Proceedings of the 2007 Winter Simulation Conference
S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds.

METAMODELING FOR CYCLE TIME-THROUGHPUT-PRODUCT MIX SURFACES USING PROGRESSIVE MODEL FITTING

Feng Yang
Jingang Liu

Industrial and Management Systems Engineering Dept.
West Virginia University
Morgantown, WV 26506, U.S.A

Mustafa Tongarlak
Bruce E. Ankenman
Barry L. Nelson

Dept. of Industrial Engineering & Management Sciences
Northwestern University
Evanston, IL 60208-3119, U.S.A.

ABSTRACT

A simulation-based methodology is proposed to map the mean of steady-state cycle time as a function of throughput and product mix for manufacturing systems. Nonlinear regression models motivated by queueing analysis are assumed for the underlying response surface. To insure efficiency and control estimation error, simulation experiments are built up sequentially using a multistage procedure to collect data for the fitting of the models. The resulting response surface is able to provide a cycle-time estimate for any throughput and any product mix, and thus allows the decision maker to instantly investigate options and trade offs regarding their production planning.

1 INTRODUCTION

Planning for semiconductor manufacturing, either at the factory or enterprise level, requires answering what-if questions involving (perhaps a very large number of) different scenarios for product mix, production targets, and capital expansion. Many man-hours are invested in developing and exercising simulation models of wafer fabs, models that include critical details that are difficult or impossible to incorporate into simple load calculations or queueing approximations. Unfortunately, simulation models can be clumsy tools for planning or decision making because even a few minutes per simulation run (which is optimistic) is too slow to allow what-if analysis in real time. In our research, we develop techniques to support strategic planning from a new perspective: we combine computing horsepower, adaptive statistical methods and queueing theory to make simulation a much more effective tool than before.

In this paper, we propose a simulation-based methodology to estimate the mean of steady-state cycle time (CT) as a function of throughput (TH) and product mix (PM). Cycle time is defined as a random variable representing the time required for a job or lot to traverse a given routing in a production system (e.g., Hopp and Spearman 2001). A fab can control cycle time by controlling the product mix and the rate at which lots are started in the factory (lot-start rate or equivalently, throughput rate). Hence, the CT-TH-PM surface can play an important role in strategic planning of semiconductor manufacturing including evaluating the mean of cycle time for a given throughput and product mix; determining the sensitivity of product cycle times to changes in throughput or product mix; determining feasible throughputs that satisfy cycle-time constraints; and finding a product mix that maximizes revenue subject to cycle-time and throughput constraints.

Our methodology is able to generate a complete CT-TH-PM surface (with the response of interest being the long-run average cycle time of products) like that provided by a tractable queueing model, but with the fidelity of simulation. Given a simulation model of a wafer fab, simulation experiments are sequentially performed at selected settings of throughput and product mix until a desired precision has been achieved on the estimation. Based on the data collected over the TH-PM space, CT-TH-PM surfaces are fitted assuming that the underlying surface can be captured by the proposed regression models, the forms of which are motivated by queueing analysis. Such a response surface is able to provide a cycle-time estimate for any throughput and any product mix, and thus allows the decision maker to investigate options and trade offs almost instantly without running additional simulations.

2 STATEMENT OF THE PROBLEM

2.1 Formulation of the Product System

As noted, our goal is to approximate the cycle-time response surface as a function of the factory throughput and product mix. In this section, we define most of the notation that

will be used in the remainder of this paper, and state the research problem in more precise terms. We formulate a stylized model of a manufacturing system to motivate the CT-TH-PM response-surface model that we ultimately fit to a detailed simulation model.

We consider an M-station manufacturing system (e.g., wafer fab) which processes K different types of products, and we define the system in a generic way as follows.

- $\{s_j, j = 1,2,\ldots,M\}$: the number of parallel resources at station j.
- μ_{ij}: the effective service rate of each resource at station j for products of type i.
- δ_{ij}: the expected number of visits by product type i to station j.

The product flow is described by:

- λ: the overall release rate of all the products into the system.
- $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)$: the product-mix vector with each element α_i representing the fraction of type i products in the flow, so that $\sum_{i=1}^{K} \alpha_i = 1$, $\alpha_i \in [0,1]$.
- $\lambda_i = \alpha_i \lambda$: the release rate of product type i to the system.

Under this formulation, we can easily calculate ρ_j, the utilization of station j ($j = 1,2,\ldots,M$). Let $\rho_{ij} = \delta_{ij}/(s_j \mu_{ij})$, then $\rho_j = \lambda \sum_{i=1}^{K} \alpha_i \rho_{ij}$. The maximum utilization $\rho_{max} = \max_j \rho_j$ is called the *system utilization* and is denoted by x in this paper. A station, say, station j_{BN} that reaches ρ_{max} is called a bottleneck (BN) station, that is,

$$j_{BN} = \text{argmax}_j \rho_j = \text{argmax}_j \sum_{i=1}^{K} \alpha_i \rho_{ij}. \tag{1}$$

The stability constraint on the system requires $x = \lambda \sum_{i=1}^{K} \alpha_i \rho_{i j_{BN}} < 1$, or equivalently,

$$\lambda < 1/\sum_{i=1}^{K} \alpha_i \rho_{i j_{BN}} = \mu^*(\alpha) \tag{2}$$

where $\mu^*(\alpha)$ is the system capacity, the upper limit on λ (or overall throughput) for stability. Obviously, capacity $\mu^*(\alpha)$ depends on the system parameters as well as α, and we assume that $\mu^*(\alpha)$ can be analytically approximated for a given system and product mix.

2.2 CT-TH-PM Surface

The response of interest is the mean of steady-state cycle time for products of type i ($i = 1,2,\ldots,K$), denoted $C_i(\lambda, \alpha)$. Different types of products follow different processing steps, and thus have different cycle-time distributions. For each type of product, we seek to estimate its average cycle time which depends on the overall product flow through the system. The product flow is characterized by starting rate/throughput λ and product mix α, and in our work, we consider λ and α as independent variables that can be controlled by the production manager (equivalently, the decision variables are $\{\lambda_i = \alpha_i \lambda, i = 1,2,\ldots,K\}$, the start rates of each product). As established in (2), the stability condition of the system is such that λ has to be less than the capacity $\mu^*(\alpha)$, which is a function of α. To normalize the range of λ across the PM region, we chose to directly estimate $C_i(x, \alpha)$ where $x = \lambda/\mu^*(\alpha)$ is the fraction of system capacity in use and x is on the scale of $[0,1)$ regardless of the value of α. Once we have obtained $c_i(x, \alpha)$, a simple transformation will give us $c_i(\lambda, \alpha)$.

To model the CT-TH-PM surface, the most straightforward way is to develop a response-surface model that incorporates x and α as independent variables. However, our investigation of analytically tractable queueing network models convinces us that a general model for $c_i(x, \alpha)$ is unlikely to be successful because the correct form of the model depends on specifics of the network topology of the factory. Therefore, we proposed a 2-step methodology for the generation of the CT-TH-PM surface, which is described in the next section.

3 OVERVIEW OF THE METHODOLOGY

Our objective is to estimate the cycle-time measure at any normalized throughput x and for any feasible product mix α. In light of the issues discussed in Section 2.2, we decided to utilize our success in estimating CT-TH curves with a fixed product mix. We propose first using simulation to fit CT-TH curves for a carefully selected range of product mixes, and then perform model fitting across the α-space. More specifically, we define

- $c_{i,x}(\alpha)$: the cycle time of product i at fixed throughput x as a function of product mix α.
- $c_{i,}\alpha(x)$: the cycle time of product i for a given product mix α as a function of throughput x.
- $\mathscr{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_n\}$: a collection of n product mixes.

Our methodology consists of two steps:

1. Use an extended version of the methodology of Yang, Ankenman, and Nelson (2007) to estimate the CT-TH curves $\{c_{i,}\alpha(x), \alpha \in \mathscr{A}\}$ for product i ($i = 1,2,\ldots,K$) over a given throughput range $x \in [x_L, x_U]$ ($0 < x_L < x_U < 1$). We take products of type 1 for example. Figure 1 shows the CT-TH curves for product 1 with each curve corresponding to a different product mix $\alpha_i \in \mathscr{A}$ ($i = 1,2,3$). Note that by making the independent variable for

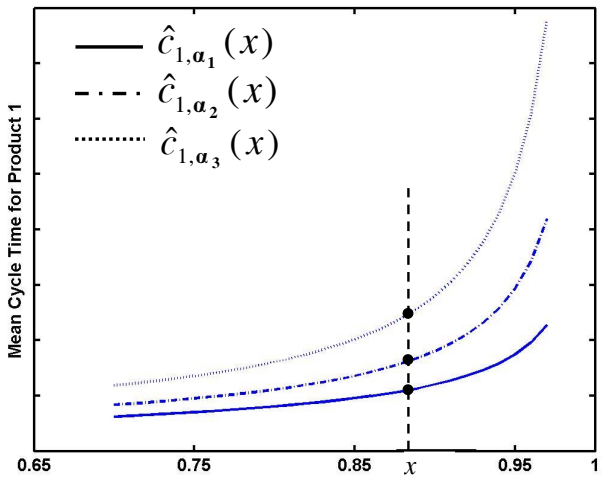

Figure 1: CT-TH curves for product 1 at different product mixes.

each product's CT-TH curve the fraction of system capacity at a given product mix, all curves run from $[x_L, x_U]$, providing a common scale.

2. With these estimated CT-TH curves for product 1 at a selected set of product mixes $\{\widehat{c}_{1,\alpha}(x), \alpha \in \mathscr{A}\}$, we can predict at any throughput x the cycle times $\widehat{c}_{1,\alpha}(x) = \widehat{c}_{1,x}(\alpha)$, $\alpha \in \mathscr{A}$. Based on data points $\{\widehat{c}_{1,x}(\alpha)\, \alpha \in \mathscr{A}\}$ denoted as black dots in Figure 1, model fitting is performed over the α-space to obtain $\widehat{c}_{1,x}(\alpha)$ for any feasible product mix α.

Next, we discuss the technical details of this approach.

4 REVIEW OF THE ESTIMATION OF CT-TH CURVES

As already illustrated, estimating CT-TH curves over a collection of product mixes $\mathscr{A}$ is the primary step for generating the CT-TH-PM response surface, which provides the basis for the estimation of cycle time across product-mix space.

In Yang, Ankenman, and Nelson (2007), a simulation-based method was proposed for the generation of CT-TH curves at a fixed product mix. A nonlinear regression model (3), which is motivated by heavy-traffic queueing analysis, was developed to represent the underlying CT-TH curve

$$C_{i,\alpha}(x) = \frac{\sum_{\ell=0}^{t} c_\ell x^\ell}{(1-x)^p}. \tag{3}$$

We fit such a model for each product i and for each product mix $\alpha \in \mathscr{A}$. To estimate the model efficiently, simulation experiments are sequentially performed at different throughput rates x for data collection. The experimentation is continued until a desired precision has been achieved for the curve estimation.

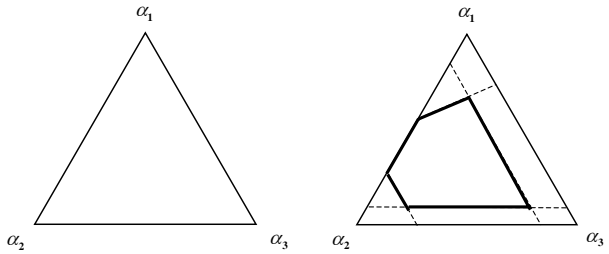

Figure 2: Feasible product-mix space: unconstrained (left) and constrained (right).

In the next section, we discuss issues related to the estimation of the CT-PM surface (modeling of the cycle time across the product-mix space with fixed utilization x).

5 CT-PM RESPONSE SURFACE

5.1 The Feasible Product-Mix Space

Obviously, product mix α has to satisfy:

$$\sum_{i=1}^{K} \alpha_i = 1, \quad \alpha_i \in [0,1]. \tag{4}$$

Figure 2 (left) illustrates the feasible product-mix region in a 3-product case defined by constraint (4). In practice, the product mix is usually subject to additional linear constraints imposed by realistic situations (e.g., lower bounds on release rates). We use the following notation to represent the linear constraints on product mix

$$\mathbf{A}\alpha \leq \mathbf{b} \tag{5}$$

where $\mathbf{A}$ is a matrix with K columns with each row representing a constraint. Figure 2 (right) gives an example of the more restricted product mix region defined by (4) and (5).

5.2 Partitioning the Product-Mix Space

Production systems are usually constrained by one or more bottleneck resources. A bottleneck (BN) is usually a facility or resource which most constrains the production flow, and it plays a key role in determining the overall performance of the manufacturing system. As we change the product mix, BN may shift from one resource to anther, which complicates the way that product mix affects the cycle time. As will be seen in Section 5.3, within an α-region where no BN shift occurs, $c_{i,x}(\alpha)$ tends to be smooth and differentiable with respect to α. For the purpose of modeling the CT-PM surface, we divide the product-mix space into a number of subregions with each one dominated by a different BN station or stations.

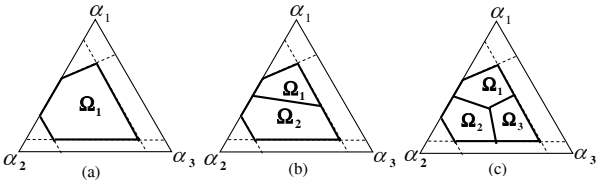

Figure 3: Division of the feasible product-mix region.

Suppose the product-mix region of feasibility is defined as $\Omega = \{\alpha | \alpha \text{ satisfies constraints (4) and (5)}\}$. Following the definition of BN station provided by (1), the subregion $\Omega_\nu = \{\alpha | \alpha \in \Omega \text{ and } \alpha \text{mix makes station } \nu \text{ a BN}\}$ is given as the collection of α that satisfies

$$\begin{aligned} \sum_{i=1}^{K} \alpha_i &= 1 \\ \mathbf{A}\alpha &\leq \mathbf{b} \\ \rho_\nu &\geq \rho_j \quad j = 1,2,\ldots,M \text{ and } j \neq \nu. \end{aligned} \tag{6}$$

Returning to the example of 3-product and 3-station system provided in Section 5.1, the feasible region as shown in Figure 2 (right) could be divided in three different ways as shown in Figure 3 depending on the system parameters.

5.3 Form of the CT-PM Model

To estimate $c_{i,x}(\alpha)$ for product i, we developed a nonlinear regression model to represent the underlying CT-PM surface, the form of which is motivated by simple queueing models such as Jackson network and M/G/1 queue.

5.3.1 Open Jackson Network Motivation

Consider a Jackson network in which each station has a single server having exponentially distributed service time with rate μ_j (independent of product type). For this network the expected cycle time for each product type can be computed analytically, and for product 1 it is

$$c_{1,x}(\alpha) = \sum_{j=1}^{M} \frac{\delta_{1j}}{\mu_j \left[1 - x\left(\frac{\sum_{k=1}^{K} \alpha_k \delta_{kj}/\mu_j}{\max_h \sum_{k=1}^{K} \alpha_k \delta_{kh}/\mu_h}\right)\right]}. \tag{7}$$

Note that in (7), a station that achieves $\max_h \sum_{k=1}^{K} \alpha_k \delta_{kh}/\mu_h$ is a BN station. Within a subregion Ω_ν defined by (6) where station ν stays BN, (7) can be written as:

$$C_{1,x}(\alpha) = \sum_{j=1}^{M} \frac{\delta_{1j}}{\mu_j \left[1 - x\left(\frac{\sum_{k=1}^{K} \alpha_k \delta_{kj}/\mu_j}{\sum_{k=1}^{K} \alpha_k \delta_{k\nu}/\mu_\nu}\right)\right]} \quad \alpha \in \Omega_\nu. \tag{8}$$

From (8), we can see that the cycle time is a continuous and differentiable function of α within a constant-BN subregion. This motivates us to separately fit a regression model to each subregion defined by (6).

Moreover, with simple mathematical manipulation, (8) can be written as:

$$c_{1,x}(\alpha) = \sum_{j=1}^{M} \frac{\sum_{k=1}^{K} a_{kj}\alpha_k}{\sum_{k=1}^{K} h_{kj}\alpha_k} = e_0 + \sum_{j=1}^{M} \frac{\sum_{k=1}^{K-1} e_{kj}\alpha_k}{h_{0j} + \sum_{k=1}^{K-1} h_{kj}\alpha_k}. \tag{9}$$

where all the coefficients a_{kj}, h_{kj}, e_{kj}, and e_0 depend on system parameters only.

5.3.2 M/G/1 Queue Motivation

Consider a multiproduct M/G/1 queue where the service rate (equivalently mean service time) of product type i is μ_i (equivalently t_i), and the variance of the service time is σ_i^2. Then using standard M/G/1 results we can show that

$$c_{i,\lambda}(\alpha) = t_1 + \frac{\frac{\lambda}{2}\sum_{i=1}^{K} \alpha_i (t_i^2 + \sigma_i^2)}{\left(1 - \lambda \sum_{i=1}^{K} \alpha_i t_i\right)}. \tag{10}$$

For this queue, $x = \lambda \sum_{i=1}^{K} \alpha_i t_i$ is the utilization, which allows us to rewrite (10) as a function of x:

$$c_{i,x}(\alpha) = t_1 + \frac{\sum_{k=1}^{K} \frac{x}{2(1-x)} \alpha_k}{\sum_{\ell=1}^{K} t_\ell \alpha_\ell} = u_0 + \frac{\sum_{k=1}^{K-1} u_k \alpha_k}{t_0 + \sum_{\ell=1}^{K-1} t_\ell \alpha_\ell} \tag{11}$$

Apparently, functional form (9) reduces to (11) in a special case of $M = 1$.

5.3.3 CT-PM Regression Model

Motivated by (9) and (11), we adopted a nonlinear regression model (12),which will be referred as CT-PM model, to approximate the CT-PM surface for product of type i within a constant-BN region.

$$\begin{aligned} c_{i,x}(\alpha) &= \mu(\overline{\alpha}) = \tau + \sum_{r=1}^{R} f(\overline{\alpha}, \mathbf{b}_r) \\ &= \tau + \sum_{r=1}^{R} \frac{\sum_{k=1}^{K-1} b_{kr}\alpha_k}{b_{0r} + \sum_{\ell=1}^{K-1} d_{\ell r}\alpha_\ell} \end{aligned} \tag{12}$$

where

- $\overline{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{K-1})$ are independent variables. We eliminate $\alpha_K = 1 - \sum_{k=1}^{K-1} \alpha_k$ from the model due to linear dependence.
- Unknown parameters are the constant term τ, and the coefficients $\mathbf{b}_r = (b_{0r}, b_{1r}, \ldots, b_{K-1,r})$ $r = 1,2,\ldots,R$.

Model (12) is the sum of R ratio models

$$f(\overline{\alpha}, \mathbf{b}_r) = \frac{\sum_{k=1}^{K-1} b_{kr}\alpha_k}{b_{0r} + \sum_{\ell=1}^{K-1} d_{\ell r}\alpha_\ell} \quad r = 1,2,\ldots,R. \tag{13}$$

R is the number of ratio models $f(\overline{\alpha}, \mathbf{b}_r)$ included in (12). The value of integer R depends on the true CT-PM surface, and the determination of R will be discussed below.

Vectors $\mathbf{d}_r = (d_{1r}, d_{2r}, \ldots, d_{K-1,r})$ $r = 1,2,\ldots,R$ are parameters estimated prior to and independent of the fitting of (12), and they are treated as known values in Model (12) (see Section 5.4).

The CT-PM model (12) is almost the same as formula (9) although it is expected that R is much smaller than M, the number of stations in the system. In the remainder of Section 5, we will further examine the shape of the CT-PM surface and detail a strategy for fitting the nonlinear response surface model (12).

5.4 Curvature of CT-PM Surface

In this subsection, we discuss the curvature (or the bending) of CT-PM surface based on Jackson networks, which is the queueing network that motivates our regression model (12). The form of (9) for a Jackson network clearly suggests an additive model which is the sum of a number of ratio functions. For the convenience of discussion, we rewrite (9) as follows

$$c_{1,x}(\overline{\alpha}) = e_0 + \sum_{j=1}^{M} \frac{g_{1j}(\overline{\alpha})}{g_{2j}(\overline{\alpha})} = e_0 + \sum_{j=1}^{M} \frac{\sum_{k=1}^{K-1} e_{kj}\alpha_k}{h_{0j} + \sum_{k=1}^{K-1} h_{kj}\alpha_k}. \tag{14}$$

For each ratio function $g_{1j}(\overline{\alpha})/g_{2j}(\overline{\alpha})$, both the numerator $g_{1j}(\overline{\alpha})$ and denominator $g_{2j}(\overline{\alpha})$ are linear functions of $\overline{\alpha}$. Geometrically speaking, $g_{2j}(\alpha)$ is a one-dimensional projection of the variable vector $\overline{\alpha}$ onto the system parameter vector $\mathbf{h}_j = (h_{1j}, h_{2j}, \ldots, h_{K-1,j})$. Since the numerator $g_{1j}(\overline{\alpha})$ is linear with respect to $\overline{\alpha}$, for response surface (14) curvature is only induced to the surface along the projections defined by $\mathbf{h}_j$ ($j = 1,2,\ldots,M$). Consider a simple case with $M = 1$, the curvature of $c_{1,x}(\overline{\alpha})$ is most pronounced along vector $\mathbf{h}_1$ whereas there is no curvature in directions orthogonal to $\mathbf{h}_1$.

Real manufacturing systems could be composed of a large number of workstations (e.g., M could be on the scale of hundreds), which implies response curvature on M directions $\{\mathbf{h}_j, j = 1,2,\ldots,M\}$. However, it is reasonable to believe that using a substantially smaller number of, say R, carefully-chosen directions, (14) could be well approximated by the sum of ratio functions along those R directions. Identifying the curvature directions of the CT-PM surface plays an important role in determining the number of ratio models R incorporated in the CT-PM model and in assisting the nonlinear fitting of (12) as will be seen later. In this paper, a method based on the quadratic polynomial approximation is used for the identification of curvature directions, namely the determination of the vectors $\{\mathbf{d}_r, r = 1,2,\ldots,R\}$ in model (12).

We assume that $c_{1,x}(\overline{\alpha})$ can be approximated by a full quadratic model:

$$\begin{aligned} c_{1,x}(\overline{\alpha}) &= \beta_0 + \alpha'\beta + \alpha'\mathbf{B}\alpha \\ &= \beta_0 + \sum_{k=1}^{K-1} \beta_k\alpha_k + \sum_{k=1}^{K-1} \beta_{kk}\alpha_k^2 + \sum_{k=1}^{K-2}\sum_{\ell=k+1}^{K-1} \beta_{k\ell}\alpha_k\alpha_\ell \end{aligned} \tag{15}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_{K-1})$, and $\mathbf{B}$ is the $(K-1)\times(K-1)$ symmetric matrix

$$\mathbf{B} = \begin{pmatrix} \beta_1 1 & \beta_1 2/2 & \cdots & \beta_{1,K-1}/2 \\ \beta_1 2/2 & \beta_2 2 & \cdots & \beta_{1,K-1}/2 \\ \vdots & & \ddots & \vdots \\ \beta_{K-1,1} & \beta_{K-1,2} & \cdots & \beta_{K-1,K-1} \end{pmatrix}. \tag{16}$$

It is our empirical experience that a quadratic model, although inadequate to accurately characterize the CT-PM surface, provides a reasonably good response surface approximation. We perform the curvature analysis based on the full quadratic model (15) following the approach in Myers and Montgomery (2002).

The curvature of the surface depends on the second-order coefficient matrix $\mathbf{B}$. Let $\mathbf{P}'\mathbf{B}\mathbf{P} = \Lambda$ where Λ is a diagonal matrix containing the eigenvalues of $\mathbf{B}$ as main diagonal elements, and $\mathbf{P}$ is the $(K-1)\times(K-1)$ matrix whose columns are the normalized eigenvectors associated with the eigenvalues of $\mathbf{B}$. Let $\mathbf{d}_{max}$ be the $(K-1)\times 1$ eigenvector of $\mathbf{B}$ associated with the maximum absolute eigenvalue λ_{max}. Then $\mathbf{d}_{max}$ represents the projection direction of $\overline{\alpha}$ along which the curvature of the surface is most marked. In model (12), $\mathbf{d}_{max}$ will be assigned to $\mathbf{d}_1$, and sequentially $\mathbf{d}_2, \mathbf{d}_3, \ldots, \mathbf{d}_R$ will be determined in the process of fitting (12) in an iterative manner. The detailed method will be given in Section 5.5, which will also show that as a by-product, the fitted quadratic model (15) serves as an approximate reference for the underlying surface to prevent overfitting of the CT-PM model.

5.5 Fitting of the CT-PM Model

There are three major difficulties involved in the nonlinear fitting of the CT-PM model:

- A high degree of correlation between model parameters is very likely to exist given the structure of the CT-PM model, and obtaining good starting

values for the unknown parameters is important to achieving a well-estimated model.
- Constraining the denominator of ratio functions $f(\overline{\alpha}, \mathbf{b}_r)$ to not be zero cause additional complications to performing the nonlinear regression.
- The number of ratio functions R characterizes the complexity of the CT-PM model and needs to be determined to avoid either underfitting or overfitting of the model.

As already mentioned, determining the curvature directions $\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_R$ independently of the nonlinear regression fitting will help alleviate the difficulties. Here, we discuss a complete model fitting strategy seeking to provide a good estimation for the CT-PM model.

Suppose that N observations have been collected within a BN-constant region Ω_v. For a given throughput x, we have $\{(\alpha_1, \widehat{c}_{1,x}(\alpha_1)), (\alpha_2, \widehat{c}_{1,x}(\alpha_2)), \ldots, (\alpha_N, \widehat{c}_{1,x}(\alpha_N))\}$. We assume that

$$\widehat{c}_{1,x}(\alpha_i) = c_{1,x}(\alpha_i) + \varepsilon_i = \mu(\overline{\alpha_i}) + \varepsilon_i \quad i = 1, 2, \ldots, N$$

where $\varepsilon_i \sim N(0, \sigma^2)$ (this assumption can be justified from estimating the CT-TH curves $\{\widehat{c}_{1,\alpha_i}(x) = \widehat{c}_{1,x}(\alpha_i), i = 1, 2, \ldots, N\}$. In addition, for convenience of the discussion, we temporarily assume that N is sufficiently large to allow for any model fitting to be performed.

To estimate the CT-PM model, we solve a constrained nonlinear least squares problem. The constraints are given as follows. For each ratio model $f(\overline{\alpha}, \mathbf{b}_r)$ incorporated in the CT-PM model, the denominator cannot be 0 over the entire subregion Ω_v. That is, for the nonlinear fitting, the unknown parameter $b_{0,r}$ has to satisfy either of the following constraints:

Constr1

$$b_{0r} + \mathbf{d}_r' \overline{\alpha} > \varepsilon \Leftrightarrow b_{0r} > \varepsilon - \min_{\overline{\alpha} \in \Omega_v} \{\mathbf{d}_r' \overline{\alpha}\};$$

Constr2

$$b_{0r} + \mathbf{d}_r' \overline{\alpha} < -\varepsilon \Leftrightarrow b_{0r} < -\varepsilon - \max_{\overline{\alpha} \in \Omega_v} \{\mathbf{d}_r' \overline{\alpha}\}.$$

where ε is a small positive value, $\overline{\alpha}$ could be any product mix within the subregion Ω_v, and $\mathbf{d}_r$ is the curvature direction. When performing the constrained nonlinear regression, only one type of constraint can be imposed on the unknown parameter b_{0r}. Let `ActiveConstr` be a $R \times 1$ array defined as: `ActiveConstr`(r)=1 if $b_{0,r}$ is subject to Constr1; `ActiveConstr`(r)=2 otherwise. For a specified `ActiveConstr` array, the constrained nonlinear least squares fitting can be formalized as:

$$\begin{aligned} \min_{\tau, \mathbf{b}_1, \ldots, \mathbf{b}_R} \quad & \sum_{i=1}^{N} [y_i - \mu(\overline{\alpha_i}, \tau, \mathbf{b}_1, \ldots, \mathbf{b}_R)]^2 \\ \text{Subject to:} \quad & b_{0,r} \text{ satisfies constraint of type} \\ & \texttt{ActiveConstr}(r) \text{for } r = 1, 2, \ldots, R \end{aligned} \tag{17}$$

We propose to estimate model (12) in three steps. Initially, we set $R = 0$ (the number of ratio functions included in the CT-PM model is 0); and the set of curvature directions $\mathbf{D} = \varnothing$.

Step 1

1. Fit the full quadratic model (15) to the data $\{(\alpha_i, y_i), i = 1, 2, \ldots, N\}$, and determine the curvature direction $\mathbf{d}_{max}$ and the corresponding eigenvalue λ_{max}. Denote the estimated quadratic model as $\widehat{\mu}_{QC}(\overline{\alpha_i})$.
2. Set $R = R + 1$, $\mathbf{d}_R = \mathbf{d}_{max}$, and $\mathbf{D} = \mathbf{D} \bigcup \mathbf{d}_R$.
3. Fit the CT-PM model with the current value of R and curvature directions $\mathbf{D}$. Two different least squares problems (17) will be solved subjecting $b_{0,R}$ to Constr1 and Constr2 respectively. Compare the sum of squared error (SSE) resulting from the two nonlinear fittings. If SSE1<SSE2, set `ActiveConstr`(R)=1; otherwise, `ActiveConstr`(R)=2.

Step 2

1. From the latest CT-PM fitting, compute the residuals $\{(\alpha_i, res_i), i = 1, 2, \ldots, N\}$, based on which perform the quadratic linear regression to identify the curvature direction $\mathbf{d}_{max}$. If the new direction $\mathbf{d}_{max}$ is nearly parallel to the directions included in $\mathbf{D}$ or the corresponding eigenvalue $|\lambda_{max}| < \lambda_0$ (λ_0 is a user-specified positive constant), then stop and declare $\widehat{\mu}_R(\overline{\alpha_i})$ obtained in Step 3(1) as the best fitted model from the current data set; Otherwise, continue.
2. Set $R = R + 1$, $\mathbf{d}_R = \mathbf{d}_{max}$, and $\mathbf{D} = \mathbf{D} \bigcup \mathbf{d}_R$.
3. Fit the partial model $\mathrm{E}[res_i] = f(\overline{\alpha}, \mathbf{b}_R)$ to $\{(\alpha_i, res_i), i = 1, 2, \ldots, N\}$. As in Step 1(3), two different nonlinear fitting will be performed subjecting b_{0R} to two types of constraints. Again, if SSE1<SSE2, set `ActiveConstr`(R)=1; otherwise, `ActiveConstr`(R)=2.

Step 3

1. Estimate the CT-PM model $\widehat{\mu}_R(\overline{\alpha_i})$ by solving (17) with the current values of R, curvature directions $\mathbf{D}$, and `ActiveConstr` array specified in the previous steps. The latest estimates of the unknown parameters $\tau, \mathbf{b}_1, \ldots, \mathbf{b}_R$ obtained from previous steps will be used as the starting values.
2. Compare $\widehat{\mu}_R(\overline{\alpha_i})$ obtained in Step 3(1) and $\widehat{\mu}_{QC}(\overline{\alpha_i})$. If the maximum relative deviation over

Ω_v is intolerably large, say,

$$\frac{\widehat{\mu}_R(\overline{\alpha_i}) - \widehat{\mu}_{QC}(\overline{\alpha_i})}{\widehat{\mu}_{QC}(\overline{\alpha_i})} > 50\%, \qquad (18)$$

then stop; reject the current fitting and declare $\widehat{\mu}_{R-1}(\overline{\alpha_i})$ as the best fitted model from the current data set. Otherwise, go back to Step 2.

6 PROCEDURE FOR ESTIMATING THE CT-TH-PM RESPONSE SURFACE

This section is devoted to construction of the experiment design and issues related to computational efficiency. To provide context, a high-level description of the procedure is provided in Figure 4. To generate the CT-TH-PM response surface, simulation experiments have to be performed at a number of TH-PM combinations for data collection. Our approach is to first select the factor levels in PM space, and then for each product mix, apply the procedure proposed by Yang, Ankenman, and Nelson (2007) to decide at what throughput rates the simulation should be carried out. As illustrated in Figure 4, the experimentation is initiated with a pilot design $\mathscr{A}_0$ consisting of N_0 product mixes. For each $\alpha \in \mathscr{A}_0$, CT-TH curves $\{C_{i,}\alpha(x), i = 1,2,\ldots,K; x \in [x_L, x_U]\}$ are generated by running simulation at different throughputs. Based on these curves, we can estimate cycle time for any product type at any throughput and product mix and evaluate the relative error obtained for the cycle-time estimates. The design is augmented by including one additional point at a time until the desired precision is achieved on the estimated response surface.

6.1 Experiment Design

Yang, Ankenman, and Nelson (2007) has provided efficient and effective experiment design strategies when product mix is fixed. In that context a "design" corresponds to settings of the normalized throughput x at which to make simulation runs, and an allocation of simulation effort to each design point. To estimate the complete CT-TH-PM surface, the design also includes $\mathscr{A}$, the collection of product mix settings at which we fit the CT-PM surface. In this section, we focus on the design in the product-mix space.

As already explained, the fitting of the CT-PM surface is based on model (12) over a constant-BN region $\Omega_j \neq \varnothing$ (j is a station that can serve as BN of the system). Hence, we discuss the allocation of the PM design points within Ω_j for the purpose of achieving well-estimated (12). Furthermore, in the following discussion regarding the design of experiments, we assume without loss of generality that products of type 1 and a given production throughput x_0 are of particular interest. Our goal is to estimate the expected cycle time at throughput x_0 for product 1 with a specified

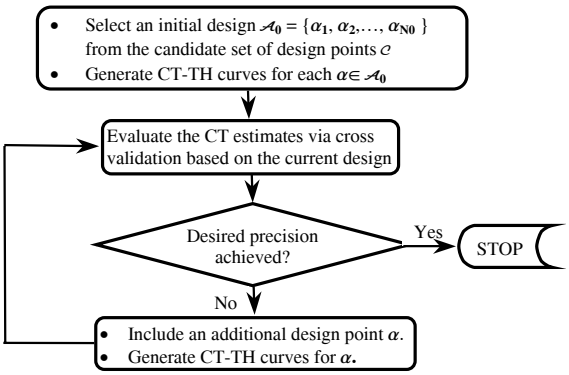

Figure 4: Flow chart for the procedure.

precision, while still well estimating the CT-PM-TH surface for all $x_L \leq x \leq x_U$ and all types of products.

Each subregion Ω_j is a simplex defined by linear constraints (6), so what we have is a K-component mixture design problem within Ω_j for the estimation of model (12) at x_0 for products of type 1.

6.1.1 Initial Design

For such constrained mixture designs, Myers and Montgomery (2002) recommended selecting design points from a candidate set, say $\mathscr{C} = \{\alpha_1^*, \alpha_2^*, \ldots, \alpha_{\overline{N}}^*\}$, which provides a good coverage of the feasible space. They claim that the set of candidate points to use for designing experiments should depend upon the form of the model the experimenter wishes to fit, and they recommended three different sets for linear, quadratic, and cubic models based on their practical experience. Our model (12) does not fall into the category of polynomials with which they have experimented. However, our empirical experience with the CT-PM surface suggests that a quadratic model is able to provide an approximate fit for the response surface, although obviously inferior compared to (12). Thus, in our experiments, we chose to use the set Myers and Montgomery (2002) recommended for quadratic models, that is, the candidate set of design should include the following points of the simplex Ω_j: extreme vertices, edge centers, constraint plane centroids, overall centroid and axial points.

Given the constraints (6) that defines Ω_j, we can use the CONVRT and CONAEV algorithms developed by Piepel (1988) to find the vertices, edge centers, and all other centroids of the simplex. In our procedure, the initial design points will be selected as a subset of these candidate points in $\mathscr{C}$. Let $\mathscr{A}_0$ denote the set of initial design points of size N_0 within constant-BN region Ω_j. To avoid extrapolation, we include all the N_v extreme vertices of Ω_j in $\mathscr{A}_0$. Besides the vertices, $\mathscr{A}_0$ must include some inner points so that we can use cross-validation to estimate the prediction error at points in $\mathscr{C} \backslash \mathscr{A}_0$ as described in Section 6.1.3. In addition, N_0 should be sufficiently large to allow for the fitting of the

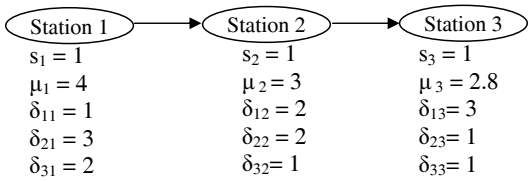

Figure 5: Three-station Jackson queueing model.

full quadratic model given by (15). In our experiments, the additional $N_0 - N_v$ non-vertex points are selected from $\mathscr{C}$ using a *maxmin* criterion which maximizes the minimum distance between any two points.

6.1.2 Design Augmentation

As illustrated in Figure 4, we initiate the experiments with a pilot design as discussed in Section 6.1.1. The design points will be sequentially added one at a time until the stopping rule is satisfied. We propose to perform the design augmentation following the method described in Seber and Wild (2003). After observations have been made at n design points, the $(n+1)^{st}$ additional design point α_{n+1} is introduced to minimize the determinant of the asymptotic variance-covariance matrix of estimated unknown parameters (D-optimality criterion).

6.1.3 Stopping Criterion

We allow the user to specify a desired precision, say $\gamma\%$, defined as the relative error on the cycle-time estimates. The sequential experimentation is terminated when the pre-specified precision is achieved.

We use cross-validation to estimate the relative error of cycle-time estimates obtained from the fitted model. For each non-vertex design point $\alpha \in \mathscr{A}$, the relative error can be estimated as

$$RE_{1,x_0}(\alpha) = \frac{\widehat{C}_{1,x_0}(\alpha) - \widetilde{C}^{(-k)}_{1,x_0}(\alpha)}{\widehat{C}_{1,x_0}(\alpha)} \qquad (19)$$

where $\widehat{C}_{i,x}(\alpha)$ is a CT observation used for fitting (12), and $\widetilde{C}^{(-k)}_{i,x}(\alpha)$ is the cycle-time predictor obtained at α based on the design set $\mathscr{A}\backslash\alpha$. Let $\text{WRE} = \max_\alpha RE_{1,x_0}(\alpha)$, and we terminate the procedure when $\text{WRE} < \gamma\%$.

6.2 An Example

In Yang, Ankenman, and Nelson (2007), the efficiency of the proposed methodology for generating CT-TH curves has been demonstrated through extensive numeric experiments. Thus the primary focus of this paper is on the modeling of the CT-PM surface $c_{i,x}(\alpha)$. Here, we illustrate the estimation of CT-PM model (12) through a simple Jackson queueing

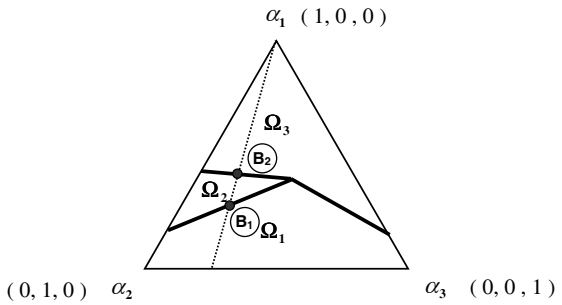

Figure 6: Division of the product-mix space and a constant-ratio product-mix path.

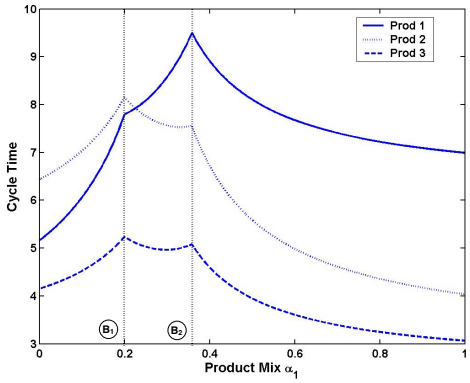

Figure 7: CT-PM curves.

model, for which the true CT-PM surface is known from queueing theory and hence provides a benchmark to evaluate the results obtained from our procedure. We consider a 3-station system that processes 3 different types of products. The system parameters are given in Figure 5. Figure 6 shows the division of product-mix space for this example. Each station can serve as a BN and the product-mix region is divided into 3 subregions with Ω_j being dominated by BN-station j ($j = 1,2,3$).

For Jackson queueing systems, the CT-PM surface (for a given x) is given by (7). Before we apply our procedure on this system, we first examine the CT-PM surface through (7). If we fix $\alpha_2 : \alpha_3 = 3 : 1$, and vary α_1 from 0 to 1, we obtain a PM path as the dotted line in Figure 6. Along this path, we plot $c_{i,0.8}(\alpha)$ ($i = 1,2,3$), the cycle time at throughput $x = 0.8$, against α_1, and the resulting CT-PM curves are given in Figure 7. Obviously in Figure 7, the CT-PM curves are smooth and differentiable except at BN-shift points B_1 and B_2, which are also marked in Figure 6. We can change the ratio of $\alpha_2 : \alpha_3$, and plot CT-PM curves similar to those obtained in Figure 7. This graphically demonstrates our conclusion in Section 5.3.1 that the CT-PM surface are smooth and differentiable in constant BN subregions, which motivate us to model each subregion Ω_j separately.

The proposed experiment design and model fitting methods have been applied on the three constant-BN subregions given $x = 0.8$. In our experiments, the desired estimation

precision $\gamma\%$ is set at 5%. Due to space constraint, we chose to present the estimation results for Ω_1, and the fitted and true CT-PM models are given in (20) and (21) respectively. The fitted model is able to approximate the true response surface to a desired precision (in this case, the maximum deviation of the fitted model from the true model is $3\% < 5\%$), although the estimated parameters are not the same as the parameters of the true surface model, as can be seen from comparing (20) and (21).

$$\begin{aligned} \widehat{c}_{1,0.8}(\overline{\alpha}) &= 5.2562 + \frac{5.9886\alpha_1 - 1.1809\alpha_2}{0.3579 - 0.9889\alpha_1 + 0.1487\alpha_2} \\ &+ \frac{-2.0462\alpha_1 + 0.8385\alpha_2}{0.4498 - 0.9944\alpha_1 - 0.1054\alpha_2} \end{aligned} \quad (20)$$

$$\begin{aligned} c_{1,0.8}(\overline{\alpha}) &= 5.1786 + \frac{1.1054\alpha_1 + 0.3685\alpha_2}{0.4514 - 0.9995\alpha_1 - 0.0322\alpha_2} \\ &+ \frac{2.0797\alpha_1 - 1.0399\alpha_2}{0.2496 - 0.9567\alpha_1 - 0.2912\alpha_2} \end{aligned} \quad (21)$$

The design strategy and model evaluation methods described above are simply one of many possible choices that we have. In future research, we will investigate and evaluate various such options. Experiments will also be performed on real manufacturing systems to demonstrate the efficiency of the proposed methodology.

ACKNOWLEDGMENTS

This research was supported by Semiconductor Research Corporation Grant 2004-OJ-1225. Additional thanks go to Professors John Fowler and Gerald Mackulak from Arizona State University.

REFERENCES

Hopp, W. J. and M. L. Spearman. 2001. *Factory Physics: Foundations of Manufacturing Management*. 2nd edition. Chicago: Irwin.

Myers, R. H. and D. C. Montgomery. 2002. *Response Surface Methodology: Process and Product Optimization Using Designed Experiment.* 2nd edition. Wiley-Interscience.

Piepel, G. F. 1988. Programs for generating extreme vertices and centroids of linearly constrained experimental regions. *Journal of Quality Technology* 20: 125–139.

Seber, G. A. F., and C. J. Wild. 2003. Nonlinear regression. Hoboken, New Jersey: John Wiley & Sons, Inc.

Yang, F., B. E. Ankenman, and B. L. Nelson. 2007. Efficient generation of cycle time-throughput curves through simulation and metamodeling. *Naval Research Logistics* 54: 78–93.

AUTHOR BIOGRAPHIES

FENG YANG is an assistant professor in the Industrial and Management Systems Engineering Department at West Virginia University. Her research interests include simulation and metamodeling, design of experiments, and applied statistics. Her e-mail and web addresses are <feng.yang@mail.wvu.edu> and <www2.cemr.wvu.edu/~yang/>.

JINGANG LIU is a PhD student in the Industrial and Management Systems Engineering Department at West Virginia University. His research work has been focused on simulation and metamodeling. His e-mail address is <jliu7@mix.wvu.edu>.

MUSTAFA TONGARLAK is a PhD student in the Department of Industrial and Management Sciences at Northwestern University. His research interests include simulation and metamodeling. His advisors are Prof. Barry L. Nelson and Prof. Bruce E. Ankenman. His email and web addresses are <m-tongarlak@northwestern.edu> and <users.iems.northwestern.edu/~mustafa>.

BRUCE E. ANKENMAN Bruce Ankenman received a BS in Electrical Engineering from Case Western Reserve University and an MS and PhD. in Industrial Engineering from the University of Wisconsin-Madison. Prior to his graduate work, he worked for five years as a design engineer for an automotive supplier in Ohio. He is currently an Associate Professor in the Industrial Engineering Department at Northwestern University. His research interests include the statistical design and analysis of experiments. He serves as a department editor for *IIE Transactions* and as an associate editor for *Naval Research Logistics*. His e-mail address is <ankenman@northwestern.edu>, and his web page is <www.iems.northwestern.edu/~bea>.

BARRY L. NELSON is the Charles Deering McCormick Professor of Industrial Engineering and Management Sciences at Northwestern University. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems. Nelson has served the profession as the Simulation Area Editor of *Operations Research* and President of the INFORMS (then TIMS) College on Simulation and is currently Editor in Chief of *Naval Research Logistics*. He has held many positions for the Winter Simulation Conference, including Program Chair in 1997 and current membership on the Board of Directors. His e-mail and web addresses are <nelsonb@northwestern.edu> and <www.iems.northwestern.edu/~nelsonb>.