A real-time production scheduling method for RFID-enabled semiconductor back-end shopfloor environment in industry 4.0

Mingyao Sun and Tianhua Zhang
College of Business Administration, Capital University of Economics and Business, Beijing, China

Received 8 May 2023
Revised 19 June 2023
Accepted 17 July 2023

Abstract

Purpose – A real-time production scheduling method for semiconductor back-end manufacturing process becomes increasingly important in industry 4.0. Semiconductor back-end manufacturing process is always accompanied by order splitting and merging; besides, in each stage of the process, there are always multiple machine groups that have different production capabilities and capacities. This paper studies a multi-agent based scheduling architecture for the radio frequency identification (RFID)-enabled semiconductor back-end shopfloor, which integrates not only manufacturing resources but also human factors.

Design/methodology/approach – The architecture includes a task management (TM) agent, a staff instruction (SI) agent, a task scheduling (TS) agent, an information management center (IMC), machine group (MG) agent and a production monitoring (PM) agent. Then, based on the architecture, the authors developed a scheduling method consisting of capability & capacity planning and machine configuration modules in the TS agent.

Findings – The authors used greedy policy to assign each order to the appropriate machine groups based on the real-time utilization ration of each MG in the capability & capacity (C&C) planning module, and used a partial swarm optimization (PSO) algorithm to schedule each splitting job to the identified machine based on the C&C planning results. At last, we conducted a case study to demonstrate the proposed multi-agent based real-time production scheduling models and methods.

Originality/value – This paper proposes a multi-agent based real-time scheduling framework for semiconductor back-end industry. A C&C planning and a machine configuration algorithm are developed, respectively. The paper provides a feasible solution for semiconductor back-end manufacturing process to realize real-time scheduling.

Keywords Real-time scheduling, Multi-agent, Capability & capacity planning, Machine configuration, Semiconductor

Paper type Research paper

1. Introduction

The semiconductor manufacturing process includes wafer fabrication, probing, assembly and final testing steps. The back-end process of semiconductor manufacturing consists of steps that follows water fabrication (Guo, Chiang, & Pai, 2007; Tu & Chen, 2009; Lin & Chen, 2015). Because semiconductor back-end manufacturing process usually operates with a short

This work was funded in part by the National Natural Science Foundation of China under No. 72201178; Research start-up fund project for new hired faculty of Capital University of Economics and Business under No. XRZ2022023.

IIMBG Journal of Sustainable Business and Innovation
Vol. 1 No. 1, 2023
pp. 39-57
Emerald Publishing Limited
2753-4022
DOI 10.1108/IJSBI-05-2023-0027

lead-time and absorbs the effect of lead time variance in the front-end process, the back-end manufacturing environment is of highly uncertainties and dynamics (Lin & Chen, 2015; Fu, Askin, Fowler, & Zhang, 2015). A typical back-end manufacturing process consists of multiple stages, including die bond (DB), wire bond (WB), molding, marking and final vision. Among these stages, die bond, wire bond and molding, are the potential bottlenecks of the manufacturing process (Lin & Chen, 2015). On-time delivery is considered as one of the most important performance indicators for semiconductor industries. Hence, how to develop a real-time scheduling framework to minimize the total weighted tardiness is critical for business success of semiconductor industries in industry 4.0.

The production process of back-end shopfloor belongs to hybrid flow shop (HFS) environment with the following features: (1) Machine heterogeneity. In order to satisfy the precision requirements of different orders, there are multiple types of machines in each back-end manufacturing stage (Lin & Chen, 2015; Lin, Chen, Chiu, & Fang, 2013). The same type machines in one stage are referred to as "machine group", while all types of machines in one stage are referred to as "machine family" in our study; (2) Order splitting and merging during the production. DB process picks the die from wafers and put it on the substrate, after which the substrate is placed into a magazine (Lin & Chen, 2015; Park, Ahn, & Hur, 2018). Magazines are carriers that transfer the jobs from one stage to the next stage. The process is repeated until the quantity requirement of one magazine is satisfied. As a result, the order is split into jobs for the WB and molding stages. All of the jobs split from the order are processed on the same machine group at each stage and are merged after completion. Hence, the flow time of the order begins from the order releasing and ends when all subsequent jobs are completed. The order splitting and merging makes the orders scheduling be more complex in semiconductor back-end industries (Hung, Liang, & Chen, 2013; Chiu, Lai, & Chen, 2023; Lin & Chen, 2015).

Despite of the significant research progress in semiconductor back-end shopfloor scheduling (e.g. Lin & Chen, 2015; Fu *et al.*, 2011; Wang, Lin, Liang, & Wang, 2023), the following questions are still unsettled:

(1) Development of industry 4.0 technologies (e.g. IIoT, AI and big data analytics) makes real-time scheduling possible (Ghaleb, Zolfagharinia, & Taghipour, 2020; Zhang, Tang, Li, Liu, & Zhang, 2021; Hu, Jia, He, Fu, & Liu, 2020). Since the semiconductor back-end manufacturing process has high degree of dynamics and complexity, it is essential to develop a real-time scheduling framework for the dynamic decision-making and adaptive control capabilities to deal with the production changes rapidly. However, existing literature that focused on the real-time scheduling framework (e.g. Zhang, Huang, Sun, & Yang, 2014; Negri *et al.*, 2021; Wang, Liu, Ren, Wang, & Wang, 2021) did not characterize the features of semiconductor back-end manufacturing process, e.g. machine heterogeneity and order splitting and merging.

(2) Few studies pay attention to the machine capability and capacity planning in the semiconductor back-end manufacturing process. On the one hand, different IC products require different process precisions in each manufacturing stage; thus multiple machine capabilities are needed in every stage. On the other hand, there are multiple types of machine groups that have different capabilities in each stage (Lin & Chen, 2015). Therefore, to design a real time dispatching architecture and method for assigning each order to a proper machine group, according to the products' capability requirements and spare capacity of the machine groups, has significant value and wild application foreground.

(3) Due to the lack of workers' real-time status information capturing, human control and integration in manufacturing environment is neglected in the current multi-agent

based scheduling literature, e.g. Zhang *et al.* (2014) and Renna (2011). Unlike other resources, human control is always hard to be integrated due to the positive initiative of human (Lin, Li, Ma, Yao, & Lu, 2020; Marichelvam, Geetha, & Tosun, 2020). In the semiconductor back-end manufacturing process, the positions of workers are not fixed because each of them operates with several machines (Wong *et al.*, 2010). Owing to the development of real-time monitoring and manufacturing execution systems (MES), it becomes realistic to integrate human factors (e.g. idle or not, technique levels and capabilities) in the manufacturing process. Nevertheless, to the best of our knowledge, few studies have focused on designing the real-time scheduling framework with human factors in the setting of semiconductor back-end industry, especially considering the features of machine heterogeneity and order splitting and merging.

To address these problems, this study establishes a flexible and adaptive production planning and control system for real-time scheduling of semiconductor back-end manufacturing processes based on the multi-agent systems. We first develop a multi-agent manufacturing system framework, including a machine group (MG) agent, information management center (IMC), Task scheduling (TS) agent, staff instruction (SI) agent, task management (TM) agent and production monitoring (PM) agent. Then, based on the developed architecture, we propose an order assignment method using greedy policy in the TD agent according to the machine groups' capabilities and spare capacities, to optimally balance the workload of each machine group. Lastly, we develop a global scheduling model for machine configuration of bottleneck stages to guarantee the delivery accuracy in semiconductor back-end industries.

The rest of the paper is arranged as follows: Section 2 reviews the relevant literature. Section 3 establishes the multi-agent based scheduling framework of the semiconductor back-end manufacturing process. Section 4 develops the C&C method and machine configuration algorithms in the TS agent. Section 5 uses a case study to illustrate our framework and models intuitively. Section 6 concludes the study and presents the future research directions.

2. Literature review

Three streams of literature are relevant to our study: those related to (1) multi-agent systems in manufacturing environment; (2) scheduling methods of HFS; and (3) work allocation in semiconductor back-end manufacturing process. In this section, we briefly review the studies related to these streams and discuss how our study differs from them to highlight our contributions.

Our research is first related to studies on multi-agent systems in manufacturing environment. As a branch of artificial intelligence (AI), agent technology has been broadly adopted and developed in manufacturing applications (Zhang *et al.*, 2014; Kamali, Banirostam, Motameni, & Teshnehlab, 2023; Popper & Ruskowski, 2022). Zhang and Wang (2016) developed a multi-agent based hierarchical collaborative scheduling system, which outperforms the FCFS rule in terms of daily movement and machine utilization. Chol and Gun (2023) studied a multi-agent based scheduling method for tandem automated guided vehicle systems. Mishra, Singh, Kumari, Govindan, and Ali (2016) developed a self-reactive cloud-based multi-agent architecture for distributed manufacturing system, which can help manufacturing industries to establish real-time information exchange between the autonomous agents, clients, suppliers and manufacturing unit. Wang, Zhang, Zhang, Cui, and Zhang (2022) proposed an independent double deep-q-network-based multi-agent reinforcement learning (MA-IDDQN) approach to produce an adaptive rule for batch forming and scheduling for a two-stage HFS scheduling. Our paper differs from these studies in that we consider a C&C allocation agent for MGs in each manufacturing stage. Moreover, despite the machines and materials in the shopfloor, our multi-agent based manufacturing system also integrates the human factors.

Our research is also related to the literature on scheduling methods of HFS. Semiconductor back-end manufacturing process is a typical case of HFS (Lin & Chen, 2015). Ruiz and Vázquez-Rodríguez (2010) reviewed the heuristic and metaheuristic methods that have been proposed for scheduling of HFS. In recent years, Shao, Shao, and Pi (2020) combined the features of distributed flow shop scheduling and parallel machine scheduling to develop a solution for the distributed hybrid flowshop scheduling problem (DHFSP) with makespan criterion. Cai, Lei, Wang, and Wang (2023) developed reinforcement learning method for distributed hybrid flowshop scheduling problem (HFSP). Lu, Liu, Zhang, and Yin (2022) studied a Pareto-based hybrid iterated greedy algorithm for a HFSP with objectives of minimization the makespan and total energy consumption. Similarly, Zheng, Zhou, Xu, and Chen (2020) also studied the scheduling method for HFS with energy consumption consideration. Gheisariha, Tavana, Jolai, and Rabiee (2021) developed a simulation–optimization model for solving multi-objective HFS scheduling problem with rework and transportation.

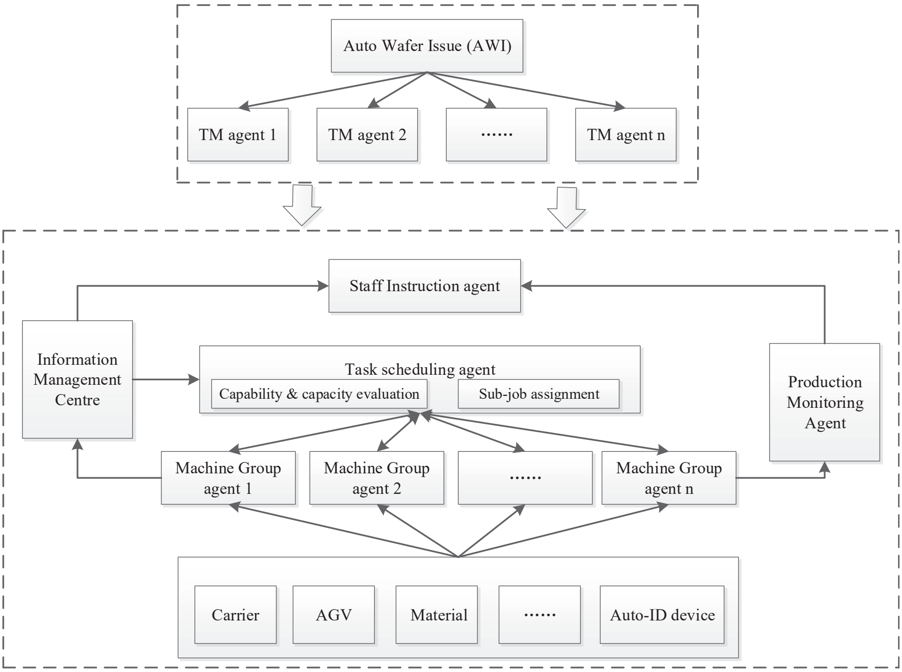
Finally, our research is also related to studies on work allocation in semiconductor back-end manufacturing process. Weigert, Klemmt, and Horn (2009) designed heuristic algorithms for simulation-based scheduling of a semiconductor back-end facility with the consideration of multiple objectives. Kress and Müller (2022) presented decomposition-based heuristic solution approaches and a mixed integer program to minimize the total weighted tardiness for a semiconductor final-test scheduling problem, by considering human operators with setup operations. Deenen, Adan, and Akcay (2020) studied the problem of allocating semiconductor wafers to customer orders with the objective of minimizing the overallocation prior to assembly, by considering that a wafer can contain dies from several different die classes. Fu *et al.* (2011) developed a new mixed-integer-linear-programming (MILP) model for the batch production scheduling of a semiconductor back-end facility with serial production stages. Wang *et al.* (2023) presented a multi-subpopulation parallel computing genetic algorithm for the semiconductor packaging scheduling problem with auxiliary resource constraints. Our paper is most related to Lin and Chen (2015), who presented a simulation–optimization approach for a hybrid semiconductor back-end manufacturing process scheduling problem by considering job splitting and merging. While also focusing on order splitting and merging, we aim to develop a multi-agent based real-time scheduling framework for semiconductor back-end manufacturing process. In addition, capability and capacity planning is also involved in our study.

In summary, (1) different from most of previous studies on multi-agent based scheduling systems, we involved a C&C planning allocation agent and a human control agent to suit the semiconductor back-end manufacturing process; (2) Unlike previous studies of HFS scheduling, we proposed a new coding mechanism for partial swarm optimization (PSO) algorithm that involves both MG and machine configuration decisions; and (3) We make the first attempt to combine order splitting and merging and HFS scheduling, which is more consistent with the semiconductor back-end manufacturing process than other studies.

3. Multi-agent based semiconductor back-end manufacturing environment

3.1 Multi-agent system architecture

The overall architecture of the multi-agent based real-time scheduling system for the semiconductor back-end manufacturing process is shown in Figure 1. The proposed



Source(s): Figure by authors

Figure 1.
System architecture

architecture includes a TM agent, a SI agent, a TS agent, an IMC, a MG agent and aPM agent. First, the dynamic manufacturing data (e.g. resources, staff and equipment) will be captured by the auto ID technologies, such as radio frequency identification (RFID) and Bluetooth. Then, the information management agent will analyze these data to extract more valuable information for further analysis. Lastly, the tasks could be well scheduled to the identified machines according to the real-time status of machines. MG agent and PM agent are used to manage the machines and monitor the manufacturing process. We will illustrate the architecture and each agent clearly in Section 3.2.

Based on the system architecture in Figure 1, we analyze the information interactive mechanism of the multi-agent based real-time scheduling method in semiconductor manufacturing process in Figure 2. First, auto wafer issue system (AWI) releases the order and a corresponding TM agent will be created and it will send the key information of the released order to the IMC. Then, IMC will process the information with real-time shopfloor data to acquire value-added information and provide them to the SI agent and TD agent. Based on the value-added information, operator and sub-job scheduling results will be transferred to MG agent, respectively, by SI agent and TD agent to instruct the production process. PM agent will track and monitor the manufacturing process whole time, when disturbance happens, SI agent will be informed and repair/maintenance activities will be arranged.

3.2 Multi-agent models

The multi-agent system consists of five parts, which are briefly described as follows:

(1) Task management (TM) agent

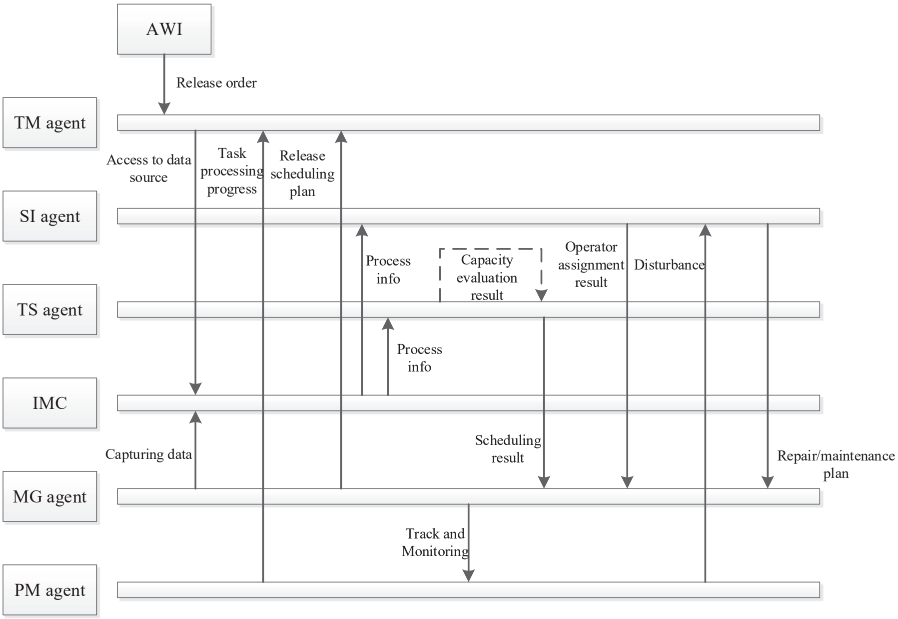


Figure 2. Information interactive mechanism for the system architecture

Source(s): Figure by authors

A TM agent is created once an order is issued by AWI, which is an important industrial application widely used in semiconductor back-end industry. The TM agent records the key information of the order, e.g. priority, quantity, required machine capability in each stage and due date. The TM agent for the order will be cancelled by the AWI as soon as the order finishes all the manufacturing processes.

(2) Staff instruction (SI) agent

"Unmanned shopfloor" becomes increasingly achievable because of the development of mobile network and wearable devices. Hence, SI agent is aimed for integrating human factors into the manufacturing system but keeping operators away from the shopfloor by the use of industrial hardware and software, e.g. central computer, smart glasses, mobile internet. Service-oriented architecture (SOA) is adopted in our architecture for the real-time and multi-source data transmission. The information flow can be described as follows: first, the new released order information will be formed an XML-based schema that contains all the key parameters needed to be monitored. Then, based on the SOA, the related XML files will be sent to the central web server and interact with operator calling system database (OCS DB). OCS DB stores all the status information of workers, including their ID, profession and skill levels, idle or not. Finally, web server will send the optimized results to the appropriate operators according to their profession and skill levels through Hypertext Transport Protocol (HTTP), which can help workers or shop managers make order scheduling decisions, monitor the production performance or equipment health status. By this way, staffs of the factory are well integrated into the manufacturing system, which contributes to the decline of error costs caused by human.

(3) Task scheduling (TS) agent

There are two main functions of TS agent, namely C&C evaluation and machine configuration. The CC evaluation is aimed for assigning each order to the proper machine family in each stage, according to the real-time data captured by machine agents. The workflow of this module can be described as follows: first, based on the precision requirement of the orders, the capability & capacity evaluation module will select the appropriate machine groups and forms a potential set according to machines' capabilities. Then, the total loading of the MGs in the potential set will be calculated by the agent, and the MG with least loading rate will be chosen for the order. The algorithm of the C&C evaluation is shown in section 4.2.

The CC evaluation result will be transferred to the machine configuration module, which assigns each operation to the identified machine definitely. In order to improve the satisfaction of customers, delivery accuracy has been considered as one of the most important performance indicators in semiconductor industries (Lin & Chen, 2015). Thus, we develop a global optimization mechanism aimed for minimizing the total tardiness and earliness of all orders in machine configuration module. The detailed algorithm is illustrated in section 4.3.

(4) Information management center (IMC)

IMC consists of two modules, namely data storage module (DSM) and real-time data processing module (DPM). DSM aims to provide unified data management platform for other agents. Three main functions are realized: (1) providing data storage for the captured real-time data and historical data; (2) mapping relation between the physical resources and virtual resources, e.g. matching RFID tags and shopfloor resources; and (3) providing information interactive and integration platform for varieties sources of information systems, e.g. ERP, SCM, CRM and MES. DPM aims to do data analytics based on the DSM module, such as data cleaning, data classification & fusion and information integration, so as to extract more value-added information from the massive data generated from shopfloor and information systems.

(5) Machine group (MG) agent

MG agent is developed for real-time data collecting and product recipe management. Real-time data collecting module is responsible for providing standard methods for heterogeneous auto-ID devices so that their perception functions can be easily invoked under a uniform model. Two standard methods, namely "reading data (Parameter [1], Parameter [i])" and "writing data (Parameter [1], Parameter [i])" can be found in Zhang, Huang, Sun, and Yang (2014).

Recipe management is designed to enhance the intelligence of the machine agent. Recipe means the processing spec of each product in semiconductor industries. Once the machine agent receives a production task, it can autonomously recognize the type of the product and release the corresponding recipe. In addition, recipe management can check whether the current operator is qualified to finish the operation.

(6) Production monitoring (PM) agent

PM agent is adopted to collect real-time information of manufacturing resources to track and monitor the production process. When any disturbance happens, the PM agent will warn and identify the root cause of the disturbance and send it to staff agent; based on which, the staff agent will assign the repair/maintenance task to appropriate operator according to their skills and levels.

4. Scheduling models and algorithms in TS agent

In this section, an order scheduling rule is designed for production planning in the shopfloor of semiconductor back-end industries. As described in section 3.3, the whole production planning process consists of C&C planning and machine configuration steps. C&C planning aims to choose an optimal MG according to the real-time status of machines for an order, and machine configuration aims to sequence each order on the identified machines (including order splitting and merging). C&C planning method and machine configuration method in the TS agent are discussed, respectively, in the following text. Some terminologies should be distinguished again before we establish the scheduling model.

Machine group: Refers to a specific machine type that has one or multiple capabilities for different product precision. Various MGs are existed in each stage.

Product type: Refers to a specific device type which requires specific processing precision and counter-stochastic processing time caused by machine availability.

Order: Refers to a specific customer demand for a specific product type with a scheduled release time and due date.

Jobs: An order is spitted into several equivalent jobs. For example, if an order splitting with the way "1/3", it means that the order is equally spitted into three jobs.

4.1 Notations and variables

First, we introduce some notations and variables in this paper.

4.1.1 Indices and set.

n: Index of orders, $n \in \{1, 2, \ldots, N\}$;

s: Index of stages, $s \in \{1, 2, 3\}$. In this text, "1" represents die bond stage, and "2" and "3" represent wire bond and molding stage, respectively;

(s, k): MG k in stage s, $k \in \{1, 2, \ldots, K_s\}$, where K_s is the total MG in stage s.

(s, k, m): Machine m of group k in stage s, $m \in \{1, 2, \ldots, M_k\}$, where M_k is the total number of machines in group k;

h: Index of order splitting way, $h \in \{1, 2, \ldots, H\}$, where H depends on the machine numbers in a MG;

$C_{s,k}$: Capability set of MG k in stage s, $C_{s,k} = \{C^1_{s,k}, C^2_{s,k}, \ldots, C^r_{s,k}\}$.

4.1.2 Input data.

Q_n: Batch quantity of order n, which is measured by total number of magazines;

$PT_{(s,k)}$: Unit processing time of MG k in stage s;

$ST_{n',n}$: Sequence-dependent setup time when order n' is a direct predecessor of work order n;

$A_{n,s}$: Processing capability requirement of order n in stage s;

D_1: Penalty coefficient when an order is finished after due date;

D_2: Penalty coefficient when an order is finished before due date;

DD_n: Due date of order n.

4.1.3 Decision variables.

$C_{n,s,(s,k,m)}$: Completion time of order n on machine (s, k, m) in stage s;

$X_{n,s,(s,k,m)}$: Processing rate of order n on machine (s, k, m) in stage s;

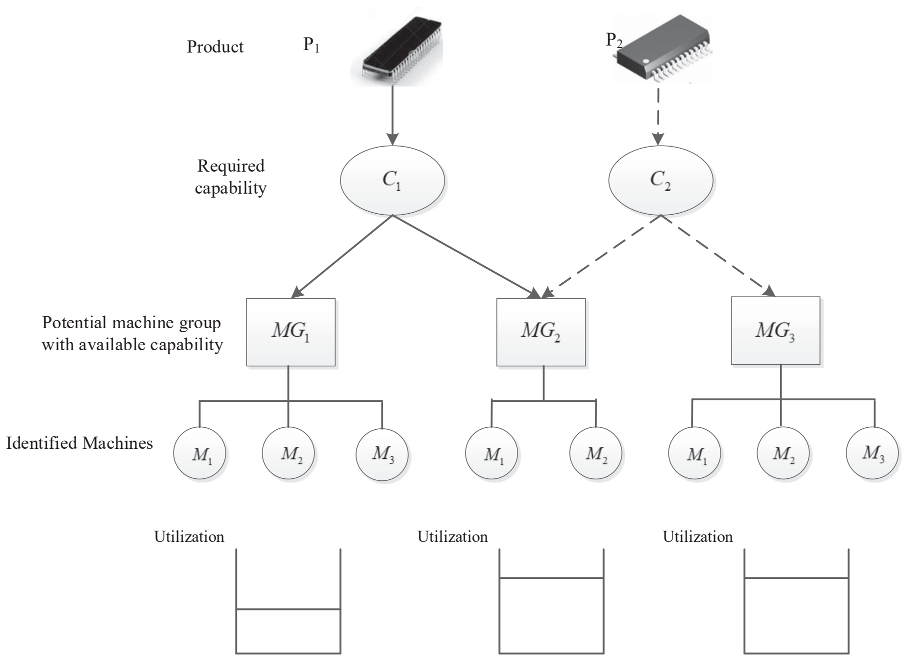
$$Y_{n,s,(s,k,m)} = \begin{cases} 1 & \textit{In stage } s, \textit{order } n \textit{ is processed on machine } (s,k,m) \\ 0 & \textit{Otherwise} \end{cases};$$

$$Z_{n,s,h} = \begin{cases} 1 & \textit{In stage } s, \textit{order } n \textit{ is processed with order splitting way } h \\ 0 & \textit{Otherwise} \end{cases};$$

$$O_{n,n',s,(s,k,m)} = \begin{cases} 1 & \textit{In stage } s, \textit{order } n \textit{ is processed before order } n' \textit{ on same machine} \\ 0 & \textit{Otherwise} \end{cases}.$$

4.2 Capability and capacity planning with greedy policy in TS agent

In semiconductor back-end shopfloor, there are many MGs in each stage and machine groups have various capabilities that are suitable to the precision requirements of different orders. Hence, C&C planning of these machine groups are critical to the whole scheduling process. A MG with multiple capabilities is likely to become the bottleneck of the whole production process. An example of MG selection can be shown in Figure 3. In one manufacturing stage, machine groups MG_1, MG_2 and MG_3 belong to a machine family. Product P_1 requires machine capability C_1 and Product P_2 requires machine capability C_2, where C_1 can be provided by machine group MG_1 and MG_2 and C_2 can be provided by machine group MG_2 and MG_3. Because machine group MG_2 can process both products, capacity planning is necessary in case that MG_2 becomes a bottleneck of the whole process. In Figure 3, machine groups MG_1 and MG_2 are selected as potentials of product P_1 in this stage. After comparing the remaining capacities of MG_1 and MG_2, product P_1 is finally assigned to MG_1.



Source(s): Figure by authors

Figure 3. An example of machine group selection

The detailed C&C planning algorithm based on the greedy policy is shown in the following, where *potential* represents the set of selectable machine groups (according to machines' capabilities and products' capability requirements):

$\forall s$;
For $n = 1$ to N
Read WIP related data from IMC.
Get $A_{n,s}$ from the TM agent
For $k = 1$ to K_s
Get $C_{s,k} = \{C^1_{s,k}, C^2_{s,k}, \ldots, C^r_{s,k}\}$ from machine capability set;
If $A_{n,s} \in C_{s,k}$
Add MG k to potential MGs set *potential*;
Calculate the total amount of elements in *potential* as I;
End if.
For $i = 1$ to I
Calculate spare capacity of MG i with $AC_{(s,i)} = \sum_{(s,i,m)=1}^{(s,i,M_k)} AC_{(s,i,m)}$, in which $AC_{(s,i,m)}$ is captured by MA agent;
Calculate available capacity ratio of MG i, $ACR_{(s,i)} = AC_{(s,i)}/TC_{(s,i)}$, where $TC_{(s,i)}$ represents total capacity of MG i;
Select $max\{ACR_{(s,i)} | i \in potential\}$ as the proper MG for order n;
Next.

4.3 Machine configuration rule in the TS agent

4.3.1 Machine configuration model based on C&C planning result. As mentioned in section 1, order splitting and merging are the main shopfloor features of semiconductor back-end industries. The way of order splitting may be stage-dependent. For example, in WB stage, an order is divided into two jobs and processed on two identified machines; however, in molding stage, the same order may be divided into three jobs and processed on three identified machines.

In machine configuration stage, C&C planning results have been already known. Thus, we use $\delta_{(n,s)}$ represents for the MG that order n selects in stage s. The objective of the scheduling problem is to minimize the total weighted earliness/tardiness. Let F_1 represent the total tardiness of all orders and F_2 stands for the total earliness, we can get:

$$F_1 = D_1 \sum_{n=1}^{N} max\left(\max_{1 \le (3,k,m) \le M_k} C_{n,3,(3,k,m)} - DD_n, 0\right); \tag{1}$$

$$F_2 = D_2 \sum_{n=1}^{N} max\left(DD_n - \max_{1 \le (3,k,m) \le M_k} C_{n,3,(3,k,m)}, 0\right). \tag{2}$$

Therefore, the machine configuration model can be written as:

$$min\, F = F_1 + F_2; \tag{3}$$

s.t.

$$C_{n,s-1,\left(s-1,k',m'\right)} + Q_n \times X_{n,s,(s,k,m)} \times PT_{(s,k)} \le C_{n,s,(s,k,m)} \forall \mathrm{n, s};\ m' \in \left\{1, 2, \ldots, M_{\delta_{(n,s-1)}}\right\}; \tag{4}$$

$$m \in \left\{1, 2, \ldots, M_{\delta_{(n,s)}}\right\}; \tag{5}$$

$$\sum_{h=1}^{H} Z_{n,s,h} = 1\ \forall n, s; \tag{6}$$

$$X_{n,s,(s,k,m)} = \frac{1}{\sum_{(s,k,m)=1}^{M_k} Y_{n,s,(s,k,m)}}\ \forall n, s,\ (s,k,m) \in \delta_{(n,s)}; \tag{7}$$

$$\sum_{(s,k,m)=1}^{M_k} X_{n,s,(s,k,m)} = 1\ \forall n, s,\ (s,k,m) \in \delta_{(n,s)}; \tag{8}$$

$$\begin{aligned} C_{n',s,(s,k,m)} \geq\ & C_{n,s,(s,k,m)} + Q_{n'} \times X_{n',s,(s,k,m)} \times PT_{(s,k)} + ST_{n',n} - L\left(2 - Y_{n,s,(s,k,m)} - Y_{n',s,(s,k,m)}\right) \\ & - L\left(1 - O_{n,n',s,(s,k,m)}\right); \end{aligned} \tag{9}$$

$$\begin{aligned} C_{n,s,(s,k,m)} \geq\ & C_{n',s,(s,k,m)} + Q_{n} \times X_{n,s,(s,k,m)} \times PT_{(s,k)} + ST_{n,n'} - L\left(2 - Y_{n,s,(s,k,m)} - Y_{n',s,(s,k,m)}\right) \\ & - L \times O_{n,n',s,(s,k,m)}; \end{aligned} \tag{10}$$

Equation (3) states that the objective is to minimize the total tardiness/earliness of all orders. Equation (4) indicates that finishing time of order n in stage s is equal to or great than the completion time of order n in stage $s-1$. Equation (5) represents that an order has to adopt one splitting way in every stage. Equation (6) ensures an order has a unique order splitting way in one stage. Equation (7) shows the relationship between the order production rate and machine quantities the order used. Equation (8) states that each order has to be processed completely in each stage. Equation (9) and Equation (10) are sequencing constraints and must be imposed on a pair of work orders n' and n only if they are allocated to the same machine (s, k, m).

4.3.2 PSO-based algorithm in machine configuration. When using PSO-based algorithm to deal with scheduling problems, it is critical to design an encoding and decoding schema to construct particles for the specific problem. In this paper, we develop a new matrix coding method to generate feasible particles. In particular, the specific coding schema can be seen as following:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}, \tag{11}$$

where a_{ij} is uniformly distributed in $\left(a, a + \frac{2^{R_{(s,a)}-1} + 2^{R_{(s,a)}-2} + \ldots + 2^0}{10^{\theta}}\right)$. a denotes the MG and $R_{(s,a)}$ denotes the total number of machines in the MG and $\theta = 2^{R_{(s,a)}-1} + 2^{R_{(s,a)}-2} + \ldots + 2^0$. C&C planning results have obtained in section 4.2 and machine quantities of each MG is fixed; hence, the distribution of a_{ij} is determined. In our encoding method, the integral part of a_{ij} stands for the selected machine group of order i in stage j, and the fractional part of a_{ij} represents the machine configuration (order splitting way) of order i in stage j on MG a. For example, if an order is encoded as 2.7 in WB stage, the order will be processed on the MG 2; besides, 7 equals 111 in binary, indicating that the order is decomposed into three jobs and processed on three parallel machines of machine group 2.

The steps of the PSO algorithm are shown as follows. Information of each particle can be represented by d dimension vector, in which position can be represented as $X_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ and velocity can be shown as $V_i = (v_{i1}, v_{i2}, \ldots, v_{id})$.

Step 1: Initialization. According to the C&C planning results and machine details, generating the initial particles randomly.

Step 2: Fitness. Evaluating the fitness of each particle in the swarm using the fitness function as shown in equation (3) and find each particle's best value *pbest* and swarm's best value *gbest*.

Step 3: Update. Calculating the velocity and current position of each particle using the following equations:

$$v_{id}(t+1) = w \cdot v_{id}(t) + c_1 \cdot rand_1(\,) \cdot [pbest_{id}(t) - x_{id}(t)] + c_2 \cdot rand_2(\,) \cdot [gbest_{id}(t) - x_{id}(t)]; \tag{12}$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1). \tag{13}$$

Step 4: Selection. If a particle's current fitness is superior to optimal history fitness, substituting it as the optimal fitness and recording the current position as the optimal position. Finding the current optimal solution in each individual swarm and global optimal solution, updating *pbest* and *gbest*.

Step 5: Termination. Stopping the algorithm until the stopping criteria is satisfied; otherwise, returning to step 2.

5. Case study

In this section, we use a case study to demonstrate our multi-agent based scheduling framework for semiconductor back-end manufacturing process. DB, WB and molding stages as bottlenecks of semiconductor back-end shopfloor are considered in this paper.

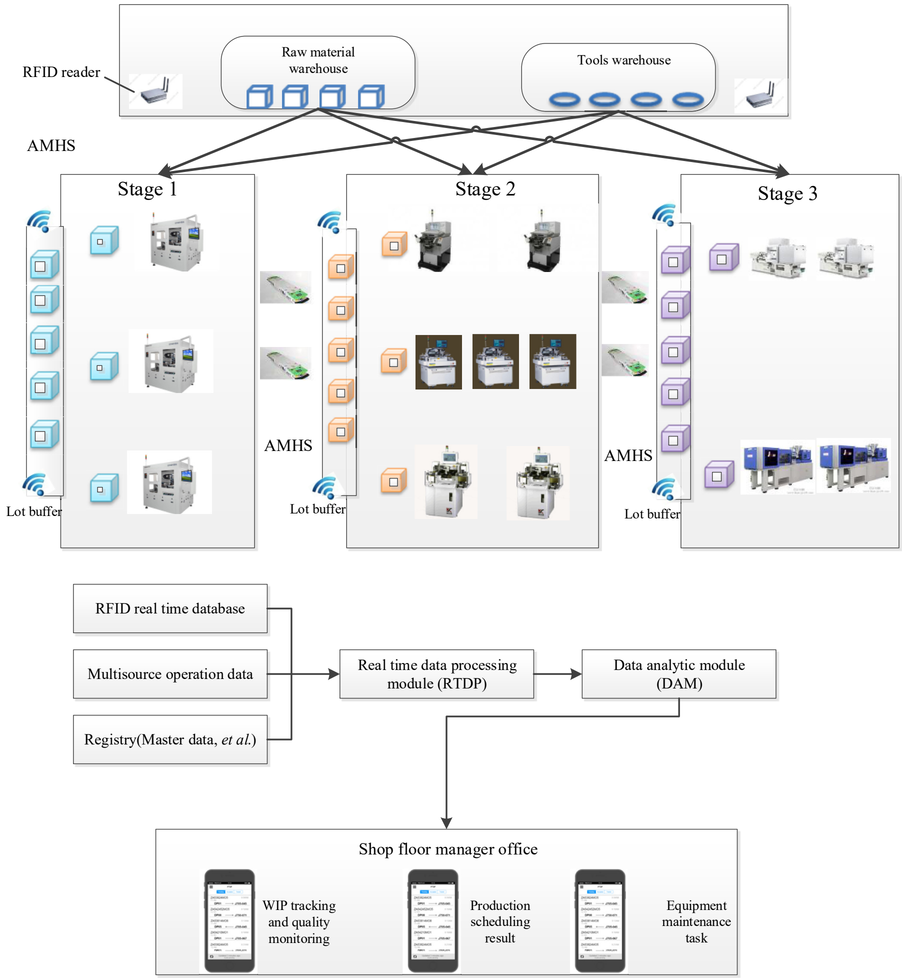
5.1 RFID-enabled intelligent manufacturing environment

Figure 4 shows the overall deployment of the intelligent manufacturing environment considering DB, WB and molding stage.

(1) Each machine in the shopfloor is embedded with RFID readers. This reader is multi-functional and is responsible for reading the tags attached on different passive objects. No matter what materials, tools or WIPs are monitored, the RFID readers will capture the data recorded in the RFID tags immediately.

(2) RFID readers are deployed in the material and tool warehouse, while RFID tags are deployed on each identified material and tool. Thus, when materials or tools are required by shopfloor managers through mobile devices, they will be transported to the production line in time. In addition, the key information recorded in the RFID tags will be captured by the RFID readers and the original data will be transmitted to upper-level applications like MES for scheduling.

(3) In shopfloor production line, RFID readers are deployed optimally at some fixed place so that the radiation range can cover the whole shopfloor. In our framework, RFID readers are deployed in the entrance and exit of the job buffer in each stage. WIPs arrives in the job buffer of each stage continuously, and the deployed RFID readers will record all important information through the RFID tags on the WIPs. The C&C planning results and machine configuration decisions will be displayed on the mobile devices in the shopfloor manager's office, by data analytic model embedded in the TS agent.



Source(s): Figure by authors

Figure 4. RFID-enabled intelligent manufacturing environment

(4) Materials, tools and WIPs are transported by automatic material handling system (AMHS) between each stage. AMHS consists of 3D warehouse, automatic gripping robotic arms, intelligent robots and automatic guided vehicles. By the analytic of multi-source real-time data, managers will give instructions to AMHS remotely according to the real time scheduling results.

5.2 C&C planning and machine configuration results

We conduct a small-scale process simulation to validate our algorithms in C&C planning and machine configuration. In reality, the semiconductor back-end manufacturing process is much more complex due to the large quantities of machines and orders. We make the following assumptions to justify our model.

(1) The order size measured by number of magazines has a uniform distribution between [10,000 and 300,000].

(2) Machine group is pre-determined, which can cover all the order capabilities' requirements. Besides, machine capacity and unit processing time are both randomly chosen according to machine groups' characteristics.

(3) Machine numbers of each group are randomly determined by different uniform distribution.

(4) Each order has only one type of products and all the orders in this simulation are different from each other.

The designed problem has ten orders and each order must undergo three processing stages (i.e. DB, WB and molding). The details of each order can be seen from Table 1. Note that we have normalized the start time and due date of each order in Table 1. Table 2 shows detailed information of machines in each stage, including MGs, machine numbers, machine capabilities, unit processing time of each MG.

The procedure of real time scheduling in this case includes two main steps, they are described as flows:

(1) At first, orders are assigned to different MGs according to the real status of different machine groups. For each process, a potential machine groups set will be firstly established according to the orders' required capability. Then, utilization of each MG in the potential set will be calculated and evaluated by the method in the TS agent. Furthermore, the machine group with the maximum available capacity ratio in the potential set will be selected. The C&C planning results are shown in Table 3. Based on the results of Table 3, we make a comparison of loading rate of each MG between the C&C planning method and stochastic dispatching method, as shown in Figure 5.

(2) After all the process of all the orders optimally assigned to MGs, TS agent is ready for scheduling the tasks (i.e. machine configuration) using PSO based algorithm designed in section 4.3.2. Table 4 shows the scheduling results of the data in Table 1.

In Table 4, the first row represents the processing stages machine configuration of semiconductor back-end manufacturing and the first column represents the order numbers. In columns 2~4, the data (x, y, z) of the 'j' row and 'i' column means the order i is manufactured in MG x from time y to time z at manufacturing stage j. For example, (WB_T03, 2, 10) of row 3 and column 3 represents that order 2 is manufactured in MG WB_T03 from time eight to ten

Orders	Earliest start time/day	Due date/day	Order size (mag.)	Required capability in each stage
1	0	96	168420	C_{11}, C_{24}, C_{32}
2	4	13	11403	C_{13}, C_{21}, C_{31}
3	9	17	23910	C_{11}, C_{23}, C_{32}
4	15	63	68970	C_{12}, C_{25}, C_{33}
5	19	73	138960	C_{11}, C_{26}, C_{31}
6	21	90	170351	C_{12}, C_{22}, C_{34}
7	28	55	15503	C_{13}, C_{26}, C_{31}
8	32	81	19921	C_{12}, C_{25}, C_{32}
9	33	87	52160	C_{11}, C_{24}, C_{34}
10	45	90	18630	C_{12}, C_{24}, C_{33}

Source(s): Table by authors

Table 1.
Detailed information of ten orders

Stages	Machine group	Numbers	Capabilities	Capacity/(mag.)	Unit processing time (day/mag)
DB	DB_T01	10	C_{11}	40,000	2/(60*24)
	DB_T02	18	C_{12}	20,000	2/(60*24)
	DB_T03	5	C_{13}	20,000	3/(60*24)
WB	WB_T01	51	C_{23}, C_{25}	10,000	3/(60*24)
	WB_T02	45	C_{24}	10,000	5/(60*24)
	WB_T03	47	C_{21}, C_{24}	10,000	2/(60*24)
	WB_T04	40	C_{22}, C_{25}, C_{26}	10,000	2/(60*24)
	WB_T05	50	C_{26}	10,000	3/(60*24)
MD	MD_T01	8	C_{32}	50,000	5/(60*24)
	MD_T02	5	C_{31}, C_{32}, C_{33}	40,000	2/(60*24)
	MD_T03	8	C_{31}, C_{34}	50,000	2/(60*24)

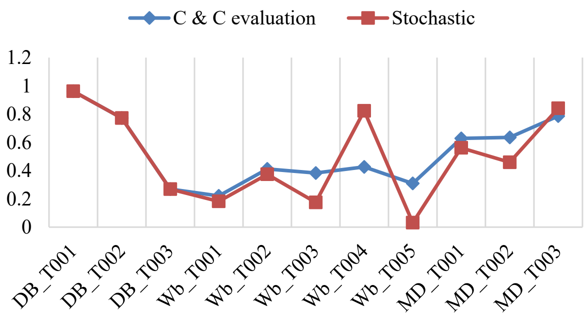
Source(s): Table by authors

Table 2. Machine group information

	1	2	3	4	5	6	7	8	9	10
DB	T01	T03	T01	T02	T01	T02	T03	T02	T01	T02
WB	T03	T03	T01	T01	T05	T04	T05	T01	T02	T02
MD	T01	T03	T02	T02	T03	T03	T02	T01	T03	T03

Source(s): Table by authors

Table 3. Machine group dispatching result according to the procedure in section 3.3



Source(s): Figure by authors

Figure 5. Loading rate comparison between C&C planning and stochastic dispatching

times at the WB stage. In the last column, the data (x, y, z) represents the machine configuration results of order i. For example, (1, 23, 3) of row 4 and column 5 represents that the order 3 uses 1 machine of group DB_T01, 23 machines of group WB_T01 and 3 machines of group MD_T02.

6. Conclusions

6.1 Concluding remarks

Real-time scheduling becomes increasingly important for semiconductor back-end manufacturing process in industry 4.0. Different from other manufacturing processes, order splitting and merging are common in semiconductor back-end manufacturing processes. Besides, in the shopfloor, there are multiple types of machines in each processing

Orders	DB stage	WB stage	MD stage	Machine configuration
1	(DB_T01, 0, 55)	(WB_T03, 55, 68)	(MD_T01, 70, 96)	(5, 17, 6)
2	(DB_T03, 4, 8)	(WB_T03, 8, 10)	(MD_T03, 10, 11)	(2, 20, 1)
3	(DB_T01, 9, 13)	(WB_T01, 13, 15)	(MD_T02, 15, 18)	(1, 23, 3)
4	(DB_T02, 15, 25)	(WB_T01, 30, 50)	(MD_T02, 50, 65)	(4, 6, 2)
5	(DB_T01, 19, 43)	(WB_T05, 43, 50)	(MD_T03, 55, 73)	(4, 26, 4)
6	(DB_T02, 21, 40)	(WB_T04, 40, 53)	(MD_T03, 54, 98)	(9, 30, 4)
7	(DB_T03, 28, 33)	(WB_T05, 33, 40)	(MD_T02, 40, 58)	(5, 24, 3)
8	(DB_T02, 32, 41)	(WB_T01, 41, 62)	(MD_T01, 62, 88)	(9, 10, 2)
9	(DB_T01, 33, 60)	(WB_T02, 60, 70)	(MD_T03, 70, 84)	(2, 10, 2)
10	(DB_T02, 49, 63)	(WB_T02, 63, 72)	(MD_T03, 72, 90)	(13, 15, 3)

Source(s): Table by authors

Table 4.
Machine configuration results

stage. Hence, real-time scheduling for semiconductor back-end manufacturing process requires efficient information exchange among physical resources (e.g. machines, tools and materials) and digital resources (e.g. information systems and RFID devices).

This paper proposes a referenced multi-agent based real-time scheduling architecture for semiconductor back-end manufacturing process, in which C&C planning and machine configuration algorithms are developed, respectively. The contributions are summarized as follows. First, an efficient information exchange mechanism is realized based on our multi-agent system architecture. For example, the auto-ID devices deployed at machine side can capture the real-time data of machines, materials, WIPs and tools, which is transmitted to the IMC for further analysis. Second, a TS agent is established for C&C evaluation and machine configuration with the real-time data in the shopfloor. In particular, a C&C planning method based on the real-time utilization of each MG is studied to optimally assign the orders to machine groups in the TS agent. Furthermore, a machine configuration algorithm based on PSO is developed to schedule the tasks with order splitting and merging in the TS agent. Third, a production agent is designed for monitoring and tracking the manufacturing disturbances during the production process. Finally, we conducted a case study to validate the proposed multi-agent architecture and the scheduling algorithms.

6.2 Managerial implications

The theoretical implications of this paper are as follows. First, we proposed a multi-agent based scheduling framework for the semiconductor back-end manufacturing process, which integrates the C&C planning of machine groups and human control that were seldom considered by previous studies. Second, we developed a C&C planning algorithm based on the greedy policy to sub-optimally assign each order to the most appropriate MG. The method balances the workload for each machine group in each stage. Third, to schedule each order and its splitting jobs to the identified machine in each manufacturing stage (i.e. machine configuration decision), we propose a PSO-based algorithm with a new coding mechanism, which determines both order assignment and order splitting and merging ways.

Our study also provides practical implications for the scheduling problems of the semiconductor back-end manufacturing process. On the one hand, the proposed multi-agent based scheduling architecture enables a seamless information flow among different manufacturing stages and information systems, which realizes the real time scheduling and monitoring of all the orders. On the other hand, our proposed algorithms for C&C planning (based on greedy policy) and machine configuration (based on PSO) enable the semiconductor back-end shop floor to balance the workload of each MG and realize the order splitting and merging conveniently.

6.3 Future research directions

The current work can be extended from the following aspects in the future. First, the multi-agent architecture in this paper focuses on the real-time scheduling of orders and machines. Future research should involve real time internal logistics planning and scheduling in the semiconductor back-end shopfloor, especially when AGVs are widely used. Second, the proposed multi-agent architecture and algorithms is specifically developed for semiconductor back-end manufacturing process. Future works should extend the framework in other manufacturing systems that have different process features, e.g. flexible manufacturing systems or lean production systems.

References

Cai, J., Lei, D., Wang, J., & Wang, L. (2023). A novel shuffled frog-leaping algorithm with reinforcement learning for distributed assembly hybrid flow shop scheduling. *International Journal of Production Research*, *61*(4), 1233–1251.

Chiu, C. C., Lai, C. M., & Chen, C. M. (2023). An evolutionary simulation-optimization approach for the problem of order allocation with flexible splitting rule in semiconductor assembly. *Applied Intelligence*, *53*(3), 2593–2615.

Chol, J., & Gun, C. R. (2023). Multi-agent based scheduling method for tandem automated guided vehicle systems. *Engineering Applications of Artificial Intelligence*, *123*. doi:10.1016/j.engappai.2023.106229.

Deenen, P. C., Adan, J., & Akcay, A. (2020). Optimizing class-constrained wafer-to-order allocation in semiconductor back-end production. *Journal of Manufacturing Systems*, *57*, 72–81, distributed manufacturing, International Journal of Production Research, 54(23): 7115-7128.

Fu, M., Askin, R., Fowler, J., Haghnevis, M., Keng, N., Pettinato, J. S., & Zhang, M. (2011). Batch production scheduling for semiconductor back-end operations. *IEEE Transactions on Semiconductor Manufacturing*, *24*(2), 249–260.

Fu, M., Askin, R., Fowler, J., & Zhang, M. (2015). Stochastic optimization of product–machine qualification in a semiconductor back-end facility. *IIE Transactions*, *47*(7), 739–750.

Ghaleb, M., Zolfagharinia, H., & Taghipour, S. (2020). Real-time production scheduling in the Industry-4.0 context: Addressing uncertainties in job arrivals and machine breakdowns. *Computers and Operations Research*, *123*. doi:10.1016/j.cor.2020.105031.

Gheisariha, E., Tavana, M., Jolai, F., & Rabiee, M. (2021). A simulation–optimization model for solving flexible flow shop scheduling problems with rework and transportation. *Mathematics and Computers in Simulation*, *180*, 152–178.

Guo, R. S., Chiang, D. M., & Pai, F. Y. (2007). Multi-objectives exception management model for semiconductor back-end environment under turnkey service. *Production Planning and Control*, *18*(3), 203–216.

Hu, H., Jia, X., He, Q., Fu, S., & Liu, K. (2020). Deep reinforcement learning based AGVs real-time scheduling with mixed rule for flexible shop floor in industry 4.0. *Computers and Industrial Engineering*, *149*. doi:10.1016/j.cie.2020.106749.

Hung, Y. F., Liang, C. H., & Chen, J. C. (2013). Sensitivity search for the rescheduling of semiconductor photolithography operations. *The International Journal of Advanced Manufacturing Technology*, *67*, 73–84.

Kamali, S. R., Banirostam, T., Motameni, H., & Teshnehlab, M. (2023). An immune-based multi-agent system for flexible job shop scheduling problem in dynamic and multi-objective environments. *Engineering Applications of Artificial Intelligence*, *123*. doi:10.1016/j.engappai.2023.106317.

Kress, D., & Müller, D. (2022). Semiconductor final-test scheduling under setup operator constraints. *Computers & Operations Research*, *138*. doi:10.1016/j.cor.2021.105619.

Lin, J. T. & Chen, C. M. (2015). Simulation optimization approach for hybrid flow shop scheduling problem in semiconductor back-end manufacturing. *Simulation Modelling Practice and Theory*, *51*, 100–114.

Lin, J. T., Chen, C. M., Chiu, C. C., & Fang, H. Y. (2013). Simulation optimization with PSO and OCBA for semiconductor back-end assembly. *Journal of Industrial and Production Engineering*, *30*(7), 452–460.

Lin, G., Li, H., Ma, H., Yao, D., & Lu, R. (2020). Human-in-the-loop consensus control for nonlinear multi-agent systems with actuator faults. *IEEE/CAA Journal of Automatica Sinica*, *9*(1), 111–122.

Lu, C., Liu, Q., Zhang, B., & Yin, L. (2022). A Pareto-based hybrid iterated greedy algorithm for energy-efficient scheduling of distributed hybrid flowshop. *Expert Systems with Application*, *204*. doi:10.1016/j.eswa.2022.117555.

Marichelvam, M. K., Geetha, M., & Tosun, Ö. (2020). An improved particle swarm optimization algorithm to solve hybrid flowshop scheduling problems with the effect of human factors–A case study. *Computers and Operations Research*, *114*. doi:10.1016/j.cor.2019.104812.

Mishra, N., Singh, A., Kumari, S., Govindan, K., & Ali, S. I. (2016). Cloud-based multi-agent architecture for effective planning and scheduling of distributed manufacturing. *International Journal of Production Research*, *54*(23), 7115–7128.

Negri, E., Pandhare, V., Cattaneo, L., Singh, J., Macchi, M., & Lee, J. (2021). Field-synchronized Digital Twin framework for production scheduling with uncertainty. *Journal of Intelligent Manufacturing*, *32*(4), 1207–1228.

Park, Y. J., Ahn, G., & Hur, S. (2018). Optimization of pick-and-place in die attach process using a genetic algorithm. *Applied Soft Computing*, *68*, 856–865.

Popper, J. & Ruskowski, M. (2022). Using multi-agent deep reinforcement learning for flexible job shop scheduling problems. *Procedia CIRP*, *112*, 63–67.

Renna, P. (2011). Multi-agent based scheduling in manufacturing cells in a dynamic environment. *International Journal of Production Research*, *49*(5), 1285–1301.

Ruiz, R. & Vázquez-Rodríguez, J. A. (2010). The hybrid flow shop scheduling problem. *European Journal of Operational Research*, *205*(1), 1–18.

Shao, W., Shao, Z., & Pi, D. (2020). Modeling and multi-neighborhood iterated greedy algorithm for distributed hybrid flow shop scheduling problem. *Knowledge-based Systems*, *194*. doi:10.1016/j.knosys.2020.105527.

Tu, Y. M., & Chen, H. N. (2009). Capacity planning with sequential two-level time constraints in the back-end process of wafer fabrication. *International Journal of Production Research*, *47*(24), 6967–6979.

Wang, J., Liu, Y., Ren, S., Wang, C., & Wang, W. (2021). Evolutionary game based real-time scheduling for energy-efficient distributed and flexible job shop. *Journal of Cleaner Production*, *293*. doi:10.1016/j.jclepro.2021.126093.

Wang, M., Zhang, J., Zhang, P., Cui, L., & Zhang, G. (2022). Independent double DQN-based multi-agent reinforcement learning approach for online two-stage hybrid flow shop scheduling with batch machines. *Journal of Manufacturing Systems*, *65*, 694–708.

Wang, H. K., Lin, Y. C., Liang, C. J., & Wang, Y. H. (2023). Multi-subpopulation parallel computing genetic algorithm for the semiconductor packaging scheduling problem with auxiliary resource constraints. *Applied Soft Computing*, *142*. doi:10.1016/j.asoc.2023.110349.

Weigert, G., Klemmt, A., & Horn, S. (2009). Design and validation of heuristic algorithms for simulation-based scheduling of a semiconductor backend facility. *International Journal of Production Research*, *47*(8), 2165–2184.

Wong, S.B. & Richardson, S. (2010). Assessment of working conditions in two different semiconductor manufacturing lines: Effective ergonomics interventions. *Human Factors and Ergonomics in Manufacturing and Service Industries*, *20*(5), 391–407.

Zhang, Y., Huang, G. Q., Sun, S., & Yang, T. (2014). Multi-agent based real-time production scheduling method for radio frequency identification enabled ubiquitous shopfloor environment. *Computers and Industrial Engineering*, *76*, 89–97.

Zhang, J., & Wang, X. (2016). Multi-agent-based hierarchical collaborativescheduling in re-entrant manufacturing systems. *International Journal of Production Research*, *54*(23), 7043–7059.

Zhang, S., Tang, F., Li, X., Liu, J., & Zhang, B. (2021). A hybrid multi-objective approach for real-time flexible production scheduling and rescheduling under dynamic environment in Industry 4.0 context. *Computers and Operations Research*, *132*. doi:10.1016/j.cor.2021.105267.

Zhang, Y., Huang, G., Sun, S., & Yang, T. (2014). Multi-agent based real-time production scheduling method for radiofrequency identification enabled ubiquitous shopfloor environment. *Computers and Industrial Engineering*, *76*, 89–97.

Zheng, X., Zhou, S., Xu, R., & Chen, H. (2020). Energy-efficient scheduling for multi-objective two-stage flow shop using a hybrid ant colony optimization algorithm. *International Journal of Production Research*, *58*(13), 4103–4120.

Corresponding author
Tianhua Zhang can be contacted at: tianhuabjtu@163.com