Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

INTEGRATING CRITICAL QUEUE TIME CONSTRAINTS INTO SMT2020 SIMULATION MODELS

Denny Kopp

Department of Mathematics and Computer Science
University of Hagen
Universitätsstraße 1
Hagen, 58097, GERMANY

Michael Hassoun

Department of Industrial Engineering and Management
Ariel University
Milken Campus (Upper), Building 8
Ariel, 40700, ISRAEL

Adar Kalir

Intel Corporation and Department of Industrial Engineering, Ben-Gurion University
Mailstop: LC2-3-ME
Qiriat-Gat, 82109, ISRAEL

Lars Mönch

Department of Mathematics and Computer Science
University of Hagen
Universitätsstraße 1
Hagen, 58097, GERMANY

ABSTRACT

In this paper, we study the impact of critical queue time (CQT) constraints in semiconductor wafer fabrication facilities (wafer fabs). Process engineers impose CQT constraints that require wafers to start a subsequent operation within a given time window after a certain operation is completed to prevent native oxidation and contamination effects on the wafer surface. We equip dataset 2 of the SMT2020 testbed with production control logic to avoid CQT constraint violations. Therefore, two different CQT-aware dispatching rules and a combination of a lot stopping strategy with these rules are proposed. The effect of the production control strategies is investigated by means of a simulation study. We show that the number of CQT violations can be reduced without large deteriorations of global performance measures such as cycle time and throughput.

1 INTRODUCTION

Semiconductor wafer fabrication facilities (wafer fabs) are complex manufacturing systems that contain hundreds of complicated and expensive machines. Wafer fabs can be modeled as complex job shops with a number of unusual characteristics, including reentrant product flows, very large numbers of unit operations, and complex technological constraints (Mönch et al. 2013). In particular, in order to prevent native oxidation and contamination effects on the wafer surface, process engineers impose CQT constraints that require wafers to start a subsequent operation within a given time window after a certain operation is completed (Scholl and Domaschke 2000). While there is a broad agreement that the proliferation of CQT constraints results in much more difficult scheduling problems (Klemmt and Mönch 2012; Jung et al. 2014), their effect on the performance of wafer fabs is hard to quantify.

CQT constraints are already included into the routes of the simulation models that belong to the SMT2020 testbed (Kopp et al. 2020). However, no production control logic is proposed so far that is able to treat the CQT constraints in an appropriate way. In the present paper, we propose simple production control strategies for this situation and assess their performance using a simulation model of the SMT2020 testbed. The proposed strategies can serve as benchmarks for more sophisticated production planning and control strategies that take into account the CQT constraints in wafer fabs.

The rest of the paper is organized as follows. The problem is described in the next section. This includes a discussion of related work. The proposed production control approaches are described in Section 3. The simulation model used for the experiments is discussed in Section 4. The results of the conducted simulation experiments are presented and analyzed in this section too. Conclusions and future research directions are discussed in Section 5.

2 PROBLEM DESCRIPTION

2.1 Problem Statement

CQT process segments are a disruptive feature of many modern wafer fabs. In this paper, a CQT segment is defined by an entrance operation and an exit operation for the set of all operations of a given product. A CQT violation occurs if the time span between the completion of the entrance step and the begin of processing of the exit step is longer than the prescribed CQT. It is likely that lots with violations of the prescribed CQT constraints have to be scrapped. Several CQT constraints are possible for the same route, i.e., the same product. We assume in this paper that overlapping, i.e. nested CQT constraints are not possible. However, the exit step of one CQT constraint can serve as entrance step of another CQT constraint. There is often an intra- and inter-CQT resource sharing, i.e., lots may share tools with other lots that are not constrained by CQTs or which are subject of other CQTs.

There are several approaches to deal with CQT constraints. One extreme approach is to release a lot only in a CQT segment if all required tools are available. This approach, also known as reservation (Zhang et al. 2016), ensures of course that under several conditions no violation occurs. At the same time, expensive tool capacity is eventually wasted. A slightly more general approach allows only a prescribed number of lots within the CQT segment. If the given threshold number is reached the lots are stopped at the begin of the entrance operation. Since this approach is similar to the Kanban approach in manufacturing systems, we refer to it as Kanban-type stopping approach. But again due to the intra- and inter resource sharing and machine breakdowns it is not obvious how to set an appropriate threshold value. A low threshold leads to wasting capacity whereas a large value might result in many CQT constraint violations.

Taking these challenges into account, we are mainly interested in proposing fairly simple dispatching strategies that take into account the length of the remaining CQTs. Moreover, when this approach still leads to many CQT violations, we are interested in combining it with a Kanban-type stopping approach. The strategies will be assessed by means of simulation experiments for dataset 2 of the SMT2020 testbed (Kopp et al. 2020). They can be used as a non-trivial benchmark for other more advanced approaches to respect CQT constraints. In a certain sense, this contributes to a more complete version of the SMT2020 testbed.

2.2 Related Work

There are only a few papers that address CQT constraints between process steps in wafer fabs. CQT constraints are considered in detailed fab scheduling approaches. Constraint programming is used by Choung et al. (2000) in a situation where CQT constraints between two consecutive process steps including a batching operation exist. Mason et al. (2007) design and assess a non-dominated sorting genetic algorithm to address scheduling problems where CQT constraints occur. Yurtsever et al. (2009) propose problem-specific iterative heuristics to solve scheduling problems with CQT constraints for the diffusion area in wafer fabs. Klemmt and Mönch (2012) consider flexible flow shop scheduling problems with CQT constraints. A decomposition approach based on mixed integer linear programming (MILP) is proposed. MILP approaches are also considered by Jung et al. (2014) for diffusion processes with CQT constraints in semiconductor manufacturing. Kalir and Tirkel (2016) study the problem of scheduling preventive maintenance activities on tools within CQT such that CQT restrictions are not violated and the

overall throughput is maximized. A MILP formulation for this problem is provided. A cross-entropy heuristic is designed for the efficient solution of this scheduling problem.

A list scheduling-type approach which also takes into account batching decisions is proposed for CQT-constrained wafer fabs by Zhang et al. (2016) and Pappert et al. (2016). Schedules are determined in an incremental manner on a lot-by-lot basis. Lots are stopped if a CQT constraint is violated according to the schedule. However, the interaction of local schedules for different tool groups in case of nested time constraints is not considered. CQT constraints between wet etch and furnace operations are discussed by Scholl and Domaschke (2000). A Kanban-type mechanism is proposed to form and start batches on the wet etch bench. However, only two consecutive operations are involved and determining an appropriate number of Kanban cards is challenging.

Sadeghi et al. (2015) propose a sampling technique to estimate whether a given lot may satisfy the CQT constraints or not. A list scheduling technique with random selection of the operations to be scheduled next based on a disjunctive graph representation of the complex job shop is used to predict the completion time of a lot. In a series of papers, Lima et al. (2017), (2019), and (2020) refine this approach to estimate the probability that multiple lots released at the entrance of a given CQT segment leave the segment without violating the CQT constraints. While the approach itself is interesting, up to now it cannot answer the question which is the preferred combination of lots to be released when the lots have different types, priorities, and belong to different CQT constraints.

3 PRODUCTION CONTROL APPROACHES FOR DEALING WITH CQT CONSTRAINTS

3.1 CQT-aware Dispatching Schemes

The main idea for the dispatching schemes consists in prioritizing lots that are within a CQT segment. If all lots in front a of certain tool group are not within a CQT segment the selection will be not affected. Otherwise, if there are also lots with an active CQT constraint these lots will be preferred. The choice of the next lot depends on how urgent is the lot with respect to the applied dispatching strategy.

In the following, two dispatching strategies are proposed. Both schemes are similar to strategies that focus on meeting global due dates. However, the main difference is that both strategies refer to local dates, i.e. deadlines, caused by CQT-constrained process steps.

The first approach works similar to critical ratio (CR)-type rules (Mönch et al. 2013). Once a lot enters a CQT segment by completing the entrance step (start date) s, the CQT will be added to determine the latest possible start time of the exit step n (end date) without a CQT violation. We refer to this rule as queue time critical ratio (QTCR) rule. The QTCR index of lot j is given as follows:

$$QTCR_j(t,i) := \begin{cases} \left(d_j^Q - t\right) \Big/ \sum_{k=i}^{n} p_{jk}, \text{if } t \le d_j^Q \\ \left(d_j^Q - t\right) \sum_{k=i}^{n} p_{jk}, \text{otherwise,} \end{cases} \tag{1}$$

where t is the current time and d_j^Q is the end date according to the CQT of lot j. Here, $s < i \le n$ and n are the current and the exit step of the CQT segment of lot j, respectively, and p_{ji} is the processing time of process step i of lot j. If $t \le d_j^Q$ lots with only a small amount of slack are preferred, while lots that are already strongly delayed with many remaining process steps are preferred for $d_j^Q > t$. A smaller index value of lots subject to a CQT constraint refers to a higher priority, while lots not belonging to a CQT constraint get a very high value by using for them an artificial d_j^Q value that is huge.

The second dispatching scheme is slack-based. It prioritizes the lot with the smallest slack to start the next process step. This rule is called QTS rule. Here, for all process steps belonging to a CQT segment,

the dates for starting the steps are determined for a lot that enters the segment. The latest possible start time of the exit step n (end date) is calculated as for the QTCR rule. If there exist one or more steps between the entrance step s and the exit step n, the dates for starting these steps will be computed as follows:

1. Estimate the total waiting (TW) time for all steps, i.e. the sum of waiting, transport, setup, and unloading/loading time, based on the processing time using local process step-specific flow factors (FF). The FF_{ji} values are calculated as the ratio of the average cycle time of all lots of product j and step i and the corresponding processing time p_{ji}. The TW value of process step $i, s < i \le n$ of lot j is given by

$$TW_{ji} := \lfloor FF_{ji} - 1 \rfloor p_{ji}. \tag{2}$$

2. Estimate the time span between the completion time of the entrance step s until the latest start time of the exit step n of lot j by

$$TT_j := \sum_{k=s+1}^{n-1} \left[TW_{jk} + p_{jk} \right] + TW_{jn} = \sum_{k=s+1}^{n-1} FF_{jk} \cdot p_{jk} + TW_{jn}. \tag{3}$$

3. Determine the ratio of the cycle time per step and the time span from Step 2 for all process steps before the exit step, i.e., we obtain for process step $s < i < n$

$$Ratio_{ji} := \left(FF_{ji} \cdot p_{ji} \right) / TT_j. \tag{4}$$

Moreover, we set $Ratio_{jn} = TW_{jn} / TT_j$ for the exit step. We clearly have $0 \le Ratio_{ji} \le 1$.

4. Compute the fraction of CQT, abbreviated by FCQT, that is allocated to process step $s < i \le n$ of lot j by multiplying the ratio and CQT, i.e.

$$FCQT(j,i) := CQT \cdot Ratio_{ji}. \tag{5}$$

5. Calculate the latest starting time for process step i by

$$d_{ji} := C_{js} + \sum_{k=s+1}^{i} FCQT_{jk} - p_{ji} \quad i = s+1, \ldots, n-1, \tag{6}$$

without violating the CQT. Here, C_{js} is the completion time of the entrance step s. We observe that we have $d_{ji} := C_{js} + CQT$ for the exit step n

The local FF values are required to incorporate appropriate estimated TW times. Thus, they ensure that an appropriate portion of the overall CQT is allocated to a given process step. The local FF values are determined from long simulation runs. The smallest d_{ji} value for all lots waiting for a tool group refers to the highest priority, and lots that are not subject to a CQT constraint get a very high value.

3.2 Combining Stopping Strategies with CQT-aware Dispatching

A stopping strategy makes sure that no further lots will be released into specific CQT segments if certain conditions are fulfilled. That means that lots waiting in front of a tool group corresponding to an entrance step of a CQT segment of the lots are not chosen for processing, i.e., they are on hold. The proposed stopping strategy does not rely on a single threshold number of lots for each individual CQT segment since such an approach does not account for the possible intra- and inter-CQT resource sharing.

Longer waiting times are caused by a larger number of lots waiting in front of a tool group. Moreover, there are also waiting lots that are not subject to CQT constraints, i.e., they are irrelevant for the CQT segments. Therefore, only lots subject to a CQT are counted for the threshold. The threshold is the maximum number of CQT lots before a tool group that can be processed there to make a trade-off between CQT violations and wasting capacity. Based on these insights, our stopping strategy considers maximum threshold numbers of lots subject to a CQT constraint individually for each tool group of a CQT segment. It must be combined with a CQT-aware dispatching strategy since the tools are shared by CQT-constrained and –unconstrained lots.

Two threshold values are used for each tool group. The first one refers to the number of constrained lots directly in front of the tool group including lots that are currently processed on the tools. The second threshold adds to the first threshold the number of all lots which are currently in CQT segments on steps before the process step that corresponds to the tool group. That means that the second threshold incorporates also upstream lots. If at least one threshold value is reached or exceeded for a tool group, no lot will be released into all CQT segments related to this tool group, i.e., before the entrance step of a CQT segment is processed for a lot, all tool groups that belong to this segment are checked. The threshold values will be ignored if the lot leaves a CQT segment, i.e., the exit step of the current segment is the entrance step of the next consecutive segment. Setting appropriate threshold values is a non-trivial task, the values depend on factors like number of tools in a tool group or the processing times.

4 SIMULATION EXPERIMENTS

4.1 Simulation Model

Dataset 2 of the SMT2020 testbed is used within the experiments. It represents a low-volume/high-mix wafer fab. Ten products with routes having 242 up to 583 process steps are included. The model contains 105 tool groups with a total of 913 tools. Exponentially distributed tool breakdowns and time-based as well as counter-based preventive maintenance (PM) are considered. Important additional features of modern wafer fabs are modeled, such as cascading tools, batch processing tools, tools with significant sequence-dependent setup times, and lot-to-lens dedications for the steppers. Moreover, reentrant process flows with up to 44 mask layers exist in the dataset. In addition, for all ten products CQT segments are included in the model with up to 42 segments per product. All lots have 25 wafers. 1,000 wafers per product are released per week, i.e. a total of 400 lots, that leads to a target bottleneck utilization of around 97% for some tool groups in the lithography area. 2.5% of all lots are hot lots with a higher priority than regular lots (cf. SMT2020 Testbed 2020 for a detailed description of dataset 2).

A baseline dispatching strategy, abbreviated by BASE, is applied that considers several criteria. It does not take into account CQTs. First, lots with a higher priority are preferred, while the priority of hot lots and regular lots are 20 and 10, respectively. Among lots with the same priority, lots with the smallest setup time are preferred. Hence, hot lots can trigger setups. Finally, the CR rule is applied. BASE is extended to deal with CQT constraints by incorporating the CQT-aware dispatching schemes from Subsection 3.1. However, especially hot lots will be chosen first regardless of lots with existing CQT constraints. The different components to determine which lot should be processed next are applied in the following order:

1. Lot priority
2. Least setup time
3. QTCR/QTS
4. CR.

Batches are formed by determining a first lot according to the dispatching scheme. If more appropriate lots are available they will be used to fill the batch. The minimum batch size is respected. On the one hand, it is likely that a dispatching strategy that prefers CQT-constrained lots works well if among

the lots waiting in front of a tool group there are many lots without CQT constraints. On the other hand, if almost all lots in front of a tool group are CQT-constrained, a stopping strategy is desirable since a pure CQT-aware dispatching strategy is not able to differentiate between the lots and avoid CQT violations. To create such a more complex situation, new CQT segments are included in dataset 2 in addition to default segments of dataset 2 (cf. SMT2020 Testbed 2020). The additional segments contain process steps on the LithoTrack_FE_95 and LithoTrack_FE_115 tool groups with 40 and 41 tools, respectively. Both tool groups are formed by steppers and have a target bottleneck utilization of around 97%. All process steps related to these tool groups are within a CQT segment. Hence, the total number of segments for all products increases from 274 to 441 segments. The stopping strategy from Subsection 3.2 is applied together with the BASE, QTCR, and QTS dispatching schemes. The number of CQT segments of the ten products is summarized in Table 1.

Table 1: Number of CQT segments per product.

#CQT segments	**Product**									
	1	2	3	4	5	6	7	8	9	10
default	30	39	42	26	12	28	25	21	24	27
complex	50	61	65	43	23	41	39	37	39	43

4.2 Design of Experiments

The goal of a first series of simulation experiments is to study the impact of the CQT-aware dispatching schemes with respect to global performance measure values and CQT violations. We expect that the performance depends on the number and the complexity of interrelated CQT segments. The design of experiments is summarized in Table 2.

Table 2: Design of experiments.

Factor	**Level**	**Count**
dispatching scheme	BASE, QTCR, QTS	3
CQT constraint setting	default, complex	2
# of independent simulation replications		10
total # of simulation runs		60

In a second set of experiments, we are interested in analyzing the impact of the applied stopping strategy in a situation with significantly more CQT segments, especially for the tool groups Litho-Track_FE_95 and LithoTrack_FE_115. We expect that the performance depends on the threshold setting for these two tool groups. A high threshold value of 1,000 lots is used for all tool groups except for Litho-Track_FE_95 and LithoTrack_FE_115 where tailored values are used. The applied threshold values are collected in Table 3.

Table 3: Configuration of the stopping scenarios.

Tool group	**Stopping scenario (first threshold value/second threshold value)**			
	none	**high**	**medium**	**small**
LithoTrack_FE_95	1,000/1,000	90/130	60/95	50/85
LithoTrack_FE_115	1,000/1,000	90/220	70/150	55/125
remaining tool groups	1,000/1,000	1,000/1,000	1,000/1,000	1,000/1,000

The provided numbers refer to the first and second threshold value, respectively. Not a single stop can be observed when high threshold values are applied for all tool groups (none scenario). A total of 90

simulation runs is conducted since the stopping strategy is combined with all dispatching schemes. Again, ten independent simulation runs are performed for each factor combination.

We report throughput (TH) values, i.e. the number of completed lots within the simulation horizon, the average cycle time (ACT) in days, and the percentage of on-time lots (%ONTIME). The fraction of violations relative to the total number of lots that complete a CQT segment is also reported. This quantity is abbreviated by %VL. In addition, the percentage of violations longer than 1 hour (%VL1h), 2 hours (%VL2h), and 4 hours (%VL4h) are gathered. Moreover, we measure the average duration of the violations (AVL) in hours and also the average earliness of lots with respect to the CQT constraints (AONT) in hours.

A simulation horizon of two years is applied in all experiments. This includes a warm-up phase of one year which is excluded from the statistics for ACT values and %ONTIME. The simulation starts with initial WIP lots that align with the target bottleneck utilization. We provide average values based on the ten independent simulation replications. The experiments are performed using the commercial simulation engine AutoSched AP 11.3.0. The QTCR and QTS schemes and the stopping strategy are coded in the C++ programming language using customization features of the AutoSched AP software.

4.3 Simulation Results

The detailed simulation results for the different dispatching schemes for the default CQT situation are shown in Table 4. We group the results according to products and lot types. Production regular lots (PRL) for product i are abbreviated by Lot i, whereas production hot lots (PHL) of product i are indicated by Hotlot i.

Table 4: Detailed simulation results for the default CQT constraint situation.

Product/Lot type			BASE			QTCR			QTS		
			TH	ACT (days)	% ONTIME	TH	ACT (days)	% ONTIME	TH	ACT (days)	% ONTIME
PL	PRL	Lot 1	3745	47.8	97.7	3746	47.5	98.8	3741	47.8	96.8
		Lot 2	3738	49.8	98.0	3739	49.5	99.0	3735	49.7	96.8
		Lot 3	3709	53.5	97.9	3710	53.2	99.0	3706	53.4	97.0
		Lot 4	3846	30.0	98.9	3844	29.9	98.8	3842	30.1	98.0
		Lot 5	3888	22.1	97.6	3889	21.9	98.1	3888	22.0	96.2
		Lot 6	3848	29.2	71.3	3844	29.1	71.6	3841	29.3	67.1
		Lot 7	3830	32.9	98.0	3833	32.4	99.9	3831	32.5	99.8
		Lot 8	3820	35.6	97.9	3818	35.5	98.8	3814	35.8	96.8
		Lot 9	3793	38.7	70.9	3793	38.5	71.0	3789	38.7	66.7
		Lot 10	3808	37.9	99.1	3807	37.8	98.8	3805	38.0	97.8
	PHL	Hotlot 1	99	30.2	73.1	99	30.3	72.5	99	30.4	74.7
		Hotlot 2	99	32.6	73.7	99	32.7	73.8	99	32.6	75.6
		Hotlot 3	98	34.4	79.6	99	34.8	78.0	98	34.5	77.6
		Hotlot 4	100	20.2	82.8	101	20.3	82.0	101	20.3	81.1
		Hotlot 5	101	14.4	85.7	101	14.3	86.1	101	14.4	85.2
		Hotlot 6	101	17.9	85.1	101	18.1	82.7	101	18.1	83.9
		Hotlot 7	100	21.2	78.1	100	21.2	78.4	100	21.3	75.4
		Hotlot 8	100	22.7	78.2	100	22.7	78.8	100	22.7	78.0
		Hotlot 9	100	23.5	79.1	100	23.6	79.9	100	23.6	78.0
		Hotlot 10	100	24.8	77.7	100	24.7	79.7	100	24.6	78.0

Next, we provide aggregated, not product-specific simulation results for both the default and the complex CQT situations with respect to the lot types in Table 5. The results are grouped according to

PRL and PHL. We also provide results for the CQT violations where we group them into the Litho, rest, and total CQT segment sets. The latter consists of all CQT segments, while Litho refers to all segments that include the LithoTrack_FE_95 and LithoTrack_FE_115 tool groups. Rest is the set difference of total and Litho.

Table 5: Simulation results for the default and complex CQT situation.

Scheme	Lot type	TH	ACT (days)	% ONTIME	CQT Setting	% VL	% VL1h	% VL2h	% VL4h	AVL (h)	AONT (h)
Default CQT situation – None stopping scenario											
BASE	PRL	3802	37.7	92.7	Litho	18.6	16.4	14.5	11.5	1.46	3.22
					Rest	17.4	15.4	14.0	11.7	1.99	2.17
	PHL	100	24.2	79.3	Total	17.5	15.6	14.1	11.7	1.92	2.31
QTCR	PRL	3802	37.5	93.4	Litho	1.2	0.8	0.6	0.4	0.06	4.19
					Rest	10.6	8.8	7.4	5.5	0.81	2.31
	PHL	100	24.3	79.2	Total	9.4	7.7	6.5	4.8	0.71	2.56
QTS	PRL	3799	37.7	91.3	Litho	1.1	0.7	0.6	0.4	0.05	4.19
					Rest	10.5	8.8	7.4	5.5	0.83	2.32
	PHL	100	24.3	78.8	Total	9.3	7.7	6.5	4.9	0.73	2.56
Complex CQT situation – None stopping scenario											
BASE	PRL	3802	37.7	92.7	Litho	25.6	21.8	19.4	16.0	2.71	1.96
					Rest	17.2	15.3	13.8	11.5	1.95	2.18
	PHL	100	24.2	79.3	Total	21.1	18.3	16.4	13.6	2.30	2.07
QTCR	PRL	3804	37.6	94.0	Litho	16.7	9.9	6.4	3.4	0.57	1.87
					Rest	10.9	9.1	7.7	5.7	0.85	2.27
	PHL	100	24.4	77.6	Total	13.6	9.4	7.1	4.6	0.72	2.09
QTS	PRL	3803	37.5	93.8	Litho	17.2	10.2	6.6	3.4	0.58	1.89
					Rest	10.8	9.0	7.6	5.6	0.84	2.28
	PHL	100	24.4	77.6	Total	13.8	9.6	7.1	4.6	0.72	2.10

Moreover, the results of the simulation runs for the different stopping scenarios based on the complex CQT situation are shown in Table 6. The aggregated ACT values of all dispatching schemes for PRL are provided in Figure 1 where we group the results according to both CQT situations and all stopping scenarios. The percentage of violations for the lithography CQT segments is shown in Figure 2.

4.4 Discussion of the Simulation Results

We see from Table 4 that the obtained performance measure values are very similar for all dispatching schemes. At the same time, considering the CQT violation results from Table 5 we observe much better results when QTCR or QTS are applied instead of BASE. We have to state that a deterioration for these performance measure is observed from some preliminary simulation experiments for dataset 1, a high-volume/low-mix wafer fab, with only two products when QTCR or QTS is used. This behavior is mainly caused by applying the First In First Out (FIFO) dispatching rule that replaces the CR rule applied in dataset 2.

In the complex CQT situation with considerably more CQT segments, we see from Table 5 that for all dispatching schemes more violations occur, especially for the lithography segments. The ACT values are still very similar when QTCR or QTS are used (see Table 5 or Figure 1). Moreover, the CQT-related performance is worse with BASE. We observe from Figure 2 that the percentage of violations for lithography segments is much higher in the complex situation. For instance, the %VL values for QTS increases from 1.1% in the default situation to 17.2% in the complex situation for the lithography CQT

segments, while the corresponding value only increases from 10.5% to 10.8% for the remaining segments (rest) (see Table 5). This behavior is caused by the fact that the treatment of lots within these segments is more difficult since all process steps on LithoTrack_FE_95 or LithoTrack_FE_115 are within a CQT segment. Thus, long waiting times due to long queues in front of these tool groups are caused by lots that are subject of CQT constraints. The stopping strategy can be used to limit the number of waiting lots. Table 6 and Figure 2 show that the defined threshold values of the different stopping scenarios lead to better results with respect to the number of violations for all dispatching schemes.

Table 6: Simulation results for the different stopping scenarios in the complex CQT constraint situation.

Scheme	Lot type	TH	ACT (days)	% ONTIME	CQT setting	% VL	% VL1h	% VL2h	% VL4h	AVL (h)	AONT (h)
High stopping scenario											
BASE	PRL	3806	37.6	93.1	Litho	23.6	19.4	16.9	13.5	2.14	1.99
					Rest	17.0	15.0	13.6	11.3	1.92	2.18
	PHL	100	24.2	79.8	Total	20.0	17.1	15.2	12.4	2.02	2.09
QTCR	PRL	3803	37.6	93.3	Litho	10.7	5.5	4.1	2.9	0.45	1.95
					Rest	11.0	9.2	7.8	5.8	0.85	2.27
	PHL	100	24.4	77.6	Total	10.9	7.5	6.1	4.4	0.67	2.12
QTS	PRL	3800	37.5	94.1	Litho	12.6	5.8	4.0	2.8	0.46	1.95
					Rest	10.8	9.1	7.7	5.8	0.85	2.28
	PHL	100	24.4	78.5	Total	11.6	7.5	6.0	4.4	0.67	2.12
Medium stopping scenario											
BASE	PRL	3732	47.0	0.5	Litho	17.1	13.5	11.4	8.7	1.36	2.26
					Rest	13.0	11.5	10.3	8.6	1.45	2.26
	PHL	100	24.4	77.3	Total	14.9	12.4	10.8	8.6	1.41	2.26
QTCR	PRL	3803	37.7	93.8	Litho	6.9	4.7	3.7	2.7	0.41	2.21
					Rest	11.0	9.2	7.8	5.8	0.85	2.27
	PHL	100	24.4	77.6	Total	9.1	7.1	5.9	4.3	0.64	2.24
QTS	PRL	3800	37.6	93.6	Litho	7.1	4.7	3.8	2.6	0.41	2.21
					Rest	10.8	9.0	7.6	5.6	0.83	2.28
	PHL	100	24.4	77.6	Total	9.0	7.0	5.8	4.2	0.63	2.24
Small stopping scenario											
BASE	PRL	3598	63.5	0.1	Litho	13.7	10.3	8.5	6.4	1.05	2.40
					Rest	12.3	10.8	9.8	8.1	1.42	2.28
	PHL	100	24.5	76.7	Total	13.0	10.6	9.2	7.3	1.24	2.34
QTCR	PRL	3800	38.4	89.5	Litho	6.1	4.3	3.5	2.5	0.39	2.43
					Rest	10.7	9.0	7.6	5.7	0.83	2.27
	PHL	100	24.5	77.2	Total	8.6	6.8	5.7	4.2	0.63	2.35
QTS	PRL	3797	38.4	89.9	Litho	6.1	4.3	3.5	2.5	0.39	2.43
					Rest	10.7	8.9	7.6	5.7	0.84	2.28
	PHL	100	24.5	76.8	Total	8.5	6.8	5.7	4.2	0.63	2.35

The ACT values are almost the same in the case of a fairly high, but restricting threshold value. However, the applied dispatching scheme is crucial if smaller threshold values are used. We can see from Figure 1 and Table 6 that the ACT values for PRL using BASE are very large in the medium scenario, while the ones from QTCR and QTS are still almost on the same level. In the small stopping scenario, the global performance is the worst for QTCR and QTS compared to the remaining stopping scenarios. But the smallest percentage of violations is observed in this situation (see Figure 2). Note that violations might be caused by other factors, for instance the batch minimum. If the threshold values are even smaller

than in the last situation, this results in very high ACT values for QTCR and QTS since too long queues occur before the entrance step of the corresponding CQT segments.

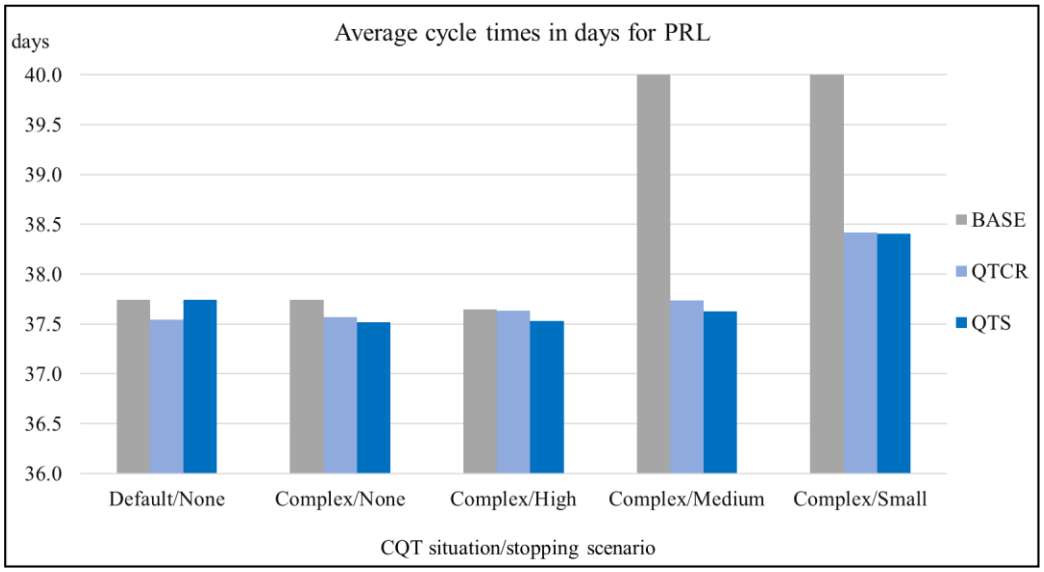

Figure 1: Average cycle time in days for PRL.

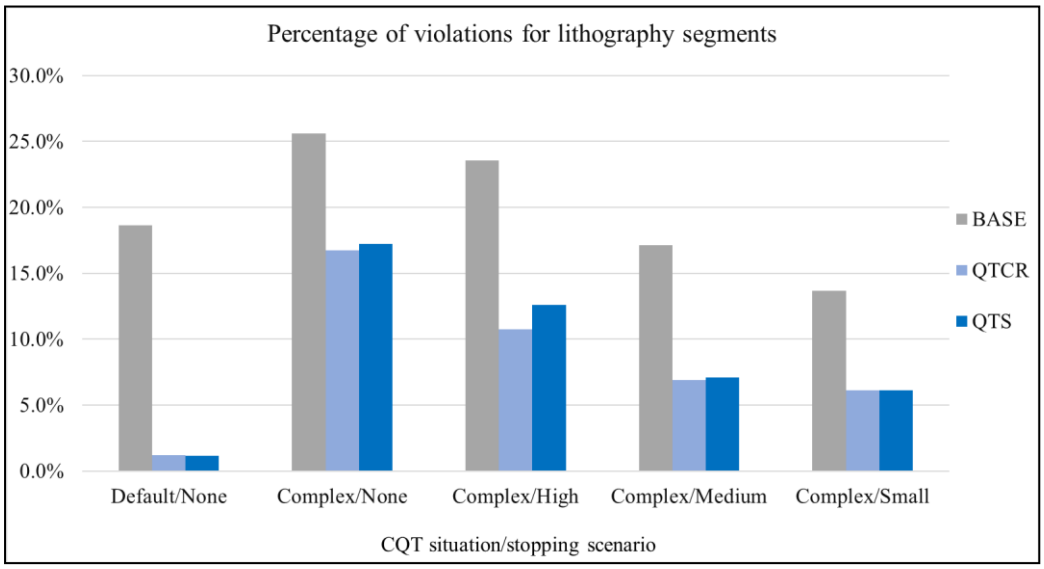

Figure 2: Percentage of violations for lithography CQT segments.

We can see from the results depicted in Figures 1 and 2 that the combination of the stopping mechanism with the CQT-aware dispatching schemes works better than with the BASE strategy that neglects CQT constraints. Because, on the one hand, the percentage of violations is better due to the prioritization of the lots. On the other hand the latter treatment leads to an accelerated resolution of stopping situations. Therefore, smaller threshold values can be used by ensuring appropriate ACT values at the same time. For instance, the ACT values for QTCR are very similar for the high and the medium stopping scenario, e.g. for PRL it merely raises from 37.6 to 37.7 days (see Table 6). Even with small threshold values, the ACT values of PRL only increases to 38.4 days for QTCR (see Figure 1). In contrast, if the BASE strategy is applied in combination with the stopping strategy we see from Table 6 that the

ACT values of PRL increase from 37.6 days to 63.5 days for the high and the small stopping scenario, respectively. Finally, we see from Tables 5 and 6 that the ACT values of hot lots are almost the same in all situations. This is an expected behavior that is caused by the described dispatching strategy that prefers hot lots. Overall, we see from the tables and the figures that the CQT-aware dispatching schemes lead to very similar results in both CQT situations and for all stopping scenarios with respect to global performance measure values and CQT violations.

5 CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In this paper, we studied the behavior of a wafer fab with CQT constraints under different CQT-aware dispatching and lot stopping strategies. Therefore, we used one of the simulation models of the SMT2020 testbed. The two proposed dispatching strategies take into account the length of the remaining time window. They were able to eliminate the CQT constraint violations to a large extent if the number of CQT constraints is not too large. We demonstrated by simulation experiments that an additional stopping strategy in front of the work centers where CQT constraints arise was beneficial if the number of CQT constraints is large. In this situation, the combination of stopping and the proposed dispatching strategies was beneficial, i.e., we were able to avoid a large portion of the CQT constraint violations while the corresponding ACT and TH values were only slightly changed.

There are several directions for future research. First, it is desirable to tackle the problem of setting an appropriate number of Kanbans for the CQT segments of a wafer fab. We believe that this can be done based on simulation-based optimization. Moreover, the SMT2020 simulation models and the proposed production control strategies can be used to assess deterministic scheduling approaches that are able to consider CQT constraints in a rolling horizon setting, i.e., the schedules are executed in the simulation model.

ACKNOWLEDGMENTS

The research was partially supported by the iDev 4.0 project. The iDev40 project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 783163. The JU receives support from the European Union's Horizon 2020 research and innovation programme. It is co-funded by the consortium members, grants from Austria, Germany, Belgium, Italy, Spain and Romania. The authors gratefully acknowledge this financial support.

REFERENCES

Choung, Y.-I., K.-S. Jun, D.-S. Han, Y.-C. Jang, T.-E. Lee, and R. C. Leachman. 2001. "Design of a Scheduling System for Diffusion Processes". In *Proceedings International Conference on Modeling and Analysis of Semiconductor Manufacturing (MASM)*, 69-73.

Jung, C., D. Pabst, M. Ham, M. Stehli, and M. Rothe. 2014. "An Effective Problem Decomposition Method for Scheduling of Diffusion Processes Based on Mixed Integer Linear Programming". *IEEE Transactions on Semiconductor Manufacturing* 27(3): 357-363.

Kalir, 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, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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, 2173-2182. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc..

Kopp, D., M. Hassoun, A. Kalir, and L. Mönch. 2020. "SMT2020 – A Semiconductor Manufacturing Testbed". *IEEE Transactions on Semiconductor Manufacturing*, in press.

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, New Jersey: Institute of Electrical and Electronics Engineers, Inc..

Lima, A., V. Borodin, S. Dauzère-Pérès, and P. Vialletelle. 2019. "Sampling-based Release Control of Multiple Lots in Time Constraint Tunnels*". Computers in Industry* 119:3-11.

Lima, A., V. Borodin, S. Dauzère-Pérès, and P. Vialletelle. 2020. "A Sampling-based Approach for Managing Lot Release in Time Constraint Tunnels in Semiconductor Manufacturing". *International Journal of Production Research*, in press.

Mason, S. J., M. Kurz, L. M. Pohl, J. W. Fowler, and M. E. Pfund. 2007. Random Keys Implementation of NSGA-II for Semiconductor Manufacturing Scheduling. *International Journal of Information Technology and Intelligent Computing* 2(3).

Mönch, L., J. W. Fowler, and S. J. Mason. 2013. *Production Planning and Control for Semiconductor Wafer Fabrication Facilities: Modeling, Analysis, and Systems*. New York: Springer.

Pappert, F., T. Zhang, F. Suhrke, J. Mager, T. Frey, and O. Rose. 2016. "Impact of Time Bound Constraints and Batching on Metallization in an Opto-semiconductor Fab". In *Proceedings of the 2009 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 2947-2957. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc..

Sadeghi, R., S. Dauzère-Pérès, C. Yugma, and G. Lepelletier. 2015. "Production Control in Semiconductor Manufacturing with Time Constraints". In *Proceedings 26th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC*), 29-33.

Scholl, W., and J. Domaschke. 2000. "Implementation of Modeling and Simulation in Semiconductor Wafer Fabrication with Time Constraints Between Wet Etch and Furnace Operations". *IEEE Transactions on Semiconductor Manufacturing* 13(3): 273-2277.

SMT2020 Testbed. 2020. SMT2020 Semiconductor Manufacturing Testbed General Data specification. http://p2schedgen-.fernuni-hagen.de/index.php?id=simulation&L=1. accessed 15th May.

Yurtsever, T., E. Kutanoglu, and J. Johns. 2009. "Heuristic Based Scheduling System for Diffusion in Semiconductor Manufacturing". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 1677-1685. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc..

Zhang, T., F. Pappert, and O. Rose. 2016. "Time Bound Control in a Stochastic Dynamic Wafer Fab". In *Proceedings of the 2009 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 2903-2911. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc..

AUTHOR BIOGRAPHIES

DENNY KOPP is is a Ph.D. student at the Chair of Enterprise-wide Software Systems, University of Hagen. He received MS degree in Business Economics from the Otto-von-Guericke-University Magdeburg, Germany. His research interests include applied optimization and simulation-based production control. His email address is Denny.Kopp@fernuni-hagen.de.

MICHAEL HASSOUN is a senior lecturer at the Industrial Engineering Department of the Ariel University, Israel. His research interests focus on modeling and management of production systems, with a special interest in Semiconductor manufacturing. He earned his PhD and MSc in Industrial Engineering from Ben-Gurion University of the Negev, Israel, and his BSc in Mechanical Engineering from the Technion, Israel. He was a postdoctoral fellow at the University of Michigan in 2009. His email address is michaelh@ariel.ac.il.

ADAR KALIR received his B.S. and M.S. degrees in industrial engineering and management from Tel-Aviv University, Israel, and his Ph.D. degree in industrial and systems engineering from Virginia Tech. He is a Sr. Principal Engineer at the Fab/Sort Manufacturing network of Intel Corp., responsible for the application of operational optimization in high volume manufacturing across Intel's factories, driving improvements in WIP management, production capacity and cycle time, equipment and capital productivity. He is also an Adjunct Associate Professor at Ben-Gurion University, Israel and serves as a co-chair of the IEEE Technical Committee on Semiconductor Manufacturing Automation (TC-SMA). His email is kalira@post.bgu.ac.il.

LARS MÖNCH is Professor in the Department of Mathematics and Computer Science at the University of Hagen, Germany. He received a master's degree in applied mathematics and a Ph.D. in the same subject from the University of Göttingen, Germany. His current research interests are in simulation-based production control of semiconductor wafer fabrication facilities, applied optimization and artificial intelligence applications in manufacturing, logistics, and service operations. He is a member of GI (German Chapter of the ACM), GOR (German Operations Research Society), and INFORMS. He is an Associate Editor for the European Journal of Industrial Engineering, Journal of Simulation, Business & Information Systems Engineering, IEEE Robotics and Automation Letters, IEEE Transactions on Semiconductor Manufacturing, and IEEE Transactions on Automation Science and Engineering. His email address is lars.moench@fernuni-hagen.de.