ANALYTIC ERROR BOUNDS
FOR APPROXIMATIONS OF QUEUEING NETWORKS
WITH AN APPLICATION TO ALTERNATE ROUTING

NICO M. VAN DIJK

Abstract

A general condition is provided from which an error bound can be concluded for approximations of queueing networks which are based on modifications of the transition and state space structure. This condition relies upon Markov reward theory and can be verified inductively in concrete situations. The results are illustrated by estimating the accuracy of a simple throughput bound for a closed queueing network with alternate routing and a large finite source input. An explicit error bound for this example is derived which is of order M^{-1}, where M is the number of sources.

1. Introduction

Ever since Erlang's and Engset's classical results in the early twenties, queueing theory has been extensively involved in teletraffic and communication theory. Particularly, motivated by Jackson's celebrated product form results in the late fifties (cf. [11]), the theory of queueing networks has gained a wide popularity in telecommunication and computer performance evaluation. Part of this success can be attributed to the various product form extensions and their robustness with respect to the underlying distributional assumptions (insensitivity properties), (e.g. [2-6], [8-10], [13], [20], [24], [37], [41]). Most unfortunately, typical practical features such as blocking phenomena, dynamic routing, overflow, breakdowns and job-priorities usually destroy the appealing product form (e.g. [9], [25], [29], [42]).

---

[1] Free University, Amsterdam, The Netherlands.

Another Part of their success, however, can be explained by the fact that simplifying assumptions (such as infinite and independent stations) which guarantee product forms, tend to give "reasonable" approximate results in various practical situations, especially when the system is large, (cf. [25], [42]). These simplications can often be seen as minor though critical modifications of the underlying transition structure such as by adding or deleting particular transitions. Despite numerical support, however, analytic a priori error bounds for the accuracy of such "product form" approximations do not seem to be available.

Also other types of approximate modeling issues are typically concerned with networks of queues. One of these is the issue of a closed (finite source input) or open (Poissonian input) description (cf. [40]).with advantages (e.g. computational, finiteness) and disadvantages (e.g. complexity, station dependence) for either of them. Convergence results for closed approximations of open systems have been established (cf. [26], [40]). But (error) bounds of this form are limited to simple Erlang type systems (cf. [40]) or robust bounds for state space truncations which do not secure an order of accuracy (cf. [26]).

Another approximate or modelling issue is the exactness of system input parameters such as the mean arrival and service rates, as in practice these are usually subject to randomness (e.g. resulting from confidence intervals for statistical estimates or external fluctuations). To this end, perturbation results with error bounds have recently been developed in [35], with one dimensional queueing applications.

All of the above "approximations" come down to some kind of modification or perturbation of the transition structure and/or a truncation or an extension of the state space. This paper, therefore, aims to provide a general tool for concluding error bounds for such approximations. It thereby extends the perturbation error bound results from [35] in that it

(i) allows modifications of the state space such as a truncation for closed or an infinite extension for open modeling and

(ii) particularizes to networks of queues rather than one-dimensional queueing applications.

A pair of simple conditions is provided from which error bounds can be concluded. The actual verification of these conditions, however, is the crucial part for practical application. To this end, an inductive verification technique as also used in [32], [34], and [36] will be illustrated and applied to a particular non-product form system of practical interest (cf. [1], [21]). This concerns a queueing network with alternate routing upon saturation of a primary access station and a large finite source input. A simple throughput

is proposed and an explicit error bound is derived of order M^{-1}, with M the number of sources.

Though the example is to be seen as mainly generic as it is relatively simple from a practical point of view, it includes the essential phenomenon of a finite capacity constraint (or blocking) and a state dependent (thus dynamic) routing. For more complex situations, such as with more capacity constraints and blocking phenomena, similar though more complicated proofs can be expected along essentially the same lines.

First, the general models are presented in Section 2. Next, the corresponding error bound result is developed in Section 3. Finally, an illustrative network with alternate routing is analyzed in Section 4. An evaluation concludes the paper.

2. Comparative models

Consider an arbitrary open or closed single class exponential queueing network with N service stations (hereafter called the original model), such as illustrated in Figure 1.

The state of the network is described by $\bar{n} = (n_1, \ldots n_N)$ where n_i is the number of jobs at station i, $i = 1, \ldots N$. By $n + e_i$ (or $\bar{n} - e_i$), we denote the state equal to $\bar{n}$ except for one job more (respectively less) at station i where we also allow $i = 0$ with the convention that $\bar{n} \pm e_0 = \bar{n}$. Consequently, by $\bar{n} - e_i + e_j$ we denote the state equal to $\bar{n}$ with one job moved from station i into station j, where $i = 0$ corresponds to an external arrival at station j and $j = 0$ to a departure from the system at station i. Let $q(\bar{n}, \bar{n} - e_i + e_j)$ for $i, j = 0, 1, \ldots, N$ be the transition rate for a change from state $\bar{n}$ into state
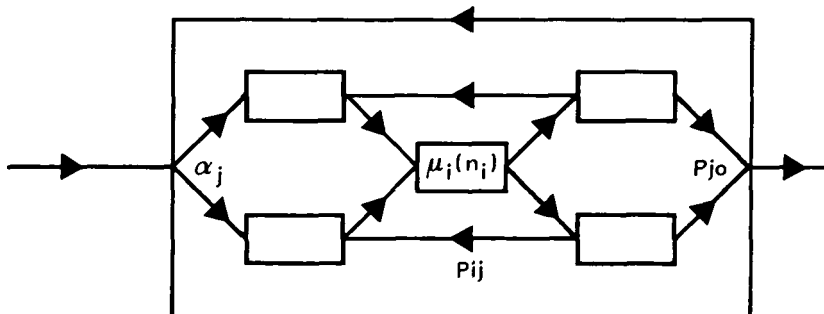

FIGURE 1

$\overline{n} - e_i + e_j$, while transition rates for changes not of this form are assumed to be 0. For example, for a standard Jackson network we have

$$q(\overline{n}, \overline{n} - e_i + e_j) = \mu_i p_{ij},$$

with μ_i the service rate at station i and p_{ij} the routing probability from station i to j, while an additional capacity constraint N_j yielding a reflective blocking (communication protocol) is parametrized by

$$q(\overline{n}, \overline{n} - e_i + e_j) = \mu_1 p_{ij} 1(n_j < N_j),$$

where $1(A)$ or $1_{\{A\}}$ denotes an indicator of event A, i.e., $1(A) = 1_{\{A\}} = 1$ if event A is satisfied and $1(A) = 1_{\{A\}} = 0$ otherwise. Without restriction of generality, the following assumptions are made:

(1) The underling Markov jump process is irreducible at some set S of admissible states $\overline{n}$, with a unique stationary distribution $\pi(.)$.

(2) The transition rates are uniformly bounded. That is, we can choose a finite Q such that

$$Q \geq \sup_{\overline{n} \in S} \sum_{i,j} q(\overline{n}, \overline{n} - e_i + e_j). \tag{2.1}$$

(3) For some given reward rate $r(\overline{n})$ the value g is finite and well-defined by

$$g = \sum_{\overline{n}} \pi(\overline{n}) r(\overline{n}). \tag{2.2}$$

The value g then represents some performance measure of interest, such as the throughput of a particular station j by $r(\overline{n}) = \mu_j(n_j)$ or the steady state probability of a particular subset B by $r(\overline{n}) = 1(\overline{n} \in B)$.

Comparative model. Now consider a modified version of the single class exponential queueing network (hereafter called the modified model) with a description as above, but with $q(\overline{n}, \overline{n} - e_i + e_j)$ replaced by $\overline{q}(\overline{n}, \overline{n} - e_i + e_j)$, the assumptions 1, 2, and 3 adopted with S, π, r and g replaced by $\overline{S}, \overline{\pi}, \overline{r}$ and $\overline{g}$, but Q kept the same, and most essentially

$$\overline{S} \subset S. \tag{2.3}$$

3. Comparison result

We now wish to evaluate the difference $|\overline{g} - g|$, that is the difference of the performance measure for the original and modified queueing network, without having to compute the stationary probabilities $\pi(.)$ and $\overline{\pi}(.)$.

To this end, as justified by the boundedness assumption (2.2), we first apply the standard uniformization technique (e.g. [29], p. 110) in order to

transform the continuous-time description into a discrete-time formulation. More precisely, let Q be any arbitrary finite number satisfying (2.2) and define one-step transition probabilities $p(\bar{n}, \bar{n} - e_i + e_j)$ and $\bar{p}(\bar{n}, \bar{n} - e_i + e_j w)$ by

$$p(\bar{n}, \bar{n} - e_i + e_j) = q(\bar{n}, \bar{n} - e_i + e_j)/Q,$$
$$\bar{p}(\bar{n}, \bar{n} - e_i + e_j) = \bar{q}(\bar{n}, \bar{n} - e_i + e_j)/Q,$$

$$p(\bar{n}, \bar{n}) = 1 - \left[\sum_{i,j=0}^{N} q(\bar{n}, \bar{n} - e_i + e_j)/Q \right], \tag{3.1}$$

$$\bar{p}(\bar{n}, \bar{n}) = 1 - \left[\sum_{i,j=0}^{N} \bar{q}(\bar{n}, \bar{n} - e_i + e_j)/Q \right],$$

while transition probabilities $p(.,.)$ and $\bar{p}(.,.)$ for any other type transition are assumed to be 0. From now on, we always use an upper bar "-" symbol to indicate an expression for the modified system and the symbol "(-)" to indicate that the expression is to be read for both the original and modified system. Further, let operators $\overset{(-)}{T}$ and $\{\overset{(-)}{T}_t | t = 0, 1, 2, \ldots\}$ upon arbitrary functions $f \colon \overset{(-)}{S} \to R$ be defined by

$$\overset{(-)}{T} f(\bar{n}) = \sum_{i,j=0}^{N} \overset{(-)}{p}(\bar{n}, \bar{n} - e_i + e_j) f(\bar{n} - e_i + e_j),$$

$$\overset{(-)}{T}_{t+1} f = \overset{(-)}{T} \overset{(-)}{T}_t f (t \geq 0), \tag{3.2}$$

$$\overset{(-)}{T}_0 = I.$$

And define the reward functions $\{\overset{(-)}{V} | t = 0, 1, \ldots\}$ at $\overset{(-)}{S}$ by

$$\overset{(-)}{V}_N = Q^{-1} \sum_{t=0}^{N-1} \overset{(-)}{T}_t r. \tag{3.3}$$

Then by virtue of the uniformization technique (cf. [29], p. 110) and the irreducibility assumptions of $\overset{(-)}{S}$, by standard Markov reward theory (cf. [16]) we conclude

$$\overset{(-)}{g} = \lim_{N \to \infty} \frac{Q}{N} \overset{(-)}{V}_N(\bar{l}) \tag{3.4}$$

for arbitrary $\bar{l} \in \overset{(-)}{S}$. This leads to the following key-theorem which guarantees an error bound for the difference $|\bar{g} - g|$. Its conditions will be discussed later on. Herein, we use the abbreviation

$$\Delta(\bar{n}, \bar{n} - e_i + e_j) = [\bar{q}(\bar{n}, \bar{n} - e_i + e_j) - q(\bar{n}, \bar{n} - e_i + e_j)].$$

THEOREM 3.1. (General Conditions) *Suppose that for some constants β, δ, $\varepsilon > 0$, some state $\bar{l} \in \bar{S}$, some nonnegative functions $\Phi(.)$ and all $t \geq 0$, $\bar{n} \in \bar{S}$:*

$$\overline{T}_t \Phi(\bar{l}) \leq \beta, \tag{3.5}$$

$$|\bar{r}(\bar{n}) - r(\bar{n})| \leq \delta \Phi(\bar{n}), \tag{3.6}$$

$$\left| \sum_{i,j=0}^{N} \Delta(\bar{n}, \bar{n} - e_i + e_j)[V_t(\bar{n}, \bar{n} - e_i + e_j) - V_t(\bar{n})] \right| \leq \varepsilon \Phi(\bar{n}). \tag{3.7}$$

Then

$$|(\overline{V}_N - V_N)(\bar{l})| \leq \beta[\delta + \varepsilon]N/Q \tag{3.8}$$

and

$$|\bar{g} - g| \leq \beta[\delta + \varepsilon]. \tag{3.9}$$

PROOF. Clearly, (3.9) immediately follows from (3.4) and (3.8). To prove (3.8) first conclude from (3.2) and (3.3) that for any $t \geq 0$:

$$\overset{(-)}{V}_{t+1} = \overset{(-)}{r}/Q + \overset{(-)}{T}_t \overset{(-)}{V}_t. \tag{3.10}$$

As the transition probabilities $\bar{p}(.,.)$ remain restricted to $\overset{(-)}{S}$ while also $\overset{(-)}{S} \subset S$, we can thus write for arbitrary $\bar{n} \in \bar{S}$:

$$
\begin{aligned}
(\overline{V}_N - V_N)(\bar{n}) &= (\bar{r} - r)(\bar{n})/Q + (\overline{TV}_{N-1} - TV_{N-1})(\bar{n}) \\
&= (\bar{r} - r)(\bar{n})/Q + (\overline{T} - T)V_{N-1}(\bar{n}) + \overline{T}(\overline{V}_{N-1} - V_{N-1})(\bar{n}) \\
&= \sum_{t=0}^{N-1} \overline{T}_t([\bar{r} - r]/Q + [(\overline{T} - T)V_{n-t-1}])(\bar{n}) + \overline{T}_N(\overline{V}_0 - V_0)(\bar{n}),
\end{aligned}
\tag{3.11}
$$

where the latter equality follows by iteration. Now note that the last term in the last right hand side is equal to 0 as $\overline{V}_0(.) = V_0(.) = 0$ by definition. Further, from (3.1) and (3.2) we find for any s and $\bar{n} \in \bar{S}$:

$$
\begin{aligned}
(\overline{T} - T)V_s(\bar{n}) &= \sum_{i,j=0}^{N} \bar{q}(\bar{n}, \bar{n} - e_i + e_j)[V_s(\bar{n}, \bar{n} - e_i + e_j) - V_s(\bar{n})]/Q \\
&\quad - \sum_{i,j=0}^{n} q(\bar{n}, \bar{n} - e_i + e_j)[V_s(\bar{n}, \bar{n} - e_i + e_j) - V_s(\bar{n})]/Q \\
&= \sum_{i,j=0}^{N} \Delta(\bar{n}, \bar{n} - e_i + e_j)[V_s(\bar{n} - e_i + e_j) - V_s(\bar{n})]/Q. \tag{3.12}
\end{aligned}
$$

Further, note that $\overline{T}_t f_1 \leq \overline{T}_t f_2$ for any $f_1 \leq f_2$ in component-wise sense as $\overline{T}_t$ is an expectation operation. As a result, by substituting (3.12) in (3.11),

substituting $\bar{n} = \bar{l}$, taking absolute values and applying (3.5) (3.7), we obtain:

$$|(\overline{V}_N - V_N)(\bar{l})| \leq [\delta + \varepsilon]Q^{-1} \sum_{t=0}^{N-1} \overline{T}_t \Phi(\bar{l}) \leq \beta[\delta + \varepsilon]N/Q. \qquad (3.13)$$

REMARK 3.2. (Discussion of the theorem). In the above theorem one must typically think of β and/or $[\delta + \varepsilon]$ to be small. The following physical insights and steps of verification are underlying to the conditions.

1) **Physical interpretations and related literature.** The essential conditions (3.5) and (3.7) are based upon the following two physical insights, which are related to numerical analysis:

  (i) Consistency: Comparison of local behaviour (condition 3.7).
  (ii) Stability: Boundedness of marginal expectations (condition 3.5).

  (i) (Consistency) The left hand side of condition (3.7) can be seen as $\|(\overline{A} - A)V_t\|$ where $\overline{A}$ and A denote the infinitesimal generators of the underlying stochastic processes. Intuitively, the difference in (infinitesimal) local behaviour is equivalent to the difference per unit of time and thus grows linearly in time, say linearly in N after N time units as in (3.8). This is a well-established fact in numerical analysis (cf. [18], [19]) and semigroup theory (cf. [31], [15], [16], [17]) provided one deals with uniformly bounded functions V_t. For cumulative values V_t, however, a quadratic growth of local differences comes out naturally (cf. [7], [38]) as the functions V_t themselves grow linearly in t. They key-step to overcome this problem by noting that differences $V_t(j) - V_t(i)$ are uniformly bounded in time is applied earlier in [35] as related to results in Markov decision theory. In this reference, however, only small differences in local behaviour were considered, as due to small perturbations. This basically comes down to Theorem 3.1 without function $\Phi(.)$.

  (ii) (Stability) By including a scaling function $\Phi(.)$, however, also significant differences of local behaviour such as due to a totally modified transition or truncation can be taken into account provided:

  • Either the difference remains small up to a scaling function Φ which has a uniformly bounded expected value over time (as guaranteed by (3.5) with $\beta < \infty$) (stability).
  • Or the occurrence of significant local differences has only a small probability (as guaranteed by (3.5) with small β and $\Phi(.)$ an indicator function).

2) **Verification of conditions.** In view of the above insights the conditions can be verified along the following lines:

Step 1(Bounded bias-terms) As first and most essential step one has to estimate (bound) the so-called bias terms $V_t(\bar{n} - e_i + e_j) - V_t(\bar{n})$ as:

$$|V_t(\bar{n} - e_i + e_j) - V_t(\bar{n})| \leq B_{i,j}, \qquad (3.14)$$

uniformly in t. From Markov reward theory it is standardly known that such terms are bounded uniformly in t for any given i and j as based upon mean first passage time results (cf. [22], [35]) and assuming $r(.)$ to be bounded. For finite networks a bound B uniformly in t and i, j can be concluded. The actual computation of such bounds by means of mean first passage times, however, becomes practically impossible for multi-dimensional applications such as considered in this paper (see [12] or [35] for simple one-dimensional cases). In the next section, therefore, we will illustrate how estimates for these bias-terms can be derived analytically.

Step 2 (Transition differences) Secondly, one has to find out whether the differences in the transition structure $\Delta(.,.)$ are small or just bounded up to a state dependent scaling function $\Phi(.)$. For illustration, think of $\Phi(.) = 1$ and consider the following examples.

Example 1 Consider a standard single-server queue with arrival rate λ and service rate μ as original model and the same model with arrival rate $\lambda+\tau$ (perturbation), resulting from a statistical confidence interval, as modified model, where τ is small. Then $|\Delta(.,.)| \leq \tau$.

Example 2 Consider the same original model as in the example above but now with rejection of arrivals (state space truncation) if upon arrival the number n of jobs present is equal to some limit L. Then $|\Delta(.,.)| \leq \lambda 1(n = L)$.

Step 3 (Bounding functions Φ) By Comparing the transition structures, candidates for an appropriate bounding function $\Phi(.)$ come up naturally. Here one may typically think of polynomial type functions, for example, $\Phi(\bar{n}) = n$ with n the total number of jobs present. One may thus have various options. As illustration, in example 2 above, condition (3.7) will be satisfied with some constant B resulting from (3.14) and

$$\begin{cases} \varepsilon = \lambda B/L & \text{if } \Phi(.) = n \\ \varepsilon = \lambda B & \text{if } \Phi(.) = 1(n = L). \end{cases}$$

Step 4 (Stability) Which option of $\Phi(.)$ is appropriate will eventually depend on whether we can easily verify (3.5), requiring that its expected value

over time remains bounded (stability) by either a small or just a finite number. As illustration, again for example 2 from above, we have

$$
\begin{cases}
\beta = \beta_1 = 1 & \text{if } \Phi(.) = 1 \\
\beta = \beta_2 = (\lambda/\mu)^L / \sum_{k=1}^{L} (\lambda/\mu)^k & \text{if } \Phi(.) = 1(n = L) \\
\beta = \beta_3 = \sum_{k=0}^{L} k(\lambda/\mu)^k / \sum_{k=0}^{L} (\lambda/\mu)^k & \text{if } \Phi(.) = n.
\end{cases}
$$

3) **Summary.** Roughly speaking. Theorem 3.1 can thus be applicable in a twofold manner given that the bias-terms can be sufficiently estimated:

(i) By showing that the impact of the difference Δ in the transition structures upon the state-dependent estimates for the bias terms is sufficiently small, such as for example 1 with $\varepsilon = \tau B$ and $\beta = 1$ by using $\Phi(.) = 1$, or example 2 with $\varepsilon = \tau B/L$ and $\beta = \beta_3$ by using $\Phi(.) = n$.

(ii) By showing that the expected value of the scaling function or the probability of being in states where this difference is significant, is sufficiently small, such as for example 2 with $\varepsilon = \tau B$ and $\beta = \beta_2$ by using $\Phi(.) = 1(n = L)$.

REMARK 3.3. (Unbounded rewards) Note that no assumption has been made as to any boundedness of the reward rate. For example, we can have $r(\overline{n}) = n_j$ so as to calculate the mean queue length at a particular infinite station j.

REMARK 3.4. (Unbounded intensities) The boundedness assumption (2.1) is made in order to apply the uniformatization technique (3.1) yielding a recursive formulation. This, however, can be avoided in a technical manner similarly to [33] so as to allow unbounded intensities, such as from infinite server stations.

4. Application: A simple throughput and explicit error bound for a closed queueing network with alternate routing

4.1 Model

This section investigates an application of the preceding results which contains both a perturbation of the transition structure and a state space truncation. Moreover, the performance measure of interest, the system throughput, involves an unbounded reward structure. The application concerns a generic example of a practical phenomenon in teletraffic analysis: blocking with alternate routing.

Consider a queueing network with N service stations, as illustrated below (see Figure 2), of which the first station is a primary entrance station which allows no more than some finite number L of jobs and where upon saturation

of this station jobs have to take an alternate route according to routing prob-
abilities p_{ij}, $i, j = 0, 2, \ldots, N$, starting at station j with probability $p_{oj} = \alpha_j$.
Upon service completion at station i a job leaves the system with probability
$p_{i0} = 1 - \sum_{j=2}^{N} p_{ij}$ for $i \neq 1$ and $p_{io} = 1$, where the transition matrix (p_{ij}) for
$i, j = 0, 2, \ldots, N$ is assumed to be irreducible.

   The service rate at station i is $\mu_i(n_i)$ when n_i jobs are present, where $\mu_i(n_i)$
is assumed to be nondecreasing in n_i and $\mu_i(0) = 0$. Jobs arrive at the system
according to a finite source exponential input with M sources and exponential
idle times with parameter γ. That is, if n jobs are present in the system the
arrival rate is $(M - n)\gamma$.

   The system under consideration is not of product-form due to the dynamic
routing feature upon saturation. This feature naturally arises in teletraffic
applications for which various alternate routing schemes are of actual interest
(cf. [1], [21]). Here a large finite source input is most realistic, so that often a
Poissonian input approximation is used to simplify analysis or avoid complex
computations of a performance a measure such as the throughput. Below we
will investigate the accuracy of such an approximation, or more precisely, of
the throughout bound (it can be shown to be indeed an upper bound):
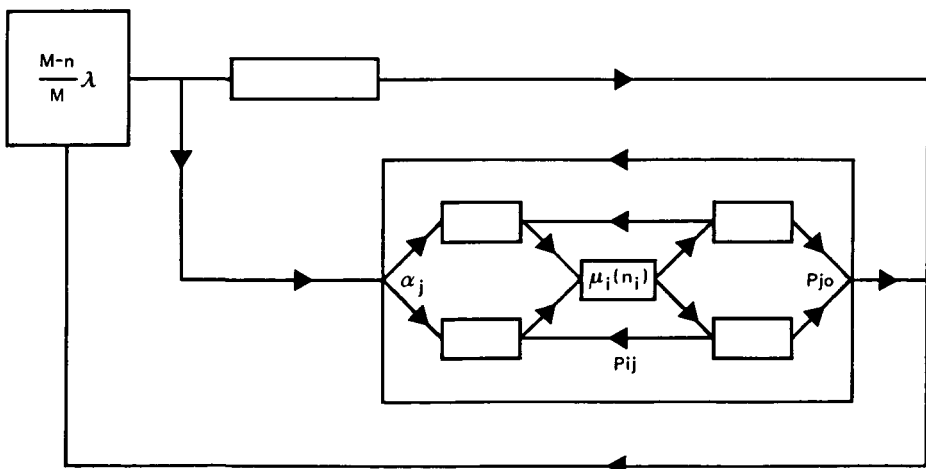
$$\lambda = \gamma M. \tag{4.1}$$

FIGURE 2

4.2. Parametrization

As we require $\overline{S} \subset S$, we consider the open Poisson input case as the original model and the closed finite source case as the modified model. Then with

$$Q \geq \lambda + \sum_i \mu_i(M), \tag{4.2}$$

and choosing

$$\mu_i(n_i) = \mu_i(M)$$

for $n_i \geq M$, as well as

$$S = \{\overline{n}|n_1 \leq L\} \qquad \overline{S} = \{\overline{n}|n_1 \leq L, n_1 + \cdots + n_N = M\}, \tag{4.3}$$

the assumption 1, 2 and 3 of section 2 are guaranteed for both the closed and open version with respective transition rates $\overline{q}(.,.)$ and $q(.,.)$ given by

$$\overset{(-)}{q}(\overline{n}, \overline{n} - e_i) = \mu_i(n_i)p_{i0}, \qquad (i = 1, \ldots, N)$$

$$\overset{(-)}{q}(\overline{n}, \overline{n} - e_i + e_j) = \mu_i(n_i)p_{ij}, \qquad (i, j = 2, \ldots, N)$$

but

$$\begin{cases} q(\overline{n}, \overline{n} + e_1) = \lambda 1(n_1 < L) \\ q(\overline{n}, \overline{n} + e_j) = \lambda 1(n_1 = L)\alpha_j \qquad (j = 2, \ldots, N) \end{cases}$$

and

$$\begin{cases} \overline{q}(\overline{n}, \overline{n} + e_1) = (M - n)\gamma 1(n_1 < L) \\ \overline{q}(\overline{n}, \overline{n} + e_j) = (M - n)\gamma 1(n_i = L)\alpha_j \qquad (j = 1, \ldots N). \end{cases} \tag{4.4}$$

The uniformization (3.1) is thus justified and with

$$\overset{(-)}{r}(\overline{n}) = \sum_{i=1}^N \mu_i(n_i)p_{i0}, \qquad (\overline{n} \in S) \tag{4.5}$$

the values $\overline{g}$ and g, as per (2.2) or equivalently (3.4), represent the throughput of the closed and open system. Since, however, $g = \lambda = \gamma M$ with $\overline{g}$ cannot be computed easily, it is of interest to investigate Theorem 3.1 so as to estimate the difference $|\overline{g} - \lambda|$.

4.3. Comparison result

We adopt all notation from section 3. As per the discussion in remark 3.2, the following lemma is the most crucial step. Herein, for arbitrary functions $f: S \to R$ and $j = 1, 2, \ldots, N$ we use the notation:

$$\Delta_j f(\overline{n}) = f(\overline{n} + e_j) - f(\overline{n}). \tag{4.6}$$

LEMMA 4.1. *For all $t \geq 0$ and j and $\overline{n}$ such that $\overline{n} + e_j \in S$:*

$$0 \leq \Delta_j V_t(\overline{n}) \leq 1. \tag{4.7}$$

PROOF. This will be given by induction to t. For $t = 0$, (4.7) trivially holds as $V_0(.) = 0$. Suppose that (4.7) holds for $t \leq m$ and for convenience write $h = Q^{-1}$.

The equation that will be presented below will at first glance be difficult to understand. It can be checked by explicitly writing out all terms, by adding and subtracting extra terms at appropriate places and collecting corresponding terms. As these steps will be cumbersome without providing insight they are omitted. The relation can also be understood rather directly along with following line of thinking.

Heuristic explanation of (4.8). "When comparing the transition mechanism in two adjacent states $\overline{n} + e_i$ and $\overline{n}$ most of the possible transitions and corresponding transition probabilities are the same in both states, such as a job-movement from a station $j \neq i$ to station k with probability $h\mu_j(n_j)p_{jk}$. The effect of these transitions in the two states can thus be compared pairwise for each transition separately. Certain transitions, however, will have a larger probability in one state than in the other, which includes transitions that are possible in one but not the other state. Say, a certain transition has probability γ more in state 1 than in state 2. Then, with probability γ one can compare the effect of that transition in state 1 with that of a "dummy transition" in state 2. The remaining probability for "no transition" then becomes equal for both states and the effect of transitions can be investigated per transition which is possible in either of the two states".

Keeping this arguing in mind and noting that this lemma concerns the open case with arrival parameter λ and $q(.,.)$ defined by (4.4), from (3.10), (4.4), (4.5) and (4.6) we conclude for $t = m + 1$ and $i = 1, \ldots, N$:

$$\begin{aligned}
\Delta_i V_{m+1}(\overline{n}) &= h[r(\overline{n} + e_i) - r(\overline{n})] + TV_m(\overline{n} + e_i) - TV_m(\overline{n}) \\
&= [\mu_i(n_i + 1) - \mu_i(n_i)]hp_{i0} + \lambda h 1_{\{i \neq 1, n_1 < L\}} \Delta_i V_m(\overline{n} + e_i) \\
&\quad + \lambda h 1_{\{i=1, n_1+1<L\}} \Delta_i V_m(\overline{n} + e_i) \\
&\quad + \lambda h 1_{\{i \neq 1, n_1 = L\}} \sum_{j=2}^{N} \alpha_j \Delta_i V_m(\overline{n} + e_j) \\
&\quad + \lambda h 1_{\{i=1, n_1+1=L\}} \sum_{j=2}^{N} \alpha_j \Delta_j V_m(\overline{n} + e_j)
\end{aligned}$$

(continues)

$$+ \sum_{j=1}^{N} \mu_j(n_j) h \sum_{k=0}^{N} p_{jk} \Delta_i V_m(\overline{n} - e_j + e_k)$$

$$+ [\mu_i(n_i + 1) - \mu_i(n_i)] h \sum_{k=2}^{N} p_{ik} \Delta_k V_m(\overline{n})$$

$$+ [1 - \lambda h - [\mu_i(n_i + 1) - \mu_i(n_i)] h - \sum_{j=1}^{N} \mu_j(n_j) h] \Delta_i V_m(\overline{n}). \qquad (4.8)$$

The lower estimate $\Delta_i V_{m+1}(\overline{n}) \geq 0$ now directly follows from substituting the induction hypothesis $\Delta_j V_m(.) \geq 0$ for all j, is equal to 0, recalling that the service rate $\mu_i(.)$ is nondecreasing by assumption and observing that the last term is nonnegative by virtue of $h = Q^{-1}$ satisfying (4.3).

The upper estimate $\Delta_i V_{m+1}(\overline{n}) \leq 1$ is concluded similarly by substituting the induction hypothesis $\Delta_i V_m(\overline{n}) \leq 1$ for all j, and observing that all coefficients together then sum up to 1.

We are now able to verify condition (3.8). To this end, recall that the transition structures $q(.,.)$ and $\overline{q}(.,.)$ as according to (4.4) differ only in their arrival rates. With (4.7) and $\gamma = \lambda M^{-1}$ as per (4.1), we then find

$$\left| \sum_{i,j=0}^{N} \Delta(\overline{n}, \overline{n} - e_i + e_j)[V_i(\overline{n} - e_i + e_j) - V_i(\overline{n})] \right|$$

$$= \left| [(M - n)\lambda M^{-1} - \lambda]\{1_{\{n_1 < L\}}[V_i(\overline{n} + e_1) - V_i(\overline{n})] \right.$$

$$\left. + 1_{\{n_1 = L\}} \sum_{j=2}^{N} \alpha_j[V_i(\overline{n} + e_j) - V_i(\overline{n})]\} \right| \leq n\lambda/M. \qquad (4.9)$$

The following choice thus seems appropriate:

$$\Phi(\overline{n}) = n. \qquad (4.10)$$

Lemma 4.2 below investigates whether (3.5) can then be verified.

LEMMA 4.2. *Let W be the sojourn time of a job, in the open version, and $0 = (0, \ldots, 0)$ the empty state. Then for all $t \geq 0$:*

$$\overline{T}_t \Phi(\overline{0}) \leq T_t \Phi(\overline{0}) \leq \lambda W. \qquad (4.11)$$

PROOF. First we will prove that for all $t \geq 0$:

$$\overline{T}_t f(\overline{0}) - T_t f(\overline{0}) \leq 0 \qquad (4.12)$$

for any f such that for all $\bar{n}, \bar{n} + e_j \in S$:

$$f(\bar{n} + e_j) - f(\bar{n}) \geq 0, \qquad (j = 1, \ldots, N). \qquad (4.13)$$

To this end, from (3.10) and the fact that $\bar{S} \subset S$, we obtain similarly to (3.11) or by direct telescoping:

$$(\bar{T}_t - T_t)f(\bar{0}) = \sum_{s=0}^{t-1} \bar{T}_s(\bar{T} - T)T_{t-s-1}f(\bar{0}). \qquad (4.14)$$

As per (3.12) and (4.9) however we also have for any $\bar{n} \in \bar{S}$ and function V

$$(\bar{T} - T)V(\bar{n}) = -[n\lambda/M]\{1_{\{n_1 < L\}}[V(\bar{n} + e_1) - V(\bar{n})]$$

$$+ 1_{\{n_1 = L\}} \sum_{j=2}^{N} \alpha_j[V(\bar{n} + e_j) - V(\bar{n})]\}/Q. \qquad (4.15)$$

Since the operators $\bar{T}_s$ remain restricted to $\bar{S}$ while $\bar{T}_s \psi \geq 0$ whenever $\psi \geq 0$ componentwise, from (4.14) and (4.15) inequality (4.12) is concluded, provided (4.13) holds with f replaced by $T_s f$ for any s, where f itself also satisfies (4.13).

This will be proven by induction to s. For $s = 0$ it is satisfied by definition. Suppose that T_s satisfies (4.13) for $s \leq m$. Then similarly to (4.8) with $h = Q^{-1}$:

$$\Delta_i(T_{m+1}f)(\bar{n}) = T(T_m f)(\bar{n} + e_i) - T(T_m f)(\bar{n})$$
$$= \lambda h 1_{\{i \neq 1, n_1 < L\}} \Delta_i(T_m f)(\bar{n} + e_1) + \lambda h 1_{\{i=1, n_1+1 < L\}} \Delta_i(T_m f)(\bar{n} + e_i)$$

$$+ \lambda h 1_{\{i \neq 1, n_1 = L\}} \sum_{j=2}^{M} \alpha_j \Delta_i(T_m f)(\bar{n} + e_j)$$

$$+ \lambda h 1_{\{i=1, n_1+1=L\}} \sum_{j=2}^{N} \alpha_j \Delta_j(T_m f)(\bar{n} + e_1)$$

$$+ \sum_{j=1}^{N} \mu_j(n_j)h \sum_{k=0}^{N} p_{jl} \Delta_i(T_m f)(\bar{n} - e_j + e_k) \qquad (4.16)$$

$$+ [\mu_i(n_i + 1) - \mu_i(n_i)]h \sum_{k=2}^{N} p_{ik} \Delta_k(T_m f)(\bar{n})$$

$$+ \left[1 - \lambda h - [\mu_i(n_i + 1) - \mu_i(n_i)]h - \sum_{j=1}^{N} \mu_j(n_j)h \right] \Delta_i(T_m f)(\bar{n}).$$

The induction hypothesis $\Delta_j(T_m f) \geq 0$ for all j, now yields as in the proof of Lemma 4.1: $\Delta_i(T_{m+1} f) \geq 0$.

Inequality (4.12) is hereby proven and particularly, since $\Phi(\overline{n}) = n$ satisfies (4.13), also the first inequality of (4.11). To prove the second, we will now inductively prove that, again for f satisfying (4.13), for all $t \geq 0$:

$$T_t f(\overline{0}) \leq T_{t+1} f(\overline{0}). \tag{4.17}$$

For $t = 0$, we have:

$$T f(\overline{0}) = \lambda Q^{-1} f(\overline{0} + e_1) + [1 - \lambda Q^{-1}] f(\overline{0}) \geq f(\overline{0}). \tag{4.18}$$

Assume that (4.17) holds for $t \leq m$ for any f satisfying (4.13). Then from this induction hypothesis and, as proven above, the fact that (4.13) also holds f replaced by $Tf = T_1 f$ when f satisfies (4.13), inequality (4.17) is proven for $t = m + 1$ by:

$$(T_{m+1} f - T_{m+2} f)(\overline{0}) = (T_m - T_{m+1})(T f)(\overline{0}) \leq 0. \tag{4.19}$$

With L the mean number of jobs in the open systems, finally, we conclude from (4.12) and (4.17) with $f(\overline{n}) = \Phi(\overline{n}) = n$ and Little's result:

$$\overline{T}_t(\Phi(\overline{0}) \leq T_t \Phi(\overline{0}) \lim_{t \to \infty} T_t \Phi(\overline{0}) = L = \lambda W. \tag{4.20}$$

From $\overline{r}(.) = r(.)$ as per (4.5). Lemma 4.1, inequality (4.9) and Lemma 4.2, we now directly obtain by applying Theorem 3.1:

THEOREM 4.2. (*Throughput error bound*) *With $\overline{\lambda}$ the throughout of the (γ, M)-finite source system, W the sojourn time of a job in the open version and $\lambda = \gamma M$:*

$$|\overline{\lambda} - \lambda| \leq \lambda W / M. \tag{4.21}$$

Example 4.3 (Deterministic alternate routing) Let all stations be infinite server stations with service parameters μ_i at station i, and assume that $\alpha_2 = 1$, $p_{i,i+1} = 1$ for $i = 2, \ldots, N - 1$ and $p_{N,0} = 1$. Then

$$|\overline{\lambda} - \lambda| \leq \lambda M^{-1} \max \left\{ \mu_1^{-1}, \sum_{i=2}^{N} \mu_i^{-1} \right\}. \tag{4.22}$$

REMARK 4.4. ($\overline{\lambda} \leq \lambda$) By using the lower estimates $\Delta_j V_m \geq 0$ from Lemma 4.1 in (3.11) and (4.9), rather than upper estimates after taking absolute values,

as in (3.13) and noting that $(\bar{r} - r)(.) = 0$, from (3.11), (3.12), (4.9) and (3.4) we can also conclude: $\bar{\lambda} \leq \lambda$. Intuitively, this may seem trivial. Counter-intuitively, however, as per counterexamples of related situations in [32] and [36], such monotonicity results will not generally hold.

Evaluation. Approximations for queueing networks are often based on modifications of the original transition structure and/or the set of admissible states. An analytical tool is provided in order to estimate the accuracy of such approximations. Particularly, as scaling functions such as polynomials are allowed, the results do not require the modifications themselves to be small. The necessary conditions are generally verifiable by inductive Markov reward arguments. A typical application is an open approximation of a large closed system. Extensions to multi-class and nonexponential networks seem possible.

Acknowledgement

References

[1] G. R. Ash, A. H. Kafker and K. R. Krishnan, "Intercity dynamic routing architecture and feasibility". *Proc. 10th Int. Teletraffic Congress, Paper 3.2.2* (Montreal June 1983).

[2] A. Barbour, "Networks of queues and the method of stages", Adv. in Appl. Prob. **8** (1976) 584–591.

[3] F. Baskett, M. Chandy, R. Muntz and J. Palacios, "Open, closed and mixed networks of queues with different classes of customers", *J. A. C. M.* **22** (1975).

[4] D. Y. Burman, J. P. Lehoczky and Y. Lim, "Insensitivity of blocking probabilities in a circuit switching network", *J. of Appl. Prob.* **21** (1982) 850–859.

[5] K. M. Chandy, J. H. Howard and D. F. Towsley, "Product form and local balance in queueing networks", *J. A. C. M.* **24** (1977) 250–263.

[6] K. M. Chandy and A. J. Martin, "A characterization of product-form queueing networks", *J. A. C. M.* **30** (1983) 286–299.

[7] K. Hinderer, "On approximate solutions of finite-stage dynamic programs", in *Dynamic Programming and its Application* (ed. M. Puterman), (Academic Press, New York, 1978) 289–318.

[8] A. Hordjik and N. M. van Dijk, "Networks of queues. Part I: Job-local-balance and the adjoint process. Part II: General routing and service characteristics", *Lecture Notes in Control and Informational Sciences* (Springer Verlag, **60** 1983) 158–205.

[9] ——, "Networks of queues with blocking", in *Performance '81* (ed. K. J. Kylstra), (North Holland, 1983) 51–65.

[10] _____ , "Adjoint process, job-local-balance and insensitivity of stochastic networks", *Bull. of 44th Session of the Int. Stat. Inst.* **50** (1983) 776–788.

[11] J. R. Jackson, "Jobshop-like queueing systems", *Manag. Sci.* **10** (1963) 131–142.

[12] J. G. Kemeny, J. L. Snell and A. W. Knapp, *Denumerable Markov chains* (Van Nostrand, Princeton, NJ, 1966).

[13] F. P. Kelly, *Reversibility and stochastic networks* (Wiley, 1979).

[14] J. Kohlas, *Stochastic methods of operations research* (Cambridge University Press, 1982).

[15] T. G. Kurtz, "Extensions of Trotter's operator semigroup approximations theorems", *J. Functional Analysis* **3** (1969) 111–132.

[16] _____ , "Semigroups of conditional shifts and approximation of Markov processes", *Ann. Prob.* **3** No. 4 (1975) 618–642.

[17] H. J. Kushner, *Probability methods for approximations in stochastic control and for elliptic questions* (Academic Press, New York, 1977).

[18] P. D. Lax and R. D. Richtmeyer, "Survey of the stability of linear finite difference equations", *Comm. Pure Appl. Math.* **9** (1956) 267–293.

[19] T. Meis and U. Marcowitz, *Numerical solution of partial differential equations* (Springer-Verlag, Berlin, 1981).

[20] A. S. Noetzel, "A generalized queueing discipline for product form network solutions", *J. of ACM* **26** (4) (Oct., 1979) 779–793.

[21] T. J. Ott and K. R. Krishnan, "State dependent routing of telephone traffic and the use of separable routing schemes", *Proc. 11th Int. Teletraffic Cong.*, Tokyo, 1985.

[22] S. M. Ross, *Applied probability models with optimization applications* (Holden-Day, San Francisco, 1970).

[23] J. R. Rohlicek and A. S. Willsky, "The reduction of perturbed Markov generators: an algorithm exposing the role of transient states", *J. A. C. M.* **35** (1988) 675–696.

[24] R. Schassberger, "Insensitivity of steady state distributions of generalized semi-Markov processes with speed", *Adv. Appl. Prob.* **10** (1978) 836–851.

[25] M. Schwartz, *Telecommunication networks* (Addison Wesley, 1987).

[26] E. Seneta, *Non-negative matrices and Markov chains* (Springer Verlag, 1980).

[27] J. G. Shanthikumar and D. Yao, "Stochastic monotonicity of the queue length in closed queueing networks", *Oper. Res.* **35** (1987) 583–588.

[28] D. Stoyan, *Comparison methods for queues and other stochastic models* (Wiley, New York, 1983).

[29] H. C. Tijms, *Stochastic modelling and analysis. A computational approach* (Wiley, New York, 1986).

[30] H. F. Trotter, "Approximations of semigroups of operators", *Pac. J. Math.* **8** (1958) 887–919.

[31] N. M. van Dijk, *Controlled Markov processes: time-discretization*, TWI Tract **11** (Amsterdam, 1984).

[32] _____ , "Simple bounds for queueing systems with breakdowns", *Performance Evaluation* **8** (1988) 117–128.

[33] _____ , "On the finite horizon Bellman equation for controlled Markov jump models with unbounded characteristics: Existence and approximation", *Stochastic Proc. Appl.* **28** (1988) 141–157.

[34] _____ , "A formal proof for the insensitivity of simple bounds for finite multiserver non-exponential tandem queues based on monotonicity results", *Stochastic Proc. Appl.* **27** (1988) 261–277.

[35] N. M. van Dijk and M. L. Puterman, "Perturbation theory for Markov reward processes with applications to queueing systems", *Adv. Appl. Prob.* **20** (1988) 79–98.

[36] N. M. van Dijk and J. van der Wal, "Simple bounds and monotonicity results for multi-server exponential tandem queues", to appear in *Queueing Systems*.

[37] J. Walrand, *Introduction to queueing networks* (Prentice-Hall, 1988).

[38] W. Whitt, "Approximations of dynamic programs", *J. Math. Oper. Res.* **3** (1978) 231–243.
[39] ——, "Comparing counting processes and queues", *Adv. Appl. Prob.* **33** (1981) 207–220.
[40] ——, "Open and closed models for networks of queues", *AT&T Bell Lab. Tech. J.* **63** (1984) 1911–1979.
[41] P. Whittle, *Systems in stochastic equilibrium* (1984).
[42] J. Zahorjan, K. C. Sevcik, D. L. Eager and B. Galler, "Balanced job bound analysis of queueing networks", *Commun. of the Assoc. for Comput. Mach.* **25** (1982) 134–141.