Efficient Simulation-Based Composition of Scheduling Policies by Integrating Ordinal Optimization With Design of Experiment

Bo-Wei Hsieh, Chun-Hung Chen, *Senior Member, IEEE*, and Shi-Chung Chang, *Member, IEEE*

***Abstract*—Semiconductor wafer fab operations are characterized by complex and reentrant production processes over many heterogeneous machine groups with stringent performance requirements. Efficient composition of good scheduling policies from combinatorial options of wafer release and machine dispatching rules has posed a significant challenge to competitive fab operations. In this paper, we design a fast simulation-based methodology by an innovative integration of ordinal optimization (OO) and design of experiments (DOEs) to efficiently select a good scheduling policy for fab operations. Instead of finding the exact performance among scheduling policies, our approach compares their relative orders of performance to a specified level of confidence. Our new approach consists of three stages: performance estimation model construction using DOE, policy option screening process, and final simulation evaluation with intelligent computing budget allocation. The exponential convergence of OO is integrated into all the three stages to significantly improve computational efficiency. Simulation results of applications to scheduling wafer fabrications not only screen out good scheduling policies but also provide insights about how factors such as wafer release and the dispatching of each machine group may affect production cycle times and smoothness under a reentrant process flow. Most of the OO-based DOE simulations require 2–3 orders of magnitude less computation time than those of a traditional approach. Such a high speedup enables decision makers to explore much larger problems.**

***Note to Practitioners*—This paper designs a fast simulation-based methodology to compose a good scheduling policy from various dispatching rules of fab operations. The methodology innovatively applies DOE to estimate performance of dispatching rule combinations (policies) over various machines groups in a fab, screens out good enough policy options by using OO over the performance estimation, and allocates computation time intelligently to simulate potentially good options. Our study shows that OO-based DOE simulations require 2–3 orders of magnitude less computation time than those of a traditional approach. The high speedup enables fab managers to identify good scheduling policies from the many combinations of wafer release and dispatching rules.**

***Index Terms*—Design of experiment (DOE), ordinal optimization (OO), scheduling policy composition, semiconductor wafer fab.**

I. Introduction

SEMICONDUCTOR wafer fabrication faces stringent challenges of volatile product demands, very short time to market, complex but fast evolving process technology, skyrocketing capital investment, and highly cost-sensitive competition. Market demands for consumer electronic products, for example, not only are diversified but may also have a time window as short as 3–6 months. Capital cost of a 300 mm wafer fabrication factory (fab) of subwavelength technology can easily go beyond 2 billion U.S. dollars, let alone the investment needed for gigafabs [1], [2], where the cost of capital equipment now constitutes more than 75% of the total cost. Flexible and efficient use of fab production resources for quick response to market demands has been essential for improving manufacturing productivity and critical to the competitiveness of fab operations.

Effective factory scheduling and dispatching plays a key role in improving equipment reliability and utilization and in cycle time reduction and on-time delivery [3]. This production control function decides how wafers should be released into a fab and how they should be dispatched among machines for processing, which will be referred to as the scheduling problems hereafter. In order to fully utilize the expensive production equipment, it is imperative that effective scheduling and dispatching tools be utilized. Several factors complicate fab scheduling.

1) Diverse product types: There may be hundreds of product types, such as analog, digital, and mixed-signal circuits, memory, or central processing units, involving various process technologies and hundreds of processing steps.
2) Reentrant production flows: The production flow of each product type may revisit the same type of machines, i.e., the same machine group, a few times, which leads to capacity competition among processing steps of a product.
3) Diverse and failure-prune machines: There are hundreds of sophisticated, delicate, and costly machines with di-

Manuscript received August 1, 2006; revised February 28, 2007. This paper was recommended for publication by Associate Editor T. Chen and Editor N. Viswanadham upon evaluation of the reviewers' comments. This work was supported in part by the National Science Council, Taiwan, R.O.C., under Grant NSC89-2213-E-002-119, Grant NSC 94-2213-E-002-015, and Grant NSC 95-2811-E-002-009, in part by the Semiconductor Research Corporation and International SEMATECH under FORCe Project 090E1092, in part by the National Science Foundation under Grant IIS-0325074 and Grant DMI-0323220, in part by the NASA Ames Research Center under Grant NNA05CV26G, and in part by FAA under Grant 00-G-016.

B.-W. Hsieh was with the Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan, R.O.C. He is now with VIA Technologies, Inc., Fremont, CA 94539 USA (e-mail: bwhsieh@ntu.edu.tw).

C.-H. Chen is with the Department of Systems Engineering and Operations Research, George Mason University, Fairfax, VA 22030 USA (e-mail: cchen9@gmu.edu).

S.-C. Chang is with the Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan, R.O.C. (e-mail: scchang@cc.ee.ntu.edu.tw).

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/TASE.2007.906342

verse characteristics and constraints/limitations and unpredictable machine failures.

4) Emergent automatic material handling systems (AMHSs): The operation of emerging AMHS needs to be characterized and the dispatching of AMHS needs full integration with fab scheduling.

Compounded with the dynamic and uncertain market demands, these complicating factors have posed unique challenges to effective production scheduling in a fab.

Literature on fab scheduling has long supported that both wafer release and lot dispatching have significant impacts on fab performance. In academic research, Wein [4] conducted simulation studies and showed that a good selection of wafer release policy leads to 44.9% improvement in mean cycle time. Simulation analyses of Lu *et al.* [5] demonstrated that a good dispatching rule can reduce mean cycle time by 22% over the simple first-come–first-serve (FCFS) rule. The standard deviation of cycle time reduction up to 52% can also be achieved using a proper dispatching rule. Practitioners have confirmed such academic assessments. An implementation of a good wafer release policy in a GaAs fab achieved a 40%–60% reduction in mean cycle time [6], and by using a good dispatching rule, a Motorola fab reported 32.9% reduction in mean cycle time [7]. Reference [8] reported that a good scheduling rule is beneficial for production smoothness and equipment utilization. The researchers of [9] performed simulation studies of AMHS vehicle and machine dispatching rules over SEMATECH fab data of actual production fabs. Their analyses show that vehicle and machine dispatching rules as well as their interactions have significant impact on cycle time, wafer-in-process (WIP) and on-time-delivery.

There exist many decision rules for wafer release (input regulation) [5], [10], [11] and lot dispatching [5], [12]–[14]. Fab managers often select from such existing rules, compose them into a fab scheduling policy to support the manufacturing philosophy and implement the policy via a software suite for real-time dispatching. Simulation evaluation has been widely adopted by the industry for fab scheduling policy composition [4], [5], [15]–[17].

The simulation approach has the advantages of high fidelity and modeling flexibility in coping with the fast-changing characteristic of wafer fabrications, especially for foundry fabs, which are characterized by high-variety, low-volume, and make-to-order production. However, the fidelity and flexibility often comes at a computational expense. To obtain a sound statistical estimate of performance, extensive simulation experiments are required for the candidate rule options [19].

Rule selection by using the traditional simulation approaches is not fast enough in computation for short-term scheduling of fab operations. Hsieh *et al.* [18] explored the application of the ordinal optimization (OO)-based simulation technique [20] to improve simulation efficiency of selecting good rules for scheduling wafer fabrications under different factors such as initial state, performance index, and time horizon. The major strength of OO is that the relative order comparison of performance measures converges much faster than the performance measures themselves do. Under certain conditions, the rate of convergence for OO can be exponential [21], [22].

However, the number of candidate rule and policy options grows in a combinatorial way with various control factors of production. Given a fab operation goal, a set of heterogeneous machines, an initial state of the fab, and a specific time horizon, the composition of scheduling policies determines how wafer lots should be released into the fab and how they should be dispatched among machines for processing so that the operation goal can be effectively achieved. The number of candidate policy options grows in a combinatorial way with the number of machine groups and over the planning time horizon if rules are changed dynamically according to fab states. With this new challenge, a brute-force application of the OO method to selecting a good scheduling policy is still infeasible.

The DOE method has been effective in reducing the number of options to be evaluated for combinatorial problems [23]. With the assumption that high-order interaction effects on a performance function of interest caused by several factors are not significant and negligible, performance evaluation of a large portion of redundant options can be saved through the use of orthogonal array [24].

In this paper, we design a fast simulation-based methodology by an innovative integration of OO and DOE to efficiently select a good scheduling policy for fab operation. The new method is called OO-based DOE simulation, which consists of three major stages: performance estimation model construction using DOE, policy option screening process, and final evaluation using efficient simulation. The fast convergence of OO is utilized by integrating the notion of ordinal comparison into all the three stages. Furthermore, an OO-based technique called optimal computing budget allocation [25] is applied in the final evaluation stage. By intelligently determining the effective allocation of samples/replications for each option, the efficiency can be further improved beyond the exponential convergence of OO.

We also investigate the application of the OO-based DOE simulation to efficiently select scheduling policies among machine groups, where different rule sets are adopted by different machine groups based on machine characteristics. Policy options are selected from the prominent ones of [4], [5], [12], [26], whose properties are also explored. The single product model of [5] is adopted for benchmarking. Properties of scheduling rules for individual machine groups are studied, which include performances under different factors of operation objectives, fab initial states, and scheduling horizons. The effectiveness of the proposed OO-based DOE simulation for such an application is also investigated. Simulation results of applications to scheduling wafer fabrications show that most of the OO-based DOE simulations require 2–3 orders of magnitude less computation time than those of a traditional approach, and the speedup is up to 7000 times in certain cases for such a small-to-medium-size problem. We anticipate the speedup can be even higher when larger problems are considered.

The remainder of this paper is organized as follows. Section II describes the problems of selection among rules for various machine groups. Challenges of applying simulation approaches to handle the combinatorial complexity of the problems are also given. Motivated by the deficiencies of applying traditional simulation approaches to scheduling rule selection problems, OO and DOE methods that can significantly reduce the required

simulation time are described in Section III. Section IV presents the design of an efficient simulation methodology, which integrates the notions of OO and DOE to handle combinatorial option selection problems. Rule selection experiments for a fab are conducted in Section V and effectiveness of the OO-based DOE simulation method is analyzed in Section VI. Finally, Section VII concludes this paper.

II. Composition of Scheduling Policy Among Machine Groups

Scheduling a semiconductor wafer fab is challenging because of the complexity and uncertainty of a fab. For example, hundreds of process steps on 50–120 different types of equipment are required to fabricate an integrated circuit on a silicon wafer. Due to the diversity of equipment in a fab, scheduling rules for each machine group should be designed based on the specific characteristics and operation goals of the machine group. Empirical or heuristic scheduling rules are collected for individual machine groups. Such rule collections are built into a scheduling rule library of a fab. There are two major challenges for simulation-based approaches: i) *uncertainty*: a large number of simulation replications must be performed for each policy option to have a good estimate of the system performance and ii) *combinations*: the number of candidate policies grows in a combinatorial way with respect to the number of machine groups and the number of alternative policies at one machine group. The simulation time required for such an enormous number of policies is usually costly formidable.

Specifically, denote $h(\Theta_i; w)$ as the performance measure of a policy i, where Θ_i is a vector of design parameters for policy i and w is a random vector that represents uncertain factors in the system. Our objective is to find a best policy with maximum expected performance measure

$$\max_{\Theta_i \in \Theta} E_w[h(\Theta_i; w)] \tag{1}$$

where Θ, the search space for candidate policies, is an arbitrary, huge, structureless but finite set. Note that for complex systems considered in this paper, $h(\Theta_i; w)$ is available only in the form of a complex calculation via simulation. The system constraints are implicitly involved in the simulation process, and so are not shown in (1).

If we conduct independent simulations and the variance is finite, the strong law of large numbers dictates that the following property holds with probability 1:

$$\frac{1}{n}\sum_{j=1}^{n} h(\Theta_i; w_j) \to E_w[h(\Theta_i; w)], \text{ as } n \to \infty. \tag{2}$$

While it is not possible to conduct an infinite number of simulation replications, one would have to perform a large number of simulations in order to ensure the estimation accuracy is sufficient, i.e., n must be large. Adding the complexity that the size of Θ is extremely large and the search space can be structureless, the total simulation time can be prohibitively expensive.

Consider the fab model of [5] as an example, where various products are aggregated into one product type and there are 12 stations (machine groups) and 60 production stages. Fig. 1 depicts the reentrant process flow. A lot is the basic unit of material handling and transportation, usually consisting of 25 wafers. There are three factors typically considered for scheduling policy composition: wafer release rule, dispatching rule of a machine group, and performance index. These factors are elaborated as follows.

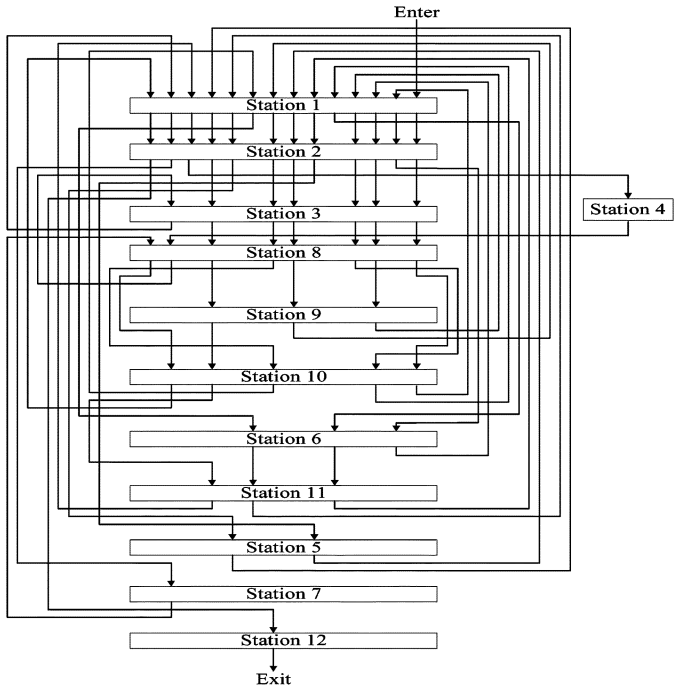


Fig. 1. Process flow of a fab model.

A. Wafer Release and Machine Dispatching Rules

Wafer release rules apply to wafer entry into the production model of Fig. 1 at station 1. Four frequently used wafer release rules include deterministic release, workload regulation, batch deterministic release, and batch workload regulation. Batch release rules are designed for regulating batch processing at station 8. Photolithography and furnace are two of the most critical machine types in a fab. Due to their specific process characteristics, different dispatching rule sets are applied. There are four sets of dispatching rules in this example: lot assignment rules for photolithography (LAP_PH), dispatching rules for general machine groups (LDR), dispatching rules for furnace machine group (LDR_F), and dispatching rules for upstream of furnace machine group (LDR_UF), where each set has four rules. Lot assignment policies for photolithography machine group dedicate a lot to a photo machine by means of the notion of workload balancing. Dispatching rules for furnace machine group first select a stage with certain criteria and then choose a batch of lots from the stage for processing. To avoid furnace machine group from starving, special rules are designed for the immediately upstream machine groups to the furnace machines. The dispatching of all other machine groups applies the set of general rules. Please refer to Tables X–XV of Appendix A for the detailed descriptions of machine groups and the sets of wafer release and dispatching rules.

B. Performance Indices

Mean cycle time, variance of cycle time, and smoothness of a fabrication line are three scheduling performance indices in industrial practice. Details of these indices are described as follows.

- Mean cycle time (MCT): The cycle time of a lot is the time elapsed from lot release to completion of entire processing. Many fab operation practitioners pursue MCT reduction while meeting the output volume requirement. For wafer fabrication lines, especially for foundry fabs, a short cycle time leads to customers' short time to market and corresponds to low WIP level.
- Variance of cycle time (VCT): Reduction of VCT leads to accuracy in delivery commitment and facilitates synchronization among operations. The VCT measurement also reveals the stability of the manufacturing processes.
- Smoothness (SM): Ideally, it is desirable that individual stages in a fab be run at a constant production rate. In reality, this can hardly be achieved because of fabrication constraints and uncertainties. Let r_i be a prespecified target production rate at stage i and $[0, t]$ be a time interval of observation. One definition of smoothness is $\mathrm{SM} \equiv \sum_{i=1}^{N_s} \theta_i / N_s$, where N_s is total number of stages and θ_i is the score at stage i and is defined to have value 1 if the completed number of wafer moves of stage i up to time t is greater than $0.9 r_i t$ and have value 0, otherwise.

A scheduling policy is a composition of a wafer release rule and lot dispatching rules of individual machine groups. It corresponds to Θ_i defined in (1). It is assumed that all the general machine groups will adopt the same dispatching rule. In this fab model, there are, therefore, 1024 ($4 \times 4 \times 4 \times 4 \times 4$) scheduling policies by combining wafer release rule options and dispatching rule options of four distinct types. Such combinatorial composition of policy options must be evaluated by simulation to idenfify good policy options over individual performance indices. Although each rule set has the same number of rules and only three specific performance indices are considered in this example, they are not limited to the simulation-based composition method we shall develop below.

III. Ordinal Optimization (OO) and Design of Experiments (DOEs)

Our scheduling policy composition method is developed by integrating two techniques: i) OO and ii) DOE. This section gives an overview of each technique.

A. Ordinal Optimization (OO)

Although the estimate $\hat{E}_w[h(\Theta_i; w)] \equiv (1/n) \sum_{j=1}^{n} h(\Theta_i; w_j)$ converges very slowly as n goes to infinity, recent research has shown that comparing relative orders of performance measures converges much faster than the performance measures themselves do. Ordinal optimization refers to the general approach that selects a subset of alternatives from the design space based on the ordinal ranking of the designs, and focus on selection of a good or best alternative rather than on accurate estimation of all designs' performance. Suppose we select an option b using the following criterion:

$$\{b\} \equiv \arg\max_i \hat{E}_w[h(\Theta_i; w)].$$

Definition 1: Define *correct selection*-1 (CS_1) as the event that the selected option b is actually the best option. Define the *confidence probability* $P\{\mathrm{CS}_1\} \equiv \Pr\{\text{The current top-ranking option } b \text{ is actually the best option.}\}$

Based on the results from OO, it is possible to establish the relative order of $\hat{E}_w[h(\Theta_i; w)]$ efficiently, i.e., to achieve high $P\{\mathrm{CS}_1\}$, although the value of $\hat{E}_w[h(\Theta_i; w)]$ may converge slowly. In particular, Dai [21] showed that $P\{\mathrm{CS}_1\}$ can converge to 1.0 exponentially. A critical issue in the application of OO is the estimation of the $P\{\mathrm{CS}_1\}$ itself.

Theorem 1: [27] Let $\tilde{J}_i, i \in \{1, 2, \ldots, b-1, b, b+1, \ldots, R\}$, denote the random variable whose probability distribution is the posterior distribution of the expected performance for option i under a Bayesian model. For a maximization problem

$$P\{\mathrm{CS}_1\} \geq \prod_{i=1, i \neq b}^{R} P\{\tilde{J}_b > \tilde{J}_i\} \equiv \text{Approximate Probability of Correct Selection-1}(\mathrm{APCS}_1). \quad (3)$$

The posterior distribution $p(\tilde{J}_i)$ consists of information from both the prior distribution and the simulation results $\{h(\Theta_i; w_j), j = 1, 2, \ldots, n\}$. Chen [27] showed that APCS_1 is a good approximation to $P\{\mathrm{CS}_1\}$.

B. Design of Experiments (DOEs)

The DOE method plays an important role in quality design and control in recent decades. It is used to experiment with various combinations of design factors for the purpose of identifying the particular combination that optimizes certain design criteria or performance measure. Among different DOE methods, *fractional factorial design of experiments* (FFDOE) is particularly useful in experiments with several design factors simultaneously [23], [28]. By exploiting orthogonal arrays (OAs), the FFDOE method is utilized to construct a good model to estimate performance measures of all options by simulating only a very small fraction of policy options, having the potential to significantly reduce computation time. The first step is to identify factors that affect the performance function of interest and choose different levels of design parameters for each factor. An OA is then used to determine a small set of test samples for evaluation by simulation. Simulation runs are conducted for the set of test samples. We construct a performance estimation model for all options using the simulated performance measures of the test samples.

An OA specifies which level combinations among options should be used to determine the test samples. The rows of an array represent the experiments or test samples to be performed. The columns of the OA correspond to the different factors whose effects are being analyzed. The entries in the array specify the levels of factors. Denote an OA as $\mathrm{OA}(n_S, n_F, n_L, s)$, where n_S is the size of the test sample set to be performed, n_F is the number of factors, n_L is the number of levels of each factor, and s is the number of columns where we are guaranteed to see all the possible combinations of levels an equal number of times. The number s is called strength.

In practice, an OA of strength 2 is commonly adopted concerning about the cost of experiments. Sampling using OAs, especially those of lower strength, may reduce computation time dramatically. Fig. 2 compares the numbers of all possible

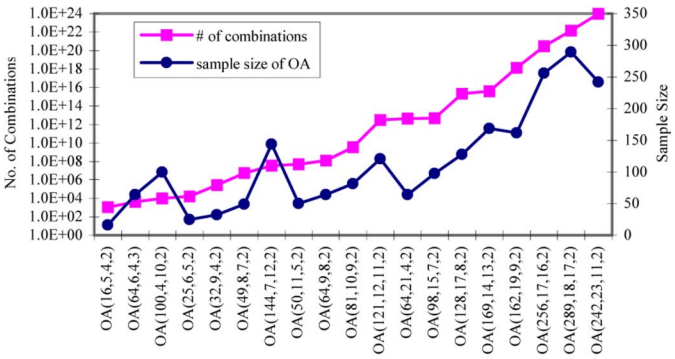


Fig. 2. Time saving by OA sampling.

combinations (number of options) with the sizes of several frequently used OAs of strength 2. The left vertical axis is the number of possible combinations in log scale and the right vertical axis is the OA sizes in linear scale. The number of combinations increases in an exponential way as the number of factors and levels increase while the OA size increases in a linear way. Multiple orders of computation time savings might be achieved for large problem if an appropriate OA is adopted.

IV. New Development of OO-Based DOE Simulation

In this section, an efficient simulation methodology is designed to handle scheduling rule selection problems for fab operation, which integrates OO with DOE methods. As the scheduling policy composition has a combinatorial nature in possible options both over time and among machine groups, efficiency is the most critical concern in the development of our methodology. We adopt the DOE method to reduce the number of scheduling policies to be simulated. Instead of simulating all policy options to obtain their performance measures, DOE is exploited to find an OA and construct a performance estimation model of all policies with only the policies in the OA simulated. This eliminates the need of simulating a very big number of options.

To further improve the simulation efficiency of DOE, we integrate the notion of OO into our approach. Traditional DOE intends to establish a performance estimation model which is uniformly good over the entire design space by, for example, minimizing the model parameter estimation variance. However, we are interested in finding a good policy option rather than to have a good estimation for *all* policies. The model accuracy is more critical near good policies. The compromise in the model accuracy near bad policies can lead to another substantial reduction in computation time.

Since the DOE estimation model is an approximation, the best policy option selected using the estimation model is not necessarily the true best option. To reduce the chance of selecting a bad option, we screen a subset of policies for more extensive simulation evaluation and the true best option in the subset is chosen as the final solution. The way of screening subsets has significant impact on the final solution quality and the associated computation load. Ideally, this subset should contain a best option or at least a good option. A bigger subset will increase the probability of containing a good option. On the other hand, to avoid long computation time, this subset should be as small as possible. We develop an OO-based screening procedure to deal with this tradeoff effectively.

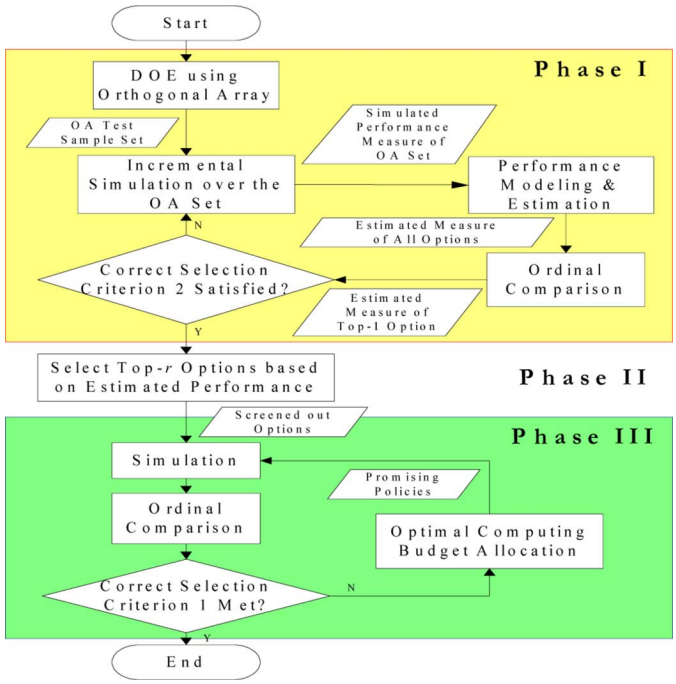


Fig. 3. Flowchart of the OO-based DOE algorithm.

Once a subset of policy options are determined, all the options in this subset are simulated more extensively to identify the best one. An OO-based technique, called *optimal computing budget allocation* (OCBA) [25], [29], is applied for this purpose. OCBA intelligently allocates simulation budget as simulation proceeds and can find the best policy among a given set of options with a minimum computation time.

By integrating the aforementioned design ideas, a fast simulation methodology is developed, which will be referred to as the OO-based DOE simulation methodology hereafter. This new method consists of three innovative phases as shown in Fig. 3. Phase I makes use of DOE technique to establish an estimation model by simulating only a very small fraction of policy options. By utilizing the exponential convergence of OO, simulation can be stopped much earlier by focusing on good policy area. Phase II screens some top-ranking policies for further evaluation under an OO criterion. Phase III applies OO and OCBA techniques to top-ranking policies screened in phase II and identifies the simulated best policy using a minimum computation time while a prespecified confidence level is satisfied. Further details for each component are given in Sections IV-A–C.

A. Phase I: OO-Based DOE Estimation Model

The DOE method is exploited to find an OA and estimate performance measures of all policies with only policies in the OA simulated. Once the test options in an OA are evaluated by simulation, the next step is to construct an estimation model of the performance measures of all options. Denote

$h_{ij} \equiv h(\Theta_i; w_j)$	the jth sample of option i;
$\overline{h}_i \equiv \overline{h}(\Theta_i; w)$	the sample mean of option i, $\overline{h}_i = (1/n)\sum_{j=1}^{n} h_{ij}$;
$\hat{h}_i \equiv \hat{h}(\Theta_i; w)$	estimated performance of option i;

$S(n_S)$	the orthogonal set of test options;
θ_{ij}	the design parameter of factor j in option i.

Unlike traditional DOE, to integrate the notion of OO, we have to construct models for estimation of both means and variances for all policy options.

1) Estimation of Mean Value: A regression model is commonly used to estimate the response surface for a factorial design. Since factors of scheduling rule selection problems have discrete choices without quantitative relationship, general regression models cannot be directly applied, i.e., the relations among levels of a factor and the performance function of interest cannot be modeled as a polynomial function. The mathematical model of the performance measure is thus represented as

$$h(\Theta_i;w) = h_0 + \sum_{p=1}^{n_F} h_p(\theta_{ip};w) + \sum_{p=1}^{n_F}\sum_{q=p+1}^{n_F} h_{pq}(\theta_{ip}\theta_{iq};w) + \sum\sum_{p\neq q\neq r}\sum h_{pqr}(\theta_{ip}\theta_{iq}\theta_{ir};w) + \cdots + \varepsilon \quad (4)$$

where $h_p(\theta_{ip};w)$ is the performance (main) effect caused by factor p, $h_{pq}(\theta_{ip}\theta_{iq};w)$ is the two-factor interaction effect caused by factors p and q and ε is the error term.

There is a tradeoff between modeling higher order interaction and computational efficiency. To achieve the best efficiency without compromising too much on modeling accuracy, we assume that third-factor and higher interaction effects are not significant and can be neglected. Thus, only the first three terms in (4) must be kept. To remedy the loss of modeling accuracy, instead of selecting the best policy using the estimation model, we shall extensively simulate selected policies in the last phase and present in Section IV-C.

The best approach of estimating the effects is the method of least squares [30], which minimizes

$$\sum_{k\in S(n_s)} [h(\Theta_k;w) - \hat{h}(\Theta_k;w)]^2. \quad (5)$$

The grand mean of the performance is thus estimated by

$$\hat{E}[h_0] = \frac{1}{n_S}\sum_{k\in S(n_s)} \bar{h}_k. \quad (6)$$

Main effects and two-factor interactions of $\hat{h}(\Theta_i;w)$ can be estimated by

$$\hat{E}[h_p(\theta_{ip}=\alpha;w)] = \underset{k}{\text{Average}}(\bar{h}(\Theta_k;w|\theta_{kp}=\alpha)) - \hat{E}[h_0] \quad (7)$$

where k is the index of simulated options with the design parameter of factor p equal to α, and

$$\hat{E}[h_{pq}(\theta_{ip}=\alpha,\theta_{iq}=\beta;w)] = \underset{k'}{\text{Average}}(\bar{h}(\Theta_{k'};w\,|\theta_{k'p}=\alpha,\theta_{k'\theta}=\beta)) - \hat{E}[h_p(\theta_{ip}=\alpha;w)] - \hat{E}[h_q(\theta_{iq}=\beta;w)] - \hat{E}[h_0] \quad (8)$$

where k' is the index of simulated options with the design parameter of factors p and q equal to α and β, respectively.

2) Estimation of Variance: It is quite difficult to accurately estimate the variance of an option without making any simulation. Without attempting to find accurate estimates of variances, a conservative estimation approach is taken, where the variance of an unsimulated policy is estimated by the maximum residual variation of simulated policies. The residual variation of a performance model, defined as portion of the total variation in the experiment data that is not explained by the model, is composed of two parts: that due to pure error among test samples (replications) and that due to model lack of fit [23]. The residual sum of squares (SSs) for policy k is, therefore, composed of two parts, the pure error SSs due to replication and the remainder of the variation in the data that is neither accounted for in the model nor a result of the replication in the data

$$\begin{aligned}\text{SS}_k(\text{residual}) &= \sum_{j=1}^{n}(h_{kj}-\hat{h}_k)^2 \\ &= (n-1)S_k^2 + n(\bar{h}_k-\hat{h}_k)^2 \\ &= \text{SS}_k(\text{pure error}) + \text{SS}_k(\text{lack of fit}) \\ &\text{for } k\in S(n_s)\end{aligned} \quad (9)$$

where $S_k^2 = \left(\left(\sum_{j=1}^{n}(h_{kj}-\bar{h}_k)^2\right)/(n-1)\right)$ is the sample variance of policy k.

The variance of $\hat{h}_k$ is estimated by

$$\hat{S}_k^2 \equiv \frac{\text{SS}_k(\text{residual})}{n-1} = S_k^2 + \frac{n(\bar{h}_k-\hat{h}_k)^2}{n-1} \quad (10)$$

which can be easily calculated by the sample mean and sample variance of the simulated policy k. The variance of an unsimulated policy i corresponding to the estimated performance measure $\hat{h}_i$ is estimated by

$$\hat{S}_i^2 \approx \max_k \hat{S}_k^2, \text{ for } i\notin S(n_s). \quad (11)$$

The estimation of the distribution of $\tilde{J}_i$ then becomes

$$\tilde{J}_i \sim N\left(\hat{h}\left(\Theta_i;w\right), \hat{S}_i^2\right), \text{ for } i\notin S(n_s). \quad (12)$$

While this conservative estimation of variance may lead to more computation efforts to obtain a good estimate, it is one simple heuristic trying not to overestimate.

3) Integration With Ordinal Optimization (OO): Traditional DOE simulation does not stop until the simulation accuracy for all simulated policies is sufficiently high. As discussed in Section III-A, the overall simulation efficiency can be dramatically improved by focusing on ordinal comparison due to exponential convergence of OO. To work within DOE setting, we consider a new format of correct selection defined as CS_2.

Definition 2: Define *correct selection-2* (CS_2) as the event that the true performance of the observed rank-r option is not worse than β fraction of the performance of the true best option. Define $P\{\text{CS}_2(r,\beta)\} \equiv \Pr\{\text{The true performance of the observed rank-}r \text{ option is not worse than } \beta \text{ fraction of the performance of the true best option.}\}$

β can be considered as an optimality expectation. Higher value of β demands higher simulation accuracy, resulting in higher confidence on the optimality, i.e., the performance measure of the selected option is closer to that of the true best option.

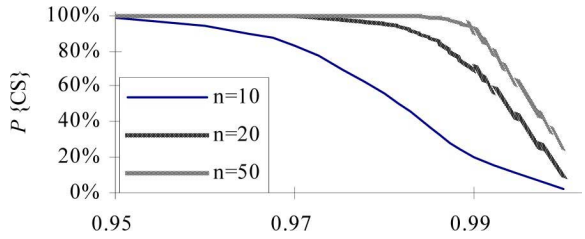


Fig. 4. $P\{\mathrm{CS}_2\}$ versus fraction parameter β.

On the other hand, lower β implies that we are willing to compromise the optimality, which may lead to substantial reduction of computation time, as shown later in this section. In this paper, we apply $P\{\mathrm{CS}_2(r,\beta)\}$ as the stopping criterion for DOE simulation and set $r=1$. When $P\{\mathrm{CS}_2(1,\beta)\}$ is sufficiently high, the best option or a very good option can be identified with high probability, implying that the simulation accuracy near good options are high and that the constructed performance estimation model near good options is reliable.

Similar to the estimation of $P\{\mathrm{CS}_1\}$, $P\{\mathrm{CS}_2(r,\beta)\}$ is estimated by $\mathrm{APCS}_2(r,\beta)$ given in the following theorem.

Theorem 2: Let $\tilde{J}_i$, $i \in \{1,2,\ldots,R\}$, denote the random variable whose probability distribution is the posterior distribution of the expected performance for option i under a Bayesian model. For a maximization problem

$$P\{\mathrm{CS}_2(r,\beta)\} \geq \prod_{i=1,i\neq r}^{R} P\{\tilde{J}_r > \beta\tilde{J}_i\} \equiv \mathrm{APCS}_2(r,\beta). \tag{13}$$

Proof: Please refer to Appendix B.

The posterior distribution for option i, $\tilde{J}_i$, is estimated by (12). $\mathrm{APCS}_2(r,\beta)$ can then be calculated by

$$\mathrm{APCS}_2(r,\beta) = \prod_{i=1,i\neq r}^{R} \Phi\left(\frac{\hat{h}(\Theta_r;w) - \beta\cdot\hat{h}(\Theta_i;w)}{\sqrt{\hat{S}_r^2 + \beta^2\hat{S}_i^2}}\right) \tag{14}$$

where Φ is the standard normal cumulative distribution. $P\{\mathrm{CS}_2(1,\beta)\}$ approximated by $\mathrm{APCS}_2(1,\beta)$ is used as the stopping criterion for DOE simulation.

To give some insight about setting the value of β, we consider a scheduling rule selection problem with 1024 policies. Fig. 4 shows $P\{\mathrm{CS}_2(1,\beta)\}$ as we vary β from 0.95 to 1.0. Three cases are considered, where the numbers of simulation replications (n) are 10, 20, and 50, respectively. In all three cases, $P\{\mathrm{CS}_2(1,\beta)\}$ is very close to 1 when β is lower than 0.95, even in the case that only ten simulation replications are performed. When β is raised to 0.97, $P\{\mathrm{CS}_2(1,\beta)\}$ still stays close to 1 provided that the number of simulation replications $n \geq 20$. This reveals that a bit sacrifice in optimality may lead to substantial reduction in computation time. In our numerical experiments, β is set to be 0.97.

B. Phase II: OO-Based Screening Procedure

Since the estimation model constructed using our very fast DOE simulation is an approximation, the best policy selected using this model is not necessarily the true best policy. To reduce the probability of selecting a not-so-good policy, we screen out a subset of policies for further evaluation in order to determine the final solution. We want to select the subset to ensure that a best policy or at least a good policy is included with high probability. Since the performance estimation model offers a good approximation around good options, it is a good idea to choose this subset as the set containing the observed top-r options from the estimated model. Then, a remaining question is how big this subset should be.

A bigger subset (bigger r) will increase the probability of containing a good policy. On the other hand, we do not want r too big in order to avoid high-computation cost. In this paper, we determine r by utilizing the ordinal comparison probability $P\{\mathrm{CS}_2(r,\beta)\}$ given in previous sections. First, we set the values for $P\{\mathrm{CS}_2(r^*,\beta)\}$ requirement, P_r^*, and the optimality expectation, β. Then, the subset is the set containing the observed top-r^* options from the performance estimation model, where r^* is the minimum r such that $P\{\mathrm{CS}_2(r,\beta)\} > P_r^*$. This setting intends to ensure the probability that the true performance of the observed rank-r^* option is not worse than β fraction of the performance of the true best option is no less than P_r^*. In the procedure, we typically set β here the same as or lower than that in Section IV-A. Choosing a smaller β or P_r^* implies that we are willing to include more options in the screened subset.

C. Phase III: Optimal Computing Budget Allocation (OCBA) for Final Selection

Most policy options screened out of Phase II are good options. In Phase III, we extensively simulate all the screened options to find the best among them. With the advantage of such an exponential convergence, instead of equally simulating all screened policy options, OCBA further improve the performance of OO by intelligently determining the effective allocation of samples/replications for each option within a given simulation budget. Intuitively, to ensure a high $P\{\mathrm{CS}_1\}$, a larger portion of the computing budget should be allocated to those options that are critical in the process of identifying the best option. Chen *et al.* [25] show the application of OCBA can attain speedup factors of an additional order of magnitude above and beyond the exponential convergence of OO.

Definition 3: Define $\mathrm{EPCS}(N_1,N_2,\ldots,N_{s-1},N_s+\tau,N_{s+1},\ldots,N_R)$ as an estimated $P\{\mathrm{CS}_1\}$ if additional τ simulations are performed on rule option s. EPCS is computed using the statistical information after $N_1,N_2,\ldots,N_R$ replications are completed for rule options $1,\ldots,R$, respectively.

If $s \neq b$

$$\mathrm{EPCS}(N_1,N_2,\ldots,N_{s-1},N_{s+\tau},N_{s+1},\ldots,N_R) = P\{\tilde{J}_b > \hat{J}_s\}\cdot\prod_{i=1,i\neq b,i\neq s}^{R} P\{\tilde{J}_b > \tilde{J}_i\} \tag{15}$$

where $\tilde{J}_i \sim N\left((1/N_i)\sum_{j=1}^{N_i} h_i(w_j), \left(\sigma_i^2/N_i\right)\right)$ and

$$\hat{J}_s \sim N\left(\frac{1}{N_s}\sum_{j=1}^{N_s} h_s(w_j), \frac{\sigma_s^2}{N_s+\tau}\right).$$

Else, if $s = b$, $\tilde{J}_b$ is replaced by

$$\hat{J}_b \sim N\left(\frac{1}{N_b}\sum_{j=1}^{N_b} h_b(w_j), \frac{\sigma_b^2}{N_b+\tau}\right). \tag{16}$$

Note that the expression of EPCS is similar to that of APCS_1 in (3) and can be easily calculated. EPCS provides sensitivity information about how APCS_1 will change if additional τ simulations are performed on rule s. The promising index (PI) is defined as follows:

$$\begin{aligned} PI(s) \equiv \mathrm{EPCS}(&N_1, N_2, \ldots, N_{s-1}, \\ &N_{s+\tau}, N_{s+1}, \ldots, N_R) \\ - \mathrm{APCS}_1(&N_1, N_2, \ldots, N_{s-1}, \\ &N_s, N_{s+1}, \ldots, N_R). \end{aligned} \tag{17}$$

The "promising" rules are those rules which can maximize the improvement of APCS_1 if they are further simulated. OCBA technique intelligently allocates incremental simulation replications to "more promising" scheduling policies. As the simulation continues, the ranking estimates improve and promising policies are redetermined. This procedure continues until a prespecified confidence level is obtained. In summary, we develop an iterative experimentation procedure which is summarized as follows:

Step 0. Perform n_0 simulations for all rule options, $l \leftarrow 0$, $N_1^l = N_2^l = \cdots = N_R^l = n_0$.

Step 1. If $\mathrm{APCS}_1\left(N_1^l = N_2^l = \cdots = N_R^l\right) \geq$ "A prespecified confidence level", stop.

Step 2. Calculate $PI(s)$ for all rules $s = 1, 2, \ldots, R$.

Step 3. Find the set $S(m) \equiv \{s : PI(s)$ is among the highest $m\}$

Step 4. Perform additional τ simulations for rule $i, i \in S(m)$. Set $N_i^{l+1} \leftarrow N_i^l + \tau$, for $i \in S(m)$, and $N_i^{l+1} \leftarrow N_i^l$, for $i \notin S(m)$, $l \leftarrow l + 1$, go to Step 1.

Please refer to [31] for details about the selection of the required parameters in the procedure.

V. Applications to Scheduling Policy Composition

To assess the application potential of the new methodology, we apply our OO-based DOE simulation to the fab scheduling policy composition problem defined in Section II. The fab model of [5] serves as a baseline model for simulation. Our simulation study investigates fab scheduling policy composition to achieve good performance in mean and VCTs and production smoothness of a fabrication line (SM) under different factors of fab initial state and scheduling horizon.

Similar to the setting of [5], our fab model has a release rate of 0.52 lot/h, under which utilizations of the machine groups are mostly greater than 90%, namely, a highly loaded situation. Bottleneck machine group is Station 1, of which the utilization rate is 94.23%. Detailed machine information is listed in Table X of Appendix A, including the machine type, the number of homogeneous machines in a group, number of visits to a machine group by a lot, number of lots per batch processing, time per processing, mean times to failure and repair, and utilization. Machine groups of photolithography, implanter, and furnace are specifically identified and the remaining ones are all labeled as general machines.

In the following experiment study of OO-based DOE simulation, the number of simulation runs for test samples in an OA is set to 10, and the number of simulation increment is set to 1. The fraction parameter β is set to 0.97, and the confidence probabilities are all set to 0.9. The number of top top-ranking policies screened in phase II is set to be $\min(16, r^*)$, where r^* is defined in Section IV-B. Simulations are conducted on an 800 MHz AMD Duron™ personal computer.

A. Composition for Long-Term Performance

This set of experiments is designed to study the long-term MCT, VCT, and SM performance of individual scheduling policies. A simulation run of 8.2 years starting with an empty fab is conducted for each scheduling policy. The first half-year of simulated time serves as a warm-up period. The rest of the simulated time is divided into 100 intervals. Each interval represents four weeks. At each interval, we sample MCT, VCT, and SM for each scheduling policy. Top-10 scheduling policies of individual performance indices are listed in Tables I–III. Note that the variances of cycle time are so high (as compared with the differences of the MCTs among different scheduling policies) that multiple simulation replications is needed as motivated in Section II. The total simulation time of this set of experiments takes about 28 CPU hours. Analyses are then performed to investigate the effects of individual factors.

1) Top-Ranking Policies: Scheduling policies combining DET release policy, FSMCT dispatching rule for general machine groups and LTWL loading policy for photolithography machine groups outperforms other policies in MCT performance. Nine out of the top-10 policies have DET-FSMCT-LTWL combinations. FSMCT is the best dispatching rule for VCT reduction. LLS dispatching rule for furnace machine groups also has good performance in VCT since LLS, as well as FSMCT, is slack time-based dispatching rule. These results are consistent with the findings of [5]. FSMCT obtains as good performance in index SM as LDF although it is not specially designed for improving SM. Among the 1024 combinations, DET-FSMCT-LTWL-LLS-FSMCT, which ranks 2, 5, and 2 in MCT, VCT, and SM, respectively, is good across the three indices consider.

2) Analysis of Single-Factor Effects: In addition to investigation on top-ranking scheduling policies, an analysis is performed to investigate the effects of individual design factors. Fig. 5 shows the main effects of the five factors that affect the three performance indices. Interested readers may refer to Table XVI for detailed data. The effect of DET release policy on MCT performance is 7.19, which is calculated by using (7) in Section IV-A. A positive value in Fig. 5 means that the rule is worse than the average performance of all rules and the magnitude represents the absolute difference. FSMCT dispatching rule along with LTWL lot assignment policy are

TABLE I
LONG-TERM MCT PERFORMANCE

WRP	LDR	LAP PH	LDR F	LDR UF	MCT	VCT	SM
DET	FSMCT	LTWL	LNGQ	FIFO	441.18	140.10	0.851
DET	FSMCT	LTWL	LLS	FSMCT	445.10	129.63	0.856
DET	FSMCT	LTWL	LLS	SAF1	446.01	137.74	0.843
DET	FSMCT	LTWL	LAS	SAF2	447.60	137.72	0.833
DET	FSMCT	LTWL	LAS	SAF1	450.56	111.16	0.838
DET	FSMCT	LTWL	LNGQ	SAF1	451.99	150.57	0.833
DET	FSMCT	LTWP	LNGQ	SAF1	455.14	146.56	0.835
DET	FSMCT	LTWL	LLS	FIFO	455.56	121.43	0.844
DET	FSMCT	LTWL	LLS	SAF2	456.48	141.66	0.835
DET	LDF	LTWP	LLS	FSMCT	456.52	211.10	0.857

TABLE II
LONG-TERM VCT PERFORMANCE

WRP	LDR	LAP PH	LDR F	LDR UF	MCT	VCT	SM
DET	FSMCT	LTWL	LAS	SAF1	450.56	111.16	0.838
DET	FSMCT	LTWL	LLS	FIFO	455.56	121.43	0.844
DET-B	FSMCT	LTWP	LLS	FSMCT	471.33	123.77	0.839
DET	FSMCT	LGWP	LLS	FIFO	497.08	128.84	0.818
DET	FSMCT	LTWL	LLS	FSMCT	445.10	129.63	0.856
DET-B	FSMCT	LTWP	LLS	FIFO	457.28	129.95	0.839
DET-B	FSMCT	LTWP	LLS	SAF2	480.87	130.66	0.819
DET-B	FSMCT	LTWL	LLS	SAF1	460.25	131.24	0.833
DET	FSMCT	LTWP	LAS	SAF1	469.61	131.48	0.830
DET	FSMCT	LTWP	LLS	SAF1	467.79	131.50	0.829

TABLE III
LONG-TERM SM PERFORMANCE

WRP	LDR	LAP PH	LDR F	LDR UF	MCT	VCT	SM
DET	LDF	LTWP	LLS	FSMCT	456.52	211.10	0.857
DET	FSMCT	LTWL	LLS	FSMCT	445.10	129.63	0.856
WR-BH	LDF	LTWP	LNGQ	FIFO	467.88	305.83	0.855
WR	LDF	LTWL	LNGQ	FSMCT	468.12	295.71	0.854
DET-BH	LDF	LTWL	LAS	FSMCT	460.09	218.60	0.853
WR-BH	LDF	LTWL	LAS	FSMCT	461.80	275.32	0.852
WR-BH	LDF	LTWP	LNGQ	SAF1	469.14	479.29	0.852
DET-BH	LDF	LTWL	LLS	FIFO	467.67	211.03	0.852
WR-BH	FSMCT	LTWL	LAS	SAF1	464.60	184.67	0.852
WR-BH	FSMCT	LTWL	LLS	FSMCT	465.43	184.91	0.852

superior to other rules in MCT reduction. The VCT and SM performances are significantly influenced by dispatching rules for general machine groups. These results are consistent with aforementioned results on top-ranking policy composition.

However, DET release policy performs worst among the four release policies, which seems contradict with the results in the previous subsection. To explain this phenomenon, we calculate standard deviations of MCTs of scheduling policies with DET and WR as the release rules, respectively. The standard deviation of DET is 23.43, which is higher than that of WR, 15.09. This means that the MCT performance of DET varies significantly with the dispatching rules used while that of WR does not; namely, the MCT performance of a policy with WR

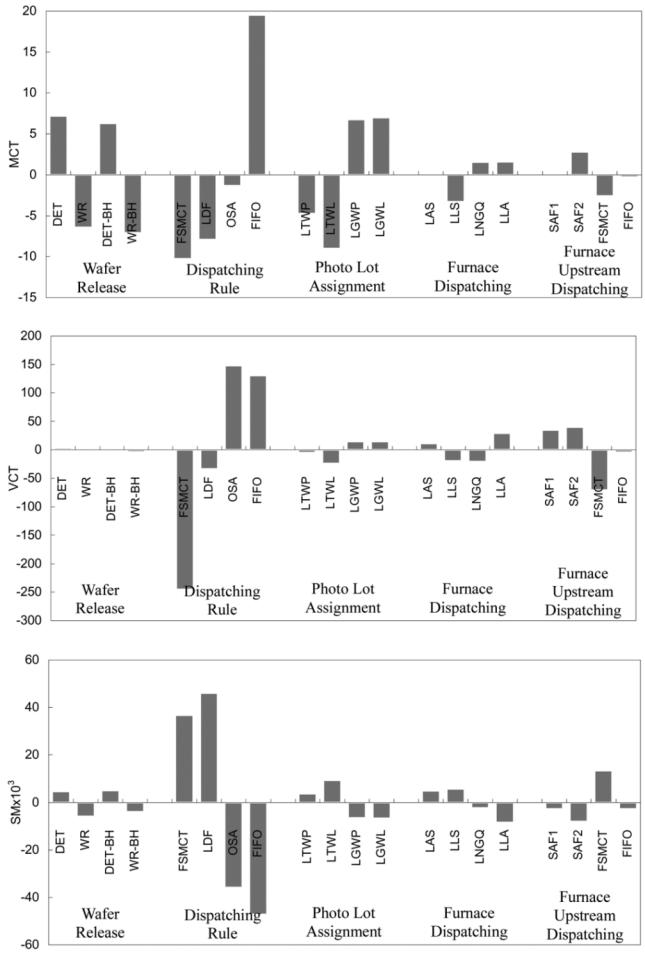


Fig. 5. Long-term performance analysis: main effects.

TABLE IV
EXPERIMENT SET OF SHORT-TERM POLICY SELECTION

Exp. ID	Time Horizon	Performance Index	Initial State	Exp. ID	Time Horizon	Performance Index	Initial State
1	1 week	VCT	IS_1	5	4 weeks	MCT	IS_1
2	1 week	VCT	IS_2	6	4 weeks	MCT	IS_2
3	1 week	SM	IS_1	7	4 weeks	SM	IS_1
4	1 week	SM	IS_2	8	4 weeks	SM	IS_2

release rule is insensitive to dispatching rules. Consequently, many top-ranking policies have DET as the release rule, while scheduling policies with WR as the release rule have a good across-policy average performance in MCT.

B. Short-Term Policy Composition

In this set of experiments, the OO-based DOE simulation is exploited to facilitate short-term scheduling policy selection. Table IV lists the set of short-term policy selection experiments under different operation objectives, initial fab states and planning horizons. In the current fab practice of short-term operation scheduling, performance is measured and reviewed daily, weekly, and monthly. The two commonly used time horizons, one week and four weeks are, therefore, considered in this study. MCT performance considers only a four-week

horizon since one week is relatively short with respect to cycle time; the effect of changing scheduling policies needs more than one week of delay to appear on MCT. However, VCT measurements can easily reveal fabrication variability within a shorter time horizon. Smoothness (SM) follows industrial practice that it is reviewed weekly and monthly. The two initial states, IS_1 and IS_2, are generated by simulating the fab model for one year starting from an empty fab with scheduling policies DET-FSMCT-LTWL-LLS-FSMCT and WR-LDF-LGWP-LLA-SAF1, respectively. The findings are as follows.

1) FSMCT Dominates MCT and VCT Performances: By analyzing the MCT performance of scheduling policies over a four-week horizon and the VCT performance over a one-week horizon under the two initial states, it is observed that dispatching rules for general machine groups of the top-ranking policies are mostly FSMCT for MCT reduction and are all FSMCT for VCT reduction. Such superiority of FSMCT to other dispatching rules in MCT and VCT reductions are consistent with the findings of [5]. Details are listed in Tables XVII and XVIII of Appendix C.

2) Policies for SM Vary With Time Horizons and Initial States: Table V lists the SM performance of scheduling policies over one week and four weeks, where the top-ranking policies for maximizing SM vary with time horizons and initial states. Under the initial state IS_2, WR or WR-BH release rules combining with FIFO or LDF dispatching rules and LLS dispatching rule for furnace have good SM performance over a one-week horizon. However, DET-FSMCT combinations are superior to other rule combinations in SM over a four-week horizon under the initial state IS_2. Under the initial state IS_1, LDF for general machine group is preferred over a four-week horizon, whereas FSMCT is preferred under the initial state IS_2.

TABLE V
SCHEDULING POLICY RANKING FOR SM

(a) Over 1 Week under IS_1

Rank	WRP	LDR	LAP_PH	LDR_F	LDR_UF	SM
1	DET	LDF	LTWL	LNGQ	SAF1	0.933
2	DET	LDF	LGWP	LLS	SAF1	0.906
3	DET-BH	LDF	LTWL	LLS	SAF2	0.878
4	DET-BH	LDF	LTWL	LLS	SAF1	0.861
5	DET	LDF	LTWL	LLS	SAF1	0.858
6	DET-BH	LDF	LGWP	LLS	SAF1	0.839
7	DET-BH	LDF	LTWL	LNGQ	SAF1	0.820
8	DET	LDF	LTWL	LLS	SAF2	0.818
9	DET-BH	OSA	LGWP	LLS	SAF1	0.807
10	DET-BH	FIFO	LTWL	LLS	SAF2	0.802

(b) Over 4 Weeks under IS_1

Rank	WRP	LDR	LAP_PH	LDR_F	LDR_UF	SM
1	DET	LDF	LTWL	LLS	SAF1	0.852
2	DET	LDF	LTWL	LLA	SAF1	0.850
3	DET	LDF	LGWL	LLS	SAF1	0.847
4	DET	LDF	LTWL	LNGQ	SAF1	0.842
5	WR	LDF	LTWL	LLS	SAF2	0.841
6	DET	LDF	LTWL	LLS	SAF2	0.840
7	WR	LDF	LTWL	LLS	SAF1	0.839
8	WR	LDF	LTWL	LNGQ	SAF1	0.836
9	DET	LDF	LGWP	LLS	SAF1	0.832
10	DET-BH	LDF	LTWL	LAS	SAF1	0.831

(c) Over 1 Week under IS_2

Rank	WRP	LDR	LAP_PH	LDR_F	LDR_UF	SM
1	WR-BH	FIFO	LTWL	LLS	SAF2	0.574
2	WR	FIFO	LTWP	LLS	FIFO	0.571
3	WR-BH	FIFO	LTWP	LLS	SAF2	0.570
4	WR	FIFO	LTWL	LLS	SAF2	0.563
5	WR	FIFO	LTWP	LLS	SAF2	0.560
6	WR-BH	FIFO	LTWP	LLS	FIFO	0.558
7	DET	LDF	LTWL	LLS	SAF1	0.514
8	WR-BH	LDF	LTWL	LNGQ	SAF1	0.484
9	WR	LDF	LTWP	LLS	SAF1	0.476
10	WR	LDF	LTWL	LNGQ	SAF1	0.468

(d) Over 4 Weeks under IS_2

Rank	WRP	LDR	LAP_PH	LDR_F	LDR_UF	SM
1	DET	FSMCT	LTWP	LLS	SAF1	0.753
2	DET	FSMCT	LTWP	LLS	FSMCT	0.740
3	DET	FSMCT	LTWL	LLS	FSMCT	0.733
4	DET	FSMCT	LTWP	LAS	FSMCT	0.728
5	DET	FSMCT	LTWL	LNGQ	SAF1	0.725
6	DET	FSMCT	LTWP	LNGQ	FSMCT	0.723
7	DET	FSMCT	LTWP	LLA	SAF1	0.722
8	DET	FSMCT	LTWL	LNGQ	FSMCT	0.720
9	DET	FSMCT	LTWP	LLA	FSMCT	0.712
10	DET	FSMCT	LTWL	LLA	SAF1	0.711

VI. EFFECTIVENESS OF THE OO-BASED DOE SIMULATION

To further investigate the effectiveness of the OO-based DOE simulation in fab scheduling policy composition, simulation experiments are designed and conducted for analyses by the effects of DOE approximation, computation efficiency, and the quality of the policies obtained. In specific, effects of two-factor interactions on individual performance indices are first analyzed to support our approximation in the DOE estimation model (4). Estimation models with and without two-factor interactions are then compared over their applications to good policy selection. The gain in computation by using the OO-based DOE approach is finally assessed.

A. Analysis of Two-Factor Interaction Effects

Table VI lists the two-factor interaction effects among the five factors of wafer release and dispatching rules for MCT. The first two columns represent the levels of the two factors to be analyzed, respectively. Observe that the interaction effect of FSMCT dispatching rule (first level of F2) and LTWL lot assignment policy (second level of F3) on MCT performance is -7.83 as listed in the second row of the table, which is calculated by using (8). Tests of significance of the interaction effects are performed by using the analysis of variance (ANOVA) method. Two-factor interactions WRP-LDR, WRP- LAP_PH and LDR-LAP_PH are significant for MCT with 97.5% confidence. This explains why nine out of the top-10 scheduling

TABLE VI
TWO-FACTOR INTERACTION EFFECTS FOR MCT

L_A[1]	L_B	Two-factor Interaction Effects[2] (MCT)									
		F1-F2	F1-F3	F1-F4	F1-F5	F2-F3	F2-F4	F2-F5	F3-F4	F3-F5	F4-F5
1	1	-5.72	-2.77	-0.36	1.26	-3.95	-0.18	-1.33	0.77	0.96	2.89
1	2	8.81	-6.67	-1.34	-0.43	-7.83	1.38	-0.44	-0.40	0.63	-1.68
1	3	-4.13	4.63	-0.49	-1.23	6.53	-0.80	1.72	-0.89	-1.41	-1.30
1	4	1.04	4.81	2.19	0.41	5.25	-0.41	0.05	0.52	-0.19	0.09
2	1	2.65	1.61	0.52	-1.27	-3.58	0.43	-1.16	-0.39	2.09	-0.79
2	2	-9.50	7.11	0.88	0.63	-1.60	0.78	0.16	0.08	-0.73	1.26
2	3	7.83	-4.13	0.73	1.45	1.92	0.28	1.64	-0.64	-0.76	0.64
2	4	-0.98	-4.59	-2.13	-0.80	3.26	-1.48	-0.64	0.95	-0.59	-1.11
3	1	-0.08	-1.07	-0.38	0.62	4.27	0.71	0.88	-0.62	-1.49	-0.77
3	2	11.05	-7.00	-0.12	0.32	7.14	-0.36	-1.28	0.76	0.17	0.02
3	3	-11.6	3.88	-1.13	-1.71	-5.62	0.36	-0.61	1.22	0.39	0.44
3	4	0.69	4.20	1.63	0.78	-5.79	-0.71	1.01	-1.36	0.93	0.31
4	1	3.16	2.24	0.22	-0.60	3.26	-0.96	1.61	0.24	-1.56	-1.33
4	2	-10.3	6.56	0.57	-0.51	2.29	-1.80	1.56	-0.44	-0.07	0.40
4	3	7.95	-4.37	0.89	1.50	-2.83	0.16	-2.75	0.31	1.77	0.22
4	4	-0.75	-4.42	-1.69	-0.39	-2.72	2.60	-0.41	-0.11	-0.14	0.72

[1] L_A: the level of the first factor; L_B: the level of the second factor
[2] F1: WRP; F2: LDR; F3: LAP_PH; F4: LDR_F; F5: LDR_UF

policies of MCT have the same combinations of WRP, LDR, and LDR-LAP. MCT is dominated by two-factor interactions between two of WRP, LDR, and LDR-LAP.

To keep the presentation of the main text concise, simulation results of the two-factor interaction effects for VCT and SM are given in Tables XIX and XX of Appendix C. Two factor interactions of LDR-LAP_PH and LDR-LDR_UF have significant effects on VCT. Both the MCT and VCT performance measures are, therefore, modeled by main effects and two-factor interactions. None of the two-factor interactions are significant to SM. The SM performance measures can, therefore, be accurately modeled by main effects without considering two-factor or higher order interactions.

B. Evaluation of DOE Estimation Models

Now, we evaluate the value added to good policy selection by capturing two-factor interactions in the performance modeling and estimation method described in Section IV-A. The estimation of long-term MCT, VCT, and SM performance of individual scheduling policies serves as the study problem.

1) Single-Factor Only Model: First, consider an estimation model with only the first two terms of (4). Instead of performing simulations over all the 1024 scheduling policies, an orthogonal set of 16 test samples is determined by the orthogonal array, OA(16,5,4,2), which has five factors and four levels of each factor and is the smallest OA that meets the experiment requirements. Main (single-factor) effect estimation models of individual factors for all the 1024 policies are built based on the simulated performance measures of the test samples. Interested readers may find the details in Table XXI of Appendix C.

Table VII lists the top-16 scheduling policies selected from the single-factor estimation model of MCT performance. Differences between the estimated and the simulated performance

TABLE VII
TOP-16 POLICIES FOR LONG-TERM MCT BY ESTIMATION OF SINGLE-FACTOR EFFECTS ONLY

WRP	LDR	LAP_PH	LDR_F	LDR_UF	Est. MCT(E)	Sim. MCT(s)	Rank	(E)-(S)
WR	FSMCT	LTWL	LAS	FSMCT	443.74	468.16	122	-24.42
WR	LDF	LTWL	LAS	FSMCT	444.94	470.15	158	-25.21
WR	FSMCT	LTWP	LAS	FSMCT	446.31	467.88	114	-21.57
WR-BH	FSMCT	LTWL	LAS	FSMCT	447.04	467.56	106	-20.52
WR	LDF	LTWP	LAS	FSMCT	447.51	459.83	18	-12.32
WR-BH	LDF	LTWL	LAS	FSMCT	448.24	461.80	29	-13.56
WR-BH	FSMCT	LTWP	LAS	FSMCT	449.61	470.79	173	-21.18
WR	FSMCT	LTWL	LLS	FSMCT	450.08	468.96	136	-18.88
WR-BH	LDF	LTWP	LAS	FSMCT	450.81	461.36	27	-10.55
WR	FSMCT	LGWP	LAS	FSMCT	450.83	489.71	445	-38.88
WR	LDF	LTWL	LLS	FSMCT	451.28	463.35	43	-12.07
WR	LDF	LGWP	LAS	FSMCT	452.03	465.91	77	-13.88
DET	FSMCT	LTWL	LAS	FSMCT	452.11	459.47	15	-7.36
WR	FSMCT	LTWP	LLS	FSMCT	452.65	466.17	80	-13.52
DET	LDF	LTWL	LAS	FSMCT	453.31	477.00	261	-23.69
WR-BH	FSMCT	LTWL	LLS	FSMCT	453.38	465.43	71	-12.05

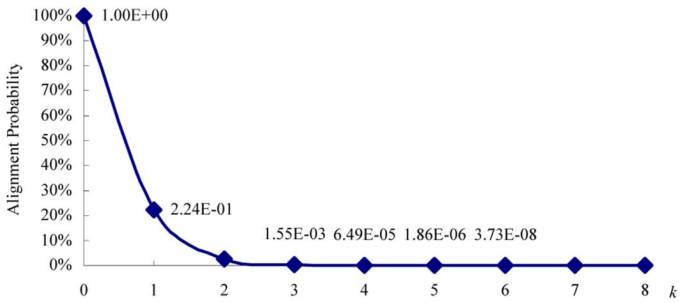


Fig. 6. Alignment probability of blindly picking.

measures range from -38.88 to -7.36 and the relative differences are within -8%. The value estimates on the performance are poor as compared with 0.5% level of the maximum relative standard error, i.e., the ratio of the standard deviation of an estimator to the absolute value of the estimator of the 16 test samples. However, from ordinal perspective, one of the estimated top-16 scheduling policies is indeed within the top-16 based on the simulated performance measures. In contrast to blindly picking, consider the k-alignment probability defined as the probability that at least k of 16 blindly picked scheduling policies belong to the real top-16 and Fig. 6 depicts the k-alignment probability out of 16 [20]. In the case of $k = 1$, the alignment probability is about 0.224. Using the DOE estimation model is much more effective in identifying good policies than blindly picking.

2) Model With Two-Factor Interactions: The DOE method may lead to better performance estimation if two-factor interactions are modeled. A larger orthogonal array, OA(32, 9, 4, 2), is used to determine an orthogonal set of 32 test samples, where the first five columns of OA(32, 9, 4, 2) represent the five factors. Table VIII lists the top-16 policies based on MCT performance estimation modeling both single-factor effects and two-factor interactions. Differences between the estimated and the simulated performance range from -25.47 to 6.19 and absolute values of relative differences are within 5.5%, which are a

TABLE VIII
TOP-16 POLICIES FOR LONG-TERM MCT BY ESTIMATION OF SINGLE-FACTOR EFFECTS AND TWO-FACTOR INTERATIONS

WRP	LDR	LAP _PH	LDR _F	LDR _UF	Est. MCT(E)	Sim. MCT(s)	Rank	(E)-(S)
DET	FSMCT	LTWL	LLA	FSMCT	445.25	470.72	169	-25.47
DET	FSMCT	LTWL	LLA	SAF1	445.54	469.38	144	-23.84
DET	FSMCT	LTWL	LLA	FIFO	445.72	470.80	176	-25.08
DET	FSMCT	LTWL	LNGQ	FSMCT	446.90	457.94	13	-11.04
DET	FSMCT	LTWL	LNGQ	SAF1	447.19	451.99	6	-4.80
DET	FSMCT	LTWL	LNGQ	FIFO	447.37	441.18	1	6.19
DET	FSMCT	LTWL	LLA	SAF2	449.88	464.00	52	-14.12
DET	FSMCT	LTWL	LAS	FSMCT	450.45	459.47	15	-9.02
DET	FSMCT	LTWL	LAS	SAF1	450.74	450.56	5	0.18
DET	FSMCT	LTWL	LAS	FIFO	450.92	470.32	160	-19.40
DET	FSMCT	LTWL	LNGQ	SAF2	451.53	457.82	12	-6.29
WR	LDF	LTWL	LLA	FSMCT	453.77	468.91	135	-15.14
WR	LDF	LTWL	LLA	SAF1	454.06	473.53	227	-19.47
WR-B	LDF	LTWP	LLA	FSMCT	454.14	469.96	156	-15.82
WR	LDF	LTWL	LLA	FIFO	454.24	474.98	243	-20.74
WR-B	LDF	LTWP	LLA	SAF1	454.42	479.05	280	-24.63

little bit better than using estimation modeling only single-factor effects. It is observed that six of the top-16 scheduling policies of the estimated performance measures are within the top-16 of the simulated performance measures. For $k = 6$, the alignment probability of blindly picking is about $3.73*10^{-8}$ (Fig. 6). From the perspective of ordinal comparison, the performance estimation model of scheduling policies considering both single-factor and two-factor interaction effects is much superior to that considering only single-factor effects.

Our simulation study and analyses also show that in terms of VCT performance, when modeling with only single-factor effects, only one of the top-16 policies selected by estimation is within the simulated (true) top-16, while there are four when two-factor effects are modeled. In terms of SM performance, four of the top-16 policies selected by estimation are within the top-16 of the simulated top-16. This relatively high probability can be explained by an earlier observation in Section VI-A that two-factor interaction effects of SM performance are not significant and negligible. Readers may refer to Table XXII for the OO-based DOE rankings of the top-16 policies selected by using the two estimation models for the three performance metrics.

C. Computational Efficiency of the OO-Based DOE Simulation

To examine the efficiency of the OO-based DOE simulation, traditional simulation approach serves as a benchmark. The number of simulation replications is increased for each rule until the value estimation of the mean performance measure is sufficiently stable, i.e., the variation of the estimator is sufficiently small as compared with the absolute value of the estimator. Define ρ as the ratio of the standard deviation of the estimator to the absolute value of the estimator. ρ is also called the relative standard error. An estimation is considered good when ρ is sufficiently small. While a good selection of ρ is problem specific, a number smaller than 0.5% is usually a reasonable selection. Since traditional simulation approach is formidable for a huge amount of options, the number of simulation replications needed by traditional simulation approach is approximated by the number of simulation replications of OA test samples multiplied by the ratio of the total number of combinations, 1024, to the sample size of the OA (32 or 16 in this set of experiments).

TABLE IX
EFFICIENCY OF OO-BASED DOE SIMULATION

Exp. ID	Perform. Index	Simulation		Time Saving Factor
		Traditional (approx.)	OO-based DOE	
1	VCT	7,023,865	972	7,226
2	VCT	5,520,734	783	7,051
3	SM	1,099,417	576	1,909
4	SM	1,994,606	435	4,585
5	MCT	22,524	858	26
6	MCT	20,583	1,053	20
7	SM	381,547	546	699
8	SM	476,888	417	1,144

Computation times of the two simulation approaches are listed in Table IX. Time saving factor is defined as the ratio between simulation replications of traditional approach and that of the OO-based DOE approach. In experiments 1 and 2, the traditional simulation approach requires up to 7000 times more computation time than the OO-based DOE approach. However, in experiments 5 and 6, the time saving factors are about 20. This is because MCT performance measures of scheduling policies converge faster and require fewer simulation replications to obtain a good approximation than VCT and SM. Most of the OO-based DOE simulations require three to four orders of magnitude less time than the traditional approach.

VII. CONCLUSION

Semiconductor wafer fabrication faces stringent challenges of volatile product demands, very short time to market, complex but fast evolving process technology, skyrocketing capital investment and highly cost-sensitive competition. In this paper, we present a fast OO-based DOE simulation technique to efficiently select a good composition of scheduling policy from combinatorial options of wafer release and machine dispatching rules, which plays a key role in improving equipment reliability and utilization and in cycle time reduction, and on-time delivery. We develop the new technique by integrating the notions of DOE and OO. DOE is used to construct a performance estimation model by simulating only a small fraction of options, while the exponential convergence of OO is fully utilized through the entire algorithm to further improve computational efficiency. Although we cannot guarantee an optimal policy will be determined, we are able to identity a good policy using a very minimum computation time. Simulation results of applications to scheduling wafer fabrications not only screen out good scheduling policies but also provide insights about how factors such as wafer release and the dispatching of each machine group may affect production cycle times and smoothness under a reentrant process flow. In specific, DET-FSMCT is a good rule combination in MCT and VCT reductions and in SM performance over a four-week horizon. In terms of SM

over a four-week horizon, WR or WR-BH release rules combined with FIFO or LDF dispatching rules and LLS dispatching rule for furnace is superior. Our factor effect analyses show that MCT and VCT performance measures need to be modeled by main effects and two-factor interactions, while the SM performance measures can be accurately modeled by main effects without considering two-factor or higher order interactions. The new OO-based DOE simulations require 2–3 orders of magnitude less computation time than those of a traditional approach for small to medium-size problems. We anticipate the speedup factor will be much higher when applied to larger problems.

APPENDIX A
WAFER RELEASE AND DISPATCHING RULES

TABLE X
MACHINE GROUP DATA

Station	Mach. Type	No. of Mach.	No. of Visits	MPT[3]	MTBF[4]	MTTR[5]	Batch Size	Util.
1	Photo	4	14	0.500	150	5	1	94.23%
2	General	3	12	0.375	200	9	1	82.31%
3	Implante	10	7	2.500	200	5	1	93.44%
4	General	1	1	1.800	200	1	1	94.10%
5	General	1	2	0.900	200	1	1	94.10%
6	General	2	3	1.200	200	1	1	94.10%
7	General	1	1	1.800	200	1	1	94.10%
8	Furnace	4	8	4.800	150	5	6	86.43%
9	General	1	3	0.580	200	5	1	92.92%
10	General	9	5	3.000	130	5	1	90.37%
11	General	2	3	1.100	200	5	1	88.24%
12	General	2	1	2.500	200	5	1	67.44%

[3]MPT: Mean Processing Time (by hours)
[4]MTBF: Mean Time between Failures (by hours)
[5]MTTR: Mean Time to Repair (by hours)

TABLE XI
WAFER RELEASE RULES (WRP)

SYMBOL	DESCRIPTION
DET	Inter-arrival times of lots are constant.
WR(C)	Workload regulation release for one bottleneck system. When the expected work in fab for bottleneck machine drops to C hours, then release a new lot.
DET-BH(B)	Inter-arrival times of batches of lots are constant, where the batch size is B.
WR-BH(C, B)	Workload regulation release for one bottleneck system. When the expected work in fab for bottleneck machine drops to C hours, then release a batch of lots, where the batch size is B.

TABLE XII
LOT ASSIGNMENT RULES FOR PHOTOLITHOGRAPHY MACHINE GROUP (LAP_PH)

SYMBOL	DESCRIPTION
LTWP	Choose a machine with least dedicated WIPs in the re-entrant line.
LTWL	Choose a machine with least total workload in the re-entrant line.
LGWP	Choose a machine with least dedicated WIPs among photo machines.
LGWL	Choose a machine with least workload among photo machines.

TABLE XIII
DISPATCHING RULES FOR GENERAL MACHINE GROUP (LDR)

SYMBOL	DESCRIPTION
FSMCT	Choose the lot with smallest $(n_p/\lambda_p + C_p - \zeta_i)$, where p represents the index of product type, n_p is the sequence number of the lot under consideration, C_p is the mean cycle time, λ_p is the throughput rate, and ζ_i is the estimate of the remaining cycle time from buffer i.
LDF	Choose a stage with the largest deviation of completed moves from the desired moves and then choose a lot from the stage by FSVCT rule, which chooses a lot with the smallest $(a_n + C_p - \zeta_i)$ and a_n is the release time of lot n.
OSA	Choose a stage according to the following priorities: Priority I: stage i such that $N_i(t) > \bar{N}_i$ and $N_{i+1}(t) < \bar{N}_{i+1}$; Priority II: stage i such that $N_i(t) < \bar{N}_i$ and $N_{i+1}(t) < \bar{N}_{i+1}$; Priority III: stage i such that $N_i(t) > \bar{N}_i$ and $N_{i+1}(t) > \bar{N}_{i+1}$; Priority IV: stage i such that $N_i(t) < \bar{N}_i$ and $N_{i+1}(t) > \bar{N}_{i+1}$, where $N_i(t)$ is the WIP at time t at stage i, $\bar{N}_i$ is the average WIP at stage i. Choose a lot with the same priority using FSVCT.
FIFO	Select the lot which arrived in the queue at the earliest time.

TABLE XIV
DISPATCHING RULES FOR FURNACE GROUP (LDR_F)

SYMBOL	DESCRIPTION
LAS	Choose a stage with least average slack time of lots in queue and then choose a batch of lots from the stage with least slack time.
LLS	Choose a stage that has the lot with least slack time and then choose a batch of lots from the stage with least slack time.
LNGQ	Choose a stage with longest queue and then choose a batch of lots from the stage with least slack time.
LLA	Choose a stage that has the lot, which arrives the machine group at the earliest time, and then choose a batch of lots from the stage with least slack time.

TABLE XV
RULES FOR UPSTREAM OF FURNACE GROUP (LDR_UF)

SYMBOL	DESCRIPTION
SAF1	If the queue size is smaller than a threshold for all downstream furnace stages, select a stage that has the largest sum of lots at current stage and downstream furnace stage, and then choose a lot from the stage using FIFO. If not, use FIFO.
SAF2	If the queue size is smaller than a threshold for all downstream furnace stages, select a stage that has the largest lots at downstream furnace stage, and then choose a lot from the stage using FIFO. If not, use FIFO.
FSMCT	As FSMCT in Table XIII.
FIFO	As FIFO in Table XIII.

APPENDIX B
PROOF OF APCS_2 AS LOWER BOUND OF $P\{\text{CS}_2\}$

Lemma 1: Let $X_1, X_2, \ldots, X_R$ be R random variables and $X_2, \ldots, X_R$ are mutually independent. Then

$$\Pr\left\{X_r > \max_{j\neq r}[X_1, X_2, \ldots, X_j, \ldots, X_R]\right\} \geq \prod_{j=1, j\neq r}^{R} \Pr\{X_r > X_j\}.$$

Proof: Apply Lemma 4 in [27] and exchange X_1 with X_r.

Definition 2: Define *correct selection-2* (CS_2) as the event that the true performance of the observed rank-r option is not worse than β fraction of the performance of the true best option. Define $P\{\text{CS}_2(r,\beta)\} \equiv \Pr\{$The true performance of the observed rank-r option is not worse than β fraction of the performance of the true best option$\}$.

Theorem 2: Let $\tilde{J}_i$, $i \in \{1, 2, \ldots, R\}$, denote the random variable whose probability distribution is the posterior distribution of the expected performance for option i under a Bayesian model. For a maximization problem, $P\{\text{CS}_2(r,\beta)\} \geq \prod_{i=1, i\neq r}^{R} P\{\tilde{J}_r > \beta\tilde{J}_i\} \equiv \text{APCS}_2(r,\beta)$.

Proof:

$$\begin{aligned} &P\{\text{CS}_2(r,\beta)\} \\ &\quad = \Pr\left\{\tilde{J}_r > \beta \cdot \max[\tilde{J}_1, \tilde{J}_2, \ldots, \tilde{J}_j, \ldots, \tilde{J}_R]\right\} \\ &\quad = \Pr\left\{\tilde{J}_r > \beta \cdot \max_{j\neq r}[\tilde{J}_1, \tilde{J}_2, \ldots, \tilde{J}_j, \ldots, \tilde{J}_R]\right\}. \end{aligned}$$

Apply Lemma 1 and replace X_j with $\beta\tilde{J}_j$ for $j = 1, \ldots, R$ to have

$$\Pr\left\{\tilde{J}_r > \beta \cdot \max_{j\neq r}[\tilde{J}_1, \ldots, \tilde{J}_j, \ldots, \tilde{J}_R]\right\} \geq \prod_{j=1, j\neq r}^{R} \Pr\{\tilde{J}_r > \beta\tilde{J}_r\} = \text{APCS}_2(r,\beta).$$

APPENDIX C
DETAILED SIMULATION RESULTS

TABLE XVI
LONG-TERM PERFORMANCE ANALYSIS: SINGLE-FACTOR

Factor	Levels	MCT[6]	VCT	SM
WRP(F1)	DET	7.19	3.16	4.67
	WR	-6.43	-0.36	-5.76
	DET-BH	6.30	1.05	5.05
	WR-BH	-7.07	-3.85	-3.95
LDR(F2)	FSMCT	-10.25	-245.31	36.74
	LDF	-7.91	-33.68	46.18
	OSA	-1.35	148.19	-35.79
	FIFO	19.52	130.80	-47.13
LAP_PH(F3)	LTWP	-4.76	-5.20	3.80
	LTWL	-9.01	-24.18	9.39
	LGWP	6.76	15.00	-6.48
	LGWL	7.01	14.38	-6.70
LDR_F(F4)	LAS	0.12	11.49	4.96
	LLS	-3.26	-19.73	5.77
	LNGQ	1.55	-21.05	-2.36
	LLA	1.59	29.28	-8.37
LDR_UF(F5)	SAF1	0.02	34.77	-2.83
	SAF2	2.81	40.27	-7.95
	FSMCT	-2.56	-70.80	13.49
	FIFO	-0.27	-4.24	-2.71

[6]average performance of a specific factor level - average performance of all policies

TABLE XVII
SCHEDULING POLICY RANKING FOR MCT OVER FOUR WEEKS

(a) Under IS_1

Rank	WRP	LDR	LAP_PH	LDR_F	LDR_UF	MCT(hrs)
1	DET-B	FSMCT	LGWP	LNGQ	FIFO	401.46
2	DET	FSMCT	LGWP	LLA	SAF1	402.64
3	DET-B	FSMCT	LGWP	LNGQ	SAF2	402.84
4	DET-B	FSMCT	LGWP	LNGQ	SAF1	403.37
5	DET	LDF	LTWP	LNGQ	FIFO	403.88
6	DET-B	FSMCT	LTWP	LNGQ	SAF1	404.24
7	DET	FSMCT	LGWP	LNGQ	SAF2	405.22
8	DET	FSMCT	LGWP	LLS	SAF1	405.86
9	DET	FSMCT	LGWP	LNGQ	FIFO	405.90
10	DET	LDF	LTWP	LNGQ	SAF2	406.68

(b) Under IS_2

Rank	WRP	LDR	LAP_PH	LDR_F	LDR_UF	MCT(hrs)
1	WR	FSMCT	LGWP	LLS	SAF1	406.89
2	WR	FSMCT	LGWL	LLS	FIFO	408.09
3	WR	LDF	LGWP	LLS	FIFO	408.44
4	DET	FSMCT	LGWP	LLS	FIFO	408.53
5	DET-BH	FSMCT	LGWL	LLS	SAF1	409.46
6	DET	LDF	LGWL	LLS	SAF1	410.33
7	DET	LDF	LGWP	LLS	SAF1	410.64
8	DET-BH	FSMCT	LGWP	LNGQ	FIFO	410.77
9	DET-BH	FSMCT	LGWP	LLS	SAF2	411.14
10	DET	FSMCT	LGWL	LLS	FIFO	411.88

TABLE XVIII
SCHEDULING POLICY RANKING FOR VCT OVER ONE WEEK

(a) Under IS_1

Rank	WRP	LDR	LAP_PH	LDR_F	LDR_UF	VCT(hrs^2)
1	DET-BH	FSMCT	LGWP	LNGQ	FSMCT	74.32
2	WR	FSMCT	LTWP	LNGQ	FSMCT	80.94
3	DET	FSMCT	LGWP	LLA	FSMCT	82.64
4	DET	FSMCT	LGWP	LNGQ	FSMCT	86.43
5	WR-BH	FSMCT	LTWP	LNGQ	SAF2	88.93
6	DET	FSMCT	LTWP	LLA	FSMCT	89.39
7	DET	FSMCT	LGWP	LLA	SAF2	89.84
8	DET	FSMCT	LTWP	LNGQ	SAF2	98.57
9	DET-BH	FSMCT	LTWP	LNGQ	FSMCT	104.68
10	DET-BH	FSMCT	LTWP	LNGQ	SAF2	105.91

(b) Under IS_2

Rank	WRP	LDR	LAP_PH	LDR_F	LDR_UF	VCT(hrs^2)
1	WR-BH	FSMCT	LTWL	LLA	SAF1	200.20
2	WR	FSMCT	LTWL	LLA	SAF1	200.42
3	WR	FSMCT	LTWL	LNGQ	SAF1	201.05
4	WR	FSMCT	LTWL	LLA	FIFO	207.45
5	WR	FSMCT	LTWL	LNGQ	SAF2	208.09
6	WR	FSMCT	LGWL	LLA	SAF1	217.04
7	WR-BH	FSMCT	LTWL	LNGQ	SAF1	218.26
8	WR	FSMCT	LGWP	LNGQ	SAF1	221.53
9	DET-BH	FSMCT	LTWL	LLA	SAF1	226.37
10	WR-BH	FSMCT	LTWL	LLA	SAF2	227.23

TABLE XIX
TWO-FACTOR INTERACTION EFFECTS FOR VCT

L_A	L_B	Two-factor Interaction Effects (VCT)									
		F1-F2	F1-F3	F1-F4	F1-F5	F2-F3	F2-F4	F2-F5	F3-F4	F3-F5	F4-F5
1	1	-22.2	-20.9	18.89	23.83	0.30	-15.9	-36.9	21.91	6.47	50.75
1	2	33.38	-3.17	-7.20	-13.6	18.42	12.46	-34.6	-9.43	7.56	3.43
1	3	16.74	10.81	-6.51	-2.74	-8.32	27.52	69.21	-4.85	-11.7	-19.5
1	4	-27.8	13.35	-5.18	-7.47	-10.4	-24.0	2.39	-7.63	-2.26	-34.6
2	1	20.26	16.51	-7.82	-9.43	-81.7	8.79	-33.9	27.85	31.83	-13.4
2	2	-43.7	16.49	3.15	-0.34	-42.9	-0.71	-7.25	-12.9	-14.5	-5.45
2	3	5.65	-14.2	5.39	4.07	61.93	15.10	66.15	-8.62	-5.63	8.76
2	4	17.81	-18.7	-0.72	5.70	62.78	-23.1	-24.9	-6.26	-11.6	10.12
3	1	-18.3	-14.9	6.89	4.91	38.21	26.67	58.07	-27.9	-20.5	-25.3
3	2	49.60	-23.4	4.03	-0.66	23.37	-10.4	6.10	9.11	-0.63	-5.46
3	3	-24.1	14.43	-9.66	-0.04	-31.1	-29.6	-68.9	9.11	9.71	17.52
3	4	-7.17	24.00	-1.26	-4.20	-30.4	13.49	4.76	9.71	11.46	13.26
4	1	20.35	19.44	-17.9	-19.3	43.27	-19.4	12.81	-21.8	-17.7	-12.0
4	2	-39.2	10.15	0.02	14.63	1.15	-1.27	35.84	13.29	7.61	7.48
4	3	1.72	-10.9	10.78	-1.29	-22.4	-12.9	-66.4	4.36	7.70	-6.75
4	4	17.19	-18.6	7.16	5.97	-21.9	33.68	17.78	4.18	2.45	11.26

TABLE XX
TWO-FACTOR INTERACTION EFFECTS FOR SM

L_A	L_B	Two-factor Interaction Effects (SM) $x10^3$									
		F1-F2	F1-F3	F1-F4	F1-F5	F2-F3	F2-F4	F2-F5	F3-F4	F3-F5	F4-F5
1	1	-2.27	-0.13	-0.78	-0.41	3.85	-3.45	1.67	-1.94	0.46	0.73
1	2	1.48	-2.83	-1.03	0.54	7.80	-2.13	3.70	-0.43	-1.30	-0.29
1	3	-1.69	2.07	0.35	-2.20	-6.16	0.47	-8.45	1.32	0.45	-1.59
1	4	2.48	0.89	1.46	2.06	-5.49	5.11	3.08	1.05	0.39	1.16
2	1	5.02	1.11	1.54	0.07	1.48	-7.40	3.93	-0.01	-0.29	-1.00
2	2	-1.44	3.02	-0.44	1.32	-4.13	-1.66	0.99	-0.63	-1.22	1.58
2	3	-0.28	-2.64	-0.43	0.93	1.15	3.51	-8.12	-0.11	0.84	-1.43
2	4	-3.30	-1.49	-0.67	-2.32	1.49	5.55	3.19	0.74	0.67	0.85
3	1	-4.52	-0.10	0.41	-0.28	-2.06	5.06	-1.96	0.81	0.06	0.93
3	2	2.01	-3.40	0.00	0.82	-2.78	-0.15	-1.17	0.02	1.68	-0.05
3	3	1.08	0.99	-0.80	-1.30	3.24	-0.88	6.29	-0.34	-0.90	0.08
3	4	1.42	2.51	0.39	0.77	1.59	-4.02	-3.15	-0.49	-0.85	-0.97
4	1	1.77	-0.89	-1.17	0.62	-3.28	5.79	-3.63	1.14	-0.23	-0.66
4	2	-2.05	3.21	1.47	-2.68	-0.89	3.95	-3.52	1.04	0.84	-1.24
4	3	0.89	-0.42	0.88	2.57	1.76	-3.09	10.28	-0.88	-0.40	2.94
4	4	-0.61	-1.91	-1.19	-0.51	2.41	-6.64	-3.13	-1.30	-0.21	-1.04

TABLE XXI
SAMPLING USING OA(16,5,4,2)

TS[7] No	WRP	LDR	LAP _PH	LDR _F	LDR _UF	MCT	VCT	SM
1	DET	FSMCT	LTWP	LAS	SAF1	469.61	131.48	0.830
2	DET	LDF	LTWL	LLS	SAF2	469.44	253.19	0.842
3	DET	OSA	LGWP	LNGQ	FSMCT	489.50	387.29	0.763
4	DET	FIFO	LGWL	LLA	FIFO	540.56	627.75	0.718
5	WR	FSMCT	LTWL	LNGQ	FIFO	474.55	208.47	0.835
6	WR	LDF	LTWP	LLA	FSMCT	462.20	283.75	0.835
7	WR	OSA	LGWL	LAS	SAF2	496.45	574.22	0.749
8	WR	FIFO	LGWP	LLS	SAF1	502.40	591.83	0.723
9	DET	FSMCT	LGWP	LLA	SAF2	489.12	166.57	0.800
10	DET	LDF	LGWL	LNGQ	SAF1	523.64	451.43	0.830
11	DET	OSA	LTWP	LLS	FIFO	490.25	577.25	0.763
12	DET	FIFO	LTWL	LAS	FSMCT	487.85	369.87	0.778
13	WR	FSMCT	LGWL	LLS	FSMCT	484.69	185.82	0.827
14	WR	LDF	LGWP	LAS	FIFO	467.49	342.64	0.792
15	WR	OSA	LTWL	LLA	SAF1	488.28	604.53	0.751
16	WR	FIFO	LTWP	LNGQ	SAF2	508.36	609.99	0.704

[7]TS: Test Sample

TABLE XXII
OO-BASED DOE RANKINGS OF TOP-16 POLICIES SELECTED BY ESTIMATION MODELS

Rank	Model: Single-factor Effect			Model: Single and two-factor effects	
	MCT	VCT	SM	MCT	VCT
1	18	5	2	1	1
2	15	35	5	5	15
3	43	25	15	13	16
4	27	23	13	12	9
5	29	40	56	6	28
6	77	44	62	15	23
7	80	37	42	160	29
8	106	49	63	144	59
9	71	99	32	135	113
10	136	104	93	176	120
11	114	146	134	52	45
12	122	70	160	156	56
13	173	114	115	243	128
14	261	121	136	280	89
15	158	157	114	227	99
16	445	153	178	169	178

REFERENCES

[1] M. Liu, "Advanced foundry in the consumer electronics era," in *Proc. 2nd ISMI Symp. Manufact. Effectiveness, Keynote Speech*, Austin, TX, Oct. 24–26, 2005. [Online]. Available: http://ismi.sematech.org/ismisymposium/past/II/abstracts/K1Liu_Mark.pdf

[2] SEMATECH, "Consumer electronics may spawn gigafabs and 2X tool productivity increase," SEMATECH News, Austin, TX, 2005. [Online]. Available: www.sematech.org/corporate/news/releases/20051101.htm

[3] ITRS, "Factory integration." [Online]. Available: http://www.itrs.net/Links/2005ITRS/Factory2005.pdf 2005
[4] L. M. Wein, "Scheduling semiconductor wafer fabrication," *IEEE Trans. Semiconduct. Manufact.*, vol. 1, pp. 115–130, Aug. 1988.
[5] S. C. H. Lu, D. Ramaswamy, and P. R. Kumar, "Efficient scheduling policies to reduce mean and variance of cycle-time in semiconductor manufacturing plants," *IEEE Trans. Semiconduct. Manufact.*, vol. 7, pp. 374–388, Aug. 1994.
[6] J. W. Lawton, A. Drake, R. Henderson, L. M. Wein, R. Whitney, and D. Zuanich, "Workload regulating wafer release in a GaAs fab facility," in *Proc. IEEE/SEMI Int. Semiconduct. Manufact. Sci. Symp.*, 1990, pp. 33–38.
[7] D. W. Collins, K. Williams, and F. C. Hoppensteadt, "Implementation of minimum inventory variability scheduling 1-step ahead policy(R) in a large semiconductor manufacturing facility," in *Proc. 6th Int. Conf. Emerging Technol. Factory Autom.*, 1997, pp. 497–504.
[8] M. Thompson, "Using simulation-based finite capacity planning and scheduling software to improve cycle time in front end operations," in *Proc. IEEE/SEMI 1995 Adv. Semiconduct. Manufact. Conf. Workshop*, 1995, pp. 131–135.
[9] J. Christopher, M. E. Kuhl, and K. Hirschman, "Simulation analysis of dispatching rules for automated material handling systems and processing tools in semiconductor fabs," in *Proc. 2005 IEEE Int. Symp. Semiconduct. Manufact.*, San Jose, CA, Sep. 2005, pp. 84–87.
[10] C. R. Glassey and M. G. C. Resende, "Closed-loop job release control for VLSI circuit manufacturing," *IEEE Trans. Semiconduct. Manufact.*, vol. 1, pp. 36–46, Feb. 1988.
[11] C. R. Glassey, J. G. Shanthikumar, and S. Seshadri, "Linear control rules for production control of semiconductor fabs," *IEEE Trans. Semiconduct. Manufact.*, vol. 9, pp. 536–549, Nov. 1996.
[12] S. Li, T. Tang, and D. W. Collins, "Minimum inventory variability schedule with applications in semiconductor fabrication," *IEEE Trans. Semiconduct. Manufact.*, vol. 9, pp. 145–149, Feb. 1996.
[13] M. Mittler and A. K. Schoemig, "Comparison of dispatching rules for semiconductor manufacturing using large facility models," in *Proc. 1999 Winter Simulation Conf.*, Dec. 1999, vol. 1, pp. 709–713.
[14] G. S. Baweja and H. T. La, "Real-time lot dispatching for semiconductor fab," *Future Fab Int.*, vol. 11, 2001.
[15] Y.-D. Kim, J.-U. Kim, S.-K. Lim, and H.-B. Jun, "Due-date based scheduling and control policies in a multiproduct semiconductor wafer fabrication facility," *IEEE Trans. Semiconduct. Manufact.*, vol. 11, pp. 155–164, Feb. 1998.
[16] N. G. Pierce and T. Yurtsever, "Dynamic dispatch and graphical monitoring system," in *Proc. 1999 IEEE Int. Symp. Semiconduct. Manufact. Conf.*, Oct. 1999, pp. 65–68.
[17] M. Shickmair, M. Graml, C. Pichler, and W. Laure, "Simulation-based synthesis of composite dispatching rules," in *Proc. 16th Eur. Simulation Symp.*, Budapest, Hungary, Oct. 17–20, 2004, pp. 267–272.
[18] V. S. Kouikoglou and Y. A. Phillis, "Discrete event modeling and optimization of unreliable production lines with random rates," *IEEE Trans. Robot. Autom.*, vol. 10, pp. 153–159, Apr. 1994.
[19] B. W. Hsieh, C. H. Chen, and S. C. Chang, "Scheduling semiconductor wafer fabrication by using ordinal optimization-based simulation," *IEEE Trans. Robot. Autom.*, vol. 17, pp. 599–608, Oct. 2001.
[20] Y. C. Ho, R. S. Sreenivas, and P. Vakili, "Ordinal optimization of DEDS," *J. Discrete Event Dynamic Syst.*, vol. 2, no. 2, pp. 61–88, 1992.
[21] L. Dai, "Convergence properties of ordinal comparison in the simulation of discrete-event dynamic systems," *J. Opt. Theory Appl.*, vol. 91, no. 2, pp. 363–388, Nov. 1996.
[22] L. Dai and C.-H. Chen, "Rate of convergence for ordinal comparison of dependent simulations in discrete-event dynamic systems," *J. Opt. Theory Appl.*, vol. 94, no. 1, pp. 29–54, Jul. 1997.
[23] R. E. DeVor, T.-H. Chang, and J. W. Sutherland, *Statistical Quality Design and Control: Contemporary Concepts and Methods*. New York: Macmillan, 1992.
[24] A. S. Hedayat, N. J. A. Sloane, and J. Stufken, *Orthogonal Arrays: Theory and Applications*. New York: Springer, 1999.
[25] H.-C. Chen, L. Dai, C.-H. Chen, and E. Yucesan, "New development of optimal computing budget allocation for discrete event simulation," in *Proc. 1997 Winter Simulation Conf.*, Dec. 1997, pp. 334–341.
[26] C. Y. Lin, "Shop floor scheduling of semiconductor wafer fabrication using real-time feedback control and predictions," Ph.D., Ind. Eng. Oper. Res., Univ. California at Berkeley, Berkeley, CA, 1996.
[27] C.-H. Chen, "A lower bound for the correct subset selection probability and its application to discrete-event system simulations," *IEEE Trans. Autom. Contr.*, vol. 41, pp. 1227–1231, Aug. 1996.
[28] D. C. Montgomery, *Design and Analysis of Experiments*, 5th ed. New York: Wiley, 2001.
[29] H.-C. Chen, C.-H. Chen, and E. Yucesan, "Computing efforts allocation for ordinal optimization and discrete event simulation," *IEEE Trans. Autom. Contr.*, vol. 45, pp. 960–964, May 2000.
[30] R. L. Plackett and J. P. Burman, "The design of optimum multifactorial experiments," *Biometrika*, vol. 33, no. 4, pp. 305–325, Jun. 1946.
[31] C. -H. Chen, J. Lin, E. Yücesan, and S. E. Chick, "Simulation budget allocation for further enhancing the efficiency of ordinal optimization," *J. Discrete Event Dynamic Syst.: Theory and Appl.*, vol. 10, pp. 251–270, Jul. 2000.

Bo-Wei Hsieh received the B.S., M.S., and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, in 1997, 1999, and 2002, respectively.

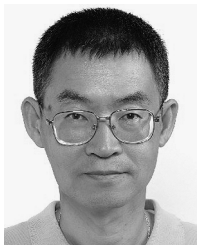
He is currently a Senior Logic Design Engineer at VIA Technologies, Inc., Fremont, CA.

Dr. Hsieh is one of the recipients of the Kayamori Best Automation Paper Award at the 2003 IEEE International Conference on Robotics and Automation.

Chun-Hung Chen (S'91–M'94–SM'01) received the B.S. degree in control engineering from National Chiao-Tung University, Hsinchu, Taiwan, R.O.C., in 1987, the M.S. degree in electrical engineering from National Taiwan University, Taipei, in 1989, and the Ph.D. degree in simulation and decision from Harvard University, Cambridge, MA, in 1994.

He is a Professor of Systems Engineering and Operations Research at George Mason University (GMU), Fairfax, VA. He was as Assistant Professor of Systems Engineering, University of Pennsylvania, Philadelphia, PA, before joining GMU. He was on leave with the Department of Electrical Engineering, National Taiwan University, during the Spring semester of 2006. His research interests include stochastic systems modeling and simulation, optimization, network management, systems design under uncertainty, and air traffic management. He currently leads NSF-sponsored research in developing efficient network simulation methodologies, and leads NASA-sponsored research in national air transportation simulation modeling.

Dr. Chen won the 1994 Harvard University Eliahu I. Jury Award for the Best Thesis in the Field of Control. He is one of the recipients of the 1992 MasPar Parallel Computer Challenge Award and the Kayamori Best Automation Paper Award at the 2003 IEEE International Conference on Robotics and Automation. He is a Program Co-Chair for the 2007 Informs Simulation Society Workshop. He is a listed in *Who's Who in America*, *Who's Who in Finance and Business*, *Who's Who in Science and Engineering*, and *Who's Who in Education*.



Shi-Chung Chang (S'83–M'87) received the B.S.E.E. degree from National Taiwan University, Taipei, in 1979, and the M.S. and Ph.D. degrees in electrical and systems engineering from the University of Connecticut, Storrs, in 1983 and 1986, respectively.

From 1979 to 1981, he served as an Ensign in the Chinese Navy, Taiwan. He worked as a Technical Intern at the Pacific Gas and Electric Company, San Francisco, CA, in the summer of 1985. During 1987, he was a member of the Technical Staff, Decision Systems Section, ALPHATECH, Inc., Burlington, MA. He has been with the Electrical Engineering Department, National Taiwan University since 1988 and was promoted to Professor in 1994. During 2001–2002, he served as the Dean of Student Affairs and a Professor of Electrical Engineering, National Chi Nan University, Pu-Li, Taiwan. He was a Visiting Scholar at the Electrical and Computer Engineering Department, University of Connecticut during his sabbatical leave in the 2003–2004 and 2006–2007 academic years. Besides the Electrical Engineering Department, he is now jointly appointed by the Graduate Institute of Industrial Engineering and the Graduate Institute of Communication Engineering, National Taiwan University. His research interests include optimization theory and algorithms, production scheduling and control, network management, Internet economics, and distributed decision making. He has been a principal investigator and consultant to many industry and government funded projects in the above areas, and has published more than 130 technical papers.

Dr. Chang is a member of Eta Kappa Nu and Phi Kappa Phi. He received the award for Outstanding Achievements in University–Industry Collaboration from the Ministry of Education for his pioneering research collaborations with the Taiwan semiconductor industry on production scheduling and control in 1996.