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.

PERIODIC WORKLOAD CONTROL: A VIABLE ALTERNATIVE FOR SEMICONDUCTOR MANUFACTURING

Philipp Neuner
Stefan Haeussler
Quirin Ilmer

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

ABSTRACT

This paper analyzes a rule based workload control model applied to a scaled-down semiconductor simulation model. We compare two well established continuous order release models from the semiconductor domain, namely the Starvation Avoidance (SA) and the CONstant LOAD (ConLOAD) approach, with the COrrected aggregate Load Approach (COLA) which was originally developed for small and medium enterprises in make-to-order companies. The main difference between these order release approaches is that the former two (SA and ConLOAD) release orders continuously whereas the latter (COLA) releases orders at periodic intervals. Contrary to earlier research on order release models for semiconductor models we show that the periodic order release model outperforms the other two continuous mechanisms by yielding lower costs and better timing performance. Thus, this paper highlights that *periodic* rule based order release models are a viable alternative which was largely neglected in recent semiconductor literature.

1 INTRODUCTION

Workload control (WLC) is a manufacturing planning and control concept which originated from the idea to control the mean cycle times by controlling order releases and thus the level of Work-In-Process (WIP) and the output (Bertrand and Wortmann 1981). The core of WLC is certainly the *order release* decision which determines the amount and time when orders are released to the shop floor. Literature on order release mechanisms can be divided into two parts which categorize the approaches by the timing convention which can be either *continuous*, at any moment in time, or *discrete* meaning that orders are released at periodic intervals (Bergamaschi et al. 1997). WLC was also widely applied to the semiconductor environment (Fowler et al. 2002) where the main focus is on continuous methods (Glassey and Resende 1988; Wein 1988; Spearman et al. 1990). Glassey and Resende (1988) introduce Starvation Avoidance (SA) which limits the number of orders within a certain lead time window. More precisely, they introduce a virtual inventory which includes all work in the system that could reach the bottleneck work center within this lead time. Furthermore, Wein (1988) develop the Workload Regulation approach which controls the workload that is released to the bottleneck. At the same time, other bottleneck oriented approaches like the famous Drum-Buffer-Rope concept and its Optimized Production Technology (OPT) were used frequently (Jacobs 1984; Goldratt and Cox 1986). Additionally, Spearman, Woodruff, and Hopp (1990) introduced Constant Work-In-Process (CONWIP) which regulates the WIP on the whole shop floor. Over the years, CONWIP was adapted to focus on the contributed load of each job to the bottleneck work center to keep it at a desired level, and is denoted as ConLOAD (Rose 1999; Mönch et al. 2013).

Regarding the application of *periodic* order release models in the semiconductor domain, literature has neglected promising advancements of such approaches that were mainly applied to small and medium

enterprises in make-to-order environments. This is quite surprising, since a very vibrant research domain on periodic WLC models evolved with a focus on small and medium enterprises in make-to-order environments (Thuerer et al. 2011; Hutter et al. 2018). However, the only periodic order release models which were applied to semiconductor manufacturing were optimisation based models (Hackman and Leachman 1989; Hung and Leachman 1996; Albey and Uzsoy 2015). Please note that a review of optimisation based order release models as well as scheduling in the semiconductor industry is beyond the scope of this paper, we refer the interested reader to Uzsoy et al. (1994), Missbauer and Uzsoy (2011) and Haeussler et al. (2020).

Therefore, this paper is the first to analyze the performance of a well-established periodic rule based order release mechanism - the Corrected Aggregate Load Approach (COLA) (Oosterman et al. 2000; Thuerer et al. 2012) - in the semiconductor domain. Furthermore, we also contribute to an ongoing discussion within this stream of research with regard to the length of the so called "Time Limit" which determines how many orders from the order pool should be considered for release, i.e., only orders are considered whose due dates are within the time limit (Wiendahl 1995; Thuerer et al. 2012; Haeussler and Netzer 2019). Therefore, we analyze the influence of tightening this time limit on the performance of this method. This is done by using a simulation model of a scaled down semiconductor fab (Kayton et al. 1997) and we will compare it to two well-known continuous approaches - SA and ConLOAD (Glassey and Resende 1988; Rose 1999). Performance will be measured twofold: First, by cost-based measures consisting of WIP, Finished Goods Inventory (FGI) and backorder costs and, second, by the mean and absolute due date deviation representing the timing performance.

The remainder of this paper is structured as follows: The next Section introduces the tested order release approaches and thereafter in Section 3 we outline the used simulation model and the experimental design. In Section 4 we present our results before we summarize and conclude in Section 5.

2 ORDER RELEASE APPROACHES

2.1 Starvation Avoidance (SA)

SA is a purely continuous release mechanism and focuses on releasing new orders whenever the aggregate load (direct and indirect) of the bottleneck drops below a pre-determined lower bound (Fowler et al. 2002; Glassey and Resende 1988). The indirect load is hereby defined as all orders that are upstream of the bottleneck prior to their first visit to the bottleneck work center and all orders that are upstream of the bottleneck within a defined time frame. This time frame represents the lead time of the bottleneck, i.e., the time required for an order to arrive at the bottleneck for the first time once being released (Glassey and Resende 1988). Since SA only controls the release of orders that undergo processing at the bottleneck work center, all non-bottleneck-products are released immediately at order arrival in the order pool (Thuerer et al. 2017).

With regard to the implementation of SA (Glassey and Resende 1988), Table 1 shows what needs to be specified first.

Table 1: Implementing SA (Glassey and Resende 1988).

Notion	Description
B	Bottleneck work center
m	Number of machines at B
K_i	Current number of orders at step i (in queue or in process)
w_i	Work center corresponding to step i
d_i	Processing time at step i
i_0	Process step corresponding to first visit to B
$S_B = (i\|w_i = B)$	Set of all bottleneck work center steps
$F = (1,...,i_0 - 1)$	Set of process steps prior to first visit to B

Moreover, the aforementioned lead time L of the bottleneck B is calculated as the sum of the expected processing times over the process steps of F:

$$L = \sum_{i=1}^{i_0-1} d_i. \tag{1}$$

In addition, n_i defines the process step number corresponding to the next visit to the bottleneck work center considering that the respective order is currently at step i. Consequently, set P represents all those process steps whose expected processing time plus the expected processing time of the subsequent process steps prior to n_i is less than L. Thus, P is given by

$$P = (i| \sum_{j=i}^{n_i-1} d_j < L). \tag{2}$$

Subsequently, the set of critical steps $Q = F \cup P \cup S_B$ can be determined. In this regard, Q contains all steps that are performed prior to the first bottleneck work center visit (i.e., set F) or that are within the lead time of the bottleneck (i.e., set P) and additionally contains all bottleneck steps (i.e., set S_B). Now, the virtual inventory W represents all work in the system at the critical steps Q and thus represents the aggregate load of the bottleneck work center B (and the respective number of machines m), which can be calculated as follows:

$$W = \frac{\sum_{i \in Q} K_i * d_{n_i}}{m}. \tag{3}$$

The virtual inventory W is continuously updated as orders traverse through the system and whenever this virtual inventory W drops below $\alpha * L$ with $\alpha > 0$, the bottleneck work center is likely to starve and hence, SA pulls orders forward from the order pool to the shop floor until this critical threshold level is exceeded. In this regard, different safety levels can be set by altering the α-value to take into account manufacturing uncertainties, such as processing time variability. More precisely, a higher α-value represents a higher safety level thus releasing orders earlier compared to a lower α-value. From above it can be concluded that one needs to calculate and continuously update W and determine an appropriate α-value when applying SA (Glassey and Resende 1988).

2.2 CONstant LOAD (ConLOAD)

CONstant LOAD (ConLOAD) is a continuous order release approach that aims at holding the bottleneck workload at a pre-determined level. ConLOAD extends Workload Regulation which calculates the load contribution of a job simply by summing up the processing times at the bottleneck work center. By contrast, ConLOAD divides this sum of bottleneck processing times by an estimate of the shop floor time of the corresponding product type. Thus, when using ConLOAD, one needs to estimate the average shop floor time of each product type (Rose 1999).

Regarding the order release decision under ConLOAD, arriving orders are collected in an order pool and whenever the current total bottleneck load drops below a pre-defined lower limit (hereinafter denoted as ConLOAD limit), ConLOAD pulls orders forward from the order pool to the shop floor until this critical threshold level is exceeded. If an order is released, its bottleneck load contribution is added to the current bottleneck load. Moreover, the total load contribution of an order to the bottleneck work center is not removed from the total bottleneck load until the underlying order has finished processing (i.e., all process steps have been performed). However, if the order has left the shop floor, the total bottleneck load is reduced by the total bottleneck load contribution of the underlying order (Rose 1999).

Similar to SA, CONLOAD only controls orders that undergo processing at the bottleneck work center and therefore all non-bottleneck-products are released immediately at the moment of arrival in the order pool.

2.3 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):

1. All arriving orders are collected in an order pool.
2. At the beginning of each period, the release procedure checks whether the first order in the order pool violates the predefined workload norm (upper bound for aggregate workload, wherein the contributed workload is calculated by dividing the processing time by the respective process step number in the routing of a job) of any work center.
3. If no workload norm is violated, the order is released, the corresponding corrected workloads (processing times corrected for positions of respective work centers in routing of that order) are added to the work centers on its routing and the next order is selected. If the workload norm of at least one work center is exceeded, the order is kept in the order pool.
4. This procedure is either repeated until all orders were checked (hereinafter *COLA unlimited*) or is stopped when the due date of the respective order is beyond a certain time frame which is called the "Time Limit". Thus, in the latter case (denoted as *COLA tight* hereinafter) the procedure only checks a limited amount of orders in the order pool which, on the one hand reduces the possibilities for balancing, but, on the other, is beneficial regarding timing performance (Land 2006; Haeussler and Netzer 2019).
5. The whole procedure is started again after a certain time which is defined by the release frequency which may be a planning period (e.g., a day).

Note that it is sufficient to set only one workload norm, which is the same for all work centers (Thuerer et al. 2011), and that the workload contribution of an order to a work center is not removed from the current workload of this work center until the respective process step is completed (Oosterman et al. 2000).

3 Simulation Model and Experimental Design

We use a simulation model of a re-entrant bottleneck system which was built with attributes of a real-world semiconductor wafer fab previously studied in WLC research (Kayton et al. 1997; Kacar et al. 2012; Ziarnetzky et al. 2015). The major characteristics of wafer fabrication are multiple products with re-entrant and varying product routings and number of operations, unreliable machines and batch processing machines. This model has one re-entrant bottleneck work center (photolithography process) and includes batching work centers (work centers 1 and 2) early in the process which represent furnaces for diffusion and oxidation processes. All other remaining work centers process one lot at a time. The target utilization for the bottleneck work center is 90%. The model is shown in Figure 1.

The simulation model is made up of 11 work centers, each with one server except the bottleneck work center (work center 4) that has two servers. The processing times for the work centers are log-normally distributed with the standard deviation less than or equal to 10 percent of the mean. Table 2 shows the specific work center processing times and the respective batch sizes. In addition, machines 3 and 7 include machine failures, wherein the mean time to failure (MTTF) and the mean time to repair (MTTR) of these machines are characterized by gamma distributions with the following parameters that are the same for both machines:

- 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

In the following, we denote each visit to a work center as "process step" for consistency and three products are produced which have a varying number of process steps: Product 1 has 22 process steps including 6 visits to the bottleneck work center, product 2 has 14 process steps with 4 visits to the bottleneck

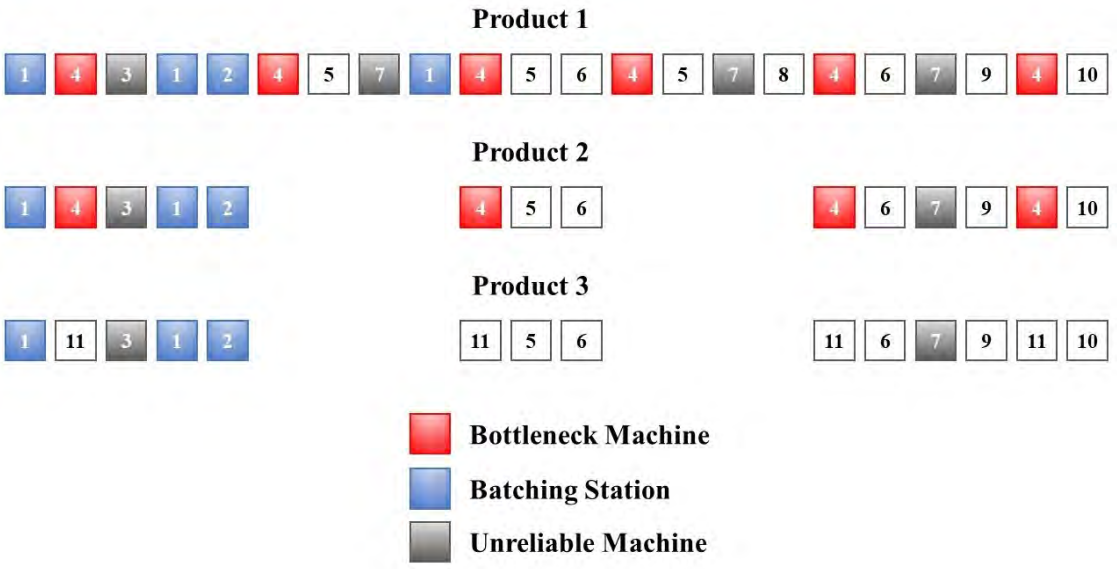


Figure 1: Re-entrant bottleneck model process chart for products (Kacar et al. 2012).

Table 2: Processing times and batch sizes.

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

work center and product 3 has 14 process steps and does not visit the bottleneck. The system is required to produce a product mix that is $3:1:1$ of Product 1, 2, and 3 respectively.

The semiconductor model has two work centers with low reliability that create most of the starvation at the bottleneck. One is situated at the beginning of the process routing and is only visited once by each product which constitutes a "gateway operation" since it opens and closes the flow into the system. The second work center capable of starving the bottleneck is a re-entrant work center that is visited multiple times by the products and occurs later in the processing steps. This machine represents the Chemical Vapor Deposition (CVD) process which produces with high output rates. These two unreliable work centers have the ability to produce many products very quickly, but may also starve the bottleneck due to their poor availability.

In our analysis, we use a stochastic demand with exponentially distributed inter–arrival times. Orders arrive with a mean of one order per 98 minutes and the due dates are set as follows (Kutanoglu 1999; Thuerer et al. 2011; Land 2006; Gupta and Sivakumar 2007; Bahaji and Kuhl 2008): On arrival the product type is randomly assigned by a discrete uniform distribution $dunif\{1,5\}$ (1-3: product type 1, 4: product

type 2, and 5: product type 3). For due date setting we use two different parameterizations: a loose and a tight due date slack which is determined by adding a random allowance where, under the loose due date setting, the minimum slack is defined as 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):

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

For the tight due date setting, the minimum and the maximum slack are set to five times and eleven times the total processing time of product type 1 which is 5,620 minutes and 12,364 minutes respectively:

$$DD_j = AT_j + unif\{5,620\,;12,364\}. \tag{5}$$

Here, DD_j denotes the due date and AT_j the arrival time of order j and the random allowance was set such that an Immediate Release strategy yields a percentage of tardy jobs between 0% and 5% under a loose due date slack and between 5% and 10% under a tight due date slack. Having described the simulation model, now the focus is on the experimental design which is depicted in Table 3.

Table 3: Experimental Design.

Loose Due Date Slack		**Tight Due Date Slack**	
Order Release Model	**Tested Parameters**	**Order Release Model**	**Tested Parameters**
Starvation Avoidance (SA)	α (2.5; 3; 3.5; 4; 4.5)	Starvation Avoidance (SA)	α (3.5; 4; 4.5; 5; 5.5)
ConLOAD	ConLOAD limit (2.0; 2.1; 2.2; 2.3; 2.4)	ConLOAD	ConLOAD limit (2.3; 2.4; 2.5; 2.6; 2.7)
COLA unlimited	workload norm (1,900; 2,000; 2,100)	COLA unlimited	workload norm (1,900; 2,000; 2,100)
COLA tight	workload norm (1,900; 2,000; 2,100) time limit (8,640; 10,080; 11,520; 12,960)	COLA tight	workload norm (1,900; 2,000; 2,100) time limit (5,760; 7,200; 8,640; 10,080)

The three above described order release approaches, SA, ConLOAD and COLA are analyzed with different set of parameters (see Table 3). Note that we used earliest due date as the pool sequencing rule in all tested scenarios. Therefore, the most urgent order based on the external due date is considered first for order release. Nevertheless, for both due date slacks, SA and ConLOAD are analyzed with five different α-values or ConLOAD limits respectively. Since ConLOAD requires an estimation of the average shop floor time for each product type, pilot simulations runs using an Immediate Release scenario were conducted. Moreover, to apply COLA without a time limit (i.e., COLA unlimited) we test three workload norms and for COLA with a time limit (i.e., COLA tight) we test three workload norms each together with four time limits for both due date slacks respectively. Thus, in total 50 different scenarios are simulated, wherein the parameters have been specified based on pilot simulations runs. The used dispatching rule is First-In First-Out throughout all investigated scenarios. Therefore, the results are solely dependent on the specific Order Release approach and the respective parameterization.

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. A cost function was defined to evaluate the results 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:

$$\text{Total Costs} = \sum_{t=1}^{T}\sum_{n=1}^{N} \omega WIP_{n,t} + \sum_{t=1}^{T}(\pi FGI_t + \kappa BO_t) \tag{6}$$

We set the cost parameters ω, π and κ 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).

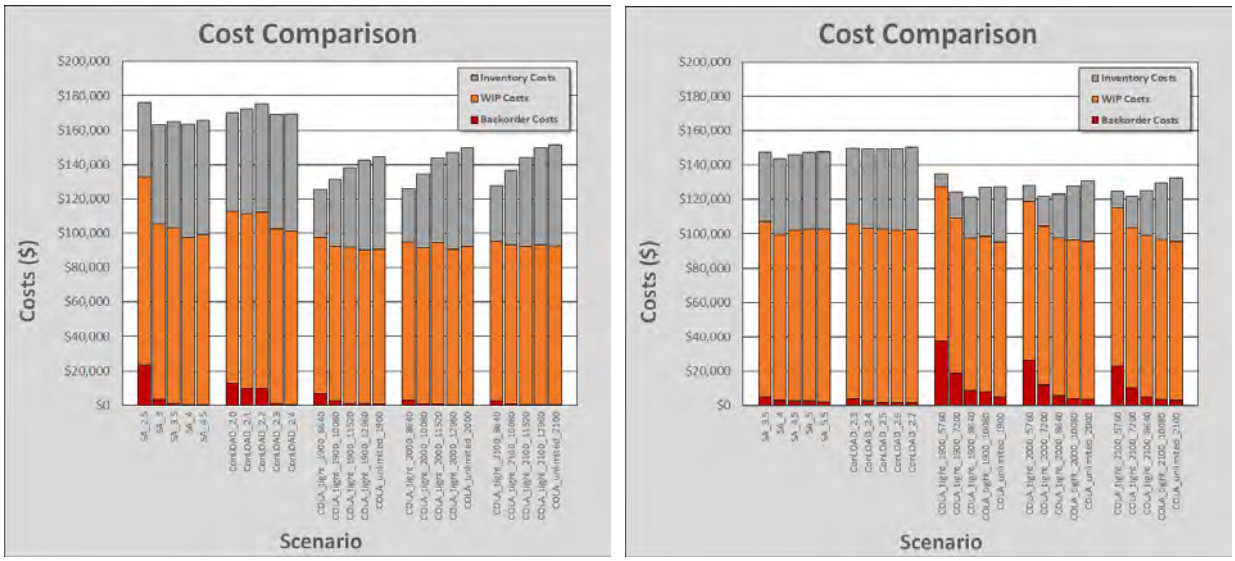
4 RESULTS

In this section, the results for the different order release approaches under a loose and a tight due date slack are discussed. Table 4 shows the timing and cost measures over all replications for the simulated scenarios. The first column denotes the tested order release approach and the corresponding parameterization. For brevity, we use a double for SA and ConLOAD, and a triple or quadruple to denote each COLA scenario: The first component corresponds to the order release mechanism (SA, ConLOAD or COLA), the second component denotes the tested α value (for SA) or the ConLOAD limit or specifies whether a time limit for COLA is used (unlimited $\rightarrow$ without time limit, tight $\rightarrow$ with time limit). Moreover, regarding COLA, in the third component the numbers between $1,900-2,100$ represent the analyzed workload norms for all the work centers and in the fourth component the numbers between $8,640-12,960$ represent the time limit (only for COLA tight). The remaining columns depict the Mean Due Date Deviation (MDDD), Absolute Due Date Deviation (ADDD), and the average Backorder (BOC), WIP (WIPC), Inventory (INVC) and Total Cost (TOTC) values over all replications. All results are compared to the best performing scenario (lowest total costs) which is *COLA_tight_1900_8640* under both tight and loose due date slack which is highlighted in bold in Table 4 . Differences are tested at a significance level of $p = 0.05$ using a Wilcoxon/Mann-Whitney-U Test where all values marked with an asterisk are not significantly different from the corresponding best performing model. Furthermore, to preserve readability we indicated the best performing scenario from each order release model for both due date slacks in italics. Regarding a loose due date slack, one can see that compared to the best scenario *COLA_tight_1900_8640*, the best SA scenario *SA_3* yields \$37,697.75, the best ConLOAD scenario *ConLOAD_2.3* yields \$43,571.24 and the best COLA unlimited scenario *COLA_unlimited_1900* yields \$19,183.26 higher total costs on average. In the following, to preserve readability, we denote each of the best performing scenarios simply as "COLA_tight", "COLA_unlimited", "SA" and "ConLOAD". More precisely, for a *loose due date slack* COLA_tight yields slightly but significantly higher backorder costs but significantly lower WIP and inventory costs than SA and ConLOAD. Similarly, COLA_tight yields slightly higher backorder costs, no significantly higher WIP costs but significantly lower inventory costs than COLA_unlimited. Interestingly, the latter shows that a tighter time limit for COLA_tight does not deteriorate balancing performance (WIP costs). When focusing on the cost measures for a *tight due date slack* it can be seen that the due date tightness has little impact on the (relative) performance of the order release approaches: COLA_tight still yields significantly lower total costs compared to SA and ConLOAD, mainly due to a significant WIP and inventory cost reduction that outweighs the slightly higher backorder costs. Similar to above, compared to COLA_unlimited, COLA_tight yields slightly higher backorder costs, no significantly lower WIP costs but significantly lower inventory costs.

Table 4: Timing and Cost Measures for different order release scenarios.

Loose Due Date Slack:						
Scenario	**MDDD**	**ADDD**	**BOC**	**WIPC**	**INVC**	**TOTC**
SA_2.5	-3,520.29	4,881.94	$23,218.01	$110,011.62	$42,887.75	$176,117.38
SA_3	*-5,562.16*	*5,745.40*	*$3,140.52*	*$102,310.47*	*$57,706.59*	*$163,157.59*
SA_3.5	-6,039.66	6,102.85	$1,074.55	$101,816.54	$61,919.96	$164,811.06
SA_4	-6,489.67	6,519.10	$504.12	$96,918.36	$66,305.75	$163,728.23
SA_4.5	-6,511.93	6,527.80	$270.90	$99,061.73	$66,585.06	$165,917.68
ConLOAD_2.0	-5,273.04	6,001.37	$12,434.76	$100,159.19	$57,436.83	$170,030.77
ConLOAD_2.1	-5,689.44	6,249.46	$9,614.96	$101,870.69	$60,947.16	$172,432.81
ConLOAD_2.2	-5,916.31	6,470.65	$9,548.94	$102,682.23	$63,135.81	$175,366.98
ConLOAD_2.3	*-6,475.87*	*6,542.39*	*$1,137.69*	*$101,560.62*	*$66,332.76*	*$169,031.08*
ConLOAD_2.4	-6,706.48	6,735.34	$492.17	$100,792.86	$68,478.14	$169,763.17
COLA_tight_1900_8640	**-2,553.84**	**2,943.59**	**$6,870.62**	**$90,575.60**	**$28,013.61**	**$125,459.84**
COLA_tight_1900_10080	-3,746.17	3,883.40	$2,436.64	$90,055.32*	$38,780.08	$131,272.03
COLA_tight_1900_11520	-4,531.36	4,602.48	$1,260.24	$90,496.82*	$46,583.76	$138,340.82
COLA_tight_1900_12960	-5,085.68	5,138.59	$944.55	$89,406.03*	$52,071.94	$142,422.52
COLA_unlimited_1900	*-5,264.94*	*5,308.63*	*$782.59*	*$89,994.71**	*$53,865.80*	*$144,643.10*
COLA_tight_2000_8640	-2,999.39	3,159.32	$2,838.20	$91,872.66*	$31,311.80	$126,022.66*
COLA_tight_2000_10080	-4,187.86	4,230.80	$771.19	$90,538.76*	$42,850.73	$134,160.67
COLA_tight_2000_11520	-4,817.63	4,854.72	$672.12	$93,749.36	$49,227.74	$143,649.21
COLA_tight_2000_12960	-5,509.13	5,522.83	$252.71	$90,394.07*	$56,233.39	$146,880.16
COLA_unlimited_2000	-5,631.60	5,646.84	$279.01	$91,896.98*	$57,396.08	$149,572.06
COLA_tight_2100_8640	-3,106.31	3,251.19	$2,521.89	$92,745.85*	$32,399.44	$127,667.18
COLA_tight_2100_10080	-4,235.68	4,273.76	$669.21	$92,416.60*	$43,299.36	$136,385.16
COLA_tight_2100_11520	-5,071.01	5,087.11	$288.38	$92,130.30*	$51,718.49	$144,137.16
COLA_tight_2100_12960	-5,541.49	5,553.48	$214.41	$92,963.21*	$56,506.26	$149,683.89
COLA_unlimited_2100	-5,761.90	5,772.47	$188.48	$92,455.83*	$58,735.24	$151,379.54
					* not significant ($p < 0.05$)	
Tight Due Date Slack:						
Scenario	**MDDD**	**ADDD**	**BOC**	**WIPC**	**INVC**	**TOTC**
SA_3.5	-3,791.66	4,096.38	$5,182.94	$101,816.54	$40,219.80	$147,219.28
SA_4	*-4,241.67*	*4,414.73*	*$2,949.46*	*$96,918.36*	*$44,123.99*	*$143,991.81*
SA_4.5	-4,263.93	4,417.28	$2,619.46	$99,061.73	$44,323.75	$146,004.94
SA_5	-4,288.99	4,431.14	$2,419.70	$100,242.95	$44,502.64	$147,165.29
SA_5.5	-4,351.99	4,485.76	$2,276.51	$100,278.95	$45,088.10	$147,643.56
ConLOAD_2.3	-4,227.87	4,468.17	$4,087.53	$101,560.62	$44,308.65	$149,956.81
ConLOAD_2.4	-4,458.48	4,602.85	$2,460.37	$100,792.86	$46,162.51	$149,415.74
ConLOAD_2.5	*-4,539.01*	*4,636.18*	*$1,647.43*	*$101,013.07*	*$46,720.00*	*$149,380.51*
ConLOAD_2.6	-4,625.13	4,731.11	$1,810.02	$100,083.11	$47,697.49	$149,590.62
ConLOAD_2.7	-4,648.00	4,742.81	$1,609.93	$100,622.71	$47,880.10	$150,112.74
COLA_tight_1900_5760	322.09	1,841.54	$37,230.77	$90,027.79*	$7,747.79	$135,006.35
COLA_tight_1900_7200	-988.81	2,059.60	$18,587.69	$90,396.14*	$15,511.65	$124,495.47*
COLA_tight_1900_8640	**-2,072.70**	**2,571.09**	**$8,705.66**	**$89,020.65**	**$23,610.51**	**$121,336.82**
COLA_tight_1900_10080	-2,529.59	3,003.74	$8,299.94*	$90,357.40*	$28,171.63	$126,828.97
COLA_unlimited_1900	*-3,016.94*	*3,314.61*	*$5,215.24*	*$89,994.71**	*$32,240.79*	*$127,450.73*
COLA_tight_2000_5760	-117.90	1,657.61	$26,551.13	$92,383.60	$9,039.01	$127,973.74
COLA_tight_2000_7200	-1,374.75	2,054.55	$11,818.59	$92,407.57	$17,438.26	$121,664.42*
COLA_tight_2000_8640	-2,359.82	2,676.59	$5,548.86	$91,973.31	$25,615.89	$123,138.06
COLA_tight_2000_10080	-2,971.11	3,202.37	$4,078.04	$92,459.50	$31,418.74	$127,956.28
COLA_unlimited_2000	-3,383.60	3,583.30	$3,514.40	$91,896.98*	$35,444.94	$130,856.31
COLA_tight_2100_5760	-294.00	1,626.06	$22,888.80	$92,195.51	$9,752.48	$124,836.78*
COLA_tight_2100_7200	-1,503.90	2,092.88	$10,184.39	$93,305.37	$18,278.50	$121,768.26*
COLA_tight_2100_8640	-2,430.04	2,733.60	$5,279.17	$93,780.02	$26,281.41	$125,340.61
COLA_tight_2100_10080	-3,100.53	3,301.53	$3,509.15	$93,243.25	$32,593.14	$129,345.54
COLA_unlimited_2100	-3,513.90	3,696.22	$3,176.65	$92,455.83	$36,694.88	$132,327.36
					* not significant ($p < 0.05$)	

Figures 2a and 2b illustrate the total costs and the cost distribution between the simulated scenarios regarding backorder, WIP and inventory costs for a loose and a tight due date slack respectively. One can see that, under both due date slacks, all COLA_unlimited scenarios yield lower total costs than all SA and ConLOAD scenarios. Moreover, under a loose due date slack and for a given workload norm, all COLA_tight scenarios lead to lower total costs on average than the corresponding COLA_unlimited scenario. By contrast, under a tight due date slack and for a given workload norm, tightening the time limit improves the cost performance until a certain time limit is reached. By further tightening the time limit, the cost performance gets worse. In general, tightening the time limit results in a later release of orders (on average) which reduces FGI costs, but this positive effect might be outweighed by increasing backorder costs.



a) Comparison of the costs between different parameterized order release models for a loose due date slack.

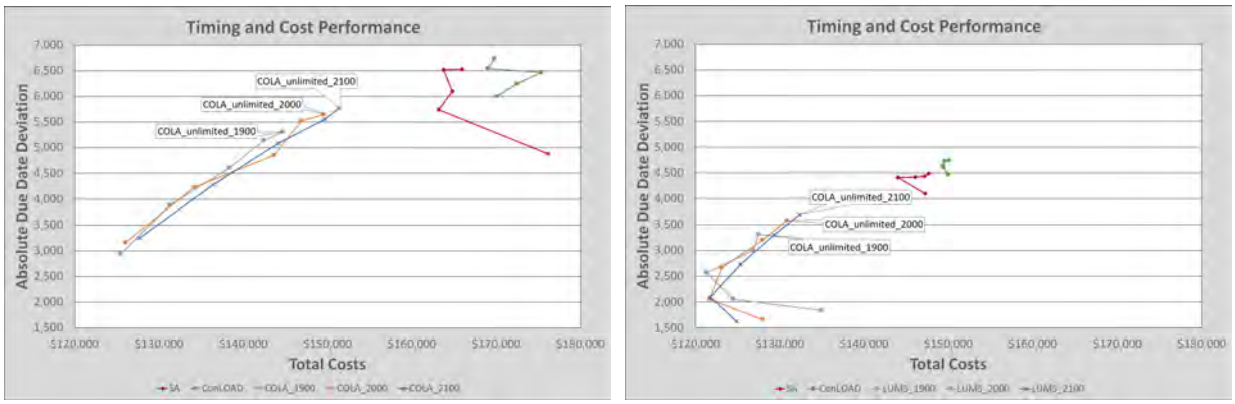
b) Comparison of the costs between different parameterized order release models for a tight due date slack.

Figure 2: Comparison of the costs for a loose and a tight due date slack.

Regarding timing performance, it can be seen that, as expected, including a time limit into the order release decision under COLA significantly improves the timing performance in terms of mean and absolute due date deviation. More precisely, under a *loose due date slack*, COLA_tight yields a 3,008.32, 3,922.03 and 2,711.10 minutes lower mean due date deviation and a 2,801.81, 3,598.80 and 2,365.04 minutes lower absolute due date deviation than SA, ConLOAD and COLA_unlimited respectively. Similarly, under a *tight due date slack*, COLA_tight yields a 2,168.97, 2,466.31 and 944.24 minutes lower mean due date deviation and a 1,843.64, 2,065.09 and 743.53 minutes lower absolute due date deviation than SA, ConLOAD and COLA_unlimited respectively.

Figures 3a and 3b illustrate the absolute due date deviation and the total costs for the simulated scenarios under both due date slacks. One can see that applying a time limit significantly reduces the total costs while at the same time reducing the mean and absolute due date deviation thus leading to a much better due date performance. However, the due date performance already improves when including a rather loose time limit. Nevertheless, a step-wise tightening of the time limit on the one hand improves the due date as well as the cost performance but, on the other, introduces more tardy jobs which results in higher backorder costs that are outweighed by lower inventory costs due to more orders being delivered on time or at least

closer to their external due date. However, under a tight due date slack, although the due date performance improves when tightening the time limit, the cost performance gets worse when the time limit is too tight.



a) Timing and Cost Performance for different parameterized order release models for a loose due date slack.

b) Timing and Cost Performance for different parameterized order release models for a tight due date slack.

Figure 3: Comparison of timing and cost performance for a loose and a tight due date slack.

5 CONCLUSION

This paper compared a rule based workload control approach, the COrrected aggregate Load Approach (COLA), with two well established order release mechanisms, the CONstant LOAD (ConLOAD) (Rose 1999) and the Starvation Avoidance (SA) models (Glassey and Resende 1988), where the former was developed for small and medium enterprises in make-to-order companies and the latter two were developed for semiconductor manufacturing. One of the main differences between these approaches is that SA and ConLOAD are purely continuous and COLA is a purely periodic approach. Since periodic rule based order release models were largely neglected in semiconductor manufacturing, this paper is the first to investigate whether such periodic workload control approaches can improve the cost and timing performance compared to purely continuous order release models that are well established in the semiconductor domain. Therefore, we use a simulation model of a scaled-down semiconductor model (Kayton et al. 1997) and find that, independent of the due date slack, COLA yields lower total costs than SA and ConLOAD. This is mainly due to lower WIP and inventory costs, without worsening the due date performance by much. Furthermore, we also analyzed the impact of including a time limit into the release decision of COLA and how this affects its balancing and due date performance. In this regard, the results showed that, for the simulated case, including a time limit into COLA improves the due date performance without significantly increasing the WIP and thus also reduces the average total costs mainly due to significantly lower inventory costs. These insights are also consistent under a tight due date slack although the absolute performance advantage decreases when tightening the due date slack. However, these findings show that the application of a time limit within the COLA order release approach is a viable alternative for order release in semiconductor manufacturing.

Furthermore, the fact that COLA is a purely periodic order release mechanism should ease its implementation in practice, since it was often argued that periodic decision making is thought to be a better fit with the behavior of planners who typically make release decisions once a shift or day (Hendry and Kingsman 1991; Sabuncuoglu and Karapınar 1999; Stevenson et al. 2011; Thuerer et al. 2012). Furthermore, in comparison to earlier purely periodic WLC methods, the COLA model has the advantage of setting only one initial WLC norm.

The study provides important insights, but we are aware of its limitations. Firstly, the results are limited to

the simulated case and the validity of the results for e.g. large-scale semiconductor fabs must be assessed in future studies. Secondly, adding further experimental factors would be beneficial like analyzing different demand patterns or including different pool sequencing and scheduling rules. Furthermore, future studies should also compare the considered COLA tight model to the hybrid LUMS-COR model also in combination with a time limit, and also 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).

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.

Bergamaschi, D., R. Cigolini, M. Perona, and A. Portioli. 1997. "Order Review and Release Strategies in a Job Shop Environment: A Review and a Classification". *International Journal of Production Research* 35(2):399–420.

Bertrand, J. W. M., and J. C. Wortmann. 1981. *Production control and information systems for component manufacturing shops*. New York: Elsevier Science Inc.

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. 2019. "Comparison between Rule- and Optimization based Workload Control Concepts: A Simulation Optimization approach". *International Journal of Production Research*:1–20. forthcoming.

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.

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.

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

Mönch, L., J. Fowler, and S. Mason. 2013. *Production Planning and Control for Semiconductor Wafer Fabrication Facilities: Modeling, Analysis, and Systems*. Springer Science & Business Media.

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.

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.

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., 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.

Thuerer, M., M. Stevenson, C. Silva, and T. Qu. 2017. "Drum-buffer-rope and workload control in High-variety flow and job shops with bottlenecks: An assessment by simulation". *International Journal of Production Economics* 188:116–127.

Uzsoy, R., C.-Y. Lee, and L. A. Martin-Vega. 1994. "A review of production planning and scheduling models in the semiconductor industry part II: Shop-floor control". *IIE Transactions* 26:44–55.

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 currently 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. philipp.neuner@uibk.ac.at

STEFAN HAEUSSLER is currently Assistant Professor at the Department of Information Systems, Production and Logistics Management at the University of Innsbruck. He received his PhD from the University of Innsbruck, School of Management. His areas of interest include manufacturing planning and control, simulation modeling, workload control, optimization models, forecasting, regression and behavioral operations management. stefan.haeussler@uibk.ac.at

QUIRIN ILMER is currently working as assistant researcher 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 2017 and is studying for his PhD degree in Management at the University of Innsbruck. quirin.ilmer@uibk.ac.at