Scheduling Semiconductor Manufacturing Operations in Research and Development Environments

Valeria Borodin[1], Vincent Fischer[2], Agnès Roussy[3], Claude Yugma[3]
[1]*IMT Atlantique, LS2N-CNRS, La Chantrerie, 4, rue Alfred Kastler, Nantes cedex 3, F-44307, France*
valeria.borodin@imt-atlantique.fr

[2]*Technology Platforms Dept., Univ. Grenoble Alpes, CEA, LETI, 17 Rue des Martyrs 38054 Grenoble, Cedex 9, France*
Vincent.FISCHER@cea.fr

[3]*Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, F-13541 Gardanne, France*
{roussy, yugma}@emse.fr

2024 35th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC) | 979-8-3503-8455-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ASMC61125.2024.10545483

***Abstract*—This paper focuses on a scheduling problem encountered in shop floors of Research and Development (R&D) semiconductor manufacturing facilities. R&D facilities are characterized by a large product mix in very small quantities with unique/non-standard/varying processing routes, little process control of engineering experiments, dynamic prioritization of research activities, and pre-process checks. In contrast to typical scheduling problems found in semiconductor manufacturing systems, we provide and discuss the implications of factors of complexity (unknown parameters, evolving settings, R&D fab characteristics, etc.) on operations' scheduling specific to R&D environments. An existing dispatching rule-based heuristic, running in R&D settings, is challenged, investigated, and improved. Numerical experiments are conducted on a real-life instance and analyzed in terms of: (i) the sequence performance and quality, and (ii) the approximation accuracy of uncertain processing times and its impact on the decision performance.**

***Index Terms*—Semiconductor manufacturing, Research and Development facility, Scheduling problem, Optimization, Dispatching rule**

I. Context and motivations

Semiconductor manufacturing ranks among the largest, most competitive, highly complex, and capital-intensive industries in the world [1]. Efficient production control and operations management are vital under the cost pressure of building and maintaining semiconductor facilities[1], enabling: cycle time improvement, increased utilization of expensive machines, reduction of queueing time, acceleration research and innovation, throughput improvement, and on-time delivery [2].

Given the intricacies of semiconductor manufacturing processes, scheduling the full set of operations of wafer facilities (fabs in short) is generally a highly complex problem, including a rich set of constraints and optimization criteria [3, 4]. While many studies have been dedicated to smart operations management of production fabs in the related literature, very few studies are available on the improvement of the production performance of R&D facilities [5]. In this paper, we focus on scheduling operations in a Research and Development (R&D) semiconductor manufacturing facility. Semiconductor R&D systems are different from production fabs in various aspects [5, 6, 7, 8, 9]:

- *Scales of production:* Production is characterized by a great variety of products in very small quantities. Specific to R&D systems, the concept of a campaign can be used to launch production events.
- *Set of operations management instruments:* No bill of material is available to manage material procurement and usage operations. The concept of a product line is generally used to express and manage the processes' requirements.
- *Level of production control:* Production flows are subject to very little control. Engineering experiments can take many unpredictable trajectories leading to a variable number of process changes and inspections, lot holdings and releases, reworks, etc. In contrast to production fabs, no detailed information related to process routes (type of machine/recipe to use, etc.) is always available in R&D shop floors [9]. Little historical data compatible with approaches dedicated to characterizing new processes is available. In cases where process capabilities are unavailable, operation processing is outsourced, inducing thus additional variability.
- *Automation:* The deployment of automation is not so straightforward as in production fabs, where operating conditions are known beforehand and manufacturing schemes are predictable (i.e., "unknowns are known"). More complicated and highly customized automation models are necessary to be developed to bring added value to decision makers operating under "unknown unknowns", in conditions where their intervention is difficult to automate [9].

[1]https://www.fabtime.com/fabtime-volume-24.php#24.04

In the context of the automation of R&D fabs, this paper focuses on production scheduling in R&D semiconductor manufacturing based on an existing dispatching rule-based heuristic. Despite their weaknesses compared to scheduling methods using advanced optimization techniques (e.g., sub-optimality, hyper-parameter obsolescence), rule-based dispatching systems remain prevalent even on production shop floors due to the inherent complexity of scheduling problems in general, and specifically in wafer fabrication [3, 10, 11, 12].

In addition to challenging and improving an existing solution approach, this paper raises and aims to shed light to the following research questions when scheduling semiconductor manufacturing operations in R&D systems:

- How to appropriately deal with "(un)known unknowns" (processing times, changes in the predetermined sequence of operations, etc.) when scheduling operations in an R&D fab?
- What are the costs and benefits of the quality of the problem input data given the dynamic production settings?
- What is the appropriate balance between proactive and reactive prescriptions in such a time-varying production context?

The remainder of this paper is organized as follows. Section II introduces the problem under study, provides the dispatching rule applied to sequence the operations after a simple priority-based assignment phase, and presents several improvement directions. The numerical experiments conducted on a real-life instance are analyzed in Section III. Concluding remarks are provided in Section IV.

II. Problem Statement, Industrial Solution Approach, and Improvement Lines

a) Problem statement: Consider a set J of jobs (i.e., lots) and a set of known n_j operations $O_j = \{o_{1,j}, o_{2,j}, \dots, o_{n_j,j}\}$ (i.e., steps) associated with each job $j \in J$. In production fabs, the routes are fixed and known beforehand, while in R&D environments, they have a prospective character and may include multiple possible trajectories depending on process capabilities [9]. Operations belonging to the same job must be performed in the order specified by the route of the job.

b) Current solution approach: To schedule operations in its R&D centers, the practitioner partner of this study operates with a dispatching rule-based heuristic including two main phases:

1) **Assignment:** In R&D centers, no analytical specifications are available to indicate which machines are qualified and which recipes are adapted for a given job and operation [9]. Engineers typically rely on their expertise and knowledge of product requirements to select machines and recipes. Instead of explicitly considering individual resources, a set of resource bins M is defined. A resource bin corresponds to a combination of a priority compartment and an operation type. Each resource bin $\mu \in M$ has a limited and fixed capacity. Before the sequencing phase, operations are assigned to resource bins based on the priority of their jobs as predetermined by R&D specifications. Let $\mu(j)$ be the resource bin assigned to operations of jobs $j \in J$.
2) **Sequencing:** Once the assignment of jobs to resource bins is performed, jobs are sorted according to a compound coefficient calculated as follows:

$$c_j = v(\mu(j), \dots) \cdot \frac{\rho_j}{scale_\rho} \cdot \left(1 - \frac{\min\{0, \theta_j\}}{scale_\theta}\right) \quad (1)$$

where

- $v(\mu(j), \dots)$: denotes the adjusted speed required to meet the job deadline. This adjusted speed is clamped by the capacity of resource bin $\mu(j)$ and is a function of other parameters, including the remaining and total number of operations in the job route known at the time of dispatching, and the sign of the job slack time.
- ρ_j: represents the number of operations in the route of job j that remains to reach full job achievement, scaled by a predefined constant $scale_\rho$. Coefficient ρ_j gradually decreases from the first operation to the last one and pushes products to progress toward their last operations.
- θ_j: represents the scaled minimal remaining overtime (i.e., slack time) to release job j, scaled by a predefined constant $scale_\theta$. If the slack time is positive, the term associated with θ_j is constant and insensitive. In the case of overdue jobs, those with minimal slack time are prioritized for execution.

By virtue of formula (1), note that the dispatching rule-based heuristic aims to maximize the on-time delivery rate while overlooking the machine rate utilization. This is consistent with the ultimate goal of R&D, which is mainly oriented at expanding the range of process capabilities.

c) Lines of improvement: The main goal of this study is to investigate the performance of the current solution approach and to challenge it for further improvement purposes. To do this, three levers are considered in this paper:

- **Monitoring of the variability of Key Performance Indicators (KPIs):** With the expansion of automation in R&D fabs motivated, inter alia, by the fast chip development and demand soaring [9], the automated management of Work-In-Process (WIP) become critical to improve KPIs such as cycle time, throughput, and on-time delivery [11, 12, 13]. To gain insights related to Work-In-Process management, a detailed dashboard is provided.
- **Monitoring of the variability of processing times:** As a variable of coefficients $v(\mu(j), \dots)$ and θ_j, the processing time represents an important contributor in the calculation of the compounded coefficients c_j associated with jobs $j \in J$. In the current approach, processing times are considered deterministic and correspond to average values. To study the relevance of representing processing times via empirical average values, we extract

and analyze the distribution characteristics of empirical processing time data via a moment-based approximation as done in [14].

- **Analysis of the sequence quality provided by the dispatching rule-based heuristic:** Special attention is given to the sequence generated by the dispatching rule-based heuristic itself, aiming to gain a deeper understanding of the terms used in calculating the job coefficients and to identify the potential obsolescence of heuristic hyper-parameters.

Fig. 1. Work-In-Process (WIP) at March 14, 2023, and downstream processing of lots

Fig. 2. Monitoring of WIP and throughput per hour

III. Numerical Experiments

The numerical experiments have been conducted on an illustrative industrial instance provided by CEA-Leti, one of the three European Research and Technology Organizations in semiconductor manufacturing. This instance has been extracted on March 14, 2023, and includes 952 jobs (i.e., Work-In-Process lots) of two types, namely engineering (ENG) and standard (STD). Fig. 1 traces the processed downstream operations of WIP jobs. The lengths of jobs' routes vary between 1 to 220 operations and extend over large time horizons. Before sequencing, jobs are assigned to one among 5 resource bins, named according to the priority compartment: $M = \{\texttt{top20}, \texttt{high}, \texttt{standard}, \texttt{medium}, \texttt{without commitment}\}$.

a) Performance assessment: Fig. 1 illustrates the level of WIP over time and underscores the intractability of manual decision-making to manage increasing WIP levels. As observed in Fig. 2, the output rate increases as WIP levels rise, and it must be monitored closely to prevent the bottleneck of the R&D fab from full utilization (which, in turn, may lead to infinite WIP and cycle time).

Let us now turn our attention to cycle times over the job routing illustrated in Fig. 3. The cycle time is one of the main performance metrics in capacity planning [13], measured as the time from when a job is released into the production line to when it exits [15]. Without taking into account atypical extreme values, the variability of cycle times remains high and needs to be minimized to ensure steady production. As a first step in this sense, Fig. 4 provides the absolute errors, i.e., the differences between the true cycle times and those predicted by the dispatching rule-based heuristic over the job routing. No sufficient historical data are available to conclude about the prediction accuracy in the case of `without commitment` jobs. Particularly noticeable are the discrepancies in cycle times for `standard` jobs, that may subsequently poorly support capacity planning decisions. The dispatching rule-based heuristic tends to underestimate the processing times largely. To fix this issue, the next paragraph is dedicated to characterizing the empirical processing times.

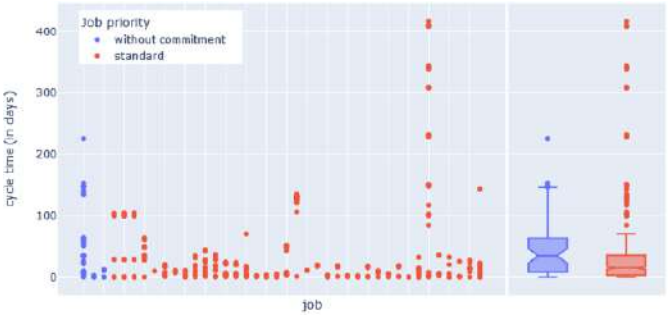

Fig. 3. Cycle time (expressed in days)

b) Processing times: The current dispatching rule operates with average processing times. Fig. 6 confronts the true processing times and those planned by the dispatching rule, and illustrates the propagation of errors over time. The range of error magnitude is particularly wide for operations

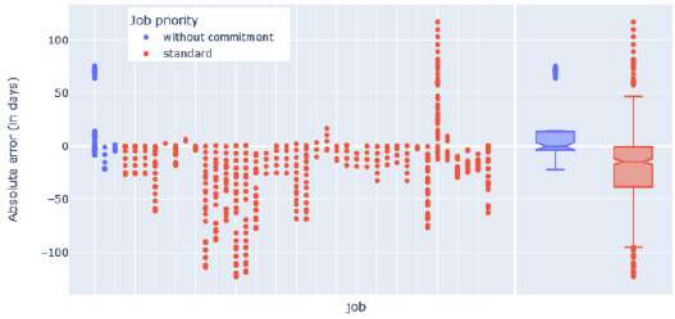

Fig. 4. Cycle time (expressed in days): *True - Planned*

corresponding to `standard` jobs (68% of total number of observations) and `without commitment` jobs (6% of total number of observations), including a large number of values unusually far from the mean value. Except for operations of `medium` priority, the dispatching rule tends to underestimate processing times, in particular, those of `without commitment` operations.

To calculate the empirical probability distributions associated with observed processing times, consider a cloud of N points $\{p_i\}_{i=1}^{N}$ drawn independently from an unknown probability measure ν on $\mathbb{R}$ with compact support S corresponding to a random variable ξ. The main-mass, tails, and shape approximation of ν can be derived from the sequence of moments associated with $\{p_i\}_{i=1}^{N}$ based on the Christoffel function (see e.g., [14, 16]).

Let us analyze the empirical probability distributions associated with observed processing times illustrated in Fig. 5. Probability distributions are asymmetric and exhibit very long right tails. The long-tail processing times may together drastically overestimate the true release dates and have an even greater impact than atypical particularly long operations.

TABLE I
APPROXIMATION OF PROCESSING TIMES AND PREDICTION OF RELEASE DATES r_j AND CYCLE TIMES CL_j OVER JOB ROUTING: *Accuracy improvement (i.e., error minimization) compared to the current approximation*.

Approximation approach	**RMSE** (r_j)	**RMSE** (CL_j)
mean$\{p_i, i = \overline{1,N} : p_i \leq Q_\xi(0.95)\}$	-1%	21%
mean$\{p_i, i = \overline{1,N} : p_i \leq Q_\xi(0.97)\}$	8%	22%
mean$\{p_i, i = \overline{1,N} : p_i \leq Q_\xi(0.99)\}$	16%	21%
min$\{x \in \mathbb{R} : F_\xi(x) \geq q\}$	**22%**	**30%**

As previously explained, a large portion of atypical values of processing times are normal for R&D environments. Let us investigate to what extent the variability of processing times can be characterized. To do this, we applied a quantile-based outlier detection approach. Q_ξ denotes the quantile of random variable ξ. While using the mean as a summary operator, the quality of prediction of release dates (denoted by $r_j, \forall j \in J$) can be improved by 16% by removing the most 1% extreme values, measured in terms of Root Mean Squared Error (RMSE). However, when removing the most 5% extreme values, the prediction accuracy suffers.

Given the non-linear behavior of the prediction accuracy based on quantile-based outlier detection and the mean value as a summary operator, we replaced the mean values by resource bin-dependent quantiles $q \in [0.64, 0.95]$. Quantile q is chosen depending on the length of right tails and associated probability weights. This improves by 22% the quality of prediction of operations' release times. As illustrated in Fig. 7, using quantile instead of mean to summarize empirical values enables us: **(i)** to reduce noticeably the variability of error distributions for `standard`, `medium` and `without commitment` processing times, and **(ii)** to reduce some snowball effects of errors in the case of `standard` and `without commitment` processing times. In terms of RMSE, the quantile-based summary improves the release date accuracy by 22% compared to the current mean-based approximation of processing times.

As an immediate positive outcome, the improvement of prediction quality of processing times enables us to improve the prediction quality of cycle times (denoted by CL_j) by 30%, as provided in the third column of Tab. I. Further investigation is necessary to determine the acceptable margin of error for processing times without affecting processing time-dependent decisions.

c) Sequence quality: Fig. 8 highlights the relationship between the job positions in the sequence provided by the dispatching rule-based heuristic and their slack times. The size of the points in Fig. 8 and Fig. 9 is proportional to the total number of operations in the job route. Roughly, for negative slack times, the more the slack time the greater the position in the sequence. A significant drawback, highlighted within the gray box in Fig. 8, is that jobs with a large number of operations are prioritized over those with a small negative slack time or those nearing completion, which consequently impacts the throughput. As WIP levels continue to rise in R&D centers, conducting a comprehensive analysis of the cycle time *versus* throughput curve becomes imperative due to its crucial contribution to efficient capacity management [12].

One of the primary drawbacks of dispatching rules is their reliance on hyper-parameters, which evolve over long time horizons [3]. Fig. 9 illustrates the importance of updating deprecated hyper-parameters. For example, this update enabled long jobs, highlighted within the gray box, to start before jobs with comparable total numbers of operations but with larger slack time.

IV. CONCLUSIONS AND PERSPECTIVES

This paper focuses on shop floor scheduling in Research and Development (R&D) semiconductor manufacturing facilities. After discussing the particular features of operations management in R&D environments, an industrial dispatching rule-based heuristic is presented and several improvement directions are proposed and evaluated on a real-life instance.

Through the prism of an illustrative real-life instance and as a starting point, we highlighted several lines of improvement of a scheduling solution approach used in practice. Some results have been shown in this paper but further research

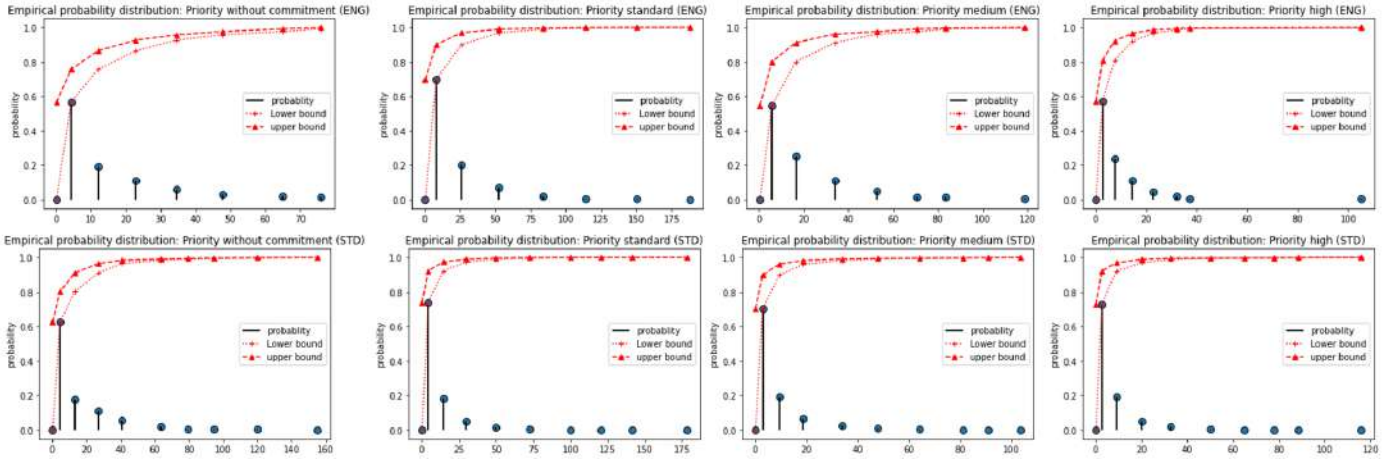

Fig. 5. Processing times per priority compartment and operation type (expressed in days): *Empirical probability distribution*

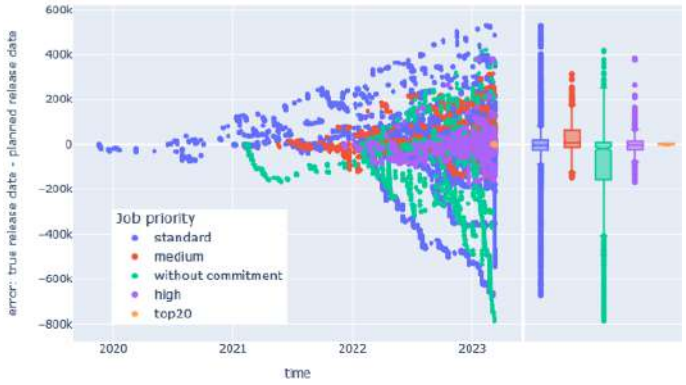

Fig. 6. Release dates per operation: *True versus Planned*. Points correspond to the operation release dates.

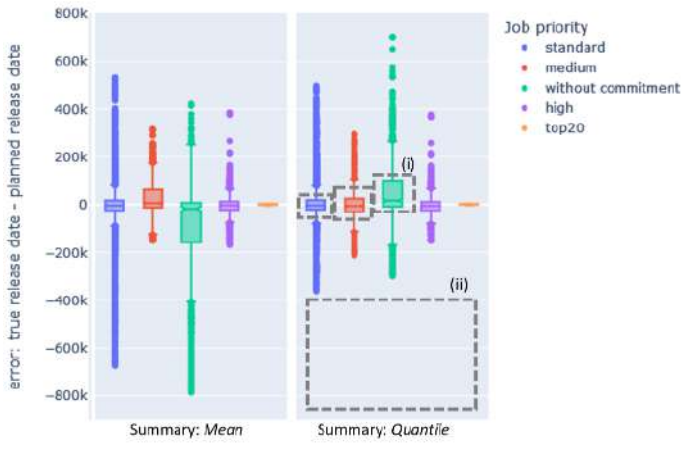

Fig. 7. Approximation of processing times (expressed in minutes): *Mean summary versus Quantile summary*

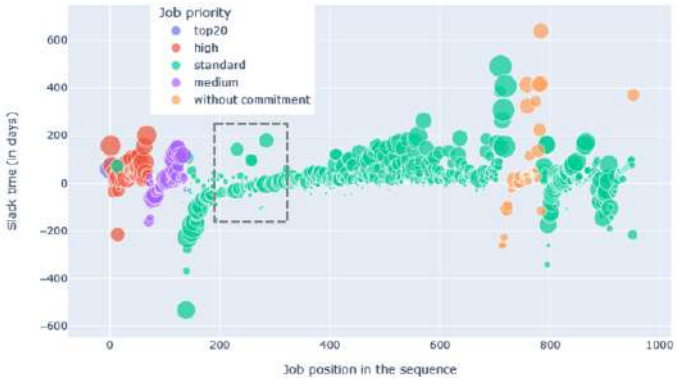

Fig. 8. Focus on the sequence provided by the dispatching rule-based heuristic.

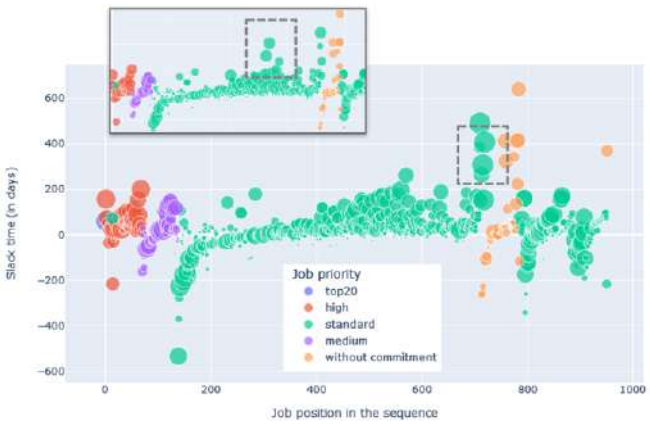

Fig. 9. On the importance of the coefficient updating of the dispatching rule-based scheduling heuristic.

remains to be done, such as: **(i)** to characterize explicitly the processing of lots by machines in order to improve the quality of decisions, by levering the available knowledge about product routes at the time of operations scheduling, **(ii)** to study the relevance of conventional KPIs (commonly used in manufacturing systems) in R&D environments, and **(iii)** to measure the production efficiency in R&D environments characterized by a low level of production control conditioned by technological reasons. Simulation and combined simulation-optimization approaches [17] effectively tackle these questions (see e.g., [11]), paving the way for subsequent development of more sophisticated optimization solvers. Unlike industrial systems, in R&D environments, decision makers typically operate with relatively limited WIP levels while dealing with longer processing times. The pronounced snowball effects of errors become increasingly apparent and must be meticulously controlled as WIP levels expand.

ACKNOWLEDGMENTS

The present work was partly conducted in the framework of the project ACCURATE (project ID: 101138269), supported by the Horizon Europe Framework Programme, under grant agreement number 101138269, HORIZON-CL4-2023-TWIN-TRANSITION-01-07.

REFERENCES

[1] L. Mönch, J. W. Fowler, S. Dauzère-Pérès, S. J. Mason, and O. Rose, "A survey of problems, solution techniques, and future challenges in scheduling semiconductor manufacturing operations," *Journal of scheduling*, vol. 14, pp. 583–599, 2011.

[2] S. Elaoud, D. Xenos, and T. O'Donnell, "Deploying an integrated framework of fab-wide and toolset schedulers to improve performance in a real large-scale fab," in *2023 34th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*. IEEE, 2023, pp. 1–6.

[3] S. C. Sarin, A. Varadarajan, and L. Wang, "A survey of dispatching rules for operational control in wafer fabrication," *Production Planning and Control*, vol. 22, no. 1, pp. 4–24, 2011.

[4] M. Flores-Gómez, V. Borodin, and S. Dauzère-Pérès, "Maximizing the service level on the makespan in the stochastic flexible job-shop scheduling problem," *Computers & Operations Research*, vol. 157, p. 106237, 2023.

[5] V. Ramamurthi, M. E. Kuhl, and K. D. Hirschman, "Analysis of production control methods for semiconductor research and development fabs using simulation," in *Proceedings of the Winter Simulation Conference, 2005.* IEEE, 2005, pp. 9–pp.

[6] D.-Y. Liao, S.-C. Chang, K.-W. Pei, and C.-M. Chang, "Daily scheduling for R&D semiconductor fabrication," *IEEE transactions on semiconductor manufacturing*, vol. 9, no. 4, pp. 550–561, 1996.

[7] Y.-H. Kim, J.-H. Lee, and D.-S. Sun, "The operational optimization of semiconductor research and development fabs by fab-wide scheduling," *The Transactions of the Korean Institute of Electrical Engineers*, vol. 57, no. 4, pp. 692–699, 2008.

[8] Y. Chen, D. Kayarat, and J. Binford, "Multi objective optimization of process R&D goals in automated semiconductor manufacturing systems," in *IIE Annual Conference. Proceedings*. Institute of Industrial and Systems Engineers (IISE), 2017, pp. 524–529.

[9] V. Fischer, O. Landré, and M. Duranton, "Automation in r&d: complying with contradictory constraints of seemingly incompatible world," in *2023 34th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*. IEEE, 2023, pp. 1–4.

[10] A. Lima, V. Borodin, S. Dauzère-Pérès, and P. Vialletelle, "Analyzing different dispatching policies for probability estimation in time constraint tunnels in semiconductor manufacturing," in *2017 Winter Simulation Conference (WSC)*. IEEE, 2017, pp. 3543–3554.

[11] F. Barhebwa-Mushamuka, "Novel optimization approaches for global fab scheduling in semiconductor manufacturing," *Doctoral dissertation, Université de Lyon*, 2020.

[12] J. W. Fowler, S. Park, G. T. MacKulak, and D. L. Shunk, "Efficient cycle time-throughput curve generation using a fixed sample size procedure," *International Journal of Production Research*, vol. 39, no. 12, pp. 2595–2613, 2001.

[13] J. Robinson, J. Fowler, and E. Neacy, "Capacity loss factors in semiconductor manufacturing," *FabTime Inc. http://www. fabtime. com/abs_CapPlan. shtml*, 2003.

[14] V. Borodin, N. G. Chembu, A. Khemiri, J. v. Heugten, and C. Yugma, "The impact of data fragmentation on processing time modeling in semiconductor manufacturing : *topic: Ie: Industrial engineering," in *2023 34th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 2023, pp. 1–6.

[15] W. J. Hopp and M. L. Spearman, *Factory physics*. Waveland Press, 2011.

[16] P. Gavriliadis and G. Athanassoulis, "Moment information for probability distributions, without solving the moment problem, ii: Main-mass, tails and shape approximation," *Journal of computational and applied mathematics*, vol. 229, no. 1, pp. 7–15, 2009.

[17] V. Borodin, J. Bourtembourg, F. Hnaien, and N. Labadie, "COTS software integration for simulation optimization coupling: case of ARENA and CPLEX products," *International Journal of Modelling and Simulation*, vol. 39, no. 3, pp. 178–189, 2019.