Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

An asymptotic analysis of closed queueing networks with branching populations

N. Bayer, E.G. Coffman Jr. and Y.A. Kogan

Department of Operations Reasearch, Statistics, and System Theory

BS-R9524 1995

Report BS-R9524
ISSN 0924-0659

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

An Asymptotic Analysis of Closed Queueing Networks with Branching Populations

N. Bayer
CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

E. G. Coffman, Jr.
AT&T Bell Laboratories
Murray Hill, NJ 07974, USA

Y. A. Kogan
AT&T Bell Laboratories
Holmdel, NJ 07733, USA

Abstract

Closed queueing networks have proven to be valuable tools for system performance analysis. In this paper, we broaden the applications of such networks by incorporating populations of *branching* customers: whenever a customer completes service at some node of the network, it is replaced by $N \geq 0$ customers, each routed independently to a next node, where N has a given, possibly node-dependent branching distribution. Applications of these branching and queueing networks focus on *stable* regimes, where the customer population iteratively dies out in finite expected time; at customer depletions the network is immediately supplied with a new set of customers according to a given initial-state distribution. In our first result, we give a necessary and sufficient condition for stability. An analysis of queue-length distributions seems much more difficult, even under exponential assumptions. Homogeneous, infinite-server networks and n-server single node networks are exceptions for which we give exact results, but our main response to the added difficulty is to give an asymptotic analysis of exponential networks with slowly varying population sizes. We model the behavior of such networks by that of a sequence of closed queueing networks differing only in population size, in analogy with the approach of nearly completely decomposable systems. The theoretical foundations of this approximation technique require limit laws, which we provide for throughputs and for stationary queue-length distributions.

AMS Subject Classification (1991): 60K25, 60J80, 60G10, 60F99.

Keywords and Phrases: Closed queueing networks, branching processes, nearly completely decomposable systems.

Note: The first author is supported by the European grant BRA-QMIPS of CEC DG XIII.

1. Introduction

By generalizing customer behavior in closed queueing networks, this paper extends the usefulness of such networks as tools for the performance analysis of complex systems. A classical network structure is considered in which a set of nodes and a stochastic routing matrix are specified, with each node defined by a service time distribution and a number, possibly infinite, of identical servers. In the initial state, all servers are idle and the network queue lengths are sampled from a given joint distribution. Thereafter, the customer population evolves according to a branching process: whenever service is completed at some node, the departing customer is replaced by a number $N \geq 0$ of new customers. The new customers, if any, are immediately routed to other nodes of the network according to the given routing matrix, with each new customer being routed independently of the others. The customer branching determined by the random variables N is governed by given *branching distributions* which may vary from node to node.

The new networks are called *branching and queueing* networks, or simply BQ networks. In applications, the branching distributions are such that the network is transient; the population eventually dies out in finite expected time. When this happens, the network is instantaneously supplied with a new sample from the initial-state distribution, and the process regenerates. The iterations, each beginning with a fresh initial state and ending with a depletion of customers, partition the process into an sequence of cycles. The cycle initializations are i.i.d., so the sequence is i.i.d., and the count of cycles (customer depletions) is a renewal process.

There are several important applications of BQ networks, in which a served customer triggers a random number of new customers. These include fundamental processes of parallel computers, communication networks, and manufacturing systems such as those in the semiconductor industry. As this paper concentrates on the theory of BQ networks, we will not go into details; these are discussed at length in a companion paper [8].

Even with exponential service-time distributions, BQ networks will normally lack the properties that lead to explicit formulas for stationary queue-length distributions, in particular those having product form. Accordingly, we turn to an approximation scheme, one that exploits the theory of closed queueing networks to approximate BQ networks with slowly varying population sizes. The main results of this paper are obtained from an asymptotic analysis of this approximation.

To develop the approximation, we first introduce a parameter θ which varies the branching distribution of any given BQ network, $\mathcal{B}$. At a service completion in the parametrized network $\mathcal{B}_\theta$, the number of new customers is generated by the branching distribution of $\mathcal{B}$ with

probability θ, and is 1 with probability $1-\theta$, i.e., the transition is that of a closed queueing network with probability $1-\theta$. As θ is made smaller, the rate at which the population size of $\mathcal{B}_\theta$ varies becomes slower. We compute the limit $\theta \to 0$ of the networks $\mathcal{B}_\theta$, which gives us the desired approximation for networks having slowly varying population sizes. In analogy with nearly completely decomposable networks, the behavior in the limit $\theta \to 0$ is modeled by that of a sequence of closed queueing networks $\{\mathcal{C}_\ell\}$ indexed by population size, where $\mathcal{C}_\ell$ is simply $\mathcal{B}$ with a fixed population of ℓ customers; the branching distributions are all concentrated at 1. As θ approaches 0, the trajectories of the process can be partitioned into intervals such that, during each, the state transitions are governed by a network $\mathcal{C}_\ell$ for some $\ell \geq 1$. The transition rates from one $\mathcal{C}_\ell$ to another are $O(\theta)$.

As discussed later in the paper, the value $\theta = 0$ is a singular perturbation point; the small θ behavior is as above, but the network $\mathcal{B}_\theta$ with θ set to 0 is in fact simply a single closed queueing network. We remark also that, in terms only of estimating throughputs, the above approximation has proven to be quite useful for many networks that do *not* have slowly varying population sizes. Experiments supporting this fact are reported in [5].

The next section introduces notation and basic assumptions; it then proves a stability result that characterizes the BQ networks of interest in the remainder of the paper. Section 3 contains a number of exact results for specialized BQ networks. Sections 4 and 5 deal with the approximation to networks with slowly varying populations. Section 4 introduces further notation and presents the basic limit laws. The proofs of the results in Section 4 can be found in Section 5.

2. Preliminaries

A *BQ network* $\mathcal{B}$ is defined by seven quantities. The first four give its *configuration*, which also defines the structure of a conventional closed queueing network.

- $k \geq 1$ counts the number of nodes.
- $\boldsymbol{\Gamma} = (\gamma_{ij})$ is a $k \times k$ indecomposable stochastic routing matrix.
- $\mathbf{n} = (n_1, \ldots, n_k)$ is the vector of server numbers, i.e., $n_i \geq 1$ is the number of parallel identical servers at node i; $n_i = \infty$ denotes an infinite-server node.
- $\mathbf{F}_S = (F_S^1, \ldots, F_S^k)$ is a vector of absolutely continuous service-time distributions, each having a strictly positive, finite first moment.

The next two determine its initial state:

- r is a positive integer denoting the initial population size.
- F_I, the *initial spread distribution*, is a k-dimensional distribution function concentrated on the vectors with nonnegative integer components summing to r.

With r and F_I determining the initial number and locations of customers, the six quantities above define a closed queueing network which we call a CQ network, for short. The definition of the more general BQ network requires that one more vector be specified.

- $\mathbf{F}_N = (F_N^1, \ldots, F_N^k)$ is a vector of distribution functions concentrated on the nonnegative integers, each having a strictly positive first moment. F_N^i is the branching distribution at node i, $1 \leq i \leq k$. Thus, samples from F_N^i give the number of new customers generated by departures at node i; each such customer is routed independently according to the distribution in row i of $\boldsymbol{\Gamma}$.

Introduce $\mathbf{b} = (b_1, \ldots, b_k)$, where b_i is the mean of F_N^i, and note the assumption above that $b_i > 0$ for all i, $1 \leq i \leq k$; BQ networks having absorbing nodes ($b_i = 0$) that never generate new customers can be handled by an easy extension of the analysis in this paper.

Embedded in a BQ network is a multitype Galton-Watson branching process (see e.g. [10]) corresponding to each initial customer, where the type of a customer is the identity of the node where it is served. If $\mathbf{Z}_0 = \mathbf{e}_i$, where $\mathbf{e}_i$ denotes a unit k-vector with the 1 in the i^{th} component, then $\{\mathbf{Z}_n, n \geq 0\}$ denotes such a process for an initial customer at node i. $\mathbf{Z}_n = (Z_n^1, \ldots, Z_n^k)$ is the n^{th} generation with Z_n^j denoting the number of customers of type j, $1 \leq j \leq k$. Because of the applications being modeled, our interest is confined to networks in which the branching process $\{\mathbf{Z}_n\}$ generates, for each of the initial r customers, a total population with finite expected size; we say that such networks are *proper*. Proper networks become depleted of customers in finite expected time, and induce an over-all regenerative process in which the count of depletions is a renewal process.

An application of the theory of multitype Galton-Watson processes gives a necessary and sufficient condition for a BQ network to be proper. Almost-sure extinction is a standard, very similar property of the processes $\{\mathbf{Z}_n\}$, but it has a slightly different characterization. Nevertheless, it will be useful in proving the theorem below. Let $\rho(\cdot)$ denote the spectral radius of a matrix.

Theorem 2.1. *A BQ network is proper if and only if* $\rho(\text{diag}\,(\mathbf{b})\boldsymbol{\Gamma}) < 1$.

Proof: First observe that $\text{diag}\,(\mathbf{b})\boldsymbol{\Gamma}$ is the matrix $\mathbf{H} = (h)_{ij}$ of the associated Galton-Watson process defined by $h_{ij} = E[Z_1^j \mid \mathbf{Z}_0 = \mathbf{e}_i]$. It follows from Proposition 4.1 and Theorem 10.1 in

[10, p. 46] that the extinction probability is 1 if and only if $\rho(\mathbf{H}) \leq 1$. Thus, since a population extinction has probability 1 in a proper BQ network, we may restrict the proof to BQ networks with $\rho(\mathbf{H}) \leq 1$.

Let f^i be the (k-dimensional) probability generating function of $\mathbf{Z}_1$ given $\mathbf{Z}_0 = \mathbf{e}_i$, and let g^i be the probability generating function of the total population given the same initial condition. For the purpose of this proof it suffices to assume that both are real functions, g^i being defined for every $z \in [0,1]$, and f^i for every $\mathbf{z} \in [0,1]^k$. Define $\mathbf{g} : [0,1] \to [0,1]^k$ and $\mathbf{f} : [0,1]^k \to [0,1]^k$ as the vector functions whose coordinates are the g^i and f^i. Observe that $\mathbf{g}$ is a nondefective generating function by virtue of the almost-sure extinction; therefore its derivatives can be used for the calculation of moments.

The familiar relation

$$\mathbf{g}(z) = z\mathbf{f}(\mathbf{g}(z)), \quad z \in [0,1] \ , \tag{2.1}$$

is obtained by observing that the total population is one plus the sum of the populations headed by the first generation customers. Differentiation of (2.1) yields

$$\mathbf{g}'(z) = \mathbf{f}(\mathbf{g}(z)) + z\mathbf{f}'(\mathbf{g}(z))\mathbf{g}'(z), \quad z \in [0,1) \ , \tag{2.2}$$

where $\mathbf{f}'$ is the Jacobian. The matrix $\mathbf{I} - z\mathbf{f}'(\mathbf{g}(z))$ is nonsingular for every $z < 1$, since

$$\rho(z\mathbf{f}'(\mathbf{g}(z))) < \rho(\mathbf{f}'(\mathbf{g}(z))) \leq \rho(\mathbf{f}'(\mathbf{g}(1))) = \rho(\mathbf{H}) \leq 1 \ .$$

Hence $\mathbf{g}'(z)$ can be extracted from (2.2). This gives

$$\mathbf{g}'(z) = [\mathbf{I} - z\mathbf{f}'(\mathbf{g}(z))]^{-1}\mathbf{f}(\mathbf{g}(z)) = \left\{\sum_{n=0}^{\infty} z^n [\mathbf{f}'(\mathbf{g}(z))]^n\right\} \mathbf{f}(\mathbf{g}(z)), \quad z \in [0,1) \ . \tag{2.3}$$

The expectations are given by the limit of (2.3) as $z \uparrow 1$. If $\rho(\mathbf{H}) < 1$, the limit is the finite and positive vector

$$(\mathbf{I} - \mathbf{H})^{-1}\mathbf{1} = \left(\sum_{n=0}^{\infty} \mathbf{H}^n\right) \ \mathbf{1} \ .$$

It can be seen that, otherwise, the rightmost expression in (2.3) diverges. ∎

Remark. The b_i's and the spectral radius of $\mathrm{diag}\,(\mathbf{b})\boldsymbol{\Gamma}$ have an intuitive interpretation in an associated fluid network model. Let the nodes be k pumping stations interconnected by pipelines circulating a fluid in the closed network. Suppose there is also a reservoir connected to all of the stations. Assume that the network is in steady state, so that the flows and reservoir level are constants. Let w_i be the total output flow from station i to the other stations, and let γ_{ij} be the fraction of w_i directed to station j. The total flow, say u_i, into station i from

other stations may differ from w_i because of a flow to or from the reservoir. Assume that $w_i = (b_i/a)u_i$; that is, the means b_i of the branching distributions act in this fluid model as amplification factors, up to some normalization constant a. Then $a = \rho(\mathrm{diag}\,(\mathbf{b})\boldsymbol{\Gamma})$. This follows from the above formulation through the relation

$$\mathbf{u} = \frac{1}{a}\mathbf{u}\ \mathrm{diag}\,(\mathbf{b})\boldsymbol{\Gamma}\ ,$$

which says that $\mathbf{u} = (u_1, \ldots, u_k)$ is a nonnegative left eigenvector of $\mathrm{diag}\,(\mathbf{b})\boldsymbol{\Gamma}$ with eigenvalue a. Further, it can be checked directly that if $\mathbf{u}$ is normalized ($\mathbf{u}\cdot\mathbf{1} = \mathbf{1}$), then $a = \mathbf{u}\cdot\mathbf{b}$. In the asymptotic analysis of Section 4, we require that $\mathbf{v}\cdot\mathbf{b} < 1$, where $\mathbf{v}$ is the Perron-Frobenius eigenvector of $\boldsymbol{\Gamma}$. This can be viewed as a limiting version of the condition $\mathbf{u}\cdot\mathbf{b} < 1$ for a proper network, since the vector $\mathbf{u}$ of flows into the stations approaches $\mathbf{v}$, and its scalar product with $\mathbf{b}$ is kept less than unity. ∎

Next, consider a BQ network $\mathcal{B}$ with a slowly varying population size. To obtain an approximation for $\mathcal{B}$ that is supported by a limit law, i.e., one that is asymptotically exact, we introduce a family of networks $\{\mathcal{B}_\theta, 0 < \theta \leq 1\}$; $\mathcal{B}_\theta$ has the same parameters as $\mathcal{B}$ except for the branching distributions, which are given by the vector $\theta\mathbf{F}_N + (1-\theta)\mathbf{F}_1$, where $\mathbf{F}_1$ is a k-vector of distributions F_1, each concentrated at 1. For small θ, a proper network $\mathcal{B}_\theta$ is likely to dwell over long time intervals in states having the same population size. When the intervals are long enough, the behavior of $\mathcal{B}_\theta$ during each such interval can be approximated by the stationary behavior of a CQ network with the same configuration as $\mathcal{B}$; the population size of the CQ network is the same as the one that prevails during the interval to which it applies.

Although the BQ networks that have been used to model practical systems have all been uniformly proper in θ, examples of proper BQ networks can be constructed that do not have this property. In particular, families of networks $\{\mathcal{B}_\theta\}$ can be found such that the value of the spectral radius of the matrix $\mathbf{H}_\theta = \mathrm{diag}\,(\theta\mathbf{b}+(1-\theta)\mathbf{1})\boldsymbol{\Gamma}$ (cf. Theorem 2.1) alternates in regions where it exceeds 1 and is less than 1 as $\theta \to 0$ [8]. We leave as an open problem the analysis of BQ networks not uniformly proper in θ; *in the remainder of the paper, all BQ networks are assumed to be uniformly proper in θ. Furthermore, exponential service times at each node are assumed.* The latter assumption simplifies the asymptotic analysis of Sections 4 and 5. But it has an equally important role in applying the results of these sections, as it allows one to exploit the computational tools in the existing theory of CQ networks. We return to this point in Section 4.

3. Exact Results

Explicit formulas are obtainable in interesting special cases, as shown in this section. We say that a BQ network is uniform in one of the parameters $\boldsymbol{\Gamma}$, $\mathbf{n}$, $\mathbf{F}_S$, $\mathbf{F}_N$ if all components of that parameter are identical. Though interesting in its own right, our first result will be most useful in the proofs of later theorems, both in this section and in the subsequent sections on the asymptotic analysis of the networks $\{\mathcal{B}_\theta, 0 < \theta \leq 1\}$.

Let $\{p_i, i \geq 1\}$ be the steady-state distribution of the process $\{L(t), t \geq 0\}$, where $L(t)$ gives the population size at time t; recall that $L(0) = r$ is the initial population size. Let L_j denote the population size just after the j^{th} service completion, with $L_0 = r$ as before. The stationary distribution of the embedded chain $\{L_j\}$ is denoted by $\{\tilde{p}_i, i \geq 1\}$, and the generating functions for the p_i and $\tilde{p}_i$ are denoted by $g_L(z)$ and $\tilde{g}_L(z)$, respectively.

The state transitions of the above processes are clearly state-dependent. Moreover, in general, these processes are not Markovian. However, if the BQ network is uniform in $\mathbf{F}_N$, then the chain $\{L_j\}$ is Markovian, and the state dependence exists only in the state with exactly one customer; if no new customers are generated by a departure in this state, then the transition will be to a new initial state.

Theorem 3.1. *If $\mathcal{B}$ is uniform in $\mathbf{F}_N$, then*

$$\tilde{g}_L(z) = \frac{1-b}{r} \frac{z^r - 1}{1 - \frac{1}{z} g_N(z)} ,$$

where g_N is the generating function and b the mean of the branching distribution F_N common to all nodes.

Proof: The evolution equation of $\{L_j\}$ is

$$L_{j+1} = L_j - 1 + N_{j+1} + r \cdot 1_{\{L_j - 1 + N_{j+1} = 0\}} , \tag{3.1}$$

where N_j is the number of new customers generated at the j^{th} service completion. The generating function of the two rightmost terms on the right of (3.1), conditioned on $L_j = \ell$, can be written separately for $\ell = 1$ and for $\ell > 1$, and then recombined to give

$$\begin{aligned} &E(z^{N_{j+1} + r \cdot 1_{\{L_j - 1 + N_{j+1} = 0\}}} \mid L_j = \ell) \\ &\quad = E(z^{N_{j+1}} \mid L_j = \ell) + 1_{\{\ell = 1\}} (z^r - 1) P(N_{j+1} = 0 \mid L_j = \ell), \quad \ell \geq 1 . \end{aligned} \tag{3.2}$$

Introducing (3.2) into the generating function for (3.1) conditioned on L_j then gives

$$E(z^{L_{j+1}} \mid L_j) = z^{L_j - 1} E(z^{N_{j+1}} \mid L_j) + 1_{\{L_j = 1\}} (z^r - 1) P(N_{j+1} = 0 \mid L_j) . \tag{3.3}$$

A comparison of the expectations of the two sides of (3.1) conditioned on L_j shows that

$$1_{\{L_j=1\}}P(N_{j+1}=0 \mid L_j) = \frac{1}{r}[E(L_{j+1} \mid L_j) - L_j + 1 - E(N_{j+1} \mid L_j)] \ ,$$

so that (3.3) can be written

$$E(z^{L_{j+1}} \mid L_j) = z^{L_j-1}E(z^{N_{j+1}} \mid L_j) + \frac{1}{r}(z^r-1)[E(L_{j+1} \mid L_j) - L_j + 1 - E(N_{j+1} \mid L_j)] \ .$$

Thus, since the number of customers generated by a departure is state independent (by the uniformity in $\mathbf{F}_N$),

$$E(z^{L_{j+1}} \mid L_j) = z^{L_j-1}g_N(z) + \frac{1}{r}(z^r-1)[E(L_{j+1} \mid L_j) - L_j + 1 - b] \ .$$

Removing the conditioning and taking the limit $j \to \infty$ then gives

$$\tilde{g}_L(z) = \frac{1}{z}\tilde{g}_L(z)g_N(z) + \frac{1}{r}(z^r-1)(1-b) \ .$$

Solving for $\tilde{g}_L(z)$ proves the theorem. (See also [6] for a special, state-dependent case that leads to a closed-form solution.) ∎

Under the conditions of the theorem, one obtains that, although the parameter θ affects the *rate* of transitions from one population size to another, it does not affect the stationary distribution of population size at service completion epochs. Indeed, substituting $\theta g_N(z) + (1-\theta)z$ and $\theta b + (1-\theta)$ for $g_N(z)$ and b in Theorem 3.1 proves

Corollary 3.2. *Let $\mathcal{B}$ be uniform in F_N. Then the stationary population-size distribution of $\mathcal{B}_\theta$ at service completion epochs is independent of θ and given by Theorem 3.1.*

The results for special networks given below rely on the embedded-chain probabilities $\tilde{p}_i$, whose generating function is calculated in Theorem 3.1. First, the single-node network is a simple, but useful special case (see [5] for applications); in this case, the population size alone gives the network state. The proof of the following result is a standard argument, and hence omitted.

Corollary 3.3. *For a BQ network consisting of a single n-server node, let μ be the parameter of the exponential service-time distribution, and define $\tau_j = j\mu$ if $1 \le j \le n$, and $\tau_j = n\mu$ if $j > n$. Then the stationary distribution is*

$$p_j = \frac{\tilde{p}_j/\tau_j}{\sum_{j\ge 1}\tilde{p}_j/\tau_j} \ , \quad j \ge 1 \ ,$$

where the $\tilde{p}_j$ are given by Theorem 3.1.

A solution is also available for an infinite-server BQ network if the initial spread distribution F_I satisfies a special condition and if the network is homogeneous, i.e., if the network is uniform in the parameters $\boldsymbol{\Gamma}$, $\mathbf{F}_S$, and $\mathbf{F}_N$, as well as $\mathbf{n}$, so that the nodes of the network are indistinguishable. Because of the restrictive assumptions, the result below is primarily of theoretical interest; it illustrates the assumptions needed for product form solutions.

In what follows, $\mathbf{0}$ and $\mathbf{1}$ denote the k-vectors of all 0's and all 1's, respectively. $\mathbb{Z}^k$ denotes the space of k-vectors with integer components, and $\mathbb{Z}_+^k$ denotes its nonnegative orthant. Define $\mathbb{Z}_\oplus^k = \mathbb{Z}_+^k - \{\mathbf{0}\}$, the state space of a BQ network. Finally, $\mathbb{Z}_+^{k,m} = \{\mathbf{q} \in \mathbb{Z}_+^k : \mathbf{q}\cdot\mathbf{1} = m\}$ is the m^{th} simplex of $\mathbb{Z}_+^k$. Note that $\mathbb{Z}_+^{k,r}$ is the support of F_I.

Theorem 3.4. *Let $\mathcal{B}$ be an infinite-server, homogeneous BQ network, and suppose that, on $\mathbb{Z}_+^{k,r}$, F_I is parallel to the distribution*

$$\pi(\mathbf{q}) = \frac{(\mathbf{q}\cdot\mathbf{1})!\tilde{p}_{\mathbf{q}\cdot\mathbf{1}}}{(\mathbf{q}\cdot\mathbf{1})k^{\mathbf{q}\cdot\mathbf{1}}\sum_{j=1}^{\infty}\tilde{p}_j/j}\cdot\frac{1}{\prod_{i=1}^k q_i!}, \quad \mathbf{q} = (q_1,\ldots,q_k) \in \mathbb{Z}_\oplus^k\ , \tag{3.4}$$

with the $\tilde{p}_j$ given by Theorem 3.1. Then (3.4) is the stationary distribution of $\mathcal{B}$.

Remark. Note that the first factor on the right of (3.4) depends on $\mathbf{q}$ only through the population size $\mathbf{q}\cdot\mathbf{1}$. Thus, given the population size, $\{\pi(\mathbf{q})\}$ has a conditional product form, and is parallel to the stationary distribution of a CQ-network with the same configuration as $\mathcal{B}$. ∎

Proof: For convenience, we take the service rate to be 1. It is readily checked that the distribution (3.4) is normalized, so it remains to verify that the global balance equations

$$\pi(\mathbf{q})\sum_{\mathbf{q}'\neq\mathbf{q}}\pi(\mathbf{q},\mathbf{q}') = \sum_{\mathbf{q}'\neq\mathbf{q}}\pi(\mathbf{q}')\pi(\mathbf{q}',\mathbf{q}), \quad \mathbf{q}\in Z_\oplus^k \tag{3.5}$$

are satisfied. Note that local balance does not hold, and that the network is not reversible.

Let $\{\sigma_m; m\geq 0\}$ be the probability mass function of the branching distribution F_N common to all nodes. Suppose that a transition from state $\mathbf{q}'$ to $\mathbf{q}$ is possible following a departure at node j. Then it is possible to write $\mathbf{q} = \mathbf{q}' - \mathbf{e}_j + \mathbf{a}$ for some $\mathbf{a}\in\mathbb{Z}_+^k$. This contribution $\pi_j(\mathbf{q}',\mathbf{q})$ to $\pi(\mathbf{q}',\mathbf{q})$ is given by

$$\pi_j(\mathbf{q}',\mathbf{q}) = (q_j+1-a_j)\sigma_{\mathbf{a}\cdot\mathbf{1}}\frac{(\mathbf{a}\cdot\mathbf{1})!}{k^{\mathbf{a}\cdot\mathbf{1}}\prod_{i=1}^k a_i!}\ . \tag{3.6}$$

The three factors in (3.6) are the service rate at node j, the probability of generating the appropriate number of new customers, and the probability of their spread, respectively. Define

$\ell \equiv \mathbf{q} \cdot \mathbf{1}$. The cumulative departure rate on the left of (3.5) is ℓ, so substitution of (3.4) and (3.6) into (3.5) and slight rearrangement leads to

$$\frac{\tilde{p}_\ell \ell!}{k^\ell \prod_{i=1}^k q_i!} =$$

$$\sum_{m=0}^{\ell} \sum_{\{\mathbf{a} \in \mathbb{Z}_+^{k,m} \mid \mathbf{a} \le \mathbf{q}\}} \underbrace{\sum_{j=1}^{k} \frac{\tilde{p}_{\ell-m+1}(\ell-m+1)!}{(\ell-m+1)k^{\ell-m+1}[\prod_{i=1}^k (q_i-a_i)!](q_j-a_j+1)}}_{\pi(\mathbf{q}')} \cdot \tag{3.7}$$

$$\underbrace{(q_j-a_j+1)\sigma_m \frac{m!}{k^m \prod_{i=1}^k a_i!}}_{\pi_j(\mathbf{q}',\mathbf{q})} + \underbrace{1_{\{\ell=r\}} \tilde{p}_1 \sigma_0 \frac{r!}{k^r \prod_{i=1}^k q_i!}}_{\text{reinitialization}}, \quad \mathbf{q} = (q_1, \ldots, q_k) \in \mathbb{Z}_\oplus^k ,$$

where $\mathbf{a} \le \mathbf{q}$ is interpreted component-wise. But by a standard property of multinomially distributed random variables (see, e.g., [12, p. 659]),

$$\sum_{\{\mathbf{a} \in \mathbb{Z}_+^{k,m} \mid \mathbf{a} \le \mathbf{q}\}} \frac{(\ell-m)!}{\prod_{i=1}^k (q_i-a_i)!} \cdot \frac{m!}{\prod_{i=1}^k a_i!} = \frac{\ell!}{\prod_{i=1}^k q_i!}.$$

Thus, (3.7) reduces to

$$\tilde{p}_\ell = 1_{\{\ell=r\}} \tilde{p}_1 \sigma_0 + \sum_{m=0}^{\ell} \tilde{p}_{\ell-m+1} \sigma_m,$$

which is the balance equation of the embedded chain $\{L_j\}$ in Theorem 3.1. ∎

4. Asymptotic Results

In this section we state the asymptotic results, deferring the proofs to Section 5. We comment briefly on the ideas and techniques of the proofs, focusing on the over-all argument leading to the main result, the limit law in Theorem 4.4. We then close this section with a brief discussion of computational issues.

Our treatment is analogous to eigenvector approximation for nearly completely decomposable Markov chains as developed by Courtois [9, p. 20]. A nearly completely decomposable Markov chain is described by a transition-probability matrix having the form

$$\mathbf{K}^\epsilon = \mathbf{K}^* + \epsilon \mathbf{C} = \begin{bmatrix} \mathbf{K}_1^* & \epsilon \mathbf{C}_{12} & \ldots & \epsilon \mathbf{C}_{1n} \\ \epsilon \mathbf{C}_{21} & \mathbf{K}_2^* & \ldots & \epsilon \mathbf{C}_{1n} \\ \vdots & \vdots & & \vdots \\ \epsilon \mathbf{C}_{n1} & \epsilon \mathbf{C}_{n2} & \ldots & \mathbf{K}_n^* \end{bmatrix} \tag{4.1}$$

where $\mathbf{K}^*$ is a stochastic block-diagonal matrix and ϵ is a small positive parameter. As ϵ approaches zero, the exit rates out of the groups of states corresponding to the blocks of $\mathbf{K}^*$ diminish. The system then tends to dwell in these groups of states, which we call *macro-states*. The objective here is the limiting invariant vector of $\mathbf{K}^\epsilon$, $\epsilon \to 0$. As indicated in

[9], the calculation of this limit has two ingredients: The first calculates the invariant vectors of the blocks $\mathbf{K}_i^*$; these vectors give the limiting invariant probabilities conditioned on the macro-state. The second ingredient calculates the limiting macro-state stationary probability distribution.

In our setting, the state space is infinite, and the small parameter is θ, which modifies the branching distributions in a family $\{\mathcal{B}_\theta, 0 < \theta \leq 1\}$ of BQ networks. A macro-state is thus characterized by some fixed total population. The transition matrix of the network $\mathcal{B}_\theta$ can be written explicitly in a form similar to (4.1), with θ playing the role of ϵ. However, this is of little help here, since the analysis available in the literature (see, e.g., Aven et al. [3, p. 163]) relies heavily on the fact that $\mathbf{K}^\epsilon$ in (4.1) is a *finite* matrix. In that case one can express the invariant probability vector in terms of the elements of $\mathbf{K}^\epsilon$ and use this expression to obtain the limiting probability vector as ϵ tends to zero. When $\mathbf{K}^\epsilon$ is infinite, as here, we first have to prove the existence of the limiting invariant vector; we then adopt a probabilistic approach rather than an algebraic one.

Let $\mathbf{Q}^{\mathcal{B}}(t)$ denote the k dimensional queue-length process of the BQ network $\mathcal{B}$; the i^{th} component, $1 \leq i \leq k$, gives the number of customers present at node i (waiting or being served) at time t. As in the previous section, $L^{\mathcal{B}}(t) = \mathbf{Q}^{\mathcal{B}}(t) \cdot \mathbf{1}$ denotes the total-population process, or equivalently, the macro-state process. We emphasize that, except where noted otherwise, *the results and analysis of this and the next section are restricted to stationary versions of the queue-length and macro-state processes in uniformly proper BQ networks.* Thus, we assume the necessary and sufficient condition established by Theorem 2.1 for proper networks, and hence for the existence of a stationary regime. The stationary processes $\mathbf{Q}^{\mathcal{B}}(t)$ and $L^{\mathcal{B}}(t)$ are defined over $-\infty < t < \infty$. The chains embedded at service completion epochs are denoted by $\mathbf{Q}_j^{\mathcal{B}}$ and $L_j^{\mathcal{B}}$, $j = 0, \pm 1, \pm 2, \ldots$; by convention the indexing is chosen so that $\mathbf{Q}_1^{\mathcal{B}}$ is the queue length at the first service completion after time 0. The $\mathbf{Q}_j^{\mathcal{B}}$ give the queue lengths just after samples are drawn from the branching distributions at service completion epochs. Although sampled from continuous-time stationary processes, the discrete sequences $\{\mathbf{Q}_j\}$ and $\{L_j\}$ need not themselves be stationary. In fact, the very need to select an indexing scheme dictates that the distribution of $\mathbf{Q}_0$, for example, must differ from the steady-state distribution of $\mathbf{Q}_j$ as $j \to \infty$. When dealing with distributions of discrete–time processes, we will be interested solely in their steady-state.

Throughout the remainder of the paper, $\mathcal{C}_\ell$ denotes a CQ network with population size ℓ and the same configuration as some given BQ network $\mathcal{B}$; the underlying network $\mathcal{B}$ will always be clear in context. The results to follow all describe the asymptotic behavior of BQ networks

$\mathcal{B}_\theta$ in the limit of slowly varying population sizes; this limit, $\theta \to 0$, is a tacit assumption in all of the convergence results of this and the next section.

The first ingredient of the analysis determines the asymptotics of the state distribution at service completion epochs, conditioned on population size; this ingredient establishes the fundamental connection between BQ and CQ networks in the limit $\theta \to 0$.

Lemma 4.1. *For every* $\ell \geq 1$ *and* $\mathbf{q} \in \mathbb{Z}_+^{k,\ell}$,

$$\lim_{j\to\infty} P\{\mathbf{Q}_j^{\mathcal{B}_\theta} = \mathbf{q} \,|\, L_j^{\mathcal{B}_\theta} = \ell\} \to \lim_{j\to\infty} P\{\mathbf{Q}_j^{\mathcal{C}_\ell} = \mathbf{q}\}.$$

The proof (in Section 5) makes use of the fact that the distribution of $\mathbf{Q}_i^{\mathcal{C}_\ell}$, given any initial distribution, converges geometrically fast in i to the stationary distribution.

The second ingredient of the analysis determines the probability of being in any macro-state. We first consider the stationary distribution of the chain $\left\{L_j^{\mathcal{B}_\theta}\right\}$. When the network is uniform in the branching distributions $\mathbf{F}_N$, this chain is Markovian and the generating function of the desired distribution is given by Theorem 3.1. In the generalization to networks which are not necessarily uniform in $\mathbf{F}_N$, the *visit ratio* vector $\mathbf{v}$ appears; $\mathbf{v}$ is the normalized invariant vector of the routing matrix $\boldsymbol{\Gamma}$, i.e., $\mathbf{v}$ satisfies $\mathbf{v}\boldsymbol{\Gamma} = \mathbf{v}$ and $\mathbf{v} \cdot \mathbf{1} = 1$. Recall that $\mathbf{g}_N(z)$ is the vector generating function for $\mathbf{F}_N$.

Theorem 4.2. *Assume that* $\mathbf{v} \cdot \mathbf{b} < 1$. *Then the stationary distribution of the chain* $\left\{L_j^{\mathcal{B}_\theta}\right\}$ *converges element-wise and in expectation to the distribution* $\{\tilde{p}_\ell\}$ *generated by*

$$\tilde{g}_L(z) = \frac{1 - \mathbf{v}\cdot\mathbf{b}}{r} \frac{z^r - 1}{1 - \frac{1}{z}\mathbf{v}\cdot\mathbf{g}_N(z)} \ .$$

The proof introduces a transition matrix under which the stationary distribution of $\left\{L_j^{\mathcal{B}_\theta}\right\}$ is invariant. This transition matrix converges to a limit. We then rely on the continuity of the invariant vector of an infinite stochastic matrix as a function of the matrix elements, a property that holds when the first moment of the invariant vector is bounded.

It remains to combine the two ingredients in Lemma 4.1 and Theorem 4.2 and to extend the distributions limited to service completion epochs to those valid at any fixed time. To this end, we need first a limit theorem for the over-all rate of departures from a BQ network $\mathcal{B}_\theta$, i.e., the intensity of the stationary point process $\left\{T_j^{\mathcal{B}_\theta}\right\}$, where $T_j^{\mathcal{B}_\theta}$ is departure epoch j, $j = 0, \pm 1, \pm 2, \ldots$. By our indexing convention, $T_1^{\mathcal{B}_\theta}$ is the first positive point of the process. The intensity, denoted by $\lambda_{\mathcal{B}_\theta}$, represents the over-all throughput of the network.

Theorem 4.3. *If* $\mathbf{v} \cdot \mathbf{b} < 1$, *then* $\lambda_{\mathcal{B}_\theta} \to \tilde{\lambda}_{\mathcal{B}}$ *with*

$$\tilde{\lambda}_{\mathcal{B}} = \frac{1}{\sum_{\ell=1}^{\infty} (\tilde{p}_\ell / \lambda_{C_\ell})},$$

where $\{\tilde{p}_\ell\}$ is the distribution determined by Theorem 4.2, and where λ_{C_ℓ} is the overall throughput in the CQ network C_ℓ.

The following theorem gives the limiting state distribution. It combines the two ingredients of the near-complete decomposability analysis.

Theorem 4.4. *If $\mathbf{v} \cdot \mathbf{b} < 1$, and if we let $\ell = \ell(\mathbf{q}) = \mathbf{q} \cdot \mathbf{1}$, then*

$$P\{\mathbf{Q}^{\mathcal{B}_\theta}(t) = \mathbf{q}\} \to \tilde{\lambda}_{\mathcal{B}} \tilde{p}_\ell \frac{1}{\lambda_{\mathcal{C}_\ell}} P\{\mathbf{Q}^{\mathcal{C}_\ell}(t) = \mathbf{q}\} \ ,$$

for every $\mathbf{q} \in \mathbb{Z}_\oplus^k$. (Note that such an element-wise convergence of distributions over a countable space is equivalent to uniform convergence [2, p. 306]).

In our setting, the inversion formula of the Palm calculus [4, p. 23] expresses the continuous-time state distribution of a stationary queueing process in terms of the corresponding distribution restricted to the epochs of a given embedded chain. The proof of Theorem 4.4 exploits this relationship by converting arguments about the continuous-time distribution to arguments about Palm probabilities.

Note that in Theorem 4.2 we had the macro-state probabilities only at service completion epochs; by summing the result in Theorem 4.4 over all $\mathbf{q}$, we have them at an arbitrary time.

Corollary 4.5. *The limiting probabilities of the process $L^{\mathcal{B}_\theta}(t)$ are given by*

$$p_\ell = \tilde{\lambda}_{\mathcal{B}} \tilde{p}_\ell \frac{1}{\lambda_{\mathcal{C}_\ell}}, \quad \ell \geq 1 \ .$$

The next section also proves the following result on throughputs.

Corollary 4.6. *The throughput at any node i, $1 \leq i \leq k$, converges to $v_i \tilde{\lambda}_{\mathcal{B}}$, where v_i is the i^{th} component of the visit ratio vector $\mathbf{v}$.*

The results of this section provide an approximation for the state distributions and throughputs of BQ networks with slowly varying population sizes. The approximation is computed in terms of (a) the corresponding, well–known results for a series of product-form CQ networks, and (b) a single one-dimensional generating function. The throughput of a product-form CQ network can be computed by the recursive method described in [11, p. 170]. This technique is well suited for handling a series of networks; when applied to the network with maximum population size, the results for the remaining networks are obtained as a by-product. The inversion of the one-dimensional generating function required in (b) is a standard, computationally cheap numerical procedure (see e.g. [1]).

5. Proofs

Let M_j^θ, $j = 0, \pm 1, \ldots$, denote i.i.d. Bernoulli random variables with parameter $\theta = P(M_j^\theta = 1)$. It is convenient to suppose that, in generating the trajectories of a network $\mathcal{B}_\theta$, the samples from the branching distribution $\theta \mathbf{F}_N + (1-\theta)\mathbf{F}_1$ are constructed as follows from the M_j^θ: if service completion j takes place at node i, then the number of new customers replacing the departure is a sample from F_N^i if $M_j^\theta = 1$, and is 1 if $M_j^\theta = 0$. Let Y_j^θ denote the length of the current run of unsuccessful trials with $M_i^\theta = 0$, i.e., $Y_j^\theta = j - \sup\{i \le j | M_i^\theta = 1\}$, and note that Y_j^θ is geometrically distributed with parameter $1-\theta$. In the steady-state, $L_j^{\mathcal{B}_\theta}$ and M_j^θ are independent since

$$\lim_{j\to\infty} P\{L_j^{\mathcal{B}_\theta} = \ell | M_j^\theta = 0\} = \lim_{j\to\infty} P\{L_{j-1}^{\mathcal{B}_\theta} = \ell\} = \lim_{j\to\infty} P\{L_j^{\mathcal{B}_\theta} = \ell\} \ .$$

But then $L_j^{\mathcal{B}_\theta}$ and Y_j^θ must also be independent, since

$$\begin{aligned} \lim_{j\to\infty} P\{L_j^{\mathcal{B}_\theta} = \ell | Y_j^\theta = y\} &= \lim_{j\to\infty} P\{L_{j-y}^{\mathcal{B}_\theta} = \ell | M_{j-y}^\theta = 1, M_{j-y+1}^\theta = \cdots = M_j^\theta = 0\} \\ &= \lim_{j\to\infty} P\{L_{j-y}^{\mathcal{B}_\theta} = \ell | M_{j-y}^\theta = 1\} = \lim_{j\to\infty} P\{L_{j-y}^{\mathcal{B}_\theta} = \ell\} = \lim_{j\to\infty} P\{L_j^{\mathcal{B}_\theta} = \ell\} \ . \end{aligned}$$

Lemma 4.1 is restricted to the queueing process at departure epochs. However, to prove the remaining results of Section 4, it is convenient to have the extension of Lemma 4.1 to the joint process $(\mathbf{Q}_j^{\mathcal{B}_\theta}, \mathbf{R}_j^{\mathcal{B}_\theta})$, $j = 0, \pm 1, \pm 2, \ldots$, where the structure $\mathbf{R}_j^{\mathcal{B}_\theta}$ gives the residual service times of the customers still being served at departure epoch j. If $(\mathbf{R}_j^{\mathcal{B}_\theta})_{i_1,i_2}$ is positive, then it gives the residual service time of the customer at server i_2, $1 \le i_2 \le n_{i_1}$, of node i_1, $1 \le i_1 \le k$. Otherwise element (i_1, i_2) has the value zero, meaning that server i_2 at node i_1 is idle. For the sake of definiteness, assume that a customer always chooses the free server with the lowest index. The desired extension is: if $\mathbf{v} \cdot \mathbf{b} < 1$ then

$$\lim_{j\to\infty} P\{\mathbf{Q}_j^{\mathcal{B}_\theta} = \mathbf{q}, \mathbf{R}_j^{\mathcal{B}_\theta} \in \cdot | L_j^{\mathcal{B}_\theta} = \ell\} \to \lim_{j\to\infty} P\{\mathbf{Q}_j^{\mathcal{C}_\ell} = \ell, \mathbf{R}_j^{\mathcal{C}_\ell} \in \cdot\} \ , \tag{5.1}$$

where convergence is in the total variation norm. Since we are assuming exponential service times, (5.1) is an easy extension of Lemma 4.1.

Proof of Lemma 4.1. We prove that, for any $\ell \ge 1$ and state $\mathbf{q}$,

$$\Delta^{\mathcal{B}_\theta}(\mathbf{q}) = \left| \lim_{j\to\infty} P\{\mathbf{Q}_j^{\mathcal{B}_\theta} = \mathbf{q} | L_j^{\mathcal{B}_\theta} = \ell\} - \lim_{j\to\infty} P\{\mathbf{Q}_j^{\mathcal{C}_\ell} = \mathbf{q}\} \right| \to 0 \ , \text{ as } \theta \to 0. \tag{5.2}$$

First, condition on the run length $Y_j^{\mathcal{B}_\theta}$ and write

$$P\{\mathbf{Q}_j^{\mathcal{B}_\theta} = \mathbf{q} | L_j^{\mathcal{B}_\theta} = \ell\} = \sum_{y=0}^{\infty} P\{\mathbf{Q}_j^{\mathcal{B}_\theta} = \mathbf{q} | L_j^{\mathcal{B}_\theta} = \ell, Y_j^{\mathcal{B}_\theta} = y\} P\{Y_j^{\mathcal{B}_\theta} = y | L_j^{\mathcal{B}_\theta} = \ell\}. \tag{5.3}$$

Now apply the geometric law

$$P\{Y_j^{\mathcal{B}_\theta} = y\} = (1-\theta)\theta^y,$$

the independence of $L_j^{\mathcal{B}_\theta}$ and $Y_j^{\mathcal{B}_\theta}$, and the geometrically fast convergence of $\{\mathbf{Q}_j^{\mathcal{C}_\ell}\}$ to obtain

$$\left|\lim_{j\to\infty} P\{\mathbf{Q}_j^{\mathcal{B}_\theta} = \mathbf{q} | L_j^{\mathcal{B}_\theta} = \ell,\ Y_j^{\mathcal{B}_\theta} = y\} - \lim_{j\to\infty} P\{\mathbf{Q}_j^{\mathcal{C}_\ell} = \mathbf{q}\}\right| \le \alpha\beta^y \tag{5.4}$$

for some α and for some $\beta < 1$. Then

$$\Delta^{\mathcal{B}_\theta}(\mathbf{q}) \le \sum_{y=0}^{\infty} \alpha\beta^y(1-\theta)^y\theta\ ,$$

or

$$\Delta^{\mathcal{B}_\theta}(\mathbf{q}) \le \frac{\alpha\theta}{1-\beta(1-\theta)}\ ,$$

which converges to 0 as desired. ∎

Proof of Theorem 4.2. Before we begin the proof, we need a technical result that specializes a theorem of Kuijper [13, Theorem 6.56]. If $\mathbf{B}$ and $\mathbf{C}$ are matrices with the same number of rows, then $[\mathbf{BC}]$ denotes the matrix which is the concatenation of the columns of $\mathbf{B}$ and $\mathbf{C}$, where those of $\mathbf{C}$ are rightmost. Define $\ker\mathbf{B} = \{\mathbf{x}|\mathbf{Bx} = \mathbf{0}\}$ and $\text{im }\mathbf{B} = \{\mathbf{y}|\mathbf{y} = \mathbf{Bx} \text{ for some } \mathbf{x}\}$.

Lemma 5.1. *Let $\mathbf{C}$ and $\mathbf{E}$ be two square matrices of the same finite dimension. Let $\mathbf{B}$ be a matrix, not necessarily square but with the same number of rows, such that the matrix $[\mathbf{EB}]$ has full rank. Assume that the matrix $s\mathbf{E} - \mathbf{C}$ is invertible for every sufficiently large real s. Then the elements of*

$$\limsup_{s\to\infty}(s\mathbf{E} - \mathbf{C})^{-1}\mathbf{B}$$

are all finite if and only if the following condition holds: The intersection of $\ker\mathbf{E}$ with $\mathbf{C}^{-1}(\text{im }\mathbf{E}) = \{\mathbf{x}|\mathbf{Cx} \in \text{im }\mathbf{E}\}$ contains the zero vector only.

Returning to the proof of Theorem 4.2, recall that our objective is the limit $\theta \to 0$ of the stationary distribution of the chain $\{L_j^{\mathcal{B}_\theta}\}$. Let $\mathcal{U}$ be defined as the 'uniformization' of $\mathcal{B}$ in which the branching distribution is $\mathbf{v}\cdot\mathbf{F}_N$ for all nodes; all other parameters of $\mathcal{B}$ and $\mathcal{U}$ are the same. Note that $\mathcal{U}$ is proper by Theorem 2.1, since its branching distribution has mean $\mathbf{v}\cdot\mathbf{b} < 1$. Let $\tilde{\mathbf{p}} = (\tilde{p}_1, \tilde{p}_2, \ldots)$ denote the stationary distribution in vector format of the macro-state chain $\{L_j^{\mathcal{U}}\}$, and let $\tilde{\mathbf{p}}(\theta)$ be similarly defined for $\{L_j^{\mathcal{B}_\theta}\}$. Theorem 3.1 shows that the $\tilde{p}_\ell$ are generated by

$$\tilde{g}_L(z) = \frac{1-\mathbf{v}\cdot\mathbf{b}}{r}\frac{z^r-1}{1-\frac{1}{z}\mathbf{v}\cdot\mathbf{g}_N(z)}\ ,$$

so the theorem will be proved once we show that $\tilde{\mathbf{p}}(\theta) \to \tilde{\mathbf{p}}$ element-wise and in expectation.

Define the two indecomposable stochastic matrices, $\mathbf{A}$ and $\mathbf{A}(\theta)$ having $\tilde{\mathbf{p}}$ and $\tilde{\mathbf{p}}(\theta)$ as their respective stochastic invariant vectors. The matrix $\mathbf{A}$ is simply the transition matrix of the chain $\{L_j^{\mathcal{U}}\}$ with elements $a_{i\ell} = P\{L_j^{\mathcal{U}} = \ell | L_{j-1}^{\mathcal{U}} = i\}, i, \ell = 1, 2, \ldots$. We define $\sigma_m^w = F_N^w(m) - F_N^w(m-1)$ for every integer m and node number w, so by definition of the network $\mathcal{U}$,

$$a_{i\ell} = \sum_{w=1}^{k} (\sigma_{\ell-i+1}^w + 1_{\{i=1,\ell=r\}}\sigma_0^w)v_w \ .$$

The second matrix, $\mathbf{A}(\theta)$, is defined by

$$a_{i\ell}(\theta) = \lim_{j\to\infty} P\{L_j^{\mathcal{B}_\theta} = \ell | L_{j-1}^{\mathcal{B}_\theta} = i, M_j^\theta = 1\}, \quad i, \ell = 1, 2, \ldots \ .$$

The ℓ^{th} component of the invariant vector of this matrix is $\lim_{j\to\infty} P\{L_j^{\mathcal{B}_\theta} = \ell | M_j^\theta = 1\} = \lim_{j\to\infty} P\{L_j^{\mathcal{B}_\theta} = \ell\} = \tilde{p}_\ell(\theta)$ by the independence of $L_j^{\mathcal{B}_\theta}$ and M_j^θ at steady-state. We now prove a lemma giving conditions for the desired convergence; once we have shown that these conditions hold, we will have proved the theorem.

Lemma 5.2. *Assume that $\mathbf{A}(\theta) \to \mathbf{A}$ element-wise, and that the first moment of the stochastic invariant vector $\tilde{\mathbf{p}}(\theta)$ is a bounded function of θ. Then $\tilde{\mathbf{p}}(\theta) \to \tilde{\mathbf{p}}$ both element-wise and in expectation.*

Proof: Assume by contradiction that there exists a sequence of θ values for which $\tilde{\mathbf{p}}(\theta)$ does not converge to $\tilde{\mathbf{p}}$, despite the fact that $\mathbf{A}(\theta) \to \mathbf{A}$ and the first moment of $\tilde{\mathbf{p}}(\theta)$ is bounded. By Helly's extraction principle [7, p. 83], there is a subsequence of this sequence of θ values for which $\tilde{\mathbf{p}}(\theta)$ converges to some substochastic vector $\mathbf{x}$. It is easy to verify that $\mathbf{x}$ must satisfy $\mathbf{xA} = \mathbf{x}$. Moreover, since the first moment of $\tilde{\mathbf{p}}(\theta)$ is bounded, $\mathbf{x}$ must be strictly stochastic. The fact that $\mathbf{x} \neq \tilde{\mathbf{p}}$ then contradicts the uniqueness of the stochastic invariant vector. ∎

Next, we verify the conditions of the lemma, starting with $\mathbf{A}(\theta) \to \mathbf{A}$. Define $W_j^{\mathcal{B}_\theta}$ as the index of the node where service completion j occurred. Note that the $W_j^{\mathcal{B}_\theta}$ are well defined (a.s.), since the service-time distributions are absolutely continuous. Conditioning on this index, we have

$$\begin{aligned} a_{i\ell}(\theta) \quad = \quad & \sum_{w=1}^{k} \lim_{j\to\infty} P\{L_j^{\mathcal{B}_\theta} = \ell | L_{j-1}^{\mathcal{B}_\theta} = i, M_j^\theta = 1, W_j^{\mathcal{B}_\theta} = w\} \\ & \cdot \lim_{j\to\infty} P\{W_j^{\mathcal{B}_\theta} = w | L_{j-1}^{\mathcal{B}_\theta} = i, M_j^\theta = 1\} \ . \end{aligned}$$

The first factor of the summand is clearly equal to $\sigma_{\ell-i+1}^w + 1_{\{i=1,\ell=r\}}\sigma_0^w$, so it remains to check that the second factor converges to v_w. The index $W_j^{\mathcal{B}_\theta}$ is independent of M_j^θ, so this second

factor is just $\lim_{j\to\infty} P\{W_j^{\mathcal{B}_\theta} = w | L_{j-1}^{\mathcal{B}_\theta} = \ell\}$. But $W_j^{\mathcal{B}_\theta}$ is a deterministic function of the state $(\mathbf{Q}_{j-1}^{\mathcal{B}_\theta}, \mathbf{R}_{j-1}^{\mathcal{B}_\theta})$ (it is the node index of the smallest positive element of $\mathbf{R}_{j-1}^{\mathcal{B}_\theta}$), so by (5.1), the second factor indeed converges to the CQ network counterpart $\lim_{j\to\infty} P\{W_j^{\mathcal{C}_\ell} = w\} = v_w$.

It remains to show that the first moment of $\tilde{\mathbf{p}}(\theta)$ is bounded. The total population at any time consists of *primary* customers and *copies*; the former are created in the initial state or in samples from branching distributions (after successful trials $M_j^\theta = 1$), and the latter are the replacements created at unsuccessful trials (copies of a primary customer continue to be generated up to the next successful trial). At most one customer in the set consisting of a primary customer and its copies, if any, can be counted in any given state. It follows that the population size observed at a service completion is stochastically no larger than the total number of primaries generated from an initial state. By the analysis in Section 2 (cf. Theorem 2.1), we have that the j^{th} component of $(\mathbf{I} - \mathbf{H}_\theta)^{-1}$ is the expected size of the total population (primaries plus copies) created by an initial customer at node j, where $\mathbf{H}_\theta = (1-\theta)\mathbf{\Gamma} + \theta \text{diag}\,(\mathbf{b})\Gamma$. Since the rate of successful trials is θ, and since there are at most r customers per queue in the initial state, the scalar product $(r\theta\mathbf{1}) \cdot ((\mathbf{I} - \mathbf{H}_\theta)^{-1}\mathbf{1})$ bounds the first moment of $\tilde{\mathbf{p}}(\theta)$.

The parameter r is just a constant factor, so to show that the scalar product remains bounded as $\theta \to 0$, it is enough to verify that the elements of the vector

$$\limsup_{\theta\to 0} \left\{\theta^{-1}(\mathbf{I} - \boldsymbol{\Gamma}) - [\mathbf{I} - \text{diag}\,(\mathbf{b})]\boldsymbol{\Gamma}\right\}^{-1} \mathbf{1}$$

are all finite. For this, we apply Lemma 5.1; with the associations $s = \theta^{-1}$, $\mathbf{E} = \mathbf{I} - \boldsymbol{\Gamma}$, $\mathbf{C} = [\text{diag}\,(\mathbf{b}) - \mathbf{I}]\boldsymbol{\Gamma}$, and $\mathbf{B} = \mathbf{1}$, we need only check that the conditions of the lemma hold. First, we know that $s\mathbf{E} - \mathbf{C}$ is invertible for every $s > 1$, because the BQ network is uniformly proper. Second, we need to verify that $[\mathbf{EB}]$ has full rank. This will follow if we show that no linear combination of its rows can be zero, so there is no solution $\mathbf{x} \neq \mathbf{0}$ of the equation $\mathbf{x}[\mathbf{EB}] = 0$. Substituting for $\mathbf{E}$ and $\mathbf{B}$ and decomposing into blocks, this equation reads

$$\begin{aligned} \mathbf{x}(\mathbf{I} - \boldsymbol{\Gamma}) &= \mathbf{0} \\ \mathbf{x} \cdot \mathbf{1} &= 0\ . \end{aligned}$$

The solution space of the first block consists of the span of $\mathbf{v}$, the normalized left Perron-Frobenius vector of $\boldsymbol{\Gamma}$. But $\mathbf{v}$ is a stochastic vector, so it cannot satisfy the second block.

Finally, we must verify the main condition $\mathbf{C}^{-1}(\text{im}\ \mathbf{E}) \cap \ker \mathbf{E} = \{\mathbf{0}\}$. We see immediately that $\ker \mathbf{E}$ is the span of the vector $\mathbf{1}$. In order to show that $\mathbf{1}$ is not in $\mathbf{C}^{-1}(\text{im}\ \mathbf{E})$, we have to check that there is no vector $\mathbf{y}$ such that $\mathbf{C1} = \mathbf{Ey}$. Suppose by contradiction that there is

such a $\mathbf{y}$. Substituting for $\mathbf{C}$ and $\mathbf{E}$ gives $(\mathrm{diag}\,(\mathbf{b})-\mathbf{I})\boldsymbol{\Gamma}\mathbf{1} = (\mathbf{I}-\boldsymbol{\Gamma})\mathbf{y}$. Putting this in the form $\mathbf{b}-\mathbf{1} = \mathbf{y}-\boldsymbol{\Gamma}\mathbf{y}$, and multiplying from the left by $\mathbf{v}$, we get $\mathbf{v}\cdot\mathbf{b}-1=0$, which contradicts our assumption that $\mathbf{v}\cdot\mathbf{b}<1$. ∎

Proof of Theorem 4.3. The intensity of a stationary point process is the reciprocal of the mean distance between points. Thus, recalling that $T_j^{\mathcal{B}_\theta}$ denotes departure epoch j, we have

$$\lambda_{\mathcal{B}_\theta} = \lim_{j\to\infty} \frac{1}{E[T_j^{\mathcal{B}_\theta} - T_{j-1}^{\mathcal{B}_\theta}]} ,$$

and we are required to prove that

$$\lim_{j\to\infty} E[T_j^{\mathcal{B}_\theta} - T_{j-1}^{\mathcal{B}_\theta}] \to \sum_{\ell=1}^{\infty} \tilde{p}_\ell/\lambda_{\mathcal{C}_\ell} . \tag{5.5}$$

Conditioning on the population size, we have

$$1/\lambda_{\mathcal{B}_\theta} = \lim_{j\to\infty} \sum_{\ell=1}^{\infty} E[T_j^{\mathcal{B}_\theta} - T_{j-1}^{\mathcal{B}_\theta} | L_{j-1}^{\mathcal{B}_\theta} = \ell] P(L_{j-1}^{\mathcal{B}_\theta} = \ell) ,$$

or, in terms of scalar products,

$$1/\lambda_{\mathcal{B}_\theta} = \tilde{\mathbf{e}}(\theta)\cdot\tilde{\mathbf{p}}(\theta) ,$$

where $\tilde{\mathbf{e}}(\theta) = (\tilde{e}_1(\theta), \tilde{e}_2(\theta), \ldots)$ with $\tilde{e}_\ell(\theta) = \lim_{j\to\infty} E[T_j^{\mathcal{B}_\theta} - T_{j-1}^{\mathcal{B}_\theta} | L_{j-1}^{\mathcal{B}_\theta} = \ell]$, and where $\tilde{\mathbf{p}}(\theta)$ is the stationary distribution of the chain $\{L_j^{\mathcal{B}_\theta}\}$. Note that $\tilde{\mathbf{p}}(\theta)$ converges to $\tilde{\mathbf{p}} = (\tilde{p}_1, \tilde{p}_2, \ldots)$ as shown in Theorem 4.2. Also, the duration $T_j^{\mathcal{B}_\theta} - T_{j-1}^{\mathcal{B}_\theta}$ is contained in the state $(\mathbf{Q}_{j-1}^{\mathcal{B}_\theta}, \mathbf{R}_{j-1}^{\mathcal{B}_\theta})$ (it is the minimum of the entries of $\mathbf{R}_{j-1}^{\mathcal{B}_\theta}$), so by (5.1),

$$\lim_{j\to\infty} E[T_j^{\mathcal{B}_\theta} - T_{j-1}^{\mathcal{B}_\theta} | L_{j-1}^{\mathcal{B}_\theta} = \ell] \to \lim_{j\to\infty} E[T_j^{\mathcal{C}_\ell} - T_{j-1}^{\mathcal{C}_\ell}] = 1/\lambda_{\mathcal{C}_\ell} .$$

Finally, $\tilde{e}_\ell(\theta)$ is bounded uniformly in ℓ and θ, since it is bounded by $\max\limits_{1\le i\le k}(1/\mu_i)$, the maximum of the means of the service-time distributions $F_S^1, \ldots, F_S^k$. Then (5.5) follows from the fact that the product of a convergent stochastic vector and a convergent nonnegative bounded vector is convergent. ∎

Proof of Theorem 4.4. It is convenient to work with the joint process $(\mathbf{Q}, \mathbf{R})$. With the help of (5.1), we prove that, if $\mathbf{v}\cdot\mathbf{b}<1$, then with $\ell \equiv \mathbf{q}\cdot\mathbf{1}$ we have

$$P\{\mathbf{Q}^{\mathcal{B}_\theta}(t) = \mathbf{q}, \mathbf{R}^{\mathcal{B}_\theta}(t) \in \cdot\} \to \tilde{\lambda}_{\mathcal{B}}\tilde{p}_\ell \frac{1}{\lambda_{\mathcal{C}_\ell}} P\{\mathbf{Q}^{\mathcal{C}_\ell}(t) = \mathbf{q}, \mathbf{R}^{\mathcal{C}_\ell}(t) \in \cdot\} \tag{5.6}$$

uniformly in $\mathbf{q} \in \mathbb{Z}_\oplus^k$ and the measurable sets of residual service-time structures. Theorem 4.4 will follow at once.

Fix $t = 0$ without loss of generality. Our approach begins by expressing the left-hand side of (5.6) in terms of a Palm probability induced by the over-all process $D_{\mathcal{B}_\theta}$ of departures from the network $\mathcal{B}_\theta$. Let the points of $D_{\mathcal{B}_\theta}$ be marked with elements from the space Ω of sample paths of $\mathcal{B}_\theta$. The markings are expressed in terms of time shifts $\{\vartheta_\tau, -\infty < \tau < \infty\}$ operating on sample paths $\omega \in \Omega$; in the construction here, the mark given to a point T on the sample path ω is the left-shifted version $\vartheta_T\omega$. These are called the *universal marks* (cf. [4, p. 11]). The Palm probability $\tilde{P}(V)$ of an event V is defined to be the relative intensity of markings taken from V. The event $\vartheta_\tau \in U$ designates the set of sample paths whose shifted versions belong to U.

A basic tool of the Palm calculus is its inversion formula; the version to be used here reads (cf. [4, formula 4.1.2b])

$$P(U) = \lambda \int_{\tau=0}^{\infty} \tilde{P}(\tau < T_1, \vartheta_\tau \in U) d\tau \ , \tag{5.7}$$

where λ is the intensity and T_1 the first positive point of the underlying point process. Adapting this formula to our setting and then applying it to the left-hand side of (5.6) with $t = 0$ yields (recall that, by convention, $T_1^{\mathcal{B}_\theta}$ is the first positive point of $D_{\mathcal{B}_\theta}$)

$$P\{\mathbf{Q}^{\mathcal{B}_\theta}(0) = \mathbf{q}, \mathbf{R}^{\mathcal{B}_\theta}(0) \in \cdot\} = \lambda_{\mathcal{B}_\theta} \int_{\tau=0}^{\infty} \tilde{P}\{\tau < T_1^{\mathcal{B}_\theta}, \vartheta_\tau \in \{\mathbf{Q}^{\mathcal{B}_\theta}(0) = \mathbf{q}, \mathbf{R}^{\mathcal{B}_\theta}(0) \in \cdot\}\} d\tau \ . \tag{5.8}$$

Observe that if the first point $T_1^{\mathcal{B}_\theta}$ has not occurred till time τ, then the condition $\vartheta_\tau \in \{\mathbf{Q}^{\mathcal{B}_\theta}(0) = \mathbf{q}, \mathbf{R}^{\mathcal{B}_\theta}(0) \in \cdot\}$ holds for just those sample paths with

$$\{\mathbf{Q}^{\mathcal{B}_\theta}(0) = \mathbf{q}, \mathbf{R}^{\mathcal{B}_\theta}(0) - \tau \cdot \mathbf{1}_{\{\mathbf{R}^{\mathcal{B}_\theta}(0) > 0\}} \in \cdot\};$$

here $\mathbf{1}_{\{\mathbf{R}^{\mathcal{B}_\theta}(0) > 0\}}$ is defined as the structure $\mathbf{R}^{\mathcal{B}_\theta}(0)$ with 1's replacing the positive entries. Note that if $\tau < T_1^{\mathcal{B}_\theta}$, then the least residual service time exceeds τ, so subtracting τ from the positive entries of $\mathbf{R}^{\mathcal{B}_\theta}(0)$ leaves them nonnegative. Thus, conditioning on the population size, we can write

$$\begin{aligned} &P\{\mathbf{Q}^{\mathcal{B}_\theta}(0) = \mathbf{q}, \mathbf{R}^{\mathcal{B}_\theta}(0) \in \cdot\} \\ &= \lambda_{\mathcal{B}_\theta} \textstyle\int_{\tau=0}^{\infty} \tilde{P}\{\tau < T_1^{\mathcal{B}_\theta}, \mathbf{Q}^{\mathcal{B}_\theta}(0) = \mathbf{q}, \mathbf{R}^{\mathcal{B}_\theta}(0) - \tau \cdot \mathbf{1}_{\{\mathbf{R}^{\mathcal{B}_\theta}(0) > 0\}} \in \cdot | L^{\mathcal{B}_\theta}(0) = \ell\} d\tau \cdot \tilde{P}\{L^{\mathcal{B}_\theta}(0) = \ell\} \ . \end{aligned} \tag{5.9}$$

Now $\lambda_{\mathcal{B}_\theta}$ converges to $\tilde{\lambda}_{\mathcal{B}}$ by Theorem 4.3, and $\tilde{P}\{L^{\mathcal{B}_\theta}(0) = \ell\} = P\{L_j^{\mathcal{B}_\theta} = \ell\}$ by definition of Palm probability (more specifically, by the fact that in sample paths qualifying as markings, a point of the underlying point process occurs at time 0). Then $\tilde{P}\{L^{\mathcal{B}_\theta}(0) = \ell\}$ converges to $\tilde{p}_\ell$ by Theorem 4.2, so (5.6) will be proved if we can verify that the integral in (5.9) converges to

$$\frac{1}{\lambda_{\mathcal{C}_\ell}} P\{\mathbf{Q}^{\mathcal{C}_\ell}(0) = \mathbf{q}, \mathbf{R}^{\mathcal{C}_\ell}(0) \in \cdot\} \ . \tag{5.10}$$

To prove this, we again resort to the construction of the Palm probability and observe that the integral in (5.9) can be expressed as follows in terms of the embedded process:

$$(5.11)\qquad \int_{\tau=0}^{\infty} \lim_{j\to\infty} P\{\tau < T_{j+1}^{\mathcal{B}_\theta} - T_j^{\mathcal{B}_\theta}, \mathbf{Q}_j^{\mathcal{B}_\theta} = \mathbf{q}, \mathbf{R}_j^{\mathcal{B}_\theta} - \tau \cdot \mathbf{1}_{\{\mathbf{R}_j^{\mathcal{B}_\theta}>0\}} \in \cdot | L_j^{\mathcal{B}_\theta} = \ell\} d\tau \ .$$

The integrand is at most $\lim_{j\to\infty} P\{T_{j+1}^{\mathcal{B}_\theta} - T_j^{\mathcal{B}_\theta} > \tau\}$, where $T_{j+1}^{\mathcal{B}_\theta} - T_j^{\mathcal{B}_\theta}$ is a smallest positive element of $\mathbf{R}_j^{\mathcal{B}_\theta}$, and is dominated stochastically by a largest service time, i.e., a service time sampled from a distribution F_S^i with smallest rate parameter. Thus, the integrand in (5.11) is bounded by an integrable function of τ. Then by (5.1) and the dominated convergence theorem, we conclude that (5.11) converges to its CQ-network counterpart,

$$(5.12)\qquad \int_{\tau=0}^{\infty} \lim_{j\to\infty} P\{\tau < T_{j+1}^{\mathcal{C}_\ell} - T_j^{\mathcal{C}_\ell}, \mathbf{Q}_j^{\mathcal{C}_\ell} = \mathbf{q}, \mathbf{R}_j^{\mathcal{C}_\ell} - \tau \cdot 1_{\{\mathbf{R}_j^{\mathcal{C}_\ell}>0\}} \in \cdot\} d\tau \ .$$

The convergence is uniform over the sample space of $(\mathbf{Q}^{\mathcal{C}_\ell}, \mathbf{R}^{\mathcal{C}_\ell})$, since (5.1) holds in the total variation sense. To complete the proof, we consider (5.10). In analogy with the BQ-network case, the inversion formula gives

$$(5.13)\qquad \frac{1}{\lambda_{\mathcal{C}_\ell}} P\{\mathbf{Q}^{\mathcal{C}_\ell}(0) = \mathbf{q}, \mathbf{R}^{\mathcal{C}_\ell}(0) \in \cdot\}$$
$$= \int_{\tau=0}^{\infty} \tilde{P}\{\tau < T_1^{\mathcal{C}_\ell}, \vartheta_\tau \in \{\mathbf{Q}^{\mathcal{C}_\ell}(0) = \mathbf{q}, \mathbf{R}^{\mathcal{C}_\ell}(0) \in \cdot\}\} d\tau \ ,$$

and our earlier observations show that the Palm probability on the right-hand side of (5.13) can be rewritten as in (5.12). ∎

Proof of Corollary 4.6. In networks with exponential service, a natural way of calculating throughputs is by aggregating service rates. Theorem 4.4 and Corollary 4.5 show that, given the population size $L^{\mathcal{B}_\theta}(t) = \ell$, the conditional queue-length distribution converges to that of the CQ network $\mathcal{C}_\ell$. The service rate at node i in macro-state ℓ thus converges to the CQ network counterpart, $v_i \lambda_{\mathcal{C}_\ell}$. Performing the aggregation, we have that the limiting throughput at node i is $\sum_{\ell=1}^{\infty} (v_i \lambda_{\mathcal{C}_\ell}) p_\ell$. By the formulas for $\tilde{\lambda}_{\mathcal{B}}$ (Theorem 4.3) and p_ℓ (Corollary 4.5), this sum is equal to $v_i \tilde{\lambda}_{\mathcal{B}}$ as stated in the corollary.

An alternative proof of this result stays with the theory of stationary marked point processes applied in the proof of Theorem 4.4. Let $D_{\mathcal{B}_\theta}^i$ be the departure process at node i, and let $\lambda_{\mathcal{B}_\theta}^i$ be its intensity. Consider the points of the over-all, superposition process $D_{\mathcal{B}_\theta}$ marked with the indices $\{W_j^{\mathcal{B}_\theta}\}$ of the nodes where departures take place. A restriction of $D_{\mathcal{B}_\theta}$ to the points with marking i gives $D_{\mathcal{B}_\theta}^i$ so $\lambda_{\mathcal{B}_\theta}^i$ is the marginal intensity $\lim_{j\to\infty} P\{W_j^{\mathcal{B}_\theta} = i\} \lambda_{\mathcal{B}_\theta}$. The first factor converges to v_i as noted in the proof of Theorem 4.2, and the second converges to $\tilde{\lambda}_{\mathcal{B}}$ by Theorem 4.3. Thus, $\lambda_{\mathcal{B}}^i = v_i \tilde{\lambda}_{\mathcal{B}}$ as before. ∎

Acknowledgement. We are pleased to acknowledge the guidance of professor Johannes M. Schumacher (CWI, Amsterdam) in directing us to Lemma 5.1.

REFERENCES

[1] J. Abate and W. Whitt. The fourier-series method for inverting transforms of probability distributions. *Queueing Systems*, 10:5–88, 1992.

[2] Soren Asmussen. *Applied Probability and Queues.* John Wiley & Sons, 1987.

[3] O. I. Aven, E. G. Coffman, Jr., and Y. A. Kogan. *Stochastic Analysis of Computer Storage.* Reidel, 1987.

[4] François Baccelli and Pierre Brémaud. *Elements of Queueing Theory.* Springer-Verlag, 1994.

[5] N. Bayer and Y. Kogan. Branching models of MIMD architectures. Operations Research Statistics and Economics Mimeograph Series 404, Technion—Israel Institute of Technology, Faculty of Industrial and Management Engineering, September 1992.

[6] N. Bayer and Y. Kogan. A branching process with conditionally-binomial offsprings. In preparation.

[7] Kai Lai Chung. *A Course in Probability Theory.* Academic Press, 1974.

[8] E. G. Coffman Jr., N. Bayer, and Kogan Y. A. A new concept for modeling MIMD architectures. In preparation.

[9] P. J. Courtois. *Decomposability: Queueing and Computer System Applications.* Academic Press, 1977.

[10] Theodore. E. Harris. *The Theory of Branching Processes.* Springer-Verlag / Prentice-Hall, 1963.

[11] Hisashi Kobayashi. *Modeling and Analysis: An Introduction to System Performance Evaluation Methodology.* Addison-Wesley, 1978.

[12] S. Kotz, N. L. Johnson, and C. B. Read (Eds). *Encyclopedia of Statistical Sciences, Vol. 5.* Wiley & Sons, New York, 1985.

[13] Margreet Kuijper. *First-Order Representations of Linear Systems.* Birkhäuser, Stuttgart, 1994.