

Transient Analysis of General Queueing Systems via Simulation-Based Transfer Function Modeling

Feng Yang*
Jingang Liu

July 15, 2011

Abstract

This paper is concerned with characterizing the transient behavior of general queueing systems, which is widely known to be notoriously difficult. The objective is to develop a statistical methodology, integrated with extensive offline simulation and preliminary queueing analysis, for the estimation of a small number of transfer function models (TFMs) that quantify the input-output dynamics of a general queueing system. The input here is the time-varying arrival rate of jobs to the system; the time-dependent output performances include the departure rate of jobs and the mean of the work in process (i.e., number of jobs in the system). The resulting TFMs are difference equations, like the discrete approximations of the ordinary differential equations provided by an analytical approach, while possessing the high fidelity of simulation. Our method is expected to overcome the shortcomings of the existing transient analysis approaches, i.e., the computational burden of simulation and the lack of fidelity of analytical queueing models.

Key words: metamodeling; discrete event simulation; transfer function modeling; design of experiments; queueing

1 Introduction

This paper is concerned with the transient behavior of queueing systems (Kleinrock 1975) which may arise in manufacturing and service industries.

1.1 Motivation

The primary motivation of this work stems from production planning in manufacturing, which can be loosely defined as the problem of finding a best release plan of jobs so that the ac-

tual outputs over time satisfy, as closely as possible, the predetermined requirements with the minimum cost (Missbauer and Uzsoy 2010). This definition clearly suggests that the ability to accurately quantify the relationship between the input release and the system output is the key to generating a good or near optimum plan. However, this input-output relationship has never been adequately addressed in existing production planning models, from the Material Requirement Planning (MRP) developed in the 1960's to the mathematical programming-based approaches that are widely used in industry and academia. This obvious flaw is due to the simple fact that such input-output relationships are notoriously difficult to quantify for manufacturing systems: They are not only nonlinear but also time-dependent due to the variability inherent in manufacturing processes (e.g., random machine failures and stochastic processing times). Although it has been recognized that manufacturing systems, which can be considered as general queueing networks (Whitt 1983; Hopp et al. 2002; Shanthikumar et al. 2007), exhibit nonlinear time-dependent (transient) behaviors (Papadopolous et al. 1993; Uzsoy et al. 1992; Zäpfel and Missbauer 1993; Riano 2003; Missbauer and Uzsoy 2010), no existing method is able to efficiently describe such nonlinear dynamics.

1.2 Literature Review

In the literature, both discrete-event simulation and analytical methods have been used to address the transient behavior of queueing systems. The former approach can incorporate, at least in theory, any details that are important to the system, but frequently becomes too computationally demanding for real-time “what-if” analysis. Analytical methods, on the other hand, are subject to restrictive assumptions (e.g., the Markov property) that may well not hold in practice, and thus fall short in providing an accurate description of real systems. These two approaches are briefly reviewed as follows.

For Markov queueing models, the mainstay of the analytical work has been the development and solving of the time-dependent ordinary differential equations (ODEs) which describe the system's dynamic behavior. Analytical solutions to these ODEs are rare with a few exceptions including the known solutions for the $M(t)/G/\infty$ and $M/M/1$ systems (Kleinrock 1975; Gross and Harris 1985), and the $Ph(t)/Ph(t)/\infty$ systems investigated by Nelson and Taaffe (2004a, 2004b). Substantial research effort has been devoted to developing numerical solutions of the ODEs, and Ingolfsson et al. (2007) provides a fairly complete review of these methods including

Rothkopf and Oren (1979), Clark (1981), Gross and Miller (1984), Taaffe and Ong (1987), Green and Kolesar (1991a, 1991b), Eick et al. (1993a, 1993b), Jennings et al. (1996), and Massey and Whitt (1997). Note that the construction of these ODEs relies on the Markov property, and numerically solving the typically large set of ODEs is also computationally challenging. Besides the ODEs-based method, other techniques for approximating the transient behavior of queues include fluid approximations, which are accurate when there is little variability, and diffusion models, which are good for heavily loaded systems (Chen and Mandelbaum 1994; Mandelbaum and Massey 1995; Kelly et al. 1996). The analytical method developed in Riano (2003) can be considered as a parallel to the fluid and diffusion approximations. All these methods can be roughly divided into two categories: those that are highly accurate but computationally intensive (comparable to detailed simulation), and those that are fast but inaccurate. Nevertheless, a common limitation of these methods is that they rely on analytical assumptions of one sort or another, and thus are inadequate to capture many features of realistic manufacturing systems such as non-Markovian interarrival/service times, server failures, re-entrant job flows, etc.

Computer simulation is an alternative approach to address the transient behavior of queueing systems because of its high fidelity and flexibility, and increasingly also because of its ease of use and wide acceptance among practitioners. The shortcoming of simulation is that many replication runs are required to obtain good estimates of time-dependent performance measures, and thus simulation is frequently too computationally expensive for real-time decision making.

In light of the discussions above, the existing methods fall short in terms of the ability to accurately and timely predict the transient behavior of a general queueing system that may involve features of realistic manufacturing systems.

1.3 Research Objectives and Contributions

The objective of this work is to develop a statistical methodology, integrated with extensive offline simulation and preliminary queueing analysis, to generate a number of transfer function models (TFMs) that can adequately characterize a general queueing system's transient behavior. The TFMs are difference equations quantifying the input-output dynamics of the system. The input to the TFMs is the arrival rate of jobs (i.e., the first moment of the arrival process), which may vary with time. The outputs of the TFMs include the departure rate of jobs from the system (i.e., the first moment of the departure process), and the mean of the number of

jobs in the system; both outputs are usually time-dependent. For a given system of interest, the proposed method assumes the availability of its discrete-event simulation (DES) model, and fully utilizes offline simulation time, which is typically plentiful in practice, to estimate a number of TFMs.

Our simulation-based transfer function modeling approach combines the advantages of both existing transient analysis methods, i.e., computer simulation and pure analytical methods, while avoiding their shortcomings. (i) The TFMs embody the high fidelity of simulation to real systems since they are estimated from detailed simulation data. (ii) The TFMs are difference equations, the discrete-time counterpart of the ODEs provided by an analytical approach; supposing that a certain input is fed to the system under given initial conditions, the TFMs can be used to recursively compute the system's future output performance in a timely manner. Hence, the TFMs resulting from our method are able to accurately describe the transient behavior of realistic systems and, at the same time, they allow for prompt "what-if" analysis. These advantages make our TFMs-based approach a solid and sound basis to support responsive decision making such as production planning (or re-planning) in manufacturing.

The proposed approach falls into the category of metamodeling (Chapter 18 in Henderson and Nelson 2006), which refers to the techniques that utilize simulation to generate mathematical approximations quantifying the relationships implied by the simulation. However, to the best of our knowledge, this is the first attempt to develop a metamodel that takes the form of difference equations. Our approach is suitable to application contexts where metamodeling can realize the maximum potential. Such a context is articulated in Ankenman et al. (2010) as follows: the time to exercise the simulation model in advance of the decision making is relatively plentiful, whereas the decision-making or decision-maker's time is relatively scarce or expensive. Responsive production planning represents one such context: The simulation model of a manufacturing system can be developed and kept running for weeks (or even months) as soon as the system configuration has been established; while in case of production disruptions such as supplier failures, demand-forecast mismatches and natural disasters (Sheffi and Rice 2005; Datta et al. 2007; Pinedo 2007; Stadtler and Kilger 2007), a decision needs to be made quickly—as soon as possible—on how to adjust the production plan for that system. The metamodel, i.e., the TFMs in this work, fully utilizes the plentiful offline simulation time and allows for responsive decision making in time of urgency.

1.4 Applicability and Limitations

This paper focuses on characterizing a system (or subsystem) by a single set of TFMs, for which the inputs/outputs involved are all system-level variables: The TFMs take as input the arrival rate of jobs to the system, and output the departure rate of jobs from the system and the expected number of jobs in the system. As a metamodeling approach, the method developed in this paper is applicable to approximate the transient behavior of a queueing system that satisfies the following three conditions.

Constraint 1: The system involves a single class of jobs. Although preliminary work suggests that the metamodeling approach can be extended to multi-class systems, a full investigation of which is reserved for future research.

Constraint 2: The arrival process does not depend on the state of the system. This push, as opposed to pull, logic (Hopp 2007) is commonly assumed in the existing production planning models (Missbauer and Uzsoy 2010) and also frequently encountered in service systems.

Constraint 3: The third limitation also lies in the requirement on the arrival process to the system, and it stems from the functional form of the proposed TFMs: The arrival rate is the sole input taken by the TFMs for the prediction of a system's output evolution. For such TFMs to adequately describe system performance, the assumption needed is that, loosely speaking, the arrivals can be fully characterized by its rate (first moment measure). More specifically, our method is applicable to two types of system arrivals:

- (a) A non-stationary Poisson process (NSPP) whose arrival rate can be any function of time. An NSPP is known to be completely characterized by its time-varying rate, and it has the memoryless property.
- (b) A non-stationary non-Poisson process (NSNP) whose arrival rate can be any function of time, and that satisfies the conditions given below.

Describing an NSNP itself is not a trivial task (Daley and Vere-Jones 2002; McKenzie 2003), and we first sketch out some background. Recall that this metamodeling work is built on the availability of a DES model that captures the real system of interest with high fidelity. The TFMs are estimated from simulation data, and thus are directly associated with the simulated arrivals which are modeled from real data. Modeling NSNP arrivals is a challenging research topic beyond the scope of this paper, and here, we simply discuss the

conditions required by our metamodeling on NSNP arrivals in the context of simulation input modeling. To derive characteristic features of real NSNP arrivals, it is a common practice to divide the period of concern into a number of time intervals (Leemis 1991; Harrod and Kelton 2006; Gerhardt and Nelson 2009). Denote the number of arrivals in the i^{th} interval as A_i . The sequence $\{A_i; i = 1, 2, \dots\}$ constitutes a discrete-variate time series, which is typically characterized by the marginal distribution (MD) of A_i , denoted as Π_i , and the covariance between A_i and A_j , denoted as $\text{Cov}[A_i, A_j]$ with $i, j = 1, 2, \dots$ and $i \neq j$. From real data, the MDs (or statistics of the MDs) and correlation structure can be derived to build the appropriate simulation model for arrivals. The time discretization for arrivals aligns with the discrete-time TFMs. Inheriting the notations/practices above, our TFMs-based metamodeling imposes the following conditions on an NSNP, in order for it to be fully characterized by its arrival rate over time.

- (i) For any interval pair i and j ($i \neq j$), the difference in the MDs Π_i and Π_j is completely captured by $E[A_i]$ and $E[A_j]$, the first-moment measures that can be interpreted as the arrival rates per time interval. We provide two examples. The simplest is one where Π_j differs from Π_i only by a horizontal shift of $E[A_j] - E[A_i]$. In the second example, suppose that the MDs follow negative binomial distribution with two parameters: the stopping-time parameter r and the probability of success p . Varying p while holding r constant across time intervals leads to a sequence of MDs that satisfy the condition given here.
- (ii) For positive integers i and j ($i \neq j$), $\text{Cov}[A_i, A_{i+j}]$ decays quickly to 0 as j increases, and can be typically considered 0 for $j > 2$ as the covariance (or autocorrelation) for continuous-variate autoregressive time series (Box et al. 2008); plus, for i and j that have $\text{Cov}[A_i, A_{i+j}] \neq 0$, $\text{Cov}[A_i, A_{i+j}]$ can be approximated as an i -independent and j -dependent function of $E[A_i]$ and $E[A_{i+j}]$, the arrival rates. The simplest case that satisfies this condition is a discrete-variate time process with $\text{Cov}[A_i, A_{i+j}]$ being 0 for any i and j .

Apparently, the NSPP arrivals exactly apply to the modeling of TFMs with the arrival rate being the sole input. As to NSNP, the conditions (i) and (ii) above depict a non-stationary discrete-variate time series whose autocorrelation decays quickly and whose major statistical

characteristics can be fully captured by the time-varying arrival rate. Such arrivals represent a fair proportion of the non-Poisson arrivals that are modeled/simulated by the existing literature (McKenzie 2003; Gerhardt and Nelson 2009). Section 4.2.1 provides the algorithms used to simulate the arrivals in this paper.

Aside from the three constraints above, an empirical recommendation is made here as to the system configuration suitable for the transfer function modeling. A single set of TFMs, with inputs and outputs both given at the system level, is most suitable to approximate the transient behavior of a system that is dominated by one bottleneck (BN) station. Specific definitions and discussions in this regard will be given in Section 6.2.2.

The remainder of the paper is organized as follows. Section 2 provides an overview of the proposed method. The preliminary queueing analysis is performed in Section 3. The metamodeling approach, i.e., the simulation-based transfer function modeling, is described in Sections 4 and 5. In Section 6, the proposed method is applied to a range of queueing systems including a scale-down semiconductor manufacturing system, and the empirical results are provided.

2 Overview of the Methodology

We consider the system of interest as a queueing system that involves three major time-dependent processes.

$Q(t)$: the state process representing the number of jobs in the system at time t , with $t \in (-\infty, \infty)$, the whole time axis.

$A(u, v)$: the random variable counting the number of arrivals in the system within the time interval $(u, v]$; $u, v \in (-\infty, \infty)$ and $u < v$.

$D(u, v)$: the random variable counting the number of departures in the system within the time interval $(u, v]$; $u, v \in (-\infty, \infty)$ and $u < v$.

Let $\mathcal{H}_0 = \{Q(t), A(-\infty, t), D(-\infty, t); t \in (-\infty, 0]\}$ denote the history of the system evolution up to time 0. The question we intend to address here is: Standing at time 0, how do we predict the system's behavior from time 0 onward given the history \mathcal{H}_0 ? In a push system where arrivals are independent of the system status (Section 1.4), $\{A(0, t); t > 0\}$ is considered as the imposed input flow, and $\{Q(t), D(0, t); t > 0\}$ the dependent variables representing the

output performance of the system. The objective of this work is to establish the time-dependent relationship between the first moment measures of these three processes, which are defined as follows.

$m(t) = \mathbb{E}[Q(t)]$, the first moment of the number of jobs in the system at time t .

$x(t) = \lim_{\delta \rightarrow 0^+} \delta^{-1} \mathbb{E}[A(t, t + \delta)]$.

$d(t) = \lim_{\delta \rightarrow 0^+} \delta^{-1} \mathbb{E}[D(t, t + \delta)]$.

It is assumed that $x(t)$ and $d(t)$ exist and are finite. Usually, $x(t)$ and $d(t)$ are also referred to as the intensity or rate of the corresponding point process. In this work, $x(t)$ is the input variable that is able to fully characterize the arrival process (Section 1.4), and $\mathbf{y}(t) = (m(t), d(t))$ is the 2×1 vector including the output performance variables. The objective is to capture the input-output dynamics of a queueing system by the TFMs:

$$\mathbf{y}(t) = \mathbf{F}(x(t-1), x(t-2), \dots, \mathbf{y}(t-1), \mathbf{y}(t-2), \dots), \quad (1)$$

which is a discrete-time functional approximation. The time t in (1) denotes discrete time points. In the rest of this paper, t will be used to represent both continuous and discrete time index without causing any confusion.

The vector function \mathbf{F} in the TFMs (1) includes two functions, and is of the same dimension as $\mathbf{y}(t)$. Each component of \mathbf{F} is a difference equation relating an output performance at time t to the input and output history of the system. Suppose that we stand at the current time 0 and that the future time horizon is $(0, P]$. Given the seed values of $\{x(t), \mathbf{y}(t), t \leq 0\}$, which can be derived from \mathcal{H}_0 , the TFMs are able to compute recursively the system's future performance $\{\mathbf{y}(t), t \in (0, P]\}$ under any input $\{x^*(t), t \in (0, P]\}$.

It is difficult to obtain the TFMs that can accurately characterize the transient dynamics of a general queueing system, and our method is three fold.

- *Queueing analysis* (Section 3): We perform queueing analysis under fairly general assumptions. Such a theoretical analysis, although inadequate to address the time-dependent behavior of realistic systems, sheds lights on the functional forms of the TFMs.
- *Data collection via offline simulation* (Section 4): Under selected input processes, we run simulations to obtain paired input-output time series for the fitting of the TFMs. We

emphasize that our simulation is carried out offline in advance of the need to make a decision.

- *Transfer function modeling* (Section 5): From the simulation data, we develop statistical methods to obtain the parsimonious TFMs (1) that are adequate to capture the system’s dynamic behavior.

3 Non-Stationary Queueing Analysis

In this part, we perform analytical analysis on some simple queueing systems to gain insights to their non-stationary behavior. These analytical results are also what primarily motivated the functional forms of the TFMs.

3.1 An $M(t)/M/\infty$ Example

For the purpose of intuition and motivation, we consider the input-output dynamics of the simple queueing model $M(t)/M/\infty$, which is one of the very few models whose transient behavior can be characterized analytically. Suppose that the service rate for each job is μ . From the Kolmogorov forward equations for the state probabilities (Ross 1995), we can easily derive the following equations for the $M(t)/M/\infty$:

$$\begin{aligned} m'(t) &= dm(t)/dt = x(t) - \mu \cdot m(t) \\ d(t) &= \mu \cdot m(t). \end{aligned} \tag{2}$$

These ODEs describe the evolution of the outputs $\mathbf{y}(t) = (d(t), m(t))$ as functions of $x(t)$. Given the initial state of the system at time 0, the numeric solution of $\{\mathbf{y}(t); t > 0\}$ can be obtained for any given input $\{x^*(t); t > 0\}$.

Unfortunately, the situation becomes much more complicated as soon as a finite number of servers is introduced or the Markovian assumption is relaxed. Closed (or solvable) ODEs like (2) cannot be obtained even for single-station systems such as $M(t)/M/s$ and $M(t)/G/s$ (Rothkopf and Oren 1979).

3.2 A General Queue

In this subsection, we perform non-stationary analysis on a general queue to understand the transient behavior of non-Markovian systems and to obtain a theoretical basis for the ability of TFMs (1) to model general queueing systems.

We consider a single-server system with the service time following a general distribution, say $G(\tau)$, where $\tau \in (\tau_L, \tau_U)$ with $0 \leq \tau_L < \tau_U \leq \infty$. Jobs are served on a first come first served basis. The arrivals are independent of the system state. The additional assumptions made solely for the analytical analysis of this subsection are: Both the arrival and departure processes are orderly point processes (Daley and Vere-Jones 2002), the condition for which can be mathematically stated as

$$\Pr\{A(t, t + \delta) > 1\} = o(\delta); \quad \Pr\{D(t, t + \delta) > 1\} = o(\delta). \quad (3)$$

The arrivals $A(u, v)$ and departures $D(u, v)$ are as defined in Section 2, along with the state process $Q(t)$. Under condition (3), the first moment measures of the arrivals and departures can be written as

$$x(t) = \lim_{\delta \rightarrow 0^+} \delta^{-1} \Pr\{A(t, t + \delta) = 1\}; \quad d(t) = \lim_{\delta \rightarrow 0^+} \delta^{-1} \Pr\{D(t, t + \delta) = 1\}. \quad (4)$$

For such a single-server queue we seek to derive the dependence of $\mathbf{y}(t) = (d(t), m(t))$ upon $x(t)$, as we did for M(t)/M/ ∞ in Section 3.1.

Following the notation given in Section 2, let

$$\begin{aligned} p_n(t) &= \Pr\{Q(t) = n\} \\ x_n(t) &= \lim_{\delta \rightarrow 0^+} \delta^{-1} \Pr\{A(t, t + \delta) = 1, Q(t) = n\} \\ d_n(t) &= \lim_{\delta \rightarrow 0^+} \delta^{-1} \Pr\{D(t, t + \delta) = 1, Q(t) = n\}. \end{aligned}$$

Here, $x_n(t)$ denotes the arrival rate at time t while there are n jobs in the system, and $d_n(t)$ represents the departure rate at time t with n jobs (including the one that is about to leave) in the system. In Appendix A.1, we have derived the dynamic equations for the $x(t)$ - $\mathbf{y}(t)$ relationship, which is given as follows:

$$\begin{aligned} m'(t) &= x(t) - d(t) \\ d(t) &= \int_{\tau_L}^{\tau_U} x_0(t - \tau) dG(\tau) + \int_{\tau_L}^{\tau_U} (d(t - \tau) - d_1(t - \tau)) dG(\tau) \end{aligned} \quad (5)$$

Unlike equations (2) for the $M(t)/M/\infty$ system, (5) for the general queue are not closed, and thus not solvable: Aside from the input $x(t)$ and the outputs of interest $m(t)$ and $d(t)$, (5) also involves unknown time-dependent functions $x_0(t)$ and $d_1(t)$. However, for mediumly/heavily loaded queues, it is reasonable to assume that $x_0(t)$, the arrival rate when no job in the system, and $d_1(t)$, the departure rate when no job is in the waiting queue, are relatively small and can be approximated by:

$$x_0(t) \approx p_0(t) \times x(t) \approx e_1 x(t) \text{ and } d_1(t) \approx p_1(t) \times d(t) \approx e_2 d(t), \quad (6)$$

with e_1 and e_2 being small fractional constants. If we plug (6) into (5) and take the finite-difference approximation of the derivative and integrals in (5), it is clear that the discrete approximations of (5) fall into the category of TFMs (1).

Replacing the single server by multiple identical servers in the general queue investigated above, we have also derived equations of the same nature as (5) and of more complicated functional forms. The analytical results obtained for these single-station queues are what motivated us to adopt the TFMs (1) in the first place. With general arrivals and services, the single-station systems considered in this subsection provide a good representation of an aggregate queueing system. The $x(t)$ - $\mathbf{y}(t)$ equations derived for these queues, along with the ODEs given by (2), strongly suggest that the TFMs as (1) have the potential to provide discrete-time approximations adequate to describe the aggregation performance of a real queueing system.

4 Data Collection via Offline Simulation

For a given queueing system, we assume the availability of its DES model, and seek to estimate from simulation data a set of TFMs describing that system's transient behavior. In this part, we are concerned with collecting via simulation the input-output data needed for metamodeling. However, as will become clear later in this section, to specify the simulation experiments, an upfront capacity/bottleneck analysis is necessary (especially for a system consisting of multiple stations) and thus briefly described in Section 4.1 before simulation-based data collection is discussed.

4.1 Capacity and Bottleneck Analysis of a System

For a system consisting of, say J stations, existing analytical models in the literature can be used to (i) obtain the system capacity and (ii) identify the bottleneck (BN) station(s). The capacity of station j ($j = 1, 2, \dots, J$) is defined as the upper limit on that station's arrival rate for long-term stability, and is denoted as μ_j . The capacity of the system is defined as $\mu = \min\{\mu_j; j = 1, 2, \dots, J\}$; and the station j_{BN} that achieves the system capacity is called the BN station, which is the most heavily utilized station in the system.

The analytical models that can provide such capacity/bottleneck analysis include Kumar and Kumar (2001), Hopp et al. (2002), Meng and Heragu (2004), etc. These models can accurately or even exactly calculate the station capacity $\{\mu_j; j = 1, 2, \dots, J\}$ for real systems that involve batching, re-entrant flows, and machine failures and setups. In this paper, the analytical model developed in Hopp et al. (2002) is used to identify the capacity and BN station(s) for a system of interest, before any simulation experiments are performed.

4.2 Simulation

For a given system, its DES model is assumed available to our metamodeling and will be run for data collection. Section 4.2.1 briefly discusses the simulation algorithms involved in the DES models, and Section 4.2.2 describes how the sample data are obtained from simulation.

4.2.1 Stochastic Simulation Algorithms

A DES model representing a real queueing system can be considered as comprised of two algorithms: the one that mimics the system configuration and logic, and the one that simulates real arrivals. The former algorithm is known to be able to accommodate any complexities inherent in a real system (Law and Kelton 2003), and is implemented in Microsoft C++ for the empirical examples provided in Section 6.

We next discuss the latter algorithm for simulating input arrivals. In this work, the arrivals to the system are allowed to be an NSPP as well as a restricted NSNP (Section 1.4). In our simulation experiments, an NSPP is generated using the thinning algorithm proposed by Lewis and Shedler (1979), and it applies exactly to the modeling of TFMs, as pointed out in Section 1.4; NSNP arrivals, on the other hand, are generated by the algorithms in Gerhardt and

Nelson (2009) and deserve more discussions below. In Gerhardt and Nelson (2009), a complete set of input modeling methods was proposed: Techniques were developed to derive the first two-moment features of real arrivals; algorithms were developed to generate an NSNP that mimics the real data by transforming a stationary renewal process. In their paper, it has been shown that the pre-specified time-varying arrival rate (that is, the rate estimated from the real data) can be achieved by the simulated NSNP, and the variability of the NSNP can be controlled to reflect that of the real arrivals by controlling the coefficient of variation (CV) of the renewal base process.

Recall that for an NSNP to be eligible for our metamodeling, it has to satisfy the two conditions specified in Constraint 3, Section 1.4. Through Monte Carlo simulation, we have empirically verified that those conditions are satisfied by the NSNP arrivals generated by the algorithms in Gerhardt and Nelson (2009). Specifically, following the notations in Section 1.4, (i) the second moment of the marginal distribution of A_i , the number of arrivals in the i^{th} time interval, can be approximated as a function of the arrival rate $E[A_i]$; (ii) $Cov[A_i, A_{i+j}]$ decays exponentially with the increase of j , and can be approximated as a function of $E[A_i]$ and $E[A_j]$. Thus, the generated NSNP arrivals are eligible to the TFMs-based modeling, which is also evident from the metamodeling results presented in Section 6.1. Lastly, it is worth mentioning that McKenzie (2003) also discusses the modeling of $\{A_1, A_2, \dots\}$ based on discrete-variate time series, which can lead to simulation algorithms for non-Poisson arrivals eligible to our metamodeling.

4.2.2 Sampling via Simulation

Given that arrivals are adequately characterized by its first moment measure $x(t)$, Section 4.3 will discuss the specification of $x(t)$ in our simulation experiments. Here, assuming a selected $x(t)$, we describe how sample data are collected from, say I , simulation replications under the specified (and likely random) input flow. During each replication, sample data are taken at discrete, equispaced intervals of time, and the basic sampling interval is denoted as Δt . The sampling interval Δt is a user-specified parameter, and is selected out of the following considerations: Δt should be sufficiently small to allow all the systematic variation occurred in the inputs/outputs to be taken account of, and at the same time be large enough to smooth out some random variations. We recommend setting Δt as $1/10 \sim 1/5$ of $1/\mu$, where μ is the system capacity

defined in Section 4.1.

Let the sampling interval Δt serve as the unit of time. The simulation length of each replication is denoted as T time units, and T is a positive integer whose value will be specified in Section 4.3.2. For replication i ($i = 1, 2, \dots, I$), the arrival, departure and state processes $\{A_i(t - 0.5, t + 0.5), D_i(t - 0.5, t + 0.5), Q_i(t); t = 1, 2, \dots, T\}$ are recorded; note that the 0.5 here represents half of the time unit. From the I simulation replications, the paired time series $\{(X(t), \mathbf{Y}(t)), t = 1, 2, \dots, T\}$ are estimated as follows.

$$\begin{aligned} X(t) &= \hat{x}(t) = I^{-1} \sum_{i=1}^I A_i(t - 0.5, t + 0.5) \\ Y_1(t) &= \hat{m}(t) = I^{-1} \sum_{i=1}^I Q_i(t) \\ Y_2(t) &= \hat{d}(t) = I^{-1} \sum_{i=1}^I D_i(t - 0.5, t + 0.5) \end{aligned} \tag{7}$$

Both the arrival rate $X(t)$ and the departure rate $Y_2(t)$ are given in terms of the average number of occurrences per time unit (i.e., per Δt).

4.3 Design of Simulation Experiments

In this part, we discuss the design of experiments issues, which in our metamodeling work amount to specifying the input rate function $x(t)$ and determining the number of replications I for the simulation experiments.

To collect simulation data for the TFMs estimation, we adopt a piecewise constant (PWC) form for the rate function $x(t)$ out of the following considerations. The TFMs are required to be able to describe the system's steady-state behavior with the input $x(t)$ held unchanged for $t \rightarrow \infty$ (as will be articulated in Section 5.1). A PWC input $x(t)$ can provide some steady-state simulation data necessary to ensure the steady-state convergence of the fitted TFMs. It is worth pointing out that the PWC arrival rate is for the simulation performed to estimate the TFMs only; once the TFMs have been obtained, they can be used to predict the system's dynamics under an input rate of arbitrary forms (including continuously changing functions).

4.3.1 Input Range of Interest

For a PWC input rate $x(t)$, we denote its distinct constant levels as $\{x_1, x_2, \dots, x_M\}$, with $x_m \in [x_L, x_U]$. Here we intend to establish $[x_L, x_U]$, the range of input rates for the simulation experiments carried out for the TFMs estimation. For convenience of discussions, the concept of utilization is first introduced as follows. Given a system, the system capacity μ can be obtained using preliminary analytical methods (Section 4.1). Suppose that the arrival rate to the system is x_m , then the long-run system utilization is $\rho = x_m/\mu$ representing the fraction of busy time.

A queueing system designed for a real-world application is typically mediumly/heavily utilized the vast majority of the time (if not all the time). For instance, semiconductor manufacturing systems are operated 24/7 with a high utilization. Also, due to queueing effects, a system's performance under high utilizations is usually of much more concern than that under low utilizations. Thus, in this work, particular interest is given to the performance of a system over a range of relatively high utilizations (≥ 0.5). Denote the utilization range as $[\rho_L, \rho_U]$. The values of ρ_L and ρ_U are specified by the user depending on the real system's utilization in practice. In this paper, we set $[\rho_L, \rho_U] = [0.5, 0.95]$, which is suitable for semiconductor manufacturing. Given the system capacity μ and the selected utilization range $[\rho_L, \rho_U]$, the range of the input rate is set as

$$[x_L, x_U] = \mu[\rho_L, \rho_U]. \quad (8)$$

4.3.2 Experimental Design

For the TFMs modeling, the design of experiments (DOE) needs to specify the rate $x(t)$ of the input process, which is a PWC function like that in Figure 1(a), and the number of simulation replications I . Specifically, there are four questions that need to be addressed.

First, how should $\{x_1, x_2, \dots, x_M\}$ be selected? This includes determining the size M and the value of x_m ($m = 1, 2, \dots, M$) with $x_m \in [x_L, x_U]$. The selection here is guided by the system's steady-state behavior, which has been thoroughly investigated in the literature and a recent sequence of papers by Yang and co-authors (Yang et al. 2007, Yang 2010). Based on their extensive empirical experience, five different arrival rates, approximately evenly spread over mediumly/heavily utilized range $[x_L, x_U]$, are sufficient to characterize the steady-state input-output relationships. Thus, in our simulation experiments, function $x(t)$ is set as one with

$M = 5$ constant levels over the range $[x_L, x_U]$:

$$x_L, x_L + (x_U - x_L)/4, x_L + 2(x_U - x_L)/4, x_L + 3(x_U - x_L)/4, x_U. \quad (9)$$

Second, how to determine the sequence of these five arrival rates? It is desirable to be able to examine the interaction effects of $x(t)$ and $\mathbf{y}(t) = (m(t), d(t))$. For instance, does the same $x(t)$ cause $m(t)$ to respond in a different manner when $m(t)$ is at different values? In light of this, we determine the sequence of $\{x_1, x_2, \dots, x_5\}$ in such a way that the minimum jumps (up or down) between the successive levels of arrival rates is maximized. Initially at time 0^- , the system is empty and the arrival rate is $x_{0^-} = 0$, and the sequence $\{x_1, x_2, \dots, x_5\}$ is determined by

$$\max_{x_1, x_2, \dots, x_5} \min\{|x_m - x_{m-1}|, i = 1, 2, \dots, 5\} \quad (10)$$

The solution to (10) can be easily obtained by permutating the five different levels.

Third, how to determine the simulation length ℓ_m at each input level x_m ($m = 1, 2, \dots, M$)? We set $\ell_m = 2\ell_{tr}$, and ℓ_{tr} is the length of transient period for the initially empty system under the average input level $x_L + (x_U - x_L)/2$. The transient period ℓ_{tr} is determined based on preliminary simulation experiments following the methods in Chapter 9 of Law and Kelton (2000). Note that the length ℓ_m is selected to ensure that sufficient data is collected in steady state for reasons already mentioned. The total length of a simulation replication T is given as $T = \sum_{m=1}^M \ell_m$.

Fourth, how many replications should be performed at the selected time-varying input process? We use $\gamma\%$, the desired precision of $Y_1(t) = \widehat{m}(t)$ in equations (7) to determine the number of replications through a two-step process. In the first step, I_0 replications are performed under the specified input $x(t)$. Denote $\widehat{m}_0(t)$ as the estimate from the I_0 replications. The maximum sample standard deviation of $\widehat{m}_0(t)$ over the simulation length, say $[0, T]$, is given as:

$$\widehat{\sigma}_{\max}^{(0)} = \max_{t=1, 2, \dots, T} \widehat{\sigma}(\widehat{m}_0(t)) \quad (11)$$

Let t_{\max} be the time index that achieves $\widehat{\sigma}_{\max}^{(0)}$. The number of replications I that is likely to provide the desired precision $\gamma\%$ is estimated as: $n = \lceil (\widehat{\sigma}_{\max}^{(0)})^2 / (\widehat{m}_0(t_{\max}) \times \gamma\%)^2 \rceil$. In the second step, $I - I_0$ simulation replications are performed, and based on all the I replications carried out, the time series estimates are obtained following equations (7) and will be used for the TFMs fitting discussed in Section 5.

To conclude this section, we remind the readers that the goal of DOE is to provide representative sample data so that the TFMs fitted from the data can accurately predict the performance

of a real system. Thus, the input $x(t)$ used in the sampling should be specified based on the real arrivals to the system. The DOE strategies proposed here are particularly suitable to manufacturing. If the real arrivals for some other application differ substantially from those in manufacturing, then the DOE should be adjusted accordingly. For instance, as will be mentioned later in Section 6.2.3, if a system's overloaded behavior is of particular interest, then in the DOE, $x(t)$ should include arrival rates greater than the system capacity.

5 Statistical Modeling Issues of the TFMs

The modeling of the system dynamic behavior is based on the pair estimates $\{X(t), \mathbf{Y}(t), t = 1, 2, \dots, T\}$ obtained from simulation experiments. These estimates are subject to random errors, and we use the following parametric model to represent the stochastic correspondent of the TFMs (1):

$$\mathbf{Y}(t) = \mathbf{F}(\boldsymbol{\theta}; X(t-1), X(t-2), \dots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \dots) + \mathbf{e}(t), \quad (12)$$

where $X(t) = \hat{x}(t)$ and $\mathbf{Y}(t) = (\hat{m}(t), \hat{d}(t))$ as given in (7). The term $\mathbf{e}(t) = (e_1(t), e_2(t))$ denotes the disturbance. The parameter vector $\boldsymbol{\theta}$ includes all the unknown parameters involved in the vector function \mathbf{F} . For convenience of the discussion, we also write model (12) as:

$$\begin{aligned} Y_1(t) &= F_1(\boldsymbol{\theta}_1; X(t-1), X(t-2), \dots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \dots) + e_1(t) \\ Y_2(t) &= F_2(\boldsymbol{\theta}_2; X(t-1), X(t-2), \dots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \dots) + e_2(t), \end{aligned} \quad (13)$$

with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Our task is to obtain the TFMs that are of the simplest functional form and adequate to describe the system's dynamic evolution based on the paired simulation data $(X(t), \mathbf{Y}(t))$.

5.1 Estimation of the TFMs

In this part, we discuss the fitting of the TFMs assuming that a specific functional form (model structure) has been obtained.

5.1.1 Error Term

As a generalization of the i.i.d (identically and independently distributed) assumption made on error terms in regular regression (or regular transfer function modeling), the correlation among

the temporal errors $\{e_i(t); t = 1, 2, 3, \dots\}$ is accommodated in this work (for $i = 1, 2$). The temporal errors are modeled by a stationary autoregressive moving average (ARMA) process (Box et al. 2008), which can be expressed as

$$e_i(t) = \frac{C_i(q)}{D_i(q)} w_i(t) \quad i = 1, 2. \quad (14)$$

The white noise $w_i(t)$ is normally distributed with a mean of zero and a variance of σ_i^2 . The backward shift operator q^{-1} is defined by $q^{-1}z(t) = z(t - 1)$, and we have $q^{-m}z(t) = z(t - m)$. The operators $C_i(q)$ and $D_i(q)$ are defined as:

$$C_i(q) = \sum_{k=0}^{\infty} c_i(k)q^{-k} \quad (15)$$

$$D_i(q) = \sum_{k=0}^{\infty} d_i(k)q^{-k} \quad i = 1, 2. \quad (16)$$

For an ARMA process, the coefficient $c_i(k)$ (or $d_i(k)$) is typically zero for $k \geq 2$ (Box et al. 2008).

5.1.2 Fitting Process of the TFMs

Since the error terms $\{e_i(t), i = 1, 2\}$ are not i.i.d normal, we propose the following process to estimate the TFMs.

Step 0: Fit the TFMs to the $\{X(t), \mathbf{Y}(t), t = 1, 2, \dots, T\}$ data using least-square methods (Ljung 1999) as if the errors were i.i.d normal. Based on the fitted TFMs, compute the residuals $\{\hat{e}_i(t), i = 1, 2\}$. Identify the most appropriate ARMA structure (say, AR(1)) for $\hat{e}_i(t)$ following the graphical approach in Chapter 3 of Box et al. (2008).

Step 1: Fit the ARMA model of the identified structure to residuals $\{\hat{e}_i(t), i = 1, 2\}$ using Matlab function GARCHFIT. Note that the fitted ARMA model for residuals completely specifies $C_i(q)$ in (15) and $D_i(q)$ in (16), and also leads to the estimated variance of the white noise $w_i(t)$, denoted as $\hat{\sigma}_i^2$.

Step 2: Use the least square methods to refit the TFMs assuming that the errors are given as the ARMA models obtained from the previous step. Specifically, the TFMs (13) will be transformed as follows to achieve additive white noise $w_i(t)/\sigma_i$ ($i = 1, 2$) with constant

variance 1.

$$\frac{D_1(q)}{\sigma_1 C_1(q)} Y_1(t) = \frac{D_1(q)}{\sigma_1 C_1(q)} F_1(\boldsymbol{\theta}_1; X(t-1), X(t-2), \dots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \dots) + \frac{w_1(t)}{\sigma_1} \quad (17)$$

$$\frac{D_2(q)}{\sigma_2 C_2(q)} Y_2(t) = \frac{D_2(q)}{\sigma_2 C_2(q)} F_2(\boldsymbol{\theta}_2; X(t-1), X(t-2), \dots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \dots) + \frac{w_2(t)}{\sigma_2} \quad (18)$$

The least-square fitted parameters $\hat{\boldsymbol{\theta}}$ is the solution to the following optimization problem:

$$\begin{aligned} & \min_{\boldsymbol{\theta}} \text{SSE}(\boldsymbol{\theta}) \\ & = \sum_{t=1}^T \sum_{i=1}^2 \left[\frac{D_i(q)}{\sigma_i C_i(q)} Y_i(t) - \frac{D_i(q)}{\sigma_i C_i(q)} F_i(\boldsymbol{\theta}_i; X(t-1), X(t-2), \dots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \dots) \right]^2, \end{aligned}$$

given time-series data $\{X(t), \mathbf{Y}(t), t = 1, 2, \dots, T\}$. With the newly fitted TFMs, the residuals are computed and updated.

Steps 1 and 2 can be repeated until there is no significant changes in the resulting TFMs. In our experience, it suffices to perform one round of the fitting process to achieve well-estimated TFMs.

5.1.3 Other Modeling Issues

Stability

The dynamic TFMs should be stable: The TFMs should be able to converge to the system's steady-state input-output relationships. Suppose that the input rate $x(t)$ is held constant at the fixed level x . We write the outputs as $\mathbf{y}(x, t)$ emphasizing the dependence of the outputs on time t as well as on x . With the input held fixed at x , the dynamic outputs described by the TFMs should eventually converge to $\mathbf{y}(x, \infty)$, the steady-state equilibrium. To ensure the stability of the resulting TFMs, in our method, the simulation data are collected in such a way that a substantial amount of steady-state time series (Section 4.3.2) are included. Hence, the TFMs fitted from the data also well reflect the system's steady-state behavior.

Statistical Inference

In Ljung (1999), the asymptotic normality of the least-square parameters $\hat{\boldsymbol{\theta}}$ has been proved, and the statistical inference on the estimated TFMs is discussed in Appendix A.2.

5.2 Model Selection

The estimation of TFMs in Section 5.1 is based on a given functional form. In this section, we discuss the selection of the most appropriate structure for the target model.

5.2.1 Identification of the Model Family

Achieving the parsimonious TFMs that can accurately describe the system's transient performance is difficult, and we resort to a number of venues in search of the model family for TFMs.

Transient Queueing Analysis

As pointed out in Section 3, which is devoted to non-stationary queueing analysis, the analytical results obtained for some simple queueing systems are what primarily motivated the development of the TFMs. Equations (2) for the $M(t)/M/\infty$ and equations (5) for the general single-server queue both suggest that the time-dependent difference equations (1) have the potential to provide an accurate description of the system behavior in terms of $m(t)$, the expected number of jobs in the system, and $d(t)$, the departure rate from the system.

In addition, the transient queueing analysis in Section 3.2 suggests that the simplest possible form for the TFMs is likely to be:

$$\begin{aligned}y_1(t) &= b_0 + b_1x(t-1) + b_2y_1(t-1) + b_3y_2(t-1) \\y_2(t) &= c_0 + c_1x(t-1) + c_2y_2(t-1).\end{aligned}$$

Recall that $(y_1(t), y_2(t)) = (m(t), d(t))$.

Steady-State Behavior

The previous study performed in Yang et al. (2007) suggests that models of the following form can adequately approximate the steady-state behavior of real manufacturing systems.

$$y_1(\infty) \approx \frac{\mathbf{P}(x/\mu)}{1 - x/\mu} \approx \mathbf{P}(x/\mu)[1 + x/\mu + (x/\mu)^2 + \dots] \quad (19)$$

$$y_2(\infty) = x \quad (20)$$

Here, x denotes the constant rate of arrivals into the system, and μ the system capacity (Section 4.1). The stability condition requires: $x/\mu < 1$. The term $\mathbf{P}(x/\mu)$ represents a polynomial function of x/μ . As mentioned earlier, the TFMs are supposed to converge to the stationary

equations (19) and (20) in steady state. While equation (20) simply means that the departure rate is equal to the arrival rate in the long run, equation (19) indicates the possible existence of higher-order polynomial terms such as $x(t-1) \cdot y_2(t-1)$, $x^3(t-1)$, and $y_2^3(t-1)$ in the TFMs.

Empirical Experience

The analysis above assists to identify the model family for TFMs, which may contain main effects, higher-order polynomials or interactions of historical inputs/outputs:

$$y_i(t) = \sum_{q=1}^Q \sum_{j=0}^R \sum_{k=0}^{R-j} \sum_{l=0}^{R-j-k} b_{ijkl} \cdot y_1^j(t-q) \cdot y_2^k(t-q) \cdot x^l(t-q) \quad \text{for } i = 1, 2. \quad (21)$$

In (21), R represents the highest polynomial order, and Q the highest time order in the model. In our metamodeling of queueing systems (including the ones involving non-Markovian arrivals and services, machine failures, re-entrant flows, etc.), R turned out to be no higher than 3, and Q no higher than 1. Hence, we recommend considering (21) with $(R, Q) = (3, 1)$ as the most complicated functional form for the target TFMs. The values of (R, Q) can certainly be augmented if needed.

5.2.2 Backward Model Selection

Given the model family identified in equation (21), we developed a backward model selection strategy to obtain the parsimonious TFMs that are adequate to provide a good fit. We start with the most complicated model (21), and then seek to downsize the model by eliminating the functional terms that do not make a significant contribution in terms of describing the dynamic evolution. During the backward selection, the estimation of a candidate model is performed following the fitting process in Section 5.1.2. The specifics of the backward model selection method are given in Appendix A.3.

6 Empirical Examples

Given a queueing system whose simulation model is available, we apply the metamodeling method and seek to describe its transient behavior by a set of TFMs estimated from simulation data. Using such TFMs, the system's future dynamics can be predicted in a timely manner

without running additional simulation. In this section, the following examples are presented to illustrate the effectiveness of the proposed method:

- Single-station systems (Section 6.1): As pointed out in Section 3, transient analysis of even single-station systems is analytically intractable when general arrivals or services are involved.
- A scale-down semiconductor fabrication system (Section 6.2): This system consists of nine stations, and involves re-entrant flows, machine failures, and batch processing.

For each system being investigated, our metamodeling method was applied to generate a set of TFMs that describe the system dynamics. Simulation experiments were performed following the design strategies in Section 4.3, and the resulting data set, which will be referred to as the estimation data set (EDS), was used to estimate the TFMs following the statistical modeling approaches in Section 5. With the fitted TFMs, the future evolution of the system can be predicted under any input rate given the history of the system. To evaluate the prediction provided by the TFMs, a validation data set (VDS) was collected which contains simulation data different than and independent of those in the EDS, and the system dynamics directly described by the VDS was compared to that predicted by the TFMs. For all the numeric examples that we have investigated, the resulting TFMs are able to accurately predict the future evolution of the system, judging from the VDS-based cross validation.

Before discussing the empirical results, it is worth mentioning that in our discrete TFMs, one time unit represents the sampling interval Δt , which is introduced in Section 4.2.2. To avoid possible confusion, in the examples below we specify all the time periods (interarrival time, service time, simulation length, and future horizon) in terms of the time unit Δt . The arrival/service rates are also defined in terms of number per Δt .

6.1 Single-Station Systems

Table 1 summarizes the single-station queueing systems on which the metamodeling method has been successfully applied. These systems are intended to show that the proposed method can handle a wide range of service time distributions under non-stationary arrivals. In Table 1, SCV denotes the coefficient of variation (CV) for the distribution of service times. We have investigated systems with an SCV ranging from 0.1 to 2. In some cases, random failures of

servers are also considered. The arrivals to the single-station systems are modeled as both NSPP and NSNP (see Section 4.2.1 for the algorithms of generating non-stationary arrivals). The NSNP arrivals considered include those that are more and less variable than an NSPP.

Table 1: Configurations for single-station systems.

# Servers	Service Time Distribution	SCV	Failures	Arrivals
1 – 3	gamma, hyper-exponential	0.1 – 2	Yes/No	NSPP, NSNP

As an example for illustration, we chose to present the modeling results of a single-server system with the the service time following a hyper-exponential distribution with the probability density function (pdf) given as

$$g(t, p_1, p_2, \lambda_1, \lambda_2) = p_1 h(t, \lambda_1) + p_2 h(t, \lambda_2), \quad (22)$$

where function h denotes the pdf of an exponentially distributed random variable. The distribution parameters are given as $(p_1, p_2, \lambda_1, \lambda_2) = (0.1, 0.9, 0.022, 0.165)$, which results in an SCV of 2. The mean service time is 10 time units, and the service rate is 0.1 per time unit.

The arrivals in this example are modeled as an NSNP and generated using the thinning algorithm in Gerhardt and Nelson (2009). For the thinning algorithm, the renewal base process used to generate the NSNP arrivals has Erlang interarrival times with shape parameter $k = 25$ and rate parameter $\lambda = 2.5$. To obtain the EDS for the TFMs estimation, the system was simulated with NSNP arrivals following the rate function depicted in Figure 1(a). The input rate for the EDS is specified following the DOE strategies in Section 4.3. Specifically, the five constant rates are evenly-spaced to cover the system utilization range of $[0.5, 0.95]$, and they are sequenced in such a way that (10) is achieved; the simulation length of each constant-rate period is set as suggested in Section 4.3.2, and the total simulation length turns out to be $T = 12,000$ time units; the number of simulation replications performed in this case is $I = 51,500$, which is determined following the two-step process in Section 4.3.2 to achieve a relative precision level of $\gamma = 5\%$ for the estimate $Y_1(t) = \widehat{m}(t)$ in equations (7). From the multiple replications, the paired estimates $\{X(t), \mathbf{Y}(t), t = 1, 2, \dots, T\}$ were obtained using equations (7). With the collected EDS, the TFMs are obtained by applying the backward selection method (Appendix A.3) with the embedded fitting process (Section 5.1.2). The resulting TFMs for this single-server system

are given as follows:

$$\begin{aligned}
 \widehat{m}(t) &= 0.9929m(t-1) - 0.4264d(t-1) + 0.5293x(t-1) \\
 &\quad - 0.0015d(t-1)d(t-1) + 0.0721m(t-1)x(t-1) \\
 \widehat{d}(t) &= 0.0055m(t-1) + 0.5927d(t-1) + 0.3471x(t-1) \\
 &\quad - 0.2937x(t-1)x(t-1) + 0.0007d(t-1)d(t-1) - 0.0544m(t-1)x(t-1)
 \end{aligned} \tag{23}$$

Apparently, given the history $\{x(t), \mathbf{y}(t) = (m(t), d(t)), t \leq 0\}$, the fitted TFMs (23) can be used to recursively compute the future performance for any input $\{x^*(t), t = 1, 2, \dots, P\}$ with P time units representing the length of the future horizon. The computational effort required by the recursive computation is negligible.

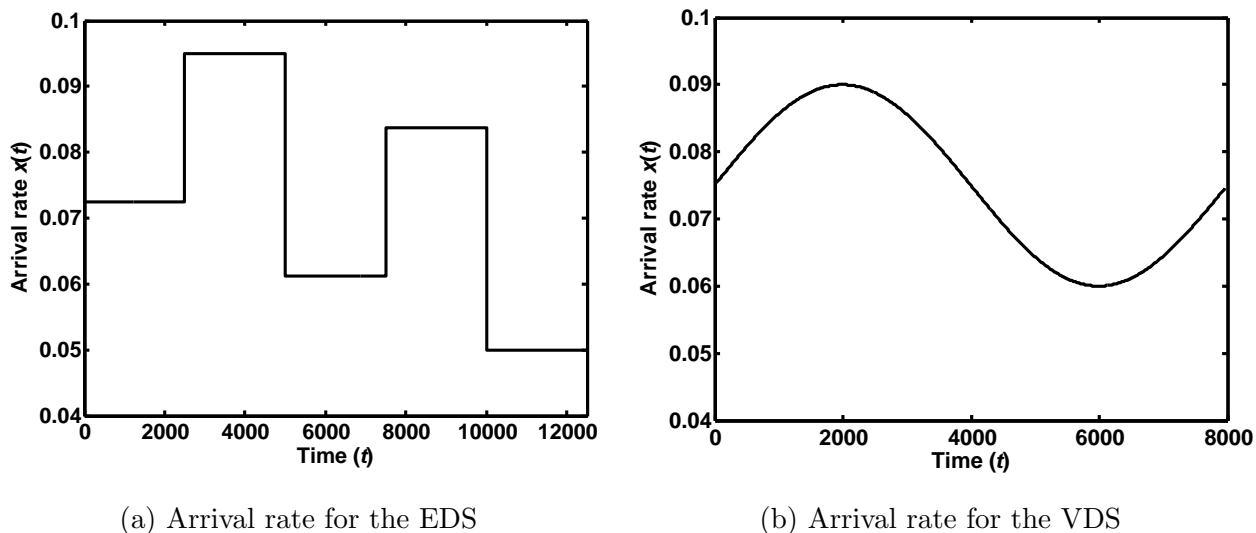


Figure 1: Arrival rate for the EDS and VDS for the single-server system

To evaluate the accuracy of the TFMs (23), the VDS were collected by running simulation with the NSNP arrivals following the rate function given in Figure 1(b). The arrivals are also generated using the thinning algorithm in Gerhardt and Nelson (2009). For the VDS, 75,000 simulation replications were performed, and highly accurate time series $\{\tilde{\mathbf{y}}(t) = (\tilde{m}(t), \tilde{d}(t)); t = 1, 2, \dots, P\}$ with $P = 8,000$ time units were obtained and considered as the “true” dynamic outputs with “zero” variance under the specified input flow. In Figure 2, the “true” outputs $\tilde{m}(t)$ and $\tilde{d}(t)$ are plotted as the dotted curves in Figure 2(a) and (b) respectively. The solid curves in Figure 2 represent the predicted dynamic outputs resulting from the fitted TFMs (23).

To obtain the predicted curves, the TFMs-based recursive computation was initiated by using the first pair of time-series points in the VDS as the seed values, and iteratively it leads to the prediction of the system evolution over the entire period given that the arrival rate follows Figure 1(b). Figure 2 shows that the predicted dynamics from the TFMs almost coincide with the “true” system evolution.

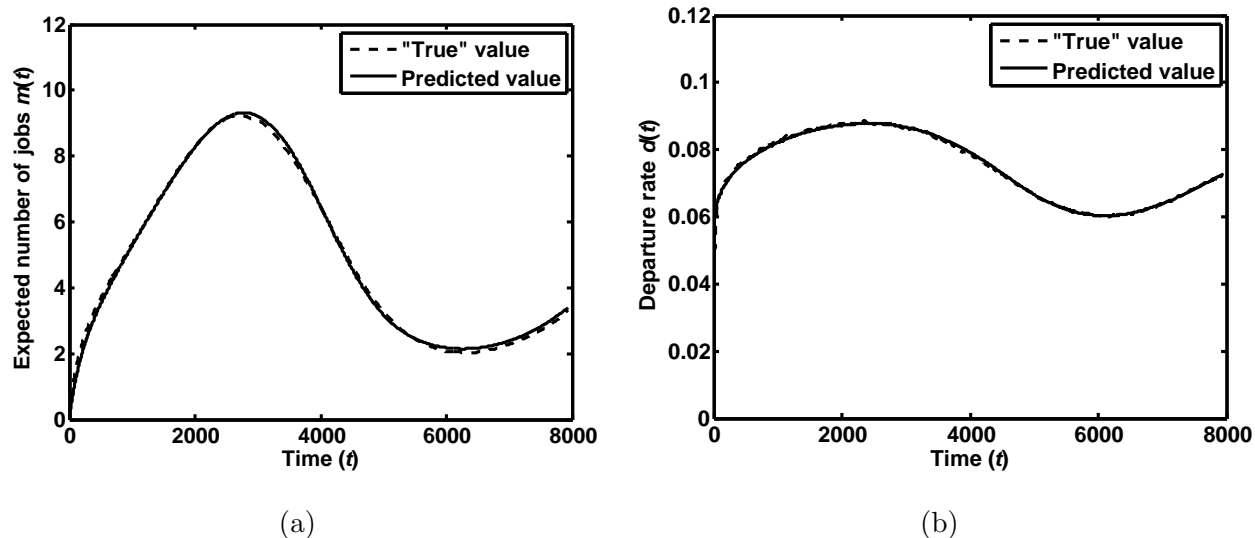


Figure 2: Comparison of the predicted dynamic outputs and their “true” values for the single-server system

6.2 A Multi-Station System

The metamodeling method has been applied on a scale-down semiconductor wafer fabrication (fab) system which has been used in the literature to evaluate new production planning methods and optimal control heuristics (Kayton et al. 1997; Asmundsson et al. 2009; Irdem et al. 2010). One type of products is considered, and the arrivals are modeled as a Poisson process. The system consists of nine workstations and involves the representative features of real wafer fab systems such as re-entrant flows, random failures, and batch processing. The specifics of the system are given in Appendix A.4. Before the detailed results for the scale-down wafer fab are presented in Section 6.2.3, we first devote Sections 6.2.1 and 6.2.2 to the analysis of multi-station systems in general.

6.2.1 Decomposition v.s. Aggregation View

There are two perspectives to approach a multi-station system: the decomposition and aggregation views, both of which have been used in the steady-state analysis of queueing systems. Examples for decomposing a queueing system into individual stations or subsystems include Whitt (1983), Bitran and Tirupati (1988), Hopp et al. (2002), and van Vuuren and Adan (2009). For this stream of work, appropriate recomposition is required to obtain the overall performance of the system from the behavior of each individual station. The aggregation approach has been explored by Atherton and Dayhoff (1986), Cheng and Kleijnen (1999), Park et al. (2002), Yang et al. (2007, 2008), etc. In those work, a multi-station system is considered as a whole and directly characterized by system-level performance profiles.

For the transient analysis of multi-station systems, both perspectives, decomposition and aggregation, can also be adopted. The basis of the decomposition method lies in the ability to accurately describe the transient behavior of a single station (or group), which has been demonstrated through the empirical examples in Section 6.1. In Appendix A.5, a five-station tandem system is considered and used to illustrate the use of our metamodeling method in the decomposition framework: The system is decomposed into five individual stations with each one characterized by a distinct set of TFMs as those in (23); the ensemble of the five sets of TFMs is able to predict the system's transient behavior recognizing that the departures from an upstream station serve as the arrivals to its downstream station. The details and the pros and cons of the decomposition method for transient analysis are discussed in Appendix A.5.

The aggregation approach, on the other hand, is the focus of this paper. The scale-down wafer fab will be used to illustrate that a multi-station system (or subsystem) can be considered as an aggregate group and its transient behavior can be accurately described by a single set of TFMs. It is worth mentioning that the aggregation system can potentially serve as a decomposed group in the decomposition approach for analyzing large-scale systems.

6.2.2 Suitability of the Aggregate Transient Analysis Method

As mentioned in Section 1.4, a single set of TFMs is most suitable to approximate the transient behavior of a system that is dominated by one bottleneck (BN) station. Here, we discuss in detail this empirical recommendation.

First, we specify the meaning of a system being dominated by a single BN station. Following

the concepts in Section 4.1, a BN station is the one with the heaviest utilization (i.e., the largest fraction of busy time); it most constrains the job flow, and it is where the queueing time and job congestion are most pronounced. We consider a system as dominated by a single BN if the utilization at the second heaviest utilized station does not exceed 80% of the BN utilization. The percentage 80% is an empirical value recommended based on the authors' experience with queueing and real manufacturing systems (Johnson et al. 2004; Yang et al. 2007, 2008; Yang 2010): The queueing effects at a station are likely to be significant when the station utilization is above 0.80.

A system is operated typically with the BN utilization well below one (Hopp 2007), at most at or slightly above one on a temporary basis. Hence, for a system dominated by one BN, the BN is the station where the jobs tend to accumulate and that largely determines the departure process from the system; whereas at the other stations, the waiting time for jobs in the queue are likely to be negligible (Hung and Leachman 1996). For such a system where the allocation of jobs among stations is roughly given, the system-level performance metrics imply a good depiction of the performance at the BN and other individual stations. This serves as a good basis for the aggregation approach, which seeks to characterize the behavior of a system by a single set of TFMs with outputs being system-level performances only: the expected number of jobs in the system and the departure rate from the system.

Needless to say, the aggregate metamodeling is a statistical approximation method, and its effectiveness is demonstrated through the scale-down wafer fab.

6.2.3 TFMs-Based Prediction Results

Given the scale-down wafer fab described in Appendix A.4, the analytical method (Section 4.1) is first applied to perform the capacity and BN analysis of the system. The analyzing results, which are summarized in Appendix A.4, show that the wafer fab is dominated by a single BN station, and assist to specify the design of simulation experiments (Section 4.3). Also, based on the analytically-derived system capacity μ , the time unit Δt is set as $1/8$ of $1/\mu$ (Section 4.2.2). The metamodeling procedure is then applied to estimate the TFMs for the system.

To obtain the EDS, Poisson arrivals were fed to the simulation model with the input rate being a piecewise constant function similar to the one in Figure 1(a). The five distinct constant arrival rates are also selected in such a way that they correspond to five evenly-spaced system

(i.e., BN) utilizations $\{0.5, 0.6125, 0.725, 0.8375, 0.95\}$. The length of a simulation replication is set as $T = 5,000$ time units. A total of 56,200 replications were performed for the EDS, from which the TFMs are estimated. The fitted TFMs take as input the arrival rate $x(t)$ into the system, and output $\widehat{m}(t)$, the expected number of jobs in the system, and $\widehat{d}(t)$, the departure rate from the system.

To evaluate the ability of the TFMs to predict the system evolution, two cases were considered with the input rate following (i) a sine function as shown in Figure 3(a), and (ii) a piece-wise constant function as shown in Figure 3(b). In each of the two cases, a large number of simulation replications were carried out to obtain the corresponding VDS, and the outputs estimated from the VDS are considered as the “true” system performance and used to evaluate the TFMs-based prediction results.

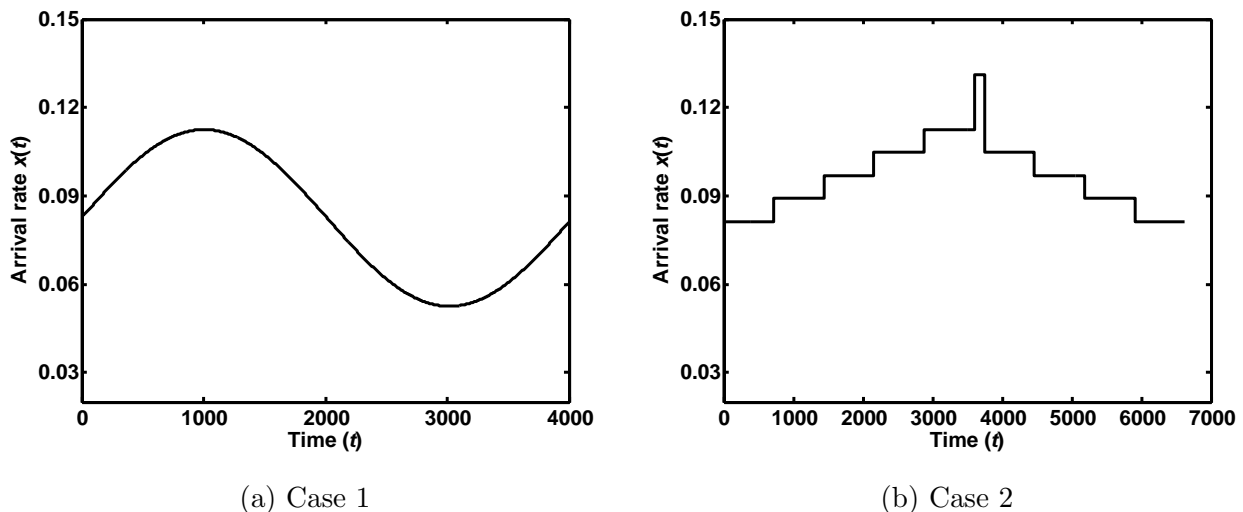


Figure 3: Arrival rate of jobs in the VDS.

Case 1

The TFMs were used to predict the system performance under Poisson arrivals with the rate following the sine function in Figure 3(a). In the sine function, the highest and lowest arrival rates correspond to a system utilization of 0.9 and 0.66 respectively. To obtain the VDS that provides the “true” system behavior, 75,000 simulation replications were carried out by feeding to the system the non-stationary Poisson arrivals that follow the sine rate. Figure 4 compares the “true” dynamic behavior $(\widetilde{m}(t), \widetilde{d}(t))$, which is obtained from the VDS and represented by

the dotted curves in the graphs, with the system evolution $(\widehat{m}(t), \widehat{d}(t))$ predicted by the TFMs, which is depicted by the solid curves. As can be seen from Figure 4, the curves predicted by the TFMs are able to accurately describe the system's evolution under the input rate depicted in Figure 3(a).

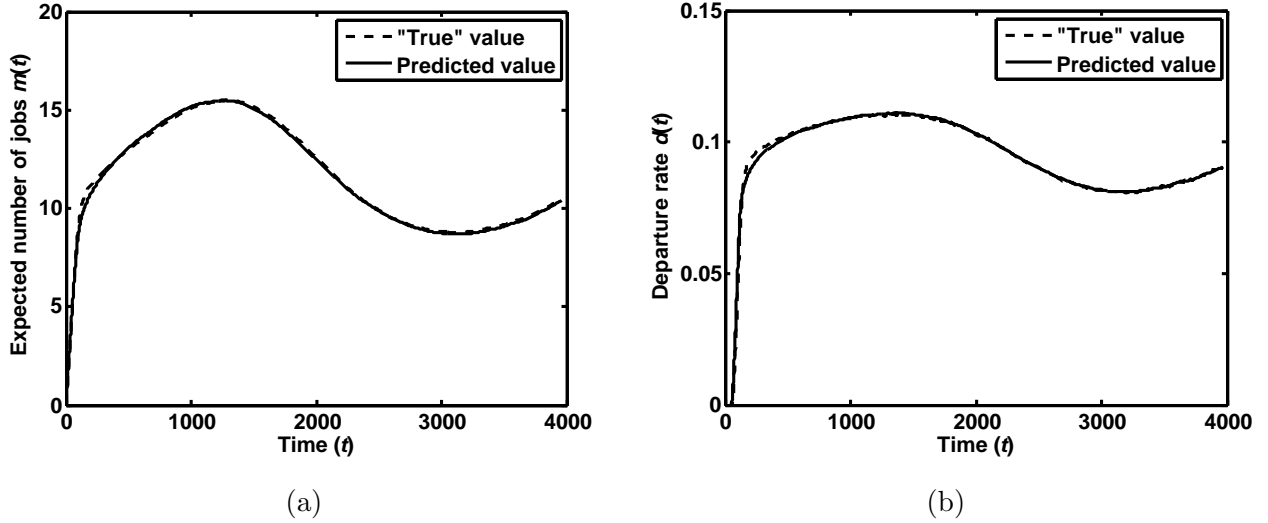


Figure 4: Comparison of the predicted dynamic outputs and their “true” values for the multi-station system (Case 1).

To quantify the prediction quality, the relative absolute errors (RAE) are also obtained over the time period $t = 1, 2, \dots, P$ shown in Figure 4. The RAE for $\widehat{m}(t)$ is calculated as

$$\text{RAE}[m(t)] = |\widehat{m}(t) - \widetilde{m}(t)| / \widetilde{m}(t); \quad t = 1, 2, \dots, P. \quad (24)$$

The statistics are summarized as follows for the error set $\{\text{RAE}[m(t)]; t = 1, 2, \dots, P\}$: The maximum is 7.05%, the 95th percentile is 3.00%, and the 50th percentile is 0.96%. Similarly, the RAE set for $\widehat{d}(t)$ is calculated as

$$\text{RAE}[d(t)] = |\widehat{d}(t) - \widetilde{d}(t)| / \widetilde{d}(t); \quad t = 1, 2, \dots, P, \quad (25)$$

and the statistics are given as: The maximum equals to 9.17%, the 95th percentile 2.06%, and the 50th percentile 0.33%.

Case 2

The TFMs were used to predict the system performance under Poisson arrivals with the rate

following Figure 3(b), in which the ten constant rates (from the left to right) correspond to the ten system utilizations

$$\{0.65, 0.71, 0.78, 0.84, 0.90, 1.04, 0.90, 0.84, 0.78, 0.71, 0.65\} \quad (26)$$

with the highest utilization being 1.04. To obtain the VDS, 75,000 simulation replications were performed with arrival rate following Figure 3(b). This case is designed to illustrate the TFMs’ ability to predict (i) the transient behavior, (ii) the steady-state behavior, and (iii) the temporarily overloaded behavior of the system.

As in Case 1, Figure 5 compares the “true” dynamic behavior, which is obtained from the VDS and represented by the dotted curves in the graphs, with the system evolution predicted by the TFMs, which is depicted by the solid curves. As shown in Figure 5, some segments of the curve have reached steady state and some have not, and in both situations, the TFMs provide good prediction. The RAE sets were also obtained for this case over the time period $t = 1, 2, \dots, P$ as shown in Figure 5: For $\{\text{RAE}[m(t)]; t = 1, 2, \dots, P\}$, the maximum, 95th percentile, and 50th percentile are 3.70%, 1.78%, and 0.43% respectively; for $\{\text{RAE}[d(t)]; t = 1, 2, \dots, P\}$, the maximum, 95th percentile, and 50th percentile are 6.76%, 2.26%, and 0.45% respectively.

The overloaded (highest) segment in Figure 5 deserves more discussions here. The meta-modeling method is applicable to accommodate overloaded system behavior. As can be seen in Figure 5, the fitted TFMs are able to track the outputs of the system when it is temporarily overloaded. However, compared to the non-overloaded segments, the prediction results for the overloaded part is worse; and in our experience, the prediction quality of the TFMs deteriorates with the increase of the overloaded period. We believe that the inadequacy of our TFMs to predict overloaded behavior is due to the following reasons. The TFMs are estimated from the EDS which only includes non-overloaded data, and predicting overloaded behavior requires extrapolation (as opposed to interpolation), which is recognized as risky in statistical prediction in general (Tamhane and Dunlop 2000). It is known and also evident from Figure 5 that overloaded behaviors have distinct features that are not shared by non-overloaded ones. Hence, the resulting TFMs here fall short in terms of extrapolating the overloaded performance. We do not recommend including overloaded data in the EDS unless overloaded behaviors are of particular interest; overloaded data are highly variable, and extremely large number of simulation replications are needed to obtain good performance estimates when a system is overloaded, which may

not be an efficient use of computation resources.

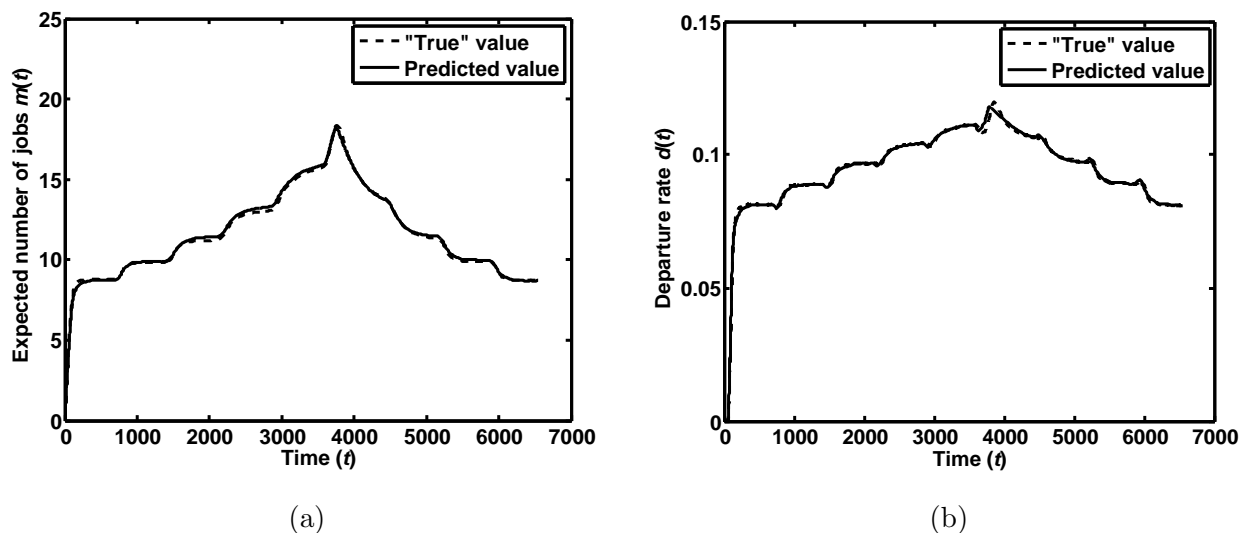


Figure 5: Comparison of the predicted dynamic outputs and their “true” values for the multi-station system (Case 2).

7 Summary

The originality of this work lies in the integration of statistical methods, computer simulation, and queueing theory to tackle the ever-difficult yet critical research problem of characterizing the transient behavior of general queueing systems. Such an approach is expected to overcome the computational burden of simulation and the intractability of analytical methods for realistic systems, and thereby to support responsive decision making.

Our simulation-based metamodeling method supplements the current literature of transient analysis in the following aspects. First, it can accommodate the various features of manufacturing systems including non-Markovian arrivals and services, multi-servers, multi-stations, re-entrant flows, batch processing, and random failures. Second, the TFMs resulting from our methods can be used to timely predict the system’s performance since they are difference equations relating the future output performance to history. In addition, it is worth mentioning that the metamodeling method can be straightforwardly extended to describe the second and higher moments of the system’s output performance (that is, the number of jobs in the system and the departures from the system). The set of TFMs considered in this paper can be expanded to include one

additional transfer function model for each new higher-order moment measure. The sample data required for estimating all the transfer function models can be collected simultaneously through simulation experiments. Since high-moment data are much more variable than its first-moment counterpart, a larger number of simulation replications are likely to be needed for estimating the expanded TFMs.

ACKNOWLEDGMENTS

This research was supported by National Science Foundation Grant CMMI-1068131. Sincere thanks go to Professor Barry Nelson from Northwestern University, Professor Reha Uzsoy from North Carolina State University, and the editors and referees.

REFERENCES

- [1]. Ankenman, B. E., B. L. Nelson, and J. Staum. 2010. Stochastic Kriging for Simulation Metamodeling. *Operations Research*, forthcoming.
- [2]. Atherton, R. W., and J. E. Dayhoff. 1986. Signature analysis: simulation of inventory, cycle time, and throughput trade-offs in wafer fabrication. *IEEE Transactions On Components, Hybrids, Manufacturing Technology* CHMT-9(4): 498–507.
- [3]. Asmundsson, J. M., R. L. Rardin, C. H. Turkseven and R. Uzsoy. 2009. Production planning with resources subject to congestion. *Naval Research Logistics* **56** (2): 142–157.
- [4]. Bitran, G. R. and D. Tirupati. 1988. Multiproduct Queueing Networks with Deterministic Routing: Decomposition Approach and the Notion of Interference. *Management Science* **34** (1): 75–100.
- [5]. Box, G., G. M. Jenkins and G. Reinsel. 2008. *Time Series Analysis: Forecasting & Control* (4rd Edition). Wiley & Sons.
- [6]. Chen, H. B. and A. Mandelbaum. 1994. *Hierarchical Modelling of Stochastic Networks Part I: Fluid Models, Stochastic Modeling and Analysis of Manufacturing Systems* (eds. Yao, D. D.), New York: Springer.
- [7]. Cheng, R. C. H. and J. P. C. Kleijnen. 1999. Improved design of queueing simulation experiments with highly heteroscedastic responses. *Operations Research* **47**: 762–777.

- [8]. Clark, G. M. 1981. Use of Polya Distributions in Approximate Solutions to Nonstationary M/M/s Queues. *Communications of the ACM* **24**: 206–217.
- [9]. Datta, P. P., M. Christopher and P. Allen. 2007. Agent-based Modelling of Complex Production/Distribution Systems to Improve Resilience. *International Journal of Logistics* **10** (3): 187–203
- [10]. Daley, D. J. and D. Vere-Jones. 2002. *An Introduction to the Theory of Point Processes, Vol I: Elementary Theory and Methods*. 2nd Ed. Springer.
- [11]. Eick, S. G., W. A. Massey and W. Whitt. 1993a. The physics of the Mt/G/infty Queue. *Operations Research* **41** (4): 731–742.
- [12]. Eick, S. G., W. A. Massey and W. Whitt. 1993b. Mt/G/infty Queues with Sinusoidal Arrival Rates. *Management Science* **39** (2): 241–252.
- [13]. Gerhardt, I. and B. L. Nelson. 2009. Transforming Renewal Processes for Simulation of Nonstationary Arrival Processes. *INFORMS Journal on Computing*. **21** (4): 630–640.
- [14]. Golub, G. H. and V. L. F. Charles. 1996. *Matrix Computations*, 3rd edition. Johns Hopkins.
- [15]. Green, L. V. and P. J. Kolesar. 1991a. The Pointwise Stationary Approximation with for Queues with Nonstationary Arrivals. *Management Science* **37** (1): 84–97.
- [16]. Green, L. V., P. J. Kolesar and A. Svornos. 1991b. Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems. *Operations Research* **39** (3): 502–511.
- [17]. Green, L. V., P. J. Kolesar and W. Whitt. 2007. Coping with Time-varying Demand when Setting Staffing Requirements for a Service System. *Production and Operations Management* **16** (1): 13–39.
- [18]. Gross, D. and C. Harris. 1985. *Fundamentals of Queueing Theory*. New Jersey: John Wiley & Sons.
- [19]. Gross, D. and D. Miller. 1984. The Randomization Technique as a Modeling Tool and Solution Procedure for Transient Markov Processes. *Operations Research* **32** (6): 926–944.
- [20]. Harrod, S. and W. D. Kelton. 2006. Numerical methods for realizing nonstationary Poisson processes with piecewise-constant instantaneous-rate functions. *Simulation* **82**: 147–157.
- [21]. Henderson, S. G. and B. L. Nelson. 2006. *Handbooks in Operations Research and Management Science: Simulation*. Amsterdam, Netherlands: Elsevier Science.

- [22]. Hopp, W. 2007. *Supply Chain Science*. Irwin/McGraw-Hill, New York.
- [23]. Hopp, W. J. M. L. Spearman, S. Chayet, K. Donohue and E. Senturk. 2002. Using an Optimized Queueing Network Model to Support Wafer Fab Design. *IIE Transactions* **34** (2): 119-130.
- [24]. Hung, Y. F. and R. C. Leachman. 1996. A Production Planning Methodology for Semiconductor Manufacturing Based on Iterative Simulation and Linear Programming Calculations. *IEEE Transactions on Semiconductor Manufacturing* **9** (2): 257–269.
- [25]. Ingolfsson, A., E. Akhmetshina, S. Budge, Y. Li and X. Wu. 2007. A Survey and Experimental Comparison of Service Level Approximation Methods for Non–Stationary M/M/s Queueing Systems. *INFORMS Journal on Computing* **19** (2): 201–214.
- [26]. Irdem, D. F., N. B. Kacar and R. Uzsoy. 2010. An Exploratory Analysis of Two Iterative Linear Programming–Simulation Approaches for Production Planning. *IEEE Transactions on Semiconductor Manufacturing* **23** (3): 442–455.
- [27]. Jennings, O. B., A. Mandelbaum, W. A. Massey and W. Whitt. 1996. Server Staffing to Meet Time–Varying Demand. *Management Science* **42** (10): 1383–1394.
- [28]. Johnson, R., F. Yang, B. E. Ankenman and B. L. Nelson. 2004. Nonlinear Regression Fits for Simulated Cycle Time vs. Throughput Curves for Semiconductor Manufacturing. *Proceedings of the 2004 Winter Simulation Conference* 1951–1955.
- [29]. Kayton, D., T. Teyner, C. Schwartz and R. Uzsoy. 1997. Focusing Maintenance Improvement Efforts in a Wafer Fabrication Facility Operating under Theory of Constraints. *Production and Inventory Management Journal* **38** (4): 51–57.
- [30]. Kelly, F. P., S. Zachary and I. Ziedins. 1996. *Stochastic Networks: Theory and Applications* (*Royal Statistical Society Lecture Note Series*). New York: Oxford University Press.
- [31]. Kleinrock, L. 1975. *Queueing Systems*. New York: John Wiley & Sons.
- [32]. Kumar, S. and P. R. Kumar. 2001. Queueing Network Models in the Design and Analysis of Semiconductorwafer Fabs. *IEEE Transactions on Robotics and Automation* **17** (5): 548–561.
- [33]. Law, Kelton, A. M. and W. D. Kelton. 2000. *Simulation Modeling and Analysis*, third edition. New York: McGraw-Hill.
- [34]. Leemis, L. M. 1991. Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process. *Management Sciences*. **37**: 886–900.

- [35]. Lewis, P. A. W. and G. S. Shedler. 1979. Simulation of nonhomogenous Poisson processes by thinning. *Naval Research Logistics Quarterly* **26** (3): 403–413.
- [36]. Ljung, L. 1999. *System Identification: Theory for the User*, second edition. Prentice Hall.
- [37]. Mandelbaum, A. and W. A. Massey. 1995. Strong Approximations for Time-Dependent Queues. *Mathematics of Operations Research* **20** (11): 33–64.
- [38]. Massey, W. A. and W. Whitt. 1997. Peak Congestion in Multi-Server Service Systems with Slowly Varying Arrival Rates. *Queueing Systems* **25**: 157–172.
- [39]. McKenzie, E. 2003. Discrete variate time series. In D. N. Shanbhag and C. R. Rao (Eds.), *Handbook of statistics 21, stochastic processes: modelling and simulation*. Amsterdam: North-Holland.
- [40]. Meng, G. and S. Heragu. 2004. Batch Size Modeling in a Multi-item, Discrete Manufacturing System via an Open Queueing Network. *IIE Transactions* **36**: 743–753.
- [41]. Missbauer, H. and R. Uzsoy. 2010. Optimization Formulations of Production Planning Problems. In: *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook* (eds. Kempf, K. G., P. Keskinocak and R. Uzsoy.) Springer: New York.
- [42]. Nelson, B. L. and M. R. Taaffe. 2004a. The $Ph_t/Ph_t/\infty$ Queueing System: Part I – the Single Node. *INFORMS Journal on Computing* **16** (3): 266–274.
- [43]. Nelson, B. L. and M. R. Taaffe. 2004b. The $[Ph_t/Ph_t/\infty]^K$ Queueing System: Part II – the Multiclass Network. *INFORMS Journal on Computing* **16** (3): 275–283.
- [44]. Papadopolous, H. T., C. Heavey and J. Browne. 1993. *Queueing Theory in Manufacturing Systems Analysis and Design*, 1st edition. New York: Springer.
- [45]. Park, S., J. W. Fowler, G. T. Mackulak, J. B. Keats, and W. M. Carlyle. 2002. D-optimal sequential experiments for generating a simulation-based cycle time-throughput curve. *Operations Research* **50**: 981–990.
- [46]. Pinedo, M. L. 2007. *Planning and Scheduling in Manufacturing and Services*. New York: Springer
- [47]. Riano, G. 2003. Transient Behavior of Stochastic Networks: Application to Production Planning with Load-Dependent Lead Times. Dissertation, School of Industrial and Systems Engineering, Georgia Institute of Technology. Atlanta, Georgia.

- [48]. Ross, S. M. 1995. *Stochastic Processes*, 2nd edition. NJ: John Wiley & Sons.
- [49]. Rothkopf, M. H. and S. S. Oren. 1979. A Closure Approximation for the Nonstationary M/M/s Queue. *Management Science* **25**: 522–534.
- [50]. Shanthikumar, J. G., S. Ding and M. T. Zhang. 2007. Queueing Theory for Semiconductor Manufacturing Systems: A Survey and Open Problems. *IEEE Transactions on Automation Science and Engineering* **4** (4): 513–522.
- [51]. Sheffi, Y. and J. B. Rice. 2005. A Supply Chain View of the Resilient Enterprise. *Sloan Management Review* **47** (1): 41–48.
- [52]. Stadtler, H. and C. Kilger. 2007. *Supply Chain Management and Advanced Planning: Concepts, Models, Software, and Case Studies*, 4th edition. New York: Springer
- [53]. Taaffe, M. R. and K. L. Ong. 1987. Approximating Nonstationary Ph(t)/M(t)/S/C Queueing Systems. *Annals of Operations Research* **8**: 103–116.
- [54]. Tamhane, A. C. and D. D. Dunlop. 2000. *Statistics and Data Analysis from Elementary to Intermediate*. Prentice Hall.
- [55]. Uzsoy, R., C. Y. Lee and L. A. Martin–Vega. 1992. A Review of Production Planning and Scheduling in the Semiconductor Industry, Part I: System Characteristics, Performance Evaluation, and Production Planning. *IIE Transactions* **24** (4): 47–61.
- [56]. van Vuuren, M. and I. J. B. F. Adan. 2009. Performance analysis of tandem queues with small buffers. *IIE Transactions* **41** (10): 882–892.
- [57]. Whitt, W. 1983. The Queueing Network Analyzer. *Bell System Technical Journal* **62** (9): 2779–2815.
- [58]. 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.
- [59]. Yang, Y., B. E. Ankenman and B. L. Nelson. 2008. Cycle Time Percentile Curves for Manufacturing Systems. *INFORMS Journal on Computing* **20** (4): 628–643.
- [60]. Yang, F. 2010. Neural Network Metamodeling for Cycle–Time Based Performance Profiles in Manufacturing. *European Journal of Operational Research*. **205** (1): 172–185.
- [61]. Zäpfel, G. and H. Missbauer. 1993. Production Planning and Control (PPC) Systems Including Load–Oriented Order Release–Problems and Research Perspectives. *International Journal of Production Economics* **30–31**: 107–122.

A Appendix

A.1 Analytical Results for a General Single-Server Queue

Following the notations in Sections 2 and 6.2, we derive the equations (5) for a general single-server queue with orderly arrivals and departures. The service time of jobs follows distribution $G(\tau)$, $\tau \in (\tau_L, \tau_U)$ with $0 \leq \tau_L < \tau_U \leq \infty$.

The state probabilities $p_n(t) = \Pr\{Q(t) = n\}$ satisfy the differential equations:

$$\begin{aligned} p'_n(t) &= x_{n-1}(t) + d_{n+1}(t) - x_n(t) - d_n(t); \quad n \geq 1 \\ p'_n(t) &= d_{n+1}(t) - x_n(t); \quad n = 0. \end{aligned} \tag{27}$$

Multiplying both sides of (27) by n and taking the sum across all values of n , we have

$$m'(t) = \frac{dE[m(t)]}{dt} = \sum_{n=0}^{\infty} n \cdot p'_n(t) = x(t) - d(t), \tag{28}$$

which is the first equation in (5). Recall that $x(t)$ is the independent input variable representing the arrival rate of jobs, and $m(t)$ and $d(t)$ characterize the output processes of interest. Next, we proceed to derive the dynamic evolution of $d(t) = \sum_{n=1}^{\infty} d_n(t)$.

We first consider $d_n(t) = \lim_{\delta \rightarrow 0^+} \delta^{-1} \Pr\{D(t, t + \delta) = 1, Q(t) = n\}$. A departure will occur during the interval $(t, t + \delta]$ with n jobs in the system at time t if one of the two following conditions holds:

- (i) The system was empty at time $t - \tau$, and one job entered during the instant $(t - \tau, t - \tau + \delta]$. During the service of this job, which lasted for a period of τ , there were $n - 1$ new arrivals to the system.
- (ii) A departure occurred during the instant $(t - \tau, t - \tau + \delta]$ while there are $k \geq 2$ jobs in the system at time $t - \tau$. Immediately after the departure, the service for the first job in the queue was initiated and lasted for a period of τ . During the service of this job, $n - k + 1$ new jobs entered the system.

Thus,

$$\begin{aligned} &\Pr\{D(t, t + \delta) = 1, Q(t) = n\} \\ &= \int_{\tau_L}^{\tau_U} \Pr\{A(t - \tau, t - \tau + \delta) = 1, Q(t - \tau) = 0, A(t - x, t) = n - 1\} dG(\tau) \\ &+ \int_{\tau_L}^{\tau_U} \sum_{k=2}^{n+1} \Pr\{D(t - \tau, t - \tau + \delta) = 1, Q(t - \tau) = k, A(t - \tau, t) = n - k + 1\} dG(\tau), \end{aligned}$$

and we further have

$$\begin{aligned}
\Pr\{D(t, t + \delta) = 1\} &= \int_{\tau_L}^{\tau_U} \sum_{n=1}^{\infty} \Pr\{A(t - \tau, t - \tau + \delta) = 1, Q(t - \tau) = 0, A(t - x, t) = n - 1\} dG(\tau) \\
&+ \int_{\tau_L}^{\tau_U} \sum_{n=1}^{\infty} \sum_{k=2}^{n+1} \Pr\{D(t - \tau, t - \tau + \delta) = 1, Q(t - \tau) = k, A(t - \tau, t) = n - k + 1\} dG(\tau) \\
&= \int_{\tau_L}^{\tau_U} \Pr\{A(t - \tau, t - \tau + \delta) = 1, Q(t - \tau) = 0\} dG(\tau) \\
&+ \int_{\tau_L}^{\tau_U} (\Pr\{D(t - \tau, t - \tau + \delta) = 1\} - \Pr\{D(t - \tau, t - \tau + \delta) = 1, Q(t - \tau) = 1\}) dG(\tau).
\end{aligned}$$

Therefore,

$$\begin{aligned}
d(t) &= \lim_{\delta \rightarrow 0^+} \delta^{-1} \Pr\{D(t, t + \delta) = 1\} \\
&= \int_{\tau_L}^{\tau_U} x_0(t - \tau) dG(\tau) + \int_{\tau_L}^{\tau_U} (d(t - \tau) - d_1(t - \tau)) dG(\tau),
\end{aligned}$$

which is the second equation in (5).

A.2 Statistical Inference on the TFMs

The notation used in Section 5 is inherited here. For convenience of the discussion, we rewrite the transformed models (17) and (18) as follows:

$$\tilde{Y}_1(t) = \tilde{F}_1(\boldsymbol{\theta}_1, \tilde{\mathcal{H}}_t) + w(t) \quad (29)$$

$$\tilde{Y}_2(t) = \tilde{F}_2(\boldsymbol{\theta}_2, \tilde{\mathcal{H}}_t) + w(t) \quad (30)$$

where $\tilde{Y}_i(t) = \frac{D_i(q)}{\sigma_i C_i(q)} Y_i(t)$ ($i = 1, 2$), the transformed function $\tilde{F}_i = \frac{D_i(q)}{\sigma_i C_i(q)} F_i$ ($i = 1, 2$), and $w(t)$ is the white noise with variance 1. The history of the system prior to time t is denoted as $\tilde{\mathcal{H}}_t = \{X(\tau), \mathbf{Y}(\tau), \tau < t\}$.

Suppose that T time-series pairs $\{X(t), \mathbf{Y}(t), t = 1, \dots, T\}$ have been obtained for the estimation of the models (29) and (30). We define the additional notation as follows:

- $\mathbf{f}_1(\boldsymbol{\theta}_1) = (\tilde{F}_1(\boldsymbol{\theta}_1, \tilde{\mathcal{H}}_1), \tilde{F}_1(\boldsymbol{\theta}_1, \tilde{\mathcal{H}}_2), \dots, \tilde{F}_1(\boldsymbol{\theta}_1, \tilde{\mathcal{H}}_T))'$ is a $T \times 1$ vector function of $\boldsymbol{\theta}_1$.
- $\mathbf{f}_2(\boldsymbol{\theta}_2) = (\tilde{F}_2(\boldsymbol{\theta}_2, \tilde{\mathcal{H}}_1), \tilde{F}_2(\boldsymbol{\theta}_2, \tilde{\mathcal{H}}_2), \dots, \tilde{F}_2(\boldsymbol{\theta}_2, \tilde{\mathcal{H}}_T))'$ is a $T \times 1$ vector function of $\boldsymbol{\theta}_2$.
- $\mathbf{D}_1(\hat{\boldsymbol{\theta}}_1) = \partial \mathbf{f}_1(\hat{\boldsymbol{\theta}}_1) / \partial \boldsymbol{\theta}'_1$ is a $T \times N_1$ first-derivative matrix, where N_1 is the dimension of $\boldsymbol{\theta}_1$.

- $\mathbf{D}_2(\widehat{\boldsymbol{\theta}}_2) = \partial \mathbf{f}_2(\widehat{\boldsymbol{\theta}}_2) / \partial \boldsymbol{\theta}'_2$ is a $T \times N_2$ first-derivative matrix, where N_2 is the dimension of $\boldsymbol{\theta}_2$.
- The design matrix \mathbf{D} is defined as:

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_1(\widehat{\boldsymbol{\theta}}_1) & 0 \\ 0 & \mathbf{D}_2(\widehat{\boldsymbol{\theta}}_2) \end{pmatrix} \quad (31)$$

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, then the estimated parameters $\widehat{\boldsymbol{\theta}}$ is approximately normally distributed with the variance-covariance matrix:

$$\widehat{\text{Var}}[\widehat{\boldsymbol{\theta}}] = \sigma^2 (\mathbf{D}'\mathbf{D})^{-1}, \quad (32)$$

where $\sigma^2 = \text{Var}[w(t)] = 1$ for models (29) and (30).

A.3 Backward Model Selection

Starting from the most complicated functional form (21), we search for a parsimonious model in the backward manner. Due to the non-i.i.d errors associated with the original model, the model selection is based on the corresponding transformed forms (17) and (18) in Section 5.1 (or equivalently, (29) and (30) in Appendix A.2). Since the starting model (21) includes a large number of regressors, we propose the following two steps to complete the backward model selection: the initial *ad hoc* screening and the hypothesis test-based model selection.

Initial *Ad Hoc* Screening

Following the notations in Appendix A.2, we consider the redundancy of the model $\widetilde{\mathbf{F}} = (\widetilde{F}_1, \widetilde{F}_2)$, which is transformed from the starting model. The transformed model $\widetilde{\mathbf{F}}$ involves parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)$, an $(N_1 + N_2) \times 1$ vector. Denote the transformed sample responses as $\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{Y}}_1, \widetilde{\mathbf{Y}}_2)$. Since $\widetilde{\mathbf{F}}$ is linear with respect to the model parameters $\boldsymbol{\theta}$, it can be written as $\widetilde{\mathbf{Y}} = \mathbf{D} \times \boldsymbol{\theta}$. For such linear models, the design matrix \mathbf{D} , as defined in (31), only depends on the data observations and does not involve any unknown parameters. Each column vector in \mathbf{D} corresponds to a linear regressor in the TFMs. The least square fitting is to project the response vector $\widetilde{\mathbf{Y}}$ onto the subspace spanned by the columns of \mathbf{D} . The dimension of this subspace is implied by the rank of \mathbf{D} , say r . If $r < N_1 + N_2$, it means that $N_1 + N_2 - r$ regressors can

be eliminated from the model due to linear redundancy. The rank r can be obtained using the single value decomposition method (Golub 1996).

Hypothesis Test-Based Model Selection

Denote the model resulting from the previous screening as \mathcal{M}_0 . For \mathcal{M}_0 , the backward stepwise regression method (Tamhane and Dunlop 2000) is used to remove potentially redundant regressors one at a time, and is detailed as follows. (i) Let $\mathcal{M}_F = \mathcal{M}_0$, estimate the model $\widehat{\mathcal{M}}_F$ using the fitting method in Section 5.1.2, and obtain the sum of square error (SSE) denoted as $\text{SSE}(\widehat{\mathcal{M}}_F)$. (ii) Drop a regressor from \mathcal{M}_F , and denote the resulting model as \mathcal{M}_P ; estimate $\widehat{\mathcal{M}}_P$ using the fitting method in Section 5.1.2, and obtain $\text{SSE}(\widehat{\mathcal{M}}_P)$. (iii) Evaluate the significance of the dropped regressor by testing the statistic

$$F = \frac{\text{SSE}(\widehat{\mathcal{M}}_P) - \text{SSE}(\widehat{\mathcal{M}}_F)}{\text{SSE}(\widehat{\mathcal{M}}_F)/\text{DF}_F}, \quad (33)$$

where DF_F is the degree of freedom for model $\widehat{\mathcal{M}}_F$. (iv) If statistic F is significant, then go to Step (v); otherwise, set $\mathcal{M}_F = \mathcal{M}_P$. (v) Repeat steps (ii)-(iv) until all the regressors in \mathcal{M}_0 have been evaluated.

The $\widehat{\mathcal{M}}_F$ resulting from the backward selection is the TFMs that will be used to describe the system behavior.

A.4 Configuration for the Multi-Station System

The system being investigated is a scale-down semiconductor wafer fabrication (fab) facility developed in Kayton et al. (1997). One type of wafer (product) flow is considered whose processing routing is illustrated in Figure 6. It takes 13 steps to process the wafers, and the numbers 1-13 in Figure 6 show the sequence of stations that each wafer has to visit. Wafers are released into the fab in a fixed lot size of 50. A “lot” is a number of wafers of the same type being processed and traveling as a whole between workstations.

As pointed out in Kayton et al. (1997), this system includes the major features in real wafer fabs:

- Re-entrant flows: As shown in Figure 6, Stations 1, 4, and 6 are revisited by products.

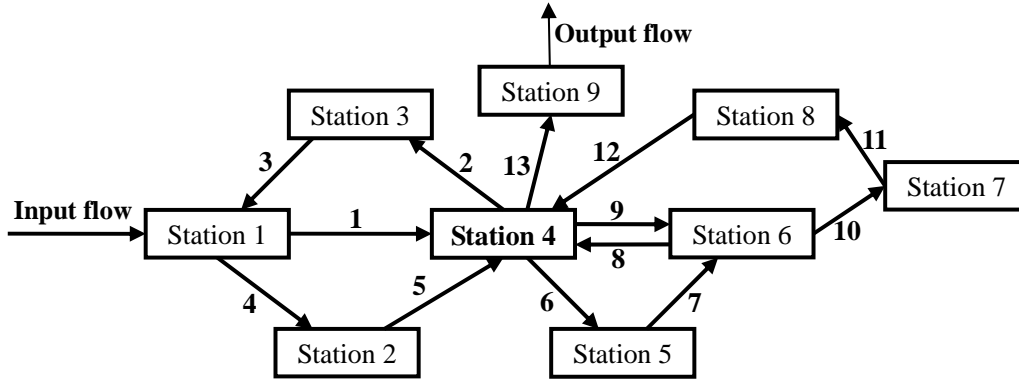


Figure 6: Product flow through the scale-down wafer fab.

- Machine failures: Stations 3 and 7 involve machine failures and repairs. Denote the gamma distribution as

$$g(t, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} \exp(-\beta t) \quad t \geq 0. \quad (34)$$

For the machines at both stations, the time to fail follows the gamma distribution with parameters being $(\alpha, \beta) = (720, 1)$; the time to repair also follows the gamma distribution with $(\alpha, \beta) = (120, 1.5)$. The times are given in terms of time units.

- Batch processing: Stations 1 and 2 process products in batches; at both stations, the minimum and maximum batch sizes allowed for processing are 2 and 4 lots respectively.

The processing times at all the workstations follow lognormal distributions. Table 2 gives the number of machines, and the mean and standard deviation (Stdev) of the processing time (in terms of time units) at each workstation.

Table 2: The processing time at each workstation.

Station #	1	2	3	4	5	6	7	8	9
# of Machines	1	1	1	2	1	1	1	1	1
Mean	8	22	4.5	4	2.5	2.2	2	5	5
Stdev	0.7	1.6	0.4	0.4	0.2	0.24	0.2	0.4	0.5

Before the simulation-based metamodeling method is applied on a multi-station system, analytical capacity/bottleneck analysis (Section 4.1) can be performed. Using the analytical models in Hopp et al. (2002), Station 4 was identified as the BN with a capacity of 0.125 lots per time unit, which is also the system capacity. The second heaviest utilized station is Station

2 with a capacity of 0.1815 lots per time unit. Thus, this wafer fab is considered as being dominated by a single BN station.

A.5 The Decomposition Approach for the Transient Analysis of A Multi-Station System

As pointed out in Section 6.2.1, the aggregation approach for multi-station systems is the focus of this paper. However, for the sake of completeness, here we provide a preliminary exploration of the decomposition version of the transient analysis method by applying it on a simple multi-station system.

The system consists of five tandem stations. Each station includes three identical servers, and all the service times follow gamma distribution (34) with parameters given as in Table 3. Each server in the system is subject to random failures. Both the time to failure and the time to repair follow exponential distributions with mean time to failure being 900 time units and mean time to repair being 100 time units. The arrivals are modeled as Poisson.

Table 3: Distribution parameters for the processing times in the five-station system.

	Station 1	Station 2	Station 3	Station 4	Station 5
Parameter α	9	9	9	9	9
Parameter β	0.9	0.828	0.63	0.756	0.9

The basic idea of the decomposition method is to decompose the system into five stations (or groups), and to characterize each of them by its own set of TFMs. Recall that each set of TFMs takes forms as those in models (23). The dynamic behavior of the entire system can be described by the multiple sets of TFMs as shown in Figure 7. Specifically, Station i is characterized by the TFMs^[i], with the superscript [i] denoting station i ($i = 1, 2, \dots, 5$). The input rate to the first station $x^{[1]}(t)$ is the input rate to the entire system $x(t)$, and the input rate to a downstream station i ($i = 2, 3, \dots, 5$) is the departure rate from its upstream station $d^{[i-1]}(t)$. The ensemble of TFMs $\{\text{TFMs}^{[i]}; i = 1, 2, \dots, 5\}$ can be used to predict the system's dynamic performance over the time period $1, 2, \dots, P$ under any input $\{x^*(t), t = 1, 2, \dots, P\}$. The prediction process is given as follows.

Step 1: With $\{x^{[1]}(t) = x^*(t); t = 1, 2, \dots, P\}$ and the history for Station 1 over $(-\infty, 0]$, the

TFMs^[1] are used to recursively compute $\hat{\mathbf{y}}^{[1]}(t) = (\hat{m}^{[1]}(t), \hat{d}^{[1]}(t))$ for $t = 1, 2, \dots, P$. Set $i = 2$.

Step 2: Given $\{x^{[i]}(t) = \hat{d}^{[i-1]}(t); t = 1, 2, \dots, P\}$ and the history for Station i over $(-\infty, 0]$, the TFMs^[i] are then used to recursively compute $\hat{\mathbf{y}}^{[i]}(t) = (\hat{m}^{[i]}(t), \hat{d}^{[i]}(t))$ for $t = 1, 2, \dots, P$. Set $i = i + 1$.

Step 3: If $i > 5$, then stop; otherwise, go to Step 2.

As the empirical examples in Section 6, EDS were collected for the fitting of the five TFMs. For the EDS, the arrival rate in the simulation follows a piecewise constant function like the one in Figure 1 (a). Note that when simulating the system, the data needed for estimating each set of TFMs can be obtained for the five stations simultaneously. To evaluate the prediction ability of the sequence of five TFMs, VDS were also obtained by running simulation under the sine arrival rate like that in Figure 1 (b). Plots were made to compare the “true” system evolution from the VDS and that predicted by the five TFMs. Figures 8–12 show the comparisons for station 1–5 respectively.

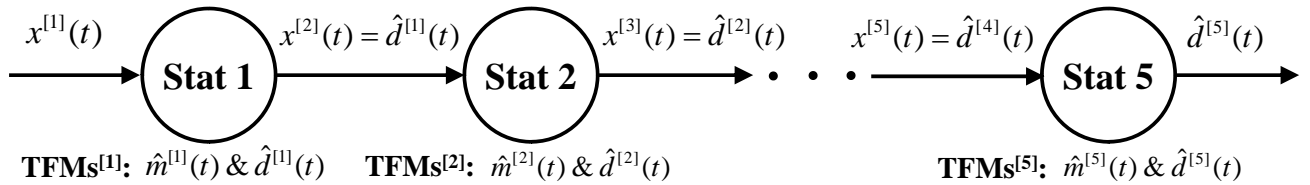


Figure 7: The five tandem stations characterized by five sets of TFMs

We next discuss the pros and cons of the decomposition approach compared to the aggregation one. Obviously, the decomposition approach intends to describe the system’s behavior on an individual station basis, and provides more information regarding the performance of each station in the system. However, the decomposition approach relies on an additional assumption aside from the three constraints presented in Section 1.4. Recall that applying a single set of TFMs on an individual station requires the arrivals to that station to be fully characterized by the arrival rate (Constraint 3 in Section 1.4). As can be seen from Figure 7, the departures from station $i - 1$ serve as the arrivals to station i ($i = 2, 3, \dots, 5$). Hence, to describe the downstream station i by TFMs^[i], the assumption needed is that the departures from station $i - 1$ (equivalently, the arrivals to station i) can be characterized by its first moment measure, departure

(arrival) rate. This assumption apparently introduces an additional layer of approximation to the transient analysis of the system.

To evaluate the impact of the departure approximation on the accuracy of the transient behavior represented by the ensemble of multiple sets of TFMs, extensive real system experiments need to be performed, which is beyond the scope of this paper. Here, we intend to empirically demonstrate the potential of the decomposition method through the simple system with five tandem stations.

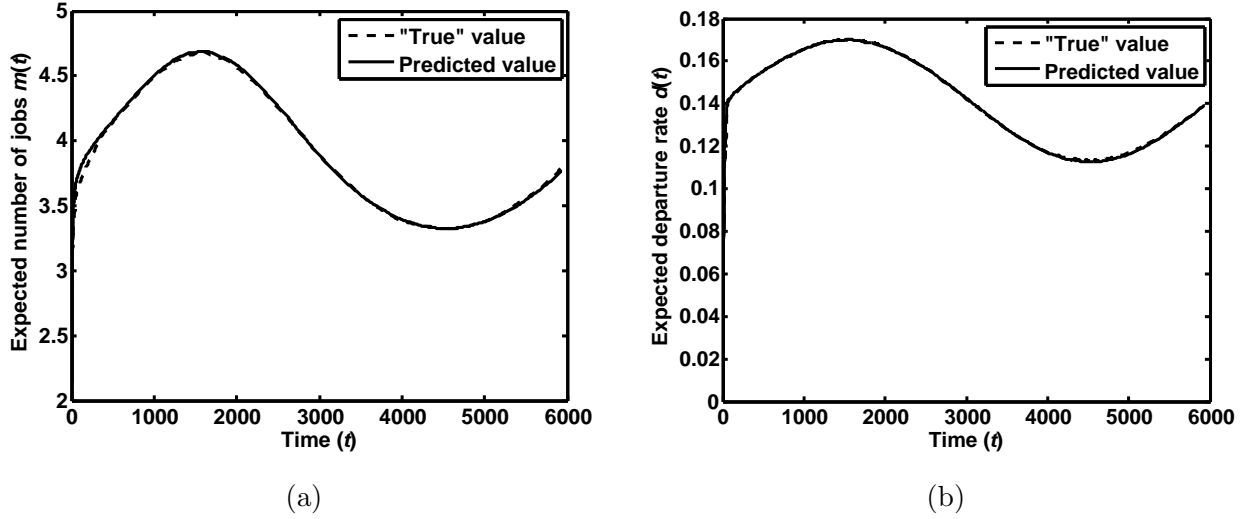


Figure 8: Comparison of the predicted dynamic outputs and their “true” values for Station 1.

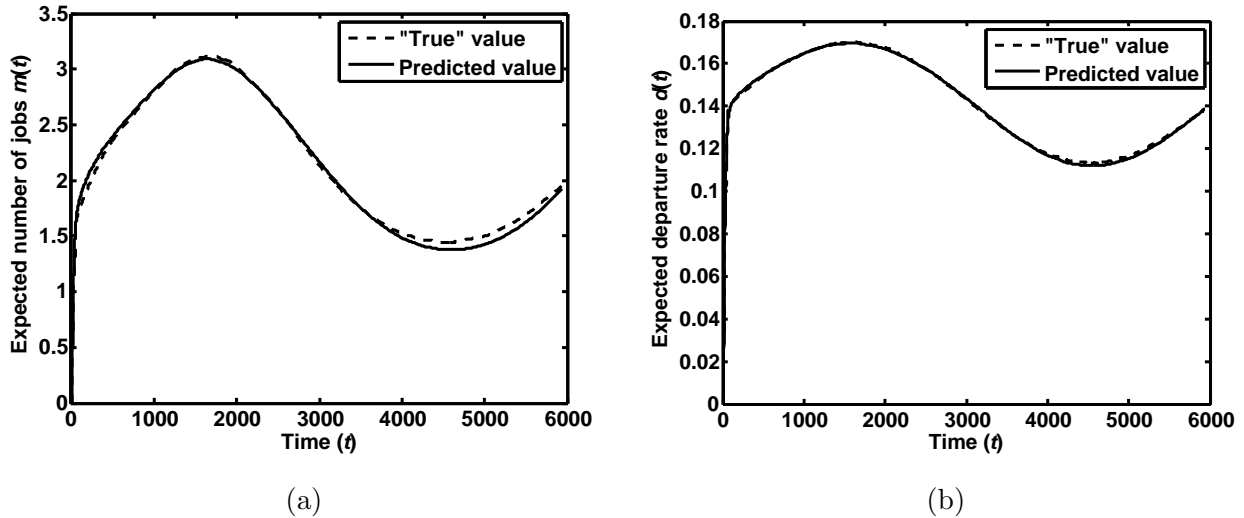
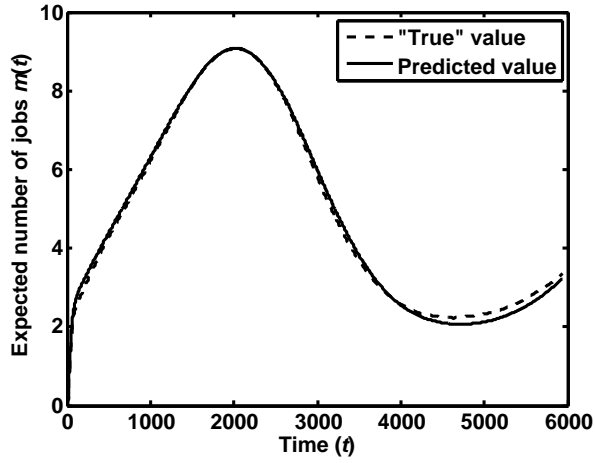
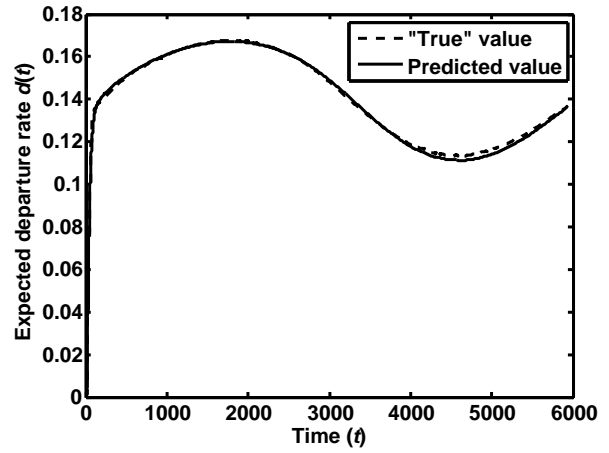


Figure 9: Comparison of the predicted dynamic outputs and their “true” values for Station 2.

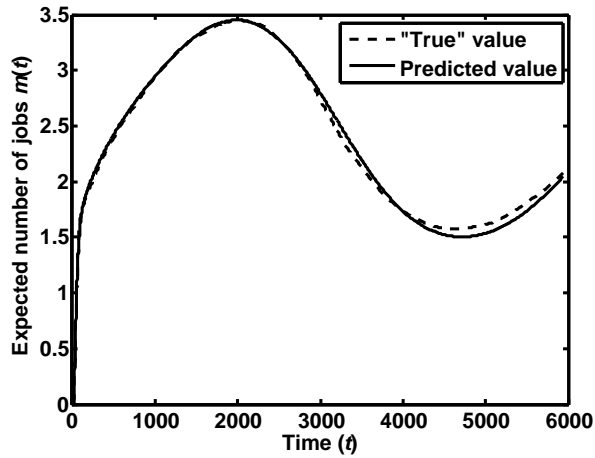


(a)

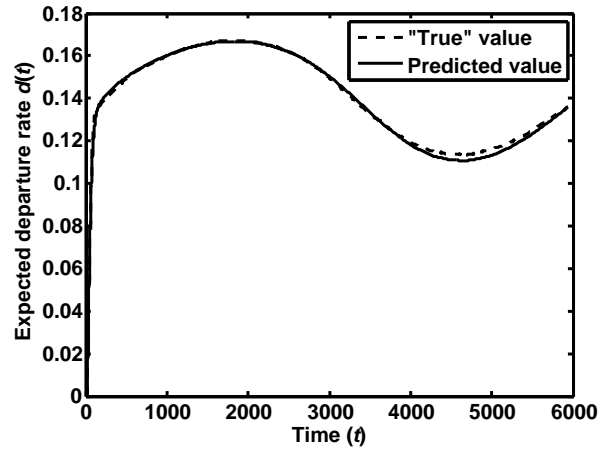


(b)

Figure 10: Comparison of the predicted dynamic outputs and their “true” values for Station 3.

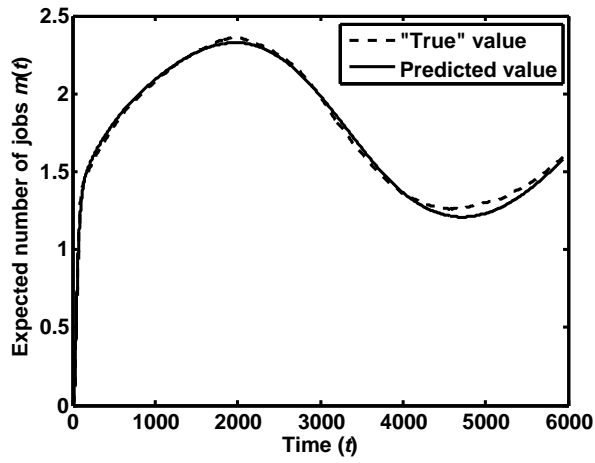


(a)

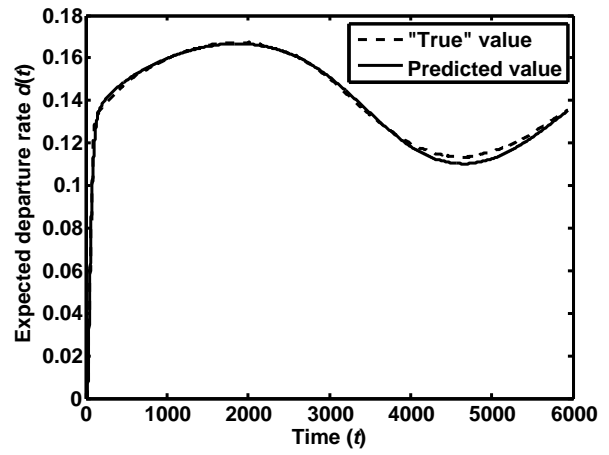


(b)

Figure 11: Comparison of the predicted dynamic outputs and their “true” values for Station 4.



(a)



(b)

Figure 12: Comparison of the predicted dynamic outputs and their “true” values for Station 5.