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

ADAPTIVE RULE BASED ORDER RELEASE IN SEMICONDUCTOR MANUFACTURING

Philipp Neuner

Department of Information Systems, Production and Logistics Management
University of Innsbruck
Innsbruck, 6020, AUSTRIA

ABSTRACT

This paper analyzes two periodic order release mechanisms and their promising extensions by using a simulation model of a scaled-down wafer fabrication facility. One extends the Backward Infinite Loading (BIL) approach by dynamically adjusting lead times and considering safety lead times, and the other extends the COrrected aggregate Load Approach (COLA) by incorporating a dynamic time limit into its release procedure (Overload). Both are periodic approaches aiming at improving the timing performance and can react to the dynamics on the shop floor, where semiconductor manufacturing provides a very challenging environment. The results show that Overload outperforms all other mechanisms by yielding less total costs mainly due to a more balanced shop which results in the lowest WIP costs. Further, Overload reduces inventory costs compared to BIL and COLA. These results reinforce the finding of previous research that *periodic* rule based order release models are a viable alternative for semiconductor manufacturing.

1 INTRODUCTION

A key task in manufacturing planning and control is the order release decision which determines when and which orders should be released to the shop floor. In this regard, the Workload Control (WLC) concept was developed based on the idea of controlling order releases to control shop floor throughput times, the level of Work-In-Process (WIP) and the output (Kingsman et al. 1989; Wiendahl 1995). In general, order release mechanisms can be divided into *periodic* and *continuous* models. While the former approaches release orders at periodic intervals only, the latter models release orders at any moment in time (Thuerer et al. 2014). Although WLC approaches have been tested in semiconductor manufacturing, the main focus was on continuous mechanisms (Glassey and Resende 1988; Wein 1988; Spearman et al. 1990). In this regard, Fowler et al. (2002) reviewed promising WLC approaches applied to semiconductor manufacturing. More precisely, Wein (1988) introduced Workload Regulation (WR) which focuses on controlling the bottleneck workload on the shop floor. Similarly, Glassey and Resende (1988) developed Starvation Avoidance (SA) which aims at keeping the number of orders on their way to the bottleneck at a pre-defined level to avoid the bottleneck from starving. Different to WR, SA only considers the work in the system that is within the lead time of the bottleneck. Later, the Drum-Buffer-Rope concept and the related Optimized Production Technology were introduced (Jacobs 1984; Goldratt and Cox 1986). Another famous approach is Constant Work-In-Process (ConWIP) which was developed by Spearman et al. (1990) and controls the number of orders released to the shop floor, i.e. the WIP. Rose (1999) adapted ConWIP such that the workload of the bottleneck is regulated rather than simply the number of orders, which resulted in the Constand Load (ConLOAD) approach. While WR simply sums up the bottleneck processing times of the released orders, ConLOAD divides each order's sum of bottleneck processing times by the average cycle time (hereinafter shop floor throughput time SFTT) of the respective product type to calculate the load contributions (Wein 1988; Rose 1999).

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

With regard to *periodic* order release approaches, almost only optimization based models were tested in semiconductor manufacturing (Hackman and Leachman 1989; Hung and Leachman 1996; Albey and Uzsoy 2015). Note that a review of such models is out of the scope of this paper, thus the interested reader is referred to Missbauer and Uzsoy (2011) and Haeussler et al. (2020). However, previous research analyzed and improved *periodic* approaches mainly in the domain of small and medium enterprises (SME) in make-to-order (MTO) environments (Thuerer et al. 2011; Hutter et al. 2018) which led to significant enhancements (Oosterman et al. 2000; Thuerer et al. 2012). One noteworthy exception is the paper by Neuner et al. (2020) who showed that the periodic rule-based COrrected aggregate Load Approach (COLA) is applicable to the semiconductor domain. The authors also demonstrated the potential of a static tight time limit within the COLA approach, which further restricts the set of orders in the order pool that are considered for order release, i.e. only orders whose planned release dates are within the time limit (Wiendahl 1995; Haeussler and Netzer 2020; Neuner et al. 2020; Haeussler et al. 2021). Regarding the use of a time limit, Haeussler et al. (2021) conceptualized an adaptive time limit policy for the hybrid LUMS-COR approach (periodic and continuous release elements) in the SME-MTO domain. Their developed Overload approach, which only applies a tight time limit during high load periods, outperforms all other tested time limit policies in terms of load balancing and timing performance. Another interesting approach to improve the timing performance was developed by Haeussler et al. (2019), who introduced a dynamic Backward Infinite Loading (BIL) approach based on Exponential Smoothing including Safety Lead Times (i.e ESSLT) which considers dynamic lead times to calculate planned release dates.

The adaptive Overload and ESSLT approaches are both able to react to changes on the shop floor. However, an analysis of which approach performs better was not performed until now. In this regard, due to its unique characteristics (e.g. bottleneck system, multiple re-entrant product routings, machine failures and work centers performing batch processing), semiconductor manufacturing provides a very challenging environment. Therefore, this paper analyzes the adaptive Overload (Haeussler et al. 2021) and the dynamic ESSLT approach (Haeussler et al. 2019) in the semiconductor domain by using a simulation model of a scaled down wafer fabrication facility (Kayton et al. 1997). As benchmark scenarios, BIL and COLA are considered, where COLA was already shown to yield superior performance compared to two well-established continuous approaches from the semiconductor domain - SA and ConLOAD (see Neuner et al. 2020). Performance will be measured threefold: First, by cost-based measures consisting of WIP, Finished Goods Inventory, and backorder costs, second, by the mean and standard deviation of shop floor throughput time and bottleneck queue time which capture the load balancing performance among the work centers and production routes, and third, by the percentage of tardy orders indicating the service level.

The remainder of this paper is structured as follows: Section 2 introduces the tested order release mechanisms, and in Section 3 the used simulation model and the experimental design are outlined. In Section 4 the results are presented before final conclusions and limitations are provided in Section 5.

2 ORDER RELEASE APPROACHES

2.1 Backward Infinite Loading (BIL)

With Backward Infinite Loading (BIL), a planned release date (PRD) is calculated for all orders in the order pool considering planned lead times (LT) for these orders. At periodic intervals, the release procedure checks whether the PRD of an order is reached and releases these orders to the shop floor. The sequence in which orders are checked is hereby determined by the PRDs of the orders, which means that the order with the earliest PRD is considered first for order release. While this overall procedure is the same for the considered static and dynamic approaches, they vary with regard to how LTs are calculated.

2.1.1 Static BIL

BIL is commonly applied using fixed or static LTs to calculate the PRDs of orders. This means that a fixed allowance is subtracted from the External Due Date (ExDD) of an order j to calculate its PRD (Ragatz and

Mabert 1988):

$$PRD_j = ExDD_j - LT, \tag{1}$$

where $ExDD_j$ is the External Due Date of order j and LT is the planned Lead Time for all orders. Note that the PRDs are rounded down to the next period value such that all orders whose PRD falls within the upcoming period are released at the beginning of this period.

2.1.2 Dynamic BIL

To account for workload and other variations on the shop floor, a dynamic BIL approach which relies on dynamic LTs for determining PRDs seems promising. In this regard, two dynamic mechanisms are considered: (i) BIL based on Exponential Smoothing denoted as ES, and (ii) BIL based on Exponential Smoothing including Safety Lead Times denoted as ESSLT. The latter approach was introduced by Haeussler et al. (2019) and is motivated by the study of Enns and Suwanruji (2004). For ES and ESSLT, the actual SFTTs, i.e. time from order release until order completion, of all orders need to be tracked. Relying on the forecasting method of exponential smoothing, ES and ESSLT calculate the lead time $LT_{i,t}$ for the current time t and product type i as follows:

$$LT_{i,t} = \alpha * SFTT_{i,t} + (1-\alpha) * LT_{i,t-1}, \tag{2}$$

where $0 \leq \alpha \leq 1$ is the smoothing factor, $SFTT_{i,t}$ represents the last observed SFTT of the respective product type i given the current time t, and $LT_{i,t-1}$ corresponds to the previous LT of product type i. Different to ES, ESSLT additionally considers a safety lead time (SLT). Assuming normally distributed deviations of the planned LTs from the actual SFTTs, Haeussler et al. (2019) use the cost ratio between backorder and inventory costs to determine the SLT, i.e. backorder cost parameter divided by (backorder + inventory cost parameters). This approach is inspired by safety stock calculations in inventory systems (Silver et al. 1998; Haeussler et al. 2019). The cost ratio is then used to determine the z-quantile of the normal distribution. The safety lead time $SLT_{i,t}$ of product type i is then calculated by multiplying the z-quantile with the standard deviation over the deviations between actual SFTTs and planned LTs for the given product type i. $SLT_{i,t}$ is subsequently added to $LT_{i,t}$, and the resulting value is used as new planned lead time $LT_{i,t}$ for determining the PRDs. For ES and ESSLT, the PRD of an order j is calculated as follows:

$$PRD_j = ExDD_j - LT_{i,t}, \tag{3}$$

where $LT_{i,t}$ is the Lead Time for the respective product type i of order j at time t. Again, the PRDs are rounded down to the next period value such that all orders whose PRD falls within the upcoming period are released at the beginning of this period. Note that due to the dynamic LT, the PRD can also be in the past. These orders are also released at the beginning of the respective period. Further, ES and ESSLT require that the PRDs of the orders in the order pool are calculated before each periodic release as the lead time forecasts and safety lead times change with every completed order.

However, reactively adjusting lead times solely based on actual SFTTs can lead to the lead time syndrome. When SFTTs are higher than expected, then the LTs are updated and increase. Orders are released earlier to the shop floor which results in a higher WIP level at the work centers. Due to this higher WIP level, SFTTs increase again which requires another update of the LTs and an even earlier release of orders (Mather and Plossl 1978). To avoid this, an upper bound for the planned LT is used as suggested by Haeussler et al. (2019). If the lead time $LT_{i,t}$ exceeds this upper bound, then the upper bound is used for the calculation of the PRDs.

2.2 COrrected aggregate Load Approach (COLA)

The COrrected aggregate Load Approach (COLA) is a purely periodic approach, whose release procedure works as follows (Oosterman et al. 2000; Thuerer et al. 2012; Haeussler et al. 2021):

1. All arriving orders are collected in an order pool and sorted according to Planned Release Dates (PRD). The first order in the order pool is then considered for order release.
2. If a time limit is currently applied, then the release procedure checks whether the PRD of this order is within the time limit, i.e. PRD $\leq$ current time + time limit. If this is true, then continue with the next step, otherwise skip the next step and continue with step 4. If no time limit is currently applied, then continue with the next step.
3. If this order does not violate any workload norm of the work centers in its routing, i.e. current workload of work center + load contribution of order $\leq$ workload norm, then the order is released and its load contributions are actually added to the current workloads of the work centers. The load contribution to a work center is calculated by dividing the processing time at this work center by the position of the work center in the routing of the order, and the load contributions are not removed until the respective operation at a work center is completed (Oosterman et al. 2000). However, if at least one workload norm is violated, the order is kept in the order pool.
4. If there are further orders in the order pool that need to be considered for order release, then the next order based on PRD is selected and step 2 is performed again. Otherwise, if all orders were considered, then the periodic release procedure is completed and is repeated after a certain time depending on the release frequency (e.g. a day).

The PRDs of the orders in the order pool are calculated by subtracting an allowance for the SFTT (i.e. lead time) from the external due date (ExDD). The allowance is given by the cumulative moving average over all realized SFTTs of the respective product type of the underlying order. Note that only one workload norm is applied to all work centers (Thuerer et al. 2011). Regarding the adaptive time limit policy under COLA, referred to as Overload, a so-called load level is used to discriminate between low and high load periods. In this regard, the total corrected shop load, i.e. the sum of the current corrected workloads over all work centers, is compared to the load level and the application of the time limit is formalized as follows:

- In low load periods (i.e. total corrected shop load $<$ load level), an unlimited time limit is applied, and
- in high load periods (i.e. total corrected shop load $\geq$ load level), a tight time limit is applied.

In case an order is released during the periodic release procedure, the load contributions are not only added to the current workloads of the work centers but they are also immediately added to the current total corrected shop load. If the load level is exceeded, then the respective time limit is activated and only the most urgent orders are still considered for order release, if any further orders in the order pool fall within the time limit at all (Haeussler et al. 2021).

3 SIMULATION MODEL AND EXPERIMENTAL DESIGN

To enable comparability with previous research on semiconductor manufacturing, a simulation model of a scaled-down wafer fab model is used which includes a re-entrant bottleneck and which was built with attributes of a real semiconductor manufacturing facility previously studied in WLC research (Kayton et al. 1997; Kacar et al. 2012; Ziarnetzky et al. 2015; Neuner et al. 2020; Neuner and Haeussler 2020). The main characteristics of semiconductor manufacturing are multiple products with varying, re-entrant product routings of different lengths, machine failures and work centers performing batch processing. The simulation model has one re-entrant bottleneck work center performing the photolithography process, and two work centers that are able to process multiple orders of different product types at once, i.e. batching work centers 1 and 2 whose batch size can vary between 2 and 4 orders at a time. The latter two work centers represent the furnaces performing the diffusion and oxidation processes. The remaining work centers process only one order at a time and the model is shown in Figure 1. As can be seen, the simulation model includes 11 work centers, where only the bottleneck work center (work center 4) has two servers and all other work centers have only one server. The respective processing times at each work center

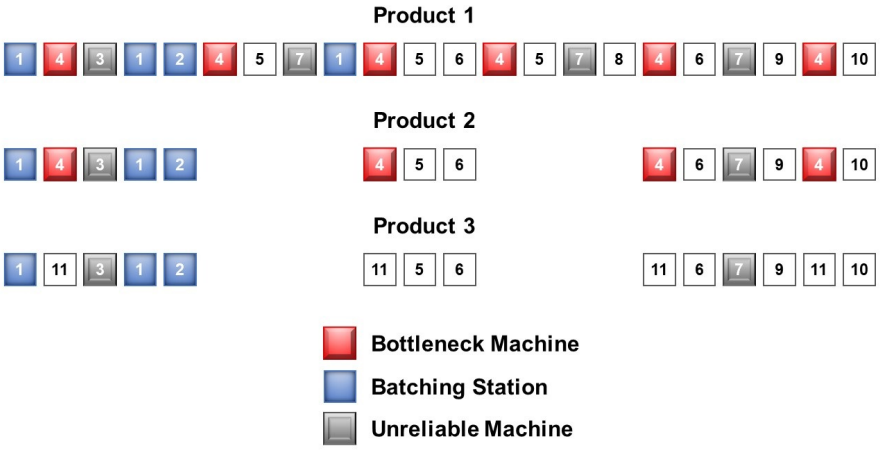


Figure 1: Product routings of re-entrant bottleneck model (Kacar et al. 2012).

follow log-normal distributions whose parameters are set such that the standard deviation $\leq$ 10 percent of the corresponding mean. Table 1 shows the detailed processing times of each work center.

Since semiconductor manufacturing typically includes machine failures, work centers 3 and 7 are subject to machine breakdowns. The mean time to failure (MTTF) and mean time to repair (MTTR) of these work centers follow gamma distributions whose parameters are set as follows:

- MTTF: α = 7,200, β = 1 $\rightarrow$ mean = 7,200, Std. Dev. = 84.9
- MTTR: α = 1,200, β = 1.5 $\rightarrow$ mean = 1,800, Std. Dev. = 52.0

Table 1: Processing times of the work centers.

Work center #	Mean	Std. Dev.	Work center #	Mean	Std. Dev.
1	80	7	2	220	16
3	45	4	4	40	4
5	25	2	6	22	2.4
7	20	2	8	100	12
9	50	4	10	50	5
11	70	2.5			

In the following, each visit to a work center is referred to as "process step". The model includes three products with the following varying product routings: Product 1 has 22 process steps and visits the bottleneck work center 6 times, product 2 has 14 process steps and visits the bottleneck work center 4 times, and product 3 has 14 process steps and does not visit the bottleneck. The product mix in the model is set to 3 : 1 : 1 of Product 1, 2, and 3 respectively, and two demand levels (high and low demand) are tested. In this regard, the bottleneck utilization equals either approximately 90% or 80%. The two unreliable work centers 3 and 7 create most of the starvation at the bottleneck work center. One of them, i.e. work center 3, is visited only once by all products and acts as gateway operation due to its location early in the process routing, which means that this work center opens and closes the flow of the products into the production system. The other work center 7 is a re-entrant work center that performs the Chemical Vapor Deposition process and is situated later in the product routings. Although processing many orders very quickly, the two unreliable work centers might cause starvation at the bottleneck due to poor availability.

In the analysis, a stochastic demand with exponentially distributed inter–arrival times is used. In the high demand setting, orders arrive with a mean of one order per 98 minutes, while in the low demand setting, orders arrive with a mean of one order per 110 minutes. Independent of the demand level, the due dates of the orders are set as follows (Kutanoglu 1999; Thuerer et al. 2011; Land 2006; Gupta and

Sivakumar 2007; Bahaji and Kuhl 2008): On order arrival the product type is randomly assigned based on a discrete uniform distribution *dunif*{*1,5*} (1-3: product type 1, 4: product type 2, and 5: product type 3). The due date slack is then determined by adding a random allowance where the minimum slack equals seven times the total processing time of product type 1 (7,868 minutes) and the maximum slack is set to 14,612 minutes (13 times total processing time of product type 1):

$$DD_j = AT_j + unif\{7,868\,;14,612\}. \tag{4}$$

While DD_j represents the due date, AT_j corresponds to the arrival time of order j. The random allowance was set such that an Immediate Release strategy yields approximately 0-5% tardy orders. Table 2 depicts the experimental design of the study. The above described order release approaches BIL, ES, ESSLT, COLA and Overload are tested with different set of parameters which have been specified in the course of preliminary simulations runs based on total cost measures (as defined below). Note that for BIL, the fixed lead time was set to 5 times the period length for the high demand setting and to 4 times the period length for the low demand setting. Regarding pool sequencing, Planned Release Date (PRD)

Table 2: Experimental design.

90% Bottleneck Utilization		**80% Bottleneck Utilization**	
Order Release Model	**Tested Parameters**	**Order Release Model**	**Tested Parameters**
Backward Infinite Loading (BIL)	Fixed Lead Time = 5 * Period Length	Backward Infinite Loading (BIL)	Fixed Lead Time = 4 * Period Length
Exponential Smoothing (ES)	α (0.1; 0.2; 0.3)	Exponential Smoothing (ES)	α (0.1; 0.2; 0.3)
Exponential Smoothing with Safety Lead Time (ESSLT)	α (0.1; 0.2; 0.3) z-value (0.736; 1.282; 1.645)	Exponential Smoothing with Safety Lead Time (ESSLT)	α (0.1; 0.2; 0.3) z-value (0.736; 1.282; 1.645)
COLA	workload norm (1,900; 2,000; 2,100)	COLA	workload norm (1,700; 1,800; 1,900)
Overload	workload norm (1,900; 2,000; 2,100) time limit = 2,880 load level (3,500; 4,000; 4,500)	Overload	workload norm (1,700; 1,800; 1,900) time limit = 2,880 load level (2,000; 2,500; 3,000)

is applied to all investigated scenarios. Therefore, the most urgent order based on PRDs is considered first for order release. However, for both demand levels, ES and COLA are analyzed with three different α-values or workload norms respectively. Since ESSLT additionally requires a safety lead time, three different z-quantiles of the normal distribution are tested. The lowest z-quantile is based on the cost ratio as suggested by Haeussler et al. (2019). However, since they only tested a low demand with 80% machine utilization, two additional z-quantiles are included in the analysis. The respective safety levels equal 0.769, 0.9 and 0.95. In addition, ES and ESSLT require an upper bound for the lead time which, based on preliminary analysis, was set to 5 times the period length and 4 times the period length for a high and low demand respectively. Regarding Overload, based on pilot simulation runs, the time limit was set to 2,880 minutes and three different load levels are analyzed for both demand levels. As indicated above, the load levels were determined based on total cost measures in the course of preliminary simulation runs and are used in the experiments to discriminate between low and high load periods depending on the current total corrected shop load. Thus, in total 50 different scenarios are simulated. Machine dispatching is done according to First-In First-Out throughout all tested scenarios. Thus, the results are solely dependent on the specific Order Release approach and the respective parameterization.

Furthermore, the period length was set to 1,440 minutes (one day), each scenario was replicated 80 times, the warm-up phase was set to 800 periods and data was collected over 1,000 periods. For parameterization

and performance analysis, a cost function was defined which consists of the sum of WIP ($WIP_{n,t}$) at each work center n, finished goods holding FGI (FGI_t) and backorder (BO_t) costs over all periods t:

$$\text{Total Costs} = \sum_{t=1}^{T}\sum_{n=1}^{N} \omega WIP_{n,t} + \sum_{t=1}^{T}(\pi FGI_t + \kappa BO_t) \tag{5}$$

The cost parameters ω, π and κ were set in the following relation: $2\frac{1}{3} : 1 : 3\frac{1}{3}$ which is taken from earlier WLC studies in semiconductor industry (Kacar et al. 2012; Kacar et al. 2013; Albey and Uzsoy 2015; Ziarnetzky et al. 2015; Neuner et al. 2020; Neuner and Haeussler 2020).

4 RESULTS

In this section, the results for the investigated order release approaches under a high and a low demand are discussed. For brevity, numerical results are only presented for the timing and cost measures under the high demand setting, but all numerical results are provided in the following data repository: http://dx.doi.org/10.17632/hdww5ffy8j.1. Table 3 shows the timing and cost measures for the simulated scenarios under a high demand. The first column denotes the tested order release approach and the corresponding parameterization. For brevity, a single is used for classic BIL, a double for BIL based on exponential

Table 3: Timing and cost measures for different order release scenarios for a high demand.

Scenario	Percentage Tardy Orders	Backorder Costs	WIP Costs	FGI Costs	Total Costs
BIL	*6.95%*	*$2,396.03*	*$95,940.69*	*$32,580.52*	*$130,917.24*
ES_0.1	*50.80%*	*$34,037.16*	*$122,909.83*	*$6,696.84*	*$163,643.83*
ES_0.2	53.65%	$38,346.32	$128,234.42	$6,252.95	$172,833.68
ES_0.3	54.78%	$39,535.92	$130,019.36	$6,094.38	$175,649.66
ESSLT_0.1_0.736	26.04%	$12,455.63	$109,670.68	$13,173.18	$135,299.49
ESSLT_0.2_0.736	28.44%	$14,122.03	$112,649.86	$12,536.66	$139,308.55
ESSLT_0.3_0.736	29.21%	$14,631.55	$114,058.28	$12,364.61	$141,054.44
ESSLT_0.1_1.282	14.58%	$5,947.05	$103,004.62	$18,258.46	$127,210.12
ESSLT_0.2_1.282	16.18%	$6,786.17	$105,234.82	$17,600.54	$129,621.54
ESSLT_0.3_1.282	16.86%	$7,184.87	$106,462.38	$17,351.01	$130,998.27
ESSLT_0.1_1.645	*10.49%**	*$3,992.04**	*$100,130.77*	*$21,284.05**	*$125,406.86*
ESSLT_0.2_1.645	11.60%*	$4,492.14*	$101,932.42	$20,654.51	$127,079.07
ESSLT_0.3_1.645	12.18%*	$4,819.74	$103,060.07	$20,415.87	$128,295.68
COLA_1900	*1.82%*	*$825.21*	*$91,112.93**	*$52,382.23*	*$144,320.36*
COLA_2000	0.77%	$186.21	$93,427.01	$56,403.90	$150,017.13
COLA_2100	0.91%	$194.94	$94,547.91	$57,728.33	$152,471.17
Overload_1900_2880_3500	20.25%	$10,359.53	$89,391.89*	$16,867.84	$116,619.26*
Overload_2000_2880_3500	12.24%	$4,340.75	$91,436.40*	$19,942.07	$115,719.22*
Overload_2100_2880_3500	9.81%*	$3,156.11*	$92,573.79	$21,239.55	$116,969.45
Overload_1900_2880_4000	19.41%*	$10,229.29*	$88,108.37*	$18,566.24	$116,903.90*
Overload_2000_2880_4000	**11.82%**	**$4,248.95**	**$89,673.87**	**$21,624.50**	**$115,547.32**
Overload_2100_2880_4000	9.67%	$3,173.62	$90,572.04*	$22,844.00	$116,589.66
Overload_1900_2880_4500	14.33%	$7,774.18	$86,130.52	$27,692.59*	$121,597.30*
Overload_2000_2880_4500	8.72%	$3,239.09	$87,034.28*	$30,535.34	$120,808.72*
Overload_2100_2880_4500	7.13%	$2,421.44	$87,523.58*	$31,572.41	$121,517.44

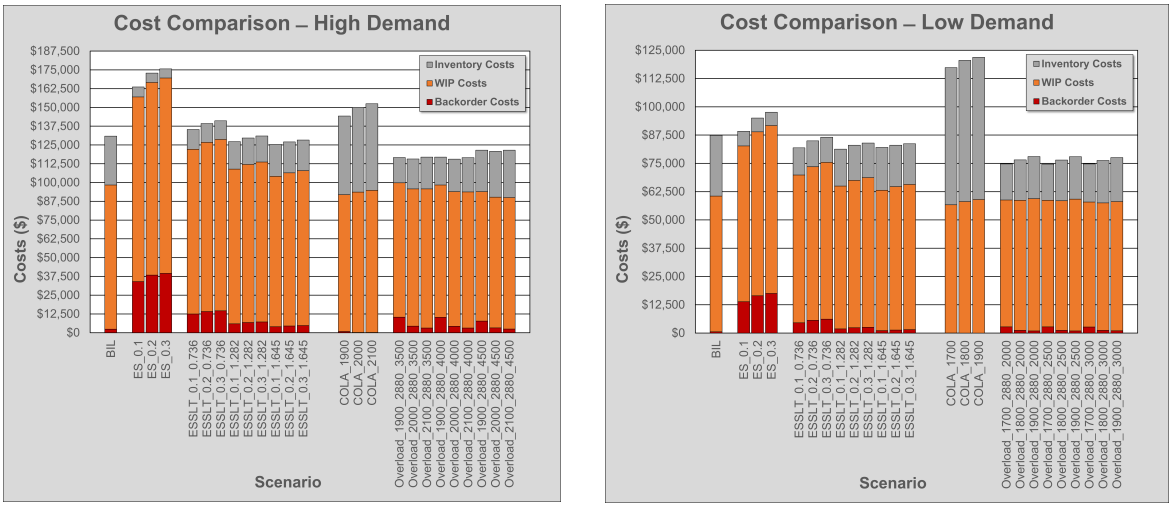
* not significant ($p < 0.05$)

smooting (i.e. ES) and the classic COLA approach without a time limit, i.e. with an unlimited time limit, a triple for BIL based on exponential smoothing with a safety lead time (i.e. ESSLT), and a quadruple for the Overload scenarios: The first component corresponds to the order release mechanism (BIL, ES, ESSLT, COLA or Overload), the second component denotes the tested α-value (for ES and ESSLT) or the workload norm (for COLA and Overload), and the third component represents the z-value for calculating the safety lead time (for ESSLT) or the time limit (for Overload). Regarding Overload, the numbers in the fourth component denote the load level which is used to distinguish between high and low load periods.

The remaining columns show the Percentage of Tardy Orders and the average Backorder, WIP, Finished Goods Inventory (FGI) and Total Cost values over all replications.

For a high demand, the best performing scenario in terms of total costs is *Overload_2000_2880_4000* which is highlighted in bold in Table 3. All other scenarios are compared to this best scenario and differences are tested at a significance level of $p = 0.05$ based on a Wilcoxon/Mann-Whitney-U Test. All values marked with an asterisk are not significantly different from *Overload_2000_2880_4000*. To preserve readability, the best performing scenario from each order release mechanism is highlighted in italics. It can be seen that compared to the best scenario *Overload_2000_2880_4000*, *BIL* yields \$15,369.92, the best ES scenario (*ES_0.1*) yields \$48,096.51, the best ESSLT scenario (*ESSLT_0.1_1.645*) yields \$9,859.54 and the best COLA scenario (*COLA_1900*) yields \$28,773.05 higher total costs on average. To preserve readability, each of the best performing scenarios is simply denoted as "BIL", "ES", "ESSLT", "COLA", and "Overload" in the following. Compared to the static approaches BIL and COLA, Overload yields slightly higher backorder costs which are outweighed by a significant inventory cost reduction. Further, Overload yields the lowest WIP costs (not significant compared to COLA). Focusing on the dynamic approaches, Overload yields higher inventory costs than ES which are outweighed by a significant WIP and backorder cost reduction. With regard to ESSLT, Overload results in a similar timing performance due to yielding similar backorder and inventory costs, but again Overload yields lower WIP costs. Therefore, Overload yields the lowest total costs on average. Exactly the same conclusions can be drawn for a low demand, which means that the demand level has no impact on the relative performance of the order release approaches. Thus, Overload also yields the lowest total costs under a low demand. Regarding the percentage of tardy orders, COLA yields the best service level followed by BIL under both demand levels.

Figures 2a and 2b illustrate the total costs and the cost distribution between the simulated scenarios regarding backorder, WIP and inventory costs for a high and a low demand respectively. Focusing on BIL, ES and ESSLT, it can be seen that, especially for a high demand level, ES results in a drastic total cost



a) Comparison of the costs between different parameterized order release models for a high demand.

b) Comparison of the costs between different parameterized order release models for a low demand.

Figure 2: Comparison of the costs for different demand levels.

increase compared to BIL. Thus, solely relying on lead time forecasts based on exponential smoothing is not recommendable. Only when a safety lead time is included into the lead time forecasts (i.e. ESSLT) then total costs can be reduced compared to BIL, which confirms the findings of Haeussler et al. (2019) also for a high demand level. Interestingly, using the cost ratio between backorder and inventory costs for determining the z-quantile as suggested by Haeussler et al. (2019) is only reasonable for a low demand.

Here total costs are lower compared to BIL and additionally, are quite insensitive with regard to the z-value. However, when switching from a low (i.e. 80% bottleneck utilization) to a high demand (i.e. 90% bottleneck utilization), Figure 2a shows that relying on the cost ratio is no longer justifiable since total costs are higher compared to BIL (see ESSLT scenarios with z-value=0.736). However, total costs can be drastically reduced when using a greater z-quantile in the calculation of the safety lead time. Therefore, depending on the parameterization, ESSLT outperforms ES and also BIL.

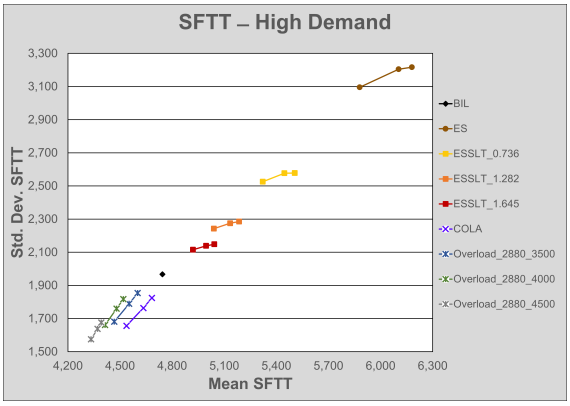
When focusing on all order release approaches, it can be seen that, for both demand levels, all Overload scenarios yield lower total costs than the best parameterization of BIL, ES, ESSLT and COLA. Moreover, for a given workload norm, all Overload scenarios lead to lower total costs on average than the corresponding COLA scenario. In this regard, applying a time limit during high load periods delays the release of non-urgent orders which reduces FGI costs but, on the contrary, more likely results in tardy orders which yield backorder costs. However, during low load periods an unlimited time limit is applied, which means that all orders in the order pool are considered for order release. In such low load periods the order pool is cleared which enables that urgent orders fit within the workload norms during upcoming high load periods. Thus, Overload enables production smoothing (Haeussler et al. 2021) and consequently, reduces WIP costs.

Focusing on the load balancing performance in greater detail, Figures 3a and 3c illustrate the mean and standard deviation of shop floor throughput time (SFTT), and Figures 3b and 3d depict the mean and standard deviation of bottleneck queue time (BQT) for a high and a low demand. The left-hand starting points of the COLA and Overload curves represent the lowest workload norm which increases when moving along the curve. Similarly, the left-hand starting point of the ES and ESSLT curves represent the lowest α-value which increases stepwise when moving to the right-hand points of the curves. For both demand levels, COLA yields lower means and standard deviations of shop floor throughput time and bottleneck queue time compared to BIL, ES and ESSLT, with ES resulting in the worst overall balancing performance. Further, ESSLT yields higher load balancing measures than BIL which explains the higher WIP costs of ESSLT compared to BIL. The reason for the worse load balancing performance of the dynamic ES and ESSLT approaches lies in the reinforcing effect of the lead time syndrome. When SFTTs increase, also LT forecasts are updated and increase until the upper bound is reached. The same logic applies to decreasing SFTTs. Regarding ESSLT, this approach further includes a safety lead time based on the standard deviation over the deviations between actual SFTTs and planned LTs. This safety lead time was observed to reduce the reinforcing effect, which enables ESSLT to yield intermediate load balancing measures compared to BIL and ES. In general, by dynamically adjusting the lead times and consequently, the release times of the orders in the order pool under ES and ESSLT, additional variability besides demand variability is brought to the shop floor which leads to a worse load balancing performance compared to BIL.

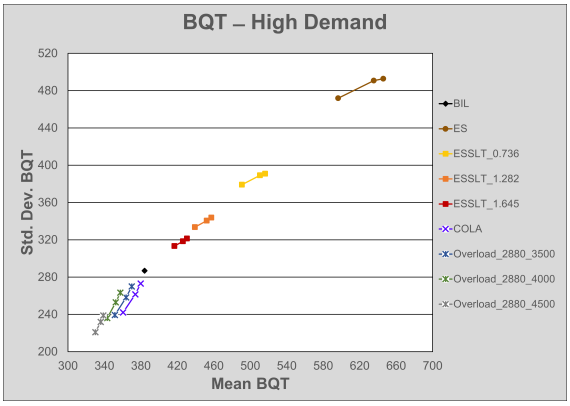
Nevertheless, for a high demand, a given workload norm and depending on the load level, Overload allows to further reduce the means and standard deviations of shop floor throughput time and bottleneck queue time compared to COLA while still yielding lower total costs than all other order release approaches. While the same holds for the mean shop floor throughput time and the mean bottleneck queue time under a low demand, here the standard deviations of shop floor throughput time and bottleneck queue time are not significantly different compared to COLA. The reason for the less superior load balancing performance lies in the reduced load balancing possibilities as, on average, less orders are waiting in the order pool to be released to the shop floor. However, it can be concluded that Overload yields a superior load balancing and cost performance compared to BIL, ES, ESSLT and COLA irrespective of the demand level.

5 CONCLUSION

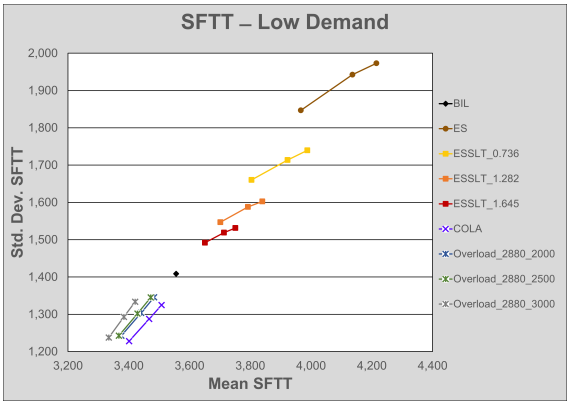
This paper compared two periodic order release approaches, the Backward Infinite Loading approach (BIL) and the COrrected aggregate Load Approach (COLA) (Oosterman et al. 2000) which is based on workload control theory. Recent studies conceptualized dynamic approaches that react to changes on the shop floor, one focused on an adaptive time limit policy (Haeussler et al. 2021) and another on adaptive lead times for determining planned release dates (Haeussler et al. 2019). In this regard, semiconductor manufacturing



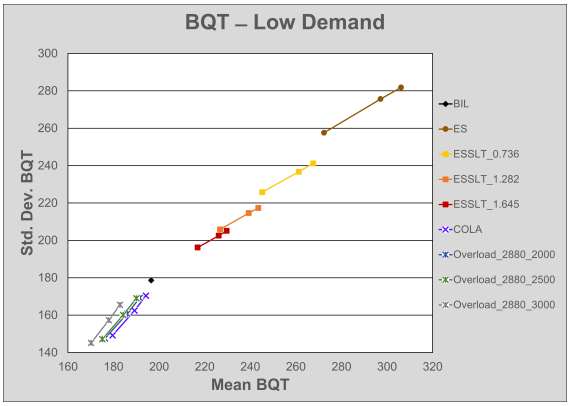
a) Mean and standard deviation of shop floor throughput time for different parameterized order release models for a high demand.



b) Mean and standard deviation of bottleneck queue time for different parameterized order release models for a high demand.



c) Mean and standard deviation of shop floor throughput time for different parameterized order release models for a low demand.



d) Mean and standard deviation of bottleneck queue time for different parameterized order release models for a low demand.

Figure 3: Comparison of load balancing performance for a high and a low demand.

provides a very challenging environment, e.g. due to machine failures and batch processing, to investigate whether a dynamic time limit policy under COLA, denoted as Overload, or the dynamic extension of the BIL approach, denoted as ESSLT, outperforms the other approaches. For this purpose, a simulation model of a scaled-down wafer fab (Kayton et al. 1997) is used, and a high and a low demand setting are applied. The findings are as follows: For both demand levels, ESSLT yields lower total costs than BIL, but Overload yields the lowest total costs on average. In this regard, Overload yields a superior load balancing performance, which results in the lowest WIP costs.

The results also show that using the cost ratio for determining the z-value for the safety lead time calculation is only reasonable for a low demand, but in the high demand scenario, a greater z-quantile has to be used, as otherwise total costs increase compared to BIL. Concluding, the findings show that a dynamic time limit policy within the COLA order release mechanism is a promising extension for order release in semiconductor manufacturing. In this regard, since the respective Overload approach is a purely periodic release model, its implementation in practice should be eased, as periodic decision making, e.g. once a day or shift, is the preferred behavior of planners (Hendry and Kingsman 1991; Sabuncuoglu and Karapınar 1999; Stevenson et al. 2011; Thuerer et al. 2012).

The study provides important insights, but also includes some limitations. Firstly, the results are limited to the experimental design and further experiments are necessary to validate the findings also for large-scale semiconductor fabs, e.g. based on MIMAC or SMT2020 models (Kopp et al. 2020). Secondly, future studies should also consider further experimental factors such as different due date slacks or different pool sequencing and scheduling rules. Thirdly, other demand patterns, e.g. with differing correlations across products, and the impact of the length of the period on the results seem also worthwhile to be investigated. Finally, future studies should also investigate the potential of the dynamic time limit policy under the hybrid LUMS-COR model in semiconductor manufacturing, and should also compare the Overload approach to the widely used periodic optimization based order release models in the semiconductor industry (Kacar et al. 2012; Kacar et al. 2013; Ziarnetzky et al. 2015).

ACKNOWLEDGMENTS

The author would like to thank Stefan Haeussler and Hubert Missbauer for their valuable comments and suggestions which greatly helped to improve the quality of the paper.

REFERENCES

Albey, E., and R. Uzsoy. 2015. "Lead Time Modeling in Production Planning". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 1996–2007. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Bahaji, N., and M. E. Kuhl. 2008. "A simulation study of new multi-objective composite dispatching rules, CONWIP, and push lot release in semiconductor fabrication". *International Journal of Production Research* 46(14):3801–3824.

Enns, S. T., and P. Suwanruji. 2004. "Work load responsive adjustment of planned lead times". *Journal of Manufacturing Technology Management* 15(1):90–100.

Fowler, J. W., G. L. Hogg, and S. J. Mason. 2002. "Workload Control in the Semiconductor Industry". *Production Planning and Control* 13(7):568–578.

Glassey, C. R., and M. G. C. Resende. 1988. "Closed-loop Job Release Control for VLSI Circuit Manufacturing". *IEEE Transactions on Semiconductor Manufacturing* 1(1):36–46.

Goldratt, E., and J. Cox. 1986. *The Goal: A Process of Ongoing Improvement*. New York: North River Press.

Gupta, A. K., and A. I. Sivakumar. 2007. "Controlling delivery performance in semiconductor manufacturing using Look Ahead Batching". *International Journal of Production Research* 45(3):591–613.

Hackman, S., and R. Leachman. 1989. "A general framework for modeling production". *Management Science* 35:478–495.

Haeussler, S., and P. Netzer. 2020. "Comparison between Rule- and Optimization based Workload Control Concepts: A Simulation Optimization approach". *International Journal of Production Research* 58(12):3724–3743.

Haeussler, S., P. Neuner, and M. Thuerer. 2021. "Balancing Earliness and Tardiness within Workload Control Order Release: An Assessment by Simulation". *Flexible Services and Manufacturing*. under review.

Haeussler, S., M. Schneckenreither, and C. Gerhold. 2019. "Adaptive order release planning with dynamic lead times". *IFAC-PapersOnLine* 52(13):1890–1895.

Haeussler, S., C. Stampfer, and H. Missbauer. 2020. "Comparison of two optimization based order release models with fixed and variable lead times". *International Journal of Production Economics* 227:107682.

Hendry, L., and B. Kingsman. 1991. "A Decision Support System for Job Release in Make-to-order Companies". *International Journal of Operations & Production Management* 11(6):6–16.

Hung, Y.-F., and R. C. Leachman. 1996. "A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations". *IEEE Transactions on Semiconductor Manufacturing* 9(2):257–269.

Hutter, T., S. Haeussler, and H. Missbauer. 2018. "Successful Implementation of an Order Release Mechanism based on Workload Control: A Case Study of a make-to-stock manufacturer". *International Journal of Production Research* 56(4):1565–1580.

Jacobs, F. R. 1984. "OPT uncovered: many production planning and scheduling concepts can be applied with or without the software". *Industrial Engineering* 16(10):32–41.

Kacar, N. B., D. F. Irdem, and R. Uzsoy. 2012. "An Experimental Comparison of Production Planning Using Clearing Functions and Iterative Linear Programming-Simulation Algorithms". *IEEE Transactions on Semiconductor Manufacturing* 25(1):104–117.

Kacar, N. B., L. Moench, and R. Uzsoy. 2013. "Planning Wafer Starts Using Nonlinear Clearing Functions: A Large-Scale Experiment". *IEEE Transactions on Semiconductor Manufacturing* 26(4):602–612.

Kayton, D., T. Teyner, C. Schwartz, and R. Uzsoy. 1997. "Focusing Maintenance Improvement Efforts in a Wafer Fabrication Facility Operating under the Theory of Constraints". *Production and inventory management journal: journal of the American Production and Inventory Control Society* 38(4):51–57.

Kingsman, B. G., I. P. Tatsiopoulos, and L. C. Hendry. 1989. "A Structural Methodology for Managing Manufacturing Lead Times in Make-to-Order Companies". *European Journal of Operational Research* 40(2):196–209.

Kopp, D., M. Hassoun, A. Kalir, and L. Mönch. 2020. "SMT2020—A semiconductor manufacturing testbed". *IEEE Transactions on Semiconductor Manufacturing* 33(4):522–531.

Kutanoglu, E. 1999. "An analysis of heuristics in a dynamic job shop with weighted tardiness objectives". *International Journal of Production Research* 37(1):165–187.

Land, M. 2006. "Parameters and sensitivity in workload control". *International Journal of Production Economics* 104(2):625–638.

Mather, H., and G. W. Plossl. 1978. "Priority fixation versus throughput planning". *Production and Inventory Management* 19:27–51.

Missbauer, H., and R. Uzsoy. 2011. *Optimization models of production planning problems*, 437–507. Norwell: Springer.

Neuner, P., and S. Haeussler. 2020. "Rule based workload control in semiconductor manufacturing revisited". *International Journal of Production Research* 0(0):1–20.

Neuner, P., S. Haeussler, and Q. Ilmer. 2020. "Periodic Workload Control: A viable Alternative for Semiconductor Manufacturing". In *Proceedings of the 2020 Winter Simulation Conference*, edited by K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 1765–1776. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Oosterman, B., M. Land, and G. Gaalman. 2000. "The influence of shop characteristics on workload control". *International Journal of Production Economics* 68(1):107–119.

Ragatz, G. J., and V. A. Mabert. 1988. "An evaluation of order release mechanisms in a job-shop environment". *Decision Sciences* 19:167–189.

Rose, O. 1999. "CONLOAD-a new lot release rule for semiconductor wafer fabs". In *Proceedings of the 1999 Winter Simulation Conference*, edited by P. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 850–855. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sabuncuoglu, I., and H. Karapınar. 1999. "Analysis of order review/release problems in production systems". *International Journal of Production Economics* 62(3):259–279.

Silver, E., D. G. Pyke, and R. Peterson. 1998. *Inventory Management and Production Planning and Scheduling*. Wiley, New York.

Spearman, M. L., D. L. Woodruff, and W. J. Hopp. 1990. "CONWIP: a pull alternative to kanban". *International Journal of Production Research* 28(5):879–894.

Stevenson, M., Y. Huang, L. C. Hendry, and E. Soepenberg. 2011. "The theory and practice of workload control: A research agenda and implementation strategy". *International Journal of Production Economics* 131(2):689–700.

Thuerer, M., T. Qu, M. Stevenson, T. Maschek, and M. Filho. 2014. "Continuous workload control order release revisited: an assessment by simulation". *International Journal of Production Research* 52(22):6664–6680.

Thuerer, M., C. Silva, and M. Stevenson. 2011. "Optimising workload norms: the influence of shop floor characteristics on setting workload norms for the workload control concept". *International Journal of Production Research* 49(4):1151–1171.

Thuerer, M., M. Stevenson, and C. Silva. 2011. "Three decades of workload control research: a systematic review of the literature". *International Journal of Production Research* 49(23):6905–6935.

Thuerer, M., M. Stevenson, C. Silva, M. J. Land, and L. D. Fredendall. 2012. "Workload Control and Order Release: A Lean Solution for Make-to-Order Companies". *Production and Operations Management* 21(5):939–953.

Wein, L. M. 1988. "Scheduling semiconductor wafer fabrication". *IEEE Transactions on Semiconductor Manufacturing* 1(3):115–130.

Wiendahl, H. 1995. *Load-Oriented Manufacturing Control*. 1st ed. Berlin: Springer.

Ziarnetzky, T., B. Kacar, L. Moench, and R. Uzsoy. 2015. "Simulation-Based Performance Assessment of Production Planning Formulations for Semiconductor Wafer Fabrication". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 2884–2895. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

PHILIPP NEUNER is working as research assistant at the Department of Information Systems, Production and Logistics Management at the University of Innsbruck. He received his M.Sc. degree in Information Systems from the University of Innsbruck in 2019 and is currently studying for his PhD degree in Management at the University of Innsbruck. His research interests include manufacturing planning and control, simulation modeling, optimization and workload control. His email address is philipp.neuner@uibk.ac.at.