

A Model Predictive Control-Based Scheduling for Optimizing both Quality and Productivity in Multistage Manufacturing Process

Sugyeong Lee¹ and Dong-Hee Lee^{2,*}

Abstract—In this paper, we present a model predictive control (MPC) based approach for scheduling multiple lots in multistage manufacturing process (MMP) to optimize both productivity and quality. We design a mixed integer linear programming (MILP) problem for hybrid flow shop MMP environment, and employ an MPC strategy which can repeatedly solves a subproblem of immediate timestep, rather than solving the scheduling problem over the entire range at once. Furthermore, we consider scenarios where a machine is temporarily unavailable or urgent lot arrivals with specific earliest starting times, which is frequently happening in real industrial settings. Simulation results demonstrate that the proposed approach can optimize both makespan and yield at the same time, which is practically applicable in real-time decision making environments such as semiconductor fabrication plant.

I. INTRODUCTION

Multistage manufacturing process (MMP) has long been recognized as challenging environments for scheduling. In the process, lots typically move through a fixed sequence of stages and each stage contains multiple parallel machines. [1] This can be modeled as a hybrid flow shop scheduling, i.e., while the processing order is the same for all lots, machine assignment at each stage must be determined individually.

A representative example of this multistage environment is semiconductor fabrication plant. Semiconductor manufacturing generally involves hundreds of stages, where unexpected events considerably occur. [2] Furthermore, since especially semiconductor manufacturing demands high quality specification while maintaining productivity, the scheduling scheme should explicitly consider quality enhancement. Although parallel identical machines within the same stage are supposed to output uniform quality characteristics, they often exhibit different quality outcomes. [3]

In other words, the scheduler must both (1) satisfy the quality standard and (2) reflect dynamically changing environment. When it comes to quality, yield is the primary measure in semiconductor manufacturing due to its importance of reducing failure rate in MMP environment. [1] However, conventional approaches have long attempted to optimize productivity as a sole goal. [4] To overcome this issue, there has been some research to address the importance

of considering quality in wafer assignment but it has not expanded its scope to dynamic scheduling. [5]

Hence, we suggest an optimization problem to consider both quality and productivity in the objective of a mixed integer linear programming (MILP) formulation. Specifically in the objective function we integrate yield as quality term to makespan as well-known productivity measure in the formula of the weighted-sum. This enables to optimize yield and productivity simultaneously.

Also to make the scheduling more realistic, we add some unforeseen events to the scheduler, which suddenly happen so that they can have a negative impact for productivity. Machine breakdowns or urgent lot inputs are typical examples. [6] Due to these events, dynamic scheduling becomes more essential rather than static or global scheduling. Model predictive control (MPC) supports step-by-step allocation of lots to machines and enables to cope with sudden changes in manufacturing environment.

The paper is organized as follows. After briefly reviewing related works in the field of scheduling which investigate the usage of the MPC, we first introduce MILP formulation of the optimization problem in hybrid flow shop MMP environment. Then we describe how MPC-based scheduler operates short-horizon planning. Simulation results are also presented to examine how well-suited our approach is. Finally, we conclude the paper discussing the future research directions.

II. RELATED WORKS

MPC is a control strategy which uses a model to predict future behavior and solves an optimization problem, and can be applicable to the systems which cannot be controlled by traditional controllers. [7] Thanks to these characteristics, MPC-based scheduling offers the advantage of direct industrial applicability. The field in which MPC is most widely used is power-related systems in general. Following papers are outstanding examples demonstrating that MPC can be effectively applied to the real-world settings. Nawaz *et al.* [8] formulate scheduling problem to mixed integer nonlinear programming in power charging area and solve it with MPC to balance charging and discharging. For similar purpose, Yang *et al.* [9] suggest mixed integer quadratic programming for MPC problem. Chen *et al.* [10] utilize MPC to minimize the energy cost of power consumption when the energy usage only depends on the request.

Although relatively few studies have adopted MPC in multistage process, one notable example addresses an open shop scheduling problem. Bozzi *et al.* [11] formulate MILP problem to minimize waiting time and total processing time,

¹Sugyeong Lee is with Department of Semiconductor and Display Engineering, Sungkyunkwan University, 2066 Seobu-ro, Suwon, South Korea and Staff Engineer in Mobile Display Division, Samsung Display Co., Ltd., 181 Samsung-ro, Asan, South Korea sugyeong@g.skku.edu

²Dong-Hee Lee is Associate Professor in Department of Industrial Engineering, Sungkyunkwan University, 2066 Seobu-ro, Suwon, South Korea dheel@skku.edu

*Corresponding author

thereby optimizing productivity. Going a step further, our objective is to include a quality term in productivity-focused scheduling and propose an approach that can also handle real-world constraints in the hybrid flow-shop scheduling. We discuss the details of this approach in the following section.

III. PROPOSED APPROACH

A. MILP Problem Formulation

We first introduce how the MMP is formulated as an MILP problem. Since our MMP is the hybrid flow shop environment and is NP-hard, we can convert it into the MILP formulation. [12]

The goal is to maximize yield and minimize makespan at the same time, thus we adopt weighted sum of yield and makespan to the objective function. Yield value of a single lot consists of base yield and predetermined additional machine-wise yield effect. As a single lot proceeds through each stage and it is assigned to a single machine every time, according yield effect is added to the base yield. In other words, the final yield is calculated by summing the base yield and the incremental yield effects resulting from the assignment. Makespan is simply the total duration of running time for all lots to be assigned.

Also, under these two assumptions, constraints for MILP problem of the MMP are determined.

- *Assumption 1:* Each lot is inputted in a serial manner with exactly one machine at a single stage.
- *Assumption 2:* When a lot is processed in a machine, no other lots can be processed on the same machine at the same time.

Parameters for the MILP problem are concisely explained in Table I (a).

Given parameters in Table I (a), we can model MILP optimization problem as below.

$$\text{Maximize } \alpha \cdot \sum_l \sum_i \sum_{m_i} (y_{l,i,m_i} \cdot x_{l,i,m_i}) - \beta \cdot z \quad (1)$$

$$\text{s.t. } \sum_{m_i} x_{l,i,m_i} = 1 \quad \forall l, i \quad (2)$$

$$z \geq f_{l,i} \quad \forall l, i \quad (3)$$

$$f_{l,i} = s_{l,i} + p_i \quad (4)$$

$$s_{l,i} \geq \max\{f_{l,i-1}, g_{l,i-1}\} \quad \forall l, i \quad (5)$$

$$s_{l_u,i} \geq f_{l_v,i} - M \cdot (1 - d_{l_u,l_v,i,m_i}) \quad (6)$$

$$- M \cdot (2 - (x_{l_u,i,m_i} + x_{l_v,i,m_i})) \quad \forall j, k \in L$$

$$s_{l_v,i} \geq f_{l_u,i} - M \cdot d_{l_u,l_v,i,m_i} \quad (7)$$

$$- M \cdot (2 - (x_{l_u,i,m_i} + x_{l_v,i,m_i})) \quad \forall j, k \in L$$

(1) is a composite objective function that maximizes total yield while minimizing the makespan. Here α and β are positive scalar weights and the importance of yield maximization and makespan reduction can be adjusted by tuning their relative magnitudes. (2) is machine assignment per stage. This ensures every lot must be assigned to exactly only one machine at each stage. (3) means that the makespan must be at least as large as every lot's finishing time. (4)

TABLE I
NOMENCLATURE USED IN THE PAPER

Symbol	Description
(a) Notation for MILP problem	
l	number of lots
L	set of lots to be inputted ($= \{1, 2, \dots, l\}$)
i	stage number in multistage manufacturing process
I	set of stages ($= \{1, 2, \dots, i\}$)
m_i	machine identification number at stage i
M_i	set of available machines at stage i ($= \{m_1, \dots, m_i\}$)
α	quality weight
β	productivity weight
y_{l,i,m_i}	effect on yield when lot l is processed in machine m_i at stage i
z	makespan
p_i	processing time of stage i
$s_{l,i}$	stage i starting time of lot l
$f_{l,i}$	stage i finishing time of lot l
$g_{l,i}$	earliest starting time of lot l of entering stage i
x_{l,i,m_i}	binary variable to indicate whether lot l uses machine m_i at stage i
d_{l_u,l_v,i,m_i}	binary variable to determine which lot is processed first when l_u and l_v both use same machine m_i
M	sufficiently large constant for Big-M method
(b) Notation for realistic environment setting	
(u_s, u_e)	single case of machine unavailable time window
U_M	set of machine unavailable interval including (u_s, u_e)
U_L	set of urgent lot
(c) Notation for MPC-based scheduling algorithm	
j_l	current stage index that lot l must process next
T	current global system timestep which is updated iteratively
H	short-horizon length
k	iteration number

indicates finishing time of a lot after a single stage ends. (5) ensures sequential stages ordering, i.e., a lot can proceed next stage only after finishing the previous stage. (6) and (7) are machine disjunctive constraints, i.e., if two lots are assigned to the same machine at the same timestep, binary variable d forces for multiple lots not to occupy the machine at the same time.

Here the binary variables x_{l,i,m_i} and d_{l_j,l_k,i,m_i} are determined as follows.

$$x_{l,i,m_i} \in \{0, 1\}, \quad \begin{cases} 1, & \text{if lot } l \text{ is processed in } m_i, \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

$$d_{l_j,l_k,i,m_i} \in \{0, 1\}, \quad \begin{cases} 1, & \text{if lot } l_k \text{ precedes lot } l_j, \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

B. MPC-based dynamic scheduling

Instead of solving the entire scheduling horizon in a single optimization run, we adopt an MPC-based rolling-horizon approach.

At each iteration at the current timestep T , the model selects only the lots that are currently active, i.e., $g_{l,i}$ is less than or equal to T and solves the above MILP over small timesteps ahead. After obtaining the optimal schedule for this short horizon, only the immediate next step for each lot is fixed in the global schedule. At the same time, the system's current timestep is updated to the earliest finishing time of

the newly fixed steps. Moreover, if urgent lots newly arrive or if machine availability changes, the next iteration sees a different set of active lots or constraints.

This procedure embodies MPC by continually re-optimizing a short-horizon future schedule but only committing to the immediate step. Moreover, real-time disturbances such as machine breakdowns or emergency lots can be adapted for the real-world implementation while still optimizing both yield and makespan.

MPC-based dynamic scheduling involves 5 steps as follows.

- Step 1 (Identifying active lots): Any lot l whose earliest starting time $g_{l,i}$ is less than or equal to the current global time T is considered to be active and available for scheduling in the next short-horizon MILP.
- Step 2 (Short-horizon MILP): This step restricts each active lot l for the next horizon, defines MILP constraints, and maximizes the objective function.
- Step 3 (Partial fixing): Only the immediate next stage j for each lot l is committed to the global schedule.
- Step 4 (Update global time): The global system time T is updated based on earliest finish policy.
- Step 5 (Repeat): By rolling iteration, the MILP problem with new active lot sets and updated times is re-solved until all lots are fully processed.

Here we integrate additional parameters to make the algorithm more realistic such as (1) urgent lot arrivals and (2) machine unavailability constraints to the MILP problem. For urgent lots, if there exists any new urgent lots at current system time T , they are appended to the set of active lots at Step 1 of the algorithm. Also for machine unavailability, (10) enforces that the scheduling result can ensure any lot l must not overlap the unavailable time window (u_s, u_e) for machine m_i by utilizing big-M constraints again. Table I (b) shows brief notation for incorporating the real-world settings.

$$f_{l,i} \leq u_s + M \cdot (1 - x_{l,i,m_i}), \quad s_{l,i} \geq u_e - M \cdot (1 - x_{l,i,m_i}) \quad (10)$$

Whole procedure is shown in Algorithm 1. Table I (c) explains additional notation for the MPC-based scheduling algorithm.

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

For demonstration, we assume 6-stage environment. At each stage, 2, 3, 1, 3, 4, 2 parallel identical machines exist in over 6 stages, i.e., $M_i = \{2, 3, 1, 3, 4, 2\}$. Processing time p_i for each stage i is 2, 3, 1, 3, 5, 2 respectively. For input lot configuration, we put 10 regular lots; lots proceed from the first stage to the end without missing any single stage, and 2 urgent lots at timestep $T = 5$. Machine breakdown is also set for two stages. Detailed configuration is shown in Table II.

Based on each lot's yield value, we add additional yield effects of each machine for use in the objective function. For the additional effect values, random numbers are generated

Algorithm 1 MPC-based Scheduling Algorithm

Require: • Lot set L , Stage set I , Machine set $\{M_i\}$, processing time $\{p_i\}$, Earliest starting time $\{g_{l,i}\}$, Horizon H
• Machine unavailable set U_M , urgent lot set U_L
Ensure: Global schedule $(m_i, s_{l,i}, f_{l,i})$ of (l, i) assignments (machine, start/finish times)

- 1: Initialize:
- 2: $T \leftarrow 0$
- 3: **for each** $l \in L$ **do** $j_l \leftarrow 0$
- 4: $k \leftarrow 0$
- 5: **while** exists at least one lot l not completed **do**
- 6: **Step 1: Identify active lots**
- 7: $L^{(k)} \leftarrow U^{(k)} \cup \{l \in L \mid g_{l,i} \leq T\}$
- 8: **Step 2: Short-Horizon MILP**
- 9: **for each** $l \in L^{(k)}$ **do**
- 10: $I_l^{(k)} \leftarrow \{i \mid i \geq j_l, i < j_l + H\} \cap I$
- 11: **Define decision variables:** $x_{l,i,m_i}, s_{l,i}, f_{l,i}$
- 12: **Objective:**
- 13: Maximize
- 14:
$$\alpha \cdot \sum_l \sum_i \sum_{m_i} (y_{l,i,m_i} \cdot x_{l,i,m_i}) - \beta \cdot z$$
- 15: **Constraints:**
- 16: 1) $\sum_{m_i \in M_i} x_{l,i,m_i} = 1$
- 17: 2) $z^{(k)} \geq f_{l,i}$
- 18: 3) $f_{l,i} = s_{l,i} + p_i$
- 19: 4) $s_{l,i} \geq \max\{g_{l,i}, f_{l,i-1}\}$
- 20: 5) Disjunctive constraints (See (6) and (7))
- 21: 6) Machine unavailability (See (10))
- 22: **Solve the MILP to obtain partial schedule**
- 23: **Step 3: Partial fixing**
- 24: **for each** $l \in L^{(k)}$ **do**
- 25: $i_{\text{fix}} \leftarrow g_{l,j}$
- 26: **Extract** (m_i^*, s^*, f^*) from MILP solution
- 27: **Fix** this assignment in the global schedule
- 28: $g_{l,i} \leftarrow g_{l,i} + 1$
- 29: **Step 4: Update global time**
- 30: $T \leftarrow \min\{f_{l,i_{\text{fix}}} : l \in L^{(k)}\}$
- 31: **for each** $l \in L^{(k)}$ **do**
- 32: $j_l \leftarrow f_{l,i_{\text{fix}}}$
- 33: **Step 5: Repeat**
- 34: $k \leftarrow k + 1$
- 35: **end while**

within the interval $(-0.1, 0.1)$, and the specific values of each effect are given in Table III. This interval is based on empirical measurements of machine-specific yield variability in our production environment, which consistently fluctuates within $\pm 0.1p$ of the nominal yield, and extra effects from specific machines in Table III likewise reflects real-world production data.

B. Results

For the algorithm, the short horizon H is set to 2, and the iteration k is 6 simply corresponding to our 6-stage

TABLE II
CONFIGURATION OF MULTISTAGE ENVIRONMENT SETTING

MMP environment configuration	
total stages	6
num. machines	2, 3, 1, 3, 4, 2 (in each stage respectively)
process time	2, 3, 1, 3, 5, 2 (for each stage respectively)
Lot input configuration	
regular lots	10
urgent lots	2
urgent lot #1	start stage: 2, end stage: 2, queue request time: $T = 5$
urgent lot #2	start stage: 2, end stage: 6, queue request time: $T = 5$
Machine unavailability configuration	
unavailable #1	at Stage 1 - machine 1 during $T = 3, 4$
unavailable #2	at Stage 2 - machine 1 during $T = 6$

TABLE III
EFFECT ON YIELD OF MACHINES

Stage number	Machine number	Value
1	1	-0.0436
	2	0.0512
2	1	0.0237
	2	-0.0499
	3	0.0819
3	1	0.0966
4	1	0.0620
	2	0.0804
	3	-0.0380
5	1	0.0460
	2	0.0798
	3	0.0368
	4	-0.0056
6	1	-0.0799
	2	-0.0132

environment. A sufficiently large number M is set as 999,999. Each iteration schedules each partial stage for the active lots and after 6 iterations, all stages from stage 1 to stage 6 are fully scheduled.

To validate our scheduling algorithm and investigate the impact of quality and productivity on scheduling, we conducted experiments based on 3 scenarios: (a) firstly put same weight in the objective function as $\alpha = \beta = 1$ which balances quality and productivity then compare this by exaggerating weight parameters as (b) $\alpha = 1$ and $\beta = 10$ for productivity focus and (c) $\alpha = 10$ and $\beta = 1$ for quality focus. Illustration is shown in Fig. 1 and comparison results are as follows in Table IV.

TABLE IV
COMPARISON RESULTS AMONG SCENARIOS

	(a)	(b)	(c)
Yield effect average	+0.9428	+0.9359	+1.0667
→ Difference from (a)		-1%	+13%
Makespan	33	32	43
→ Difference from (a)		-1	+10

Table IV compares the outcomes of scenarios (a), (b), and (c) in terms of both the averaged accumulated yield effect bonus per lot and the makespan. As shown, scenario (a)

achieves an averaged accumulated yield bonus of 0.9428, which serves as a baseline. Scenario (b) marginally lags behind (a) by 1% (dropping to 0.9359), whereas scenario (c) moderately improves quality by increasing the yield effect bonus by 14% (reaching 1.0667). In terms of the makespan, scenario (a) finishes at 33 timestep. Scenario (b) slightly improves productivity upon (a) by saving one additional timestep (32), but scenario (c) extends the makespan by 10 timesteps (43) compared to the baseline result. According to Fig. 1, in scenario (c), lot assignments are mainly concentrated on machines with high yield effects, resulting in the increased makespan. However, this trade-off also leads to a noticeable gain in overall yield. Taken together, these findings demonstrate that our approach enables real-time scheduling to optimize both yield and productivity simultaneously. Depending on the primary objective at any given time, engineers can determine the appropriate weight value.

V. CONCLUSION

In many manufacturing domains, MMP follows the flexible flow shop setting where multiple parallel identical machines at each stage produce various yield levels. Our approach directly addresses this environment by modeling MMP setting to the MILP problem, accommodating practical constraints including the guarantee of disjunctive machine usage.

A key merit of our approach is to optimize both quality and productivity simultaneously. By incorporating machine-wise yield effect into the objective function along with total processing time, the scheduler can balance quality and productivity depending on whether there is a greater need to focus on production quality or output volume.

Moreover, by employing the MPC framework to solve a short-horizon MILP, the schedule can be always revised even when unexpected events happens such as urgent lot arrivals and machine breakdown. Instead of full-horizon static and global scheduling, the algorithm re-optimizes over a short horizon and applies immediate solution at a time, thus makes it possible to adapt quickly to real-time changes.

In practice, we verified through simulation that our approach adequately reflects real-world conditions, showing the trade-off between quality and productivity and the utilization by freely adjusting the weight parameters depending on the emphasis on quality or productivity.

Finally, several promising directions can further extend and strengthen the framework: First, although current scheduler deals with several real-world manufacturing challenges, it could be refined by incorporating additional disturbance types—such as sudden machine setups at particular stages and intentional lot delays—which would our approach more realistic. Incorporating additional real-world events and each environment's unique situation would achieve an even more realistic scheduling model. Second, since we have not yet assessed our scheduler's performance over a broad range of problem sizes, future research will include a systematic

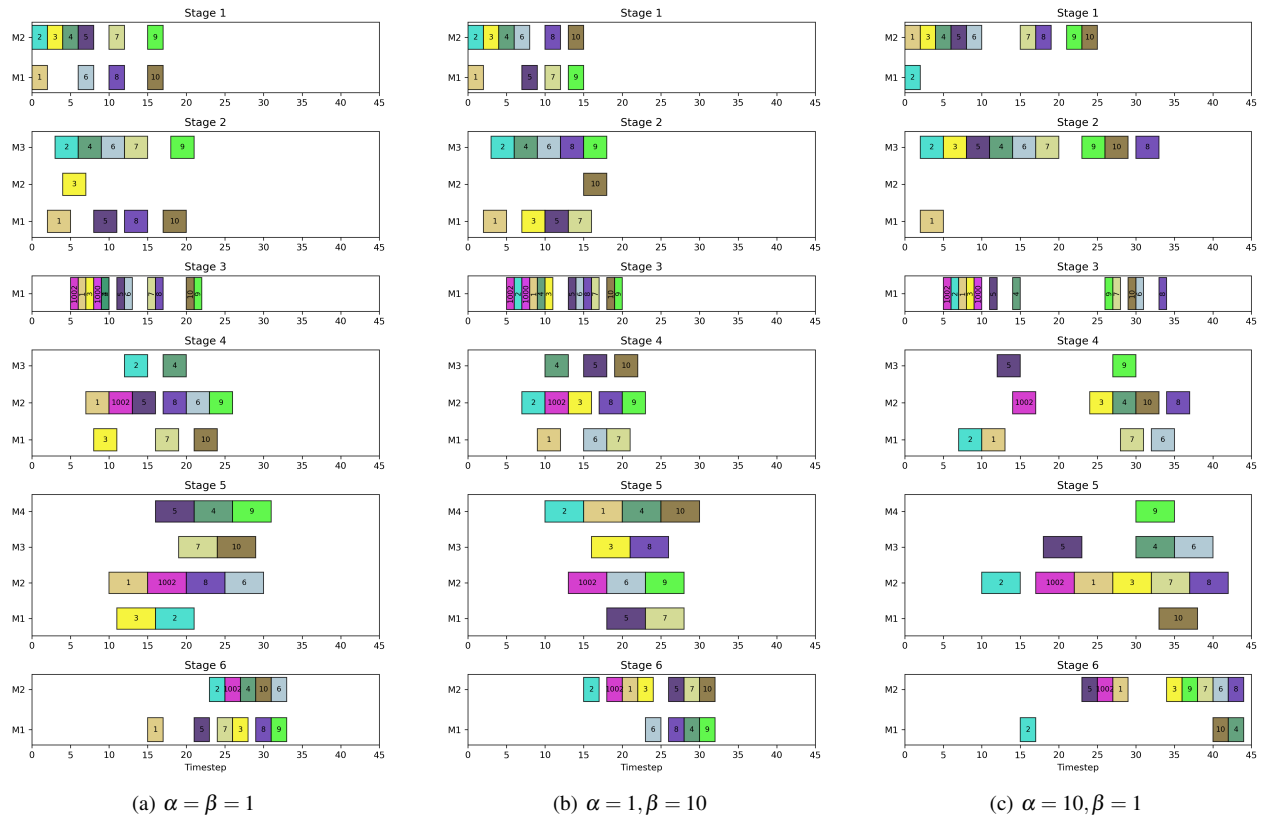


Fig. 1. Scheduling results of varying relative emphasis on quality (α) and productivity (β)

scalability evaluation across varying lot counts and stage numbers to ensure practical deployment in larger fabs.

REFERENCES

- [1] C.-H. Lee, D.-H. Lee, Y.-M. Bae, S.-H. Choi, K.-H. Kim, and K.-J. Kim, Approach to Derive Golden Paths Based on Machine Sequence Patterns in Multistage Manufacturing Process, *Journal of Intelligent Manufacturing*, vol. 33, pp. 167-183, 2022.
- [2] J. S. Jang and S. C. Park, A Reliability Model for a Wafer FAB, *Cogent Engineering*, vol. 4, 1340817, 2017.
- [3] T.-H. Pan, D. S.-H. Wong, S.-S. Jang, Chamber Matching of Semiconductor Manufacturing Process Using Statistical Analysis, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 42, part C, pp. 571-576, 2012.
- [4] L. M. Wein, Scheduling Semiconductor Wafer Fabrication, *IEEE Transactions on Semiconductor Manufacturing*, vol. 1, pp. 115-130, 1988.
- [5] D.-H. Lee, C.-H. Lee, S.-H. Choi, and K.-J. Kim, A Method for Wafer Assignment in Semiconductor Wafer Fabrication Considering both Quality and Productivity Perspectives, *Journal of Manufacturing Systems*, vol. 52, part A, pp 23-31, 2019.
- [6] F. Qiao, Y. Ma, M. Zhou and Q. Wu, A Novel Rescheduling Method for Dynamic Semiconductor Manufacturing Systems, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, pp. 1679-1689, 2020.
- [7] M. Schwenzer, M. Ay, T. Bergs, and D. Abel, Review on Model Predictive Control: an Engineering Perspective, *The International Journal of Advanced Manufacturing Technology*, vol. 117, pp. 1327-1349, 2021.
- [8] A. Nawaz, J. Wu, J. Ye, Y. Dong, and C. Long, Distributed MPC-Based Energy Scheduling for Islanded Multi-microgrid Considering Battery Degradation and Cyclic Life Deterioration, *Applied Energy*, vol. 329, 120168, 2023.
- [9] Y. Yang, H-G Yeh, R. Nguyen, A Robust Model Predictive Control-Based Scheduling Approach for Electric Vehicle Charging With Photovoltaic Systems, *IEEE Systems Journal*, vol. 17, pp. 111-121, 2022.
- [10] C. Chen, J. Wang, Y. Heo, S. Kishore, MPC-Based Appliance Scheduling for Residential Building Energy Management Controller, *IEEE Transactions on Smart Grid*, vol. 4, pp. 1401-1410, 2013.
- [11] A. Bozzi, S. Graffione, R. Sacile and E. Zero, Dynamic MPC-Based Scheduling in a Smart Manufacturing System Problem, *IEEE Access*, vol. 11, pp. 141987-141996, 2023.
- [12] D. A. Alexander, A. Pandey, and S. K. Kumar, Mathematical Models for Multi-stage Hybrid Assembly Flow-shop Scheduling with Preventive Maintenance and Release Times, *Computers & Industrial Engineering*, vol. 186, 109716, 2023.