Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

SCHEDULING FOR AN AUTOMATED GUIDED VEHICLE IN FLEXIBLE MACHINE SYSTEMS

MengChang Wang

School of Mechanical and Aerospace Engineering
Nanyang Technological University
50 Nanyang Avenue
639798, Singapore

Yaoming Zhou

School of Aeronautic Science and Engineering
Beihang University
37th Xueyuan Road, Haidian District, Bejing
100191, China

ABSTRACT

To fulfill various customer demands by the same group of machines, flexible machine systems (FMS) are commonly used in fabrication facilities with heavy investment. FMS are computer-controlled systems consisting of several stations where each station specializes in particular operations with an appropriate transport system for the movement of products. In some wafer fabrication facilities, wafers are transferred between the load ports of machines within a bay via an automated guided vehicle (AGV). Multiple transportation tasks are required to be performed in time by the vehicle in order to secure the production schedules. The intra-bay vehicle scheduling problem is studied in this paper, which is strongly NP-hard. A heuristic search approach is proposed, incorporating two heuristic rules. Computational experiments show the effectiveness of the approach.

1 INTRODUCTION

Besides production machines, vehicles on shop floors are also considered as constraint resources (Rao et al. 2011, Rao et al. 2013). Vehicle routing and scheduling (VRS) becomes important in semiconductor production environments (SPEs), especially in 300mm wafer fabrication facilities, where overhead hoist transport (OHT) systems are widely equipped. Overhead hoist vehicles, which are also a kind of automated guided vehicles (AGVs), are suspended from ceiling-mounted rail and transfer wafers between load ports of machines or load ports and stockers (Nazzal and Bodner 2003). These rails are separated into cycles, which are referred to as bays. Each bay takes charge for a certain number of load ports, where vehicles pick up or drop down wafers. Typically, vehicles that keep circuiting on a specific bay is referred to as intra-bay vehicles, while those transporting wafers from one bay to another (between stockers) are inter-bay vehicles (Kim et al. 2007). An illustration is shown in Figure 1.

Similar to VRS studies in the traditional transportation or logistics field (Foster and Ryan 1976, Bott and Ballou 1986), both mathematical programming methods and vehicle routing rules are investigated in VRS in SPEs. However, due to the complexities of the problems themselves, optimal solutions for VRS in SPEs are hard to obtain (Nazzal and Bodner 2003, Montoya-Torres 2006). Instead of searching for optimal schedules, finding efficient routing rules is the main stream in this field. Among these studies, simulation is adopted to verify the effectiveness of proposed routing rules. Montoya-Torres (2006) presents an excellent review of VRP studies in semiconductor fabrication.

Most existing works focus on the utilization or travel times of vehicles. However, according to the Just-in-Time philosophy, it would be more critical to transport wafers to destinations in right times in order to secure production schedules in a fab. This paper concentrates on the scheduling of a single intra-bay vehicle, which is to determine proper sequence of transport tasks for the vehicle in order to minimize the maximum lateness. The paper is organized as follows. In Section 2, the problems settings are introduced.

978-1-4673-9743-8/15/$31.00 ©2015 IEEE

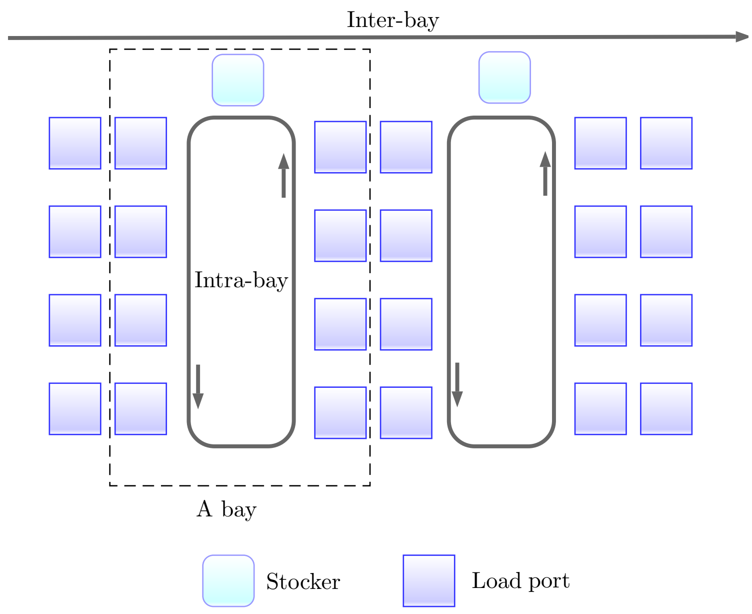

Figure 1: Bays, load ports and stockers

A mathematical formulation of the problem is presented in Section 3. An algorithm is proposed in Section 4, following which computational experiments are demonstrated and analyzed in Section 5. Section 6 gives a conclusion.

2 PROBLEM SETTINGS

A single AGV is equipped to transport wafers from one load port to another on a bay. An illustration of bays is shown in Figure 1. Without losing generality in practice, we assume that an AGV subjects to the following restrictions:

- The AGV moves at a constant speed;
- The AGV moves one way;
- The AGV can only carry one FOUP (front open unified pod) in a single transportation;
- The loading/unloading times at load ports are same and constant.

In addition, we assume that

- Capacities of load ports are large enough, so that the AGV can unload the FOUP at the time it arrives at a load port without any extra waiting;
- Wafers within a FOUP have the same production flows and scheduled times, so that FOUPs can be treated as basic units.

The transportation needs are specified according to predetermined production schedules on machines, which determine the start times and finish times of processing operations of FOUPs at machines. The vehicle has to pick up FOUPs after processing operations are finished and tries to deliver them to predefined next machines no later than scheduled times. A vehicle schedule has to determine a sequence of transportation tasks so that the maximum lateness of FOUPs is minimized.

3 THE MATHEMATICAL MODEL

Notations are listed as follows:

- I, set of all FOUPs;
- $i \in I$, FOUP index;
- t_i, earliest release time of FOUP i;
- f_i^0, pickup load port of FOUP i;
- f_i^1, drop-off load port of FOUP i;
- q_i, due time of FOUP i at its drop-off load port according to the production schedule;
- $l_{f,f'}$, travel time from load port f to f';
- $\varepsilon/2$, loading/unloading time of each FOUP;
- η_i, the time when the vehicle start to load FOUP i;
- τ_i, the time when FOUP i is unloaded from the vehicle to f_i^1;
- x_{ij}, binary variable. $x_{ij} = 1$, if task i is assigned before j;
- $\mathscr{M}$, the "big-M", a large enough positive number.

The model can be described as

$$P: \quad \min L = \max_{i \in I}\{q_i - \tau_i,\ 0\} \tag{1}$$

s.t.

$$\tau_i = \eta_i + l_{f_i^0, f_i^1} + \varepsilon,\ \forall i \in I; \tag{2}$$

$$\tau_i + l_{f_i^1, f_j^0} - \eta_j \leq (1 - x_{ij})\mathscr{M},\ \forall i, j \in I; \tag{3}$$

$$\eta_i \geq t_i; \tag{4}$$

$$x_{ij} + x_{ji} = 1,\ \forall i, j \in I. \tag{5}$$

Objective function (1) aims at minimizing the maximum lateness of all FOUPs. Equation (2) specifies the finishing time of each transportation. Constraint (3) guarantees that the vehicle has to finish the current transportation and move to the next load port before it loads the FOUP from the load port. Constraint (4) ensures that a FOUP will never be loaded before its release time. Constraint (5) makes sure that no task is assigned before and after another task at the same time.

Proposition 1 Problem (P) is strongly NP-hard.

Proof. Since the 3-Partition problem can reduce to scheduling problem $1|r_j|L_{max}$ (Pinedo 2012), which can further reduce to (P) by setting $q_i = d_i$, $l_{f_i^0, f_i^1} = p_i - \varepsilon$, $t_j = r_j$, and $l_{f_j^1, f_k^0} = 0$, for all $j, k \in I$ and $\varepsilon \geq 0$, where r_i is the release time of job i in the scheduling problem $1|r_j|L_{max}$, and p_i is the processing time of job i. Thus, problem (P) is strongly NP-hard. □

4 A HEURISTIC SEARCH APPROACH

Once a sequence of tasks are determined, the optimal solution in such case can be obtained by setting start times as $\eta_i^* = \max\left\{t_i, \tau_{i-1} + l_{f_{i-1}^1, f_i^0}\right\}$. Thus, a schedule can be constructed iteratively by selecting a remaining task in each iteration.

In this section, we present a heuristic search (HS) algorithm to construct task sequences. In order to keep the most promising selections in each iteration, two heuristic rules are incorporated to generate upper bound on the objective function L, namely, the most significant move rule (MSM) (Han and McGinnis

1989) and a new rule named the highest finish rate rule (HFR) which tries to maximize the finished tasks per unit time.

4.1 Two Heuristic Rules

MSM evaluates priorities of pickup stations according to due dates. The pickup station with the highest priority index is then selected. Considering that parallel machines with similar capacities are usually grouped under a bay in wafer fabs, we set the relative weight of processing time to 1.0 in the rule. Thus, the priority of task j at the finish time of task i is defined by

$$U_{i\to j} = \frac{1}{l_{f_i^1,f_j^0}} \exp\left(-\frac{s_j}{\bar{l}_i}\right), \forall j \in I_i, \tag{6}$$

where

- $I_i \subset I$, set of unprocessed tasks at the time τ_i;
- $s_j = q_j - l_{f_i^1,f_j^0} - \varepsilon - \tau_i$, the slack time of task j;
- $\bar{l}_i = \frac{\sum_{\forall k\in I_i} l_{f_i^1,f_k^0}}{|I_i|}$, the average travel time from f_i^1 to the next load port.

We define the *finish rate* of a task as the average number of tasks processed in unit time from the beginning to the end of the current delivery. For a sequence ended with i, the finish rate of candidate task j is specified by:

$$D_{i\to j} = \frac{|I| - |I_i| + 1}{\tau_j}, \forall j \in I_i. \tag{7}$$

Considering that a higher finish rate means more time available for less tasks in the following steps, the second rule selects the candidate task with the highest finish rate (HFR) in each step.

4.2 A Heuristic Search Algorithm

For a partially constructed task sequence ended with task i, where the maximum lateness is L_k (see Figure 2), denote B_i as the larger value between the maximum latenesses given by applying the MSM rule and the HFR rule on the remaining tasks (L_{MSM} and L_{HFR} in Figure 2), we can get a upper bound of L as $\tilde{L} = \max\{L_k, B_i\}$. For any candidate task (like j or $j^{'}$), it can be pruned if its lateness is larger than $\tilde{L}$, as it is dominated by the solution given by the rules.

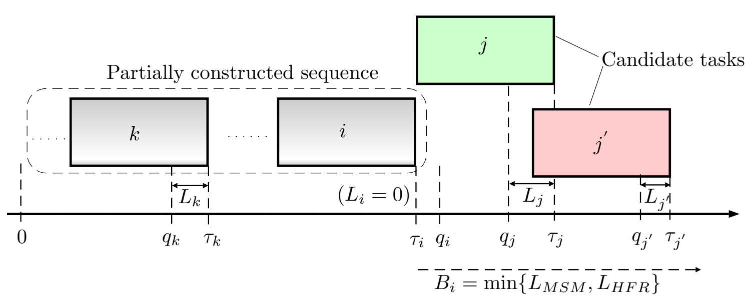

Figure 2: Evaluation of candidate tasks in a step

In addition, the following rule is introduced in the algorithm.

Proposition 2 *For a partially constructed task sequence ended with task i, unassigned task j can be excluded from candidate tasks for the task subsequent to i, when there exists an unassigned task $j^{'}$ such* *that*

$$\max\{\tau_i + l_{f_i^1 f_{j'}^0}, t_{j'}\} + l_{f_{j'}^0 f_{j'}^1} + \varepsilon + l_{f_{j'}^1 f_j^0} \le t_j. \quad (8)$$

Equation (8) means that task j' can be inserted and finished between tasks i and j before the earliest start time of j. Thus, placing task j before j' will cause unnecessary lateness to j'.

The procedure of this heuristic search approach is described as follows:

1. $S \leftarrow$ set of a dummy task indicating the initial vehicle position;
2. $\mathbb{S} = \{S\}$; $\mathbb{U} = \emptyset$; $\tilde{L} = \infty$;
3. for each $S \in \mathbb{S}$,
   3.1. $L_S \leftarrow \max_{\forall j \in S}\{\tau_j - q_j, 0\}$
   3.2. $R \leftarrow I \backslash S$; $i \leftarrow$ the last task in S; $B_i \leftarrow \min\{L_{MSM}, L_{HFR}\}$;
   3.3. if $\tilde{L}_S > \max\{L_S, B_i\}$, $\tilde{L} = \max\{L_S, B_i\}$;
   3.4. for each task $j \in R$,
      3.4.1 if j can be excluded by Proposition 2, go to step 3.4 for next task;
      3.4.2 $\delta_j \leftarrow \max\{\tau_j - q_j, 0\}$, where $\tau_j = \max\{t_j, \tau_i + l_{f_i^1 f_j^0}\} + l_{f_j^0 f_j^1} + \varepsilon$;
      3.4.3 if $\delta_j > \tilde{L}$, goto step 3.4 for next task;
      3.4.4 $S_j \leftarrow S \cup \{j\}$; $\tilde{L}_{S_j} \leftarrow \max\{L_{S_j}, B_j\}$;
      3.4.5 if $|\mathbb{U}| < K$, $\mathbb{U} \leftarrow \mathbb{U} \cup \{S_j\}$; otherwise, if $\exists S' \in \mathbb{U}$ such that $\tilde{L}_{S_j} < \tilde{L}_{S'}$, $\mathbb{U} \leftarrow \mathbb{U} \backslash \{S'\} \cup \{S_j\}$;
4. $\mathbb{S} \leftarrow \{S | L_S \le \tilde{L}, S \in \mathbb{U}\}$; $\mathbb{U} = \emptyset$;
5. if $I \backslash S = \emptyset, \forall S \in \mathbb{S}$, return $S^* \in \mathbb{S}$ such that $L_{S^*} \le L_S, \forall S \in \mathbb{S}$; otherwise, go to step 3.

In the procedure, K is the maximum number of promising sequences to be kept in the search. A larger K indicates to search the solution space more thoroughly but at the cost of more computational efforts. Its value can be set tentatively to obtain a trade-off.

5 EXPERIMENTS AND ANALYSIS

In practice, FOUPs are usually processed in batches. For example, in diffusion area, four to seven FOUPs are loaded in to a furnace at a time. Larger batch sizes mean that more FOUPs will be released at the same time. The processing times of different FOUPs at the same machine may differ from each other according to their production processes. FOUPs may be transported to another machine under the same bay or to another bay. For FOUPs that will be transported outside the bay, the intra-bay vehicle just drops them off at the stocker. Besides, there are urgent FOUPs which should be processed with higher priorities than others. The time slacks between their release times and latest finish time for the transportation are much smaller than the common FOUPs.

Considering the layout of a bay illustrated in Figure 3, scenarios with 50 transportation tasks and different processing times of machines (intervals between release times (IRT) of transportation tasks), batch sizes (BS), the percentage of intra-bay FOUPs (PIF), and the percentage of urgent FOUPs (PUF) are designed (see Table 1). In these settings, urgent FOUPs are expected to be transported within 15 time units, while the times allowed for common FOUPs are drawn from a normal distribution, $Normal(100, 10)$. Experimental instances are constructed according to the orthogonal array $L_9(3^4)$ as in Table 2, where each group includes ten instances.

The proposed approach is implemented within C++ on a computer with a 1.8 GHz CPU and a 2 GB RAM. The parameter K is set to 20. Performances of MSM, HFR, and the proposed HS are compared in Table 3

As observed from Table 3, the introduced HFR is able to improve the bounds considerably in some circumstances (see groups 4, 7, 8, and 9). By incorporating both HFR and MSM, the proposed HS obtains significant improvements to MSM for all experimental instances. As shown in Figure 4, the most

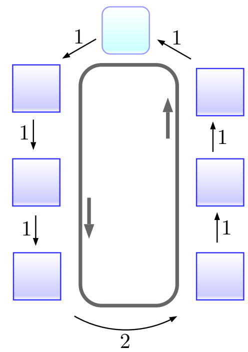

Figure 3: Layout of the bay (numbers on arrows are travel times)

Table 1: Parameters of instances

Level	IRT	BS	PIF(%)	PUF (%)
1	$Normal(50,10)$	$Uniform(1,3)$	10	5
2	$Normal(100,40)$	$Uniform(1,7)$	50	10
3	$Normal(200,80)$	$Uniform(4,7)$	80	20

Table 2: Settings of instances

Group	IRT Level	BS Level	PIF Level	PUF Level
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

significant factor that impacts the performances of HFR relative to MSM is the intervals between release times (processing times of machines). HFR tends to perform better than MSM when the machine processing times are longer. In addition, HFR is also likely to outperform MSM when there are smaller batch sizes or larger portion of urgent FOUPs. The trends of improvement to MSM by the proposed HS are shown in Figure 5, which indicates that the batch sizes impact the improvement most significantly, and larger batch sizes are likely to results in less improvements.

6 CONCLUSION

The ability to transport wafers to right machines within proper times is crucial in the execution of production schedules. In order to secure predefined production schedules, a single intra-bay vehicle routing problem aiming at minimizing the maximum lateness of transportation tasks in wafer fabrication is studied in this paper. The problem is shown to be strongly NP-hard.

A heuristic search approach is proposed, which constructs transportation task sequences by keeping the most promising candidates. Two heuristic rules is utilized to generate upper bounds, which are able to prune considerable part of inferior candidates in advance. One is MSM proposed by Han and McGinnis

Table 3: Performances (on average)

Group	L_{MSM}	L_{HFR}	L_{HS}	Improvement (to MSM) (%)	Improvement (to *HFR*)(%)
1	202.77	310.46	146.20	23.42	51.62
2	232.94	278.34	186.28	19.93	32.50
3	206.46	274.65	173.81	15.65	35.72
4	363.66	53.25	23.84	92.79	53.75
5	189.61	236.69	111.50	40.73	51.99
6	290.86	348.50	245.98	15.23	29.20
7	662.87	21.54	10.36	98.41	48.33
8	198.38	159.10	81.43	41.89	49.79
9	239.48	230.07	132.69	42.90	40.61

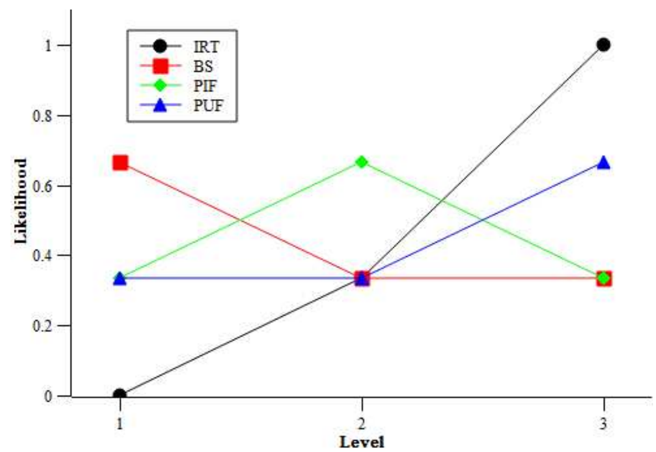

Figure 4: Trends of likelihood that HFR outperforms MSM

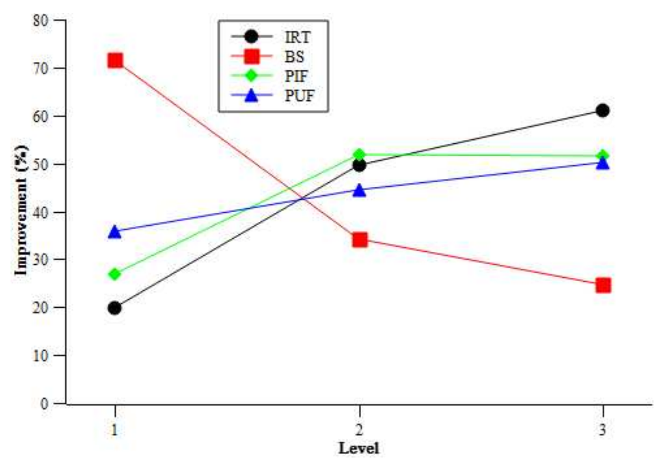

Figure 5: Trends of improvement to MSM by HS

(1989) , and the other is a novel rule which is found to improve the bounds by MSM in some situations. By incorporating these two rules, as well as another rule to further reduce the searching space, the HS obtains significant improvements to MSM in computational experiments.

In addition to our basic model, the practical situation can be more complicated: a station may have limited buffer capacity (Wu and Zhao 2015a) or discretionary priority (Zhao et al. 2015) and have parallel or serial batches (Wu 2014b, Wu et al. 2011b, Wu et al. 2014) with different impact on queue time and productivity (Wu et al. 2007). All these will complicate the model and are left for future research. The transportation delay caused by AGV will introduce addition interruptions to a station (Wu et al. 2008, Wu et al. 2011a), and affect the length of service time (Wu and Hui 2008), which will inevitably increase the variability of a manufacturing system (Wu 2005). The impact of transportation delay on queue time can be modeled as resource contention (Wu 2014a). This delay will affect the performance of the downstream stations (Wu and McGinnis 2012, Wu and Zhao 2015b) and could be captured by the intrinsic ratio (Wu and McGinnis 2013). The impact of queue time performance due to the transportation delay is left for future research.

ACKNOWLEDGMENT

This research is supported in part by GSK-Singapore Partnership for Green and Sustainable Manufacturing under Grant M406884.

REFERENCES

Bott, K., and R. H. Ballou. 1986. "Research perspectives in vehicle routing and scheduling". *Transportation Research Part A: General* 20 (3): 239–243.

Foster, B. A., and D. M. Ryan. 1976. "An integer programming approach to the vehicle scheduling problem". *Operational Research Quarterly*:367–384.

Han, M.-H., and L. F. McGinnis. 1989. "Control of material handling transporter in automated manufacturing". *IIE transactions* 21 (2): 184–190.

Kim, B.-I., S. Oh, J. Shin, M. Jung, J. Chae, and S. Lee. 2007. "Effectiveness of vehicle reassignment in a large-scale overhead hoist transport system". *International journal of production research* 45 (4): 789–802.

Montoya-Torres, J. 2006. "A literature survey on the design approaches and operational issues of automated wafer-transport systems for wafer fabs". *Production Planning & Control* 17 (7): 648–663.

Nazzal, D., and D. Bodner. 2003. "A simulation-based design framework for automated material handling systems in 300 mm fabrication facilities". In *Simulation Conference, 2003. Proceedings of the 2003 Winter*, Volume 2, 1351–1359. IEEE.

Pinedo, M. L. 2012. *Scheduling: theory, algorithms, and systems*. Springer Science & Business Media.

Rao, Y. Q., M. C. Wang, and K. P. Wang. 2011. "JIT single Vehicle scheduling in a mixed-model assembly line". In *Advanced Materials Research*, Volume 211, 770–774. Trans Tech Publ.

Rao, Y.-Q., M.-C. Wang, K.-P. Wang, and T.-M. Wu. 2013. "Scheduling a single vehicle in the just-in-time part supply for a mixed-model assembly line". *Computers & Operations Research* 40 (11): 2599–2610.

Wu, K. 2005. "An examination of variability and its basic properties for a factory". *Semiconductor Manufacturing, IEEE Transactions on* 18 (1): 214–221.

Wu, K. 2014a. "Classification of queueing models for a workstation with interruptions: a review". *International Journal of Production Research* 52 (3): 902–917.

Wu, K. 2014b. "Taxonomy of batch queueing models in manufacturing systems". *European Journal of Operational Research* 237 (1): 129–135.

Wu, K., and K. Hui. 2008. "The determination and indetermination of service times in manufacturing systems". *Semiconductor Manufacturing, IEEE Transactions on* 21 (1): 72–82.

Wu, K., and L. McGinnis. 2012. "Performance evaluation for general queueing networks in manufacturing systems: Characterizing the trade-off between queue time and utilization". *European Journal of Operational Research* 221 (2): 328–339.

Wu, K., and L. McGinnis. 2013. "Interpolation approximations for queues in series". *IIE transactions* 45 (3): 273–290.

Wu, K., L. McGinnis, and B. Zwart. 2011a. "Queueing models for a single machine subject to multiple types of interruptions". *IIE transactions* 43 (10): 753–759.

Wu, K., L. F. McGinnis, and B. Zwart. 2007. "Compatibility of Queueing Theory, Manufacturing Systems and SEMI Standards". In *Automation Science and Engineering, 2007. CASE 2007. IEEE International Conference on*, 501–506. IEEE.

Wu, K., L. F. McGinnis, and B. Zwart. 2008. "Queueing models for single machine manufacturing systems with interruptions". In *Simulation Conference, 2008. WSC 2008. Winter*, 2083–2092. IEEE.

Wu, K., L. F. McGinnis, and B. Zwart. 2011b. "Approximating the Performance of a Batch Service Queue Using the $M/M^k/1$ Model". *Automation Science and Engineering, IEEE Transactions on* 8 (1): 95–102.

Wu, K., Z. Wu, N. Zhao, and Y. Xu. 2014. "Approximating the performance of a station subject to changeover setups". In *Proceedings of the 2014 Winter Simulation Conference*, 2396–2403. IEEE Press.

Wu, K., and N. Zhao. 2015a. "Analysis of dual tandem queues with finite buffer capacity and nonoverlapping service times subject to breakdowns". *IIE Transactions*. Accepted. DOI: 10.1080/0740817X.2015.1055389.

Wu, K., and N. Zhao. 2015b. "Dependence among single stations in series and its applications in productivity improvement". *European Journal of Operational Research*. Accepted. DOI: 10.1016/j.ejor.2015.05.028.

Zhao, N., Z. Lian, and K. Wu. 2015. "Analysis of a MAP/PH/1 Queue with Discretionary Priority". *Asia-Pacific Journal of Operational Research*. Accepted. DOI: 10.1142/S0217595915500426.

AUTHOR BIOGRAPHIES

MENGCHANG WANG is a postdoctoral research fellow at Nanyang Technological University. He holds M.S. degree in Mechanical Engineering from Wuhan University of Technology and Ph.D. degree in Industrial Engineering from Huazhong University of Science and Technology. His research interests include scheduling, vehicle routing, combinatorial optimization, and manufacturing systems. His email address is wangmc@ntu.edu.sg.

YAOMING ZHOU received the Bachelor degree of Aircraft design in 2008, and the PhD degree of Aircraft design in 2013, from school of Aeronautic Science and Engineering, Beihang University. Now he serves as a Lecturer and Master Tutor at the School of Aeronautic Science and Engineering, Beihang University. His research interests include intelligent optimization, aircraft design, neural networks, fault diagnosis and state prediction, intelligent control of unmanned aerial vehicle and unmanned rotorcraft. He is the corresponding author of this paper and his email address is zhouyaoming@buaa.edu.cn.