Proceedings of the 2000 IEEE
International Conference on Robotics & Automation
San Francisco, CA • April 2000

Queueing Network Analysis for an IC Foundry

Jia-Yang Juang * and Han-Pang Huang ◆
Department of Mechanical Engineering
National Taiwan University
Taipei, TAIWAN 10674, R.O.C.
TEL/FAX: (886) 2-23633875
e-mail: hphuang@w3.me.ntu.edu.tw
◆Professor and correspondence addressee *Graduate students

Abstract

A hybrid decomposed queueing network model is developed based on actual operating data from one particular foundry fab for rapid analysis of several performance measures. The queueing model includes analyzer and predictor modules. The system analyzer module provides the analysis of arrival pattern and service pattern for each tool group. The system predictor module provides the forecast of important performance measures, such as products cycle time, lots remaining cycle time, tool utilization, queueing length, tool moves, stage moves. Utilizing the decomposition concept in the network model, there is no limitation on the number of tool groups as well as product families, and hence priority queue model can be applied. In addition, a modularized system, QFAB, is designed to implement the model proposed in this paper. A systematic analysis of arrival pattern and service pattern for tool groups is proposed. An approach to compute the effective tool number for tool groups is also addressed. Based on the analysis, the supervisors can gain more insights and choose the proper queueing models. Comparing the obtained results to the actual fab data, the accuracy of prediction of cycle time is satisfactory. The predicted results are much better than those obtained by the original approach used in the fab.

1 Introduction

Semiconductor manufacturing factories suffer from the problems of long manufacturing cycle time (or flow time), high work-in-process (WIP) level, poor due date performance, and expensive equipment. A lot of factors make the wafer fabrication process highly stochastic. This stochastic phenomenon makes the handling and prediction of important performance measures, such as cycle time, difficult tasks. One of the most important distinctions between pure foundry fabs and R&D fabs is that the former has to meet the requirements of customers. Accordingly, how to precisely handle the cycle time of all products (or orders) as well as other performance measures, such as average work-in-process at tool groups, throughput, and tool group utilization, becomes an important and challenging task for managers and engineers. In order to accomplish the above task, a detailed hybrid decomposed queueing network model for semiconductor foundries is proposed. Though the use of queueing models for performance evaluation of semiconductor manufacturing systems is not new, our models differs from others in the broader sense. The model can be utilized in more complex foundry fabs and extended to more general queueing models, including G/G/c priority queues.

The main goal of this paper is to develop an effective and efficient queueing analytical model, as opposed to simulation studies, for rapid and accurate performance evaluation of a complex semiconductor foundry. The model is aimed at attaining the following objectives:

- Predict several important performance measures, especially product cycle times.
- Provide an approach to analyze the arrival pattern and service pattern for each tool group.
- Incorporate with the database system in the fab.
- Be suitable for a complex foundry consisting of hundreds of tools and having highly product mix.

The implemented software package is called "QFAB" (*Q*ueueing *FAB*).

In the paper, we do not attempt to propose a new exact solution or approximation for specific queueing models, such as the solution of M/G/c or G/G/c queues. Instead, we intend to develop a procedure of performance estimation with available queueing models in the literature. Only the process and inspection tools are considered. Other types of machines, such as material handling systems and storage systems are not modeled here. The model developed here is not concerned with describing individual processes, and the physical and chemical principles that determine how and why a process operates. The information needed is the nature of the disturbances that influence time and quality, and particularly its frequency, duration, and pattern of occurrence. Scheduling problems, such as lots' release policy and lots' dispatching policy, are not considered.

2 Literature Review

In the literature, there are piles of papers proposed to model manufacturing facilities and to predict product cycle times. Such research can be classified into five types: direct estimation from historical operation data, computer simulation, analytical model, statistical model, and neuro-fuzzy based model. However, some research combines two or three methods of the five. Only the analytic model is described here.

The model proposed in this paper is one kind of analytical model. Different from simulation, an analytical model is used to determine system parameters by mathematical methods. The commonly used approach is queueing theory.

Whitt [10] pioneered modeling manufacturing processes using a general type queueing network, QNA. Snowdon et al. [8] gave a detailed survey of analytical-based queueing network computational tools relevant for manufacturing systems analysis. Chen et al. [2] proposed a naive BCMP queueing network models for an analysis of wafer fabrication facilities. The model is used to predict certain key system performance measures for an R&D fab.

In 1996, Connors et al. [3] addressed a benchmark paper. Based on other research in the literature, in particular QNA [10], Connors et al. proposed a sophisticated queueing network model for semiconductor manufacturing fabs. The model is designed for rapid performance analysis of semiconductor fabs. The model considers many detailed

analyses, such as scrap and rework processes, tool breakdown and PM. Similar to other models, there are some main drawbacks of this model. First, it assumes that products follow FCFS policy. Second, the fab itself is rather simple – there are only 72 tool groups; the maximum number of tools in one tool group is 5; about 80% of the tool groups have only one tool; it does not consider the tool group overlapping phenomenon; the authors do not explain how to obtain the given probability distributions. Third, the prediction results are compared with simulation model rather than actual fab data.

3 Model Formulation and Analysis

In this section, we describe the formulation and analysis of our queueing network model. First, we propose the concept of hybrid decomposed queueing model and several model assumptions. Second, we make a classification of typical tool types in a fab for the model. Then, the relevant queueing models are derived. At last, the procedure of hybrid decomposed queueing network model is addressed.

3.1 Hybrid Decomposed Queueing Network Model and Model Assumptions

A semiconductor foundry in abstraction is a set $\Gamma = \{1,...,G\}$ of distinct tool groups among which lots can be moved from one tool group to another. These lots are mainly controlled by a central transportation center between areas and by operators between tool groups. Lots belonging to a set $\Phi = \{1,...,F\}$ of different product families are released into the foundry. Each lot has a prescribed sequence of tool groups or tools to visit before completion.

We are particularly interested in the congestion measures of the foundry, for example, the number and cycle times of lots in the system and at tool groups. Typically, a semiconductor foundry can be viewed as a multiple-class open queueing network. Each tool group is modeled as a node in the network. Unlike the traditional open queueing network models in the literature, here we bypass the determination of traffic equations, and use the empirical data of lot arrivals as the input for each node. The reasons to do so are: first, traditional queueing network analysis is often accompanied with calculation of traffic equations or normalizing constant. As the number of tool groups or product families increases, this approach becomes infeasible. Second, in the literature, the queue discipline of all queueing network models is "first come and first serve". Unfortunately, almost all foundry fabs apply "priority policy" as dispatching rule. In our model, the nodes are treated as being stochastically independent and are approximated by a GI/G/c queue having a renewal arrival process independent of service times that are independently and identically distributed with a general distribution. Specifically speaking, we construct individual models for each tool group found in a fab. Obviously, this approach is indeed an approximation. The independence can be regarded as a generalization of the product-form solution that is valid for Markovian networks. Fig. 1 shows the concept of the hybrid decomposed queueing model.

In Fig. 1, each dot with one color denotes individual tool in the fab. Note that one specific tool may belong to several different tool groups. The "Raw Data" database is the original database in the fab and the "QFAB" database represents the designed database in a local sever.

The following assumptions are made in this paper.

1) System-level Assumptions

- The network is open rather than closed. Lots come from outside, receive operation at prescribed tool groups, and eventually leave the system.
- Each tool group has one queue. There are no capacity constraints of the queues (or waiting space) for tool groups, i.e., the buffer size of tool groups is infinite.
- The arrival and service pattern for each tool group are stationary.

2) Lot Assumptions

- Lot transportation times are included in the waiting times.
- Each lot is considered as an individual entity even though it often consists of several wafers.

3) Tool Assumptions

- Each tool group consists of one or several identical tools.
- The mean and variance of the duration and interarrival times of events that interrupt the operation are known. Such events are called non-available events in this paper.

4) Operation Assumptions

- The operators in fab follow a strict priority policy, i.e., when the tool become idle, the operator will choose the highest priority lot waiting in queue for operation.
- All non-available events are non-preemptive.

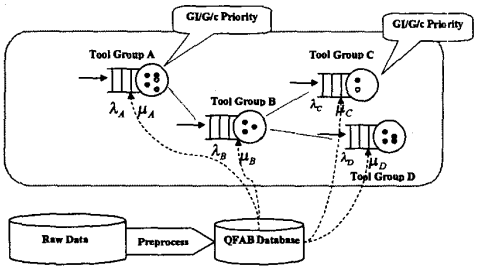

Fig. 1 Illustration of hybrid decomposed queueing network model

3.2 Classification of Tool Types

In this subsection, we focus on the classification of common tools found in the fab and derive the mean and variance expressions of processing time for each type of tools. The criterion of tool classification is according to its operation characteristics. Once the mean and variance of the processing time are derived, combined with the first two moments of interarrival times, we can determine the interested performance measures for each tool group by using the queueing formulas derived in Section 3.3.

Following the spirit of Connors et al. [3], we propose a new classification of tool groups. According to the number of wafers being process simultaneously, the tools can be roughly categorized into single-run and batch-run tools. In our classification, single-run tools consist of single-wafer, conveyor, inspection, and multi-chamber tools. Batch-run tools consist of normal batch-run and multi-stage tools.

All the operation information or probability distribution described in the following section can be obtained from empirical data in the foundry.

3.3 Derivation of Queueing Models

In this subsection, several candidate queueing formulas are examined for the corresponding tool group types. The typical queueing formulas utilized here are to calculate the

mean queueing delay (mean waiting time) or mean queue size in terms of the first two moments (mean and variance) of interarrival times and process times. Once one particular performance measure is obtained, the others can be calculated by Little's formula. In our model, we take non-available events and priority of normal products into consideration. Here we assume that non-available events are the highest priority "customer" and there are κ priority classes of normal products in the foundry.

Before the discussion of queueing model, we describe each type of non-available events that can affect the tool utilization. Then, the subsequent subsections will discuss two categories of queueing models: Single-run and batch-run tools.

1) Non-Available Events

In the foundry, a tool can stay at several possible status. The status can be roughly classified into two types: available status and non-available status. When a tool is in one of the available status, it is or has the potential to process the lots. On the other hand, when a tool is in one of the non-available status, such as PM, breakdown and etc., it cannot proceed any normal operation.

In our model, the non-available events are modeled as non-preemptive priority lots that arrive to each tool group according to renewal processes with known distributions.

Let Ω_g denotes the set of non-available events that affect tools at tool group g. The arrival rate of non-available event of type $\omega \in \Omega_g$ is denoted by $\lambda_{g,\omega}$. $\lambda_{g,\omega}$ is the reciprocal of MTBF (mean time between failures) for event ω. The processing time or MTTR (mean time to repair) for event ω at tool group g is a random variable $S_{g,\omega}$ with a known distribution, which represents the duration of the non-available event. Hence, the first two moments of $S_{g,\omega}$ can be obtained. Unfortunately, due to the absence of proper data, the second moment (or variance) of MTBF can not be obtained in our application.

2) Single-Run Tools

Single-run tools are the majority in the fabs. A great portion of tools falls into this type. For single-run tools, we derive two different queueing models: M/G/c/Priority and G/G/c/Priority. It should be noted that all the models used are priority queues in order to capture the operation traits in a real foundry.

M/G/c/Priority queues [1, 3, 6, 9]

In many real job shop systems it has been observed that the Poisson process is an adequate representation of the arrival process. Exponential distributions may not be good representations of the processing times. Therefore, M/G/c model becomes valuable in the applications of shop floor modeling.

In this case, we still assume that both lots and non-available events arrive following Poisson processes, but the assumption that the processing time of all lots is exponentially distributed is removed.

There is no exact explicit solution for the M/G/c/Priority model when c is greater than one. The approximation by Sakasegawa [9] and Connors et al. [3] is adopted. Consequently, the mean waiting time at tool group g, for Poisson arrivals, is

$$W_{q,g}^{(i)} = \frac{\rho_g^{\sqrt{2(c_g+1)}}}{c_g^2} \cdot \frac{\sum_{k=1}^{\kappa} \lambda_{g,k} E[S_{g,k}^2] + \sum_{\omega \in \Omega_g} \lambda_{g,\omega} E[S_{g,\omega}^2]}{2(1-\sigma_{g,i-1})(1-\sigma_{g,i})} \quad (1)$$

where ρ_g is the utilization of tool group g, including non-available events. Note that Eq. (1) is exact for the case c=1.

Other performance measures can be obtained by Little's formulas.

GI/G/c/Priority queues [1, 3, 10]

Queueing theory has been studied thoroughly throughout 1950s, but many problems still remain unsolved, especially the related research about GI/G/c queues. There exist several approximations for the mean queueing delay (or waiting time) of GI/G/c queues, but up to now there is no specific formula suitable for all kinds of circumstances and there is no specific approximation absolutely overreaching the others. In our model, we adopt the approximations proposed by Connors et al. [3], Whitt [10], and Buzacott et al. [1].

For GI/G/c/Priority queueing system, the mean waiting time is approximated by incorporated an adjustment factor φ into Eq. (1). The mean waiting time at tool group g is given by

$$W_{q,g}^{(i)} = \frac{\rho_g^{\sqrt{2(c_g+1)}}}{c_g^2} \cdot \frac{\sum_{k=1}^{\kappa} \lambda_{g,k} E[S_{g,k}^2] \phi_{g,k} + \sum_{\omega \in \Omega_g} \lambda_{g,\omega} E[S_{g,\omega}^2] \phi_{g,\omega}}{2(1-\sigma_{g,i-1})(1-\sigma_{g,i})} \quad (2)$$

where $\phi_{g,k} \equiv \left(c_{a,g,k}^2 + c_{s,g,k}^2\right) / \left(c_{s,g,k}^2 + 1\right)$, the quantities $c_{a,g,k}^2$ and $c_{s,g,k}^2$ represent the squares of the variation coefficients (SVC) of the interarrivals and processing times of the lots with priority k at tool group g, respectively. While $\phi_{g,\omega} \equiv \left(c_{a,g,\omega}^2 + c_{s,g,\omega}^2\right) / \left(c_{s,g,\omega}^2 + 1\right)$, the quantities $c_{a,g,\omega}^2$ and $c_{s,g,\omega}^2$ represent the squares of the variation coefficients (SVC) of the MTBF and MTTR of non-available events at tool group g, respectively.

Other performance measures can be obtained by Little's formula. This model is called QFAB_G/G/c model 1 in this paper.

Our second approach to the solution of mean waiting time of GI/G/c/Priority model is to substitute the results of GI/G/c/FCFS formula proposed by Whitt [10] into the GI/G/c/Priority model proposed by Buzacott et al. [1]. The analysis is as follows. A simple approximation for mean waiting time of a lot at tool group g based on heavy-traffic limit theorems is

$$W_{q,g} = \left(\frac{c_{a,g}^2 + c_{s,g}^2}{2}\right) \cdot W_{q,g\,M/M/c/FCFS} \quad (3)$$

then,

$$W_{q,g} = \left(\frac{r_g^{c_g}}{c_g!(c_g\mu_g)(1-\rho_g)^2}\right) \cdot \left(\sum_{n=0}^{c_g-1} \frac{r_g^n}{n!} + \frac{r_g^{c_g}}{c_g!(1-\rho_g)}\right)^{-1} \quad (4)$$

where the definitions of the variables are the same as before, $r_g = \rho_g / c_g = \lambda_g / \mu_g$, and $W_{q,g\,M/M/c/FCFS}$ is the mean waiting time at tool group g for an M/M/c/FCFS model.

Eq. (3) has been frequently used for M/G/c/FCFS queues and is known to perform quite well in that case. When the utilization of the tool group, ρ_g, approaches to 1, Eq. (3) is asymptotically correct for GI/G/c systems. Some additional study indicates that Eq. (3) is also reasonable for moderate values of ρ_g when $c_{a,g}^2 > 0.9$ and $c_{s,g}^2 > 0.9$, or $c_{a,g}^2 < 1.1$ and $c_{s,g}^2 < 1.1$ [10].

The approximation of the mean waiting time for the GI/G/c non-preemptive queue is given by

$$W_{q,g}^{(i)} = \left(\frac{1-\rho_g}{(1-\sigma_i)(1-\sigma_{i-1})}\right) \cdot W_{q,g\ GI/G/c/FCFS} \tag{5}$$

where the variables are the same as before and $W_{q,g\ GI/G/c/FCFS}$ is substituted by Eq. (4).

Eq. (5) is called QFAB_G/G/c model 2 in this paper.

3) Batch-Run Tools [3, 4, 5, 6]

Batch-run tools process in batch. This type of tool group can be modeled as a bulk service queueing system. Here, we will discuss two kinds of bulk service queues: $M/M^{[K]}/c/FCFS$ and $G/G^{[K]}/c/FCFS$. Ghare [5] obtained the steady-state joint distribution of the number in the queue and the number of busy channels. Cromie et al. [4] extends Ghare's analytical results to obtain the explicit expressions of the measures of efficiency and delay distributions. The derivation for the mean waiting time of a lot at tool group g is then given by

$$\begin{aligned} (P_{0,0})^{-1} &= \frac{(c_g r_g)^{c_g}}{c_g!}\left(1-\frac{1}{V}\right)^{-1} + \sum_{i=0}^{c_g-1}\frac{r_g^i}{i!} \\ P_{m,o} &= P_{0,0}\frac{r_g^m}{m!}, \quad m = 1,\ldots,c_g-1 \\ P_{c_g,n} &= P_{0,0}\frac{r_g^{c_g}}{c_g!}\left(\frac{1}{V}\right)^n, \quad n = 0,1,\ldots \end{aligned} \tag{6}$$

where c_g is tool number of tool group g, $r_g = \lambda_g/\mu_g$, $g \in \Gamma$. Γ is the set of tool groups, and V is the single real root, lying in the interval $(1, c_g K_g / r_g)$, of the following equation

$$f(V) = \frac{r_g}{c_g}V^{(K_g+1)} - \left(1+\frac{r_g}{c_g}\right)V^{K_g} + 1 = 0$$

where K_g is the maximum batch size. The above equation can be solved using the Secant root-finding method.

The mean waiting time is then given by

$$W_{q,M/M^{K_g}/c_g/FCFS} = \frac{P_{c_g,0}V}{\lambda_g(V-1)^2} \tag{7}$$

For $G/G^K/c_g/FCFS$ queueing systems, Connors et al. [3] mimicked the well-known approximation for GI/G/c queues as

$$W_{q,G/G^{K_g}/c_g/FCFS} \approx \frac{P_{c_g,0}V}{\lambda_g(V-1)^2}\cdot\frac{\left(c_{a,g}^2 + c_{s,g}^2\right)}{2} \tag{8}$$

where $c_{a,g}^2$ and $c_{s,g}^2$ are the squares of the variation coefficients of tool group g; $g \in \Gamma$. V is the same as the one in Eq. (6).

In this paper, we use FCFS queueing models rather than priority ones for batch-run tools. The reason is twofold: batch-run tools process and the batch operation are time-consuming. Hence, the operator is apt to make the batch size per operation as large as possible. Consequently, the priority of lots does not make much influence; and the queueing model $G/G^K/c_g$/Priority is too complicate to have any explicit form solution. In order to take into consideration of non-available events of batch-run tool groups, $r_g = \lambda_g/\mu_g$ in Eq. (6) should be adjusted as $r_g = \lambda_g \cdot E[\widetilde{S}_g]$, where $\widetilde{S}_g$ is the adjusted processing time of tool group g, which incorporates the duration of non-available events. The adjusted processing time $\widetilde{S}_g$ is defined by

$$\widetilde{S}_g = S_g + \mathrm{X}_g \tag{9}$$

where S_g is the processing time of normal lots and X_g is defined as

$$\mathrm{X}_g = \begin{cases} S_{g,\omega} & \text{w.p.} \quad \left(\lambda_{g,\omega}/\lambda_g\right), \ \omega \in \Omega_g \\ 0 & \text{w.p.} \quad \left(1 - \sum_{\omega\in\Omega_g}\lambda_{g,\omega}/\lambda_g\right) \end{cases} \tag{10}$$

The first two moments of $\widetilde{S}_g$ are then given by

$$E[\widetilde{S}_g] = E[S_g] + E[\mathrm{X}_g]$$

$$E[\widetilde{S}_g^2] = E[S_g^2] + E[\mathrm{X}_g^2] + 2E[S_g]E[\mathrm{X}_g]$$

The SVC for the adjusted process time, $c_{\widetilde{s},g}^2$, used to replace $c_{s,g}^2$ in Eq. (8), is

$$c_{\widetilde{s},g}^2 = \frac{E[\widetilde{S}_g^2] - E^2[\widetilde{S}_g]}{E^2[\widetilde{S}_g]} \tag{11}$$

3.4 Modeling Procedure

1) Input Model for Each Tool Group

Probably the least glamorous and most essential task in model building is gathering data for parameter estimation. The required raw data consists of the time at which each successive lot arrives, the time at which the lot begins and ends processing, and the tool status. Once the raw data is collected, it should be preprocessed to an appropriate form.

The number of tools for each tool group recorded in database is called nominal tool number.

Once the data has been prepared adequately, we can analyze the arrival pattern and processing pattern for each tool group. The main steps are empirical distribution, parameter estimation, and goodness-of-fit test.

2) Queueing Model for Each Tool Group

After input modeling procedure, the distributions of interarrival times and processing times for each tool group are determined. As a result, we can chose the most appropriate queueing model prepared in Section 3.3 for each tool group.

However, there is still one problem remaining unsolved. As mentioned earlier, the overlapping phenomenon between tool groups is heavy. Therefore, the nominal tool number of a tool group cannot be used as the number of servers in service facility while utilizing the queueing formulas.

Here, we propose an approach to tackle this problem by making a modification on the number of tools for each tool group. The modified number of tools for tool group g is called effective tool number and is denoted by c_g^* in contrast to the nominal one, $\overline{c}_g$. Assume that the operation of tool groups is stationary during a short time span.

Let random variable $\mathrm{W}_{q,g}$ denote the waiting time of a lot at tool group g, $f_g\left(\lambda, c_a^2, \mu, c_s^2, c\right)$ be the function of mean waiting time of tool group g, and n be the sample size. Denote the tool set vector as $\mathbf{c} = (c_1,\ldots,c_G)$. The optimal tool set vector, or effective tool number vector, $\mathbf{c}^*$ is then determined by

$$\min_{\mathbf{c}}\left(\sum_{g\in\Gamma}\left|f_g\left(\lambda, c_a^2, \mu, c_s^2, c\right) - \Sigma_{i=0}^{n} w_{q,g,i}/n\right|\right) \tag{12}$$

given $f_g(\cdot)$, $\lambda_g, c_{a,g}^2, \mu_g, c_{s,g}^2$ for $g \in \Gamma$ and n observed

waiting times $w_{q,i}$ for each tool group.

Note that Eq. (12) turns out to be a first-order linear equation with one set of independent variables, **c**, after substituting the given values. Hence, the existence and uniqueness of the solution of **c** is guaranteed. The solution of Eq. (12) is the optimal value $\mathbf{c}^*$ and it is served as the number of tools for each tool group in our queueing models.

Once the effective tool number for each tool group is obtained, the performance measures for each tool group as well as the ones for the entire system can be calculated.

3) Functions Provided

The functions provided by our queueing model, QFAB, can be partitioned into two parts. One provides the analysis of entire system. The other provides prediction functions. The former is called *System Analyzer Module*; the latter is the *Prediction Module*. The principal aim of the system analyzer module is to analyze the fab performance via the analysis of arrival and service patterns.

The main objective of Prediction Module is to forecast the cycle time of products and some other performance measure, such as WIP, utilization, tool group move, and stage move. Interested readers may refer to Juang [7].

Mean Cycle Time of Products

From our previous analysis, the mean waiting time and mean processing time are calculated for each tool group. This information, together with the routing flow information for each product family, allows us to compute the estimates of the average cycle time for each product family.

Suppose that we wish to calculate the average cycle time for product family $f \in \Phi$. Let $N^f = \{1,2,...,N\}$ denote the set of nominal operations for product family f. Then the cycle time of product family f with priority i, $CT^{f,(i)}$, is approximately given by

$$CT^{f,(i)} \approx \sum_{n \in N^f} \left(W^{(i)}_{q,g(n)} + S_{g(n),r(n)} \right) \tag{13}$$

where the mean waiting time and mean processing time is the same as the previous ones, and functions $g(n)$ and $r(n)$ are defined as

$g(n)$: $\{1,...,N\} \rightarrow \{1,...,G\}$ specifies the tool group at which operation n is performed;

$r(n)$: $\{1,...,N\} \rightarrow \{1,...,R\}$ specifies the recipe of operation n.

4 Results and Discussion

In this section, we demonstrate the results of analysis for a 200mm semiconductor foundry, Fab-X, located in Hsin-chu Science-based Industrial Park in Taiwan by using QFAB and give some discussion. The main products of Fab-X are memory and logic devices. There are about 200 distinct tool groups and 700 distinct tools. In a common condition, there are approximately 100 different products and 1000 different lots in the foundry at the same time. The study time span is from November 1998 to February 1999.

4.1 Arrival and Service Pattern

According to the analysis of actual data, we observe that the processing time distribution of many tool groups follows Erlang distribution, though we do not attempt to make a conclusion about the inherent distribution family of processing time suitable for all tool groups. One important issue should be emphasized is that the distribution of lot processing time *never* becomes exponential. This is why we do not adopt any G/M/c queueing model in our application.

Unlike the processing time distribution, the interarrival times distributions of almost all the tool groups have the "shape" of exponential distribution. As a result, the arrival process of a lot follows Poisson process. This property makes the application more powerful because of the relatively simple analytical structure. However, from the detailed χ^2 test, only a few number of tool groups really have Poisson arrivals. The result is shown in Table 1.

Table 1 The number of tool groups whose interarrivals follow exponential distribution.

Tool Type	# of Tool Groups
Single-Wafer	12
Conveyor	6
Inspection	2
Multi-Chamber	6
Normal-Batch	1
Multi-Stage	0
Total	27

4.2 Mean Cycle Time of Products

Based on the pre-analyzed results about the arrival and service patterns, many important performance measures can be forecasted by QFAB. In this section, the forecast results and accuracy are addressed and compared with actual fab data as well as FOX algorithm, which was proposed by Yeh et al. [11].

In order to validate our model and compare the forecast results with actual data and FOX ones, dozens of lots belonging to six different product families with different routes have been chosen. Among the product families selected, PROD_E has the maximum number of circuitry layers and total operation steps, while PROD_B and PROD_C have the minimum number.

Fig. 2 shows the actual and predicted cycle time of the selected products. The forecasting time instant is on Nov. 18, 1998.

In Fig. 2, six types of models are used. They are QFAB_Adaptive, QFAB_GGc model 1, QFAB_GGc model 2, QFAB_MGc, QFAB_MMc, and FOX. Different from other approaches, QFAB_Adaptive model constructs individual sub-model for each tool group, based on the analyzed information, such as arrival and service patterns. For example, if tool group 1 has Poisson arrivals and no evidence indicates that it has exponential process time distribution, it is modeled as M/G/c priority queue. If both the distribution of arrival and service pattern can not be proven as exponential ones, the tool group is modeled as a G/G/c priority queue. However, all the tool groups are modeled as G/G/c model 1, G/G/c model 2, M/G/c, and M/M/c priority queues in QFAB_G/G/c model 1, QFAB_G/G/c model 2, QFAB_M/G/c, and QFAB_M/M/c, respectively. The cycle time estimation of lots for some specific product is conducted at release time of the lots after the lots complete all the operation steps. From Fig. 2 and our analysis, a number of important points are observed as follows.

- The accuracy of QFAB_Adaptive model is better than others for most product families.
- The error of FOX is approximately 18%. The accuracy is better than QFAB_GGc model 2 and QFAB_MMc, but is worse than the other two models.

- The cycle time estimation of QFAB_Adaptive model is between QFAB_GGc model 1 and QFAB_MGc. In fact, the results of QFAB_GGc model 1, and QFAB_MGc are the upper bound and lower bound of QFAB_Adaptive, respectively.
- Due to high SVC of interarrival times of some tool groups, the cycle time estimation of QFAB_GGc modle 1 is usually greater than that of QFAB_MGc model. Remember that the former is equal to the latter multiplied by a factor. Please refer to Eq. (1) and Eq. (2).
- It is not surprised that QFAB_MMc model has the lowest cycle time estimation.
- The cycle time estimation of QFAB_GGc model 2 is underestimated.

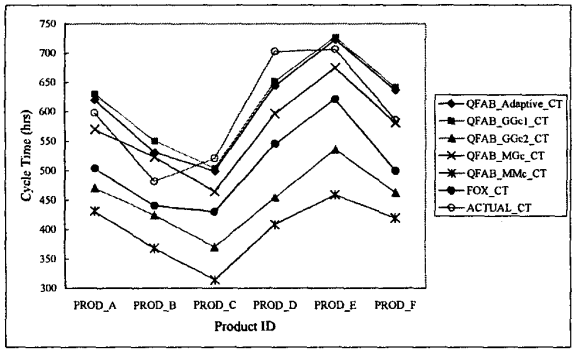

Fig. 2 Actual and predicted cycle time of six different products

Compared with the aggregated cycle time, the forecast error of waiting time for each tool group is rather large. One of the possible reasons is that discrepancies in waiting time estimation tend to cancel each other out when the aggregated cycle time is calculated.

While we cannot report any explicit comparison of cycle times for proprietary reasons, it is observed that the forecast error of QFAB_Adaptive model fall within 10% for a majority portion of products. Note that the values computed in Fig. 2 are average or expected ones for *all* priority classes of lots.

5 Conclusion

In most queueing networks model utilized in semiconductor fab in the literature, the cases under study were extremely oversimplified. The number of tool groups and tools is usually much smaller than the one in an actual fab. From the application aspects, such models may lose the representation of a real fab, which is complex inherently. The analysis of arrival and service pattern for tool groups is often ignored and there is little related research discussing the issue about the phenomenon of tool group overlapping.

According to our analysis, the empirical distribution of interarrival times for most tool groups has the "shape" of exponential distribution. However, χ^2 test indicates that only 15% of the tool groups really have Poisson arrivals. Analysis results show the distribution of service times less dispersive. Though there is no one specific distribution family can completely fit the data, Erlang or hyperexponential distribution is suggested. An approach to calculate effective tool number of tool groups is proposed to overcome the tool group overlapping problem. Comparison results indicate QFAB_Adaptive has the higher accuracy than other models, including FOX, a historical-data-based cycle time estimator. The forecast error of product cycle time is within 10% .

Acknowledgement

This work is partially supported by National Science Council under Grant number NSC 88-2218-E-002-003.

References

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

[2] H. Chen, J. M. Harrison, A. Mandelbaum, A.V. Ackere, and L.M. Wein, "Empirical Evaluation of a Queueing Network Model for Semiconductor Wafer Fabrication," *Operations Research*, vol. 36, no. 2, pp.202-215, 1988.

[3] D.P. Connors, G.E. Feigin, and D.D. Yao, "A Queueing Network Model for Semiconductor Manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 9, no. 3, pp. 412-427, 1996.

[4] M. V. Cromie and M. L. Chaudhry, "Analytically Explicit Results for the Queueing System M/M^t/C with Charts and Tables for Certain Measures of Efficiency," *Operational Research Quarterly*, vol. 27, no. 3, pp. 733-745, 1976.

[5] P. M. Ghare, "Multichannel Queueing System with Bulk Service," *Operations Research*, vol. 16, no. 1, pp. 189-192, 1968.

[6] D. Gross and C.M. Harris, *Fundamentals of Queueing Theory*, 3rd ed., New York: John Wiley & Sons, Inc., 1998.

[7] J. Y. Juang, "Development of Hybrid Decomposed Queueing Network Model for an IC Foundry," Master Thesis, Institute of Mechanical Engineering, National Taiwan University, 1999.

[8] J. L. Snowdon and J. C. Ammons, "A Survey of Queueing Network Packages for the Analysis of Manufacturing Systems," *Manufacturing Review*, vol. 1, no. 1, pp. 14-25, 1988.

[9] H. Sakasegawa, "An Approximation formula $L_q = \alpha\rho^{\beta}/(1-\rho)$," *Ann. Inst. Statist. Math.*, vol. 29, pp. 67-75, 1977.

[10] W. Whitt, "The Queueing Network Analyzer," *Bell System Technical Journal*, vol. 62, no. 9, pp. 2779-2815, 1983.

[11] C.F. Yeh, H.P. Huang, J.Y. Juang, L.R. Lin, and T. Chen, "Dynamic Average Method for Cycle Time Estimator in an IC Fab," *1998 Semiconductor Manufacturing Technology WorkShop Taiwan*, IEEE Electron Devices Society Taipei Chapter, 1998.