J. Appl. Prob. **49,** 503–520 (2012)
Printed in England

EXACT-ORDER ASYMPTOTIC ANALYSIS FOR CLOSED QUEUEING NETWORKS

DAVID K. GEORGE [*] [**] AND
CATHY H. XIA,[*] [***] *The Ohio State University*
MARK S. SQUILLANTE,[****] *IBM Thomas J. Watson Research Center*

Abstract

In this paper we study the asymptotic behavior of a general class of product-form closed queueing networks as the population size grows large. We first characterize the asymptotic behavior of the normalization constant for the stationary distribution of the network in exact order. This result then enables us to establish the asymptotic behavior of the system performance metrics, which extends a number of well-known asymptotic results to exact order. We further derive new, computationally simple approximations for performance metrics that significantly improve upon existing approximations for large-scale networks. In addition to their direct use for the analysis of large networks, these new approximations are particularly useful for reformulating large-scale queueing network optimization problems into more easily solvable forms, which we demonstrate with an optimal capacity planning example.

Keywords: Closed queueing network; large population; asymptotic analysis; exact-order asymptotics; approximation methods

2010 Mathematics Subject Classification: Primary 60K25
Secondary 68M20; 90B22

1. Introduction

Closed queueing networks are commonly used in many fields of science, engineering, and business to model a variety of real-world systems, e.g. communication networks, computer systems, transportation networks, and supply chains. In particular, closed queueing networks whose stationary distributions have a product-form solution are quite popular because many steady-state performance metrics of interest can be expressed in simple analytical forms. Such networks are referred to as product-form closed queueing networks. These product-form networks have been playing, and continue to play, an important role throughout the lifecycle of many real-world systems in science, engineering, and business from the stages of design and implementation through the stages of operational management and system capacity planning.

It is well known, however, that the computational effort of calculating exact values for the steady-state performance metrics of interest becomes prohibitively expensive as the population size grows large, even for relatively simple networks. The underlying difficulty concerns the calculation of the normalization constant of the stationary probability distribution, which

Received 12 April 2011; revision received 18 October 2011.
[*] Postal address: Department of Integrated Systems Engineering, The Ohio State University, Columbus, OH 43201, USA.
The authors gratefully acknowledge the support of the National Science Foundation under grant IIS-0916440.
[**] Email address: george.385@osu.edu
[***] Email address: xia.52@osu.edu
[****] Postal address: Mathematical Sciences Department, IBM Thomas J. Watson Research Center, PO Box 218, Yorktown Heights, NY 10598, USA. Email address: mss@us.ibm.com

involves a sum that grows combinatorially in the population size. Given the ever-increasing size and complexity of the real-world systems of today, as well as those expected to emerge in the future, it is therefore a necessary undertaking to understand more fully the behavior of these closed queueing network models as the customer population size grows arbitrarily large. Such mathematical results are important both for the theoretical insights they provide into system behavior and for their practical application in deriving improved approximations for the analysis and optimization of large-scale queueing systems.

There is a vast research literature on the analysis of closed queueing networks. The previous research that most closely relates to the present study primarily falls into three categories: exact solution methods, approximations for performance metrics, and analysis of the normalization constant. Exact solution methods consist of iterative algorithms that either calculate the normalization constant (e.g. the convolution method [4] and the local balance algorithm for normalizing constants (LBANC) [6]) or directly calculate the performance metrics of interest (e.g. mean value analysis (MVA) [21]). Each of these exact methods typically improves the computational cost over that of direct calculation but still often requires a large number of computations for networks with large populations. The second category is comprised of methods to derive bounds on the performance metrics in question [16, Chapter 5]. For example, Muntz and Wong [18] derived asymptotic bounds on the server utilizations and mean response times for a closed terminal network with only single-server and infinite-server stations. Zahorjan *et al.* [23] and Kriz [13] derived balanced job bounds on utilizations and response times by replacing the original network with a balanced network. Other studies seek to obtain increased accuracy at the cost of increased computational effort (e.g. [1], [5], and [10]). The third category focuses on analyzing the normalization constant and using such an analysis to formulate useful approximations for performance metrics of interest. Most previous studies focused on either closed-form expressions [22, Chapter 1] or asymptotic approximations [11], [12], [17] for the normalizing constant under rather restrictive assumptions on network structure or parameter values.

In this paper we directly study the *exact-order asymptotic behavior* of a general class of product-form closed queueing networks as the population size grows large. The main contributions of this paper are as follows. We first characterize the asymptotic behavior of the normalization constant for the stationary distribution of the network in exact order as the population size grows large. The asymptotics of the normalization constant are, in general, rather difficult to analyze directly, and so we develop a novel proof technique that makes use of the z-transform and associated limit theorems. We are then able to establish the exact-order asymptotic behavior of various system performance metrics, including system throughput and server utilization, and the stationary queue-length distribution, mean queue length, and mean response time at nonbottleneck stations. Our asymptotic results extend the existing results on the asymptotic behavior of closed queueing networks to exact order, yielding significant insights into system performance as a function of population size. We further demonstrate additional benefits of such exact-order asymptotics. Firstly, we combine our formulae with existing approximation methods to derive new, computationally simple approximations for system performance metrics. We show that these new approximations significantly improve the accuracy of existing methods for large-scale closed queueing network applications, especially when there are multiple bottlenecks in the network. Secondly, the simple mathematical forms and asymptotic exactness of these new approximations make them particularly useful for reformulating large-scale queueing network optimization problems into forms that are more easily solvable. We demonstrate this with an optimal capacity planning application, which is

representative of the optimization of user sessions in the network management of Internet-based data center environments. The problems of optimal fleet sizing and vehicle availability in large-scale vehicle rental systems, which we studied in [8], provide yet another interesting application area for the results presented herein.

The remainder of the paper is organized as follows. In Section 2 we discuss the details of the class of queueing networks that we study. In Section 3 we give some preliminary mathematical results on the z-transform. Section 4 presents our main exact-order asymptotic results. In Section 5 we combine our asymptotic formulae with known performance bounds to obtain improved approximations for networks with populations of arbitrary size. We demonstrate the application of these results for optimization through an optimal capacity planning example. In Section 6 we conclude and mention directions for future work.

2. Closed queueing network models

Consider a general class of closed queueing networks that consists of a fixed number, N, of customers circulating among an arbitrary, but finite, number, M, of stations. We restrict our attention to those closed queueing networks whose stationary distributions have a product-form solution. Throughout the present paper, we will often refer to this general class of product-form closed queueing networks more simply as a closed queueing network. Let $\mathcal{M}$ denote the set of all stations, so that $M = |\mathcal{M}|$. Let s_i and μ_i be the number of servers and the mean service rate of a server at station $i \in \mathcal{M}$, respectively. We categorize stations into three subsets with $\mathcal{M} = S \cup L \cup I$, where S denotes the set of single-server stations (with $s_i = 1$), L denotes the set of multi-server stations (with $1 < s_i < N$), and I denotes the set of infinite-server stations (with $s_i \geq N$). The service discipline at each station can be either first-come–first-served (FCFS), processor sharing (PS), last-come–first-served (LCFS), or infinite server. Specifically, we assume that S contains stations with PS, LCFS, and single-server FCFS disciplines, that L contains stations with multi-server FCFS discipline, and that I contains stations with infinite-server discipline. FCFS stations have exponentially distributed service times, while service times for the remaining three service disciplines can be generally distributed.

Let p_{ij} be the probability that a customer, upon completion of service at station i, moves to station j. We assume that the routeing matrix $\boldsymbol{P} = [p_{ij}]$ is irreducible. The solution to the balance equations $\pi_i = \sum_{j \in \mathcal{M}} \pi_j p_{ji}$ yields the *relative throughput* π_i of station i for all $i \in \mathcal{M}$. As these equations are linearly dependent, their solution is unique up to a multiplicative constant.

The queueing networks in which we are interested fall into the class of BCMP networks, which are known to have product-form stationary distributions [2]. For a given population size N, the continuous-time Markov process underlying the queueing network model has state space $\mathcal{S}(N, M) = \{(n_1, n_2, \ldots, n_M) : \sum_{i=1}^{M} n_i = N,\ n_i \geq 0\}$. The stationary probability $\mathrm{P}(n_1, n_2, \ldots, n_M)$ admits, for all $(n_1, \ldots, n_M) \in \mathcal{S}(N, M)$, the product-form expression [3]

$$\mathrm{P}(n_1, \ldots, n_M) = \frac{1}{G(N)} \prod_{i=1}^{M} f_i(n_i), \tag{1}$$

where

$$f_i(n_i) = \begin{cases} \gamma_i^{n_i}, & i \in S, \\ \dfrac{1}{\prod_{j=1}^{n_i} \min\{j, s_i\}} \gamma_i^{n_i}, & i \in L, \\ \dfrac{1}{n_i!} \gamma_i^{n_i}, & i \in I, \end{cases} \tag{2}$$

https://doi.org/10.1239/jap/1339878801 Published online by Cambridge University Press

and

$$\gamma_i = \frac{\pi_i}{\mu_i} \quad \text{for all } i \in \mathcal{M}. \tag{3}$$

The term $G(N)$ in (1) is the normalization constant chosen so that the stationary probabilities sum to 1. Namely,

$$G(N) = \sum_{\vec{n} \in \mathcal{S}(N,M)} \prod_{i=1}^{M} f_i(n_i). \tag{4}$$

Once $G(N)$ is obtained, many useful network performance metrics can be simply expressed in terms of this normalization constant. For example, the *actual throughput* of station i, denoted by $\Lambda_i(N)$, and the *actual utilization* of a server at station i, denoted by $U_i(N)$, can be written as

$$\Lambda_i(N) = \pi_i \frac{G(N-1)}{G(N)} \quad \text{for all } i \in \mathcal{M}, \tag{5}$$

$$U_i(N) = \frac{\pi_i}{s_i \mu_i} \frac{G(N-1)}{G(N)} \quad \text{for all } i \in S \cup L. \tag{6}$$

Unfortunately, the cardinality of the state space $\mathcal{S}(N, M)$ defines the size of the sum in (4) and it grows combinatorially as $\binom{M+N-1}{M}$, so calculating the normalization constant becomes prohibitively expensive for large-scale networks.

3. Preliminaries on z-transforms

We now present some preliminary mathematical results on the z-transform, which is a primary tool that we will use in our asymptotic analysis in Section 4. The z-transform of a sequence is defined as follows.

Definition 1. For any sequence $\{f(n),\ n \geq 0\}$, its z-transform $\mathcal{Z}\{f(n)\}$ is given by

$$\tilde{f}(z) = \sum_{n=0}^{\infty} f(n) z^{-n}, \qquad z \in \mathbb{C}.$$

We present some important mathematical properties of the z-transform next, starting with four well-known properties of the z-transform. For further details on these four results, we refer the interested reader to [20].

Properties of the z-transform. Let $\{f(n),\ n \geq 0\}$ and $\{g(n),\ n \geq 0\}$ be two sequences of real numbers with z-transforms $\tilde{f}(z)$ and $\tilde{g}(z)$, respectively. Then the following properties hold.

Time shifting. If $f(n) = g(n-k)$ for an integer k then $\tilde{f}(z) = z^{-k}\tilde{g}(z)$.

Differentiation. If $f(n) = ng(n)$ then $\tilde{f}(z) = -z\mathrm{d}\tilde{g}(z)/\mathrm{d}z$.

Convolution. If $f(n) = g_1(n) * g_2(n)$ then $\tilde{f}(z) = \tilde{g}_1(z) \cdot \tilde{g}_2(z)$.

First difference. If $f(n) = g(n) - g(n-1)$ then $\tilde{f}(z) = ((z-1)/z)\tilde{g}(z)$.

We will also use two well-known theorems for studying numerical sequences: the *final value theorem* and the *Stolz–Cesàro theorem*. We present the precise statements of both theorems next, referring the interested reader to [14, p. 581] and [19, p. 85], respectively, for additional details.

https://doi.org/10.1239/jap/1339878801 Published online by Cambridge University Press

Theorem 1. (Final value theorem.) *Let $\{p_n\}_{n\geq 0}$ be a sequence of real numbers. Its z-transform is given by $p(z) = \sum_{n=0}^{\infty} p_n z^{-n}$. Then $\lim_{n\to\infty}(1/n)\sum_{i=0}^{n} p_i = \lim_{z\to 1}(z-1)p(z)$ if either one of the limits exists.*

Theorem 2. (Stolz–Cesàro theorem.) *Let $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ be real sequences, both tending to ∞ as $n \to \infty$, and suppose that $\{b_n\}$ is asymptotically strictly monotonic. Define $\Delta a_n = a_n - a_{n-1}$ and Δb_n similarly. Then*

$$\lim_{n\to\infty} \frac{a_n}{b_n} = \lim_{n\to\infty} \frac{\Delta a_n}{\Delta b_n},$$

provided that the limit on the right-hand side exists.

In addition, we obtain the following extension of the final value theorem, which will be convenient for our proofs in Section 4.

Theorem 3. *Let $\{f(n),\ n \geq 0\}$ be a sequence of real numbers with z-transform as in Definition 1. For a fixed integer $l \geq 1$,*

$$\lim_{n\to\infty} \frac{f(n)}{n^{l-1}} = \frac{1}{(l-1)!} \lim_{z\to 1} \frac{(z-1)^l}{z^{l-1}} \tilde{f}(z),$$

provided that the limit on the right-hand side exists.

Proof. Let the sequence $\{f^{(j)}(n),\ n \geq 0\}$ denote the jth order difference of the sequence $\{f(n),\ n \geq 0\}$, defined as

$$f^{(0)}(n) := f(n), \qquad f^{(j)}(n) := \Delta f^{(j-1)}(n) := f^{(j-1)}(n) - f^{(j-1)}(n-1) \quad \text{for all } j \geq 1.$$

Upon applying the final value theorem to the sequence $\{f^{(l-1)}(n),\ n \geq 0\}$, we have

$$\lim_{n\to\infty} \frac{\sum_{i=0}^{n} f^{(l-1)}(i)}{n} = \lim_{z\to 1} (z-1)\tilde{f}^{(l-1)}(z) = \lim_{z\to 1} \frac{(z-1)^l}{z^{l-1}} \tilde{f}(z), \tag{7}$$

where the second equality follows from a recursive application of the first-difference property of the z-transform.

Applying the Stolz–Cesàro theorem, and using the facts that $\Delta f^{(j)}(n) = f^{(j+1)}(n)$ and $\Delta n^m = n^m - (n-1)^m \sim m n^{m-1}$ for large n, we have

$$\begin{aligned} \lim_{n\to\infty} \frac{f^{(0)}(n)}{n^{l-1}} &= \lim_{n\to\infty} \frac{f^{(1)}(n)}{(l-1)n^{l-2}} \\ &= \cdots \\ &= \lim_{n\to\infty} \frac{f^{(l-2)}(n)}{(l-1)!\,n} \\ &= \frac{1}{(l-1)!} \lim_{n\to\infty} \frac{\sum_{i=0}^{n} f^{(l-1)}(i)}{n}, \end{aligned} \tag{8}$$

where the last equality follows because $\sum_{i=0}^{n} f^{(l-1)}(i) = f^{(l-2)}(n)$, and we know that the limit on the right-hand side exists and is given by (7). Since $f^{(0)}(n) = f(n)$, combining (7) and (8) gives the desired result.

https://doi.org/10.1239/jap/1339878801 Published online by Cambridge University Press

4. Exact-order asymptotic results for closed queueing networks

We now apply the z-transform results presented in Section 3 to our closed queueing network model. We first note that, for any M-station product-form closed queueing network, the normalization constant $G(N)$ has z-transform $\tilde{G}(z)$ given by

$$\tilde{G}(z) = \prod_{i=1}^{M} \tilde{f}_i(z). \tag{9}$$

Equation (9) follows from the fact that the function $G(N)$, by its structure, is the convolution of the functions f_i, and, hence, its generating function is the product of the generating functions of the f_i over all $i \in \mathcal{M}$ [22, Chapter 1]. Here there are no assumptions on the form of the $\tilde{f}_i$, and, therefore, this is a general result which holds for any product-form closed queueing network.

When the structure of the network is known, we can use (9) to write explicit forms for $\tilde{G}(z)$. As a specific example, for networks of the form described in Section 2, we can easily derive $\tilde{f}_i$ from (2) as follows:

$$\tilde{f}_i(z) = \begin{cases} \dfrac{z}{z-\gamma_i}, & i \in S, \\ \mathrm{e}^{\gamma_i/z}, & i \in I, \\ F\left(\dfrac{\gamma_i}{z}, s_i\right), & i \in L. \end{cases} \tag{10}$$

Here

$$F(u,k) = \left(1 + u + \frac{u^2}{2!} + \cdots + \frac{u^{k-1}}{(k-1)!} + \frac{u^k}{k!(1-u/k)}\right). \tag{11}$$

The derivation of the first two cases is obvious. For details on how to derive the third case, see [9].

Note that in (10) the expression for the single-server and infinite-server cases of $\tilde{f}_i(z)$ can be viewed as special cases of the multi-server case, obtained by setting $s_i = 1$ or letting $s_i \to \infty$, respectively. We therefore rewrite (10) as

$$\tilde{f}_i(z) = F\left(\frac{\gamma_i}{z}, s_i\right) \quad \text{for all } i \in \mathcal{M}, \tag{12}$$

in which s_i can take on values 1 and $+\infty$.

We now present our exact-order asymptotic results for closed queueing networks of the form detailed in Section 2. Formally, we say that $G(N) \sim cf(N)$ if

$$\lim_{N\to\infty} \frac{G(N)}{f(N)} = c \quad \text{where } c \in (0, \infty) \text{ is a constant.}$$

Thus, an exact-order asymptotic result gives us the rate at which $G(N)$ grows as N becomes large.

Define

$$\rho_i := \frac{\pi_i}{s_i \mu_i} \quad \text{for all } i \in \mathcal{M},$$

where ρ_i is referred to as the *relative utilization* of station i. We next introduce the following assumption to simplify the exposition.

https://doi.org/10.1239/jap/1339878801 Published online by Cambridge University Press

Assumption 1. *Throughout the rest of the paper, we assume that the relative throughputs* $\{\pi_i,\ i \in \mathcal{M}\}$ *are chosen so that* $\max\{\rho_i,\ i \in \mathcal{M}\} = 1$ *and that stations are labeled by their relative utilizations such that* $1 = \rho_1 \geq \rho_2 \geq \cdots \geq \rho_M$.

Note that $\rho_i = 0$ for any infinite-server station $i \in I$. To avoid the trivial case, we assume that there exists at least one station $i \notin I$ with $\rho_i > 0$. Define $B := \{i \in \mathcal{M}\colon \rho_i = 1\}$, which is the set of bottleneck stations. Clearly, $|B| \geq 1$.

We now present our main theorem that characterizes the asymptotic behavior of the normalization constant $G(N)$ in exact order.

Theorem 4. *For any M-station product-form closed queueing network, the normalization constant* $G(N)$*, defined as in (1)–(4), satisfies the exact-order asymptotics*

$$G(N) \sim C_B N^{|B|-1},$$

where C_B *is a constant given by*

$$C_B = \frac{1}{(|B|-1)!} \prod_{i=1}^{|B|} \frac{\gamma_i^{s_i}}{s_i!} \prod_{j=|B|+1}^{M} \tilde{f}_j(1).$$

Proof. Let $l = |B|$. Using the multi-server form (12) for $\tilde{f}_i$, the z-transform for the sequence $\{G(n),\ n \geq 0\}$ can be written as

$$\tilde{G}(z) = \prod_{j=1}^{M} \tilde{f}_j(z) = \prod_{i=1}^{l} \left(1 + \cdots + \frac{\gamma_i^{s_i-1}}{(s_i-1)!} z^{-(s_i-1)} + \frac{\gamma_i^{s_i} z}{(s_i)!\,(z-1)} z^{-s_i}\right) \prod_{j=l+1}^{M} \tilde{f}_j(z).$$

Applying Theorem 3 to the sequence $\{G(n),\ n \geq 0\}$, we have

$$\lim_{n\to\infty} \frac{G(n)}{n^{l-1}} = \frac{1}{(l-1)!} \lim_{z\to 1} \frac{(z-1)^l}{z^{l-1}} \tilde{G}(z) = \frac{1}{(l-1)!} \prod_{i=1}^{l} \frac{\gamma_i^{s_i}}{s_i!} \prod_{j=l+1}^{M} \tilde{f}_j(1).$$

The desired result follows immediately.

Qualitatively, Theorem 4 states that, as N grows large, the rate at which $G(N)$ grows is on the order of $N^{|B|-1}$, which is closely related to the number of bottleneck servers. With this result we can now derive the asymptotic behavior of a number of system performance metrics in exact order.

The following corollary is immediate from Theorem 4 and (5) and (6).

Corollary 1. *For station* $i \in \mathcal{M}$*, the actual throughput* $\Lambda_i(N)$ *and actual utilization* $U_i(N)$*, as defined in (5) and (6), satisfy the following exact-order asymptotics:*

$$\Lambda_i(N) \sim \pi_i \left(1 - \frac{1}{N}\right)^{|B|-1} \quad \textit{for all } i \in \mathcal{M}, \tag{13}$$

$$U_i(N) \sim \rho_i \left(1 - \frac{1}{N}\right)^{|B|-1} \quad \textit{for all } i \in S \cup L. \tag{14}$$

https://doi.org/10.1239/jap/1339878801 Published online by Cambridge University Press

Note that by letting $N \to \infty$ in (13) and (14) we have $\lim_{N\to\infty} \Lambda_i(N) = \pi_i$ for all $i \in \mathcal{M}$ and $\lim_{N\to\infty} U_i(N) = \rho_i$ for all $i \in S \cup L$. Both limits agree with results that have been previously established in the literature (see, e.g. [18]). Our results in (13) and (14) demonstrate that the convergence is at the rate of $(1 - 1/N)^{|B|-1}$, where the larger the bottleneck station set B, the slower the convergence rate.

Let $(X_1^N, \ldots, X_M^N)$ denote a random vector that represents the number of customers at the stations in steady state when the closed queueing network has a total population size N. Let $X_{B^c}^N = (X_i^N,\ i \in B^c)$ be the subvector on the number of customers at all nonbottleneck stations. We can further characterize the asymptotic behavior of the stationary distribution of $X_{B^c}^N$ for all nonbottleneck stations as follows.

Corollary 2. *For a nonbottleneck station $i \notin B$ and a fixed integer $m \geq 0$, the marginal queue-length distribution at station i satisfies the exact-order asymptotics*

$$\mathrm{P}(X_i^N = m) \sim \frac{f_i(m)}{\tilde{f}_i(1)} \left(1 - \frac{m}{N}\right)^{|B|-1} \quad \textit{for all } i \notin B. \tag{15}$$

Consequently, given an integer vector $x_{B^c} = (x_{|B|+1}, \ldots, x_M) \geq 0$, we have

$$\mathrm{P}(X_{B^c}^N = x_{B^c}) \sim \prod_{i \in B^c} \frac{f_i(x_i)}{\tilde{f}_i(1)} \left(1 - \frac{x_i}{N}\right)^{|B|-1}. \tag{16}$$

Proof. We can express $\mathrm{P}(X_i(N) = m)$ as (see [15, Chapter 3])

$$\mathrm{P}(X_i(N) = m) = f_i(m) \frac{G_{\mathcal{M}-(i)}(N - m)}{G(N)}, \tag{17}$$

where $G_{\mathcal{M}-(i)}(x)$ is the normalizing constant for the stationary distribution of the closed queueing network with station i removed and a total population size x. Applying Theorem 4, we have, for large N,

$$G_{\mathcal{M}-(i)}(N - m) \sim C_B^{(i)} (N - m)^{|B|-1}, \tag{18}$$

where

$$C_B^{(i)} = \frac{1}{(|B| - 1)!} \prod_{k=1}^{|B|} \frac{\gamma_k^{s_k}}{s_k!} \prod_{j=|B|+1,\ j\neq i}^{M} \tilde{f}_i(1).$$

Combining (17) and (18), we then have (15).

Finally, we can easily establish (16) based on the product-form distribution and (15).

Note that by letting $N \to \infty$ in (15) we obtain the limiting behavior

$$\lim_{N\to\infty} \mathrm{P}(X_i(N) = m) = \frac{f_i(m)}{\tilde{f}_i(1)}.$$

Plugging in the values of f_i and $\tilde{f}_i$ for each station in different categories, it is readily verified that

$$\lim_{N\to\infty} \mathrm{P}(X_i(N) = m) = \begin{cases} (1 - \rho_i)\rho_i^m, & i \in S, \\ \mathrm{e}^{-\gamma_i} \dfrac{\gamma_i^m}{m!}, & i \in I, \\ \dfrac{1}{F(\gamma_i, s_i)} \dfrac{\gamma_i^m}{\prod_{j=1}^m \min\{j, s_i\}}, & i \in L, \end{cases}$$

https://doi.org/10.1239/jap/1339878801 Published online by Cambridge University Press

where $1/F(\gamma_i, s_i)$, defined as in (11), gives the steady-state probability of having zero customers in an M/M/s_i queue. Namely, as $N \to \infty$, the distribution of $X_i(N)$ converges to the queue-length distribution of station i in a corresponding open BCMP network that has the bottleneck stations removed. This coincides with a previous result that has been established in the literature (see, e.g. [1]). We refer the reader to, e.g. [22, Chapter 1] for the proof of a special case in the setting of a closed Jackson network.

The next corollary further characterizes the exact-order behavior of the mean queue length and mean response time at nonbottleneck stations.

Corollary 3. *For a nonbottleneck station $i \notin B$, the mean queue length $L_i(N)$ and the mean sojourn time $R_i(N)$ at station i respectively satisfy the following exact-order asymptotics:*

$$L_i(N) \sim -\frac{\tilde{f}_i'(1)}{\tilde{f}_i(1)}\left(1 - \frac{1}{N}\right)^{|B|-1} \quad \text{for all } i \notin B, \tag{19}$$

$$R_i(N) \sim -\frac{\tilde{f}_i'(1)}{\tilde{f}_i(1)\pi_i} \quad \text{for all } i \notin B. \tag{20}$$

Proof. Given the queue length distribution $\mathrm{P}(X_i(N) = m)$ for $m = 0, 1, \ldots, N$, we can compute the expected queue length as

$$\begin{aligned} L_i(N) = \sum_{m=1}^{N} m\,\mathrm{P}(X_i(N) = m) &= \sum_{m=1}^{N} m f_i(m) \frac{G_{M-(i)}(N-m)}{G(N)} \\ &= \frac{1}{G(N)} \sum_{t=0}^{N-1} (t+1) f_i(t+1) G_{M-(i)}(N-1-t) \\ &= \frac{1}{G(N)} \sum_{t=0}^{N-1} g_i(t) G_{M-(i)}(N-1-t) \\ &= \frac{H(N-1)}{G(N)}, \end{aligned} \tag{21}$$

where $g_i(t) = (t+1) f_i(t+1)$, and $H(n)$ is the convolution of the two sequences $\{g_i(n)\}$ and $\{G_{M-(i)}(n)\}$. Hence, the z-transform of $H(n)$ is given by the product

$$\tilde{H}(z) = \mathcal{Z}\{g_i(n)\}\mathcal{Z}\{G_{M-(i)}(n)\} = \tilde{g}(z)\tilde{G}_{M-(i)}(z).$$

Using the time-shifting and differentiation properties of the z-transform, we have

$$\tilde{g}(z) = \mathcal{Z}\{g_i(n)\} = \mathcal{Z}\{(n+1) f_i(n+1)\} = z\mathcal{Z}\{n f_i(n)\} = -z^2 \tilde{f}_i'(z).$$

Applying Theorem 3 to the sequence $\{H(n),\ n \geq 0\}$ with $l = |B|$, we obtain

$$\begin{aligned} \lim_{n\to\infty} \frac{H(n)}{n^{l-1}} &= \frac{1}{(l-1)!} \lim_{z\to 1} \frac{(z-1)^l}{z^{l-1}} \tilde{H}(z) \\ &= \frac{1}{(l-1)!} \lim_{z\to 1} \frac{(z-1)^l}{z^{l-1}} \tilde{G}_{M-(i)}(z)(-z^2 \tilde{f}_i'(z)) \\ &= -C_B^{(i)}\, \tilde{f}_i'(1). \end{aligned} \tag{22}$$

https://doi.org/10.1239/jap/1339878801 Published online by Cambridge University Press

Then (19) follows immediately because

$$\frac{H(N-1)}{G(N)} \sim \frac{-C_B^{(i)} \tilde{f}_i'(1)(N-1)^{|B|-1}}{C_B N^{|B|-1}} = -\frac{\tilde{f}_i'(1)}{\tilde{f}_i(1)}\left(1-\frac{1}{N}\right)^{|B|-1}.$$

For the asymptotic result on the mean response time at station i, we apply Little's law to obtain

$$R_i(N) = \frac{L_i(N)}{\Lambda_i(N)} = \frac{H(N-1)}{G(N-1)\pi_i},$$

where the second equality comes from (5) and (21). Applying Theorem 4 and (22), we then have the desired result (20).

Note that by letting $N \to \infty$ in (19) we obtain the limiting behavior

$$\lim_{N\to\infty} L_i(N) = -\frac{f_i'(1)}{\tilde{f}_i(1)}.$$

Plugging in the values of f_i' and $\tilde{f}_i$ for each station in different categories, it is readily verified that

$$\lim_{N\to\infty} L_i(N) = \begin{cases} \dfrac{\rho_i}{(1-\rho_i)}, & i \in S, \\ \dfrac{\pi_i}{\mu_i}, & i \in I, \\ \dfrac{1}{F(\gamma_i, s_i)} \displaystyle\sum_{n=0}^{s_i-1} n\frac{1}{n!}\gamma_i^n + \frac{s_i^{s_i}}{s_i}\rho_i^{s_i}\left[\frac{s_i}{1-\rho_i} + \frac{\rho_i}{(1-\rho_i)^2}\right], & i \in L. \end{cases}$$

Namely, as $N \to \infty$, $L_i(N)$ converges to the mean queue length of station i in the corresponding open BCMP network that has the bottleneck stations removed, which again coincides with a previous result that has been established in the literature (see, e.g. [1]). Our result in (19) shows that the convergence of the mean queue length at each nonbottleneck station is again at rate $(1-1/N)^{|B|-1}$. Note from (20) that the mean response time at each nonbottleneck station, however, converges to the limit at a constant rate.

Remark 1. We will refer to (13), (14), (15), and (19) as *asymptotically exact* (AE) performance approximations.

5. Performance approximations and optimization

We now combine our exact-order asymptotic results from Section 4 with existing bounding methods to obtain new, improved approximations for the mean performance metrics of a closed queueing network with arbitrary population size. In Sections 5.1 and 5.2, we present two new approximations that arise when combining our exact-order asymptotic formulae with the bounding methods of asymptotic bound analysis (ABA) [18] and balanced job bounds (BJBs) [13], [23], respectively, two of the most widely used methods for approximating throughput and response time in closed queueing networks. In Section 5.3 we demonstrate the usefulness of these new approximations for optimization applications through an optimal capacity planning example.

https://doi.org/10.1239/jap/1339878801 Published online by Cambridge University Press

5.1. Improving ABA bounds

ABA was developed by Muntz and Wong [18] to provide a computationally simple method for determining bounds on throughput and response time in closed queueing networks. These bounds require only the assumption that the service time of a customer does not depend on the number of customers in the system, or at which stations the other customers are located. ABA bounds are thus applicable to a wide variety of closed queueing networks, including the networks we defined in Section 2.

System throughput, denoted by $\Lambda(N)$, is equivalent to the throughput of the bottleneck station(s), which is given by $\Lambda_1(N)$ under our labeling. The ABA bounds on system throughput are derived by considering the two asymptotically extreme cases of light and heavy loads. When in light load, queueing effects are minimal and system throughput grows linearly with population size. When in heavy load, system throughput is constrained by the maximum service rate of the bottleneck station(s). Hence, the ABA upper bound on system throughput is given by

$$\Lambda(N) \leq \min\left\{\frac{N}{D+Z}, s_1\mu_1\right\} =: \Lambda^{\mathrm{ABA}}(N), \tag{23}$$

where $D = \sum_{i\in S\cup L} \pi_i/\pi_1\mu_i$ and $Z = \sum_{i\in I} \pi_i/\pi_1\mu_i$. Intuitively, D corresponds to the expected total service time spent at single-server (SS) and multi-server (MS) stations, and Z corresponds to the expected total time spent at infinite-server (IS) stations, and, therefore, $D + Z$ yields a lower bound on the cycle time obtained by ignoring queueing effects. The upper bound in (23) is a piecewise-linear function of N, which reduces to the flat horizontal line at $s_1\mu_1$ for N greater than the 'saturation point'. This saturation point is defined as the population size at which the bottleneck station(s) reaches 100% utilization.

Incorporating our AE formula for $\Lambda_1(N)$ from (13) should then improve the accuracy of (23) near and beyond the saturation point. The AE-based approximation for system throughput is then given by

$$\Lambda(N) \approx \min\left\{\frac{N}{D+Z}, s_1\mu_1\left(1-\frac{1}{N}\right)^{|B|-1}\right\}. \tag{24}$$

We note that in the case of a network with a single bottleneck, (24) reduces to the previous ABA bound. In particular, we see that our proposed approximation provides the greatest improvements in accuracy for networks with a larger number of bottleneck stations.

System response time, denoted by $R(N)$, is related to system throughput via Little's law and is given by the equation (see [7])

$$R(N) = \frac{N}{\Lambda(N)} - Z. \tag{25}$$

This definition of the response time is particularly common in the modeling of computer systems [5], [7], [13], [16] in which users submit jobs to a network of computers for processing, where the 'think time' in between job submissions is modeled by IS stations. In this setting, Z corresponds to the expected time that users spend in between job submissions, and $R(N)$ corresponds to the expected response time of submitted jobs through the network of computers.

Combining (23) and (25), we obtain the ABA lower bound on the system response time as

$$R(N) \geq \max\left\{D, \frac{N}{s_1\mu_1} - Z\right\} =: R^{\mathrm{ABA}}(N).$$

Similarly, combining (24) and (25), we obtain the AE-based approximation for the system

https://doi.org/10.1239/jap/1339878801 Published online by Cambridge University Press

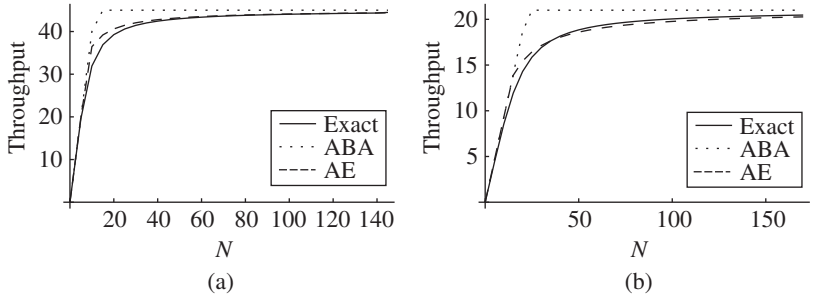

FIGURE 1: Comparison of the ABA and AE-based approximations for networks with (a) four MS stations and (b) 10 MS stations.

TABLE 1: Error comparison for the ABA and AE-based approximations for MS network examples.

	Network	Average error (%)		Maximum error (%)	
		ABA	AE	ABA	AE
Throughput	4 MS stations	5.04	1.19	25.30	14.18
	10 MS stations	8.65	2.29	32.40	16.35
Response time	4 MS stations	4.53	1.11	20.19	12.42
	10 MS stations	7.55	2.18	24.47	14.05

response time as

$$R(N) \approx \max\left\{D, \frac{N}{s_1\mu_1(1-1/N)^{|B|-1}} - Z\right\}. \tag{26}$$

Finally, noting that $\Lambda_i(N) = (\pi_i/\pi_1)\Lambda(N)$, we obtain an AE-based approximation for the throughput of an arbitrary station $i \in \mathcal{M}$ as

$$\Lambda_i(N) \approx \frac{\pi_i}{\pi_1} \min\left\{\frac{N}{D+Z}, s_1\mu_1\left(1-\frac{1}{N}\right)^{|B|-1}\right\}.$$

We illustrate and quantify the accuracy improvements of these approximations with two examples. First, consider a four-station network where each station has three identical servers. Each server at station $i \in \{1, 2, 3\}$ operates at rate $\mu_i = 15$ customers per unit time. The servers at station 4 each operate at rate $\mu_4 = 20$ customers per unit time. Routeing is symmetric, i.e. $p_{ij} = 0.25$ for all $i, j \in \{1, 2, 3, 4\}$. In Figure 1(a) we plot the system throughput obtained from different methods as a function of N. The solid line represents the exact solution values obtained from MVA, the dotted line corresponds to the ABA bound, and the dashed line corresponds to the AE-based approximation (24). Note that our AE-based approximation and the ABA approximation overlap for N less than the saturation point.

In addition, consider a second, larger sample network with 10 stations. Once again, each station has three identical servers. Each server at station $i \in \{1, \ldots, 6\}$ operates at rate $\mu_i = 7$ customers per unit time. Each server at station $i \in \{7, \ldots, 10\}$ operates at rate $\mu_i = 12$ customers per unit time. Routeing is symmetric, i.e. $p_{ij} = 0.1$ for all $i, j \in \{1, \ldots, 10\}$. In Figure 1(b) we plot the system throughput obtained via the same three methods above as a function of N.

In Table 1 we summarize the error percentages associated with the AE-based and ABA methods for calculating system throughput and system response time. For each of the four

https://doi.org/10.1239/jap/1339878801 Published online by Cambridge University Press

scenarios, in Table 1 we give each method's average and maximum percentage away from the exact, MVA-obtained solution. Specifically, we find the average and maximum errors of each method for N in $\{0, 5, 10, \ldots, 145\}$ for the four MS station scenario and for N in $\{0, 5, 10, \ldots, 170\}$ for the 10 MS station scenario. Note that if we continue further to the right of either of these intervals, the error for both methods decreases.

In both examples, the ABA bounds provide a tighter approximation for small N over the AE (second) terms in (24) and (26), but as N moves closer to the saturation point and then beyond, the AE terms provide greater accuracy over the ABA bounds. This is as expected, since our AE-based approximations are asymptotically exact. Our new AE-based approximations (24) and (26) therefore provide tighter approximations than the ABA bounds in all cases, as demonstrated by both Figure 1 and Table 1.

5.2. Improving BJBs

BJBs are bounds on the throughput and response time of a network derived by considering related networks that are 'balanced' in the sense that all stations have equal relative utilizations. The performance metrics of these related balanced systems are used to provide bounds on the performance of the original system. These bounds were first derived by Zahorjan *et al.* [23] for product-form closed queueing networks with SS stations and were then extended in [13] to include IS stations. Thus, BJBs are applicable to a smaller class of networks than the ABA bounds. For the networks to which they are applicable, however, BJBs always provide a tighter bound than the ABA bounds, at the cost of additional computational effort.

Let us consider a network consisting of only SS and IS stations, since BJBs are not applicable to networks with MS stations. The BJB formula for system throughput is given by

$$\Lambda(N) \leq \min\left\{\frac{N}{D + Z + (N-1)D/[|S|(1 + Z/D)]}, \mu_1\right\} =: \Lambda^{\mathrm{BJB}}(N). \tag{27}$$

This upper bound is a piecewise function of N, which reduces to the flat horizontal line at μ_1 for N greater than the saturation point. Hence, as with the ABA bound, we can incorporate our AE formula for $\Lambda_1(N)$ from (13) to improve accuracy near and beyond the saturation point. The new AE-based approximation is given by

$$\Lambda(N) \approx \min\left\{\frac{N}{D + Z + (N-1)D/[|S|(1 + Z/D)]}, \mu_1\left(1 - \frac{1}{N}\right)^{|B|-1}\right\}. \tag{28}$$

Note that our approximation given in (24) can still be used for this case. However, it is obvious that since the BJBs are always tighter than the ABA bounds, (28) will always provide a better approximation than (24), at the cost of slightly more computational effort.

Analogous to Section 5.1, we use (25) and (27) to obtain the lower bound for the system response time as

$$R(N) \geq \max\left\{D + \frac{(N-1)D}{|S|(1 + Z/D)}, \frac{N}{\mu_1} - Z\right\} =: R^{\mathrm{BJB}}(N),$$

and use (25) and (28) to derive the AE-based approximation for the system response time as

$$R(N) \approx \max\left\{D + \frac{(N-1)D}{|S|(1 + Z/D)}, \frac{N}{\mu_1(1 - 1/N)^{|B|-1}} - Z\right\}. \tag{29}$$

https://doi.org/10.1239/jap/1339878801 Published online by Cambridge University Press

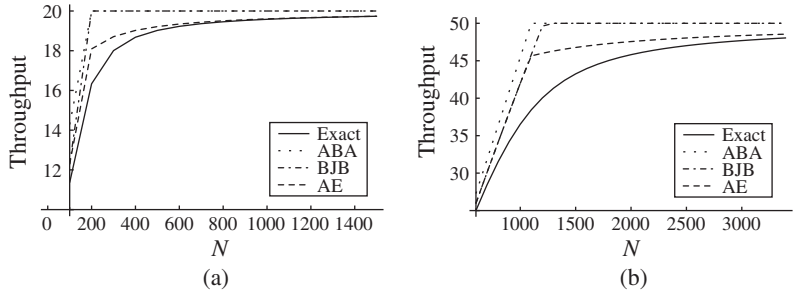

FIGURE 2: Comparison of the ABA, BJB, and AE-based approximations for networks with (a) 50 SS and 50 IS stations and (b) 100 SS and 100 IS stations.

TABLE 2: Error comparison for the BJB and AE-based approximations for SS + IS network examples.

	Network	Average error (%)		Maximum error (%)	
		BJB	AE	BJB	AE
Throughput	50 SS+50 IS stations	4.11	1.38	22.44	10.76
	100 SS+100 IS stations	8.35	4.86	23.94	18.72
Response time	50 SS+50 IS stations	5.41	2.24	30.98	17.70
	100 SS+100 IS stations	19.63	13.78	58.07	52.52

Again, noting that $\Lambda_i(N) = (\pi_i/\pi_1)\Lambda(N)$, we obtain an AE-based approximation for the throughput of an arbitrary station $i \in \mathcal{M}$ as

$$\Lambda_i(N) \approx \frac{\pi_i}{\pi_1} \min\left\{ \frac{N}{D + Z + (N-1)D/[|S|(1 + Z/D)]}, \mu_1\left(1 - \frac{1}{N}\right)^{|B|-1} \right\}.$$

We illustrate and quantify the accuracy improvements of these approximations with two examples. First, consider a network with 50 SS stations and 50 IS stations labeled (as in Section 4) such that $S = \{1, \ldots, J\}$ are the SS nodes and $I = \{J + 1, \ldots, M\}$ are the IS nodes. Station $i \in \{1, \ldots, 20\}$ operates at rate $\mu_i = 20$ customers per unit time, station $i \in \{21, \ldots, 50\}$ operates at rate $\mu_i = 30$ customers per unit time, and station $i \in I$ operates at rate $\mu_i = 10$ customers per unit time. Routeing is symmetric, i.e. $p_{ij} = 0.02$ for all $i, j \in \{1, \ldots, 100\}$. In Figure 2(a) we plot the system throughput obtained from different methods as a function of N. The solid line represents the exact solution values obtained from MVA, the dotted line corresponds to the ABA bound, the dash–dot line corresponds to the BJB, and the dashed line corresponds to the AE-based approximation (28). Note that the AE-based and BJB approximations overlap for N less than the saturation point.

In addition, consider a second network with 100 SS stations and 100 IS stations labeled as above where $S = \{1, \ldots, J\}$ (SS nodes) and $I = \{J + 1, \ldots, M\}$ (IS nodes). Station $i \in S$ operates at rate $\mu_i = 50$ customers per unit time, and station $i \in I$ operates at rate $\mu_i = 5$ customers per unit time. Routeing is symmetric, i.e. $p_{ij} = 0.005$ for all $i, j \in \{1, \ldots, 200\}$. In Figure 2(b) we plot the system throughput obtained via the same four methods above as a function of N.

In Table 2 we summarize the error percentages associated with the AE-based and BJB methods for calculating system throughput and system response time. For each of the four

https://doi.org/10.1239/jap/1339878801 Published online by Cambridge University Press

scenarios, in Table 2 we give each method's average and maximum percentage away from the exact, MVA-obtained solution. Specifically, we find the average and maximum errors of each method for N in $\{0, 100, 200, \ldots, 2000\}$ for the network with 50 SS and 50 IS stations and for N in $\{0, 100, 200, \ldots, 3400\}$ for the network with 100 SS and 100 IS stations. Again, note that, if we continue further to the right of either of these intervals, the error for both methods decreases.

In both examples, the BJBs provide a tighter approximation for small N over the AE (second) terms in (28) and (29), but as N moves closer to the saturation point and then beyond, the AE terms provide increased accuracy. Our new AE-based approximations (28) and (29) therefore provide tighter approximations than the ABA bounds and BJBs in all cases, as demonstrated by Figure 2 and Table 2.

5.3. Optimization-based application

We now apply our approximations developed in Sections 5.1 and 5.2 to the optimization of closed queueing networks. In particular, we use the motivating problem of determining optimal session capacity for a terminal computer network, though the methodology is applicable to a variety of problems concerned with optimization based on closed queueing networks. We first provide the mathematical formulation of the problem and then present the results of a corresponding set of numerical experiments.

A terminal network is a type of computer system in which a fixed number of user sessions each submit jobs to be processed by a network of computer resources. In Figure 3 we illustrate a general terminal computer network. Such terminal computer networks have been used, for example, to model the performance impact and optimization of the number of user sessions in the network management of Internet-based data center environments. These systems are often modeled as closed queueing networks with one IS station (station 1) representing connected user sessions and an arbitrary number of SS or MS stations (stations $\{2, \ldots, M\}$) representing the pool of computer resources. When a user submits a job, it is processed by the resource stations in a probabilistic order given by the routeing matrix of the network. Returning to the IS station represents completion of the job. The delay time at the IS station represents the think time of a user between job submissions. Thus, the model is a closed queueing network with a population of size N. A critical decision in designing such a system is: How many simultaneous user sessions should be allowed in the system?

In the closed queueing network model of this system, these operational decisions are equivalent to asking for the optimal population size N. Let r_i be the processing reward per customer for station $i \in \{2, \ldots, M\}$. Let m_i be a utilization threshold for station $i \in \{2, \ldots, M\}$. For example, if a station overheats or slows down past 95% utilization, we place a constraint on the utilization of that station. Let c be an overhead-type cost incurred for

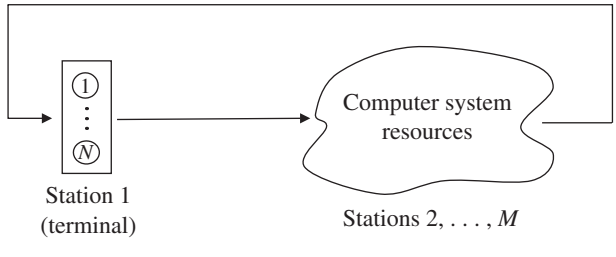

FIGURE 3: A general terminal network.

https://doi.org/10.1239/jap/1339878801 Published online by Cambridge University Press

maintaining a session per unit time. We can formulate a system-profit maximization problem as

$$\max_{N\in\mathbb{Z}_+} \sum_{i=2}^{M} r_i \Lambda_i(N) - cN \quad \text{such that} \quad U_i(N) \le m_i \quad \text{for } i = 2, \ldots, M. \tag{30}$$

The exact solution to (30) can be obtained with an iterative algorithm, but this becomes computationally expensive for large N due to the calculation of $\Lambda_i(N)$ and $U_i(N)$. Using our approximations from Section 5.2, we formulate an approximate optimization problem as

$$\max_{N\in\mathbb{R}_+,\, q\in\mathbb{R}_+^{|S|}} \sum_{i=2}^{M} r_i q_i - cN$$

such that

$$q_i \le \min\left\{ \Lambda_i^{\mathrm{BJB}}(N), \pi_i \left(1 - \frac{1}{N}\right)^{|B|-1} \right\} \quad \text{for } i = 2, \ldots, M, \tag{31}$$

$$m_i \ge \min\left\{ \mu_i^{-1} \Lambda_i^{\mathrm{BJB}}(N), \frac{\pi_i}{\mu_i} \left(1 - \frac{1}{N}\right)^{|B|-1} \right\} \quad \text{for } i = 2, \ldots, M. \tag{32}$$

Constraints (31) and (32) provide our approximations for $\Lambda_i(N)$ and $U_i(N)$, respectively. Note that we have also relaxed the integrality requirement on N since our solution method no longer requires N to be discrete. Thus, our problem can now be solved using common nonlinear programming methods.

We illustrate our approximate optimization solution with the following example. Consider a terminal network with 100 SS computer stations. The think time between job submissions for each session is exponentially distributed with a mean of 13 seconds. After service at station i, a job completes with probability $p_{i1} = 0.05$. Otherwise, it is routed to another station in the server pool, with each station being equally likely. The service time at each station is exponentially distributed with a mean of 10 seconds. Let $r_i = \$30$ for all stations and $c = \$0.2$.

In Figure 4 we show a plot of profit versus N for our proposed AE-based approximation, a BJB-only optimization, and an exact MVA solution. In Table 3 we summarize the solutions found by each method. In this example, the exact solution is $N^* = 3759$ with an optimal profit value of \$28 477. The BJB-only approximation gives $N^* = 2778$, which yields a 26.1%

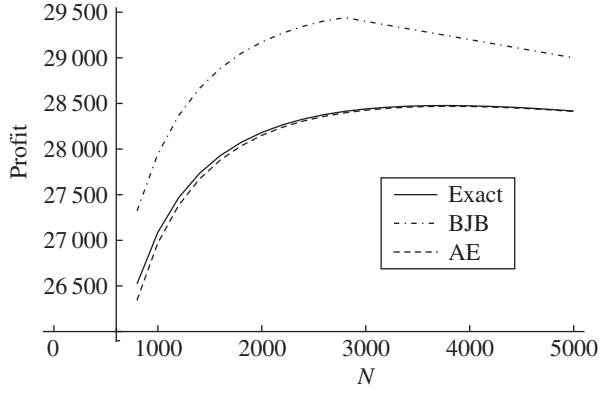

FIGURE 4: Comparison of solution methods for a 100-station terminal network.

https://doi.org/10.1239/jap/1339878801 Published online by Cambridge University Press

TABLE 3.

	N^*	% of relative error
Exact	3759	0
AE based	3804	1.2
BJB	2778	26.1

error margin. Our proposed AE-based approximation gives $N^* = 3804$, which renders only a 1.2% error. Thus, our proposed AE-based approximation very clearly offers greatly improved accuracy over the widely used BJB approximation.

6. Conclusion

In this paper we derived the exact-order asymptotic behavior of a general class of product-form closed queueing networks as the population size grows large. We first derived the exact-order asymptotic behavior of the normalization constant of the stationary distribution, obtained through a novel proof technique that makes use of the z-transform and final value limit theorems, and used this to further establish the exact-order asymptotic behavior of a number of performance metrics, including system throughput and server utilization, and the stationary queue-length distribution, mean queue length, and mean response time at nonbottleneck stations. As we established in Sections 4 and 5, our exact-order asymptotic results significantly improve upon the quality and accuracy of existing bounds and limits for large-scale networks, thus providing great benefit for practitioners. In addition to their direct use for the analysis of large networks, the simple mathematical forms and asymptotic exactness of our new approximations makes them particularly useful for reformulating large-scale queueing network optimization problems into more easily solvable forms, which we demonstrated with an optimal capacity planning example.

For future work, we plan to extend our analysis to study the exact-order asymptotics of closed queueing networks with multiple routeing chains, which allow a different routeing matrix for each different type of customer flow in the network.

Acknowledgements

This paper is based in part upon work supported by the National Science Foundation under grant IIS-0916440. The authors would also like to thank the anonymous referee and editor for their helpful comments and suggestions.

References

[1] ANSELMI, J. AND CREMONESI, P. (2008). Bounding the performance of BCMP networks with load-dependent stations. In *16th Internat. Symp. Modeling, Analysis, and Simulation of Computer and Telecommunication Systems* (Baltimore, MA, 2008), eds E. L. Miller and C. L. Williamson, IEEE Computer Society, pp. 171–178.

[2] BASKETT, F., CHANDY, K. M., MUNTZ, R. R. AND PALACIOS, F. G. (1975). Open, closed, and mixed networks of queues with different classes of customers. *J. Assoc. Comput. Mach.* **22,** 248–260.

[3] BRUELL, S. C. AND BALBO, G. (1980). *Computational Algorithms for Closed Queueing Networks*. North-Holland, New York.

[4] BUZEN, J. P. (1973). Computational algorithms for closed queueing networks with exponential servers. *Commun. Accoc. Comput. Mach.* **16,** 527–531.

[5] CASALE, G., MUNTZ, R. R. AND SERAZZI, G. (2008). Geometric bounds: a noniterative analysis technique for closed queueing networks. *IEEE Trans. Comput.* **57,** 780–794.

[6] Chandy, K. M. and Sauer, C. H. (1980). Computational algorithms for product form queueing networks. *Commun. Assoc. Comput. Mach.* **23,** 573–583.
[7] Denning, P. J. and Buzen, J. P. (1978). The operational analysis of queueing network models. *ACM Comput. Surveys* **10,** 225–261.
[8] George, D. K. and Xia, C. H. (2011). Fleet-sizing and service availability for a vehicle rental system via closed queueing networks. *Europ. J. Operat. Res.* **211,** 198–207.
[9] Gerasimov, A. I. (1995). On normalizing constants in multiclass queueing networks. *Operat. Res.* **43,** 704–711.
[10] Hsieh, C. T. and Lam, S. S. (1987). Two classes of performance bounds for closed queueing networks. *Performance Evaluation* **7,** 3–30.
[11] Knessl, C. and Tier, C. (1990). Asymptotic expansions for large closed queuing networks. *J. Assoc. Comput. Mach.* **37,** 144–174.
[12] Kogan, Y. (1992). Another approach to asymptotic expansions for large queueing networks. *Operat. Res. Lett.* **11,** 317–321.
[13] Kriz, J. (1984). Throughput bounds for closed queueing networks. *Performance Evaluation* **4,** 1–10.
[14] Kulkarni, V. G. (1995). *Modeling and Analysis of Stochastic Systems*. Chapman and Hall, London.
[15] Lavenberg, S. S. (1983). *Computer Performance Modeling Handbook*. Academic Press, Orlando, FL.
[16] Lazowska, E. D., Zahorjan, J., Graham, G. S. and Sevcik, K. C. (1984). *Quantitative System Performance: Computer System Analysis Using Queueing Network Models*. Prentice-Hall.
[17] McKenna, J. and Mitra, D. (1982). Integral representations and asymptotic expansions for closed Markovian queueing networks: normal usage. *Bell System Tech. J.* **61,** 661–683.
[18] Muntz, R. R. and Wong, J. W. (1974). Asymptotic properties of closed queueing network models. In *Proc. 8th Annual Princeton Conf. Information Sciences and Systems* (Princeton University, Princeton), pp. 348–352.
[19] Mureşan, M. (2008). *A Concrete Approach to Classical Analysis*. Springer.
[20] Ogata, K. (1987). *Discrete-Time Control Systems*. Prentice-Hall.
[21] Reiser, M. and Lavenberg, S. S. (1980). Mean-value anlaysis of closed multichain queueing networks. *J. Assoc. Comput. Mach.* **27,** 313–322. (Correction: **28**(1981), 629.)
[22] Serfozo, R. (1999). *Introduction to Stochastic Networks*. Springer, New York.
[23] Zahorjan, J., Sevcik, K. C., Eager, D. L. and Galler, B. (1982). Balanced job bound analysis of queueing networks. *Commun. Assoc. Comput. Mach.* **25,** 134–141.