Probability in the Engineering and Informational Sciences, **13**, 1999, 251–273. Printed in the U.S.A.

HEAVY TRAFFIC APPROXIMATIONS FOR A SYSTEM OF INFINITE SERVERS WITH LOAD BALANCING

PHILIP J. FLEMING
Motorola, Inc.
1501 West Shure Drive
Arlington Heights, Illinois 60004
e-mail: fleming@cig.mot.com

BURTON SIMON
Department of Mathematics
University of Colorado at Denver
Denver, Colorado 80217
e-mail: bsimon@math.cudenver.edu

We consider an exponential queueing system with multiple stations, each of which has an infinite number of servers and a dedicated arrival stream of jobs. In addition, there is an arrival stream of jobs that choose a station based on the state of the system. In this paper we describe two heavy traffic approximations for the stationary joint probability mass function of the number of busy servers at each station. One of the approximations involves state-space collapse and is accurate for large traffic loads. The state-space in the second approximation does not collapse. It provides an accurate estimate of the stationary behavior of the system over a wide range of traffic loads.

1. INTRODUCTION

We consider a queueing system which has k stations. Each station has an infinite number of servers and all jobs have independent, identical, exponentially distributed service times. Each station has a dedicated Poisson arrival stream, and, in addition to these dedicated arrival streams, there is another Poisson arrival stream of jobs to the system we call “smart” jobs, which are routed to stations based on the number of

0269-9648/99 $12.50

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

busy servers at each station. Let the positive real numbers $\alpha_1, \dots, \alpha_k$ denote weights, and let Q_i denote the number of busy servers at station i. The smart jobs join station i if $\alpha_i Q_i < \alpha_j Q_j, j \neq i$, when they arrive and choose (as a convention) the station with the smallest index among the best if there is a tie. (The tie-breaking rule does not matter in our analysis.) The system is therefore a continuous-time Markov chain that can be solved either numerically or by simulation. We refer to the exact solution as the Markov chain solution.

In the case $k = 1$, the model reduces to the $\mathrm{M}/\mathrm{M}/\infty$ system, and it is well known that the stationary number of busy servers is a Poisson random variable. It is also well known that the sample paths converge weakly to an Ornstein–Uhlenbeck process, $U(t)$, under a suitable scaling, as the arrival rate increases (Iglehart [5]). More precisely, let $Q^\lambda(t)$ denote the number of busy servers at time t in an $\mathrm{M}/\mathrm{M}/\infty$ system with arrival rate λ and service rate one. If we define

$$X^\lambda(t) = \frac{Q^\lambda(t) - \lambda}{\sqrt{\lambda}},$$

then $X^\lambda(t) \to_d U(t)$, where $\to_d$ denotes convergence in distribution. Furthermore, as $t \to \infty$, $U(t)$ converges in distribution to a standard Normal random variable, χ, and if we define X^λ to be the stationary random variable associated with $X^\lambda(t)$, then as $\lambda \to \infty$, X^λ also converges in distribution to χ. Thus, the following diagram holds:

$$\begin{tikzcd}
& U(t) \arrow[dr, "d"] & \\
X^\lambda(t) \arrow[ur, "d"] \arrow[dr, "d"] & & \chi . \\
& X^\lambda \arrow[ur, "d"] &
\end{tikzcd} \tag{1.1}$$

The heavy traffic analysis therefore implies that for sufficiently large λ we can use $\lambda + \sqrt{\lambda}\chi$ as an approximation to the stationary random variable, Q^λ. This approximation is quite accurate, even for relatively small values of λ. For example, we obtain 18, 20, and 21 as the 95th, 98th, and 99th percentiles, respectively, when $\lambda = 12$, which are all exact. Although this example is anecdotal, it indicates that this type of approximation may be useful when computing the capacity of physical systems with quality constraints in an appropriate range.

Using the above analysis of the $\mathrm{M}/\mathrm{M}/\infty$ queueing system as a guide, this paper is concerned with the asymptotic behavior of the joint distribution of the number of busy servers, in the case $k > 1$, as the overall arrival rate approaches infinity. We present two approximations for the transient and stationary joint distributions of the number of busy servers at each station. These approximations are based on diffusion processes obtained by taking limits of drift and variance parameters of a sequence of Markov chains. We do not prove weak convergence, so we view our diffusions as "conjectured limits." The conjectures, however, are consistent with the known limits of similar systems under similar scalings (Iglehart [5], Borovkov [2], Whitt [10],

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

Krichagina and Puhalskii [6], Turner [8]), and our numerous simulations and numerical experiments are entirely consistent with the implications of the conjectured limits. A certain discontinuity in the infinitesimal drift function for the diffusion (derived in Section 2) prevents us from directly proving the weak convergence from well-known theorems. After deriving the drift and variance for the limiting diffusion process we solve the differential equation for the stationary density function (equation (2.22)). We then show how to use statistics from (2.22) to estimate performance measures associated with real systems by "reversing the scaling." Our approximations are very accurate over wide ranges of parameter settings, including systems that are only "moderately" loaded.

Let β be a parameter that we use to scale the arrival rates. For convenience in the following discussion we set the arrival rate of customers dedicated to station i to $\lambda_i = \beta$, and fix $\alpha_i = 1$, $i = 1,2,\ldots,k$, so that smart customers join the station with the fewest customers (this will be relaxed in Section 2). Our two approximations correspond to two ways to scale the arrival rate λ_0 of the smart customers, with respect to the arrival rate of the dedicated customers. Let $p > 0$ be a fixed constant and let

$$Y^\beta(t) = \frac{Q^\beta(t) - (1 + p/k)\beta e}{\sqrt{\beta}},$$

where $Q^\beta(t)$ is the (row) vector of queue lengths at time t, and $e = (1,1,\ldots,1)$. When $\lambda_0 = p\beta$ we argue that $Y^\beta(t)$ collapses to a one-dimensional diffusion process, $\left(\sqrt{k + p/k}\right)U(t)e$ as $\beta \to \infty$. In this case, the intensity of λ_0 is high enough to keep the queue lengths at the individual stations approximately equal. Our second approximation sets $\lambda_0 = p\sqrt{\beta}$, and we argue that, as $\beta \to \infty$,

$$X^\beta(t) = \frac{Q^\beta(t) - \beta e}{\sqrt{\beta}}$$

converges in distribution to a k-dimensional diffusion process, $X(t)$, with a "piecewise Normal" stationary density, given by

$$f(x_1, x_2, \ldots, x_k) = C \exp\left(-\frac{1}{2}\sum_{i=1}^{k} x_i^2 + p \min(x_1, \ldots, x_k)\right),$$

where C is a normalizing constant. The main focus of the paper is on the latter scaling, which turns out to offer superior accuracy over most settings of the system parameters.

Our model has applications in systems with "soft" capacity such as Code Division Multiple Access (CDMA) cellular systems (Viterbi [9]), in which signal quality degrades as users are added to the system. In this case, the model represents a single cell with multiple wideband radio channels in a multiple cell system. The dedicated jobs represent soft hand-off requests from neighboring cells and smart jobs represent new call requests. The following is an explanation of the notion of soft handoff in a CDMA system. A station in the model represents a radio carrier that carries many individual phone calls (or data sessions) on channels determined by individual

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

codes. A mobile user stays on the same radio carrier as it moves through the various cells in its path. It can be communicating with many different cells simultaneously. This is referred to as "being in soft handoff" with several cells. As it moves through the system it adds and drops these links to cells. A soft hand-off request is performed each time the mobile can benefit from adding a link in another cell. The system does not have the freedom to assign the soft hand-off request to any radio carrier because of the large benefit of soft handoff (due to the gain in signal processing derived from macrodiversity) and risk of dropping the call in an intercarrier (or hard) handoff. On the other hand, requests for a channel from new calls can be assigned to any radio carrier. Although each carrier has a finite capacity (which may vary from carrier to carrier), its capacity is typically determined by interference, not a hard equipment limit, as in classic blocking systems, making an indefinite number of servers per radio carrier a realistic model. Reducing the variance in the number of links assigned to each radio carrier reduces the time during which interference is high. When the carriers have different capacities and/or different dedicated loads, the weighted least-load algorithm would be used.

Other applications are given in Alanyali and Hajek [1] and Hariharan, Kulkarni, and Stidham [4]. In these types of systems, a routing scheme that attempts to minimize the maximum number of users on a transmission device will improve each user's quality of service. In general, the least-load algorithm can be used when one is attempting to minimize a convex cost function of the number of busy servers at each station. Our model can be applied to systems with hard blocking by the method of truncation as in an Erlang-B type system.

In the next section we derive the asymptotics for the two scalings of our systems just described and then for a more general (asymmetric) model, where the arrival rates of the dedicated customers have the form $\lambda_i = \alpha_i^{-1}\beta + \gamma_i\sqrt{\beta}$, $i = 1,2,\ldots,k$. In Section 3, we show how the stationary density of the diffusions can be used to approximate the marginal means, variances, and quantiles, as well as joint statistics for the Markov chain, Q^{β}. In Section 4, we present closed form expressions for our approximations of the marginal mean, variance, and density in the case $k = 2$, and a discussion of an efficient Monte-Carlo importance sampling technique for estimating quantities of interest when $k > 2$. In Section 5, we present numerical work indicating the accuracy of the approximations. We summarize our results and conjectures in Section 6.

2. DERIVATION OF THE ASYMPTOTIC DISTRIBUTIONS

Consider the service system shown in Figure 1. We define a family of such service systems indexed by an "arrival rate scaling parameter," β. We assume for notational convenience (and without loss of generality) that the service rate of each server in the system is unity. Let $Q^{\beta}(t)$ be the row vector of queue lengths at time t for the βth system. Clearly $Q^{\beta}(t)$ is a positive-recurrent, continuous-time Markov chain, so

$$Q^{\beta}(t) \to_d Q^{\beta}$$

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

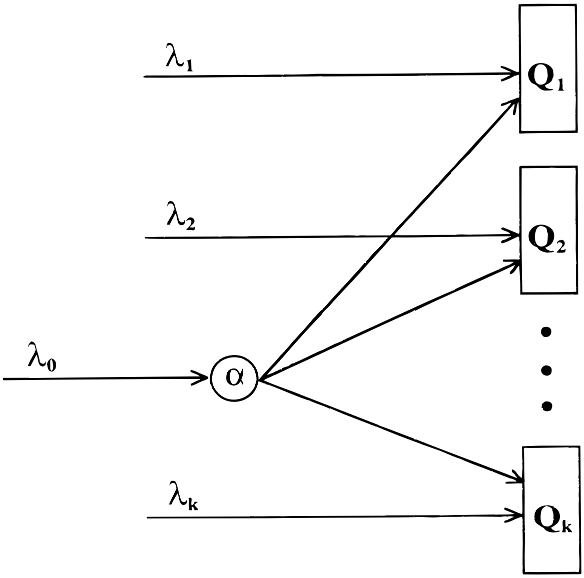

Figure 1. Join the weighted least-loaded station.

as $t \to \infty$, where Q^β is the stationary random vector. The distribution of Q^β can be readily calculated from the infinitesimal generator when β is small (with some truncation error). As β grows, the exact solution becomes increasingly difficult to compute, while at the same time, the asymptotic distributions as $\beta \to \infty$ become increasingly accurate. If β is large enough, we expect the asymptotic distribution to approximate the real system well.

We denote the system state by

$$x = (x_1, x_2, \ldots, x_k) \in \mathcal{R}^k,$$

where $\mathcal{R}^k$ is k-dimensional Euclidean space, x_i is the (scaled) number of customers at station i, and for a fixed set of positive "weights", $\alpha_1, \alpha_2, \ldots, \alpha_k$, we define

$$L_i = \{x : \alpha_i x_i < \alpha_j x_j, j \neq i\}, \qquad i = 1, 2, \ldots, k.$$

Let

$$L^* = \mathcal{R}^k - \bigcup_{i=1}^{k} L_i$$

be those x where there is no unique minimum to the set

$$S(x) = \{\alpha_1 x_1, \alpha_2 x_2, \ldots, \alpha_k x_k\},$$

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

and let L be the line

$$L = \{x : \alpha_1 x_1 = \alpha_2 x_2 = \cdots = \alpha_k x_k\}. \tag{2.1}$$

Finally, define

$$\ell \equiv \ell(x) = \begin{cases} i & \text{if } x \in L_i \\ \min\{j \in \operatorname{argmin} S(x)\} & \text{if } x \in L^* \end{cases} \tag{2.2}$$

to be the index of the station smart jobs join when the system is in state x. Let $\gamma_1, \gamma_2, \ldots, \gamma_k$ be fixed constants. The arrival rate of dedicated customers to the ith station has the form

$$\lambda_i = \alpha_i^{-1}\beta + \gamma_i\sqrt{\beta}, \qquad i = 1,2,\ldots,k.$$

The arrival rate of smart customers is also a function of β. We will consider the cases $\lambda_0 = p\sqrt{\beta}$ and $\lambda_0 = p\beta$, where $p > 0$ is a fixed constant. Note that, for any "real system" with arrival rates λ_0 and $\lambda_1, \ldots, \lambda_k$ and weights $\alpha_1, \ldots, \alpha_k$, there are settings of the constants $\gamma_1, \ldots, \gamma_k$ and p (not unique) so that for some $\beta > 0$, the βth system matches the real system. (In our numerical examples we use Eqs. (3.11), (3.12), and (3.13) as an assignment of the constants.)

We begin by considering the symmetric case, $\alpha_i = 1$ and $\gamma_i = 0$, $i = 1,2,\ldots,k$, with $\lambda_0 = p\sqrt{\beta}$. Define

$$X^\beta(t) = \frac{Q^\beta(t) - \beta e}{\sqrt{\beta}}, \tag{2.3}$$

where

$$e \equiv (1,1,\ldots,1)$$

is a row vector of k ones. As a convention, we set $Q^\beta(0) = \lfloor \beta \rfloor e$. We conjecture that

$$X^\beta(t) \to_d X(t) \tag{2.4}$$

as $\beta \to \infty$, where $X(t)$ is some k-dimensional diffusion process. In view of results proved by Iglehart [5], Borovkov [2], Whitt [10], Krichagina and Puhalskii [6], and Turner [8], the conjecture is not surprising. The proof in our case is made difficult, however, because the drift vector, which we derive next, is not continuous on the set L^*.

Consider the change in $Q^\beta(\cdot)$ in the small time interval $(t, t+h)$ when $Q^\beta(t) = (n_1, n_2, \ldots, n_k)$. The number of dedicated arrivals to station i in that interval has a Poisson distribution with mean βh. Likewise, to $o(h)$ (where $o(h)/h \to 0$ as $h \to 0$), the mean and variance of the number of departures from station i in the interval will

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

match those of a Poisson random variable with mean $n_i h$. The covariance of the change in the queue lengths at station i and j will also be $o(h)$ when $i \neq j$. Let

$$x_i = \frac{n_i - \beta}{\sqrt{\beta}}, \qquad i = 1,2,\ldots,k$$

be the scaled queue lengths. We therefore have

$$\begin{aligned} d_i^\beta(x) &\equiv \lim_{h\to 0} h^{-1} E(X_i^\beta(t+h) - X_i^\beta(t) | X_i^\beta(t) = x) \\ &= p1_{(i=\ell)} - x_i \end{aligned} \tag{2.5}$$

and

$$\begin{aligned} A_{ij}^\beta(x) &\equiv \lim_{h\to\infty} h^{-1} \operatorname{Cov}(X_i^\beta(t+h) - X_i^\beta(t), X_j^\beta(t+h) - X_j^\beta(t) | X^\beta(t) = x) \\ &= \begin{cases} 2 + \beta^{-1/2}(p1_{(i=\ell)} + x_i) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \end{aligned} \tag{2.6}$$

where ℓ is the index of the station smart customers join, given by Eq. (2.2), and $1_{(\cdot)}$ is the indicator function. It follows that, if $X(t)$ is a diffusion process, then its infinitesimal drift and covariance functions are

$$d(x) = p\delta_\ell - x \tag{2.7}$$

and

$$A(x) \equiv A = 2I, \tag{2.8}$$

respectively, where

$$\delta_\ell = (0,\ldots,1,\ldots,0)$$

is the ℓth unit basis vector, and I is the identity matrix. Furthermore, $X(t) \to_d X$ as $t \to \infty$, and the density of X satisfies the equation

$$\begin{aligned} \sum_{i=1}^k \frac{\partial}{\partial x_i}[d_i(x) f(x)] &= \frac{1}{2}\sum_{i=1}^k \sum_{j=1}^k \frac{\partial^2}{\partial x_i \partial x_j}[A_{ij}(x) f(x)] \\ &= \sum_{i=1}^k \frac{\partial^2}{\partial x_i^2} f(x) \end{aligned}$$

(Gardiner [3]), which has solution

$$f(x) = C\exp\left(-\frac{1}{2}\sum_{i=1}^k x_i^2 + px_\ell\right), \tag{2.9}$$

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

where C is a normalizing constant. Exploiting the symmetry in this case, the normalizing constant has the form

$$C^{-1} = k\int_{-\infty}^{\infty}\int_{x_1}^{\infty}\cdots\int_{x_1}^{\infty} f(x_1,x_2,\ldots,x_k)\,dx_k\,dx_{k-1}\cdots dx_1$$
$$= k\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}u^2+pu\right)\Psi(u)^{k-1}\,du \qquad \textbf{(2.10)}$$

where

$$\Psi(u) \equiv \int_u^{\infty}\exp\left(-\frac{1}{2}z^2\right)dz.$$

Let e' denote a column vector of ones. We know that $Q^{\beta}(t)e'$ is the number of customers in an M/M/∞ system with arrival rate $k\beta + p\sqrt{\beta}$ at time t, so from Eq. (1.1) we have

$$\frac{Q^{\beta}(t)e' - \left(k\beta + p\sqrt{\beta}\right)}{\sqrt{k\beta + p\sqrt{\beta}}} \to_d U(t)$$

as $\beta \to \infty$. Thus,

$$X^{\beta}(t)e' \to_d \sqrt{k}U(t) + p,$$

and by symmetry we expect the density of $X^{\beta}(t)$ to be centered at $(p/k)e$ as $\beta \to \infty$. Define

$$\hat{X}_p^{\beta}(t) = X^{\beta}(t) - \frac{p}{k}e.$$

Clearly, Eq. (2.4) implies that $\hat{X}_p^{\beta}(t) \to_d \hat{X}_p(t)$ as $\beta \to \infty$, where $\hat{X}_p(t)$ has stationary density

$$\hat{f}_p(x) = f\left(x + \frac{p}{k}e\right),$$

which is centered at the origin. Let $\hat{X}_p$ be the stationary random vector associated with $\hat{X}_p(t)$. Since $\hat{X}_p$ and X differ by a constant vector, we can bound the distance between $\hat{X}_p$ and the line $x_1 = x_2 = \cdots = x_k$ (which is L in the symmetric case) by

$$\Delta_p = \max_{1\le i\le k} X_i - \min_{1\le i\le k} X_i.$$

Choose $\epsilon > 0$ and use Eqs. (2.9) and (2.10) to write

$$P(\Delta_p > \epsilon) = k(k-1)\int_{-\infty}^{\infty}\int_{x_1+\epsilon}^{\infty}\int_{x_1}^{x_2}\cdots\int_{x_1}^{x_2} f_p(x_1,\ldots,x_k)\,dx_k\ldots dx_1$$
$$< \frac{(k-1)\displaystyle\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}u^2+pu\right)\Psi(u+\epsilon)\Psi(u)^{k-2}\,du}{\displaystyle\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}u^2+pu\right)\Psi(u)^{k-1}\,du}. \qquad \textbf{(2.11)}$$

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

We write Eq. (2.11) in the form

$$P(\Delta_p > \epsilon) < \frac{a_1(p) + a_2(p)}{b_1(p) + b_2(p)}, \tag{2.12}$$

where

$$a_1(p) = (k-1)\int_{-\infty}^{-p+1-\epsilon/k} \exp(-u^2/2)\Psi(u+p+\epsilon)\Psi(u+p)^{k-2}\,du,$$

$$a_2(p) = (k-1)\int_{-p+1-\epsilon/k}^{\infty} \exp(-u^2/2)\Psi(u+p+\epsilon)\Psi(u+p)^{k-2}\,du,$$

$$b_1(p) = \int_{-\infty}^{-p+1} \exp(-u^2/2)\Psi(u+p)^{k-1}\,du,$$

and

$$b_2(p) = \int_{-p+1}^{\infty} \exp(-u^2/2)\Psi(u+p)^{k-1}\,du.$$

One can verify that for any $r > 0$ we have

$$\frac{1}{r}\exp\left(-\frac{1}{2}(r+r^{-1})^2\right) < \Psi(r) < \frac{1}{r}\exp\left(-\frac{1}{2}r^2\right),$$

so that for $r > 1$ and $0 < s < r$,

$$\frac{\Psi(r+s)}{\Psi(r)} \le c_1 \exp(-rs) \tag{2.13}$$

and

$$\frac{\Psi(r-s)}{\Psi(r)} \le c_2 \exp(rs), \tag{2.14}$$

where c_1 and c_2 are constants. It follows from Eqs. (2.13) and (2.14) that $a_1(p)/b_1(p) \to 0$ and $a_2(p)/b_2(p) \to 0$ as $p \to \infty$, and since all four terms in (2.12) are positive we conclude that

$$P(\Delta_p > \epsilon) \to 0. \tag{2.15}$$

Combining Eqs. (1.1) and (2.15) yields

$$\hat{X}_p \to_d k^{-1/2}\chi e \tag{2.16}$$

as $p \to \infty$. In other words, if there are enough smart customers, the k queue lengths will be essentially the same in steady state. This phenomenon is known as "state-space collapse" (Reiman [7]).

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

The limit (2.16) suggests that if we let β and p go to infinity together we may encounter space collapse. In particular, if we define

$$Y^{\beta}(t) \equiv \hat{X}^{\beta}_{p\beta^{1/2}}(t) = \frac{Q^{\beta}(t) - \left(1 + \dfrac{p}{k}\right)\beta e}{\sqrt{\beta}}, \tag{2.17}$$

then we conjecture that $Y^{\beta}(t)$ converges in distribution to a diffusion on the line L. Since $Y^{\beta}(t)e'$ is the number of customers in an M/M/∞ queue with arrival rate $(k+p)\beta$, we conclude from (1.1) that

$$Y^{\beta}(t) \rightarrow_d \frac{\sqrt{k+p}}{k}\, U(t)e \tag{2.18}$$

as $\beta \to \infty$; therefore the stationary random variable is

$$Y = \frac{\sqrt{k+p}}{k}\, \chi e.$$

Another heuristic explanation of the state space collapse (2.18) is to consider the drift function (2.5) when $\lambda_0 = p\beta$. One finds that if x is not on L then the drift toward L has magnitude $O(\sqrt{\beta})$, so that the limiting process cannot leave the line.

We now consider the general model, specified by weights $\alpha_i > 0$, $i = 1,2,\ldots,k$, and constants γ_i, $i = 1,2,\ldots,k$. The arrival rate of dedicated customers to station i is $\lambda_i = \alpha_i^{-1}\beta + \gamma_i\sqrt{\beta}$. The factor of α_i^{-1} is chosen so that, as $\beta \to \infty$, the smart customers will find the stations "comparable" in size, i.e., $P(\ell = i) > 0$, $i = 1,2,\ldots,k$.

We first consider the case $\lambda_0 = p\sqrt{\beta}$. For this scaling we define

$$\tilde{\alpha} = (\alpha_1^{-1}, \alpha_2^{-1}, \ldots, \alpha_k^{-1}),$$

and set

$$X^{\beta}(t) = \frac{Q^{\beta}(t) - \tilde{\alpha}\beta}{\sqrt{\beta}}. \tag{2.19}$$

Similar to (2.7) and (2.8) we obtain a drift function

$$d(x) = p\delta_{\ell} + \gamma - x \tag{2.20}$$

and covariance function

$$A(x) \equiv A = \begin{bmatrix} 2\alpha_1^{-1} & & \\ & \ddots & \\ & & 2\alpha_k^{-1} \end{bmatrix}. \tag{2.21}$$

We therefore conjecture that $X^{\beta}(t) \rightarrow_d X(t)$ as $\beta \to \infty$, where $X(t)$ is the diffusion specified by Eqs. (2.20) and (2.21). The stationary density of $X(t)$ satisfies

$$\begin{aligned} \sum_{i=1}^{k} \frac{\partial}{\partial x_i}[d_i(x) f(x)] &= \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k} \frac{\partial^2}{\partial x_i \partial x_j}[A_{ij}(x) f(x)] \\ &= \sum_{i=1}^{k} \alpha_i^{-1} \frac{\partial^2}{\partial x_i^2} f(x), \end{aligned}$$

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

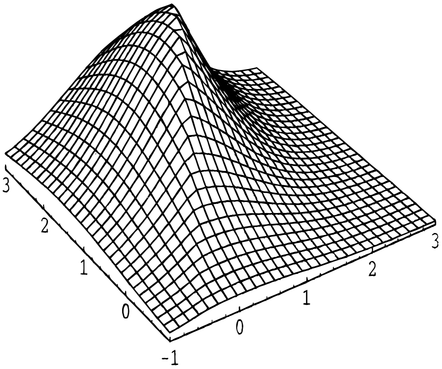

FIGURE 2. The two-dimensional joint density with $p = 1$, $\alpha = 0.5$, and $\gamma = 1.5$.

which has the solution

$$f(x) = C \exp\left(-\frac{1}{2}\sum_{j=1}^{k} \alpha_j (x_j - \gamma_j)^2 + p\alpha_\ell x_\ell\right). \tag{2.22}$$

The density $f(x)$ in Eq. (2.22) is "piecewise Normal" in the sense that, on each $L_i \subset \mathcal{R}^k$, $i = 1, 2, \ldots, k$, it is proportional to a Normal density function. The density is not differentiable on L^* (see Fig. 2) but is continuous and solves the stationary Fokker-Planck equation. In the symmetric case, (2.22) reduces to (2.9).

We next consider the scaling $\lambda_0 = p\beta$ in the general model. In the absence of smart customers, station i typically has $\alpha_i^{-1}\beta + \mathcal{O}(\sqrt{\beta})$ customers, so by redoing (2.5) we find that the drift toward L is $\mathcal{O}(\sqrt{\beta})$ and, as before, we conjecture that the smart customers cause the system state to collapse to L, which, this time, is the line through the origin in the direction of the vector $\tilde{\alpha}$. From Eq. (1.1) we see that

$$Y^\beta(t)e' = \frac{Q^\beta(t)e' - (p + \tilde{\alpha}e')\beta}{\sqrt{\beta}} \to_d \sqrt{p + \tilde{\alpha}e'}\,U(t),$$

and the state space collapse to L implies that

$$Y^\beta(t) \equiv \frac{Q^\beta(t) - (\tilde{\alpha}e')^{-1}(p + \tilde{\alpha}e')\beta\tilde{\alpha}}{\sqrt{\beta}} \to_d \frac{\sqrt{p + \tilde{\alpha}e'}}{\tilde{\alpha}e'}\,U(t)\tilde{\alpha}. \tag{2.23}$$

The stationary random variable is therefore

$$Y = \frac{\sqrt{p + \tilde{\alpha}e'}}{\tilde{\alpha}e'}\,\chi\tilde{\alpha}. \tag{2.24}$$

3. APPROXIMATIONS FOR THE MOMENTS AND QUANTILES OF Q^β

In this section we show how the heavy traffic limits can be used to derive approximations for the general queueing model, by "reversing the scaling." We are in-

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

terested in statistics associated with the stationary random vector Q^β of the continuous-time Markov chain $Q^\beta(t)$, when $\beta < \infty$. It is important to see that there are an infinite number of scalings of the parameters that yield the same limiting processes, and different scalings typically yield different approximations. Thus, any approximation for a quantity associated with Q^β based solely on X or Y is necessarily heuristic. We argue that the following approximations are in some sense "natural" choices, although in some cases the natural choice was not obvious until numerical experiments were performed.

We begin by using Eq. (2.24) to approximate the queueing model. In that case, due to the state space collapse, the Y_i's are always in the same ratios. Let $\lambda = \sum_{i=0}^k \lambda_i = (p + \tilde{\alpha}e')\beta + \gamma e'\sqrt{\beta}$ denote the total arrival rate to the system. Using Eqs. (2.23) and (2.24) and the observation that $\sum_{i=1}^k Q_i^\beta$ is a Poisson random variable with mean λ, the resulting approximation is

$$Q^\beta \approx ((\tilde{\alpha}e')^{-1}\lambda + (\tilde{\alpha}e')^{-1}\sqrt{\lambda}\chi)\tilde{\alpha}, \tag{3.1}$$

so that

$$E(Q^\beta) \approx (\tilde{\alpha}e')^{-1}\lambda\tilde{\alpha}, \tag{3.2}$$

and

$$\mathrm{Var}(Q_i^\beta) \approx (\tilde{\alpha}e')^{-2}\lambda\alpha_i^{-2}, \qquad i = 1,2,\ldots,k. \tag{3.3}$$

It turns out that this approximation for Q^β is accurate when λ_0 is large. However, the approximation based on Eq. (2.22) is typically more accurate than that for Eq. (3.1), even for large values of λ_0 and far more accurate for small values of λ_0, so we turn our attention to that case.

Let X^β be the stationary random vector associated with $X^\beta(t)$ given by Eq. (2.19). Thus,

$$Q^\beta = \tilde{\alpha}\beta + X^\beta\sqrt{\beta}.$$

This formula suggests the approximation

$$E(Q^\beta) \approx \tilde{\mu}^\beta, \tag{3.4}$$

where

$$\tilde{\mu}^\beta = \tilde{\alpha}\beta + \hat{\mu}\sqrt{\beta}, \tag{3.5}$$

and

$$\hat{\mu} = E(X).$$

The approximation is exact in the symmetric model, in which case $\hat{\mu} = (p/k)e$. In the general model, $\hat{\mu}$ must be evaluated from the density function (2.22). This can be done explicitly when $k = 2$, or by an efficient numerical technique when $k > 2$. These calculations are discussed in the next section. To approximate the marginal variance, let

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

$$\sigma_i^2(\beta) = \mathrm{Var}(Q_i^\beta)$$

and

$$\hat{\sigma}_i^2 = \mathrm{Var}(X_i).$$

Again, $\hat{\sigma}_i^2$ must be evaluated from Eq. (2.22). From our scaling we see that

$$\sigma_i^2(\beta) = c_i(\beta)\hat{\sigma}_i^2, \tag{3.6}$$

for some $c_i(\beta)$ satisfying $c_i(\beta)/\beta \to 1$ as $\beta \to \infty$. When $p = 0$ we have $\sigma_i^2(\beta) = E(Q_i^\beta)$ because the stationary queue length distributions are Poisson. Continuing with the case $p = 0$, it follows from Eq. (2.22) that $\hat{\sigma}_i^2 = \alpha_i^{-1}$. Thus, we have $c_i(\beta) = \alpha_i E(Q_i^\beta)$. Using Eqs. (3.5) and (3.6) and the $p = 0$ case as a guide, we have the approximation

$$\sigma_i^2(\beta) \approx \alpha_i \tilde{\mu}_i^\beta \hat{\sigma}_i^2. \tag{3.7}$$

To obtain quantiles we need to approximate $P(Q_i^\beta \le n)$. We seek $\xi_i^\beta > 0$ and ζ_i^β such that

$$P(Q_i^\beta \le n) \approx \int_{-\infty}^{n+1/2} \xi_i^\beta f_i(\xi_i^\beta x + \zeta_i^\beta)\,dx, \tag{3.8}$$

where $f_i(x)$ is the marginal density of X_i. (Because $\xi_i^\beta > 0$, the integrand in approximation (3.8) is a density function.) We therefore have

$$E(Q_i^\beta) \approx \int_{-\infty}^{\infty} x\xi_i^\beta f_i(\xi_i^\beta x + \zeta_i^\beta)\,dx = (\xi_i^\beta)^{-1}(\hat{\mu}_i - \zeta_i^\beta)$$

and

$$\mathrm{Var}(Q_i^\beta) \approx \int_{-\infty}^{\infty} (x - \xi_i^{\beta-1}(\hat{\mu}_i - \zeta_i^\beta))^2 \xi_i^\beta f_i(\xi_i^\beta x + \zeta_i^\beta)\,dx = (\xi_i^\beta)^{-2}\hat{\sigma}_i^2.$$

Using (3.5) and (3.7), we choose

$$\xi_i^\beta = (\alpha_i \tilde{\mu}_i)^{-1/2} \tag{3.9}$$

and

$$\zeta_i^\beta = \hat{\mu}_i - \tilde{\mu}_i \xi_i^\beta. \tag{3.10}$$

The joint distribution can be approximated in the same way by

$$\begin{aligned} &P(Q_i^\beta \le n_i,\, i = 1,2,\ldots,k) \\ &\quad\approx \left(\prod_{i=1}^{k} \xi_i^\beta\right) \int_{-\infty}^{n_1+1/2} \cdots \int_{-\infty}^{n_k+1/2} f(\xi_1^\beta x_1 + \zeta_1^\beta, \ldots, \xi_k^\beta x_k + \zeta_k^\beta)\,dx. \end{aligned}$$

Given a general model specified by $\lambda_0, \lambda_1, \dots, \lambda_k$, and $\alpha_1, \dots, \alpha_k$, the value of the parameters $p, \beta, \gamma_1, \dots, \gamma_k$, are not completely determined. In our numerical work we have chosen the following scheme based on setting $\gamma_1 = 0$.

$$\beta = \alpha_1 \lambda_1, \tag{3.11}$$

$$\gamma_i = \frac{\lambda_i - \alpha_i^{-1}\alpha_1\lambda_1}{\sqrt{\alpha_1\lambda_1}}, \qquad i = 1,2,\dots,k, \tag{3.12}$$

$$p = \frac{\lambda_0}{\sqrt{\alpha_1\lambda_1}}. \tag{3.13}$$

Of course, other schema are possible. For example, if one requires $\sum \gamma_i = 0$, some algebraic manipulations are easier. There does not seem to be a compelling reason to choose one over another, however.

4. CALCULATING STATISTICS FROM $f(x)$

The approximations based on X in the previous section require $\hat{\mu}_i = E(X_i)$ and $\hat{\sigma}_i^2 = \mathrm{Var}(X_i)$, $i = 1,2,\dots,k$, which must be obtained from $f(x)$. When $k = 2$, most quantities of interest can be computed in closed form in terms of

$$\Phi(z) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{t^2}{2}\right) dt.$$

For notational convenience we set $\gamma_1 = 0$, $\gamma_2 = \gamma$, $\alpha_1 = 1$, and $\alpha_2 = \alpha$. In this case, the joint density is given by

$$f(x_1,x_2) = \begin{cases} C_2^{-1}\exp(-\frac{1}{2}((x_1 - p)^2 + \alpha(x_2 - \gamma)^2 \alpha p(p + 2\gamma))) & \text{if } x_1 \le \alpha x_2 \\ C_2^{-1}\exp(-\frac{1}{2}(x_1^2 + \alpha(x_2 - \gamma - p)^2 + p^2)) & \text{if } x_1 > \alpha x_2 \end{cases}. \tag{4.1}$$

We present expressions for the normalizing constant C_2 and the first and second moments of X_1 and X_2 and their marginal densities. Define

$$r_1 = \frac{p - \gamma\alpha}{\sqrt{1+\alpha}}, \quad r_2 = \frac{\alpha(p+\gamma)}{\sqrt{1+\alpha}}.$$

A straightforward integration of Eq. (4.1) yields

$$C_2 = \frac{2\pi}{\sqrt{\alpha}}(\exp(-\alpha p(p + 2\gamma)/2)\Phi(-r_1) + \exp(-p^2/2)\Phi(-r_2)). \tag{4.2}$$

One can derive Eq. (4.2) by noticing that (4.1) is "piecewise normal" (Fig. 2). On the set $L_1 = \{(x_1,x_2) | x_1 < \alpha x_2\}$, Eq. (4.1) is the normal density (up to a constant multiple) centered at the point (p,γ) with a diagonal covariance matrix where the diagonal is $\tilde{\alpha} = (1,\alpha^{-1})$. Similarly, on $L_2 = \{(x_1,x_2) | x_1 \ge \alpha x_2\}$, Eq. (4.1) is the normal density centered at $(0, p+\gamma)$ with the same covariance matrix. The quantities r_1 and r_2 are the respective distances from the centers of these normals to L, the line $x_1 = \alpha x_2$.

The expected values of X_1 and X_2 are given by

$$\begin{aligned}\sqrt{\alpha}C_2 E(X_1) &= 2\pi p \exp(-\alpha p(p+2\gamma)/2)\Phi(-r_1)\\ &\quad + \sqrt{\frac{2\pi}{1+\alpha}}\,(\exp(-(p^2+r_2^2)/2) - \exp(-(\alpha p(p+2\gamma)+r_1^2)/2))\end{aligned} \tag{4.3}$$

and

$$\begin{aligned}\sqrt{\alpha}C_2 E(X_2) &= 2\pi(\gamma \exp(-\alpha p(p+2\gamma)/2)\Phi(-r_1) + (\gamma+p)\exp(-p^2/2)\Phi(-r_2))\\ &\quad - \sqrt{\frac{2\pi}{1+\alpha}}\,(\exp(-(p^2+r_2^2)/2) - \exp(-(\alpha p(p+2\gamma)+r_1^2)/2)).\end{aligned} \tag{4.4}$$

The second moments of X_1 and X_2 can be obtained similarly. For the symmetric model, the formula for the marginal variance reduces to

$$\operatorname{Var}(X_i) = 1 + \frac{p^2}{4} - \frac{pe^{-p^2/4}}{4\sqrt{\pi}\Phi\big(-p/\sqrt{2}\big)}, \quad i = 1,2,$$

so that $\operatorname{Var}(X_i) = 1$ when $p = 0$ and $\lim_{p\to\infty} \operatorname{Var}(X_i) = \frac{1}{2}$.

The marginal densities of X_1 and X_2 are

$$\begin{aligned}f_1(x) &= C_2^{-1}\sqrt{2\pi\alpha^{-1}}\left(\exp(-(\alpha p(p+2\gamma)+(x-p)^2)/2)\Phi\left(-\frac{x-\alpha\gamma}{\sqrt{\alpha}}\right)\right.\\ &\qquad\left. + \exp(-(p^2+x^2)/2)\Phi\left(-\frac{x-\alpha(\gamma+p)}{\sqrt{\alpha}}\right)\right)\end{aligned} \tag{4.5}$$

and

$$\begin{aligned}f_2(x) &= C_2^{-1}\sqrt{2\pi}(\exp(-(p^2+\alpha(x-\gamma-p)^2)/2)\Phi(-\alpha x)\\ &\qquad + \exp(-(\alpha p(p+2\gamma)+(x-\gamma)^2)/2)\Phi(\alpha x - p)).\end{aligned}$$

When $k > 2$ numerical methods are required to evaluate quantities of interest from $f(x)$. For the symmetric model most quantities of interest can be computed easily from (single) integrals in terms of $\Phi(x)$. For example, integrating Eq. (2.10) by parts, the normalizing constant, C_k, for the density as given in Eq. (2.9) can be written

$$C_k^{-1} = p(\sqrt{2\pi})^k \int_{-\infty}^{\infty} e^{px}\Phi(-x)^k\,dx,$$

and, similarly, the second moment of X_i, $i = 1, \ldots, k$, can be written

$$C_k E[X_i^2] = \frac{p}{k} (\sqrt{2\pi})^k \int_{-\infty}^{\infty} (k + 1 + px) e^{px} \Phi(-x)^k \, dx, \qquad i = 1, \ldots, k.$$

These quantities can be readily obtained by numerical integration. Similar quantities of interest, for the general model, can also be reduced to single integrals, although the expressions are more complicated. To compute quantities associated with the general model we suggest a Monte Carlo importance sampling procedure based on sequence of i.i.d. k-dimensional Normal random variates with density

$$g(x) = \left((2\pi)^{-k} \prod_1^k \alpha_i \right)^{1/2} \exp\left(-\frac{1}{2} \sum_1^k \alpha_i (x_i - \gamma_i - p/k)^2 \right). \tag{4.7}$$

The computational burden is approximately linear in k due to the need to compute k psuedorandom numbers per sample (the sample variance does not appear to be sensitive to k).

For example, to evaluate the normalizing constant C in Eq. (2.22), we define $\tilde{f}(x)$ by

$$f(x) = C\tilde{f}(x),$$

and define

$$h(x) = \frac{\tilde{f}(x)}{g(x)}. \tag{4.8}$$

Then

$$C^{-1} = E(h(Z)),$$

where Z is a random variate with density $g(x)$. One can construct Normal densities that seem to match $f(x)$ more closely than $g(x)$, but we have found $g(x)$ to be quite satisfactory for calculating moments and quantiles for reasonable systems.

5. VALIDATION

In this section we compare the results obtained from the heavy traffic approximations X (referred to as DA2) and Y (referred to as DA1) to those obtained from simulations of the Markov chain model (referred to as MCS). In general, the MCS results have a relative standard deviation of at most 0.01. The tie-breaking rule in the MCS is to choose a station randomly among those whose (weighted) load is the same level. The approximations based on Y were obtained analytically. The results from X were also obtained analytically when $k = 2$, and were obtained using the Monte Carlo technique described in the previous section when $k > 2$. The Monte Carlo calculations also have a relative standard deviation of at most 0.01.

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

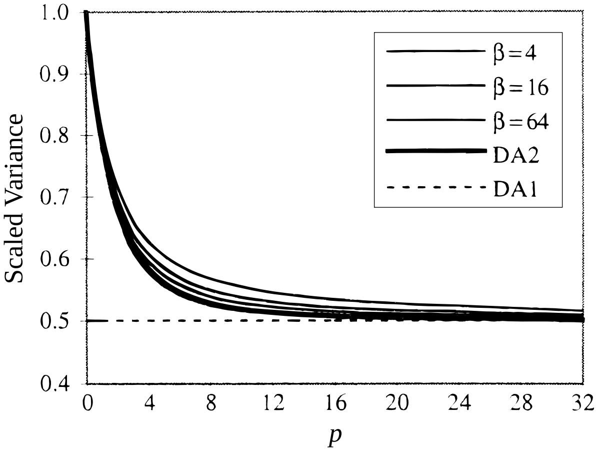

Figure 3. Scaled variance: $k = 2$.

The first set of charts (Figs. 3–8) focuses on the symmetric model, $\alpha_i = 1$, $\gamma_i = 0$, $i = 1,2,\ldots,k$. Note that in this case the means of the marginal queue length distributions are known to be λ/k, where $\lambda = \sum_{i=0}^{k} \lambda_i$ is the total arrival rate of customers to the system. Figures 3, 4, and 5 show a comparison of the marginal variance estimates (3.3) and (3.7) with the simulation for $k = 2$, 4, and 8. We have scaled the marginal variances by a factor of $(\lambda/k)^{-1}$ so that they are comparable. The figures

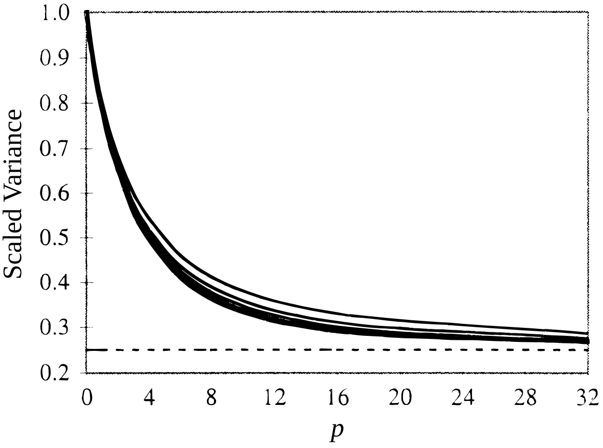

Figure 4. Scaled variance: $k = 4$.

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

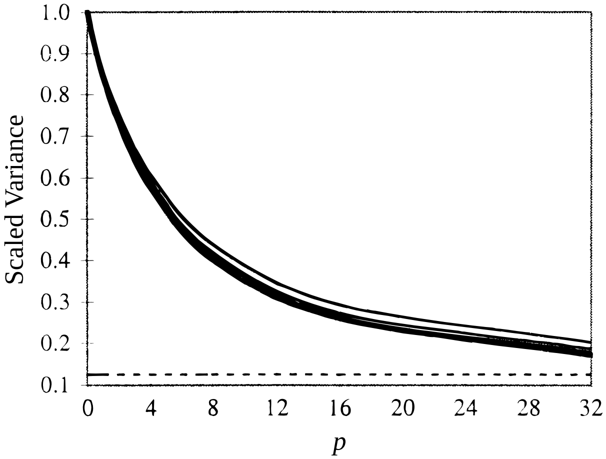

FIGURE 5. Scaled variance: $k = 8$.

each show several curves. The three curves with the largest value are obtained from the simulation (MCS) with $\beta = 4$, 16, and 64, and are shown in descending order (e.g., the scaled variance for systems with $\beta = 4$ is greater than that for $\beta = 16$). Note that, as expected, the larger β, the more accurate is (3.7), and the larger p, the more accurate is (3.3). It is also worth noting that as p gets large, (3.7) approaches (3.3). The accuracy of (3.7) for both small values of β and large values of p was un-

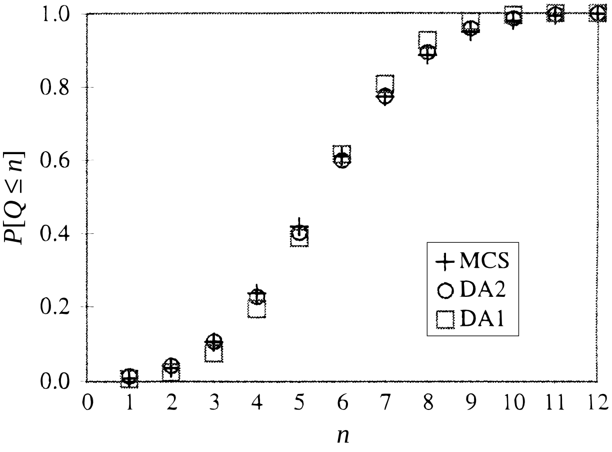

FIGURE 6. Distribution: $k = 2$, $\lambda = 12$.

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

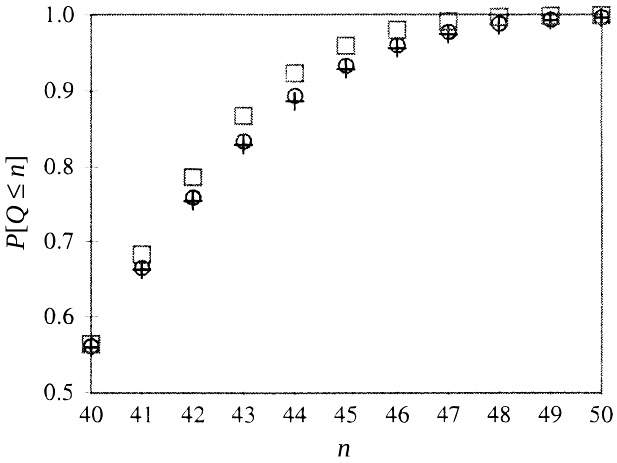

FIGURE 7. Distribution: $k = 4$, $\lambda = 160$.

expected. In addition, the accuracy does not seem to degrade as the number of stations gets large.

Figures 6, 7, and 8 show a comparison between the approximate marginal distributions derived from approximations (3.1) and (3.8) and the simulation for $k = 2$, 4, and 8. These results show a very close correspondence between (3.8) and the

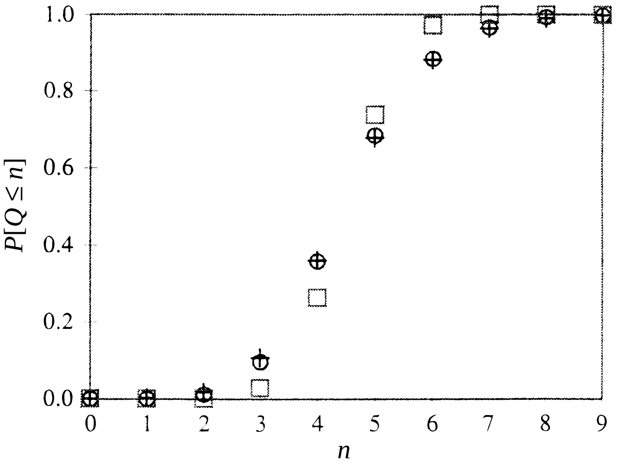

FIGURE 8. Distribution: $k = 8$, $\lambda = 40$.

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

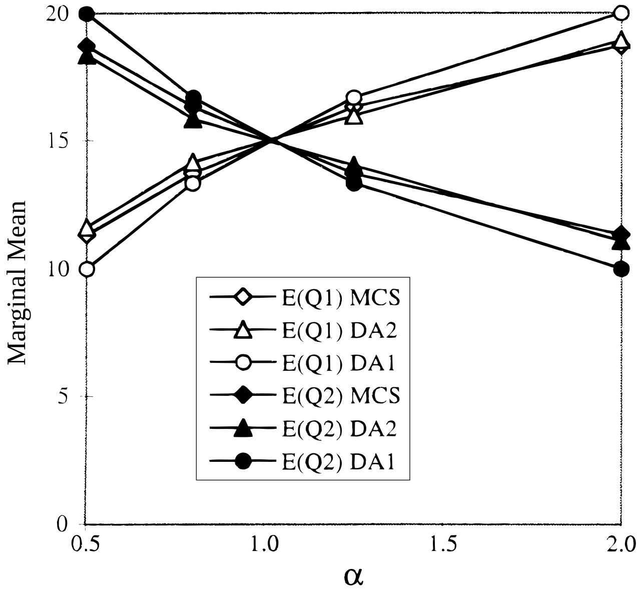

FIGURE 9. Marginal mean vs. α2 stations, $\lambda_0 = 12$, $\lambda_1 = 9$, $\lambda_2 = 9$.

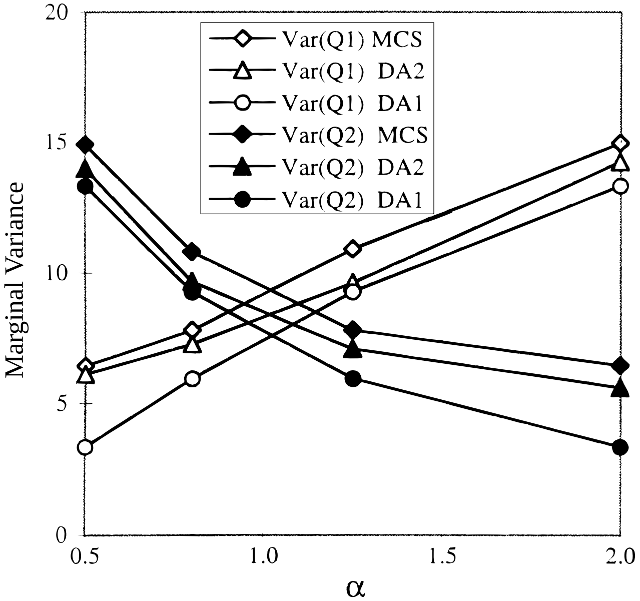

FIGURE 10. Marginal variance vs. α2 stations, $\lambda_0 = 12$, $\lambda_1 = 9$, $\lambda_2 = 9$.

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

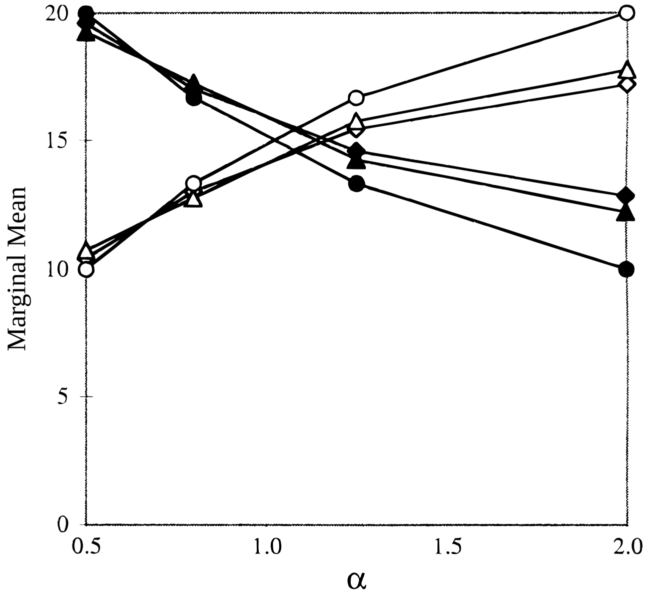

FIGURE 11. Marginal mean vs. α2 stations, $\lambda_0 = 12$, $\lambda_1 = 6$, $\lambda_2 = 12$.

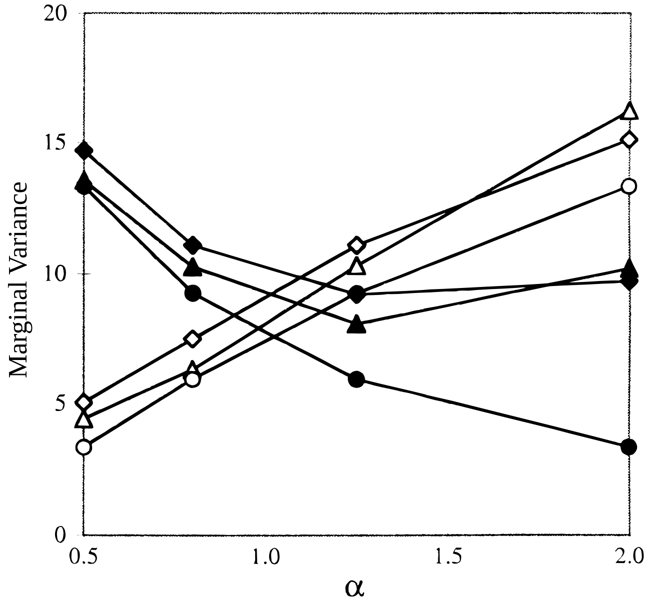

FIGURE 12. Marginal variance vs. α2 stations, $\lambda_0 = 12$, $\lambda_1 = 6$, $\lambda_2 = 12$.

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press

simulation, even for small values of λ (Figs. 6 and 8). For larger values of λ, (3.1) has improved accuracy, but (3.8) is still superior (e.g., Fig. 7).

The next set of charts (Figs. 9–12) focus on the general model in the case $k = 2$. We believe these are representative of the results for higher values of k. The legend for Figure 9 applies to Figure 11 as well, and the legend for Figure 10 also applies to Figure 12. Figures 9 and 10 show a comparison of marginal mean and marginal variance, respectively, for each station as a function of α and in a model in which $\lambda_0 = 12$ and $\lambda_1 = \lambda_2 = 9$. Figures 11 and 12 show a comparison of marginal mean and marginal variance, respectively, for each station as a function of α in a model in which $\lambda_0 = 12$, $\lambda_1 = 6$, and $\lambda_2 = 12$. These charts show that approximations (4.3)–(4.6) for these statistics derived from Eq. (2.22) can be very accurate. The state-space collapse approximations (3.2) and (3.3) are also shown for reference.

In our (unpublished) numerical investigations we have observed that the marginal densities obtained from Eq. (2.22) are well approximated by a Normal density with the same mean and variance. This could be convenient when using (2.22) in applications.

The validation results shown here are representative of many test cases the authors have investigated. As in the case of the classic heavy traffic M/M/∞ approximation (1.1), these results suggest that Eq. (2.22) is more than adequate for most engineering purposes.

6. SUMMARY OF TECHNICAL RESULTS

We have scaled the vector of queue lengths, $Q^\beta(t)$, in two ways, obtaining $X^\beta(t)$ given by Eq. (2.3) and $Y^\beta(t)$ given by Eq. (2.17). We have argued (in analogy with (1.1)) that the following diagram holds,

$$\begin{tikzcd}
 & X(t) \arrow[dr, "d"] & \\
X^\beta(t) \arrow[ur, "d"] \arrow[dr, "d"'] & & X, \\
 & X^\beta \arrow[ur, "d"'] &
\end{tikzcd} \tag{6.1}$$

where $X(t)$ is a k-dimensional diffusion process specified by Eqs. (2.20) and (2.21), X^β is the stationary random vector associated with the continuous time Markov chain, $X^\beta(t)$, and X is a random vector with density $f(x)$ given by Eq. (2.22). In the symmetric case, $f(x)$ is centered at $(p/k)e$. In that case, if we define $\hat{X}_p = X - (p/k)e$, then as $p \to \infty$,

$$\hat{X}_p \to_d k^{-1/2}\chi e, \tag{6.2}$$

where χ is a standard Normal. This is a “collapse” from a proper density on $\mathcal{R}^k$ to a density on L (the one-dimensional subspace of $\mathcal{R}^k$ given by (2.1)). State space collapse also occurs for $Y^\beta(t)$, i.e.,

$$\begin{tikzcd}
 & \frac{\sqrt{k+p}}{k}U(t)e \arrow[dr, "d"] & \\
Y^{\beta}(t) \arrow[ur, "d"] \arrow[dr, "d"] & & \frac{\sqrt{k+p}}{k}\chi e, \\
 & Y^{\beta} \arrow[ur, "d"] &
\end{tikzcd} \tag{6.3}$$

where $U(t)$ is an Ornstein–Uhlenbeck process, and Y^{β} is the stationary random vector associated with $Y^{\beta}(t)$.

The four southeast pointing limits in diagrams (6.1) and (6.3) all follow from elementary considerations. The four northeast pointing limits are conjectured, as is the fact that the upper and lower paths come together (i.e., the order of the limits $t \to \infty$ and $\beta \to \infty$ can be interchanged). The convergence in diagram (6.2) was proved in Section 2 in Eqs. (2.11)–(2.16).

References

1. Alanyali, M. & Hajek, B. (1996). On load balancing in Erlang networks. In F.P. Kelly, S. Zachary, and I. Ziedens (eds.), *Stochastic networks: Theory and applications.* Oxford: Oxford University Press.
2. Borovkov, A.A. (1967). On limit laws for service processes in multi-channel systems. *Siberian Mathematics Journal* 8: 746–763.
3. Gardiner, C.W. (1997). *Handbook of stochastic methods*. Berlin: Springer-Verlag.
4. Hariharan, R., Kulkarni, V.G., & Stidham, S. (1990). Optimal control of two infinite server queues. *Proceedings of the 29th IEEE Conference on Decision and Control*, Hawaii.
5. Iglehart, D.L. (1965). Limiting diffusion approximations for the many server queue and the repairman problem. *Journal of Applied Probability* 2: 429–441.
6. Krichagina, E.V. & Puhalskii, A.A. (1997). A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems* 25: 235–280.
7. Reiman, M.I. (1984). Some diffusion approximations with state space collapse. In F. Baccelli and G. Fayolle (eds.), *Modeling and performance evaluation methodology.* Berlin: Springer-Verlag, pp. 209–240.
8. Turner, S.R.E. (1998). A join the shorter queue model in heavy traffic. University of Cambridge Statistical Laboratory research report 1998-6. To appear in *Journal of Applied Probability*.
9. Viterbi, A.J. (1995). *CDMA: Principles of spread spectrum communication*. Reading, MA: Addison-Wesley.
10. Whitt, W. (1984). Heavy-traffic approximation for service systems with blocking. *AT&T Bell Laboratories Technical Journal* 63: 689–708.

https://doi.org/10.1017/S0269964899133011 Published online by Cambridge University Press