Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

ADVANCED PRODUCTION SCHEDULING IN A SEAGATE TECHNOLOGY WAFER FAB

Georgios M. Kopanos
Dennis Xenos
Slava Andreev

Flexciton Ltd
145 City Road
London, EC1V 1AZ, UK

Tina O'Donnell
Sharon Feely

Seagate Technology
1 Disk Drive
Londonderry, BT48 0LY, UK

ABSTRACT

This work focuses on the highly complicated scheduling problem in wafer fabs. We first provide insights on the broader impact of high quality scheduling decisions in semiconductor industries, and then we discuss traditional heuristic-based scheduling practices versus our mathematical optimization approach. The comparison between our hybrid-optimization scheduling tool against the current simulation scheduler at Seagate shows significant improvements in performance through a benchmark study that involves nine historical datasets of the metrology toolsets in Seagate Springtown. On average, our schedules report: a significant reduction of more than 43% in cycle times for high-priority wafers, a reduction of about 9% in total cycle times, and a 7% increase in throughput; compared to SimModel schedules. The large improvements on these schedule metrics are mainly due to the more balanced allocation of wafers to machines, and the better batching formations that exploit effectively wafers' release dates and priorities.

1 INTRODUCTION

Seagate is a major semiconductor industry and a world leader in data storage technology, with more than 40% share of the global Hard Disk Drive (HDD) market. The Springtown facility in Northern Ireland produces around 25% of the total global demand for recording heads, the critical sensor in a HDD. As one of only five comparable wafer fabrication sites in the world and the only site in Europe, it is ideally placed to demonstrate commercial viability of advanced analytics platforms. Seagate ships about 90 Exabytes of storage per quarter to support the 90% of the world's data that is currently stored on HDDs. Thus, the Seagate Springtown site is positioned at the heart of a $25billion worldwide HDD industry. Since 1994 the site has spent $4billion for its growth, involving 1400 employees, academic collaborations and R&D; and still continues developing as a wafer fab site, investing in areas of technology growth.

In 2019, global semiconductor industry sales totaled $412.1billion (SIA 2020) and given today's technological advancement, the industry is set to continue its growth. With the growth in demand, there is more pressure put on the industry to decrease production costs and increase the time-to-market ratio. Historically, cost reductions have been found from chip size reduction and increasing wafer size, as well as improving line yield. Cost reduction from these methods is beginning to stagnate, and improved operational processes are becoming more necessary (Mönch et al. 2000). As a result, efficient production control and advanced scheduling are vital in modern wafer fabs as a method of yielding higher production efficiency, resulting in reduced cycle time, higher throughput and reduced capital spend.

The challenge is that wafer fabrication is one of the most complicated manufacturing processes in the modern world. The thousands of process steps and complex constraints, such as batch processing, competing limited resources, and re-entrant process flows, unsurprisingly result in bottlenecks and makes wafer fab scheduling an extremely challenging task. With the cost of a single new 300mm fab now exceeding $1billion and with some tools costing in excess of $40m, fixed costs are significant and

978-1-7281-9499-8/20/$31.00 ©2020 IEEE

demand high fab utilization. The benefit of high quality scheduling is huge to factory efficiency since it enables higher utilization of expensive toolsets (e.g., photolithography and etch), reduces cycle times and ensures on-time delivery. To highlight the impact of scheduling decisions, Samsung saw that an 18.5% drop in average cycle time for DRAM devices led to $1billion extra revenues over 5 years, counteracting the economic effect of the declining DRAM selling prices back in 2001 (Leachman et al 2002).

2 DISCUSSION ON SCHEDULING APPROACHES IN WAFER FABS

2.1 Dispatching Process

Currently, the standard practice in semiconductor wafer fabs is to use dispatching systems to control Work-In-Progress (WIP) processing decisions throughout the whole fab. When a new wafer arrives at a tool, it must queue until a tool becomes available. Dispatchers examine all WIP available for a tool and determine which wafers to process next, using a complicated series of algorithmic rules that typically involve: recipe-based groupings, event-based restrictions, durables utilization, demand schedules, tool utilization, or priority ranked wafers. Dispatch systems execute quickly and are highly effective as production control tools. However, dispatching rules are fundamentally flawed since decisions are locally derived with no capability of understanding the broader factory behavior both over time and resource. This practice is disadvantageous, and the local decisions do not identify any opening for strategic action (i.e. actions that may locally appear to be optimal but have limited or no wider overall benefits). Depending on the rule in use, dispatch will attempt to force WIP through a tool to meet the local objective (e.g. priority wafers). However, this may have a negative impact downstream and on key toolsets.

Seagate uses Applied Material's Real Time Dispatcher (RTD) product which reacts fast to changes in the environment due to the output of event-based workflows being embedded within dispatch rules. Many rules are controlled via an external user parameter table to minimize maintenance. Engineers can configure the dispatch rules to suit changing production goals or expedite priority lots through the fab.

2.2 Scheduling via Heuristic Algorithms

Heuristics are popular and practical fab scheduling approaches that use techniques designed to provide a fast and feasible solution to a specific scheduling problem. A heuristic is a rules-based method that uses approximate or empirical approaches to reduce the computational burden and search space. Heuristic methods have been widely adopted in scheduling wafer fabs because the complexity of the problem has demanded simpler / approximate solutions. Heuristics are extremely fast to run, which is a key benefit of these systems. On the contrary, heuristics are fundamentally flawed when it comes to decision quality as they cannot optimize the full-scale scheduling problem. Heuristics rely on observed historical patterns and industrial engineers' knowledge to configure, which requires upkeep by experts. This continuous upkeep makes maintenance of a single solution difficult, expensive and usually highly custom to each individual fab and toolset setup. As a result, transferring a heuristic scheduler between toolsets, areas or to other fabs may not be easy and require a lot of implementation effort. Without the ability to optimize schedules over a future horizon, poor decisions are made that often results in dynamic tool bottlenecks.

2.3 Scheduling via Exact Optimization

In contrast to heuristics, exact optimization is a declarative, algorithmic problem-solving method that is capable of reliably delivering the best solution to scheduling problems (Kopanos and Puigjaner 2019). A mathematical optimization model comprises relevant objectives (production goals), variables (decisions in user control) and constraints (feasible region) to generate a feasible solution that optimizes the desired objectives. In principle, the user can specify the objective function of interest at the time. This can be a single objective or a combination of weighted objectives, such as maximize throughput whilst keeping low cycle times. Table 1 shows a comparison of the main features of scheduling approaches based on heuristics or exact optimization based on Mixed Integer Linear Programming (MILP).

Table 1: Main features of scheduling approaches based on heuristics vs. exact optimization.

Feature	**Heuristics**	**Exact Optimization**
Quick	**YES:** They reduce the solution search space by using a complicate series of empirical rules, generating feasible schedules quickly; making them highly desirable in the fab environment. Often this solution speed results in major quality sacrifices.	**NO:** It looks at all the potential scheduling options available for the decision problem. In general, MILP uses branch and bound methods to search for global-optimality. This is computationally intensive, and even with powerful computers, scheduling may take hours or days to solve a small-scale problem.
Practical	**YES:** Engineers who build and configure the heuristics can write code to constrain (tune-in) the heuristic. Almost any constraint in the fab could be modelled; however an excessive number of intricately configured constraints could over-restrict the search, and drive it in areas of very suboptimal decisions.	**NO:** It may not be possible to account for all constraints due either to the nature of the constraint or other computational burdens. As a result, mathematically optimal schedules may be infeasible in practice; especially due to the dynamic nature of wafer fabs. Schedules need to change frequently to be up-to-date, but MILP can be rather slow as discussed above.
High Quality	**NO:** A heuristic simply finds a feasible schedule, it does not have the capability of exploring all possible options to find the best solution. As a result, its solution quality depends on the configuration and often produces suboptimal solutions. It is very hard to measure how suboptimal a heuristic schedule is; unreliable behavior.	**YES:** Whilst optimization is slow to solve the full problem, the core competency of the final solution is that schedules found are either global or near-optimal in solution quality. This means that the very best solution out of the extensive range of scheduling options has been found. Also, it has reliable performance when switching between different objectives.
Easy to transfer to other fabs	**NO:** Heuristics are often tailor-made for a single fab setup, fab areas or single tools. The logic could require major changes or full rebuild if transferred to another fab.	**YES:** A model can be treated as a digital-twin of the fab, allowing it to be easily transferred from one fab to another by configuring simply the constraints, resources and objectives.

2.4 Scheduling Approaches Overview

A great amount of research on scheduling has been conducted in the last decades. Some excellent reviews on batch scheduling that relates to semiconductor manufacturing can be found in Potts and Kovalyov (2000) and Mathirajan and Sivakumar (2000). Because wafer fab scheduling problems are NP-hard, exact optimization scheduling approaches have not been very popular (Pfund et al. 2006). And, research efforts usually focus on developing approaches based on approximate methods, which typically involve discrete-event simulation, heuristic-based decomposition schemes (e.g. Ovacik and Uzsoy 1997, and Yurtsever et al. 2009), and metaheuristics based on simulated annealing, genetic algorithms, tabu search, variable neighborhood search, or greedy randomized adaptive search procedures (Sörensen 2015). Combinations of these methods could be used in the industrial practice. For instance, discrete-event simulation is usually combined with dispatching rules to generate better schedules. Since computing power is increasing quickly and commercial optimization solvers are becoming more efficient, exact optimization methods based on MILP could play a major role in providing high-quality scheduling decisions in wafer fabs.

3 ADVANCED HYBRID-OPTIMIZATION SCHEDULING APPROACH

It is evident from the discussion above that there is a lack of scheduling decision support tools that bridge the gap between heuristics and exact optimization approaches. Our proposed scheduling approach works towards closing this gap by developing a commercial hybrid-optimization scheduling tool primarily based on MILP models employed within efficient decomposition schemes and assisted by smart heuristics for solving very complex wafer fab scheduling problems. The ultimate goal of our scheduling tool is to be able to deliver reliably high quality schedules with fast solution times.

3.1 Description

As mentioned before, a single large-scale MILP scheduling model -although it is able to capture explicitly new constraints and deal with changing optimization objectives- is not suitable for the highly dynamic wafer fab production environment, because solving a full-scale MILP is very computationally expensive. For this reason, the proposed hybrid approach blends exact optimization and heuristic search methods generating an advanced scheduling tool. More specifically, we have developed and implemented advanced exact and approximate decomposition techniques to the original full MILP scheduling model. These smart decomposition techniques are able to reduce the problem complexity significantly, yet retain near-optimality by dissecting the full-scheduling problem into a series of smaller sub-problems. These sub-problems can be solved very fast with MILP or heuristics alternatives, and once solved, combined back together to generate a complete solution to the original scheduling problem (Figure 1).

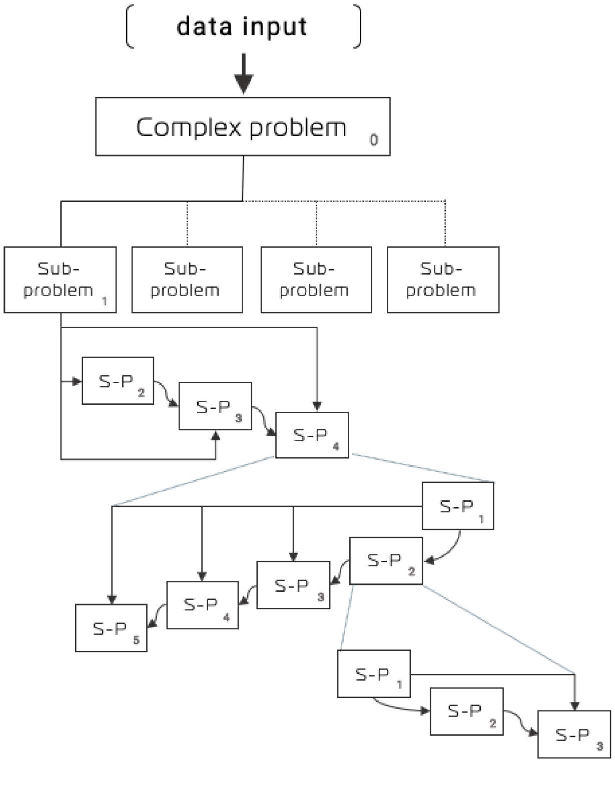

Figure 1: High-level schematic representation of the proposed scheduling approach.

One form of decomposition is to identify only the key toolsets which influence the overall throughput and then schedule these accordingly (Kim and Li 2006). By focusing only on bottleneck tools, the optimizer can focus on scheduling between the bottleneck toolsets and balance the WIP flow between them. This technique can significantly reduce the scheduling search space and minimize the required solution time, simultaneously maximizing utilization on key toolsets. Internal studies have shown that in many fabs, up to 80% of the total aggregated wafer wait time (bottlenecks) can be attributed to only 20% of the tools. Thus, by focusing on the flow of these bottleneck toolsets, you can solve a large proportion of the cycle time or throughput problems for the entire fab. The hybrid-optimization scheduling tool developed is superior in performance to pure heuristics, or exact optimization. The main features of our approach along with a brief discussion are summarized in Table 2.

Table 2: Salient features of the proposed hybrid-optimization scheduling technology.

Feature	**Discussion & Comments**
Quick	**YES:** By utilizing decomposition and other performance enhancements, a hybrid-approach can solve very complex scheduling problems in under 11 minutes. This enables the solution to function well in a dynamic fab environment and can update when there are changes on the factory floor.
Realistic Practical	**YES:** With hybrid-optimization, the model can now be easily constrained to ensure fully feasible schedules. All constraints can be added to the optimization model to ensure a true representation of all activity and limitations in the fab. The quick solution time ensures any dynamic adjustments required in the fab can be captured, ensuring consistently feasible and optimal schedules.
High Quality	**YES:** The core module used to build each schedule is based on MILP. This ensures high-quality solutions. Due to our multi-stage decomposition approach that in some parts use heuristics rules, we cannot ensure the global optimality of the end solutions; but our overall approach ensures that our solutions are near the theoretical optimal. A high quality solution is selected that best optimizes the chosen objective function.
Easy to transfer to other fabs	**YES:** The hybrid-optimization approach enables configuration of constraints and parameters, without extensive new code having to be re-written or designed (as with a heuristic scheduler). Constraints can be added or removed, with a single line of code. The significant benefit of this approach is that the scheduling solution can be scaled to multiple fabs, even when they have different production characteristics.

3.2 Interaction with Simulation Tools

Our scheduling tool uses key information derived from the simulation tool applied in the wafer fab, and optimizes further the simulated scheduling decisions. Current WIP, expected wafers arrivals, release dates, priority weights for wafers, batching restrictions and wafer-to-machine allocation preferences are some examples of simulation output data that are used as input data in our scheduling tool. Then, the scheduling tool first enables data-driven heuristics-based steps to generate quickly a feasible schedule (constructive step), and activates advanced iterative MILP-based decomposition layers (improvement step) aiming at getting the best possible schedule within reasonable time, typically a few minutes.

4 BENCHMARK STUDIES

In this section, we provide a description of our benchmark studies along with a discussion on the toolsets and datasets selected, followed by a definition of the resulting scheduling problem in terms of operational constraints, assumptions, main decisions and optimization targets.

4.1 Description

In this stage of the project, the focus has been on scheduling metrology toolsets where quality control and various measurement processes take place. Metrology tools are single load-port batching machines, i.e., several wafers can be processed sequentially in a single batch on the same machine. And, they could operate in a multi-recipe batch mode, i.e., wafers that use different recipe could be part of the same batch. As a result, the processing time of a batch depends on the recipes that are present in the batch as well as the number of wafers that use these recipes. A particular characteristic of the metrology tools is the low processing times in comparison to those in other tools. This means that under a given scheduling horizon the number of wafers that could have been processed in metrology tools is expected to be much higher than that of other tools. All the above characteristics of the metrology toolsets result in highly challenging scheduling problems which are ideal for benchmark studies.

In this study, a 3-hour scheduling horizon has been considered for the offline scheduling of up to 12 metrology tools in the Seagate Springtown facility. Nine historical datasets have been selected based on the number of wafers available for scheduling; either waiting on a storage rack or expected new arrivals in the near future. Namely, there are three groups of datasets with respect to the number of available wafers: Low, Normal and High. Here, all scheduling parameters are assumed deterministic.

For each dataset, the schedules derived by the proposed scheduling approach are compared with the schedules generated by SimModel; a simulation tool that uses the same dispatch rules as in real shop floor operations. The main schedule quality metrics are the average cycle time per high priority class wafers (i.e., P1-P3) along with the total weighted cycle time. To ensure an appropriate comparison, Seagate has implemented automatic validation tests that check that the generated schedules do follow all operational constraints. All schedules derived by the new scheduling tool have passed these schedule feasibility tests.

4.2 Problem Statement

This work focuses on the optimal single-stage scheduling of multi-recipe batching machines; such as metrology tools are. In brief, the resulting scheduling problem can be defined in terms of the next items:

- A set of available wafers with given release dates and priority weights.
- A set of unrelated machines that operate in parallel. These machines are batching machines that can process a number of wafers at the same time (aka batch size) by forming a batch. Minimum and maximum batch sizes per machine are known. Also, each machine has a given setup and teardown and unload time before and after the processing of a batch, respectively.
- A set of available recipes mapped to each machine and wafer through which the mapping of available machines per wafer can be derived. Fixed and variable processing times are known for each mapping of machine and recipe.
- A set of incompatible types of wafers that cannot belong to the same batch, i.e., wafers of type A cannot batch together with wafers of type B.
- Machines operate in a multi-recipe batching mode, i.e., multiple recipes could be present in the same batch. In other words, wafers that require different recipes could be batched together.
- A set of in-progress wafers/batches with given machine allocation and completion times that affects the ready times of the associated machines.
- A set of planned machine downtime periods with known start time and duration.

4.2.1 Assumptions

The major assumptions of the underlying scheduling problem are listed below:

- All scheduling data are assumed to be deterministic.
- A single processing step is considered per available wafer.

- Batch setup and teardown times do not depend on recipes or wafers.
- No preemption is allowed, i.e., once a batch starts, it cannot be interrupted.

4.2.2 Constraints

In this part, we provide a descriptive list of the main types of constraints that the proposed scheduling approach considers in order to ensure the feasibility of the resulting scheduling solution.

- *Wafer allocation to machines*.
  All available wafers must be scheduled, i.e., each wafer should be assigned exactly to one batch of its candidate machines.

- *Wafer timing*.
  A wafer cannot start processing earlier than its given release date, i.e., it cannot belong to a batch that has a start time earlier than the release date of the wafer.

- *Wafers timing assigned to the same batch*.
  All wafers that are part of the same batch should enter and exit the machine -where the batch has been allocated to- at the same time. Wafers cannot be removed or added in a processing batch.

- *Batch formation – allocation of wafers to batches.*
  The number of wafers assigned to a batch should not violate the given minimum and maximum batch sizes of the machine that the batch belongs to. Also, wafers that belong to incompatible wafer types should not be part of the same batch.

- *Batch timing*.
  The setup of a batch cannot start earlier than the given ready time of the machine that belongs to, or the latest release date of the wafers that are part of the batch.

- *Batches timing on the same machine.*
  A batch can be loaded to a machine only after the previous batch has been unloaded, i.e. parallel processing of batches on the same machine is not allowed.

- *Batch setup and teardown times.*
  Each machine requires a given setup time before start processing a batch, and a given teardown time after the end of processing a batch. No other operations take place in parallel in a machine that is under setup or teardown. Setup, processing and teardown of a batch cannot overlap.

- *Batch processing times.*
  The processing time of a batch consists of a fixed and a variable part that are a function of the allocated wafers, active recipes (i.e., recipes required for processing the allocated wafers) and the assigned machine of the batch. Although fixed and variable processing times data are known for each mapping of machine and recipe, active recipes are part of the optimization decisions. The fact that batch processing times are optimization variables increases significantly the complexity of the scheduling problem, as alternative batching formations could result in very different batch processing times that subsequently may affect considerably the optimization goal of interest.

- *Machine downtime.*
  A machine should not process any wafer during its given planned downtime time windows. Setup and teardown operations during machine downtimes are forbidden as well.

All the above constraints have been expressed mathematically creating optimization models that define the feasible region of the scheduling problem under study. To do so, we have introduced a set of optimization variables that are described briefly in the next paragraph.

4.2.3 Decisions & Optimization Goal

The key scheduling decisions, which are modeled as binary optimization variables, could be summarized into the following main categories:

- *Batching; the grouping of wafers into batches.*
  The decision of how many and which wafers should form a collective batch with one another substantially impacts their completion times due to the preemption behavior of processing batches of wafers. That is, no wafer may leave the tool until all wafers in the batch have completed processing. Some wafers may not be allowed to be processed at the same time as another wafer so these operational constraints around batching make this a non-trivial decision. The number of wafers to include in a batch is also of importance due to the batch size limitations of the machines.

- *Assignment; the allocation of batches onto machines*.
  A batch can be allocated to a machine if and only if all the wafers of the batch could be processed in this machine. Once a batch is allocated to a machine, the active recipes of the batch are determined, and then the processing time of the batch can be estimated. The processing times can vary significantly from machine to machine, because machines have different processing rates.

- *Sequencing and timing; the ordering of batches within each machine.*
  Once the batches with their associated wafers have been allocated to the machines, the ordering of these batches dictates which wafers will be processed in the machine first. This is often dictated by the release dates and the priority weights data of the wafers in the batch.

Note that fundamentally, the aforementioned decisions are all tightly coupled; it is not possible to influence any individual decision in isolation without compromising the overall solution quality. The proposed scheduling platform optimizes all these decisions in an integrated fashion while respecting all operational constraints. In most alternative approaches these decision layers are considered typically in a hierarchical or other forward-selection manner resulting in very poor solutions.

In this study, all the above decisions aim at minimizing the total weighted cycle time of the available wafers. The cycle time of a wafer step is the sum of its idle waiting time and the time taken to setup, process and teardown the batch to which it has been assigned to. The minimization of the total weighted cycle time aims at reducing the cycle times of the wafers giving greater importance to the cycle times of higher-priority wafers.

5 RESULTS

In this section, we compare the schedules generated by the Seagate SimModel (Sim) against the schedule produced by our new scheduling platform (Flex) across the nine historical datasets described before. The new scheduling platform was limited to 11 minutes of execution time. The main schedule metrics of interest in this benchmark study are the following ones:

- *Average cycle times for high-priority wafers (i.e., priority classes P1, P2, and P3).*
  Low cycle times are strongly preferred for high-priority wafers. The main focus is on wafers that belong to these three most important priority classes.

- *Average cycle time for all wafers.*
  The average among the times taken for all wafers to complete, from the moment they become available (i.e., wafer release date) to the point that they are unloaded from the allocated machine.

- *Throughput.*
  It has been defined as the completion rate of wafers within the scheduling horizon. In general, the bigger this number the better, since we can improve asset utilization and schedule more wafers within the same timeframe.

5.1 Aggregated Results

For the nine benchmark datasets, Table 3 provides a comparison of the above schedule metrics for each schedule derived by our scheduling platform having as a reference the related Sim schedule. More specifically, the throughput column reports the increase in throughput while the next four cycle times (CT) columns show the decrease in CT compared to the associated Sim schedule. According to Table 3, our schedules exhibit significantly better metrics across all datasets. In particular, our schedules were able to complete/process, on average, 7% more wafers while improving on both high-priority and average CTs by a considerable margin. In fact, the decrease in high priority CTs is substantial at over 43% while at the same time a decrease of around 9% on average CT for all wafers is reported. This clearly demonstrates the benefits of using a priority-weighted objective function in our model, which allows us to successfully balance the scheduling trade-offs that arise without undermining lower priority wafers.

Table 3: Comparison between Flex and Sim schedule metrics across 9 benchmark datasets.

Dataset	Throughput	P1 CT	P2 CT	P3 CT	Average CT
Low-1	13.94%	27.38%	44.88%	39.21%	7.64%
Low-2	8.84%	58.76%	26.59%	64.89%	9.44%
Low-3	5.22%	43.47%	37.44%	49.08%	15.99%
Norm-1	6.37%	36.22%	51.24%	51.79%	8.33%
Norm-2	9.20%	43.66%	44.61%	47.74%	4.99%
Norm-3	7.58%	50.47%	51.42%	30.88%	11.73%
High-1	6.14%	44.11%	35.48%	60.97%	10.03%
High-2	5.48%	20.47%	40.23%	34.26%	7.62%
High-3	3.37%	54.05%	54.52%	36.81%	3.63%
Average	**7.08%**	**43.07%**	**43.91%**	**46.43%**	**8.85%**

5.2 Further Analysis

A more detailed analysis of the obtained schedules shows that the new scheduling approach achieves better metrics due to:

- *Improved balance in wafer-to-machine allocations.*
  A better-balanced allocation scheme of wafers across the available machines is achieved by shifting wafers from busy to less utilized machines. In other words, the margins of the utilization

rates across the machines tend to decrease or be more uniform compared to SimModel schedules where some machines may report unnecessary high or low utilization rates.

- *Improved batching formations.*
  One of the basic trade-offs that arise is small batches, where all constituent wafers can be released early for downstream processing, or larger batches, where all wafers finish later but benefit from less fixed time. The optimal choice depends on a number of factors, such as wafers priorities, the operating status of other machines and the release schedule of incoming wafers for avoiding bottlenecks. Our scheduling tool captures all these aspects in an integrated fashion and optimally navigate these trade-offs which is extremely difficult to be captured explicitly via heuristic rules.

Figure 2 displays the minimum, maximum and average number of wafers allocated per machine across the 9 benchmark datasets for the schedules derived by both scheduling approaches. As can be seen, in all benchmark instances apart from High-1, the Flex schedule allocates a larger minimum number of wafers across the machines. For example, in Low-1 benchmark instance, the minimum is increased from 2 to 8 wafers, meaning that a machine that was very underutilized in the Sim schedule can substantially increase its utilization as the Flex schedule effectively showed. It is important to emphasize that even though that our schedules have on average 7% more wafers, most of these schedules (apart from those of High-2 and High-3 instances) use a maximum number of wafers per machine that is lower compared to that of Sim schedules. When averaged across all 9 cases, the minimum and maximum number of wafers allocated to a single machine increased by 29% and decreased by 16%, respectively. In other words, our schedules have a smaller range of wafers allocated compared to their Sim counterparts, indicating clearly that these schedules make much better use of the available resources and comes up with a more balanced allocation of wafers across machines.

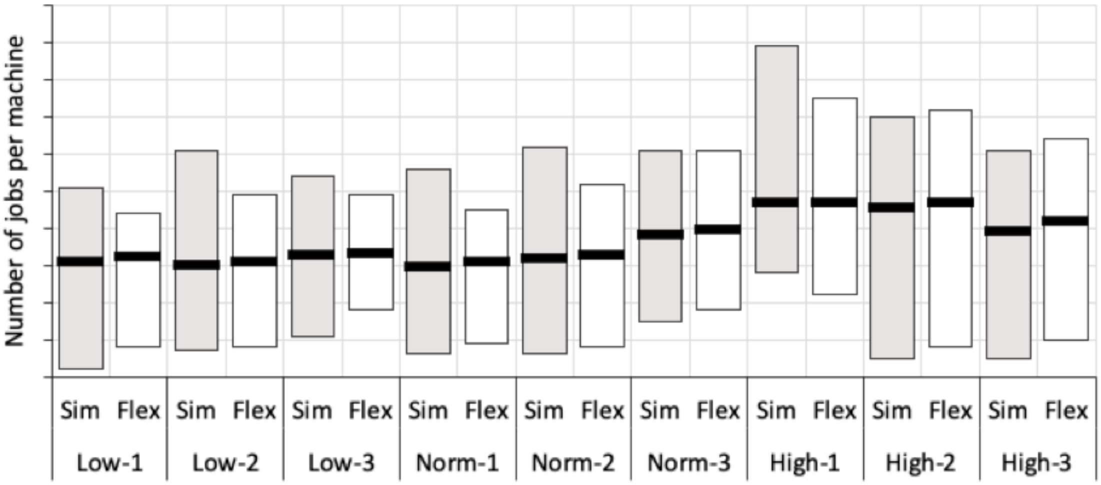

Figure 2: Range of the number of wafers per machine across 9 benchmark datasets for the schedules of both approaches. Each bar spans the min / max number of wafers and black lines show the average.

Figure 3 shows the minimum, maximum and average number of batches per machine across the 9 benchmark datasets for both scheduling approaches. In the first 4 benchmark instances which involve low and medium workload, we can see that Flex schedules result in a smaller range of batches while the average values are increased. This is, again, an indication of the better and more balanced allocation of wafers to machines. It is important to note that although these schedules involved 7% more wafers, they used a substantially higher number of batches (about 39.7% more batches). The higher number of batches is a result of the advanced capability of the proposed scheduling solution to prioritize across wafers with different priorities. In Flex schedules, generally high-priority wafers tend to batch together in smaller batches aiming at small cycle times and complete early while lower-priority wafers tend to be part of larger batches that benefit from reduced setup, teardowns and fixed processing times.

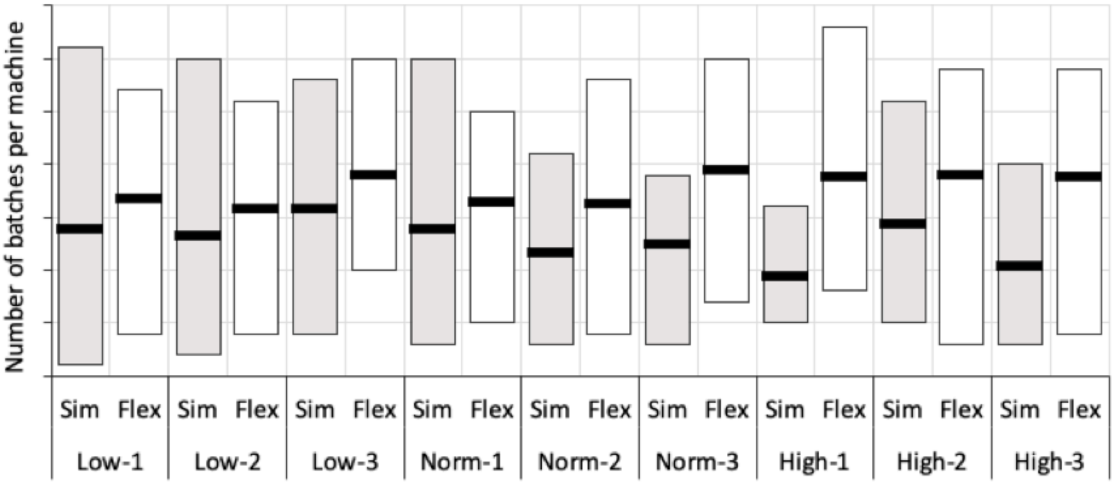

Figure 3: Range of the number of batches per machine across 9 benchmark datasets for the schedules of both approaches. Each bar spans the min / max number of batches and black lines show the average.

6 CONCLUSIONS

As the benchmark study clearly showed, our proposed MILP-based scheduling approach outperforms significantly existing traditional heuristic-based tools, such as SimModel. Our scheduling solutions achieved on average a major reduction of more than 43% on cycle times for high-priority wafers, a reduction of about 9% on total cycle times, along with an increase of 7% in throughput. In particular, these remarkable improvements on the above schedule metrics are mainly a result of the more balanced allocation of wafers to machines and the better batching formations that capture and exploit efficiently major scheduling data, such as wafers' release dates and priorities. This will vastly improve capacity utilization with potential opportunity for future capital avoidance due to increased throughput at highly utilized tools. Deploying an optimization scheduling model also ensures improvements in WIP predictability, cycle time and velocity metrics across a multi-reentrant wafer fab. Currently, we are working on schedule synchronization and adherence by performing live trials of the proposed scheduling platform in the metrology toolsets of the Springtown Seagate site. Upon successful completion of these live trials, we are going to consider more complex Seagate toolsets (such as those with multiple load ports) that involve extremely hard-to-optimize constraints such as single-recipe batches, incompatible production paths for recipes, Kanbans and reticles. As a result of these additional constraints, the complexity of the resulting optimization problems is expected to grow enormously. For this reason, a major part of our research and development activities will focus on computational performance aiming at generating high quality schedules within short timeframes.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contribution and support in this project the whole Flexciton team, and particularly the following colleagues: Begun Efeoglu, Ioannis Konstantelos, George Kopanias, Robert Moss, Anna Robson, Charles Thomas, Luke Watkins and Ruaridh Williamson, as well as the following individuals from Seagate Technology: Jen Krolnik, Richard Surman, Gerard McLaughin, Maureen Doherty and Laura McElhinney.

REFERENCES

SIA. 2020. The Semiconductor Industry Association, The 2020 SIA Factbook. https://www.semiconductors.org/the-2020-sia-factbook-your-source-for-semiconductor-industry-data/, accessed 16th May.

Kim, S. H., and Y. H. Lee. 2006. "Synchronized Production Planning and Scheduling in Semiconductor Fabrication", *Computers & Industrial Engineering* 96: 72–85.

Kopanos, G. M., and L. Puigjaner. 2019. *Solving Large-Scale Production Scheduling and Planning in the Process Industries.* Springer International Publishing

Leachman, R., J. Kang and V. Lin. 2002. "SLIM: Short Cycle Time and Low Inventory in Manufacturing at Samsung Electronics". *Interfaces* 32(1): 61–77.

Mathirajan, M., and A. Sivakumar. 2006. "A Literature Review, Classification and Simple Meta-analysis on Scheduling of Batch Processors in Semiconductor". *The International Journal of Advanced Manufacturing Technology* 29(9-10): 990–1001.

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

Ovacik, I. M., and R. Uzsoy. 1997. *Decomposition Methods for Complex Factory Scheduling Problems*. Kluwer Academic Publishers Boston.

Pfund, M. E., S. J. Mason, and J. W. Fowler. 2006. *Semiconductor Manufacturing Scheduling and Dispatching*. Handbook of Production Scheduling, 213–241. Springer, New York.

Potts, C. N., and M. Y. Kovalyov. 2000. "Scheduling with Batching: A Review". *European Journal of Operational Research* 120(2): 228–249.

Sörensen, K., 2015. "Metaheuristics - The Metaphor Exposed". *International Transactions in Operational Research* 22(1): 3–18.

Yurtsever, T., E. Kutanoglu, and J. Johns. 2009. "Heuristic Based Scheduling System for Diffusion in Semiconductor Manufacturing". In *Proceedings of the 2009 Winter Simulation Conference*, 1677–1685.

AUTHOR BIOGRAPHIES

GEORGIOS M. KOPANOS is a Lead Scientist in Quantitative Research and the R&D Lead at Flexciton. He holds a PhD in Chemical Engineering from Universitat Politècnica de Catalunya; sponsored by the Spanish Ministry of Education. His research interest is in operations management, supply chains, and optimization methods via mathematical programming. He has 15 years industrial experience in the application areas of energy systems, pharmaceuticals, chemicals, fast-moving consumer goods, food processing, and military. He is the author of more than 30 scientific articles, and a book on production planning and scheduling. He is also an editor of a book on advanced energy systems engineering. He was a research associate at Imperial College London, and hold an academic position at Cranfield University. His email address is giorgos.kopanos@flexciton.com.

DENNIS XENOS is the founding CTO at Flexciton. With over 10 year's experience in modelling, optimisation and engineering, Dionysios is responsible for overseeing the optimisation based technology being developed at Flexciton. He holds a PhD in Chemical Engineering from Imperial College London. He has innovated in the field of the operations optimisation of large industrial plants and of the most complex manufacturing environments considering data-based models and predictive information combined with mathematical programming models. He is an author of more than 15 scientific publications and one patent of the optimisation of industrial problems. His email address is dionysios.xenos@flexciton.com.

SLAVA ANDREEV is the Head of Product at Flexciton. He holds a Master's degree in Computer Science from Samara State Aerospace University. He has 20 years experience in software development and engineering, product development, research and solution design, business analysis and project management. His research interest is in artificial intelligence optimisation and heuristics methods. He has industrial experiences in the application areas of logistics, ground transportation, private hire, distribution and last mile delivery. He is the author of more than 10 scientific articles on multi-agent system and swarm intelligence. His email address is slava.andreev@flexciton.com.

TINA O'DONNELL is the Wafer Systems Engineering Manager at Seagate Technology in Springtown, Derry, Northern Ireland. She received a BEng in Electronic Engineering and Computing in 2002 and a PhD in Supply Chain Optimisation in 2007 from Ulster University. She is responsbible for Global Wafer MES Systems She has 12 years experience in developing RTD rules, automation workflow and simulation/optimisation models. Her email address is tina.odonnell@seagate.com.

SHARON FEELY works as an Industrial Engineer for Seagate Technology at the wafer fab in Springtown, Derry, Northern Ireland. She received a BEng in Industrial Engineering in 1992 from NUIG, Galway. She is responsible for managing Capacity, Cycle time and Capital Spend for Metrology Toolsets. She is also responsible for leading the operations research strategy for the global wafer fabs. Her email address is sharon.r.feely@seagate.com.