Approximating Performance and Traffic Flow in Nonstationary Tandem Networks of Markovian Queues

Ira Gerhardt
Department of Mathematics and Computer Science
Manhattan College
Riverdale, NY 10471
ira.gerhardt@manhattan.edu

Barry L. Nelson
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208-3119
nelsonb@northwestern.edu

August 7, 2009

Abstract

In this paper we examine a tandem network of queueing nodes where a nonstationary external Markovian arrival process feeds the initial upstream node. We develop methods for modeling the departure flow from *any* upstream node in its role as the arrival process to the immediate downstream node, and employ various techniques to match the interarrival moments to the downstream node to the moments of the departure count from the upstream node, thereby decoupling the network. We apply these matching techniques at variable time-steps, updating the parameters of the fit at each step and holding them constant until the next update. We test the accuracy of this approach by comparing the moments of the queue size of the downstream node against the corresponding true moments of queue size: first moments of queue size are consistently accurate, while the variance of queue size is sometimes less accurate but still useful.

Keywords: Nonstationary queues; Markovian arrival process; phase-type distribution; tandem network.

1 Introduction

To alleviate the difficulty of identifying steady-state behavior in a network of queueing nodes where both the external arrival and the nodal service processes are time stationary, Whitt [49] developed the Queueing Network Analyzer, or QNA. QNA decomposes the network into approximately equivalent independent nodes, modeling the internode traffic flow as superposed renewal processes. The decomposition technique uses local and asymptotic information from upstream nodes (i.e., any node whose departures feed a given node) to characterize the attributes of the fitted renewal arrival process to each downstream node [48].

In this paper, we propose techniques for modeling traffic flow within tandem queueing networks that have Markovian arrival and service processes that may be nonstationary. By "stationary" or "nonstationary" we mean that the *parameters* of the process do not or do change, respectively, over time. Like QNA, we utilize the idea of network decomposition, treating each node independently, and approximating the arrival process to each downstream node to provide performance analysis (specifically the mean and variance of the congestion at that node). However, unlike QNA, we do not examine steady-state performance measures; instead, we provide time-dependent performance analysis, using both local and longer-term information (identified at a particular set of times) to specify a piecewise-stationary arrival process that accurately approximates the true arrival process to each downstream node. We provide consistently accurate estimates of the time-varying mean number of entities at each queueing node, and from very accurate to rough approximations of the time varying variance of the number at each node; we also characterize the features that lead to inaccurate approximations.

The tandem network we consider is composed of general Markovian component processes (Markov arrival process, or MAP, and phase-type service distributions, or Ph), meaning that we can closely approximate non-Markovian tandem networks within this framework.

We view this work as a first step toward approximating the time-dependent behavior of queueing networks with non-stationary arrival and service processes that have more general structure than tandem, a notoriously difficult problem.

The remainder of this paper is organized as follows: We introduce the notation and modeling tools that we use in analyzing our nonstationary tandem network in Section 2. In Section 3, we present the components of the matching technique, describing methods for capturing departure count moments from the upstream node as well as explaining our algorithm for translating from these departure count moments to the fitted arrival process at the downstream node. Section 4 includes a summary of results from employing the matching technique in various network structures. We conclude with suggestions for future research in Section 5. A number of appendices support the results in these sections.

2 Background

2.1 Beyond QNA

Although decomposition-approximation in modeling queueing networks had been utilized prior to Whitt [49] (see, for example, [34, 43]), QNA is often considered the industry standard in its application of this technique. In QNA, the network consists of a finite number of finite-server, general stationary service nodes (with infinite buffer space). Performance measures at each node are calculated by assuming the nodes are independent, approximating the true arrival process to each node by a single renewal process whose mean and coefficient of variation (of its generating interrenewal distribution) are chosen to yield good nodal performance approximations.

Several variations of QNA have been proposed. Alternative network structures that have been studied include those with phase-type service and finite buffers [9], multiple customer classes [50], networks under heavy traffic [18], point-to-multipoint routing [42], and several

others. Discrete interrenewal distributions [8] and Markov-modulated Poisson Processes [10, 11] have replaced general stationary renewal processes as the tool of approximation. Other approximation techniques account for correlations between traffic streams [19], target higher interrenewal moments [7], or specify the coefficient of variation as a function of the traffic intensity at the node rather than as a single variation parameter [51].

Techniques for extending QNA to queueing networks with nonstationary component processes are less well known. Whitt [52] utilizes decomposition in providing time-dependent analysis for networks with Poisson arrivals and exponential service where the arrival rate varies due to balking, reneging, and retrying. For nonstationary networks with non-Poisson arrivals and non-exponential service, Whitt cites the techniques of Taaffe and co-authors introduced in Section 2.3.

2.2 The Nonstationary Tandem Network

In this paper, we investigate a tandem queueing network where the external arrival process is a nonstationary Markovian Arrival Process (MAP_t) [20], and service times at each of $z \geq 2$ queueing nodes have stationary phase-type (Ph) distributions [30]. We utilize MAPs and Ph distributions for two reasons. First, the Markovian properties of MAP_t and Ph distributions make the resulting queueing models more analytically tractable [21]. Second, Ph distributions are dense on the set of all distributions with support on $[0, \infty)$, implying that we can closely approximate non-Markovian service processes [1].

The interarrival times in the MAP_t describe the time it takes an underlying CTMC to reach $m_c \geq 1$ absorbing phases from a finite number $m_a < \infty$ of transient phases; the chain reaching an absorbing phase triggers an arrival. Let $J(t)$ denote the current phase of the CTMC at time $t \geq 0$. We utilize a representation here for the MAP_t that characterizes the interarrival distribution by transitions within the embedded discrete-time Markov chain (DTMC) along with a vector of transition rates (one for each transient phase) and a matrix

of the initial transient phase probabilities at any time $t \geq 0$. This representation is used by Nelson and Taaffe [28] and recounted here.

We let $\mathbf{A}(t)$ denote the time-dependent, one-step transition probability matrix of the embedded DTMC:

$$\mathbf{A}(t) = \begin{pmatrix} \mathbf{A}_1(t) & \mathbf{A}_2(t) \\ \boldsymbol{\alpha}(t) & \mathbf{0} \end{pmatrix},$$

at time $t \geq 0$. The $m_a \times m_a$ matrix $\mathbf{A}_1(t)$ represents the time-dependent one-step transition probabilities between the m_a transient phases, while the $m_a \times m_c$ matrix $\mathbf{A}_2(t)$ represents the time-dependent one-step transition probabilities from the m_a transient phases to the m_c absorbing phases. "Absorbing phase" is really a misnomer in this representation, because rather than being absorbed the process is reinitialized for the next interarrival time by an $m_c \times m_a$ initial probability matrix $\boldsymbol{\alpha}(t)$.

We define the $m_a \times 1$ vector $\boldsymbol{\upsilon}(t)$, whose j^{th} argument is $\upsilon_j(t)$, the time-dependent, integrable non-negative transition rate function corresponding to phase j, for $j = 1, 2, \ldots, m_a$. We use the convention that $\upsilon_{m_a+k}(t) = \infty$, for $k = 1, 2, \ldots, m_c$ and all $t \geq 0$, corresponding to an instantaneous sojourn time in any absorbing phase. Thus, the Nelson and Taaffe representation for a MAP_t is the pair $(\mathbf{A}(t), \boldsymbol{\upsilon}(t))$.

The Nelson and Taaffe representation of the MAP_t utilizes the convention that $m_c = m_a$ and that the matrix $\mathbf{A}_2(t)$ is diagonal (i.e., each absorbing phase may only be reached in one step from a single, unique transient phase, for any $t \geq 0$). We let $a_{ij}(t)$ and $\alpha_{ij}(t)$ denote the $(i,j)^{th}$ components of $\mathbf{A}_1(t)$ and $\boldsymbol{\alpha}(t)$, respectively, for $i, j = 1, 2, \ldots, m_a$. We let $d_j(t)$ denote the $(j,j)^{th}$ component of diagonal matrix $\mathbf{A}_2(t)$, for $j = 1, 2, \ldots, m_a$; by definition, $d_j(t) = 1 - \sum_{h=1}^{m_a} a_{jh}(t)$, for $j = 1, 2, \ldots, m_a$ and for all $t \geq 0$. In practice it is often the case that $\mathbf{A}(t) = \mathbf{A}$ (i.e., $\mathbf{A}$ is not a function of t), so that the nonstationarity is captured in the transition rate vector $\boldsymbol{\upsilon}(t)$.

For a stationary MAP in steady state with representation $(\mathbf{A}, \boldsymbol{\upsilon})$, the i^{th} noncentral

moment m_i of its interarrival distribution (and therefore it squared coefficient of variation $scv \equiv m_2/m_1^2 - 1$) and lag-k autocorrelation ρ_k are known functions of $\mathbf{A}$ and $\boldsymbol{\upsilon}$ [5].

The service process at each node is a stationary Ph renewal process, indicating that its MAP_t representation is not a function of t and that every row in its initial probability matrix is equal. Since we are now describing the service process at each node, technically we should refer to each one as a stationary Markovian *Service* Process (MSP).

Let $m_b^{(n)}$ denote the number of transient phases in the stationary Ph service process at node n, characterized by one-step transition probability matrix

$$\mathbf{B}^{(n)} = \begin{pmatrix} \mathbf{B}_1^{(n)} & \mathbf{B}_2^{(n)} \\ \boldsymbol{\beta}^{(n)} & \mathbf{0} \end{pmatrix},$$

and constant $m_b^{(n)} \times 1$ service rate vector $\boldsymbol{\mu}^{(n)}$, for $n = 1, 2, \ldots, z$. We let $b_{ij}^{(n)}$, $\beta_{ij}^{(n)}$, $f_j^{(n)}$, and $\mu_j^{(n)}$ denote the elements of the Ph service process components $\mathbf{B}_1^{(n)}$, $\boldsymbol{\beta}^{(n)}$, $\mathbf{B}_2^{(n)}$, and $\boldsymbol{\mu}^{(n)}$, respectively, for $i, j = 1, 2, \ldots, m_b^{(n)}$.

We let $s^{(n)}$ denote the total number of identical servers at node n, and define random variable $N_i^{(n)}(t)$ as the number of servers performing the i^{th} phase of service at node n at time $t \geq 0$, for $i = 1, 2, \ldots, m_b^{(n)}$ and $n = 1, 2, \ldots, z$. Let $N^{(n)}(t)$ and $Q^{(n)}(t)$ be the total number of entities at node n and the number of entities waiting for service at node n at time $t \geq 0$, respectively; therefore, $N^{(n)}(t) = \sum_{i=1}^{m_b^{(n)}} N_i^{(n)}(t) + Q^{(n)}(t)$, for $n = 1, 2, \ldots, z$ and all $t \geq 0$. Finally, we define random variable $D_t^{(n)}(t+\tau)$ to be the number of departures from node n on the interval $[t, t+\tau)$, for $t \geq 0$, $\tau > 0$, and $n = 1, 2, \ldots, z-1$. Approximating the moments of $D_t^{(n)}(t+\tau)$ is key to our representation of traffic flow.

2.3 The MDE/DDE-Approach

We utilize two tools in characterizing tandem queues: nodal models and flow models. Nodal models refer to a technique, first introduced by Clarke [3], to approximate the moments of the number of entities at each node by a finite system of linear moment-differential equations

(MDEs) that replace the (potentially infinite) system of state-probability differential equations (known as Kolmogorov Forward Equations, or KFEs) that describe the evolution of the state probabilities. Nodal approximations have previously been derived for several queueing models that include nonstationary MAP or Ph renewal process components; these include the $M_t/M_t/s$ [2], $Ph_t/M_t/s/c$ [31], $Ph_t/Ph_t/1/c$ [32], $Ph_t/Ph_t/s/k$ [36, 38], $MAP_t/MSP_t/s/k$ [37], and $Ph_t/Ph_t/\infty$ [27] single-node models as well as the $[Ph_t/Ph_t/\infty]^K$ [26] and the $[MAP_t/Ph_t/\infty]^K$ [28] networks. Models of nonstationary queues with interrupted Poisson arrivals and unreliable/repairable servers have also been examined [33]. We use these results as the building blocks of our network decomposition-approximation technique.

For flow models, we employ techniques from Nasr and Taaffe [23, 24, 25] for calculating departure-count moments; these techniques are analogous to those for the nodal models, and typically yield a finite system of departure-moment differential equations (DDEs) that describe the behavior of the moments of the number of departures from a queueing node over a finite time interval.

Our work is motivated by networks of infinite-server nodes (i.e., $s^{(n)} = \infty$, for all $n = 1, 2, \ldots, z$), where the system of MDEs and DDEs is closed; thus, solutions to infinite-server MDEs and DDEs are exact, and we can use the exact results to evaluate our approximations. We provide the MDEs and DDEs necessary to perform the matching algorithm in tandem infinite-server networks in Appendix A. We are also able to utilize a result from Nelson and Taaffe [26, 28] on the analytical equivalence of a $[Ph_t/Ph_t/\infty]^K$ network to a single $Ph_t/Ph_t/\infty$ node, calculating the true time-dependent mean and variance of the size of each node in the tandem infinite-server network using (A.2) and (A.4).

The MDE/DDE approach for finite-server nodes (i.e., $s^{(n)} < \infty$, for all $n = 1, 2, \ldots, z$) has two differences from the infinite-server approach that are worth mentioning. First, the system of MDEs and DDEs for a finite-server node is not closed, and closure techniques must be employed to provide approximate values for unknown terms; for background on

closure techniques, see [31, 36] and references therein. Typically, this requires approximating state probabilities in the MDEs and DDEs using a surrogate probability distribution; techniques are employed to specify the parameters of the surrogate mass function to match current moments of the node size and departure count. Surrogate distributions that have been employed in modeling nonstationary queueing nodes include negative binomial [35] and Pólya-Eggenberger (PE) [2], as well as limiting forms of the PE distribution [6, 17]; for background on these distributions, see [16]. In Appendix B we derive the MDEs and DDEs for a finite-server Markovian queueing node with infinite-buffer space and propose a technique for utilizing the finite-support PE distribution as a surrogate in infinite-buffer models.

The second difference in the MDE/DDE approach in finite-server networks from its application in infinite-server networks is that no analogous system of network MDEs exists; therefore, to validate our approximation we must compare the mean and variance of the fitted downstream node size (from the MDE/DDE approach) to simulation results for the finite-server network model.

3 The Matching Technique

3.1 A High-Level Summary of our Approach

Our goal in this paper is to approximate the $MAP_t/Ph^{(1)}/s^{(1)} \rightarrow \cdot/Ph^{(2)}/s^{(2)} \rightarrow \ldots \rightarrow \cdot/Ph^{(z)}/s^{(z)}$ network, for $z \geq 2$. We focus on the case of $z = 2$, since once we have approximated node 2 we can repeat the technique iteratively. Node 2 is approximated as a $MECO_t(\ell)/Ph^{(2)}/s^{(2)}$ node where the fitted arrival process has piecewise-constant parameters over a varying step-size τ; at each update, the parameters are obtained from the respective interval-departure count moments from node 1 which are numerically calculated using the MDE/DDE approach.

The MECO renewal process is introduced in [15]; briefly, it consists of a mixture of two

Erlang distributions of the same order $\ell \in \mathbb{Z}^+$. Parameters of the stationary MECO(ℓ) can be specified to match any triple of first three interrenewal moments (for a sufficiently large ℓ), and therefore, the MECO$_t(\ell)$ can be both more and less variable than exponential for a given ℓ—an important requirement since the upstream departure process may be more variable than Poisson on some intervals and less variable on others. Utilizing the MECO$_t(\ell)$ as the fitted downstream arrival process allows us to have a single Ph process where only the parameters are updated on each interval (while maintaining the number of phases and structure of the fitted arrival process); this also allows for a constant number of MDEs and DDEs at the downstream node along the duration of the time horizon approximated.

We enumerate the steps performed in our matching algorithm here, and elaborate on them in the following sections. Since we apply the algorithm at each node sequentially, for simplicity we drop the '$^{(n)}$' and let $D_t(t+\tau)$ denote the number of departures from current node 1 on $[t, t+\tau)$, for $t \geq 0$, $\tau > 0$.

Algorithm 3.1. *The Matching Technique for the Two-Node Nonstationary Tandem Network*

1. Initialize MDEs and DDEs at time $t = 0$ for the upstream $MAP_t/Ph^{(1)}/s^{(1)}$ node and downstream $MECO_t(\ell)/Ph^{(2)}/s^{(2)}$ node.
2. At time t, determine the appropriate step size τ (see Section 3.3).
3. Evaluate MDEs and DDEs at upstream node to time $t + 2\tau$, calculating values for $\mathbb{E}\{D_t(t+\tau)\}$ and $\text{Corr}\{D_t(t+\tau), D_{t+\tau}(t+2\tau)\}$.
4. Determine parameters for the fitted downstream arrival process MECO$_t(\ell)$ on $[t, t+\tau)$.
   (a) Back out fitted mean interarrival time m_1 from $\mathbb{E}\{D_t(t+\tau)\}$ and fitted squared coefficient of variation scv from $\text{Corr}\{D_t(t+\tau), D_{t+\tau}(t+2\tau)\}$ (see Section 3.2).
   (b) Specify MECO$_t(\ell)$ parameters to match m_1 and scv.
5. Evaluate MDEs for $\widehat{MECO}/Ph^{(2)}/s^{(2)}$ to time $t + \tau$.
6. Set $t = t + \tau$. Go to Step 2.

3.2 Translating from Upstream Departure Count Moments to the Fitted Downstream Arrival Process

In this section, we propose a technique to translate from the upstream departure count moments $\mathbb{E}\{D_t(t+\tau)\}$ and $\text{Corr}\{D_t(t+\tau), D_{t+\tau}(t+2\tau)\}$—obtained from numerical integration of the DDEs—to the stationary parameters of the fitted $\text{MECO}_t(\ell)$ that approximates the true arrival process to the downstream node on $[t, t+\tau)$. Let $A_t(t+\tau)$ denote the number of arrivals for the fitted downstream $\text{MECO}_t(\ell)$ on $[t, t+\tau)$. Ideally we would back out the MECO parameters directly from the departure count moments by setting

$$\mathbb{E}\{A_t(t+\tau)\} = \mathbb{E}\{D_t(t+\tau)\} \tag{1}$$

and

$$\text{Corr}\{A_t(t+\tau), A_{t+\tau}(t+2\tau)\} = \text{Corr}\{D_t(t+\tau), D_{t+\tau}(t+2\tau)\}; \tag{2}$$

however, we typically cannot derive closed-form expressions for the count moments of the $\text{MECO}_t(\ell)$ renewal process as functions of the MECO parameters, for $\ell \geq 2$.

Instead, we derive the first two interrenewal moments, m_1 and scv, for the fitted $\text{MECO}_t(\ell)$ implied by the two departure count moments using a surrogate process. We first obtain m_1 using a general result for stationary renewal processes; see Equation (3). To determine scv, we utilize two stationary Ph renewal processes (the h_2b and the MECon, described below) to act as surrogates; unlike the MECO, we can identify parameter values for the h_2b and MECon parameters (as appropriate) to satisfy (1) and (2). We take as the third interrenewal moment, m_3, the third interrenewal moment of the fitted surrogate process, and choose the $\text{MECO}_t(\ell)$ parameters to match (m_1, scv, m_3). The technique described here has some similarities to one first proposed by Whitt [47], where m_1 is obtained from information on a short interval, while scv is determined from information on a longer interval.

Notice that we can calculate m_1 directly from (1), using a result from Cox and Smith [4] connecting the mean renewal count of a stationary renewal process and its mean interrenewal

time m_1. Specifically, they show $\mathbb{E}\{A_t(t+\tau)\} = \tau/m_1$, for $t \geq 0$, $\tau > 0$; therefore,

$$m_1 = \frac{\tau}{\mathbb{E}\{D_t(t+\tau)\}}. \tag{3}$$

While the value of the mean renewal count only depends on m_1, values for higher moments of $A_t(t+\tau)$ depend on the specific interrenewal distribution [44]. A useful consequence of the MDE approach is that we can easily calculate exact values for the renewal count moments for a given stationary Ph process by plugging its MAP_t representation into (A.2), setting all service rates $\mu_i = 0$ (for $i = 1, 2, \ldots, m_b$), and initializing the MDEs appropriately. A further benefit is that when the Ph order $m_a \leq 2$, we can use the MDEs to derive closed-form expressions for the renewal count moments.

Thus, our goal is to specify the parameters of the surrogate Ph renewal process to satisfy (2), given m_1 in (3); this requires selecting surrogate Ph processes that are specified by only two parameters. The particular Ph renewal processes we recommend are a mixture of two exponentials with balanced means, or h_2b, when $\text{Corr}\{D_t(t+\tau), D_{t+\tau}(t+2\tau)\} \geq 0$, and a mixture of two Erlangs of consecutive order and common rate, or MECon, when $\text{Corr}\{D_t(t+\tau), D_{t+\tau}(t+2\tau)\} < 0$. Descriptions of the h_2b and MECon are provided in Appendix C. The two surrogate processes are invoked according to the sign of the departure count correlation since the renewal count correlation is always positive for the h_2b and negative for the MECon. Benefits to using these particular Ph choices as surrogates are three-fold. First, we can uniquely identify the value for their respective interrenewal *scv* such that $\text{Corr}\{A_t(t+\tau), A_{t+\tau}(t+2\tau)\}$ satisfies (2), given m_1 in (3) and $\tau > 0$. Second, formulas for translating from m_1 and *scv* to the two respective parameters of these Ph processes are well known; we cite them in Appendix C. Third, the h_2b and MECon provide coverage over all $scv > 0$ as well as over a wide range of count correlation values.

Since $m_a = 2$ for the stationary h_2b renewal process, we can derive expressions for the

count moments as a function of τ, m_1, and the h_2b mixing probability α, namely

$$\begin{aligned}
\mathrm{Var}\{A_t(t+\tau)\} &= \frac{1}{8m_1\alpha^2(1-\alpha)^2}\left[4\alpha(1-\alpha)\left(2\alpha^2-\alpha+1\right)\tau\right. \\
&\quad \left. - m_1(2\alpha-1)^2\left(1-e^{-4\alpha(1-\alpha)\tau/m_1}\right)\right], \\
\mathrm{Cov}\{A_t(t+\tau), A_{t+\tau}(t+2\tau)\} &= \frac{(2\alpha-1)^2}{16\alpha^2(1-\alpha)^2}\left(1-e^{-4\alpha(1-\alpha)\tau/m_1}\right)^2,
\end{aligned}$$

which yields

$$\begin{aligned}
&\mathrm{Corr}\{A_t(t+\tau), A_{t+\tau}(t+2\tau)\} = \\
&\qquad \frac{m_1(2\alpha-1)^2\left(1-e^{-4\alpha(1-\alpha)\tau/m_1}\right)^2}{2\left[4\alpha(1-\alpha)\left(2\alpha^2-\alpha+1\right)\tau - m_1(2\alpha-1)^2\left(1-e^{-4\alpha(1-\alpha)\tau/m_1}\right)\right]},
\end{aligned} \tag{4}$$

for $t \geq 0$, $\tau > 0$. Thus, we find $\alpha \in [0,1]$ that satisfies (4), given $\tau > 0$, m_1 in (3), and $\mathrm{Corr}\{A_t(t+\tau), A_{t+\tau}(t+2\tau)\} = \mathrm{Corr}\{D_t(t+\tau), D_{t+\tau}(t+2\tau)\} \geq 0$; the implied *scv* is

$$scv = \frac{1+(2\alpha-1)^2}{1-(2\alpha-1)^2}.$$

In a similar fashion, we can identify the *scv* implied by $\mathrm{Corr}\{D_t(t+\tau), D_{t+\tau}(t+2\tau)\} < 0$; however, unlike the h_2b, we cannot derive closed-form expressions for the count moments of a MECon, and therefore, have no result analogous to Equation (4) that yields the appropriate MECon parameters. Instead, notice that we can use the MDEs to numerically calculate $\mathrm{Corr}\{A_t(t+\tau), A_{t+\tau}(t+2\tau)\}$ for a stationary MECon given $scv \in (0,1)$, $\tau > 0$, and $m_1 > 0$. Therefore, just as we numerically calculate a solution to (4) for the implied $scv \geq 1$ of the h_2b, we can numerically calculate the $scv < 1$ for the MECon that yields $\mathrm{Corr}\{A_t(t+\tau), A_{t+\tau}(t+2\tau)\} = \mathrm{Corr}\{D_t(t+\tau), D_{t+\tau}(t+2\tau)\} < 0$, given $\tau > 0$ and m_1 in (3). With this technique defined, we have now obtained the two fitted interrenewal moments m_1 and *scv* implied by the two departure count moments.

Recall from Section 3.1 that specifying the fitted $\mathrm{MECO}_t(\ell)$ parameters requires knowing the first *three* interrenewal moments, while the translation technique described here yields only the first *two* interrenewal moments m_1 and *scv*. To bridge this gap, we take the implied

third interrenewal moment m_3 from the fitted surrogate. The three $\text{MECO}_t(\ell)$ parameters (i.e., the means for each of the component exponentials—λ_1^{-1} and λ_2^{-1}, respectively—and the mixing probability p) are matched to the triple (m_1, scv, m_3); for the algorithm that accomplishes this, see [15]. At present we set the MECO order ℓ equal to the minimum feasible order across all time intervals, using an algorithm in [15] that identifies the minimum feasible order for a MECO given its first three interrenewal moments.

3.3 Determining the Interval Length τ

Our initial efforts modeling traffic flow in nonstationary tandem queueing networks indicated that identifying an appropriate value for the interval length τ is key to providing an accurate approximation of the downstream arrival process. If τ is too small, the matching technique degenerates, yielding the mean and variance for the interval-departure count being nearly equal (with departure count correlation near zero). Accordingly, the fitted arrival process to these departure count moments is approximately Poisson, regardless of the network components, which typically yields a poor fit for the approximated moments at the downstream node. On the other hand, setting τ too large leads to count moments that do not accurately reflect the local behavior of the upstream departure process—behavior that is particularly important in nonstationary networks where model properties may be changing rapidly.

In this section we propose a technique for dynamically identifying an appropriate value for $\tau > 0$ at time $t \geq 0$. First, we define a metamodel that provides an initial guess for τ given the MAP_t representations for the external arrival and nodal service processes at current time t; we then propose an algorithm that refines the initial prediction to account for nonstationarity near t.

The initial–prediction metamodel is the product of an experiment in which we investigated 2000 stationary $Ph/Ph^{(1)}/\infty \to \cdot/Ph^{(2)}/\infty$ networks; for each network, we identified the smallest constant interval length $\tau > 0$ such that the maximum relative error of the

Table 1: Parameter ranges for the 2000 design points in the initial–prediction metamodel.

Symbol	Description	Range
$m_a^{(1)}$	external mean interarrival time	$[1/80, 1/5]$
scv_a	external squared coeff. of variation of arrival	$[0.25, 3]$
$m_s^{(1)}/m_a^{(1)}$	offered load at node 1	$[1/2, 100]$
$scv_s^{(1)}$	squared coeff. of variation of service at node 1	$[0.5, 2]$
$m_s^{(2)}/m_a^{(1)}$	offered load at node 2	$[1/2, 100]$
$scv_s^{(2)}$	squared coeff. of variation of service at node 2	$[0.5, 2]$

fitted variance at the downstream node versus its exact counterpart, over the duration of the time horizon approximated, was less than 1%. Each network was uniquely identified by six network parameters: the mean interarrival time $m_a^{(1)}$ and squared coefficient of arrival variation scv_a, as well as the offered loads $m_s^{(n)}/m_a^{(1)}$ and squared coefficients of service variation $scv_s^{(n)}$ at nodes $n = 1$ and $n = 2$.

Thus, the metamodel is designed to predict τ given values for these six parameters, utilizing the assumption that the arrival process and service processes at both nodes are stationary and in steady state at the current time t. Table 1 includes descriptions and ranges for the six network parameters included in the experiment; we specified these ranges to encompass the wide range of process variability and offered loads that we would expect to encounter in modeling real-world systems. We utilized a stationary h_2b for the external Ph arrival process as well as for the $Ph^{(n)}$ service processes (for $n = 1, 2$) when the respective process' squared coefficient of variation was greater than 1, and utilized a MECon when the respective scv was less than 1 (see Appendix C). The 2000 design points were selected using Latin hypercube sampling [13], and the range of τ across these points was $[0.035, 0.626]$. The metamodel was fit using kriging; for background, see [39, 40].

We let $\widehat{\tau}(t)$ denote the prediction from the metamodel given the MAP_t representations

of the external arrival process and both service processes evaluated at time $t \geq 0$; that is,

$$\begin{aligned} \widehat{\tau}(t) &\equiv f\left(\mathbf{A}(t),\, \boldsymbol{v}(t),\, \mathbf{B}^{(1)}(t),\, \boldsymbol{\mu}^{(1)}(t),\, \mathbf{B}^{(2)}(t),\, \boldsymbol{\mu}^{(2)}(t)\right) \\ &= \widehat{f}\left(m_a^{(1)}(t),\, scv_a(t),\, m_s^{(1)}(t)/m_a^{(1)}(t),\, scv_s^{(1)}(t),\, m_s^{(2)}(t)/m_a^{(1)}(t),\, scv_s^{(2)}(t)\right), \quad (5) \end{aligned}$$

for $t \geq 0$, where $\widehat{f}(\cdot)$ is the functional form of the metamodel determined from kriging. Each of the six marginal moments in (5) is calculated from the MAP_t representation of its respective arrival or service process at time t; in calculating these moments, we claim that all three component processes are stationary (and in steady state) at time t. Notice that the arguments of $\widehat{f}(\cdot)$ in (5) are presented as functions of t; for our work in this paper, we could drop the '(t)' from all service-related terms, since the tandem networks we examine here include stationary service processes at all nodes.

With the model specified for predicting τ in a stationary network, we are ready to present Algorithm 3.2 for refining this prediction based on nonstationarity near t. In its k^{th} iteration, the algorithm identifies $\bar{\tau}_k$, the average value of $\widehat{\tau}(\cdot)$ over the interval $[t, t + \bar{\tau}_{k-1}]$, for $k = 2, 3, \ldots$; it terminates when the average from consecutive iterations differ by less than a prespecified tolerance. By iteratively identifying the average prediction value—rather than selecting for τ a single predicted value—we aim to account for nonstationarity near t regardless of whether the model parameters are nearly stationary or are changing very rapidly. Notice that $\bar{\tau}_1 \equiv \widehat{\tau}(t)$, the metamodel prediction for the stationary network at current time t.

Algorithm 3.2. *Determining the interval length $\tau > 0$, at time $t \geq 0$.*

1. $\bar{\tau}_0 = 10^{-8}$, $\bar{\tau}_1 = \widehat{\tau}(t)$, $tol = 10^{-3}$, and $k = 1$.
2. While $|\bar{\tau}_k / \bar{\tau}_{k-1} - 1| > tol$
   (a) $\bar{\tau}_{k+1} = \bar{\tau}_k^{-1} \int_t^{t+\bar{\tau}_k} \widehat{\tau}(u)\, du$.
   (b) $k = k + 1$.
3. $\tau = \bar{\tau}_k$.

4 Evaluating the Matching Technique

4.1 Two-node Networks with Ph_t Arrivals

We first evaluated our matching technique in approximating a collection of $Ph_t/Ph^{(1)}/s^{(1)} \rightarrow \cdot/Ph^{(2)}/s^{(2)}$ networks that are initially empty-and-idle. The nonstationary Ph arrival process was characterized by constant squared coefficient of variation $scv_a > 0$ and nonstationary arrival rate $\lambda(t) = \lambda_a [1 + b_a \sin(c_a \pi t)]$, with $\lambda_a,\ c_a > 0$, and $b_a \in [0, 1)$, for $t \geq 0$. Node n had $s^{(n)} < \infty$ servers performing phase-distributed service with mean time $m_s^{(n)} > 0$ and squared coefficient of variation $scv_s^{(n)} > 0$, for $n = 1, 2$.

Table 2 includes descriptions and ranges for the nine parameters in our analysis; specific parameter values for 200 networks were selected using a Latin hypercube design. The range of time t across which each network was approximated was $[0, 10]$. We utilized a MECO_t for the external Ph_t arrival process as well as stationary MECOs for the two $Ph^{(n)}$ service time distributions, for $n = 1, 2$. Similar to the translation technique in Section 3.2, we specified the MECO parameters to match the first two moments of the respective interarrival or service time distribution (using the h_2b or MECon renewal processes to provide the third interrenewal moment). The two target values for the arrival MECO_t process were the current arrival rate $\lambda(t)$ (at time $t \geq 0$) and scv_a, while the two target service-time moments were $m_s^{(n)}$ and $scv_s^{(n)}$, for $n = 1, 2$. The server quantities were backed out from the respective average offered loads and utilizations, such that $s^{(n)} = \lceil \text{AOL}(n)/\text{AU}(n) \rceil$, for $n = 1, 2$, where $\lceil x \rceil$ is the smallest integer greater than or equal to x, for $x \in \Re^+$. Notice that both measures of nonstationarity (i.e., the amplitude and period fractions), as well as the average offered loads and server utilizations at both nodes, were relative to the base arrival rate λ_a; therefore, without loss of generality, we set $\lambda_a = 20$. We simulated 1000 replications of each network to estimate the moments of the true node size at nodes 1 and 2; standard errors of the estimated nodal moments were less than 2% of the estimated values.

Table 2: Ranges of parameters in evaluating the matching technique in a two-node finite-server network.

Parameter Value	Description	Range
b_a	fraction of base arrival rate in amplitude of $\lambda(t)$	$[0\%, 70\%]$
$2\lambda_a/c_a$	ratio of period-to-mean interarrival time (PTM)	$[1, 100]$
scv_a	interarrival scv	$[1/4, 3]$
$\lambda_a m_s^{(1)}$	avg. offered load at node 1 (AOL(1))	$[5, 60]$
$scv_s^{(1)}$	service time scv, node 1	$[1/2, 2]$
$\lambda_a m_s^{(1)}/s^{(1)}$	avg. server utilization, node 1 (AU(1))	$[35\%, 75\%]$
$\lambda_a m_s^{(2)}$	avg. offered load at node 2 (AOL(2))	$[5, 60]$
$scv_s^{(2)}$	service time scv, node 2	$[1/2, 2]$
$\lambda_a m_s^{(2)}/s^{(2)}$	avg. server utilization, node 2 (AU(2))	$[35\%, 75\%]$

It is worth mentioning that the MDE/DDE approach failed to return feasible results for the fitted moments (e.g., yielding negative variance for the fitted node size) in 14 of the 200 networks. Typically, these network models had values for at least one maximum nodal server utilization (defined as $\text{MU}(n) \equiv (1 + b_a) \cdot \text{AU}(n)$, for $n = 1, 2$) near (or above) 100% and $\text{AU}(n) > 65\%$. Unlike in stationary queueing models, individual nodal server utilization within nonstationary networks may be larger than 100% at times without the model becoming unstable; however, our results indicated that the MDE/DDE approach may break down when this utilization is near or above 100% for a large portion of the time horizon.

To validate our matching technique, we calculated the relative error of the approximate moments versus the corresponding true moments estimated via simulation of $N^{(n)}(t)$, for $n = 1, 2$. We let $\text{EARE}(n)$ and $\text{VARE}(n)$ denote the average relative error of the mean and variance of the node size, respectively, over the entire range of $t \in [0, 10]$, for $n = 1, 2$. We also calculated $\text{EMRE}(n)$ and $\text{VMRE}(n)$, the maximum relative error of the mean and variance size at node $n = 1, 2$, respectively; notice that we identified the maximum relative error measures only after the true mean node size had become non-negligible (i.e., we obtained

EMRE(n) and VMRE(n) from the interval $[t', 10]$, where $t' = \inf\{t \geq 0 : \mathbb{E}\{N^{(n)}(t)\} \geq 0.05 \cdot \text{AOL}(n)\}$, for $n = 1, 2$).

The main barometer for evaluating success of the matching technique was the relative error of the fitted moments at node 2. Values for average measures EARE(2) and VARE(2) across the 186 networks in this analysis fell within $[0.3\%, 15.3\%]$ and $[3.1\%, 18.2\%]$, respectively. We claim the matching technique typically provides a good fit for the mean of node 2, as values of EARE(2) larger than 6% occurred in only eight of the 186 networks, while 97 networks had EARE(2) $\leq 1\%$. The quality of the fit in the node 2 variance ranged from very good to poor. Typically, networks with $b_a > 50\%$ and PTM < 10 saw the largest relative errors in the fitted downstream mean node size, while networks with MU(2) $> 85\%$ and $|\log(scv_a)| > \log(2)$ saw the largest relative errors in the downstream node-size variance; we recommend caution when utilizing our technique in approximating tandem networks with parameters in these ranges. We provide some insight into these results now.

We performed a full quadratic ANOVA on VARE(2), attempting to qualify the nine main parameter effects and 36 pairwise interaction effects in terms of their significance at the 5% confidence level; for background on ANOVA, see [29]. To model the extent of the arrival process' deviation from Poisson (as well as the respective service time distributions' deviations from exponential), we included $\log(scv_a)$, $\log(scv_s^{(1)})$, and $\log(scv_s^{(2)})$ as factors in the ANOVA rather than scv_a, $scv_s^{(1)}$, and $scv_s^{(2)}$. The effects that most significantly explained VARE(2) included the node 2 average server utilization AU(2), the arrival amplitude parameter b_a, and the logarithm of the squared coefficient of arrival variation $\log(scv_a)$. The significance of these effects are intuitively understandable; a large value for AU(2) combined with a large value for b_a yields a large value for node 2 maximum server utilization MU(2); as described above, the quality of the MDE/DDE approach at a node is typically poor when its maximum server utilization is very large. In other words, the nodal approximation itself may break down even if the traffic flow is adequately represented. We also have observed

that the sign of $\log(scv_a)$ appears to be a direct indicator of whether the true distribution of each node size is more or less variable than Poisson; thus, extreme values for $\log(scv_a)$ should lead to nodal size distributions that deviate significantly from Poisson, and we expect it to be more difficult to fit such distributions than those closer to Poisson. Specifically, we found that the largest values of VARE(2) occurred when AU(2) $> 60\%$, $b_a > 40\%$ and $|\log(scv_a)| > \log(2)$. A similar analysis revealed that the largest values of EARE(2) occurred in networks where PTM < 12 and at least one of MU(1) or MU(2) was larger than 90%; however, as mentioned earlier, EARE(2) was typically very small regardless of the network parameters.

Similarly, we find values for maximum relative error measures EMRE(2) and VMRE(2) ranged across $[1.0\%, 16.9\%]$ and $[9.8\%, 35.9\%]$, respectively; values of VMRE(2) $> 18\%$ typically corresponded to networks with VARE(2) $> 10\%$. Results from the ANOVA on VMRE(2) confirmed that similar factors affect both node-2-variance relative-error measures, as AU(2) and $\log(scv_a)$ were found to be significant (as both main and interaction effects) in also explaining VMRE(2). Main effect factors that explain EMRE(2) at the 5% level included b_a, PTM, and AU(2), while the interaction between b_a and PTM was significant as well. Specifically, the largest values of EMRE(2) occurred in networks where both $b_a > 50\%$ and PTM < 10, as well as in networks where MU(2) $> 80\%$.

Figures 1–3 include plots of the fitted moments (solid lines) and true moments (dashed lines) for three sample networks; these three figures correspond to Networks 1–3 in Table 3 (included in Appendix D), respectively. The plots on the top row in each figure (from left to right) represent the node 1 mean and variance, while those in the bottom row represent the node 2 mean and variance, respectively. Values for the network parameters are provided with each figure. Figure 1 represents a network where the matching technique was successful, as the fit in both downstream nodal moments is very good, while Figure 2 represents a network where the matching technique appears to have failed, since the fit in the downstream variance

is poor even though the MDE/DDE approach has provided an adequate approximation to the upstream moments.

The two variance plots in the right column of Figure 3 merit further discussion. They illustrate a simple result that, nevertheless, plays an important role in the validation of our matching technique: A poor fit from the MDE/DDE approach at node 1 typically yields a poor quality fit from the matching technique at node 2. Several networks in our analysis saw values of EARE(1) and VARE(1) above 9% and 15%, respectively; it is important to identify why the MDE/DDE approach may fail at node 1 since these errors will be carried to the downstream node by the matching technique. Results from a full quadratic ANOVA on EARE(1) indicated that the arrival amplitude and period parameters, b_a and PTM—both as main and interaction effects—were significant at the 5% confidence level; networks with the largest values for EARE(1) typically had $b_a > 50\%$ and/or PTM < 10. A similar analysis indicated that b_a, $\log(scv_a)$, and average utilization AU(1) were significant at the 5% confidence level in explaining VARE(1); the poorest fits for the node 1 variance typically occurred when $b_a > 50\%$, AU(1) $> 65\%$, and $|\log(scv_a)| > \log(2)$. Notice that this is approximately the same range of parameters, with respect to node 1, that yielded the largest values for VARE(2); this indicates that the largest relative errors in the fitted moments at either node typically occurred when the respective nodal maximum server utilization was high and the external arrival process deviated significantly from Poisson.

Acknowledging that accurately approximating time-dependent behavior in nonstationary networks is difficult, we also investigated whether the matching technique described in Section 3, when employed in stationary finite-server networks, provides comparably accurate approximations of the corresponding steady-state moments of node size to those from established stationary network approximation tools such as QNA [49]. To do so, we applied the MDE/DDE approach to time-stationary versions of each of the 200 networks previously analyzed in this section, setting $b_a = 0$ (i.e., $\lambda(t) = \lambda_a$, for all $t \geq 0$); we calculated the cor-

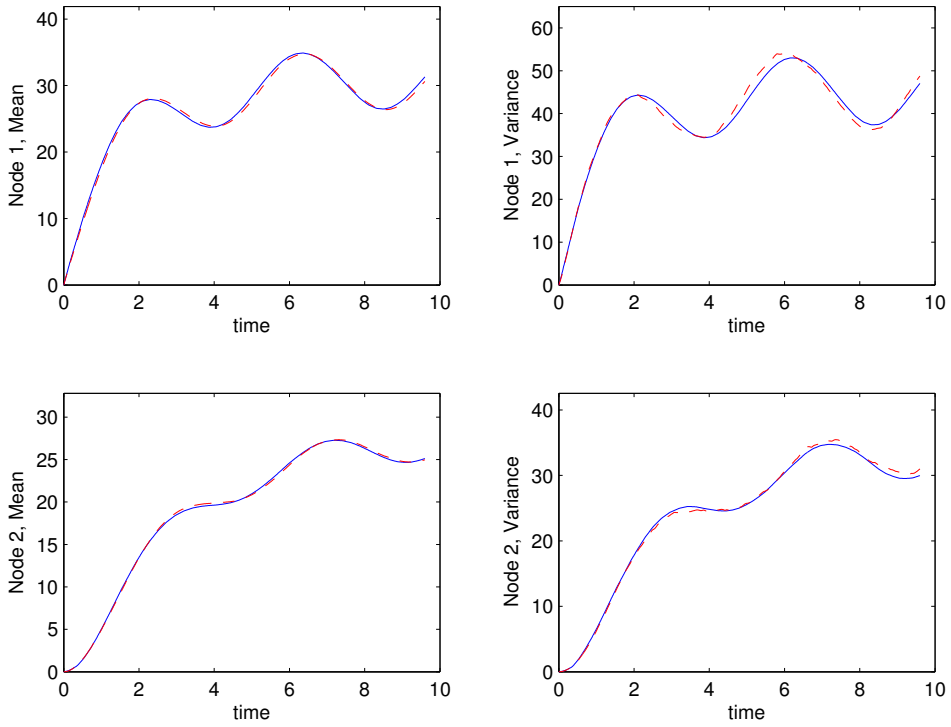

Figure 1: Network 1, with $b_a = 36\%$, $scv_a = 2.04$, $scv_s^{(1)} = 1.29$, AU(1) $= 48\%$, $scv_s^{(2)} = 1.45$, AU(2) $= 69\%$; plots indicate a good fit in both moments at both nodes.

responding node-size moment approximations from QNA using formulas in [49]. We found that the matching technique described here yields very similar relative error values (versus simulation) for each network to those from QNA for the mean size of both nodes, and that these errors are very low; the maximum mean-size relative errors from the matching technique and QNA across the 200 networks were 5.7% and 3.9% at node 1, and 4.5% and 5.3% at node 2, respectively.

However, the accuracy of the nodal variance approximations differ signficantly between our matching technique and QNA. Variance-size relative errors larger than 10% occurred (at either node) in only 15 of the 200 stationary networks using the MDE/DDE approach, with maximum errors of 32.9% at node 1 and 23.1% at node 2. As in the nonstationary analysis

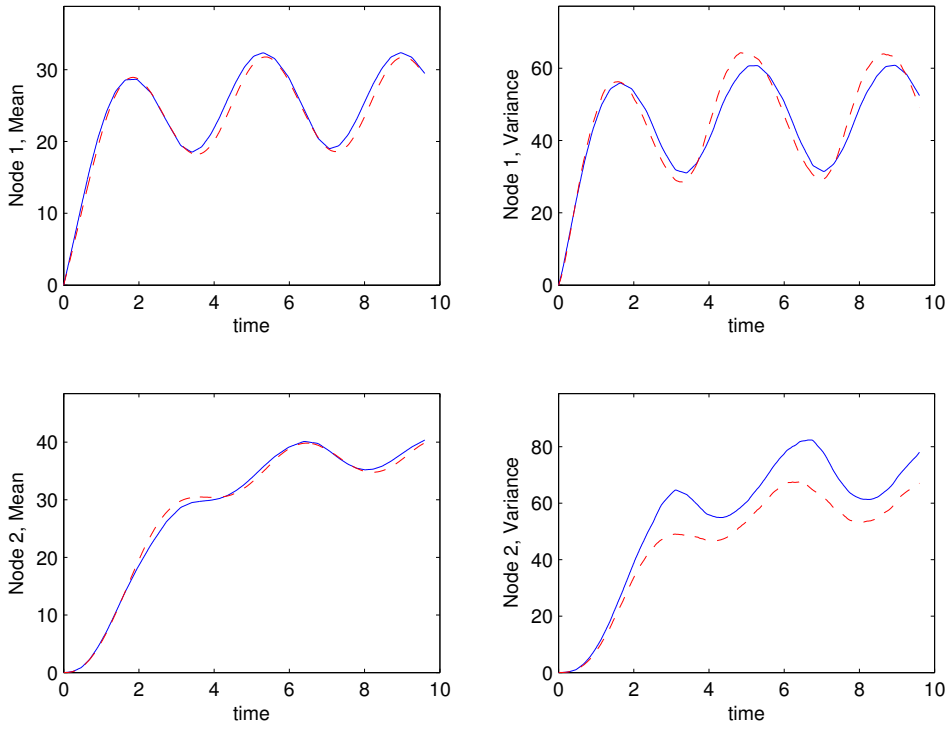

Figure 2: Network 2, with $b_a = 55\%$, $scv_a = 2.42$, $scv_s^{(1)} = 0.60$, AU(1) = 42%, $scv_s^{(2)} = 0.55$, AU(2) = 55%; plots indicate the matching technique significantly overestimates the node 2 variance (bottom right).

above, the largest errors in the steady-state node 2 variance were seen in networks where $|\log(scv_a)| > \log(2)$ and AU(2) > 65%. Maximum relative errors in the nodal variance from QNA were 97.5% and 47.7% at nodes 1 and 2, respectively. As expected, QNA yielded its poorest results when the interarrival and service-time distributions deviated significantly from exponential. That QNA yielded highly accurate approximations of mean node size yet less accurate approximations of nodal variance was not unexpected; the validity of QNA was originally assessed by evaluating the accuracy of approximations solely for the *first* moment of both steady-state node size and wait time in a wide variety of network structures [48].

We also evaluated the quality of the matching technique in two-node infinite-server networks (i.e., $s^{(n)} = \infty$, for $n = 1, 2$), using the same values for the seven non-utilization

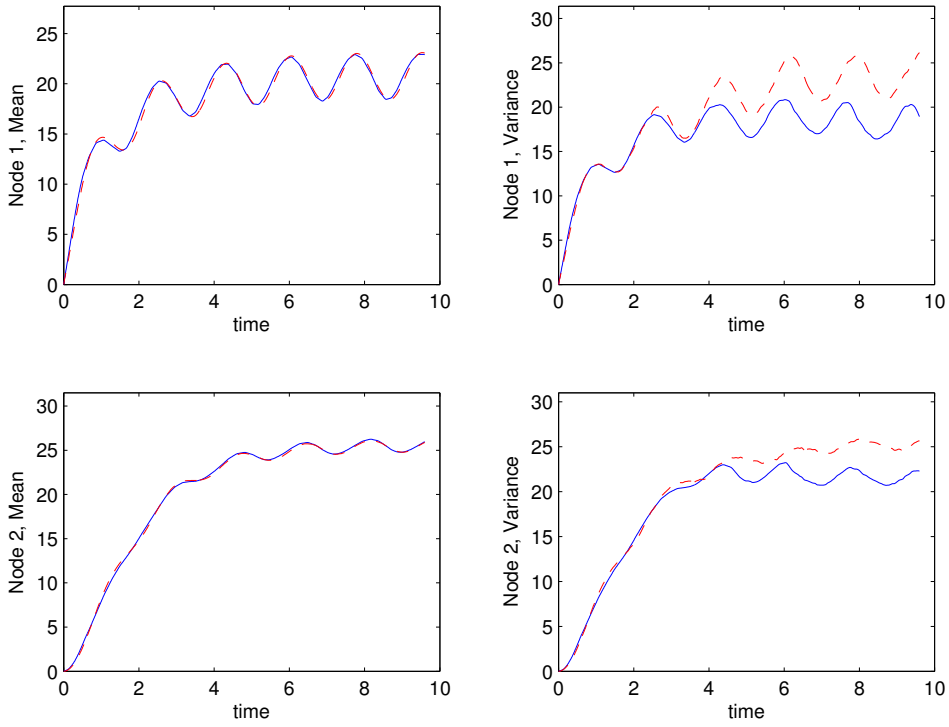

Figure 3: Network 3, with $b_a = 45\%$, $scv_a = 0.9$, $scv_s^{(1)} = 1.52$, AU(1) = 74%, $scv_s^{(2)} = 0.72$, AU(2) = 41%; plots indicate a poor fit from the MDE/DDE approach for node 1 variance (top right) yielding a poor fit at node 2 (bottom right).

parameters provided in Table 2 in the 200 networks we analyzed. Recall that the MDEs and DDEs for the infinite-server model are closed, and that the fitted downstream moments from the matching technique are evaluated versus the exact downstream moments calculated using the network MDEs. Values for EARE(2) and VARE(2) in our infinite-server analysis ranged across $[0.1\%, 5.7\%]$ and $[0.1\%, 13.6\%]$, respectively; average relative errors in individual networks were typically smaller than those in their finite-server counterparts, reinforcing the observation that high server utilization (at either node) decreases the quality of the matching technique in the finite-server arena. The largest values of EARE(2) occurred in those infinite-server networks where $b_a > 55\%$, while VARE(2) $> 10\%$ was typically observed in networks where both $\log(scv_a) > \log(2.4)$ and $\log(scv_s^{(1)}) < \log(0.65)$.

4.2 Alternative Models: Non-Empty Initial Conditions, Correlated Arrivals, and Chains with $z = 3$ Nodes

In this section we evaluate our matching technique with respect to tandem finite-server networks that have three features that are different from those analyzed in the previous section. First, we consider the effect of utilizing non-empty initial conditions at node 1. We also investigate the quality of the matching technique when the external arrival process is a correlated MAP_t. Finally, we examine how the matching technique performs as we fit the furthest downstream arrival process in a tandem chain of $z = 3$ nodes. Unlike in the previous section, we only investigated a few instances of these alternative network models; we leave a comprehensive study of these variations for future research. To better isolate the effect of these variations, we implemented them separately in five sample networks from the previous section where the matching technique was successful (i.e., EARE(2) $< 3\%$ and VARE(2) $< 5\%$); these five sample networks are Networks 1 and 4–7 in Table 3 (provided in Appendix D).

We first investigated non-empty-and-idle initial conditions, setting the initial size of node 1 equal to $\phi\lambda_a m_s^{(1)}$, for $\phi \in [1/2, 2]$. Doing this does not significantly affect either EARE(2) or VARE(2) (i.e., neither evaluation measure in any network is altered more than 1% in absolute value), nor is this result surprising. Non-empty initial conditions yield a higher departure rate and mean departure count (i.e., versus empty-and-idle initial conditions) early in the time horizon; however, this should be accurately captured by the translation technique in Section 3.2. Thus, no adjustments to our matching technique should be necessary to account for non-empty initial conditions.

Next, we employed the matching technique in networks where the external arrival process was a correlated MAP_t; correlation in these five networks was provided by varying the lag-1 autocorrelation $\rho_1 \in [-0.2, 0.4]$. To capture non-zero autocorrelation in the interarrival times, we utilize a nonstationary Markov-MECO [14] for the external MAP_t. The

Markov-MECO is a nonrenewal generalization of the MECO renewal process; formulas for specifying Markov-MECO parameters to target ρ_1 (in addition to the first three moments of the marginal distribution) are provided in [14]. Our matching technique was successful in these five instances; in fact, EARE(2) $< 1\%$ in four of the five networks, which was lower than the EARE(2) in the corresponding networks with $\rho_1 = 0$, as in Section 4.1. There was also no significant change in the values of VARE(2) from introducing correlated arrivals.

However, we do not expect that the current matching technique will be as successful for a specific network model in the presence of correlated arrivals as it is when interarrival times are uncorrelated (i.e., we would expect values for VARE(2) to be higher when $\rho_1 \neq 0$). One reason behind its success here may be the limited range of ρ_1 investigated, and we acknowledge that adjustments to the matching technique may be necessary in the presence of more significant correlation. One potential adjustment is to redefine the metamodel in Section 3.3 to predict the stationary τ based on the value of a long-term arrival variation parameter such as the index of dispersion of intervals (IDI)—equivalent to $scv_a\,(1 + 2\sum_{k=1}^{\infty} \rho_k)$ [45] (where ρ_k is the lag-k autocorrelation for interarrival times, for $k = 1, 2, \ldots$)—rather than scv_a itself; notice that IDI $= scv_a$ when interarrival times are uncorrelated. That said, it is typically difficult to specify a MAP to match an extreme value of ρ_1, particularly when $scv_a < 1$; for evidence of this claim, see [5, 12, 14] and related papers.

In the final variation investigated here, we evaluated the quality of our matching technique when we extend each of the five sample networks to a third node (i.e., $z = 3$). The nodal service parameters we model at node $n = 3$ are the same as those in Table 2 for $n = 1$, 2. We observe an initial lag in the fitted moments at node 3; Figure 4 provides an example of this observation. This lag is due to the low arrival rate to node 2 early in the time horizon (a result of empty initial conditions at node 1) yielding an inappropriately large value for τ in Step 2 of Algorithm 3.1, when employed at node $n = 2$, at time $t = 0$. In response, the fitted departure rate from node 2 is small during the interval $[0, \tau]$, leading to the underestimation

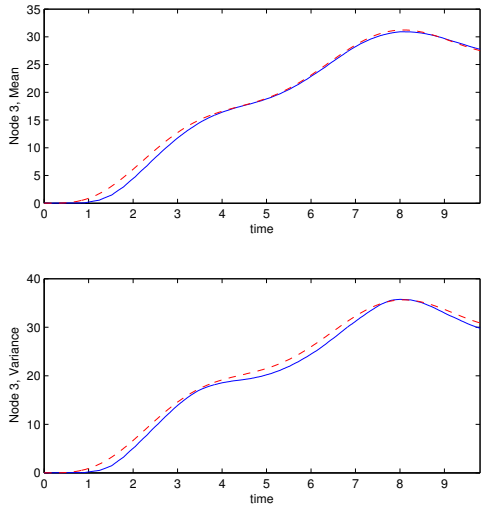

Figure 4: Network 4 extended to third node, with AOL(3) = 22.45, $scv_s^{(3)} = 1.03$, AU(3) = 48.6%: Plots of the fitted (solid) versus true (dashed) moments for mean (top) and variance (bottom) size at node 3; plots indicate the matching technique yields a satisfactory fit after an initial lag.

of the fitted moments at node 3 during this interval. Since τ was large, it took more time for this initialization effect to wear off. One potential solution is to delay the start time for initiating the matching algorithm at node 2 until the departure count moments from node 1 are non-negligible. This will yield shorter predicted values for τ in response to larger arrival rates to node 2 early in the time horizon; thus, we can expedite the approximated arrival stream to node 3, reducing the length of the time lag for the fitted moments.

Notice that we must also be cautious (when analyzing tandem chains of length $z > 2$) that the fitted mean interarrival times at node $n = 2, 3, \ldots, z-1$ fall within the range of $m_a^{(1)}$ provided in Table 1, as the kriging metamodel does not perform extrapolation; however, this may be easily resolved by expanding the range of parameters used to fit the metamodel in Section 3.3. No such action was necessary in the analysis of the five sample networks provided in this section.

5 Conclusions

In this paper we have provided an algorithm for approximating a tandem queueing network where the external arrival process is a nonstationary MAP_t. Our analysis of the matching technique, as employed in a two-node, finite-server network, indicates the technique is successful in accurately yielding the time-dependent downstream mean node size (and, to a slightly lesser extent, the time-dependent variance of the downstream node size) across a wide range of network parameters; however, the matching technique (and the MDE/DDE approach at its core) may provide a poor fit if nodal server utilizations are near (or above) 100% for a significant portion of the time horizon.

Opportunities for future research abound; these include investigating methods to improve the quality of the matching technique in the parameter ranges identified in the previous section, as well as in the presence of correlated arrivals and tandem chains of more than two nodes. Other opportunities involve investigating the necessary adjustments to employ the matching technique in general feed-forward networks, such as proposing methods for approximating the superposition of multiple streams of nonstationary internode traffic flow.

Acknowledgments

The authors thank Michael Taaffe and Jeremy Staum for helpful discussions. This work is supported by National Science Foundation Grant DMII-0521857.

References

[1] S. Asmussen. *Applied Probability and Queues*. John Wiley & Sons, New York, 1987.

[2] G. M. Clark. Use of Polya distributions in approximate solutions to nonstationary $M/M/s$ queues. *Commun. ACM*, 24(4):206–217, 1981.

[3] A. B. Clarke. A waiting line process of Markov type. *Annals of Math. Stat.*, 27(2):452–459, June 1956.

[4] D. R. Cox and W. L. Smith. On the superposition of renewal processes. *Biometrika*, 41(1/2):91–99, 1954.

[5] J. E. Diamond and A. S. Alfa. On approximating higher order MAPs with MAPs of order two. *Queueing Systems*, 34:269–288, 2000.

[6] J. W. Drane, S. Cao, L. Wang, and T. Postelnicu. Limiting forms of probability mass functions via recurrence formulas. *The American Statistician*, 47(4):269–274, 1993.

[7] M. K. Girish and J. Q. Hu. Higher order approximations for tandem queueing networks. *Queueing Systems*, 22:249–276, 1996.

[8] G. Hasslinger and E. S. Rieger. Analysis of open discrete time queueing networks: A refined decomposition approach. *The Journal of the Operational Research Society*, 47(5):640–653, 1996.

[9] B. R. Haverkort. Approximate analysis of networks of $PH/PH/1/K$ queues: Theory & tool support. In *Messung, Modellierung und Bewertung von*, pages 239–253, 1995.

[10] A. Heindl. Decomposition of general tandem queueing networks with MMPP input. *Performance Evaluation*, 44(1-4):5–23, 2001.

[11] A. Heindl. Decomposition of general queueing networks with MMPP inputs and customer losses. *Perform. Eval.*, 51(2–4):117–136, 2003.

[12] A. Heindl, G. Horváth, and K. Gross. Explicit inverse characterizations of acyclic MAPs of second order. In András Horváth and Miklós Telek, editors, *EPEW*, volume 4054 of *Lecture Notes in Computer Science*, pages 108–122. Springer, 2006.

[13] R. L. Iman, J. C. Helton, and J. E. Campbell. An approach to sensitivity analysis of computer models, Part 1: Introduction, input variable selection and preliminary variable assessment. *J. Qual. Technol.*, 13(3):174–183, 1981.

[14] M. A. Johnson. Markov MECO: A simple Markovian model for approximating nonrenewal arrival processes. *Communications in Statistics–Stochastic Models*, 14(1&2):419–442, 1998.

[15] M. A. Johnson and M. R. Taaffe. Matching moments to phase distributions: Mixtures of Erlang distributions of common order. *Communications in Statistics–Stochastic Models*, 5:711–743, 1989.

[16] N. L. Johnson and S. Kotz. *Urn Models and Their Applications*. John Wiley & Sons, New York, 1977.

[17] N. L. Johnson, S. Kotz, and A. W. Kemp. *Univariate Discrete Distributions*. John Wiley & Sons, New York, 1992.

[18] S. Kim. The heavy-traffic bottleneck phenomenon under splitting and superposition. *European Journal of Operational Research*, 157(3):736–745, 2004.

[19] S. Kim, R. Muralidharan, and C. A. O'Cinneide. Taking account of correlations between streams in queueing network approximations. *Queueing Syst. Theory Appl.*, 49(3-4):261–281, 2005.

[20] D. M. Lucantoni. New results on the single server queue with a batch Markovian arrival process. *Communications in Statistics–Stochastic Models*, 7(1):1–46, 1991.

[21] D. M. Lucantoni, K. S. Meier-Hellstern, and M. F. Neuts. A single server queue with server vacations and a class of non-renewal arrival processes. *Advances in Applied Probability*, 22:676–705, 1990.

[22] W. Nasr. Personal Communication, 2007.

[23] W. Nasr and M. R. Taaffe. Analysis and approximations for time-dependent queueing models. Technical report, Virginia Tech, Grado Department of Industrial and Systems Engineering, Blacksburg, VA, 2006.

[24] W. Nasr and M. R. Taaffe. Fitting Ph_t distributions to queueing departure count moments. Technical report, Virginia Tech, Grado Department of Industrial and Systems Engineering, Blacksburg, VA, 2006.

[25] W. Nasr and M. R. Taaffe. Fitting Ph_t tandem queues with M_t service and Ph_t arrival processes. Technical report, Virginia Tech, Grado Department of Industrial and Systems Engineering, Blacksburg, VA, 2007.

[26] B. L. Nelson and M. R. Taaffe. The $[Ph(t)/Ph(t)/\infty]^k$ queueing system: Part II – The multiclass network. *INFORMS Journal on Computing*, 16:275–83, 2004.

[27] B. L. Nelson and M. R. Taaffe. The $Ph(t)/Ph(t)/\infty$ queueing system: Part I – The single node. *INFORMS Journal on Computing*, 16:266–74, 2004.

[28] B. L. Nelson and M. R. Taaffe. The $MAP_t/Ph_t/\infty$ queueing system and multiclass $[MAP_t/Ph_t/\infty]^K$ queueing network. Technical report, Virginia Tech, Grado Department of Industrial and Systems Engineering, Blacksburg, VA, 2006.

[29] J. Neter, M. Kutner, W. Wasserman, and C. Nachtsheim. *Applied Linear Statistical Models*. McGraw-Hill/Irwin, 5th edition, 2004.

[30] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. The Johns Hopkins University Press, 1981.

[31] K. L. Ong and M. R. Taaffe. Approximating nonstationary $Ph_t/M_t/s/c$ queueing systems. *Annals of Operations Research*, 8:103–116, 1987.

[32] K. L. Ong and M. R. Taaffe. Approximating $Ph(t)/Ph(t)/1/c$ nonstationary queueing systems. *Mathematics and Computers in Simulation*, 30:441–452, 1988.

[33] K. L. Ong and M. R. Taaffe. Nonstationary queues with interrupted Poisson arrivals and unreliable/repairable servers. *Queueing Systems: Theory and Applications*, 4:27–46, 1989.

[34] M. Reiser and H. Kobayashi. Accuracy of the diffusion approximation for some queueing systems. *IBM J. Res. Dev.*, 18:110–124, 1974.

[35] M. H. Rothkopf and S. S. Oren. A closure approximation for the nonstationary $M/M/s$ queue. *Management Sci.*, 25(6):522–534, June 1979.

[36] J. E. Rueda. The $Ph(t)/Ph(t)/s/c$ queueing model and approximation. Master's thesis, Virginia Tech, Grado Department of Industrial and Systems Engineering, Blacksburg, VA, 2003.

[37] J. E. Rueda and M. R. Taaffe. The $MAP_t/MSP_t/s/k$ queueing model and approximation. Technical report, Virginia Tech, Grado Department of Industrial and Systems Engineering, Blacksburg, VA, 2005.

[38] J. E. Rueda and M. R. Taaffe. The $Ph_t/Ph_t/s/k$ queueing model and approximation. Technical report, Virginia Tech, Grado Department of Industrial and Systems Engineering, Blacksburg, VA, 2005.

[39] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989.

[40] T. J. Santner, B. Williams, and W. Notz. *The Design and Analysis of Computer Experiments*. Springer-Verlag, 2003.

[41] C. Sauer and K. Chandy. Approximate analysis of central server models. *IBM J. of Research and Development*, 19:301–313, 1975.

[42] G. Schneider, M. Schuba, and B. R. Haverkort. QNA-MC: A performance evaluation tool for communication networks with multicast data streams. In *TOOLS '98: Proceedings of the 10th International Conference on Computer Performance Evaluation: Modelling Techniques and Tools*, pages 63–74, London, UK, 1998. Springer-Verlag.

[43] J. G. Shanthikumar and J. A. Buzacott. Open queueing network models of dynamic job shops. *Int. J. Prod. Res.*, 19:255–266, 1981.

[44] R. M. Soland. A renewal theoretic approach to the estimation of future demand for replacement parts. *Operations Research*, 16(1):36–51, 1968.

[45] K. Sriram and W. Whitt. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE J. on Selected Areas in Communications, SAC*, 4(6):833–846, 1986.

[46] H. C. Tijms. *Stochastic Models: An Algorithmic Approach*. John Wiley & Sons, Inc, Chichester, England, 1994.

[47] W. Whitt. Approximating a point process by a renewal process, I: Two basic methods. *Operations Research*, 30:125–147, 1982.

[48] W. Whitt. Performance of the Queueing Network Analyzer. *Bell System Technical Journal*, 62(9):2817–2843, Nov. 1983.

[49] W. Whitt. The Queueing Network Analyzer. *Bell System Technical Journal*, 62(9):2779–2815, Nov. 1983.

[50] W. Whitt. Towards better multi-class parametric-decomposition approximations for open queueing networks. *Ann. Oper. Res.*, 48:221–248, 1994.

[51] W. Whitt. Variability functions for parametric-decomposition approximations of queueing networks. *Manage. Sci.*, 41(10):1704–1715, 1995.

[52] W. Whitt. Decomposition approximations for time-dependent Markovian queueing networks. *Operations Research Letters*, 24(3):97–103, 1999.