Journal of Manufacturing Systems
Vol. 22/No. 1
2003

Translating Overall Production Goals into Distributed Flow Control Parameters for Semiconductor Manufacturing

Ming-Der Hu and **Shi-Chung Chang,** Dept. of Electrical Engineering, National Taiwan University, Taipei, Taiwan

Abstract

As manufacturing systems are often operated under distributed control, translating overall production goals, such as system output rate and cycle time, into local flow control targets for internal operations, such as work-in-process (WIP) levels and part release intervals, has been a significant and challenging research topic. This paper studies the problem for semiconductor wafer fabrication plants, where there are multiple part types, failure-prone machines, and deterministic reentrant process flows. Such a system is first modeled as a failure-free and reentrant open queuing network with an aggregated part type. By exploiting a decomposition-based approximation, a backward queuing network analysis (BQNA) is designed to derive flow control parameters in terms of means and variations of input processes for each machine group. These derived control parameters can be utilized to obtain managerially tangible requirements for cycle times and WIP levels and to create guidelines for distributed flow control leading to the desired overall production goals. Comparisons with simulation results for two example wafer fabrication models demonstrate the accuracy and computational efficiency of BQNA and its potential for real applications.

Keywords: *Reentrant Production Lines, Queuing Models, Approximation Analysis, Flow Control Requirements, Wafer Fabrication Systems*

Introduction

A modern factory is often operated under a distributed control architecture. When shop floor performance measures are closely mapped to the desired system performance, individual machines may then be operated locally to meet their local performance requirements while effectively achieving the overall production goals. How the required overall production goals of a factory should be translated into performance requirements for individual production entities or processing steps within the factory has long been a significant research topic in production control. The translation problem has become more challenging because the competition for quick, high-quality, and low-cost responses to customer demands in manufacturing is fierce worldwide, and manufacturing system complexity is increasing.

Specifically, very large scale integrated circuit (VLSI) wafer fabrication is a business characterized by high capital investment, short product life cycles, high product add-on values, and stringent customer demands for both excellent quality and timely delivery. A wafer fabrication plant (fab) represents one of the most complex manufacturing processes. There are several hundred production steps in a process, which involves tens of delicate and expensive equipment groups. The process flow is highly reentrant because it requires multiple visits to individual equipment groups as successive circuit layers are added onto a wafer. This reentrant feature poses a unique challenge to production flow control (PFC) because wafers of different product types as well as wafers of the same product type but at different layers of fabrication may compete for the finite capacity of an equipment group (Graves et al. 1983; Kumar 1993). In addition, factors such as rework, scrap, wafer lot splitting/merging, equipment breakdown, preventive maintenance, operator availability, and setup requirements combine to make the manufacturing environment highly dynamic and uncertain. Translating the overall production goal to local performance requirements within such reentrant manufacturing systems has therefore been a significant research challenge.

In the literature, a few performance models of reentrant manufacturing systems have been presented, including: (1) discrete event simulation models (Dayhoff and Atherton 1984, 1986; Lu and Kumar 1991; Lu, Ramaswamy, Kumar 1994), (2) stochastic optimal control models (Kumar 1993; Lou and Kager 1989), and (3) queuing network models (Bitran and Tirupati 1988; Burman et al. 1986; Chen et al. 1988; Connors, Feigin, Yao 1996). Among them, a stochastic optimal control model usually leads to analysis of very small problems, although theoretic insights can

be generated from it. Discrete event simulation models are usually high-fidelity representations of a real fab and are frequently used for detailed design evaluation. However, they require significant development and maintenance efforts and are mostly suitable for descriptive/evaluative rather than prescriptive purposes. Analyses based on queuing network models have been successfully applied to both planning and performance evaluation problems of realistic scales. Fast computation times can be achieved on a personal computer. Several software tools based on approximate analyses of queuing network models are available for performance evaluation of manufacturing systems. QNA (Segal and Whitt 1988; Whitt 1983) is intended to provide approximations for various performance measures of a general open queuing network (OQN). MPX developed by Network Dynamics Inc. models manufacturing systems as general OQNs and adopts a number of approximation techniques for quick performance evaluation. CHIPS, a joint project between SEMATECH and Texas A&M University, provides performance analysis for fabs through a combination of queuing analysis and simulation.

Many queuing network model based analyses and PFC designs utilize mean values such as mean demand rates for individual operations and mean processing rates of individual machines. Inverse queuing network analysis (IQNA) proposed by Kuroda and Kawada (1995) is an optimization method for designing wafer release control. It combines a mean value analysis (MVA) of a queuing network model for quick performance evaluation and a simulated annealing algorithm to search for optimal wafer in process (WIP) levels within a fab and the corresponding wafer release control to achieve the required throughput of each product. Narahari and Khan (1996) presented an MVA-based method to predict the throughput rate and the cycle time of a reentrant line under the various queuing disciplines studied by Lu and Kumar (1991), such as First Buffer First Serve and Last Buffer First Serve. Lin and Lee (2001) proposed a queuing network-based algorithm to determine the standard WIP level of a fab by adding the WIP at all workstations in the fab. They determine the output rate of each reentrant loop through the bottleneck workstation by giving the required overall throughput rate and product mix ratio. Each workstation is treated as an M/M/m service node for WIP calculation. All three aforementioned works assume exponentially distributed part processing times.

To remove the limitation of assuming an exponential time distribution and for better handling of uncertainties, the aspect of variability, i.e., second-order statistics, as well as the mean values, has been brought into practice (Li, Tang, Collins 1996; Sattler 1996; Shiu et al. 1996; Böbel and Halmel 1998). Specifically, Li, Tang, and Collins proposed a minimum inventory variability scheduling (MIVS) scheme to determine the priority among individual operation steps. The authors clearly indicated how the mean WIP and cycle time of a machine depend on both mean and variability of the arrival and service processes by exploiting both Little's formula and Kingman's approximation (Gelenbe and Pujolle 1987) for GI/G/1 queue. MIVS calculates the difference between the measured instantaneous WIP and the historical WIP average of each operation step. Priority among operational steps is then determined by comparing the relative differences among consecutive operation steps. Field applications of MIVS have led to significant reductions of mean and variance of WIP and cycle times. However, the question of how to systematically translate overall production goals into managerially tangible local control parameters in terms of mean rate and variability remains an open challenge.

Motivated in this paper by the need for a production goal translation, the translation problem is first formulated in terms of OQN modeling. A fab with multiple part types, failure-prone machines, and reentrant process flows is considered and modeled as a failure-free and reentrant OQN of a single aggregated part type. First-come first-serve (FCFS) is assumed as the service discipline for each node, where a node corresponds to one machine group. This aggregated, reentrant OQN is analyzed by using a class of approximate decomposition methods. The decomposition methods decompose an OQN into individual network nodes and use two types of parameters to characterize the stochastic arrival, service, and departure processes of each node: one describing the rate and the other describing the variability. Various stationary network performance measures, such as work in process, utilization, and flow time, can then be derived based on these two types of parameters. There have been quite a few research results in this direction (Whitt 1983; Buzacott and Yao 1986; Bitran and

Tirupati 1988; Segal and Whitt 1988; Buzacott and Shanthikumar 1992; Connors, Feigin, Yao 1996). Specifically, applications to semiconductor manufacturing have been proposed by Bitran and Tirupati; Segal and Whitt; and Connors, Feigin, and Yao.

In conjunction with the reentrant OQN model and the decomposition method, a new solution method is designed for the translation problem—the backward queuing network analysis (BQNA). BQNA derives target mean rates and variation bounds of input processes of individual nodes from desired overall production goals. Frequently considered production goals of fabs include the output rate, inter-output time variance, mean cycle time, or cycle time variance, where cycle time is the time between job release and its completion. Specifically, BQNA takes a reverse view of relations among variables in the results of the decomposition-based approximation. BQNA leads to two sets of linear equations with system production goals and nodal service parameters such as mean and variability of service times given. Results of BQNA then serve as the basis for calculating various managerially tangible and local PFC requirements such as bounds on mean and variance of waiting times and queue lengths. Although the derivation of local performance requirements assumes FCFS at individual nodes, the requirements thus derived represent a minimum level of performance for designing PFC schemes.

Translation Problem

In fab manufacturing practice, there are two major functions for PFC of a fab: wafer release and dispatching. To meet production goals under the manufacturing constraints of a fab, wafer release regulates the arrivals of new wafers into the fab according to an output schedule. Dispatching is locally carried out at each machine or machine group. Dispatching ranks the processing priority among wafers of various product types and production steps by taking local status and performance requirements into consideration. Both release and dispatching functions should be integrated to control wafer flows for achieving the required overall production goals. A manufacturing execution system collects measurable information from the shop floor and facilitates the decision making of PFC functions. *Figure 1* depicts such a basic architecture.

Overall fab production goals, such as weekly minimum fab output volume and maximum average cycle time, need to be translated into local performance requirements for the two PFC functions. The form of local performance requirements varies with the specific PFC scheme used. Examples of frequently used forms include target number of wafers processed during one shift or maximum and minimum WIP levels at each production step. PFC parameters are then periodically extracted from both the measured fab status data and the given system production goals, and are fed into the PFC scheme. Production line managers or a PFC computer executes the control actions calculated/recommended by the PFC scheme to regulate raw wafer release and dispatch wafers to individual machines for manufacturing.

Fab Modeling

Effective translation between fab goals and local performance requirements starts with a good model of production flows in a fab. Consider a fab modeled as a general reentrant line and illustrated in *Figure 2*, into which wafers arrive and from which wafers eventually exit. In this model, there are M machine groups (MGs), I part types, and J sequential production steps. Machines in one group are identical, independent, and failure prone. The process flow of each product type consists of a subsequence of the J sequential production steps, of which the first step always starts at MG 1 while the last step always finishes at MG M. This is not just for convenience of later analysis but to capture a feature of process flows in a real wafer fab. Each production step requires one machine of a specific group for processing. It is assumed that there is an infinite buffer, denoted by b_j in *Figure 2*, for each step because buffer space for wafer storage is usually not a limiting constraint in a fab. As a MG may be revisited several times at different production steps of one process flow, the line is reentrant.

Just like other production lines, wafer scrap or rework may occur after an inspection step. Scrapped wafers leave the line and rework wafers go into respective sequences of rework depending on the step where the scrap occurs. A machine may fail at random due to either functional breakdown or unscheduled maintenance. When a failure occurs, the failed machine requires a random repair time.

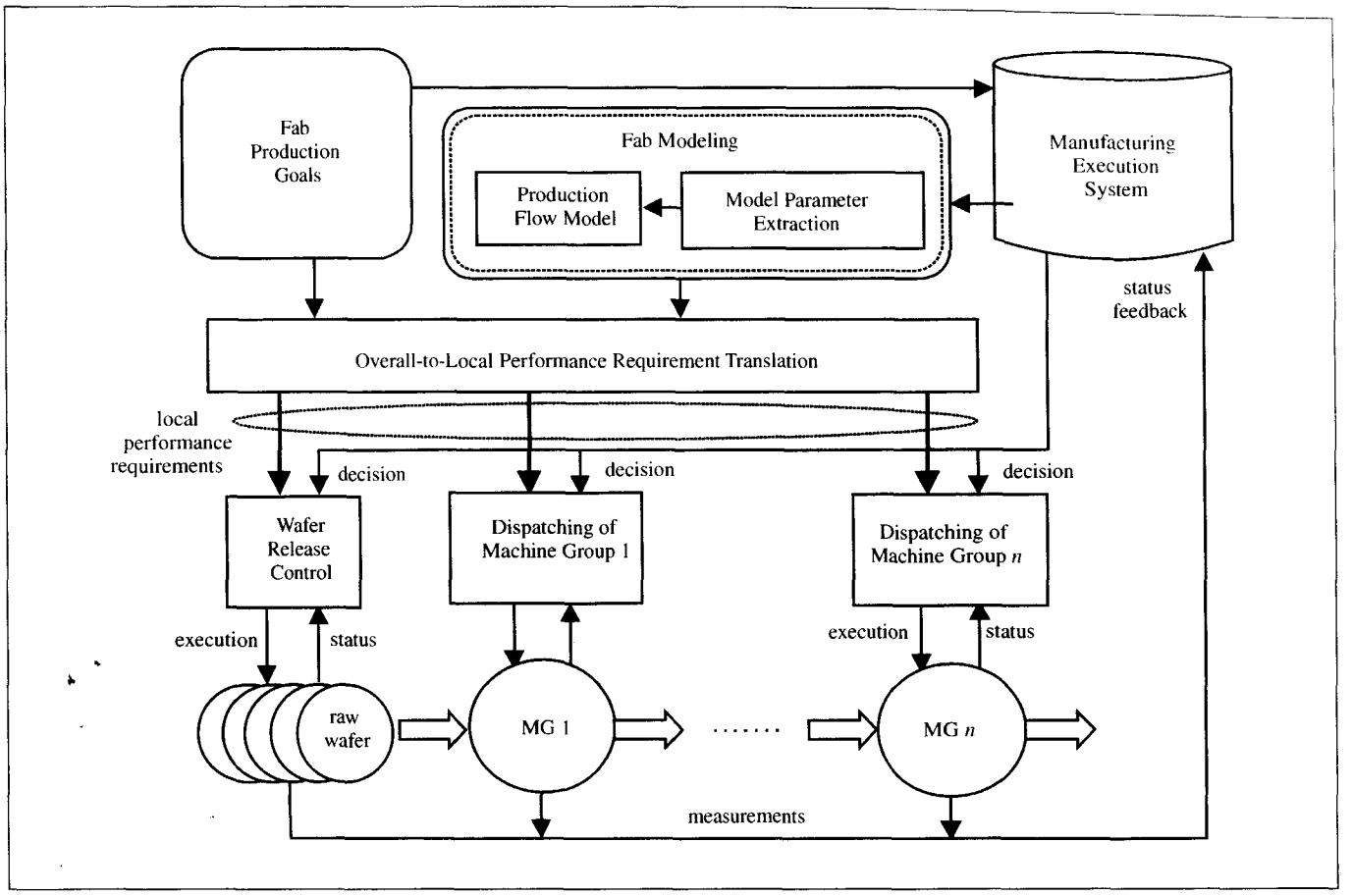

Figure 1
PFC Architecture

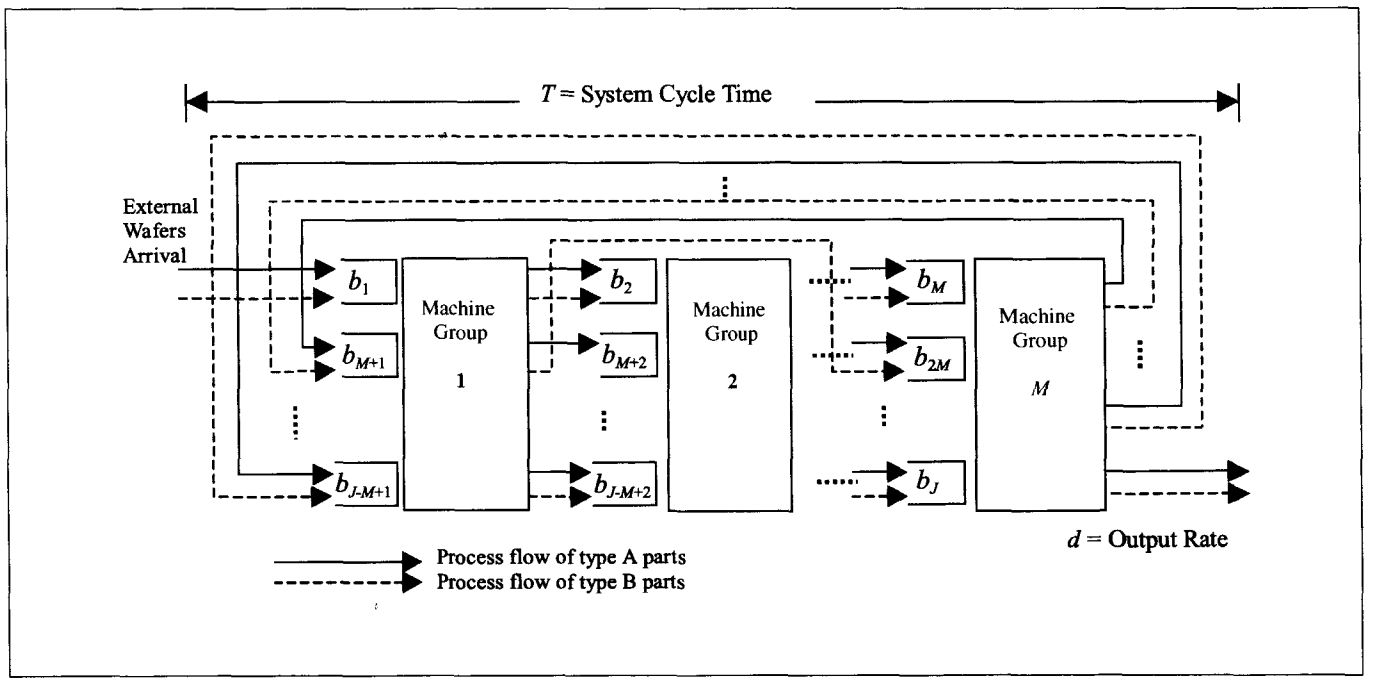

Figure 2
Reentrant Production Flow Model of a Fab

OQN Model

In modeling a fab as an open queuing network, a MG m of M_m identical machines is represented as a service node of parallel servers with an infinite and shared buffer, $m = 1, 2, \ldots, M$. The fabrication of a wafer corresponds to a job, and each part type corresponds to a job class. Each class-i job, $i = 1, 2, \ldots, I$, has a total of S_i processing steps. Let (i, k) denote the kth processing step of a class-i job. The process flow of a class-i job follows a deterministic route, $\{(i, k), k = 1, 2, \ldots, S_i\}$, which is a subsequence of $\{1, 2, \ldots, J\}$. Step (i, k) requires processing by a service node $m_{ik} \in \{1, 2, \ldots, M\}$. Although a job may visit a service node more than once, it is assumed that $m_{ik} \neq m_{i,k+1}$, $k \in \{1, 2, \ldots, S_i - 1\}$ for any class i, that is, any two consecutive steps of a job must be processed by different MGs. For simplicity of discussion, scrap and rework can be but are not included in the modeling.

The service time of each node varies among production steps, and service times of individual job classes at a given node are assumed independent of each other. Each machine has its own independent probability distributions of processing times, up time, and down time. The Appendix illustrates one procedure (Kouvelis and Tirupati 1991) that incorporates machine failure effects by modifying the service time distribution of each node and forms a statistically equivalent failure-free node. Interarrival times of individual job classes to the entry node, MG 1, are assumed independent and identically distributed and are independent among job classes. FCFS is the service discipline of each node. In the literature, it has been shown that dispatching by FCFS may lead to instability of a reentrant line in certain situations (Bramson 1994; Seidman 1994). Hasenbein (1997) provided a procedure that can be used to check if a given reentrant line is stable under FCFS dispatching. This paper assumes that a reentrant line adopts FCFS dispatching and is stable.

Translation Problem Definition

To convey key ideas of the translation problem, consider, for example, an OQN consisting of a single GI/G/1 queue shown in *Figure 3*. This single machine-group system has service time parameters, $\underline{\theta} = (\tau, C_s^2)$, and part release parameters, $\underline{\alpha} = (\lambda, C_a^2)$, where τ is the mean service time, λ is the mean release rate, and C_s^2 and C_a^2 are squared coefficients of variation (SCV) of service and interrelease times, respectively. Note that the SCV of a random variable is defined as its variance divided by the square of its mean. When the queue is in steady state, it is obvious that the mean output rate, d, should be equal to the mean release rate, i.e., $d = \lambda$. The mean system cycle time (or sojourn time), T, can be approximately expressed by Kingman's equation (Gelenbe and Pujolle 1987) as

$$T = \tau + \frac{\lambda\tau^2}{1-\lambda\tau} \cdot \frac{\left(C_a^2 + C_s^2\right)}{2}$$

Let the production goals be that the cycle time should be no greater than T and that the output rate should be no less than d. Set goal parameters as $\underline{Y} = (T, d)$. The production goal translation problem is then to find values of the wafer release parameters $\underline{\alpha}$ to achieve the goal performance specification $\underline{Y}$ under the given machine group with characteristic parameters $\underline{\theta}$. Namely, solve for $\underline{\alpha} = (\lambda, C_a^2)$ such that $\lambda \geq d$, and

$$T = \tau + \frac{\lambda\tau^2}{1-\lambda\tau} \cdot \frac{\left(C_a^2 + C_s^2\right)}{2} \leq T$$

Denote $(\bar{\lambda}, \bar{C}_a^2)$ as the solution when the equality holds, that is,

$$\bar{\lambda} = d \text{ and}$$

$$\bar{C}_a^2 = \frac{2\left(1-\bar{\lambda}\tau\right)(T-\tau)}{\bar{\lambda}\tau^2} - C_s^2$$

Then the mean output rate and mean cycle time of this queue should meet goal levels if the job release is controlled at the target mean rate, $\bar{\lambda}$, and its interarrival time SCV is restrained no greater than the value $\bar{C}_a^2$.

Now a generic definition of the translation problem is given.

Definition: Translation Problem

Let $\underline{Y} \equiv [Y_1, Y_2, \ldots, Y_L]$ be L fab production goals, $\underline{y}_m \equiv [y_{m1}, y_{m2}, \ldots, y_{ml}]$ be l local performance requirements for node (machine group) m in the OQN model, $\underline{y} \equiv [\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_M]$ be the collection of per-

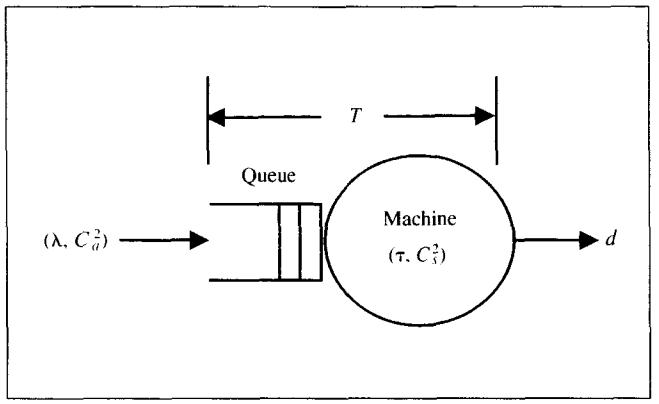

Figure 3
A GI/G/1 Queue

formance requirements for individual machine groups, $\underline{\theta}_m \equiv [\theta_{m1}, \theta_{m2}, \ldots, \theta_{mv}]$ be v service node parameters of node m, $\underline{\theta} \equiv [\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_M]$ be the collection of service parameters of individual machine groups, $\underline{\alpha} \equiv [\alpha_1, \alpha_2, \ldots, \alpha_v]$ be v parameters of the wafer release process, and $\mathbf{Q}$ be process routing data. In the OQN model, let two vector functions $\underline{Y} = \underline{F}(\underline{\alpha}, \underline{\theta}, \mathbf{Q})$ and $\underline{y} = \underline{G}(\underline{\alpha}, \underline{\theta}, \mathbf{Q})$ be the global and local performance functions, respectively, characterizing the mathematical relations among global goal, local performance targets, and model parameters. The translation problem is to solve $\underline{\alpha}$ from the global performance function, $\underline{Y} = \underline{F}(\underline{\alpha}, \underline{\theta}, \mathbf{Q})$, given the values of fab production goals, $\underline{Y}$, process routings, $\mathbf{Q}$, and all service node parameters, $\underline{\theta}$, and then to obtain the local performance targets $\underline{y} = \underline{G}(\underline{\alpha}, \underline{\theta}, \mathbf{Q})$ with the derived $\underline{\alpha}$.

Decomposition-Based Approximation

In defining the translation problem of a fab, the OQN model of the fab first needs to be analyzed for constructing both the global and local performance functions, $\underline{Y} = \underline{F}(\underline{\alpha}, \underline{\theta}, \mathbf{Q})$ and $\underline{y} = \underline{G}(\underline{\alpha}, \underline{\theta}, \mathbf{Q})$. The most frequently used analysis has been by the decomposition approximation approach. The approach (Bitran and Tirupati 1988; Whitt 1983) consists of three parts. The first part aggregates multiclass jobs into a single class. The second part is an analysis that essentially attempts to generalize the notion of independence and product form results for Jackson-type networks to more general systems. It relies on two notions:

1. The nodes can be treated independently as GI/G/m queues.
2. Two parameters, average rate and squared coefficient of variation, are used to characterize an arrival and a service process at each node.

In the decomposition approximation analysis, traffic equations characterize the nodal output processes in terms of input and service parameters of individual nodes. Relations among mean rates and inter-event time SCVs of processes at individual nodes are summarized by two sets of linear equations. Finally, the third part derives mean rate per job class and interarrival time SCVs at individual nodes via a disaggregation procedure.

Class Aggregation

In this paper, the procedure of aggregating the multiclass OQN into a single-class OQN follows that of Whitt (1983). Now defined are some notations for a brief description of the aggregation procedure:

- I: total number of classes;
- i: class index, $i = 1, \ldots, I$;
- λ_i: external arrival rate of class i to entry node;
- C_i^2: interarrival time SCV of class-i arrivals to entry node;
- S_i: total number of steps in the process flow of class i;
- k: operation step index, $k = 1, \ldots, S_i$;
- τ_{ik}: mean processing time at kth operation step of a class-i job;
- C_{ik}^2: service time SCV of a class-i job at its kth processing step;
- $\mathbf{1}H(\bullet)$: a membership function defined on set H ($\mathbf{1}H(x) = 1$ if $x \in H$, and 0 otherwise);
- M: total number of nodes;
- m, n: node indices, $m, n = 1, \ldots, M$.

The overall external arrival rates to the entry node and the flow rates within the network are obtained by adding up average flow rates of individual classes:

$$\lambda_e = \sum_{i=1}^{I} \lambda_i ,$$

$$\lambda_{mn} = \sum_{i=1}^{I} \sum_{k=1}^{S_i - 1} \lambda_i \mathbf{1}\{(i,k): m_{ik} = m, m_{i,k+1} = n\}, \forall m \neq$$

n, and $m, n = 1, 2, \ldots, M$.

As the external arrivals of different classes are independent, the interarrival time SCV of aggregate external arrivals to the entry node is as follows:

$$C_e^2 = \sum_{i=1}^{I} C_i^2 \left(\lambda_i \Big/ \sum_{j=1}^{I} \lambda_j \right)$$

The ratio of routing from node m to node n can be calculated as follows:

$$q_{mn} = \lambda_{mn} \Big/ \sum_{i=1}^{I} \sum_{k=1}^{S_i} \lambda_i \mathbf{1}\{(i,k): m_{ik} = m\}$$

The aggregate service time of a step at node m is a composite of service times of individual steps of various classes that require processing by server node m. The average aggregate service time at node m is then

$$\tau_m = \sum_{i=1}^{I} \sum_{k=1}^{S_i} \lambda_i \tau_{ik} \mathbf{1}\{(i,k): m_{ik} = m\} \Big/ \sum_{i=1}^{I} \sum_{k=1}^{S_i} \lambda_i \mathbf{1}\{(i,k): m_{ik} = m\}$$

and the corresponding SCV is

$$C_{sm}^2 = \sum_{i=1}^{I} \sum_{k=1}^{S_i} \lambda_i \tau_{ik}^2 \left(C_{ik}^2 + 1\right) \mathbf{1}\{(i,k): m_{ik} \neq m\} \Big/ \sum_{i=1}^{I} \sum_{k=1}^{S_i} \lambda_i \tau_m^2 \mathbf{1}\{(i,k): m_{ik} = m\} - 1$$

Traffic Rate Equation

When the network is in a steady state, there is a flow rate balance relationship among network nodes. Let λ_{an} be the job arrival rate of node n. Then

$$\lambda_{an} = \lambda_e \delta_n + \sum_{m=1}^{M} \lambda_{mn} = \lambda_e \delta_n + \sum_{m=1}^{M} \lambda_{am} q_{mn}, \; 1 \le n \le M \tag{1}$$

where $\delta_n = 1$ if $n = 1$ and $\delta_n = 0$ otherwise. In Eq. (1), there are M equalities for the M unknown variables $\{\lambda_{am}, m = 1, 2, \ldots, M\}$ for total arrival rate to node m. Once the value of $\{\lambda_{am}\}$ are obtained, the traffic intensity (expected utilization) of node m, ρ_m, can be calculated by

$$\rho_m = \lambda_{am} \tau_m / M_m \tag{2}$$

where M_m is the number of machines in node m.

Traffic Variability Equation

Now, a second-order characterization of the interarrival times is given in terms of SCVs. Denote C_{an}^2 as the interarrival time SCV of node n. The approach of Segal and Whitt (1988) is closely followed. Their approach approximates the SCVs of merging and splitting point processes with considerations of the deterministic feedback routing and the interference among multiple classes. Then a set of linear traffic variability equations is derived as follows:

$$C_{an}^2 = a_n + \sum_{m=1}^{M} C_{am}^2 b_{mn}^2, \; 1 \le n \le M \tag{3}$$

where

$$a_n = 1 + \varpi_n \left\{ \frac{\lambda_e C_e^2}{\lambda_{an}} \delta_n - 1 + \sum_{m=1}^{M} \left(\frac{\lambda_{mn}}{\lambda_{an}} \right) \left[q_{mn} \rho_m^2 \chi_m + (1 - q_{mn})^2 C_{em}^2 \right] \right\}$$

$$b_{mn} = \varpi_n \left(\frac{\lambda_{mn}}{\lambda_{an}} \right) q_{mn} \left[\left(1 - \rho_m^2\right) + \left(1 - q_{mn}\right) \right],$$

$$\varpi_n = \left[1 + 4(1 - \rho_n)^2 (\nu_n - 1) \right]^{-1},$$

$$\nu_n = \left[\sum_{m=0}^{M} \left(\lambda_{mn} / \lambda_{an} \right)^2 \right]^{-1},$$

$$\chi_m = 1 + \frac{\left(\max\left\{ C_{sm}^2, 0.2 \right\} - 1 \right)}{\sqrt{M_m}}, \text{ and}$$

$$C_{em}^2 = \sum_{i=1}^{I} C_i^2 \left(\sum_{k=1}^{S_i} \lambda_i \mathbf{1}\{(i,k): m_{ik} = m\} \Big/ \sum_{i=1}^{I} \sum_{k=1}^{S_i} \lambda_i \mathbf{1}\{(i,k): m_{ik} = m\} \right)$$

Note that the above formulas describe the approximate relationship among the interarrival time SCVs of all nodes, which correspond to the movement of jobs throughout the network. The term a_n captures the SCV of the external arrival processes and b_{mn} approximates the SCV due to internal transitions. The coefficients ϖ_n and v_n specify the flow-merging effects from upstream flows, C_{em}^2 is a weighted average of external arrival SCV by expected number of visits, and χ_m is used to specify the departure operation. The study of Whitt (1983) suggests empirically that χ_m be maintained above 0.2 to achieve good approximation when the service time SCV is low.

Frequently considered system performance metrics of fabs include desired system output rate, upper bounds of inter-output time variability, mean cycle time, and cycle time variance. Mean waiting time and queue length (WIP level) are commonly used metrics in local performance control. Queuing theory has provided steady-state performance equations (Whitt 1993) relating the arrival and service parameters (λ_{am}, C_{am}^2, τ_m, C_{sm}^2) at individual nodes with the aforementioned global and local performance metrics. Combining these equations with Eqs. (1), (2), and (3), one can then establish functions $\underline{Y} = \underline{F}(\underline{\alpha}, \underline{\theta}, \mathbf{Q})$ and $\underline{y} = \underline{G}(\underline{\alpha}, \underline{\theta}, \mathbf{Q})$ for the OQN model. Specific cases and the solution method will be addressed in the next section.

Backward Queuing Network Analysis

Exploiting the results of decomposition approximation in the previous section, a backward queuing network analysis (BQNA) is now designed to solve $\underline{\alpha}$ from the global performance equation $\underline{Y} = \underline{F}(\underline{\alpha}, \underline{\theta}, \mathbf{Q})$. In specific, BQNA derives wafer release parameters $\underline{\alpha}$, (λ_e, C_e^2), and the desired means and SCVs of interarrival times for individual nodes, (λ_{am}, C_{am}^2), from the specified process routing, nodal characteristics, and production goals of the system. The resultant $\underline{\alpha}$ is then substituted into the local performance function $\underline{y} = \underline{G}(\underline{\alpha}, \underline{\theta}, \mathbf{Q})$ to obtain managerially tangible local PFC references in terms of WIP or cycle time for achieving the desired overall production goals of a reentrant line.

BQNA with Target Cycle Time and Desired Output Rate

The development of the BQNA method will first be focused on using a target mean total cycle time, T, and a desired system output rate, d, as the overall production goals. That is $\underline{Y} = (T, d)$. The application of BQNA to the translation of other production goals will then be discussed.

Required Flow Control Target on Arrival Rate

When a network is in a steady state, the required overall external arrival rate at the entry node should be equal to the desired output rate, that is,

$$\lambda_e = d \tag{4}$$

Let $(\gamma_1, \gamma_2, \ldots, \gamma_I)$ be the job class mix ratio with $\gamma_1 + \gamma_2 + \ldots + \gamma_I = 1$. Then the external arrival rate of class i to the entry node can also be deaggregated as

$$\lambda_i = \gamma_i \times \lambda_e \quad \text{for } i = 1, 2, \ldots, I$$

Furthermore, with the given output rate goal, the overall arrival rates at individual nodes, λ_{am}, $m = 1, 2, \ldots, M$, can be calculated by substituting Eq. (4) into Eq. (1) and solving the set of M linear equations. The expected utilization of node m, ρ_m, is then also available by Eq. (2).

Required Flow Control Bounds on Interarrival Time SCV

Based on queuing theory, mean waiting time can be approximated as a function of variability parameters, mean service time, and loading intensity. This fact is used to establish the relationship between the mean cycle time target and system variability. Assume that all service nodes are highly utilized, which is frequently true in a fab. With heavy-traffic limit theorems, the expected waiting time of node m modeled as a GI/G/m queue, $E[W_m]$, can be approximated as follows (Whitt 1983, 1993):

$$E[W_m] = \frac{C_{am}^2 + C_{sm}^2}{2} E[W_m(\mathrm{M}/\mathrm{M}/M_m)] \tag{5}$$

where $E[W_m(\mathrm{M/M}/M_m)]$ is the expected waiting time for an M/M/M_m queue, defined as follows (Buzacott and Shanthikumar 1993):

$$E[W_m(\mathrm{M/M}/M_m)] = \frac{\tau_m (M_m \rho_m)^{M_m} \pi_m(0)}{M_m (1-\rho_m)^2 M_m!}$$

with

$$\pi_m(0) = \left\{ \sum_{t=0}^{M_m-1} \frac{(M_m\rho_m)^t}{t!} + \frac{(M_m\rho_m)^{M_m}}{(1-\rho_m)M_m!} \right\}^{-1}$$

Then the corresponding mean cycle (or sojourn) time can be expressed as follows:

$$T_m = \tau_m + E[W_m] \tag{6}$$

The expected total cycle time for a job to go through the network is a function of the average number of visits per job and average processing and waiting times of individual nodes in the network. The average number of visits per job to node m is as follows:

$$V_m = \lambda_{am}/\lambda_e, \text{ for } m = 1, 2, \ldots, M \tag{7}$$

The expected total cycle time is thus

$$T = \sum_{m=1}^{M} V_m T_m = \sum_{m=1}^{M} V_m (\tau_m + E[W_m]) \tag{8}$$

By using Eq. (5), Eq. (8) can be rewritten to obtain the mean cycle time as follows:

$$\begin{aligned} T = &\sum_{m=1}^{M} V_m \left\{ \tau_m + E\left[W_m(\mathrm{M}/\mathrm{M}/M_m)\right] C_{sm}^2/2 \right\} \\ &+ \sum_{m=1}^{M} V_m E\left[W_m(\mathrm{M}/\mathrm{M}/M_m)\right] C_{am}^2/2 \end{aligned} \tag{9}$$

Now consider the relationship among the SCVs of nodal arrival processes. Examined first is the loading of wafers into a line, that is, the wafer release process, which is fully controlled by a fab manager. It is straightforward that the wafer release rate must be no less than the desired output rate for each part type. In the manufacturing practice of fabs, there is usually an upper bound set to the lot inter-output time SCV for the smoothness of delivery. As fab operations only add variability, the inter-output time SCV imposes an upper bound on the interarrival time SCVs of individual part types. It is a common fab practice to have a fair wafer release among different part types, that is, to have equal SCVs for arrival processes of all job classes, $C_i^2 = C_j^2$ for $i \neq j$. Note that the resultant inter-output time SCVs among different types are generally different because of their different routings; in a loose sense, the longer the route, the larger the inter-output time SCV.

Under the fair wafer release, it can then be easily derived from the aggregation procedure that $C_e^2 = C_i^2$, $\forall\, i$, and that $C_{em}^2 = C_e^2$, $\forall\, m$. These results are then substituted into the traffic variability equation. Then Eq. (3) can be arranged as follows:

$$C_{an}^2 = a_n + \beta_n C_e^2 + \sum_{m=1}^{M} C_{am}^2 b_{mn} \tag{10}$$

where

$$\alpha = 1 - \varpi_n \sum_{m=1}^{M} (\lambda_{mn}/\lambda_{an}) q_{mn} \rho_m^2 \chi_m, \text{ and}$$

$$\beta_n = \varpi_n (\lambda_e/\lambda_{an}) \delta_n + \varpi_n \sum_{m=1}^{M} (\lambda_{mn}/\lambda_{an})(1-q_{mn})^2,$$

for $n = 1, 2, \ldots, M$

Equations (9) and (10) constitute a set of M+1 linear equations for M+1 variables. They can be expressed in a matrix form as follows:

$$\begin{bmatrix} b_{11}-1 & b_{21} & \cdots & b_{M1} & \beta_1 \\ b_{12} & b_{22}-1 & \cdots & b_{M2} & \beta_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{1M} & b_{2M} & \cdots & b_{MM} & \beta_M \\ b_{1,M+1} & b_{2,M+1} & \cdots & b_{M,M+1} & 0 \end{bmatrix} \begin{bmatrix} C_{a1}^2 \\ C_{a2}^2 \\ \vdots \\ C_{aM}^2 \\ C_e^2 \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_M \\ \alpha_{M+1}-T \end{bmatrix} = \mathbf{0} \tag{11}$$

where

$$\alpha_{M+1} = \sum_{m=1}^{M} V_m \left\{ \tau_m + E[W_m(\mathrm{M}/\mathrm{M}/M_m)] C_{sm}^2/2 \right\},$$

and

$$b_{m,M+1} = V_m E[W_m(\mathrm{M/M/}M_m)]/2,$$
for $m = 1, 2, \ldots, M$

Parameters C_e^2 and $\{C_{am}^2\}$ can be obtained by solving Eq. (11). These nodal SCVs can then be used to develop wafer release control and other local PFC applications, which will be described at the end of this section.

BQNA with Other Production Goals

Fab production goals vary with market changes and business objectives. For example, when the market is booming, a fab may want to maximize the throughput while setting an upper bound on cycle times. In contrast, during low business periods a fab may need to compete on short cycle times and ontime delivery for getting more orders. In this case, a fab needs to reduce mean cycle times and the variance of inter-output times to maintain a steady output. Therefore, BQNA should have the flexibility to support different system production goals to be applicable to real fabs.

In fact, BQNA can be applied to a number of system performance metrics. The basic idea is as follows. Recall that BQNA takes a reverse view of the decomposition method, deriving for an aggregate OQN the rate and SCV bound of the arrival process to a node from the given rates and SCV bounds of output (departure) and service processes of the node. It is intuitively clear that if the selected system performance metrics are functions of rates and SCVs of machine service and input processes, BQNA may then possibly be applicable. A closer examination of the rate and variability equation sets of BQNA [Eqs. (1), (4), and (11)] easily reveals that there are two degrees of freedom in the equations. That is, a two-item system production goal suffices to determine the solution.

Let d be the output rate, T be the bound on average total cycle time, and C_d^2 and C_T^2 be the bounds of inter-output time SCV and total cycle time SCV, respectively. Two-item system production goals frequently considered by semiconductor fabs include (d, T), (d, C_d^2), and (T, C_T^2). Each item can be approximated as a function of the first two statistical moments of service and input processes. For illustration, consider (d, C_d^2) as the required overall production goals. In a stable line, the desired external arrival rate should be equal to the desired output rate, that is, $\lambda_e = d$. Target arrival rates at individual nodes, λ_{am}, $m = 1, 2, \ldots, M$, can then be computed by using Eq. (1). Because the output process of an individual node is affected by the characteristics of both the arrival and the service processes, the inter-output time SCV of output (departure) node, C_{dM}^2, can be expressed as follows (Whitt 1983):

$$C_{dM}^2 = 1 + (1-\rho_M^2)(C_{aM}^2 - 1) + \rho_M^2\left(\max\{C_{sM}^2, 0.2\} - 1\right)/\sqrt{M_M} \tag{12}$$

The relationship between C_{dM}^2 and C_d^2 satisfies (Segal and Whitt 1988)

$$C_d^2 = q_{M0}C_{dM}^2 + (1-q_{M0})q_{M0}C_{aM}^2 + (1-q_{M0})^2 C_{eM}^2 \tag{13}$$

Note that q_{M0} is the ratio of jobs from node M leaving the system. Combining Eqs. (3), (12), and (13), the variability equation set constitutes a set of $M+1$ linear equations for $M+1$ variables as

$$\begin{bmatrix} b_{11}-1 & b_{21} & \cdots & b_{M-1,1} & b_{M1} & \beta_1 \\ b_{12} & b_{22}-1 & \cdots & b_{M-1,2} & b_{M2} & \beta_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{1M} & b_{2M} & \cdots & b_{M-1,M} & b_{MM}-1 & \beta_M \\ 0 & 0 & \cdots & 0 & -1 & \beta_{M+1} \end{bmatrix} \begin{bmatrix} C_{a1}^2 \\ C_{a2}^2 \\ \vdots \\ C_{aM}^2 \\ C_e^2 \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_M \\ \alpha_{M+1} \end{bmatrix} = \mathbf{0}$$

where

$$\alpha_{M+1} = \frac{C_d^2 - q_{M0}\rho_M^2\left[1 + \left(\max\{C_{sM}^2, 0.2\} - 1\right)/\sqrt{M_M}\right]}{q_{M0}\left[(1-q_{M0}) + (1-\rho_M^2)\right]}$$

and

$$\beta_{M+1} = -(1-q_{M0})^2 / [q_{M0}(2-q_{M0}-\rho_M^2)]$$

Parameters C_e^2 and $\{C_{am}^2\}$ can be obtained by solving this set of linear equations.

BQNA Results for PFC Application

One obvious way to apply BQNA results to PFC is to control the job arrival process so that its target mean rate is met and its interarrival time SCV is no greater than the value calculated from BQNA. Such a control can easily be implemented by setting the inter-wafer lot release times accordingly. When such a release control is achieved at the entry node, the overall system performance should then meet the given production goals. Furthermore, various service disciplines are usually designed for individual nodes to result in better output performances than simply using FCFS in manufacturing practice. The performance of individual nodes derived by BQNA may therefore serve as guide for real applications.

In manufacturing practice, various PFC schemes (Hopp and Roof 1998; Leachman 1994; Li, Tang, Collins 1996; Wu et al. 1998) have been proposed. Many of them use target levels and bounds on WIP and/or cycle times as the measures to control because such data is easily accessible from the shop floor and is managerially tangible to operators. Although the flow control requirements directly derived by BQNA are mean rates and SCVs of input processes for each machine group, they can be utilized to derive other measurable requirements for cycle time and WIP. For example, the calculation of expected waiting time of node m can be calculated using Eq. (5). The corresponding expected WIP of node m, $E[L_m]$, can be derived by using Little's formula, which relates the average time spent to the average number of jobs in the node, as follows:

$$E[L_m] = \lambda_{am}(\tau_m + E[W_m]) \tag{14}$$

Furthermore, the mean WIP of the entire network, L, is then

$$\begin{aligned} L &= E[L_1] + E[L_2] + \cdots + E[L_M], \\ &= \sum_{m=1}^{M} \lambda_{am}(\tau_m + E[W_m]), \\ &= \lambda_e \sum_{m=1}^{M} V_m(\tau_m + E[W_m]), \\ &= \lambda_e \times T. \end{aligned}$$

The waiting time and queue length variances may be approximately calculated as defined by Whitt (1993). BQNA may therefore provide these existing PFC schemes with effective calculation of control parameters and facilitate their realization.

Exploiting the aggregated production flow results, the per-part cycle times and/or WIP levels can be calculated to provide guidelines for achieving differentiated service levels among part types. Note that under the FCFS discipline, the average waiting times at a node for wafers of different part types are the same. So, the target mean cycle time for a job of class i is as follows:

$$E[T_i] = \sum_{k=1}^{S_i} (\tau_{ik} + E[W_{ik}])$$

where $E[W_{ik}]$ is the expected waiting time at service node m_{ik} computed by using Eq. (5), and an upper bound of cycle time variance for class i can be calculated by the following:

$$Var[T_i] = \sum_{k=1}^{S_i} \tau_{ik}^2 C_{ik}^2 + \sum_{k=1}^{S_i} Var[W_{ik}]$$

The target average total WIP for a job of class i can be obtained by Little's formula as follows:

$$E[L_i] = \lambda_i \times E[T_i]$$

These per-part type requirements can serve as part of the detailed flow control requirements at each step or machine group.

Numerical and Simulation Results

To validate the BQNA, its numerical results are compared with simulation results on two example fab models. For simplicity, these reentrant manufacturing models are dubbed FAB1 and FAB2. FAB1 is an aggregated single-product fab model while FAB2 represents a 10-product fab model for investigation of multiple product-type effect. A C language based discrete event dynamic system simulator (Hsieh et al. 1998) is used for simulation. The validation first ap-

plies the BQNA procedure to derive all the means and SCVs of interarrival times of a reentrant OQN. Parameters (λ_{am}, C_{am}^2, τ_m, C_{sm}^2) at each node can then be utilized to estimate the local performance requirements by steady-state performance equations. To validate the BQNA results, the derived release parameters $\underline{\alpha} = (\lambda_e, C_e^2)$ together with service parameters $\underline{\theta}$ and routing parameters $\mathbf{Q}$, are set as inputs to the simulation model. FCFS is the service discipline at each node. Simulation outputs include local PFC parameters such as WIP and cycle time at each node and the system performance measures such as total cycle time and desired output rate.

Single-Product Fab Model

FAB1 is adopted from an aggregated full-scale production line previously studied by Lu, Ramaswamy, and Kumar (1994). This model consists of 60 production stages, 12 different MGs, and a total of 40 machines. There is only one part type. Its sequence of processing steps and the MG used by each step are given in *Figure 4*. Machines are subject to random failures, and each failed machine requires a random repair time. All the times to failure, times to repair, and processing times have exponential distributions with various values of mean time to failure (MTTF), mean time to repair (MTTR), and mean processing time (MPT). *Table 1* lists the basic information of the model.

Overall system production goals of this example are set as statistical performance measures frequently considered in fabs, as follows:

1. desired system output rate $d = 0.52$ lots/hr, and
2. target mean cycle time $T \leq 383.25$ hours,

where the given system output rate equals the wafer release rate of Lu, Ramaswamy, and Kumar and the specified target cycle time bound is obtained from an independent simulation result with Poisson-distributed inter-lot release time. The given output rate of the line demands an average machine loading of more than 90%, and MGs 9 and 11 are two capacity bottlenecks with 96.0% loading intensity.

Results from applying BQNA to this example are obtained in 0.55 msec of CPU time on a 300 MHz PC. The derived requirements are means and SCVs of arrival processes to individual MGs. These derived results are then used to calculate the nodal performance estimates. In particular, the mean cycle time and mean WIP can be calculated for each machine group by using Eqs. (6) and (14). The standard deviations of cycle times and WIPs are obtained by the refined formulas of Whitt (1993). To validate BQNA results, the derived external arrival parameters (λ_e, C_e^2) = (0.52, 0.998) are then input to the discrete simulation model.

In the simulation study, there are six simulation runs. Each simulation run begins with an empty line and ends at the departure of the 114,000th lot out of

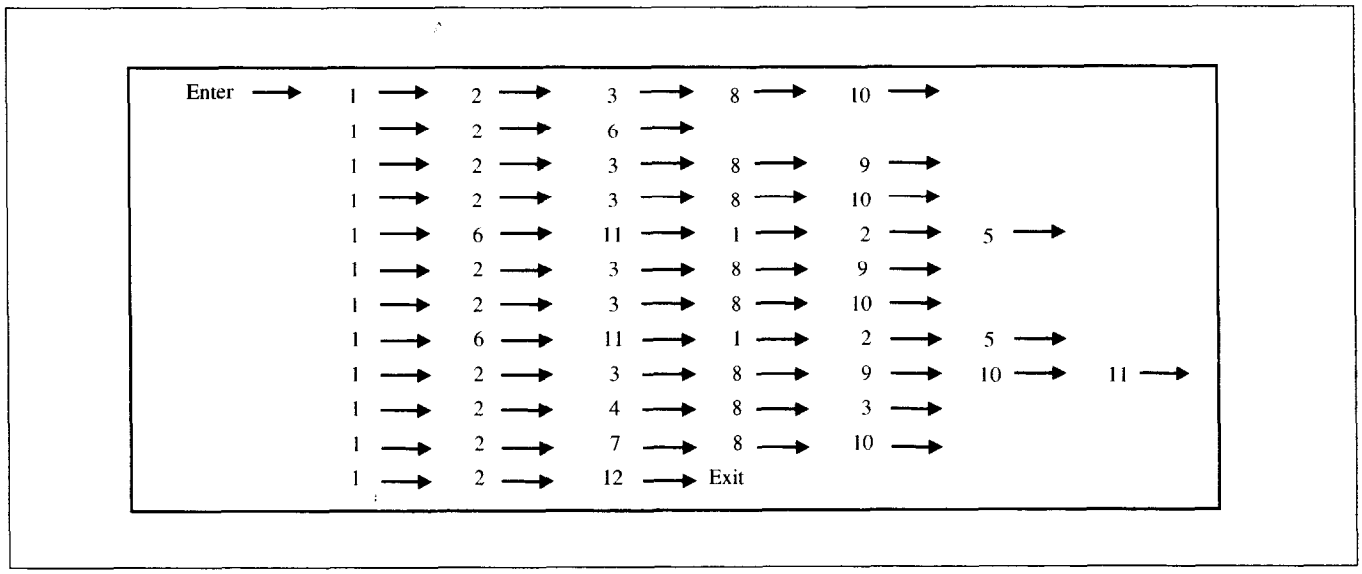

Figure 4
Process Flow of FAB1

Table 1
Machine Group Data of FAB1

MG	No. of Machines	No. of Visits	MPT (hr/lot)	MTTF (hr)	MTTR (hr)	Utilization %*
1	4	14	0.5	150	5	94.2
2	3	12	0.375	200	9	82.3
3	10	7	2.5	200	5	93.4
4	1	1	1.8	200	1	94.1
5	1	2	0.9	200	1	94.1
6	2	3	1.2	200	1	94.1
7	1	1	1.8	200	1	94.1
8	4	8	0.8	150	5	86.4
9	1	3	0.6	200	5	96.0
10	9	5	3.0	130	5	90.3
11	2	3	1.2	200	5	96.0
12	2	1	2.5	200	5	67.4

*Utilization $\% = \left[\frac{0.52(No.\ of\ Visits)(MPT)}{No.\ of\ Machines} + \frac{MTTR}{MTTF + MTTR}\right] \times 100$

the line. Because the average monthly throughput of FAB1 is 374.4 lots/month, a simulation run approximately corresponds to 25 years of fab operation. The simulation of the first 14,000 lot departures, about three years, serves as a warm-up period, which is long as compared to an average cycle time of 15.9 days. The succeeding simulation is partitioned into 20 intervals for collecting statistics, including the mean and variance of cycle time and WIP. Statistics of individual intervals are then averaged over the 120 intervals of the six simulation runs.

Figures 5 and *6* contrast node-level performances in cycle time and WIP derived by BQNA and simulation, respectively. It is observed that the differences in mean cycle time and mean WIP are both below 3%. Averaged relative differences in standard deviations of cycle time and standard deviations of WIP are about 10–20%. Although the estimates of standard deviations of WIP at the MG level are not exceptionally accurate, the maximum difference is less than 1.5 lots in this case. These results confirm that the nodal arrival parameters calculated by using BQNA results can be used to obtain the internal PFC required parameters such as mean and variance of nodal cycle time and WIP. *Table 2* shows close system-level performances in mean cycle time and throughput rate by simulation to the specified production goals. This clearly implies that the reentrant line achieves the desired system output rate and target mean cycle time when arrival parameters are controlled to match the performance requirements calculated by BQNA.

Ten-Product Fab Model

FAB2 closely follows the model of Bitran and Tirupati (1988), which is based on real-life data and models the production facility of a semiconductor company. This model is represented by 13 different machine groups with jobs representing 10 product families, and the job class mix ratio γ_i is set as 0.1 for $i = 1, \ldots, 10$. The last process step of each product finishes at MG 13. The sequence of processing steps and the MG used by each step are given in *Table 3*. *Table 4* lists the basic information of the model. Note that different machine groups have different processing time distributions in this FAB2 model.

The required system production goals of this model are set as follows:

1. desired system output rate $d = 1.0$ lots/hr, and
2. target mean cycle time $T \leq 40.64$ hours,

where the output rate requirement equals the release rate of Bitran and Tirupati and the mean cycle time target is obtained from an independent simulation of the line with $C_i^2 = 1.5$ for each product. The system output rate of the line demands an average machine loading of more than 80%, and MG 9 is the capacity bottleneck with 94.0% loading intensity.

Results of applying BQNA to this example are obtained in 0.55 msec of CPU time on a 300 MHz PC. The derived requirements of individual MGs will be used to calculate the nodal performance estimates BQNA-derived aggregated external arrival parameters $(\lambda_e, C_e^2) = (1.0, 1.5)$ are then used as the lo

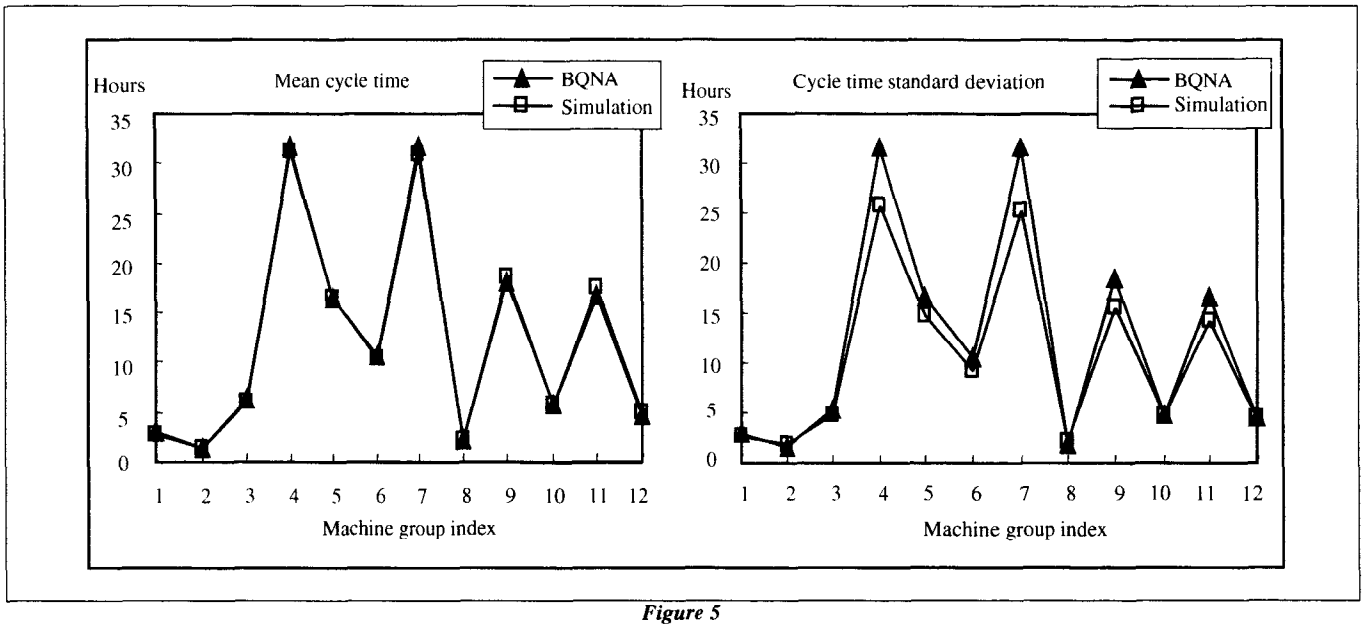

Figure 5
Cycle Time Mean and Standard Deviation of FAB1

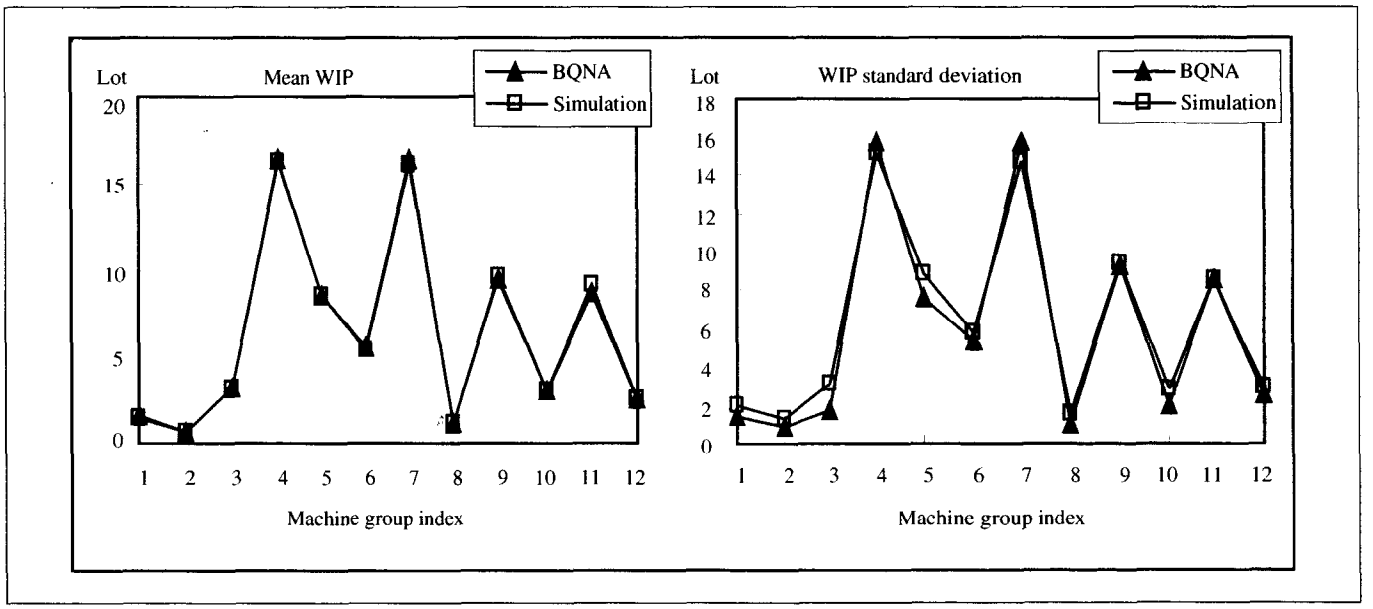

Figure 6
WIP (in Lots) Mean and Standard Deviation of FAB1

release parameters in the discrete simulation model for validation.

In the simulation study, there are six simulation runs. Each simulation run begins with an empty line and ends at the departure of the 30,000th lot out of the line. Because the average monthly throughput of FAB2 is 720 lots/month, such a simulation run approximately corresponds to 3.5 years of fab operation. The simulation of the first 5,000 lot departures, corresponding to about seven months of fab operation, serves as a warm-up period, which is long enough as compared to an average cycle time of 1.7 days. The succeeding simulation is partitioned into five intervals for collecting statistics on the mean and variance of cycle time. About 5,000 lots of each product are collected in each interval, and in total about 25,000 lots of each product are collected in each simulation run. Statistics for individual intervals are then

Table 2
System-Level Performance Comparisons of FAB1

	Mean Cycle Time	Throughput Rate
System goals	383.25	0.52
Simulation	382.0595	0.5195
Relative error %	−0.3107%	−0.0962%

Table 3
Product Routing Sequence of FAB2

Product	Number of Operations	Process Routing Sequence
1	8	1, 2, 4, 2, 9, 10, 11, 13
2	9	1, 2, 5, 2, 8, 9, 10, 11, 13
3	9	1, 2, 6, 4, 2, 9, 12, 11, 13
4	9	1, 2, 7, 4, 2, 9, 10, 11, 13
5	8	1, 2, 4, 12, 2, 9, 2, 13
6	8	1, 2, 5, 12, 2, 9, 7, 13
7	8	1, 2, 6, 12, 2, 8, 2, 13
8	12	1, 2, 3, 7, 4, 12, 2, 8, 6, 9, 2, 13
9	13	1, 2, 3, 5, 4, 6, 12, 2, 8, 2, 10, 6, 13
10	13	1, 2, 3, 6, 2, 4, 12, 7, 2, 9, 11, 5, 13

averaged over the 30 intervals of the six simulation runs.

Figures 7 and *8* provide the performance comparisons of the cycle time mean and standard deviation at the nodal level between BQNA and simulation. These results suggest that local cycle time performance also can be accurately estimated from calculated nodal arrival parameters in the 10-product system. *Table 5* shows that simulated system-level performances in mean cycle time and throughput rate are close to the specified production goals. It also suggests that the system production goals can be achieved in this 10-product fab model when lot arrivals follow the interarrival times requirement derived by BQNA.

Summary

This paper is aimed at getting local performance bounds or targets for individual nodes. BQNA is a systematic and computationally efficient way to translate system production goals into local performance bounds. In BQNA, FCFS is adopted as the PFC scheme for all nodes but wafer release control at the entry node. In the simulation, the wafer release is controlled to meet the derived external parameters, and FCFS is the service discipline of each node. It is observed that in the application of BQNA to the above two fab models, the calculated performance measures of WIP and cycle time are mostly within

Table 4
Machine Group Data of FAB2

MG	No. of Machines	Service Time Distribution	MPT (hr/lot)	Utilization %*
1	1	Uniform	0.78	78.0
2	1	Uniform	0.25	62.5
3	1	Erlang Order 2	1.667	50.0
4	1	Exponential	1.057	74.0
5	1	Erlang Order 3	1.700	68.0
6	1	Erlang Order 4	0.950	57.0
7	1	Uniform	1.775	71.0
8	1	Uniform	1.875	75.0
9	1	Erlang Order 2	1.175	94.0
10	1	Erlang Order 3	1.800	72.0
11	1	Exponential	1.430	71.5
12	1	Erlang Order 4	0.750	52.5
13	1	Uniform	0.870	87.0

$$^{*}\text{Utilization } \% = \sum_{i=1}^{10}\left[\frac{0.1(No.\ of\ Visits)(MPT)}{No.\ of\ Machines}\right]\times 100$$

Table 5
System-Level Performance Comparisons of FAB2

	Mean Cycle Time	Throughput Rate
System goals	40.64	1.0
Simulation	38.0203	1.0014
Relative error %	−6.446%	+0.14%

95% of accuracy at both node and system levels as compared to those of the simulation results. In particular, how well BQNA works depends largely on how well the adopted decomposition approximation method captures the features of the manufacturing processes.

It is also observed that it takes less than one millisecond of CPU time to apply BQNA to the abovementioned reentrant lines. Specifically, the main computation load of BQNA lies in solving the two sets of linear equations to obtain required arrival parameters of individual nodes. After aggregation, the dimension of the linear equations depends only on the number of different MGs but not on the total number of machines, the number of production steps, or the number of part types. There are many existing and efficient algorithms for solving such linear equations. Such computation efficiency warrants the application of BQNA to real-sized fabs and to fast recalculations with respect to changes in system capacity and production goals.

Note that the local performance bounds or targets obtained by application of BQNA for PFC at indi-

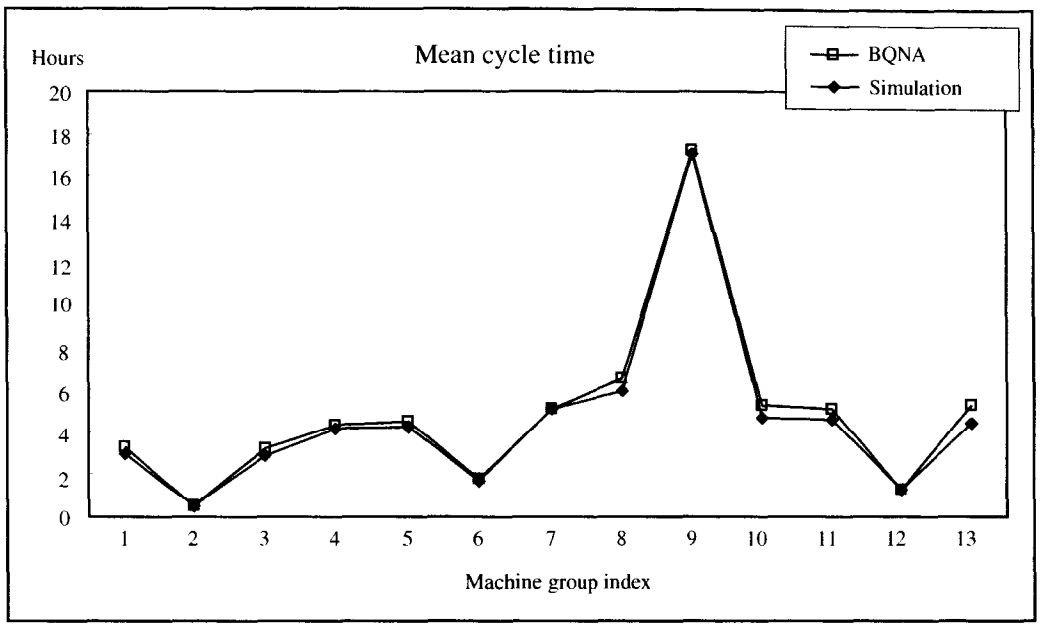

Figure 7
Mean Cycle Time of MGs in FAB2

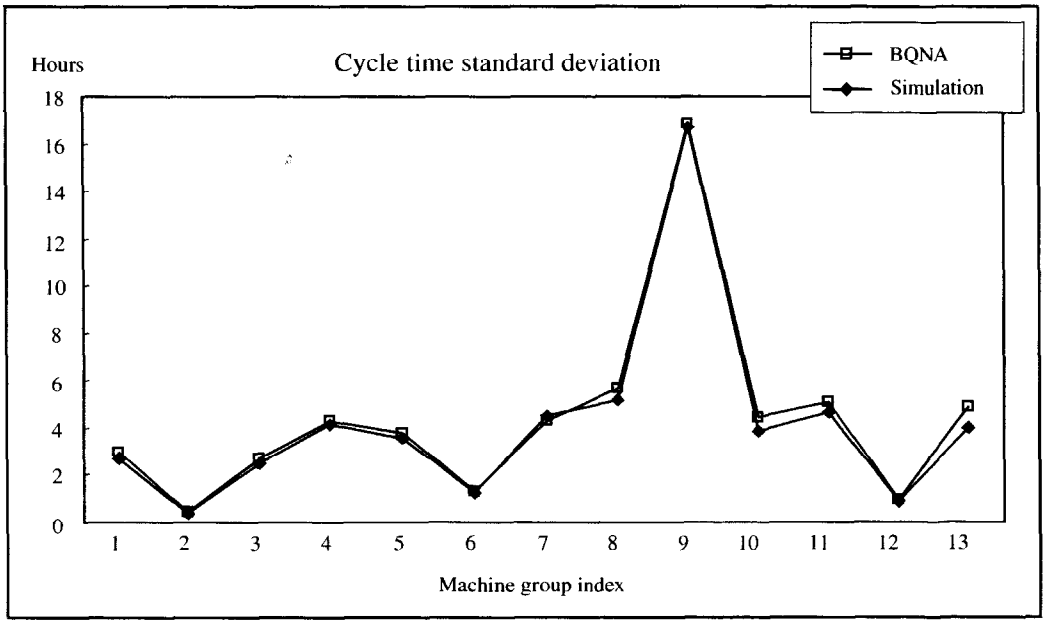

Figure 8
Standard Deviation of Cycle Times for FAB2 MGs

vidual nodes are the minimum levels of performance to be achieved by PFC. The simulations show that FCFS at each internal node in conjunction with a wafer release control at the entry node is just one possible PFC to achieve the desired performance. Because FCFS is a very simple PFC scheme, there can be many alternative local PFC or dispatching schemes that perform better within the performance bounds at individual nodes and hence achieve more beyond the overall fab goals than using FCFS. Design of such PFC or dispatching schemes, however, is not the subject of this paper.

Conclusions

In this paper, a production goal translation method has been developed for translating desired system performance into local control parameters in terms of mean rates and SCVs of input processes for reentrant manufacturing systems, specifically VLSI wafer fabrication. A multiple part type reentrant manufacturing system with failure-prone machines was first modeled as a single aggregated type, failure-free OQN. In conjunction with this model, the decomposition-based method was adopted to approximate the relationship among the external input, machine service, and output processes as two sets of linear equations. Taking a reverse viewpoint of relations among variables in the decomposition approximation, BQNA was developed to derive local control parameters from two sets of linear equations with a two-item system production goal. Results of BQNA could further be utilized to obtain the managerially tangible PFC references of both the aggregate and per-part type nodal cycle time and WIP level. The overall performance of a fab will attain the given production goals even under a FCFS dispatching rule at each node, if the job release is controlled to meet the translated control parameters. BQNA was applied to two fab models and each application took less than one millisecond of CPU time. Numerical results demonstrated that the differences between BQNA and simulation are mostly within 95% accuracy at both node and system levels. Both the accuracy and computing efficiency of BQNA support its potential for real-sized applications.

Acknowledgments

This work was supported in part by the National Science Council of the Republic of China under grants NSC 86-2622-E-002-025R, NSC 87-2212-E-002-023, NSC88-2213-E-002-065, and NSC89-2212-E-002-040. The authors would also like to thank Mr. Thomas Chen, Fab3 Manager of Manufacturing Department, Taiwan Semiconductor Manufacturing Co., and Dr. Da-Yin Liao for their valuable inputs on fab flow control practice, future requirements, and models of their manufacturing execution system.

Appendix

Machine failures result in service unavailability and interruption to customers, which in turn leads to a longer average and a larger variation in the time needed for completing a service step than when machines are failure free. To study the long-run behavior of an equilibrium network, there has been a standard procedure in the literature to convert a failure-prone OQN into a failure-free one by modifying the processing time distribution. The basic idea is to incorporate machine failure effects into the general service time distribution of each node and form a statistically equivalent failure-free node. The conversion procedure is given as follows.

Define for a service node m

τ_m: original mean processing time;
C_{sm}^2: original processing time SCV;
τ_{rm}: mean down time or mean time to repair;
C_{rm}^2: downtime SCV or repair time SCV;
τ_{um}: mean up time or mean time to failure;
$\bar{\tau}_m$: modified mean processing time;
$\bar{C}_{sm}^2$: modified service time SCV.

Then the probability that a job encounters a breakdown while being processed can be calculated as follows:

$$p_m = \tau_m / \tau_{um}$$

The average of modified service time at node m is then

$$\bar{\tau}_m = p_m(\tau_m + \tau_{rm}) + (1 - p_m)\tau_m$$

$$= \tau_m + p_m \tau_{rm},$$

and the corresponding modified SCV can be calculated as follows:

$$\overline{C}_{sm}^2 = \left\{ p_m \left[\left(C_{sm}^2 + 1 \right) \tau_m^2 + \left(C_{rm}^2 + 1 \right) \tau_{rm}^2 + 2\tau_m \tau_{rm} \right] + \left(1 - p_m \right) \left[\left(C_{sm}^2 + 1 \right) \tau_m^2 \right] \right\} \Big/ \overline{\tau}_m^2 - 1$$

$$= \left\{ \left(C_{sm}^2 + 1 \right) \tau_m^2 + p_m \left[C_{rm}^2 \tau_{rm}^2 + 2\tau_{rm} \tau_m + \tau_{rm}^2 \right] \right\} \Big/ \overline{\tau}_m - 1$$

References

Bitran, G.R. and Tirupati, D. (1988). "Multiproduct queueing networks with deterministic routing: decomposition approach and the notion of interference." *Mgmt. Science* (v34), pp75-100.

Böbel, F. and Halmel, S. (1998). "The SIEMENS productivity program methodology and advances." Lecture notes of IEEE/SEMI Advanced Semiconductor Mfg. Conf. and Workshop, Boston, MA.

Bramson, M. (1994). "Instability of FIFO queueing network." *Annals of Applied Probability* (v4), pp414-431.

Burman, D.Y.; Gurrola-Gal, F.J.; Nozari, A.; Sathaye, S.; and Sitarik, J.P. (1986). "Performance analysis techniques for IC manufacturing lines." *AT&T Bell Labs Technical Journal* (v65), pp45-56.

Buzacott, J.A. and Yao, D.D. (1986). "On queueing network models of flexible manufacturing systems." *Queueing Systems* (v1), pp5-27.

Buzacott, J.A. and Shanthikumar, J.G. (1992). "Design of manufacturing systems using queueing models." *Queueing Systems* (v12), pp135-213.

Buzacott, J.A. and Shanthikumar, J.G. (1993). *Stochastic Models of Manufacturing Systems* Englewood Cliffs, NJ: Prentice-Hall.

Chen, H.; Harrison, J.M.; Mandelbaum, A.; Van Ackere, A.; and Wein, L.M. (1988). "Empirical evaluation of a queueing network model for semiconductor wafer fabrication." *Operations Research* (v36), pp202-215.

Connors, D.P.; Feigin, G.E.; and Yao, D.D. (1996). "A queueing network model for semiconductor manufacturing." *IEEE Trans. on Semiconductor Mfg.* (v9), pp412-427.

Dayhoff, J.E. and Atherton, R.W. (1984). "Simulation of VLSI manufacturing areas." *VLSI Design*, pp84-92.

Dayhoff, J.E. and Atherton, R.W. (1986). "Signature analysis: simulation of inventory, cycle time and throughput tradeoffs in wafer fabrication." *IEEE Trans. on CHMT* (v9), pp498-507.

Gelenbe, E. and Pujolle, G. (1987). *Introduction to Queueing Networks.* New York: John Wiley & Sons.

Graves, S.C.; Meal, H.C.; Stefek, D.; and Zeghmi, A.H. (1983). "Scheduling of re-entrant flow shops." *Journal of Operations Mgmt.* (v3), pp197-207.

Hasenbein, J.J. (1997). "Necessary conditions for global stability of multiclass queueing networks." *Operations Research Letters* (v21), pp87-94.

Hopp, W.J. and Roof, M.L. (1998). "Setting WIP levels with statistical throughput control (STC) in CONWIP production lines." *Int'l Journal of Production Research* (v36), pp867-882.

Hsieh, B.-W.; Hu, M.-D.; Chen, C.-H.; and Chang, S.-C. (1998). "DEDS-discrete event dynamic system simulator." Technical Report. Taipei, Taiwan: National Taiwan Univ.

Kouvelis, P. and Tirupati, D. (1991). "Approximate performance modeling and decision making for manufacturing systems: a queueing network optimization framework." *Journal of Intelligent Mfg.* (v2), pp107-134.

Kumar, P.R. (1993). "Re-entrant lines." *Queueing Systems* (v13), pp87-110.

Kuroda, M. and Kawada, A. (1995). "Adaptive input control for job-shop type production systems with varying demands using inverse queueing network analysis." *Int'l Journal of Production Economics* (v41), pp217-225.

Leachman, R. (1994). "Production planning and scheduling practice across the semiconductor industry." Technical Report ESRC 94-29/CSM-18. Berkeley, CA: Univ. of California-Berkeley.

Li, S.; Tang, T.; and Collins, D.W. (1996). "Minimum inventory variability schedule with application in semiconductor fabrication." *IEEE Trans. on Semiconductor Mfg.* (v9), pp145-149.

Lin, Y. and Lee, C. (2001). "A total standard WIP estimation method for wafer fabrication." *European Journal of Operational Research* (v131), pp78-94.

Lou, S.X.C. and Kager, P.W. (1989). "A robust production control policy for VLSI wafer fabrication." *IEEE Trans. on Semiconductor Mfg.* (v2), pp159-164.

Lu, S.H. and Kumar, P.R. (1991). "Distributed scheduling based on due dates and buffer priorities." *IEEE Trans. on Automatic Control* (v36), pp1406-1416.

Lu, S.H.; Ramaswamy, D.; and Kumar, P.R. (1994). "Efficient scheduling policies to reduce mean and variance of cycle-times in semiconductor manufacturing plants." *IEEE Trans. on Semiconductor Mfg.* (v7), pp374-388.

Narahari, Y. and Khan, L.M. (1996). "Performance analysis of scheduling policies in re-entrant manufacturing systems." *Computers Operations Research* (v23), pp37-51.

Sattler, L. (1996). "Using queueing curve approximations in a fab to determine productivity improvements." *Proc.* of 7th Annual IEEE/SEMI Advanced Semiconductor Mfg. Conf. and Workshop, Cambridge, MA, pp140-145.

Segal, M. and Whitt, W. (1988). "A queueing network analyzer for manufacturing." *Proc.* of 12th Int'l Teletraffic Conf., Torino, Italy, pp1146-1152.

Seidman, T.I. (1994). "'First come, first served' can be unstable!" *IEEE Trans. on Automatic Control* (v39), pp2166-2171.

Shiu, J.; Hwang, T.-K.; Huang, Y.-W.; Tsai, C.-M.; Su, W.-M.; Cheng, Y.; Chang, S.-C.; and Chien, C.-C. (1996). "FASE: a scheduling environment for semiconductor fabrication." *Proc.* of 19th Int'l Electronics Mfg. Technology Symp., Austin, TX, pp34-41.

Whitt, W. (1983). "The queueing network analyzer." *Bell System Technical Journal* (v62, n9), pp2779-2815.

Whitt, W. (1993). "Approximations for the GI/G/m queue." *Production and Operations Mgmt.* (v2), pp114-161.

Wu, G.-L.; Wei, K.; Tsai, C.-Y.; Chang, S.-C.; Wang, N.-J.; Tsai, R.-L.; and Liu, H.-P. (1998). "TSS: a daily production target setting system for foundry fabs." *Proc.* of Int'l Symp. on Semiconductor Mfg., Tokyo, pp75-78.

Authors' Biographies

Ming-Der Hu received his BS degree in control engineering from National Chiao-Tung University in 1985, MS degree in information science and electronics engineering from National Central University in 1987, and PhD degree in electrical engineering from National Taiwan University in 2002. He is currently an associate scientist in the information and communication research division of Chung-Shan Institute of Science & Technology. His current research interests include distributed control of manufacturing systems, radio resource management, and radio access protocol.

Shi-Chung Chang received his BSEE degree from National Taiwan University in 1979 and his MS and PhD degrees in electrical and systems engineering from the University of Connecticut-Storrs in 1983 and 1986, respectively. He has been with the electrical engineering department of 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 at National Chi Nan University, Pu-Li, Taiwan. He is also 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, high-speed networks, Internet economics, and distributed decision making.