WHITE PAPER

Superior Scheduling: Hybrid Approach Boosts Margins

Issue one
flexciton.com

Flexciton

Introduction

If semiconductor fabs are to maintain high margins, their costs must continue to fall. Historically, this has been accomplished by shrinking the chip, shifting manufacture to larger wafers and increasing yield. Now little play is left in these levers, so efforts must turn to improving operational processes, as this trims cycle times and increases throughput.

Implementing this is far from easy. Producing semiconductor chips involves many process steps, and there are numerous constraints. This makes scheduling and simulating work-in-progress (WIP) a challenging, mathematically intensive task.

To cut costs and increase efficiency, fabs need scheduling. When fabs are capacity constrained, this increases sales; and when they are market constrained, this drives down production costs.

The two most common scheduling methods are the heuristic approach, which is quick but not optimal; and mathematical optimal results, but takes far too long.

At Flexciton, we are breaking new ground by uniting both these methods with a hybrid technique that delivers excellent results in a matter of minutes. With this approach, which we refer to as an advanced mathematical hybrid-optimization technology, it takes a matter of minutes to creates near-optimal schedules that can increase fab efficiency by up to 10 percent.

Why Is Scheduling Necessary?

To fully appreciate the benefits of our approach, it is critical to understand the complexities of scheduling in a fab. This task is not easy, as schedules need to respect all relevant production constraints that affect process decisions, including reticle constraints and those related to time control. There can be local scheduling, such as WIP movements in photolithography, as well as scheduling at the whole-fab level. To aid visualisation and planning, Gantt charts are often used to indicate lot allocation, tools and duration. When a new lot arrives at a tool it is placed in a queue, waiting until the tool becomes available. Dispatchers examine all the WIP available and determine which lot to process next, using a series of complex algorithmic rules that can be tweaked to suit changing production objectives or expedite priority lots through a fab. Dispatch systems are highly effective production control tools, adopted by many high-volume fabs. However, the rules they employ are fundamentally flawed.
The problem is that decisions are locally optimized, so they fail to account for broader factory behaviour. Actions justified at the local level can impair fab throughput, by having a negative impact on downstream tools.

Importance of scheduling WIP

To avoid downstream bottlenecking, jobs must be planned ahead of time - typically 12-24 hours in advance. This ensures the best decisions for dispatch and ultimately the best overall performance for the fab, in terms of throughput and cycle time. Driving home the importance of great scheduling is the success this delivered at Samsung's DRAM fab.
By slashing the average layer cycle time for DRAM products at this fab by more than 18%, Samsung generated $1 billion of additional revenue over five years [2]. Given that around $1 billion is now needed to build a new 300 mm fab, equipped with tools that can cost more than $40 million, there is more need than ever for great scheduling that ensures high fab utilisation.

Scheduling Essentials

Topping the list of requirements for the scheduling of semiconductor facilities are the following criteria

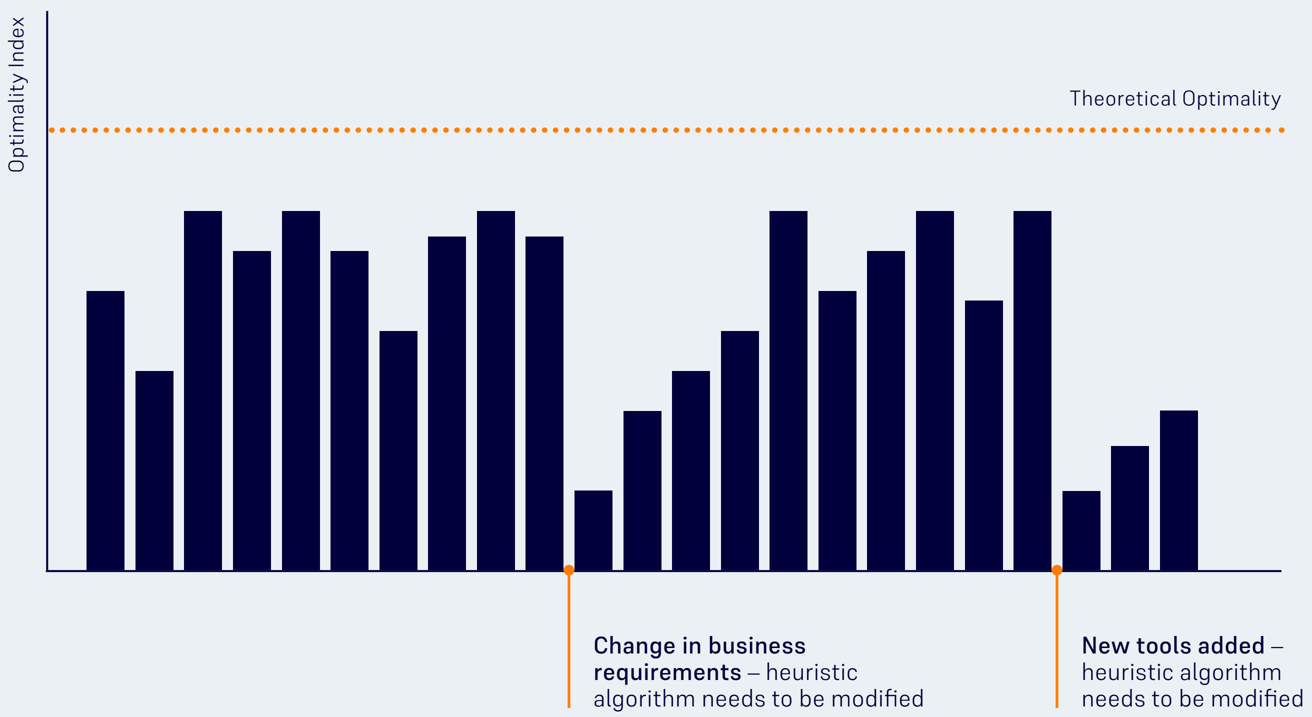
Quick	Fast schedule delivery is critical. If a tool goes down, this can wreak havoc unless a new plan is put in place to maintain high-quality WIP management.
Realistic	Fast schedule delivery is critical. If a tool goes down, this can wreak havoc unless a new plan is put in place to maintain high-quality WIP management.
Optimal	The scheduler needs to consistently obtain the best solutions, with KPIs such as cycle times, throughput and on-time delivery taking priority.
Easy to Maintain	Scheduling algorithms should require little upkeep and manual interference. When the fab workload changes, such as a shift in product mix, the scheduler must adapt automatically and maintain optimality.
Adaptable	Larger chipmakers require scheduling solutions that are easy to transfer to other fabs, and are scalable to multiple sites.

State-of-the-Art Scheduling

Heuristics

Heuristics are rules-based approaches that employ a 'best-guess' to lighten the computational load and shrink the search space. Mathematical-shortcuts are taken, speeding the time it takes to find a solution - but its quality cannot be evaluated. To yield the best results, industrial engineers have to tune and frequently maintain scheduling rules (see Figure 1). This is expensive and it hampers scaling. Yet despite these drawbacks, heuristics are widely deployed in fabs, because they offer a very fast, simple solution to a complex problem.

Figure 1: Heuristics find quick solutions, but are not optimal.



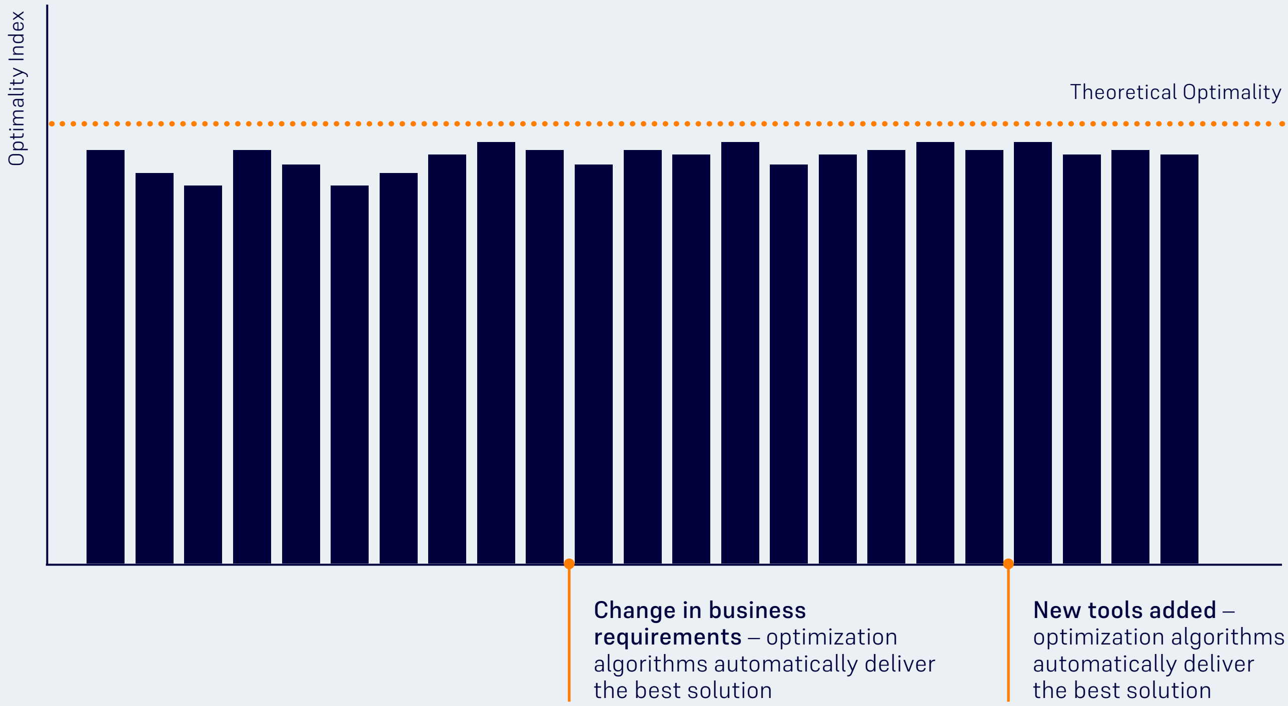
The fundamental flaw of heuristics is that they cannot optimize full-scale scheduling problems. Poor quality decisions result, leading to intermittent tool bottlenecks. The table below shows the other advantages and disadvantages to applying heuristics scheduling in a wafer fab.

Heuristics Scheduling		
Quick	Yes	Feasible schedules are found quickly, using approximations and 'rule of thumb' to shrink the solution search space.
Realistic	Yes	Experienced developers model almost any constraint in the fab. However, heavily constrained heuristics struggle to find feasible solutions, leaving fab managers with a highly suboptimal schedule.
Optimal	No	Heuristics fail to explore all possible options, so suboptimal global solutions result. It is very difficult to judge how far they are from the optimal solution.
Easy to Maintain	No	Heuristic algorithms demand constant updating, redevelopment and 'start from scratch' approaches to accommodate business objectives and production characteristics. For example, the addition of a new tool leads to the distribution of work over more tools, requiring modification of the heuristic to maintain performance.
Adaptable	No	Heuristics are often custom written and uniquely modified for a single fab setup, fab areas or an individual toolset.

Mathematical Optimization

In stark contrast to heuristics is mathematical optimization, which offers a declarative, algorithmic problem-solving approach that consistently delivers the best possible solution. This comes from considering goals and constraints, and accounting for decisions within the control of the fab [3].

Figure 2: A mathematical optimization approach delivers the best results possible, but it requires time to find them.



Optimization targets a well-defined objective, such as maximising the number of moves per hour or reducing cycle time. It is even possible to consider a combination of objectives, such as maximising global throughput.

Mathematical Optimization		
Quick	No	Searching for the global-optimal is extremely computationally intensive. Even with powerful computers, generating a fab scheduling solution to a simple problem can take days.
Realistic	No	Finding a solution in a reasonable timeframe requires shortcuts, which lead to the omission of constraints in the final schedule – and ultimately render mathematical optimization infeasible and unrealistic for day-to-day running of fabs. When solutions are eventually found, the reality has often moved on, making it impossible to implement the changes.
Optimal	Yes	Whilst optimization is slow to compute the full problem, it does uncover the very best solution from the extensive range of scheduling options.
Easy to Maintain	Yes	Although work is required to build the initial model, there is no need to adjust internal algorithms to sustain optimal results. Adding and removing constraints is easy, requiring only a line of code, and it is possible to refine production objectives as often as required. More time is needed to prepare data for constraints than with heuristics, but implementation is far quicker.
Adaptable	Yes	Transferring an optimization model is incredibly easy, because it does not require drastic redevelopment when moving to a different fab.

Which pick?

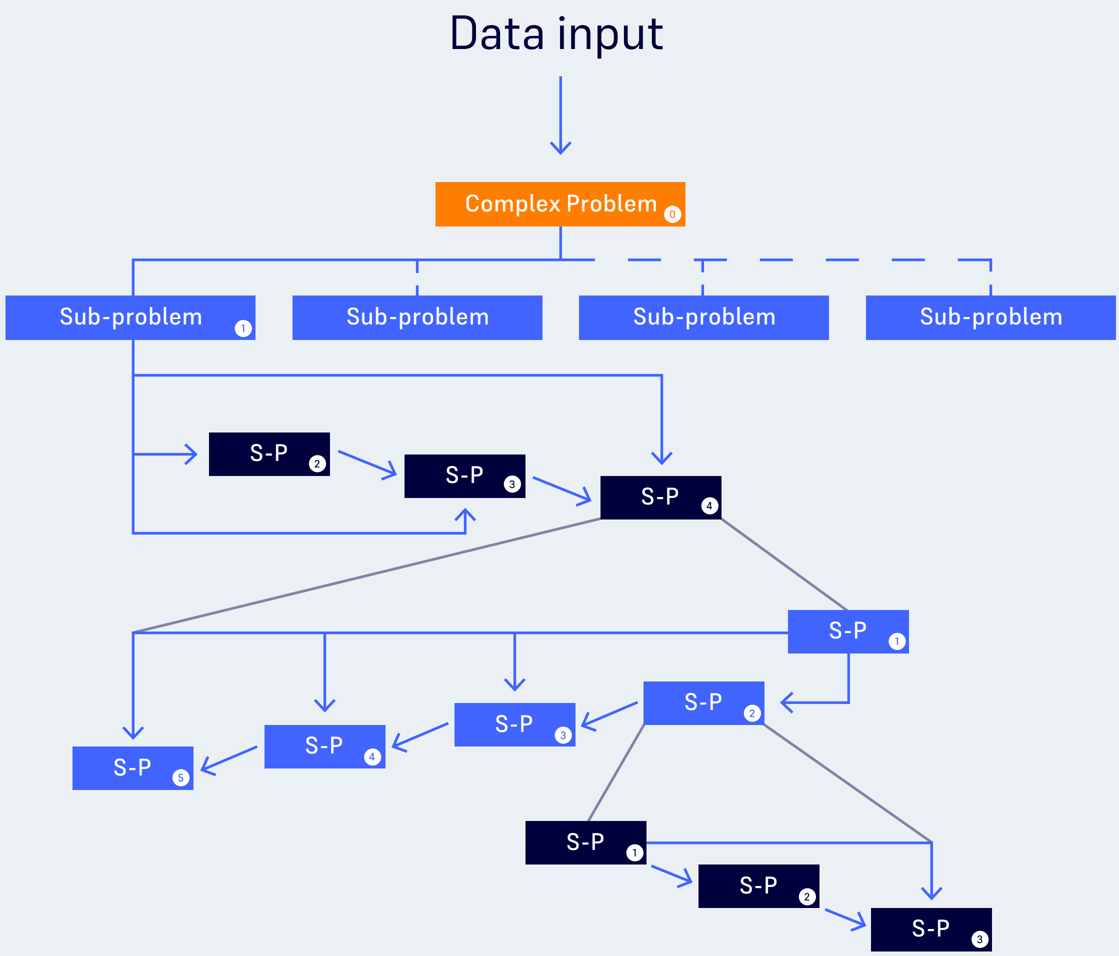
At first glance, given the complexities of wafer fabrication, it would appear that scheduling should be based on mathematical optimization. However, although this approach guarantees the best possible outcome, it is too slow for fabs. So heuristics are adopted. This is a pragmatic choice, aiding scheduling, but holding back efficient production.

A Hybrid Solution

Our approach offers the best of both worlds by providing a fast, very high quality solution to scheduling. Drawing on more than ten years of academic research and industrial studies, we have developed a hybrid-optimization model that solves wafer fab scheduling problems with mixed-integer linear programming (MILP).

By combining heuristic search techniques with optimization, we draw on MILP to obtain excellent solutions in good time. Schedules are close to optimal, thanks to the use of exact and approximate decomposition techniques, which are based on the original full MILP scheduling model.

Figure 3: Flexciton applies decomposition techniques for optimized wafer fab scheduling.



These smart decomposition techniques are critical to the success of our approach. They simplify the problem while retaining near-optimality by breaking down a full-scheduling problem into a series of smaller sub-problems, which are quickly solved by MILP. The solutions to these sub-problems are aggregated to provide a very good overall schedule.

Decomposition can involve identifying toolsets that govern overall throughput, before scheduling them accordingly. This approach, allowing the optimizer to focus on scheduling between bottleneck toolsets while balancing the WIP flow between them, slashes the scheduling search space and the time taken to generate a high-quality solution. Based on our work within this industry, bottlenecks in many fabs are associated with just one-fifth of the tools. By focusing on the flow at these critical toolsets, fabs can realise substantial reductions in cycle time and increases in throughput.

Flexciton Hybrid Optimization		
Quick	Yes	Very complex scheduling problems are solved in under 5 minutes, making this technique applicable to the dynamic fab environment, as updates can keep pace with changes on the factory floor.
Realistic	Yes	Fully accurate schedules are realised by accommodating all constraints. This ensures a true representation of all activity in the fab and its limitations.
Optimal	Yes	By employing MILP, optimization guarantees high-quality solutions. Final solutions are very close to the global-optimal, thanks to performance-enhancing decomposition.
Easy to Maintain	Yes	Since the core of the solution is MILP, very little maintenance is needed to retain high performance. The model consistently optimizes scheduling, regardless of the state and aims of the fab.
Adaptable	Yes	Constraints and parameters can be configured without requiring re-writing or the design of extensive new code — just the addition of a single line of code may suffice. The scheduling solution can be rolled out to multiple fabs, even when they have different production characteristics.

Performance

Solve Time	Using hybrid-optimization, an optimal schedule is produced within minutes.
Solution Complexity	Hard scheduling problems are solved to near-optimality, allowing multiple toolsets to be scheduled over a 6-12 hour time horizon.
Constraints	The number of constraints is unlimited, and can include those related to recipes, set-up times, maintenance, kanbans, reticles/resources, monitor/test wafers, timelinks and batching. Constraints can be configured to meet business objectives and ensure exact representation of each individual fab.
Fab Efficiency	Live implementation into a large-scale 200mm fab running a production process have revealed that hybrid-optimization delivers substantial efficiency improvements. In this fab a small number of key toolsets cause considerable loss of efficiency. Applying our hybrid-optimization scheduling method, with the objective to globally schedule these key toolsets, uncovered the opportunity to trim the cycle time by 7%. Combining this with optimizing wafer starts enables an increase in overall performance by 10%.

Conclusions

Traditionally, fab managers have had to work with second best. While desiring the fidelity of a mathematical approach, they have had to employ heuristics that provide fast, suboptimal solutions.

Offering a practical, superior solution to fab scheduling is our hybrid approach. It delivers an increase in productivity by up to 10%, accomplished by consistently optimising schedules to near-global-optimality, using a low-maintenance approach. The hybrid methodology ensures on-time delivery and superior margins, attributes that give fab owners a tremendous edge over their rivals.

	Heuristics	Mathematical optimization	Hybrid-optimization
Quick	Yes	No	Yes
Realistic	Yes	No	Yes
Optimal	No	Yes	Yes
Easy to Maintain	No	Yes	Yes
Adaptable	No	Yes	Yes

References

[1] Mönch, L., Fowler, J. and Mason, S., 2013. Production Planning And Control For Semiconductor Wafer Fabrication Facilities.

[2] Leachman, R., Kang, J. and Lin, V., 2002. SLIM: Short Cycle Time and Low Inventory in Manufacturing at Samsung Electronics. Interfaces, 32(1), pp.61-77.

[3] Gurobi. 2020. 4 Key Advantages Of Using Mathematical Optimization Instead Of Heuristics - Gurobi. [online]