Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

A STUDY ON THE INTEGRATION OF COMPLEX MACHINES IN COMPLEX JOB SHOP SCHEDULING

Karim Tamssaouet
Stéphane Dauzère-Pérès
Claude Yugma

Mines Saint-Etienne, Univ Clermont Auvergne
CNRS, UMR 6158 LIMOS
CMP, Department of Manufacturing Sciences and Logistics
Gardanne, 13120, FRANCE

Sebastian Knopp

AIT Austrian Institute of Technology
Center for Mobility Systems
Vienna, 1210, AUSTRIA

Jacques Pinaton

STMicroelectronics Rousset
Rousset, 13790, FRANCE

ABSTRACT

In this paper, we study the problem of considering the internal behavior of complex machines when solving complex job-shop scheduling problems encountered in semiconductor manufacturing. The scheduling problem in the diffusion area is presented, and the complex structures and behaviors of the different machine types in this area are described. Previous related research is reviewed, and our approach to consider the complexity of these machines with batching constraints when scheduling the diffusion area is described. The main part of the paper presents and discusses numerical experiments on industrial instances to show the benefit of choosing a suitable modeling for complex machines.

1 INTRODUCTION

The production of microelectronic devices is a highly complex and cost-intensive process, particularly in the front-end phase where the fabrication of wafers takes place. In this context, the assignment of lots of wafers and their sequencing on resources have a substantial impact on key performance indicators such as throughput and cycle time (Mönch et al. 2011). In today's semiconductor manufacturing facilities (fabs), dispatching rules are still often used for these decisions and scheduling solutions are not widely applied (Mönch et al. 2012). Because of the high degree of automation that allows for automated real-time data collection and the constantly increasing computation power of modern computers, scheduling algorithms promise better operational performance than myopic common dispatching systems (Mönch et al. 2011).

Among the complex workshops in a wafer manufacturing facility (wafer fab), the diffusion area is of critical importance. The processes in this area are performed on two types of machines: Wet benches and furnaces. These machines can process several lots simultaneously. The processing times can be very long (several hours) compared to other operations in the fab. Many other constraints have to be considered so that the proposed schedules are feasible, including maximum and minimum time constraints, usually from the end of an operation of a lot to the start of its next operation, availability constraints, sequence-dependent

978-1-5386-6572-5/18/$31.00 ©2018 IEEE

setup times and other specific constraints. Multi-objective optimization also has to be considered to deal with different conflicting criteria.

This paper deals with complex scheduling problems arising in the diffusion area. Our focus is on the modeling of furnaces and wet benches that have complex behavior. Relevant differences between these complex machines and cluster tools exist. However, as these two classes of tools share a complex behavior (Rotondo et al. 2015), studies conducted on cluster tools can provide useful insights into the behavior of diffusion machines and their modeling. Most of the scheduling literature regarding cluster tools deals with internal scheduling (Mönch et al. 2011). Our objective is to solve scheduling problems for the whole diffusion area, i.e., to simultaneously determine the assignment and the sequencing of jobs waiting in front of machines. Regarding the modeling of cluster tools, there are in the literature two ways of dealing with the external scheduling of these complex tools. The difficulty lies in the fact that different sequences of lots will lead to different operation cycle times, i.e., the time between the processing start of the operation and its processing end on the machine. The first way consists in using a detailed simulation model of a cluster tool to evaluate the cycle times for job sequences for the scheduling algorithm, such as for example in (Dümmler 1999). In the second way of modeling cluster tools, cycle time approximations are used, e.g., (Niedermayer and Rose 2004).

In order to choose a suitable modeling approach for real applications, different practical aspects should be considered. The first and direct impact of a given modeling is the prediction accuracy, that can be defined as the difference between the planned and realized times. In the context of the diffusion area, several aspects make the prediction accuracy critical. Most of the jobs in wet-etch operations must proceed for further processing to furnaces without exceeding a maximum time lag between the two batching processes. As efficient schedules for wet-etch operations are essential to ensure high productivity at the furnaces (Ham and Fowler 2008), accurate prediction of completion times on wet bench machines ensures that the satisfaction of the time constraints when the realized times are not far from what is predicted by the scheduling algorithm. Also, the prediction accuracy is crucial when considering batching constraints. Errors in job arrival times to furnace operations reduce the expected improvement from a scheduling algorithm over dispatching rules, e.g., (Kohn and Rose 2013). The second aspect that should be considered when choosing a modeling is its impact on the efficiency of the scheduling algorithm. Having only few minutes to return a solution, it is important to make sure that the solution evaluation is not too time-consuming with the adopted modeling. Third, the resulting complexity of the scheduling system should be considered. In real applications, modeling a scheduling problem and developing an optimization algorithm to solve it is only part of the story. The scheduling system has to be integrated into the information system of the enterprise, which can be a formidable task (Pinedo 2016).

In this paper, we propose a new analytical modeling of complex machines in the diffusion area by generalizing the modeling approach of (Knopp et al. 2014) to the case of batch processing machines. This modeling is compared with a data-driven analytical modeling using industrial instances. Simulation models are not considered in this study as a comparison base for different reasons. First, it is not easy to find the necessary knowledge to understand and implement the internal scheduling algorithm of complex machines, which is proprietary to the manufacturers of the machines. Second, from an industrial perspective, using simulation within a scheduling algorithm can be prohibitive regarding the algorithm efficiency as it must periodically call the simulation model to evaluate the found solutions.

The scheduling problem in the diffusion area is described in Section 2, and the structure and behavior of furnaces and wet benches are detailed. In Section 3, the main components of our solution approach are summarized. In Section 4, using a small industrial instance, the proposed modeling is compared to a data-driven analytical model of a wet bench machine, showing how it naturally leads to throughput optimization without any data preprocessing. Section 5 is devoted to the analysis of numerical experiments performed on large industrial instances. Section 6 concludes the paper and provides some perspectives.

2 PROBLEM DESCRIPTION

This work deals with complex scheduling problems arising in the diffusion area in semiconductor manufacturing, but can be applied to other industrial contexts. Let us consider a set of jobs with different sizes and different priorities, and that arrive to the work area dynamically, i.e., jobs have release dates. Each job is associated with a route that specifies the sequence of operations the job must go through. It is possible for a job to visit some machines more than once and this situation is called in literature reentrant flows (Graves et al. 1983). Each operation can only be performed on a set of qualified machines, and its processing time depends on the selected machine. There are time lag constraints, both minimum and maximum, between operations of the same job, not necessarily successive. Maximum time lags limit the waiting time between two operations of the same job. Minimum time lags prescribe a minimum waiting time between two operations of the same job. A significant constraint to consider is parallel batching, abbreviated as p-batching, which refers to the capability of machines to process more than one job at the same time. Let us focus on the version where all the jobs in a batch are processed together, start at the same time and have the same processing time. On a qualified batching machine, an operation can only be batched with operations that belong to the same batch family. Machines might be unavailable during a fixed time period in the scheduling horizon, to model for example preventive and curative maintenance. Some machines are subject to sequence-dependent setups, i.e., the required time to configure a machine to process an operation depends on the machine configuration used by the preceding operation. As in (Yugma et al. 2012), we optimize multiple criteria, among others: The number of processed wafers within the given scheduling horizon and the batching coefficient that corresponds to the filling of constituted batches in the scheduling horizon (calculated as the number of processed wafers divided by the sum of the number of batches performed on each machine, times the maximum capacity of that machine) (Yugma et al. 2012).

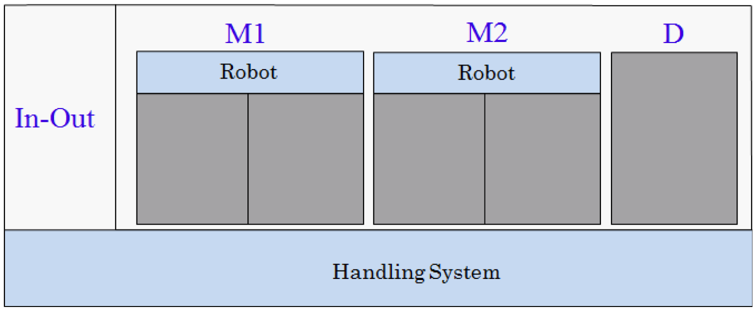
When it comes to proposing a solution for this complex scheduling problem, an important question arises: How to model the details of real-world machines and how to decide on the suitable degree of detail? In the diffusion area, there are two main types of machines: Furnaces and wet benches. Furnaces perform deposition, oxidation and annealing processes using tubes, boats, and robots. A tube is the reactor chamber within which processes take place. A boat is a movable carrier for wafers inside a tube. Boats are used to load, unload and cool wafers. Robots deal with loading and unloading of wafers. There are different kinds of furnaces: Those with one tube and one boat, furnaces with two tubes with one boat for each tube, and furnaces with two tubes and two boats for each tube. A wet bench is a tool used to carry out wet cleaning and etching operations. Such tools are capable of batching, so multiple wafers from different lots can be processed at the same time. Wet benches commonly include several modules (also called tanks or "baths"), each containing either cleaning or etching solutions, and water in a rinsing tank. As for furnaces, there are different types of wet benches, i.e., with different structures.

It is not realistic to only associate fixed processing times to operations performed on these complex machines. Different reasons make it necessary to precisely compute cycle times of operations: The possibility to have different batches running in parallel, the absence or the limited capacity of internal storage, the presence of bottleneck components, the presence of internal scheduling algorithms, more or less sophisticated, and the diversity of processes with different times. To illustrate the different modeling alternatives of these complex machines, we use the example of one type of wet bench through this paper.

A schematic representation of this machine type is given in Figure 1. The maximum batch size is two lots. The processing part consists of two processing modules (M_1, M_2) and a dryer D in which the wafers are rinsed and dried. Each module consists of its own robot and two tanks. Some processes, called *short processes* in this paper, require one of the processing modules and the dryer before unloading the wafers from the machine. *Long processes* require all the modules following this sequence ($M_1 \to M_2 \to D$). Modules M_1 or M_2 can be unavailable without making the whole machine unavailable.

In addition to the two robots dealing with the wafer handling inside the two modules, the wafer handling system consists of: A loading port and an unloading port, a robot that batches the lots after they are loaded and "un-batches" them after they are processed, and a robot that transports the wafers between the different

Figure 1: Example of the structure of a wet bench machine.



components. For ease of presentation, we only study the modeling of the processing part. In practice, instead of modeling explicitly the handling part, handling times are included in the processing times. Contrary to furnaces where the wafers are handled individually during the boat loading, all the wafers of a batch are handled together within wet benches. Also, because batches are transported over small distances, handling times are small compared to times that batches spend in the processing modules.

3 SOLUTION APPROACH

Our solution approach relies on two adaptations of the disjunctive graph, a widely used representation of scheduling problems introduced by (Roy and Sussmann 1964). In this graph, nodes model operations and conjunctive and disjunctive arcs model precedence constraints between operations. Section 3.2 briefly recalls the approach proposed by (Knopp et al. 2017) in order to integrate p-batching constraints in an efficient way. Then, Section 3.3 recalls how (Knopp et al. 2014) model the detailed behavior of complex machines through the structure of the graph. As all the machines in our context are batching machines, the two adaptations are combined in a generalized representation, briefly presented in Section 3.4. Before this, Section 3.1 gives a brief description of the data-driven modeling to which the proposed approach is compared.

3.1 DATA-DRIVEN ANALYTICAL MODEL

In the current version of the scheduling application we are developing, a data-driven analytical modeling is chosen for wet benches that are modeled as black boxes. Operation processing times are computed as the averages of historical cycle times. Due to the parallel processing of the batches, instead of directly using historical process cycle times as operation processing times, a factor $\alpha < 1$ is estimated and used to compute the *actual* processing times. Considering the wet bench taken as an example, as it is possible to have three batches inside the processing part, the actual processing time is considered to be equal to $\alpha = \frac{1}{3}$ times the processing time. Inefficiencies, mainly due to blocking constraints, that are generated by the succession of two operation are represented as sequence-dependent setup times. After classifying wet-etch processes, experts working in this area provided us with estimated inefficiencies that are generated by the succession of two process classes. However, this is not enough as the completion time of an operation is not only influenced by its direct first predecessor, but also up to its fourth predecessor. To deal with this, we use simple decision rules that are given to operators on the shop floor. A set of *efficient* sequences with a length up to five operations are displayed for operators in front of the concerned machines. *Inefficient* sequences are also identified. To make the scheduling algorithm promoting efficient sequences, setup times between successive operations are adjusted to make sure that the difference between the largest sum of setup times among efficient sequences and the smallest sum of setup times among inefficient ones is maximized. After formulating constraints on what is considered as acceptable changes, an integer linear program is used to obtain these adjustments. Finally, the operation processing time is used as the duration of the minimum time lag, introduced to force jobs to stay at wet-etch operations before moving to the next operation in their routes.

3.2 BATCH-OBLIVIOUS GRAPH

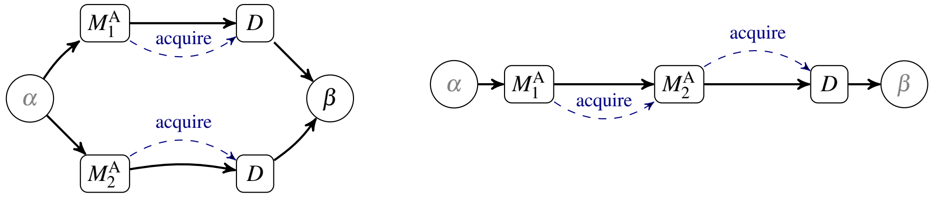
Most existing solution approaches for complex job-shop scheduling problems with batching machines rely on the disjunctive graph representation of (Ovacik and Uzsoy 2012). This representation introduces dedicated nodes to represent batching decisions explicitly. A novel modeling approach, called batch-oblivious, is presented by (Knopp et al. 2017). As in a classical conjunctive graph, the batch-oblivious conjunctive graph uses nodes to uniquely model operations and arcs to model precedence constraints on routes and resources. Instead of inserting additional nodes and arcs, batches are encoded in the arc weights. This new representation has several advantages. It reduces the structural complexity of the graph and allows reusing ideas and techniques for "simpler" problems such as the move proposed by (Dauzère-Pérès and Paulli 1997) for the flexible job-shop scheduling problem. Last but not least, an integrated construction algorithm is proposed in (Knopp et al. 2017) that simultaneously computes start dates and improves the solution during the graph traversal to compute start and completion times by filling underutilized batches through a combined re-sequencing and reassignment strategy.

3.3 ROUTE GRAPH MODELING

In the real-world problem, we are given a linear route of processing operations for each job. Each operation must be assigned to a machine that must be chosen from a subset of qualified machines. In order to model machines in detail, their internal components should be modeled explicitly. For clarity, an *operation* corresponds to the processing of a job within a whole machine like a furnace or a wet bench. The term *step* is used to describe the elementary processing of a job within an internal component of a machine. Hence, for a given operation on a machine, there is a linear sequence of steps that use the internal components of the machine. For a solution to be feasible, it is necessary to ensure that, when a step of an operation of a job is performed by a component of a machine, all the subsequent steps must be performed on the components of the same machine. To include these dependencies between internal components of complex machines, the concept of *route graph* is introduced in (Knopp et al. 2014). For each job, this graph is constructed during the preprocessing phase, i.e., before solving the problem.

If a machine is modeled in details, each of its components is represented as a resource. Each processing operation on a complex machine is decomposed into separate steps describing the different elementary transformations that take place inside the machine. Instead of dynamically assigning resources to operations and steps during the initial solution construction or in the neighborhood search, resources are statically assigned. Now, each decomposition of a processing operation into a set of paths of steps yields a partial route graph: The paths of steps are inserted between two separator nodes. If a machine is not complex, it is represented as a resource and the operations it performs are represented as one step. Finally, the partial route graphs of the job processing operations are concatenated. Instead of having a linear path of nodes representing the operations for which resources are dynamically assigned by the scheduling algorithm, each job is associated with a *two-terminal series parallel graph* (Eppstein 1992). The disjunctive graph representing the problem to solve is obtained by combining all the route graphs of jobs in the problem, in addition to the classical start and end dummy nodes. The conjunctive graph representing a solution is obtained by choosing, for each job, a path in its route graph and sequencing the selected operations on their assigned resources. Unscheduled operations, i.e., those nodes that are not part of a route selection, remain a part of the graph but are disconnected.

This basic definition is extended by an additional constraint called *resource acquisition*. This constraint is used in the case of two operations, not necessarily successive, that are part of the same route and require a common resource. It represents the situation where the concerned resource is exclusively acquired between two operations; it is prohibited for other operations to use the resource in between. Figures 2a and 2b illustrate two route graphs of two jobs that go only through a wet cleaning operation using, respectively, a short process and a long process on the wet bench machine.



(a) Short recipe

(b) Long recipe

Figure 2: Modeling of a wet bench machine.

As already specified, the chosen wet bench machine consists of two modules (M_1, M_2) and a dryer D. First, let us consider the case of a short process that uses one of the processing modules and module D. The modeling of an operation of a job that should be performed on this machine is shown in the route graph of Figure 2a. Since resources are statically assigned to operations, the possibility of performing the processing part of a short process is expressed through the flexibility of choosing one of the sequences in the route graph. Resource acquisitions are indicated by superscript “A” and a dashed line (which is not an edge of the route graph) to the release operation. The scheduling algorithm is allowed to use resource M_1 only when resource D becomes available so that the current job can start to be rinsed. Resource acquisition constraints are used to express blocking constraints that model the absence of storage capacity between machine components (Hall and Sriskandarajah 1996). As a consequence of this absence of storage capacity, a batch that completes a step remains on the machine component until a downstream component becomes available for processing. The same notations, given above for the case of short processes, can be used to understand the modeling of long processes, i.e., processes that follow this sequence ($M_1 \rightarrow M_2 \rightarrow D$) as shown in Figure 2b.

3.4 BATCH-OBLIVIOUS ROUTE-AWARE GRAPH

A more detailed presentation of the two types of graphs described above can be found in (Knopp 2016). The batch-oblivious graph is also generalized in order to take into consideration the route graph concept. However, the proposed approach cannot simultaneously handle the detailed modeling of machines and their batching capability. As batching constraints are critical in the diffusion area, the proposed approach forbids decomposing batch operations into separate steps. To remove this restriction, we propose a new solution graph, called *batch-oblivious route-aware graph*. This new graph and the associated approach are not detailed in this paper, but its impact is discussed in the numerical experiments of Section 4 and Section 5.

In order to model complex batching machines, it is worth pointing out that batches are formed outside the machines. The jobs of the same batch stay together during all the steps of the batch in the machine, i.e., they all use the same machine components and simultaneously. Knowing this, the integrated algorithm within the batch-oblivious approach is given the freedom to constitute batches in the first step in the complex batching machine and then constrained to keep the same batch during the subsequent steps. To do this, local changes must be performed in the graph to make sure that the jobs remain together during all the steps. When doing this, it is necessary to ensure that there is always a way to modify the graph in order to respect batching constraints without creating a cycle. Using this generalized representation, it is then possible to model in detail complex batching machines in our context such as wet benches and furnaces and solve the scheduling problem in the whole diffusion area. Currently, this approach can be used within the construction heuristic described in (Knopp et al. 2017). Within this heuristic, jobs are initially sorted in decreasing order of the sum of the shortest processing durations of their operations. The heuristic then iterates over the sorted list of jobs and successively inserts all operations of the current job. The operations of a job are greedily inserted, starting from the first operation, by selecting the best insertion position for each operation. The best insertion position is determined by the objective function value of the partial solution obtained by actually inserting the considered operation. The construction is completed when all operations of all jobs have been inserted. Like the batch-oblivious approach, the proposed generalized approach can be embedded in a heuristic based on the idea of the Greedy Randomized Adaptive Search Procedure (GRASP). Simulated annealing can also be used to improve initial solutions. Due to the blocking constraints that make the problem harder, moves that create cycles in a classical way and those that lead to deadlocks must be characterized. As this work is in progress, the approach is only used within the described construction heuristic in the experiments of Section 5.

4 EXPERIMENTAL RESULTS ON SMALL INSTANCES

The objective of this section is to show how using the approach summarized in Section 3.4 naturally leads to throughput optimization and accurate prediction without making the scheduling system complex. A small industrial instance is used to compare the data-driven analytical modeling and the one that uses the batch-oblivious route-aware graph. Table 1 provides four examples of recipes with the elementary processing times, given in minutes, on wet bench modules. The three first recipes (1, 2 and 3) are short processes while the last one is a long process. The first step of recipes 2 and 3 can be performed either on Module M_1 or on Module M_2. For Recipe 1, the first step can only be performed by Module M_1. The considered problem is described in Table 2. Eight jobs have to be scheduled on a wet bench machine, four of which require the same short process (Recipe 1) and four others the same long process (Recipe 4). Column *"Possible Sequences"* provides the sequence of steps each operation follows in the wet bench machine. Note that the first step of the short process can only be processed by Module M_1. In order to optimize machine throughput, the optimization criterion is the makespan.

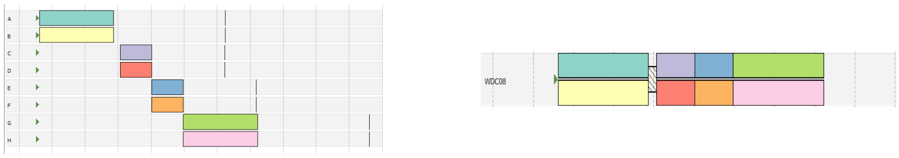
Table 1: Example of processes on a wet bench machine.

Recipe	Type	M_1	M_2	D
1	Short	13	-	15
2	Short	38		15
3	Short	21		15
4	Long	16	47	15

Table 2: Small size industrial problem instance.

LotID	**RecipeID**	**Recipe Type**	**Possible Sequences**
A	4	Long Recipe	$M_1 \rightarrow M2 \rightarrow D$
B	4	Short Recipe	$M_1 \rightarrow M2 \rightarrow D$
C	1	Short Recipe	$M_1 \rightarrow D$
D	1	Short Recipe	$M_1 \rightarrow D$
E	1	Short Recipe	$M_1 \rightarrow D$
F	1	Short Recipe	$M_1 \rightarrow D$
G	4	Long Recipe	$M_1 \rightarrow M_2 \rightarrow D$
H	4	Long Recipe	$M_1 \rightarrow M_2 \rightarrow D$

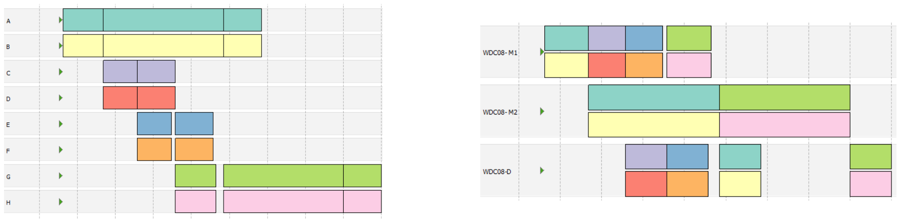
Using the described data-driven analytical modeling, the optimal solution is represented through a Gantt chart from job perspective in Figure 3a and from a resource perspective in Figure 3b. Each color represents a distinct job, rectangles represent operations and parallel rectangles batches. In this solution, four batches are constructed, each containing two jobs. The actual processing time of each operation is represented by the length of the colored rectangles. In Figure 3a, vertical lines to the right side represent the constraint that forces each job to wait during the cycle time before going to the next operation in its route. In Figure 3b, a setup time is represented as a small rectangle with a downward diagonal pattern.



(a) Job perspective Gantt Chart.

(b) Machine perspective Gantt Chart.

Figure 3: Scheduling 8 jobs on a wet bench machine, using data-driven analytical modeling.



(a) Job perspective Gantt Chart.

(b) Machine perspective Gantt Chart.

Figure 4: Scheduling 8 jobs on a wet bench machine, using route graph modeling.

Using the route graph modeling, the optimal solution is shown on a Gantt chart from a job perspective in Figure 4a and from a resource perspective in Figure 4b. Different aspects of the complex behavior can be identified in these figures. First, note that this modeling can represent the possibility of parallel processing of multiple batches. Even if the batch containing jobs A and B is the first one to be loaded on the machine, it is ready to be unloaded only when the two next batches are unloaded. Finally, the consequence of the blocking constraint can be identified in Figure 4b. Considering the step sequence on Module M_1, the last batch in the sequence, containing jobs G and H, does not start its processing directly after its predecessor even if it is available at the beginning of the scheduling horizon. This is because Module M_1 is blocked by the batch containing jobs E and F. This last batch is waiting for the batch containing jobs D and E to free the dryer so that it can free Module M_1.

The two considered models allow the same optimal loading sequence to be obtained in terms of machine throughput. However, let us compare the models according to other practical aspects. The obtained prediction accuracy using data-driven analytical modeling is not satisfactory. This can be explained by the use of rough estimations and their adjustment, time averages and a loss of a relevant information after too many data aggregations. Regarding our proposed modeling, prediction accuracy is satisfactory as the machine is modeled in detail and the processing times are explicitly considered for each process type, i.e., modifying these processing times leads to different process types. Considering the fast-changing environment, the data-driven analytical modeling is also not satisfactory regarding the scheduling system complexity. This modeling should be periodically reviewed in order to update process classification and time estimations, which could be reflected in the implementation of the database management subsystem. In the same way, efficient and inefficient sequences should be reviewed together with the estimated setup times. In addition, as the internal modules can be unavailable without making the whole machine unavailable, data extraction has to be conditioned by the production status of these internal modules. The processing and setup times have to be adapted to the current configuration: Both modules are available, one of the modules is unavailable or both modules are unavailable. This leads to a complex scheduling system and more effort to maintain it. Using our proposed modeling, this additional burden and complexity can be avoided. As operation times are adjusted to consider unavailability periods on their assigned simple machines, step times are also adjusted to consider unavailability periods on their assigned complex machine components.

5 EXPERIMENTAL RESULTS ON LARGE INDUSTRIAL INSTANCES

The objective of this section is to show the impact of complex machine modeling on the quality evaluation of solutions. As optimization approaches rely on this evaluation to explore the solution space, inaccurate evaluations can lead optimization algorithms to discard good solutions, or to return solutions that lead to poor performance on the shop floor. To do this, as in Section 4, the data-driven analytical modeling, where machines are modeled as black boxes, is compared to the batch-oblivious route-aware graph modeling. The same deterministic construction algorithm is used to propose two solutions, one for each modeling approach, for each industrial instance. The batch-oblivious approach, described in Section 3.2 allows batching decisions to be optimized during the construction of the schedule. Thanks to the proposed approach summarized in Section 3.4, that combines the batch oblivious approach with the route graph modeling, batch processing machines are modeled in detail.

To conduct the experiments, 10 large instances are extracted from the Manufacturing Execution System (MES) of a STMicroelectronics fab over a period of three months. The number of jobs per instance is between 580 and 850. For each job, between one and seven operations have to be performed. On average, there are 135 batch families. Jobs must be scheduled on 70 machines that are all capable of processing multiple operations in the same batch. When considering batch-oblivious route-aware graph modeling, only wet bench machines are concerned (9 machines over 70). The scheduling horizon is fixed to 8 hours. The considered optimization criteria are batching coefficient, number of processed wafers during the scheduling horizon and total weighted completion time. This last criterion is computed for all the jobs, unlike the first two ones that are only computed based on the operations that start their processing within the scheduling horizon of 8 hours. Also, the objective is to minimize the total weighted completion time, while the batching coefficient and the number of processed wafers within the scheduling horizon are maximized. The experiment results are summarized in Table 3. For each criterion and for each instance, the results are given in terms of the relative deviation of the value obtained using the data-driven analytical modeling from the value obtained using the batch-oblivious route-aware graph modeling.

Table 3: Comparison between analytical modeling and proposed route graph modeling.

Instance	**Batching Coefficient**	**Number of processed wafers**	**Total Weighted Completion Time**
1	-5.1%	-13.1%	94.7%
2	-8.3%	-2.6%	67.9%
3	-6.1%	-13.6%	118.5%
4	-5.1%	-5.1%	106.3%
5	-8.2%	-11.7%	137.9%
6	-10.1%	-15.1%	108.7%
7	-5.5%	-11.7%	81.9%
8	-4.8%	-1.3%	75.0%
9	-3.8%	-5.1%	88.3%
10	-2.9%	-0.4%	51.7%

Looking at relative errors on the batching coefficient, the data-driven analytical modeling leads to less filled batches for all instances. The largest error is for Instance 6 and the smallest one is for Instance 10. This can be explained by the fact that the data-driven analytical modeling returns pessimistic estimations on the completion times of wet bench operations. The pessimistic estimation is induced by different factors. The first reason is the use of artificial setup times (up to 30 minutes for some sequences) to model inefficiencies and to force the scheduling algorithm to propose schedules that optimize wet bench throughput. The second reason is the use of historical cycle times as the minimum duration a job must stay in wet bench operations. As these historical data are fixed and already include inefficiencies that are induced by the different historical job combinations, the inefficiencies are always considered, even if the job starts it processing on an empty machine. As most wet bench operations are followed by furnace operations, pessimistic completion times mean pessimistic job arrival times at furnaces, and thus the chance to complete a non-full batch is reduced. As the number of furnaces is larger than the number of wet benches, incomplete batches on furnaces degrade the batching coefficient in the whole area. These lower batching coefficients are part of the explanation of the results on the number of processed wafers. In the third column of Table 3, the data-driven analytical modeling predicts from 0.4% up to 15.1% less processed wafers than the batch-oblivious route-aware modeling. Recall that operation durations in the diffusion area are very long compared to other operations in the fab, especially those of furnace operations that can last up to 10 hours. As the scheduling horizon is fixed to 8 hours, it is important to fill batches in order to maximize the number of processed wafer. Finally, when considering the minimization of the total weighted completion time, the relative errors range from 51.7% to 137.9%. This can be explained by the combination of the previous reasons.

These results show that the chosen modeling has a significant impact on the solution evaluation. It is not the purpose of these experiments to show that it is impossible to find a data-driven analytical modeling that can lead to satisfactory results. It may be possible to propose a data-driven analytical modeling of complex machines that provides better solution evaluation. As the analyzed data-driven analytical modeling requires a lot of preprocessing, this section points out that it is probably difficult to find such a modeling and to use it in an industrial scheduling solution.

6 CONCLUSION

This paper studied the relevance of considering the detailed modeling of complex machines in solving complex job-shop scheduling problems. Two previous research works separately conducted on modeling in detail the internal behavior of machines using route graphs and on solving complex job-shop scheduling problems with batching were recalled. Then, an approach combining these two contributions was summarized. We then mainly focused on showing the relevance of the detailed modeling on small instances, to better explain the results, and on large industrial instances, to show the impact on the selected criteria. It is possible to see that wrong schedules might be obtained when replacing the detailed modeling by a simpler data-driven analytical modeling of the behavior of the machines.

In our future work, we intend to finalize the approach summarized in Section 3.3 and run our experiments on at least two cases. In the first case, the GRASP metaheuristic will be run without the batch-oblivious route-aware modeling, and the best solution found will be recomputed at the end using the batch-oblivious route-aware modeling. In the second case, the GRASP metaheuristic of Section 3.3 will be used, and thus will be slower and explore less solutions than in the first case because the batch-oblivious route-aware modeling is always considered, but the best solution will be feasible. Comparing the solutions obtained with the two cases should help to decide which approach is the most relevant.

ACKNOWLEDGMENTS

We would like to thank all contributors from STMicroelectronics who supported our work with valuable information. Also, we would like to thank the anonymous reviewers for their suggestions and comments.

REFERENCES

Dauzère-Pérès, S., and J. Paulli. 1997, Apr. “An Integrated Approach for Modeling and Solving the General Multiprocessor Job-Shop Scheduling Problem Using Tabu Search”. *Annals of Operations Research* 70(0):281–306.

Dümmler, M. A. 1999. "Using Simulation and Genetic Algorithms to Improve Cluster Tool Performance". In *Proceedings of the 31st Conference on Winter Simulation: Simulation*, edited by P. F. et al. D. Sturrock, and G. Evans, WSC '99, 875–879. New York, NY, USA: ACM.

Eppstein, D. 1992. "Parallel Recognition of Series-Parallel Graphs". *Information and Computation* 98(1):41–55.

Graves, S. C., H. C. Meal, D. Stefek, and A. H. Zeghmi. 1983. "Scheduling of Re-entrant Flow Shops". *Journal of Operations Management* 3(4):197–207.

Hall, N. G., and C. Sriskandarajah. 1996. "A Survey of Machine Scheduling Problems with Blocking and No-wait in Process". *Operations research* 44(3):510–525.

Ham, M., and J. W. Fowler. 2008. "Scheduling of Wet Etch and Furnace Operations with Next Arrival Control Heuristic". *The International Journal of Advanced Manufacturing Technology* 38(9-10):1006–1017.

Knopp, S. 2016. *Complex Job-Shop Scheduling with Batching in Semiconductor Manufacturing*. Ph. D. thesis, Université de Lyon, Department of Manufacturing Sciences and Logistics, Gardanne, France.

Knopp, S., S. Dauzère-Pérès, and C. Yugma. 2014. "Flexible Job-shop Scheduling with Extended Route Flexibility for Semiconductor Manufacturing". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. T. et al., WSC '14, 2478–2489. Piscataway, NJ, USA: IEEE Press.

Knopp, S., S. Dauzère-Pérès, and C. Yugma. 2017. "A Batch-Oblivious Approach for Complex Job-Shop Scheduling Problems". *European Journal of Operational Research* 263(1):50 – 61.

Kohn, R., and O. Rose. 2013. "The Impact of Accuracy in Lot Arrival Prediction on Solution Quality for the Parallel Batch Machine Scheduling Problem in Wafer Fabrication". In *Simulation in Produktion und Logistik 2013*, edited by D. W. et al., 121–132. Paderborn: Heinz Nixdorf Institut.

Mönch, L., J. Fowler, S. Dauzere-Peres, S. 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.

Mönch, L., J. Fowler, and S. Mason. 2012. *Production Planning and Control for Semiconductor Wafer Fabrication Facilities: Modeling, Analysis, and Systems*, Volume 52 of *Operations Research Computer Science Interfaces Series*. New York: Springer-Verlag.

Niedermayer, H., and O. Rose. 2004, May. "Approximation of the Cycle Time of Cluster Tools in Semiconductor Manufacturing". In *Proceedings of the industrial engineering research conference*, edited by R. K. et al., 1–6. Georgia: IIE.

Ovacik, I. M., and R. Uzsoy. 2012. *Decomposition Methods for Complex Factory Scheduling Problems*. Springer Science & Business Media, New York.

Pinedo, M. 2016. *Scheduling: Theory, Algorithms, and Systems*. Basel: Springer International Publishing AG.

Rotondo, A., P. Young, and J. Geraghty. 2015. "Sequencing Optimisation for Makespan Improvement at Wet-etch Tools". *Computers & Operations Research* 53:261 – 274.

Roy, B., and B. Sussmann. 1964. "Les Problemes d'Ordonnancement avec Contraintes Disjonctives". *Note DS* 9.

Yugma, C., S. Dauzère-Pérès, C. Artigues, A. Derreumaux, and O. Sibille. 2012. "A Batching and Scheduling Algorithm for the Diffusion Area in Semiconductor Manufacturing". *International Journal of Production Research* 50(8):2118–2132.

AUTHOR BIOGRAPHIES

KARIM TAMSSAOUET is a Ph.D. student at the Center of Microelectronics in Provence (CMP) of Mines Saint-Etienne in France and an engineer at ST Microelectronics Rousset. He studied Industrial Engineering at Ecole Nationale Polytechnique of Algiers where he graduated in 2014. He received an M.Sc. degree in Supply Chain Mangement at Paris Dauphine University in 2015. His email address is karim.tamssaouet@emse.fr.

STEPHANE DAUZERE-PERES is Professor at the Center of Microelectronics in Provence (CMP) of Mines Saint-Etienne in France and Adjunct Professor at BI Norwegian Business School in Norway. He received the Ph.D. degree from the Paul Sabatier University in Toulouse, France, in 1992; and the H.D.R. from the Pierre and Marie Curie University, Paris, France, in 1998. He was a Postdoctoral Fellow at the Massachusetts Institute of Technology, U.S.A., in 1992 and 1993, and Research Scientist at Erasmus University Rotterdam, The Netherlands, in 1994. He has been Associate Professor and Professor from 1994 to 2004 at the Ecole des Mines de Nantes in France. His research interests broadly include modeling and optimization of operations at various decision levels (from real-time to strategic) in manufacturing and logistics, with a special emphasis on semiconductor manufacturing. He has published more than 65 papers in international journals. He has coordinated multiple academic and industrial research projects. He was runner-up in 2006 of the Franz Edelman Award Competition, and won the Best Applied Paper of the Winter Simulation Conference in 2013. His email address is dauzere-peres@emse.fr.

SEBASTIAN KNOPP is a Scientist at the AIT Austrian Institute of Technology in Vienna, Austria. He studied computer science at the University of Karlsruhe, Germany, with a focus on algorithms and compiler construction, and obtained his Diploma degree in 2006. From 2006 to 2013, he worked as a Software Developer at PTV Group in Karlsruhe, Germany, on the implementation and design of algorithms in the field of logistics and digital maps, in particular for routing in large road networks, map matching, fault tolerant geocoding, and truck driver scheduling. From 2013 to 2016, he worked at the École des Mines de Saint-Etienne in Gardanne, France, where he completed his Ph.D. thesis on scheduling in semiconductor manufacturing. In 2016, he joined the Dynamic Transportation Systems team of the AIT Center for Mobility Systems where he works on combinatorial optimization problems in transportation and logistics. His email address is sebastian.knopp@ait.ac.at.

CLAUDE YUGMA is Professor at the Center of Microelectronics in Provence (CMP) of Mines Saint-Etienne in France. He received the Ph.D. degree from the Institut National Polytechnique of Grenoble, France, in 2003; He was a Postdoctoral Researcher at the Ecole Nationale Supérieure de Génie Industriel, Grenoble, from 2003 to 2004 and from 2005 to 2006 at the Provence Microelectronics Center. He obtained the H.D.R. from the Jean-Monnet University in Saint-Etienne, France, in 2013. He has been Associate Professor and Professor from 2007 to 2016. His research interests include modeling and simulation of operations in manufacturing with a focus on semiconductor manufacturing. He has published more than 15 papers in international journals. His email address is claude.yugma@emse.fr.

JACQUES PINATON is a manager of Process Control System group at ST Microelectronics Rousset, France. He is an engineer in metallurgy from the Conservatoire National des Arts et Metiers d'Aix en Provence. He joined ST in 1984. After 5 years in the process engineering group, he joined the device department to implement SPC and Process Control methodology and tools. He participated in the startup of 3 fab generations. He is leading various Rousset R&D programs on manufacturing science including programs on automation, APC, and diagnostics. His email address is jacques.pinaton@st.com.