Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

REAL-TIME METROLOGY CAPACITY OPTIMIZATION IN SEMICONDUCTOR MANUFACTURING

Mathis Martin[1], Stéphane Dauzère-Pérès[1], Claude Yugma[1], Aymen Mili[2], and Renaud Roussel[2]

[1]Mines Saint-Étienne, Univ. Clermont Auvergne, Gardanne, FRANCE
[2]STMicroelectronics, Crolles, FRANCE

ABSTRACT

This paper addresses a capacity management problem in semiconductor metrology, where lots are typically sent to measurement under static sampling rules. Such rigid strategies often cause bottlenecks when unexpected events occur, delaying production. To address this, we propose a corrective approach that dynamically selects lots to skip, i.e., not measured, while prioritizing the most critical ones and ensuring that metrology tools respect capacity limits. The method combines a skipping algorithm with the Iterated Min-Max (IMM) workload balancing procedure, which ensures a fair workload distribution and helps identify the most critical tools. Several performance indicators are introduced to evaluate the efficiency of this approach compared to a classical balancing strategy. Computational experiments with industrial data demonstrate that integrating IMM improves lot selection, reduces the number of skipped lots, and preserves measurement for the highest priority ones while better satisfying capacity constraints.

1 INTRODUCTION

Wafer manufacturing is a complex and highly dynamic environment with a continuously changing portfolio of products that must be produced at the lowest possible cost. Moreover, in particular in applications such as automotive and robotics, product quality and reliability are critical. To ensure both, one important operation is metrology (measurement), which consists of measuring wafers at specific steps of the process flow to detect potential defective production machines and prevent the propagation of quality issues. Various types of inspections are required, which depend on the production machines being controlled. Each inspection type relies on a limited number of expensive and complex measurement tools. In addition, inspections slow down production and do not directly add value to products. These reasons motivate semiconductor manufacturers to carefully manage measurement resources. In particular, not all lots can be sent to metrology, and thus, only a limited number of lots are sampled to be measured.

Sampling strategies used to select which lots should be measured can be broadly classified into three categories Nduhura-Munga et al. (2013): Static, adaptive, and dynamic. Static sampling consists of setting rates that specify the frequency at which lots are sampled to be measured (see for example Perez (2017)), e.g., 1 lot out of 3. While static sampling is simple to implement and widely used in industry, it does not adapt well to operational disturbances such as metrology tool breakdowns or production surges, often resulting in bottlenecks in the metrology area. Adaptive Mouli and Scott (2007) and dynamic Le Quéré et al. (2020) sampling strategies have been proposed to address this limitation, but their practical implementation remains challenging, especially in large-scale industrial environments.

This paper considers the management of capacity in a metrology area that can be overloaded with too many lots following an operational disturbance. A key consideration is that lot measurements must be carried out quickly to stop defective production machines as soon as possible, but also to avoid lots waiting in metrology instead of continuing their manufacturing routes. Therefore, it may be necessary to skip lots, i.e., to reduce the workload on metrology tools by not measuring some sampled lots, to both accelerate the detection of defective production machines and reduce the cycle times of lots.

979-8-3315-8726-0/25/$31.00 ©2025 IEEE

In the considered real-time capacity optimization problem, the objective is to ensure timely inspections without overloading the metrology tools. We proposed an iterative procedure that relies on the integration of two components: A workload balancing procedure to check if the metrology area is overloaded, and a skipping strategy to decide which lots to skip if the metrology area is overloaded. Workload balancing is performed either with a classical approach, where the largest workload of any metrology tool is minimized, or with the Iterated Min-Max (IMM) procedure proposed in Christ et al. (2019). The skipping strategy relies on a priority index derived from the "Wafers at Risk" (W@R) indicator (see Dauzère-Pérès et al. (2010)). The W@R of a production machine corresponds to the number of wafers processed on the machine since the last lot measured on a metrology tool and processed on the machine.

Although job rejection has been studied in the scheduling literature (see for instance Shabtay and Gerstl (2024), Geng et al. (2023) and Freud and Mosheiov (2021)), the goal is to optimize scheduling objectives or reduce costs. In contrast, we do not consider detailed scheduling decisions, but want to ensure that there is globally enough capacity to measure the lots currently in a metrology area. One motivation is that scheduling tools are not always available in metrology areas, where simple dispatching rules are often used. Moreover, skipping decisions in our problem rely on a risk indicator not considered in the scheduling literature to our knowledge.

The problem is described in detail and formalized in Section 2. Then, Section 3 recalls the IMM procedure, which is used in the iterated skipping procedure proposed in Section 4. Computational experiments are reported and discussed in Section 5. Some conclusions and perspectives can be found in Section 6.

2 PROBLEM DEFINITION AND MODELING

Section 2.1 describes the problem and provides an overview of the proposed solution approach, which consists of two phases. The first phase evaluates the current workload balance across the metrology tools. The second phase only required when not all lots can be measured, identifies which lots should be skipped based on their priorities. Then, Section 2.2 introduces a classical workload balancing model along with the IMM (Iterated Min-Max) procedure proposed in (Christ et al. 2019), and discusses the benefits of both approaches. Finally, Section 2.3 explains how the priorities of lots for measurement are modeled. These priorities are used in the skipping phase of the iterated skipping procedure proposed in Section 4.

2.1 Problem Definition

Let $\mathscr{N}$ be a set of sampled lots waiting for measurement in the metrology queue and $\mathscr{M}$ be the set of available metrology tools with different qualifications Q (the metrology tools are not qualified, i.e. eligible, to measure all recipes) and measurement speeds. More precisely, $Q_{l,m} = 1$ if lot $l \in \mathscr{N}$ is qualified to be measured on metrology tool $m \in \mathscr{M}$, and the measurement time of lot l on m is denoted $p_{l,m}$.

The problem is to estimate whether the lots can be assigned to metrology tools without any metrology tool exceeding a maximum capacity W. If this is not the case, critical tools have to be identified to skip lots from these tools until no metrology tool has a workload that is strictly above W. One main challenge of this problem is that the metrology tools are not qualified to measure all lot. Moreover, the measurement time of a lot on a qualified metrology tool can be different due to different factors, such as the age of the tool. Another challenge is to decide which criteria are relevant to decide which lots to skip. "In contrast to global sampling plans (see, for example, (Dilosi et al. 2022)), the objective of this paper is to introduce a fast and accurate corrective approach that operates on top of an existing sampling plan. Moreover, unlike other skipping algorithms in the literature (e.g., (Le Quéré et al. 2019)), the proposed method is combined with a workload balancing algorithm (see Section 4), enabling a more precise selection of lots while preventing excessive skipping.

2.2 Workload Balancing on Metrology Tools

We first need to evaluate if the lots in $\mathscr{N}$ can be assigned to the metrology tools without exceeding the maximum capacity W.

Let $X_{l,m} \in [0,1]$ be the fraction of lot l that is allocated to tool $m \in \mathscr{M}$. The workload of tool m is defined as:

$$W_m = \sum_{l \in \mathscr{N};\ Q_{l,m}=1} p_{l,m}\ X_{l,m}$$

We consider the continuous problem, where the measure of a lot can be split on several tools. From an industrial point of view, this hypothesis is accurate enough to estimate if a set of lots can be assigned to the metrology tools without exceeding a given maximum capacity. The first classical approach is to solve the problem of minimizing the maximum workload across all metrology tools while ensuring that all lots are assigned using the linear program $P(\mathscr{N},\mathscr{M})$ below.

$$\min S \tag{1}$$

$$W_m \leqslant S \qquad \forall m \in \mathscr{M} \tag{2}$$

$$W_m = \sum_{l \in \mathscr{N};\ Q_{l,m}=1} p_{l,m}\ X_{l,m} \qquad \forall m \in \mathscr{M} \tag{3}$$

$$\sum_{m \in \mathscr{M};\ Q_{l,m}=1} X_{l,m} = 1 \qquad \forall l \in \mathscr{N} \tag{4}$$

$$X_{l,m} \leqslant Q_{l,m} \qquad \forall l \in \mathscr{N}, \forall m \in \mathscr{M} \tag{5}$$

$$X_{l,m} \in [0,1] \qquad \forall l \in \mathscr{N}, \forall m \in \mathscr{M} \tag{6}$$

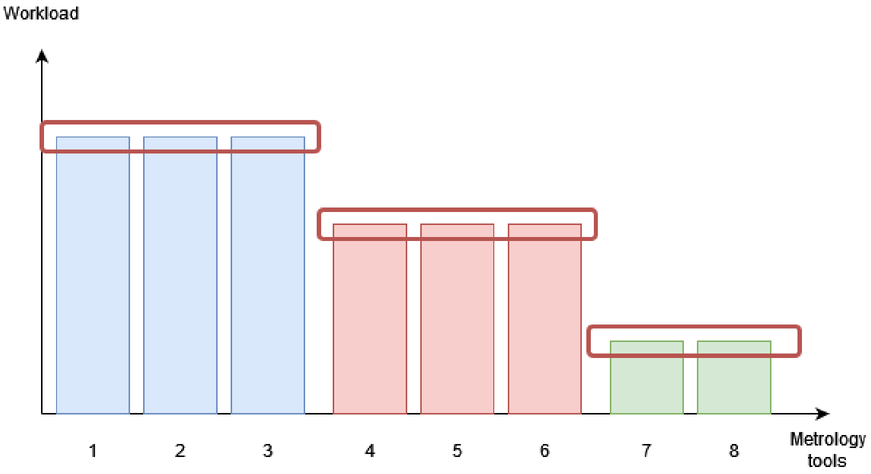
Constraints (2) and (3) ensure that variable S takes the value of the largest workload of any tool, and Constraints (4) that each lot is fully assigned to metrology tools. Constraints (5) guarantee that a lot can only be assigned to metrology tools on which the lot is qualified.

If, after solving $P(\mathscr{N},\mathscr{M})$, the optimal value S^* is lower than W, then all lots can be measured. Otherwise, lots must be selected to be skipped to decrease S^* below W. One important problem with $P(\mathscr{N},\mathscr{M})$, as shown in (Christ, Dauzere-Peres, and Lepelletier 2019), is that a metrology tool m such that $W_m = S^*$ is not always critical, i.e. it might be possible to reduce the workload of m without increasing S^*. And it is important to only reduce the workload of critical tools. However, we know that there is at least one tool m such that $W_m = S^*$ which is critical.

The IMM (Iterated Min-Max) workload balancing procedure, proposed in (Christ et al. 2019) and recalled in Section 3, is well suited to this challenge. In an optimal solution determined by the IMM procedure, the critical tools are clearly defined, and skipping lots on these tools with the highest workloads more effectively reduces the maximum workload. From an industrial point of view, correctly defining the critical tools is crucial.

Figure 1 shows the workload balance after the IMM procedure on one industrial instances (set of instances I1 presented in Section 5). The figure illustrates the characteristics of the solution determined by the IMM procedure, where the tools are divided into different groups, each with the same workload. Moreover, each recipe belongs to a single group of metrology tools with the same workload. Hence, the IMM procedure determines independent groups of tools and recipes.

Figure 1: Workload balance determined by the IMM procedure on one industrial instance.



2.3 Risk Modeling

The sampled lots are measured to verify that production processes are performed correctly, i.e. that the production machines on which the lots were processed are not defective. If the measurement of a lot shows a default, then the production machine on which the lot was processed is stopped, and corrective actions are taken. The impacted lots are either scrapped or reworked. A risk indicator widely used in semiconductor manufacturing is the "Wafer at Risk" (W@R), which provides a real-time estimate of the number of wafers processed by a production machine since the last confirmed non-defective lot.Using the W@R indicator allows us to model the measurement priority of the lots. Let introduce the following notations:

- $W@R_r$: Number of wafers at risk on production machine r,
- $W@R_r^{new}(l)$: New number of wafers at risk on production machine r if lot l is measured ($W@R_r^{new}(l) \leqslant W@R_r$),
- I_r: Maximum number of wafers at risk before production machine r is stopped,
- $\alpha > 1$: Real number that is used to increase the risk not linearly with the number of products at risk,

The current risk level on production machine r is modeled as the non-linear function $(\frac{W@R_r}{I_r})^\alpha$. Hence, the risk reduction on risk r when measuring lot l is defined by:

$$w_{r,l} = \left(\frac{W@R_r}{I_r}\right)^\alpha - \left(\frac{W@R_r^{new}(l)}{I_r}\right)^\alpha \tag{7}$$

In the context of this paper, each lot has been processed on a unique production machine i.e. if lot l has been processed on production machine r, $w_{r',l} = 0 \quad \forall r' \neq r$. Therefore we can reformulate into w_l for each lot l. In practice, w is computed considering no lot is defective because it is used decide which lot to sample. Indeed, the GSI (Global Sampling Indicator), introduced in (Dauzère-Pérès, Rouveyrol, Yugma, and Vialletelle 2010), is based on the value w and is widely used for dynamic sampling strategies in semiconductor manufacturing. A large value of w_l indicates that lot l is more likely to be defective. Therefore, lots with the highest w values should be prioritized for measurement.

Following the value of w, the sampled lots are categorized into four categories: (1) Low priority, (2) Medium priority, (3) High priority and (4) Ultra-high priority.

Low priority lots are typically lots that have been processed on a production machine quickly after a confirmed non-defective lot. These lots tend to be the first ones that are skipped. On the opposite, ultra-high priority lots are mandatory lots that cannot be skipped for other reasons than the fact that they reduce the risk on production machines.

3 ITERATED MIN-MAX PROCEDURE (IMM)

The IMM workload balancing procedure proposed in (Christ et al. 2019) assigns lots to qualified metrology tools with the goal of *fairly* balancing the workload. The IMM procedure thus helps to identify the critical tools and to support effective decision making in related procedures such as lot skipping. Let us recall the IMM procedure in the following by first rewriting the linear program $P(\mathcal{N},\mathcal{M})$ as the linear program $P(\mathcal{N},\mathcal{M},\mathcal{B},\gamma)$, where $\mathcal{B}$ is a subset of the metrology tools ($\mathcal{B} \subseteq \mathcal{M}$) and $\gamma = \{\gamma_1,\dots,\gamma_{|\mathcal{M}|}\}$ is a given workload vector (γ_m is the workload assigned to m).

$$\min S \tag{8}$$

$$W_m \leqslant S \qquad \forall m \in \mathcal{M}\backslash\mathcal{B} \tag{9}$$

$$W_m = \gamma_m \qquad \forall m \in \mathcal{B} \tag{10}$$

$$W_m = \sum_{l\in\mathcal{N};\ Q_{l,m}=1} p_{l,m}\, X_{l,m} \qquad \forall m \in \mathcal{M} \tag{11}$$

$$\sum_{m\in\mathcal{M};\ Q_{l,m}=1} X_{l,m} = 1 \qquad \forall l \in \mathcal{N} \tag{12}$$

$$X_{l,m} \leqslant Q_{l,m} \qquad \forall l \in \mathcal{N}, \forall m \in \mathcal{M} \tag{13}$$

$$X_{l,m} \in [0,1] \tag{14}$$

The objective is to minimize the maximum workload S of a specific subset of metrology tools ($\mathcal{M}\backslash\mathcal{B}$), determined through Constraints (9), while fixing the workloads of the other tools ($\mathcal{B}$) through Constraints (10). Note that Constraints (11) through (13) are equivalent to Constraints (3) through (5).

In the IMM procedure, set $\mathcal{B}$ is initially empty, and the maximum workload across all tools is minimized. The workload vector γ is constructed iteratively. Let λ_m be defined as the dual variable corresponding to Constraint (9) for tool m in $\mathcal{M}\backslash\mathcal{B}$. Algorithm 1 from (Christ, Dauzere-Peres, and Lepelletier 2019) provides an overview of the IMM procedure.

Algorithm 1 IMM($\mathcal{N},\mathcal{M}$)

```
Inputs: $\mathcal{B} = \emptyset$, $\gamma_m = 0 \ \ \forall m \in \mathcal{M}$, $S^* = 0$
while $\mathcal{B} \neq \mathcal{M}$ do
    Solve $P(\mathcal{N},\mathcal{B},\gamma)$ and determine $S^*$
    for $m \in \mathcal{M}\backslash\mathcal{B}$ do
        if $\lambda_m < 0$ then
            $\gamma_m := S^*$
        end if
    end for
    $\mathcal{B} \leftarrow \mathcal{B} \cup \{m \in \mathcal{M}\backslash\mathcal{B} \ \ s.t \ \ \lambda_m < 0\}$
end while
return Optimal solution of $P(\mathcal{N},\mathcal{M},\mathcal{B},\gamma)$
```

The algorithm uses the complementary slackness theorem and determines with condition $\lambda_m < 0$ if Constraint (9) is saturated for metrology tool m, i.e. if m is critical. If it is the case, m is added to the set $\mathcal{B}$, and W_m is fixed to the current value of S^* in the next iterations through the vector γ.

As demonstrated in (Christ et al. 2019), the solution determined by the IMM procedure has the following interesting properties (see Figure 1):

1. All the metrology tools in a solution determined by the IMM procedure are such that it is impossible to decrease the workload of tool m without increasing the workload of a tool m with a larger workload than m, i.e. such that $W_{m'} \geqslant W_m$.
2. If two metrology tools m and m' share the fraction of at least one common lot, then m and m' have the same workload.

These two properties allow us to better identify critical metrology tools that exceed the maximum capacity W. Hence, the chances to decrease the maximum workload by skipping lots assigned to these tools are larger. This is because, if the workload of a group of tools is strictly larger than W, then some lots assigned to these tools by the IMM procedure have to be skipped.

4 ITERATED SKIPPING PROCEDURE

An iterated skipping procedure is proposed in this section. This procedure combines any workload balancing procedure to determine critical tools that are overloaded with a skipping phase where lots are skipped on the critical tools to ensure that the maximum workload W is satisfied. Let us denote by $WB(\mathscr{N},\mathscr{M})$ any workload balancing procedure (P and IMM in this paper) that takes a set of lots $\mathscr{N}$ and a set of tools $\mathscr{M}$ as inputs and returns a workload balance between the tools.

In the iterated skipping procedure, WB is first ran on the initial set of sampled lots. If there is no tool with a workload strictly larger than W, then no lots are skipped. Otherwise, among the lots assigned to the most loaded critical tools, the lot with the lowest ratio "priority divided by measurement time" is skipped. $WB(\mathscr{N},\mathscr{M})$ is ran after each skipped lot to be sure that all the tools are critical. The procedure is iterated until the maximum capacity W is reached.

The following notations are introduced for a specific workload balancing procedure WB (P or IMM):

- $W_{max}(\mathscr{N},\mathscr{M})$: Maximum workload in the solution determined by $WB(\mathscr{N},\mathscr{M})$,
- $\mathscr{M}_{max}(\mathscr{N},\mathscr{M})$: Set of the tools m such that $W_m = S^*$ in the solution determined by $WB(\mathscr{N},\mathscr{M})$,
- $\mathscr{L}_m(\mathscr{N},\mathscr{M})$: Set of lots l with a strictly positive fraction assigned to tool m, i.e. such that $X_{l,m} > 0$, in the solution determined by $WB(\mathscr{N},\mathscr{M})$,
- w_l: Measurement priority of lot l (the higher the priority of l, the larger w_l).

At each iteration of Algorithm 2, the lot to skip is not only selected by its measurement priority (w_l), but also by its measurement time. Indeed, the maximum workload should be decreased as quickly as possible, and we want to skip as few lots as possible. Hence, when critical tools have several multiple assigned lots with the same priority, the lot with the largest measurement time is skipped.

Algorithm 2 IS+WB

```
Inputs: $WB$, $W$, $\mathscr{N},\mathscr{M}$
Output: Set of skipped lots $\mathscr{S}^*$
Run $WB(\mathscr{N})$
$\mathscr{N}^* \leftarrow \mathscr{N}$
$\mathscr{S}^* \leftarrow \{\}$
while $W_{max}(\mathscr{N}^*,\mathscr{M}) > W$ do
    for $m \in \mathscr{M}_{max}(\mathscr{N}^*,\mathscr{M})$ do
        $worstlot \leftarrow argmin_{l \in \mathscr{L}_m(\mathscr{N}^*,\mathscr{M})}(w_l/p_{l,m})$
    end for
    $\mathscr{N}^* \leftarrow \mathscr{N}^* \backslash worstlot$
    $\mathscr{S}^* \leftarrow \mathscr{S}^* \cup \{worstlot\}$
    Run $WB(\mathscr{N}^*,\mathscr{M})$
end while
```

The efficiency of the iterated skipping algorithm with both workload balancing procedures is analyzed in the next section on industrial instances.

5 COMPUTATIONAL EXPERIMENTS

After explaining how the experiments on industrial data were designed, numerical results are discussed to compare the efficiency of the iterated skipping algorithm with both workload balancing procedures.

5.1 Design of Experiments

The experiments were conducted on two different sets I_1 and I_2 of 15 industrial instances in terms of sampled lots, availablemetrology tools, qualifications and measurement times. The measurement priorities of the lots were assigned and distributed using the following probability distribution:

- Low priority lots ($w_l = 1$), $\approx 20\%$,
- Normal priority lots ($w_l = 5$), $\approx 60\%$,
- High priority lots ($w_l = 20$), $\approx 15\%$,
- Ultra-high priority lots ($w_l = 100$), $\approx 5\%$.

Tables 1 and 2 provide details on the two sets of instances $I1$ and $I2$, respectively. The main difference between the two sets is that fewer metrology tools are available in set $I1$, with a number of qualifications per tool which is very unbalanced compared to set $I2$. For each instance, the tables show the number of lots sampled to be measured, the number of available metrology tools and the value S^* (in seconds) corresponding to the objective function of $P(\mathscr{N},\mathscr{M},\varnothing,[0,...,0])$, i.e. the workload of the most loaded tool if all the sampled lots are assigned to the metrology tools (not lot is skipped).

Instances	$\lvert\mathscr{N}\rvert$	$\lvert\mathscr{M}\rvert$	S^*
$I1_1$	6	10	159.1
$I1_2$	15	10	541.1
$I1_3$	56	9	1637.9
$I1_4$	35	8	2300.5
$I1_5$	45	9	1677.9
$I1_6$	105	10	3948.5
$I1_7$	136	9	5466.9
$I1_8$	194	9	8810.4
$I1_9$	198	6	10893.4
$I1_{10}$	155	9	5497.3
$I1_{11}$	24	10	472.3
$I1_{12}$	32	11	771.9
$I1_{13}$	24	10	713.6
$I1_{14}$	34	9	2050.7
$I1_{15}$	158	9	5970.1

Table 1: Description of industrial instances $I1$.

Instances	$\lvert\mathscr{N}\rvert$	$\lvert\mathscr{M}\rvert$	S^*
$I2_1$	138	15	564.0
$I2_2$	97	14	482.3
$I2_3$	83	13	350.0
$I2_4$	82	15	400.3
$I2_5$	141	15	569.9
$I2_6$	149	14	502.6
$I2_7$	117	15	411.5
$I2_8$	85	12	328.5
$I2_9$	94	15	398.0
$I2_{10}$	105	15	481.6
$I2_{11}$	95	14	507.5
$I2_{12}$	113	14	512.66
$I2_{13}$	83	15	395.5
$I2_{14}$	100	15	391.8
$I2_{15}$	97	15	482.1

Table 2: Description of industrial instances $I2$.

Not that skipping some lots is required as, in some instances, the waiting queue exceeds 2.5 hours, which is not acceptable in practice. All the algorithms and the linear programs in this paper were implemented in Julia 1.8.5 on an Intel(R) Core(TM) i5-1135G7 of 2.40GHz with 16GB RAM. The linear programs were solved using IBM ILOG CPLEX 12.10.

Computational times are not given in the numerical results as they never exceed 10 seconds.

5.2 Comparison between IS+IMM and IS+P

To compare the performance of $IS+P$ and $IS+IMM$ for different values of the maximum capacity W, the number of skipped lots and the priorities of the skipped lots are provided. We would like to show that the iterated skipping algorithm is more effective when combined with the IMM workload balancing procedure, as this combination enables a more accurate identification of critical tools, and therefore a more informed selection of lots to skip to reduce the maximum workload.

For each instance, we ran both algorithms with a maximum metrology capacity $W = \alpha \cdot S^*$, where $\alpha \in \{0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1\}$. The objective is to progressively reduce the maximum capacity and observe how many lots are skipped by both methods, as well as how this affects the priority objective defined as the sum of the priorities of all lots if no lot is skipped ($\sum_{l \in \mathcal{N}} w_l$).

Let us introduce the following indicators:

- **Degradation of the priority objective.** For each instance and for capacity W, the two following values are defined:
$$D_W^{IS+IMM} = (\sum_{l \in \mathcal{N}} w_l - \sum_{l \in \mathcal{S}_W^* IS+IMM} w_l) / \sum_{l \in \mathcal{N}} w_l$$
$$D_W^{IS+P} = (\sum_{l \in \mathcal{N}} w_l - \sum_{l \in \mathcal{S}_W^* IS+P} w_l) / \sum_{l \in \mathcal{N}} w_l$$
which model, respectively, the degradations on the priority objective due to the skipping algorithms combined with IMM and with P.
- **Improvement (Impt).** The improvement is defined as $(D_{\alpha\ S*}^{IS+IMM} - D_{\alpha\ S*}^{IS+P})$, which corresponds, for a given value of α, to the difference of degradations between the two procedures. If the improvement is larger than equal to 0, then $IS+IMM$ degrades less the priorities of lots than $IS+P$.
- **Average Improvement.** It is defined as the average improvement $\sum_{\alpha}(D_{\alpha\ S*}^{IS+IMM} - D_{\alpha\ S*}^{IS+P})/10$ which corresponds, for each instance, to the average difference of degradations between the two procedures. If the average improvement is larger than equal to 0, then, on average, $IS+IMM$ degrades less the priorities of lots than $IS+P$
- **Best and Worst Improvement.** The best improvement is the largest difference of degradation when $IS+IMM$ degrades less the priorities of lots than $IS+P$ and, the worst improvement is largest difference of degradation when $IS+P$ degrades less the priorities of lots than $IS+IMM$.
- **Number of skipped lots.** NS_{α}^{IS+IMM} and NS_{α}^{IS+P} denote the number of lots skipped by $IS+IMM$ and $IS+P$ respectively, for a given value of α.

For example, looking at the first line of Table 5, an average improvement of +4.50% means that, on average (over the values of $W = \alpha\ S^*$), the iterated skipping algorithm $IS+IMM$ degrades by 4.50% the priorities of lots less than the iterated skipping algorithm $IS+P$. A best improvement of +13.50% means that the largest degradation difference between $IS+IMM$ and $IS+P$ is 13.50% over all values of α. A worst improvement of -0.00% means that $IS+P$ always degrades more the priorities of lots than $IS+IMM$.

Tables 3 and 4 show detailed results on three instances from sets $I1$ and $I2$, respectively, and for all values of α. In each instance, some groups of tools dominate the others in terms of number of qualifications Therefore, the set of critical tools with maximum workload is often the same for both workload balancing approaches when α is close to 1. This is why $IS+IMM$ and $IS+P$ skip the same sets of lots. Overall, note that $IS+IMM$ is much more efficient than $IS+P$ when α becomes small, i.e. the number of lots to skip is large. $IS+IMM$ skips fewer lots to reach the maximum allowed workload W and keeps the lots with the highest priority.

Tables 5 and 6 summarizes the benefits of using IMM workload balancing procedure in the iterated skipping algorithm. For most instances (both industrial and semi-industrial instances), the number of lots

Instance	α	Impt	NS_{α}^{IS+IMM}	NS_{α}^{IS+P}
$I1_1$	0.9	+0.00%	**1**	**1**
	0.8	+0.00%	**1**	**1**
	0.7	+0.00%	**2**	**2**
	0.6	+0.00%	**3**	**3**
	0.5	+0.00%	**3**	**3**
	0.4	+0.00%	**4**	**4**
	0.3	+13.51%	**5**	6
	0.2	+13.51%	**5**	6
	0.1	+13.51%	**5**	6
$I1_2$	0.9	+0.00%	**1**	**1**
	0.8	+0.55%	**3**	4
	0.7	+0.55%	**3**	4
	0.6	+0.55%	**3**	4
	0.5	+0.00%	**6**	**6**
	0.4	+24.63%	**9**	13
	0.3	+24.63%	**9**	13
	0.2	+24.63%	**9**	13
	0.1	+17.19%	**12**	13
$I1_6$	0.9	+0.42%	**5**	10
	0.8	+0.08%	18	**15**
	0.7	+0.60%	21	**20**
	0.6	+1.21%	**26**	27
	0.5	-0.52%	34	**32**
	0.4	+1.11%	**41**	43
	0.3	+1.20%	**53**	58
	0.2	+3.34%	**66**	74
	0.1	+5.62%	87	**84**

Table 3: Detailed improvement and number of skipped lots for each α on some instances of $I1$.

Instance	α	Impt	NS_{α}^{IS+IMM}	NS_{α}^{IS+P}
$I2_2$	0.9	+0.00%	**6**	**6**
	0.8	+0.00%	**12**	**12**
	0.7	+0.00%	**17**	**17**
	0.6	+0.45%	**23**	24
	0.5	+0.45%	**29**	30
	0.4	+0.45%	**35**	36
	0.3	+4.33%	**44**	56
	0.2	+1.17%	**51**	56
	0.1	+3.16%	**61**	72
$I2_{10}$	0.9	+0.00%	**6**	**6**
	0.8	+0.00%	**11**	**11**
	0.7	+0.00%	**16**	**16**
	0.6	+0.00%	**21**	**21**
	0.5	+0.00%	**26**	**26**
	0.4	+0.08%	**33**	34
	0.3	+0.34%	**40**	44
	0.2	+2.94%	**48**	58
	0.1	+7.43%	**58**	67
$I2_{14}$	0.9	+0.00%	**4**	**4**
	0.8	+0.00%	**9**	**9**
	0.7	+0.00%	**14**	**14**
	0.6	+0.00%	**19**	**19**
	0.5	5.36%	**25**	37
	0.4	4.55%	**30**	41
	0.3	+4.01%	**35**	48
	0.2	+6.69%	**43**	53
	0.1	+3.57%	**55**	57

Table 4: Detailed improvement and number of skipped lots for each α on some instances of $I2$.

that are skipped is lower with $IS+IMM$ than with $IS+P$. These results indicate that IMM allows for a better identification of the most critical tools and, consequently, for a more effective selection of lots to skip. Moreover, in almost all cases, $IS+IMM$ is better at not skipping the highest priority lots, as the average degradation difference between the two procedures is almost always positive. It may happen that the lots skipped with $IS+P$ are more relevant than the lots skipped with $IS+IMM$ in terms of lot priorities for some values of W in the first set of instances $I1$. This is because, in these instances, the metrology tools are all very similar, and can easily have the same workload. in this case, the tools with the maximum workload in P are often all critical, and skipping a lot assigned to any of these tools has a high probability of reducing the maximum workload.

Finally, note that $IS+IMM$ is more effective on small and medium sized instances, where each skipped lot has a significant impact on both the maximum workload and the lot priorities. For larger instances, i.e. with a large number of sampled lots, much more lots need to be skipped to reach the maximum allowed workload W, and individual skipping decisions have less impact.

Instance	Improvement			Total number of skipped lot	
	Avg	Best	Worst	$IS+IMM$	$IS+P$
$I1_1$	+4.50%	+13.51%	-0.00%	**29**	32
$I1_2$	+10.30%	+24.63%	-0.00%	**55**	71
$I1_3$	+1.07%	+9.64%	-0.00%	**170**	174
$I1_4$	+0.33%	+1.14%	-0.00%	**99**	108
$I1_5$	+1.19%	+8.10%	-11.94%	**174**	182
$I1_6$	+1.47%	+5.62%	-0.52%	**351**	363
$I1_7$	-0.08%	+1.00%	-1.30%	**426**	**426**
$I1_8$	+0.02%	+0.20%	-0.00%	**683**	**683**
$I1_9$	+0.58%	+2.73%	-0.0%	**683**	695
$I1_{10}$	+0.12%	+1.14%	-0.28%	**564**	**564**
$I1_{11}$	+1.08%	+7.84%	-0.00%	**96**	98
$I1_{12}$	+0.40%	+3.74%	-0.00%	**113**	**113**
$I1_{13}$	+1.31%	+9.80%	-0.00%	**84**	87
$I1_{14}$	+1.30%	+3.4%	-0.00%	**116**	140
$I1_{15}$	+0.02%	+1.10%	-0.84%	**613**	**613**

Table 5: Indicators for all instances of $I1$: AI, HI, WI, and total number of skipped lots for both methods and for all values of α

.

Instance	Improvement			Total number of skipped lot	
	Avg	Best	Worst	$IS+IMM$	$IS+P$
$I2_1$	+1.67%	+2.89%	-0.00%	**355**	420
$I2_2$	+1.11%	+4.33%	-0.00%	**278**	309
$I2_3$	+0.66%	+2.75%	-0.00%	**211**	231
$I2_4$	+0.53%	+1.33%	-0.00%	**207**	226
$I2_5$	+1.18%	+3.06%	-0.00%	**374**	460
$I2_6$	+0.83%	+3.65%	-0.00%	**317**	370
$I2_7$	+0.65%	+1.88%	-0.15%	**240**	256
$I2_8$	+1.26%	+2.68%	-0.00%	**213**	260
$I2_9$	+0.55%	+1.39%	-0.46%	**210**	228
$I2_{10}$	+1.20%	+7.43%	-0.00%	**259**	283
$I2_{11}$	+1.48%	+5.55%	-0.00%	**204**	242
$I2_{12}$	+0.66%	+3.22%	-0.00%	**290**	312
$I2_{13}$	+1.09%	+2.44%	-0.00%	**211**	233
$I2_{14}$	+2.67%	+6.69%	-0.00%	**234**	282
$I2_{15}$	+0.94%	+6.85%	-0.02%	**242**	262

Table 6: Indicators for all instances of $I2$: AI, HI, WI, and total number of skipped lots for both methods and for all values of α.

6 CONCLUSIONS AND PERSPECTIVES

The IMM workload balancing procedure proposed in (Christ et al. 2019) was developed for operational production planning, as shown in (Christ et al. 2023). In this paper, we show that the IMM procedure is also efficient in the context of metrology capacity management. The IMM procedure provides a reliable estimation of which tools are truly critical, enabling a more precise identification of the lots that should be skipped to satisfy metrology capacity. We argue that the IMM procedure is particularly useful for heterogeneous metrology areas with many different tools (see Section 5, particularly set of instances $I2$).

Looking forward, we believe that different sampling and skipping strategies could be combined with the IMM procedure for metrology management. For example, sampling strategies could be dynamically adjusted based on the workloads of the metrology tools determined by the IMM procedure, for example by prioritizing lots that can be measured on less critical tools.

REFERENCES

Christ, Q., S. Dauzere-Peres, and G. 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, Q., S. Dauzère-Pérès, and G. Lepelletier. 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.

Dauzère-Pérès, S., J.-L. Rouveyrol, C. Yugma, and P. Vialletelle. 2010. “A Smart Sampling Algorithm to Minimize Risk Dynamically”. In *2010 IEEE/SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 307–310. IEEE.

Dilosi, A., A. Hassan, A. Mili, and A. Siadat. 2022. “Dynamic Sampling Plans using a Metrology Situation Indicator (MSI)”. In *2022 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 1139–1143. IEEE.

Freud, D., and G. Mosheiov. 2021. “Scheduling with Competing Agents, Total Late Work and Job Rejection”. *Computers & Operations Research* 133:105329 https://doi.org/10.1016/j.cor.2021.105329.

Geng, X.-N., X. Sun, J. Wang, and L. Pan. 2023. “Scheduling on Proportionate Flow Shop with Job Rejection and Common Due Date Assignment”. *Computers & Industrial Engineering* 181:109317 https://doi.org/10.1016/j.cie.2023.109317.

Le Quéré, É., S. Dauzère-Pérès, S. Astie, C. Maufront, X. Michallet, G. Bugnon *et al.* 2019. “Dynamic Cloud-based Computation for Skipping Lots in Metrology: IE: Industrial Engineering”. In *2019 30th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, 1–5. IEEE.

Le Quéré, É., S. Dauzère-Pérès, K. Tamssaouet, C. Maufront, and S. Astie. 2020. “Dynamic Sampling for Risk Minimization in Semiconductor Manufacturing”. In *2020 Winter Simulation Conference (WSC)*, 1886–1897. IEEE.

Mouli, C., and M. J. Scott. 2007. “Adaptive metrology sampling techniques enabling higher precision in variability detection and control”. In *2007 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, 12–17. IEEE.

Nduhura-Munga, J., G. Rodriguez-Verjan, S. Dauzère-Pérès, C. Yugma, P. Vialletelle, and J. Pinaton. 2013. “A Literature Review on Sampling Techniques in Semiconductor Manufacturing”. *IEEE Transactions on Semiconductor Manufacturing* 26(2):188–195.

Perez, J. A. S. 2017. *Risk Minimization Through Metrology in Semiconductor Manufacturing*. Ph. D. thesis, Université de Lyon.

Shabtay, D., and E. Gerstl. 2024. “Coordinating Scheduling and Rejection Decisions in a Two-machine Flow Shop Scheduling Problem”. *European Journal of Operational Research* 316(3):887–898 https://doi.org/10.1016/j.ejor.2024.03.021.

AUTHOR BIOGRAPHIES

MATHIS MARTIN is a PhD student at Mines de Saint-Etienne, France. His research interests include modeling and optimization of metrology operations in semiconductor manufacturing. His email address is mathis.martin@emse.fr.

STÉPHANE DAUZÈRE-PÉRÈS is Professor at Mines Saint-Etienne, France, and Adjunct Professor at BI Norwegian Business School, Norway. He received the Ph.D. degree from Paul Sabatier University in Toulouse, France, in 1992 and the H.D.R. from Pierre and Marie Curie University, Paris, France, in 1998. He was a Postdoctoral Fellow at M.I.T., U.S.A., in 1992 and 1993, and Research Scientist at Erasmus University Rotterdam, The Netherlands, in 1994. He has been Associate Professor and Professor from 1994 to 2004 at the Ecole des Mines de Nantes, France. His research interests broadly include modeling and optimization of operations at various decision levels in manufacturing and logistics, with a special emphasis on production planning and scheduling, on semiconductor manufacturing and on railway operations. He has published more than 115 papers in international journals and contributed to more than 250 communications in national and international conferences. Stéphane Dauzère-Pérès has coordinated numerous academic and industrial research projects, including 5 European projects and more than 30 industrial (CIFRE) PhD theses, and also eight conferences. He was runner-up in 2006 of the Franz Edelman Award Competition, and won the Best Applied Paper of the Winter Simulation Conference in

2013 and the EURO award for the best theory and methodology EJOR paper in 2021. His email address is dauzere-peres@emse.fr.

CLAUDE YUGMA is Professor at the Center of Microelectronics in Provence (CMP) of Mines Saint-Etienne in France since 2016 in the Manufacturing Sciences and Logistics department. He received the Ph.D. degree from the Institut National Polytechnique of Grenoble, France, in 2003, and his H.D.R. from the Jean-Monnet University, Saint-Etienne, in December 2013. He was a Postdoctoral fellow at the École Nationale Supérieure de Génie Industriel, Grenoble, from 2003 to 2004 and from 2005 to 2006 at EMSE. He co-organized several international conferences as for example the 2013 edition of the conference Modeling and Analysis of Semiconductor Manufacturing. His research interests are modeling and scheduling in semiconductor manufacturing. He has published more than 20 papers in international journals and has contributed to more than 80 communications in conferences. His email address is yugma@emse.fr.

AYMEN MILI Aymen Mili, is currently the Control Plan Systems and Tools Manager for the Digital Front-End Manufacturing Quality System at STMicroelectronics. He has accumulated valuable experience through various roles at STMicroelectronics and focusing on quality systems, process control, and digital manufacturing. Aymen holds a Master's degree from Arts et Métiers ParisTech – École Nationale Supérieure d'Arts et Métiers (2005–2006) and brings a strong skill set in management, project management, and industrial quality. His professional profile reflects a deep engagement with the optimization of manufacturing processes in the semiconductor industry. aymen.mili@st.com.

RENAUD ROUSSEL is a Scheduling and Dispatching Full Automation Expert at STMicroelectronics in Crolles (France). He has been working for more than 2 decades in the semiconductor industry in manufacturing science at the frontier between operational management, industrial engineering and data science to make the fab as efficient as possible. His email address is is renaud.roussel@st.com.