A Simple Approximation for Modeling Nonstationary Queues*

Wei-Ping Wang, David Tipper and Sujata Banerjee
Telecommunications Program
Department of Information Science
University of Pittsburgh, Pittsburgh, PA 15260
Email: {wwang, tipper, sujata}@tele.pitt.edu

Abstract

Evaluation of the behavior of queues with nonstationary arrival processes is of importance in several applications including communication networks. However, the analysis of nonstationary queues is in general computationally complex, and seldom produces closed form expressions. Thus approximation methods may be more appropriate. In this paper, the pointwise stationary fluid flow approximation (PSFFA) for determining the mean queue length of nonstationary queues is presented. The PSFFA combines steady state queueing results with a simple fluid flow model to develop a single nonlinear differential equation model of the queue. Numerical integration techniques are used to solve the PSFFA model and the method is illustrated by several examples. The power of this approach is that it can handle very general queueing systems.

1 Introduction

In many real world queueing systems, including communication networks, the customer arrival process is nonstationary with the arrival process parameters depending on the time of day [6]. Communication networks in particular are subject to a variety of phenomena that give rise to transient/nonstationary conditions such as load sharing, changes in routing and flow control parameters, failure of links, nodes or other network resources and most commonly, nonstationary input loads. There is empirical evidence that the user demand for communication is nonstationary in many networks, varying with the time of day [3]. Furthermore, as communication networks evolve to encompass a wide range of data rates which are utilized to transport complex traffic types with various quality of service requirements, the traffic in the network is expected to be very bursty and nonstationary in nature.

The relatively scarce literature that exists on transient/non-stationary analysis can largely be grouped into four areas: i) simulation techniques ii) transient analysis techniques, iii) nonstationary analysis techniques and iv) applications of the analysis methods. Note that a distinction is made between transient behavior and nonstationary behavior since transient behavior describes the system going from one stationary load to another, whereas, nonstationary behavior occurs when the arrival and/or service rate vary continuously with time.

Simulation methods observe the behavior of the system over an ensemble of statistically identical but distinct independent replications. This is accomplished by running the simulation a large number of times and averaging the quantities of interest across an ensemble of independent runs at a particular time instant. Many such points may be obtained at different time instants and the behavior of the system studied as a function of time. The principle difficulty in conducting simulation studies of this type is the large number of independent runs that must be generated in order to get a representative ensemble from which a statistically accurate portrayal of the system behavior can be determined. Hence, very large amounts of computer run time for even moderate sized networks are required [13].

Analytical transient analyses usually involve the use of transform techniques to solve differential/difference equation models from an embedded Markov process/chain. The result of the analysis is normally a transform expression for $p^i(t)$, the time varying probability of i customers in a single queueing system. Only in some special simple cases are the transform expressions invertible to yield a closed form expression and even then the result is usually computationally complex to evaluate. Hence, there has been an effort to numerically determine the transient behavior rather than deriving a closed form expression.

Numerical approaches have largely focused on two methods: uniformization and numerical analysis techniques. The basic idea in uniformization is to convert a finite state space Markov process into an equivalent discrete time Markov chain and Poisson process [5]. One then works with the state transition matrix of the Markov Chain and a truncated version of the Poisson random variable to find the transient behavior as discussed in [5]. In contrast, for the numerical methods approach the underlying differential/difference equation model is numerically solved using standard numerical analysis techniques (e.g., Runge-Kutta method). The two approaches are compared in terms of computational complexity and accuracy in [18]. The principal disadvantage of both methods is that the computational complexity grows with the queue state space and one is limited to considering Markovian type sys-

*The work reported here was supported in part by a grant from IBM, Research Triangle Park, NC

tems. In order to determine the transient behavior of nonMarkovian queues, several approximate analysis methods have been proposed such as diffusion models [2], fluid flow models [20, 15], and service time convolution [11]. In [20] we have extended the numerical analysis method for transient Markovian queueing analysis to the more general nonstationary case. The approach is to approximate the time varying queue arrival and/or service rates by constants over a small time interval and then numerically solve the underlying differential/difference equation model. The procedure is then repeated for all the time intervals of interest. Similar approaches to extending transient numerical analysis techniques to approximate the general nonstationary behavior for Markovian systems using uniformization [21] and Floquet's method [22] have recently appeared. These methods have the drawback that they are computationally intensive, the computation required depends on the state space of the queue and they are limited to Markovian systems. For more general nonMarkovian queues a few approximations have been proposed namely; diffusion models, fluid flow models, the pointwise stationary approximation [7] and the modified offered load approximation [10, 14].

Here we are interested in identifying techniques that can be used for the design of network controls and the performance evaluation of communication networks. Since many network controls and performance studies are done on the basis of average quantities, we focus on determining the mean transient/nonstationary behavior of queueing systems. In this paper we present an approximate fluid flow modeling method for determining the mean behavior of queues with general arrival and service distributions. This work was motivated in part by the results presented by Greene et al. [7] using the Pointwise Stationary Approximation (PSA). The PSA is obtained by computing performance measures at each time point during the period of interest using the steady state (i.e., stationary) queueing formulas with the arrival rate that corresponds to each point in time. The instantaneous arrival rate is determined from the time varying arrival process. This instantaneous rate is then substituted into steady state queueing formulae for the particular queueing system under study. This process can be carried out over a desired time interval, and for periodic arrival processes the time average number in the system over a period can be computed. We describe below how the PSA method can be coupled with a fluid flow model to form the Pointwise Stationary Fluid Flow Approximation (PSFFA) modeling technique.

The PSFFA models the average number in the system at a queue by a single nonlinear differential equation which is solved numerically. The PSFFA approach derives the form of the fluid flow differential equation from the steady state queueing relationships for the model. The use of the approach to determine the nonstationary behavior of general finite and infinite capacity queueing systems is discussed below. The model is shown to be reasonably accurate for the cases considered and a considerable improvement over the PSA method. Note that we have modeled non-Markovian queues and it would appear that the approach is quite general in nature and represents a generalization of our earlier results on fluid flow modeling [20, 19]. In fact, it may be possible to develop the fluid flow model from measurement data. The principal advantages of this approach are its generality, its simplicity in modeling queueing systems and computational efficiency. Additionally, these methods could be used as the basic mathematical model for developing dynamic network control mechanisms along the lines of [16] and [9].

2 The Pointwise Stationary Fluid Flow Approximation

Consider a single server queueing system with a nonstationary arrival process. Let μ denote the average queue service rate and $\lambda(t)$ denote the ensemble average arrival rate at time t. We define $x(t)$ as the state variable representing the ensemble average number in the system at time t. Let $\dot{x}(t) = \frac{dx(t)}{dt}$ be the rate of change of the state variable with respect to time. From the flow conservation principle, the rate of change of the average number in the system is equal to the difference between the average arrival and departure rates. Let $f_{in}(t)$ and $f_{out}(t)$ denote the ensemble average flow in and flow out of the system at time t, respectively. The rate of change of the state variable can be related to the flow in and flow out by

$$\dot{x}(t) = -f_{out}(t) + f_{in}(t) \ . \tag{1}$$

This type of equation is commonly referred to as a fluid flow or dynamic flow equation [1, 9, 13, 20, 4]. The flow out of the system $f_{out}(t)$ can be related to the ensemble average utilization of the server $\rho(t)$ by $f_{out}(t) = \mu\rho(t)$. If the queue waiting space is infinite, then the flow into the system is just the arrival rate (i.e., $f_{in}(t) = \lambda(t)$) and the fluid flow model of Eqn. (1) becomes

$$\dot{x}(t) = -\mu\rho(t) + \lambda(t) \ . \tag{2}$$

The expression for $\rho(t)$ in Eqn. (2) will depend on the queueing system under study. In general, determining an exact expression for $\rho(t)$ is quite difficult even for the simplest queues. Hence, an approximate method based on the PSA method is adopted. The general idea is to determine the values for $\rho(t)$ at particular instants of time by a pointwise mapping from the current value of $x(t)$ into ρ using the steady state queueing relationships. Then the value of ρ thus obtained is used to numerically solve (2) over a small time interval to get a new $x(t)$ and the procedure is repeated for the next time step.

Considering the infinite queue case of Eqn. (2), we assume that at steady state (i.e., $\dot{x}(t) = 0$) the following functional relationship can be determined:

$$x = G_1(\rho) \ . \tag{3}$$

Additionally, we assume that the functional relationship $G_1(\rho)$ is numerically invertible, that is $\rho = G_1^{-1}(x)$. This results in the PSFFA model

$$\dot{x}(t) = -\mu(G_1^{-1}(x(t))) + \lambda(t) \ . \tag{4}$$

Note that Eqn. (4) is quite general in nature — the only requirement being that the functional relationship G_1 be determined and invertible. For many queueing systems the function G_1 is well known in closed form. Furthermore, for some queueing systems $G_1(\rho)$ is invertible and one can derive a closed form expression for the PSFFA model as per Eqn. (4). This is however not a requirement, as the function G_1 can be determined numerically or by *curve fitting from measurements* for an existing system. One advantage of determining the approximate expression for $\rho(t)$ in (2) using the approach above is that the resulting fluid flow model (4) is exact under steady state conditions. Hence, in solution of the PSFFA model for the transient response, the model will always converge to the correct steady state value.

The PSFFA model for the infinite queue (4) can easily be numerically solved to determine the time varying mean behavior of the queueing system [20]. The basic solution procedure is described here. We identify the initial condition for the state variable at time zero as $x(0)$ and assume the arrival rate to be a constant over a very small time step $[0, \Delta t]$ (i.e., $\lambda(t) = \lambda(\Delta t/2)$ for $t \in [0, \Delta t]$. Then Eqn. (4) can be numerically integrated for the value of the state variable at the end of the time interval, $x(\Delta t)$. Note that in solving the fluid flow model over a small time interval one may need to apply a numerical procedure to find $G_1^{-1}(x)$. The state variable value at the end of the time interval, $x(\Delta t)$, then becomes the initial condition for the next time step $[\Delta t, 2\Delta t]$. We then adjust the arrival rate for the new time step. This procedure is repeated for each time interval in the time horizon. For all numerical solutions to the differential equations used in this paper, the fifth order Runge-Kutta routine provided in MATLAB was utilized. Our numerical results have been validated by simulations carried out in SLAM [17] using the ensemble averaging technique of [13]. Specifically, we conducted 10,000 independent simulation runs of the system under study and determined average values across the 10,000 simulations at each time point to construct the ensemble average curves shown. For all simulation results three curves are shown, the middle curve represents the estimate from simulation and the upper and lower curves correspond to the 95% confidence intervals. As an illustration of the PSFFA method several queueing systems have been modeled in the following sections.

2.1 The M/G/1 Queue

Consider an M/G/1 queue where the arrival process is Poisson and the service time is arbitrarily distributed with successive service times being independent and identically distributed. The well-known Pollaczek-Khintchine (P-K) formula [8], gives the average number in the system at steady state, x (i.e., the state variable) as:

$$x = \rho + \frac{\rho^2(1 + C_s^2)}{2(1 - \rho)} \ . \qquad (5)$$

where C_s^2 is the squared coefficient of variation of the service time distribution. Note that Eqn. (5) corresponds to the functional relationship $x = G_1(\rho)$ of Eqn. (3) and in this case it can be inverted in a closed form to yield

$$\rho = \frac{x + 1 - \sqrt{x^2 + 2C_s^2 x + 1}}{1 - C_s^2} \ . \qquad (6)$$

Hence the PSFFA equation for the M/G/1 queue is given, using Eqns. (4) and (6), as

$$\dot{x} = -\mu \left[\frac{x + 1 - \sqrt{x^2 + 2C_s^2 x + 1}}{1 - C_s^2} \right] + \lambda(t) \ . \qquad (7)$$

M/D/1	$\dot{x} = -\mu \left[(x+1) - \sqrt{x^2+1} \right] + \lambda$
M/E_k/1	$\dot{x} = -\mu \left[\frac{k(x+1)}{k-1} - \frac{\sqrt{k^2x^2+2kx+k^2}}{k-1} \right] + \lambda$
M/M/1	$\dot{x} = -\mu \left(\frac{x}{x+1} \right) + \lambda$

Table 1: M/G/1 PSFFA Models

For a specified coefficient of variation of the service time distribution C_s^2 Eqn. (7) can be solved numerically for the time varying behavior of the average number in the system. Table 1 lists some special cases of the M/G/1 PSFFA for various common service time distributions namely: D - deterministic service times with $C_s^2 = 0$; E_k - Erlang-k distributed service times with $C_s^2 = 1/k, k \geq 1$; and M - exponentially distributed service times with $C_s^2 = 1$. Note that for the special case of the M/M/1 queue, the service distribution is exponential with $C_s^2 = 1$ which results in the expression for $\rho(t)$ in Eqn. (6) becoming an indeterminate form of 0/0 and L'Hospital's rule must be applied to obtain the expression given in Table 1.

The accuracy of the M/G/1 PSFFA model has been studied by extensive comparison with simulation, and for the sake of brevity we summarize typical results here (see [23] for additional M/G/1 results and for additional M/M/1 and M/D/1 results see our earlier work in [20] and [19]). In order to illustrate the accuracy of the PSFFA, different numerical cases were considered, for various traffic patterns. From our numerical studies (including results not given here, see [23]) we conclude that the PSFFA model transient response in general exceeds the simulation results for heavy loads, on the other hand it under estimates the simulation results for light loads.

Following the previous literature on nonstationary analysis of communication networks [3, 6, 20], we consider the nonstationary load to follow a sinusoidal mean behavior representing the cyclic load pattern over a fixed time interval period (e.g., day), specifically $\lambda(t) = A + Bsin(wt + D)$. The effects of other nonstationary arrival patterns are given in [23] and [19]. Typical results for the nonstationary behavior

of the $M/G/1$ PSFFA models of Table 1 are given in Figures 1, 2, and 3 for the $M/D/1, M/E_2/1$ and $M/M/1$ models respectively. In Figures 1, 2, and 3, the average number in the queueing system x is plotted versus time for the nonstationary traffic $\lambda(t) = 0.5+0.4\sin(0.2(t+20))$[1] with mean service rate $\mu = 1.0$ and initial condition $x(0) = 0.1$. Additional numerical results for the nonstationary behavior of other M/G/1 type models are given in [23]. It is readily seen for Figures 1, 2, and 3 that the PSFFA model produces the same form of response as the corresponding simulation (i.e., the curves have the same shape) and overshoots the magnitude of peaks and valleys in the response. Comparing the figures it can be seen that the error between the PSFFA model and the simulation results increases with increasing C_s^2. This was found to hold for the transient results as well. However, the model is reasonably accurate and has considerable computational advantage over the corresponding simulation.

2.2 The GI/M/1 Queue

In this section, we concentrate on the G/M/1 queueing model where the service time is exponentially distributed and the interarrival process is generally distributed with successive interarrival times independent and identically distributed. Let $A(t)$ denote the interarrival time distribution. Following [8] the GI/M/1 queue steady state analysis is performed by embedding a Markov chain at the customer arrival instant. The steady state distribution for the number of customers found in the system by a new arrival for the GI/M/1 queue is a geometric distribution:

$$\pi_n = (1-\sigma)\sigma^n \ .$$

The parameter σ is the unique real root in the range $0 < \sigma < 1$ of the transcendental equation

$$\sigma = f_a^*(s) \mid_{s=\mu(1-\sigma)} \tag{8}$$

where $f_a^*(s)$ is the Laplace–Stieltjes transform of the interarrival time distribution $A(t)$, that is

$$f_a^*(s) = \mathcal{L}^*(A(t)) = \int_0^\infty e^{-st} dA(t) \ . \tag{9}$$

Note, that in solving Eqn. (8), $\sigma = 1$ is always a root of the equation. From the standard GI/M/1 queueing formula [8], at steady state the average number in the system, x, is

$$x = \frac{\lambda}{\mu(1-\sigma)} = \frac{\rho}{(1-\sigma)} \ . \tag{10}$$

In determining the PSFFA model, Eqn. (10) corresponds to the needed steady state relationship (3) and inverting (10) for ρ result in

$$\rho = x(t)(1-\sigma(t)) \ . \tag{11}$$

[1] The sine wave traffic pattern is shifted in time by 20 units, to allow the corresponding simulation program to *warm up*.

Therefore, the pointwise stationary fluid flow equation for the GI/M/1 queueing model is

$$\dot{x}(t) = -\mu x(t)(1-\sigma(t)) + \lambda(t) \ . \tag{12}$$

For a GI/M/1 queue, given the interarrival time distribution $A(t)$, we can use Eqns (8) and (9) to solve for the parameter σ, and then the PSFFA Eqn. (12) can be solved numerically to get the time varying behavior of the queueing system. Note that in some special cases it is possible to solve (8) to get a closed form expression for σ and the PSFFA model of (12) (e.g., the $E_2/M/1, M/M/1, C_2/M/1$, etc. see [23] for details). In general one can not get a closed form for σ and one must numerically determine σ for each new value of $\lambda(t)$. This can either be incorporated as an additional step within the PSFFA solution procedure or σ can be precomputed over a range of λ and a table look up used to find σ given λ in solving Eqn. (12). The exact procedure for determining σ will depend upon the interarrival distribution $A(t)$, but will normally involve a root finding algorithm such as Laguerre's method. Table 2 lists the PSFFA along with the expression for σ found from (8) for several interesting cases of the GI/M/1 queue. The $D/M/1$ case in Table 2 corresponds to a deterministic arrival process where the interarrival time distribution $A(t)$ is a delta function (i.e., $dA(t) = f_a(t)dt$ and $f_a(t) = \delta(t - 1/\lambda),$). The $E_k/M/1$ entry in Table 2 corresponds to an Erlang-k interarrival distribution. The last entry in the table, the $IPP/M?/1$ queue corresponds to a Interrupted Poisson Process arrival process. The IPP is a Poisson process whose rate is a function of a two state Markov process, with the arrival rate in one state being zero. The IPP is a special case of the more general Markov-modulated Poisson Process (MMPP) [25]. The IPP is also called a a 2-state MMPP On–Off model. The IPP is characterized by the 2-state continuous-time Markov chain with infinitesimal generator Q and the Poisson arrival rate λ as shown below using the notation of [25].

$$Q = \begin{bmatrix} -\sigma_1 & \sigma_1 \\ \sigma_2 & -\sigma_2 \end{bmatrix} \text{and } \Lambda = \text{diag}(\lambda, 0) \ .$$

Here state 1 corresponds to the ON state and state 2 denotes the OFF state. The details of the derivation of the expression for σ given in Table 2 for the $D/M/1$ and $E_k/M/1$ cases can be found in [8] and for the IPP in [23].

Several different cases of the general GI/M/1 PSFFA model have been compared with simulation results in [23] for various traffic loads. Here we summarize representative results. Typical results for the transient behavior of the G/M/1 PSFFA model is shown in Figure 4 where the average number in the system is plotted versus time. Figure 4 shows the transient behavior of the D/M/1 queue with mean service rate $\mu = 1$, initial condition $x(0) = 0$ and arrival rate $\lambda = 0.4$. Notice that for the deterministic arrival process an arrival rate of $\lambda = 0.4$ results in a customer arrival every $2.5 = 1/\lambda$ time units and a jump in the number in the system by 1 at the arrival instance. Hence the system in effect goes through a series of transients rather than converging to a simple

Queueing System	PSFFA Equation	σ
D/M/1	$\dot{x}(t) = -\mu x(t)(1-\sigma) + \lambda(t)$	$\sigma = e^{\frac{\mu}{\lambda}(\sigma-1)}$
E_k/M/1	$\dot{x}(t) = -\mu x(t)(1-\sigma) + \lambda(t)$	$\sigma = \left(\frac{k\lambda}{k\lambda+\mu-\mu\sigma}\right)^k$
IPP/M/1	$\dot{x}(t) = -\mu x(t)(1-\sigma) + \lambda(t)$	$\sigma = \frac{\lambda(\mu-\mu\sigma+\sigma_2)}{(\mu-\mu\sigma)^2+(\lambda+\sigma_1+\sigma_2)(\mu-\mu\sigma)+\sigma_2\lambda}$

Table 2: GI/M/1 PSFFA Models

steady state value. One can see that the PSFFA closely tracks the actual system behavior. The nonstationary behavior of the G/M/1 PSFFA model is illustrated for the $E_k/M/1$ and $IPP/M/1$ queues in Figures 5 and 6 respectively. Figure 5 plots the nonstationary behavior of the number in the system versus time for the $E_k/M/1$ queue with $k = 2, \mu = 1, x(0) = 0$ and $\lambda = 0.3+0.2\sin(0.2(t+20))$. Figure 6 shows the behavior of the state variable $x(t)$ for the for the IPP/M/1 queue with $\mu = 1.0$, $\sigma_1 = 0.1$, $\sigma_2 = 0.15, x(0) = 0$ and $\lambda = 0.3 + 0.2\sin(0.2(t+20))$. We can see the PSFFA model results closely match the simulation results in both Figures. The accuracy of the PSFFA model for the GI/M/1 queue for both transient and nonstationary results was found to be dependent on the parameter σ. The smaller the parameter σ is, the greater the accuracy of the PSFFA model. Note that in some G/M/1 models the parameter σ is proportional to the load and the accuracy decreases as the load increases.

2.3 The GI/G/1 Queue

In this section, we concentrate on the general queueing model, where both the interarrival process and the service process are arbitrarily distributed with successive interarrival times and service times independent and identically distributed. For the GI/G/1 queueing system determining the steady state behavior is difficult and many approximations have been proposed. A well-known approximation for the expected number in the system, x, in the GI/G/1 queueing system was presented by Kramer and Lagenbach-Belz [12].

$$x \approx \rho + \frac{\rho^2 \cdot (C_a^2 + C_s^2) \cdot J(C_a^2, C_s^2, \rho)}{2(1-\rho)} \qquad (13)$$

$$\text{with } J(C_a^2, C_s^2, \rho) = \begin{cases} e^{-\frac{2(1-\rho)(1-C_a^2)^2}{3\rho(C_a^2+C_s^2)}} & C_a^2 \le 1 \\ e^{-\frac{(1-\rho)(C_a^2-1)}{C_a^2+4C_s^2}} & C_a^2 \ge 1 \end{cases}.$$

In Eqn. (13), ρ is the server utilization, C_a^2 and C_s^2 represent the squared coefficients of variation of the interarrival and service processes, respectively. Here we use (13) to approximate the steady state relationship needed in (3) to develop the PSFFA model. It is generally not possible to invert (13) in closed form for ρ and numerical techniques must be adopted to determine ρ given x. Given a value of $\rho = G^{-1}(x)$ for a particular x, the general PSFFA model given by Eqn. (4) can be solved over an appropriate time interval. A possibly more accurate approach to the G/G/1 system would be to determine the steady state functions $x = G(\rho)$ and $\rho = G^{-1}(x)$ by curve fitting either steady state measurement data from a system or steady state simulation results. The following algorithm is used to determine the behavior of the queue over a time interval $[t_0, t_f]$:

1. Initialization: set the current time, t, to $t = t_0$ and establish the initial system occupancy (e.g., $x(t_0) = 0$ etc.).
2. Numerically solve Eqn. (13), or use curve fitting to get the value $\rho = G^{-1}(x)$.
3. Solve the differential equation given by Eqn. (4) over a small time interval Δt, approximating the arrival rate by $\lambda = \lambda(t + \Delta t/2)$, and get the new system occupancy at time $t + \Delta t$, $X(t + \Delta t)$.
4. Increment time, $t = t + \Delta t$. If $t < t_f$, goto 2, else stop.

This iteration is carried out until the desired final time t_f is reached.

As a simple example of using the general GI/G/1 approximation model given by Eqn. (13) we consider the $E_k/E_k/1$ queue. For the Erlang-k distribution, the squared coefficient of variation is $1/k$ (i.e., $C_a^2 = C_s^2 = 1/k$). Using the G/G/1 PSFFA solution procedure above, the $E_k/E_k/1$ model was compared with simulation results for various traffic patterns. Some representative results for the $E_k/E_k/1$ PSFFA model are shown in Figure 7, where the average number in the $E_k/E_k/1$ queueing system is plotted versus time. The results are plotted for the $E_k/E_k/1$ queueing system with $k = 2$, mean service rate $\mu = 1.0$, and initial condition $x(0) = 0$. Figure 7 plots the nonstationary behavior of the average number in the system versus time, with the arrival rate $\lambda = 0.4+0.3\sin(0.2(t+20))$. While the PSFFA model closely tracks the actual system behavior, it is less accurate at high loads. Additional numerical results for $G/G/1$ models are in [23].

3 Modeling Finite Queues

In this section, we extend the PSFFA to model queueing systems with finite capacity. As in the derivation of the PSFFA model for infinite queues, we begin with the basic flow conservation Eqn. (1), which relates the rate of change of the state variable (i.e., the average number in the system), $\dot{x}(t)$, to the ensemble average flow into the queue $f_{in}(t)$, and the ensemble average flow out of the queue $f_{out}(t)$. Let μ denote the mean service rate, $P_0(t)$ the probability of zero customers in the system at time t, $\lambda(t)$ the ensemble average arrival rate at time t and $P_B(t)$ the customer

blocking probability at time t. Note that for a case of a finite queue of size N, the flow into the queue depends on the blocking that the queue offers to the input traffic[2]. Specifically, the actual flow into the queue will be $f_{in}(t) = (1 - P_B(t))\lambda(t)$. The flow out of the queue can be related to the probability that the server is busy by $f_{out}(t) = \mu(1 - P_0(t))$ and Eqn. (1) can then be written as

$$\dot{x}(t) = -\mu(1 - P_0(t)) + (1 - P_B(t))\lambda(t) \ . \qquad (14)$$

Determining the exact expressions for $P_0(t)$ and $P_B(t)$ is quite difficult for even simple systems and we again adopt a PSA type approximation method. The approach is to do a pointwise mapping from the current value of $x(t)$ to ρ using steady state results, then using ρ estimate P_0 and P_B using steady state relationships. The values of P_0 and P_B are then used in (14) which is numerically solved over a small time step to get a new value for $x(t)$ and the procedure is repeated for the next time interval. Specifically, we assume the following steady state functional relationships can be determined for the average number in the system, x, the probability of zero customers in the system, P_0, and the blocking probability, P_B:

$$x = G_1(\rho) \quad P_0 = G_2(\rho) \quad P_B = G_3(\rho) \ . \qquad (15)$$

Furthermore, we assume that the function $G_1(\rho)$ is numerically invertible such that $\rho = G_1^{-1}(x)$. This results in the PSFFA model for finite queues as

$$\dot{x}(t) = -\mu(1 - G_2(G_1^{-1}(x))) + \lambda(t)(1 - G_3(G_1^{-1}(x))) \ . \qquad (16)$$

This single differential equation can be solved for the time varying behavior of the queue under nonstationary arrival/service processes using the identical numerical approach as was adopted for the infinite PSFFA model of section 2.

The behavior of the finite queue case PSFFA model has been studied for a number of queueing models in [23]. Here we show typical results for the PSFFA model. Figure 8 shows the transient behavior of the average number in the system versus time for a M/M/1/20 PSFFA model with $\lambda = 0.2$, $\mu = 1.0$, and $x(0) = 1$. Note that the M/M/1/20 queue transient/nonstationary behavior can be determined exactly by integrating the Chapman Kolmogorov equations as discussed in [?].In Figure 8 the exact results along with the PSA results are shown. Figure 9 illustrates the PSFFA nonstationary behavior for the M/M/1/20 queue with $\lambda(t) = .4 + .3sin(0.2(t + 20))$, $\mu = 1$, and $x(0) = 0$. Again the PSA and exact results are included, along with the steady state results of using the average arrival rate (AVG). As shown in Figures 8 and 9 the PSFFA closely follows the exact solution and is a considerable improvement on the PSA. A detailed comparion of the PSA, MOL and PSFFA approximations for M/M/1/K queues is given in [24]. It was found that the PSFFA model is more accurate then the PSA and MOL approximations.

[2]We did not have this problem in the infinite queue case.

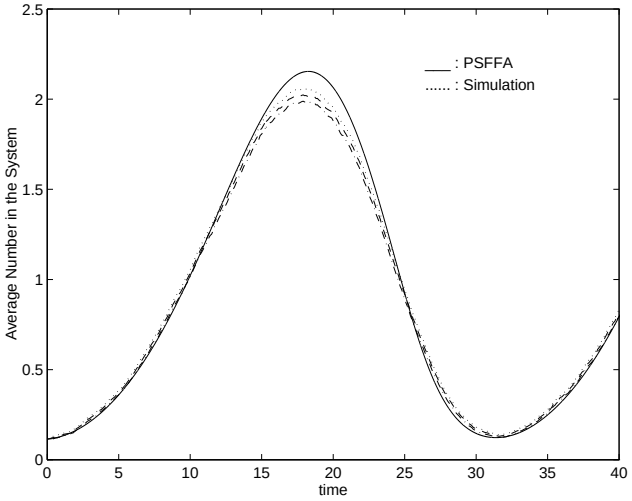

Figure 1: Comparison of the M/D/1 model with simulation for nonstationary traffic

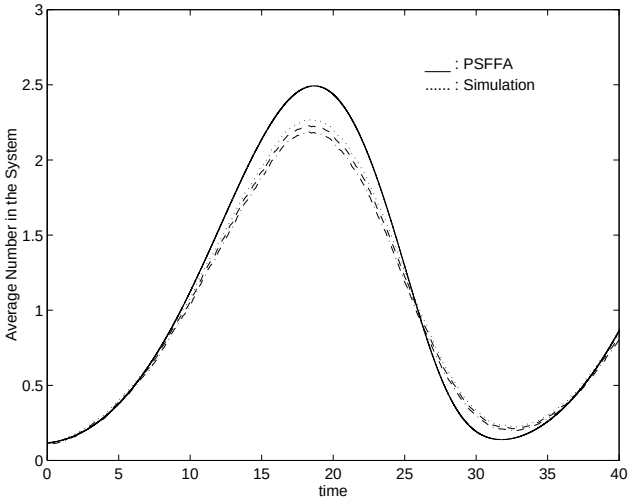

Figure 2: Comparison of the M/E_2/1 model with simulation for nonstationary traffic

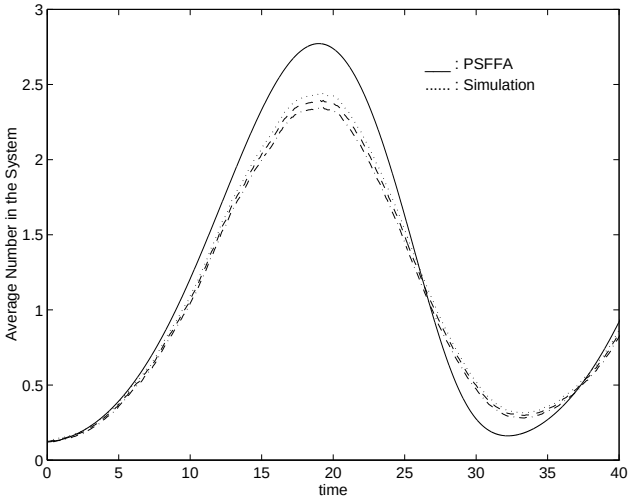

Figure 3: Comparison of the M/M/1 model with simulation for nonstationary traffic

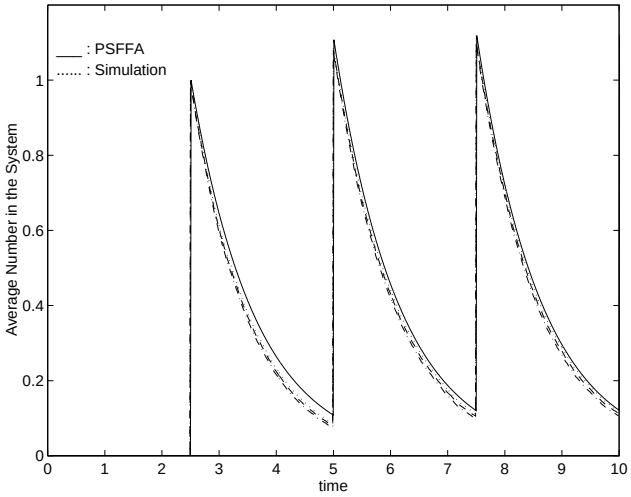

Figure 4: Comparison of the D/M/1 model with simulation for low load

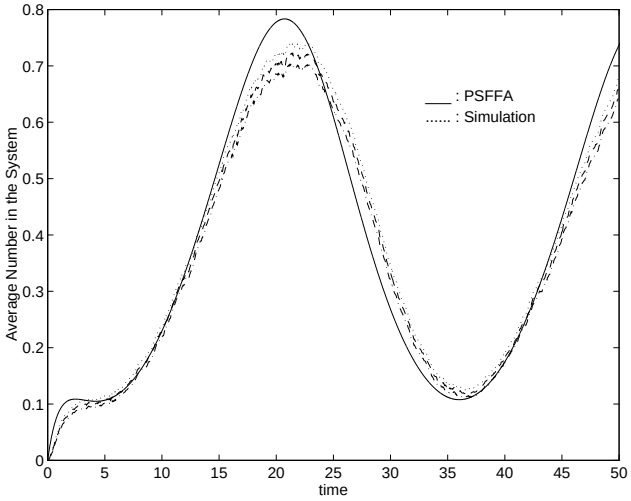

Figure 5: Comparison of the E_2/M/1 model with simulation for nonstationary traffic

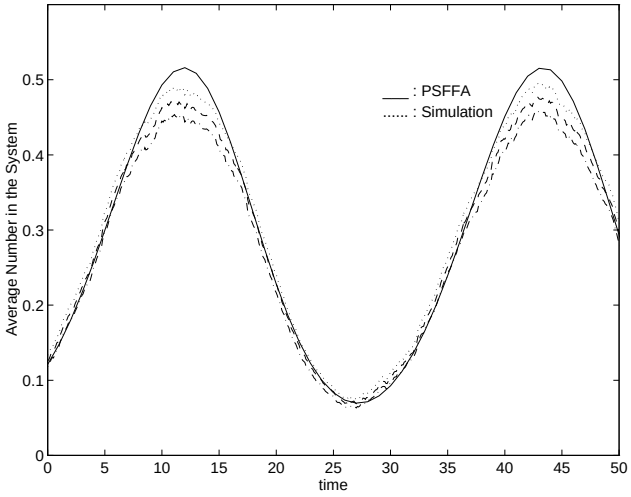

Figure 6: Comparison of the IPP/M/1 model with simulation for nonstationary traffic

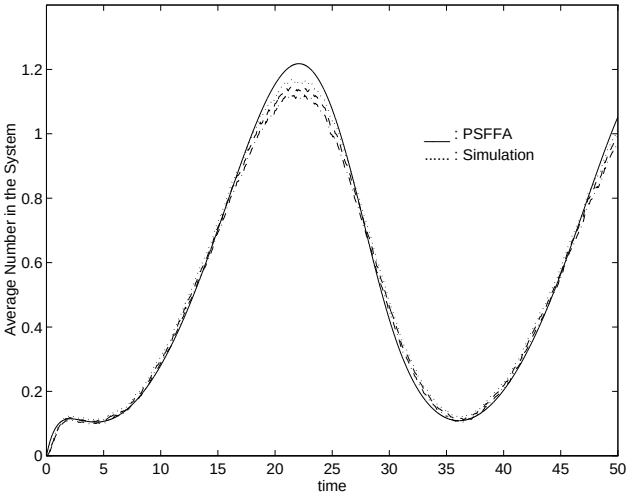

Figure 7: Comparison of the $E_2/E_2/1$ model with simulation for nonstationary traffic

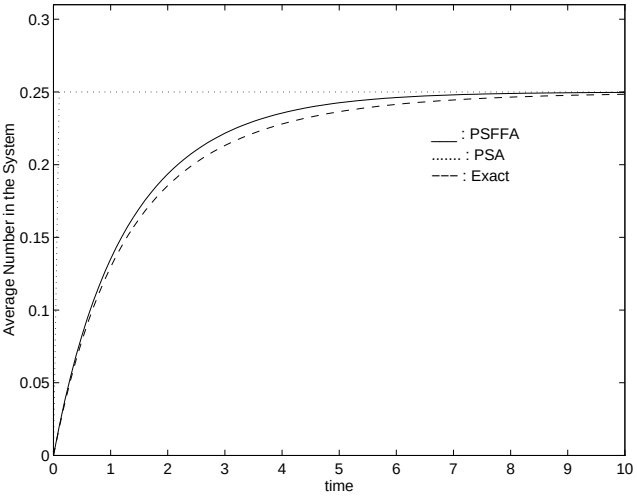

Figure 8: Comparison of the M/M/1/20 model transient behavior with exact solution

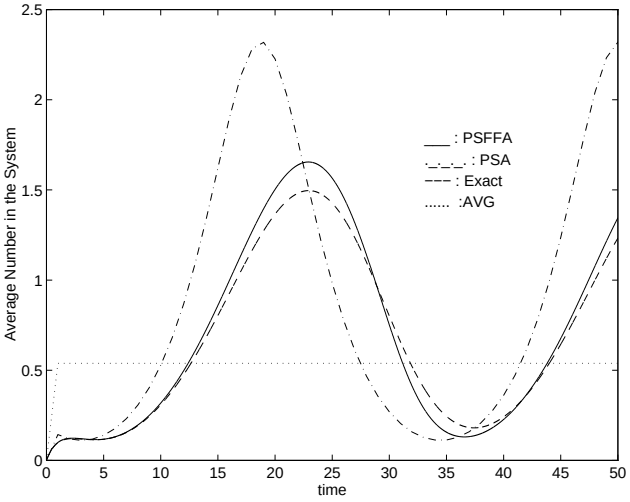

Figure 9: Comparison of the M/M/1/20 model nonstationary behavior with exact solution

4 Conclusions

In this paper, a relatively simple and computationally non-intensive approach to compute the mean behavior of various nonstationary queueing systems was presented. This approach termed the PSFFA integrates the previously proposed fluid flow model and the PSA method. In this paper, both infinite and finite queueing systems were considered, and examples for several queueing systems (M/G/1, G/M/1, G/G/1) were presented. The PSFFA models were found to be close to the results obtained by simulation, or wherever possible, exact calculation. Further, the accuracy was much higher using the PSFFA than using the PSA. One of the major advantages of the PSFFA approach (apart from the low computational overhead) is that it is a general approach requiring very few assumptions.

References

[1] C.E.Agnew, "Dynamic Modeling And Control Of Congestion-Prone Systems," *Operations Research*, 24(3):400-419, May 1976.

[2] A. Duda, "Transient Diffusion Approximations for Some Queueing Systems," *INFOR*, pp:118-28, 1983.

[3] S.G.Eick, W.A.Massey, and W.Whitt, "Nonstationarity in Offered Traffic to the AT&T Long Distance Network," *AT&T Joint Symposium on Performance Analysis and Teletraffic Research*, Dec. 1990.

[4] J.Filipiak, Real Time Network Management. *North Holland*, 1991.

[5] W. Fisher and K. Meier-Hellstern, "The Markov Modulated Poisson Process (MMPP) Cookbook", *Performance Evaluation*, 18:149-172, 1992.

[6] W.K.Grassmann, "Computational Methods in Probability Theory," Handbooks in Operations Research and Management Science,Vol. 2, *North Holland*, 1990.

[7] L.Green, P.Kolesar, and A.Svoronos, "Some Effects Of Nonstationarity On Multiserver Markovian Queueing Systems," *Operations Research*, 39(3):502-11, June 1991.

[8] L.Green and P.Kolesar, "The Pointwise Stationary Approximation for Queues With Nonstationary Arrivals," *Management Science*, 37(1):84-97, Jan. 1991.

[9] D.Gross and C.M.Harris, Fundamentals of Queueing Theory. 2nd ed, John Wiley & Sons, New York.

[10] X. Gu, K. Sohraby, and D. Vaman, Control and Performance in Packet, Circuit, and ATM Networks, *Kluwer*, 1995.

[11] D.L.Jagerman, "Nonstationary Blocking in Telephone Traffic," *Bell System Technical Journal*, 54(3): 625-661, Mar. 1975.

[12] D.L.Jagerman, "Approximate Mean Waiting Times in Transient GI/G/1 Queues," *Bell System Technical Journal*, 61(8): 2003-22, Oct. 1982.

[13] W.Kramer and M.Lagenbach-Belz, "Approximate Formulae for the Delay in the Queueing System GI/G/1," *8th International Teletraffic Congress*, 1976.

[14] W.Lovegrove, J.L.Hammond, and D.Tipper, "Simulation Methods for Studying Nonstationary Behavior of Computer Networks," *IEEE Journal on Selected Areas in Communication*, 8(9):1696-1708, Dec. 1990.

[15] W. A. Massey and W. Whitt, "On the modified-offered-load approximation for the nonstationary Erlang loss model," *Proceedings of the 14th International Teletraffic Congress -ITC 14*, pp 145-153, June 6-10, 1994.

[16] G.F.Newell, "Queues with Time Dependent Arrival Rates i, ii, iii," *Journal of Applied Probability*, 5:436-451,579-590, 591-606, 1968.

[17] A.Pitsillides, J.Lambert, and D. Tipper, "Dynamic Bandwidth Allocation In Broadband ISDN using a Multilevel Optimal Control Approach," *Telecommunication Systems* Vol. 4, No. I-II, 1995.

[18] A.B.Pritsker, Introduction to Simulation Using SLAM II. John Wiley & Sons, New York, 1986.

[19] A.Reibman and K.Trivedi, "Numerical Transient Analysis Of Markov Models," *Comput. Operations Res.*, 15(1):19-36, 1988.

[20] S.Sharma and D.Tipper, "Approximate Models for the Study of Nonstationary Queues and Their Applications," *Proceedings of 1993 IEEE International Conference on Communications*, June 1993.

[21] D.Tipper and M.K.Sundareshan, "Numerical Methods for Modeling Computer Networks Under Nonstationary Conditions," *IEEE Journal on Selected Areas in Communication*, 8(9):1682-1695, Dec. 1990.

[22] N. Van Dijk, "Uniformization for Nonhomogenous Markov Chains," *Operations Research Letters*, Vol. 12, pp. 283-291, 1992.

[23] I. Viniotis, A. Rindos, S. Woolet, and K. Trivedi, "Exact Methods for the Analysis of Nonhomogenous Continous Time Markov Chains," in Numerical Methods for Markov Chains, ed. W. Stewart, Elsevier, 1995.

[24] W.-P. Wang, "Fluid Flow Models for Analyzing the Nonstationary Behavior of High Speed Communication Networks," *M.S. Thesis, Computer Engineering,* Clemson University, Aug. 1994.

[25] W.-P. Wang, D. Tipper, and S. Banerjee, "The Pointwise Stationary Fluid Flow Approximation for Modeling Nonstationary Queues," *Technical Report* University of Pittsburgh, 1996. Available by anonymous ftp to **violet.tele.pitt.edu**.