Christ, Q., Dauzère-Pérès, S., & Lepelletier, G. (2023). A three-step approach for decision support in operational production planning of complex manufacturing systems. International Journal of Production Research, 61(17), 5860-5885.

https://doi.org/10.1080/00207543.2022.2118387

A Three-Step Approach for Decision Support in Operational Production Planning of Complex Manufacturing Systems

Quentin Christ[1,2] Stéphane Dauzère-Pérès[1,3] Guillaume Lepelletier[2]

[1]Mines Saint-Etienne, Univ Clermont Auvergne
CNRS, UMR 6158 LIMOS
CMP, Department of Manufacturing Sciences and Logistics
F-13541 Gardanne, France
E-mail: quentin.christ@st.com
E-mail: dauzere-peres@emse.fr

[2]STMicroelectronics Crolles
F-38926 Crolles, France
E-mail: guillaume.lepelletier@st.com

[3]Department of Accounting and Operations Management
BI Norwegian Business School
0484 Oslo, Norway

Corresponding author and contact details:
Quentin Christ
Mines Saint-Etienne, Univ Clermont Auvergne
CNRS, UMR 6158 LIMOS
CMP, Department of Manufacturing Sciences and Logistics
F-13541 Gardanne, France
E-mail: quentin.christ@st.com

ARTICLE HISTORY
Compiled December 18, 2022

ABSTRACT
In this paper, a practical relevant operational production planning problem in complex manufacturing systems is adressed. In this problem, lots are planned individually to provide a more detailed plan than approaches that only consider production quantities. A three-step approach, which is currently fully integrated and used in a Decision Support System, is then introduced. This work follows the one of Mhiri et al. (2018) who addressed this problem. We push the approach a step further by introducing new optimization possibilities through new smoothing rules, whose performance is studied according to different indicators. Furthermore, we present the production planning process in which the decision support tool is embedded and how it bridges the gap between the upper and lower planning levels.

KEYWORDS
Production Planning, Complex Manufacturing Systems, Semiconductor Manufacturing, Heuristic, Optimization

1. Introduction

1.1. *Semiconductor Industry and Production Planning*

The electronics industry is one of the largest industries in the world. At the heart of this industry is one of the most complex systems, namely the manufacture of integrated circuits on thin silicon discs (wafers). Semiconductor manufacturing is one of the most important and critical manufacturing sectors, particularly nowadays with the well-known shortages of integrated circuits. This manufacturing process is divided into four stages: Wafer manufacturing, wafer probe, packaging and final test (Uzsoy, Lee, and Martin-Vega (1992)). Wafer manufacturing is recognized as the most complex stage because it corresponds to a flexible job-shop manufacturing environment with hundreds of parallel non identical machines, reentrant process flows, hundreds of operations for each product and therefore very long lead times (generally two to three months). Production and development lots are often processed on the same equipment, and various process constraints must be taken into account (Uzsoy, Lee, and Martin-Vega (1992); Mönch, Fowler, and Mason (2012)). Due to their high cost, machines are usually heterogeneous and therefore the same operations can be performed on machines of different generations (Gupta et al. (2006)). Semiconductor manufacturing is a very competitive environment and it is essential for companies to maximize their service level and in particular their ability to meet their delivery dates. An optimized production management is becoming critical and therefore justifies the growing interest that these issues have had in the literature.

Production planning problems in the semiconductor industry have been studied for decades (Bitran, Haas, and Hax (1981); Leachman and Carmon (1992); Mönch, Fowler, and Mason (2012)). Planning problems can take different shapes and names, depending on the scale considered, but can generally be divided into three levels, namely strategic, tactical and operational (Anthony (1965)).

At the border between the strategic and tactical levels is the function of Master Planning. The objective of this function is, considering the future demand of customers as well as the entire supply chain, to coordinate the different production sites by defining for each of them production objectives by period (usually weeks or months). These production objectives are the inputs of the production planning function. This function, generally at the factory level, aims to determine for each period (generally days or weeks) product quantities to be started in production to meet the Master Planning objectives. Master Planning in semiconductor manufacturing is discussed in Mönch, Uzsoy, and Fowler (2018), and interesting contributions can be found in Aouam and Uzsoy (2015) and Zhang et al. (2020).

Finally, production scheduling controls the assignment and the sequencing of production lots on resources to optimize production objectives such as on-time delivery or completion times. This hierarchical process from production planning to production scheduling is sufficient in most manufacturing systems. However, when considering the manufacturing of integrated circuits (front end manufacturing or wafer manufacturing), a production lot requires on average 700 operations to complete its fabrication. The operating time of an operation varies from about 10 minutes to more than 12 hours. Considering the thousands of lots present at the same time in the factory, this leads to millions of decisions to be made by the production scheduling function to control the production flows, with the start and due dates of the lots as the main instructions. There is therefore a significant gap in decision-making between the two planning functions, production planning and production scheduling. Indeed, the latter cannot optimize the management of all the stages of the lots in the factory on the only basis of the incoming and outgoing lots of the factory. As a result, production scheduling generally only considers subsets of machines (see for instance Yugma et al. (2012) or Zhang, Lv, and Zhang (2018)) and only optimize local criteria such as the average cycle time of lots in a workshop (i.e. the average time spent by lots in the workshop), the machine utilization rate or the machine throughput. It is therefore essential to have an additional function, as proposed by Govind et al. (2008), whose aim is to ensure that products are processed at the right time in each workshop in order to meet the production objectives defined by the Master Planning. This production flow forecast also allows, for instance, a better planning of preventive maintenance operations.

1.2. *Operational Production Planning*

This additional function is little studied, as pointed out by Mönch, Uzsoy, and Fowler (2018), and only a few papers have studied this interface between Production Planning and Shop-Floor Scheduling. Horiguchi et al. (2001) seek to define the time period in which each pair (lot, operation) must be processed. Only operations related to near-bottleneck machines are considered. To define these short-term production plans, the authors use a classical forward scheduling heuristic for lots in the WIP (Work In Progress i.e. products currently in the facility) and a backward scheduling for orders not filled by the WIP. Numerical experiments using a simulation model show the value of driving workshop scheduling on the basis of data (lot, operation, period) rather than on the basis of lot delivery dates, in particular by reducing the overall delay. A similar operational production planning problem has been studied by Habenicht and Mönch (2002), using a Beam-Search based method which they tested on simplified instances, aggregating processing steps into operations, and have shown the benefit of considering this short-term production planning problem. Habla, Mönch, and Drießel (2007) model the problem as a Mixed Integer Program (MIP), and solve it using a Lagrangian Relaxation approach on a simplified version of the problem where only what they call *bottleneck steps are considered.* The method is also applied by Bard et al. (2010) to a very similar problem called "Manufacturing Planning Problem". They also evaluate Benders Decomposition, showing the computational intractability of the problem. Then, they develop a greedy approach by solving each period of the planning horizon with the MIP and using a rescheduling heuristic to better distribute the product quantities between the machines for each period. The approach shows positive results but could only be tested on instances consisting of three product families. Recently, Mhiri et al. (2018) also propose a MILP to model this operational production planning problem and develop a heuristic approach used for the WIP projection in semiconductor manufacturing systems.

In this paper, we pursue and improve the work in Mhiri et al. (2018), bringing some corrections and improvements to the presented mathematical model, which is detailed in Section 2. We also improve the heuristic approach in different ways.

First, the three-step heuristic in Mhiri et al. (2018), only tries to optimize customer criteria (i.e. due date performance-related criteria, see the end of Section 4.1). However, in the semiconductor industry, other indicators are also equally important, such as the throughput of the fab or the machine utilization rate. Thus, we present different variants of the approach to optimize these different indicators, and conduct a comparative study. Since the work presented in Mhiri et al. (2018), the approach is fully implemented and integrated in a Decision Support System which is now used to define weekly production plans in the 200mm and 300mm factories of Crolles in France (Christ et al. (2018)). Therefore, unlike Mhiri et al. (2018) for whose the objective was to simulate the evolution of the very large number of lots, in our case, the objective is to provide a production plan that can be followed (as it respects production capacity), but that is also the best possible solution (according to the various indicators analyzed) in addition to being as close as possible to the reality.

Although this problem has not been studied much, it corresponds to a concrete case of operational production planning in STMicroelectronics factories. The production flow corresponds to complex job-shop manufacturing with re-entrant flows. Each fab includes numerous (more than 300) very heterogeneous machines grouped in work-centers. The average WIP consists of more than 3,000 lots, divided into several hundred different products, with manufacturing routes involving an average of 700 operations. The planning department must, on the basis of a given weekly release plan and a given weekly delivery plan (generally several thousand wafers per week), give daily production instructions to optimize production flows. This plan is generally defined for 8 weeks, and is re-evaluated each week. The main decisions are related to the priorities of lots, and instructions on the products to perform in priority on each machine group. The eight-week plan also makes it possible to foresee future resource requirements, which can be prepared for example by adjusting preventive maintenance plans or by qualifying some machines to support the processing of critical operations (not enough capacity for the number of wafers to produce). This weekly operational plan is determined by using a decision support system that relies on the three-step approach presented in this article. The approach has been designed for high-mix low-volume semiconductor manufacturing systems, which are amongst the most complex manufacturing systems. The approach can thus be used for less complex manufacturing systems, including low-mix high-volume semiconductor manufacturing systems. Generally speaking, the approach is adapted to manufacturing systems where products have relatively long manufacturing routes (typically more than 10 operations) and cycle times (typically more than one week).

The remainder of this paper is structured as follows. In Section 2, we introduce the problem and present a Mixed Integer Linear Program (MILP) to model the problem of interest. We also compare our problem with classical production planning problems (scheduling, lot-sizing) by showing their similarities and their differences. Section 3 presents the three-step approach initially proposed in Mhiri et al. (2018) to efficiently determine detailed production plans. Section 4 details the different smoothing rules used in the module which integrates capacity constraints, and provides a comparative study of the smoothing rules based on industrial instances presented in Section 5.2. Section 6 provides details on the decision support tool utilization inside the production planning process. Finally, in Section 7, the contributions of the paper are summarized and perspectives are discussed.

2. Problem modelling

2.1. *Problem definition*

First, we present the problem tackled by the heuristic approach, and the associated mathematical model. Differences with the model in Mhiri et al. (2018) are detailed in Section 2.3.

The planning horizon is decomposed into T periods and each period $t \in \{1,\dots,T\}$ has a fixed duration p. A set of machines $\mathscr{M}$ and a set of lots $\mathscr{L}$ currently in the fab (usually called WIP) as well as lots to be started, are considered. Each lot l has a customer delivery period also called due date d_l, a size q_l and a release period r_l. Each lot l requires a set $\mathscr{O}_l$ of consecutive operations to be completed, often called a route. Operation o of a lot l can be processed by a set of machines $\mathscr{M}_{o,l,t}$ in period t with a processing time $p_{o,l,m}$ that depends on machine m. Each machine m has a production capacity $c_{m,t}$ in period t (i.e. it is available for $c_{m,t}$ time units in period t). The completion period C_l of lot l depends on the plan. The tardiness T_l of lot l is defined as $\max(C_l - d_l, 0)$. One of the objectives is to find a production plan that minimizes the total tardiness (TT), i.e., $\sum T_l$ (other objectives are also studied in Section 5.2). A solution is a plan that defines, for each operation o of lot l, the period t in which operation o is processed, defined by the variable $X_{o,l,t} \in \{0,1\}$. A number of constraints are considered, the most important being capacity constraints.

The remaining assumptions are listed below:

- The length of a period is an input parameter. In the context of the STMicroelectronics factories (and in our computational experiments), the period length is generally one week. This period length is a good compromise, as longer periods might lead to a loss of model accuracy, while periods that are too short could conflict with some long operations (more than 24 hours on some diffusion machines) and lead to an unsolvable MILP, as it might not be possible to assign to the same period the start and end dates of these long operations.
- As in Mhiri et al. (2018), it is possible to split the processing of a wafer over several machines. This hypothesis is not real, but it simplifies the problem to be solved for the heuristic approach. This hypothesis remains justifiable as the solutions of this problem are meant to be input for detailed scheduling approach which, for their part, consider the non-preemption constraints.
- The variables T_l have integer values. Indeed, in the semiconductor industry, lots are generally not shipped every day, but rather every week. If a lot is not shipped on time, it will leave at the end of the following period, hence the integer values.
- Batches and setup times are integrated as inefficiencies using discount factors for each machine m to reduce it production capacity ($c_{m,t}$). For setup times, the discount factor is based on the proportion of time in a day that the lot spent in a setup state. Regarding batching, the regular capacity is multiplied by the maximum batch size measured in number of lots, discounted by a factor that takes into account the fact that batches are not always fully loaded. This data is maintained by the Industrial Engineering team which is in charge of the machine capacity model. These simplifying assumptions again come from the desire not to consider detailed scheduling decisions and to reduce complexity.
- Storage constraints are not considered, as they are generally not critical in a semiconductor manufacturing facility.

2.2. *Notations*

The notations for the mathematical modeling of our problem are summarized in Table 1.

2.3. *Mathematical Model*

Using the notations above, the problem, as it is solved by the three-step-heuristic introduced in the following section, can be modelled as the Mixed Integer Linear Program (P_1) below.

Table 1.: Problem Notation

Sets, Indices and Parameters	Description
$\mathscr{L}$	Set of lots
$l \in \mathscr{L}$	Lot index
$\mathscr{O}_l$	Set of consecutive operations required to complete lot l
$o \in \mathscr{O}_l$	Operation index of lot l
$\mathscr{M}$	Set of machines
$m \in \mathscr{M}$	Machine index
T	Number of periods in the planning horizon
$t \in \{1,\ldots,T\}$	Period index
$\mathscr{M}_{o,l,t}$	Set of machines that can process operation o of lot l in period t
s_t	Start date of period t ($s_0 = 0$)
q_l	Number of wafers in lot l (i.e., size of lot)
$r_l \in \{1,\ldots,T\}$	Release period of lot l
$d_l \in \{1,\ldots,T\}$	Due date of lot l, i.e. period in which lot can be delivered without delay penalty
$p_{o,l,m}$	Time required to process operation o on machine m per wafer of lot l
$c_{m,t}$	Capacity of machine m in period t
Variables	**Description**
$S_{o,l} \in \mathbb{R}^+$	Start date of operation o of lot l
$E_{o,l} \in \mathbb{R}^+$	End date of operation o of lot l
$T_l \in \mathbb{N}^+$	Tardiness of lot l (in number of periods)
$Q_{o,l,m,t} \in \mathbb{R}^+$	Number of wafers of lot l at operation o processed by machine m in period t
$X_{o,l,t} \in \{0,1\}$	Is equal to 1 if operation o of lot l is processed in period t, and 0 otherwise

$$(P_1) \quad min \quad \sum_{l=1}^{L} T_l \tag{1}$$

$$s.c. \quad X_{1,l,r_l} = 1 \qquad l \in \mathscr{L} \tag{2}$$

$$S_{o,l} \geq E_{o-1,l} \qquad l \in \mathscr{L} \quad o \in \mathscr{O}_l \tag{3}$$

$$S_{o,l} + \sum_{m \in \mathscr{M}_{o,l,t}} (p_{o,l,m} Q_{o,l,m,t}) \leq E_{o,l} \qquad l \in \mathscr{L} \quad o \in \mathscr{O}_l \quad t \in \{1,\ldots,T\} \tag{4}$$

$$\sum_{t=r_l}^{T} X_{o,l,t} = 1 \qquad l \in \mathscr{L} \quad o \in \mathscr{O}_l \tag{5}$$

$$S_{o,l} \geq \sum_{t=r_l}^{T} (s_t X_{o,l,t}) \qquad l \in \mathscr{L} \quad o \in \mathscr{O}_l \tag{6}$$

$$E_{o,l} \leq \sum_{t=r_l}^{T} (s_{t+1} X_{o,l,t}) \qquad l \in \mathscr{L} \quad o \in \mathscr{O}_l \tag{7}$$

$$T_l \geq \sum_{t=r_l}^{T} (t X_{O_l,l,t}) - d_l \qquad l \in \mathscr{L} \tag{8}$$

$$\sum_{l=1}^{L} \sum_{o=1}^{O_l} (p_{o,l,m} Q_{o,l,m,t}) \leq c_{m,t} \qquad m \in \mathscr{M} \quad t \in \{1,\ldots,T\} \tag{9}$$

$$\sum_{m \in \mathscr{M}_{o,l,t}} Q_{o,l,m,t} = q_l X_{o,l,t} \qquad l \in \mathscr{L} \quad o \in \mathscr{O}_l \quad t \in \{1,\ldots,T\} \tag{10}$$

$$S_{o,l},\ E_{o,l},\ Q_{o,l,m,t} \geq 0 \qquad l \in \mathscr{L} \quad o \in \mathscr{O}_l \quad t \in \{1,\ldots,T\} \quad m \in \mathscr{M} \tag{11}$$

$$X_{o,l,t} \in \{0,1\} \qquad l \in \mathscr{L} \quad o \in \mathscr{O}_l \quad t \in \{1,\ldots,T\} \tag{12}$$

$$T_l \in \mathbb{N}^+ \qquad l \in \mathscr{L} \tag{13}$$

The objective (1) is to minimize the sum of tardiness of all lots. Constraints (2) ensure that lots must start in their release period. Constraints (3) are precedence constraints between two consecutive operations of the same lot. Constraints (4) ensure that the end period of each operation is larger than the start date plus the processing time of the operation. Each operation is guaranteed to be executed in a single period through Constraints (5), while Constraints (6) and (7) force each operation to end in the same period it started. They also link variables $S_{o,l}$ and $E_{o,l}$ with variables $X_{o,l,t}$. Constraints (8) define the tardiness for each lot, and the capacity constraints for each machine at each period are ensured through Constraints (9). Constraints (10) ensure that all the wafers (also defined as the size) in each lot are processed and allow preemption. Allowing the splitting of the wafers of a lot on multiple machines, although it is not done in practice, is common in production planning to model capacity and handle large complex problems. The discrete assignment of lots to machines is done at the scheduling level. Constraints (11) guarantee that start and end dates, as well as quantities processed on machines, are positive. Constraints (12) and (13) are integrity constraints.

This model is a rewritten version of the model in Mhiri et al. (2018), with the same characteristics and several corrections to ensure that the model is consistent with the considered operational production planning problem. The most critical correction is related to how the workload allocation on the machines is defined. In Mhiri et al. (2018), the number of wafers of lot l at operation o processed by machine m is a given parameter $a_{o,l,m}$, while this is a variable ($Q_{o,l,m,t}$) in our model. This difference is important for two reasons:

- Mhiri et al. (2018) do not explain how the parameters $a_{o,l,m}$ are defined, in particular in the computational experiments. The most likely is that the authors rely on the historical data distribution and assume that this distribution remains the same in the future. This assumption is hardly acceptable given the high variability in semiconductor manufacturing systems, as we observed for instance in our industrial instances. The workload distribution on the machines changes over time in particular because machines are heterogeneous. Constraints (4), (9) and (10) are thus different in our model than in the model in Mhiri et al. (2018).
- Workload balancing on the machines is essential in operational production planning, and enforcing the workload distribution prevents the use of a critical lever when optimizing the production plan.

Three other differences with the model in Mhiri et al. (2018) are listed below, the first one being the more significant:

(1) In Mhiri et al. (2018), the process time of a lot is counted in the capacity of the start period of the operation although the operation might end in the next period. This simplification may lead to a very poor estimation of the capacity consumption in two consecutive periods, in particular for operations that start at the very end of a period. In our model, operations must start and end in the same period. Constraints (4) are thus different in our model than in the model in Mhiri et al. (2018).

(2) As already mentioned, the due dates are integer, which better fit the industrial reality.

(3) There is no weight per lot since weights are not considered in operational planning. In Mhiri et al. (2018), all weights are identical in the computational experiments.

Although it has its own characteristics, our model is also related to previous research, in addition to Mhiri et al. (2018), as discussed below.

The definition of start and end dates for each operation, as well as precedence constraints, are characteristic of shop-floor scheduling problems such as in Singer and Pinedo (1998). But the scheduling of operations of lots on the machines is not considered in (P_1), and is replaced by a production capacity constraint per period for each machine.

This use of a capacity per period is a characteristic of capacitated lot sizing problems and its variants (Karimi, Ghomi, and Wilson (2003)). However, in (P_1), besides production quantities (lot sizes) to produce in each period, we also need to determine the start and end dates of each operation.

The model presented in Habla, Mönch, and Drießel (2007) is probably the closest to ours. Their objective is to define, for some operations (those on bottleneck machines) of each lot, the period during which they should be processed. A capacity per period for groups of machines is also taken into account. However, the model in Habla, Mönch, and Drießel (2007) does not consider all the process operations and, more importantly, does not manage the distribution of the workload on heterogeneous machines, i.e. machines which can process common process operations but with different processing speeds, which is a critical point in semiconductor manufacturing.

As highlighted in Garey and Johnson (1979) and mentioned in Mhiri et al. (2018), production planning, capacity planning and scheduling problems in complex job shops like semiconductor manufacturing are known to be strongly NP-hard. Although we do no detail the proof in this paper, it is possible to show that our problem can be reduced to the well known NP-hard bin packing problem, even with one machine and one operation per lot. Thus, this justifies the use of a heuristic for the operational production planning decision support tool which is introduced in the next section.

3. A three-step approach for operational production planning

3.1. *Overview of the approach*

As previously mentioned, semiconductor manufacturing involves planning and scheduling problems with thousands of lots, each lot requiring hundreds of unit operations. Therefore, considering the problem complexity, it is highly unlikely that exact methods could solve real life instances. Hence, a planning Decision Support System has been developed, based on an approach in which the planning problem is decomposed into three main modules as shown in Figure 1.

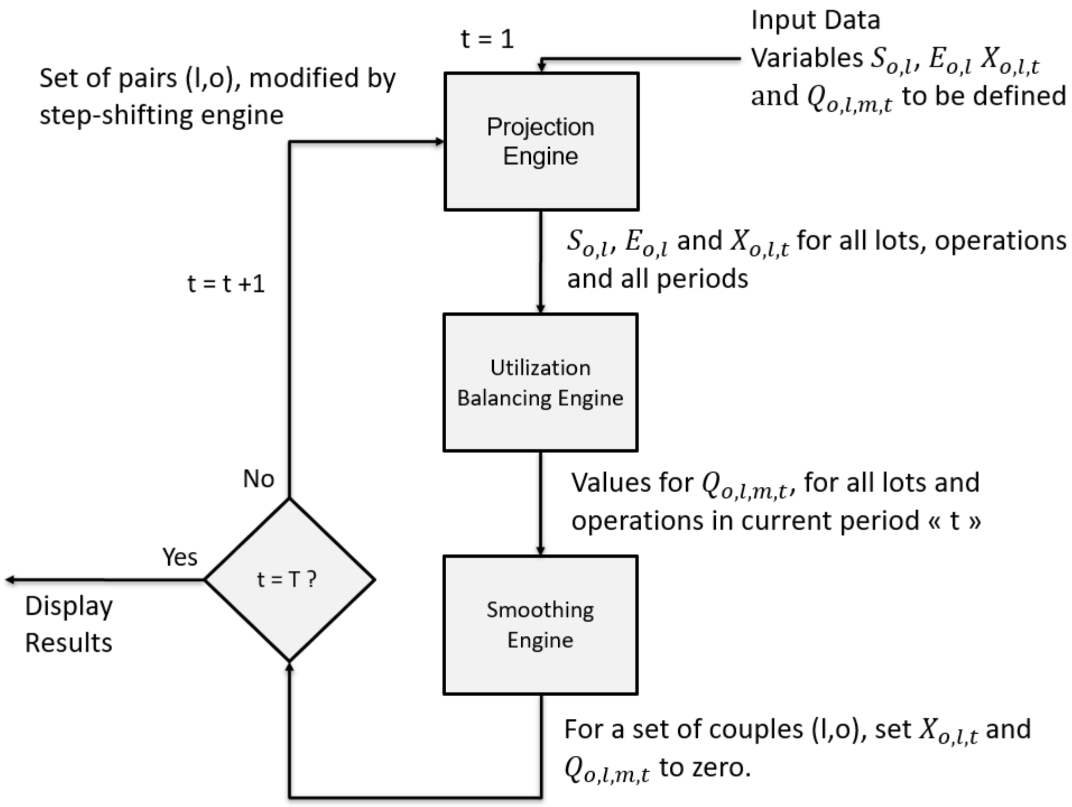
Input Data
Variables $S_{o,l}$, $E_{o,l}$ $X_{o,l,t}$ and $Q_{o,l,m,t}$ to be defined
t = 1
Set of pairs (l,o), modified by step-shifting engine
Projection Engine
$S_{o,l}$, $E_{o,l}$ and $X_{o,l,t}$ for all lots, operations and all periods
t = t +1
Utilization Balancing Engine
No
Values for $Q_{o,l,m,t}$, for all lots and operations in current period « t »
Yes
t = T ?
Display Results
Smoothing Engine
For a set of couples (l,o), set $X_{o,l,t}$ and $Q_{o,l,m,t}$ to zero.

Figure 1.: Flowchart of the three-step approach.
Alt Text: Flowchart of the three-step approach consisting of 3 successive blocks with a loop repeating for each period t. Each block fixes a part of the problem variables for the current period.

Following most of the approaches in the literature, the planning horizon is discretized into periods. The three-step approach runs each module in each period of the planning horizon. The overall idea of the approach is as follows:

(1) First, in the *Projection Engine* module, from historical cycle times that include the waiting times between operations, start and end processing dates are assigned to each operation of each lot, and thus the period in which the operation is processed (thus determine the variables $S_{o,l}$, $E_{o,l}$ and $X_{o,l,t}$). This first initialization step does not take capacity constraints into account. They are considered in the next two steps of the approach to adjust the initial theoretical cycle time of each lot.
(2) In the second step, an *Utilization Balancing Engine* is run to allocate lot processing workload to the available machines (still without taking into account capacity constraints), which gives an estimate of the equipment utilization rate. Since capacity constraints are not considered, some machines may be overloaded and the solution may therefore not be feasible.
(3) Then, a third module, the *Smoothing Engine* described in Section 3.2, is in charge of postponing lots to later periods in order to smooth the workload and satisfy capacity constraints.

After processing on all periods, the end result is a feasible solution which respects the capacity constraints.

This three-step approach is therefore a constructive heuristic which defines a solution by taking decisions iteratively until a complete solution is obtained. This approach has already been applied in semiconductor manufacturing, for instance for scheduling problems in Mason, Fowler, and Matthew Carlyle (2002) or for a problem closer to ours in Horiguchi et al. (2001), and are still studied in a variety of applications as in biopharmaceutical production (Oyebolu et al. (2017)).

The well known MRP II approach is also a constructive procedure and the three-step heuristic follows the same principles than MRP II (time phasing, capacity requirements). However, MRP II does not detail how to apply the logic to specific settings. Because of their complexity and characteristics, semiconductor manufacturing facilities, in particular the "front-end" ones considered in this article, often use locally developed approaches (Mönch, Fowler, and Mason (2012)). Our approach does for instance consider a fairly accurate modeling of the workload allocation on heterogeneous machines, with variable processing times that depend on the products and the machines (module 2). Moreover, the limited capacity of the machines leads to choices in the operations to be carried out for each lot at each period. This choice may vary according to the current company's strategy, and the smoothing approach of module 3 (Section 3.2) can be selected based on this strategy.

This paper is mainly dedicated to the development and the study of the smoothing procedure (third module), and the next section will be dedicated to its description. Thus, for the sake of brevity, we will not go into further detail on the projection and balancing modules. For the projection module, which has not been modified, we refer the reader to Mhiri et al. (2018) to get more details on how this module operates. Regarding the balancing module, several important changes have been made. The main changes concern the linear programming approach used to balance the workload on machines. Instead of a one-pass resolution via an objective function that tries to smooth the workload on the machines, an iterative approach is used where the maximum workload is minimized at each iteration, and dual variables of the capacity constraints are used to characterize the biding machines. This approach allows "Min-Max Fair" solutions to be obtained, with interesting properties for our problem. The approach and its relevance are described in details in Christ, Dauzère-Pérès, and Lepelletier (2019).

3.2. *Smoothing module*

Entering this step, we already have assigned values to all variables $S_{o,l}$, $E_{o,l}$, $X_{o,l,t}$ and $Q_{o,l,m,t}$. So, a complete production plan is defined for period t, which may not meet capacity constraints. The smoothing module builds a feasible solution from this initial plan, using a forward smoothing process, an approach used for instance in capacitated lot-sizing problems (Trigeiro, Thomas, and McClain (1989); Brahimi, Dauzère-Pérès, and Najid (2006); Lu, Zhang, and Han (2013)). The idea is to select some lots in the current period and shift them to the next period in order to reduce the induced utilization rate on the machines. Algorithm 1 describes the general smoothing procedure.

Algorithm 1 takes as inputs the period, the production and allocation plans, i.e. the $X_{o,l,t}$ and $Q_{o,l,m,1}$ variables. As long as there are overloaded machines, the algorithm first selects the most overloaded machine. Then, among its assigned operations, the algorithm selects the one of the lot with the lowest priority. This priority can be assessed in several ways, which are discussed in Section 4. In the algorithm used as an example, the lot priority is defined according to its due date. Thus, a lot with an early due date has a higher priority than one with a later due date. The lot with the latest due date will tend to be postponed, which is equivalent to the well-known Earliest Due Date scheduling rule.

Once the lower priority lot (at a given operation) is selected, it is postponed (as well as all

Algorithm 1: Smoothing Procedure

```
Input  : $t$ = Current period
         $X_{o,l,t}$, $Q_{o,l,m,t}$ = Assigned values for all lots, operations and machines
         $W_{m,t}$ = Utilization rate of machine $m$ in period $t$
Output: $\mathscr{O}_{shifted}$ = Set of operations shifted to next period
$\mathscr{O}_{shifted} \leftarrow \emptyset$
while $\max_{m \in \mathscr{M}} W_{m,t} > 1$ do
    // Select the most overloaded machine
    $m' \leftarrow \mathrm{argmax}_{m \in \mathscr{M}} W_{m,t}$
    $\mathscr{O}^{m'} \leftarrow \{o \in \mathscr{O}_l; \forall l \in \mathscr{L} \mid Q_{o,l,m',t} > 0\}$
    // Select the operation processed by the machine whose lot due date is the farthest
    $o' \leftarrow \mathrm{argmax}_{o \in \mathscr{O}^{m'}}(dd_l)$
    for $o \in [o', ..., O_l]$ do
        if $X_{o,l,t} = 1$ then
            $X_{o,l,t} \leftarrow 0$
            $\mathscr{O}_{shifted} \leftarrow \mathscr{O}_{shifted} \cup \{o\}$
        for $m \in \mathscr{M}$ do
            $Q_{o,l,m,t} \leftarrow 0$
        end
    end
    for $m \in \mathscr{M}$ do
        $W_{m,t} \leftarrow \frac{\sum_{o \in \mathscr{O}_l}(p_{o,l,m} Q_{o,l,m,t})}{c_{m,t}}$
    end
end
```

following operations of the lot) to the next period and the corresponding workload is removed from all the machines to which the lot was assigned.

The Smoothing module ends with a production plan for the current period that is feasible in terms of capacity. This plan is, however, potentially unfeasible from the point of view of respecting the due dates. It also provides the list of postponed lots and from which operation each lot is postponed. If lots are shifted, it is necessary to project again their operations from the next period. We therefore enter into the projection module again, which takes as inputs the initial plan and where new start and end dates will only be re-computed for lots that have been shifted (assignment of new values to $S_{o,l}$, $E_{o,l}$ and $X_{o,l,t}$).

Once the projection is completed, the new planning is sent to the balancing module, and the three-step approach is repeated until all the periods have been processed. The number of iterations is thus equal to the number of periods in the planning horizon (generally 8 weeks). Each iteration takes a finite amount of time as, in the worst case (which never happens in practice), all lots in a period are moved to the next period.

The output of the Smoothing module in the last period is a finite capacity production plan that aims at minimizing the lot delay, giving for each operation of each lot the period in which it is expected to be processed, an estimate of the workload balancing and its impact on the machine utilization rate.

4. Smoothing module

In the previous section, we presented the global framework of the three-step approach and detailed how the third module, i.e. the smoothing procedure, works. We mentioned that this third module ensures that the capacity of the machines is respected by postponing the lowest priority lots to following periods. We defined the lowest priority lot as the one with the latest due date, which resembles the Earliest Due Date (EDD) scheduling rule. We now propose other rules that are summarized in Table 3, and evaluate their impact on the quality of the resulting solutions. Thus, in this section, we study different ways of evaluating the priority of a lot, which thus influences the choice of lots to be postponed (shifting the processing of an operation from one period to the next) when machines are overloaded. These shifted lots are those considered as having the lowest priority according to the selected smoothing rule.

In addition to the notations already used previously, we introduce in Table 2 some additional notations in order to define the implemented smoothing rules.

Table 2.: Additional notations used in the definition of smoothing rules

Notation	Description
$p_{o,l}$	Processing time of operation o of lot l
$w_{o,l}$	Waiting time of lot l before processing operation o

4.1. *Due date oriented rules*

Among the proposed smoothing rules, the simplest one is Earliest Due Date (EDD), based only on the due dates of lots. Thus, once the most overloaded machine is determined, the lot with the largest due date (among the lots processed by this machine) is first postponed.

The problem with the EDD rule is that it does not take into account the position of the lot in its route (i.e. the fixed sequence of operations to perform on the lot before it is completed). Indeed, two lots l_1 and l_2 with the same due date will have the same priority. However, if l_1 still requires 100 operations before being completed while l_2 is only one operation from being completed, it will be more difficult to meet the due dates of l_1 than the due dates of l_2. Lot l_1 should therefore have a higher priority than lot l_2, which is not possible with the EDD rule. We therefore introduced a second rule, which is a variant of the *Operation Due Date* rule (ODD, Baker (1984)), where the *slack time* is the difference between the time remaining until the delivery date and the cumulative time of the remaining operations. Given $S_{o,l}$ the start date of operation o of lot l, d_l its delivery date and $p_{o,l}$ the time required to perform the operation o of lot l. Then, the slack time of lot l at operation o ($ST_{o,l}$) is defined by:

$$ST_{o,l} = d_l - S_{o,l} - \sum_{o' \geq o} p_{o',l} \tag{14}$$

However, in our case, this rule is modified to include the waiting time of the lot at each operation $w_{o,l}$, in addition to its processing time $p_{o,l}$. This is because, in *front end* semiconductor manufacturing facilities, waiting times in front of machines correspond to the majority of the cycle time of a lot, much more than the sum of the processing times. Thus, a good evaluation of the available margin must take into account not only the remaining processing times to complete the operations of the lot, but also the estimated waiting time before each operation. The sum of the waiting time and the process time of an operation is commonly called the cycle time of the operation. The sum of the cycle times of all operations of a lot (from its start date into the factory until its completion) is called the lot cycle time. Finally, the cumulative cycle time of the remaining operations for a given lot is called its remaining cycle time. All this data is based on statistical calculations based on historical data. This work was the topic of a PhD thesis in collaboration with STMicroelectronics (Dequeant (2017)), whose main subject is the variability in semiconductor manufacturing. Chapter 4 of Dequeant (2017) is dedicated to the definition of the cycle time model used in our work. In our definition of *slack time*, the term $\sum_{o' \geq o} p_{o',l}$ is therefore replaced by the theoretical cycle time remaining for the lot l from operation o: $\sum_{o' \geq o}(p_{o',l} + w_{o',l})$. The new formula for the *slack time* is then:

$$ST_{o,l} = d_l - S_{o,l} - \sum_{o' \geq o} (p_{o',l} + w_{o',l}) \tag{15}$$

Thanks to the *ST* smoothing rule, among the lots that can potentially be shifted, the choice will be made among those with the largest margin for on-time delivery.

In addition to the *ST* rule presented above, a variant is proposed, named “Slack Time If Postponed“ (ST_{post}). To motivate this variant, let us recall that a discretized time horizon is considered, working iteratively period by period. Periods are generally of the order of one week. As the period length is rather long, the impact of shifting an operation of a lot from one period to the next will be different if the operation was originally scheduled at the beginning or at the end of the period. If the operation was scheduled at the beginning of the period, postponing the operation to the next period means postponing the lot by several days (potentially up to 7). Let us take for instance lot l_1 having more margin than lot l_2 (for example a 10-day margin for l_1 against a 6-day margin for l_2). Assume that l_1 and l_2 are candidates to be shifted because they both have an operation on the most loaded machine in the period considered. Lot l_1 has in theory a lower priority than l_2. However, as l_1 is processed at the beginning of the period, postponing l_1 would delay it by 7 days, while l_2 is processed at the end of the period and would therefore only be postponed by 1 day. In this case, l_1 would have a new Slack Time of $10 - 7 = 3$ days, while $l2$ would have an Slack Time of 5 days. This means that shifting l_2 is actually preferable because it would leave a larger margin than if l_1 was shifted. Thus, the $ST_{post,o,l}$ variant calculates the *Slack Time* that the lot l would have if it was moved to the next period. If we always define $S_{o,l}$ as the instant when the lot l waits in front of the machine for its operation o, and t_{next} as the start date of the next period, we obtain the expression below for the calculation of this new indicator:

$$ST_{post,o,l} = ST_{o,l} - (t_{next} - S_{o,l}) \tag{16}$$

Combining with expression (15), we obtain:

$$ST_{post,o,l} = d_l - t_{next} - \sum_{o' \geq o} (p_{o',l} + w_{o',l}) \tag{17}$$

Thus, the indicator $ST_{post,o,l}$ is none other than the indicator $ST_{o,l}$ when considering the operation o of lot l performed at the beginning of the following period t_{next}.

Rules 2 and 3 take into account the margin of lots to be sent on time to customers. However, these rules do not differentiate between two lots with the same margin but at different positions on their route. If a lot is considered late, it will be easier to catch up its delay if the lot is at the beginning of its route (for example if it theoretically has 2 months left in the factory) than if it is supposed to be delivered within the week. This is why we implemented a fourth smoothing rule, which is a variation of the well-known rule *Critical Ratio* (CR, Baker (1984)). Keeping the notations used to define the indicator ST, let us define the *critical ratio* $CR_{o,l}$ of a lot l at an operation o:

$$CR_{o,l} = \frac{d_l - S_{o,l}}{\sum_{o' \geq o}(p_{o',l} + w_{o',l})} \tag{18}$$

This indicator is therefore the ratio between the time remaining until the delivery date of the lot and the target remaining time. The CR rule will therefore tend to give more weight to lots that are close to the end of their route.

This indicator is the ratio between the time remaining until the due date, and the theoretical remaining time. This rule therefore prioritizes lots near the end of their route.

In the same way as for ST rule, a variant of the CR indicator is defined, named "Critical Ratio if Postponed" ($CR_{post,o,l}$), that considers the value of the indicator of lot l if it would effectively be postponed from operation o. This variant is the fifth smoothing rule.

Finally, we have included the rule presented in Mhiri et al. (2018). In Mhiri et al. (2018), this priority rule, named RankingCoeff, illustrates "the priority of a lot in terms of its position in the process sequence on the considered toolset and the urgency of delivery". In the remainder of this paper, we refer to this rule as the "Critical Ratio and Position on Machine" (CRPM) rule. Adapted to our formulation, with $S_{o,l}$ the start date of operation o of lot l and t_{next} the start date of the next period, the $CRPM$ indicator is defined below:

$$CRPM_{o,l} = CR_{o,l} + \frac{S_{o,l}}{t_{next}} \tag{19}$$

The first term is the Critical Ratio and aims at evaluating the lot delivery urgency, while the second term is a normalized value of the position of the lot in the machine queue. This rule tends to prioritize late lots which furthermore are at the beginning of the machine queue. This rule aims above all to meet due dates, and is therefore customer-oriented. When analyzing how the three-step approach proceeds and because the second term in (19) prioritizes lots that are at the beginning of a machine queue (in the considered period), lots processed early in the period may be prioritized. This should be close to the mechanism used in the *post* variants presented earlier. Note also that the two terms $CR_{o,l}$ and $\frac{t_{o,l}}{t_{next}}$ are ratios, with values generally around 1, whereas Mhiri et al. (2018) do not detail how these two terms are balanced.

4.2. *Machine oriented rules*

All the previously introduced rules are based on lot time considerations (due dates, remaining cycle times) and mainly target the minimization of the overall delay (TT). However, there are other indicators in semiconductor manufacturing, such as the average cycle time, the average machine utilization rate, or the overall fab throughput (total number of operations performed in the fab on a time horizon). These indicators are monitored by managers with equal (or sometimes greater) importance than customer-oriented indicators. However, the previous smoothing rules do not take these other objectives into account.

Thus, it is relevant to propose new rules that take into account other non-customer-oriented indicators (either exclusively or partially). This makes it possible to evaluate how much the rules in Section 4.1 could degrade indicators such as the throughput and machine utilization rate, but also to develop an approach capable of considering different optimization criteria.

Hence, we developed a sixth rule, called Machine Impact (MI), which takes the workload generated by the lots on the machines into account. As with the previous rules, the first step is to determine the most overloaded machine, and then to identify the set of lots with at least one operation on that machine in the period. Then, we search in this set for the lot to postpone that has the smallest utilization rate on the machines that are not overloaded. Our goal is to reduce the workload on the overloaded machines, while minimizing the workload lost on the other machines, i.e. the machines whose workload does not have to be reduced. We are therefore trying to postpone the lot generating the least workload on non-overloaded machines. For this purpose, for each (lot, operation) pair, the cumulative workload generated by this lot for this operation and the subsequent operations in the period (only on non-overloaded machines) is calculated, in order to evaluate the utilization rate reduction resulting from postponing this lot. This is because a lot generally goes through several processing operations during the same period. Therefore, shifting a lot from an operation implies that any following operation of this lot planned in the period must also be shifted and therefore will lead to a smaller utilization rate or throughput loss.

The Machine Impact rule, unlike the previous rules, ignores the customer dimension. This rule should therefore normally favour objectives such as maximizing machine utilization or global throughput maximization rather than delay oriented objectives. In order to reconcile the two types of indicators, we combined the smoothing rules into two last rules called "Machine Impact and ST" (MI_{ST}) and "Machine Impact and CR" (MI_{CR}). These two rules are based on the following idea: The most overloaded machine is always identified first, then all lots with at least one operation processed by this machine during the period are identified. However, only lots in advance are considered. Lots in advance are those with a positive Slack Time, i.e. their theoretical remaining cycle time is lower than the remaining time available before being delivered to the customer. Once the lots in advance have been identified, the Machine Impact smoothing rule is applied by postponing the lot with the least impact on non-overloaded machines. If all the lots processed by the most overloaded machine are already late, the Machine Impact rule is not applied. Instead, the customer oriented smoothing rules are used, respectively ST_{Post} for the MI_{ST} rule, and CR_{Post} for the MI_{CR} rule.

Table 3 provides a summary of the different rules described above. Note that, in the column "Description", the lots are always selected from the subset of lots having at least one operation processed in the period by the most overloaded machine.

Table 3.: Rules in Smoothing module

Notation	**Shifting Rule Name**	**Description**
EDD	Earliest Due Date	Postpone the lot with the latest Due Date
ST	Slack Time	Postpone the lot with the largest margin (difference between remaining available time and remaining theoretical cycle time)
ST_{Post}	Slack Time if postponed	Consider the margin of the lot if it was postponed to the beginning of the next period from the considered operation
CR	Critical Ratio	Postpone the lot whose ratio between its remaining available time and its cycle time is the largest
CR_{Post}	Critical Ratio if postponed	Critical Ratio of the lot if it was postponed to the beginning of the next period from the operation
MI	Machine Impact	Postpone the lot whose induced workload in the period (from the operation), on machines that are not overloaded, is the lowest.
MI_{ST}	Machine Impact with Slack Time if postponed	Postpone the lot (among those in advance) whose induced workload in the period (from the operation) on machines not overloaded is the lowest. If no lots are in advance, apply the ST_{Post} rule
MI_{CR}	Machine Impact with Critical Ratio if postponed	Postpone the lot (among those in advance) whose induced workload in the period (from the operation) on machines not overloaded is the lowest. If no lots are in advance, apply the CR_{Post} rule
CRPM	Critical Ratio and Position on Machine	For all considered operations, shift the one whose Critical Ratio value plus the normalized position of the operation in the machine queue is the largest

5. Computational experiments

5.1. *Comparing the three-step approach and the MILP*

In this section, numerical results comparing the performance of the three-step approach with the MILP solved using IBM ILOG CPLEX are analysed. This study is rather short for two reasons. The first one is that a fairly comprehensive study has already been proposed in Mhiri (2016)The second reason is that, unlike Mhiri (2016) who extensively worked on the exact resolution of the problem, notably through the use of a Lagrangian relaxation approach, our primary goal was to work on an approach that provides good solutions and is flexible and fast enough to be integrated into the company's planning process. Section 5.1 shows that, although the use of advanced methods can improve the sizes of the problem instances that can be solved, they are still far too small compared to those in our industrial application. Then, Section 5.4 presents a comparative study of the different smoothing rules based on industrial instances that are described in Section 5.3.

The mathematical model in Section 2 has been solved with IBM ILOG CPLEX 12.7.1 on very small instances generated from simplified industrial instances summarized in Table 4.

Table 4.: Characteristics of simplified test instances

Parameters	**Values**
Number of lots	1 to 7
Number of machines	357
Number of products	3
Average number of operations per lot	350
Lot delivery periods distribution	Remaining time before due date is 1 (Hard) or 2 (Easy) times the lot average cycle time
Number and period length	8 periods of one week each

The instances were taken from industrial data, but some parameters were greatly reduced. The number of machines and the average number of operations are very close to those found in the factory. The number of lots is the main parameter that was reduced. This number is varied from 1 to 7, to be compared with the thousands of lots usually in the factory. The duration and number of periods (8 weeks) correspond to the industrial setting. The delivery periods have been readjusted according to two configurations. The simplest instances ("Easy") consider lots with delivery periods such that the remaining available time is twice the theoretical remaining cycle time of the lots. More difficult instances ("Hard") were also proposed with a remaining time before delivery which is equal to the average lot cycle time.

For the sake of brevity, we are not interested in analyzing the influence of each parameter on the performance of the modelThus, we chose to keep most of the parameters as in the industrial instances, focusing only on the number of lots as well as the tightness of the due dates.

The problem instances were solved using an Intel Core i5 PC with a 2.3 GHz processor and 8 GB of RAM. Some results are summarized in Table 5, providing the average CPU time required to solve each instance.

The model was able to solve instances with up to 4 or 5 lots in less than one hour, depending on the difficulty of the instances. In an attempt to improve the performance of the initial model, two techniques have been developed. First, instances often have a large number of operations with very short processing times. Thus it is possible to aggregate these operations as a fixed total time, remaining small compared to the overall planning horizon, and therefore having little impact on the final solution. The aggregation then allows the number of variables associated with the operations to be reduced and thus the calculation time to be reduced. A second technique is to first solve the problem by relaxing the capacity constraints and then using the resulting solution as a lower bound for the general problem. These improvements led to an overall reduction in the computational time, making the improved MILP able to solve instances with up to 6 or 7 lots. However, in spite of these improvements, several hours are still necessary to solve problems with only 10 lots. Therefore, it is impossible to use this model for problems with hundreds of lots.

Table 5.: Average CPU time (in seconds) of initial and improved MILPs for small size instances

	Initial MILP (sec.)		Improved MILP (sec.)		Time Reduction	
Instance Size	Easy	Hard	Easy	Hard	Easy	Hard
1 lot / 350 oper.	13.4	12	11.8	11.4	-12%	-5%
2 lots / 700 oper.	25.6	29	22.2	19	-13%	-34%
3 lots / 1050 oper.	212.4	522.8	73.2	55.8	-66%	-89%
4 lots / 1400 oper.	464	1572.2	147	424	-68%	-73%
5 lots / 1750 oper.	4100	9645	413	1547.0	-90%	-84%
6 lots / 2100 oper.	-	-	2309.7	3404	-	-
7 lots / 2450 oper.	-	-	7234	14228	-	-

In constrast to the MILP, the three-step approach has the advantage of being very fast, requiring only 7 seconds on average to solve small instances. For real-life instances, involving thousands of lots, each having hundreds of operations over dozens of weeks, the approach still delivers plans in less than 5 minutes. However, it is important to evaluate the quality of the resulting solutions. Thus, a comparative study was conducted between optimal solutions provided by the exact model (or the best upper bound if the MILP was not solved in less than one hour), and solutions provided by our approach. We compared the two approaches on 20 industrial instances, with the number of lots ranging from 1 to 20. The characteristics of the instances are the same as those in Table 4. The due date distribution is based on the configuration “Hard“. Results are summarized in Table 6. The rows *MILP* and *Approach* show the total delay in number of periods (all lots combined), and the row $Gap(\%)$ $(= \frac{APPROACH-MILP}{MILP})$ gives the gap between the solution of the three-step approach and the best feasible solution of the MILP in one hour. Note that, for instances for which the exact model determines an optimal solution, our approach achieves the same objective function for almost all cases. However, note that our approach ends with a lower quality solution in some instances with 5 and 7 lots. This is due to the greedy behavior of the smoothing algorithm used in the smoothing module and is a classical example of the drawback of greedy algorithms. Nevertheless, note that our approach finds optimal solutions in most cases, or solutions that are always better than the upper bounds determined by the MILP. It also can be noted that, for the “larger“ instances, the gap is widening between the solutions provided by the approach and those provided by the MILP (-57.2% gap for instances with 19 lots). This reflects the difficulty of the MILP in providing feasible solutions in a reasonable time frame (less than one hour), which increases rapidly with the number of lots to be managed.

5.2. *Comparison of smoothing procedures*

As we have already said, this paper is devoted to the study of smoothing rules. The experimental studies in this section aim at comparing the performances of the approach according to the different smoothing rules used and the different performance indicators considered.

Table 6.: Comparison between solutions provided by three-step approach and MILP in one hour (* when exact solution is found)

Nb lots	1	2	3	4	5	6	7	8	9	10
MILP	0*	1*	5*	7*	7*	14*	18*	26	45	45
Approach	0	1	5	7	8	14	19	22	28	30
Gap (%)	0%	0%	0%	0%	14.3%	0%	5.6%	-15.4%	-37.8%	-33.3%
Nb lots	11	12	13	14	15	16	17	18	19	20
MILP	47	55	65	80	64	100	117	120	152	160
Approach	34	37	41	42	47	50	53	61	65	71
Gap (%)	-27.7%	-32.7%	-36.9%	-47.5%	-27%	-50%	-54.7%	-49.2%	-57.2%	-55.6%

In Mhiri et al. (2018), a study is conducted on the performance of exact methods for solving the operational production planning problem. The authors solved a MILP via CPLEX and used a Lagrangian relaxation method to improve the size of the solvable problems. However, this size remains to some dozens of lots and hundreds of operations, underlining the intractability of real instances with thousands of lots and almost one thousand operations per lot.

Besides, the authors compared the plans determined by the approach and the actual production in the fab. The study aims at evaluating the ability of the approach to reliably simulate the evolution of the WIP in the factory. However, this is not the goal of our approach, which is used daily, to define production objectives. The plans are therefore not only predicting what will happen, but prescribe the path to follow to maximize the objectives. Therefore, in this work, we focus on evaluating the impact of the different smoothing rules on different performance measures.

5.3. *Industrial instances*

In this section, we compare several versions of the approach to study the impact of the smoothing rules based on 25 industrial instances directly taken from two STMicrolectronics plants. The characteristics of these instances are summarized in Table 7 whose columns are described below.

Column named WIP shows the number of lots in the plant at the beginning of the planning horizon with an average of 4244 lots. Then, another aspect that can impact the production plan is the production line imbalance. Product routes are usually broken down into 10 milestones with an equivalent number of process operations. Measuring the dispersion of the WIP on these different milestones is a good indicator of the imbalance of the production line. The greater the dispersion (measured by the standard deviation $Std(WIP)$), the more likely that some machines will be overloaded while others will be underutilized. Since volumes can vary between instances significantly, the relative standard deviation of the WIP (noted $RStd(WIP)$) is given in Table 7. The dispersion of the WIP is rather high, ranging from 18% to more than 50%, showing very unbalanced production flows.

Regarding the satisfaction of the demand, the difficulty of the instances depends on the status of the lots (being in advance or delayed) at the beginning of the planning horizon (the lots in the initial WIP). The status of a lot at a given operation o in the WIP can be measured using the slack time ($ST_{o,l}$) defined in Section 4.1. Based on this information, we use two indicators to evaluate the difficulty of an instance. Let us denote by n the number of lots currently in the WIP. The first indicator, denoted $Avg(ST) = \frac{\sum_l ST_{o,l}}{n}$, is the average initial delay of each lot. This indicator varies from 0.46 (a rather low value considering average cycle times of several months) to more difficult instances with 2.98 days of delay on average for each lot. The second indicator $\sum U_i(\%) = \frac{\sum_l [(ST_{o,l} \leq -1)?(1:0)]}{n}$ corresponds to the ratio of lots with a significant delay (more than 1 day) at the beginning of the planning horizon.

The last three columns of the table show the size of the instances, with respectively the number of machines (*Tools*), the number of different products in the WIP (*Prd*), and the average number of process operations remaining for each lot (Avg(Op)).

Table 7.: Characteristics of industrial instances

Instance	WIP (lots)	Std(WIP)	RStd(WIP)	Avg(ST)	$\sum U_i(\%)$	Tools	Prd	Avg (Op)
1	3123	71	23%	0,71	23%	322	320	149
2	3168	66	21%	1,50	29%	323	328	145
3	2877	92	32%	2,69	26%	327	268	134
4	2845	83	29%	1,65	31%	327	259	132
5	2757	77	28%	0,69	12%	326	243	138
6	2822	51	18%	0,86	16%	327	250	133
7	2785	60	21%	0,66	13%	326	244	135
8	2831	56	20%	0,46	11%	326	238	135
9	2883	60	21%	0,47	12%	326	239	134
10	4011	136	34%	1,48	36%	395	433	597
11	4002	132	33%	1,50	27%	396	423	594
12	4182	139	33%	1,52	28%	397	436	594
13	4261	156	37%	1,39	28%	397	428	589
14	4199	153	36%	0,81	16%	397	439	599
15	4190	154	37%	1,15	23%	400	413	593
16	4265	160	37%	1,71	26%	403	439	607
17	5725	300	52%	2,41	37%	417	575	579
18	5675	299	53%	2,84	37%	415	579	584
19	5542	259	47%	2,02	30%	426	574	556
20	5608	254	45%	2,35	26%	426	577	554
21	5531	251	45%	2,06	24%	429	574	559
22	5652	256	45%	2,24	28%	429	589	559
23	5652	256	45%	2,20	27%	429	589	559
24	5735	276	48%	2,98	30%	430	584	537
25	5770	288	50%	2,04	24%	439	629	519
Avg	4244	163	36%	1,62	25%	382	427	417
Max	5770	300	53%	2,98	37%	439	629	607
Min	2757	51	18%	0,46	11%	322	238	132

For each instance, we run the three-step approach for 12 periods (weeks) with the 8 smoothing rules summarized in Table 3, and compared them on six performance measures detailed in Table 8.

First, it should be noted that the first three indicators ($\sum T_l$, $\sum U_l$ and T_{max}) are based on due dates and assess whether the approaches provide quality solutions for customers. The other indicators focus on evaluating how smoothing rules affect lot cycle time (CT), machine utilization rate (Utlz%) and overall throughput (TP%).

These performance measure are standard indicators in semiconductor manufacturing. Uzsoy, Lee, and Martin-Vega (1992) classify the indicators in this context into two classes: Those based on due dates and those based on flow time. The indicators based on due dates (such as $\sum T_l$, $\sum Ul$ and T_{max}) aim at measuring customer satisfaction, which is essential given the strong competition in semiconductor manufacturing. Among the indicators based on flow time, the first is cycle time (CT), whose minimization ensures a better responsiveness of facto-

Table 8.: Performance measures

Performance Measures	**Description**
$\sum T_l$	Total Tardiness, i.e cumulative sum of the tardiness (in number of weeks) for all lots.
$\sum U_l$	Total number of lots delivered late.
T_{max}	Maximum tardiness among all lots.
CT	Average Cycle Time of lots.
Utlz%	Average utilization rate (in percentage of their maximum capacity) of machines on the planning horizon, given as the gap with the solution obtained with the EDD rule.
TP%	Throughput gap compared to the EDD rule. A positive value means that the rule provides solutions with higher throughput, i.e. performs more operations in the same period of time.

ries to market demands. The throughput of the factory (TP%) as well as the utilization rate of machines (Utlz%) are also important indicators, quantifying the profitability of the production system.

Although the indicators $\sum T_l$, $\sum U_l$ and T_{max} are due date oriented, the results can probably differ between them. In particular, if a rule tends to share the delay between lots fairly, it will tend to minimize the maximum delay, i.e. T_{max}. This will generally be to the detriment of the overall number of late lots, i.e. $\sum U_l$. The empirical work in Ovacik and Uzsoy (2012) suggests that methods that perform well for T_{max} also perform well for objectives such as $\sum T_l$, $\sum C_l$ (sum of completion times, which is strongly correlated to the *Cycle-Time* indicator (Mönch, Fowler, and Mason (2012))) and C_{max} (makespan). However, this implication does not seem to be valid for $\sum U_l$. As for the $\sum T_l$ indicator, it is not necessarily strongly correlated with $\sum U_l$ or T_{max}. However, it can be expected that a solution giving both poor results according to $\sum U_l$ and T_{max} will also give poor results for $\sum T_l$.

A positive correlation between TP% and Utlz% can also be expected since an increase in throughput TP% implies more operations to perform in a given period, and therefore a larger machine utilization rate. However, it should be noted that this correlation is not guaranteed. Indeed, since processing times vary from one operation to another and from one machine to another, it may be tempting, for example, to process long operations on machines in order to increase their utilization, to the detriment of other faster operations. This choice shows that it is possible to increase the utilization rate of machines without increasing the overall throughput. On the opposite, it is possible to favour fast operations in order to increase throughput without increasing the utilization rate on the machines.

Concerning the cycle time, studies (see for instance Kacar, Mönch, and Uzsoy (2013); Kacar, Mönch, and Uzsoy (2016)) show that it generally increases with the overall load of the fab. It can therefore be expected that if a rule significantly increases the machine utilization rate and/or fab throughput, it will be at the expense of the overall cycle time. In addition, a positive correlation between CT and $\sum T_l$ can also be expected, since a reduction in the average cycle time should allow lots to leave earlier and thus lots should more easily meet their due dates.

Note also that this correlation is strongly influenced by the way the due dates are set. If the due dates are set based on cycle time estimates, a product for which the estimated cycle time is larger than its target cycle time would have the due dates of its lots placed later than what would be originally planned. This could potentially reduce the impact on the measured tardiness. On the contrary, when due dates based on target cycle times are set by the company, which is our case, an increase of the cycle time of a product would have a direct impact on the tardiness measured for this product.

5.4. *Comparison of smoothing procedures*

5.4.1. Average performances

In this section, we compare the multiple smoothing rules according to the different indicators. The results are summarized in Table 9 and Figures 2 and 3. Table 9 shows, for each performance measure, the average value for each rule on the 25 instances. The results in bold in the table correspond to the best result among the smoothing rules. For $\sum T_l$, $\sum U_l$, T_{max} and CT, the best result is the lowest average, while for TP% and Utlz%, the best result is the one with the highest average. Additional information can be found in Tables 1 and 2 in the appendix. They give, for each smoothing rule and each performance measure, respectively the worst and the best results among the 25 instances. Detailed results are also available at the end of the appendix.

To complete the analysis, Figures 2 and 3 show box plots. Figure 2 presents the results for $\sum T_l$, $\sum U_l$, T_{max} and CT, which should be minimized, and Figure 3 presents the results for TP% and Utlz%, which should be maximized.

In order to position the different performance measures on the same figure, the values are normalized. Thus, if $v_{r,p,i}$ denotes the value of performance measure p obtained using smoothing rule r on instance i, the normalized value $v^n_{r,p,i}$ is:

$$v^n_{r,p,i} = \frac{v_{r,p,i} - \min_{\forall(r,p)} v_{r,p,i}}{\max_{\forall(r,p)} - \min_{\forall(r,p)}} \tag{20}$$

Table 9.: Comparison of shifting rules - average

Performance	Rules								
Measures	EDD	ST	ST_{Post}	CR	CR_{Post}	MI	MI_{ST}	MI_{CR}	CRPM
$\sum T_l$	828.8	1475.7	1215.5	1496.5	**624.1**	1205.6	1125.6	946.9	1366.8
$\sum U_l$	363.8	333.4	**301.3**	496.3	405.7	442.6	324.2	312.6	479
T_{max}	9.8	11.4	10.6	10.2	**6.0**	12	10.1	9.1	11.5
CT	43.4	45.9	45.00	45.5	**42.6**	44.5	44.9	43.8	44.5
TP%	0.00%	0.78%	0.82%	0.58%	0.62%	**1.01**%	0.91%	0.81%	0.23%
Utlz%	0.00%	0.81%	1.19%	0.57%	0.93%	**1.75**%	1.62%	1.42%	0.21%

First of all, note the very good results of the CR_{Post} rule which provides the best average results for performances measures $\sum T_l$, T_{max} and CT. These good performances are also visible in Figure 2 where we can see that the CR_{Post} rule has a lower dispersion and globally lower values (max, min, median, first and third quartiles) than the other smoothing rules. However, despite the quality of the CR_{Post} rule, the results are more mixed for the $\sum U_l$ indicator. Thus, the CR_{Post} rule seems to be successful in reducing the average and maximum delay, but at the expense of a larger number of late lots.

Concerning the $\sum U_l$ indicator, Table 9 and Figure 2 indicate that the best results are obtained with the ST and ST_{Post} rules. Thus, considering lots only from the point of view of their absolute delay (and not relative to their position), tends to minimize the number of late lots. But this is at the expense of the overall and maximum lot delays.

A possible explanation for these results is that considering the relative delay tends to give more weight to lots close to the end of their route (and therefore that are generally close

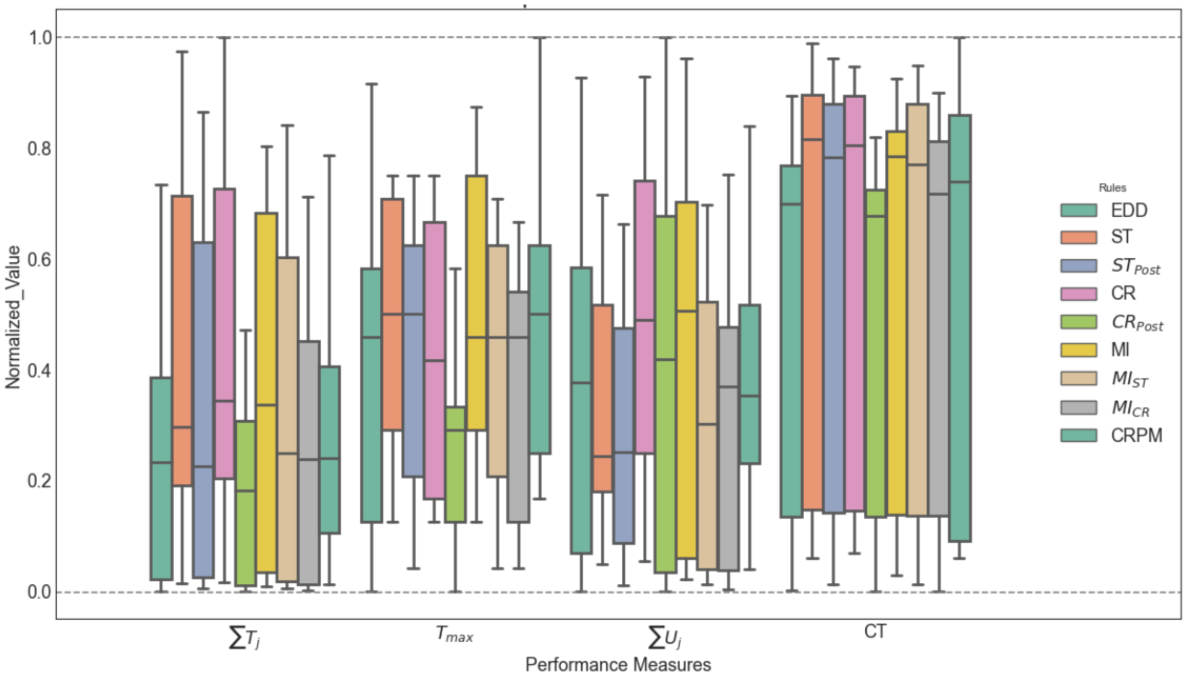

Figure 2.: Comparison of smoothing rules on performance measures to minimize.
Alt Text: Box plot showing the performance of each smoothing rule for the performance measures to be minimized, the total sum of tardiness, the maximum tardiness, the number of lots delivered late and the average cycle time of the lots. Rule CR_{Post} is the one showing globally the best results

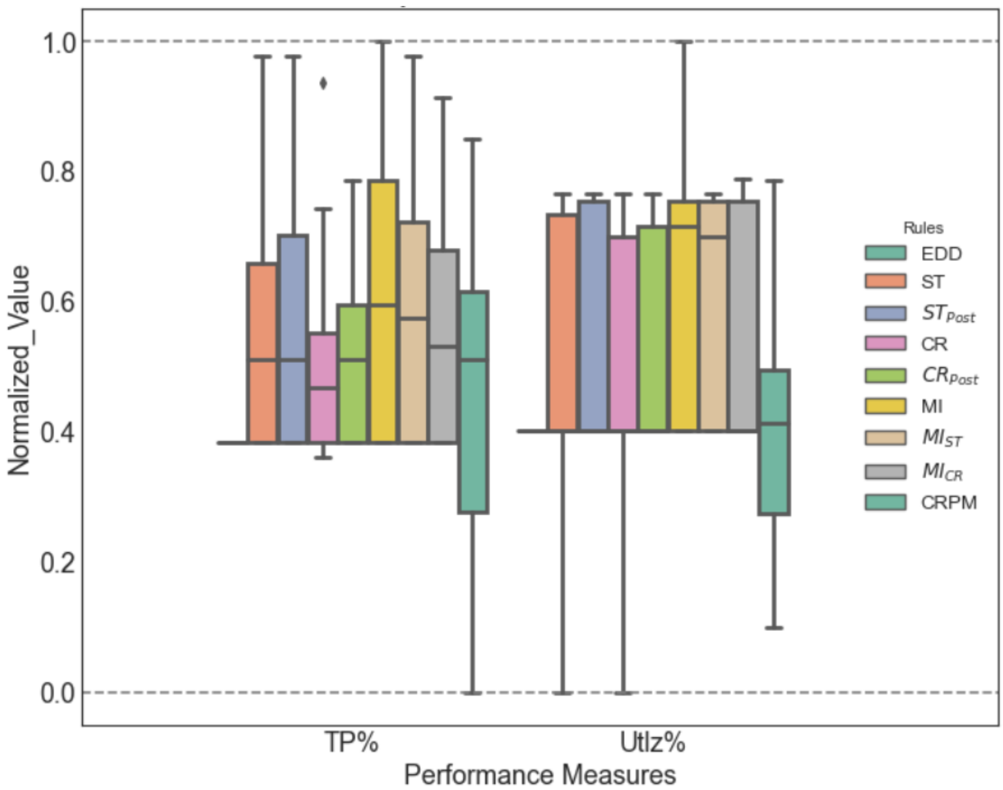

Figure 3.: Comparison of smoothing rules on performance measures to maximize.
Alt Text: Box plot showing the performance of each smoothing rule for the performance measures to be maximized, i.e. the average throughput and the average utilization rate. Rule *MI* is the one showing globally the best results

to their delivery date). Thus, the lots that are late and close to the end of their route will have a higher priority and will not be slowed down, limiting their delay. In contrast, the *ST* and ST_{post} rules will not necessarily give priority to lots at the end of their route, which will potentially have larger delays than with the *CR* or CR_{post} rules. However, since the lots at the beginning and the middle of their routes are less penalized, it is possible that the heuristic better manages potential delays upstream, thus avoiding additional delays. This should make the *ST* and ST_{post} rules more effective in reducing $\sum U_l$, but less effective than the *CR* or CR_{post} rules in reducing $\sum T_l$ and T_{max}.

Not surprisingly, rules considering only or in part the impact of lots on machines present average performances regarding cycle time and due date oriented performances indicators. We can nevertheless underline the correct performances of the MI_{CR} rule, which sometimes succeeds in achieving the best solutions, but above all makes it possible to generally obtain good results and rarely very poor results.

For non-customer oriented indicators, i.e. TP% and Utlz%, note the good overall performance of the rules considering the impact of lots on machines (MI, MI_{ST} and MI_{CR}). The best results are obtained by the totally machine oriented (MI) rule with an average throughput which is 1% larger than the EDD rule, and 1.75% larger for the average machine utilization rate. However, this domination is limited to the TP% and Utlz% indicators. Indeed, for the average cycle time of the lots, the CR_{Post} rule stands out this time by providing the best results on all instances.

Note again that the quality of the results obtained using the MI_{CR} rule, with a good compromise between the three indicators.

An important remark is that the CRPM rule from Mhiri et al. (2018) does not provide good solutions, being always dominated (average and maximum) by other rules with even the worst results on indicators T_{max}, TP% and Utlz%

When considering all indicators, the ST and CR rules are globally dominated by their MI_{ST} and MI_{CR} variants. This means that taking into account the influence of the shift of a lot to a new period really improves the quality of solutions. Therefore, the ST and CR rules will no longer be considered, as well as the CRPM rule. Besides, we do not consider the MI_{ST} rule whose performance does not significantly differ from those of the MI or CR rules in terms of quality, and is outperformed by MI_{CR} rule on most of the performance measures.

5.4.2. *Best and worst case performances*

Since no solution totally dominates the others over all six indicators, we now analyze the number of times one of the rules provides the best or worst solutions. The results of the analysis can be found in Figures 4 and 5, through two Kiviat diagrams, where each axis corresponds to one of the performance measures and each rule has a different color. In Figure 4, the number of times a rule finds the best solution for a performance measure is shown. Thus, a good rule is assumed to cover a large area. The maximum on an axis is 25, which means that the rule always provides the best solution for the associated performance measure. In contrast, in Figure 5, the number of times a rule finds the worst solution for a performance measure is shown. A good rule should therefore cover a small area. Note that the sum on each axis is not necessarily 25, since some of the best and worst solutions are obtained with the OD, RS, CRPM and MI_{ST} rules which are not deemed to be relevant in terms of solution quality and thus are not represented for clarity of representation.

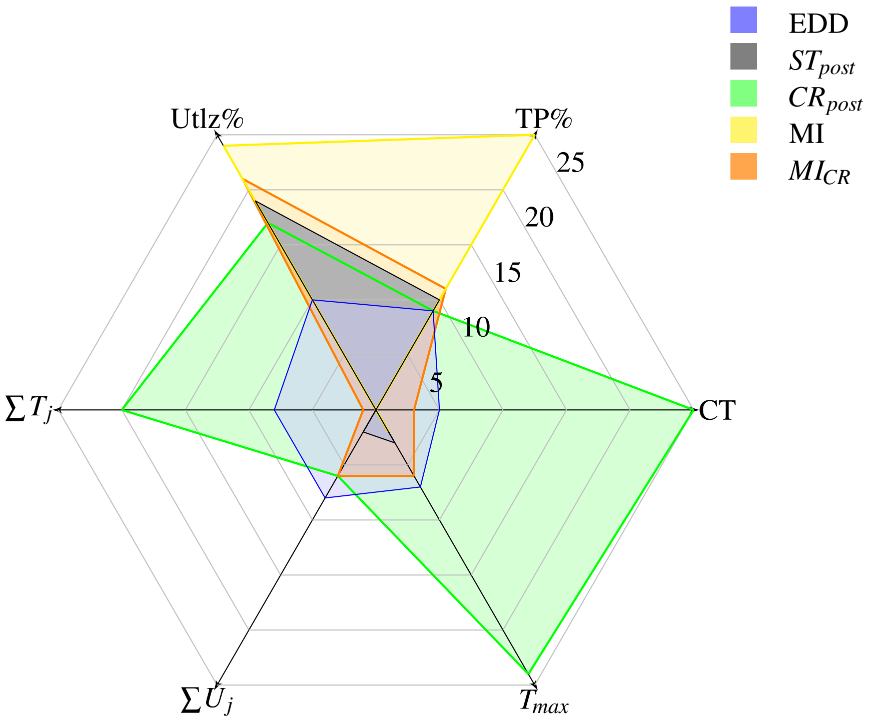

Figure 4.: Number of best solutions by performance measure and smoothing rule.
Alt Text: Radar diagram with 6 branches representing the different performance measures. The scale corresponds to the number of times a rule found the best solution. The CR_{post} and *MI* rules are the ones with the largest areas represented. CR_{post} is the first one.

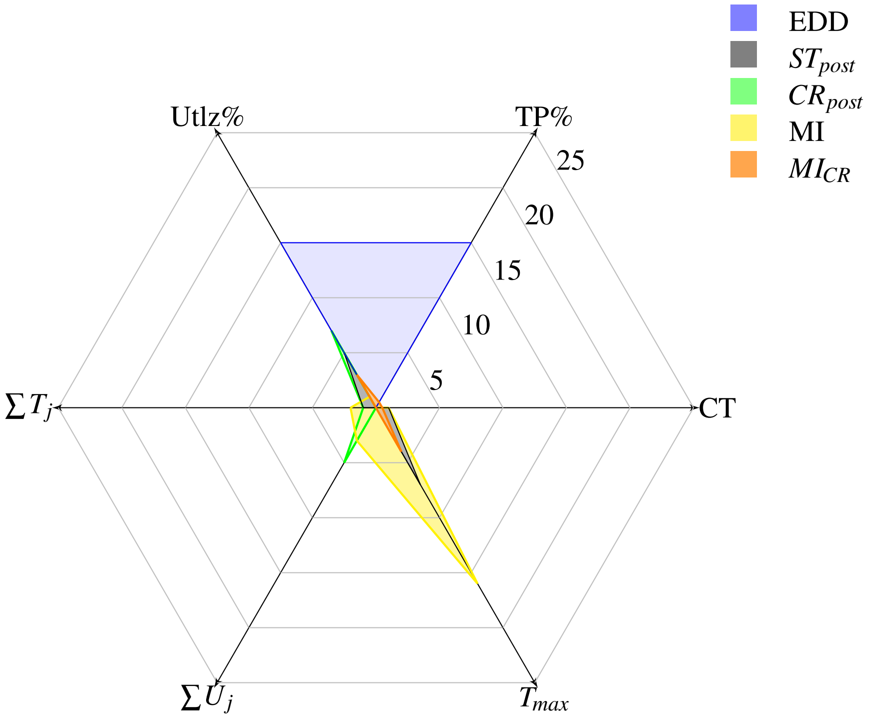

Figure 5.: Number of worst solutions by performance measure and smoothing rule.
Alt Text: Radar diagram with 6 branches representing the different performance measures. The scale corresponds to the number of times a rule found the worst solution. The *EDD* and *MI* rules are the ones with the largest areas represented. *EDD* is the first one.

Figure 4 emphasizes again the performance of the CR_{Post} rule, which often leads to the best solutions for some indicators. Moreover, Figure 5 shows that the CR_{Post} rule rarely gives worst solutions. However, note that, for the indicators for which the CR_{Post} rule gets the worst solutions, the CR_{Post} rule often also gives the best solutions, which may reflect some variability in the results.

Figure 4 shows that EDD obtains good results by sometimes finding the best solution for all indicators. In addition, Figure 5 shows that EDD provides 15 times the worst solution for the Utlz% and TP% indicators, and therefore significantly degrades the quality of solutions on these aspects.

According to Figure 4, the MI rule also obtains good results by almost always dominating the other rules for the Utlz% and TP% indicators. According to Figure 5, it also rarely gets the worst solutions, except for the T_{max} criterion where it achieves the worst results.

Finally, the analysis shows that the MI_{CR} rule performs well for the Utlz% and TP% indicators (even though it is not as effective as the MI rule which does not consider lot delays), but it rarely obtains the best solutions for the other indicators. However, Figure 5 shows that the MI_{CR} rule very rarely obtains the worst solutions (three times for T_{max} and Utlz%) and is therefore never dominated in terms of worst performance.

5.4.3. Computational times

In Table 10, Columns *Average*, *Minimum* and *Maximum* give respectively the average, minimum and maximum computational times in seconds of the three-step approach for the 25 instances and for each smoothing rule. Column *Average* also gives the percentage gap (*Gap(%)*) from the EDD rule, which is the reference rule.

Table 10.: Average computational times of three-step approach for each smoothing rule

	Average		Minimum	Maximum
Smoothing Rule	CPU (sec)	Gap(%)	CPU (sec)	CPU (sec)
DD	104	+0%	88	110
ST	90	-13%	77	109
ST_{Post}	95	-9%	81	105
CR	88	-15%	72	98
CR_{Post}	104	+0%	89	116
MI	119	+15%	98	138
MI_{ST}	101	-2%	82	125
MI_{CR}	106	+2%	85	125
CRPM	109	+5%	91	113

First, note that the ST and CR rules are on average faster than the EDD rule and their respective variants ST_{Post} and CR_{Post}. This last point can be explained in two ways. First, the a priori evaluation of indicators requires more calculations. But above all, because the ST_{Post} and CR_{Post} rules tend to prioritize shifting of lots at the end of the period, which generates less workload reduction with each shift and therefore requires more iterations to ensure that the capacity is respected. The standard rules are therefore faster, but we have seen in Sections 5.4.1 and 5.4.2 that this comes at the expense of the quality of the solutions. The CRPM rule slightly increases the computational time, although it remains close to the overall average. Then, note that the MI rule is the one that requires the longest computational time. This can be explained by the fact that the MI rule requires pre-processing in order to assess the impact that each lot would have on the machines if it was postponed. It should also be noted that the

MI_{ST} and MI_{CR} rules do not suffer a significant increase in computational time with respect to the EDD rule. The fact that the MI_{ST} and MI_{CR} rules have a much shorter computational time than the original MI rule can be explained by the fact that, in many situations, no lots available to be shifted are considered in advance. In this case, the time-consuming pre-processing phase to assess the impact of lots on machines is not performed, and the MI_{ST} and MI_{CR} rules only use the ST_{Post} and CR_{Post} rules which are faster.

6. Practical Implementations

Since Mhiri et al. (2018), work on the three-step approach and the decision support tool that contains it has continued and allows the tool to be today fully integrated into the planning process of the Crolles site of the STMicrolectronics company. In this section, we give more information on this integration as well as on the interest of the new smoothing rules implemented.

6.1. *Tool Integration in Manufacturing System*

The entire production planning process is summarized in Figure 6. Based on the classical classification of planning levels, we can see that the process presented is divided between the tactical and operational levels.

(1) The first input of this process is the delivery plan given by the central services of the company, which handle the Master Planning function. Here the indications are given for each production site, and the Crolles site therefore receives a delivery plan in the form of a delivery volume target per week and per type of product. These plans are given over 12 months and can be periodically adjusted if necessary.

(2) On the basis of these delivery plans, the internal planning department at the Crolles site aims to define the release plans to ensure the on time delivery. These release plans define the volumes per product and per week, and are re-evaluated each week according to the production evolution.

(3) Based on the delivery and release plans, the Operational Production Planning function seeks to define an optimized production plan to meet the delivery objectives. These production plans are drawn up over 8 weeks by the Industrial Engineering department and are presented to the main production managers in order to give them guidelines to follow. In addition, these weekly production plans also feed into other tools that automatically update the priority status of lots or define production targets by product and equipment, for use in Scheduling tools. These plans are then re-evaluated every week, taking into account the evolution of the release plans, but also the new situation of the factory (WIP evolution, new machine status, ...)

(4) Finally, the Scheduling, Dispatching and Human Production Management functions aim to optimize the local performance of the various workshops while ensuring on-time delivery of lots, with the help of instructions given by the higher functions.

As already mentioned, this approach is today fully implemented in a decision support tool, which has been used in a rolling horizon for over three years to plan the production of two STMicroelectronics factories in France. The tool provides production and capacity plans in a few minutes for large-size real problems and has led to several advances within the company. The ability to take into account capacity constraints has made it easier for planners to create feasible production plans. The speed to obtain the plans also allows users to run “what if“ scenarios by varying the inputs (e.g. modifications of the start plans or of machine capacities).

One of the indicators of the successful implementation of our approach is that between 68% and 73% of the quantities recommended to be produced each week by the tool are respected. This percentage is relatively high, given the high uncertainty in high-mix wafer manufacturing systems and the unavoidable operational adjustments.

The different modules of the approach also support other applications. For instance, the projection module is automatically run every morning to evaluate the earliness or lateness of lots (based on their Slack Time or Critical Ratio) and to update their priorities. The production plans are also used to support the management of masks in the photolithography workshop. Various use cases are discussed in Christ et al. (2018).

6.2. *Recommendations*

In this section, we discuss the results obtained when using the three-step approach in an industrial context.

Based on the results, the three preferred rules are CR_{Post}, MI and MI_{CR}. The CR_{Post} rule provides the best average results for the customer oriented indicators $\sum T_l$ and T_{max}, making it a preferred rule if managers are mainly concerned by customer commitments. Note that the ST_{Post} rule is better at limiting the number of late lots $\sum U_l$, however leading to a significant increase of cycle times compared to the CR_{Post} rule (see Table 9), which makes the ST_{Post} rule generally less preferable.

The primary objective is not always to minimize customer delays, but can be to maximize throughput or machine utilization.

This is in particular the case when demands are very high and the fab capacity is too low. In this context, delays can become inevitable and managers may choose to focus on maximizing the overall throughput of the plant, by maximizing the throughput and machine utilization. In this case, it is preferable to use the MI rule which, although it is not the best at minimizing customer delays (in particular the maximum delay, see Figure 5), dominates when maximizing the throughput of the facility and the average use of machines. However, this rule induces an increase in computational time compared to any other rule. However, this computational time remains short, about two minutes, which remains largely acceptable for the creation of weekly plans allowing several calculations to be quickly restarted to test several scenarios.

Finally, the use of the MI_{CR} rule can be recommended because of its good overall performance. Indeed, although this rule on average is never the best for any of the indicators, it generally remains the second or third best. Only the maximum delay seems to pose difficulties, but the MI_{CR} rule remains efficient for the global tardiness and number of late lots. Moreover, this is not at the expense of the computational time, which makes this rule a good option when balancing among all indicators.

Currently, the CR_{Post} rule is integrated in the approach as the default rule. But we are working to enable managers to easily select other rules (in particular the MI and MI_{CR} rules) according to the indicators they consider to be the most important when running the approach.

7. Conclusions and perspectives

In this paper, we presented an operational production planning problem in a complex manufacturing system. In this problem, lots are planned individually to provide a more detailed plan than approaches that only consider production quantities. To address this problem, we present a three-step approach following the one introduced in Mhiri et al. (2018). The approach is extended by introducing new optimization possibilities through new smoothing rules, whose

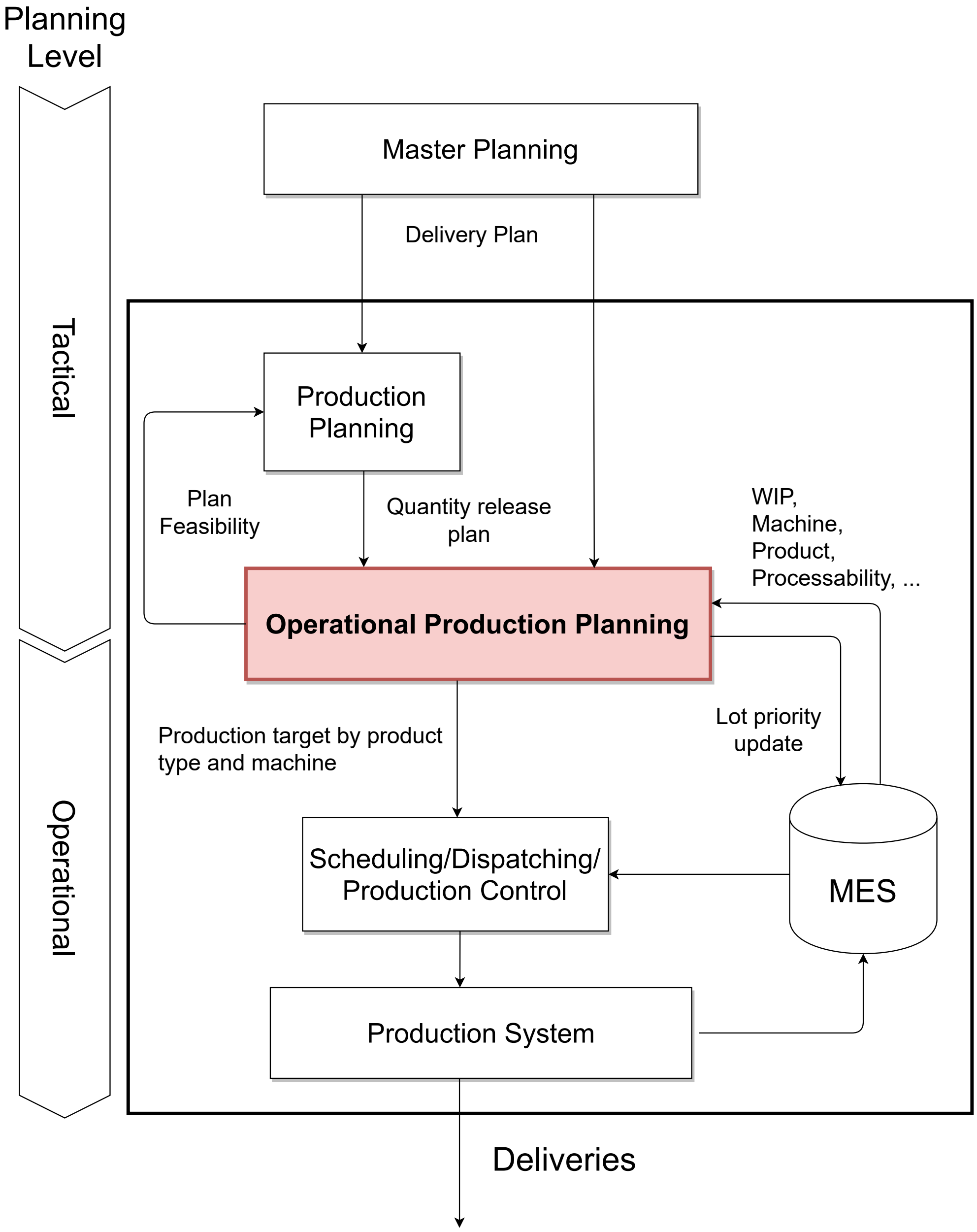

Figure 6.: Production planning system in site of Crolles of STMicroelectronics.
Alt Text: Flowchart representing the successive production planning functions from tactical to operational, namely: Master Planning, Production Planning, Operational Production Planning, Scheduling or Dispatching and finally Production System.

performances are studied according to different indicators.

The comparison of the different smoothing rules shows the influence of the choice of lots to postpone in the smoothing module, and that although no rule fully dominates the others, some rules stand out because of their overall performance or the fact that they rarely provide the worst solutions. Some recommendations are also provided to propose the best rules to use in an industrial context.

This approach is fully implemented in a decision support tool used to plan production in two STMicroelectronics factories. The tool also supports other uses such as the management of lot priorities based on the automatic evaluation of the slack time and critical ratio of lots.

Among future studies, extending the approach to consider multiple periods in the utilization balancing approach and the smoothing modules seems promising. Indeed, these modules take greedy decisions in a single period, but the utilization rate might be better balanced between periods. For example, the approach only allows the postponement of lots to a later period in case of overload of some machines, and thus to degrade the current solution. However, the approach does not allow to anticipate the realization of certain operations if the machine capacity allows it. A backward-smoothing approach, complementary to the smoothing approach already implemented, could be an interesting extension in order to improve the quality of the proposed plans. Other construction heuristics could be developed and compared to our approach. Also, the smoothing rules are compared based on the quality of the production plans determined by the three-step approach. Another interesting aspect would be to evaluate how the rules are followed in reality, and if indeed the rules leading to the best plans also lead to the best realized plans. This analysis is not easy to conduct in real-life conditions. Hence, it could be relevant to develop a simplified simulation model of the production system, which could be used to test the production guidelines resulting from the optimized plans. Another perspective is to extend our approach to handle lot due dates that are within periods instead of at ends of periods. This should be handled by modifying the third step of our approach where the capacity constraints are taken into account and lots are postponed. The third step of our approach could also be adapted to consider lots with different priorities such as hot lots.

Acknowledgements

This work has been partially financed by the ANRT (Association Nationale de la Recherche et de la Technologie) through the PhD number 2016/0421 with CIFRE funds and a cooperation contract between STMicroelectronics and Mines Saint-Etienne.

Data availability

The data used in this article have been anonymized and can be made available on request.

References

Anthony, Robert Newton. 1965. *Planning and Control Systems: A Framework for Analysis [by]*. Division of Research, Graduate School of Business Administration, Harvard University.

Aouam, Tarik, and Reha Uzsoy. 2015. "Zero-order production planning models with stochastic demand and workload-dependent lead times." *International Journal of Production Research* 53 (6): 1661–1679.

Baker, Kenneth R. 1984. "Sequencing rules and due-date assignments in a job shop." *Management science* 30 (9): 1093–1104.

Bard, Jonathan F, Yumin Deng, Rodolfo Chacon, and John Stuber. 2010. "Midterm planning to minimize deviations from daily target outputs in semiconductor manufacturing." *IEEE transactions on semiconductor manufacturing* 23 (3): 456–467.

Bitran, Gabriel R, Elizabeth A Haas, and Arnoldo C Hax. 1981. "Hierarchical production planning: A single stage system." *Operations Research* 29 (4): 717–743.

Brahimi, Nadjib, Stéphane Dauzère-Pérès, and Najib M Najid. 2006. "Capacitated multi-item lot-sizing problems with time windows." *Operations Research* 54 (5): 951–967.

Christ, Quentin, Stéphane Dauzère-Pérès, and Guillaume Lepelletier. 2019. "An Iterated Min-Max Procedure for Practical Workload Balancing on Non-Identical Parallel Machines in Manufacturing Systems." *European Journal of Operational Research* 279 (2): 419–428.

Christ, Quentin, Stéphane Dauzère-Pérès, Guillaume Lepelletier, and Philippe Vialletelle. 2018. "A multi-purpose operational capacity and production planning tool." In *2018 29th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 40–44. IEEE.

Dequeant, Kean. 2017. "Workflow variability modeling in microelectronic manufacturing." Theses, Université Grenoble Alpes. https://hal.archives-ouvertes.fr/tel-01652884.

Garey, Michael R, and David S Johnson. 1979. *Computers and intractability: A Guide to the Theory of NP-Completeness*. Vol. 174. WH Freeman New York.

Govind, Nirmal, Eric W Bullock, Linling He, Bala Iyer, Murali Krishna, and Charles S Lockwood. 2008. "Operations management in automated semiconductor manufacturing with integrated targeting, near real-time scheduling, and dispatching." *IEEE Transactions on Semiconductor Manufacturing* 21 (3): 363–370.

Gupta, Jatinder ND, R Ruiz, JW Fowler, and SJ Mason. 2006. "Operational planning and control of semiconductor wafer production." *Production Planning & Control* 17 (7): 639–647.

Habenicht, K, and Lars Mönch. 2002. "A finite-capacity beam-search-algorithm for production scheduling in semiconductor manufacturing." In *Proceedings of the Winter Simulation Conference*, Vol. 2, 1406–1413. IEEE.

Habla, Christoph, L Mönch, and R Drießel. 2007. "A finite capacity production planning approach for semiconductor manufacturing." In *2007 IEEE International Conference on Automation Science and Engineering*, 82–87. IEEE.

Horiguchi, Kazuo, N Raghavan, R Uzsoy, and S Venkateswaran. 2001. "Finite-capacity production planning algorithms for a semiconductor wafer fabrication facility." *International Journal of Production Research* 39 (5): 825–842.

Kacar, Necip Baris, Lars Mönch, and Reha Uzsoy. 2013. "Planning wafer starts using nonlinear clearing functions: A large-scale experiment." *IEEE Transactions on Semiconductor Manufacturing* 26 (4): 602–612.

Kacar, Necip Baris, Lars Mönch, and Reha Uzsoy. 2016. "Modeling cycle times in production planning models for wafer fabrication." *IEEE Transactions on Semiconductor Manufacturing* 29 (2): 153–167.

Karimi, Behrooz, SMT Fatemi Ghomi, and JM Wilson. 2003. "The capacitated lot sizing problem: a review of models and algorithms." *Omega* 31 (5): 365–378.

Leachman, Robert C, and Tali F Carmon. 1992. "On capacity modeling for production planning with alternative machine types." *IIE Transactions* 24 (4): 62–72.

Lu, Zhiqiang, Yuejun Zhang, and Xiaole Han. 2013. "Integrating run-based preventive maintenance into the capacitated lot sizing problem with reliability constraint." *International Journal of Production Research* 51 (5): 1379–1391.

Mason, Scott J, John W Fowler, and W Matthew Carlyle. 2002. "A modified shifting bottleneck heuristic for minimizing total weighted tardiness in complex job shops." *Journal of Scheduling* 5 (3): 247–262.

Mhiri, Emna. 2016. "Capacity planning in the context of high mix, application in the semiconductor industry." Theses, Université Grenoble Alpes. https://tel.archives-ouvertes.fr/tel-01485148.

Mhiri, Emna, Fabien Mangione, Mireille Jacomino, Philippe Vialletelle, and Guillaume Lepelletier. 2018. "Heuristic algorithm for a WIP projection problem at finite capacity in semiconductor manufacturing." *IEEE Transactions on Semiconductor Manufacturing* 31 (1): 62–75.

Mönch, Lars, John W Fowler, and Scott J Mason. 2012. *Production planning and control for semicon-*

ductor wafer fabrication facilities: modeling, analysis, and systems. Vol. 52. Springer Science & Business Media.

Mönch, Lars, Reha Uzsoy, and John W Fowler. 2018. "A survey of semiconductor supply chain models part III: master planning, production planning, and demand fulfilment." *International Journal of Production Research* 56 (13): 4565–4584.

Ovacik, Irfan M, and Reha Uzsoy. 2012. *Decomposition methods for complex factory scheduling problems*. Springer Science & Business Media.

Oyebolu, Folarin B, Jeroen van Lidth de Jeude, Cyrus Siganporia, Suzanne S Farid, Richard Allmendinger, and Juergen Branke. 2017. "A new lot sizing and scheduling heuristic for multi-site biopharmaceutical production." *Journal of Heuristics* 23 (4): 231–256.

Singer, Marcos, and Michael Pinedo. 1998. "A computational study of branch and bound techniques for minimizing the total weighted tardiness in job shops." *IIE transactions* 30 (2): 109–118.

Trigeiro, William W, L Joseph Thomas, and John O McClain. 1989. "Capacitated lot sizing with setup times." *Management Science* 35 (3): 353–366.

Uzsoy, Reha, Chung-Yee Lee, and Louis A Martin-Vega. 1992. "A review of production planning and scheduling models in the semiconductor industry part I: system characteristics, performance evaluation and production planning." *IIE Transactions* 24 (4): 47–60.

Yugma, Claude, Stéphane Dauzère-Pérès, Christian Artigues, Alexandre Derreumaux, and Olivier Sibille. 2012. "A batching and scheduling algorithm for the diffusion area in semiconductor manufacturing." *International Journal of Production Research* 50 (8): 2118–2132.

Zhang, Fan, Jie Song, Yingzhuo Dai, and Jie Xu. 2020. "Semiconductor wafer fabrication production planning using multi-fidelity simulation optimisation." *International Journal of Production Research* 58 (21): 6585–6600.

Zhang, Peng, Youlong Lv, and Jie Zhang. 2018. "An improved imperialist competitive algorithm based photolithography machines scheduling." *International Journal of Production Research* 56 (3): 1017–1029.

Appendix: Detailed numerical results

Table 1.: Comparison of shifting rules - Worst results

Performance	Rules								
Measures	EDD	ST	ST_{Post}	CR	CR_{Post}	MI	MI_{ST}	MI_{CR}	CRPM
$\sum T_l$	2528	3353	2981	3444	**1625**	2768	2897	2453	2711
$\sum U_l$	955	738	**682**	957	1030	991	718	775	865
T_{max}	22	18	18	18	**14**	21	17	16	24
CT	53.9	56.5	55.8	55.4	**51.9**	54.8	55.4	54.1	56.8
TP%	**0**%	**0**%	**0**%	-0.10%	**0**%	**0**%	**0**%	**0**%	-1.85%
Utlz%	**0**%	-3.45%	**0**%	-3.45%	**0**%	**0**%	**0**%	**0**%	-2.59%

Table 2.: Comparison of shifting rules - Best results

Performance	Rules								
Measures	EDD	ST	ST_{Post}	CR	CR_{Post}	MI	MI_{ST}	MI_{CR}	CRPM
$\sum T_l$	0	53	19	59	**0**	30	19	5	47
$\sum U_l$	**0**	50	12	56	**0**	23	13	3	41
T_{max}	0	3	1	3	**0**	3	1	1	4
CT	29.29	30.90	29.58	31.16	**29.25**	30.07	29.58	29.26	30.93
TP%	0.00%	2.80%	2.80%	2.60%	1.90%	**2.90**%	2.80%	2.50%	2.25%
Utlz%	0.00%	3.13%	3.13%	3.13%	3.13%	**5.13**%	3.13%	3.33%	3.32%

Table 3.: Comparison of shifting rules on Total Tardiness

Instances	Rules								
	EDD	ST	ST_{Post}	CR	CR_{Post}	MI	MI_{ST}	MI_{CR}	CRPM
1	18	133	19	151	18	30	19	19	62
2	0	439	91	157	0	60	37	34	201
3	37	53	41	59	37	57	41	37	266
4	33	73	51	144	36	73	45	42	329
5	6	134	30	215	6	119	29	7	47
6	174	892	28	793	7	98	29	13	520
7	119	596	62	414	2	79	60	5	263
8	48	1001	249	791	63	321	258	67	361
9	73	1020	265	814	76	400	265	81	430
10	2086	2279	2122	3444	1353	2355	1859	1752	509
11	1607	2489	2309	2346	1154	2586	2170	1838	536
12	1330	2630	2169	1982	1028	2352	2078	1461	815
13	1493	2104	2073	2191	884	2427	1878	1553	866
14	2528	3188	2981	2842	1625	2768	2885	1759	920
15	1254	3054	2618	3017	878	1586	2550	1540	1015
16	1610	3353	2894	2979	1061	2693	2897	2140	830
17	1637	2827	2567	2504	1280	2734	2416	2132	1104
18	802	2460	1964	2505	628	1522	1362	1335	1370
19	1004	2331	2351	2531	700	1662	2075	2453	1397
20	973	1737	1383	2182	881	1730	1114	1349	2059
21	907	1289	1301	1188	1105	1162	931	1025	2202
22	1106	761	780	700	1363	1186	862	823	2476
23	690	683	679	1169	482	700	774	813	1961
24	678	708	702	1297	496	697	723	763	1988
25	507	658	658	996	439	743	784	631	2711
Avg	829	1476	1216	1496	**624**	1206	1126	947	1010
Max	2528	3353	2981	3444	**1625**	2768	2897	2453	2711
Min	0	53	19	59	**0**	30	19	5	47

Table 4.: Comparison of shifting rules on Maximum Tardiness

Instances	Rules								
	EDD	ST	ST_{Post}	CR	CR_{Post}	MI	MI_{ST}	MI_{CR}	CRPM
1	2	6	2	4	2	3	2	2	4
2	0	4	1	3	0	3	1	1	4
3	3	3	3	3	3	3	3	3	6
4	1	3	3	4	1	5	2	2	6
5	3	5	5	4	3	7	5	3	5
6	5	7	5	5	5	5	5	5	9
7	2	6	4	4	1	6	4	1	5
8	2	7	5	4	2	7	5	4	6
9	3	7	5	4	2	7	5	3	6
10	20	17	17	17	9	20	17	10	24
11	16	17	17	16	7	17	17	16	18
12	12	17	17	17	9	18	17	14	16
13	14	18	15	18	9	18	15	14	16
14	15	17	15	15	8	20	14	12	15
15	12	17	17	16	8	19	17	13	12
16	17	17	16	17	8	19	15	14	
17	22	18	18	18	8	21	15	15	17
18	15	14	15	14	14	20	13	13	23
19	14	11	11	9	8	14	10	10	15
20	11	11	11	11	7	11	11	11	14
21	11	12	12	10	7	9	11	11	12
22	11	13	13	11	8	12	12	12	11
23	10	12	12	10	8	11	11	12	11
24	10	12	12	10	8	11	12	12	10
25	13	13	14	12	6	14	14	14	13
Avg	9.76	11.36	10.60	10.24	**6.04**	12.00	10.12	9.08	11.52
Max	22	18	18	18	**14**	21	17	16	24
Min	0	3	1	3	**0**	3	1	1	4

Table 5.: Comparison of shifting rules on number of late lots

Instances	Rules								
	EDD	ST	ST_{Post}	CR	CR_{Post}	MI	MI_{ST}	MI_{CR}	CRPM
1	17	73	18	85	17	23	18	18	55
2	0	241	91	113	0	47	37	34	139
3	35	51	39	56	35	49	39	35	144
4	33	56	45	93	36	56	42	39	210
5	2	50	15	102	2	62	14	3	41
6	169	252	12	378	2	46	13	6	363
7	82	168	39	219	2	46	37	5	180
8	47	242	197	363	62	155	209	58	238
9	71	244	202	343	75	166	202	78	291
10	698	471	478	812	888	797	493	597	252
11	712	532	490	743	829	658	576	492	309
12	587	541	474	744	698	724	539	420	361
13	508	425	398	701	491	806	420	380	454
14	955	649	605	838	1030	937	616	672	511
15	450	523	518	763	577	593	549	402	399
16	605	598	564	866	679	821	623	557	342
17	718	738	682	957	788	991	718	657	490
18	388	550	518	860	431	589	503	519	532
19	603	545	588	828	571	729	675	775	488
20	537	332	333	505	575	618	311	407	733
21	482	283	351	349	720	522	311	380	769
22	630	214	259	257	897	557	337	405	767
23	299	186	205	519	256	355	272	301	608
24	259	189	214	567	260	384	259	304	684
25	208	183	198	347	222	334	291	270	865
Avg	363.8	333.44	**301.32**	496.32	405.72	442.6	324.16	312.56	409
Max	955	738	**682**	957	1030	991	718	775	865
Min	**0**	50	12	56	**0**	23	13	3	41

Table 6.: Comparison of shifting rules on Cycle Time

Instances	Rules								
	EDD	ST	ST_{Post}	CR	CR_{Post}	MI	MI_{ST}	MI_{CR}	CRPM
1	33.68	34.01	33.69	34.04	33.68	33.7	33.68	33.68	32.82
2	32.95	34.13	33.18	33.27	32.95	33.06	33.04	33.03	32.18
3	33.29	33.33	33.3	33.35	33.29	33.33	33.3	33.29	31.76
4	31.94	32.04	31.97	32.23	31.94	32.01	31.95	31.95	31.00
5	30.54	30.9	30.6	31.16	30.54	30.77	30.6	30.54	30.95
6	30.2	32.45	30.13	32.21	30.03	30.22	30.13	30.06	31.46
7	30.67	31.98	30.48	31.5	30.31	30.47	30.48	30.32	30.93
8	29.47	32.09	29.75	31.38	29.44	30.07	29.74	29.48	31.49
9	29.29	31.91	29.58	31.25	29.25	30.07	29.58	29.26	31.28
10	51.16	51.76	51.36	54.71	49.11	51.46	50.7	50.31	53.24
11	47.57	50.4	49.99	49.9	46.41	50.22	49.54	48.68	54.16
12	48.73	52.02	50.85	50.41	47.92	50.88	50.52	49.02	56.82
13	48.53	50.2	50.14	50.32	47.09	50.91	49.62	48.8	52.78
14	53.39	56.52	55.78	55.12	51.26	54.78	55.41	51.77	55.07
15	46.89	51.72	50.63	51.56	45.87	47.76	50.51	47.69	56.59
16	51.4	55.63	54.55	54.67	50.14	54.19	54.6	52.71	56.02
17	49.1	52.22	51.58	51.42	48.15	51.84	51.26	50.44	52.95
18	49.3	53.29	51.94	53.17	47.98	49.89	50.49	50.33	46.49
19	50.46	54.19	54.26	53.91	49.19	51.6	53.26	54.06	46.23
20	50.12	53.74	53.52	53.57	49.21	52.13	54.7	51.63	48.52
21	52.71	53.94	53.14	53.36	51.16	52.63	53.52	51.28	49.63
22	53.9	55.56	54.31	54.44	51.85	53.43	54.64	52.07	50.82
23	50.37	54.61	54.09	54.62	49.57	53.22	54.51	52.4	50.08
24	50.65	55.25	54.67	55.35	49.9	53.53	54.76	52.86	50.08
25	48.27	52.3	51.17	51.53	48.09	51.13	50.84	49.56	49.75
Avg	43.38	45.85	44.99	45.54	**42.57**	44.53	44.86	43.81	44.52
Max	53.90	56.52	55.78	55.35	**51.85**	54.78	55.41	54.06	56.82
Min	29.29	30.90	29.58	31.16	**29.25**	30.07	29.58	29.26	30.93

Table 7.: Comparison of shifting rules on Throughput

Instances	Rules								
	EDD	ST	ST_{Post}	CR	CR_{Post}	MI	MI_{ST}	MI_{CR}	CRPM
1	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.2%
2	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.2%
3	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	-0.3%
4	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.1%
5	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	-0.7%
6	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.0%
7	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	-0.5%
8	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.8%
9	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.6%
10	0.00%	1.90%	1.90%	1.60%	1.20%	2.00%	2.00%	1.40%	0.3%
11	0.00%	2.00%	2.00%	1.70%	1.10%	2.40%	2.20%	1.60%	-0.6%
12	0.00%	1.80%	1.80%	1.50%	0.90%	2.00%	1.70%	1.60%	-0.2%
13	0.00%	1.60%	1.60%	1.10%	1.00%	2.00%	1.80%	1.50%	-0.6%
14	0.00%	2.80%	2.80%	2.60%	1.50%	2.90%	2.80%	2.50%	1.9%
15	0.00%	1.20%	1.30%	0.70%	0.70%	1.60%	1.40%	1.30%	1.2%
16	0.00%	1.50%	1.50%	0.80%	1.00%	1.50%	1.50%	1.50%	1.4%
17	0.00%	0.80%	1.00%	0.60%	0.60%	1.10%	0.90%	1.00%	2.2%
18	0.00%	1.30%	1.60%	0.40%	1.90%	2.10%	1.80%	1.80%	-0.3%
19	0.00%	1.10%	1.20%	1.50%	1.40%	1.90%	1.60%	1.40%	1.1%
20	0.00%	1.00%	1.20%	0.40%	1.00%	1.40%	1.10%	1.10%	-0.7%
21	0.00%	0.90%	0.90%	0.40%	1.10%	1.40%	1.40%	1.20%	1.0%
22	0.00%	0.40%	0.50%	0.40%	0.80%	1.00%	0.90%	0.70%	-0.5%
23	0.00%	0.50%	0.50%	0.50%	0.60%	0.70%	0.60%	0.60%	0.8%
24	0.00%	0.60%	0.60%	0.50%	0.60%	0.80%	0.70%	0.70%	0.6%
25	0.00%	0.10%	0.10%	-0.10%	0.20%	0.40%	0.40%	0.40%	-1.8%
Avg	0.00%	0.78%	0.82%	0.58%	0.62%	**1.01**%	0.91%	0.81%	0.45%
Max	0.00%	2.80%	2.80%	2.60%	1.90%	**2.90**%	2.80%	2.50%	2.25%
Min	0.00%	0.00%	0.00%	-0.10%	0.00%	0.00%	0.00%	0.00%	-1.85%

Table 8.: Comparison of shifting rules on Utilization rate

Instances	Rules								
	EDD	ST	ST_{Post}	CR	CR_{Post}	MI	MI_{ST}	MI_{CR}	CRPM
1	0.00%	2.86%	2.86%	2.86%	2.86%	2.86%	2.86%	2.86%	2.1%
2	0.00%	3.03%	3.03%	3.03%	3.03%	3.03%	3.03%	3.03%	3.3%
3	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	-1.1%
4	0.00%	0.00%	0.00%	0.00%	0.00%	3.33%	0.00%	3.33%	0.8%
5	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	-1.5%
6	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.3%
7	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	-1.2%
8	0.00%	-3.45%	0.00%	-3.45%	0.00%	0.00%	0.00%	0.00%	0.1%
9	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
10	0.00%	3.13%	3.13%	3.13%	3.13%	3.13%	3.13%	3.13%	0.5%
11	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	-0.6%
12	0.00%	0.00%	3.03%	0.00%	0.00%	3.03%	3.03%	3.03%	-1.5%
13	0.00%	0.00%	2.94%	2.94%	0.00%	2.94%	2.94%	2.94%	-0.5%
14	0.00%	3.03%	3.03%	3.03%	3.03%	3.03%	3.03%	3.03%	2.6%
15	0.00%	3.03%	3.03%	0.00%	3.03%	3.03%	3.03%	3.03%	2.1%
16	0.00%	3.03%	3.03%	3.03%	3.03%	3.03%	3.03%	3.03%	2.0%
17	0.00%	3.13%	3.13%	0.00%	0.00%	3.13%	3.13%	0.00%	3.3%
18	0.00%	2.56%	2.56%	2.56%	2.56%	5.13%	2.56%	2.56%	-1.1%
19	0.00%	0.00%	0.00%	0.00%	0.00%	2.50%	2.50%	0.00%	0.8%
20	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	-1.5%
21	0.00%	0.00%	0.00%	0.00%	0.00%	2.86%	2.86%	2.86%	0.3%
22	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	-1.2%
23	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.70%	2.70%	0.1%
24	0.00%	0.00%	0.00%	0.00%	2.70%	2.70%	2.70%	0.00%	0.00%
25	0.00%	0.00%	0.00%	-2.78%	0.00%	0.00%	0.00%	0.00%	-2.6%
Avg	0.00%	0.81%	1.19%	0.57%	0.93%	**1.75**%	1.62%	1.42%	0.21%
Max	0.00%	3.13%	3.13%	3.13%	3.13%	**5.13**%	3.13%	3.33%	3.32%
Min	0.00%	-3.45%	0.00%	-3.45%	0.00%	0.00%	0.00%	0.00%	-2.59%