Post print:

Scheduling batches with time constraints in wafer fabrication

Pirovano, G., Ciccullo, F., Pero, M., Rossi, T.

Published in 2020 in the International Journal of Operational Research

To cite the paper: Pirovano, G., Ciccullo, F., Pero, M., Rossi, T. (2020) Scheduling batches with time constraints in wafer fabrication International Journal of Operational Research, 37(1), pp. 1-31

Link: http://www.inderscience.com/storage/f610254791218113.pdf

DOI: 10.1504/IJOR.2020.104222

Scheduling batches with time constraints in wafer fabrication

Abstract

This work proposes and tests an algorithm for batching and dispatching lots along cleaning and diffusion operations of a wafer fab. These are characterized by (i) time constraints (i.e. the time between the end of an operation 'n' and the start of the operation 'n+q' must be lower than a time-limit, in order to guarantee the lots' quality) and (ii) absence of batching affinity between operations. Literature so far has been falling short in proposing scheduling algorithms suitable for this context. Therefore, we propose two heuristic algorithms to minimize the average flow time and the number of re-cleaned lots, maximize machine saturation, and avoid scrapped lots. Discrete-event simulation was used to test the performance of the two algorithms using real data of STMicroelectronics. The formerly proposed model outperforms the latter. Therefore, STMicroelectronics implemented the former in its fab in Catania gaining an increase in the average Overall Equipment Effectiveness of 7%.

Keywords: semiconductor manufacturing, dispatching rules, batch, scheduling, wafer fab, time constraints, diffusion, STMicroelectronics

1. Introduction

Electronics industry is one of the largest industries in the world. Key players are the semiconductor manufacturers, which produce integrated circuits on silicon wafers. According to a report by the Semiconductor Industry Association, the worldwide sales of semiconductors reached a peak of $25.53 billion in July 2013 (an increase of 5.1 % over July 2012). To remain competitive in the industry, semiconductor manufacturers must satisfy customers in terms of quality, quantity and due dates (Kim et al., 2001).

Wafer fabrication is a very complex manufacturing process (Tajan et al. 2013). It takes place in fabs, which are very large systems with tens to hundreds of machines moving hundreds to thousands of wafers (Mönch et al., 2011). The scheduling problems in a wafer fab is similar to scheduling flexible job shop but it is more complicated due to (Dabbas and Fowler 2003): (1) Re-entrance of lots in some stages (Jia et al. 2015); (2) Different types of machines, e.g. machines processing one wafer at a time, or lot (a set of wafers) or batch (a set of lots); (3) Setup times; (4) Auxiliary resources such as reticles in the photolithography stage; (5) Multiple orders per lot; (6) Uncertainties, e.g. in arrival of new orders and machine failure; (7) Time constraints: often, the time between the end of operation 'n' and the start of operation 'n+q' must be lower than a time-limit, in order to guarantee lots' quality.

Literature has approached scheduling problems of flexible job-shop differently. Differences lie in the methodologies adopted (Narayanaswami and Rangaraj, 2012), but also in the inclusion of different types of constraints that reflect different contextual situations. Until recently, authors have proposed different scheduling models in a job-shop context considering as constraints for example: no-wait time between operations (e.g. Ahani and Asyabani, 2014); batch processing machines, capable of concurrently load more than one job (e.g. Tajan et al., 2013), limited capacity of sequential processors and incompatible job families (e.g. Tajan et al., 2013). The continuously reinforcing interest in these types of problems is fostered by the more and more constrained requirements of companies in different sectors (Cheng et al., 2014). Although different scheduling problems are generally built upon relevant assumptions that combine some of the constraints mentioned above (e.g. Bilyk et al., 2014) few are the contributions (e.g. Tajan et al., 2012; 2013) that not only develop a complex

scheduling problem, but also test them in a real context, providing a model that can be easily operationalised in the wafer fab.

This work proposes an approach which attempts to solve the problem experimented by production managers in scheduling jobs the cleaning and diffusion areas of a wafer fab. The areas in which cleaning and diffusion operations are performed are the most critical ones of wafer fabs, due to, among others, the presence of multiple inhomogeneous processes, batching machines, long process time and time constraints (Jung et al., 2014). To the best of our knowledge, as it will be explained in the literature review section, none of the models proposed so far in literature can be used to solve the problem addressed by this paper.

The specific problem we are addressing is defining the batches and dispatching the lots among three consecutive operations (i.e. one cleaning and two diffusions) characterized by time constraints between them and no batches affinity between the first and the second operation. By following the classification of Brucker (2007) and considered 'j' as the lot's index, such a problem can be described as in expression (1):

$$R\,J\left|\, p_j, d_j, p-batch, incompatible-batch, T_{lags}^{min}\right| \min(\overline{FT}), \max(Saturation), \min(R_{work}), S_{scrap}=0 \quad (1)$$

Where the first part of (1) defines the machine environment, which is unrelated parallel machines (R) with different operations (J). The second part describes the product characteristics. In particular, the studied situation is characterized by different processing time of the lots (pj), relevant and different due dates for the lots (dj), the time for processing a batch on a machine is the longest processing time among all the lots in the batch (p-batch), not all the lots can be batched together (incompatible-batch), and presence of time constraints between operations (Tlagsmin). The last part represents the objectives, which are: minimizing the average flow time of the lots (FT) and the number of re-cleaned lots (Rwork), maximize machine Saturation, and avoid scrapped lots (Sscrap=0).

In particular, it presents the results of a research carried on in the framework of the EU-funded project Integrated Solutions for Agile Manufacturing in High-mix Semiconductor Fabs (INTEGRATE). The studied case is one of the STMicroelectronics Catania plants.

The remainder of the paper is organized as follows. In section 2 the literature related to batching and scheduling approaches with time constraints is reviewed. In section 3 the problem under analysis is described. Section 4 presents Model 1, while section 5 Model 2. The simulation model developed and the campaigns run to test the performance of the two models are presented in section 6. Section 7 discusses the main results. Conclusions and direction for future research are presented in section 8.

2. Literature Review

Several researchers have addressed the problem of composing and dispatching batches in a wafer fab. The presence of various elements to consider (e.g.: parallel machines, time constraints, objectives (Mönch et al., 2011)) has brought to different solving techniques, such as mathematical programming, simulation heuristic (Mathirajan and Sivakumar, 2006) and meta-heuristic (e.g. Bilyk et al., 2014; Ahani and Asyabani, 2014) algorithms.

Over the years, authors have developed methods based on the value of an index that takes into account lots' features and that is adopted as a priority rule (e.g. Saadaoui et al., 2014). To minimize the sum of weighted tardiness, Vepsalainen and Morton (1987) develop an index called Apparent Tardiness Cost (ATC), which assigns the priority on the basis of the expected tardiness cost per immediate processing requirements. An evolution of this index is presented by Cerekci and Amarnath (2010). They focus on mean tardiness performance of a batch machine in a two-stages system by including

an upstream unit capacity machine. They propose two new control strategies, called BATC-I and BATC-II, in order to compose and dispatch the batches. Akçali et al. (2000) focus on a loading policy based on the minimum batch size, and different dispatching policies, based on an index called critical ratio of lots.

Some approaches incorporate information about the state of the fab. Solomon et al. (2002) develop a dispatching policy for batch-processing machines that incorporates information about future arrivals and the status of critical machines, in order to improve the makespan. Also, Cigolini et al. (2002) consider the future arrivals for scheduling several products on parallel batching machines. They focus on the context in which there is less than a full load batch queuing at a batching machine. The procedure compares the total delay of lots in the queue when an incomplete batch is processed immediately versus the one obtained by waiting for the next arrival. Flower et al. (2000) present a dispatching policy, based on NACH (Fowler, 1992) strategy, which considers the future arrivals. The developed policy considers multi-family and parallel batch machines and has a pull decision point when a machine becomes available and a push one when lots arrive in the queue. Tajan et al. (2012) modelled a two-stage system made up by an upstream serial processor and a downstream batch processor (a diffusion furnace). In their proposed heuristic model, a batch is formed at the upstream stage considering the "job family preference" of the batch processor. A simulation campaign tested the model and revealed a substantial cycle time reduction.

Others influencing factors taken into consideration in the extant literature are: the number of considered lots family (Mönch et al., 2005), different typologies of batch processing machines (Yang et al., 2013), machines not always available in the planning horizon (Golmakani and Namazi, 2014).

Other authors consider multiple objectives functions. Sung and Kim (2003) develop three efficient polynomial time algorithms for minimizing three corresponding due-date related measures including maximum tardiness, the number of tardy jobs and total tardiness. Kim et al. (1998) develop a scheduling approach in order to face: i) lot release control; ii) mask scheduling in the photolithographic workstations and iii) batch scheduling. He et al. (2016) consider two bi-criteria problems for scheduling jobs on a single machine with parallel batching. As objectives, they set: minimisation of the makespan subject to maximum rejection cost and minimisation of total rejection cost, subject to a maximum makespan.

Some works aim at planning the entire fab. Caumond et al. (2008) introduced a framework based on a disjunctive graph to model the problem and on a mimetic algorithm for generating the lots scheduling on the machines. Chen and Tang (2012) base their heuristics algorithm on four priority rules, which are applied at the different stages of the system in order to minimize the number of tardy lots. Re-entrant flows and queue time constraints are taken into considerations.

As far as time constraints, both practitioners and researchers have recognized the relevance of the topic. Cho et al. (2014) propose a local scheduling model considering queue time constraints related to the diffusion area of a wafer fab. They seek to minimise a multiple objectives function, which comprises five components: weighted completion time of a job, queue-time penalties/reworks, penalties for admitting a queue-time constrained lot into the queue-time zone, tool idle times, and weighted throughput.

Solutions have been proposed in industries other than semiconductors and considering both tardiness and makespan as objectives. Gicquel et al. (2012) study hybrid flow-shop scheduling problems in bio-process industry, where limited waiting time between processing stages must be considered. They propose an exact solution approach for minimizing the total weighted tardiness, based on a discrete time representation and a mixed-integer linear programming formulation. Li and Li (2007) study the hybrid flow-shop scheduling problem with limited waiting time constraint in a multi-stage process, characterized by parallel unit capacity machines. The objective is to minimize the makespan for a

given set of jobs. They propose a recursive backtracking algorithm, which schedules each job from the first stage to the last.

Finally, there are some works considering both batch-processing machines and time constraints. Su (2003) proposes a heuristic algorithm to minimize the makespan on a two stages process, where at the first stage there is a batching machine and lots belong to the same family. Yugma et al. (2012) propose a solving method based on disjunctive graph representation. In addition, they develop a constructive algorithm based on iterative sampling and Simulated Annealing, in order to improve the initial solution. Table 1 re-classifies the presented contributions according to the notation of Bucker (2007).

As Table 1 shows, none of the already proposed model matches with expression (1), i.e. none of them can be successfully applied to solve the problem of the STMicroelectronics Catania fab.

<Insert Table 1 approx. here>

3. Research objective and methodology

The objective of this work is to develop an algorithm to support batching and dispatching decisions in the cleaning and diffusion areas of a wafer fab. The algorithm aims to minimize the average flow time of lots, maximize saturation of machines, minimize the occurrence of reworks, and nullifying the occurrence of scraps, while taking into account the complexity of a real industrial environment, e.g. time constraints between operations. The problem to address is the one faced by STMicroelectronics Catania fab. Therefore, the developed model must be applicable to the company.

To this aim, the following methodology has been followed:

(i) *Problem description*, i.e. detailed analysis of the needs of STMicroelectronics Catania fab in the cleaning and diffusion areas;

(ii) *Development of two models*, i.e. two heuristic algorithms, hereinafter called model 1 and model 2, to address such needs;

(iii) *Simulation campaign:* i.e. development of simulation models of the STMicroelectronics Catania fab, and running of simulation campaigns with real data of the fab to compare model 1 and model 2;

(iv) *Results analysis*, i.e. analysis of the results of the simulation campaigns and comparison of the performance of the two models;

(v) *Choice:* i.e. choice of the model to implement in the company.

The analysis of the needs of the fab allowed understanding the context and the specific characteristics of the cleaning and diffusion areas, the performance that the company aims to optimize, and the kind of model to develop. In particular, we chose to develop a heuristic algorithm. In fact, the company requested for an algorithm that can be easily implemented in their information systems, and that can be easily understood by the operators in the fab. Moreover, the company traditionally uses heuristics to schedule operations.

4. Problem description

The case study analysed refers to the cleaning and diffusion areas of STMicroelectronics Catania fab.

<Insert Figure 1 approx. here>

In each area, there are only parallel batching machines, i.e. each machine can process a maximum number of lots at the same time (Capacity of the machine). Batching machines are typical of wafer fabs (Tajan et al. 2012). The process time of each batch on each machine is variable and depends on the "recipe", i.e. the technological parameters of the process the lot has to perform on the machine. All the lots loaded together in a batch must have the same "recipe" (batch affinity). Moreover, each machine can perform a limited predefined set of recipes. The machines in the diffusion area have two load-spaces, i.e. while one batch is being processed in one load-space, another batch can be loaded in the other load-space.

Figure 1 shows the flows along the two areas. There are three types of lots' flows: lots of type A have to perform only cleaning; lots of type B have to perform cleaning and one diffusion operation; lots of type C have to perform cleaning and two consecutive diffusion operations. The lots of type B and C, after having performed the cleaning operation, must wait before performing the diffusion operation no longer than a certain time value X. If a lot waits for a time longer than X, it must be re-cleaned.

The lots of type C, after having performed the first diffusion operation, must wait before performing the second diffusion operation no longer than a certain time value Y. If a lot waits for a time longer than Y, it must be scrapped.

The waiting time is calculated from the moment, when the lots finish to be processed at one operation, to the instance when the lots are loaded on a load-space in a machine at the next operation.

Two generic lots can have the same recipe at the first operation and different recipes at the second operation. However, lots of Type C share the same recipe at the first and second diffusion operation.

Given the context, the algorithms need to address simultaneously a batching problem, i.e. which lots to load together on a machine, and a dispatching problem, i.e. when to load the batch on the machine.

The main objectives set by STMicroelectronics managers are: minimize the average flow time of lots, maximize machine saturation, avoid scrapped lots and minimize the number of re-cleaned lots.

5. Development of two models

Two heuristic algorithms have been developed. They differ in terms of how batches are formed and dispatched along operations.

In particular, in model 1, batches are composed at cleaning operation with lots sharing the same recipes in all the operations, i.e. the cleaning operation, the first diffusion operation and the second diffusion operation, so that each batch can be dispatched on the machines at different operations without any needs of splitting the batch or waiting other lots. Then, batches are dispatched using a dispatching rule. To assure the respect of the time constraint between cleaning and diffusion, batches of type B and C are loaded on the cleaning machine only if the expected queue in front of the diffusion machine, in hours, is lower than the time constraint (look ahead). As for the batches that need to perform two consecutive diffusions, i.e. type C, the respect of time constraint is assured by booking the load–space on the machine, which has to perform the second diffusion.

As far as model 2 is concerned, batches of lots sharing the same recipe are formed at each operation. Thus, at the end of each operation, the batches are split, and lots are ready to join other lots in batches for the next operation. Dispatching is based on priorities indexes similar to the one developed by Tu and Chen (2008). The respect of time constraint between cleaning and first diffusion is assured by checking, after a predefined time, whether to load the lots in the queue in front of the first diffusion on the machine immediately or wait more for other lots to arrive. As for the time constraint between the first and the second diffusion, lots of type C are dispatched at cleaning only if a machine for the

second diffusion will be available within a time threshold equals to the average cleaning time of the lots in the batch plus the diffusion process time.

Table 2 depicts the main differences between the two models.

<Insert *Table 2– comparison between model 1 and 2* approx. here>

6. Model 1

Model 1 seeks to form, at the cleaning operation, the batches that will be dispatched in the two areas. Batches have to be composed of as many lots as possible (respecting machines capacity), of the same type and characterized by batch affinity in all the operations, and – if of Type C – that will respect the time constraint Y between the two diffusion operations. This is to avoid lots waiting at the machines in diffusion area, due to lack of others lots to be included in the batch.

Such a method for forming batches is coherent with all the objectives listed in section 3. Indeed, the higher the lot waiting time at first diffusion operation, the higher the probability that the lot should be re-cleaned and the higher the lot flow time. The higher the lot waiting time at the second diffusion operation, the higher the probability that the lot is scrapped and, again, the higher the lot flow time. Finally, the lower the number of lots composing the batch, the lower the machines saturation.

The parameters used in the algorithm are the followings:

Indexes		Decision parameters	
J	Lot index	ΔT	Time within which it is possible to look for the lots that will arrive at the cleaning operation
I	Machine index	Wtmax	Maximum waiting time allowed for the lot at first stage [hours]
O	Operation index (1=cleaning, 2=first diffusion, 3=second diffusion)	Tf	The machine for the second diffusion is booked for the batch Tf hours after the cleaning operation [hours]
G	Recipe index		
K	Batch index		
Parameters			
$WIP_{g,i}$	# of waiting batches with recipe g in front of the machine i [equivalent batches]	b_i	Boolean variable which is 0 if the machine is broken 1 in the other case
$u_{g,i}$	Average service rate of the machine i for the recipe g [equivalent batches/hours]	n_k	Number of lots in a batch (when the batch has the same size for all operations)
p_{oj}	Process time of lot j at operation o [hours]	z	Number of recipes
Tnow	Time when the priority indexes are calculated [hours]	a_i	Average availability of the machine i
d_j	Due date of the lot j [hours]	TC	Time constraint between two stages [hours]

L	Look-ahead factor (1,5<K<4,5)	CAP_i	Capacity of the machine i
R_j	Set of the recipes of the lot j	W_{t1j}	Waiting time of lot j at cleaning operation [hours]
R_{jo}	Recipe of the lot j at operation o	n_{ko}	Number of lots in a batch at operation o
V_j	Cost of the delay on delivery for the lot j [€/hours]		

With reference to the dispatching rule used at the cleaning area, we test two different priority indexes: ATC by Vepsalainen and Morton (1987) (formula 2) and BATC by Cerekci and Banerjee (2010) (formula 3).

Since one of the key factors to succeed in the market is meeting due dates (Kim et al., 2001), ATC was chosen as priority index for its ability to outperform other priority indexes, such as, among the others, COVERT, FIFO, EDD, in reducing weighted tardiness and number of tardy jobs (Vepsalainen and Morton, 1987).

$$ATC_j = \frac{V_j}{p_{1,j}} e^{-\left(\frac{(d_j - p_j - tnow)}{Lp}\right)} \quad (2)$$

ATC is calculated for each lot. It is composed of two parts: (i) the ratio between the cost of delay (V_j) and the process time (p_{1j}), which represents the weighted shortest process time and gives priority to the lots with shorter process time, assuming equal cost for delay, (ii) the index slack per remaining process time, which gives priority to the lots with the lowest remaining process time.

$$BATC_k = \sum_{l=1}^{n_k} \frac{V_l}{p_{1,l}} e^{-\left(\frac{(d_l - p_l - tnow)}{Lp}\right)} \left(\frac{n_k}{CAP_i}\right) \quad (3)$$

BATC is the sum of the ATC indexes of the n_k lots l composing the batch k weighted by machine saturation with that batch. Therefore, with BATC, we consider the objective of maximizing batch saturation, while also considering tardiness.

Expression (4) presents the dispatching rule at the cleaning used for lots that have to perform at least one diffusion, i.e. type B and C. (4), in line with Little's law, aims to limit the number of batches waiting at the first diffusion in order to avoid to overstep the time constraint.

$$TC * b_i < \sum_{g=1}^{z} \frac{Wip_{g,i}}{a_i * u_{g,i}} \quad (4)$$

In particular, (4) checks if a specific machine can process all the batches in its queue without overstepping the time constraint (TC), by considering the waiting batches ($WIP_{g,i}$), the service rate ($u_{g,i}$) and machine availability (a_i), in line with Little's law. When the machine is broken the left part of expression (4) is set to zero and no lots are allowed to enter in the queue.

6.1 Model 1: algorithm

When at the cleaning area either: (i) a machine is empty, or (ii) a new lot arrives in queue and a cleaning machine is empty, or (iii) a machine at the diffusion area becomes free and a cleaning machine is empty, the priority index of all the lots in the queue and the ones arriving within ΔT is calculated. After the index calculation, the algorithm orders the lots or batches according to its value and selects the one with higher priority index among the ones that can be processed on the free cleaning machine (see Figure 2 for ATC). When BATC is used, using the lots that have the same recipes and can be processed on the free machine, all the possible batches are formed and BATC calculated for them.

See Figure 3, Figures 4 and 5 for a description of the algorithm for flow types A, B and C respectively.

If the type of the chosen lot is:

1) A: a batch is formed by grouping the selected lot with others, of the same type sharing batch affinity, that are in queue or arriving within ΔT (note that if in the batch there are lots not yet arrived but arriving within ΔT the machine is considered «busy» by the batch).
2) B or C: it evaluates if expression (4) is respected. When it is the case, lots of the same type and batch affinity, already in queue or arriving within ΔT, are grouped in a batch.

In case the number of lots in the batch is lower than the maximum capacity of the batch machine: if, among the lots in the batch formed, there are lots that have been waiting in queue for more than Wtmax, the batch is loaded. Otherwise, the next lot (or batch) with the highest priority index is selected.

After having performed the cleaning operation, the lots arrive physically in the queue in front of the diffusion, here they have to wait for all the lots of the batch.
When a load-space of a machine for diffusion is empty: using a FIFO logic, among the batches in queue, a batch that can be processed on that machine is selected. For type C lots, in order to ensure that the time constraint, between the first and second diffusion, will be respected, the batch will book the load–space on the machine, which has to perform the second diffusion. This happens when the difference between Tnow and the end of the cleaning operation is smaller than a threshold (Tf).

<Insert *Figure 2. Priority index calculation (for ATC)* here>

<Insert *Figure 3. Cleaning scheduling* here>

< Insert *Figure 4. Cleaning-diffusion scheduling* here>

< Insert *Figure 5. Cleaning-diffusion-diffusion scheduling* here>

7. Model 2

By applying Model 1, lots of type A might experience a high flow time, due to their waiting time in the queue in front of the cleaning machine (as a matter of fact they have to wait for lots with batch affinity). Therefore, Model 2 allows to overcome the likely deterioration of flow time performance, complying with the time constraint imposed between the consecutive stages.

In this model, the queue in front of the cleaning area remains physically unique, but it is logically split in:

- *stand-by queue*: it contains the B and C-type lots that are not yet allowed to be processed at cleaning;
- *free-pass queue*: it contains all the A-type lots, along with B and C-type lots, which have been authorized to enter in this queue.

"Model 2" consists of two main parts: (i) a procedure to dispatch batches at cleaning and diffusion machines (see section 7.2: dispatching at cleaning and diffusion areas), and (ii) a procedure to assign B and C-type lots to the free pass queue (see section 7.1: sub-algorithm for assigning lots to diffusion machines).

With reference to the main parameters characterizing the algorithm, they are the followings:

Decision variables

λ	Weight of the due date
ρ = 1-λ	Weight of the batch saturation
ρD1	Weight for the batch saturation on the first diffusion
ρD2= 1-ρD1	Weight for the batch saturation on the second diffusion

Parameters

$\boldsymbol{r_j}$	Remaining processing time of the lot j
$\boldsymbol{\psi}$	Safety parameter set to 0,8
$\boldsymbol{\Delta t_{io}}$	Remaining process time of machine i on operation o
$\boldsymbol{\Delta t_c}$	Average process time at the cleaning operation for the Z recipe

With reference to the dispatching rules, they are based on two different priority indexes: ATC_j for cleaning, and Id_{ki} for diffusion. Id_{ki} is similar to the priority index proposed by Tu and Chen (2008), but it is computed in two ways depending on the lots type included in the batch. All the lots in the batch have the same recipe.

In case the batch considered for the calculation includes only lots of type B, Id_{ki} is computed as expression (5):

$$Id_{ki} = \lambda * \frac{\sum_{l=1}^{n_{k2}} \frac{Tnow + r_l}{d_l}}{n_{k2}} + \rho * \frac{n_{k2}}{CAP_i} \qquad (5)$$

If the batch considered for the calculation includes at least one lot of type C, the priority index is computed as expression (6):

$$Id_{ki,m} = \lambda * \frac{\sum_{l=1}^{n_{k2}} \frac{Tnow + r_l}{d_l}}{n_{k2}} + \rho * \left(\rho_{D1} * \frac{n_{k2}}{CAP_i} + \rho_{D2} * \frac{n_{k3}}{CAP_m}\right) \qquad (6)$$

Where CAPm=the Capacity of the machine performing the second diffusion

7.1 Model 2: sub-algorithm for assigning lots to diffusion machines

Sub-algorithm (Figure 6) starts when one machine at the diffusion operation is expected to become free within a Δt_c . The sub-algorithm allows for:

- choosing all those lots that can be processed on the selected machine (lots of type B and type C);
- batching them considering the recipe at the first diffusion operation;
- checking, for each C-type lot, if there would be an empty machine that can process it at the second diffusion, after a slack of time equal to the average cleaning time plus the diffusion process time. If this condition becomes true, it is included in the batch. The machine is assigned to the lot. Therefore it is possible to take into account that this lot will be processed on the machine when checking machine idleness;
- calculating, for all the batches formed, the priority index $Id_{ki,}$ $Id_{ki,m}$;
- choosing the batch with the highest priority index and moving the composed lots of to the free-pass queue.

<Insert *Figure 6. Sub-algorithm for creating the free-pass queue* here>

7.2 Model 2: dispatching at the cleaning and diffusion areas

The trigger event for starting the dispatching at the cleaning area is: there is an empty machine in the cleaning area. The lots to be dispatched are the ones belonging to the free-pass queue. For each lot, the priority index ATC_j is computed. A batch is then formed by including those lots that can be processed on the available machine with higher ATC_j.

Lots in the queue at the first diffusion have to wait for all the lots having the same index (Id_{ki} or Id_{kim}). When the first lot of a batch with a certain Id_{ki} or Id_{kim} arrives in the queue in front of the first diffusion, it waits for the other lots to arrive. The algorithm, at a time called "Time control", checks where are the lots of the batch that have not yet arrived in the queue in front of the first diffusion.

Time control is calculated using the expression (7):

$$Time\ control = (TC - \max(p_{1j})) * \psi \quad (7)$$

If there is at least one lot already performing the cleaning, lots in the queue wait for this lot to finish the cleaning operations, and then the batch is loaded, when possible, on the first diffusion machine. Otherwise, lots in the diffusion queue are processed, as soon as possible, on the first diffusion machine. In this case, the batch will be incomplete. The lots that were expected to be part of the batch, but did not arrive on time, are re-assigned to the stand-by queue. Lots of type C are then loaded on the machine that has been assigned to them for the second diffusion, if free. Otherwise, they are scrapped.

If one machine at either the first or the second diffusion is down and the TC is going to be overtaken, the batch in front of these machine seeks for another eligible machine in order to avoid the re-cleaning or the scrapping.

8. Simulation campaigns

In order to test and compare Model 1 and Model 2, simulation is used. As a matter of fact, simulation is often used to check in advance production plans (see among others Rossi and Pero (2011)). A sound

simulation study has been put in place according to the SMDP approach developed by Manuj et al. (2009). Therefore, the logic model representing the flows in the areas has been validated with company's experts. Then, two different logic models and simulation models (one for each model) have been built in Arena™ (discrete event simulation and, in particular, Arena™ are used since they are commonly applied to represent production and logistics contexts. See among others Cigolini et al. (2011)).

The input data came from the real situation of STMicroelectronics' Catania plant. Data concern: 9 days of production, the characteristics of processed lots, e.g.re-entrant flow, various routes and different arrival time at the operations and the historical paths of machines' availability. In particular, the latter data have been analysed, to obtain the empirical distribution of failures. Analysis of the data of production in the fab has shown that the quantity of lots in the areas and the product mix (in terms of type of flows) is constant along the year. Therefore the 9 days considered are representative of the average situation in the areas.

Due to confidentiality reasons, no information can be provided on the number of machines and process times of the lots.

For both models, the performance analysed are: daily batch saturation, flow time of each lot, total daily number of scrapped lots, percentage of re-cleaned lots (i.e. the ratio between the total number of re-cleaned lots and the total number of lots processed at cleaning).

Daily batch saturation has been calculated as follows:

$$Daily\ batch\ saturation = \frac{\sum_{k=1}^{w} \frac{n_k}{\max capacity\ of\ the\ machine\ processing\ batch\ k}}{w}$$

Where w=number of batches processed in a day.

With reference to the design of experiments, the initial situation of the system in terms of availability of lots and machines has been recreated in the simulation model based on real data. Therefore, there is no need to consider a warm-up period. Each simulation run lasts 8.3 days, i.e. 200 hours (the plant works 24 hours/day), since after 200 hours the system starts to get empty.

The experimental design encompasses the comparison of the two models on average performance. To define the average performance, a simulation campaign for each model has been developed.

The simulation campaign for Model 1 is composed of 56 scenarios. In fact, the following levels were tested: for priority index: ATC and BATC, for Wtmax: 0 and 12 hours, for ΔT: 0,2,4,6,8,10 and 12 hours, and for Tf: 0 and 4 hours. 32 scenarios were discarded after the first simulation results, with 6 runs for each scenario. In particular, in the scenarios with $T_f = 4$, the number of scrapped lots was higher than zero and the scenarios with $\Delta T = 12$ have an average flow time too high to be acceptable by STMicroelectronics and no significant difference in terms of the other performance measures with respect to the scenarios with $\Delta T=10$. It should be noted that: Wtmax and ΔT levels were chosen considering the average queue time at cleaning, while Tf levels were chosen considering that on average a diffusion needs 5 hours to be performed.

The simulation campaign for Model 2 is composed of 5^2 scenarios, since ρ and λ, and $\rho D2$ and $\rho D1$ are complementary, it is possible to test 5 different levels of only two variables (λ and $\rho D1$) to have a complete campaign. The levels tested of both variables are: 0; 0,25; 0,5; 0,75;1.

Table 3 and 4 report the correspondence between the different scenarios adopted in the experimental phase and the values of the decision variables.

<Insert *Table 3. Scenarios adopted for model 1 experimental phase* approx. here >

<Insert *Table 4. Scenarios adopted for model 2 experimental phase* approx. here >

9. Results analysis

This section compares the output of the simulation runs applying model 1 and model 2 to the company data. Figures 7 and 8 show the results. For confidentiality reasons, we cannot provide the information regarding the values of the data obtained.

<Insert *Figure 7. Results of simulation campaign with model 1* approx. here>

<Insert *Figure 8. Results of the simulation campaign with model 2* approx. here>

9.1. Comparison of model 1 and model 2

Since the data from Model 1 and 2 lack homoscedasticity and they are not normally distributed, Mann–Whitney U test has been performed to compare the two models performance. In particular, the test allows testing whether the two samples come from the same population (η).

Based on the test, the following can be inferred from the comparison of the performance of the models:

- Average Flow Time: Test of $\eta 1 = \eta 2$ vs $\eta 1 \neq \eta 2$ is significant at 0,1499, therefore we cannot infer that the flow time of the simulated scenarios applying model 1 and the ones applying model 2 are different.
- Daily batch saturation: Test of $\eta 1 = \eta 2$ vs $\eta 1 > \eta\ 2$ is significant at 0,0000. Therefore, we might infer that the daily batch saturation of model 2 tends to be lower than the batch saturation of model 1.
- % of Recleaned lots: Test of $\eta 1 = \eta 2$ vs $\eta 1 < \eta 2$ is significant at 0,0000. Therefore, we might infer that the % of recleaned lots of model 2 tends to be higher than one in model 1.
- Scrapped lots: Test of $\eta 1 = \eta 2$ vs $\eta 1 < \eta 2$ is significant at 0,0000. Therefore, we might infer that the number of scrapped lots of model 2 tends to be higher than the one in model 1.

It must be noted that model 2 in all the simulation runs presents some scrapped lots, while the same does not apply to model 1. Since one of the objectives of the model is to assure zero scraps, we chose to implement model 1 in the case study. Moreover, since the performance of model 1 are better than the ones of model 2, we have focused our further analysis on model 1.

<Insert *Figure 9. Qualitative comparison between model 1 and model 2 considering performance* approx. here>

9.2. Model 1 performance analysis

To identify whether and how the decision variables (i.e. Wtmax, ΔT, Priority index) affect the performance (i.e. daily batch saturation, flow time, % of recleaned lots), ANOVA analysis and Tuckey pairwise comparison (with 95% confidence) have been performed on the results. We did not analyse the number of scrapped lots, since it is always zero (Tf is set to zero in all the scenarios).

In particular, tests show that data lacks homoscedasticity, but the ones related to recleaned lots. Therefore, one-way ANOVA test was used to identify whether recleaned lots was influenced by the decision variables, whereas for the other, Welch one-way ANOVA was used.

As for recleaned lots, the normality of residuals has been tested (pvalue = 0,05). Afterwards, by Tuckey's and Games-Howell's tests, the differences among mean values of the analysed performance for each level of the variables have been verified.

Table 5 presents the main results of the tests performed.

<Insert *Table 5. Tests' results for model 1* approx. here>

From Figure 10, it can be inferred that all the decision variables affect at least one performance. The daily number of recleaned lots is affected only by ΔT. Moreover, since daily batch saturation and daily number of recleaned lots are affected by ΔT in the same way, we will consider only daily batch saturation in the following analyses.

<Insert *Figure 10. The impact of decision variables on performance* approx. here>

As reported in Table 5 and as we could expect, there is a trade-off between daily batch saturation and flow time. In those scenarios where daily batch saturation is higher, also flow time is higher. Therefore, there is not the best scenario for all the performance, but to identify the values of the decision variables to use in real applications, methods for decision making with multiple objectives (e.g. Analytic Hierarchy Process) have to be used. One company can assign different priorities to the different objectives depending on some contextual variables. For example, when demand is low, companies can opt to prioritise daily batch saturation at the expenses of flow time. When instead demand is high, companies might seek to endorse flow time reduction, thus adopting a higher ΔT in order to cope with the high demand with an adequate service level. As shown in Figure 10, in this last case, companies should be aware that high ΔT can have more recleaned lots as a drawback.

In order to verify that using model 1, we can obtain better performance than those obtainable using models consolidated in literature, the results of scenarios with ATC or BATC priority index (Vepsalainen and Morton, 1987; Cerekci and Banerjee, 2010), and with Wtmax and ΔT set to null, have been confronted with the scenarios where the same priority index is used, but Wtmax and ΔT are not null, and model 1 outperforms the use of only ATC or BATC. In fact, first of all, results in Figure 6 show that scenario 1 presents repetitions with scrapped lots, that are not present in other scenarios. Moreover, with the same level of % of recleaned lots, scenarios 3 and 9, outperform scenario 1 in terms of average batch saturation, by also having zero scrapped lots. Moreover, it can be noted that scenario 19 is better than scenario 13 for what concerns daily batch saturation, but the same does not apply for flow time. However, if we compare the average values of the performance, we can observe that daily batch saturation of scenario 19 is 17,4% higher than the same performance in the scenario 13, while the flow time is only 2% higher. Therefore, in those cases when the importance of daily batch saturation is higher than the one of the flow time, scenario 19 might outperform scenario 13.

As a whole, the value to assign to the decision variables that determine different scenarios, can be affected by contextual conditions, which can lead to privilege efficiency (i.e. daily batch saturation) at the expenses of service level (i.e. longer flow time) or vice versa, targeting a flow time reduction at the expenses of inefficiencies caused by possible re-cleaned lots.

10. Conclusions

In this paper, we propose and compare two algorithms for batching and dispatching lots among three consecutive operations (i.e. one cleaning and two diffusions) of a wafer fab, characterized by time constraints between them and by the fact that all the lots can be batched together and loaded on the same machine at the first and second operation. The algorithm aims to minimize the average flow time of the lots and the number of re-cleaned lots, maximize machine saturation, and avoid scrapped lots.

Literature so far has falling short in proposing approaches to solving such a problem. Therefore, this paper fills this gap in the literature. Interestingly, the specific problem has not been debated in the literature despite its practical relevance. Indeed, this problem was faced by the managers of STMicroelectronics Catania plant. Thus, we believe our work can represent a contribution to the scheduling literature, by proposing a solution to a real complex problem and testing it using data from a real complex environment.

Two different models, i.e. model 1 and model 2, have been developed. The presented models share some major characteristics, such as the distinction in different flow types, the use of priority indexes, The major differences between the two models are: the batching rule, the dispatching rule at the cleaning operation, the dispatching rule at the diffusion operation, the way time constraints are managed between cleaning and diffusion and between first and second diffusion.

In model 2 is the decision to not batch in front of the cleaning area aims to prioritize flow time instead of batch saturation as performance outcome.

The results of the simulation campaigns show the two models are comparable in terms of batch saturation and flow time, but this is not true for scrapped and re-cleaned lots. In fact, in model 2 there are scrapped lots in each of the tested scenarios, thus violating the main STMicroelectronics target. Moreover, in model 2 there is a higher percentage of re-cleaned lots even in the best scenario than the ones present in the worst scenario of model 1. Therefore as a whole, model 2 is dominated by model 1.

Given this consideration, model 1 has been successfully implemented in STMicroelectronics Catania plant for scheduling a set of cleaning and of diffusion machines – that are bottleneck machines. The results are promising: the area observed an increase of 7% of the Overall Equipment Effectiveness.

We believe this work can bring interesting managerial implications worth considering also by a wider set of operations within the fab as well as different industries, which share the same set of constraints in the production environment and that are dealing with short term production planning decisions. Proposed models respond well to the specific problem exigencies and can be operationalized in a tool for short term production planning. Some degrees of freedom are however left to planners, assessing the current situation of the fab, and taking decisions consequently.

Despite the result of a predominance of model 1 over model 2 cannot be generalized, it can be noted that model 2 prioritize reduction of flow time rather than batch saturation. Therefore, by analysing the results in more general terms, for bottleneck machines or products with a loose due date, model 1, that results in large batches and efficient capacity utilisation, can be the preferred choice (Hopp and Spearman, 2008). For other contexts instead, e.g. non- bottleneck machines or products with a tight due date, flow time minimisation can be more relevant than batch saturation, thus making model 2 preferable.

This resembles the typical situation for simultaneous or parallel batching machines problems in which the trade-off is between effective capacity utilisation (for which batches with high saturation are required) and minimal wait-to-batch time (for which small batches are desirable) (Hopp and

Speearman, 2008). The relevance of batch saturation or flow time can depend on different contingent variables (e.g. customers' pressure on delivery time, the importance of efficiency). Moreover, the results are strongly affected by the value of the system parameters, such as: machine capacity, processing time, machine availability, time constraint duration. Therefore, further researches might be devoted to investigating the application of the two models in different contexts.

Our approach is not without limitations, given the heuristic nature of the algorithms presented, and the typical ad hoc nature of rules in these types of models, the derived schedule can be far from optimal (Hopp and Spearman, 2008). Despite this fact, we believe our model to be straightforward and easily understandable by our stakeholders.

References

Ahani, G, Asyabani, M. (2014) 'A tabu search algorithm for no-wait job shop scheduling problem', *International Journal of Operational Research*, Vol 19 No. 2, pp.246-258.

Akçali, E, Uzsoy, R D G, Hiscock, A L, Moser, T J, Teyner (2000) 'Alternative loading and dispatching policies for furnace operations in semiconductor manufacturing: a comparison by simulation', *Proceedings of the 32nd conference on Winter simulation. Society for Computer Simulation International*.

Attar, S F M, Mohammadi R, Tavakkoli-Moghaddam (2013) 'Hybrid flexible flowshop scheduling problem with unrelated parallel machines and limited waiting times', *The International Journal of Advanced Manufacturing Technology*, pp.1-17.

Azizoglu, M, & Webster, S. (2001) 'Scheduling a batch processing machine with incompatible job families.' *Computers & Industrial Engineering*, Vol. 39 No. 3, pp.325-335.

Bilyk, A, Mönch, L, Almeder, C. (2014) 'Scheduling jobs with ready times and precedence constraints on parallel batch machines using metaheuristics.' *Computers & Industrial Engineering*, 78, pp.175-185.

Brucker, P. 2007. Scheduling algorithms. 5th ed., Springer.

Caumond A, Lacomme, P, Tchernev, N. (2008) 'A memetic algorithm for the job-shop with time-lags', *Computers & Operations Research*, Vol. 35 No. 7, pp.2331-2356.

Cerekci, A, Banerjee, A. (2010) 'Dynamic control of the batch processor in a serial-batch processor system with mean tardiness performance', *International Journal of Production Research*, Vol. No. 48 (5), pp.1339-1359.

Chen, C L, Tang, T. I. (2012) 'Flexible flow line scheduling problems with re-entrant flows and queue-time constraints.', *Proceeding of the Automatic Control and Artificial Intelligence International Conference on IET (ACAI 2012).*

Cheng B, Cai, J, Yang S, Hu, X. (2014) 'Algorithms for scheduling incompatible job families on single batching machine with limited capacity', Computers & Industrial Engineering, Vol. 75, pp.116-120.

Chihyun J, Pabst, D, Myoungsoo H, Stehli, M, Rothe, M. (2014) 'An Effective Problem Decomposition Method for Scheduling of Diffusion Processes Based on Mixed Integer Linear Programming', *IEEE Transactions on Semiconductor Manufacturing*, Vol. 27 No.3, pp.357-363.

Cho, L, Park, H M, Ryan, J K, Sharkey, T C, Jung, C, Pabst, D. (2014). 'Production Scheduling with Queue-time Constraints: Alternative Formulations'. Paper presented at *IIE Annual Conference. Proceedings (p. 282). Institute of Industrial Engineers-Publisher, January, 2014*.

Cigolini, R, Pero, M, Rossi, T. (2011) 'An object-oriented simulation meta-model to analyse supply chain performance', *International Journal of Production Research*, Vol. 49 No.19, pp.5917-5941.

Cigolini, R, Perona, M, Portioli, A, Zambelli, T. (2002) 'A new dynamic look ahead scheduling procedure for batching machines', *Journal of scheduling*, Vol. 5 No.2, pp.185-204.

Dabbas, R M., Fowler J. W (2003) 'A new scheduling approach using combined dispatching criteria in wafer fabs', *IEEE Transactions on Semiconductor Manufacturing*, Vol.16 No.3, pp.501-510.

Fowler, J W, Hogg G L, Phillips D. T. (1992),'Control of multi product bulk service diffusion/oxidation processes', *IEEE Transactions*, Vol.24 No.4, pp.84-96.

Fowler, J W, Hogg G L, Phillips D. T. (2000), 'Control of multiproduct bulk server diffusion/oxidation processes. Part 2: multiple servers', *IEEE Transactions*, Vol. 32 No.2: pp.167-176.

Gicquel, C, Hege, L, Minoux, M, Van Canneyt, W. (2012) 'A discrete time exact solution approach for a complex hybrid flow-shop scheduling problem with limited-wait constraints', *Computers & Operations Research*, Vol. 39 No.3, pp.629-636.

Golmakani, H R, Namazi, A. (2014) 'An artificial immune algorithm for multiple-route job shop scheduling problem with preventive maintenance constraints', *International Journal of Operational Research*, Vol.19 No. 4, pp.457-478.

Gupta, A K, Sivakumar, A. I. (2006) 'Job shop scheduling techniques in semiconductor manufacturing', *The International Journal of Advanced Manufacturing Technology*, Vol. 27, No.11-12, pp. 1163-1169.

He, C, Leung, J Y T, Lee K, Pinedo, M. L. (2016), 'Scheduling a single machine with parallel batching to minimize makespan and total rejection cost', *Discrete Applied Mathematics*, Vol. 204, pp.150-163.

Hopp, W. J., and Spearman M. L. (2008). *Factory physics*, 2nd ed., Waveland Press.

Jia, W, Jiang, Z, Li, Y. (2015) 'Combined scheduling algorithm for re-entrant batch-processing machines in semiconductor wafer manufacturing', *International Journal of Production Research*, Vol. 53 No.6, pp.1866-1879

Jung, C, Pabst, D, Ham, M, Stehli, M, Rothe, M. (2014) 'An effective problem decomposition method for scheduling of diffusion processes based on mixed integer linear programming', *IEEE Transactions on Semiconductor Manufacturing*, Vol. 27 No.3, pp.357-363.

Kim, Y D, Kim, J G, Choi, B, Kim, H. (2001) 'Production scheduling in a semiconductor wafer fabrication facility producing multiple product types with distinct due dates', *IEEE Transactions on Robotics and Automation*, Vol. 17 No.5, pp. 589-598

Kim, Y D Lee, D H, Kim, J U, Roh, H-K (1998) 'A simulation study on lot release control, mask scheduling, and batch scheduling in semiconductor wafer fabrication facilities', *Journal of Manufacturing Systems*, Vol. 17 No.2, pp.107-117.

Kim, Y D, Joo, B J, Shin, J. H. (2009) 'Heuristics for a two-stage hybrid flowshop scheduling problem with ready times and a product-mix ratio constraint', *Journal of Heuristics*, Vol. 15 No.1, pp.19-42.

Li, T., Li, Y. (2007) 'Constructive backtracking heuristic for hybrid flowshop scheduling with limited waiting times', *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing (WiCom 2007).*

Mathirajan, M, Sivakumar, A. I. (2009) 'Scheduling a burn-in oven with non-agreeable release times and due dates', *International Journal of Operational Research*, Vol.4 No.3, pp.231-249.

Mathirajan, M, Sivakumar, A. I. (2006) 'A literature review, classification and simple meta-analysis on scheduling of batch processors in semiconductor', *The International Journal of Advanced Manufacturing Technology*, Vol. 29, No.9-10, pp.990-1001.

Mönch, L, Fowler, J W, Dauzère-Pérès, S, Mason, S J, Rose, O. (2011) 'A survey of problems, solution techniques, and future challenges in scheduling semiconductor manufacturing operations', *Journal of Scheduling*, Vol. 14 No.6, pp.583-599.

Narayanaswami, S, Rangaraj, N. (2012) 'A framework for dynamic dispatch decision-making applied in transportation scheduling', *International Journal of Operational Research*, Vol.15 No. 4, pp. 448-465.

Neale, J J, Duenyas, I. (2000) 'Control of manufacturing networks which contain a batch processing machine', *IEEE Transactions*, Vol.32, No. 11, pp.1027-1041.

Rossi, T, Pero, M. (2011) 'A simulation-based finite capacity MRP procedure not depending on lead time estimation', *International Journal of Operational Research*, Vol. 11, No.3, pp.237-261.

Saadaoui, S., Dhiaf, M M, Kamoun, H, Barqawi, B. (2014) 'Solving scheduling problems with earliness and tardiness penalties using priority rules and linear programming', *International Journal of Operational Research*, Vol. 20 No.4, pp.369-395.

Solomon, A, Fowler, J, Jensen P. H. (2002) 'The inclusion of future arrivals and downstream setups into wafer fabrication batch processing decisions', *Journal of Electronics Manufacturing*, Vol. 1 No.2, pp.149-159.

Su, L. H. (2003) 'A hybrid two-stage flowshop with limited waiting time constraints', *Computers & Industrial Engineering*, Vol. 44 No. 3, pp.409-424.

Sung, C S, Kim, Y. H. (2003) 'Minimizing due date related performance measures on two batch processing machines', *European Journal of Operational Research*, Vol. 147 No.3, pp.644-656.

Sung, C S, Yoon, S. H. (1997) 'Minimizing maximum completion time in a two-batch-processing-machine flowshop with dynamic arrivals allowed', *Engineering Optimization*+ A35, Vol. 28, No.3, pp.231-243.

Tajan, J B C, Sivakumar, A I, Gershwin, S. B. (2012) 'Evaluating the performance of look-ahead policies for upstream serial processor with downstream batch processor serving incompatible job families and finite buffer sizes', *International Journal of Operational Research*, Vol. 15, No.3, pp.260-289.

Tajan, J B C, Sivakumar, A I, Gershwin, S. B. (2013) 'Modelling and analysis of a multi-stage system involving batch processors with incompatible job families', *International Journal of Operational Research*, Vol. 17 No.4, pp.449-482

Tu, Y M, Chen, H. N. (2008) '*Shop-Floor Control Model in Batch Processes of Wafer Fabrication with Time Constraints*', Paper presented at: POMS 19th Annual Conference, La Jolla, California.

Vepsalainen, P J, Morton T. E. (1987)' Priority rules for job shops with weighted tardiness costs', *Management science*, Vol. 33 No.8 pp.1035-1047.

Yang, D, W, Jiang, J Z , Li, Y. (2013) 'Variable Time Windows-Based Three-Phase Combined Algorithm for On-Line Batch Processing Machine Scheduling with Limited Waiting Time

Constraints‘, Paper presented at the Institute of Industrial Engineers Asian Conference, pp. 559-567. Springer Singapore.

Yugma, C, Dauzère-Pérès, T E, Artigues, C, Derreumaux, A, Sibille, O. (2012) ‘A batching and scheduling algorithm for the diffusion area in semiconductor manufacturing’, *International Journal of Production Research*, Vol. 50 No.8, pp.2118-2132.

Tables

Reference	Time constraints	Problem addressed according to the notation by Baker (2007)
Cigolini et al. 2002	N	$PM\ G\vert\ p_j, p-batch\vert \min(FT)$
Kim et al. 1998	N	$PM\ J\vert\ p_j, p-batch, incompatible\ batch\vert \min(T)$
Akçali et al. 2000	N	$o\vert\ p_j, p-batch, incompatible-batch\vert \min(C)$
Neale and Duenyas 2000	N	$PM\ J\vert\ p_j, p-batch,\vert \min(C)$
Cerekci and Banerjee 2010	N	$PM\ J\vert\ p_j, d_j, p-batch, incompatible-batch\vert \min(T)$
Sung and Kim 2003	N	$o\ J\vert\ d_j, p-batch,\vert \min(T)$
Sung and Yoon 1997	N	$o\ J\vert\ p, p-batch,\vert \min(C)$
Gupta and Sivakumar 2006	N	$o\vert\ p, d_j, p-batch,\vert \min(T)\ , \min(\max(T)\ ), \min\overline{(T)}$
Solomon et al. 2002	N	$o\vert\ p_j, d_j, p-batch, incompatible-batch\ \vert \min(C)$
Mönch et al. 2005	N	$PM\vert\ p_j, d_j, p-batch, incompatible-batch\ \vert \min(T*w)$
Azizoglu and Webster 2001	N	$o\vert\ p_j, p-batch, incompatible-batch\ \vert \min(C*w)$
Fowler et al. 2000	N	$PM\vert\ p_j, p-batch, incompatible-batch\ \vert \min(C)$
Vepsalainen and Morton 1987	N	$o\vert\ p_j, d_j\vert \min(T_j*w_j)$
Kim et al. 2001	N	$o\vert\ p_j, p-batch, incompatible-batch\vert \min(C)$
Yugma et al. 2012	Y	$R\ J\vert\ p_j, d_j, p-batch, incompatible-batch, T_{lags}^{min}\vert \min(\mathrm{C}), \min(Number\ of\ moves)$
Su 2003	Y	$o\ J\vert\ p_j, p-batch, T_{lags}^{min}\vert \min(C)$
Attar et al. 2013	Y	$PM\ J\vert\ p_j, T_{lags}^{min}\vert \min(C)$
Chen and Tang 2012	Y	$PM\ J\vert\ p_j, T_{lags}^{min}\vert \min(T)$
Gicquel et al. 2012	Y	$FF\ PM\ \vert size, T_{lags}^{min}\vert \min(T*w)$
Kim et al. 2009	Y	$R\ J\vert\ p_j, p-batch, incompatible-batch, T_{lags}^{min}\vert \min(\mathrm{C})$
Li and Li 2007	Y	$R\ J\vert\ p_j, p-batch, T_{lags}^{min}\vert \min(\mathrm{C})$
Caumond et al. 2008	Y	$J\vert\ p_j, p-batch, T_{lags}^{min}\vert \min(\mathrm{C})$
Yang et al. 2013	Y	$R\ J\vert\ p_j, d_j, p-batch, incompatible-batch\vert \min(\mathrm{T}*\mathrm{w})$
Mathirajan and Sivakumar 2009	N	$1\vert\ p_j, p-batch, rj; dj\ \vert \min(C)$
Ahani and Asyabani 2014	Y	$R\ J\vert\ p_j, T_{lags}^{min}\vert \min(C)$
Tajan et al. 2012	Y	$R\ J\vert\ p_j, p-batch, incompatible-batch\vert \min(\mathrm{C})$

Tajan et al. 2013	Y	$J, ni = 2 \mid , p_j, p - batch, s - batch, incompatible - batch \mid \min(C)$
Bilyk et al. 2014	Y	$PM \mid p_j, p - batch, incompatible - batch \mid \min(T * w)$
Golmakani and Namazi 2014	N	$R\ J, Ai \mid p_j \mid \min(C)$
He et al. 2016	N	$1 \mid p_j, p - batch, rj = r \mid \min(C)$, min (RJCT)
Cho et al., 2014	Y	$J \mid T_{lags}^{min} \mid \min(C * w)$

Table1 – Classification of analysed scheduling problems adapted from Brucker, 2007

Key: PM= identical parallel machine; R= unrelated parallel machine; o=open job shop with no precedence relation; J= job shop with relations between operations; FF=flexible flow shop;1=single machine system; Ai=different availability of machine i; ni=number of machines i; G=general shop; pj=different process time; dj= different due dates; rj= different release date (where rj is set to "r" it means the problem considers identical release dates; p-batch=parallel batching machine (the process time of the batch is the longest between the lots that form the batch); s-batch= serial batching machine, incompatible batch=there are lots that cannot be batched together; T_{lags}^{min} =presence of time constraints between operations; size=fixed size batches; C=makespan; T=tardiness; $\bar{T}$ =average tardiness; w= generic weight assigned to a performance; RJCT=rejection costs.

Model	Batching	Dispatching rule at cleaning operation	Dispatching rule at diffusion operation	Time constraint management between cleaning and diffusion	Time constraint management between diffusion and diffusion
Model 1	Constant: the same batch size and composition in all operations	ATC from Vepsalainen and Morton (1987) / BATC from Cerekci and Banerjee (2010)	Fist In First Out	Look ahead	Look ahead
Model 2	Variable: batch size and composition might be different depending on the operation	ATC	Priority indexes that consider both batch saturation and due date of lots (adapted from Tu and Chen (2008))	Booking	Look back

Table 2. Comparison between model 1 and 2

	Priority index	Wtmax	ΔT
Scenario 1	ATC	0	0
Scenario 2	ATC	0	10
Scenario 3	ATC	0	2
Scenario 4	ATC	0	4
Scenario 5	ATC	0	6
Scenario 6	ATC	0	8
Scenario 7	ATC	12	0
Scenario 8	ATC	12	10
Scenario 9	ATC	12	2
Scenario 10	ATC	12	4
Scenario 11	ATC	12	6
Scenario 12	ATC	12	8
Scenario 13	BATC	0	0
Scenario 14	BATC	0	10
Scenario 15	BATC	0	2
Scenario 16	BATC	0	4
Scenario 17	BATC	0	6
Scenario 18	BATC	0	8
Scenario 19	BATC	12	0
Scenario 20	BATC	12	10
Scenario 21	BATC	12	2
Scenario 22	BATC	12	4
Scenario 23	BATC	12	6
Scenario 24	BATC	12	8

Table 3. Scenarios adopted for model 1 experimental phase

	λ	ρ	ρD1
Scenario1	0	1	0
Scenario2	0	1	0,25
Scenario3	0	1	0,5
Scenario4	0	1	0,75
Scenario5	0	1	1
Scenario6	0	1	0
Scenario7	0,25	0,75	0,25
Scenario8	0,25	0,75	0,5
Scenario9	0,25	0,75	0,75
Scenario10	0,25	0,75	1
Scenario11	0,5	0,5	0
Scenario12	0,5	0,5	0,25
Scenario13	0,5	0,5	0,5
Scenario14	0,5	0,5	0,75
Scenario15	0,5	0,5	1
Scenario16	0,75	0,25	0
Scenario17	0,75	0,25	0,25
Scenario18	0,75	0,25	0,5
Scenario19	0,75	0,25	0,75
Scenario20	0,75	0,25	1
Scenario21	1	0	0
Scenario22	1	0	0,25
Scenario23	1	0	0,5
Scenario24	1	0	0,25
Scenario25	1	0	1

Table 4. Scenarios adopted for model 2 experimental phase

			Model 1		
			Decision variable		
			ΔT	**Wt_{max}**	**Priority index**
Performance	**Daily batch saturation**	Welch one-way ANOVA	Verified impact (p-value = 0)	Verified impact (p-value = 0)	Verified impact (p-value = 0)
		Games-Howell's test results	The longer the ΔT, the higher the Daily batch saturation Scenarios with ΔT = 10 have the best performance	The longer the Wtmax, the higher the Daily batch saturation Scenarios with Wtmax = 12 have the best performance	BATC is associated with the higher Daily batch saturation
	Daily recleaned lots	One-way ANOVA	Verified impact (p-value = 0)	Impact not verified (p-value = 0,29)	Impact not verified (p-value = 0,31)
		Tuckey's test results	The longer the ΔT, the lower the Daily recleaned lots Scenarios with ΔT = 10 have the best performance	Not Applicable	Not Applicable
	Flow time	Welch one-way ANOVA	Verified impact (p-value = 0)	Verified impact (p-value = 0)	Verified impact (p-value = 0)
		Games-Howell's test results	The longer the ΔT, the higher the flow time Scenarios with ΔT = 0 have the best performance	The longer the Wtmax, the longer the Flow time Scenarios with Wtmax = 0 have the best performance	ATC is associated with the lowest flow time

Table 5. Tests' results for model 2

Figures

Figure 1. System description

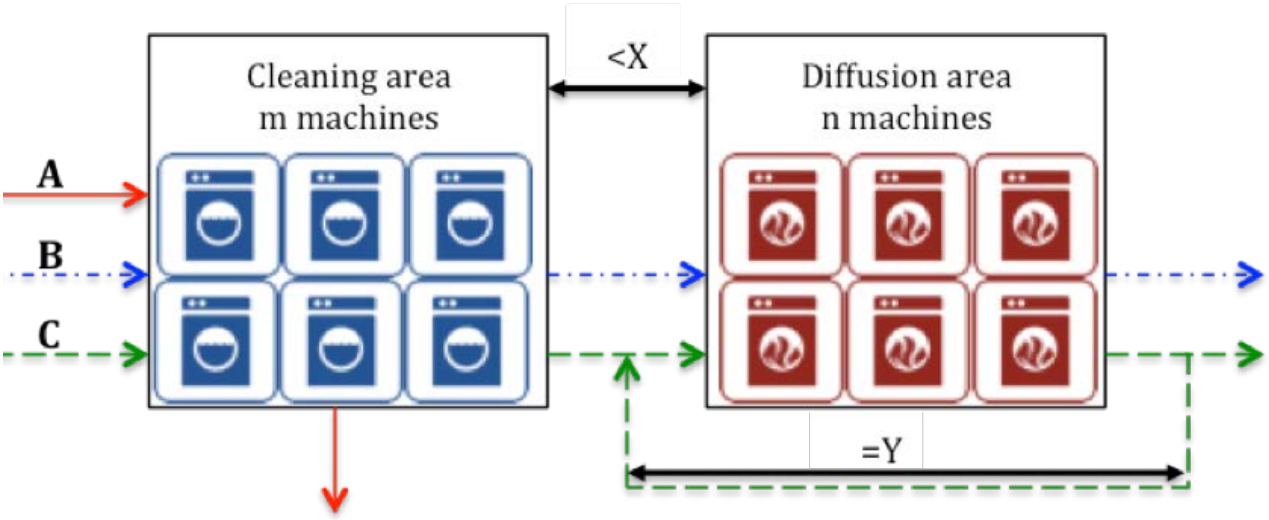


Figure 2. Priority index calculation for ATC

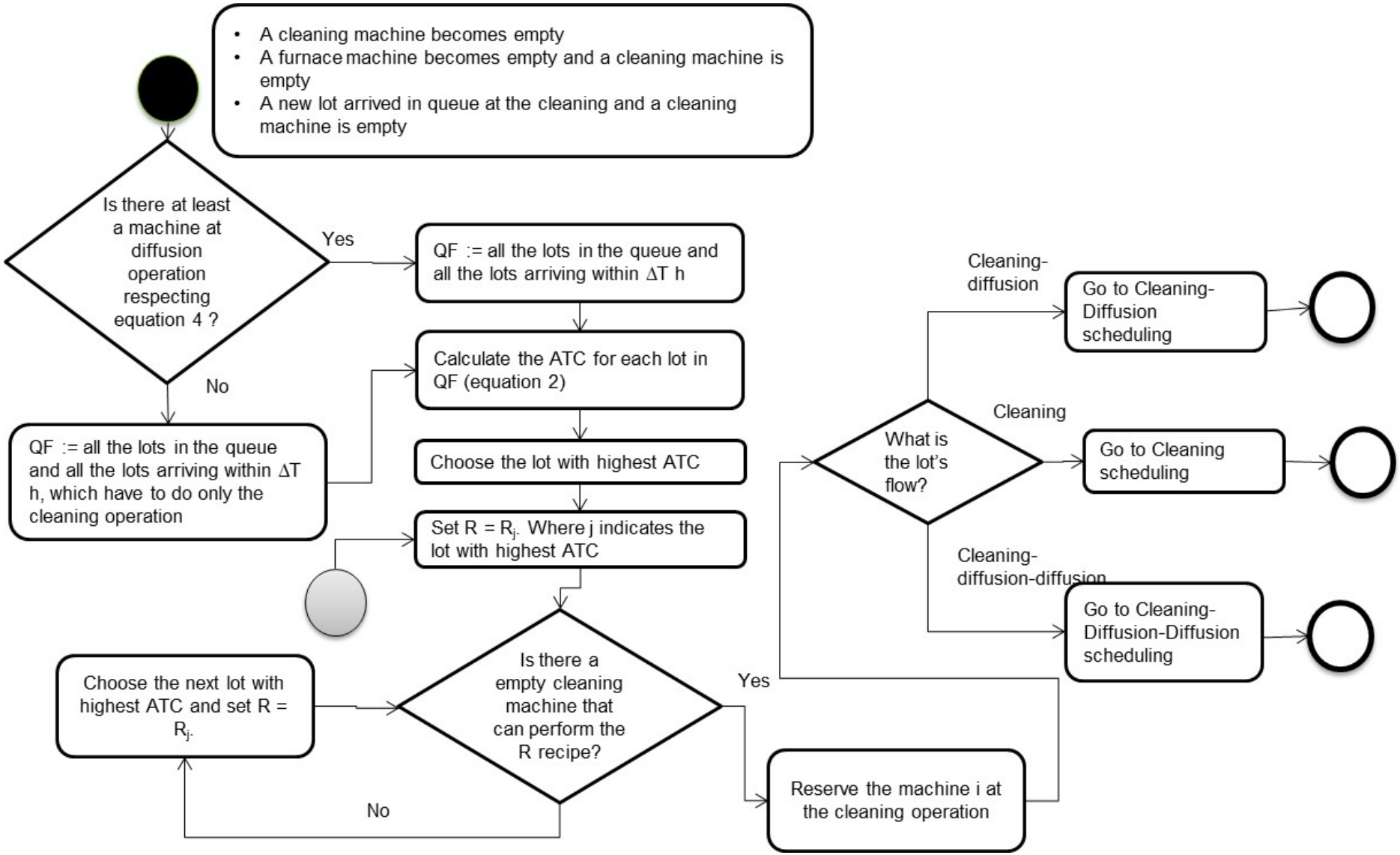


Figure 3. Cleaning scheduling

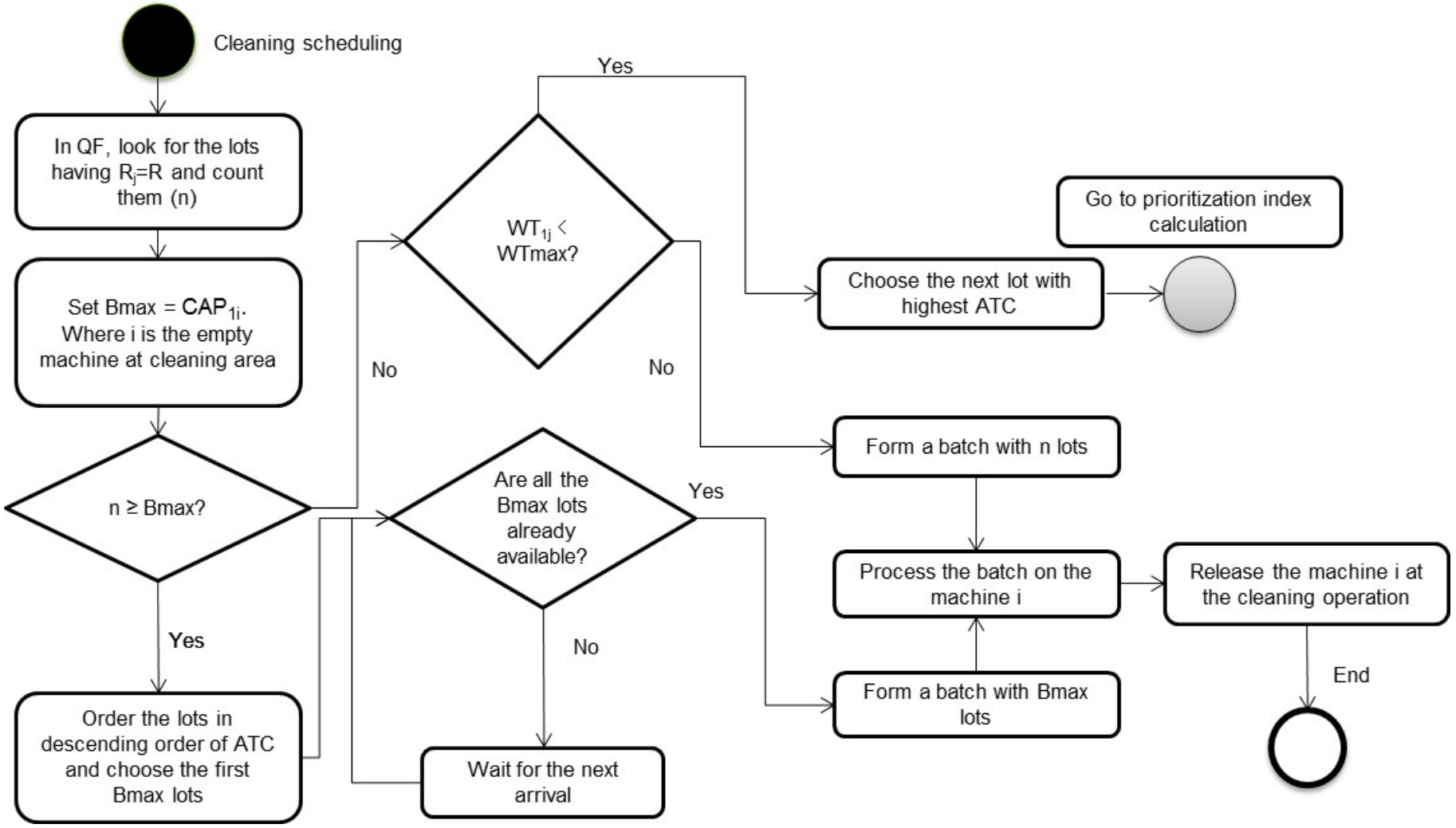


Figure 4. Cleaning-diffusion scheduling

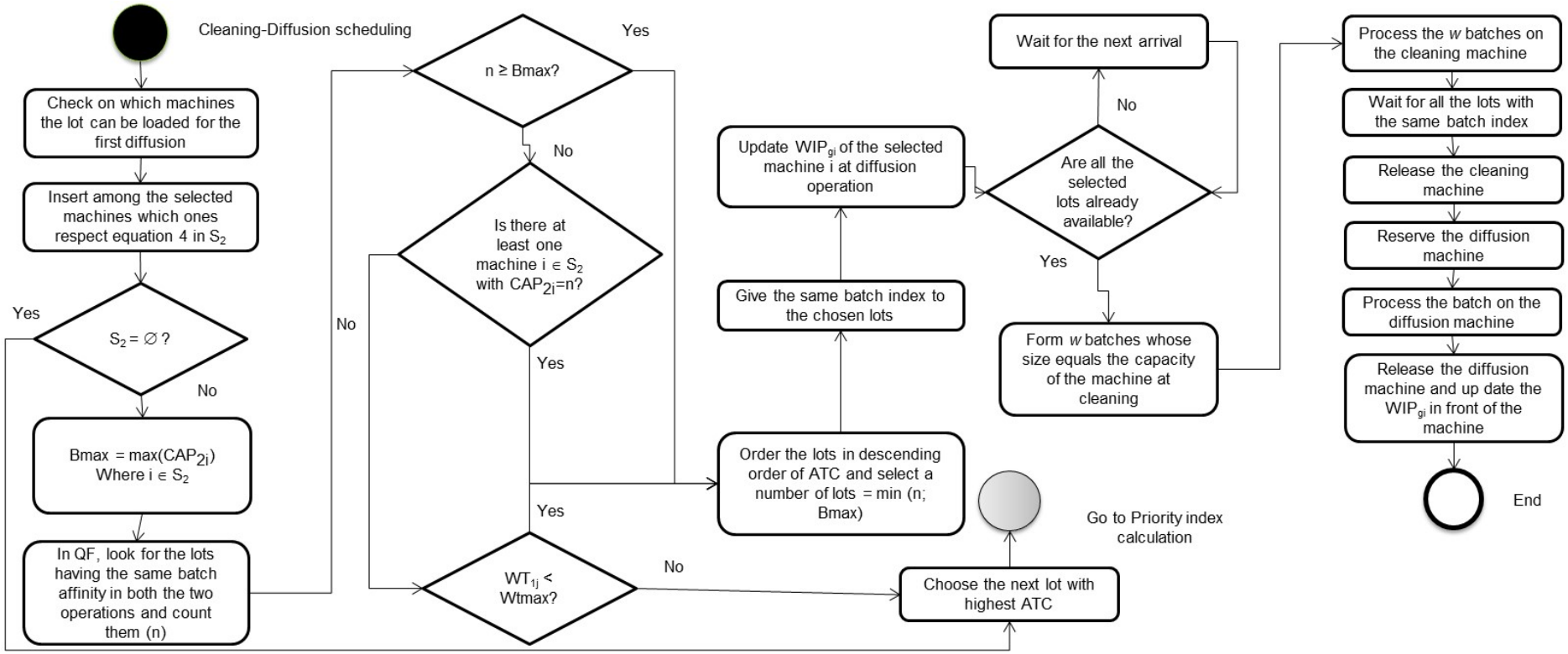


Figure 5. Cleaning-diffusion-diffusion scheduling

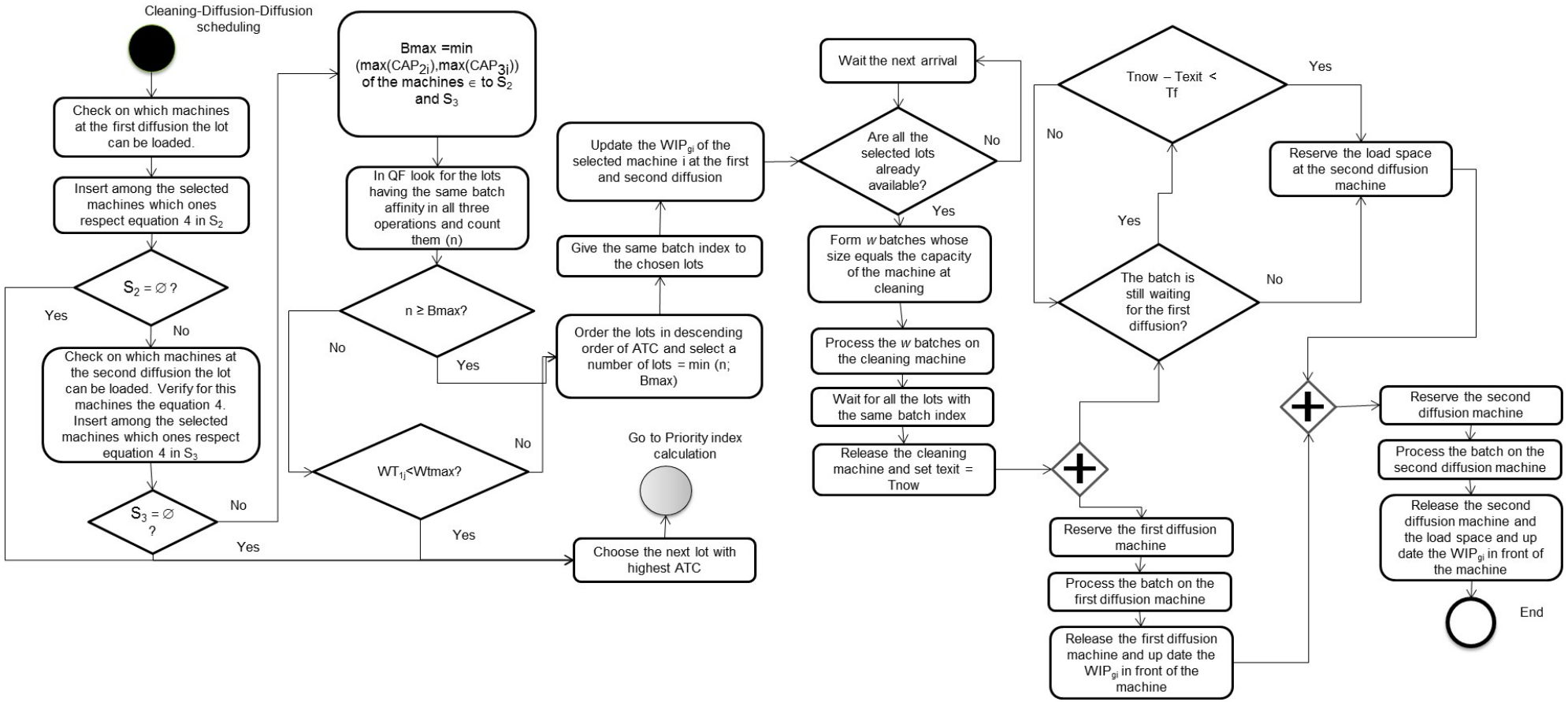


Figure 6. Sub-algorithm for creating the free-pass queue

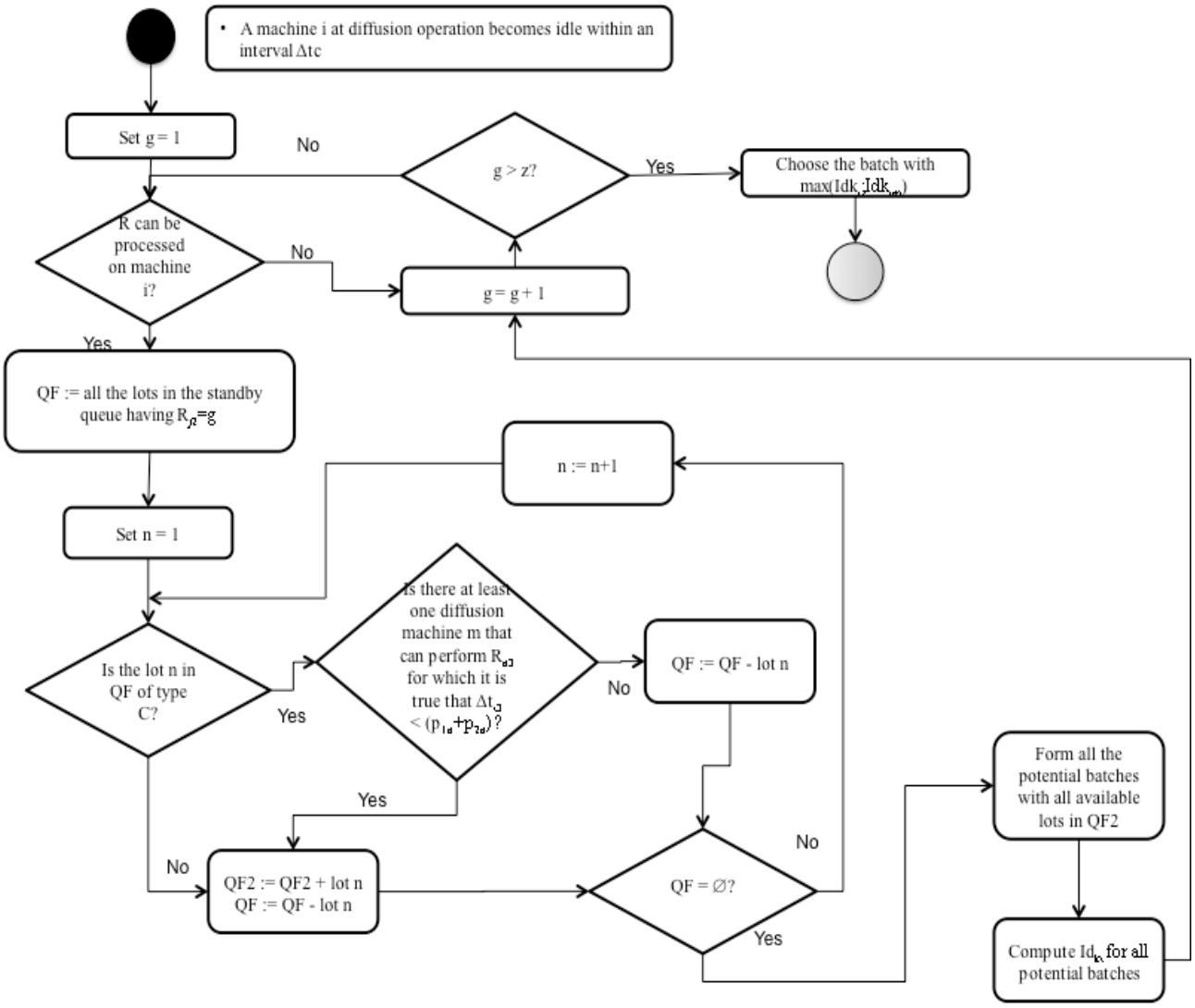
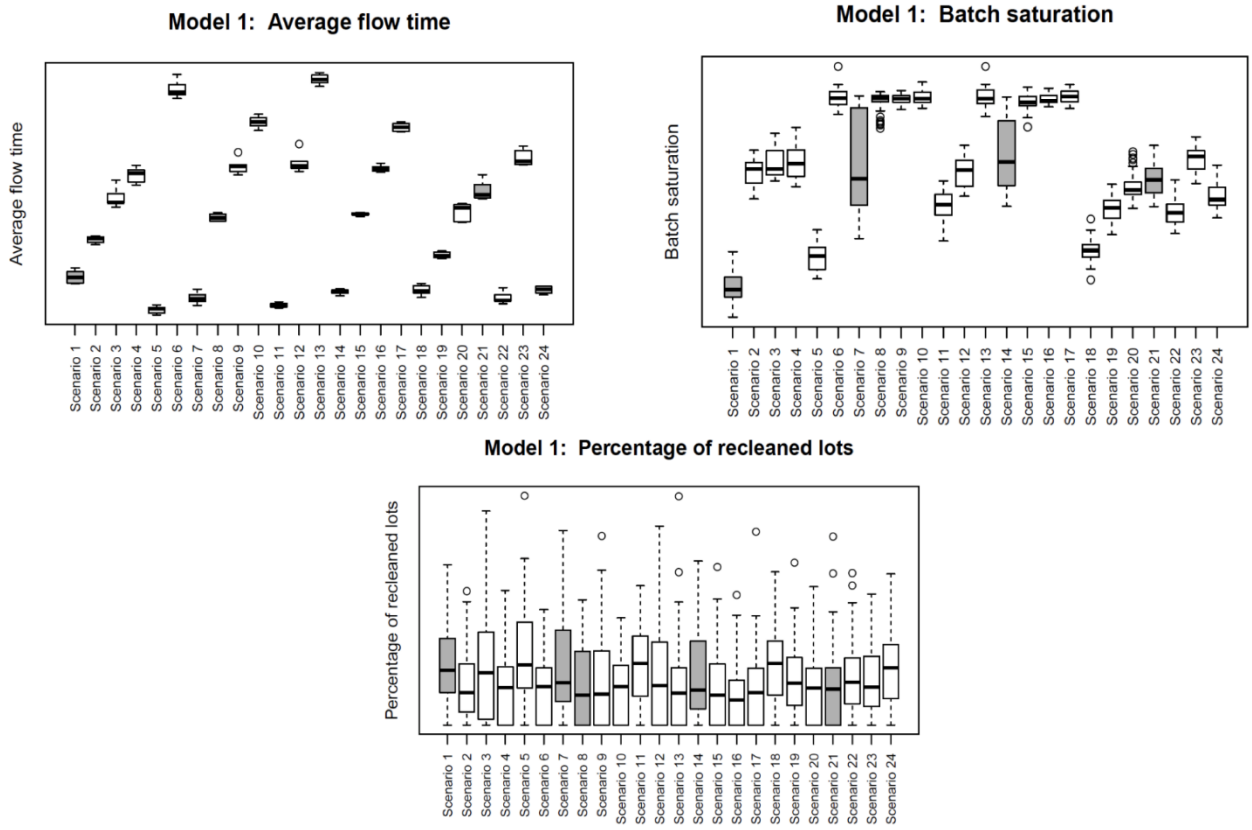
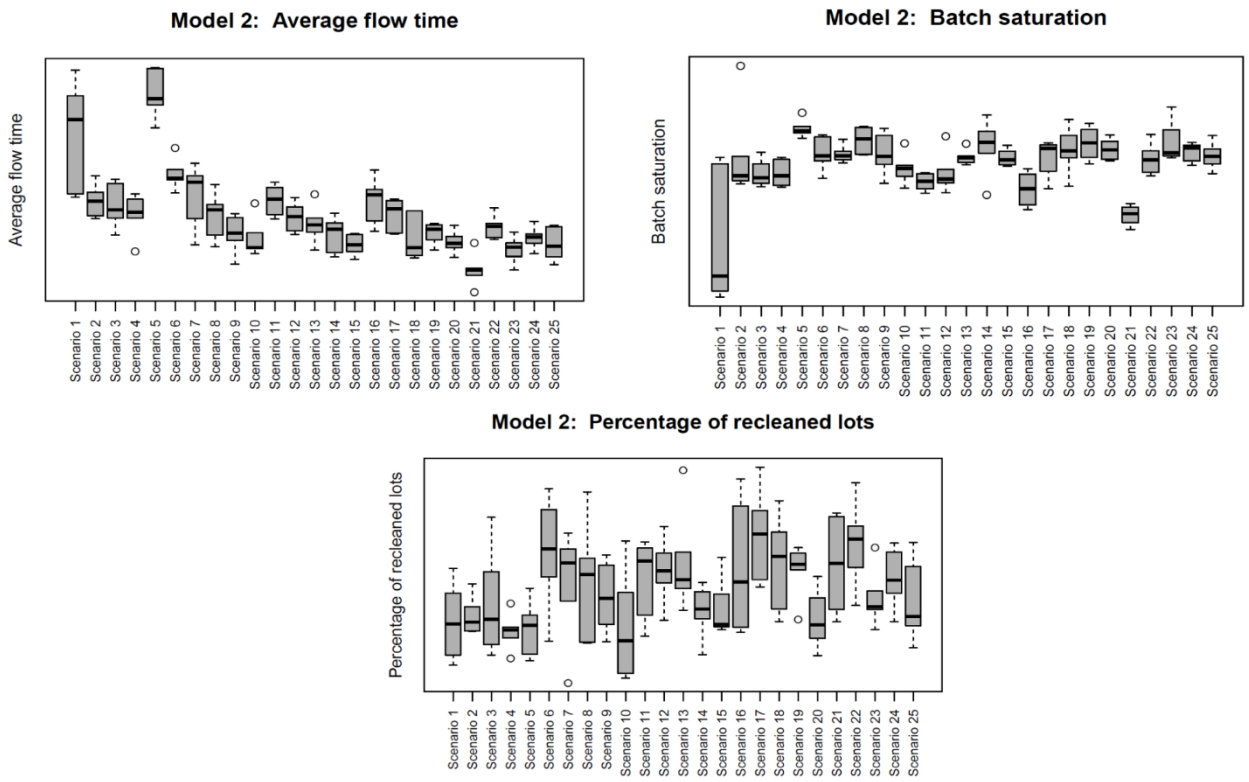


Figure 7. Results of the simulation campaign with model 1



Key: when a box plot is coloured in grey = there is at least one scrapped lot is one repetition of the scenario

Figure 8. Results of the simulation campaign with model 2



Key: when a box plot is coloured in grey = there is at least one scrapped lot is one repetition of the scenario

Figure 9. Qualitative comparison between model 1 and model 2 considering performance

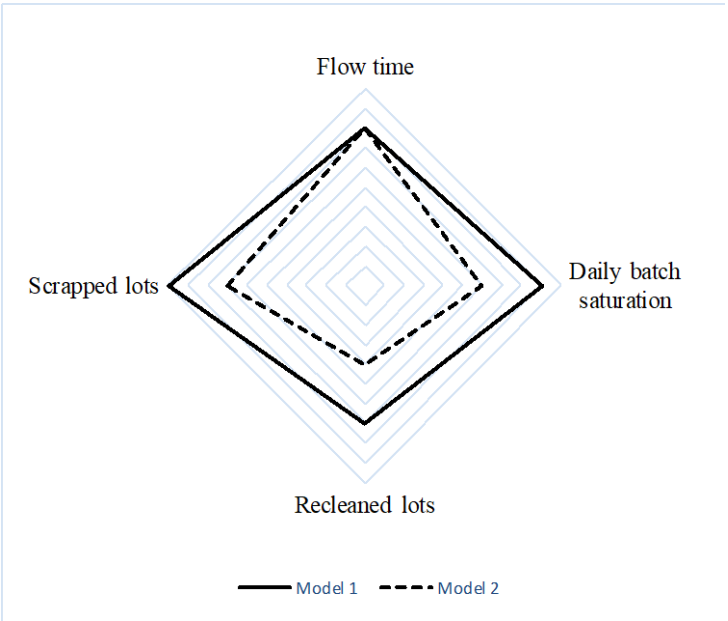


Figure 10. Impact of decision variables on performance