Updated State-of-the-Art analysis

Project acronym	**AISSI** - 20212
Project title	**A**utonomous **I**ntegrated **S**cheduling in **S**emiconductor **I**ndustry
Programme	ITEA – Call AI 2020
Start date of the project	01-06-2021
Duration	36 months

Deliverable reference number	D1.1
Deliverable title	Updated State-of-the-Art analysis
WP contributing to the deliverable	WP1
Due date	31.10.2021
Actual submission date	21.10.2021

Responsible organisation	Karlsruhe Institute of Technology (KIT)
Authors	Christoph Jacobi, Benedikt Schulz
Peer reviewer	Karlheinz Leonardi, Tobias Sprogies, Holger Brandl, Boon Ping Gan
Abstract	This deliverable provides an overview of both the scientific as well as the industrial state-of-the-art of production scheduling and maintenance planning in the semiconductor industry.
Keywords	Cluster tool scheduling, job shop scheduling, maintenance planning, literature review, questionnaire

Dissemination level	Public
Revision	Version 1.0

ITEA 3
GEFÖRDERT VOM Bundesministerium für Bildung und Forschung
With the support of Enterprise Singapore

TABLE OF CONTENTS

1 VERSION CONTROL

Date	Author	Version	Notes
03.09.2021	Christoph Jacobi, Benedikt Schulz	0.1	First version by KIT
11.10.2021	Benedikt Schulz, Christoph Jacobi	0.2	Integration of reviewer comments
21.10.2021	Kai Schelthoff	1.0	Final version by Kai Schelthoff

2 INTRODUCTION

Semiconductor manufacturing processes are among the world's most complex industrial processes[1]. There are approximately 300–400 process steps in a wafer production with up to 80 different types of machines[2]. The re-entry properties and the diversity of equipment suppliers complicate the processes even more. This is the case for the semiconductor industry in particular, which focuses on a high product mix to supply the market with customised and specialised products used, e.g. in the automotive and home appliance industry. Additionally, the cost of a single machine, such as lithography tools, may exceed US$20 million, higher than most other industries[3]. Raising manufacturing efficiency via synchronising the scheduling of production lot flows and maintenance is therefore a critical and demanding task to maximise the return on investments. Moreover, it has been demonstrated that a good maintenance policy reduces the production loss of machines as unexpected machine failures can be reduced[4,5,6]. Conventionally, machine condition-based job assignment or scheduling mainly relies on domain experience and knowledge of experienced employees. In the current state, existing theoretical approaches cannot easily replace experience and human knowledge required for holistic scheduling and planning of production and maintenance activities. However, the evolution towards Industry 4.0 standards with easy access to terabytes of data made available through IoT opens new perspectives for training AI approaches. Our aim in this deliverable is to provide the current state-of-the-art of job shop scheduling and maintenance planning in the semiconductor industry, taking both scientific literature and industrial practice of the industry partners in the consortium into account.

2.1 Overview and scope of the deliverable

Chapter 3 gives an overview of the academic state-of-the-art. We first describe the methodology of our literature review and present the results regarding production scheduling, maintenance planning and approaches that integrate these two aspects.
In Chapter 4, we present the industrial state-of-the-art based on a survey among the project partners from industry. First, we describe the questionnaire used for the survey and second, we present the results.

2.2 Achievements compared to the project objectives

All objectives concerning the scientific state of the art analysis have been achieved with this deliverable. Furthermore, this document provides an overview of the industrial state of the art of the project's use case providers.

[1] Mönch, L., Fowler, J. W., Mason, S. J. (2013). Production Planning and Control for Semiconductor Wafer Fabrication Facilities. Springer, New York.
[2] Chung, S.-H., Huang, H.-W. (2002). Cycle time estimation for wafer fab with engineering lots. IIE Transactions 34 (105-118).
[3] Johnzén, C., Dauzère-Pérès, S., Vialletelle, P. (2006). Flexibility measures for qualification management in wafer fabs. Production Planning & Control 22(1) (81–90).
[4] Luo, M., Yan, H. C., Hu, B., Zhou, J. H., Pang, C. K. (2015). A data-driven two-stage maintenance framework for degradation prediction in semiconductor manufacturing industries. Computers & Industrial Engineering 85 (414–422).
[5] Tag, P. H., Zhang, M. T. (2006). E-Manufacturing in the semiconductor industry. IEEE Robotics and Automation Magazine 13(4) (25–32).
[6] Yu, H.-C., Lin, K.-Y., Chien, C.-F. (2014). Hierarchical indices to detect equipment condition changes with high dimensional data for semiconductor manufacturing. Journal of Intelligent Manufacturing 25(5) (933–943).

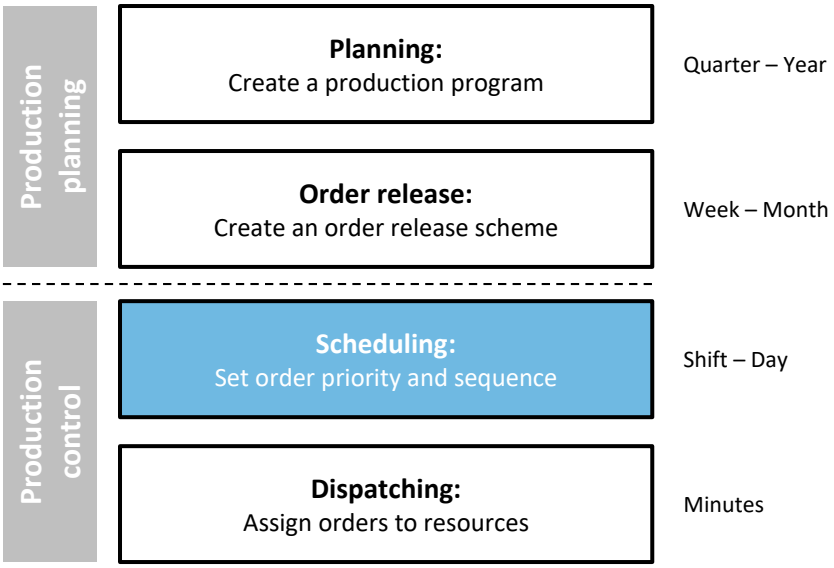

Figure 1. Production planning and production control in semiconductor manufacturing (adapted from Mönch, Fowler and Mason (2013)[1]).

3 SCIENTIFIC STATE-OF-THE-ART ANALYSIS

Production scheduling in a complex job shop environment such as wafer fabrication facilities has been of research interest since the 1960s[7]. Herein, scheduling is considered the planning process that deals with allocating resources to tasks over given periods[8]. In the context of job shop scheduling, jobs will be assigned to machines for a specific period in the future[9], with the primary aim to ensure effective and efficient use of available resources[10]. The challenges and complexities described above necessitate effective scheduling policies to maintain a competitive advantage and remain profitable in operational terms. Since Kim et al.[11] divided scheduling into release and dispatch in semiconductor fabrication systems, the dominance of job release policies based on simple heuristics has been efficaciously demonstrated by several authors in the literature[12,13,14].

Figure 1 shows an overview of production planning and production control in semiconductor manufacturing, according to Mönch, Fowler, and Mason (2013)[1]. Production planning is considered a long-term oriented process with a time horizon from weeks to a year where a

[7] Conway, R. W., Maxwell, W. L., Miller, L. W. (1967). Theory of scheduling. Wesley Publishing Company.
[8] Pinedo, M. (2016). Scheduling. Springer International Publishing.
[9] Aytug, H., Lawley, M. A., McKay, K., Mohan, S., Uzsoy, R. (2005). Executing production schedules in the face of uncertainties: A review and some future directions. European Journal of Operational Research 161(1) (86-110).
[10] Branke, J., Nguyen, S., Pickardt, C. W., Zhang, M. (2016). Automated design of production scheduling heuristics: A review. IEEE Transactions on Evolutionary Computation 20(1) (110-124).
[11] Kim, Y.-D., Kim, J. G., Choi, B., Kim, H.-U. (2001). Production scheduling in a semiconductor wafer fabrication facility producing multiple product types with distinct due dates. IEEE Transactions on Robotics and Automation 17(5) (589–598).
[12] Rose, O. (1999). CONLOAD-a new lot release rule for semiconductor wafer fabs. Proceedings of the 1999 Winter Simulation Conference (850–855).
[13] Qi, C., Sivakumar, A. I., Gershwin, S. B. (2009). An efficient new job release control methodology. International Journal of Production Research 47(3) (703–731).
[14] Li, Y., Jiang, Z., Jia, W. (2014). An integrated release and dispatch policy for semiconductor wafer fabrication. International Journal of Production Research 52(8) (2275–2292).

production program is created first and then refined in a more detailed order release scheme, which defines sets of jobs with associated release times. In contrast, production control is short-term oriented. Scheduling is defined as "the process of allocation of scarce resources over time"[1] with the goal to optimise one or more objectives (i.e. throughput, utilisation, or lead-time). Dispatching is the activity to assign the next job to be processed from a set of jobs awaiting service on an available machine in the manufacturing area.
In our literature review, we focus on scheduling production activities and maintenance planning in semiconductor manufacturing processes.

3.1 Methodology

Our state-of-the-art analysis provided in the Full Project Proposal forms the basis of this literature review. We extend it with a literature research using the Scopus database. We searched for all papers with keywords containing "Scheduling" or "Maintenance" combined with "Semiconductor", "Wafer", or "Fab" (Appendix A provides the entire search string). We only considered English-language papers published in high-quality journals (indicated by a Journal Impact Factor of 1 or higher) between 2009 and 2021. Further, we only included articles published in the research areas "Engineering", "Mathematics", and "Decision Science" to exclude publications concerned with chemical process engineering, medicine, and energy. This procedure led to a set of 344 papers that were analysed in detail.

3.2 High-performance scheduling

The highly dynamic production environment and the rapid changes in product mix ratio in semiconductor manufacturing require the implementation of efficient scheduling mechanisms, not only for elected equipment tools but also for the entire fab. Many studies[15,16,17,18] indicate that wafer fabrication needs to be understood as a make-to-order manufacturing system in which bottlenecks often shift with unbalanced workloads caused by complex product-mix orders[19]. In combination with random events like sudden machine breakdowns, rework, line incidents and the like, this results in increased variability of the performance measures of interest (e.g. cycle time, throughput, and number of tardy orders) and decreases the delivery reliability of the fab.
Beyond that, there are different areas in semiconductor manufacturing systems where different processes take place. To some extent, these necessitate different scheduling approaches as well. The following processes can be differentiated in frontend and backend of semiconductor manufacturing fabs:

[15] Ma, Y., Qiao, F., Zhao, F., Sutherland, J. (2017). Dynamic scheduling of a semiconductor production line based on a composite rule set. Applied Sciences 7(10).
[16] Priore, P., Ponte, B., Puente, J., Gómez, A. (2018). Learning-based scheduling of flexible manufacturing systems using ensemble methods. Computers & Industrial Engineering 126.
[17] Chung S.-H., Huang, C.-Y. (2003). The design of rapid production planning mechanism for the product mix changing in a wafer fabrication, Journal of the Chinese Institute of Industrial Engineers 20(2).
[18] Chien, C.-F., Hsu, C.-Y., Hsiao, C.-W. (2012). Manufacturing intelligence to forecast and reduce semiconductor cycle time. Journal of Intelligent Manufacturing 23(6).
[19] Shiue, Y.-R., Lee, K.-C., Su, C.-T. (2020). A Reinforcement Learning Approach to Dynamic Scheduling a Product-Mix Flexibility Environment. IEEE Access 8.

Table 1. Processes in semiconductor fabs according to Mönch et al. (2013)[1].

	Process	Description
Frontend	Oxidation / diffusion	The surface of a wafer is deposited with a layer of material via oxidation or diffusion. The used furnaces are usually batch machines.
	Film deposition	Dielectric or metal layers are deposited onto wafers.
	Photolithography	Wafers are coated with a photosensitive polymer and a pattern is produced by projecting ultraviolet light through a reticle. This process can be repeated many times to build circuits on the wafer.
	Etching	After photolithography, leftover photoresist is removed from the wafer.
	Ion implantation	The surface of the wafer is selectively deposited with dopant ions.
	Planarisation	The wafer surface is cleaned and levelled.
Backend	Assembly	The main assembly covers dicing saw, die attach, wire bonding and optical inspection. In areas with less strict clean-room conditions, packaging, molding, lid sealing and environmental testing area carried out.
	Wafer test	Electrical and heat-stress tests are performed.

Mathematical programming approaches can be used to solve deterministic job shop scheduling problems optimally. However, real-world systems often exhibit a high level of complexity, making these methods unsuitable for practical problems, mainly due to a high implementation effort and long computational runtimes[20]. Especially in stochastic and dynamic environments, the required computing time to get a solution becomes crucial. In semiconductor wafer fabrication facilities, these stochastic and dynamic events might be machine breakdowns, new job arrivals, stochastic processing times or changes of due dates, which make job shop scheduling an NP-hard problem[21] and the application of heuristics common[22]. Therefore, one approach in the literature is to select an optimal scheduling strategy by comparative experimentation: Singh and Mathirajan[23] investigate the impact of 15 release policies and three dispatching policies on the performance of a fictional but representative semiconductor facility in a simulation study. Different scheduling policies and their impact on the performance of a multi-product manufacturing system with finite buffers and sequence-dependent setup times are analysed in another study using continuous-time Markov chain models[24]. In particular, the impact on system throughput is investigated and conditions that characterise the single policies' superiority are identified. Min and Yih[25] develop a scheduler for selecting dispatching rules for dispatching decision variables to obtain the desired performance measures given by a user for each production interval.

[20] Branke, J., Nguyen, S., Pickardt, C. W., Zhang, M. (2016). Automated design of production scheduling heuristics: A review. IEEE Transactions on Evolutionary Computation 20(1) (110-124).
[21] Garey, M. R., Johnson, D. S., Sethi, R. (1976). The complexity of flow-shop and jobshop scheduling. Mathematics of operations research 1(2) (117-129).
[22] Burke, E. K., Hyde, M., Kendall, G., Ochoa, G., Özcan, E., Woodward, J. R. (2010). A classification of hyper-heuristic approaches. Handbook of Metaheuristics (449-468).
[23] Singh, R., Mathirajan, M. (2018). Experimental investigation for performance assessment of scheduling policies in semiconductor wafer fabrication – a simulation approach. International Journal of Advanced Manufacturing Technology 99.
[24] Feng, W., Zheng, L., Li, J. (2012). Scheduling policies in multi-product manufacturing systems with sequence-dependent setup times and finite buffers. International Journal of Production Research 50(24).
[25] Min, H.-S., Yih, Y. (2003). Selection of dispatching rules on multiple dispatching decision points in real-time scheduling of a semiconductor wafer fabrication system. International Journal of Production Research 41(16) (3921-3941).

However, both approaches – mathematical programming and experimental selection – are not suitable for the problem instances we face in AISSI. Therefore, in the following, we will limit ourselves to approaches that seem appropriate for our project. The remainder of this section is structured as follows. Cluster tools are the core production equipment in a wafer fab. Scheduling cluster tools is a major sub-problem when it comes to efficient scheduling of wafer fabs. We give a review of the main methods in section 3.2.1. However, the sole focus on scheduling cluster tools produces local optimisation minima that do not consider the overall fab performance (e.g. in terms of throughput, cycle time, and work-in-progress). Therefore, we also provide an overview of approaches for batch scheduling in section 3.2.2 and for WIP and line balancing in section 3.2.3. Moreover, we present approaches based on artificial intelligence (AI) and adaptive methods in section 3.2.4 and approaches that are specified for certain areas in semiconductor manufacturing fabs in section 3.2.5.

3.2.1 Cluster tool scheduling

Cluster tools are automated robotic manufacturing systems used for wafer fabrication that provide a reconfigurable, flexible, and efficient production environment[26,27,28] and are adopted for almost all fabrication processes[28]. A cluster tool consists of process modules, a wafer handling robot, which serves all process modules, and so-called loadlocks for wafer cassette loading and unloading[28]. Scheduling cluster tools is a major sub-problem for the production planning of semiconductor manufacturers since they require high investment, which necessitates efficient operation. Moreover, scheduling a cluster tool is a non-trivial task since it requires concurrently scheduling the robot and wafer processing, while buffer space is rare and constraints regarding the production process occur[28]. Therefore, a plethora of publications that address the topic of scheduling cluster tools can be found.
For operating cluster tools, the robot and the processing modules need to be scheduled simultaneously[28] where both tasks heavily depend on each other. For this purpose, Petri Nets have been used by many authors[28,29,30,31]. Some studies assume that a wafer can stay in the processing module for unlimited time, an assumption that is, however, found to be not always tenable in practice concerning wafer surface quality[28]. Adding wafer residency time constraints further complicates the scheduling problem. Branch-and-bound algorithms compensate for this requirement[32,33] but are computationally time-consuming[28], which justifies the adoption of

[26] Bader, M., Hall, R., Strasser, G. (1990). Integrated processing equipment. Solid State Technology 33(5).
[27] Burggraaf, P. (1995). Coping with the high cost of wafer fabs. Semiconductor International 38.
[28] Pan, C., Zhou, M., Qiao, Y., Wu, N. (2018). Scheduling Cluster Tools in Semiconductor Manufacturing: Recent Advances and Challenges. IEEE Transactions on Automation Science and Engineering 15(2).
[29] Chen, Y. F., Li, Z. W., Barkaoui, K., Giua, A. (2015). On the enforcement of a class of nonlinear constraints on Petri nets. Automatica 55(5).
[30] Chen, Y. F., Li, Z. W., Barkaoui, K., Wu, N. Q., Zhou, M. C. (2017). Compact supervisory control of discrete event systems by Petri nets with data inhibitor arcs. IEEE Transactions on Systems Man Cybernetics-Systems 47(2).
[31] Chen, Y., Li, Z., Zhou, M. (2014). Optimal supervisory control of flexible manufacturing systems by Petri nets: A set classification approach. IEEE Transactions on Automation Science and Engineering 11(2).
[32] An, Y. J., Kim, Y. D., Choi, S. W. (2016). Minimizing makespan in a two-machine flowshop with a limited waiting time constraint and sequence-dependent setup times. Computers & Operations Research 71.
[33] Bouquard, J. L., Lenté, C. (2006), Two-machine flow shop scheduling problems with minimal and maximal delays. 4OR 4(1).

heuristic approaches[34,35,36]. For example, a genetic algorithm combined with a simulation model was used to generate optimal processing sequences of lots at cluster tools[37].
In real manufacturing environments, robot activity time and wafer processing time are not deterministic but rather subject to random variation and sometimes abrupt random disturbances[28,34]. Unpredictable variations arise, such as wafer alignment failure and retrial with a failure probability between 3% and 10% depending on the cluster tool, process module delay, communication delay, and computer processing delay[28]. Therefore, it is necessary to find an efficient scheduling and control method such that a cluster tool is robust to random disturbances at task time[28]. To address this problem, an earliest start strategy with time variation within finite intervals[38] and a dynamic adjustment of the robot waiting time to offset the effect of activity time variation is proposed[39,40,41,42]. Another approach applies a timetabling technique for real-time scheduling and allows to take wafer transfer delay into account[43]. An algorithm for short-term scheduling of cluster tools based on arrival time estimation has been proposed as well[44]. While random variations and disturbances can occur and affect the system at any time, transient behaviour occurs in start-up and close-down phases, where a cluster tool is filled and emptied due to, e.g. maintenance activities, and appropriate scheduling approaches are required[45,46,47].
Beyond single cluster tools, several of these can be connected to multi-cluster tools with different structures and these need to be scheduled as well. Time constraints and multiple

[34] Kim, J. H., Lee, T. E., Lee, H. Y., Park, D. B. (2003). Scheduling analysis of timed-constrained dual-armed cluster tools. IEEE Transactions on Semiconductor Manufacturing 16(3).
[35] Lee, T. E., Park, S. H. (2005). An extended event graph with negative places and tokens for timed window constraints. IEEE Transactions on Automation Science and Engineering 2(4).
[36] Yoon, H. J., Lee, D. Y. (2005). Online scheduling of integrated single-wafer processing tools with temporal constraints. IEEE Transactions on Semiconductor Manufacturing 18(3).
[37] Dümmler, M. (1999). Using simulation and genetic algorithms to improve cluster tool performance. Proceedings of the 1999 Winter Simulation Conference (875–879).
[38] Kim, J. H., Lee, T. E. (2008). Schedulability analysis of time-constrained cluster tools with bounded time variation by an extended Petri net. IEEE Transactions on Automation Science and Engineering 5(3).
[39] Wu, N. Q., Zhou, M. C. (2010). Analysis of wafer sojourn time in dual-arm cluster tools with residency time constraint and activity time variation. IEEE Transactions on Semiconductor Manufacturing 23(1).
[40] Qu, N. Q., Zhou, M. C. (2012). Modeling, analysis and control of dual-arm cluster tools with residency time constraint and activity time variation based on Petri nets. IEEE Transactions on Automation Science and Engineering 9(2).
[41] Wu, N. Q., Zhou, M. C. (2012). Schedulability analysis and optimal scheduling of dual-arm cluster tools with residency time constraint and activity time variation. IEEE Transactions on Automation Science and Engineering 9(1).
[42] Qiao, Y., Wu, N., Yang, F., Zhou, M., Zhu, Q., Qu, T. (2019). Robust Scheduling of Time-Constrained Dual-Arm Cluster Tools with Wafer Revisiting and Activity Time Disturbance. IEEE Transactions on Systems Man Cybernetics-Systems 49(6).
[43] Lim, S.-Y., Park, Y.-J., Lee, H., Hur, S. (2014). A real-time scheduling method for the cluster tool with wafer transfer delay. International Journal of Production Research 52(4).
[44] Tu, Y.-M. (2021). Short-term scheduling model of cluster tool in wafer fabrication. Mathematics 9.
[45] Kim, T.-K., Jung, C., Lee, T.-E. (2012). Scheduling start-up and close-down periods of dual-armed cluster tools with wafer delay regulation. International Journal of Production Research 50(10).
[46] Zhu, Q., Zhou, M., Qiao, Y., Wu, N. (2018). Petri net modeling and scheduling of a close-down process for time-constrained single-arm cluster tools. IEEE Transactions on Systems Man Cybernetics-Systems 48(3).
[47] Yang, F., Qiao, Y., Gao, K., Wu, N., Zhu, Y., Simon, I.W., Su, R. (2020). Efficient Approach to Scheduling of Transient Processes for Time-Constrained Single-Arm Cluster Tools with Parallel Chambers. IEEE Transactions on Systems Man Cybernetics-Systems 50(10).

wafer product types have been taken into account to schedule[48] such systems. One-wafer cyclic schedules with minimal cycle time, conditions for their existence, and algorithms to determine the schedules are investigated for linear[49,50] and treelike[51] multi-cluster tools.
Another important characteristic of cluster tools is the number of arms, i.e. the differentiation between single-arm and dual-arm cluster tools. While in a single-arm cluster tool, only one arm has to be scheduled, a dual-arm cluster tool enables to handle more wafers at once but at the same time scheduling approaches which are adapted to this characteristic are necessary. In the literature, approaches for both scheduling single-arm cluster tools[52] and dual-arm cluster tools[53] are presented.
Despite the efforts and advances in the problem of cluster tool scheduling, two aspects prevent the provided results from being considered sufficient for scheduling an entire semiconductor manufacturing fab. On the one hand, as technology improves and market demand changes, new requirements (e.g. lot switching scheduling, multiple wafer type processing, chamber cleaning requirements, and failure response policies) are made on tool scheduling[28]. On the other hand, cluster tools are integrated into a manufacturing system with revisiting product flows, which further complicates the planning process and – since guaranteeing deadlock-freedom is mandatory – makes well-known scheduling algorithms infeasible[28].
In summary, methods of mathematical programming, as well as heuristics, prove to be an efficient and computational light approach to schedule cluster tools. However, the focus on scheduling cluster tools results in local optima, as the overall flow in the fab is not taken into consideration.

3.2.2 Batch scheduling

Some processes in semiconductor manufacturing allow to process lots in batches – e.g. in the diffusion and oxidation area. In contrast to processes where single lots have to be scheduled, this poses additional requirements for the scheduling task. In the scientific literature, batch

[48] Liu, M.-X., Zhou, B.-H. (2013). Modelling and scheduling analysis of multi-cluster tools with residency constraints based on time constraint sets. International Journal of Production Research 51(16).
[49] Bai, L., Wu, N., Li, Z., Zhou, M. (2016). Optimal One-Wafer Cyclic Scheduling and Buffer Space Configuration for Single-Arm Multicluster Tools with Linear Topology. IEEE Transactions on Systems Man Cybernetics-Systems 46(10).
[50] Yang, F., Wu, N., Qiao, Y., Zhou, M. (2017). Optimal One-Wafer Cyclic Scheduling of Time-Constrained Hybrid Multicluster Tools via Petri Nets. IEEE Transactions on Systems Man Cybernetics-Systems 47(11).
[51] Yang, F.J., Wu, N.Q., Qiao, Y., Zhou, M.C. (2018). Optimal One-Wafer Cyclic Scheduling of Hybrid Multirobot Cluster Tools With Tree Topology. IEEE Transactions on Systems Man Cybernetics-Systems 48(2).
[52] Yang, F., Wu, N., Qiao, Y., Zhou, M., Su, R., Qu, T. (2020). Modeling and Optimal Cyclic Scheduling of Time-Constrained Single-Robot-Arm Cluster Tools via Petri Nets and Linear Programming. IEEE Transactions on Systems Man Cybernetics-Systems 50(3).
[53] Zhu, Q., Zhou, M., Qiao, Y., Wu, N., Hou, Y. (2020). Multiobjective Scheduling of Dual-Blade Robotic Cells in Wafer Fabrication. IEEE Transactions on Systems Man Cybernetics-Systems 50(12).

scheduling is approached by mathematical optimisation[54,55,56,57], genetic algorithms[58,59] or with other heuristics[60,61]. In particular, Trindade et al.[56] develop mixed-integer linear programming formulations for single and parallel processing machines with and without job release times. Jula and Leachman[57] take into account resource constraints when scheduling parallel batch machines. Wang and Uzsoy[62] consider the problem of scheduling a single batch machine with job release dates to minimise maximum lateness. They use a genetic algorithm coupled with dynamic programming techniques to solve the problem. Mönch et al.[63] use genetic algorithms coupled with time window techniques and decision theory approaches. Yugma et al.[64] present a scheduling approach dedicated to the diffusion area in semiconductor manufacturing and which utilises simulated annealing. Mönch et al.[65] use a neural network to adjust the look-ahead parameter in the Apparent Tardiness Cost (ATC) dispatching rule for parallel batch machines.

3.2.3 Scheduling focused on line balancing and WIP balancing

Apart from batch scheduling, Lee et al.[66] present a systematic approach for assigning wafers to machines in semiconductor manufacturing to maximise wafer yield while satisfying a pre-determined target level of productivity. In this course, line balancing is addressed as well. Line

[54] Lee, J.-H., Kim, S.H., Lee, Y.H. (2013). Discrete lot sizing and scheduling problem under batch processing constraints in the semiconductor manufacturing. International Journal of Advanced Manufacturing Technology 69.
[55] Lee, Y.H., Lee, Y.H. (2013). Minimising makespan heuristics for scheduling a single batch machine processing machine with non-identical job sizes. International Journal of Production Research 51(12).
[56] Trindade, R.S., de Araújo, O.C.B., Fampa, M.H.C., Müller, F.M. (2018). Modelling and symmetry breaking in scheduling problems on batch processing machines. International Journal of Production Research 56(22).
[57] Jula, P., Leachman, R.C. (2010). Coordinated multistage scheduling of parallel batch-processing machines under multiresource constraints. Operations Research 58(4).
[58] Su, G., Wang, X. (2011). Weighted nested partitions based on differential evolution (WNPDE) algorithm-based scheduling of parallel batching processing machines (BPM) with incompatible families and dynamic lot arrival. International Journal of Computer Integrated Manufacturing 24(6).
[59] Noroozi, A., Mokhtari, H., Kamal Abadi, I.N. (2013). Research on computational intelligence algorithms with adaptive learning approach for scheduling problems with batch processing machines. Neurocomputing 101.
[60] Chen, H., Du, B., Huang, G.Q. (2010). Metaheuristics to minimise makespan on parallel batch processing machines with dynamic job arrivals. International Journal of Computer Integrated Manufacturing 23(10).
[61] Zhou, S., Chen, H., Xu, R., Li, X. (2014). Minimising makespan on a single batch processing machine with dynamic job arrivals and non-identical job sizes. International Journal of Production Research 52(8).
[62] Wang, C., Uzsoy, R. (2002). A Genetic Algorithm to Minimize Maximum Lateness on a Batch Processing Machine. Computers & Operations Research 29(12).
[63] Mönch, L., Balasubramanian, H., Fowler, J. W., Pfund, M. E. (2005). Heuristic scheduling of jobs on parallel batch machines with incompatible families and unequal ready times of the jobs. Computers & Operations Research 32 (2731–2750).
[64] Yugma, C., Dauzère-Pérès, S., Artigues, C., Derreumaux, A., Sibille, O. (2012). A batching and scheduling algorithm for the diffusion area in semiconductor manufacturing. International Journal of Production Research 50(8).
[65] Mönch, L., Zimmermann, J., Otto, P. (2006). Machine learning techniques for scheduling jobs with incompatible families and unequal ready times on parallel batch machines. Engineering Applications of Artificial Intelligence 19(3) (235-245).
[66] Lee, D.-H., Lee, C.-H., Choi, S.-H., Kim, K.-J. (2019). A method for wafer assignment in semiconductor wafer fabrication considering both quality and productivity perspectives. Journal of Manufacturing Systems 52.

balancing is also focused by Yu et al.[67], applying a prediction-based dynamic scheduling method with a multi-layer perceptron. They determine a prediction model based on a simulation dataset of empirical industrial manufacturing facilities and incorporate the predictions in a dynamic dispatching rule for optimal load balancing based on the queue length at each workstation.
Production processes are usually constrained by machines which are a bottleneck. Consequently, scheduling these bottleneck machines is of great importance with regard to the performance of the whole fab. Chiou and Wu[68] present and test seven metaheuristics for scheduling bottleneck machines in semiconductor manufacturing facilities. In particular, an appropriate allocation of jobs to various machines is considered in their work. Another approach to control and schedule bottlenecks is presented by Hu et al.[69]. They propose a dynamic WIP control strategy comprising offline target WIP level setting and online WIP control. Based on detected and classified bottlenecks, target WIP levels are allocated to the bottlenecks to avoid process fluctuations because of unpredictable events. During real-time dispatching, upstream machines of bottlenecks modify their dispatching order to adjust the deviation of WIP levels at the bottlenecks. The current WIP distribution of the entire system is taken into account by Siebert et al.[70] as well. They propose a fluid model lot dispatching policy that considers travel times and iteratively optimises lot selection based on the current WIP distribution.

3.2.4 AI-based and adaptive scheduling and dispatching rules

Since modelling scheduling rules can be tedious and time-consuming, automated generation of heuristics is frequently performed using machine learning techniques. These so-called hyper-heuristics are defined as an automated methodology for selecting or generating heuristics to solve complex computational search problems[22]. Hyper-heuristics can be trained and applied on static, deterministic and stochastic instances[71].
Huang and Chen[72] propose an online rescheduling mechanism combined with theory of constraints (TOC) and deploy a genetic algorithm for searching dispatching rule sets. Lin and Chen[73] combine a genetic algorithm with simulation to optimise scheduling decisions in a semiconductor backend assembly facility. Hwang and Jang[74] deploy an AI approach based on Q-learning for scheduling automated transport systems. Waschneck et al.[75] propose applying

[67] Yu, Q., Yang, H., Lin, K.-Y., Li, L. (2020). A predictive dispatching rule assisted by multi-layer perceptron for scheduling wafer fabrication lines. Journal of Computing and Information Science in Engineering 20(3).
[68] Chiou, C.-W., Wu, M.-C. (2014). Scheduling of multiple in-line steppers for semiconductor wafer fabs. International Journal of Systems Science 45(3).
[69] Hu, H., Jiang, Z., Zhang, H. (2010). A dynamic WIP control strategy for bottlenecks in a wafer fabrication system. International Journal of Production Research 48(17).
[70] Siebert, M., Bartlett, K., Kim, H., Ahmed, S., Lee, J., Nazzal, D., Nemhauser, G., Sokol, J. (2018). Lot targeting and lot dispatching decision policies for semiconductor manufacturing: optimisation under uncertainty with simulation validation. International Journal of Production Research 56(1-2).
[71] Hildebrandt, T., Heger, J., Scholz-Reiter, B. (2010). Towards improved dispatching rules for complex shop or scenarios: a genetic programming approach. Proceedings of the 12th annual conference on Genetic and evolutionary computation (257–264).
[72] Huang, H., Chen, T. (2006). A New Approach to On-Line Rescheduling for a Semiconductor Foundry Fab. 2006 IEEE International Conference on Systems, Man and Cybernetics (4727-4732).
[73] Lin, J. T., Chen, C.-M. (2015). Simulation optimization approach for hybrid flow shop scheduling problem in semiconductor back-end manufacturing. Simulation Modelling Practice and Theory 51.
[74] Hwang, I., Jang, Y. J. (2019). Q(λ) learning-based dynamic route guidance algorithm for overhead hoist transport systems in semiconductor fabs. International Journal of Production Research (1-23).
[75] Waschneck, B., Reichstaller, A., Belzner, L., Altenmüller, T., Bauernhansl, T., Knapp, A., Kyek, A. (2018). Optimization of global production scheduling with deep reinforcement learning. Procedia CIRP 72(1) (1264-1269).

Deep Q Network agents, a deep reinforcement learning approach for global job shop scheduling across several production workcentres. Based on the insight that approaches of reinforcement learning allow to set up powerful decision-making systems (e.g. for playing games like chess or Go), Kuhnle et al.[76] present an adaptive production control system based on reinforcement learning. In particular, they address the design of a reinforcement learning approach with regard to state, action and reward function. Additionally, they identify robust designs for reinforcement learning systems and present a real-world example of a semiconductor manufacturer. With regard to reinforcement learning methods usually being considered as 'black box' models, Kuhnle et al.[77] propose an approach to increase the plausibility of reinforcement learning based control strategies. They combine methods with high prediction accuracy (e.g. neural networks) on the one hand and high explainability (e.g. decision trees) on the other hand.

Shiue et al.[78] propose a reinforcement learning-based dynamic scheduling method that applies the multiple dynamic scheduling rule selection (MDSR) mechanism to effectively respond to product-mix ratio variations in semiconductor wafer fabrication systems. The proposed reinforcement learning (RL) approach is based on Q-learning. It uses a dynamic and multi-pass approach for deciding MDSRs, which is applied following the status of the production system at the beginning of the scheduling interval. The system then decides the most suitable MDSRs for the next scheduling interval.

A fuzzy neural network (FNN) based rescheduling decision model is implemented by Zhang et al.[79], which can rapidly choose an optimised rescheduling strategy to schedule the semiconductor wafer fabrication lines according to current system disturbances.

A self-adaptive agent-based fuzzy-neural system is constructed by Chen (2011)[80] to enhance the performance of scheduling jobs in a wafer fabrication factory. The system integrates dispatching, performance evaluation and reporting, and scheduling policy optimisation. Unlike in past studies, a single pre-determined scheduling algorithm is used for all agents. In this study, every agent develops and modifies its own scheduling algorithm to adapt to local conditions.

Zhang et al.[81] present an imperialist competitive algorithm incorporating remaining cycle time prediction for photolithography machines' scheduling problem with the objective of total completion time minimisation. A deep autoencoder neural network is presented at first to predict remaining cycle time, responding to the environmental changes. Secondly, an imperialist competitive algorithm in the framework of a rolling horizon strategy is proposed to address the scheduling problem, incorporated with the accurately predicted remaining cycle time.

To minimise the makespan for a multichip product (MCP) scheduling problem, Park et al.[82] propose a setup change scheduling method using reinforcement learning (RL) in which each agent determines setup decisions in a decentralised manner and learns a centralised policy by

[76] Kuhnle, A., Kaiser, J.-P., Theiß, F., Stricker, N., Lanza, G. (2021). Designing an adaptive production control system using reinforecement learning. Journal of Intelligent Manufacturing 32.
[77] Kuhnle, A., May, M. C., Schäfer, L., Lanza, G. (2021). Explainable reinforcement learning in production control of job shop manufacturing system. International Journal of Production Research.
[78] Shiue, Y.-R., Lee, K.-C., Su, C.-T. (2020). A Reinforcement Learning Approach to Dynamic Scheduling in a Product-Mix Flexibility Environment. IEEE Access 8.
[79] Zhang, J., Qin, W., Wu, L.H., Zhai, W.B. (2014). Fuzzy neural network-based rescheduling decision mechanism for semiconductor manufacturing. Computers in Industry 65, pp. 1115 – 1125.
[80] Chen, T. (2011). A self-adaptive agent-based fuzzy-neural scheduling system for a wafer fabrication factory. Expert Systems with Applications 38, (7158 – 7168).
[81] Zhang, P., Zhao, X., Sheng, X., Zhang, J. (2018). An Imperialist Competitive Algorithm Incorporating Remaining Cycle Time Prediction for Photolithography Machines Scheduling. IEEE Access 6.
[82] Park, I.-B., Huh, J., Kim, J., Park, J. (2020). A Reinforcement Learning Approach to Robust Scheduling of Semiconductor Manufacturing Facilities. IEEE Transactions on Automation Science and Engineering 17(3) (1420–1431).

sharing a neural network among the agents to deal with the changes in the number of machines. Furthermore, novel definitions of state, action, and reward are proposed to address the variabilities in production requirements and initial setup status.
Lee et al.[83] formulate the fab scheduling problem as a semi Markov decision process and propose a reinforcement learning method combined with a fab simulator to determine a dispatching policy.

3.2.5 Area-specific scheduling approaches

In semiconductor manufacturing, there are different areas where different production steps are executed (see Table 1). As different technologies and characteristics hallmark these areas, area-specific scheduling approaches can be differentiated. One core area in semiconductor manufacturing is photolithography. For scheduling purposes within the photolithography area, a framework for rolling horizon scheduling[84], a dynamic scheduling method based on a Kohonen neural network[85], a metaheuristic that takes auxiliary resources into account[86] and a mixed-integer programming model, as well as a heuristic to optimise scheduling of photolithography processes with both individual and cluster tools,[87] are proposed. Another publication is concerned with the rescheduling problem in the photolithography area and applies simulated annealing, a genetic algorithm and tabu search to approach it. Additionally, an approach for sensitivity search is proposed[88].
Another area in semiconductor manufacturing is the final testing stage. The final testing scheduling problem in semiconductor manufacturing has been addressed in several publications where both heuristics[89] and genetic algorithms[90,91] are utilised. Different methodological approaches have been proposed for the wafer sorting scheduling problem that has to be solved in the course of testing wafers[92,93,94].

[83] Lee, W., Ko, K., Shin, H. (2019). Simulation-Based Multi-Objective Fab Scheduling by using Reinforcement Learning. Proceedings of the 2019 Winter Simulation Conference.
[84] Zhang, P., Lv, Y., Zhang, J. (2018). An improved imperialist competitive algorithm based photolithography machines scheduling. International Journal of Production Research 56(3).
[85] Zhou, B.-H., Li, X., Fung, R.Y.K. (2015). Dynamic scheduling of photolithography process based on Kohonen neural network. Journal of Intelligent Manufacturing 26.
[86] Bitar, A., Dauzère-Pérès, S., Yugma, C., Roussel, R. (2016). A memetic algorithm to solve an unrelated parallel machine scheduling problem with auxiliary resources in semiconductor manufacturing. Journal of Scheduling 19.
[87] Chalil Madathil, S., Nambiar, S., Mason, S.J., Kurz, M.E. (2018). On scheduling a photolithography area containing cluster tools. Computers & Industrial Engineering 121.
[88] Hung, Y.-F., Liang, C.-H., Chen, J.C. (2013). Sensitivity search for the rescheduling of semiconductor photolithography operations. International Journal of Advanced Manufacturing Technology 67.
[89] Wang, S., Wang, L., Liu, M., Xu, Y. (2015). A hybrid estimation of distribution algorithm for the semiconductor final testing scheduling problem. Journal of Intelligent Manufacturing 26.
[90] Wang, S., Wang, L. (2015). A knowledge-based multi-agent evolutionary algorithm for semiconductor final testing scheduling problem. Knowledge-Based Systems 84.
[91] Zheng, X.-L., Wang, L., Wang, S.-Y. (2014). A novel fruit fly optimization algorithm for the semiconductor final testing scheduling problem. Knowledge-Based Systems 57.
[92] Ying, K.-C. (2012). Scheduling identical wafer sorting parallel machines with sequence-dependent setup times using an iterated greedy heuristic. International Journal of Production Research 50(10).
[93] Ying, K.-C., Lin, S.-W. (2014). Efficient wafer sorting scheduling using a hybrid artificial immune system. Journal of the Operational Research Society 65(2).
[94] Lin, S.-W., Lee, Z.-J., Ying, K.-C., Lin, R.-H. (2011). Meta-heuristic algorithms for wafer sorting scheduling problems. Journal of the Operational Research Society 62(1).

3.3 Maintenance planning

Typically, the status of machines is assumed to be independent of the production schedule. However, advanced process technologies require high conformance to process specifications. Even though a machine is shown as available in the Manufacturing Execution System (MES), the process quality may not be guaranteed due to machine deterioration[95]. Additionally, from a long-term viewpoint, the probability of machine failures naturally increases with the age of a machine[96]. However, as time-based maintenance often causes over-maintenance[97], while still not fully being able to address the problem of unplanned downtime, predictive maintenance is gaining popularity. In the Machine Learning (ML) community, Predictive Maintenance (PdM) is typically modelled either as a problem of Failure Prediction (FP)[98] or of estimating the Remaining Useful Life (RUL)[99]. Guan et al.[100] present a framework for throughput-driven condition-based maintenance aiming at predictive maintenance for heavily utilised and reconfigurable production equipment.

Another common method is to evaluate the machine condition in the Advanced Process Control (APC) framework through calculating an Equipment Health Indicator (EHI)[101,102]. To evaluate EHI, several methodologies have been developed in the literature. The multivariate process capability index is commonly used to integrate multiple parameters into an overall EHI. A recipe-independent EHI and its hierarchical monitoring scheme are further proposed to evaluate the machine health and to diagnose faults systematically[103].

Susto et al.[104] generate so-called "health factors" or quantitative indicators of a system's status associated with a given maintenance issue and determine their relationship to operating costs and failure risk. They train multiple classification modules with different prediction horizons to provide different performance tradeoffs in terms of frequency of unexpected breaks and unexploited lifetime and then employ this information in an operating cost-based maintenance

[95] Sloan, T. W., Shanthikumar, J. G. (2002). Using in-line equipment condition and yield information for maintenance scheduling and dispatching in semiconductor wafer fabs. IIE Transactions 34(2) (191–209).

[96] Kamien, M. I., Schwartz, N. L. (1971). Optimal maintenance and sale age for a machine subject to failure. Management Science 17(8) (B495–B504).

[97] Ahmad, R., Kamaruddin, S. (2012). An overview of time-based and condition-based maintenance in industrial application. Computers & Industrial Engineering 63(1) (135–149).

[98] Babu, G.S., Zhao, P., Li, X. L. (2016). Deep convolutional neural network based regression approach for estimation of remaining useful life. International Conference on Database Systems for Advanced Applications (214–228).

[99] El-Koujok, M., Gouriveau, R., Zerhouni, N. (2011). Reducing arbitrary choices in model building for prognostics: An approach by applying parsimony principle on an evolving neuro-fuzzy system. Microelectronics Reliability 51(2) (310–320).

[100] Guan, C.S., Kuang, Y.C., Ooi, M.P.-L. (2013). Throughput-driven condition-based maintenance for frequently reconfigured mass production equipment. International Journal of Advanced Manufacturing Technology 65.

[101] Holfeld, A., Barlovic, R., Good, R. P. (2007). A fab-wide APC sampling application. IEEE Transactions on Semiconductor Manufacturing 20(4) (393–399).

[102] Obeid, A., Dauzère-Pérès, S., Yugma, C. (2012). Scheduling on parallel machines with time constraints and equipment health factors. IEEE Conference on Automation Science and Engineering (401–406).

[103] Blue, J., Gleispach, D., Roussy, A., Scheibelhofer, P. (2013). Tool condition diagnosis with a recipe-independent hierarchical monitoring scheme. IEEE Transactions on Semiconductor Manufacturing 26(1) (82–91).

[104] Susto, G.A., Schirru, A., Pampuri, S., McLoone, S. (2015). Machine Learning for Predictive Maintenance: A Multple Classifier Approach. IEEE Transactions on Industrial Informatics 11(3).

decision system to minimise expected costs. Luo et al.[105] employ back-propagation Neural Networks and evolvable NN for the degradation prediction and multiple regression forecasting to check prediction accuracies. Zhang et al.[106] propose a purely data-driven approach for solving the Health Indicator Learning (HIL) problem based on Deep Reinforcement Learning (DRL). They find the HIL problem can be mapped to a credit assignment problem. DRL learns from failures by naturally back-propagating the credit of failures into intermediate states.
Ramírez-Hernández et al.[107] present architecture and implementation of a software for preventive maintenance optimisation, which is based on algorithms for optimal scheduling of preventive maintenance tasks in semiconductor manufacturing. Additionally, results from applying this software in simulation case studies based on real industrial data are reported.

3.4 Integrated production and maintenance planning

In highly flexible and highly integrated manufacturing systems – such as semiconductor manufacturing – dynamic interactions between equipment conditions, operations executed on the tools and product quality necessitate joint decision-making in maintenance scheduling and production operations[108].
Lee and Ni[109] present a decision-making architecture to determine maintenance and production dispatching policies based on condition monitoring information and the relationship between machine degradation and product quality. They apply a Markov decision process for long-term decision-making and mathematical programming for short-term decision-making. Celen and Djurdjanovic[110] address the aspect that the condition of equipment is usually not perfectly observable by applying a partially observable Markov decision process to model the interaction of production and maintenance activities. Chung et al.[111] propose a binary integer programming model and a heuristic for a single machine scheduling problem with maintenance activities and irregular intervals in between. In another study, Tonke and Grunow[112] investigate simultaneous scheduling of preventive maintenance, shutdowns and production of robotic cells in semiconductor manufacturing and show that integrating production and maintenance scheduling has substantial advantages. Li and Ma[113] develop a particle swarm optimisation algorithm to solve an integrated preventive maintenance and production scheduling problem

[105] Luo, M., Yan, H.-C., Hu, B., Zhou, J.-H., Pang, C. K. (2015). A data-driven two-stage maintenance framework for degradation prediction in semiconductor manufacturing industries. Computers & Industrial Engineering 85 (414–422).
[106] Zhang, C., Gupta, C., Farahat, A., Ristovski, K., Ghosh, D. (2019). Equipment Health Indicator Learning Using Deep Reinforcement Learning. Machine Learning and Knowledge Discovery in Databases, Springer International Publishing (488-504).
[107] Ramírez-Hernández, J.A., Crabtree, J., Yao, X., Fernandez, E., Fu, M.C., Janakiram, M., Marcus, S.I., O'Connor, M., Patel, N. (2010). Optimal preventive maintenance scheduling in semiconductor manufacturing systems: Software tool and simulation case studies. IEEE Transactions on Semiconductor Manufacturing 23(3).
[108] Celen, M., Djurdjanovic, D. (2012). Operation-dependent maintenance scheduling in flexible manufacturing systems. CIRP Journal of Manufacturing Science and Technology 5, (296 – 308).
[109] Lee, S., Ni, J. (2013). Joint decision making for maintenance and production scheduling of production systems. International Journal of Advanced Manufacturing Technology 66.
[110] Celen, M., Djurdjanovic, D. (2020). Integrated maintenance and operations decision making with imperfect degradation state observations. Journal of Manufacturing Systems 55.
[111] Chung, T.-P., Xue, Z., Wu, T., Shih, S.C. (2019). Minimising total completion time on single-machine scheduling with new integrated maintenance activities. International Journal of Production Research 57(3).
[112] Tonke, D., Grunow, M. (2018). Maintenance, shutdown and production scheduling in semiconductor robotic cells. International Journal of Production Research 56(9).
[113] Li, R., Ma, H. (2017). Integrating Preventive Maintenance Planning and Production Scheduling under Reentrant Job Shop. Mathematical Problems in Engineering.

of a re-entrant job shop and evaluate the model with simple simulation experiments. Ao et al.[114] propose a two-step strategy to approach the integrated decision problem of production and maintenance. First, a dynamic maintenance plan including time points for maintenance is determined based on a Markov decision process. Second, an integrated decision model for production and maintenance and to balance a production line when times for production and maintenance are conflicting is set up.

Applying EHI in production control helps identify machine failures that prolong production cycle times and improve the production schedule's effectiveness. Many studies in the literature have proposed and discussed condition-based maintenance models[115,116,117]. A two-level maintenance methodology for manufacturing systems is proposed by Xia et al.[118], in which machine-level predictive maintenance schedules are considered first and then a variable maintenance time window is used to optimise system-level maintenance. Such condition-based maintenance and health management can further be considered in batch production with variable lot sizes[119]. Both long and short-term machine deterioration and condition-based maintenance motivate the integration of machine conditions in scheduling decisions.

Information on the condition of machines facilitates the quality improvement of scheduling decisions. Exploiting degradation modelling and monitoring, Cholette et al.[120] consider preventive maintenance events and production sequencing jointly to design an integrated decision policy, achieving higher expected profits than a traditional maintenance policy. Kao et al.[121] adopt a Markov decision process model to include machine deterioration and determine equipment maintenance and production schedules for maximising the long-run expected average profit. These works mainly focus on a single tool or a set of homogeneous tools. For scheduling a cluster tool taking into account chamber conditions and maintenance activities a genetic algorithm is presented[122].

Some works addressed equipment condition related scheduling problems. For example, machine condition parameters are considered by Doleschal et al.[123] in the optimal schedule to improve yield. However, the machine condition is modelled as a constant over the whole scheduling horizon, ignoring that the machine condition changes after processing wafers. To

[114] Ao, Y., Zhang, H., Wang, C. (2019). Research of an integrated decision model for production scheduling and maintenance planning with economic objective. Computers & Industrial Engineering 137.

[115] Cheng, G. Q., Zhou, B. H., Li, L. (2017). Joint optimization of lot sizing and condition-based maintenance for multi-component production systems. Computers & Industrial Engineering 110 (538–549).

[116] Cui, W., Lu, Z., Li, C., Han, X. (2018). A proactive approach to solve integrated production scheduling and maintenance planning problem in flow shops. Computers & Industrial Engineering 115 (342–353).

[117] Luo, M., Yan, H. C., Hu, B., Zhou, J. H., Pang, C. K. (2015). A data-driven two-stage maintenance framework for degradation prediction in semiconductor manufacturing industries. Computers & Industrial Engineering 85 (414–422).

[118] Xia, T., Tao, X. Y., Xi, L. (2017). Operation process rebuilding (OPR)-oriented maintenance policy for changeable system structures. IEEE Transactions on Automation Science and Engineering 14(1) (139–148).

[119] Xia, T., Jin, X., Xi, L., Ni, J. (2015). Production-driven opportunistic maintenance for batch production based on MAM–APB scheduling. European Journal of Operational Research 240 (781–790).

[120] Cholette, M. E., Celen, M., Djurdjanovic, D., Rasberry, J. D. (2013). Condition monitoring and operational decision making in semiconductor manufacturing. IEEE Transaction on Semiconductor Manufacturing 26(4) (454–464).

[121] Kao, Y. -T., Zhan, S.-C., Chang, S.-C., Ho, J.-H., Wang, P., Luh, P. B., Chang, J. (2011). Near optimal furnace tool allocation with batching and waiting time constraints. IEEE Conference on Automation Science and Engineering (108–113).

[122] Lee, S., Ni, J. (2012). Genetic algorithm for job scheduling with maintenance consideration in semiconductor manufacturing process. Mathematical Problems in Engineering.

[123] Doleschal, D., Weigert, G., Klemmt, A. (2015). Yield integrated scheduling using machine condition parameter. Proceedings of the 2015 Winter Simulation Conference (2953–2963).

model and illustrate the integration of EHI in scheduling decisions to balance between productivity and quality risk, Kao et al.[124] present two mixed-integer linear programs to schedule jobs on heterogeneous parallel batching machines. They demonstrate that both problems are NP-hard, meaning that solving the problems for large instances in a highly dynamic manufacturing environment requires fast heuristic algorithms, which are not yet available.

Geurtsen et al.[125] study a new scheduling problem on unrelated parallel machines with simultaneous scheduling of jobs and resource-constrained preventive maintenance activities. They develop a mathematical model to investigate preventive maintenance activities that are known in advance and that have to be scheduled in one of its given discrete time windows within the scheduling horizon. Cui et al.[126] investigate the integration of production planning and maintenance planning to optimise the quality robustness and solution robustness of schedules for flow shops with failure uncertainty.

The multi-objective flexible job shop scheduling problem with maintenance activities is approached with a novel discrete artificial bee colony algorithm by Li et al.[127]. First, a schedule without maintenance activities is generated. They are inserted dynamically afterwards based on a heuristic. The algorithm results in highly effective and efficient performance for a set of well-known benchmark instances from literature.

Kaihara et al.[128] present a method for re-entrant production floor optimisation using Lagrangian decomposition coordination. By regarding maintenance as jobs that are limited by a starting and finishing time, the proposed approach produces a schedule that can facilitate proper maintenance.

[124] Kao, Y.-T. , Dauzère-Pérès, S., Blue, J., Chang, S.-C. (2018). Impact of integrating equipment health in production scheduling for semiconductor fabrication. Computers & Industrial Engineering 120 (450-459).

[125] Geurtsen, M., Adan, J., Stokkermans, J., Adan, I. J. B. F., Akcay, A. (2020). Integrated Maintenance & Production Scheduling (in preparation).

[126] Cui, W., Lu, Z., Li, C., Han, X. (2018). A proactive approach to solve integrated production scheduling and maintenance planning problem in flow shops. Computers & Industrial Engineering 115 (342-353).

[127] Li, J., Pan, Q., Tasgetiren, M. (2014). A discrete artificial bee colony algorithm for the multi-objective flexible job-shop scheduling problem with maintenance activities. Applied Mathematical Modelling 38 (1111-1132).

[128] Kaihara, T., Fujii, N., Tsujibe, A., Nonaka, Y. (2010). Proactive maintenance scheduling in a re-entrant flow shop using Lagrangian decomposition coordination method. CIRP Annals 59(1) (453–456)

4 INDUSTRIAL STATE-OF-THE-ART ANALYSIS

In chapter 3, a broad overview of methods and approaches for production scheduling and maintenance planning in the semiconductor industry was presented. While these approaches provide various relevant thoughts for practical applications, most of them have been tested and demonstrated using exemplary data or small examples from practice. At the interface of scientific approaches and practical applications, a common understanding of challenges and solution approaches is required. As the project AISSI aims at implementing sophisticated scientific approaches for production and maintenance scheduling in industry, a survey has been conducted to grasp the understanding, definitions and terminology of the partners from industry in this domains. This provides the counterpart of the scientific perspective in chapter 3 and aims at establishing a common base and understanding for the work throughout the project.

4.1 Methodology

The core of the survey was a questionnaire that was distributed to all project partners from industry. The questionnaire covers different aspects that are relevant to grasp the industrial state-of-the-art regarding production scheduling and maintenance planning. It is composed of nine questions, each focusing on an aspect of relevance. The covered aspects are:

- Scheduling vs. dispatching and the corresponding time horizon
- Focus of scheduling on single machines vs. holistic/factory-wide perspective
- Criteria taken into account for scheduling decisions
- Manual vs. automated scheduling decision-making
- Methodological approaches for scheduling decision-making
- Characterisation of the digital twin that is used/developed to support planning
- Categorisation of maintenance activities (planned, unplanned etc.)
- Relation between production planning and maintenance planning (separately, integrated etc.)
- Availability and accessibility of data to use in digital twin and other planning approaches

The questions provide a structure for the questionnaire and some options which are given for most of the questions define a framework so that the answers of the project partners that are asked to fill the questionnaire are comparable. Additionally, a section for free text was added to each question so that each project partner could describe their approaches as detailed as necessary and specify unique characteristics beyond the given pre-formulated options.
During the creation of the questionnaire, a draft was created first and presented to and discussed with the project partners from industry to ensure that it is both comprehensible and pertinent. Based on the feedback from the project partners, the questionnaire was adjusted and finalised. A blank version of the final questionnaire is attached in the annex of this deliverable.
The questionnaire was sent to and filled by all project partners from industry – Bosch, Nexperia, SYSTEMA and D-SIMLAB. Due to the limited number of project partners, it was neither possible nor intended to provide a representative study of the industrial state-of-the-art regarding production scheduling and maintenance planning. Instead, the survey was intended to grasp the current approaches applied by the project partners from industry and based on this, to establish a common understanding, terminology, and starting point for developing AI-based approaches for production scheduling and maintenance planning.
To do so, the filled questionnaires have been evaluated and aggregated. The aggregated results are presented in the following section.

4.2 Results

To provide an overview of the state-of-the-art in semiconductor manufacturing regarding the aspects listed in the section above, the feedback of the project partners in the filled questionnaires was aggregated and is presented below, following the structure of the questions in the questionnaire.

While there are definitions of and differentiation between the terms scheduling and dispatching and the corresponding time horizons in the scientific literature (see Figure 1), experience shows that these terms and the corresponding time horizons are not unique and unambiguous in industrial practice. Therefore, the first question in the questionnaire addressed these two terms, their use and the respective time horizon.
On the one hand, the feedback from the project partners corresponds with the structure proposed by Mönch, Fowler and Mason[1] shown in Figure 1 with regard to the fact that scheduling refers to a longer time horizon than dispatching. However, the project partners state exact time horizons of scheduling and dispatching that differ between hours, shifts and a day (scheduling) and between real-time and seconds (dispatching) respectively. Additionally, it is pointed out that the time horizon of scheduling depends on the tool group and the number of production steps involved.
The stated time horizons implicitly indicate that scheduling refers to determining a (production and/or maintenance) plan while dispatching is concerned with real-time decision-making and execution of such a plan. This insight provides the base for another dimension to differ scheduling and dispatching along. Scheduling is usually approached based on some objective function and by methods from the operations research domain. A common objective is to optimise an assignment or ordering problem mathematically. On the other hand, dispatching tends to be either based on heuristic rules or refers to a dispatch list that comprises certain lots in a certain order and is used as a prescription for operators and machines. The creation of a dispatch list may or may not be based on scheduling outcomes.

The second question of the questionnaire refers to the focus of scheduling: Does it address single machines independently or are several machines, equipment groups or the like taken into account holistically? This differentiation is motivated by the assumption that scheduling decisions only taking into account single machines independently might result in locally optimal decisions (for the single machine of interest) but do not necessarily result in globally optimal decisions (for the whole fab, an equipment group or several machines). Scheduling of several single machines independently might even result in local decisions that negatively affect each other.
The feedback from the project partners indicates that single machines are the primary focus of dispatching. However, the incorporation of criteria that exceed the scope of a single machine but are associated with the whole factory is emphasised. On the other hand, the broader scope of equipment groups and sequences of equipment groups is usually addressed by scheduling. In particular, scheduling is common for lithography, implantation and furnace areas. However, the feedback from the partners points out that neither dispatching nor scheduling addresses a whole factory at once.

Regarding different criteria that are taken into account to come to a dispatching and/or scheduling decision, the questionnaire also aims to record the criteria applied by the project partners from industry. The corresponding question proposes several common criteria and allows to add and describe other criteria as well.
The aggregated feedback from the project partners is shown in Figure 2. It becomes apparent that multiple criteria are taken into account for dispatching and/or scheduling by the project partners from industry (multiple answers were allowed). Moreover, the results show that

dynamic and lot-specific criteria like batching, priorities and due dates are of higher relevance in industry than machine-specific and more static criteria like the queue length and the processing time. It is also worth noting that the approach of 'First-Come-First-Served' (FCFS) – which corresponds to not taking any decision/not making any change at a given production and/or maintenance plan – is of minor relevance in industry.

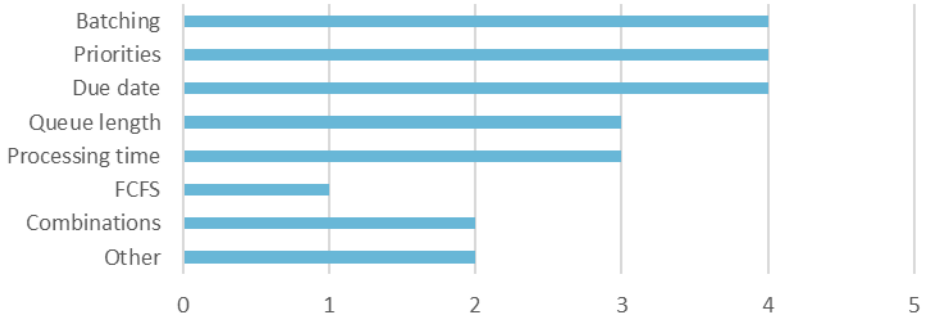

Figure 2. Criteria taken into account for scheduling by the project partners from industry.

Beyond that, some project partners reported to use combinations of the proposed criteria and other than the proposed criteria. As other criteria, they stated to take the setup time of equipment in case of recipe changes, the queue time, and the stream size into account.

While the criteria taken into account for scheduling provide an insight into which aspects and numbers scheduling decisions are underlying, it is not clear how these are utilised from a methodological point of view to come to a scheduling decision. Therefore, another question in the questionnaire addresses this aspect by proposing some common methodological approaches and the option to specify other approaches which are not proposed in the questionnaire.
From the feedback of the project partners (see Figure 3), it becomes clear that decision rules are most widely used in industrial practice, followed by mathematical optimisation approaches. While decision rules tend to be heuristic, mathematical optimisation approaches allow to determine an optimal solution for a given set of parameters and a given objective function. Beside mathematical optimisation approaches, simulation-based optimisation approaches are mentioned by a project partner as another methodological approach. In contrast, approaches based on machine learning or artificial intelligence are rarely applied by the project partners from industry at the moment. Having the approaches proposed by researchers (see chapter 3) in mind, this provides a good starting point to develop ML- and AI-based approaches for scheduling and to transfer them to industrial practice.

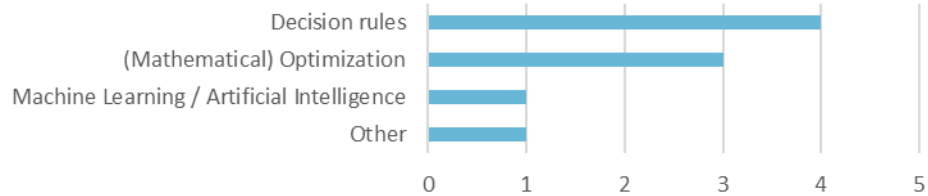

Figure 3. Methodological approaches applied for scheduling by the project partners from industry.

Beyond the criteria and methodological approaches, the process of scheduling can be executed manually or automated. The fact that the methodological approaches described above cannot be (completely) executed manually but require computational support corresponds with the feedback from the project partners that scheduling and dispatching are automated to a large extent. Apart from that, automated decisions can be manually overruled.

Beside scheduling decision making, the physical processes in the manufacturing system can be executed manually or automated as well. With regard to different wafer sizes, a higher degree of automation is reported for 300mm fabs than for 200mm fabs. Different processes like lot delivery and lot loading and unloading are reported to be executed both manually and automated in industry.

A digital twin represents another core component in the project AISSI. Therefore, the state-of-the-art regarding a digital twin at the project partners from industry with regard to purpose, scope and level of detail is inquired in the questionnaire. The feedback indicates that there is no unique comprehension of the functional range of a digital twin. While some partners define it as a state representation of a physical asset only, for others, it is a kind of simulation model which reproduces the manufacturing system with high fidelity. In general, it should cover aspects regarding men, machines, methods and materials and should allow to aggregate and investigate the underlying manufacturing system in different levels of detail. The project partners mention various aspects where a digital twin could provide support and insights:

- Simplified and streamlined KPI calculation
- Check requested production volume vs. fab capacity
- Time-based and usage-based scheduling of maintenance activities
- Forecast of WIP profile and production output
- Forecast of lot journey through the fab
- Load mix optimisation
- Operator resource planning

Another important aspect in semiconductor manufacturing systems – beside production scheduling – is the scheduling of maintenance activities. As different kinds of maintenance activities can be distinguished – planned, unplanned, preventive and predictive maintenance, where each has a different impact on maintenance scheduling – another question in the questionnaire focuses on the respective shares of these kinds of maintenance activities in the manufacturing systems of the project partners from industry.
Both planned and unplanned, preventive and predictive maintenance are reported to be applied in industry. The respective shares, however, differ heavily. While planned and unplanned maintenance is reported to occur in equal shares for one project partner, a considerably higher share is attached to planned maintenance by another project partner. Moreover, the feedback reveals differing comprehensions regarding the intersection of planned and unplanned maintenance on the one hand and preventive and predictive maintenance on the other hand. There is no consistent comprehension of whether these are mutually exclusive terms and activities or whether they coincide to some extent. This insight provides an important starting point for further discussions with the project partners. In the further course of the project, it is necessary to define a common comprehension of the different kinds of maintenance and their interrelation to provide a common base for the development efforts throughout the project.
Furthermore, the project partners provided some details regarding preventive maintenance. It can be based on:

- Time: Maintenance has to take place in regular intervals,
- Volume: Maintenance has to take place after a certain number of wafers or when some material is used up,
- Process parameters: Maintenance has to take place if observed process parameter(s) exceed some predefined threshold(s).

In particular, the feedback from the project partners indicates that maintenance is primarily based on decision rules instead of optimisation approaches or concepts from the domain of machine learning and artificial intelligence.
Finally, predictive maintenance is only touched upon in the feedback of the project partners and it is mentioned to be in a proof-of-concept phase. This establishes another aspect that can be addressed by the development efforts in the project.

In research, there are approaches that focus on either production scheduling or maintenance scheduling only and approaches that integrate production and maintenance scheduling (see chapter 3). As both production and maintenance take place on each machine, they depend on each other and have to share the limited time. For this reason, an integrated approach for production and maintenance scheduling seems reasonable from a theoretical point of view. To grasp the state-of-the-art in industry regarding this issue, a corresponding question is also contained in the questionnaire.
The feedback reveals that production planning and maintenance planning are exclusively executed separately. Maintenance activities are then manually integrated into the production schedule. Doing so, the expected WIP for the single machines is taken into account and maintenance activities are preferably planned for points in time when there is low WIP for a machine. Consequently, production is implicitly prioritised over maintenance.

While all planning approaches require some input data to base their decisions on, approaches from the domain of machine learning and artificial intelligence tend to have higher requirements regarding the amount of input data. Therefore, data availability and accessibility are essential for implementing planning approaches based on machine learning and artificial intelligence. Hence, the extent of data availability and accessibility at the project partners is inquired in the last question of the questionnaire.
The project partners report to have data, e.g. lot trace data, stored and available for use in a digital twin and other planning approaches. However, challenges might occur with regard to very detailed and specific data and appropriate consolidation and aggregation procedures of the data.

5 REFERENCED DOCUMENTS

The listing below provides an overview of documents and other sources of information that have been referenced in this document.

Short name	Full name
Questionnaire	Questionnaire Industrial SOTA

6 NOTES

6.1 Abbreviations

A list of used abbreviations.

Abbreviation	Meaning
ML	Machine learning
AI	Artificial intelligence
RL	Reinforcement learning
DRL	Deep reinforcement learning
MDSR	Multiple dynamic scheduling rules
FMS	Flexible manufacturing system
FAB	Semiconductor wafer fabrication
NN	Neural network
FNN	Fuzzy neural network
MCP	Multichip products
EHI	Equipment health indicator
PdM	Predictive maintenance
RUL	Remaining useful life
APC	Advanced process control
WIP	Work in process

6.2 Terminology

A list of used terminology.

Term	Explanation
SotA	State-of-the-Art

APPENDIX A. SEARCH STRING

Using the "advanced search" function in the Scopus database[129], we deployed the following search string for the literature research:

KEY (("Scheduling" OR "Maintenance") AND ("Semiconductor" OR "Wafer" OR "Fab")) AND (LIMIT-TO (PUBSTAGE,"final")) AND (LIMIT-TO (DOCTYPE,"ar")) AND (LIMIT-TO (SUBJAREA,"ENGI") OR LIMIT-TO (SUBJAREA,"COMP") OR LIMIT-TO (SUBJAREA,"MATH") OR LIMIT-TO (SUBJAREA,"DECI") OR LIMIT-TO (SUBJAREA,"BUSI") OR LIMIT-TO (SUBJAREA,"ECON")) AND (LIMIT-TO (LANGUAGE,"English")) AND (LIMIT-TO (SRCTYPE,"j"))

Additionally, we did only took into account publications from journals with a Journal Impact Factor[130] greater than 1 and which have been published between 2009 and 2021.

[129] https://www.scopus.com/
[130] https://impactfactorforjournal.com/