Heuristic algorithm for a WIP projection problem at finite capacity in semiconductor manufacturing

Emna Mhiri, Fabien Mangione, Mireille Jacomino, Philippe Vialletelle, Guillaume Lepelletier

Heuristic algorithm for a WIP projection problem at finite capacity in semiconductor manufacturing

Emna Mhiri, Fabien Mangione, Mireille Jacomino, Philippe Vialletelle, and Guillaume Lepelletier

***Abstract*—In this paper, we propose a heuristic approach for fixing work-in-progress (WIP) projection issues in the semiconductor industry especially for High Mix Low Volume (HMLV) facilities. The considered problem consists of estimating the start and end dates for each remaining process step of the production lots in the WIP and anticipating the fab loading taking into account the constraints of the maximum throughput of machines considered as capacity constraints and customer delivery commitments. The objective being to guarantee on-time delivery, we focus on minimizing the total weighted tardiness (TWT). We have formulated the problem into a mixed-integer programming (MIP) and we have empirically shown its computational intractability. Due to the computational intractability using actual production data, a heuristic algorithm is proposed. It is an iterative finite capacity planning system that considers as inputs lots due dates and equipment capabilities and capacities. The performance of the heuristic is assessed using industrial instances. It turns out that it achieves targeted objectives with satisfactory results in terms of quality of the solution and computation time.**

***Index Terms*—WIP projection; finite capacity planning; semiconductor industry; mixed integer programming; iterative algorithm.**

I. Introduction

WORK-in-progress (WIP) projection is a mid-term capacity planning activity. The objective is to compute a mid-term target schedule in order to drive factory execution, to anticipate production issues and to calculate net demand and net resource capacities. In our study, the outcome is a weekly-released schedule that depicts the start and end dates of each remaining processing step as well as the expected workload accumulated on each equipment per time bucket over the planning horizon.

In this study, the WIP projection problem is considered in one of the most dynamic industries in the world, the semiconductor industry. The semiconductor manufacturing process is extremely complex and constantly innovating. The considered wafer production plant is a High Mix Low Volume (HMLV) production line: there are several hundreds of products, different technologies and heterogeneous toolsets i.e. collections of nonidentical multi-purpose parallel machines (or tools). Moreover, typical semiconductor fabrication processes require several hundreds of different steps. As, for obvious reasons, HMLV fabs cannot multiply machines, their production flows are re-entrant: the same machine can process products at different stages of their fabrication. This also means that according to the decisions taken on the production line, products may experience various cycle times depending on the priority given either to a given product, to satisfy customer demand, or to a certain technological level for the purpose of line balancing. Hence, the capacity planning issue is difficult to solve and it is particularly more complex than in other industries [1].

Semiconductor manufacturing is composed of four major phases: wafer fabrication (fab), wafer probe, assembly, and final test. Wafer fabrication, often referred to as "front end", represents the most complicated, expensive and time-consuming phase of all four stages [2]. In this phase, hundreds of circuits are layered through successive operations on a silicon wafer. The manufacturing process in wafer fabs involves a highly complex sequence of processing operations which can be classified into various types, as for example: oxidation and thermal treatment, film deposition, planarization, photolithography, etching and ion implantation. These operations are repeated for each layer of circuitry on the wafer. Figure 1 presents a simplified view of the wafer fabrication process.

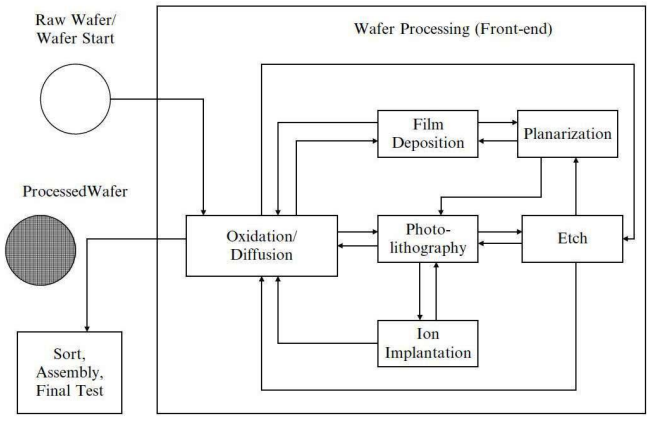


Fig. 1: Wafer fabrication Process. [1]

Each operation shown in Figure 1 can include multiple elementary steps (cleaning, process, measurement). The total number of steps per flow typically ranges between 400 and 800 for current production technologies and up to 1200 steps for latest generations. Some of the processing steps in a flow are performed on individual wafers, others on groups of wafers (lots), and still others on groups of lots (batches).

Steps performed on individual wafers or lots of wafers are referred to as serial steps, while those performed on groups of lots are called batch steps. A lot is generally composed of

E. Mhiri is with Univ. Grenoble Alpes, CNRS, G-SCOP, 38 000 Grenoble, France (e-mail: emna.mhiri@grenoble-inp.fr).

F. Mangione and M. Jacomino are with Univ. Grenoble Alpes, CNRS, G-SCOP, 38 000 Grenoble.

P. Vialletelle and G. Lepelletier are with STMicroelectronics, F-38926 Crolles Cedex, France.

25 wafers, while a typical batch contains up to six lots. In the considered case study, the lot requires 8 to 10 weeks to be processed. Steps are executed on more than one hundred workstations called "toolset" [1]. Due to flow re-entrance, lots visit the same toolset more than once during the manufacturing process.

For each step, the wafer has to be processed on various types of tools using a well-defined recipe. The recipe contains the detailed instructions to be used at the machine level in order to proceed the intended physical transformations or measurements. The identification of the candidate tools to be used is made through qualification of recipes on the tools. However, in HMLV fabs, because of multiple differences in hardware and software configurations, hence variety of recipes to be used, it is not possible to qualify all recipes on every machine. Qualification is one of the characteristics of the HMLV semiconductor manufacturing. It determines the processing authorization of a product on a machine. It acts like an eligibility constraint that allows production volume allocation of a product to a machine. It is known also as the process capability constraint [3].

Besides, each toolset has an identified throughput considered as its capacity which refers to its upper loading threshold under a given product mix condition. To establish a feasible production schedule over a planning horizon of several weeks, thus requires to consider capability and capacity constraints. Moreover, as for other industries, semiconductor manufacturing facilities must respect customers delivery commitments to survive in competitive business environments. For HMLV fabs, actual cycle time is widely spread and skewed due to large variability of numerous sources: equipment heterogeneity, product priorities, low redundancy, steps qualifications, etc. It is then crucial to consider also variable cycle times while defining a production plan. In practice, fab's historical data and various applications of the queuing theory are often used.

In this paper, a mixed integer program (MIP) and a heuristic algorithm are proposed to project current inventory and new wafer starts throughout the remaining processing sequence, taking into account all the cited constraints. The objective is to establish a feasible midterm schedule, in a fast execution time (less than 5 minutes, the required computation time of capacity planners of the industrial partner), while minimizing lots delivery delays and optimizing workload balance among all toolsets. This study is applied to the Crolles 300 mm wafer fab of STMicroelectronics. Thus, data from actual production process are collected and used to evaluate the performance of the developed approaches.

This paper is organized as follows. This section introduces the main characteristics of the considered industrial environment. In section 2, some background on existing related capacity planning problems is provided. In section 3, the problem is stated and the MIP formulation is presented. The proposed iterative heuristic algorithm is explained in section 4. Next, in section 5, experiments conducted and analyses carried out are discussed. Finally, section 6 draws conclusions and provides suggestions for future work.

II. Previous related work

As the semiconductor industry is considered as one of the most complex manufacturing processes, many researchers have paid attention to the capacity planning problems encountered in this environment.

The various problems investigated have considered different phases of the manufacturing process of integrated circuits, different constraints, different methods and techniques used for capacity planning and different performance measures. Mönch et al. [1], Uzsoy et al. [2], [4] and Gupta et al. [5] have mentioned in their reviews different capacity planning techniques used in the semiconductor environment which can be classified in infinite and finite capacity planning techniques. They can also be divided, according to the length of the planning horizon, into long-term (strategic), mid-term (tactical) and short-term (operational) planning tools.

Among the methods used for capacity planning, classical techniques were successfully used in many industries especially for tactical and operational production planning, such as Material Requirement Planning (MRP) developed by Orlicky [6], Manufacturing Resource Planning (MRPII) [7], Just In Time (JIT) [8] and Theory Of Constraints (TOC) [9]. The application of these traditional techniques for capacity planning in semiconductor industry presents some shortcomings. Indeed, it is proven that MRP method can be inefficient and may produce unrealistic production schedules when used in field applications. It ignores capacity constraints and assumes fixed cycle times ([10], [11], [12], [13]). However, in semiconductor facilities, cycle times depend on many factors, such as machine utilization rate, lot size, inventory and dispatching rules, and are thus variable. Either shortcoming above leads to infeasible production schedules, fluctuating workloads over time and significant users effort to adjust the plans. The JIT technique proves its strengths [14]; however, it presents some limitations in the high-mix low-volume production systems. It seems to be more suitable for a repetitive production environment with stable demand and low product mix [15]. The TOC seems an efficient capacity planning technique in semiconductor industry [16] but it considers only bottleneck resources and it can not deal with changes in the bottlenecks.

In addition to these classical industrial methods, authors use discrete event simulation models, queueing theory, linear programming and heuristics for capacity planning applied to semiconductor industry. Discrete event simulation is often used for capacity planning decisions in wafer fabs [17] in order to evaluate the performance of production planning strategies ([18], [19], [20], [21], [22]). Indeed, discrete-event simulation is considered as the only practical method that explicitly calculates the cycle time as a function of resource availability and production rate. The simulation model can be used also to determine bottlenecks under a given product mix and to make strategic decisions concerning equipment purchase [23]. However, simulation models used for capacity planning in the semiconductor industry present some severe limitations. Their set-up is very time-consuming due to the volume and often complexity of the data required for the models involved. Moreover, these models do not provide a

means for optimization of the plan ([24], [25]).

Concerning queueing network models, Shanthikumar et al. [26] presented a survey of the different applications of queueing theory for semiconductor manufacturing systems. They recognized that in spite of fast computing time compared with simulation models, the accuracy of classical queueing models is not satisfactory due to the complexity of the semiconductor manufacturing process.

The linear programming (LP) approach is widely applied to specific issues encountered in capacity planning for the semiconductor industry. Mixed-integer programming (MIP) models are developed for strategic planning in order to maximize the profit ([27], [28], [29]) or to minimize the machine tool operating costs, new tool acquisition costs, and inventory holding costs taking into account capacity constraints [30]. A good source of previous work related to more strategic capacity planning is provided by Geng and Jiang [31].

LP (sometimes in combination with discrete-event simulation) is also used to solve medium-term finite capacity planning problems. The work of Hung and Leachman [32] is an example of such an approach. Leachman [33] used LP for production planning and presents a corporate capacity planning model, which includes multiple facilities integrated with the production process. Habla et al. [34] suggested a MIP formulation to determine completion time targets for the lots on bottleneck steps. Bermon et al. [35] introduced a linear programming model to analyze the capacity of large and complex manufacturing production lines.

Due to the intractability of LP models, they are generally combined with heuristics such as genetic algorithms [36] or decomposition techniques as Benders [37] or Lagrangian relaxation ([30], [34]) to reduce execution time. Besides, approximate methods have also been widely used to develop either infinite or finite capacity planning systems for the semiconductor industry.

Infinite capacity planning systems are developed to estimate the future loading of equipment in order to identify bottleneck resources and to balance the loading of each production resource over the planning horizon ([38], [39], [40]).

Bearing in mind the importance of capacity constraints, many authors developed finite capacity planning systems using algorithmic approaches. Fargher et al. [41] used a beam-search algorithm in combination with backtracking steps for lot release and for the determination of schedules in an aggregated sense.

Horiguchi et al. [42] proposed an algorithm that estimates the start and finish date of each job scheduled on each critical resource: their algorithm considers the available time for all the feasible combinations of time bucket and critical resource, and it reduces the available time whenever a new production order is added to the schedule. This approach, due to the high aggregation level in modeling resources and relationships, might lead to orders overlapping on the same resource in the same time bucket (i.e. infeasible plans).

Habenicht and Mönch [43] used also a beam-search algorithm to determine planned start and completion dates for the macro operations (sets of consecutive process steps) of a lot.

Chua et al. [44] developed an intelligent multi-constraint finite capacity-based lot release system. This system has been designed, developed and implemented to solve the lot release problems in a discrete manufacturing environment with a huge product mix and multiple capacity constraints.

In this study, we are interested in operations research related (or mathematical) optimization approaches and we consider a medium-term finite capacity planning problem, applied to a semiconductor production line. So far, the literature review has pointed out that the debate about this problem is still open, and the proposed approaches by several authors still have some limits. Table II presents a taxonomy of studies considering the same problem and using operations research solving tools. Steps cycle time can be either defined as a fixed input by the proposed approach or a variable output of the procedure. Capability constraints are relevant, since they can be embedded or not in the proposed procedure. Lots due dates, relevant as well, can be considered as input parameters or not. Finally, the model can be tested via data generated by authors (random instances) or through data from real-life production systems (real case).

Whilst capacity and cycle time are tightly linked one another through the Little's law [45], cycle time is considered by most approaches as a fixed input parameter. Moreover, some methods ignore capability constraints thus leading to infeasible production plans. Finally, the applicability in field to real-life companies has not been reported for all the anterior studies.

Furthermore, Table II presents, for each study, its algorithmic and operational objectives. As one can notice, the main issues treated in the existing studies are generally limited to dispatching rules and release control policies which are outside the scope of this paper.

In the literature, there are few studies considering the WIP projection problem in the semiconductor industry ([46], [47], [48]). In these works, authors consider different objectives and do not take into account all the cited constraints.

Kim and Leachman [46] proposed a LP formulation and a decomposition heuristic method to determine net demand and net resource capacities taking into account capacity constraints. They tested their approaches using random data. Lee et al. [47] employed deterministic linear programming techniques for the WIP projection problem in the wafer fab, that explicitly considers the variable cycle time. Govind and Fronckowiak [48] consider WIP projection problem to measure production performance at IBM's 300 mm wafer fab by computing productivity and WIP targets at infinite capacity. As one can see, even if some papers tackle similar planning problems, none of the already proposed models explicitly address our specific problem.

The research work outlined here tried to overcome some of the limits above: the proposed finite capacity planning algorithm does not consider fixed steps cycle time, it takes into account lots due dates and it has been tested in a real-life industrial context. Furthermore, it meets the key requirement of semiconductor industrials, consisting on fast computing of feasible production plans (in five minutes at most on a personal computer) to facilitate "what-if" analysis.

III. Mathematical formulation of the problem

A. Problem description

The WIP projection problem may be defined as follows. A set of lots $l \in \{1, \ldots, L\}$, composed of Q_l wafers each, is considered. For each lot l, it remains an identified number of steps S_l to be processed on a time horizon discretized in T periods $t \in \{1, \ldots, T\}$ of equal length P_t. Each lot of a weight w_l indicating its priority, has a release date r_l and a due date d_l.

The performance measurement to be minimized in this problem is total weighted tardiness TWT. TWT is a measure that incurs a penalty for each lot that finishes processing after its promised delivery date. This penalty increases with the magnitude of the tardiness, and therefore schedules that minimize the weighted (by lot priority) sum of penalties provide good on-time delivery performance, whereas higher values of total weighted tardiness indicate that many important lots are not being delivered on time. Indeed, a processing schedule will provide a completion time, C_l , for each lot. The tardiness, T_l, of lot l is then defined as $T_l = \max(0, C_l - d_l)$. The weighted tardiness of the lot l (WT_l) is defined as WT_l = ($w_l \times T_l$). Total weighted tardiness computes the weighted sum of tardiness values: $TWT = \sum_l WT_l$.
Each remaining step $s_l \in \{1, \ldots, S_l\}$ of the lot l is processed on one or several qualified toolsets $i \in \{1, \ldots, I\}$. The quantity of wafers of a lot l assigned to the toolset i, processing the step s_l during the period t, is denoted $a_{s_l,l,i,t}$. It has a waiting time $wt_{s_l,l}$ and it consumes a unit processing time $p_{s_l,l,i}$ on each of its qualified processing toolsets. It also has a start date $s_{s_l,l}$ and an end date $e_{s_l,l}$ (Figure 2). Each toolset i has a finite capacity $C_{i,t}$, which gives the maximal loading $L_{i,t}$ over a period t.

Table I summarizes the notation.

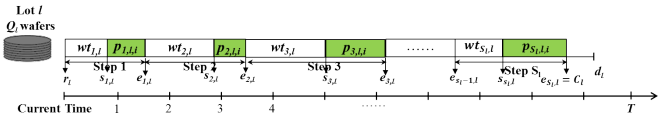


Fig. 2: Problem description.

TABLE I: Summary of problem notation

Indices	**Description**
L	Number of lots
$l = 1..L$	Lot index
S_l	Number of remaining steps of lot l
$s_l = 1..S_l$	Lot's step index
I	Number of toolsets
$i = 1..I$	Toolset index
T	Number of time buckets
$t = 1..T$	Period index
Parameters	**Description**
P_t	Length of period t
Q_l	Quantity of wafers of lot l
r_l	Release date of lot l
w_l	Weight of lot l
d_l	Due date of lot l
$p_{s_l,l,i}$	Unit processing time of step s_l of lot l on qualified toolset i, 0 on non-qualified toolset i
$C_{i,t}$	Capacity of toolset i in period t
$a_{s_l,l,i}$	Quantity of wafers of lot l in step s_l processed by the toolset i
Decision variables	**Description**
$s_{s_l,l}$	Start date of step s_l of lot l
$e_{s_l,l}$	End date of step s_l of lot l
C_l	Completion date of lot l
T_l	Tardiness of lot l
$L_{i,t}$	Loading of toolset i in period t
$y_{s_l,l,t}$	= $s_{s_l,l}$ if the step s_l of lot l is released in period t, 0 otherwise
$x_{s_l,l,t}$	=1 if step s_l of lot l is processed in period t, 0 otherwise

B. Mixed-Integer Program

In this subsection, an appropriate MIP formulation is presented for the multi-product, multi-period and multi-resource capacity planning problem. The proposed MIP is similar to LP capacity planning models that can be found in standard textbooks, with some variations and extensions. Using the notation presented above, the MIP is as follows:

$$min \quad \sum_l w_l T_l \tag{1}$$

$$s.c. \quad s_{1,l} \geq r_l \qquad l = 1, \ldots, L \tag{2}$$

$$s_{s_l,l} + \sum_i p_{s_l,l,i} \times a_{s_l,l,i,t} \times x_{s_l,l,t} = e_{s_l,l}$$
$$s_l = 1, \ldots, S_l,\ l = 1, \ldots, L \tag{3}$$

$$s_{s_l,l} \geq e_{s_l-1,l} \quad s_l = 2, \ldots, S_l,\ l = 1, \ldots, L \tag{4}$$

$$\sum_t y_{s_l,l,t} = s_{s_l,l} \quad s_l = 1, \ldots, S_l,\ l = 1, \ldots, L \tag{5}$$

$$\sum_t x_{s_l,l,t} = 1 \quad s_l = 1, \ldots, S_l,\ l = 1, \ldots, L \tag{6}$$

$$C_l = e_{S_l,l} \qquad l = 1, \ldots, L \tag{7}$$

$$T_l \geq C_l - d_l \qquad l = 1, \ldots, L \tag{8}$$

$$T_l \geq 0 \qquad l = 1, \ldots, L \tag{9}$$

$$t \times P_t \times x_{s_l,l,t} \leq \quad y_{s_l,l,t} \quad s_l = 1, \ldots, S_l,$$
$$l = 1, \ldots, L,\ t = 1, \ldots, T \tag{10}$$

$$(t+1) \times P_t \times x_{s_l,l,t} > y_{s_l,l,t} \quad s_l = 1, \ldots, S_l,$$
$$l = 1, \ldots, L,\ t = 1, \ldots, T-1 \tag{11}$$

$$L_{i,t} = \sum_l \sum_{s_l} p_{s_l,l,i} \times x_{s_l,l,t} \times a_{s_l,l,i,t}$$
$$i = 1, \ldots, I,\ t = 1, \ldots, T \tag{12}$$

$$L_{i,t} \leq C_{i,t} \qquad i = 1, \ldots, I,\ t = 1, \ldots, T \tag{13}$$

$$x_{s_l,l,t} = \{0, 1\} \qquad s_l = 1, \ldots, S_l,\ l = 1, \ldots, L,$$
$$t = 1, \ldots, T \tag{14}$$

TABLE II: Taxonomy of the literature of finite capacity planning approaches applied to semiconductor industry.

Approach	Reference	Objectives		Constraints and assumptions				Test
		Algorithmic	Operational	Capacity constraints	Capability constraints	Due dates	Variable cycle times	
Linear programming based	Hung and Leachman [32], Bermon et al. [35]	Maximize the profit	Determine wafer release quantities	✓			✓	Real case study
	Leachman [33]	Maximize the profit	Generate capacity-feasible start and out schedules	✓	✓	✓	✓	Real case study
Algorithms /heuristics based	Habla et al. [34]	Minimize total weighted tardiness	Determine completion time targets for the bottleneck steps	✓		✓		Example
	Fargher et al. [41]	Reduce cycle time and the variance of cycle time	Determine the work to release into the factory at any time	✓		✓	✓	Real case study
	Horiguchi et al. [42]	Estimate the start and finish date of each job scheduled on each critical resource	Improve delivery performance and system predictability	✓		✓		Example
	Habenicht and Mönch [43]	Determine the start date and the end date of each operation of the lot	Establish a feasible production plan	✓		✓	✓	Example
	Chua et al. [44]	Compute orders release dates for semiconductor back end assembly	Solve lot release problem	✓				Real case study
	Our study	Minimize total weighted tardiness	Establish a feasible target production plan	✓	✓	✓	✓	Real case study

The objective function (1) minimizes the total weighted tardiness (TWT). The MIP constraints can be classified in two kinds: temporal constraints ((2)...(11)) and cumulative constraints (constraints (12)-(13)). Constraints (2) define the start date of the first remaining step for each lot. The end date of each remaining step of each lot is computed using constraints (3). Constraints (4) present precedence constraints of processing steps. Constraints (5) guarantee that each remaining step of each lot is released once. Constraints (6) verify that each remaining step of each lot is processed once during the planning horizon. Constraints (7) define the lots completion date. Constraints (8) and (9) compute the tardiness for each lot. Constraints (10) and (11) indicate that each remaining step of each lot is processed in one period. Constraints (12) calculate the workload accumulated by each toolset over each period taking into account the qualification of the toolset to the processed step and the quantity of wafers assigned to the considered toolset. Constraints (13) are the capacity constraints. Constraints (14) are the binary constraints for the decision variable.

The mathematical model presented above has been solved by `ILOG CPLEX` solver. Experiments were run on an `Intel® Core™ i5` PC running a 2.7 GHz processor and 4 GB of RAM. Tests have been performed on 30 randomly generated instances of the problem in order to highlight the main characteristics of the industrial data and to maintain a certain degree of generality in order to preserve all the difficulty of the problem. Indeed, based on the observation made in the literature, we identified seven important problem parameters which could affect the performance of the proposed approach: number of lots (L), maximum number of remaining steps for each lot ($\max S_l$), number of toolsets (I), length of the planning horizon (T), lots steps unit processing times ($p_{s_l,l,i}$), lots due dates (d_l) and machines capacities ($C_{i,t}$).

We consider the cases of 3, 5, 10, 20, 100 and 300 parallel toolsets with a fixed capacity corresponding to the maximum equipment utilization rate which is equal to 100%. The lots weights w_l are chosen from a uniform distribution over (0,1). The lots release dates r_l and lots quantity of wafers Q_l are supposed equal to 0 and 25 for all lots, respectively. The range of lots due dates d_l and steps unit processing times $p_{s_l,l,i}$ is extracted from real data. d_l are ranging from 1 to 210 days relative to the release date and $p_{s_l,l,i}$ range between 0.0005 and 0.5 hours. The planning horizon is set to 24 periods (weeks). Table III presents the different tests parameters generating 30 instances.

Optimal results were obtained in reasonable execution time while testing the MIP on instances of reduced size. Further increasing the size of the tested instances (up to about 4000 steps plan), the resolution of MIP was halted as it required a very large amount of time and computer memory (Figure 3). Indeed, the whole real problem presents 70 742 400 constraints and 69 371 200 variables. It corresponds to a WIP composed of 2000 lots, each lot having a maximum of 680 remaining steps to process on 300 toolsets over a planning horizon composed of 24 periods (weeks). Thus, the size of real instances is obviously too large to be solved using the proposed MIP (Figure 3).

TABLE III: Summary of tests parameters

Problem parameter	Values used
Number of lots (L)	2, 3, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 240, 1000, 1700, 2000
Maximum number of remaining steps of lot l ($\max S_l$)	1, 2, 5, 6, 8, 10, 20, 30, 40, 50, 60, 100, 150, 200, 250, 680
Number of toolsets (I)	3, 5, 10, 20, 100, 300
Number of time buckets (T)	24
Weight per lot (w_l)	Uniform (0,1)
Lots release dates (r_l)	0
Lots due dates (d_l)	r_l+[1..210]
Lots quantity of wafers (Q_l)	25
Steps unit processing times ($p_{s_l,l,i}$)	[0.0005..0.5]

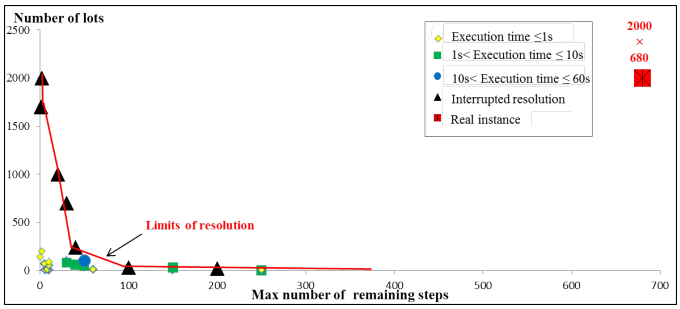


Fig. 3: Limits of MIP resolution.

From the empirical evidence on the computational difficulties in getting optimal schedule considering lots due dates and capacity constraints, it is obvious that the problem of WIP projection applied to the real case study will be computationally intractable. Furthermore, Garey and Johnson [49] highlighted in their study that production planning, capacity planning and scheduling problems in complex job shops like semiconductor manufacturing are known as strongly NP-hard problems. This has motivated us to develop a heuristic algorithm for the research problem considered in this study to provide near optimal solutions and/or efficient solution in a reasonable time. The proposed heuristic algorithm is presented in the next section.

IV. Heuristic Algorithm

An alternative methodology for the above problem should be accurate and, at the same time, fast and small enough to be stored and implemented in a mainframe or work station computer system. Bearing this in mind, a heuristic approach for WIP projection problem in HMLV semiconductor manufacturing line has been developed. It is an iterative algorithm composed of three main modules: (*i*) WIP projection at infinite capacity, (*ii*) workload accumulation and capacity analysis and (*iii*) workload and capacity balancing. The algorithm is executed by iterations on periods of the planning horizon. The principle of iterative running of the algorithm is inspired from the literature [50] and the detailed scheduling in the commercial ERP/APS. For each defined period, WIP projection module estimates the evolution of the WIP, lot by lot, based on lots due dates. Then, workload accumulation module calculates the expected equipment loading. In case of toolsets oversaturation i.e. the loading of toolsets exceeds their maximal

capacity, workload and capacity balancing module is employed to reduce toolsets loading by shifting their affected steps to subsequent periods. The following sections will detail the three major modules. Figure 4 depicts the flow of the developed system.

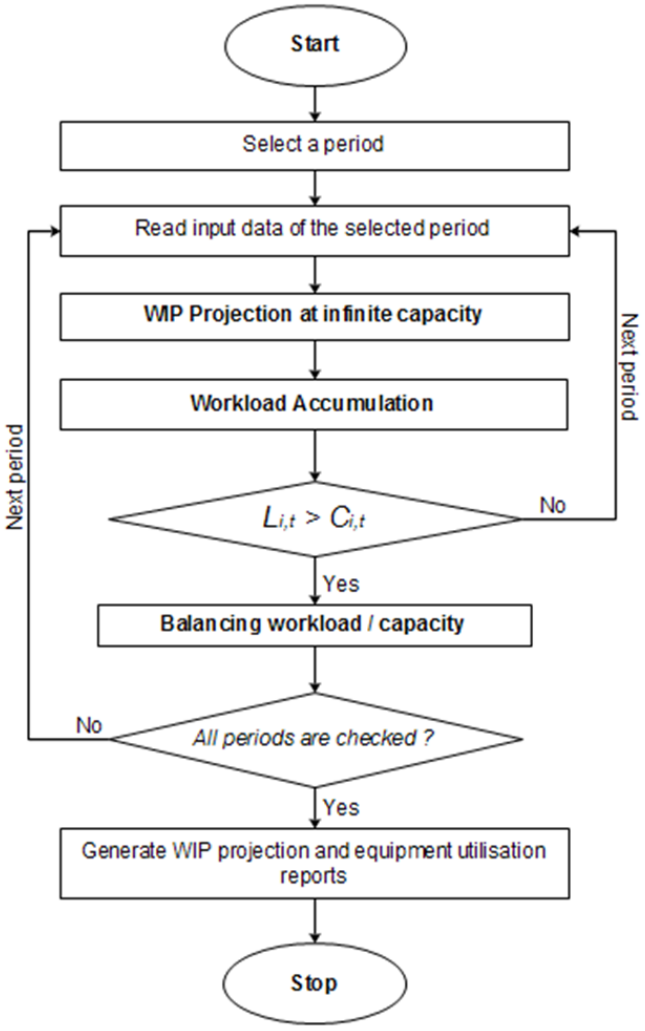


Fig. 4: Finite Capacity Planning Algorithm Flow.

A. WIP projection module

The objective of this module is to push lots, one by one, forward along their routes, from their current positions up to their due dates. It also aims to compute over a period the activity required by step to ensure the delivery plan.

For each selected time bucket t of the planning horizon, this module requires the following data input:

1) WIP status and wafer starts at the beginning of the considered projection period (position r_l, quantity Q_l),
2) Routing information, including a partition of each route into consecutive steps,
3) Steps unit processing times $p_{s_l,l,i}$,
4) Lots due dates d_l and weights w_l,
5) Flow factor $Xfactor_{s_l,l}$ that reflects possible waiting times between consecutive process steps to achieve the target cycle time, extracted from historical data. It is defined as the step mean cycle time divided by the step raw processing time RPT [51].

WIP projection module includes three steps. Step 1 computes, for each lot, from its position in the route, four parameters which are remaining process time $RemPT_l$, remaining reference cycle time $RemRefCT_l$, remaining expected cycle time $RemExpCT_l$ and cycle time coefficient $CTCoeff_l$.
$RemPT_l$ is equal to the sum of lot remaining steps unit process time multiplied by lot quantity of wafers Q_l.

$$RemPT_l = \sum_{s_l=1}^{S_l} \sum_{i=1}^{I} p_{s_l,l,i} \times Q_l \tag{15}$$

$RemRefCT_l$ corresponds to the sum of the reference cycle times of lot remaining steps $RefCT_{s_l,l}$.

$$RemRefCT_l = \sum_{s_l=1}^{S_l} RefCT_{s_l,l} \tag{16}$$

In the industrial context considered, each step has a reference cycle time, extracted from historical data, named $RefCT_{s_l,l}$. $RefCT_{s_l,l}$ corresponds to the product of the unit step process time with the quantity of lot wafers Q_l and the flow factor $Xfactor_{s_l,l}$. It is the maximum amount of time that a lot would spend at that step, including waiting and processing times.

$$RefCT_{s_l,l} = \sum_{i=1}^{I} p_{s_l,l,i} \times Q_l \times Xfactor_{s_l,l} \tag{17}$$

$RemExpCT_l$ is equal to the maximum between the difference between the due date and the current time t and $RemPT_l$.

$$RemExpCT_l = \max(d_l - t, RemPT_l) \tag{18}$$

The lot cycle time coefficient $CTCoeff_l$ identifies the necessary and sufficient speed for lots to achieve their due date according to the reference cycle time. It is equal to the ratio between lot remaining expected cycle time $RemExpCT_l$ and lot remaining reference cycle time $RemRefCT_l$:

$$CTCoeff_l = \frac{RemExpCT_l}{RemRefCT_l} \tag{19}$$

In step 2, the $RemExpCT_l$ is split on the elementary steps of each lot l to compute an expected cycle time per step $ExpCT_{s_l,l}$ which is equal to the product of $ObjCT_{s_l,l}$ and $CTCoeff_l$.

$$ExpCT_{s_l,l} = RefCT_{s_l,l} \times CTCoeff_l \tag{20}$$

Equation (20) gives a rough estimation of queuing time at each step. Hence, waiting time by step $wt_{s_l,l}$ can be computed:

$$wt_{s_l,l} = ExpCT_{s_l,l} - \sum_{i=1}^{I} p_{s_l,l,i} \times Q_l \tag{21}$$

In step 3, having the waiting time by step, start dates and end dates for all lots remaining steps, decision variables $x_{l,s_l,t}$, lots completion date and tardiness are computed.
Figure 5 illustrates an example of 2 lots with different due dates, having 3 remaining steps each. The first one has an

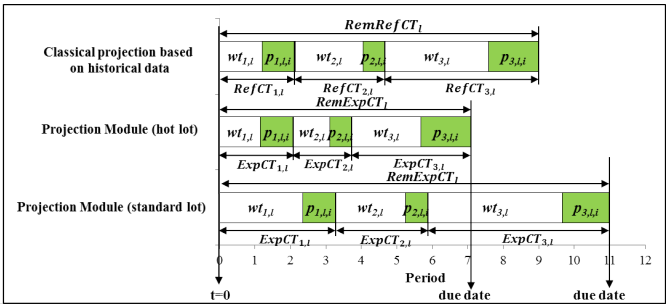


Fig. 5: Principle of cycle time computation for each processing step of a lot.

earlier due date i.e. a higher priority and less $RemExpCT_l$ than the second.

Using the classical projection based on historical data, the two lots have the same distribution of remaining steps waiting times over the planning horizon, independently of their due dates, because they have the same remaining process time $RemPT_l$. However, the proposed projection module allocates steps expected cycle times taking into account lots priorities i.e. due dates. Indeed, there are multiple priority levels of production lots. Production priorities can be divided into two levels according to the urgency of delivery: hot and standard. So, to respect these priorities, the projection module shrinks steps waiting times in order to satisfy the hot lot's due date. However, for a standard lot, it extends steps waiting times respecting the lot due date.

To further explain the concept of WIP projection, a simple random instance is tested. The considered WIP consists of 10 lots of 25 wafers each, following different routes, and having different due dates. Table IV presents, for each lot, the number of remaining steps, $RemPT_l$, $RemRefCT_l$, $RemExpCT_l$ and $CTCoeff_l$.

Figure 6 illustrates projection results of the 10 lots during the first period of the planning horizon. For some lots, a sequence of steps is repeated twice (lots 1,4,5,7,8,9) i.e. lots visit the same toolset twice which illustrates the re-entrant flows. Figure 6 shows start and end dates, waiting time and processing time for each remaining process step during the considered period. Some steps (step 4.5 and step 10.4) start in the first period and finish in the subsequent periods of the planning horizon. This figure demonstrates that the projection engine allows the extension of steps waiting times for lots having a far due date which is the case of lots 1 and 6 and it shrinks steps cycle times in case of close due date for lots 2, 4 and 8. Lots 2, 4 and 8 are not delivered on time. Their due dates are not reachable so their fab-out dates are equal to the sum of the current date ($t = 0$) and the remaining process time $RemPT_l$.

B. Workload accumulation and capacity analysis module

After WIP projection, the loading of toolsets, over each considered period $L_{i,t}$, is computed based on the assumption of infinite capacities.

To optimize the computation time, toolsets are distributed in balancing groups. A balancing group is a set of toolsets

TABLE IV: Data for simple instance

Lot l	Weight w_l	Number of remaining steps S_l	$RemPT_l$ in days	$RemRefCT_l$ in days	$RemExpCT_l$ in days	$CTCoeff_l$
Lot 1	0.33	6	1.1	1.6	5	3.125
Lot 2	1	4	0.8	1.1	0.5	0.45
Lot 3	0.5	2	0.25	0.41	1.5	3.65
Lot 4	0.5	8	1.7	2.3	1.5	0.65
Lot 5	0.5	6	1	1.4	1.5	1.07
Lot 6	0.33	4	0.75	1.02	5	4.9
Lot 7	0.5	8	0.86	1.05	1.5	1.43
Lot 8	1	4	0.8	1.05	0.5	0.48
Lot 9	0.5	4	0.8	1.05	1.5	1.43
Lot 10	0.5	6	1.4	1.9	1.5	0.79

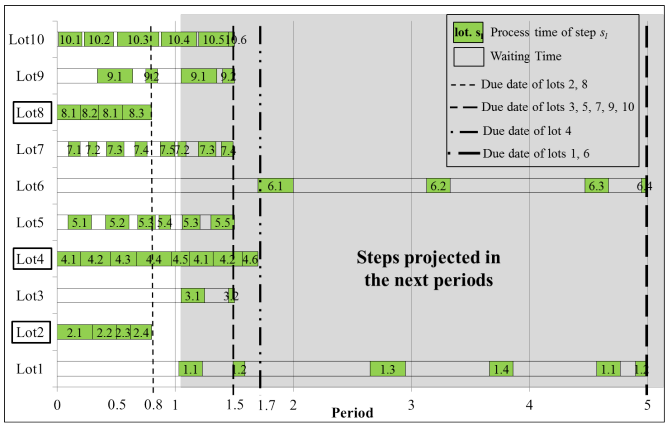


Fig. 6: Simple instance: Production schedule at infinite capacity.

that have same qualifications and share same recipes. This approach enables to decompose the problem into small sub-problems. It is a linear program used to optimize workload balancing of toolsets, belonging to the same balancing group, over a selected time bucket. The formulation of the linear program, for each balancing group and over each period, is as follows:

Notations

Indices:

B	Number of balancing groups
$b = 1..B$	Balancing group index
R_b	Number of recipes related to the balancing group b
$r = 1..R_b$	Recipe index
I_b	Number of toolsets of the balancing group
I_r	Number of toolsets qualified for recipe r, $I_r \subseteq I_b$
$i = 1..I_b$	Toolset index

Parameters:

$x_{s_l,l,t}$	Decision variables values, results of WIP projection module in period t
$v_{s_l,l,r}$	=1 if recipe r corresponds to step s_l of lot l, 0 otherwise
$a_{s_l,l,i}$	Quantity of wafers of lot l in step s_l processed by the toolset i
$p_{r,i}$	Processing time of recipe r on toolset i
Q_l	Quantity of wafers of lot l

Decision variables:

$L_{i,t}$	Loading of toolset i over period t
$W_{r,i}$	Quantity of wafers produced by toolset i qualified for recipe r
$Lmax$	Workload of the most loaded toolset in the balancing group
$Lmin$	Workload of the least loaded toolset in the balancing group
$Lmax_r$	Loading, for a given recipe r, of the most loaded toolset among those on which r is qualified
$Lmin_r$	Loading, for a given recipe r, of the least loaded toolset among those on which r is qualified

Using the above parameters, and decision variables, the linear program formulation can be represented as follows:

$$\begin{cases} \text{Minimize} & \alpha \cdot Lmax - \beta \cdot Lmin + \gamma \cdot \sum_r Lmax_r \\ & -\delta \cdot \sum_r Lmin_r + \delta \cdot (\sum_i^{I_b} L_{i,t} - Lmin) \\ \text{with} & \alpha = I_b^2, \beta = I_b, \gamma = 1, \delta = 1/I_b \end{cases}$$

s.t.

$$L_{i,t} = \sum_r p_{r,i} \times W_{r,i} \quad i = 1, \ldots, I_b \tag{22}$$

$$\sum_{i=1}^{I_r} W_{r,i} = \sum_l \sum_{s_l} x_{s_l,l,t} \times v_{s_l,l,r} \times a_{s_l,l,i} \quad r = 1, \ldots, R_b \tag{23}$$

$$L_{i,t} \geq Lmin_r \quad r = 1, \ldots, R_b, i = 1, \ldots, I_r \tag{24}$$

$$L_{i,t} \leq Lmax_r \quad r = 1, \ldots, R_b, i = 1, \ldots, I_r \tag{25}$$

$$L_{i,t} \geq Lmin \quad i = 1, \ldots, I_b \tag{26}$$

$$L_{i,t} \leq Lmax \quad i = 1, \ldots, I_b \tag{27}$$

The linear program seeks to :

- Minimize the workload of the most loaded toolset in the balancing group $Lmax$.
- Maximize the workload of the least loaded toolset in the balancing group $Lmin$.
- Minimize the total workload of toolsets $\sum_i^{I_b} L_i$ and maximize the total workload of the least loaded toolset per recipe $\sum_r Lmin_r$, with the same degree of priority.
- Minimize the total workload of the most loaded toolsets per recipe $\sum_r Lmax_r$.

For the example cited above, the remaining steps of 10 lots are considered to be processed by 6 toolsets $\{M1, M2, M3, M4, M5, M6\}$. These toolsets are classified into 4 balancing-groups $\{M1, M6\}, \{M2, M4\}, \{M3\}$ and $\{M5\}$. Figure 7 illustrates the saturation percentage of toolsets i.e. the ratio of the loading to the available capacity during the first period of the planning horizon ($\frac{L_{i,1}}{C_{i,1}}, i = 1..6$) while processing the remaining steps ordered in increasing order of the start date.
In this example, the capacity of all the considered toolsets ($C_{i,1}, i = 1..6$) is equal to 24 hours/day. Figure 7 shows that there are two over-saturated toolsets ($M2$ and $M6$) which workloads exceed saturation threshold.

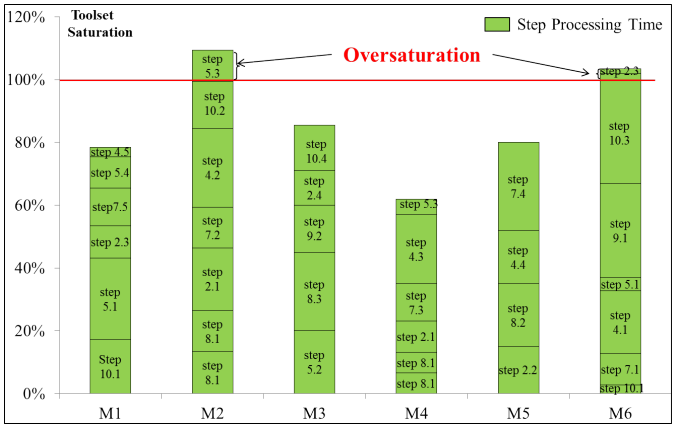


Fig. 7: Workload accumulation at infinite capacity for the first period of the planning horizon.

C. Workload/Capacity Balancing Module

As a result of the workload accumulation module, loading of some toolsets may exceed their maximal capacities i.e. constraints (13) are not satisfied. In this case, the toolset is unable to process all its affected steps during the considered period so its loading should be balanced over subsequent periods. The principle of this module is to postpone additional lots in order to bring back workload of over-saturated toolsets below their maximal saturation and to smooth the activity over the planning horizon. The algorithm for workload/capacity balancing module is as follows:

1) Sort toolsets in decreasing order of saturation.
2) Select lots executed on over-saturated toolsets.
3) Sort selected lots in increasing order of a computed ranking coefficient ($rankingCoeff_l$). The $rankingCoeff_l$ illustrates the priority of the lot in terms of its position in the process sequence of the considered toolset and the urgency of delivery. The position of a lot in the process sequence of a toolset is determined by the processing date of its last remaining step treated by the considered toolset denoted $s_{S_l,l,t}$.
The urgency of delivery is defined by the lot cycle time coefficient ($CTCoeff_l$). To compute the $rankingCoeff_l$, the lot position in the process se-

quence is normalized by the period length P_t. Hence, the $rankingCoeff_l$ is equal to :

$$rankingCoeff_l = \frac{1}{CTCoeff_l} + \frac{s_{S_l,l,t}}{P_t} \quad (28)$$

4) For the first selected lot in the sorted list, shift the last step executed in the considered over-saturated toolset and its successors to the next period.
5) Remove the processing time of shifted steps s'_l from the loading of its qualified processing toolsets while considering the quantity of wafers $a_{s'_l,l',i}$ processed by each toolset. The new value of the loading of the considered toolsets $L'_{i,t}$ is, then, equal to :

$$L'_{i,t} = L_{i,t} - \sum_{l'} \sum_{s'_l} (a_{s'_l,l',i} \times p_{s'_l,l',i} \times x_{l',s'_l,t}) \quad (29)$$

6) Repeat steps 2, 3, 4 and 5 for all toolsets until the saturation criterion is satisfied for all toolsets over the period t.

Hence, this module modifies steps projection at period t as well as the WIP for the beginning of period $t+1$.
For instance, to balance the capacity and the workload of the over-saturated toolsets $M2$ and $M6$ in the considered example, the balancing module selects $M2$ as the most over-saturated toolset ($\frac{L_{2,1}}{C_{2,1}} = 109.3\%$). Then, it selects lots 2, 4, 5, 7, 8 and 10 processed by this resource (Figure 7). These lots are sorted in increasing order of $rankingCoeff_l$ as it is mentioned in Table V.

TABLE V: Order of lots processed on $M2$ according to $rankingCoeff_l$

Lot l	**CTCoeff$_l$**	**Step** s_l	**s$_{S_l,l,t}$**	**RankingCoeff$_l$**
Lot 5	1.07	Step 5.3	0.68	1.25
Lot 7	1.43	Step 7.2	0.265	1.43
Lot 10	0.79	Step 10.2	0.23	2.04
Lot 4	0.65	Step 4.2	0.2	2.34
Lot 8	0.48	Step 8.1	0	3.08
Lot 2	0.45	Step 2.1	0	3.22

In order to decrease the loading of the toolset $M2$, steps 5.3 and 7.2 and its successors are shifted to the next period of the planning horizon. Hence, the loading of $M2$ becomes less than its maximum capacity: $\frac{L_{6,1}}{C_{6,1}} = 96.4\%$. $M4$ is also qualified for step 5.3, so its loading decreases by 5.06%. Step 5.4 projected in the first period is also postponed as it is the successor of the shifted step 5.3. Thus, the loading of $M1$ processing step 5.4 becomes equal to 58.83%. Shifting the successors of step 7.2 (steps 7.3, 7.4 and 7.5) leads to decreasing the loading of toolsets $M1$, $M4$ and $M5$. The same algorithm is applied to the toolset $M6$ by shifting step 9.1 and its successor step 9.2. So, its loading decreases to 72%.

The toolsets workload obtained after steps shifting is illustrated in Figure 8. Table VI presents the WIP and the computed parameters ($RemPT_l$, $RemObjCT_l$, $RemExpCT_l$ and $CTCoeff_l$) in the beginning of the next period.

The proposed approach is tested over a five-day planning horizon. Indeed, as mix variations were present in industrial dataset used for this study, it was decided to focus on a very short planning horizon to evaluate the proposed approach.

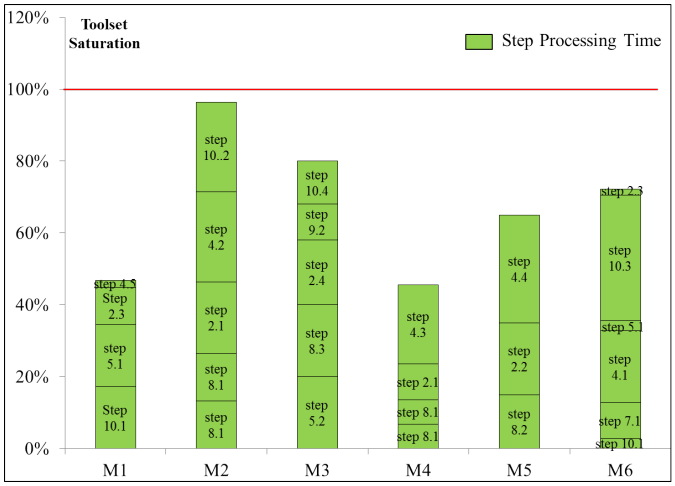


Fig. 8: Workload accumulation at finite capacity.

TABLE VI: WIP parameters in the beginning of the second period

Lot l	Weight w_l	Number of remaining steps S_l	$RemPT_l$ in days	$RemRefCT_l$ in days	$RemExpCT_l$ in days	$CTCoeff_l$
Lot 1	0.33	6	1.1	1.6	4	2.5
Lot 3	0.5	2	0.25	0.41	0.5	1.22
Lot 4	0.5	3	0.58	0.76	0.7	0.92
Lot 5	0.5	4	0.6	0.83	0.5	0.6
Lot 6	0.33	4	0.75	1.02	4	3.92
Lot 7	0.5	7	0.76	0.91	0.5	0.55
Lot 9	0.5	4	0.8	1.05	0.5	0.47
Lot 10	0.5	2	0.3	0.41	0.5	1.22

Clearly, the shorter the length of the period, the more accurate the results of the approach. The final obtained schedule for this instance is illustrated in Figure 9. For this instance, the TWT is equal to 1.46 days and we have five delayed lots.

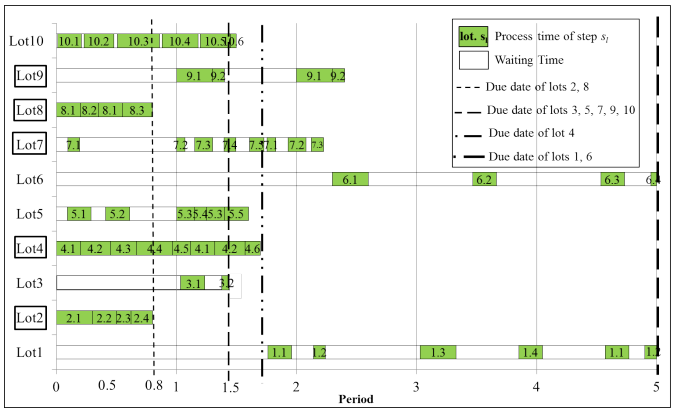


Fig. 9: The obtained schedule using heuristic approach.

V. Results and Discussion

The proposed algorithm is coded in Java and it is tested on a 4 GigaOctet RAM and 2.7 GigaHertz processor computer. We conducted two types of experiments to evaluate the performance of the proposed approach. The first type corresponds to a comparison between the exact method and the heuristic

using a set of randomly generated instances. In the second type of experiment, we compare the projected schedule obtained by the proposed approach using real data with what is really going on in the wafer fab following this schedule.

A. Evaluation of the proposed heuristic algorithm in comparison with optimal solution

For this evaluation, random instances were generated and solved using the MIP and the proposed heuristic algorithm. The parameter, number of lots (L), assumes only five levels (L=10, 15, 20, 50 and L=100). The parameters generated for the proposed instances are presented in Table VII. So, three random problem instances for each fixed parameter combination are obtained, giving a total of 270 test problems. Each of the 270 problem instances generated has been solved using the `ILOG CPLEX` solver and the proposed heuristic algorithm.

TABLE VII: Test data parameters

Problem Parameter	Values used	Total values
Number of lots (L)	10, 15, 20, 50, 100	5
Maximum number of remaining steps of lot l ($\max S_l$)	10, 20, 30, 40, 50, 100	6
Number of toolsets (I)	5, 10, 20	3
Number of time buckets (T)	24	1
Weight per lot w_l	uniform (0,1)	1
Lots release dates r_l	0	1
Lots due dates d_l	uniform(1,30)	1
Lots quantity of wafers Q_l	25	1
Steps unit process times $p_{s_l,l,i}$	0.0001× uniform(5,50)	1
	Total parameter combinations	90
	Number of problem instances	3
	Total problems	270

The results on TWT obtained for each instance using MIP model and using the proposed iterative algorithm are recorded. Based on these results, the heuristic solution matched exactly with the optimal solution 53 times.
Furthermore, for each instance with a size $L \times \max S_l \times I$, we compute:

- The absolute deviation= |TWT value from a heuristic algorithm - optimal TWT value|
- The relative deviation= $\frac{\text{absolute deviation value}}{\text{optimal TWT value}}$

Figure 10 shows the relative deviation over 270 instances plotted against the absolute deviation.

In this figure, we can define four zones or classes according to the size of the instance:

- The first zone (corresponding to absolute deviation values $\in [0..30]$ days and relative deviation values ≤ 1): Around 92% of the tested instances are situated in this zone. Hence, for most of the instances, the heuristic solution is close to the optimal one.
- The second zone (corresponding to absolute deviation values $\in]30..140]$ days and relative deviation values ≤ 1): The 8 instances ($\simeq 3\%$ of the total of tested instances) belonging to this category are instances of large size ($\geq$ 10000). For example, we find the instance with a size equal to 10000 (L=100, $\max S_l$=10, I=10) which has an absolute deviation equal to 79 days and a relative deviation equal to 0.36. This instance has an optimal solution TWT equal to 218 days. The important value of the absolute deviation is thus not significant because of high values of TWT.
- The third zone (corresponding to absolute deviation values $\in [0..30]$ days and relative deviation values >1): 14 instances ($\simeq$ 5% of the total of tested instances) are located in this zone. We can cite the example of the instance with a size equal to 15000 (L=50, $\max S_l$=30, I=10), a low value of absolute deviation equal to 2.23 days and a high value of relative deviation equal to 4.74. For this instance, both of the optimal and the approximate solutions present a low value of TWT. Hence, in this zone, the importance of the relative deviation has no significance.
- The fourth zone (corresponding to absolute deviation values $>$ 30 days and relative deviation values >1): No instance is located in this zone characterized by high values of absolute and relative deviations.

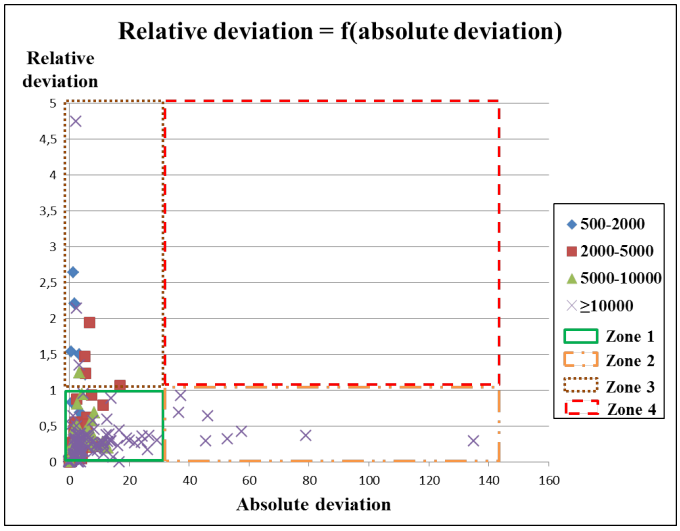


Fig. 10: Comparaison between the optimal and the approximate solution.

B. Experimental tests on real fab data

The aim of this section is to evaluate the ability of the proposed approach to tackle real world problems. The test of the real instance (L=2000, $\max S_l$=680, I=300, T=24), unsolved in reasonable execution time using the MIP approach in Section 3, is treated. The execution time of this instance with the proposed algorithm is around 30 seconds. In the calculated production schedule, 80 % of projected lots are delivered on time. Furthermore, the saturation of toolsets is kept below the pre-defined saturation threshold while minimizing lots lateness. Figure 11 illustrates the obtained weekly saturation at infinite and finite capacity of a photo-lithography toolset considered as a bottleneck. In semiconductor fabs, several indicators are used to measure performance [52]. Jointly with

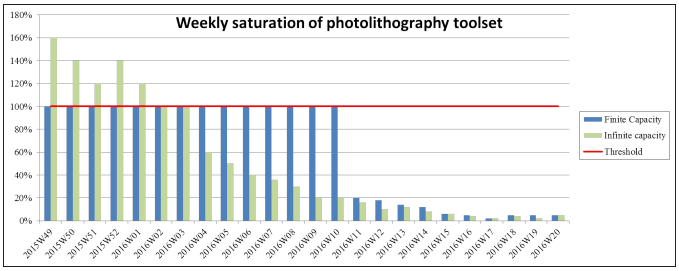


Fig. 11: Weekly saturation of a photo-lithography toolset at infinite and finite capacity.

managers of the fab, we identified three relevant indicators for our study, as described below:

- *Number of moves*: This corresponds to the number of completed steps on each period of the planning horizon, which can be compared to the real number in the production line.
- *Number of moves by usage*: It is the number of processed steps by set of toolsets belonging to the same area named "usage" in each period of the planning horizon.
- *Total Weighted Tardiness TWT*: This indicator is used to evaluate the waiting times of lots for processing.

In this section, we compare the cited indicators of performance of the heuristic solution with the indicators determined in the real production line. To ensure this experiment, six tests have been performed on actual instances issued over four months of production: September, October, November and December 2015. We have made projections in six different periods (week1, week2, week3, week4, week5 and week6) and we have determined the three indicators for each projection. For confidential reasons, we are not allowed to provide the real values of the fab. This is why, we compute the relative deviation between the predicted value and the real one for each period of the planning horizon:

$$\text{Relative deviation} = \frac{|\text{Estimated value} - \text{real value}|}{\text{real value}}$$

1) Analysis based on the performance measure: number of moves: Figure 12 shows relative deviations of *number of moves* over 15 time buckets (weeks) of the planning horizon. It illustrates that in the first 6 periods for the different instances, the relative deviation between the real number of processing steps and the calculated value is low. The average of the average relative deviations over six periods for the different tests is equal to 12.7%, reflecting a small difference between the estimated number of moves and the achieved one. Further being away from the beginning of the projection, the relative deviation between the obtained solution and the real number of moves increases which is explained by the variability of the process. Hence, there is a convergence between what is estimated and what is achieved in terms of periodic activity for a short-term planning horizon.

2) Analysis based on the performance measure: number of moves by usage: To evaluate how the heuristic solution anticipates the fab loading, we compute the absolute deviation of the number of moves by set of toolsets sharing the same qualifications named "*usage*" over the six instances for each period of the planning horizon. Figure 13 shows the difference

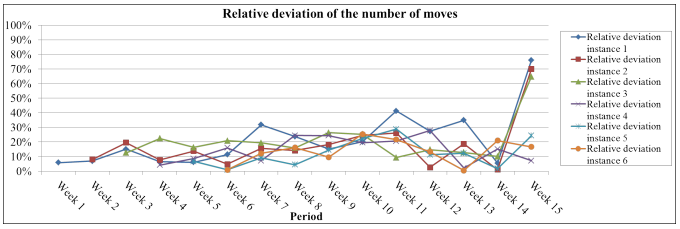
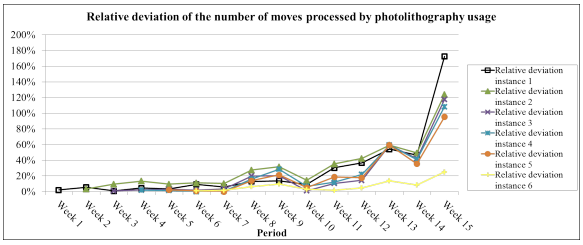
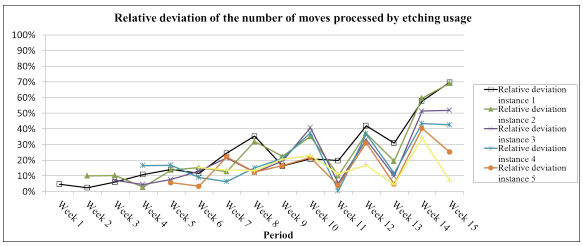


Fig. 12: Number of moves comparison actual versus forecast.

between the total number of completed steps processed by two types of bottleneck usages (photolithography and etching). For this indicator also, we observe a convergence between the planning and the real process for the first 6 periods with an average of the average relative deviations over these periods equal to 6.5% for the usage of photolithography and 12.3% for the usage of etching. Therefore, the heuristic provides good estimations of the tools loading close to the real workload while respecting capacity constraints.



(a) Photolithography usage



(b) Etching usage

Fig. 13: Total number of moves processed by photolithography and etching usages comparison actual versus forecast.

3) Analysis based on the performance measure: TWT: To compare between the real total weighted tardiness and the obtained value of this indicator using the iterative algorithm for the six tests, absolute and relative deviations are computed and reported in Table VIII. From Table VIII, we note that the estimated value of TWT is close to the real tardiness while respecting lots due dates. Indeed, the average of relative deviations over six instances is equal to 4%.

VI. Conclusions and Perspectives

This paper has examined the problem of WIP projection at finite capacity to minimize the TWT, and has proven empirically the computational complexity in obtaining optimal solution and suggested a simple, fast and efficient heuristic.

TABLE VIII: TWT comparison actual versus forecast

Instance	Absolute deviation (days)	Relative deviation (%)
Instance 1	228.255	6
Instance 2	98.305	2.62
Instance 3	108.86	3.39
Instance 4	47.77	1.83
Instance 5	50.13	2.08
Instance 6	146.23	7.92

The motivation for this research is to compute a feasible production plan to drive the execution of wafer fabs. This problem is of considerable practical value because the heuristic, proposed in this paper, can be used in planning of a large number of production lots while respecting lots due dates and capacity and capability constraints.

The computational tests, made on real production instances, showed that acceptable solutions are obtained in reasonable execution time. Indeed, the TWT could be minimized and the average tool utilization rate could be balanced significantly by using the developed system. Besides, the computation for real instances is achieved in around 30 seconds which is efficient for planning problems with a horizon of weeks up to months in real situations. Hence, this decision support tool outperforms simulation and analytic models for establishing a feasible production schedule rapidly. Finally, it is observed (as well as statistically verified) from the results of the comparison of the different criteria (total number of moves, number of moves by usage and TWT) an obvious convergence between what is predicted using the developed approach and what is achieved in the real process over a short-term planning horizon. These results show that the implementation of the finite capacity planning system in real fabs seems very interesting to minimize lots lateness and to establish a feasible production schedule. There are a number of interesting extensions of the problems that can be pursued. The first important issue would be to perform a more thorough multi-criteria analysis while shifting lots to balance toolsets loadings. Besides, it is necessary to implement the developed finite capacity planning system in the production plant to guarantee the performance of the solution. Considering other specificities of semiconductor industry such as batching or sequence dependent setup times may be interesting to enhance the accuracy of the developed system.

ACKNOWLEDGMENT

This work is supported by the ENIAC European Project INTEGRATE. The authors also gratefully acknowledge STMicroelectronics for their support on the knowledge of the semiconductor industry.

REFERENCES

[1] L. Mönch, J. Fowler, and S. Mason, *Production planning and control for semiconductor wafer fabrication facilities*. Springer New York, 2013.

[2] R. Uzsoy, C.-Y. Lee, and L. Martin-Vega, "A review of production planning and scheduling models in the semiconductor industry part I: system characteristics, performance evaluation and production planning," *IIE Transactions*, vol. 24, no. 4, pp. 47–60, 1992.

[3] M. Rowshannahad and S. Dauzère-Pérès, "Qualification management with batch size constraint," in *Proceedings of the 2013 Winter Simulation Conference*, (Washington, United States), pp. 3707–3718, 2013.

[4] R. Uzsoy, C.-Y. Lee, and L. Martin-Vega, "A review of production planning and scheduling models in the semiconductor industry part II: shop-floor control," *IIE Transactions*, vol. 26, no. 5, pp. 44–55, 1994.

[5] J. N. D. Gupta, R. Ruiz, J. W. Fowler, and S. J. Mason, "Operational planning and control of semiconductor wafer fabrication," *Production Planning and Control*, vol. 17, no. 7, pp. 639–647, 2006.

[6] J. Orlicky, *Material requirements planning*. McGraw-Hill Professional, 1975.

[7] P. J. Rondeau and L. A. Litteral, "Evolution of manufacturing planning and control systems: from reorder point to enterprise resource planning," *Production and Inventory Management Journal*, vol. 42, no. 2, p. 17, 2001.

[8] D. Y. Golhar and C. L. Stamm, "The just-in-time philosophy: a literature review," *International Journal of Production Research*, vol. 29, no. 4, pp. 657–676, 1991.

[9] E. M. Goldratt, *Theory of constraints: What is this thing called Theory of Constraints and how should it be implemented*. North River Press, 1990.

[10] T. Rossi and M. Pero, "A simulation-based finite capacity mrp procedure not depending on lead time estimation," *International Journal of Operational Research*, vol. 11, no. 3, pp. 237–261, 2011.

[11] H. Jodlbauer and S. Reitner, "Material and capacity requirements planning with dynamic lead times," *International Journal of Production Research*, vol. 50, no. 16, pp. 4477–4492, 2012.

[12] L. Sun, S. S. Heragu, L. Chen, and M. L. Spearman, "Comparing dynamic risk-based scheduling methods with mrp via simulation," *International Journal of Production Research*, vol. 50, no. 4, pp. 921–937, 2012.

[13] T. Aouam and R. Uzsoy, "Zero-order production planning models with stochastic demand and workload-dependent lead times," *International Journal of Production Research*, vol. 53, no. 6, pp. 1661–1679, 2015.

[14] M. E. Levitt and J. A. Abraham, "Just-In-Time methods for semiconductor manufacturing," in *Proceedings of the 1990 Advanced Semiconductor Manufacturing Conference*, (Danvers, MA), pp. 3–9, 1990.

[15] J. G. Carlson and A. C. Yao, "Mixed model assembly simulation," *International Journal of Production Economics*, vol. 26, no. 1-3, pp. 161–167, 1992.

[16] C. Rippenhagen and S. Krishnaswamy, "Implementing the theory of constraints philosophy in highly reentrant systems," in *Proceedings of the 1998 Winter Simulation Conference*, (Piscataway, New Jersey), pp. 993–996, 1998.

[17] M.-G. Resende, "A program for simulation of semiconductor wafer fabrication," tech. rep., University of California, Berkeley, Operations Research Center, 1985.

[18] B. Tullis, V. Mehrotra, and D. Zuanich, "Successful modeling of a semiconductor R & D facility," in *Proceedings of the 1990 IEEE/SEMI International Semiconductor Manufacturing Science Symposium*, (Burlingame, California, United States), pp. 26–32, 1990.

[19] M. Thompson, "Using simulation-based finite capacity planning and scheduling software to improve cycle time in front end operations," in *Proceedings of 1995 IEEE/SEMI Advanced Semiconductor Manufacturing Conference Workshop*, pp. 131–135, 1995.

[20] J. Fowler, H. Brown, S.and Gold, and A. Schoemig, "Measurable improvements in cycle-time-constrained capacity," in *Proceedings of IEEE International Symposium On Semiconductor Manufacturing Conference*, (San Francisco, United States), pp. 21–24, 1997.

[21] A. J. Weintraub, A. Zozom Jr, T. J. Hodgson, and D. Cormier, "A simulation-based finite capacity scheduling system," in *Proceedings of the 29th conference on Winter simulation*, pp. 838–844, IEEE Computer Society, 1997.

[22] N. Grewal, A. Bruska, T. Wulf, and J. Robinson, "Integrating targeted cycle-time reduction into the capital planning process," in *Proceedings of the 1998 Winter Simulation Conference– WSC 1998*, (Washington, United States), pp. 1005–1010, 1998.

[23] K. Potti and S. J. Mason, "Using simulation to improve semiconductor manufacturing," *Semiconductor International*, vol. 20, no. 8, pp. 289–292, 1997.

[24] A. A. B. Pritsker and K. Snyder, "Production scheduling using FACTOR," in *The Planning and Scheduling of Production Systems*, pp. 337–358, Springer US, 1997.

[25] J. P. Ignizio and H. Garrido, "Fab simulation and variability," *Future Fab International*, vol. 41, pp. 41–45, 2012.

[26] J. G. Shanthikumar, S. Ding, and M. T. Zhang, "Queueing theory for semiconductor manufacturing systems: A survey and open problems," *IEEE Transactions on Automation Science and Engineering*, vol. 4, no. 4, pp. 513–522, 2007.
[27] S. Bermon and S. Hood, "Capacity optimization planning system (CAPS)," *Interfaces*, vol. 29, no. 5, pp. 31–50, 1999.
[28] J. Swaminathan, "Tool capacity planning for semiconductor fabrication facilities under demand uncertainty," *European Journal of Operational Research*, vol. 120, no. 3, pp. 545–558, 2000.
[29] F. Barahona, S. Bermon, O. Günlük, and S. Hood, "Robust capacity planning in semiconductor manufacturing," *Naval Research Logistics (NRL)*, vol. 52, no. 5, pp. 459–468, 2005.
[30] B. Çatay, c. Erengüç, and A. Vakharia, "Tool capacity planning in semiconductor manufacturing," *Computers & Operations Research*, vol. 30, no. 9, pp. 1349 – 1366, 2003.
[31] N. Geng and Z. Jiang, "A review on strategic capacity planning for the semiconductor manufacturing industry," *International Journal of Production Research*, vol. 47, no. 13, pp. 3639–3655, 2009.
[32] Y.-F. Hung and R. C. Leachman, "A production planning methodology for semiconductor manufacturing based on iterative simulation and linear programming calculations," *IEEE Transactions on Semiconductor Manufacturing*, vol. 9, no. 2, pp. 257–269, 1996.
[33] R. C. Leachman, "Modeling techniques for automated production planning in the semiconductor industry," in *Optimisation in Industry: Mathematical Programming and Modeling* (T. Ciriani and R. Leachman, eds.), (Wiley, New York), pp. 1–30, 1993.
[34] C. Habla, L. Mönch, and R. Drissel, "A finite capacity production planning approach for semiconductor manufacturing," in *Proceedings of the 3rd Annual IEEE Conference on Automation Science and Engineering*, (Scottsdale, United States), pp. 82–87, 2007.
[35] S. Bermon, G. Feigin, and S. Hood, "Capacity analysis of complex manufacturing facilities," in *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, vol. 2, pp. 1935–1940, 1995.
[36] Y. Hsiung, M.-C. Wu, and H.-M. Hsu, "Tool planning in multiple product-mix under cycle time constraints for wafer foundries using genetic algorithm," *Journal of the Chinese Institute of Industrial Engineers*, vol. 23, no. 2, pp. 174–183, 2006.
[37] J. Bard, Y. Deng, R. Chacon, and J. Stuber, "Midterm planning to minimize deviations from daily target outputs in semiconductor manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 23, no. 3, pp. 456–467, 2010.
[38] J. C. Chen, C. W. Chen, C. J. Lin, and H. Rau, "Capacity planning with capability for multiple semiconductor manufacturing fabs," *Computers and Industrial Engineering*, vol. 48, no. 4, pp. 709–732, 2005.
[39] J. C. Chen, Y.-C. Fan, and C.-W. Chen, "Capacity requirements planning for twin fabs of wafer fabrication," *International Journal of Production Research*, vol. 47, no. 16, pp. 4473–4496, 2009.
[40] J. C. Chen, L.-H. Su, C.-J. Sun, and M.-F. Hsu, "Infinite capacity planning for IC packaging plants," *International Journal of Production Research*, vol. 48, no. 19, pp. 5729–5748, 2010.
[41] H. E. Fargher, M. A. Kilgore, P. J. Kline, and R. A. Smith, "A planner and scheduler for semiconductor manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 7, no. 2, pp. 117–126, 1994.
[42] K. Horiguchi, N. Raghavan, R. Uzsoy, and S. Venkateswaran, "Finite-capacity production planning algorithms for a semiconductor wafer fabrication facility," *International Journal of Production Research*, vol. 39, no. 5, pp. 825–842, 2001.
[43] K. Habenicht and L. Mönch, "A finite-capacity beam-search-algorithm for production scheduling in semiconductor manufacturing," in *Simulation Conference, 2002. Proceedings of the Winter*, vol. 2, pp. 1406–1413, IEEE, 2002.
[44] T. J. Chua, M. W. Liu, F. Y. Wang, W. J. Yan, and T. X. Cai, "An intelligent multi-constraint finite capacity-based lot release system for semiconductor backend assembly environment," *Robotics and Computer-Integrated Manufacturing*, vol. 23, no. 3, pp. 326–338, 2007.
[45] J. D. Little, "A proof for the queuing formula: L= λ w," *Operations research*, vol. 9, no. 3, pp. 383–387, 1961.
[46] J. S. Kim and R. C. Leachman, "Decomposition method application to a large scale linear programming wip projection model," *European Journal of Operational Research*, vol. 74, no. 1, pp. 152–160, 1994.
[47] Y. Lee, S. Kim, S. Yea, and B. Kim, "Production planning in semiconductor wafer fab considering variable cycle times," *Computers & Industrial Engineering*, vol. 33, no. 3-4, pp. 713–716, 1997.
[48] N. Govind and D. Fronckowiak, "Setting performance targets in a 300mm wafer fabrication facility," in *Proceedings of Advanced Semiconductor Manufacturing Conference and Workshop*, pp. 75–79, 2003.
[49] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the Theory of NPCompleteness*. New York, NY, USA: W. H. Freeman & Co., 1979.
[50] T. Winkler, P. Barthel, and R. Sprenger, "Modeling of complex decision making using forward simulation," in *Proceedings of the 2016 Winter Simulation Conference*, pp. 2982–2991, IEEE Press, 2016.
[51] D. Martin, "Key factors in designing a manufacturing line to maximize tool utilization and minimize turnaround time," in *Semiconductor Manufacturing Science Symposium, 1993. ISMSS 1993., IEEE/SEMI International*, pp. 48–53, 1993.
[52] J. Montoya-Torres, "Manufacturing performance evaluation in wafer semiconductor factories," *International Journal of Productivity and Performance Management*, vol. 55, no. 3/4, pp. 300–310, 2006.

Emna Mhiri is PhD student in G-SCOP Laboratory (www.g-scop.grenoble-inp.fr). She received industrial engineering degree from engineering school of Tunisia, in 2012 and completed her masters degree in industrial engineering from the University of Grenoble, France in 2013. Her research interests include capacity planning in semiconductor industry. Her email address is Emna.Mhiri@grenoble-inp.fr.

Fabien Mangione is assistant professor in G-SCOP Laboratory. He received the Ph.D. degree in industrial engineering from the University of Grenoble, France and works on production planning, particularly on industrial case studies. His research also deals with lot sizing problems on supply chain modeling. His email address is fabien.mangione@grenoble-inp.fr.

Mireille Jacomino is professor in G-SCOP Laboratory. She is carrying out her research in combinatorial optimization of systems. Her application fields are manufacturing and energy systems, used particularly in execution context of systems. Her works aim at computing control decisions that guaranty performance during execution. Robust control and robust decision are the key research interests of professor Jacomino to address the uncertainties. Her email address is Mireille.Jacomino@grenoble-inp.fr.

Philippe Vialletelle is principal staff engineer at the Industrial Engineering department of STMicroelectronics Crolles300. He is in charge of the definition and follow-up of collaborative projects in Manufacturing Sciences. His fields of interest cover production planning and management techniques, process control and Big data. His email address is philippe.vialletelle@st.com.

Guillaume Lepelletier is senior project leader at STMicroelectronics. He received Engineering degree in Operations and Production Management from INSA de Lyon, France and Master of Science in "Advanced Modeling Systems" from Brunel University, Uxbridge, UK in 1997. He has 15 years of professional experience in Industrial Engineering in the semiconductor industry. He is working on capacity planning, cycle time management, discrete event simulation, industrial reporting and equipment performance tracking. His email address is guillaume.lepelletier@st.com.