Review of models for large scale manufacturing networks

Kai FURMANS 1
1 Institute for Material Handling and Logistics, Karlsruhe Institute of Technology, Karlsruhe, Germany, kai.furmans@kit.edu
George LIBEROPOULOS 2
2 Department of Mechanical Engineering, University of Thessaly, Volos, Greece, glib@mie.uth.gr
Marion RIMMELE 1
marion.rimmele@kit.edu
Olaf ZIMMERMANN 1
olaf.zimmermann@kit.edu

Productive4.0 is a large-scale research project, which aims at using Industry 4.0 ideas and technologies to improve the design, manufacturing and distribution of high-tech products. Within this project, two work-packages provide models, which shall be used to design, optimize and control large supply chains, which consist of manufacturing and distribution elements. In our contribution, we give an overview about the current state of the usage and usability of stochastic, especially queueing models for the above mentioned purpose.

Key words: Productive4.0; Control and Optimization; Queueing Models

1. Challenges of Industry 4.0 for the modelling of large scale manufacturing networks

Industry 4.0 describes a set of research programs, which try to evaluate, explain and understand the impact of digitization on the industry and then succeedingly try to shape and develop new processes, products and services for the industry. The program has been originated in Germany but the ideas have spread out to other regions and countries.

Cyber-physical Production Systems (CPPS) are an important part of future Smart Factories (see Kagermann et al. (2012)). CPPS are a special case of Cyber-physical systems. The term Cyber-Physical System has been created by the NSF in 2006, describing systems, which operate in the physical world, in real-time, in a continuous state-space. They are controlled by a digital representation (sometimes called digital twin) which is used to control the CPS and it's interaction with the surrounding world. CPPS are a representation of CPS in a production environment.

Control of a system requires the specification of a to-be state, the necessary measurements to check the current state of the system, if the is-state deviates from the to-be state, the generation of the necessary control actions, which bring the system back in the desired state. These control actions have to be executed by the actors of the system.

This creates the following applications models of large scale manufacturing networks:

Design a system, that fulfils the requirements of the manufacturing system, this includes the section of machinery and tools, the mapping of products to machinery, the positioning of buffers and the selection of planning and control algorithms.

Optimize the parameters, like number of parallel machinery, the size of buffers, the control parameters.

Control and planning based on the current state of the system, check whether a control action is required, devise the control action and execute it. This also includes the planning actions, which makes the decision about sequence and quantity of products to be produced at every stage.

Using the same or a very closely related model in all three cases would reduce the modelling work and lead to more consistent decision making. The usage of a model for control purposes requires the possibility to quickly generate control actions, which would bring the system back to its desired state. The more frequently this can be done, the better the properties of the control. Therefore we are looking for models and model-types , which provide the capability of a repeated, fast evaluation of possible control actions in a trial-and-error fashion or - even better - the direct deviation of control measures from the inversion of known functions, which connect the performance measures with the control actions.

In Productive4.0 we want to evaluate and demonstrate, that the idea of a CPS in a large scale manufacturing environment is feasible. A major issue in production systems nowadays is the capability to not only guarantee average cycle time - resulting in an average quoted lead time to the customer, but to guarantee a lead time with a high probability, which allows the customer to reliably calculate his supply lead times and inventory levels. Therefore, an emphasis will be put on the confidence level of the computed lead times of the production network. The control models will therefore be evaluated on their capability, to generate stable, reliable lead times.

Figure 1 shows the setup of the part of the project concerning the control of the production system. The basis is the value stream of the real, physical production system. For this system, real data will be collected, validated, evaluated for outliers and made available for the models. Additionally data of customer demand and supplier data will be added to the projects' data pool.

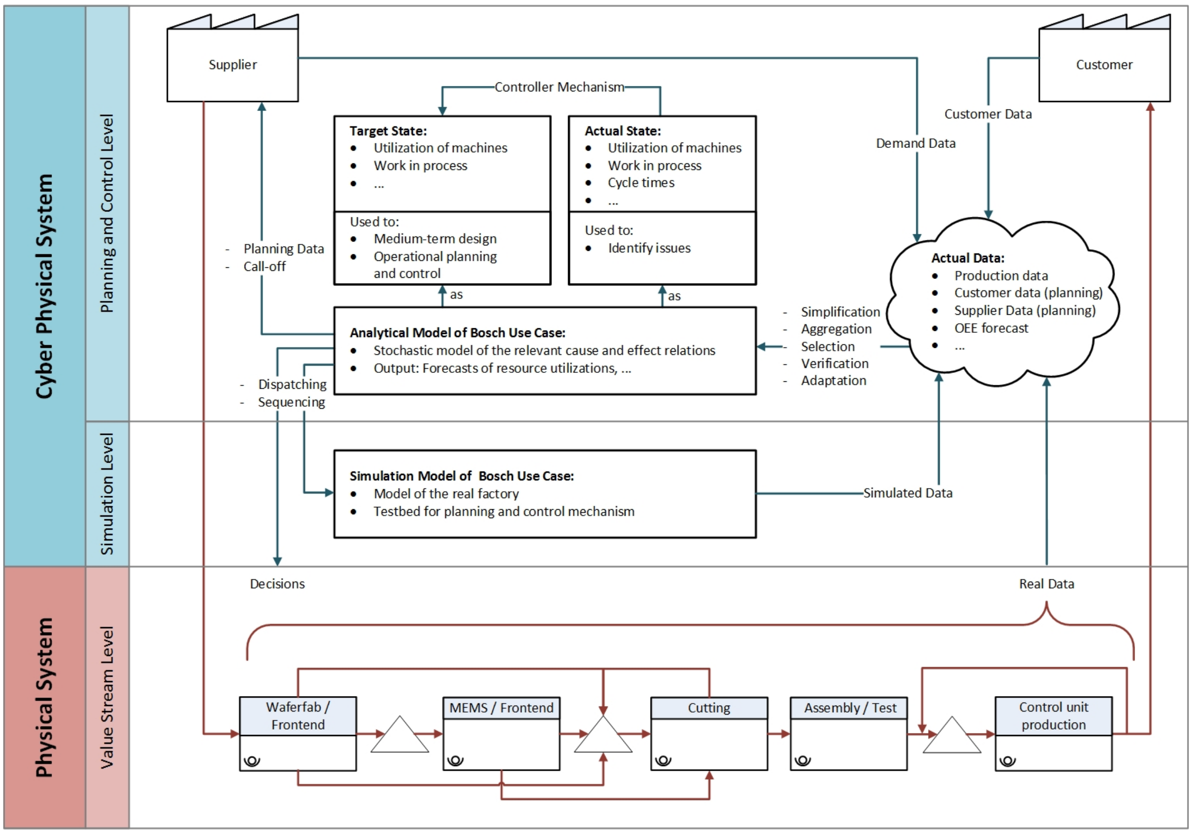

Figure 1. Use case of Productive4.0: Virtual production planning and control of a semiconductor supply chain

Based on this data and knowledge about the value stream, a simulation model will be built, which will act as a testbed for the planning and control algorithms. This is required, since different models and decision making algorithms have to be evaluated under comparable conditions. The testbed itself has to be validated against real data, in order to make sure, that it is representing the real production system accurately enough for the subsequent evaluation of the planning and controlling algorithms.

The models, which are used for decision making, are supplied with the required verified, aggregated, simplified and suitably adapted data from the real system. Then, selected status information and performance information will be captured from the simulation testbed (as it would later be from the operation) and used for decision making in planning and control, based on input from the stochastic model. These decisions will then be fed back to the testbed. Thus, the simulation system will use input from another model influencing the model at run-time. This way, it will be evaluated, which decision making and controlling methods are influencing the production system in a favourable way.

At least potentially, queueing models are able to provide the capability of a fast evaluation of system performance parameters under a certain setting. However, there were shortcomings in the level of detail, which such models could provide. On the other hand, the larger a system the more statistics of large numbers might possibly apply. Here, in this paper, we want to review the state of the art of using queueing models for the design, optimization, planning and control of manufacturing systems.

2. Literature review

In this section we discuss models used for performance evaluation, production planning and control of large scale manufacturing networks. First, we give a literature overview of models for performance evaluation in manufacturing systems. Then different reviews are presented in the area of planning and control of large scale manufacturing networks and selected models are briefly described. Table 1 lists the discussed models of manufacturing systems and their characteristics reported in the literature.

2.1. Performance evaluation of a large scale manufacturing networks

In recent years, researchers have focused on providing solutions for performance evaluation of large scale manufacturing systems, such as semiconductor factories. In this context, cycle time estimation is an essential step during the planing of those systems, especially for the analyses of performance indicators, capacity planning, and the assignments of due dates (Pearn et al. (2009)). Methods for cycle time estimation can be classified into the following four categories: analytical, simulation, statistical analysis and hybrid methods (Chung and Huang (2002)).

Analytical methods are primarily based on queuing models. Shanthikumar et al. (2007) presented a survey regarding queuing theory for semiconductor manufacturing systems (SMS). They pointed out that practical applications of queuing theory in SMS are rather limited due to unsatisfactory results. Morrison and Martin (2007) proposed extensions to the popular and intuitive closed-form approximations for the mean cycle time in G/G/m-queues of Whitt (1993) considering tool failure, tools with parallelism and idle times despite available work. The G/G/m cycle time of failure prone tools was considered by Kim and Morrison (2011), which model the system as an absorbing Markov chain. Also based on a Markov chain model, Ding et al. (2007) analysed a queuing system with batching service and heterogeneous tools using the example of a lithography tool. They evaluated the impact of tool up and down time on the cycle time of the system. Akhavan-Tabatabaei et al. (2012) developed a Markov chain framework that attempts to approximate the cycle time of a tool set in the presence of informal operational rules which creates dependence between the arrival and service processes.

A decomposition approach is an appropriate method for large scale manufacturing networks. The classical decomposition method called queuing network analyser (QNA) was developed by Whitt (1983). A decomposition without aggregation (DWOA) method derived Grosbard et al. (2013) to model a Semiconductor Manufacturing Queuing Network (SMQN) and yield cycle time approximations. Sagron et al. (2015) presented modifications for decomposition approximation methods, especially to account for the heavy-traffic bottleneck phenomenon in open-queuing networks with deterministic routing and downtime induced variability.

Simulation is a common approach in cycle time estimation and performance analysis of SMS (Shanthikumar et al. (2007)). Scholl and Domaschke (2000) used a discrete event simulation to investigate

Reference	Target	Model		Considered Characteristics								Result
		Type	Time	Re-entrance	Blocking	Arrival Process	Service Process	Multi-Server	Service	Batches	Downtimes	
Morrison and Martin (2007) Kim and Morrison (2011)	P	Queuing station	AC			G	G	x	FCFS		x	Cycle Time
Ding et al. (2007)	P	Markov chain model	AC			M	M	x	GEO	x	x	Cycle Time
Akhavan-Tabatabaei et al. (2012)	P	State-dependent Markov chain model	AC			G	G	x	FCFS		x	Work-in-Process Cycle Time
Grosbard et al. (2013)	P	Open queuing network	AC	x		G	G	x	FCFS	x	x	Cycle Time Waiting Time
Sagron et al. (2015)	P	Open queuing network	AC			G	G		FCFS TP		x	Waiting Time
Scholl and Domaschke (2000)	P	Discrete event simulation	S	x	x	G	G	x	ESD		x	Cycle Time
Sivakumar and Chong (2001)	P	Discrete event simulation	S	x	x	G	G	x	ESD		x	Cycle Time
Pearn et al. (2009)	P	Fitted Gamma distribution	AC			-	-		-			Waiting Time Cycle Time
Tai et al. (2012)	P	Fitted Weibull distribution	AC			-	-		-			Waiting Time Cycle Time
Akhavan-Tabatabaei et al. (2009)	P	Flow Analysis Model	AC			-	-		-		x	Work-in-Process Cycle Time
Veeger et al. (2010a, 2010b)	P	EPT-based aggregate simulation	AC			G	G	x	FCFS		x	Cycle Time
Veeger et al. (2011)	P	EPT-based aggregate simulation	S			G	G		FCFS LCFS		x	Cycle Time
Can and Heavey (2016)	P	EPT-based discrete event simulation	S		x	M	M	x	FCFS		x	Cycle Time
Yang et al. (2011)	P	Simulation-based metamodelling	S			-	-	x	TP			Cycle Time
Yang (2010)	P	Neural network metamodelling	S			-	-	x	TP			Cycle Time
Hsieh et al. (2014)	P	Progressive simulation metamodelling	S			-	-		TP	x		Cycle Time
Schelasin (2011, 2013)	P	Static capacity model Queuing station	AC			G	G	x	FCFS		x	Waiting Time Cycle Time
Zisgen and Meents (2008) Brown et al. (2010)	P	Open queuing network	AC	x		G	G	x	FCFS	x	x	Cycle Time Queue Lengths
Zarifoglu et al. (2013)	C	Queuing Model	AC			G	G		FCFS	x		Optimal Lot Size
Chang (2016)	C	Simulation	S	x		G	G	x	TP	x		Optimal Product Mix
Wang and Wang (2007)	C	Simulation	S	x		G	P	x	FCFS	x		Optimal Lot Size
Akhavan-Tabatabaei et al. (2011)	C	Simulation	S	x		M	M		FCFS		x	Optimal Lot Release
Fowler et al. (2002)	C	Queuing Station	AC			G	G	x	FCFS	x		Optimal Batch Size
Kuo et al. (2011)	C	Neural Network	AC			G	G	x	FCFS	x	x	Reduce Cycle Time
Li et al. (2010)	C	Queuing Network	AC		x	M	M		FCFS		x	Optimal WIP level
Wang et al. (2013)	C	Neural Network	AC			G	G		OPT			Optimize Service Discipline
Chien et al. (2012)	C	Neural Network	AC			-	-		-			Reduce Cycle Time
Yao et al. (2004)	C	Modelling Framework	AC			-	-		-		x	Maintenance Scheduling
Ramirez-Hernandez et al. (2010)	C	Simulation	AC			-	-		-		x	Maintenance Scheduling
Kalir (2013)	C	Queuing Station	AC			G	G	x	FCFS		x	Optimize FM Splitting
Morrison (2014)	C	Queuing Station	AC			G	G	x	FCFS		x	Maintenance Scheduling

Table 1: Characteristics of modelling a large scale manufacturing networks
(Target: P - Performance Evaluation, C - Control; Time: AC - Analytical Continuous, S - Simulation)

the actual situation in the factory of Infineon Technologies in Dresden, and to identify recommendations to eliminate or to reduce the impact of time constraints. The relationship between selected input and output variables in semiconductor backend manufacturing systems was analysed by Sivakumar and Chong (2001), using a data driven discrete event simulation model.

Regression analysis is a typical statistical method to determine the relationship between the cycle time and other related parameters (Chung and Huang (2002)). Pearn et al. (2009) used a Gamma distribution to model waiting times for single operations of each product type in semiconductor plastic ball grid array (PBGA) packaging factories. They present a combined cycle time estimation model which incorporates the reproductive property of the Gamma distribution to estimate the whole factory cycle time. A Weibull-distributed waiting time for a single-layer testing operation was used by Pearn et al. (2009) to estimate the cycle time distribution in the re-entrant semiconductor final testing process flow. Akhavan-Tabatabaei et al. (2009) incorporated the existing correlations in arrival and service processes by observing the flow of lots through the tool set. This correlation is then used in a flow analysis model to generate more accurate forecasts of work-in-process levels which results in more accurate cycle time estimations.
There are an increasing number of hybrid methods, which combine analytical, simulation and statistical analysis methods to generate cycle time estimation. Schelasin (2011) used a backward calculation algorithm to calculate factory variability based on historical data. The author used the variability to predict the cycle time by the Kingman equation and later extended this approach by an improved G/G/m formula (Schelasin (2013)). In aggregate simulation models an effective process time (EPT) distribution is often used to determine the cycle time of the manufacturing system. The EPT distribution is estimated based on lot arrival and lot departure data that is measured at the workstation in operation (Hopp and Spearman (2011)). Outages of the machines are thus taken into account. Veeger et al. (2010a) developed an aggregate simulation model to predict the mean cycle times of a workstation for different throughputs and product mixes (CT-TH-PM) using the EPT distribution to aggregate the various workstation details. They used a curve fitting procedure to deal with the typically limited number of arrival and departures encountered in semiconductor manufacturing practice (Veeger et al. (2010b)). To model a lithography workstation Veeger et al. (2011) extended their approach and took into account the order in which lots are processed. Genetic programming (GP) and EPT distributions were combined by Can and Heavey (2016) to predict cycle time using a discrete event simulation model of a production line. A machine learning approach exploits the positional and granular information from the simulation model used to imitate a production system. Yang et al. (2011) integrated queuing theory and simulation-based metamodelling. The queuing analysis leads to the division of the feasible region into a number of subregions. This allows for the fitting of a smooth CTTHPM surface within each subregion. The author developed also an alternative fitting approach, which is based on neural networks (Yang et al. (2011)). Hsieh et al. (2014) derived a progressive simulation metamodel that can characterise the relationship between the response variable and the input variables of interest. Based on an experimental design they calculated the mean cycle time depending on the percentage of hot lot.
A integrated system which combines the approaches of queuing network modelling, simulation and statistical analysis was developed by IBM (Brown et al. (2010)). The Enterprise Production Planning and Optimization System (EPOS) is a queuing network based simulation system for tactical and operational production planning and production management which can be integrated with the fab MES to capture routes, tools, raw process times, rework rates, and WIP (Zisgen et al. (2008)). Tools are modelled as $G^X/G(b, b)/c$ server queues. The manufacturing system is considered as an open queuing network. A decomposition approach is used to determine the queue lengths for each equipment in the network.

2.2. Production planning and control of a large scale manufacturing network

A second topic that has been researched in large scale manufacturing networks is the production planning and control of existing networks under different aspects. We want to give an overview over some of the works which have been published in this field. We start by discussing literature reviews which have been published on this topic.
The literature on production planning and control in the manufacturing systems dates back to the mid

1980s. It is extensive and includes several reviews and a handful of monographs. The earliest comprehensive review is the two-part paper by Uzsoy et al.. In the first part, Uzsoy et al. (1992) looked at long-term production planning approaches. In the second part of the paper, Uzsoy et al. (1994) reviewed research on shop-floor control in semiconductor manufacturing facilities, focusing on dispatching rules and input regulation strategies, deterministic scheduling algorithms, control-theoretic approaches, and knowledge-based approaches. They also discussed the relationship between shop-floor control and production planning. Gupta et al. (2006) provided an updated glimpse of the developments in planning and control issues in semiconductor wafer production. Mönch et al. (2011) reported on the breadth of scheduling problems in the semiconductor manufacturing industry. They presented scheduling problems for entire job shops and reviewed the current state of practice, where they referred to the large number of dispatching systems that were in place in wafer fabs. Besides the aforementioned references, there have also been few other literature reviews on specific operational problems in semiconductor manufacturing. Examples are the review of Mathirajan and Sivakumar (2006) on the scheduling of batch processors and the review of Geng and Jiang (2009) on strategic capacity planning. Mönch et al. (2013) published a monograph on production planning and control for semiconductor wafer fabrication facilities. The authors introduced production control schemes that are based on dispatching rules as they are predominately used in practice. They described various aspects of decision support provided by manufacturing execution systems and advanced planning systems.

All of the published papers in this area try to achieve similar goals in either decreasing the cycle time of the manufacturing network or increasing the throughput of it. The methods used to achieve the goals can be differentiated in several areas. They have in common that they take a specific control procedure of the manufacturing network and try to optimize it. The procedures include dispatching rules, lot size, product mix, lot release, batch size, work-in-progress level and maintenance. The range of optimization goes from decisions which only influence single machines or process steps up to decisions which influence the whole network.

One approach to improve cycle time is the development of new dispatching rules. Here for example Wang et al. (2013) propose a new dispatching rule. They state that the methods they used to improve their new rule could also be applied to existing rules to achieve an improvement of their performance. Wang and Wang (2007) use a simulation model to show that the cycle time can be reduced by reducing the lot size if the available bottleneck capacity is high enough to handle smaller lots. They also show that this only works until a critical lot size after which the cycle time increases again. The work of Zarifoglu et al. (2013) is based on the work of Wang and Wang. They however come to the conclusion that smaller lot sizes are not always the better option and the optimal lot size has to be determined for every manufacturing network based on the manufacturing technology available. They also state that extended research has to be done on more complex networks to validate their results.

Chang (2016) develops a simulation based optimization model including a new solution method to solve the problem of obtaining the optimal product mix. The new quantile-based method AGLS-QC can also control the upside risk that comes with selecting the product mix.

Akhavan-Tabatabaei and Salazar (2011) apply a WIP dependent lot release strategy based on the WIP of the static bottleneck and compare it to a CONWIP lot release strategy. They show that for some cases the new strategy is better than CONWIP but generally CONWIP gives better results, especially when not only cycle time but also throughput is taken into consideration. They still recommend their new strategy because they state that a CONWIP strategy is not applicable in practice. Fowler et al. (2002) develop a $G/G^{(b_p)}/C$-Queue and then show the use of the queue to solve the problem of optimal batch building in a tool set. This queue can then be used to minimize the cycle time of a production network when combined with a Genetic Algorithm to reduce execution time.

Li et al. (2010) optimize the Multi-CONWIP level of a manufacturing system. For this the system was modelled as a closed queueing network with several loops. They evaluated the model with a Genetic Algorithm and a cost-oriented optimization problem and achieved satisfying results.

Another topic which has been researched intensively is preventive maintenance in a large production

network. Yao et al. (2004) developed a two-level hierarchical modelling framework for planning and scheduling maintenance. They also focused on optimizing the scheduling process of maintenance and reached an increase in throughput with their model. Ramírez-Hernández et al. (2010) used this model among others to create a software tool for planning and optimizing preventive maintenance. Kalir (2013) discussed the question if a planned preventive maintenance can be split in several parts. He achieved a further improvement of cycle time with his method even though the sum of maintenance times of all parts is larger than the original maintenance time. Morrison et al. (2014) extended this model again to not only look at a single preventive maintenance but to consider multiple maintenance cycles on the same tool set.
Finally, another way of optimizing large scale manufacturing networks which we focus on in this review is the use of Neural Networks. While they are not yet used to improve single aspects of controlling manufacturing networks, it has been shown that the results of using Neural Networks can lead to identifying factors which are important for reducing the cycle time. Kuo et al. (2011) for example use a Neural Network to predict WIP levels and have shown that the implementation of their models has led to a significant reduction in cycle time in the evaluated fab by deriving action plans from the results of the models. Chien et al. (2012) use their Neural Network model in a similar way. But they also include an adaptive model which changes the model parameters based on the relative forecast error when input parameters are changed.

3. Direction for future work

The usage of simulation and analytical models for the purpose of performance evaluation and especially cycle time estimation has been documented and tried several times. However, the focus was usually set on an average cycle time, not on a percentile of the cycle time distribution. This leads to a gap to be filled. Dispatching and sequencing rules have also been studied, however, the impact on guaranteed performance figures is still open. With neural networks and statistical regression methods new approaches have been tried, which seem to be useful, when some data points for the system are available and the gaps between these data points have to be filled. This is an opportunity for the combination of data points generated from time consuming methods like simulation with real time decision making. Hybrid methods make use of the higher computing power available today, using combined knowledge from several sources.

However, the integration of data collection, planning algorithms and performance perediction of guaranteed cycle times with a high probability is an open task.

Acknowledgments
This research is expected to be supported by the research project ‘Productive4.0 - Electronics and ICT as enabler for digital industry and optimized supply chain management covering the entire product lifecycle’; a conditionally approved project by the EU and BMBF (Call: H2020-ECSEL-2016-2-IA-two-stage).

References

Akhavan-Tabatabaei, R., S. Ding, J. G. Shanthikumar. 2009. A method for cycle time estimation of semiconductor manufacturing toolsets with correlations. *Proceedings Winter Simulation Conference 2009*. 1719–1729.

Akhavan-Tabatabaei, R., Y. Fathi, J. G. Shanthikumar. 2012. A markov chain framework for cycle time approximation of toolsets. *IEEE Transactions on Semiconductor Manufacturing* **25**(4) 589–597.

Akhavan-Tabatabaei, R., C. Salazar. 2011. Effective wip dependent lot release policies: a discrete event simulation approach. *Proceedings Winter Simulation Conference 2011*. 1976–1985.

Brown, S. M., T. Hanschke, I. Meents, B. R. Wheeler, H. Zisgen. 2010. Queueing model improves ibm's semiconductor capacity and lead-time management. *Interfaces* **40**(5) 397–407.

Can, B., C. Heavey. 2016. A demonstration of machine learning for explicit functions for cycle time prediction using mes data. *Proceedings Winter Simulation Conference 2016*. 2500–2511.

Chang, K. 2016. Risk-controlled product mix planning in semiconductor manufacturing using simulation optimization. *IEEE Transactions on Semiconductor Manufacturing* **29**(4) 411–418.

Chien, C., C. Hsu, C. Hsiao. 2012. Manufacturing intelligence to forecast and reduce semiconductor cycle time. *Journal of Intelligent Manufacturing* **23**(6) 2281–2294.

Chung, S., H. Huang. 2002. Cycle time estimation for wafer fab with engineering lots. *IIE Transactions* **34** 105–118.

Ding, S., R. Akhavan-Tabatabaei, J. G. Shanthikumar. 2007. Stochastic modeling for serial-batching workstations with heterogeneous machines. *Automation Science and Engineering, 2007. CASE 2007. IEEE International Conference on*. IEEE, 77–81.

Fowler, J. W., N. Phojanamongkolkij, J. K. Cochran, D. C. Montgomery. 2002. Optimal batching in a wafer fabrication facility using a multiproduct g/g/c model with batch processing. *International Journal of Production Research* **40**(2) 275–292.

Geng, N., Z. Jiang. 2009. A review on strategic capacity planning for the semiconductor manufacturing industry. *International journal of production research* **47**(13) 3639–3655.

Grosbard, D., A. Kalir, I. Tirkel, G. Rabinowitz. 2013. A queuing network model for wafer fabrication using decomposition without aggregation. *Automation Science and Engineering (CASE), 2013 IEEE International Conference on*. 717–722.

Gupta, J. N. D., R. Ruiz, J. W. Fowler, S. J. Mason. 2006. Operational planning and control of semiconductor wafer production. *Production Planning & Control* **17**(7) 639–647.

Hopp, W. J., M. L. Spearman. 2011. *Factory physics*. Waveland Press.

Hsieh, L., K. Chang, C. Chien. 2014. Efficient development of cycle time response surfaces using progressive simulation metamodeling. *International Journal of Production Research* **52**(10) 3097–3109.

Kagermann, H., W. Wahlster, J. Helbig. 2012. Umsetzungsempfehlungen fr das zukunftsprojekt industrie 4.0. Report, Forschungsunion.

Kalir, A. A. 2013. Segregating preventive maintenance work for cycle time optimization. *IEEE Transactions on Semiconductor Manufacturing* **26**(1) 125–131.

Kim, W., J. R. Morrison. 2011. On cycle time approximations for the failure prone g/g/m queue: Theoretical justification of a practical approximation. *Control, Automation and Systems (ICCAS), 2011 11th International Conference on*. 1558–1563.

Kuo, C., C. Chien, J. Chen. 2011. Manufacturing intelligence to exploit the value of production and tool data to reduce cycle time. *IEEE Transactions on Automation Science and Engineering* **8**(1) 103–111.

Li, N., S. Yao, G. Liu, C. Zhuang. 2010. Optimization of a multi-constant work-in-process semiconductor assembly and test factory based on performance evaluation. *Computers & industrial engineering* **59**(2) 314–322.

Mathirajan, M., A. I. Sivakumar. 2006. A literature review, classification and simple meta-analysis on scheduling of batch processors in semiconductor. *The International Journal of Advanced Manufacturing Technology* **29**(9-10) 990–1001.

Mönch, L., J. W. Fowler, S. Dauzère-Pérès, S. J. Mason, O. Rose. 2011. A survey of problems, solution techniques, and future challenges in scheduling semiconductor manufacturing operations. *Journal of Scheduling* **14**(6) 583–599.

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

Morrison, J. R., H. Kim, A. A. Kalir. 2014. Mean cycle time optimization in semiconductor tool sets via pm planning with different cycles: Ag/g/m queueing and nonlinear programming approach. *Proceedings Winter Simulation Conference 2014*. 2466–2477.

Morrison, J. R., D. P. Martin. 2007. Practical extensions to cycle time approximations for the $g/g/m$-queue with applications. *IEEE Transactions on Automation Science and Engineering* **4**(4) 523–532.

Pearn, W. L., Y. Tai, J. H. Lee. 2009. Statistical approach for cycle time estimation in semiconductor packaging factories. *IEEE transactions on electronics packaging manufacturing* **32**(3) 198–205.

Ramírez-Hernández, J. A., J. Crabtree, X. Yao, E. Fernandez, M. C. Fu, M. Janakiram, S. I. Marcus, M. O'Connor, N. Patel. 2010. Optimal preventive maintenance scheduling in semiconductor manufacturing systems: Software tool and simulation case studies. *IEEE Transactions on Semiconductor Manufacturing* **23**(3) 477–489.

Sagron, R., D. Grosbard, G. Rabinowitz, I. Tirkel. 2015. Approximation of single-class queueing networks with downtime-induced traffic variability. *International Journal of Production Research* **53**(13) 3871–3887.

Schelasin, R. 2011. Using static capacity modeling and queuing theory equations to predict factory cycle time performance in semiconductor manufacturing. *Proceedings Winter Simulation Conference 2011*. 2045–2054.

Schelasin, R. 2013. Estimating wafer processing cycle time using an improved g/g/m queue. *Proceedings Winter Simulation Conference 2013*. 3789–3795.

Scholl, W., J. Domaschke. 2000. Implementation of modeling and simulation in semiconductor wafer fabrication with time constraints between wet etch and furnace operations. *IEEE Transactions on Semiconductor Manufacturing* **13**(3) 273–277.

Shanthikumar, J. G., S. Ding, M. T. Zhang. 2007. Queueing theory for semiconductor manufacturing systems: A survey and open problems. *IEEE Transactions on Automation Science and Engineering* **4**(4) 513–522.

Sivakumar, A. ., C. S. Chong. 2001. A simulation based analysis of cycle time distribution, and throughput in semiconductor backend manufacturing. *Computers in Industry* **45**(1) 59–78.

Uzsoy, R., C. Lee, L. A. Martin-Vega. 1992. A review of production planning and scheduling models in the semiconductor industry part i: system characteristics, performance evaluation and production planning. *IIE transactions* **24**(4) 47–60.

Uzsoy, R., C. Lee, 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**(5) 44–55.

Veeger, C. P. L., L. F. P. Etman, E. Lefeber, I. J. B. F. Adan, J. van Herk, J. E. Rooda. 2011. Predicting cycle time distributions for integrated processing workstations: an aggregate modeling approach. *IEEE Transactions on Semiconductor Manufacturing* **24**(2) 223–236.

Veeger, C. P. L., L. F. P. Etman, J. van Herk, J. E. Rooda. 2010a. Generating ct-th-pm surfaces using ept-based aggregate modelling. *Journal of Simulation* **4**(4) 242–254.

Veeger, C. P. L., L. F. P. Etman, J. van Herk, J. E. Rooda. 2010b. Generating cycle time-throughput curves using effective process time based aggregate modeling. *IEEE Transactions on Semiconductor Manufacturing* **23**(4) 517–526.

Wang, C., C. Wang. 2007. A simulated model for cycle time reduction by acquiring optimal lot size in semiconductor manufacturing. *The International Journal of Advanced Manufacturing Technology* **34**(9-10) 1008–1015.

Wang, Y., C. Wu, T. Chen. 2013. A fuzzy-neural approach for optimizing the performance of job dispatching in a wafer fabrication factory. *The International Journal of Advanced Manufacturing Technology* **67**(1-4) 189–202.

Whitt, W. 1983. The queueing network analyzer. *Bell Labs Technical Journal* **62**(9) 2779–2815.

Whitt, W. 1993. Approximations for the gi/g/m queue. *Production and Operations Management* **2**(2) 114–161.

Yang, F., J. Liu, B. L. Nelson, B. E. Ankenman, M. Tongarlak. 2011. Metamodelling for cycle time-throughput-product mix surfaces using progressive model fitting. *Production Planning and Control* **22**(1) 50–68.

Yao, X., E. Fernández-Gaucherand, M. C. Fu, S. I. Marcus. 2004. Optimal preventive maintenance scheduling in semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing* **17**(3) 345–356.

Zarifoglu, E., J. J. Hasenbein, E. Kutanoglu. 2013. Lot size management in the semiconductor industry: Queueing analysis for cycle time optimization. *IEEE Transactions on Semiconductor Manufacturing* **26**(1) 92–99.

Zisgen, H., I. Meents, B. R. Wheeler, T. Hanschke. 2008. A queueing network based system to model capacity and cycle time for semiconductor fabrication. *Proceedings Winter Simulation Conference 2008*. 2067–2074.