Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

TRANSIENT ANALYSIS OF GENERAL QUEUEING SYSTEMS VIA SIMULATION–BASED TRANSFER FUNCTION MODELING

Feng Yang
Jingang Liu

Industrial and Management Systems Engineering Dept.
West Virginia University
Morgantown, WV 26506, USA

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 release rate of entities to the system; the time-dependent output performances include the output rate of entities and the mean of the work in process (i.e., number of entities 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. The proposed 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.

1 INTRODUCTION

This work is concerned with the transient behavior of general queueing systems. The primary motivation stems from production planning in manufacturing, for which one of the major difficulties encountered is uncertainty, such as demand-forecast mismatches, unexpected interruptions in production, and natural disasters (Blackhurst et al. 2005; Datta, Christopher, and Allen 2007; Koh 2004; Pinedo 2007; Sheffi and Rice 2005; Stadtler and Kilger 2007; Tang 2006). In case of such disruptions, an immediate reaction is required and a new production plan needs to be generated responsively for the next few weeks or months. A good plan recognizes the current status of the situation, takes into account future evolution of the system, which can be treated as a queueing system, and achieves the best overall performance (e.g., low cost, high customer service level, etc.). The key for responsively generating such a good plan lies in the ability to capture the evolution of the system, which is the focus of the proposed research.

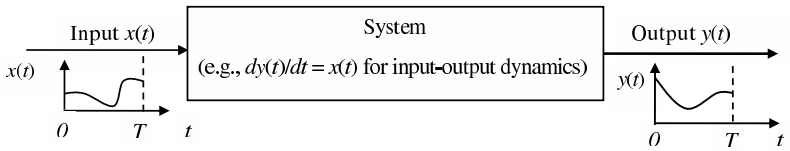

Figure 1: Time-dependent input-output dynamics of a system.

This work intends to characterize the input-output dynamics of a general queueing system, which is illustrated in Figure 1. Specifically, in our work, the input $x(t)$ represents the time-varying arrival rate of entities to the system, and the dynamic outputs $\mathbf{y}(t)$ include two time-dependent performance metrics: the expected departure rate of entities and the mean of the work in process (number of jobs in the system). The objective of this paper is to develop a statistical methodology, integrated with

extensive offline simulation and preliminary queueing analysis, for the generation of a number of transfer function models (TFMs) that quantify the time-dependent (transient) input-output relationships for general queueing systems. The resulting TFMs have two major advantages. (i) The TFMs embody the high fidelity of simulation since they are estimated from detailed simulation data representing a wide range of system operating conditions. (ii) The TFMs are difference equations, like the discrete approximations of the ordinary differential equations (ODEs) provided by an analytical approach; supposing that a certain input $x(t)$ is fed to the system under given initial conditions, the TFMs can be used to recursively compute the system's future performance $\mathbf{y}(t)$ in a timely manner with no need to run additional simulations. Hence, the TFMs resulting from the proposed work are able to accurately describe the transient dynamics of systems, as well as provide prompt "what-if" analysis.

It is worth mentioning that the simulation-based transfer function modeling falls into the category of metamodeling (Chapter 18, Henderson and Nelson 2006), which refers to the techniques that utilize simulation to generate mathematical approximations quantifying the relationships implied by the simulation. This work, to the best of our knowledge, is the first attempt to develop a metamodel that takes the form of difference equations, but nevertheless applies to the context 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 time is relatively scarce or expensive. The responsive production planning mentioned above represents one of such contexts: Simulation models for the 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, a decision needs to be made quickly–as soon as possible–regarding how to adjust the production plan for that system. The metamodel, i.e., the TFMs in this paper, fully utilizes the plentiful offline simulation time and allows for responsive decision making in time of urgency.

2 LITERATURE REVIEW

In the literature, both analytical methods and computer simulation have been used to address the time-dependent behavior of queueing systems.

For Markov queueing models, time-dependent ODEs can be developed to represent their input-output dynamics. However, analytical solutions to these ODEs are rare. A few exceptions include the known solutions for the M(t)/G/∞ and M/M/1 systems (Gross and Harris 1985; Kleinrock 1975), and the Ph(t)/Ph(t)/∞ systems investigated by Nelson and Taaffe (2004a, 2004b). The mainstay of the analytical work on transient analysis has been the development of numerical solutions of the time-dependent ODEs characterizing the transient behavior of the Markov models. 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 (1991), Green, Kolesar, and Svornos (1991), Eick, Massey, and Whitt (1993a, 1993b), Jennings et al. (1996), Massey and Whitt (1997). 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, Zachary, and Ziedins 1996). The analytical method developed in Riano (2003) can be considered as a parallel to the fluid and diffusion approximations. Green, Kolesar, and Whitt (2007) also reviews various queueing methods for approximating the transient performance of service systems such as call centers. 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, machine failures, reentrant 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 demanding for real-time "what-if" analysis.

The proposed work integrates 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. The remainder of the paper is organized as follows. Section 3 provides an overview of the proposed methods. The preliminary queueing analysis is performed in Section 4. The approach of the simulation-based transfer function modeling is described in Sections 5 and 6. In Section 7, the proposed methods are applied to a range of general queueing systems including the one that mimics the features of real semiconductor manufacturing systems; the transient analysis of multi-station systems is particularly discussed in Section 7.2.

3 OVERVIEW OF THE METHODOLOGY

We consider the system of interest as a queuing 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 $(u,v]$, $u < v \in (-\infty, \infty)$.
- $D(u,v)$: the random variable counting the number of departures in the system within the time interval $(u,v]$, $u < v \in (-\infty, \infty)$.

Both $A(u,v)$ and $D(u,v)$ are general point processes (Cox and Isham 1980) that count the number of event occurrences over a time interval, special examples for which include Poisson, renewal, self-exciting processes, and marked point processes (Daley and Vere-Jones 2002). In this work, it is assumed that neither the arrival pattern nor the service times depend on the state of the system. Hereby (until Section 4.2), we restrict our discussion to a single-station system. The extension to multi-station environment will be discussed in Section 7.2.

Let $\mathscr{H}_0 = \{Q(t), A(-\infty,t), D(-\infty,t), t \in (-\infty, 0]\}$ 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 $\mathscr{H}_0$? Note that $\{A(0,t), t > 0\}$ is considered as the independent variable (the input flow imposed on the system), 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. We define the following notations.

$m(t) = \mathbb{E}[Q(t)]$, the expectation (first moment) of the number of jobs in the system at time t.
$a(t) = \lim_{\delta \to 0^+} \delta^{-1} \mathbb{E}[A(t, t+\delta)]$.
$d(t) = \lim_{\delta \to 0^+} \delta^{-1} \mathbb{E}[D(t, t+\delta)]$.

It is assumed that $a(t)$ and $d(t)$ exist and are finite. Usually, $a(t)$ and $d(t)$ are also referred to as the intensity or rate of the corresponding point process. Denoting $\mathbf{y}(t) = (m(t), d(t))$ as the 2×1 vector including the output performance variables, and $x(t) = a(t)$ the input variable, we aim at characterizing the input-output dynamics of a queueing system by a number of TFMs:

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

which is a discrete-time functional approximation that describes the dynamics of the queueing system. 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, and any possible confusion is avoidable at the price of a negligible amount of mental energy.

The vector function $\mathbf{F}$ in the TFMs (1) includes two equations, 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 $(0,T]$. Given the seed values of $\{x(t), \mathbf{y}(t)\}$, which can be derived from $\mathscr{H}_0$, we can use the TFMs to compute recursively the system's future performance $\{\mathbf{y}(t), t \in (0,T]\}$ under any input $\{x(t), t \in (0,T]\}$.

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 4): 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 for the TFMs.
- *Data collection via offline simulation* (Section 5): Under selected input processes, we run simulations to obtain paired input-output time series. We emphasize that our simulation is carried out offline in advance of the need to make a decision.
- *Transfer function modeling* (Section 6): From the simulation data, we develop statistical methods to obtain the parsimonious TFMs (1) that are adequate to capture the system's dynamic behavior.

4 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 our transfer function metamodeling approach.

4.1 An M(t)/M/∞ Example

For the purpose of intuition and motivation, we consider the input-output dynamics of the simple queueing model M(t)/M/∞, 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/∞:

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

These equations characterize the system evolution in terms of $m(t)$ and $d(t)$. Given the initial state of the system at time 0, the numeric solution of $\{\mathbf{y}(t) = (m(t), d(t)), t > 0\}$ can be obtained for any input $\{x(t), t > 0\}$.

Unfortunately, the situation becomes much more complicated as a finite number of servers is introduced or the Markovian assumption is relaxed. The objective of the proposed work is to obtain a discrete-time approximation of equations like (2) for a general queueing system so that its dynamic behavior can be characterized.

4.2 A General Queue

We consider a single-station queueing process $Q(t)$ with arrivals $A(t)$ and departures $D(t)$, as described in Section 3. The arrival and departure rates are denoted as $a(t)$ and $d(t)$ respectively. The additional assumptions made solely for the analytical analysis of this section are:

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

Conditions (3) imply that there are no multiple simultaneous arrivals or departures, i.e., both $A(t)$ and $D(t)$ are orderly point processes (Dalye and Vere-Jones 2002).

Following the notation given in Section 3, let

$$\begin{aligned} a_n(t) &= \lim_{\delta \to 0^+} \delta^{-1} \Pr\{A(t, t+\delta) = 1, Q(t) = n\} \\ d_n(t) &= \lim_{\delta \to 0^+} \delta^{-1} \Pr\{D(t, t+\delta) = 1, Q(t) = n\} \end{aligned} \tag{4}$$

Here, $a_n(t)$ denotes the arrival rate at time t while there are n jobs (not including the one that is about to enter) 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. Apparently, we have $a(t) = \sum_{n=0}^{\infty} a_n(t)$ and $d(t) = \sum_{n=1}^{\infty} d_n(t)$.

Suppose that the system consists of a single server with service time following a general distribution, say $G(\tau)$, where $\tau \in (\tau_L, \tau_U)$, the feasible time range for the service time. It is required that $0 < \tau_L < \tau_U$. Jobs are served on a first come first served basis. For this general queue, we have derived the $x(t)$-$\mathbf{y}(t)$

relationship with $x(t) = a(t)$ and $\mathbf{y}(t) = (m(t), d(t))$:

$$\begin{aligned} m'(t) &= a(t) - d(t) \\ d(t) &= \int_{\tau_L}^{\tau_U} a_0(t-\tau) dG(\tau) + \int_{\tau_L}^{\tau_U} (d(t-\tau) - d_1(t-\tau)) dG(\tau) \end{aligned} \tag{5}$$

Unlike equations (2) for the M(t)/M/∞ system, equations (5) for the general queue are not closed, and thus not solvable: Aside from the input process $a(t)$ and the output processes of interest $m(t)$ and $d(t)$, (5) also involves unknown time-dependent functions $a_0(t)$ and $d_1(t)$. However, for mediumly/heavily loaded queues, it is reasonable to assume that $a_0(t)$, the arrival rate when no job is 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:

$$a_0(t) \approx p_0(t) \times a(t) \approx e_1 a(t) \text{ and } d_1(t) \approx p_1(t) \times d(t) \approx e_2 d(t).$$

Both e_1 and e_2 are small fractional constants. Further, if we take the finite-difference approximation of the derivative and integrals in (5), it is clear that the discrete approximations of equations (5) fall into the category of TFMs (1). Similar dynamic equations as (5) have also been obtained for single-station systems with multiple servers.

The analytical results (5) serve three purposes here. First, it shows that even for a single-server queue with general arrivals and services (a very simple queue), its non-stationary behavior is analytically intractable. Hence, the approach of TFMs-based discrete approximation may be appropriate for investigating the time-dependent behavior of general queueing systems. Second, as will become clear in Section 7.2, the basis of describing the dynamics of a multi-station system lies in the use of TFMs (1) to approximate the transient behavior of a single station (or a group of stations that can be considered as a whole), and the single-station queues (with one or multiple servers) considered above give a fairly general representation. Thus, equations (5) strongly suggest that the TFMs as (1) are likely to be successful in terms of capturing the system dynamics. As a matter of fact, it was these analytical results that motivated us to adopt the TFMs (1) in the first place. Third, equations (5) provide some valuable insights as to the specific functional forms of the target TFMs, which is very useful in the statistical fitting of the parametric models.

As already noted, the additional assumptions (3) were made here solely for the analytical analysis. Whereas the TFM modeling, as evident in Section 7, is expected to be able to describe the dynamic behavior for general queueing systems with failures and re-entrant flows. Next, in Sections 5 and 6, we discuss in detail the issues associated with the simulation-based TFM modeling.

5 DATA COLLECTION VIA OFFLINE SIMULATION

In this part, we discuss how to obtain the simultaneous pairs of the input-output observations $\{(X(t), \mathbf{Y}(t)), t = 1, 2, \ldots, T\}$ by running simulation. Note that the capital letters here are used to represent the estimated time series obtained from simulation. In this work, discrete event simulation models are constructed to represent the queueing systems of interest. The input flow of entities $A(t)$ is modeled as a point process which is characterized by its first moment measure, i.e., the input rate $a(t)$ (Section 3). As will be seen in Section 7, in our experiments two types of input processes are fed to the system: Poisson and equilibrium renewal processes with $a(t)$ being a piece-wise constant function over time t (e.g., Figure 2).

For a given queueing system, a number of, say I, simulation replications are performed with the input flow being a stochastic process characterized by a time-varying rate. For replication i ($i = 1, 2, \ldots, I$), the arrival, departure and state processes $\{A_i(t), D_i(t), Q_i(t); t = 1, 2, \ldots, T\}$ are recorded. It is assumed that the system is observed at discrete, equispaced intervals of time, and that the basic sampling interval Δt serves as the unit of time. The paired time series $\{(X(t), \mathbf{Y}(t)), t = 1, 2, \ldots, T\}$ are estimated as

follows.

$$\begin{aligned} X(t) &= \widehat{a}(t) = \frac{I^{-1}\sum_{i=1}^{I} A_i(t-\Delta t/2, t+\Delta t/2)}{\Delta t} \\ Y_1(t) &= \widehat{m}(t) = I^{-1}\sum_{i=1}^{I} Q_i(t) \\ Y_2(t) &= \widehat{d}(t) = \frac{I^{-1}\sum_{i=1}^{I} D_i(t-\Delta t/2, t+\Delta t/2)}{\Delta t} \end{aligned} \tag{6}$$

It can be seen from (6) that both the arrival rate $X(t)$ and and the departure rate $Y_2(t)$ are defined in terms of the average number of occurrences per Δt. The sampling interval should be sufficiently small to allow all the systematic variation which occurred in the inputs/outputs to be taken account of. In our experiments, we set Δt to be one tenth of the expected processing time of the server, which is typically smaller than the average interarrival time of entities.

Although the simulation involved in the proposed work is performed offline and the simulation time is assumed sufficient, it remains important to design simulation experiments so that accurate TFMs can be obtained at high computational efficiency. In this context, the design of experiments is concerned with the following questions. How to specify the piece-wise constant function $x(t) = a(t)$ for the input arrivals of the simulation experiments? How many replications should be performed at the selected time-varying input process? Due to the space constraint, no specifics for the design strategies will be given here, and the readers can obtain some idea from the empirical examples in Section 7.

6 STATISTICAL MODELING ISSUES OF THE TFMS

The modeling of the system dynamic behavior is based on the pair estimates $\{X(t), \mathbf{Y}(t), t = 1, 2, \ldots, T\}$ 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):

$$\mathbf{Y}(t) = \mathbf{F}(\theta; X(t-1), X(t-2), \ldots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \ldots) + \mathbf{e}(t), \tag{7}$$

where $X(t) = \widehat{a}(t)$ and $\mathbf{Y}(t) = (\widehat{m}(t), \widehat{d}(t))$ as given in (6). The term $\mathbf{e}(t) = (e_1(t), e_2(t))$ denotes the disturbance. The parameter vector θ includes all the unknown parameters involved in the vector function $\mathbf{F}$. For convenience of the discussion, we also write model (7) as:

$$\begin{aligned} Y_1(t) &= F_1(\theta_{\mathbf{1}}; X(t-1), X(t-2), \ldots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \ldots) + e_1(t) \\ Y_2(t) &= F_2(\theta_{\mathbf{2}}; X(t-1), X(t-2), \ldots, \mathbf{Y}(t-1), \mathbf{Y}(t-2), \ldots) + e_2(t), \end{aligned} \tag{8}$$

with $\theta = (\theta_{\mathbf{1}}, \theta_{\mathbf{2}})$. Our task here 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 $(X(t), \mathbf{Y}(t))$. Due to space constraint, next we briefly explain the model estimation and selection issues without providing the details.

Estimation of the TFMs. Assuming that a specific functional form (model structure) has been selected, the TFMs can be fitted from the simulation data. In this work, we assume that each disturbance $e_i(t)$ $(i = 1, 2)$ can be approximated by a stationary autoregressive moving average (ARMA) process (Box, Jenkins, and Reinsel 1994), and the least square methods are used to obtain the fitted TFMs. In Ljung (1999), the asymptotic normality of the least-square parameters $\widehat{\theta}$ has been proved, and the statistical inference can be performed on the estimated TFMs.

Model Selection. The estimation of the TFMs above is based on a given functional form, but how do we select the most appropriate structure for the target TFMs? 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 best TFMs, which will not be detailed here.

7 EMPIRICAL EXAMPLES

For a given simulation representing a general queueing system, we seek to describe its transient behavior by generating a number of TFMs from simulation data. Using such TFMs, a system's future dynamics can be predicted in a timely manner without running additional simulation, which could be very time consuming. In this part, two examples are presented to illustrate the effectiveness of the proposed methods: an $E_k(t)/G/1$ system (Section 7.1), and a system with six different stations and re-entrant flows (Section 7.2). The first two single-station cases are selected from a number of queueing models (Table 1) on which the TFM modeling methods have been successfully applied, and these queueing models are intended to show that the proposed methods can handle a wide range of input flows and different types of service time distribution. Note that in Table 1, ACV and SCV denote the coefficient of variation (CV) for the distribution of interarrival times and service times respectively. Accurately characterizing the transient behavior of a single station (or a group of stations that can be well approximated as a single one) serves as the basis for capturing the dynamics of a multi-station system. In Section 7.2, the six-station example also involves re-entrant flows, one of the main features of real semiconductor fabrication systems, and the specifics of extending the TFM modeling to multi-station systems are detailed through this example.

For each queueing system, the proposed methods were applied for the generation of the TFMs describing the system dynamics. The TFMs were fitted from the estimation data set (EDS), and can be used to predict the future evolution of the system under any input flow. To evaluate the prediction provided by the TFMs, a validation data set (VDS), which contains simulation data different than and independent of those in the EDS, was collected and the system dynamics estimated from 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.

Table 1: Single-Station System Configurations.

# Servers	Interarrival Time	ACV	Service Time	SCV	Failures
$1 \sim 3$	exponential,Erlang,deterministic	$0 \sim 1$	gamma	$0.1 \sim 1$	Yes/No

Before discussing the results, it is worth mentioning that in our discrete TFMs, one time unit represents the sampling interval Δt, which is set as about one tenth of the expected service time of the most heavily utilized server (Section 5). 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.

7.1 An $E_k(t)/G/1$ System

We consider a single-server system whose service time follows a gamma distribution with a mean of 10 time units (i.e., $10\Delta t$) and standard deviation of 5 time units. The interarrival time of entities follow an Erlang distribution with $k = 25$ stages (denoted as E_{25}), corresponding to a CV of 0.2.

To collect the time series data $\{X(t), \mathbf{Y}(t)\}$ for the TFM modeling, simulation experiments were carried out by feeding to the system the arrivals with the piece-wise arrival rate shown in Figure 2(a). Each piece in Figure 2(a) corresponds to a stationary renewal process with a certain first moment measure (rate), the simulation methods of which are discussed in Daley and Vere-Jones (2002) and implemented in our simulation model. The five selected arrival rates are evenly-spaced to cover the system utilization range of $[0.5, 0.98]$. The simulation length of each constant-rate period is selected to ensure that sufficient data is obtained in steady state. The number of simulation replications performed in this case is 10000. From the multiple replications, the paired estimates $\{X(t), \mathbf{Y}(t)\}$ were calculated using equations (6). With the collected EDS, the statistical modeling methods were applied and the resulting TFMs for this $E_k(t)/G/1$ system are given as follows:

$$\begin{aligned}
\widehat{m}(t) &= 1.0045m(t-1) - 0.0798d(t-1) + 0.0902x(t-1) \\
&\quad -0.0512m(t-1)x(t-1) + 0.0452m(t-1)d(t-1)x(t-1) \\
\widehat{d}(t) &= 0.0032m(t-1) + 0.9474d(t-1) + 0.0431x(t-1) \\
&\quad +0.0286m(t-1)x(t-1) - 0.0304m(t-1)d(t-1)x(t-1)
\end{aligned} \tag{9}$$

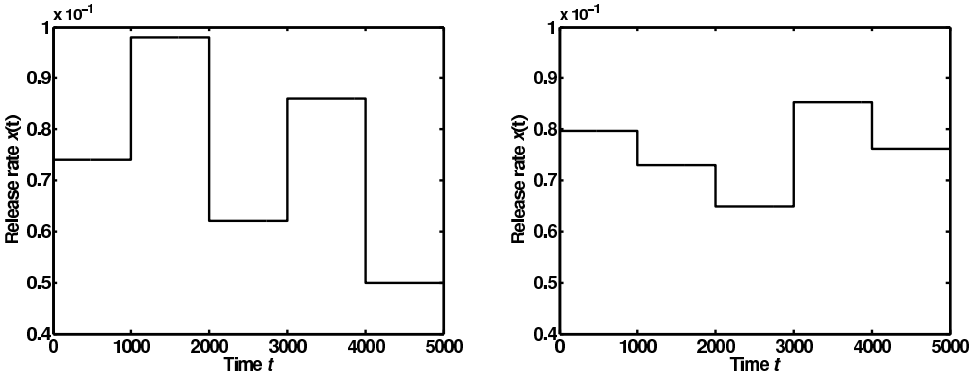

Figure 2: Release rate for estimation and validation data set for $E_k(t)/G/1$ system

.

Apparently, given the history $\{x(t), \mathbf{y}(t) = (m(t), d(t)), t \leq 0\}$, the fitted TFMs (9) can be used to recursively compute the future performance for any input $\{x^*(t), t \in (0, T]$, and the computational effort required is negligible.

To evaluate the accuracy of the TFMs (9), the VDS were collected by running simulation with the interarrival time following E_{25} and the time-varying arrival rate given in Figure 2(b). To avoid TFMs-based extrapolation, the arrival rates in the VDS are set within the rate range $[x_L, x_U]$ used in the EDS. For the VDS, 50000 simulation replications were performed, and highly accurate time series $\mathbf{y}(t) = (m(t), d(t))$ were obtained and considered as the "true" dynamic outputs with "zero" variance under the specified input flow. In Figure 3, the "true" outputs $m(t)$ and $d(t)$ are plotted as the dotted curves in Figure 3(a) and (b) respectively. The solid curves in Figure 3 represent the predicted dynamic outputs resulting from the fitted TFMs (9). 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 2(b). Figure 3 shows that the predicted dynamics from the TFMs almost coincide with the "true" system evolution.

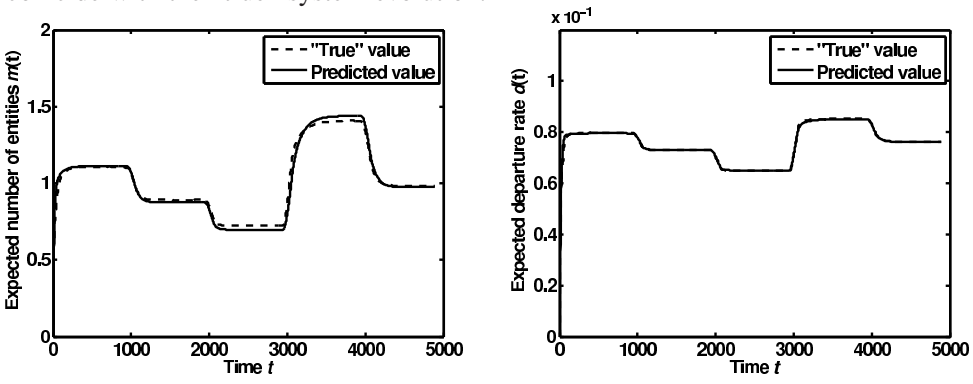

Figure 3: Comparison of the predicted dynamic outputs and their "true" values for the $E_k(t)/G/1$ system

.

7.2 A Multi-Station System with Re-entrant Flows

The TFMs that well characterize the transient behavior of a single station (or a group of stations) provide the building blocks for describing the dynamics of multi-station systems, as will become clear

in this subsection. We illustrate the TFMs-based modeling methods through the system depicted in Figure 4, which includes re-entrant flows, one of the main features of real semiconductor fabrication systems. The system consists of six stations with two re-entrant cycles: $2 \rightarrow 3$, and $4 \rightarrow 5$. Each entity has to visit the first cycle twice before it enters the second cycle, which also needs to be repeated by an entity for two times. Each station consists of three identical servers, and all the service times follow Gamma distribution with a CV of 0.5. The mean service time at each of the six stations is given in Table 2.

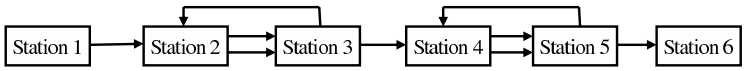

Figure 4: The six-station system

.

7.2.1 Extension to Multi-Station Systems

The basic idea to analyze a multi-station system is to decompose the system into a number of subgroups, treat each subgroup as a single station, and characterize each of them by its TFMs, like those in equations (9). The dynamic behavior of the entire system can be described by the multiple sets of TFMs with each set corresponding to a subgroup. The specifics are discussed as follows.

The decomposition of a target system is based on the identification of the most heavily utilized stations (HUSs). A bottleneck station (BNS) is defined as a station that has the maximum utilization in the system. We consider a station whose utilization is above 80% of that of the BNS(s) as a HUS. The HUSs are the stations that most constrain the entity flows and thus play a key role in determining the overall performance of the system. For a given system, analytical queueing models in the literature (e.g., Hopp et al. 2002; Kumar and Kumar 2001; Meng and Heragu 2004) are available to perform utilization analysis for even the most complicated manufacturing systems (i.e., semiconductor manufacturing systems), and thus the HUSs can be identified analytically prior to the simulation-based transfer function modeling. Denoting G as the number of HUSs in a system, we suggest formulating G subgroups: each subgroup includes one HUS, which dominates the queueing behavior of the group, and some upstream/downstream non-HUSs of that HUS.

The system decomposition has to be made on a case-by-case basis. Here, we provide a simple illustration through the example in Figure 4. We decompose the six-station system into two subgroups, mainly based on the utilization analysis discussed above. Stations 3 and 5 are considered as HUSs, and the remaining stations are non-HUSs. Hence, Subgroup 1 contains Stations 1, 2, and 3; and Subgroup 2 includes Stations 4, 5, and 6. As illustrated in Figure 5, in our transient analysis, Subgroup i is characterized by the TFMs$^{[i]}$, a set of TFMs like the one in (9), with the superscript $^{[i]}$ denoting the group i ($i = 1,2$). The input rate to the first group $x^{[1]}(t)$ is the input rate to the entire system $x(t)$, and the input rate to the second group $x^{[2]}(t)$ is the departure rate from the first group $d^{[1]}(t)$. The two sets of TFMs$^{[i]}$ ($i = 1,2$), will be used to characterize the transient behavior of the system and to predict the system dynamics under any input $x(t)$.

The approach of decomposing a system into subgroups and characterizing each group by a set of TFMs is obviously approximate. The rationale behind this approximation is two fold. First, the transient effects at non-HUSs are negligible, that is, the time it takes for a non-HUS to reach steady state is negligible. Thus a subgroup can be considered as a whole with its behavior dominated by the sole HUS. Second, the implicit assumption made in modeling a subsequent group is that the departures from the previous group (i.e., the arrivals to this subsequent group) are approximately completely characterized by the first moment measure, the departure rate. This approximation is supported by the departure analysis in Buzacott and Shanthikumar (Chapter 3, 1992), and also works empirically well in our experiments.

Table 2: Configuration parameters for the six-station system.

	Station 1	Station 2	Station 3	Station 4	Station 5	Station 6
Mean Service Time	10	10	7	10	7.8	10

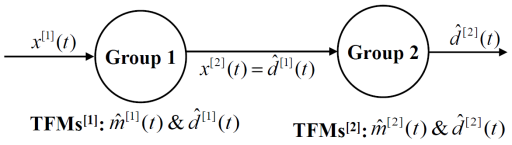

Figure 5: Decomposition of 6 station in tandem system

.

7.2.2 Modeling Results for the Multi-Station System

We present the modeling results of the six-station system which is decomposed into two subgroups as shown in Figure 5. The EDS was obtained by simulating the system with Poisson arrivals at a piecewise-constant rate $x(t)$ similar to that in Figure 2(b). From the simulation experiments performed, the time series data $\{X^{[i]}(t), \mathbf{Y}^{[i]}(t), t = 1,2,\ldots\}$ were collected for the fitting of TFMs$^{[i]}$ $(i = 1,2)$. The resulting two sets of TFMs$^{[i]}$ $(i = 1,2)$ can be used to predict the system performance under any input $\{x^*(t), t \in (0,T]\}$, and the prediction consists of two steps corresponding to two subgroups.

1. With $x^{[1]}(t) = x^*(t)$ and the identified history for Group 1, the TFMs$^{[1]}$ are used to recursively compute $\widehat{\mathbf{y}}^{[1]}(t) = (\widehat{m}^{[1]}(t), \widehat{d}^{[1]}(t))$ for $t \in (0,T]$.
2. Given $x^{[2]}(t) = \widehat{d}^{[1]}(t)$ and the identified history for Group 2, the TFMs$^{[2]}$ are then used to recursively compute $\widehat{\mathbf{y}}^{[2]}(t) = (\widehat{m}^{[2]}(t), \widehat{d}^{[2]}(t))$ for $t \in (0,T]$.

The goodness of the fitted TFMs$^{[i]}$ $(i = 1,2)$ is evaluated based on the VDS, which is obtained by simulating the system with Poisson arrivals following a piece-wise constant rate function. From the VDS, time series $\mathbf{y}^{[i]}(t) = (m^{[i]}1(t), d^{[i]}(t))$ $(i = 1,2)$ were obtained, and considered as the "true" dynamic outputs. In Figure 6, we compare $\widehat{\mathbf{y}}^{[i]}(t)$, the predicted outputs from the TFMs which are represented by the solid curves, and the "true" system evolution $\mathbf{y}^{[i]}(t)$ $(i = 1,2)$ which are denoted as the dotted curves. Evidently, the TFMs$^{[i]}$ $(i = 1,2)$ can accurately predict the dynamic outputs of this six-station system.

8 SUMMARY

The originality of the proposed 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 general queues.

The resulting TFMs from the proposed method are able to describe system dynamics and have two advantages. First, the TFMs embody the high fidelity of simulation since they are estimated from detailed simulation data. Second, the TFMs are difference equations, like the discrete approximations of the ordinary differential equations 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 performance in a timely manner. To efficiently generate such TFMs for queueing systems, analytical queueing analysis were performed to suggest appropriate functional forms of the TFMs; experimental design strategies were developed to efficiently collect data via offline simulation; and statistical TFM fitting methods were developed to obtain well-estimated TFMs from simulation data.

REFERENCES

Ankenman, B. E., J. M. Bekki, J. W. Fowler, G. T. Mackulak, B. L. Nelson, and F. Yang. 2010. Simulation in Production Planning. Chapter 6 in: *Planning Production and Inventories in the Extended Enterprise: A State of the Art Handbook* (eds. Kempf, K. G., P. Keskinocak and R. Uzsoy), to be published by Springer, New York.

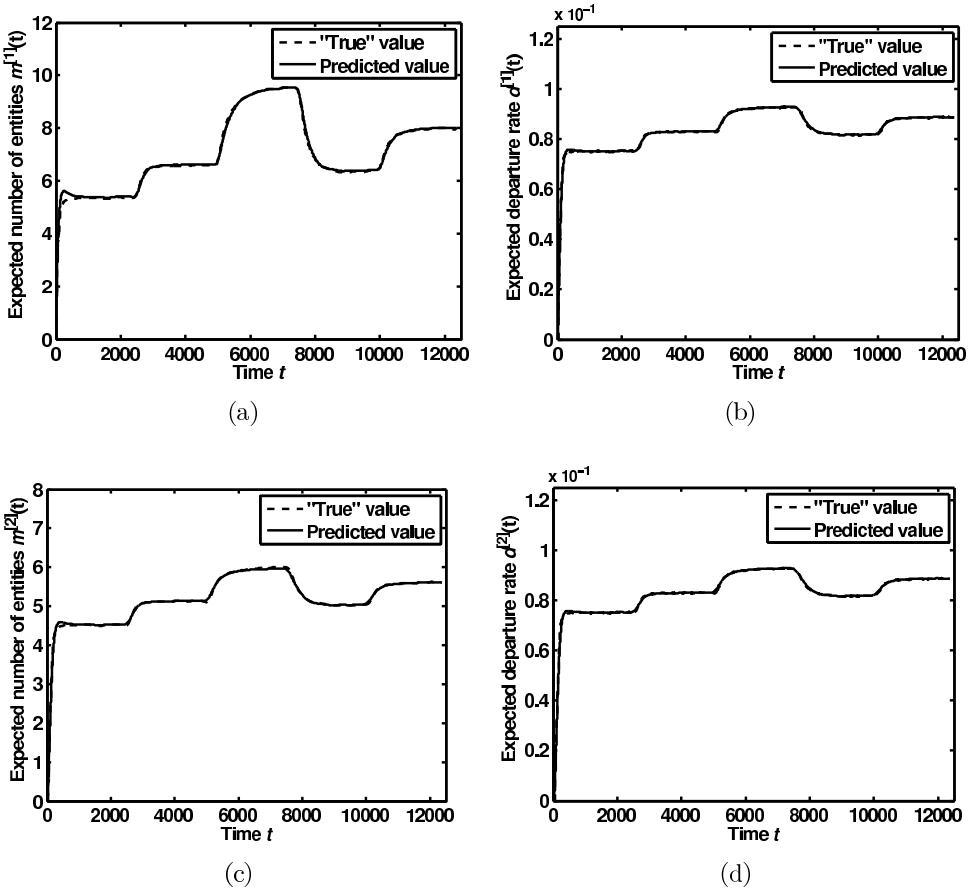

Figure 6: Comparison of the predicted dynamic outputs and their "true" values for the six-station system.

Blackhurst, J., C. W. Craighead, D. Elkins, and R. B. Handfield. 2005. An Empirically Derived Agenda of Critical Research Issues for Managing Supply–Chain Disruptions. *International Journal of Production Research* 43(19):4067-4081.

Blake, J. T., and M. W. Carter. 1996. An Analysis of Emergency Room Wait Time Issues via Computer Simulation. *Information Systems and Operational Research* 34(4):263-273.

Box, G., G. M. Jenkins, and G. Reinsel. 1994. *Time Series Analysis: Forecasting & Control* (3rd Edition). New Jersey: Prentice Hall.

Buzacott, J. A., and J. G. Shanthikumar. 1992. *Stochastic Models of Manufacturing Systems*. New Jersey: Prentice–Hall.

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.

Clark, G. M. 1981. Use of Polya Distributions in Approximate Solutions to Nonstationary M/M/s Queues. *Communications of the ACM* 24:206-217.

Coats, T. J., and S. Michalis. 2001. Mathematical Modelling of Patient Flow Through an Accident and Emergency Department. *Emergency Medicine Journal* 18:190-192.

Cox, D. R., and V. Isham. 1980. *Point Processes*. Chapman & Hall.

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.

Davies, R., and H. T. O., Davies. 1994. Modelling Patient Flows and Resource Provision in Health Systems. *Omega: The International Journal of Management Science* 22:123-131.

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.

Eick, S. G., W. A. Massey, and W. Whitt. 1993a. The physics of the Mt/G/infty Queue. *Operations Research* 41(4):731-742.

Eick, S. G., W. A. Massey, and W. Whitt. 1993b. Mt/G/infty Queues with Sinusoidal Arrival Rates. *Management Science* 39(2):241-252.

Fu, M. C., S. I. Marcus, and I-J. Wang. 2000. Monotone Optimal Policies for a Transient Queueing Staffing Problem. *Operations Research* 48:327-331.

Golub, G. H., and V. L. F. Charles. 1996. *Matrix Computations*, 3rd edition. Johns Hopkins.

Green, L. V., and P. J. Kolesar. 1991. The Pointwise Stationary Approximation with for Queues with Nonstationary Arrivals. *Management Science* 37(1):84-97.

Green, L. V., P. J. Kolesar, and A. Svornos. 1991. Some Effects of Nonstationarity on Multiserver Markovian Queueing Systems. *Operations Research* 39(3):502-511.

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.

Gross, D., and C. Harris. 1985. *Fundamentals of Queueing Theory*. New Jersey: John Wiley & Sons.

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.

Harris, C. M., K. L. Hoffman, and P. B. Saunders. 1987. Modeling the IRS Telephone Taxpayer Information System. *Operations Research* 35:504-523.

Henderson, S. G., and B. L. Nelson. 2006. *Handbooks in Operations Research and Management Science: Simulation*. Amsterdam, Netherlands: Elsevier Science.

Hopp, J. W., 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.

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.

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.

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.

Kleinrock, L. 1975. *Queueing Systems*. New York: John Wiley & Sons.

Koh, S. C. L. 2004. MRP-Controlled Batch–Manufacturing Environment under Uncertainty. *The Journal of the Operational Research Society* 55(3):219-232.

Koole, G., and A., Mandelbaum. 2002. Queueing Models of Call Centers - an Introduction. *Annals of Operations Research* 113:41-59.

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.

Lane, D., C. Monefeldt, and J. Rosenhead. 1998. Emergency-but no Accident-a Systems Dynamics Study of an Accident and Emergency Department. *OR Insight* 11:2-10.

Law, A. M., and W. D. Kelton. 2000. *Simulation Modeling and Analysis*, 3rd edition. New York: McGraw-Hill.

Mandelbaum, A., and W. A. Massey. 1995. Strong Approximations for Time–Dependent Queues. *Mathematics of Operations Research* 20(11):33-64.

Massey, W. A., and W. Whitt. 1997. Peak Congestion in Multi-Server Service Systems with Slowly Varying Arrival Rates. *Queueing Systems* 25:157-172.

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.

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.

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.

Pinedo, M. L. 2007. *Planning and Scheduling in Manufacturing and Services*. New York: Springer.

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.

Ross, S. M. 1995. *Stochastic Processes*, 2nd edition. NJ: John Wiley & Sons.

Rothkopf, M. H., and S. S. Oren. 1979. A Closure Approximation for the Nonstationary M/M/s Queue. *Management Science* 25:522-534.

Stadtler, H., and C. Kilger. 2007. *Supply Chain Management and Advanced Planning: Concepts, Models, Software, and Case Studies*, 4th edition. New York: Springer.
Sze, D. Y. 1984. A Queueing Model for Telephone Operator Staffing, *Operations Research* 32:229-249.
Tamhane, A. C., and D. D. Dunlop. 2000. *Statistics and Data Analysis from Elementary to Intermediate*. Prentice Hall.
Tang, C. S. 2006. Robust Strategies for Mitigating Supply Chain Disruptions. *International Journal of Logistics: Research and Applications* 9(1):33-45.
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.
Yang, F. 2008. Neural Network Metamodeling for Cycle–Time Based Performance Profiles in Manufacturing, submitted to *Naval Research Logistics*.

AUTHOR BIOGRAPHIES

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

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