Decomposition Methods for Manufacturing System Scheduling: A Survey

Fajun Yang, Kaizhou Gao, Ian Ware Simon, Yuting Zhu, and Rong Su, *Senior Member, IEEE*

***Abstract*—Manufacturing is the application of labor, tools, machines, chemical and biological processing, to an original raw material by changing its physical and geometrical characteristics, in order to make finished products. Since the first industrial revolution, to accommodate the large-scale production, tremendous changes have happened to manufacturing through the innovations of technology, organization, management, transportation and communication. This work first reviews the high-volume low-mix process by focusing on the quantity production, transfer line and single model assembly line. Then, it reviews the high-volume high-mix process. For such a process type, mixed/multi model assembly line is usually adopted. Hence, two main decisions on them, i.e., balancing and, sequencing are reviewed. Thereafter, it discusses the low-volume high-mix process in detail. Then, technology gap and future work is discussed, and at last, conclusions are given.**

***Index Terms*—Assembly line, high-mix, high-volume, low-mix, low-volume, manufacturing processes.**

I. Introduction

MANUFACTURING processes can be classified into two basic types: machining a component and assembling a product [1], [2]. In the machining process, a material is transformed into a desired product by a controlled material-removal process. For example, wood is transformed into desks and chairs. And in the assembling process, more than two components are assembled into a new entity. For instance, thousands of parts are joined together to form a new car. By changing its physical and geometry characteristics or properties, both processes can add value to the original material [3].

During the first Industrial Revolution, the development of machine tools or equipment climbed to a high climax such that mass production as a "technology" started to evolve [4], which involves the manufacture of large quantities of standardized parts, such as the use of stamping process to shape or cut metal in large amounts by deforming it with a die [5]. This type of mass production is usually called as quantity production. If it involves more complex items, such as computers, they cannot be produced by just one tool and need a sequence of production facilities. Thus, for this type, we call it flow production.

Depending on the ways of material fed to the machine, the quantity production can be divided into manual operation and mechanization, as illustrated in Fig. 1.

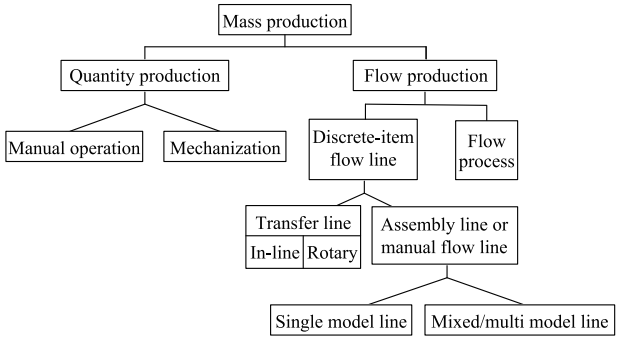


Fig. 1. Mass production system (adapted from [4]).

For the flow production, it has two subsections. One is flow process that is designed for the manufacture of some products such as food [6], chemicals [7], and steel [8] that are fluid or semifluid. If the items are discrete, it is a discrete-item flow line which can be further divided into transfer line (sometimes known as transfer machines) and assembly line (or manual flow line). Note that, in this work, we only focus on the discrete items. Hence, the flow process is not reviewed.

Transfer line, which can be either "in-line" or "rotary" type, is defined as "a set of automatic workstations arranged in a serial configuration and connected by an automatic transfer device" [4], it is a fully automated line. And a transfer line can also be a rotary machine. In a transfer line, the parts to be processed are synchronized in terms of their movement from one workstation to the next, and on each workstation, a part has to be processed in a limited time [9]. Prior to the 1970s, the transfer line has been designed mainly for mono-product type and from the 1980s onward, the mixed model style is adopted by the manufacturers [9].

An assembly line is composed of a set of workstations on which a set of limited and repetitive tasks with precedence relations are performed during a certain amount of time [10]. The workstations are usually linked by a conveyor or moving belt. Based on the number of product models, the assembly line can be categorized as single and mixed/multi model line.

It is known that mass production thrived in an era when customers can be totally satisfied by low prices, and thanks

Manuscript received July 1, 2017; accepted November 9, 2017. This work was conducted within the Delta-NTU Corporate Lab for Cyber-Physical Systems with funding support from Delta Electronics Inc and the National Research Foundation (NRF) Singapore under the Corp Lab@University Scheme. Recommended by Associate Editor Zhiwu Li. *(Corresponding author: Yuting Zhu.)*

Citation: F. J. Yang, K. Z. Gao, I. W. Simon, Y. T. Zhu, and R. Su, "Decomposition methods for manufacturing system scheduling: a survey," *IEEE/CAA J. of Autom. Sinica*, vol. 5, no. 2, pp. 389−400, Mar. 2018.

The authors are all affiliated with the School of Electrical and Electronic Engineering at Nanyang Technological University, Singapore 639798, Singapore (e-mail: fjyang@ntu.edu.sg; kzgao@ntu.edu.sg; sware@ntu.edu.sg; yuting002@e.ntu.edu.sg; rsu@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JAS.2017.7510805

to the rapid development of machine tools in the mid of 19th century, modern mass production became possible [11]. And it reached a peak after the end of Word War II since demand for products was extremely high at that time. Before the middle of the 20th century, the main characteristics for mass production have been high volume and low mix, an illustrative example is the Ford Motor Company founded in 1903. However, with the domestic and global market competition, it is especially arduous even impossible for the manufacturing companies to seize and occupy market share just by producing large volumes of standardized products, instead, they provide a variety of customized products to accommodate diversified customer needs [12], such that the production environment moved from the initial high-volume low-mix (HVLM) to high-volume high-mix (HVHM) and eventually to low-volume high-mix (LVHM). Aiming to these 3 production types, this work analyzes the main challenges and methods adopted to tackle them.

The remainder of this paper is organized as follows. Section II reviews the HVLM process with concentration on the quantity production, transfer line and single model assembly line. And Section III reviews the HVHM process by focusing on the mixed/multi model assembly line. Then, the LVHM process is reviewed in Section IV. Thereafter, Section V presents the technology gap and future work for the LVHM and, the conclusions are given in Section VI.

II. High-volume Low-mix Manufacturing Process

In the HVLM production environment, usually, quantity production, transfer line and single model assembly line are adopted, and they are discussed in the following.

A. Quantity Production

As one category of high volume manufacturing, quantity production focuses on one single product by using a standard tool or equipment [13]. For instance, the adoption of injection-molding to process items from the smallest plastic components to whole body panels of cars in large quantities, or the use of stamping process to shape or cut metal in large amounts by deforming it with a die [5]. Material is fed into the machine manually or automatically with which the final product can be obtained. And by repeating these processes the desired quantity will be achieved.

A general injection-molding consists of four main components: the base, the hopper, the barrel and, the clamping unit. The base is used to support all the other components. The hopper is where the material is poured into and, it connects the barrel that heats the material into a molten state. The clamping unit holds a mold which composes of two parts. After the molten material injected into the mold cool down to be solid, the two parts separate and a product falls out. For maintaining the barrel temperature effectively, some model-predictive-control approaches can be found in [14]−[16].

For a reliable and high quality stamping process, continuous detection, including crack, scratching, and wrinkling detection, is necessary. In [17], based on the fact that the mechanical energy is converted to the thermal energy during the stamping, the authors derive a tool to identify various malfunctions in sheet metal stamping processes by analyzing the thermal distribution of the part. In [18], the authors develop an approach of detecting cracks caused by the automotive stamping process by using a nondestructive, acoustic emission test. And in [19], a concurrent control of the failure and spring-back of a workpiece is performed.

B. Transfer Line

Transfer line is a fully automated line, it is defined as a set of automatic workstations arranged in a serial configuration and connected by an automatic transfer device [4]. It can also be a machine, for example, a cluster tool.

For an automatic transfer line, it is proved that its efficiency can be improved by dividing it into a few stages and providing buffer stores between them [20]. And in [20], it is shown how the improvement in efficiency depends on the capacity of the buffers, the number of stages, the relative failure rate of the stages, and the distribution of stage repair times. In [21], in order to increase machine utilization and reduce the total cycle time, an investigation on the line balancing of an automated cylinder block production transfer line is presented.

For a transfer line that is a machine, we can take a typical case, i.e., a cluster tool, as an example, since many problems in scheduling it could also occur in scheduling other transfer lines.

A cluster tool, as shown in Fig. 2, is widely used to process wafers in semiconductor manufacturing. It can be divided into single cluster tools (shown in Fig. 2(a)) and multi-cluster tools (Fig. 2(b)). A single cluster tool composes of a few process modules (PMs), two loadlocks for wafer loading/unloading, and a robot which can be single-arm or dual-arm. Correspondingly, it is called a single-arm or dual-arm cluster tool. A multi-cluster tool consists of some single cluster tools connected by buffer module. If all the robots are single-arm or dual-arm, we call it a single-arm or dual-arm multi-cluster tool. Otherwise, if there exists both, we call it a hybrid multi-cluster tool.

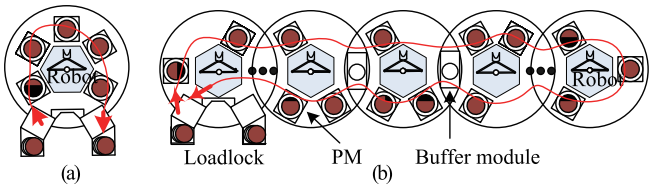


Fig. 2. Cluster tools: (a) single cluster tool and (b) multi-cluster tool.

In scheduling a cluster tool, sometimes, wafer residency time constraint [22] needs to be taken into account, which means that a processed wafer has to be removed from the PM within a limited time. Also, in operating a cluster tool, the activity time may subject to variation. By using a Petri net model, references [23] and [24] analyze the effects of activity time variation on wafer residency time delay in a PM for dual and single-arm cluster tools, respectively. As they focus on one type of wafer with an infinite number, thus, it is a HVLM manufacturing. They present methods to analytically find the upper bound of wafer residency time delay in a PM. Nevertheless, the obtained upper bound is not the exact one but is

enlarged such that the schedulability conditions are sufficient, not necessary. The exact upper bound is obtained in [25] for single-arm cluster tool by some polynomial algorithms.

For a hybrid multi-cluster tool, the conditions under which the system can reach the lower bound of cycle time are developed in [26]. Based on the conditions, an efficient algorithm is proposed to find a one-wafer cyclic schedule with the lower bound as its cycle time. This problem is further investigated in [27] if the system cannot reach the lower bound of cycle time, and methods are developed to find an optimal one-wafer cyclic schedule. For a hybrid treelike multi-cluster tool, [28] develops an efficient method to obtain the optimal one-wafer cyclic schedule. For a multi-cluster tool with wafer residency time constraint, [29] and [30] derive the necessary and sufficient conditions under which there is an optimal and feasible one-wafer cyclic schedule.

C. Single Model Assembly Line

A simple single model assembly line is shown in Fig. 3. As we can see in Fig. 3, in such a line, only one single product is handled and for a certain work station, the operations are standard all the time.

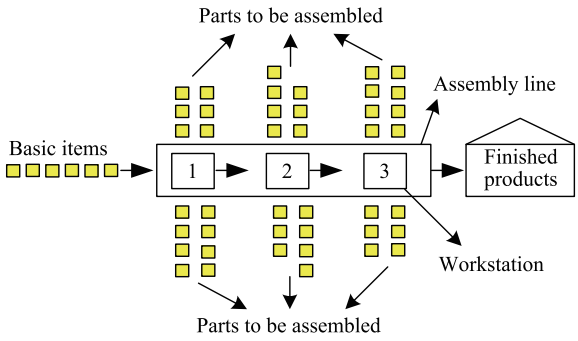


Fig. 3. Single model line (adapted from [4]).

For the single model line, its main challenge is how to optimally partition the assembly work among the stations, i.e., the assembly line balancing problem (ALBP). The first mathematical formalization of assembly line balancing is proposed by Salveson [31]. Since then, great attention has been paid to the assignment of tasks to stations. As numerous simplifying assumptions are taken into account for this problem, it is labeled as the simple assembly line balancing problem (SALBP) in [32]. Such a problem is known to be NP-hard in general [33].

Base on different objectives, SALBP can be commonly classified into three groups:

SALBP-1: minimize the number of workstations with a given cycle time;

SALBP-2: minimize cycle time with deterministic number of workstations; and

SALBP-E: maximize line efficiency by changing cycle time and number of workstations simultaneously.

To address the SALBP optimally or approximately, techniques such as evaluation of heuristics [34] and exact methods [35] are derived.

For SALBP-1, the exact methods can be divided into dynamic programming (DP) approaches and branch and bound (B&B) procedure. The first DP approach for this topic is developed by [36] and improved by [37]. And the B&B procedure can be found in [38]. For SALBP-2, two B&B procedures are also derived in [39].

However, to the best knowledge of the authors, the procedures to directly solve SALBP-E are not available due to its inherent complexity. Instead, an ant colony optimization algorithm is proposed in [40] for a two-sided assembly line balancing problem, and it is the first one that intends to minimize two conflicting objectives (cycle time and, number of workstations) at the same time on a parallel two-sided assembly line configuration. Tabu search and simulated annealing procedure for SALBP-2 and SALBP-1 can be found in [41] and [42], respectively.

Though there is much work on SALBP, as pointed by [43], SALBP still remains to be a challenging topic for researchers. And for future work, the search for new and more practical and challenging benchmarks is an interesting path [10].

Quantity production, transfer line and single model assembly line are adopted for HVLM manufacturing. However, global market competition and customers' different demands make the enterprises begin to provide a variety of customized products to seize and occupy market share. Consequently, HVHM gradually became a new trend from the middle of the 20th century. As one category of HVHM manufacturing, mass customization (MC) mode became a trend since 1980s [44]. It is also known as the concept of "build to order" or "made to order" [45], implying that manufacturers start to produce products only after they have already known what the customers need [46]. Flexibility and quick responsiveness are essential for MC. Strictly speaking, MC belongs to HVHM.

III. High-volume High-mix Manufacturing Process

In the HVHM manufacturing environment, it will take significant cost to introduce a new assembly line for any single model, thus, manufacturers try to produce one product with different characteristics or manufacture a few models on one assembly line which is called to be a mixed/multi model assembly line. Such a line is handed by Thomopoulos for the first time [47]. As illustrated in Figs. 4 and 5, the main difference between the mixed and multi model assembly lines is that, for the former, products are produced in a mixed order. Whereas, for the latter, products are produced on the same assembly line in a "one set of products" by "one set of products" way.

In the manufacturing industry, to process high-mix products, job shop is also adopted. A classical job shop problem (JSP) is usually defined as: there are n jobs, each composed of a specific set of operations which need to be done by m machines/work stations during a given time period according to a given order with the aim of finding a schedule to minimize one or multiple measures of performance. In JSP, for each operation of every job, it is assumed that there is no flexibility of the resources. As an extension of the classical JSP, flexible JSP allows one operation to be done by any machine from a given set [48]. And it consists of two sub-problems: machine

selection and operations sequencing. However, due to the space limit, we do not intend to review them in detail in this work. Instead, we focus on the mixed/multi model assembly line for the HVHM manufacturing.

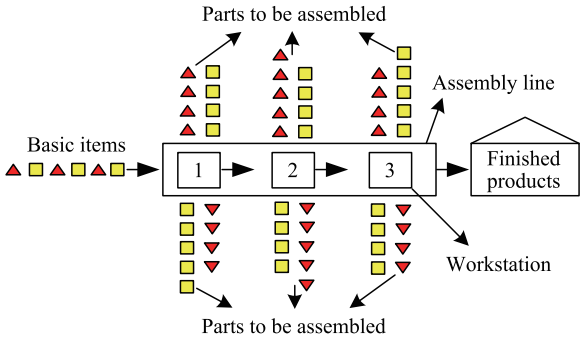


Fig. 4. Mixed model line (adapted from [4]).

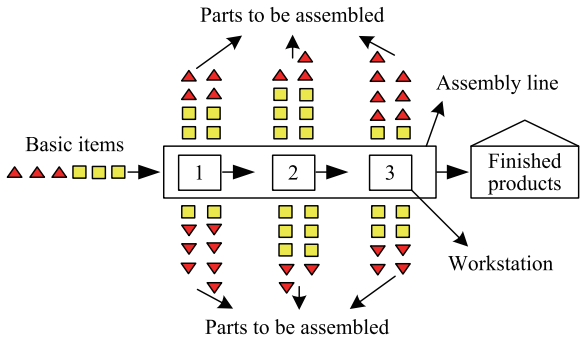


Fig. 5. Multi model line (adapted from [4]).

In general, there are two dimensions in managing a mixed/multi model assembly line, one is to assign tasks to the stations (i.e., balancing) and the other is to determine the best production order or sequence of different product models at each station (i.e., sequencing). For a long time, they were dealt separately by many researchers.

A. Balancing

The assembly line balancing problem (ALBP) is to assign tasks to workstations while optimizing one or more objectives with some constraints imposed on the line satisfied. To address the ALBP, simulated annealing approaches are developed in [49], [50], genetic algorithm techniques are derived in [51], [52], and ant colony optimization (ACO) techniques are proposed in [53] and [54].

Two-sided assembly lines with workstations locating on both sides of a straight line are introduced by Bartholdi [55]. Usually, such lines are established to produce large-sized items. In [56], a tabu search algorithm is developed to solve the two-sided assembly line balancing problem and it is proved that the method owns a good performance.

In [57], a line parallelization idea is introduced and, the parallel assembly line balancing problem is defined. By locating two straight lines in parallel the authors showed that the total number of workstations can be minimized.

Much work on the scheduling and optimization of mixed-model lines, two-sided lines, and parallel lines has been published individually. However, binary combinations of them, such as parallel two-sided lines, are very few, and the study on mixed-model parallel two-sided lines is fairly limited.

In parallel two-sided assembly line configuration, more than two lines are located in parallel and to improve productivity, workers can operate on both adjacent lines. It is first introduced by [58]. Also, such a line is commonly applied to produce large-sized items (for example cars and buses). The mixed-model parallel two-sided assembly line balancing problem is first introduced by [59] and further investigated by [60], where a flexible agent-based ant colony optimization algorithm is proposed to address this problem. The test results of a conducted experimental study show that such a line can reduce the workforce compared with separately balanced mixed-model two-sided lines and, the derived method outperforms the six heuristics and tabu search algorithm usually adopted for the assembly line balancing problem.

Recently, the lexicographic bottleneck mixed-model assembly line balancing problem is introduced by Pastor [61] with the purpose of hierarchically minimizing the most heavily loaded workstation, followed by the second most heavily loaded one and so on. Compared with the traditional smoothness index objectives, the lexicographic bottleneck objective can obtain a more smoothly distributed workload. This problem is further investigated by [62] by proposing artificial bee colony and tabu search algorithms to deal with the variations in processing times of tasks for each product model. With their methods, it is found that CPU times demanded are quite reasonable, particularly for the large-sized test problems.

B. Sequencing

In a mixed/multi model assembly line, setup operations have been reduced such that products of different models can be produced in intermixed order or sequences. However, the observed variety of model lines makes a thorough sequence planning necessary for investigating the benefits of assembly line production. Thus, researchers started to pay close attention to mixed/multi model sequencing since its first formulation is introduced by [63].

Basically, the model sequencing problem decides the production order of each product model assembled on the identical line. It aims at avoiding sequence-dependent work overload or minimizing sequence-dependent work based on a detailed scheduling with operation times, worker movements, station borders and other operational characteristics of the line being taken into consideration.

A great deal of research work has been done for the mixed/multi model assembly line sequencing problem. Bolat [64] derived a mathematical model for sequencing selection according to the objective of minimizing costs related to the delivery date. In this work, the orders that own least earliness and tardiness costs are adopted. In [65], by incorporating order selection, capacity adjustment and sequencing decision, one can minimize delivery dependent cost. Reference [66] tackled integrated planning problem of sequencing and master production scheduling and derived their interdependencies. Manavizadeh *et al.* [67] proposed order acceptance and rejection policies according to the given priorities of the orders.

These priorities are decided by the profit value that can be obtained from the orders. Also, there are some heuristic algorithms for the mixed/multi model sequencing problem, such as artificial bee colony algorithm [68], and particle swarm optimization [69].

By [70], it is known that the performance of line balancing depends on the sequencing of the produced models and the optimality of the model sequencing is affected by the line balancing, or they are heavily interdependent. Hence, to implement a mixed-model assembly line successfully, these two dimensions need to be treated and tackled together. In the next, we will introduce some work that deals with these two problems simultaneously.

C. Simultaneous Balancing and Sequencing

Regarding the simultaneous balancing and sequencing solution approaches, Kim *et al.* [70] developed an endosymbiotic evolutionary algorithm for solving the balancing and sequencing problem in mixed-model U-lines. This algorithm imitates the natural evolution process of endosymbionts and outperforms the existing symbiotic algorithms and hierarchical approaches in finding quality solutions.

In Kara *et al.* [49], for balancing and sequencing mixed-model U-lines, a multi-objective approach to simultaneously minimize the workloads of workstations, part usage rate, and cost of setups is derived. As the formalized problem has multiple conflicting objectives, it cannot be easily addressed by traditional mathematical techniques such that a simulated annealing approach is developed to minimize the weighted sum of the performance measures mentioned above. The authors also extended the derived approach to consider the stochastic completion times of tasks.

The balancing and sequencing problem of mixed-model U-lines is further investigated in Özcan *et al*. [71] and Hamzadayi *et al*. [52], where two different genetic algorithm-based approaches are developed by considering different features or constraints. Specifically, stochastic times are considered for the former and parallel workstations-zoning constraints are taken into account in the latter.

For the balancing and sequencing problem of parallel mixed-model assembly lines, Özcan *et al*. [72] developed a simulated annealing based solution approach with the purpose of maximizing the line efficiency and distributing the workloads smoothly across stations. The effectiveness of the proposed approach is proved by two numerical examples.

In [73], the mixed-model parallel two-sided assembly line balancing and sequencing problem is introduced for the first time, also, a framework of agent based ant colony optimization algorithm is developed for solving the problem. However, it did not test the efficiency of the proposed approach, and give no quantitative result about the benefits of considering sequencing and line balancing simultaneously. This problem is further studied by [74] in which a mathematical model is derived to describe the problem and a novel agent-based ant colony optimization algorithm is developed to find a near optimal solution for the problem. Experimental results show that the proposed methods can minimize line length and total number of workstations that required.

In [75], aiming to the mixed-model parallel U-shaped assembly line, the design which combines the efficiency of parallel and U-shaped lines with the flexibility of mixed-model lines is addressed by using a heuristic solution approach. By considering the model sequencing, this approach explores effective balancing solutions for different combinations. Experimental results indicate that the developed line design approach needs fewer numbers of workstations compared with independently balanced mixed-model U-lines.

For a transfer line with HVHM, [76] explores a scheduling problem of single-arm cluster tool in which two types of wafers are processed concurrently in a cyclic operational sequence for utilizing the process modules (PMs) better. In that work, different wafer types share no PMs and, a mixed integer programming model is developed to determine the optimal robot task sequencing.

IV. Low-volume High-mix Manufacturing Process

In the past few decades, customers have become more demanding for various and innovative products with shorter life-cycles in low price and quantity. Such an increasing demand on diversity of products with low quantity has resulted in LVHM manufacturing process. For a LVHM manufacturing environment, the number of different products can be more than 600, and for each product, its incoming order quantities may range from $0-1000$ over a particular time horizon [77]. In such a kind of business environment, due to the frequent production changeover, the conversion time must be increased, leading to decreased production efficiency. Hence, for LVHM manufacturing enterprises, it is increasingly important to enhance productivity and profitability through effective production scheduling.

In the LVHM manufacturing environment, there are three primary objectives as shown in the following [77]:

1) minimize lead times and, ensure on time delivery;
2) minimize makespan or maximize throughput;
3) maximize production resources utilization.

In the following, we will give some literature review on these three topics, respectively.

A. Minimize Lead Times and Ensure on Time Delivery

In [78], unexpected events and disruptions such as rush orders, machine failures, quality problems, processing time delays, and unavailable materials are taken into consideration in the LVHM manufacturing environment. Also, the scheduling of manpower which involves deciding the break time and the number of employees in each shift is taken into account. For tackling such a complex production environment, the initial design and development of a prototype system is discussed, with which enterprises could be able to enhance their delivery performance.

It is known that if unexpected events occur, rescheduling is practically mandatory and usually, three common methods are applied to reschedule an infeasible scheduling, that are right shift rescheduling, partial rescheduling, and regeneration. Similar to these methods, in [78], an initial design of interactive Gantt chart for dynamic schedule adjustment is

developed. However, it gives no idea about how to reschedule an infeasible scheduling specifically.

In [79], a case study is done from the perspective of management of high level in Schlumberger Houston Product Center that has a typical LVHM manufacturing environment. In that case, a systematic lean transformation is implemented with the purposes of improving the quality, reducing the cost and lead time. Its success heavily relies on three strategies, i.e., zero-defect-out strategy, manufacturing system design and, total employee involvement.

For the first strategy, it consists of corrective actions which strive to achieve initial improvement by fault diagnosis and correction, preventive actions and sustainable actions which seek continuous improvements. For the second strategy, it is implemented to solve the problems of excessive work-in-process, prioritization conflicts, etc. The last strategy is developed to improve employee's performance on quality and productivity.

In [80], 6 dispatching policies are combined to form a combined dispatching approach to allow the company to set up a few objectives and make the best compromise among several different requirements. These 6 dispatching policies are:

1) The first in first out rule. This rule is fair for each requirement.

2) The shortest processing time first rule with the objective of maximizing the throughput.

3) The critical ratio rule considering the due date of the lot and the remaining processing time of the current stage. The purpose of this rule is to optimize the throughput and on-time delivery.

4) The earliest due date rule with the aim of optimizing the on-time delivery from a global point of view.

5) The operation due date rule. It defines local due dates per each lot and stage. Its aim is on-time delivery.

6) The line balance rule. It is used to balance the work in process and avoid starvation of tools.

With a given weight which is obtained by a factory model using a genetic optimization algorithm for each rule, these rules are combined in a linear way. The combined dispatching policy shows an average improvement on cycle time and on-time delivery in comparison with the currently available reference dispatching policy based on an extended first-in-first-out rule.

B. Minimize Makespan or Maximize Throughput

In the lithography process, for a new mask, its exposure condition is traditionally determined by the evaluation result of focus-exposure-matrix wafer which involves exposure, metrology and data analysis, leading to a low productivity, particularly for a LVHM specific integrated circuit manufacturing. To address this problem, [81] develops a virtual lithography system for the elimination of exposure.

In [82], aiming to the design of flow paths in LVHM flow manufacturing environment, a mathematical programming which considers the sharing of machines among diverse product families with the objectives of minimizing the number of shared machines and maximizing the throughput in a predefined planning period is presented.

Owing to dynamic market and uncertain manufacturing environment, an efficient material allocation is extremely necessary in a LVHM manufacturing. Therefore, in [83], multi-constraints-based genetic algorithm is proposed for material allocation to determine an optimal combination of product types and quantity in any manufacturing system with the objective of maximizing the throughput. The constraints include capacity constraints, demand constraints, material constraints, line capacity constraints, component supply constraints, products-to-materials constraints, and materials-to-materials constraints. In the genetic algorithm, one chromosome denotes a solution of the problem and it is represented as a matrix, where the rows and columns represent the number of products and the available material, respectively. Since the chromosome is represented as a matrix, the crossover is conducted only between respective elements within two matrices.

In [84], aiming at LVHM production of electric drives, the author developed a single minute exchange of die-based (SMED-based) approach to analyze set-up procedures in order to reduce set-up times. Potential positive impacts of lower set-up times include lower stocks, smaller lot sizes, decreased lead times and minimized makespan.

It is known that there are numerous approaches for reducing set-up times, which are mainly divided into production technology, production planning and control, and organizational aspects on the shop floor [84]. For the production technology aspect, it comprises the installation of devices and the optimization of machine parameters. In addition, changing the product design or replacing the material being used by some substitution also has significant effect on set-up time reduction. For the production planning and control aspect, sequencing using minimal setup times is one of the main methods. As a swift and cost-saving approach towards reducing set-up times, organizational changes to the procedures on the shop floor are adopted in [84]. It uses a SMED-based approach which consists of 5 steps to reduce set-up times.

The first step of SMED-based approach in [84] is recording and analysis which involves distinguishing internal operations from external operations. For the former, they can be done only after the machine is shut down, for example, the operations of attaching a new die. However, for the latter, they are done while the machine is running, e.g., the preparation of the die. In the second step, in order to reduce the idle time, internal and external operations are separated. The purpose of step 3 is to transform the internal operations to external ones and it is followed by step 4 which involves the optimization of internal and external operations. At last, in step 5, to further decrease the set-up time, operations are parallelized where possible such that they can be handled by two or more operators.

In [85], to find a production schedule that minimizes the total manufacturing time of the set of jobs-makespan, three heuristic production scheduling algorithms are proposed to determine the number of machines at each stage for the jobs of a certain products family. All three algorithms can balance the machines assignment with the selected criterion. However, with their methods, it needs to reallocate the number of machines when new order enters or when any of the order is completely finished at any of the stage.

The application of conventional leveling methods is only suitable for high volume production. In [86], [87], based on the principles of group technology, the authors develop a systematic procedure for leveling LVHM production. In such a production environment, there exists a large number of product types such that a great deal of setup time is primarily caused by material changeovers. Hence, clustering techniques are adopted to subsume them into a manageable number of product families. Generally speaking, grouping criteria includes required staff and equipment, operation sequences, process times, setup times for changeover and the share of identical components, parts, or raw material, and so on. With the obtained families, a family-oriented leveling pattern can be realized by which one can minimize the changeover time or get an optimal makespan.

In [88], for enhancing current assembly processes and facility layouts, the authors focus on lean transformation in a LVHM electronics assembly environment which is characterized by long cycle times and high fall-out rates. In it, for different assembly lines, Kanban sizes are estimated to integrate and implement a "pull-system" into the lean framework. An "iterative-MAIC" (measure, analyze, improve, and control) approach, is applied to implement lean principles. With this method, it is found that the cycle time of the pilot line products is decreased by 40 %.

Recently, the wafer size has been increasing from 200 to 300 even to 450 mm, meanwhile, the circuit width has been reducing, such that the wafer lot size is decreasing from 25 wafers to even $7-8$ wafers, resulting in frequent lot switching operation [89]$-$[91].

In [89], the authors present a mixed integer programming model for obtaining an optimal robot task sequencing with the minimum makespan for both single and dual-arm cluster tools with lot switching. This problem is further addressed by [90] by proposing closed-form expressions of the makespan of the lot switching period for a given robot task sequencing.

With lot switching, a cluster tool always performs start-up and close-down operations [91]$-$[94]. Specifically, reference [92] proposes a branch and bound procedure for noncyclic scheduling problems of both single and dual-arm cluster tools to minimize the makespan. Reference [91] examines the schedulability analysis for noncyclic operation of both single and dual-arm cluster tool by considering both wafer residency time constraint and activity time variation with given robot task sequencing. References [93] and [94] derive linear programming models to determine the optimal schedules for start-up and closedown processes of single-arm cluster tools considering wafer residency time constraint. For a cluster photolithography tool with multi-robot that can process a few different wafers at the same time, [95] explores how incremental scheduling, a technique based on prioritized planning, can be used to schedule the system.

C. Maximize Production Resources Utilization

In LVHM manufacturing environment, various products always involve diversified constraints and parameters. To denote different decision outputs and objectives, a layered-matrix-encoding structure which composes of several separated layers is usually used. However, in [96], instead of using complicated multidimensional encoding, a layered-encoding cascade optimization structure with a two-layered 2-D matrix is developed to represent the product-mix schedule, resulting in a simplified problem representation. In the proposed structure, the first layer is used for selecting the items, and the second layer is for determining the volume of items that are going to be tested or generated. These two layers act as two agents that can communicate with each other to obtain the best decision and pass it to the other layer for optimization. In this work, four combinations of genetic algorithm (GA) and particle swarm optimization (PSO), i.e., GA-GA, GA-PSO, PSO-PSO, and PSO-GA are used as optimizers in cascade and it is found that GA-PSO model is an effective approach for dealing with LVHM product-mix problems since it exhibits the advantages of low changeover time, high productivity, and high equipment efficiency or resource utilization.

For a modern foundry where varying cast products made of different alloys are simultaneously processed with relatively low volume, to decide the volume of each product within a casting shift, [97] derived a linear programming model by considering some real constraints, for example the product quantities need to match with the customer order and, the working time of casting machines cannot exceed the specified limitation of each shift. The objective of the model is to maximize the average efficiency of melting furnaces, i.e., the average efficiency of alloy usage throughout the shifts. The efficiency of the algorithm is checked on a real-world dataset of a real foundry and the experimental results prove that by using the derived approach, optimality of alloy usage can be achieved. In comparison with the previous work, [97] has the following contributions.

1) Instead of assuming that there is just one single furnace for melting a specific alloy as previous studies did, [97] supposes that more than one furnace can be used to supply the same kind of alloy.

2) The previous work prescribed that one kind of cast product is assigned to just one casting machine. However, [97] denotes a more general case that more than one casting machine can be adopted to manufacture the same product type.

3) Lastly, in [97], a single die casting machine can be installed with a large die with several cavities that have identical or different impressions, where the former is called to be a multi-cavity die, and the latter, combination die. In the previous work, such dies were not considered.

In a LVHM production environment, effective equipment utilization and sustainable production capacities that are partly ensured by appropriate maintenance methods act a significant role in the competitiveness of semiconductor industry. Currently, the maintenance strategies are mainly derived by following experts' historical knowledge. However, such a way is not always efficient to deal with an evolving nature of machine failure behaviors. Hence, in order to effectively support maintenance actions and experts' knowledge renewal, [98] introduces a new methodology by combining the Bayesian approach and an extended FMECA (failure mode, effect and criticality analysis) method. In this methodology, FMECA files are used to model experts' historical knowledge as an

operational Bayesian network to offer real-time feedback on bad maintenance actions. And it has three steps.

Step 1: Use FMECA files to define functions and a set of associated objective fulfillment criteria which have to be respected by technicians during executing maintenance actions. The concentration of this step is to capitalize, unify and share experts' knowledge on best maintenance practices, and it also gives a description about the potential consequences led by poorly executed maintenance actions.

Step 2: Use historical data to evaluate the accuracy of operational Bayesian network model such that one can control the relevance of existing experts' knowledge.

Step 3: Update knowledge if model accuracy drifts beyond the defined limitations. The triggering events to update the experts' knowledge include inaccuracy of operational Bayesian network and drift of maintenance performance measurement indicator.

Except these three objectives, there is also some work which focuses on integrating customer flexibility in the order commitment process [12], and minimizing work in process inventory level [99].

In the LVHM manufacturing environment, demands for on time delivery, shortened lead time, frequent customer orders and various customer requirements have made the order commitment process, which plays an increasingly crucial role in satisfying individual customer's needs, especially challenging. To address this process efficiently, [12] integrates the flexibility from both manufacturing and customer side. Customer flexibility is characterized by his/her indifference to some attributes such as price, due date, order quantity, product specification, delivery schedule, and so on. Furthermore, under some conditions, the customer would like to make trade-offs on these attributes. For example, the customer would accept a delayed delivery with cheaper price. Intuitively, by considering customer flexibility, the interests of both customers' and manufacturers' can be better satisfied because the solution space of meeting supply and demand will be broadened in comparison with the traditional domain solely from a manufacturing perspective.

In [12], to describe the customer's tolerance and the sensitivity of customer's satisfaction to different attributes, the developed customer flexibility is characterized by two dimensions. One is range which considers the acceptable range of customer among different alternatives, and the other is response which considers the correlations among different attributes. In customer flexibility, such correlations form the trade-off constraints. The obtained customer flexibility representation is then incorporated into the order commitment model. During this process, first, it needs to map the customer requirements into material and process requirements. Then, a mixed-integer-programming model is proposed to offer optimal order commitment decisions. Experimental study shows that multi-attribute customer flexibility has huge positive impacts on system performances.

Aiming to a LVHM manufacturing environment with significant setup times, Srinivasan and Viswanathan [99] derived an approach to obtain the optimal work-in-process inventory level of each product such that the required demand could be satisfied. There are two decision variables for each product type, i.e., the number of pallets and the number of products a pallet can contain. To describe the congestion in the system, all product types are modelled as a closed queueing network with multiple customer classes, where a customer class represents a product type. Then, with the development of state equations for the closed queueing network by using mean value analysis, the optimization problem can be formulated. However, it is a complicated non-linear integer programming with a non-convex objective function. To solve it, the upper and lower bounds on the number of pallets for each individual product is developed first, within which, exhaustive enumeration is used to obtain the optimal solution for this problem. This method works well for the case that has no more than 20 product types. Otherwise, the computational complexity increases exponentially with the number of products. Hence, a simple heuristic algorithm is proposed for the case with large number of products.

In the LVHM manufacturing environment, an energy efficiency benchmarking (E2BM) which acts as a method to identify the best practices that serve as possible benchmarks for measuring and managing efficiency improvement of energy within a company (including the plant, production line and machine) is developed in [100]. It composes of five steps.

Step 1: Energy and Material Flows Modelling. With the model, one can define the data and parameters requirements for further analysis.

Step 2: Data Collection. This can be realized either by estimation based on electricity bill, production records and, machine specifications, or by measurement with suitable sensors.

Step 3: Metrics Determination.

Three metrics are proposed and they are:

1) Energy intensity. Its definition is "the energy consumption per unit dollar".

2) Specific energy consumption. Its definition is "the energy consumption per production volume".

3) Energy efficiency with reference to the economic performance and production activities.

Step 4: Benchmarking and Analysis. It includes the following four methods.

1) Internal benchmarking using linear regression analysis.

2) Internal benchmarking using data development analysis.

3) External benchmarking with energy intensity as the key metric.

4) External benchmarking with energy efficiency as the key metric.

Step 5: Improvement Planning.

With the obtained benchmarking result, improvement planning is then implemented to allocate the responsibilities for the improvement of energy efficiency.

As it is extremely challenging for implementing overall equipment effectiveness (OEE) in LVHM manufacturing environment, a new effectiveness method, i.e., machining equipment effectiveness (MEE), is developed and evaluated for LVHM in [101], considering the three factors (availability, performance and quality) proposed by Nakajima [102], for measuring the effectiveness of equipment.

The definition of availability factor in original OEE considers up and downtime of the equipment whose concept is the same for both LVHM and LMHV. Hence, the availability factor of OEE can be directly applied for MEE with no redefinition.

In the original OEE, the performance factor is calculated by dividing the net operating time by the operating time. However, for LVHM, in a given time interval, there is a variety of different products being manufactured, furthermore, for each product type, its optimal operating time may be different. Thus, the performance needs to be calculated for each product and then summation is done.

In the original OEE, by dividing the valuable operating time by the net operating time, one can obtain the quality factor. In LVHM, as there are multiple categories (including concession, rework and reject) for defective products, to calculate the quality factor, a possible way is to take the costs of defects into account since the defective products are reworked with extra costs which have direct relation with the quality deviation of the produced part.

V. Technology Gap and Future Work

For the complex and challenging LVHM manufacturing, limited work has been published such that the technology gap is still very obvious and much work remains to be done.

1) As the biggest source of uncertainty to any company, demand of the customer is unpredictable [103]. For an existing production scheduling, when priority order newly comes and is inserted, how to response quickly or rescheduling is a very challenging problem. Furthermore, if the stability metrics which evaluate how many operations of existing jobs will be remained on the same processing machine in rescheduling are taken into account, the problem will be more challenging, particularly considering the changeover time and on time delivery.

2) In the LVHM manufacturing environment, thousands of different parts need to be delivered to a large number of workstations even in just one shift. On one hand, material shortages will result in stoppage and idleness of machines and workers. On the other hand, enlarged safety stocks near the machine occupy the scarce space of workstations, leading to an extra inventory cost at the workstations. As the volume for each product is very low, how to deliver the parts just in time or how to determine the optimal work in process is very challenging.

3) Since the volume of each product is low, changeover is very frequent in the LVHM manufacturing, leading to low productivity. To address this problem, usually, some different products first form a part family based on process similarity analysis by using group technology. Then, for these families, the optimal sequence that has the minimal setup time needs to be determined. However, such a method could result in heavy work in processing. Thus, to reduce the setup time, more work is still needed.

4) In LVHM manufacturing, order commitment speed is becoming one of the main competitive differentiators among manufacturing companies. Reference [12] derives an approach for representing customer flexibility and, connects it with manufacturing flexibility by using a proposed mixed integer programming model. However, for large scale problems, such a method is unsuitable. Therefore, it needs to develop efficient heuristic algorithms to facilitate fast order commitments.

5) For the custom products, complexity and variations often result in mistake on bill of material, such that low quality and high cost is inevitable. Frequent changeover also brings in low resource utilization. Further, long lead time and delayed delivery are still nightmares for the customers. Hence, efficient multi-objective programming is desired to solve these problems.

6) There is much work focusing on single cluster tool with lot switching. However, a multi-cluster tool is structurally more complex than a single cluster tool, and for it, there is no work for lot switching with consideration of wafer residency time constraint and activity time variation. If taking the PM cleaning requirements, PM failure and wafer revisiting process into account, the problem will be more challenging.

VI. Conclusions

This work reviews the manufacturing evolution from high-volume low-mix to high-volume high-mix and at last to low-volume high-mix. Due to the development of machines tools or equipment has reached a high climax during the first Industrial Revolution, mass production as a "technology" started to evolve. In that era, customers can be totally satisfied by low prices such that low-mixed products were processed in the factory. Thus, quantity production, transfer line and single model assembly line for high-volume low-mix manufacturing is reviewed first.

With the domestic and global market competition, the production environment moved to high-volume high-mix and eventually to low-volume high-mix since the customers have become more demanding for various and innovative products with shorter product life-cycles in low price and quantity. For the high-volume high-mix manufacturing, this work reviews mixed/multi model assembly line by focusing on its two main problems, balancing and sequencing. And for the low-volume high-mix manufacturing, papers on how to minimize lead times and makespan, maximize production resources utilization, etc., are reviewed. After reviews, the technical gap and future work for LVHM is pointed out.

References

[1] M. P. Groover, *Fundamentals of Modern Manufacturing: Materials, Processes, and Systems*. Third edition. New Jersey, USA: John Wiley and Sons, Inc., 2007.

[2] P. Sivasankaran and P. Shahabudeen, "Literature review of assembly line balancing problems," *Int. J. Adv. Manuf. Technol.*, vol. 73, no. 9−12, pp. 1665−1694, Aug. 2014.

[3] N. M. Z. N. Mohamed and M. K. Khan, "Decomposition of manufacturing processes: a review," *Int. J. Automot. Phys. Mechan. Eng.*, vol. 5, pp. 545−560, Jan.−Jun. 2012.

[4] F. de P Hanika, "Mass production management," *Journal of the Operational Research Society*, vol. 25, no. 2, pp. 330, 1974.

[5] X. C. Liu, K. Ji, O. El Fakir, H. M. Fang, M. M. Gharbi, and L. L. Wang, "Determination of the interfacial heat transfer coefficient for a hot aluminium stamping process," *J. Mater. Process. Technol.*, vol. 247, pp. 158−170, Sep. 2017.

[6] J. A. Brierley, C. J. Cowton, and C. Drury, "A comparison of product costing practices in discrete-part and assembly manufacturing and continuous production process manufacturing," *Int. J. Prod. Econ.*, vol. 100, no. 2, pp. 314−321, Apr. 2006.

[7] R. L. Tousain and O. H. Bosgra, "Market-oriented scheduling and economic optimization of continuous multi-grade chemical processes," *J. Process Control*, vol. 16, no. 3, pp. 291−302, Mar. 2006.

[8] L. X. Tang and G. S. Wang, "Decision support system for the batching problems of steelmaking and continuous-casting production," *Omega*, vol. 36, no. 6, pp. 976−991, Dec. 2008.

[9] K. Dhouib, A. Gharbi, and N. Landolsi, "Throughput assessment of mixed-model flexible transfer lines with unreliable machines," *Int. J. Prod. Econ.*, vol. 122, no. 2, pp. 619−627, Dec. 2009.

[10] O. Battaïa and A. Dolgui, "A taxonomy of line balancing problems and their solution approaches," *Int. J. Prod. Econ.*, vol. 142, no. 2, pp. 259−277, Apr. 2013.

[11] D. A. Hounshell, *From the American System to Mass Production, 1800-1932*. Baltimore, Maryland, USA: Johns Hopkins University Press, 1984.

[12] Q. Zhang and M. M. Tseng, "Modelling and integration of customer flexibility in the order commitment process for high mix low volume production," *Int. J. Prod. Res.*, vol. 47, no. 22, pp. 6397−6416, Aug. 2009.

[13] L. E. Cárdenas-Barrón, "Economic production quantity with rework process at a single-stage manufacturing system with planned backorders," *Comput. Ind. Eng.*, vol. 57, no. 3, pp. 1105−1113, Oct. 2009.

[14] C. H. Lu and C. C. Tsai, "Adaptive decoupling predictive temperature control for an extrusion barrel in a plastic injection molding process," *IEEE Trans. Ind. Electron.*, vol. 48, no. 5, pp. 968−975, Oct. 2001.

[15] T. L. Chia, "Model predictive control helps to regulate slow processes-robust barrel temperature control," *ISA Trans.*, vol. 41, no. 4, pp. 501−509, Oct. 2002.

[16] S. N. Huang, K. K. Tan, and T. H. Lee, "Adaptive GPC control of melt temperature in injection moulding," *ISA Trans.*, vol. 38, no. 4, pp. 361−373, Nov. 1999.

[17] Y. M. H. Ng, M. L. Yu, Y. Huang, and R. X. Du, "Diagnosis of sheet metal stamping processes based on 3-D thermal energy distribution," *IEEE Trans. Autom. Sci. Eng.*, vol. 4, no. 1, pp. 22−30, Jan. 2007.

[18] J. Song, S. Kim, Z. Y. Liu, N. N. Quang, and F. Bien, "A real time nondestructive crack detection system for the automotive stamping process," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 11, pp. 2434−2441, Nov. 2016.

[19] F. Z. Oujebbour, A. Habbal, and R. Ellaia, "Optimization of concurrent criteria in the stamping process," in *Proc. 2013 Int. Conf. Industrial Engineering and Systems Management*, Rabat, Morocco, 2013.

[20] J. A. Buzacott, "Automatic transfer lines with buffer stocks," *Int. J. Prod. Res.*, vol. 5, no. 3, pp. 183−200, Jan. 1967.

[21] S. Masood, "Line balancing and simulation of an automated production transfer line," *Assembly Autom.*, vol. 26, no. 1, pp. 69−74, Jan. 2006.

[22] S. Rostami, B. Hamidzadeh, and D. Camporese, "An optimal periodic scheduler for dual-arm robots in cluster tools with residency constraints," *IEEE Trans. Robot. Autom.*, vol. 17, no. 5, pp. 609−618, Oct. 2001.

[23] N. Q. Wu and M. C. Zhou, "Analysis of wafer sojourn time in dual-arm cluster tools with residency time constraint and activity time variation," *IEEE Trans. Semicond. Manuf.*, vol. 23, no. 1, pp. 53−64, Feb. 2010.

[24] Y. Qiao, N. Q. Wu, and M. C. Zhou, "Petri net modeling and wafer sojourn time analysis of single-arm cluster tools with residency time constraints and activity time variation," *IEEE Trans. Semicond. Manuf.*, vol. 25, no. 3, pp. 432−446, Aug. 2012.

[25] C. R. Pan. Y. Qiao, N. Q. Wu, and M. C. Zhou, "A novel algorithm for wafer sojourn time analysis of single-arm cluster tools with wafer residency time constraints and activity time variation," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 45, no. 5, pp. 805−818, May 2015.

[26] F. J. Yang, N. Q. Wu, Y. Qiao, and M. C. Zhou, "Petri net-based optimal one-wafer cyclic scheduling of hybrid multi-cluster tools in wafer fabrication," *IEEE Trans. Semicond. Manuf.*, vol. 27, no. 2, pp. 192−203, May 2014.

[27] F. J. Yang, N. Q. Wu, Y. Qiao, and M. C. Zhou, "Petri net-based polynomially complex approach to optimal one-wafer cyclic scheduling of hybrid multi-cluster tools in semiconductor manufacturing," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 44, no. 12, pp. 1598−1610, Dec. 2014.

[28] F. J. Yang, N. Q. Wu, Y. Qiao, and M. C. Zhou, "Optimal one-wafer cyclic scheduling of time-constrained hybrid multicluster tools via petri nets," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 47, no. 11, pp. 2920−2932, Nov. 2017. doi: 10.1109/TSMC.2016.2531697.

[29] F. J. Yang, N. Q. Wu, Y. Qiao, and M. C. Zhou, "Optimal one-wafer cyclic scheduling of hybrid multirobot cluster tools with tree topology," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 48, no. 2, pp. 289–298, Feb. 2018.

[30] Q. H. Zhu, N. Q. Wu, Y. Qiao, and M. C. Zhou, "Scheduling of single-arm multi-cluster tools with wafer residency time constraints in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 28, no. 1, pp. 117−125, Feb. 2015.

[31] M. E. Salveson, "The assembly line balancing problem," *J. Ind. Eng.*, vol. 6, no. 3, pp. 18−25, Jan. 1955.

[32] I. Baybars, "A survey of exact algorithms for the simple assembly line balancing problem," *Manage. Sci.*, vol. 32, no. 8, pp. 909−932, Aug. 1986.

[33] T. S. Wee and M. J. Magazine, "Assembly line balancing as generalized bin packing," *Oper. Res. Lett.*, vol. 1, no. 2, pp. 56−58, Apr. 1982.

[34] S. G. Ponnambalam, P. Aravindan, and G. M. Naidu, "A comparative evaluation of assembly line balancing heuristics," *Int. J. Adv. Manuf. Technol.*, vol. 15, no. 8, pp. 577−586, Jul. 1999.

[35] A. Scholl and C. Becker, "State-of-the-art exact and heuristic solution procedures for simple assembly line balancing," *Eur. J. Oper. Res.*, vol. 168, no. 3, pp. 666−693, Feb. 2006.

[36] J. R. Jackson, "A computing procedure for a line balancing problem," *Manage. Sci.*, vol. 2, no. 3, pp. 261−271, Apr. 1956.

[37] M. Held, R. M. Karp, and R. Shareshian, "Assembly-line balancing-dynamic programming with precedence constraints," *Oper. Res.*, vol. 11, no. 3, pp. 442−459, Jun. 1963.

[38] A. Sprecher, "Dynamic search tree decomposition for balancing assembly lines by parallel search," *Int. J. Prod. Res.*, vol. 41, no. 7, pp. 1413−1430, May 2003.

[39] R. Klein and A. Scholl, "Maximizing the production rate in simple assembly line balancing — a branch and bound procedure," *Eur. J. Oper. Res.*, vol. 91, no. 2, pp. 367−385, Jun. 1996.

[40] I. Kucukkoc and D. Z. Zhang, "Type-E parallel two-sided assembly line balancing problem: Mathematical model and ant colony optimisation based approach with optimised parameters," *Comput. Ind. Eng.*, vol. 84, 56−69, Jun. 2015.

[41] A. Scholl and S. Voß, "Simple assembly line balancing-heuristic approaches," *J. Heuristics*, vol. 2, no. 3, pp. 217−244, Dec. 1997.

[42] P. R. McMullen and G. V. Frazier, "Using simulated annealing to solve a multiobjective assembly line balancing problem with parallel workstations," *Int. J. Prod. Res.*, vol. 36, no. 10, pp. 2717−2741, Dec. 1998.

[43] A. C. Nearchou, "Balancing large assembly lines by a new heuristic based on differential evolution method," *Int. J. Adv. Manuf. Technol.*, vol. 34, pp. 9−10, pp. 1016−1029, Oct. 2007.

[44] S. M. Davis, "From future perfect: Mass customizing," *Plan. Rev.*, vol. 17, no. 2, pp. 16−21, Dec. 1989.

[45] D. Pollard, S. Chuo, and B. Lee, "Strategies for mass customization," *J. Bus. Econ. Res.*, vol. 6, no. 7, pp. 77−86, Jul. 2011.

[46] H. Tieng, C. F. Chen, F. T. Cheng, and H. C. Yang, "Automatic virtual metrology and target value adjustment for mass customization," *IEEE Rob. Autom. Lett.*, vol. 2, no. 2, pp. 546−553, Apr. 2017.

[47] N. T. Thomopoulos, "Line balancing-sequencing for mixed-model assembly," *Manage. Sci.*, vol. 14, no. 2, pp. B59−B75, Oct. 1967.

[48] M. Yazdani, M. Amiri, and M. Zandieh, "Flexible job-shop scheduling with parallel variable neighborhood search algorithm," *Expert Syst. Appl.*, vol. 37, no. 1, pp. 678−687, Jan. 2010.

[49] Y. Kara, U. Ozcan, and A. Peker, "Balancing and sequencing mixed-model just-in-time U-lines with multiple objectives," *Appl. Math. Comput.*, vol. 184, no. 2, pp. 566−588, Jan. 2007.

[50] H. Mosadegh, M. Zandieh, and S. M. T. F. Ghomi, "Simultaneous solving of balancing and sequencing problems with station-dependent assembly times for mixed-model assembly lines," *Appl. Soft Comput.*, vol. 12, no. 4, pp. 1359−1370, Apr. 2012.

[51] S. Akpinar and G. M. Bayhan, "A hybrid genetic algorithm for mixed model assembly line balancing problem with parallel workstations and zoning constraints," *Eng. Appl. Artif. Intell.*, vol. 24, no. 3, pp. 449−457, Apr. 2011.

[52] A. Hamzadayi and G. Yildiz, "A genetic algorithm based approach for simultaneously balancing and sequencing of mixed-model U-lines with parallel workstations and zoning constraints," *Comput. Ind. Eng.*, vol. 62, no. 1, pp. 206−215, Feb. 2012.

[53] A. S. Simaria and P. M. Vilarinho, "2-ANTBAL: an ant colony optimisation algorithm for balancing two-sided assembly lines," *Comput. Ind. Eng.*, vol. 56, no. 2, pp. 489−506, Mar. 2009.

[54] B. Yagmahan, "Mixed-model assembly line balancing using a multi-objective ant colony optimization approach," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 12453−12461, Sep. 2011.

[55] J. J. Bartholdi, "Balancing two-sided assembly lines: a case-study," *Int. J. Prod. Res.*, vol. 31, no. 10, pp. 2447−2461, Oct. 1993.

[56] U. Özcan and B. Toklu, "A tabu search algorithm for two-sided assembly line balancing," *Int. J. Adv. Manuf. Technol.*, vol. 43, no. 7−8, pp. 822−829, Aug. 2009.

[57] H. Gökçen, K. Ağpak, and R. Benzer, "Balancing of parallel assembly lines," *Int. J. Prod. Econ.*, vol. 103, no. 2, pp. 600−609, Oct. 2006.

[58] U. Özcan, H. Gökçen, and B. Toklu, "Balancing parallel two-sided assembly lines," *Int. J. Prod. Res.*, vol. 48, no. 16, pp. 4767−4784, Aug. 2010.

[59] D. Z. Zhang and I. Kucukkoc, "Balancing mixed-model parallel two-sided assembly lines," in *Proc. Int. Conf. Industrial Engineering and Systems Management*, Rabat, Morocco, 2013.

[60] I. Kucukkoc and D. Z. Zhang, "Mixed-model parallel two-sided assembly line balancing problem: A flexible agent-based ant colony optimization approach," *Comput. Ind. Eng.*, vol. 97, 58−72, Jul. 2016.

[61] R. Pastor, "LB-ALBP: the lexicographic bottleneck assembly line balancing problem," *Int. J. Prod. Res.*, vol. 49, no. 8, pp. 2425−2442, Apr. 2011.

[62] K. Buyukozkan, I. Kucukkoc, S. I. Satoglu, and D. Z. Zhang, "Lexicographic bottleneck mixed-model assembly line balancing problem: Artificial bee colony and tabu search approaches with optimised parameters," *Expert Syst. Appl.*, vol. 50, pp. 151−166, May 2016.

[63] L. Wester and M. Kilbridge, "The assembly line model-mix sequencing problem," in *Proc. 3rd Int. Conf. Operation Research*, Oslo, Norway, 1964, pp. 247−260.

[64] A. Bolat, "A mathematical model for selecting mixed models with due dates," *Int. J. Prod. Res.*, vol. 41, no. 5, pp. 897−918, 2003.

[65] T. Volling and T. S. Spengler, "Modeling and simulation of order-driven planning policies in build-to-order automobile production," *Int. J. Prod. Econ.*, vol. 131, no. 1, pp. 183−193, May 2011.

[66] J. Dörmer, H. O. Günther, and R. Gujjula, "Master production scheduling and sequencing at mixed-model assembly lines in the automotive industry," *Flex. Serv. Manuf. J.*, vol. 27, no. 1, pp. 1−29, Mar. 2015.

[67] N. Manavizadeh, A. H. Goodarzi, M. Rabbani, and F. Jolai, "Order acceptance/rejection policies in determining the sequence in mixed model assembly lines," *Appl. Math. Modell.*, vol. 37, no. 4, pp. 2531−2551, Feb. 2013.

[68] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," Tech. Rep.-TR06, Oct. 2005.

[69] Q. Y. Dong, J. S. Lu, and Y. K. Gui, "Integrated optimization of production planning and scheduling in mixed model assembly line," *Proced. Eng.*, vol. 29, pp. 3340−3347, Jan. 2012.

[70] Y. K. Kim, J. Y. Kim, and Y. Kim, "An endosymbiotic evolutionary algorithm for the integration of balancing and sequencing in mixed-model U-lines," *Eur. J. Oper. Res.*, vol. 168, no. 3, pp. 838−852, Feb. 2006.

[71] U. Özcan, T. Kellegöz, and B. Toklu, "A genetic algorithm for the stochastic mixed-model U-line balancing and sequencing problem," *Int. J. Prod. Res.*, vol. 49, no. 6, pp. 1605−1626, Mar. 2011.

[72] U. Özcan, H. Çerçioğlu, H. Gökçen, and B. Toklu, "Balancing and sequencing of parallel mixed-model assembly lines," *Int. J. Prod. Res.*, vol. 48, no. 17, pp. 5089−5113, Sep. 2010.

[73] I. Kucukkoc and D. Z. Zhang, "Simultaneous balancing and sequencing of mixed-model parallel two-sided assembly lines," *Int. J. Prod. Res.*, vol. 52, no. 12, pp. 3665−3687, Jan. 2014.

[74] I. Kucukkoc and D. Z. Zhang, "Mathematical model and agent based solution approach for the simultaneous balancing and sequencing of mixed-model parallel two-sided assembly lines," *Int. J. Prod. Econ.*, vol. 158, pp. 314−333, Dec. 2014.

[75] I. Kucukkoc and D. Z. Zhang, "Balancing of mixed-model parallel U-shaped assembly lines considering model sequences," *Int. J. Prod. Res.*, vol. 55, no. 20, pp. 5958−5975, Apr. 2017. doi: 10.1080/00207543.2017.1312586.

[76] J. H. Lee, H. J. Kim, and T. E. Lee, "Scheduling cluster tools for concurrent processing of two wafer types," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 2, pp. 525−536, Apr. 2014.

[77] R. M. Mahoney, *High-Mix Low-Volume Manufacturing*. Loveland, Colorado, USA: Hewlett-Packard Company, 1997.

[78] T. J. Chua, T. X. Cai, and J. M. W. Low, "Dynamic operations and manpower scheduling for high-mix, low-volume manufacturing," in *Proc. 2008 IEEE Int. Conf. Emerging Technologies and Factory Automation*, Hamburg, Germany, 2008.

[79] P. Pandian, L. Yang, and X. Y. Liu. Lean transformation for high mix low volume production: a case study," in *Proc. 2010 Industrial Engineering Research Conf.*, Cancun, Mexico, 2010.

[80] O. Rose, "Development and introduction of a combined dispatching policy at a high-mix low-volume ASIC facility," in *Proc. 2012 Winter Simulation Conf. (WSC)*, Berlin, Germany, 2012.

[81] K. Yoshida, T. Sato, T. Kono, E. Yamanaka, M. Kariya, A. Inoue, and S. Mimotogi, "Virtual lithography system to improve the productivity of high-mix low-volume production," in *Proc. SPIE*, Japan, vol. 6607, 2007.

[82] Y. F. Peng, Z. L. Guan, L. Ma, C. Y. Zhang, and P. G. Li, "A mathematical programming method for flow path design in high-mix and low-volume flow manufacturing," in *Proc. IEEE Int. Conf. Industrial Engineering and Engineering Management*, Singapore, 2008.

[83] A. Ali, H. Seifoddini, and J. Lee, "Efficient material allocations in high-mix low-volume manufacturing," *J. Adv. Manuf. Syst.*, vol. 9, no. 2, pp. 101−116, Feb. 2010.

[84] R. Fritsche, "Reducing set-up times for improved flexibility in high-mix low-volume electric drives production," in *Proc. 1st Int. Electric Drives Production Conf. (EDPC)*, Nuremberg, Germany, 2011.

[85] J. Svancara and Z. Kralova, "High-mix low-volume flow shop manufacturing system scheduling," *IFAC Proc. Vol.*, vol. 45, no. 6, pp. 145−150, May 2012.

[86] F. Bohnen, T. Maschek, and J. Deuse, "Leveling of low volume and high mix production based on a group technology approach," *CIRP J. Manuf. Sci. Technol.*, vol. 4, no. 3, pp. 247−251, Jan. 2011.

[87] F. Bohnen, M. Buhl, and J. Deuse, "Systematic procedure for leveling of low volume and high mix production," *CIRP J. Manuf. Sci. Technol.*, vol. 6, no. 1, pp. 53−58, Dec. 2013.

[88] V. A. Raghavan, S. Yoon and K. Srihari, "Lean transformation in a high mix low volume electronics assembly environment," *Int. J. Lean Six Sigma*, vol. 5, no. 4, pp. 342−360, Oct. 2014.

[89] J. H. Lee, H. J. Kim, and T. E. Lee, "Scheduling lot switching operations for cluster tools," *IEEE Trans. Semicond. Manuf.*, vol. 26, no. 4, pp. 592−601, Apr. 2013.

[90] J. H. Lee and H. J. Kim, "Makespan analysis of lot switching period in cluster tools," *IEEE Trans. Semicond. Manuf.*, vol. 29, no. 2, pp. 127−136, May 2016.

[91] H. J. Kim, J. H. Lee, and T. E. Lee, "Schedulability analysis for non-cyclic operation of time-constrained cluster tools with time variation," *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 3, pp. 1409−1414, Jul. 2016.

[92] H. J. Kim, J. H. Lee, and T. E. Lee, "Noncyclic scheduling of cluster tools with a branch and bound algorithm," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 2, pp. 690−700, Apr. 2015.

[93] Y. Qiao, M. C. Zhou, N. Q. Wu, and Q. H. Zhu, "Scheduling and control of startup process for single-arm cluster tools with residency time constraints," *IEEE Trans. Control Syst. Technol.*, vol. 25, no. 4, pp. 1243−1256, Jul. 2017.

[94] Q. H. Zhu, M. C. Zhou, Y. Qiao, and N. Q. Wu, "Petri net modeling and scheduling of a close-down process for time-constrained single-arm cluster tools," *IEEE Trans. Syst. Man Cybern. Syst.*, 2016. doi: 10.1109/TSMC.2016.2598303.

[95] S. Ware and R. Su, "An application of incremental scheduling to a cluster photolithography tool," in *Pro. 2017 IFAC World Congress*, Toulouse, France, vol. 50, no. 1, pp. 1114−1120.

[96] S. C. Neoh, N. Morad, C. P. Lim, and Z. A. Aziz, "A layered-encoding cascade optimization approach to product-mix planning in high-mixClow-volume manufacturing," *IEEE Trans. Syst. Man Cybern. Part A Syst. Humans*, vol. 40, no. 1, pp. 133−146, Jan. 2010.

[97] Y. K. Park and J. M. Yang, "Scheduling of die casting operations including high-mix low-volume and line-type production," *Int. J. Prod. Res.*, vol. 51, no. 6, pp. 1728−1744, Jan. 2013.

[98] A. B. Said, M. K. Shahzad, E. Zamai, S. Hubac, and M. Tollenaere, "Experts' knowledge renewal and maintenance actions effectiveness in high-mix low-volume industries, using Bayesian approach," *Cogn. Technol. Work*, vol. 18, no. 1, Feb. 2016.

[99] M. M. Srinivasan and S. Viswanathan, "Optimal work-in-process inventory levels for high-variety, low-volume manufacturing systems," *IIE Trans.*, vol. 42, no. 6, pp. 379−391, Mar. 2010.

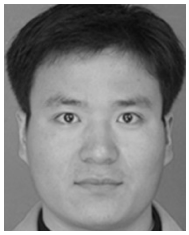
[100] Y. S. Tan, T. B. Tjandra, and B. Song, "Energy efficiency benchmarking methodology for mass and high-mix low-volume productions," *Proced. CIRP*, vol. 29, pp. 120−125, Dec. 2015.

[101] J. M. J. Becker, J. Borst, and A. van der Veen, "Improving the overall equipment effectiveness in high-mix-low-volume manufacturing environments," *CIRP Ann.*, vol. 64, no. 1, pp. 419−422, May 2015.

[102] S. Nakajima, *TPM Development Program: Implementing Total Productive Maintenance*. Cambridge, MA, USA: Productivity Press, 1989.

[103] A. Jain, P. K. Jain, F. T. S. Chan, and S. Singh, "A review on manufacturing flexibility," *Int. J. of Prod. Res.*, vol. 51, no. 19, pp. 5946−5970, Oct. 2013.

Fajun Yang received the B.S. degree in industrial engineering from Hunan University of Science and Technology, Hunan, China, in 2011, the Ph.D. degree in mechanical engineering from Guangdong University of Technology, China, in 2016. From 2015−2016, he was a Visiting Student with New Jersey Institute of Technology, Newark, NJ, USA. He is currently a Research Fellow with Nanyang Technological University, Singapore. He has 10+ international journal papers (majority in the IEEE Transactions). His research interests include Petri nets, production planning, discrete event systems, scheduling and control. He has served as a reviewer for a number of journals.



Kaizhou Gao received the B.Sc. and master degrees from China in 2005 and 2008, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, 2016. From 2008 to 2012, he was with the School of Computer, Liaocheng University, China. He was a Research Associate in the School of Electronic and Electrical Engineering (EEE), NTU, Singapore, from Feb. 2012 to Sep. 2013. From Oct. 2013 to Mar. 2015, he was a software engineer in Singapore Institute of Manufacturing Technology (SIMTech), A*star, Singapore. Since Apr. 2015, he was a Research Fellow in the School of Electronic and Electrical Engineering, NTU, Singapore. His research interests include intelligent computation, optimization, scheduling, and intelligent transportation. He has published over 50 refereed papers.

Ian Ware Simon received the B.S. degree in software development from Waikato University, New Zealand, in 2007, the Ph.D. degree in computer science from Waikato University in 2014. He is currently a Research Fellow with Nanyang Technological University, Singapore. His research interests include discrete event systems, optimization, scheduling, and control. He has served as a reviewer for a number of journals.

Yuting Zhu received the B.S. degree from Southeast University, Jiangsu, China, in 2016. Currently, she is a Ph.D. candidate at Nanyang Technological University, Singapore. Her research interests include discrete event systems, and supervisory control.

Rong Su (M'11-SM'14) received the B.E. degree in automatic control from University of Science and Technology of China, Hefei, China, in 1997, and the M.A.Sc. and Ph.D. degrees both in electrical engineering from the University of Toronto, Toronto, ON, Canada, in 2000 and 2004, respectively. Since then he was affiliated with University of Waterloo and Technical University of Eindhoven before he joined Nanyang Technological University, Singapore, in 2010. His research interests include discrete event systems, supervisory control, model-based fault diagnosis, multiagent systems, optimization and scheduling with applications in green buildings, flexible manufacturing, power management, and intelligent transportation systems. In the aforementioned areas, he has more than 130 publications in journals, book chapters, and conference proceedings, and two patents. Dr. Su is an Associate Editor of *Automatica, Journal of Discrete Event Dynamic Systems: Theory and Applications, Journal of Control and Decision,* and *Transactions of the Institute of Measurement and Control*. He is also the Chair of IEEE Control Systems Society Technical Committee on Smart Cities.