Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

FAB-WIDE SCHEDULING OF SEMICONDUCTOR PLANTS: A LARGE-SCALE INDUSTRIAL DEPLOYMENT CASE STUDY

Ioannis Konstantelos
Johannes Wiebe
Robert Moss
Sebastian Steele
Dennis Xenos

Flexciton Ltd.
Churchill House, 142 Old St
London, EC1V 9BW, UK

Tina O'Donnell
Sharon Feely

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

ABSTRACT

This paper presents a novel fab-wide scheduling approach for large semiconductor manufacturing plants. The challenge lies in the scale and complexity of the problem at hand; thousands of wafers must be scheduled on hundreds of machines across several steps while respecting a wide array of operational constraints such as the use of photolithography reticles, timelinks and flow control limits. We begin by surveying the state-of-the art and presenting some key opportunities that arise in the context of global scheduling. The proposed approach is presented, highlighting its hierarchical structure and how it can interface with local toolset schedulers. We present some illustrative examples and aggregate statistics obtained during ongoing trials at Seagate Springtown. We demonstrate that the proposed approach can result in substantial cycle time improvements when compared to myopic dispatch methods and a marked reduction in the need of manual intervention for controlling flows.

1 INTRODUCTION

The semiconductor industry is one of the largest and most complex industries in the world. The critical factors in semiconductor manufacturing are the ability to rapidly develop and test novel technologies, improve manufacturing processes to reduce rework and waste, as well as meet production targets in terms of prescribed volumes and due dates. In recent years, the industry has been affected by various well-documented issues such as increasing demand, lack of production capacity and shortage of raw materials (Voas et al. 2021). Staying competitive in this evolving market is a constant challenge for manufacturers. In view of the high capital costs and long lead times involved in expanding production capacity, efficient operation management is important to ensure that the available resources are utilised in an optimal manner.

High quality scheduling is of paramount importance since it enables higher utilization of expensive toolsets, enables faster research and innovation, reduces cycle times and ultimately enables higher throughput and on-time delivery. However, producing quality schedules is a difficult task and an active research topic attracting high academic and industrial interest across the world. A comprehensive review of problems and solution techniques that arise in the context of semiconductor manufacturing operations can be found in (Mönch, Fowler, Dauzère-Pérès, Mason, and Rose 2011). In brief, the characteristics that render it such a challenge are, among others, the large scale of some factories involving hundreds of jobs and machines, the re-entrant flows, batching constraints, timelinks and the long cycle time of products that require hundreds of process steps.

Due to the long cycle times, decision-making in semiconductor fabrication plants (fabs) is typically framed as a multi-level problem. As illustrated in (Barhebwa-Mushamuka 2020), scheduling decisions can be ascribed to two levels; global and local. On one hand, global scheduling (or fab-wide) is tasked with the strategic management of factory assets while considering all work-in-progress, incoming and outgoing flows across the fab, expected resource availability and other constraints. On the other hand, local (or toolset-level) scheduling focuses on the operation of individual work centres. It is typically tasked with identifying the best immediate dispatch decisions i.e. which jobs waiting for dispatch should be assigned to which available machine. A review of typical dispatch approaches is in (Sarin et al. 2011).

The aim of this paper is to demonstrate a novel fab-wide scheduling approach that can provide guidance and orchestrate the operation of multiple toolset-level schedulers to materially improve factory-wide Key Performance Indicators (KPIs). In the next sections we will present the global scheduling state-of-the-art, explain the proposed methodology and present examples of successful operation and some initial results from recent trials in the Seagate Springtown plant.

2 LITERATURE REVIEW

A number of approaches to global scheduling for semiconductor fabs has been proposed. They can generally be categorized by the type of global scheduling model used (simulation, optimization or heuristic) and by the interface used between the global and local (toolset) scheduling layers (priorities or production targets). Using production targets to drive scheduling decisions is generally seen as a direct approach for staying in line with the true commitments of the business, whereas updating priorities can be more indirect since they typically also capture other aspects such as customer preferences. Approaches using production targets generated by the global scheduler to guide the local layer include work by (Kao and Chang 2018), who use a heuristic approach for setting production targets, (Barhebwa-Mushamuka 2020) who propose a linear programming-based global scheduling framework, and (Hwang and Chang 2003), who also use an optimisation-based global layer combining Lagrangian relaxation with network flow optimization. Approaches with a priorities-based interface include work by (Bureau et al. 2007), who use a simulation-based local and global model, and (Zhong, Liu, and Bao 2021) who use a meta-heuristic approach to generate priorities which guide a local dispatcher. In addition, there are some popular heuristic techniques that can operate at a large scale, such as the Shifting Bottleneck Heuristic presented in (Mason et al. 2002). However, as a general rule these techniques suffer either in the level of fidelity they use to represent actual operation (warranting the need for a "translation" layer to reconcile between what is being instructed and what is operationally feasible), or in their searching capability to identify meaningful strategic actions i.e. actions whose benefit becomes apparent only when considering later steps. For example, deprioritising a high-priority wafer that will be stuck behind a bottleneck in its next step (see for example Section 5.3). To date there has not been a published case study of a fab-wide scheduler successfully deployed in a large semiconductor manufacturing facility.

The framework proposed in this paper falls into the category of priority-based interfaces; the Fab-Wide Scheduler (FWS) module combines a high-fidelity operational modelling engine with search techniques for identifying strategic opportunities for wafer re-prioritisation. This communicates with a local scheduler that combines exact Mixed Integer-Linear Programming (MILP) techniques and heuristics. We note that the presented work is part of an existing scheduling platform that until now has focused on toolset-level scheduling. In (Kopanos et al. 2020) a substantial improvement of 43% on the cycle time of high priority wafers was reported. In a follow up publication by (Elaoud et al. 2021), the hybrid approach was taken further, with the inclusion of multi-step timelink constraints, showing timelink violations were decreased by up to 72%. In the present paper we take this approach even further, and demonstrate a novel solution strategy capable of looking ahead, identifying strategic decisions across the entire factory and successfully communicating them to local toolset schedulers.

3 IDENTIFYING OPPORTUNITIES

3.1 Bottlenecks and Flow Diversity

The cause, impact and management of bottlenecks in factories is of great interest to researchers and operators, since they impede operation on many levels. In semiconductor fabs, the majority of work largely involves repetition of process loops that build consecutive layers upon a silicon wafer. This re-entrant flow combined with the fact that some types of machines are more expensive, have less capacity or are more constrained due to physical (e.g. single recipe batching or reliance on a busy secondary resource) or operational constraints, leads to the formation of bottlenecks. At times, some machines become very busy and completing all outstanding processes in the queue can take several days. Predicting when, where and how severe a bottleneck will be is not a straightforward task. Although outright resolving a bottleneck is in most cases not possible, since it arises due to real capacity and technical constraints, what can be managed intelligently is bottleneck avoidance.

When a bottleneck arises, a myopic approach would be to keep applying the standard dispatch; high-priority lots (groups of wafers) are expedited while standard priority lots wait for their turn, perhaps subject to a periodic priority-increase rule as wait time is accumulated. In contrast, a strategic approach looks at the bigger picture and understands that there is little point in the early dispatch of wafers that will end up in a bottleneck area; what makes more sense is to instead process wafers that can continue along their downstream path unobstructed. The end effect is that although queuing in front of bottlenecked tools may not reduce, throughput across the entire fab can increase substantially.

3.2 WIP Flow Control Management

WIP (work-in-progress) flow control mechanisms (also referred to as kanbans) are an important part of a factory's operation. They typically apply as limits to the number of lots that can be be processed in a group of machines or across a series of route steps. They are primarily used for quality control purposes and can dynamically change based on process quality feedback. These limits can materially prevent efficient use of capacity, particularly when they block high-priority wafers from progressing downstream due to a kanban being at its limit. Fab-wide scheduling can materially improve this aspect of operation by quantifying the wider impact that a kanban at its limit has and taking appropriate action (e.g. speeding up a low priority lot) which would otherwise go against heuristic dispatch rules.

3.3 Timelink Constraints

Timelinks are one of the most challenging aspects of modelling a wafer fab. Timelinks define the minimum or maximum amount of time that can elapse between two or more consecutive process steps of a lot to prevent oxidation and contamination of wafers while waiting between processes. Violating a timelink constraint can lead to wafers being scrapped or requiring costly and time-consuming re-work (Klemmt and Mönch 2021). Due to their very nature, timelinks lead to a conundrum: on the one hand we wish to keep downstream machines idle, so as to guarantee that there will be spare capacity to process them in time. On the other hand, keeping machines idle has a detrimental impact on other wafers and overall fab performance. Typically, time-linked lots will be subject to conservative dispatch rules. More advanced approaches have been presented in the past including MILP decomposition (Klemmt and Mönch 2021) and sampling-based approaches (Lima, Borodin, Dauzère-Pérès, and Vialletelle 2021). Fab-wide scheduling can greatly assist in this respect, by accurately predicting arrival times of incoming lots, weighting all priorities and objectives and deciding when to trigger time-linked lots.

4 METHODOLOGY

The scheduling framework proposed in this paper is hierarchical and consists of two main components which run independently and at different frequency — the Toolset Scheduler (TS) and Fab-Wide Scheduler

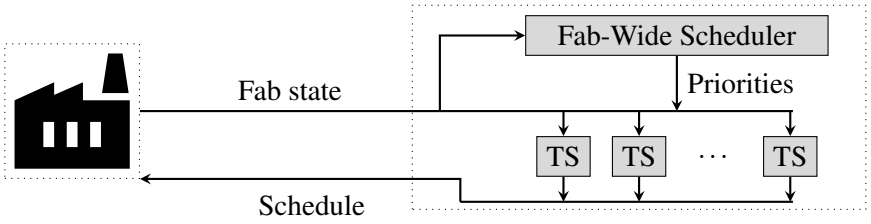


Figure 1: Overview of the solution strategy. FWS and TS receive fab state data. FWS computes priorities, used as guidance by the toolset schedulers. Toolset schedulers produce a schedule sent to the fab.

(FWS). We receive data representing the current state of the fab at a 5-minute frequency. By "Fab state" we denote a data snapshot containing information on the state and location of all in-process lots and reticles, the state of all machines, all active kanban rules and all other dynamic constraints that impact operation. As illustrated in Figure 1, the arrival of this data triggers a FWS run and toolset scheduler runs for each toolset cluster. The toolset schedulers finish within 5 minutes and return a full schedule containing all current steps to the fab. The FWS run time depends on how busy the fab is, but generally completes within 7 minutes. Each TS run utilizes guidance from the most recent FWS run. Depending on the FWS computation time, this means that the TS may use priority changes computed by the FWS on the basis of slightly "older" data (typically by just a few minutes). Consistency issues that could arise due to this time delay (e.g. a wafer requiring rework) is handled by the permissive interface between FWS and TS. Most importantly, the impact of this time delay is, in most cases, minimal given that fundamental fab conditions (kanbans, bottlenecks, timelinks etc.) and wafer flows do not abruptly change in a few minutes across the entire plant, and priority changes are re-computed at a high frequency. Also note that due to the high fidelity of FWS and TS in capturing operational constraints (reinforced by trial periods and automatic validation testing on the fab's side), there is no "reconciliation" layer necessary; the schedules are directly actionable in the fab.

4.1 Toolset Scheduler (TS)

The Toolset Scheduler considers the currently in-process and/or upcoming process step of all wafers in the cluster. This means that the look-ahead horizon is a single step for each in-process wafer, which in terms of time typically corresponds to several days of fab operation. Given the state of the fab and the ongoing work, we are able to split machines into independent clusters that have zero interaction and can be solved independently. After solving all clusters, the final schedule can be aggregated and sent back to the factory. Our toolset scheduler implements all constraints commonly found in wafer fabs, e.g.: single recipe and multiple recipe tools, secondary resources (such as reticles, reticle pods, and monitor wafers), internal processing path conflicts, timelinks, batching (including dynamic maximum batch sizes) and WIP flow control management (such as kanbans). The objective of this optimization-based approach is largely cycle time minimization weighted by wafer priority, but also includes other components such as reducing lot lateness, the number of batches, reticle and pod moves, and timelink violations. A key strength of the tool's solution strategy is the combined use of fast heuristics for warm starting, and decomposition methods for effectively handling MILP problem size, allowing the solution of large problems within a few minutes. For a detailed discussion of the toolset scheduler see (Kopanos et al. 2020) and (Elaoud et al. 2021).

4.2 Fab-Wide Scheduler (FWS)

Unlike the toolset scheduler, our fab-wide scheduler takes a view of the entire fab at once and considers multiple future steps for each wafer. It is simulation-based and focuses on improving schedule quality by considering the flow of wafers through the fab, something the toolset scheduler cannot do due to its single-step, toolset-level nature. The main purpose is to redirect flow through the fab and thereby improve flow linearity, reduce bottlenecks, improving WIP flow control management, and reducing timelink violations.

Our fab-wide scheduling approach achieves this by predicting wait/cycle times for multiple future steps, analysing those predicted wait/cycle times with respect to the different areas of potential improvement, and re-prioritising wafer steps in a way that guarantees improved (weighted) cycle times. In brief, FWS combines two main elements: (i) an operational module that captures in full detail all relevant constraints e.g. detailed process time modeling, machine maintenance, shift changes, dynamic batching constraints, kanbans etc. (ii) a search module that identifies beneficial priority changes given the evolving fab conditions. Although a single iteration can already identify beneficial priority changes, the search process can be embedded in an iterative scheme to enhance the search and prevent it from getting trapped in local minima. Naturally, this gives rise to a trade-off between benefit and computational speed, which is highly dependent on the fab characteristics.

Regarding the FWS look-ahead horizon, there is no limit and is actually user-defined in terms of either time or number of steps. Again, this ultimately depends on the operation characteristics of the fab such as average step velocity, number of wafers in processing, number of machines, number of time-linked steps etc. In principle, it is possible to extend the horizon several months ahead or until all in-process wafers complete their processing. A longer time horizon should in principle result in better decisions. However, it can increase computation time, to the point where once the FWS output has been computed, the fab state has changed considerably and the computed scheme is no longer beneficial or relevant. During trials, a number of different look-ahead horizons were tested to further explore this trade-off, ranging from a few hours up to several weeks.

FWS communicates with the toolset schedulers via priority weights for individual steps of a wafer. An advantage of our approach is that, while FWS always schedules all tools in the fab, users can specify which toolsets are subject to guidance; FWS adjusts its search accordingly. This is particularly useful for gradually rolling out FWS in a fab and evaluating its impact, as described in Section 6.1.

5 TRIAL OPERATION EXAMPLES

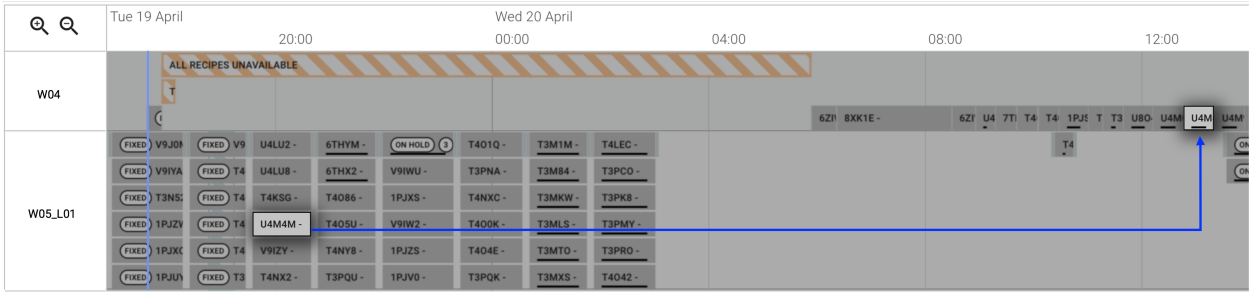
Seagate is a major semiconductor industry and a world leader in data storage technology, with more than 40% share of the global Hard Disk Drive (HDD) market. The Springtown facility in Northern Ireland produces around 25% of the total global demand for recording heads, the critical component in a HDD.

In this section we present a few examples of beneficial FWS guidance, as captured during the trials in Seagate Springtown. They are illustrative of the overall FWS-TS interaction principle and show how KPIs can be improved by look-ahead guidance.

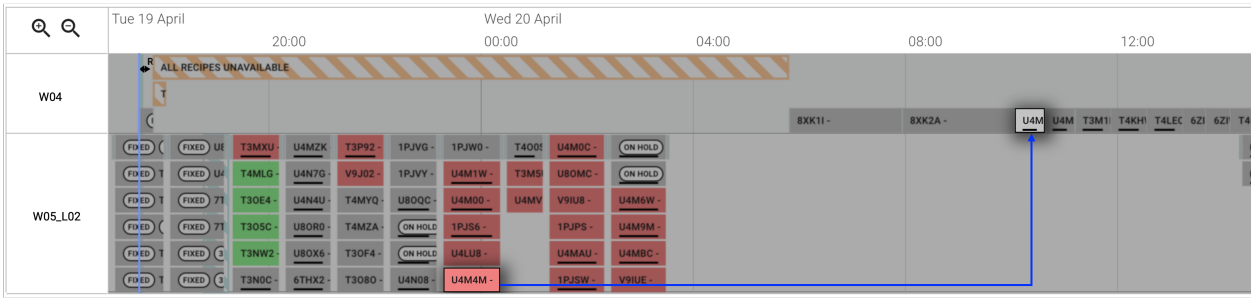
5.1 Case Study 1: Managing Maintenance and Outages

The first example shows the benefit of predicting queue times. Whether a toolset is bottlenecked or not can change rapidly; typical reasons include high influx of work as well as planned or unplanned maintenance. An unexpected tool outage can drastically reduce toolset capacity, instantly creating a bottleneck that may last days or weeks. Due to the high run frequency of both the FWS and TS in our framework, our solution strategy reacts quickly to these outages and starts re-directing flow within minutes of a tool going down. By considering predicted repair times when generating expected queue times, we can also account for the sudden increase in toolset capacity once the tool comes online again. It is important to highlight that these opportunities are uncovered by the FWS module due to the high fidelity when modelling operation; this would not be possible if aggregate representations of flows and capacity were used.

Figure 2 shows how FWS can make better decisions around bottlenecks caused by maintenance. Without FWS (Figure 2a), the highlighted wafer U4M4M (see black border) is scheduled for processing on machine W05_L01 at around 20:00, while its subsequent step on machine W04 (top of the Gantt chart) is stuck behind a long maintenance period (time block with orange stripes) and get dispatched after 1pm the next day, since a long queue builds up. With FWS (Figure 2b), this wafer is deprioritised (deprioritised and prioritised wafers are shown in red and green color respectively) and runs several hours later. This



(a) Without fab-wide scheduling: wafer runs early despite having to wait on its next step due to machine maintenance



(b) With fab-wide scheduling: wafer runs later to avoid queue time at the next step where a machine is undergoing maintenance

Figure 2: Deprioritising wafers due to machine maintenance.

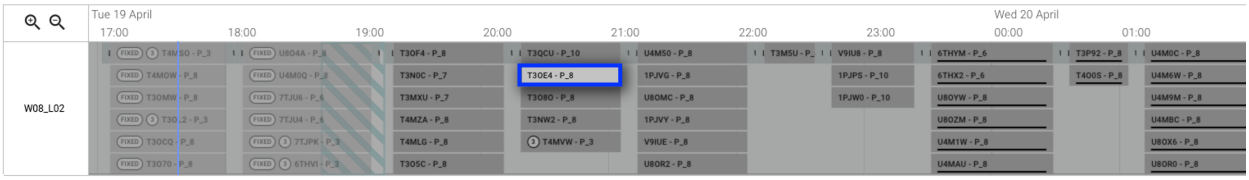
deprioritisation allows wafers along other routes to continue unobstructed, increasing fab throughput, while the 4 hour delay on U4M4M has no material impact on cycle time.

5.2 Case Study 2: Managing WIP Control Mechanisms

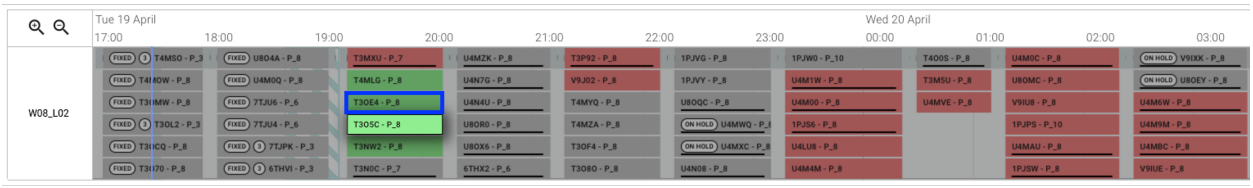
As outlined above, WIP control mechanisms are an important way of managing flow through the fab and ensuring product quality is high. However, these constraints can also lead to considerable queue times. Consider a kanban that is currently full but only contains low priority wafers; this means that one of these low priority wafers will have to exit the kanban to allow a new wafer to enter. However, due to their low priority, it may take a long time for these wafers to move through the kanban. This is particularly the case if dispatch decisions are left entirely to a local scheduler that only considers the wafer's next step and ignores later knock-on effects.. As such we have low-priority wafers blocking other, potentially higher priority, wafers from entering the kanban leading to increased (weighted) cycle time. Figure 3 shows how our solution strategy reduces (weighted) cycle time by speeding up kanban wafers that are blocking other, high priority wafers from entering the kanban. The highlighted wafers T3OE4 in Figure 3a has a lower priority and only runs after a batch of higher priority wafers. However, this wafer is on the exit step of a kanban and is blocking a high-priority lot from entering the kanban. As shown in Figure 3b, FWS increases the priority of wafer T3OE4 on the exit step (highlighted green), which is dispatched an hour earlier (around 19:00), thus allowing high priority wafers to enter the kanban earlier.

5.3 Case Study 3: Managing Secondary Resources and Long Processing Times

Finally, Figure 4 shows an example of how our approach handles secondary resources such as reticles, reticle pods (loading groups of reticles to a machine using specialised containers) or internal processing path conflicts (wafers being processed in a machine constrain the type of processes that can be done in parallel when loading a different load port). The wafer T4LIW highlighted in Figure 4a can go onto either load port of W09, however it is blocked by the fixed batch on W09_L01 due to an internal processing path (IPP) conflict (which due to long processing time is set to last over 5 hours). This means that, even



(a) Without fab-wide scheduling: Wafer runs later, even though it is blocking high priority wafers from entering kanban



(b) With fab-wide scheduling: Wafer runs earlier to free up full kanban blocking other high priority wafers

Figure 3: Prioritising wafers because they are blocking other wafers from entering a kanban.

though the other load port W09_L02 is idle, the second step cannot start until the batch on W09_L01 has finished. As Figure 4b shows, FWS recognises this and deprioritises the first step of the wafer (marked in light red), moving it later by over 4 hours. This allows other wafers, not hindered by this bottleneck, to be processed first.

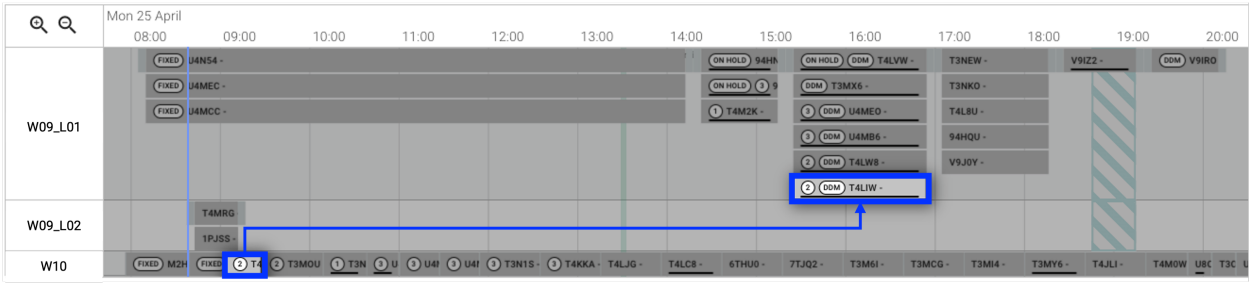
6 FAB-WIDE SCHEDULING PERFORMANCE

In this section the deployment and testing strategy adopted for trials in Seagate Springtown is described. Note that the trial took place between March-May 2022 and the system is currently operational 24/7 (June 2022). Understanding the overall impact of a fab-wide scheduling approach to a factory's KPIs is a non trivial task. This is due to several factors: (i) long timelines, dynamic and uncertain operational conditions (ii) inability to have a well-defined counterfactual scenario to compare against (iii) preexisting dispatch rules, heuristics and human behaviour that may be working against the new approach (iv) reluctance to test the new approach directly at a grand scale.

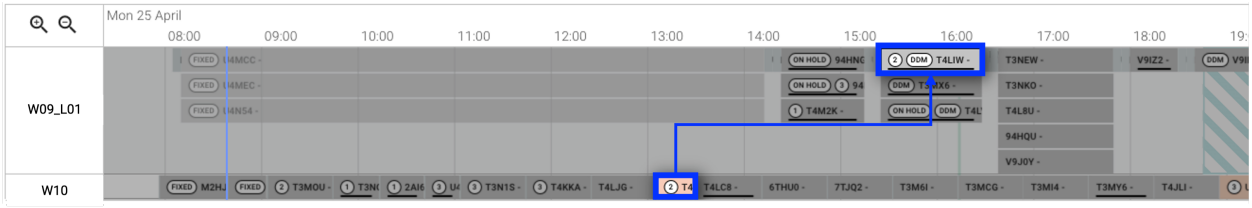
6.1 Deployment and Testing

Deploying and testing a novel piece of technology in a large factory that runs around the clock presents many practical challenges. In this vein, a key strength of the proposed fab-wide scheduling approach is that it can easily facilitate controlled deployments via two main features. Firstly, although the FWS module always schedules the entire fab, the manager can easily configure it to only control a subset of toolsets. This allows for localised trials to be conducted at problematic areas, where the scope for large-scale adverse effects is limited, the desired behaviour may be known a priori or can be kept under intense monitoring for an extended period of time. Secondly, the "strength" of the guidance can be changed, enabling deployment to begin with only making small "nudges" to the local scheduler. In other words, we can allow only for high-value suggestions which should be easier to validate, as opposed to changes that may provide less obvious benefits and be more involved in their justification. Note that although all decisions of the fab-wide scheduler are fully explainable, some can be more robust than others due to the inherent uncertainty and prediction error around processing and transfer times. As the decisions are validated in the factory with the aid of operators and confidence in the system increases, this strength can be increased as required, eventually allowing for full control of an area.

The early phases of the trial were supported by specially-designed diagnostic tools. In particular, the scheduling platform can produce heat maps of queue times across the fab and job flow diagrams. These were instrumental in identifying areas of interest and subsequently justifying the identified strategic decisions.



(a) Without fab-wide scheduling: Wafer runs early, even though its next step has to wait due to a conflicting IPP.



(b) With fab-wide scheduling: Wafer runs later allowing other wafers to go first.

Figure 4: Deprioritising wafers because they are blocked by a secondary resource.

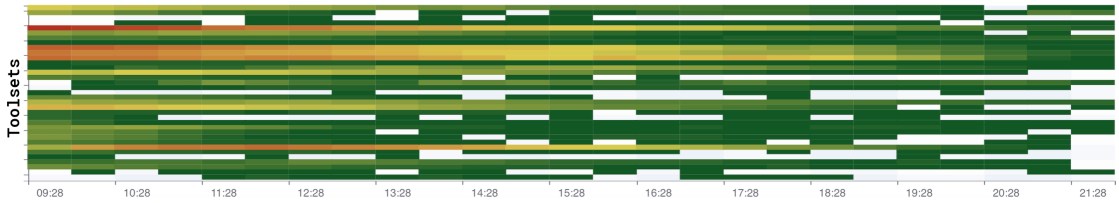
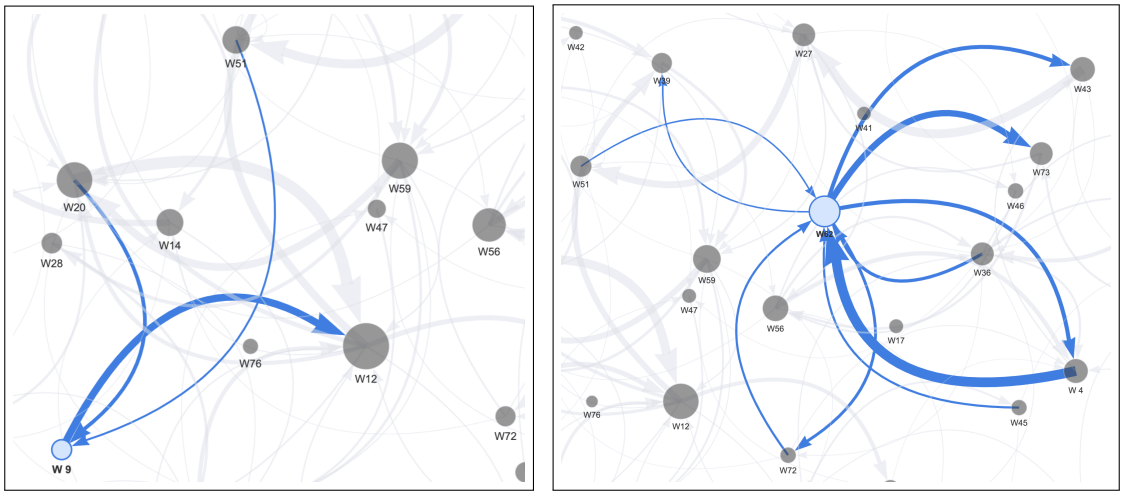


Figure 5: Heatmap of projected queuing time across a subset of toolsets over time. Red indicates long queuing times i.e. presence of a bottleneck, while green means that jobs can be started after little or no waiting.

The heatmap of queue times, shown in Figure 5 is plotted for some toolsets across the planning horizon. Since the FWS always schedules the entire fab, it is possible to use this to identify existing and predicted future bottlenecks. Once bottlenecks or future bottlenecks are identified, network flow diagrams (as shown in Figure 6 can be used to show which toolsets feed those bottlenecks; these are good candidate areas at which to trial the FWS's bottleneck avoidance capability. By controlling these areas, the system can then prevent predicted bottlenecks from actually occurring. It became apparent during the deployment that focusing on areas with high "flow diversity" was very valuable. For example, metrology work stations were good candidates since most do not have recipe-based batching constraints and have many downstream toolsets. As such, there are many opportunities to prioritise lots that have unobstructed downstream paths.

6.2 Theoretical Benefit Quantification

A theoretical measure of the improvement due to using the presented FWS strategy can be measured by comparing the cycle times obtained when using different scheduling methods. Table 1 shows the improvement of the proposed FWS scheme with respect to three other methods, as computed on snapshots obtained across a 14 day horizon during trials in the Seagate Springtown fab. Positive values denote a reduction in cycle times i.e. superior performance by FWS. The improvement is broken down per priority (with P1 and P10 denoting the highest and lowest respectively). The three other methods are (i) first-in-first-out (FIFO) i.e. wafers are dispatched according to how long they have been waiting on the rack (ii) last-in-first-out (LIFO) i.e. newly-arrived wafers are dispatched earlier (iii) Heuristic Dispatch Rules



(a) (b)

Figure 6: Network flow diagrams focusing on a toolset with (a) low and (b) high diversity flow. Each node represents a toolset (size is proportional to number of queuing lots), while arrow thickness indicates the number of wafers that are likely to move between toolsets within the current shift. Thicker arrows are indicative of an ongoing/impending bottleneck on the destination toolset.

(HDR): a custom collection of heuristic priority-based dispatch rules that have been used in TS showing good performance Note that this analysis solely captures the improvement due to re-prioritisation and flow re-direction achieved by the FWS. It does not include any improvements made by the toolset scheduler, e.g. improved batching or reduced reticle moves, which are considerations reserved for the local level.

Table 1: Cycle time improvements of FWS compared to other methods averaged across 12-hour snapshots. The weighted improvement uses the TS relative priority weights.

Model	Weighted	P1	P2	P3	P4	P6	P7	P8	P10
FIFO	9.79%	21.00%	15.32%	11.95%	10.03%	5.00%	-0.94%	-10.64%	-18.27%
LIFO	8.70%	18.39%	12.27%	10.25%	7.24%	5.25%	2.03%	-11.74%	-17.33%
HDR	1.15%	1.79%	0.39%	0.55%	-4.23%	-1.88%	2.20%	0.80%	3.61%

The table shows measurements from an early phase of the trials where FWS was controlling hundreds of tools accounting for more than 60% of the fab, while the planning horizon was 12 hours, corresponding to scheduling multiple steps of more than 4,000 lots. It is expected that controlling more tools and using a longer horizon would result in an even more pronounced benefit. When compared to the most advanced method (HDR) FWS achieves better weighted cycle times on all scenarios analysed. As can be seen in the table, FWS outperforms the other three methods, across almost all priority classes - note that the slightly penalised P4 and P6 medium-priority classes are the least populous. It is important to highlight that the potential for improvement changes with the fab conditions. At certain times large improvements can be achieved. An example of this would be a scenario where low priority wafers are blocking a large number of high priority wafers from entering a kanban. Other states may have very little room for improvement e.g. a low WIP situation with no bottlenecked tools and largely free kanbans. Most importantly, this snapshot-based analysis captures only a small part of the benefit. Sustained use of FWS has a compounding effect, where strategic actions combine over time to bring the fab to a better state. However, by looking at individual 12-hour snapshots cannot capture the effect of FWS interventions beyond the 12-hour horizon, since each snapshot "starts from scratch" (i.e. does not build upon the improved fab state produced by the FWS). As such, this compounding effect cannot be quantified.

A proxy we can use for this benefit is the volume of ad hoc control flow rules activated/deactivated in the fab. Every day, specialists have to define numerous, in some cases even hundreds, of ad hoc control flow rules to better manage operations given the prevalent conditions. For example, setting a "hard down" rule, where lots are manually placed on hold so as not to continue to a downstream bottleneck. In Figure 7, we show the number of ad hoc operational rules implemented in the Seagate Springtown fab between weeks 2 and 26 of the year 2022 (i.e. from early January until late June). Light gray, dark gray and black colours mark the weeks before deployment, during trials and after FWS became operational 24/7 respectively. As can be seen in the final weeks, the number of ad hoc rule transactions averaged less than 150 per week, a decrease of over 300% compared to the pre-deployment period. This is strong evidence that FWS deployment reduced massively manual interventions required to control flows within the fab.

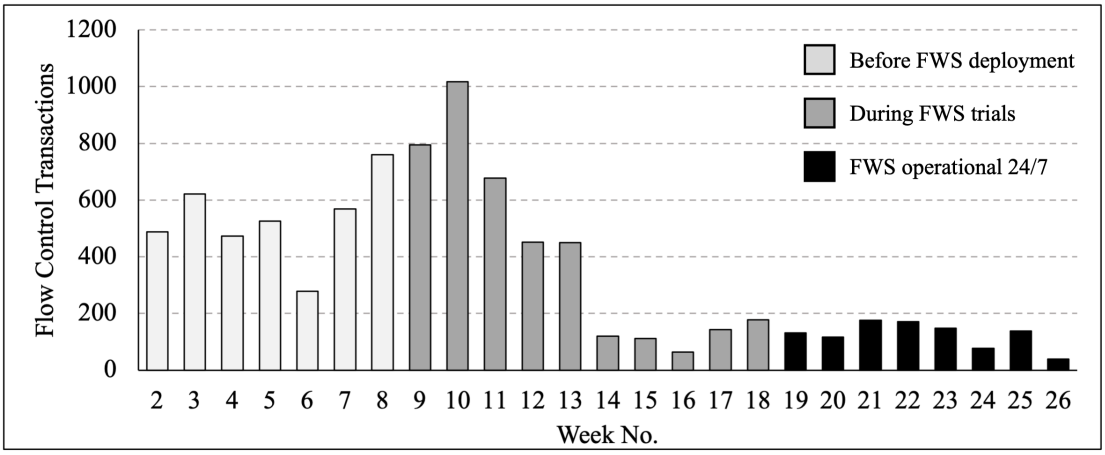


Figure 7: Weekly volume of ad hoc flow management rule transactions.

6.3 Measuring Success in the Fab

The challenges of measuring improvements to fab performance are infamous due to the dynamic and complex nature of fab flows. Initial feedback from the deployment has been positive; manual interrogation of the schedules and the actual lots dispatched has provided confidence in the quality of the decisions made, and a key result has been that the number of "hard down" rules has been greatly reduced. In fact, for the first time in many years at the fab, there was a 24 hour period without the need for this manual intervention. Nevertheless, measuring real-world benefits of the proposed FWS approach requires a systematic evaluation strategy based on actual KPI changes. Beyond the advantages that FWS brings in terms of situational awareness, due to e.g. the improved ability to predict future flows and allocate specialists in an anticipatory manner, we are currently tracking four main indicators for measuring impact:

1. An average increase in moves per shift compared to the prediction made by the fab's retrospective simulation model. This model can provide a counterfactual scenario with which to compare, since it simulates the existing fab dispatch rules while considering retrospective information like actual tool utilisation and availability. As such, we can measure the realised moves against the moves that would be achieved had FWS and TS not been enabled. Nevertheless, this method may well underestimate the benefit, since continuous use of FWS is expected to drive the fab to a more favourable state over time, preventing bottlenecks that may have otherwise occurred.
2. Measuring changes to the distribution of lot lateness (or inversely, the average velocity required to meet due dates) before and after FWS deployment. The bottleneck avoidance behaviour and consideration of lot lateness in prioritisation are expected to reduce the mean and variance of required velocity for production lots across the fab. This is synonymous with an improvement to fab linearity, enabling greater predictability and allowing for more reliable on-time delivery.

3. The distribution of tool time spent in an idle state should have lower mean and variance. We expect to reduce the "starving" of tools by prioritising lots that can go to idle tools instead of bottlenecks.
4. Finally, we expect the mean and variance of queue time at each toolset to decrease, indicating a possible increase in overall throughput.

7 CONCLUSIONS & FUTURE WORK

In this paper we have presented a novel fab-wide scheduling approach for semiconductor fabs. The main idea is to identify wafers that can be slowed down or sped up in order to improve factory KPIs. The solution has been deployed in the Seagate Springtown semiconductor plant with some encouraging first results; some particular cases studies were presented in detail to showcase the scheduler's capabilities. Compared to traditional toolset-level dispatch approaches with no look-ahead, we can see that the FWS module can improve KPIs substantially. However full benefit quantification is an ongoing task. The present paper also focuses on the practical aspects of deploying a novel solution in a critical production environment. We presented the deployment approach, designed with controllability and minimum disruption in mind and discussed how success is measured. We close with a discussion of the future work directions being pursued:

- A key strength of the presented approach is that it can materially benefit from compute resource parallelization to accommodate parallelizable local search heuristics (Steinhöfel et al. 2002). This is particularly applicable to cloud-based platforms that elastic fleets of computational resources.
- Another strategic decision that has a large impact on factory KPIs is determining when to release new lots for processing to achieve prescribed production targets. The FWS model presented can form the basis of a wafer release planning tool by substantially extending the look-ahead horizon.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contribution and support in this project by the whole Flexciton team and particularly Dominic Bealby-Wright, Daniel Cifuentes, Sudesh Lutchman, Yingbao Sun, Diogo Santos, Marcus Vitelli, Charles Thomas and Semya Elaoud for their contributions to this paper. We would also like to thank and acknowledge the following individuals from Seagate Technology for their support during trials: Fergal Hamilton, Carol Brown, Paul McLaughlin, Maureen Doherty and Laura McElhinney.

REFERENCES

Barhebwa-Mushamuka, F. 2020. *Novel Optimization Approaches for Global Fab Scheduling in Semiconductor Manufacturing*. Ph. D. thesis, Université de Lyon.

Bureau, M., S. Dauzere-Peres, C. Yugma, L. Vermarien, and J.-B. Maria. 2007. "Simulation Results and Formalism for Global-Local Scheduling in Semiconductor Manufacturing Facilities". In *Proceedings of the 2007 Winter Simulation Conference*, edited by J. Tew, R. Barton, S. Henderson, B. Biller, M. Hsieh, and J. Shortle, 1768–1773. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Elaoud, S., R. Williamson, B. E. Sanli, and D. Xenos. 2021. "Multi-objective Parallel Batch Scheduling in Wafer Fabs with Job Timelink Constraints". In *Proceedings of the 2021 Winter Simulation Conference*, edited by S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, 1–11. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Hwang, T.-K., and S.-C. Chang. 2003. "Design of a Lagrangian Relaxation-Based Hierarchical Production Scheduling Environment for Semiconductor Wafer Fabrication". *IEEE Transactions on Robotics and Automation* 19(4):566–578.

Kao, Y.-T., and S.-C. Chang. 2018. "Setting Daily Production Targets with Novel Approximation of Target Tracking Operations for Semiconductor Manufacturing". *Journal of Manufacturing Systems* 49:107–120.

Klemmt, A., and L. Mönch. 2021. "Scheduling Jobs with Time Constraints Between Consecutive Process Steps in Semiconductor Manufacturing". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 1–10. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Kopanos, G. M., D. Xenos, S. Andreev, T. O'Donnell, and S. Feely. 2020. "Advanced Production Scheduling in a Seagate Technology Wafer Fab". In *Proceedings of the 2020 Winter Simulation Conference*, edited by K.-H. G. Bae, B. Feng,

S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 1954–1965. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lima, A., V. Borodin, S. Dauzère-Pérès, and P. Vialletelle. 2021. "A Sampling-based Approach for Managing Lot Release in Time Constraint Tunnels in Semiconductor Manufacturing". *International Journal of Production Research* 59(3):860–884.

Mason, S. J., J. W. Fowler, and W. Matthew Carlyle. 2002. "A Modified Shifting Bottleneck Heuristic for Minimizing Total Weighted Tardiness in Complex Job Shops". *Journal of Scheduling* 5(3):247–262.

Mönch, L., J. W. Fowler, S. Dauzère-Pérès, S. J. Mason, and O. Rose. 2011. "A Survey of Problems, Solution Techniques, and Future Challenges in Scheduling Semiconductor Manufacturing Operations". *Journal of Scheduling* 14(6):583–599.

Sarin, S. C., A. Varadarajan, and L. Wang. 2011. "A Survey of Dispatching Rules for Operational Control in Wafer Fabrication". *Production Planning and Control* 22(1):4–24.

Steinhöfel, K., A. Albrecht, and C.-K. Wong. 2002. "Fast Parallel Heuristics for the Job Shop Scheduling Problem". *Computers & Operations Research* 29(2):151–169.

Voas, J., N. Kshetri, and J. F. DeFranco. 2021. "Scarcity and Global Insecurity: The Semiconductor Shortage". *IT Professional* 23(5):78–82.

Zhong, H., M. Liu, and L. Bao. 2021. "A Job-priority Based Soft Scheduling Approach for Uncertain Work Area Scheduling in Semiconductor Manufacturing". *International Journal of Production Research* 60(16):1–17.

AUTHOR BIOGRAPHIES

IOANNIS KONSTANTELOS is a Principal Optimisation Engineer at Flexciton. He holds a PhD in Electrical & Electronic Engineering from Imperial College London. His expertise lies in optimisation and machine learning methods applied to complex manufacturing and energy systems. He is the author of more than 50 journal and conference papers. His email address is ioannis.konstantelos@flexciton.com.

JOHANNES WIEBE is an Operations Research Scientist at Flexciton. He holds a PhD in Computer Science from Imperial College London. His research has focused on optimization under uncertainty and combining data-driven models with mathematical optimization and has been applied to industrial problems including equipment degradation and drill scheduling. He has authored more than 10 scientific publications and multiple open source software tools. His email address is johannes.wiebe@flexciton.com.

ROBERT MOSS is a Senior Technical Product Manager at Flexciton and Deployment Lead for Flexciton's optimised scheduling solution at Seagate Technology. He has considerable software engineering experience with simulation, optimisation, AI and automation in a wide range of areas including wafer fab scheduling and autonomous vehicles. He has a Master of Physics degree (First Class) from the University of Oxford. His email address is robert.moss@flexciton.com.

SEBASTIAN STEELE is a Product Manager at Flexciton. He holds an MEng in Mechanical Engineering from the University of Southampton, focused primarily on optimization and design of experiments. He is responsible for overseeing the FWS project at Flexciton, alongside broader product design and coordination. His email address is sebastian.steele@flexciton.com.

DENNIS XENOS is the founding CTO at Flexciton. With over 10 years' experience in modelling, optimization and engineering, Dennis is responsible for overseeing the technology for optimising semiconductor wafer fabs at Flexciton. He holds a PhD in Chemical Engineering from Imperial College London. He has innovated in the field of the operations optimization of large industrial plants considering data-based and mathematical programming models. He is the author of more than 15 scientific publications, and with Flexciton tech team filed three patents on the optimisation of semiconductor fabs. His email address is dennis.xenos@flexciton.com.

TINA O'DONNELL is a Senior Manager at Seagate Technology in Springtown, Derry, Northern Ireland. She is a critical contributor in the information systems leadership team that has made a transformative impact on wafer manufacturing processes with their cutting edge innovation in automation workflow and simulation/optimisation systems. She holds a BEng degree in Electronic Engineering and Computing and a PhD in Supply Chain Optimisation from Ulster University. Her email address is tina.odonnell@seagate.com.

SHARON FEELY is the Industrial Engineer Manager 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 all Toolsets. She is also responsible for leading and driving the operations research strategy for the global wafer fabs. Her email address is sharon.r.feely@seagate.com.