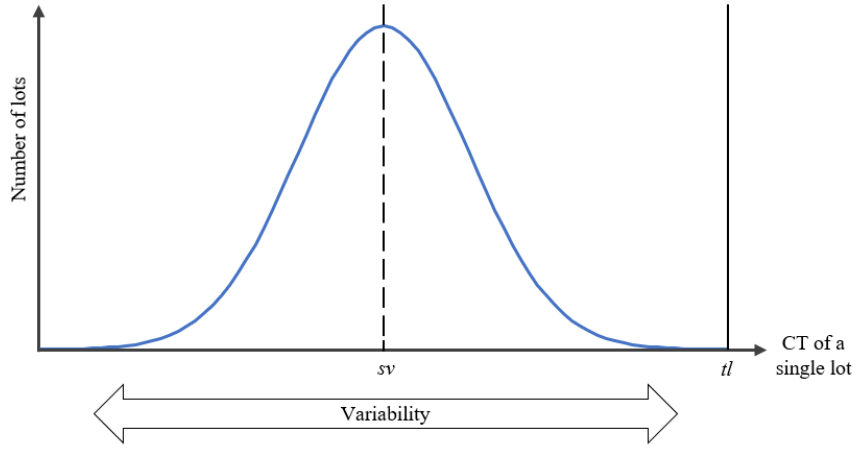

Figure 2: Cycle time distribution of lots within a timelink.

As we wrote in Section 3, a lot of factors are influencing the performance of the timelink. If we shift the distribution very far to the left, we avoid timelink violations to a high degree. On the other hand, depending on the variability within the timelink, we might lose capacity hence bottlenecks run empty due to slow reaction time for increasing WIP feed until arrival at the bottleneck.

Depending on the factors we described in Section 3, a good value for *sv* might be

- within a wide range (Case 1 in Figure 3) and setting the value somewhere within this range will neither affect timelink violation risk to a significant degree, nor reduce the capacity of the timelink area or
- capacity loss cannot be avoided and an optimal value needs to be estimated that reduces timelink violation to a very high degree and at the same time minimizes the capacity loss (Case 2 in Figure 3).

As a first shot, we have only an intuitive approach to set *sv* by our experience and starting with a low value. In Section 5 a simulation model is proposed that is able to estimate this value. However, a very good modelling of the timelink chain is crucial. Alternative calculation methods for capacity impact are to be found in Ono et al. (2006).

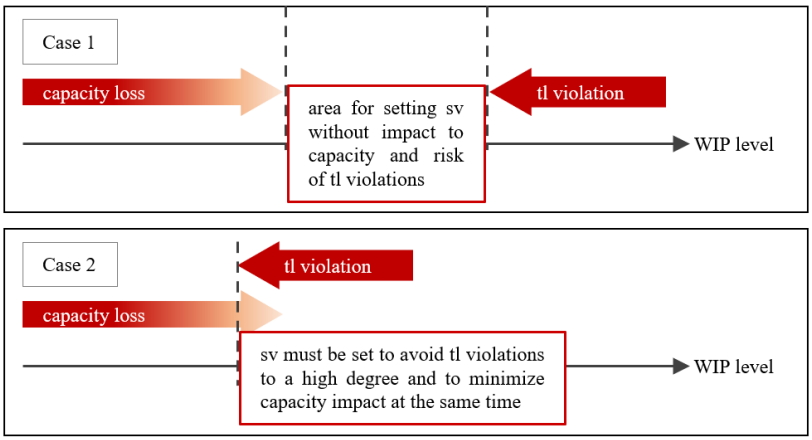

Figure 3: Timelink violation risk vs capacity loss.

5 ANALYSIS AND OPTIMIZATION OF A TIMELINK WITHIN THE METAL LAYERS

In this section, one special area that is part of the Backend of Line where the metal layers are placed on top of the transistor is discussed. This part of the line is characterized by recurrent process flows, that pass the same machine groups and the same type of timelink a few times. Beside of timelinks, only one part of this area has a set WIP limit to reduces scrap risk if the cleanroom air is contaminated. This part is not a timelink but is not that much different from it hence the WIP level is limited.

5.1 Description of Copper Timelink Area

The area consists of several machine groups and starts with an etch process p_m. This is at the same time the gate of the chain. After etch, a wet process p_{m+1} and a sputter process p_n are following. The sputter process is characterized by complex dedications that limit the ability of lots to be processed on the tools. Between etch and sputter, only a certain quantity of wafers are allowed at the same time. After sputter, a timelink starts that ends at wet p_o and allows only a few hours within this part of the line. The wet process p_o has the lowest move capability within the copper area. Hence, the aim must be to ensure that enough material is waiting in front of this process to avoid standby times but also to ensure that the timelink restriction is kept.

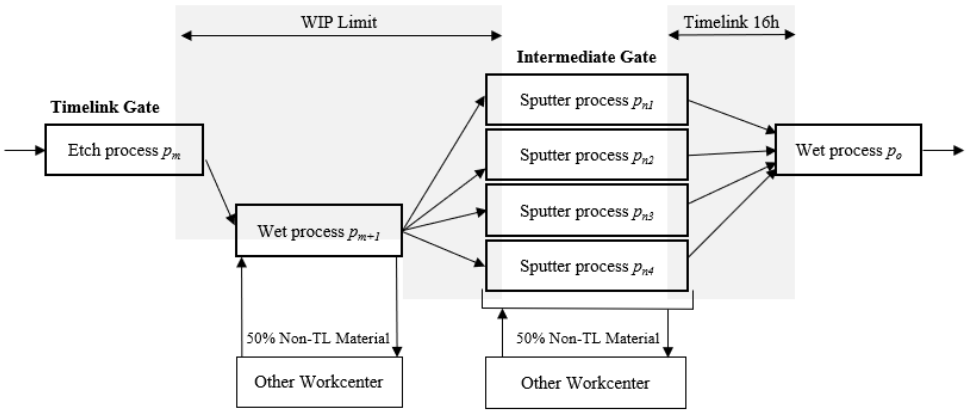

Figure 4: Copper timelink area.

5.2 Decision Support System

As described in the previous subsection, different factors have an influence on how critical a timelink regarding timelink violations or capacity impact is. The real influence of each single factor is not very transparent. Especially very complex timelink areas require a monitoring that allows evaluation of problems that currently persist, will arise within the next hours and also continues tracking of topics that are often hurting (long term root causes analysis as continuous improvement).

In Figure 5, a dashboard for a Decision Support System (DSS, see Turban et al. (2010) for the general concept) is shown that we developed for monitoring the described timelink chain. APF is used to aggregate the data from the MES system and push it to a database on a pre-aggregated level to allow fast access. Also historization is managed by APF. Data is updated every 15 minutes by an AM-Job (part of APF). Tableau is used for the final visualization as shown in Figure 5.

In contrast to the manual analysis with locking at single machine group WIP, Move and Uptime levels that takes up to an hour and gave us only a blurry picture of the current problem, the Decision Support System enables us to see within a few seconds if the chain is running well or where the problems are.

The Decision Support System Dashboard in Figure 5 shows

1. Graphs with WIP, Moves and Machine Availability on hourly level for the three relevant machine groups and provides a crisp view on the dependencies within one machine group, like how much WIP increase after a major availability drop and between the three different machine groups. When one is running empty it can be expected that the following one will follow within hours and potential counteractions can be triggered. The purple line indicates how much of the WIP is blocked by the timelink gate controller. Figure 6 gives a more detailed description of this part of the dashboard.
2. Cycle time Violation for the Workcenters and for the whole chain. This allows for current and long term root cause analysis.
3. Filling level between etch and sputter compared to the constant WIP limit.
4. Split between timelink and non timelink WIP for the machine groups.
5. Detailed availability with WIP level per machine separated by the four different process groups at sputter.
6. WIP level over time of the four different process groups at sputter.

Anthouard et al. (2022) are currently working on a Decision Support System for manual decision making on which lots to let into a timelink area. A simulation based approach is described and discussed. However, they mainly focus on the guidance of the operators to make a concrete decision about a lot whereas we focus on leveraging the DSS for parametrizing the existing dispatch rules.

Figure 5: Decision Support System (DSS) dashboard.

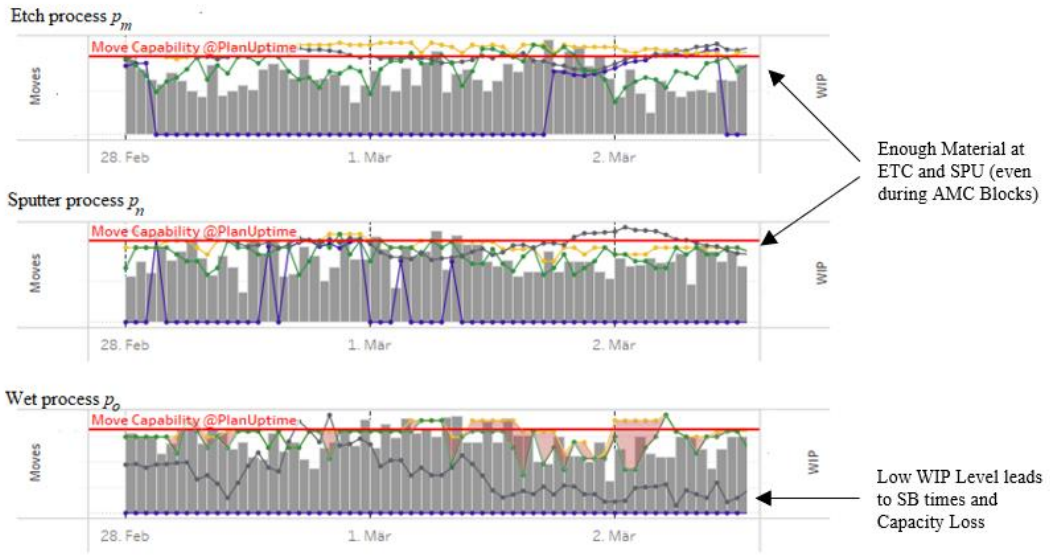

Figure 6: Example for detailed view on the three critical machine groups.

6 SIMULATION STUDY

During our analysis, we found out that the first part of the timelink area is not that critical and is not very often the limiting factor. The second part from sputter to wet is more critical. From a static capacity point of view, the wet step is more utilized than the other ones (even if they are also highly utilized). Therefore, we decided to simulate the second part using the AnyLogic Simulation Tool.

For simplification, we decided to model this area as follows:

- Sputter and wet machines are modelled with their respective throughput.
- Material is started in front of the sputter machines. We start more material than sputter can absorb hence we want to model a fully utilized system. Material will leave after being processed at wet.
- Processing of additional material is blocked at sputter if the allowed WIP limit in front of the wet machines is reached or exceeded.
- We vary the following parameters. The current performance in our Fab is marked in bold. We vary a bit into each direction to see the potential or risk if a parameter changes:
    - sv is set to
        - 0,18/ WIP for ~3h continuous processing
        - 0,26/ WIP for ~ 4,2h continuous processing
        - 0,42/ WIP for ~ 6,7h continuous processing
    - Uptime at sputter 79%, **83%**, 87%
    - Mean time offline (MTOL) at sputter: 5h, **8h**, 11h
    - Mean time offline (MTOL) at wet: 6h, **9h**, 12h
- KPIs to evaluate performance are the timelink violations (should go close to zero) and the uptime utilization (UU) of the bottleneck process wet (should go close to 100%)
- One run of the model is stopped after 100 days of simulation time and 10 replications are made.
- Hence the material start rate is fixed the only source of uncertainty are the modelled down times of the sputter and wet machines.

Model validation has been done comparing capacity loss materializing as machine stand-by at wet at times the sputter tools are full of WIP. Results are shown in Table 1. It gets clear that the allowed WIP within the chain has a significant influence onto the uptime utilization of the bottleneck. If the mean time offline values get worse, a higher WIP is required to prevent wet from starvation.

It is clear that this behavior is somehow expected but this model enables us to see that even high WIP levels will not lead to timelink violations but are required, especially if the mean time offline and uptime get worse. When we were initially evaluating this timelink, the allowed WIP level were set to a very low level with the argumentation to minimize the risk. First, we started with measuring and visualizing the real data that we derived from our real lots for argumentation. But this simulation model holds additional advantages in contrast to simply setting parameters and test on a real wafer fab:

- Long term risk assessment, e.g. how much lots we will violate per year. Even if we run a real-world test and will not find any violation within a few weeks, line situation might change and a few lots are at risk for scrap within a longer timeframe.
- Estimation what will happen if performance, e. g. mean time offline or uptime will deteriorate.
- Model can be extended for every kind of timelink structure, even if a machine group handles timelink and non timelink material, the implemented prioritization strategies can be evaluated.
- What if we add/reduce a timelink at this point in the line can be evaluated.

In Tu and Liou (2006) and Tu and Chen (2009) an interesting approach is shown that calculates the capacity loss using a queueing model. However, there are two major differences. First, any kind of down events are not considered and secondly, no control approach is proposed. The capacity loss is due to possible rework of material that forces a second pass on the same operations.

Table 1: Result of simulation experiments.

Safety Value	SPU Uptime	SPU MTOL	WET MTOL	UUm	SB	PR	TL Violations	Avg Queue time	Safety Value	SPU Uptime	SPU MTOL	WET MTOL	UUm	SB	PR	TL Violations	Avg Queue time	Safety Value	SPU Uptime	SPU MTOL	WET MTOL	UUm	SB	PR	TL Violations	Avg Queue time
0.18	79	11	12	98.62%	1.19%	84.92%	0	1.51	0.26	79	11	12	99.22%	0.67%	85.44%	0	2.47	0.42	79	11	12	99.65%	0.30%	85.81%	0	4.01
0.18	79	11	9	98.70%	1.12%	85.05%	0	1.50	0.26	79	11	9	99.27%	0.63%	85.54%	0	2.50	0.42	79	11	9	99.70%	0.26%	85.91%	0	4.02
0.18	79	11	6	98.86%	0.98%	85.22%	0	1.45	0.26	79	11	6	99.37%	0.54%	85.66%	0	2.40	0.42	79	11	6	99.71%	0.25%	85.95%	0	3.94
0.18	79	8	12	99.00%	0.86%	85.25%	0	1.52	0.26	79	8	12	99.50%	0.43%	85.68%	0	2.51	0.42	79	8	12	99.80%	0.17%	85.94%	0	4.03
0.18	79	8	9	99.08%	0.80%	85.37%	0	1.58	0.26	79	8	9	99.57%	0.37%	85.80%	0	2.56	0.42	79	8	9	99.84%	0.14%	86.03%	0	4.02
0.18	79	8	6	99.08%	0.80%	85.41%	0	1.43	0.26	79	8	6	99.58%	0.37%	85.84%	0	2.41	0.42	79	8	6	99.86%	0.12%	86.08%	0	3.91
0.18	79	5	12	99.46%	0.47%	85.64%	0	1.54	0.26	79	5	12	99.79%	0.18%	85.93%	0	2.53	0.42	79	5	12	99.92%	0.07%	86.04%	0	4.07
0.18	79	5	9	99.50%	0.43%	85.74%	0	1.55	0.26	79	5	9	99.81%	0.17%	86.00%	0	2.52	0.42	79	5	9	99.92%	0.07%	86.10%	0	4.01
0.18	79	5	6	99.56%	0.38%	85.83%	0	1.47	0.26	79	5	6	99.86%	0.12%	86.09%	0	2.44	0.42	79	5	6	99.94%	0.06%	86.15%	0	3.93
0.18	83	11	12	99.21%	0.68%	85.43%	0	1.61	0.26	83	11	12	99.58%	0.37%	85.74%	0	2.61	0.42	83	11	12	99.82%	0.16%	85.95%	0	4.19
0.18	83	11	9	99.28%	0.62%	85.54%	0	1.59	0.26	83	11	9	99.60%	0.34%	85.82%	0	2.60	0.42	83	11	9	99.84%	0.14%	86.03%	0	4.20
0.18	83	11	6	99.37%	0.54%	85.66%	0	1.56	0.26	83	11	6	99.68%	0.28%	85.93%	0	2.56	0.42	83	11	6	99.86%	0.12%	86.09%	0	4.05
0.18	83	8	12	99.43%	0.49%	85.62%	0	1.60	0.26	83	8	12	99.72%	0.24%	85.87%	0	2.58	0.42	83	8	12	99.87%	0.12%	86.00%	0	4.16
0.18	83	8	9	99.56%	0.38%	85.78%	0	1.56	0.26	83	8	9	99.80%	0.17%	86.00%	0	2.55	0.42	83	8	9	99.90%	0.09%	86.08%	0	4.12
0.18	83	8	6	99.56%	0.38%	85.83%	0	1.56	0.26	83	8	6	99.79%	0.18%	86.03%	0	2.55	0.42	83	8	6	99.92%	0.07%	86.13%	0	4.08
0.18	83	5	12	99.76%	0.21%	85.90%	0	1.59	0.26	83	5	12	99.89%	0.10%	86.02%	0	2.58	0.42	83	5	12	99.93%	0.06%	86.05%	0	4.17
0.18	83	5	9	99.79%	0.18%	85.99%	0	1.58	0.26	83	5	9	99.91%	0.08%	86.09%	0	2.58	0.42	83	5	9	99.94%	0.05%	86.11%	0	4.12
0.18	83	5	6	99.79%	0.18%	86.03%	0	1.54	0.26	83	5	6	99.92%	0.07%	86.14%	0	2.51	0.42	83	5	6	99.94%	0.05%	86.16%	0	4.08
0.18	87	11	12	99.79%	0.18%	85.93%	0	1.64	0.26	87	11	12	99.87%	0.12%	86.00%	0	2.65	0.42	87	11	12	99.91%	0.07%	86.04%	0	4.26
0.18	87	11	9	99.84%	0.14%	86.03%	0	1.64	0.26	87	11	9	99.92%	0.07%	86.10%	0	2.65	0.42	87	11	9	99.94%	0.05%	86.11%	0	4.24
0.18	87	11	6	99.84%	0.14%	86.07%	0	1.62	0.26	87	11	6	99.91%	0.08%	86.13%	0	2.60	0.42	87	11	6	99.94%	0.06%	86.15%	0	4.21
0.18	87	8	12	99.85%	0.13%	85.98%	0	1.64	0.26	87	8	12	99.92%	0.07%	86.04%	0	2.65	0.42	87	8	12	99.94%	0.05%	86.06%	0	4.24
0.18	87	8	9	99.85%	0.13%	86.03%	0	1.63	0.26	87	8	9	99.91%	0.07%	86.09%	0	2.63	0.42	87	8	9	99.94%	0.05%	86.12%	0	4.24
0.18	87	8	6	99.85%	0.13%	86.07%	0	1.61	0.26	87	8	6	99.92%	0.07%	86.14%	0	2.61	0.42	87	8	6	99.94%	0.05%	86.15%	0	4.20
0.18	87	5	12	99.90%	0.09%	86.02%	0	1.64	0.26	87	5	12	99.93%	0.06%	86.05%	0	2.65	0.42	87	5	12	99.94%	0.05%	86.06%	0	4.24
0.18	87	5	9	99.92%	0.07%	86.10%	0	1.64	0.26	87	5	9	99.94%	0.05%	86.11%	0	2.65	0.42	87	5	9	99.94%	0.05%	86.12%	0	4.26
0.18	87	5	6	99.91%	0.08%	86.13%	0	1.61	0.26	87	5	6	99.94%	0.05%	86.15%	0	2.60	0.42	87	5	6	99.94%	0.05%	86.16%	0	4.21

7 SUMMARY AND OUTLOOK

We were able to prove that the implementation of a Decision Support System can help to optimize given timelinks. Finally, we optimized this critical timelinks in our Fab and reduced the capacity loss from around 10% close to zero. Furthermore, the cycle time in this area has significantly reduced. We were able to find the main weak spots and limitations with optimization potential. The main improvement points using this DSS is improved transparency and monitoring capability. The estimated root causes for 10% capacity loss at our main bottleneck within the chain are as follows:

- Uptime Stability at sputter: ~5%
- WIP limit between etch and sputter: ~0.5%
- Feed material from etch to sputter that matches the current process availability (according to the machine up state: ~1.5%
- Increase $\mathrm{WIP}_{allowed}$ of wet: ~1.5%
- Uptime stability at wet: ~1.5%

Besides the creation and deployment of this DSS, which is only necessary for critical timelinks with very complex interconnections between the influencing factors, we also set up standard reports for all timelinks that:

- track the cycle time of lots within timelinks and gives us a distribution in relation to the timelink restriction in order to evaluate the criticality of the current WIP limits/safety value and the possibility to increase sv,
- track the ratio of timelink gate blocked lots for each machine group and add this information to the charts in Figure 6 and
- use this data within our cycle time reporting to get clarity for root causes of cycle time peaks that are caused by machine groups within a timelink. You might notice that the WIP level and cycle time of a bottleneck within a timelink is very low hence the first step of the timelink chain is holding all of the WIP. By identifying the root cause, we can shift this WIP to the responsible bottleneck and get a clear estimation of the machine groups that generate cycle time.

We were also able to proof the procedure of using a DSS for increasing the WIP limits within the timelink. The simulation underlined the decisions and findings we made with help of the DSS. Of course, a DSS cannot prevent or forecast all influences, that may lead to an extension of the allowed time restriction. Therefore we still need buffer time (and additional capacity buffer) for the existing variability. Further simulation approaches with additional variability factors or forecasting methods e. g. for unscheduled tool downs are needed. Digital twins or Artificial Intelligence might also help to adjust limits more accurate and additional controlling strategies like scheduling.

A more operational focus point is the bottleneck management by using additional forecasting methods for identifying future bottleneck tools or evaluate capacity increase options.

We described different aspects of operative controlling possibilities for timelinks and also define a simulation model to proof our assumptions regarding the shown timelink system. As we already explained, the points mentioned should be considered in more detail because we see additional potential for further investigation.

REFERENCES

Anthouard, Benjamin, V. Borodin, Q. Christ, S. Dauzère-Pérès, and R. Roussel. 2022. “A Simulation-Based Approach for Operational Management of Time Constraint Tunnels in Semiconductor Manufacturing”. In *Proceedings of the Advanced Semiconductor Manufacturing Conference*, May 2nd – 5th, Saratoga Springs, New York, USA, 1-6

Arima, S., A. Kobayashi, Y. -F. Wang, K. Sakurai, and Y. Monma. 2015. “Optimization of Re-Entrant Hybrid Flows With Multiple Queue Time Constraints in Batch Processes of Semiconductor Manufacturing”. *IEEE Transactions on Semiconductor Manufacturing* 28(4):528-544.

Ignizio, J. P. 2009. *Optimizing Factory Performance: Cost-Effective Wasy to archive Significant and Sustainable Improvement*. New York: The McGraw-Hill Companies.

Kalir, A. A., and I. Tirkel. 2016. "Scheduling preventive maintenance within a queue time for maximum throughput in semiconductor manufacturing". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 2750–2761. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Klemmt, A, and L. Mönch. 2012. “Scheduling jobs with time constraints between consecutive process steps in semiconductor manufacturing”. In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A.M. Uhrmacher, 1-10. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Lima, A., V. Borodin, S. Dauzère-Pérès, and P. Vialletelle. 2017. “Analyzing Different Dispatching Policies for Probability Estimation in Time Constraint Tunnels in Semiconductor Manufacturing”. In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan, A. D’Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 3543–3554. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Ono, A., S. Kitamura, and K. Mori. 2006. “Risk Based Capacity Planning Method for Semiconductor Fab with Queue Time Constraints". *IEEE International Symposium on Semiconductor Manufacturing*, Tokyo, Japan, 49-52.

Robinson, J. K., and, R. Giglio. 1999. "Capacity Planning for Semiconductor Wafer Fabrication with Time Constraints between Operations". In *Proceedings of the 1999 Winter Simulation conference*, edited by P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 880-887. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Sadeghi, R., S. Dauzère-Pérès, C. Yugma, and G. Lepelletier. 2015. "Optimization of Re-Entrant Hybrid Flows With Multiple Queue Time Constraints in Batch Processes of Semiconductor Manufacturing". In *26th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, Saratoga Springs, New York, USA, 29-33.

Scholl, W., and J. Domaschke. 1999. "Implementation of Modeling and Simulation in Semiconductor Wafer Fabrication with Time Constraints between Wet Etch and Furnace Operations". In *1999 IEEE International Symposium on Semiconductor Manufacturing Conference Proceedings*, 61-63.

Tu, Y.-M., and C.-S. Liou. 2006. "Capacity Determination Model with Time Constraints and Batch Processing in Semiconductor Wafer Fabrication". *Journal of the Chinese Institute of Industrial Engineers* 23(3):192-199.

Tu, Y.-M., and H.-N. Chen. 2009. "Capacity Planning with Sequential two-level Time Constraints in the back-end Process of Wafer Fabrication". *International Journal of Production Research* 47(24):6967-6979.

Turban, E, R. Sharda, and D. Delen.(2010). Decision Support and Business Intelligence Systems. International Edition. 9th ed. New Jersey: Pearson.

Wang, M., S. Srivathsan, E. Huang, and K. Wu. 2018. "Dispatch Control for Production Lines With Overlapped Time Window Constraints". *IEEE Transactions on Semiconductor Manufacturing* 31(2):206-214.

Winkler T., Barthel P., and Sprenger, R. 2016. “Modelling of Complex Decision Making using Forward Simulation”. In *Proceedings of the Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick,, 2982–2991. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

AUTHOR BIOGRAPHIES

NINA DYBOWSKI is Industrial Engineer in the Industrial Engineering Department at Infineon in Dresden. She received a master’s degree in Industrial Engineering from the University of Applied Science Mittweida. Her research interests include industrial engineering in semiconductor manufacturing, timelinks and visualization. Her email address is Nina.Dybowski@infineon.com

MARIA SANDER is head of the Line Analysis team that is part of the Line Control Department at Infineon in Dresden. She received a master’s degree in Industrial Engineering from Dresden University of Technology. Her research interests include optimization of production lines in semiconductor manufacturing. Her email address is Maria.Sander@infineon.com

RALF SPRENGER is director of the Industrial Engineering Department at Infineon in Dresden. He received a Ph.D. from the Department of Mathematics and Computer Science at the University of Hagen, and a master’s degree in computer science at Dresden University of Technology. His research interests include industrial engineering in semiconductor manufacturing and optimization. His email address is Ralf.Sprenger@infineon.com