Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

PUTTING A PRICE TAG ON HOT LOTS AND EXPEDITING IN SEMICONDUCTOR MANUFACTURING

Philipp Neuner
Stefan Haeussler
Julian Fodor

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

Gregor Blossey

Department of Information Systems,
Production and Logistics Management
University of Innsbruck
Innsbruck, 6020, AUSTRIA
and
Department of Business Informatics and
Operations Research
European University Viadrina
Frankfurt, 15230, Germany

ABSTRACT

A common practice in semiconductor manufacturing is to give higher priority to certain “hot lots” to reduce their cycle time and deliver them on time. Despite good performance of these high priority lots, expediting might worsen the overall performance of the fab due to decelerating all other lots. Thus, this paper uses a simulation model of a scaled-down wafer fabrication facility, to put a price-tag on hot lots and expediting measures to derive suggestions for decision makers on (i) how much additional profit per hot lot is required to compensate for increasing cost due to introducing hot lots, and (ii) the allowable maximum expediting cost per period.

1 INTRODUCTION

In the age of digitalization, the availability of semiconductors is of strategic importance as they are critical for economic success and even national security. Due to their essential role, a reliable and timely delivery of these components is crucial for customer satisfaction. However, this is a challenging task for decision makers in the industry given the complex manufacturing environment, a high degree of product differentiation, and a multitude of product-specific routings through the manufacturing system. Therefore, sophisticated coordination approaches are required, which take these aspects into account and, if needed, re-sequence production orders to meet important deadlines. In semiconductor manufacturing some lots have higher processing priority resulting in two classes of lots, namely high priority orders which are often denoted as “hot lots” and regular orders which are hereinafter denoted as “cold lots”. Hot lots appear for different reasons: They can be lots for developing new processes, pilot lots for new products, sample lots, and lots for urgent orders. However, the presence of hot lots affects the performance of cold lots and thus impacts overall performance (Ehteshami et al. 1992; Narahari and Khan 1997; Trybula 1993).

The paper by Ehteshami et al. (1992) studies the effect of hot lots on the cycle time of cold lots in a semiconductor fab by using a dispatching rule that always prioritizes hot lots. Their simulation results show that with increasing proportion of hot lots the average cycle time and the corresponding standard deviation of cold lots increases. Zhou and Rose (2012) propose a dispatching rule to reduce tardiness of hot

978-1-6654-7661-4/22/$31.00 ©2022 IEEE

lots, as well as the work-in-process (WIP) balance of cold lots. They combine WIP balance and due date oriented rules in one global dispatching rule and show that their approach results in low tardiness of hot lots without harming WIP measures too much. Crist and Uzsoy (2011) investigate the impacts of several policies for allocating resources to cold and hot lots on the shop floor using a scaled-down simulation model of a wafer fab. Their focus is on engineering lots which are special lots that need the presence of an engineer. They tested static and dynamic batching policies (with trigger when to switch to engineer lots) at the bottleneck machine and showed that several of their tested scheduling policies could potentially be the best fit depending on the goals and priorities of each corporation.

Furthermore, several capacity reservation approaches were introduced that reserve capacity for hot lots that have not yet arrived. This may cause a negative influence on the flow of cold lots by decreasing tool utilization, which is one of the major limitations of the reservation policy (Seo et al. 2015; Chung et al. 2017). Seo et al. (2015) introduce a two-step dispatching rule which in the first step tries to minimize waiting times of hot lots and secondly aims for high utilization of the used tools. By using simulation they show that their two-step dispatching rule outperforms static rules regarding on-time delivery of hot lots. Chung et al. (2017) extend the former paper and present a dispatching rule which considers tool utilization as well as the on-time delivery of hot lots. They use a simulation model to compare their approach with static dispatching and conventional capacity reservation policies. Their results show that their dispatching rule outperforms the benchmarks especially regarding the performance of cold lots. Finally, some papers also include transport and material handling system into their analysis of how to handle the different classes of lots: Wang and Chen (2012) propose a heuristic preemptive dispatching rule to ensure that the production of hot lots will not be hindered by its automated material handling system. Ho et al. (2016) analyze the performance of hot and cold lots in a thin-film-transistor liquid-crystal display fab. They use a simulation model to compare different lot selection approaches for inter bay automated guided vehicles, intrabay machines and the photo bay selection of lots and compare their results to an industry benchmark. They find that if their newly introduced fuzzy-based dynamic bidding (for the lot selection problems) and the earliest possible time method (for photo bay selection) are combined they perform better than the industry benchmark.

However, all of these above papers neither evaluate the cost for introducing hot lots nor has any paper evaluated the cost benefit of expediting hot lots instead of non-expediting. Thus, this paper aims to put a price tag on hot lots and expediting which should help decision makers to determine (i) the needed extra-profit for these special orders and (ii) the allowable increase in cost for prioritizing them. Therefore, we use a simulation model of a scaled down wafer fabrication facility and use the Constant Load approach (Rose 1999) for releasing orders over time. We test different utilization levels and ratios of hot lots and compare the cost performance (consisting of holding and backorder costs) of using a First-In First-Out dispatching rule with and without expediting of hot lots.

The paper is organized as follows. In Section 2, we describe the used simulation model and Section 3 outlines the used experimental design. In Section 4 we present the results which is followed by our Conclusion in Section 5.

2 SIMULATION MODEL

Regarding our simulation model we use a standard model from the semiconductor literature which was developed based on attributes of a real-world semiconductor wafer fabrication facility (Kayton et al. 1997; Kacar et al. 2012; Ziarnetzky et al. 2015, see Figure 1). According to Crist and Uzsoy (2011) the simulation model was validated by management of the respective real world fab at the time when the simulation model was developed. The major characteristics of semiconductor manufacturing are multiple products with re-entrant product routings which vary in their length and visits to different work centers. Moreover, it includes unreliable machines and work centers that perform batching operations. The respective model has one re-entrant bottleneck work center which performs the photolithography process and includes two batching work centers (WC1 and WC2). The latter are located at the beginning of the product routings and represent furnaces for diffusion and oxidation processes (see Figure 1).

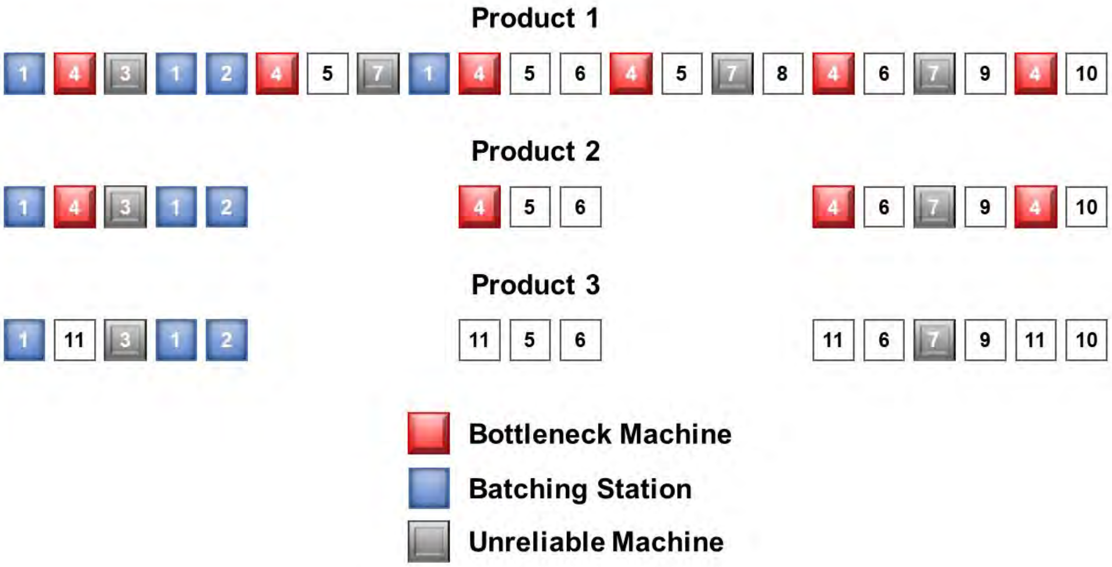


Figure 1: Re-entrant bottleneck model process chart for products (Kacar et al. 2012).

The simulation model includes eleven work centers with one server each except the bottleneck work center 4 that has two. Table 1 depicts the mean processing times and required batch sizes of the work centers.

Table 1: 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

All processing times follow a log-normal distribution where the standard deviation is less than or equal to 10 percent of the corresponding mean. Work centers 1 and 2 can process at least 2 and at most 4 different types of products at once, while all other remaining work centers process only one lot at a time. Machine failures are also included in the simulation model, as machines 3 and 7 break according to the following gamma distributions:

- Mean Time To Failure: α = 7,200, β = 1 → mean = 7,200, Std. Dev. = 84.9;
- Mean Time To Repair: α = 1,200, β = 1.5 → mean = 1,800, Std. Dev. = 52.0.

In the model, there are three different products with varying routings: 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 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 two unreliable work centers 3 and 7 create most of the starvation at the bottleneck. As work center 3 is additionally situated near the beginning of the process routings and is only visited once by each product, it represents a gateway operation by opening and closing the flow of lots into the system. The other unreliable work center can also lead to starvation at the bottleneck work center as it is a re-entrant work center that is visited multiple times by product 1. This machine represents the Chemical Vapor Deposition process which produces with high output rates.

3 EXPERIMENTAL DESIGN

This section introduces the experimental design of our study in terms of demand, order release, machine dispatching and the respective parameterizations.

3.1 Demand

The demand is generated based on exponentially distributed inter–arrival times. This stochastic demand is varied at two levels: in the low demand setting the bottleneck utilization was parameterized based on an Immediate Release strategy to yield approximately 90%, i.e. orders arrive with a mean of one order per 98 minutes, and in the high demand setting the bottleneck utilization is approximately 95%, i.e. an order arrives on average every 93 minutes. To represent the desired product mix of 3:1:1 of product types 1, 2 and 3, the product type is randomly assigned based on a discrete uniform distribution, i.e. *dunif*{*1,5*} where 1-3 represents product type 1, 4 represents product type 2, and 5 represents product type 3. With regard to due date setting the due dates are set according to previous research (Mosley et al. 1998). First, depending on the proportion *x%* of hot lots, orders are randomly defined as ”hot” or ”cold” based on a uniform distribution (*unif*{*0,100*}). If the random number is smaller or equal to *x*, the order is defined as hot irrespective of its product type. Otherwise, the order is a cold lot. The due date of hot lots is then defined as follows:

$$DD_j = AT_j + unif\{0.5\,;1\} * CT_i, \tag{1}$$

where DD_j represents the due date of order j, AT_j is the arrival time of order j and CT_i denotes the average cycle time of the corresponding product type i of order j. On the other hand, the due date of cold lots is defined as follows:

$$DD_j = AT_j + unif\{1.25\,;1.75\} * CT_i. \tag{2}$$

CT_i was specified based on preliminary simulation runs using an Immediate Release strategy.

3.2 Order Release

Besides Immediate Release, we also apply the so-called Constant Load (ConLOAD) approach in our simulation study. ConLOAD is a continuous order release approach which seeks to keep the bottleneck workload at a pre-defined level. This means that orders are released from an order pool until a certain threshold level is reached which is denoted as ConLOAD limit. Once the ConLOAD limit is reached an order can only be released from the order pool to the shop floor if an order has finished processing. The workload contribution of an order to the bottleneck workload is hereby defined by the ratio of the sum of bottleneck processing times and the mean cycle time of the respective product type (specified like above). An order contributes to the bottleneck workload until its last processing step was performed. Once processing is finished, the total workload contribution of the underlying order is removed from the bottleneck workload. In that case, a new order can be released from the order pool if the ConLOAD limit is not violated (Rose 1999).

Since ConLOAD only controls orders that include a processing step at the bottleneck work center, all other orders (i.e. product type 3) are released immediately when they arrive in the system.

3.3 Machine Dispatching

The sequence in which orders waiting in front of a work center are processed is based on the First-In First-Out logic. However, within this FIFO approach hot lots are handled in two different ways: In the first case, no difference is made between hot and cold lots. But in the second case, hot lots are given priority over cold lots, i.e. an expediting mechanism is applied. Here, given the FIFO-sequence, hot lots are always processed before cold lots.

3.4 Tested Parameters

An overview of the experimental design is provided in Table 2. The bottleneck utilization is varied at two levels, the proportion of hot lots at three levels and two order release approaches are applied where five ConLOAD limits are tested for ConLOAD. While a proportion of 40% hot lots seems quite high, we want to reveal the boundaries of expediting measures. Of course, these boundaries also depend on other factors such as the cost parameters. Note that Immediate Release can also be interpreted as a ConLOAD scenario with an infinite ConLOAD limit. Regarding machine dispatching, two alternatives are included in the study. Based on a full factorial design, 60 scenarios have been simulated. Note that FIFOEXP only makes sense for a proportion of hot lots greater than 0%. Regarding pool sequencing we also relied on First-In First-Out. Therefore, orders are considered for release in the sequence of arrival.

Table 2: Overview of Tested Parameters.

Experimental Factor	**Tested Parameters**
Bottleneck Utilization	Low (90%), High (95%)
Proportion of Hot Lots	0%, 20%, 40%
Order Release	Immediate Release (IMRE) ConLOAD (2, 2.25, 2.5, 2.75, 3)
Machine Dispatching	FIFO (without expediting), FIFOEXP (including expediting of hot lots)

The period length was set to 1,440 minutes (one day), each scenario was replicated 100 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) cost over all periods t:

$$\text{Total Cost} = \sum_{t=1}^{T}\sum_{n=1}^{N} \omega WIP_{n,t} + \sum_{t=1}^{T}(\pi FGI_t + \kappa BO_t) \tag{3}$$

The relation of the cost parameters ω, π and κ were set according to previous research on semiconductor manufacturing (Kacar et al. 2012; Kacar et al. 2013; Albey and Uzsoy 2015; Ziarnetzky et al. 2015; Neuner and Haeussler 2021; Neuner 2021): $2\frac{1}{3}$: 1 : $3\frac{1}{3}$. Although it might appear more reasonable to discriminate the backorder costs for cold and hot lots, we want to provide a method which, independent of the cost structure, can be used to monetize expediting measures taken in practice. Note that we provide a very conservative analysis since all effects would be larger when we set different costs for hot and cold lots (e.g., higher backorder costs for hot lots).

4 RESULTS

In this section we analyze the results for the simulated scenarios under a low and high demand. For brevity, we only present the results for the best performing scenarios in terms of total cost measures but provide the results for all simulated scenarios in our data repository available under: http://dx.doi.org/10.17632/zvm9b7mgvv.1. Table 3 shows the cost measures for the best performing scenarios under a low (upper part) and high demand (lower part). The first column denotes the respective scenario and its parameterization. A quadruple is used for each scenario: The first component corresponds to the machine dispatching approach, i.e. FIFO → without expediting and FIFOEXP → with expediting of high priority orders. The second component denotes the ConLOAD limit (note that a ConLOAD limit of 1000 in the data repository means Immediate Release), the numbers 0, 20 or 40 in the third component represent the proportion of hot lots in % and the fourth component denotes the demand level. The second column then highlights which results are referred to in the remaining columns, i.e. the results over all orders in a given scenario or only the results for hot or cold lots. The remaining columns show the mean Backorder, WIP, Finished Goods Inventory (FGI), Timing (Backorder + FGI) and Total cost values over all replications.

For a given proportion of hot lots, here 20% or 40%, differences between the respective FIFO and FIFOEXP scenarios 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 the corresponding FIFOEXP values. For example, *FIFO_3_20_90* does not yield significantly higher WIP cost than *FIFOEXP_2.75_20_90*.

It can be seen that, for a low demand, expediting hot lots improves the overall cost performance compared to handling all lots the same way. While the WIP cost between the respective FIFO and FIFOEXP scenarios do not significantly differ, expediting reduces total cost due to improving the timing performance of orders as indicated by the lower backorder and inventory cost. Focusing on the cost measures only for hot lots, we can say that expediting hot lots significantly reduces backorder and WIP cost as those high priority lots rush through the shop floor which leads to higher inventory cost due to finishing them earlier than their due date. Overall, the total cost of hot lots can be drastically reduced by expediting measures. However, by reviewing the cost measures for cold lots, the downside of expediting becomes obvious. By speeding up hot lots, all other lots are given less priority which decelerates them at the shop floor level resulting in higher backorder and WIP cost but lower inventory cost. Nevertheless, this detrimental effect on cold lots is outweighed by the drastic cost reduction for hot lots which results in a cost performance improvement for all lots.

While the qualitative effect is exactly the same for a high demand and a proportion of 20% hot lots, the overall cost performance is deteriorated by expediting for a high demand and 40% hot lots. Again, hot lots rush through the production resulting in lower backorder and WIP cost but higher inventory cost and cold lots are decelerated resulting in higher backorder and WIP cost but lower inventory cost. But in this case, the overall cost performance is no longer improved. Each of the cost measures for all lots is not significantly lower for *FIFOEXP_3_40_95* compared to *FIFO_3_40_95*. This indicates that expediting becomes obsolete for a higher demand level in conjunction with a higher proportion of hot lots.

While the findings so far have demonstrated the potential of expediting hot lots under given circumstances, we now want to focus on putting a price tag on hot lots and expediting measures. This issue is tackled from two perspectives, namely a sales/profit and a operations/cost viewpoint.

Regarding the sales/profit perspective, the relevant question is ”what does it cost if hot lots are introduced in the wafer fab?”. Thus, decision makers need to know how much additional margin hot lots need to generate. This can be relevant for example for a sales department, where this information can be used in the negotiation process with the customer. Therefore, the focus is on the FIFO scenarios, where Figure 2 shows the mean WIP and timing (backorder + inventory) cost measures for the above best performing scenarios for a low and high demand, respectively. We illustrated the respective relationships by the solid (blue and orange) arrows: Since all best performing scenarios without expediting, i.e. FIFO, have the same ConLOAD limits for a given demand level, the WIP cost are exactly the same. However, the higher the proportion of hot lots, the worse the timing performance of orders and hence the higher the timing

Table 3: Cost Measures for the best performing scenarios for different proportions of hot lots and demand levels.

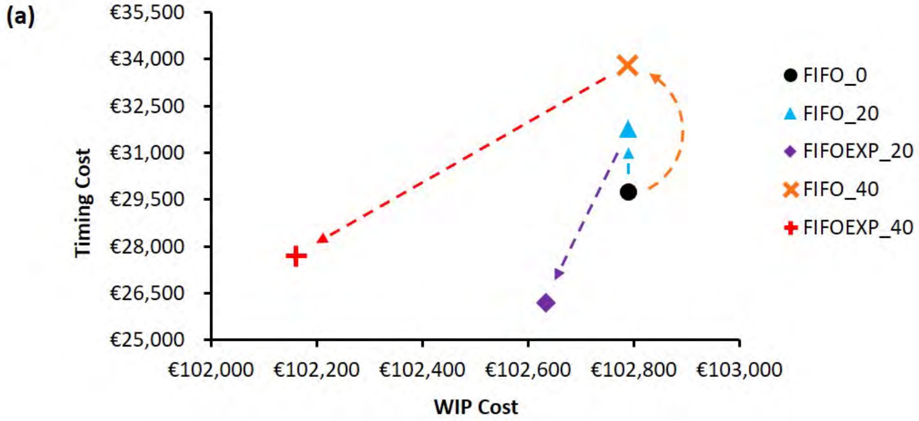
Low Demand (90% Bottleneck Utilization):						
Scenario	**Lot Type**	**Backorder Cost**	**WIP Cost**	**Inventory Cost**	**Timing Cost**	**Total Cost**
FIFO_3_0_90	All (14703)	1090.26	102,789.16	28,681.29	29,771.55	132,560.70
FIFO_3_20_90	All (14703)	8231.08	102,789.16*	23,556.34	31,787.42	134,576.57
	Hot (2945)	7359.60	20,578.55	616.86	7976.46	28,555.01
	Cold	871.48	82,210.60	22,939.48	23,810.96	106,021.56
FIFOEXP_2.75_20_90	All (14702)	4051.36	102,634.64	22,135.88	26,187.24	128,821.88
	Hot (2945)	425.82	9042.92	3433.82	3,859.64	12,902.56
	Cold	3,625.54	93,591.72	18,702.06	22,327.60	115,919.32
FIFO_3_40_90	All (14703)	15,358.20	102,789.16*	18,456.76	33,814.96	136,604.11
	Hot (5880)	14,706.33	41,088.80	1,235.02	15,941.35	57,030.15
	Cold	651.87	61,700.36	17,221.74	17,873.61	79,573.97
FIFOEXP_2.75_40_90	All (14702)	10,622.27	102,160.72	17,073.46	27,695.73	129,856.45
	Hot (5879)	1081.86	19,239.29	6419.44	7501.30	26,740.59
	Cold	9,540.41	82,921.42	10,654.02	20,194.43	103,115.85
						* not significant ($p < 0.05$)
High Demand (95% Bottleneck Utilization):						
Scenario	**Lot Type**	**Backorder Cost**	**WIP Cost**	**Inventory Cost**	**Timing Cost**	**Total Cost**
FIFO_3_0_95	All (15492)	9757.02	133,340.40	19,806.59	29,563.61	162,904.01
FIFO_3_20_95	All (15492)	23,020.56	133,340.40*	16,139.64	39,160.20	172,500.60
	Hot (3099)	15,228.91	26,677.46	288.74	15,517.65	42,195.11
	Cold	7791.65	106,662.94	15,850.90	23,642.55	130,305.49
FIFOEXP_3_20_95	All (15492)	18,301.74	132,983.40	15,075.23	33,376.97	166,360.37
	Hot (3099)	874.81	9602.44	3340.34	4215.15	13,817.59
	Cold	17,426.92	123,380.96	11,734.89	29,161.81	152,542.78
FIFO_3_40_95	All (15492)	36,236.48*	133,340.40*	12,472.42*	48,708.90*	182,049.29*
	Hot (6194)	30,398.63	53,293.93	579.71	30,978.34	84,272.27
	Cold	5837.85	80,046.47	11,892.71	17,730.56	97,777.03
FIFOEXP_3_40_95	All (15492)	34,733.83	133,058.83	12,092.28	46,826.11	179,884.94
	Hot (6193)	2247.91	20,535.98	6152.91	8400.82	28,936.79
	Cold	32,485.92	112,522.85	5939.37	38,425.29	150,948.14
						* not significant ($p < 0.05$)

cost. These differences in timing cost illustrated by the solid (blue and orange) arrows are summarized in Table 4. The absolute cost deviations need to be put in relation to the number of hot lots (see Table 3).

Table 4: Cost deviations from sales/profit perspective for different hot lot proportions and demand levels.

Hot Lots	Low Demand	High Demand
20%	2015.87	9596.59
40%	4043.41	19,145.28

Therefore we conclude for a low demand level, that profits for hot lots need to increase at least by about 20% of the unit backorder cost (e.g. 2015.87 divided by 2945 divided by 3.33) to outweigh the total cost increase compared to having no hot lots in the wafer fab. Additionally, increasing the demand level means that the marginal profit needs to increase even more per hot lot to at least 93% of the unit backorder cost.



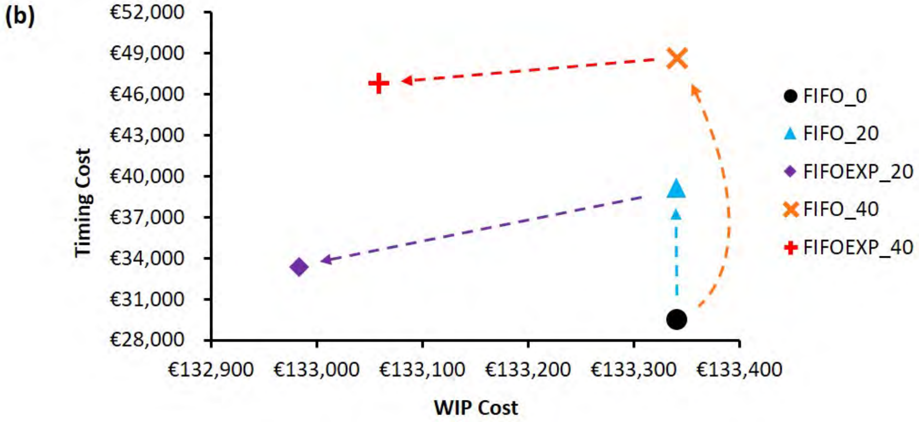


Figure 2: Comparison of the cost for (a) low demand levels and (b) high demand levels.

Regarding the second perspective, we focus on the operations/cost viewpoint: In this regard, we are now asking ”what is the most we can pay for expediting measures without losing money?”. Here the dashed (violet and red) arrows are relevant, i.e. the relationships between FIFO and FIFOEXP for a given level of hot lots. It can be clearly seen in Figure 2 that expediting measures improve the cost performance compared to FIFO as WIP and timing cost are decreased. Although the WIP cost reduction is insignificant, the timing cost can be reduced significantly with one exception of high demand and 40% hot lots (see Table 5). For the operations/cost perspective, we assume that expediting is performed by an operator on

Table 5: Cost deviations from operations/cost perspective for different hot lot proportions and demand levels.

Hot Lots	**Low Demand**	**High Demand**
20%	5600.18	5783.24
40%	6119.23	not significant

the shop floor which means that we are now calculating the maximum cost for this operator per period (1000 periods were simulated). Regarding both demand levels and 20% hot lots, the operator can cost up to about 170% of the unit backorder cost per period, while for a low demand and 40% of hot lots the maximum cost go up to about 180% of the unit backorder cost. Finally, for a high demand and 40% hot lots we do not recommend to execute expediting measures as the higher cost of introducing hot lots can no longer be outweighed by speeding up those high priority lots.

5 CONCLUSION

Earlier semiconductor manufacturing research on high priority – so-called ”hot lots” – shows that speeding up these orders at the machine dispatching level has positive effects on their cycle time and delivery performance. Despite that, these positive effects might be outweighed by deteriorating effects on all other lots as they are decelerated on the shop floor. However, literature focused mainly on the impact of hot lots on the cycle time and the corresponding standard deviation of cold lots (Ehteshami et al. 1992; Zhou and Rose 2012). The questions of how much more profit a hot lot needs to generate and secondly how much the increased effort of expediting may cost is hard to assess and is to the best of the authors’ knowledge not addressed in literature. Thus, the main contribution of our study is to put a price tag on hot lots and expediting measures. Therefore, we use a simulation model of a scaled down wafer fabrication facility and use ConLOAD for making order release decisions. We test different demand scenarios and ratios of hot lots and compare the cost performance (consisting of holding and backorder costs) of using a First-In First-Out dispatching rule with and without expediting of hot lots.

We tackle this issue from a sales/profit and a operations/cost perspective. Starting with the sales/profit viewpoint, we find that for a low demand level profits for hot lots need to increase at least by about 20% of the unit backorder cost to outweigh the total cost increase compared to having no hot lots in the wafer fab. Additionally, for a high demand level this marginal profit per hot lot needs to be increased at least by about 93%. Regarding, the operations/cost view, for both demand levels and 20% hot lots, the operator to expedite hot lots can cost up to about 170% of the unit backorder cost per period. For a low demand and 40% of hot lots the maximum cost go up to about 180%. Finally, for a high demand and 40% hot lots we do not recommend to execute expediting measures as the higher cost of introducing hot lots can no longer be outweighed by speeding up those high priority lots.

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). However, the simulation model still captures the major characteristics of semiconductor manufacturing and also the cost structure is fitted to the wafer fab environment. Nevertheless, the details of the procedures (e.g., defining hot lots based on their average cycle time) depends on the specific wafer fab and need to be parameterized accordingly. Still the study provides a method of how to monetize expediting . Secondly, while expediting measures are often based on a static definition of hot lots, future research could also focus on dynamic hot lot assignments. This seems reasonable assuming that some hot lots are delivered prior to their due date and thus, should have been decelerated at some point at the shop floor. Thirdly, future studies should also consider expediting in conjunction with other order release approaches (Neuner and Haeussler 2021; Schneckenreither et al. 2021), especially workload control mechanisms (Haeussler and

Netzer 2020; Neuner and Haeussler 2021). Fourthly, adding further experimental factors such as different due date slacks, pool sequencing and scheduling rules might be beneficial.

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.

Chung, Y. H., J. C. Seo, C. M. Kim, B. H. Kim, and S. C. Park. 2017. "Reservation-based dispatching rule for make-to-order wafer FAB with high-priority lots". *Concurrent Engineering* 25(1):68–80.

Crist, K., and R. Uzsoy. 2011. "Prioritising production and engineering lots in wafer fabrication facilities: a simulation study". *International Journal of Production Research* 49(11):3105–3125.

Ehteshami, B., R. G. Petrakian, and P. M. Shabe. 1992. "Trade-offs in cycle time management: hot lots". *IEEE Transactions on Semiconductor manufacturing* 5(2):101–106.

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.

Ho, Y.-C., J.-W. Lin, H.-C. Liu, and Y. Yih. 2016. "A study on the lot production management in a thin-film-transistor liquid-crystal display fab". *Journal of Manufacturing Systems* 40(1):9–25.

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.

Kopp, D., M. Hassoun, A. Kalir, and L. Mönch. 2020. "SMT2020A semiconductor manufacturing testbed". *IEEE Transactions on Semiconductor Manufacturing* 33(4):522–531.

Mosley, S. A., T. Teyner, and R. M. Uzsoy. 1998. "Maintenance scheduling and staffing policies in a wafer fabrication facility". *IEEE Transactions on Semiconductor Manufacturing* 11(2):316–323.

Narahari, Y., and L. Khan. 1997. "Modeling the effect of hot lots in semiconductor manufacturing systems". *IEEE Transactions on Semiconductor Manufacturing* 10(1):185–188.

Neuner, P. 2021. "Adaptive rule based order release in semiconductor manufacturing". In *Proceedings of the 2021 Winter Simulation Conference*, edited by S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C.Szabo, and M. Loper, 1–12. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Neuner, P., and S. Haeussler. 2021. "Rule based workload control in semiconductor manufacturing revisited". *International Journal of Production Research* 59(19):5972–5991.

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.

Schneckenreither, M., S. Haeussler, and C. Gerhold. 2021. "Order release planning with predictive lead times: a machine learning approach". *International Journal of Production Research* 59(11):3285–3303.

Seo, J., Y. Chung, and S. Park. 2015. "On-Time Delivery Achievement of High Priority Orders in Order-driven Fabrications.". *International Journal of Simulation Modelling (IJSIMM)* 14(3):475–484.

Trybula, W. J. 1993. "" Hot" jobs, bane or boon". In *Proceedings of 15th IEEE/CHMT International Electronic Manufacturing Technology Symposium*, 317–322. Piscataway, New Jersey: Institute of Electrical and Electronic Engineers, Inc.

Wang, C.-N., and L.-C. Chen. 2012. "The heuristic preemptive dispatching method of material transportation system in 300 mm semiconductor fabrication". *Journal of Intelligent Manufacturing* 23(5):2047–2056.

Zhou, Z., and O. Rose. 2012. "WIP balance and due date control in a wafer fab with low and high volume products". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 1–8. Piscataway, New Jersey: Institute of Electrical and Electronic Engineers, Inc.

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 Associate Professor at the Department of Information Systems, Production and Logistics Management at the University of Innsbruck. His main research focus is on quantitative methods for decision support in the areas of production planning and supply chain management. His research focuses on order release, lead time management, dispatching and their practical application. Methodologically, he focuses on discrete event simulation, optimization, economic experiments and machine learning methods. stefan.haeussler@uibk.ac.at

JULIAN FODOR 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 2022 and is currently studying for his PhD degree in Management at the University of Innsbruck. julian.fodor@student.uibk.ac.at

GREGOR BLOSSEY 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 Business Administration for Engineers and Natural Scientists from the University of Jena in 2017 and is currently a doctoral candidate at the European University Viadrina Frankfurt (Oder). gregor.blossey@uibk.ac.at