Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

A PULL/PUSH CONCEPT FOR TOOLGROUP WORKLOAD BALANCE IN WAFER FAB

Zhugen Zhou
Oliver Rose

Institute of Applied Computer Science
Dresden University of Technology
Dresden, 01187, GERMANY

ABSTRACT

In this paper, a pull/push concept is proposed in order to balance toolgroup workload in a wafer fab. This is accomplished by using a so-called WIP Control Table. Each upstream toolgroup maintains a WIP Control Table which contains current WIP information of downstream toolgroups such as target WIP, actual WIP and WIP difference. In case of lot move in/out and tool status change, the WIP Control Table is updated. Therefore, the upstream toolgroup is able to detect WIP distribution and pull request of downstream toolgroups dynamically, then push optimal lots with consideration of lot status and local tool constraint to the downstream toolgroup which runs short of WIP. The simulation results demonstrate that the proposed pull/push concept is superior over First-in-First-out (FIFO) and Operation Due Date (ODD) with regard to average cycle time and on time delivery.

1 INTRODUCTION

In a real wafer fab, the material flow is generally nonlinear due to different events such as unpredictable tool failure, batch processing, setup change, process dedication and so on. Thus WIP fluctuation occurs frequently, especially lots pile up in front of toolgroup during downtime period. In that case lots spend hours, even days in the queue if there is no appropriate scheduling. This accumulated WIP in front of a certain toolgroup causes WIP imbalance for the whole line, which has great impact on cycle time as well as on time delivery.

Many WIP balance approaches have been already developed to correct for WIP imbalance. Minimum Inventory Variability Scheduling (MIVS) (Li, Tang, and Collins 1996) and Line Balance (LB) (Dabbas and Fowler 2003) are the representatives. MIVS considers both upstream operation and downstream operation. It gives the highest priority to an operation which has a high WIP and its downstream operation has a low WIP, in order to avoid starvation at downstream operation. In contrast, it gives the lowest priority to an operation which has a low WIP and its downstream operation has a high WIP. MIVS tries to keep the WIP of each operation close to the average target WIP level. Similarly, Line Balance algorithm intends to minimize the deviation between the actual WIP and target WIP for each process stage. Through calculating throughput, signals, cumulative signals and unconstrained quantities, LB determines proportion of WIP at all stages which are required to be pushed downstream, so as to balance the line. In both approaches, the target WIP used to correct the actual WIP is determined either by a simulation model or from experience of industrial engineer. They both look at WIP balance at the viewpoint of process operation/stage. However, there are two potential problems with them. First of all, they fall short of taking lot status like due date and local tool constraint like utilization, workload, batch and setup requirement into consideration. Therefore, they have a shortage of optimizing cycle time and on time delivery simultaneously. Secondary, the toolgroup which is shared by different operations has a high chance to be congestion. If the high WIP toolgroup is a non-bottleneck, which makes the bottleneck starved and degrades wafer fab's performance.

978-1-4244-9864-2/10/$26.00 ©2010 IEEE

This study is motivated by considering WIP balance at the viewpoint of toolgroup instead of process operation/stage like MIVS and LB. Thereby, a pull/push concept is proposed to balance the WIP of toolgroup. With the help of a WIP Control Table, the upstream toolgroup is capable of detecting a pull request from downstream toolgroup dynamically. Among the lots which fulfill the demand from downstream toolgroup with pull request, the upstream toolgroup pushes optimal lots based on lot status and local tool constraint, which on one hand maintains actual WIP level close to target WIP level of downstream toolgroup, on the other hand optimizes other performance indicators like cycle time and on time delivery.

This paper is organized as follows. In Section 2, we describe the proposed pull/push concept in detail. In Section 3, we present and compare the simulation results with First-in-First-out (FIFO) and Operation Due Date (ODD) with regard to average cycle time, cycle time variance, on time delivery and so on. Section 4 gives conclusion and further work.

2 THE PULL/PUSH CONCEPT FOR WORKLOAD BALANCE OF TOOLGROUP

2.1 WIP Control Table

To deal with WIP imbalance of toolgroup, a WIP Control Table is proposed for each toolgroup in the fab. Each upstream toolgroup maintains a WIP Control Table which contains current WIP information of all its downstream toolgroups e.g. target WIP level, actual WIP level, WIP difference and utilization. Figure 1 and Table 1 describe an example of WIP Control Table.

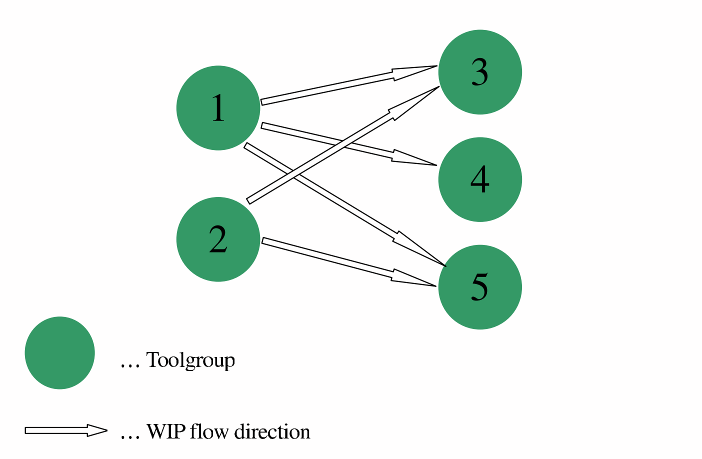


Figure 1: WIP flow direction in wafer fab.

Table 1: WIP Control Table of toolgroup 1.

Toolgroup	Target WIP (lot)	Actual WIP (lot)	WIP Difference (%)	Utilization (%)
3	12	8	-33.3	90
4	16	10	-37.5	70
5	10	14	40	50

- Target WIP: the desired WIP level of toolgroup needed to be maintained. It is based on simulation model, historical data or engineer experience.
- Actual WIP: the current WIP level of toolgroup including lots in queue and in process.
- WIP Difference: the deviation of actual WIP to target WIP.

$$WIPDiff = \frac{ActualWIP - TargetWIP}{TargetWIP}. \tag{1}$$

- Utilization: toolgroup utilization from lot release to current time.
- The Actual WIP, WIP Difference and Utilization will be updated in case of lot move in/out and tool status change of toolgroup 1.

The objectives of WIP Control Table are: (1) Measuring the pull effect of downstream toolgroup; (2) Minimizing the deviation of actual WIP to target WIP of downstream toolgroup. In case of lot move in/out and tool status change, the WIP Control Table will be updated. Therefore, the upstream toolgroup detects the WIP distribution in downstream toolgroups dynamically. Through comparison between actual WIP and target WIP, the upstream toolgroup is able to push lot to a downstream toolgroup which runs out of WIP instead of pushing lot to a downstream toolgroup with congestion.

2.2 Measure Pull Request of Downstream Toolgroup

Each upstream toolgroup can dynamically find out the current WIP situation of all its downstream toolgroups with assistance of WIP Control Table. The WIP difference represents the deviation of actual WIP to target WIP of toolgroup. The higher the WIP difference is, the stronger pull request the downstream toolgroup has. In addition, in order to better distinguish the pull request of toolgroup, toolgroup utilization is also taken into account. High utilized toolgroup implies a high pull request. Two layers to measure the pull request of toolgroup is explained below.

- (1) For each downstream toolgroup, if WIP Difference is less than Delta (Delta is a predefined negative value, which represents the downstream toolgroup is running short of WIP), the downstream toolgroup has a pull request;
- (2) Compare the utilizations of toolgroups which have pull request from (1), the one with higher utilization has stronger pull request. Rank the toolgroups in descending order according to their pull request, e.g. the first rank toolgroup has the strongest pull request.

2.3 Proportion of WIP Required to be Pushed

After determining which downstream toolgroup runs out of WIP and has a strong pull request, the upstream toolgroup has to push lots to it to balance the WIP. Using the WIP Control Table, the upstream toolgroup identifies the proportion of WIP needed to be pushed to the downstream toolgroup.

$$ProportionofWIP = ceiling(\frac{|ActualWIP - TargetWIP|}{NumbersofUpstreamToolgroups}). \qquad (2)$$

where actual WIP and target WIP are described in the WIP Control Table above, numbers of upstream toolgroups is the quantity of upstream toolgroups which the downstream toolgroup has.

In this study, this Proportion of WIP ignores lots availability in the upstream toolgroup. We just assume the upstream toolgroup theoretically has to push average quantity of lots to its downstream toolgroup, no matter whether the upstream toolgroup has lots in queue or not.

2.4 Push Optimal Lot to Downstream Toolgroup

As we mentioned in section 2.1, the WIP Control Table is used to measure the pull effect of toolgroup. In order to optimize other wafer fab's performance indicators such as average cycle time and on time delivery, lot status like due date and local toolgroup constraint, e.g. batch processing and setup change, have to be taken into consideration. Lots are considered as optimal lots and assigned descending priorities according to their characteristics below:

- Priority 1: The lot needs to be accelerated if it is delayed.
- Priority 2: If the time lot experiences in queue exceeds time boundary, the lot needs to be accelerated. This could avoid: (1) the situation that upstream toolgroup processes a lot for a particular high utilized downstream toolgroup as soon as a lot for this particular downstream toolgroup arrives. Other lots for the low utilized downstream toolgroups could not be processed but only waiting remarkable time in queue; (2) high variability of time lot spends in operations.
- Priority 3: Try to make batch as full as possible, which can reduce total setup time and capacity loss.

- Priority 4: If lots do not belong to these 3 categories above, the lot which enters the queue first is favored.

The lots, which queue in the upstream toolgroup and fulfill the demand of downstream toolgroup with pull request, are sorted into different groups in accordance with the rank of downstream toolgroup. For instance, the lots of first group fulfill the demand of first rank downstream toolgroup with strongest pull request. If the first lots group is not empty, the upstream toolgroup pushes proportion of WIP belonged to those 4 categories above to the first rank downstream toolgroup. Otherwise, the upstream toolgroup searches from the next lots group until it is not empty, and pushes optimal lots to the downstream toolgroup of corresponding rank.

3 SIMULATION RESULTS AND PERFORMANCE ANALYSIS

3.1 Simulation Model

The small whole wafer fab dataset MIMAC6 from Measurement and Improvement of MAnufacturing Capacities (MIMAC) (Fowler and Robinson 1995) is used to test the proposed pull/push concept for workload balance of toolgroup. MIMAC6 is a typical complex wafer fab model including:

- 9 products, 9 process flows, maximum 355 process steps.
- 24 wafers in a lot. 2777 lots are released per year under fab loading of 100%. All lots have the same priority of 1 when they are released in the fab.
- 104 toolgroups, 228 tools. 46 single processing toolgroups, 58 batching processing toolgroups.
- Setup avoidance, rework, MTTR (mean time to repair) and MTBF (mean time between failures) of toolgroup.

The simulation experiments are carried out with Factory eXplorer (FX) from WWK. The proposed method is not provided by the FX simulation package, but FX supports customization via a set of user-supplied code and dispatch rules. We used a customized FX interface developed by René Wolf (Wolf 2008) to implement the pull/push concept.

3.2 Simulation Results

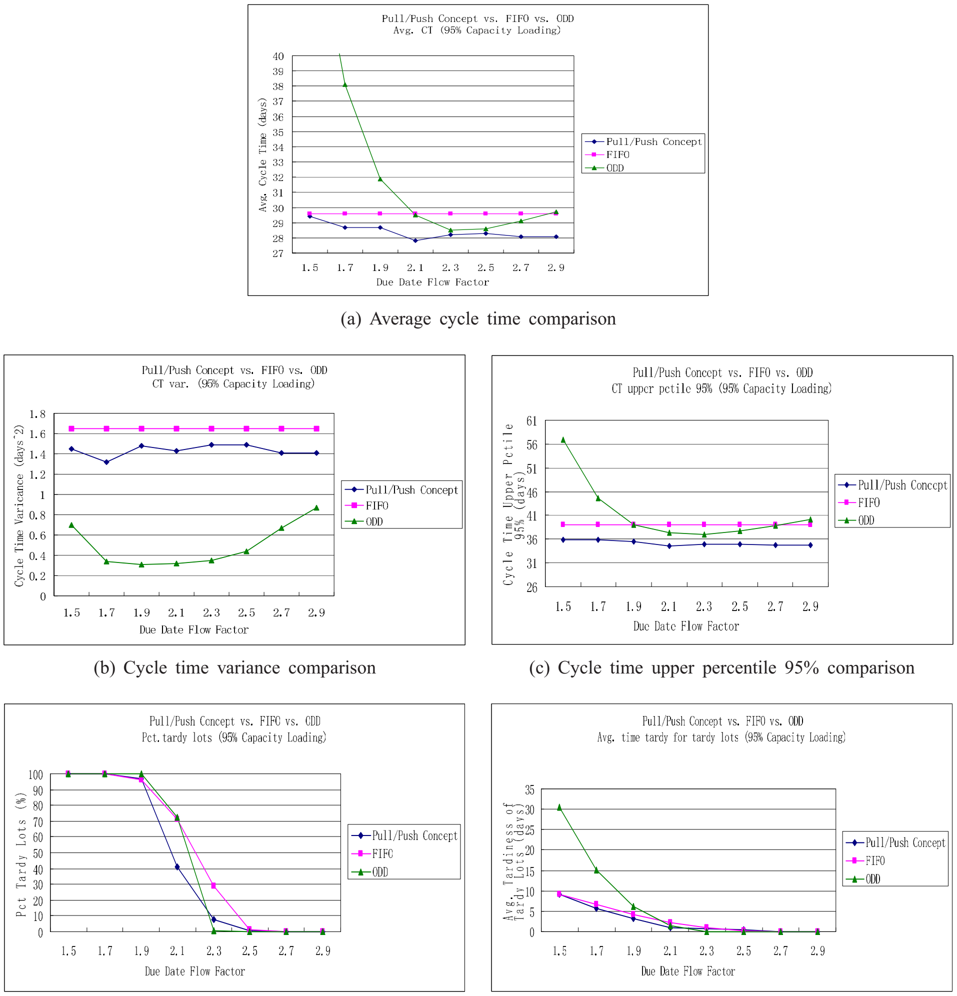
The simulation length of MIMAC6 was carried out for 18 months. The first 6 months were considered as warm-up periods, and not taken into account for statistic. We simulated the fab with 3 dispatching rules: the pull/push concept, FIFO and ODD, under 95% fab capacity loading and with due date flow factor ranging from 1.5 to 2.9 in steps of 0.2.

Here we introduce ODD simply. ODD is a due date oriented dispatching rule like Critical Ratio (CR). It considers a due date for each operation which is defined as the release time plus the sum of raw processing time (RPT) up to this operation times the target due date flow factor (FF), and the lot with the closest operation due date is favored. Rose (Rose 2003) presented a detailed performance evaluation of ODD from tight due date to loose due date, and showed that ODD outperforms CR for tight due date in wafer fab.

$$ODD(i) = ReleaseTime + RPT(i) * FF. \tag{3}$$

Where RPT(i) denotes the RPT for a sequence of operations from operation 1 to operation i (including operation i). FF is defined as the target cycle times divided by the raw processing time.

At first we consider average cycle time, cycle time variance, cycle time upper percentile 95%, on time delivery percentage and average tardiness for tardy lots as major performance measures. Figure 2 shows these 5 performance results. The pull/push concept's average cycle time curve has a similar trend like ODD. The maximum average cycle time exits in tight target due date 1.5, then it becomes smaller and reaches its minimum at due date flow factor 2.1, after that it goes into larger again with loose target due date. The most important thing is no matter with tight due date or loose due date, it outperforms FIFO and ODD, which is a promising result. With respect to cycle time variance, ODD has absolute predominance. Because each lot strictly follows its operation due date and keep the similar pace through the fab. In this study, we mostly consider toolgroup behavior instead of lot status. Therefore, the pull/push concept can not provide cycle time variance as good as ODD, but it is still better than FIFO. When considering cycle time variance, cycle time upper percentile 95% is



(a) Average cycle time comparison

(b) Cycle time variance comparison

(c) Cycle time upper percentile 95% comparison

(d) On time delivery percentage comparison

(e) Average time tardy for tardy lots comparison

Figure 2: 5 performance measures comparison of MIMAC6 with 3 dispatching rules, fab loading of 95% and due date flow factor ranging from 1.5 to 2.9.

an auxiliary reference to see how cycle time distributes. As we can see figure(c), 95% of lots' cycle times of the pull/push concept are smaller than in the case of FIFO and ODD. Concerning average tardiness of tardy lots, the pull/push concept is superior than FIFO and ODD. Moreover, with regard to the on time delivery percentage, the pull/push concept is smoother than FIFO and ODD with the reaction on small due date flow factor change.

Secondly, we focus on simulation results of due date flow factor 2.1 and take a close look at the workload of 3 bottleneck toolgroups which are under high capacity loading. Table 2 lists basic information of these 3 toolgroups and the simulation results are listed in Table 3. Sufficient WIP to achieve high utilization of the bottleneck is important. However, if the WIP exceeds the required level to protect bottleneck from starvation, the cycle time will be increased because lots experience more waiting time in queue. The workload of toolgroup ”20540_CAN_0.43_MII” and ”12553_POSI_GP” can not get improved compared with in the case of FIFO and ODD, even worse. However, the workload of toolgroup ”11026_ASM_B2” is improved considerably. Figure 3 demonstrates the workload shift among ”11026_ASM_B2” and other toolgroups. As we can see, ”11026_ASM_B2” has two upstream toolgroups which are ”12021_ AUTO-CL_undot” and ”12022_AUTO-CL_dot”. These two upstream toolgroups have more than 10 downstream toolgroups, here we only list the major 3 downstream toolgroups which have the most workload shift. The pull/push concept plays an important role in shifting certain WIP from a high WIP toolgroup to a low WIP toolgroup. As long as detecting a high

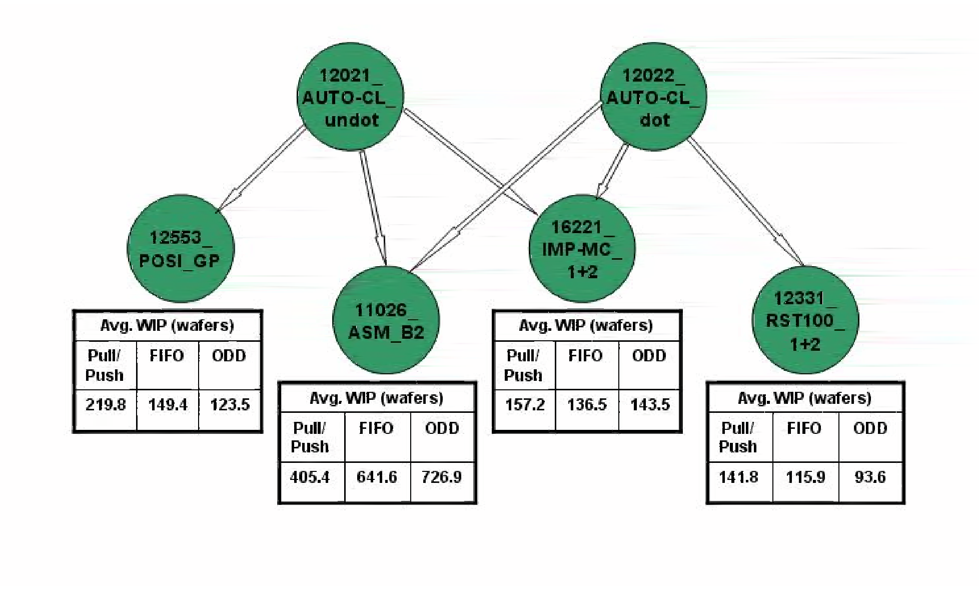
WIP taking place in ”11026_ASM_B2”, the pull/push concept stops pushing lots to it. In contrast, the lots are pushed to other downstream toolgroup, e.g. ”12553_POSI_GP”. This is reason why the average WIP level of ”12553_POSI_GP” is higher than in the case of FIFO and ODD, and each lot averagely spends 12.1 hours in queue which is 4 and 5.6 hours higher than FIFO and ODD cases respectively. However, look at the average queue delay of ”11026_ASM_B2”, for the pull/push concept, each lot has 15.5 hours queue delay on average that is a half of ODD case, which gets improved hugely. Therefore, the pull/push concept successfully balances the workload among different toolgroups.

Table 2: Basic information of 3 high capacity loading toolgroups.

Toolgroup	Processing type	Number of tools	Capacity loading
20540_CAN_0.43_MII	Single, Photo	7	95%
12553_POSI_GP	Batch, Wet Etch	1	93.54%
11026_ASM_B2	Batch, Furnace	1	90.56%

Table 3: WIP and queue delayed comparison of 3 bottleneck toolgroups.

Bottleneck toolgroup		Avg. WIP (wafers)	Avg. Queue Delayed (hours)	Avg. Batch Size (wafers/batch)
20540_CAN_0.43_MII	Pull/Push	338.4	2.5	0
	FIFO	256.5	1.7	0
	ODD	231.8	1.4	0
12553_POSI_GP	Pull/Push	219.8	12.1	98.4
	FIFO	149.4	8.1	98.4
	ODD	123.5	6.5	98.4
11026_ASM_B2	Pull/Push	405.4	15.5	90.8
	FIFO	641.6	27.4	90
	ODD	726.9	31.6	89.4



Avg. WIP (wafers)		
Pull/ Push	FIFO	ODD
219.8	149.4	123.5

Avg. WIP (wafers)		
Pull/ Push	FIFO	ODD
405.4	641.6	726.9

Avg. WIP (wafers)		
Pull/ Push	FIFO	ODD
157.2	136.5	143.5

Avg. WIP (wafers)		
Pull/ Push	FIFO	ODD
141.8	115.9	93.6

Figure 3: WIP shift among different downstream toolgroups.

4 CONCLUSIONS AND FURTHER WORK

In this study we proposed a pull/push concept to balance the workload of toolgroup in wafer fab. This is accomplished by a WIP Control Table of each upstream toolgroup, which contains the current WIP information of all downstream toolgroups. Through comparison between the actual WIP and target WIP of downstream toolgroup, upstream toolgroup is capable of measuring pull request and pushing lot to it dynamically, so as to minimize the deviation of actual WIP to target WIP of downstream toolgroup. In order to optimize other wafer fab's performance indicators, we also considered lot status and local toolgroup constraint to schedule optimal lots to downstream toolgroup. The promising simulation results demonstrated that the pull/push concept can achieve better workload balance compared with FIFO and ODD cases.

For the future research, more datasets have to be tested with the pull/push concept. The accuracy of measuring pull request of toolgroup highly relies on the setting of target WIP. In order to acquire accurate target WIP of each toolgroup, more WIP determination methods have to be investigated, e.g. using neural network and queue model.

5 ACKNOWLEDGEMENTS

This work was supported by Infineon Technologies AG Dresden, Germany and the research project V-Fab 2008-2010, financed by the SAB, Germany. The authors would like thank Frank Lehmann and Sebastian Werner for their fruitful discussion.

REFERENCES

Dabbas, R. M., and J. W. Fowler. 2003. A new scheduling approach using combined dispatching criteria in wafer fabs. *IEEE Transactions on Semiconductor Manufacturing* 16:501–509.

Fowler, J., and J. Robinson. 1995. Measurement and improvement of manufacturing capacities (mimac): Final report. Technical Report 95062861A-TR, SEMATECH, Austin, TX.

Li, S., T. Tang, and D. W. Collins. 1996. Minimum inventory variability schedule with applications in semiconductor fabrication. *IEEE Transactions on Semiconductor MAnufacturing* 9:1–5.

Rose, O. 2003. Comparison of due-date oriented dispatch rules in semiconductor manufacturing. *In Proceedings of the 2003 Industrial Engineering Research Conference*.

Wolf, R. 2008. *Entwicklung einer steuerungsschnittstelle für den simulator factory explorer einschließlich ausführlichem test am beispiel der abfertigungsregel "operation due date (odd)"*. M.S. thesis, Department of Computer Science, Dresden University of Technology, Dresend, Germany.

AUTHOR BIOGRAPHIES

ZHUGEN ZHOU is a PhD student at Dresden University of Technology. He is a member of the scientific staff of Prof. Dr. Oliver Rose at the Chair of Modeling and Simulation. He received his M.S. degree in Computational Engineering from Dresden University of Technology. His research interests include dispatching concepts for complex production facilities and workcenter modeling for wafer fab. His email address is <zhugen.zhou@tu-dresden.de>.

OLIVER ROSE holds the Chair for Modeling and Simulation at the Institute of Applied Computer Science of the Dresden University of Technology, Germany. He received an M.S. degree in applied mathematics and a Ph.D. degree in computer science from Würzburg University, Germany. His research focuses on the operational modeling, analysis and material flow control of complex manufacturing facilities, in particular, semiconductor factories. He is a member of IEEE, INFORMS Simulation Society, ASIM, and GI, and General Chair of WSC 2012. His web address is <www.simulation-dresden.com>.