Sensitivity results in open, closed, and mixed product-form queueing networks

Zhen Liu, Philippe Nain

UNITÉ DE RECHERCHE
INRIA-SOPHIA ANTIPOLIS

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

Rapports de Recherche

N° 1144

Programme 3
Réseaux et Systèmes Répartis

SENSITIVITY RESULTS IN OPEN, CLOSED, AND MIXED PRODUCT-FORM QUEUEING NETWORKS

Zhen LIU
Philippe NAIN

Décembre 1989

RR-1144

Résultats de Sensibilité pour les Réseaux de Files d'Attente à Forme Produit Ouverts, Fermés et Mixtes

Zhen LIU and Philippe NAIN
INRIA–Sophia Antipolis
2004, Route des Lucioles
06565 Valbonne Cedex
France

Résumé

Dans cet article, nous établissons des formules générales quantifiant l'impact sur les performances d'une modification des paramètres d'un réseau BCMP [3]. Ces formules montrent que la dérivée par rapport à n'importe quelle intensité de service/arrivée de l'espérance mathématique de n'importe quelle fonction Φ de l'état du réseau, s'exprime simplement en terme de fonctions connues de l'état du réseau. Des résultats de sensibilité portant, en particulier, sur les débits et sur les distributions des longueurs des files d'attente sont alors facilement obtenus par différents choix de la fonction Φ.

Mots-Clès: Théorie des Files d'Attente; Réseaux de Files d'Attente à Forme Produit; Sensibilité; Monotonicité.

Sensitivity Results in Open, Closed, and Mixed Product-Form Queueing Networks

Zhen LIU and Philippe NAIN
INRIA–Sophia Antipolis
2004, Route des Lucioles
06565 Valbonne Cedex
France

Abstract

General formulas are proposed to quantify the effects of changing the arrival and service rates in the so-called BCMP network [?]. These formulas relate the derivative of the expectation of any function Φ of the state of the network with respect to any arrival/service rate in the network, to known functions of the state of the network. As an application, sensitivity results of interest bearing on throughputs and on the moments of queue lengths can be derived by appropriate choices of the function Φ.

Keywords: Queueing Theory; Queueing Networks; Product-Form; Sensitivity; Monotonicity.

1 Introduction

Although the use of queueing networks in the modeling and analysis of telecommunication networks and computer systems was initiated with the work of A. K. Erlang [9] at the beginning of the century, it became widespread only after the pioneering work of Jackson [11], Gordon and Newell [10], Baskett, Chandy, Muntz, and Palacios [3], and Kelly [14] on a special class of queueing networks known as *product-form queueing networks*. The nice feature of product-form queueing networks is that for certain classes of Markovian networks, the solution of the balance equations is in the form of a product of simple factors. Afterwards, various generalizations were obtained that extend the product-form property to state-dependent routing [26], non-differentiable service time distribution functions [21], stationary dependent service times [12], and concurrent classes of customers [6], [15]. A fairly complete survey on product-form queueing networks can be found in [8].

Besides these theoretical results, efficient computational algorithms for computing the main performance measures (expected number of customers at a given node, mean waiting times, mean sojourn times, utilization factor of each node, throughputs, etc.) have been proposed by Buzen [4], Reiser and Kobayashi [18], Chandy and Sauer [5], Reiser and Lavenberg [17], and Conway and Georganas [7].

More recently, "first-order qualitative properties" of queueing networks are receiving attention in the literature. These studies aim to determine the sensitivity of various performance measures of the network with respect to particular parameters such as arrival rates, service rates, number of servers, number of customers for closed networks, etc. For closed product-form queueing networks with a single class of customers, Stewart and Stohs [25] have shown that if the service rates are load independent, then the system throughput increases when the service rate of one of the queues increases. This result has been generalized by Shanthikumar and Yao [22] to the case where the service rates are nondecreasing functions of the queue lengths. For the same network, Shanthikumar and Yao [23] have also investigated the effect of increasing the customer population on the queue lengths. Monotonicity properties in product-form queueing networks with loss of customers have been established by Nain [16] and Ross and Yao [19]. Monotonicity results have also been derived lately for non-Markovian queueing networks by Adan and Van der Wal [1], Shanthikumar and Yao [24], and Tsoucas and Walrand [27].

These properties have been obtained using stochastic comparison techniques involving different stochastic orderings, coupling and pathwise arguments. However, these probabilistic methods do not provide formulas enabling one to quantify the impact of a model parameter modification on the network behavior (e.g., rate of increase/decrease of any monotonic function of the state of the

network, etc.).

In this paper, we consider the network studied by Baskett, Chandy, Muntz, and Palacios [3], referred to as the BCMP network, and we analyze the sensitivity of an arbitrary function Φ of the state of the network with respect to the arrival and service rates at any node. More precisely, we show that the derivative of the expectation of any function Φ of the state of the network with respect to any arrival/service intensity, provided this quantity is well-defined, can be expressed simply in terms of known functions of the state of the network. A similar approach has been employed by Jordan and Varaiya [13] to get sensitivity results in a generalized Erlang loss system.

Our results can be used to analyse both quantitative and qualitative effects of modifying model parameters. For instance, monotonicity or nonmonotonicity properties for the moments of queue lengths at a given node, the throughput of a given class of customers at a given node, etc., can be found. In particular, some of the monotonicity properties obtained in [22] can easily be derived from this approach. Furthermore, these results can also be used for optimization purposes by appropriately choosing the "cost function" Φ.

The paper is organized as follows. In section 2 we recall the main features of the BCMP network and introduce some definitions and notation. Section 3 contains the key results of the paper. In section 4 we present some applications.

2 The Model

The network considered in this paper is similar to that analyzed in [3], the only difference being in the modeling of the exogeneous arrivals (see below).

There are $N \geq 1$ stations and $R \geq 1$ different classes of customers. Customers travel through the network and change class according to transition probabilities. Thus a customer of class r which leaves station i upon its service completion will enter station j as a customer of class s with the probability $p_{i,r;j,s}$. The transition matrix $[p_{i,r;j,s}]$ defines a Markov chain whose states are labeled by the pairs (i,r). This Markov chain is assumed to be decomposable into L ergodic subchains. Denote by $E_1, E_2, \ldots, E_L$ the sets of states in each of these subchains. A customer of class r at station i is called a customer of type (i,r). A customer of type E_l is a customer whose type belongs to E_l.

Customers may arrive at the network from NR external sources according to independent Poisson processes. To be more specific, define $M_l(S)$ to be the number of customers of type E_l when the state of the network is S (to be made more precise). Then, the exogeneous arrival rate

of customers of type $(i,r) \in E_l$ is $\lambda_{ir}\gamma_l(M_l(S))$, where $\lambda_{ir} \geq 0$ and that γ_l is an arbitrary mapping $\mathbb{N} \to [0,+\infty)$ that does not depend on $\{\lambda_{ir}\}_{ir}$ (here $\mathbb{N} := \{0,1,2,\ldots\}$).

If $\lambda_{ir} = 0$ for all $1 \leq i \leq N$, $1 \leq r \leq R$, then the network is *closed*. If there exists a partition (L_A, L_B) of the set $\{1,2,\ldots,L\}$, such that

$$\forall (i,r) \in \bigcup_{l \in L_A} E_l \qquad \lambda_{ir} = 0,$$

and

$$\forall l \in L_B \;\; \exists (i,r) \in E_l \qquad \lambda_{ir} > 0,$$

then the network is *mixed*. If for all $1 \leq l \leq L$ there exists $(i,r) \in E_l$ such that $\lambda_{ir} > 0$, then the network is *open*. We say that E_l is closed if $\lambda_{ir} = 0$ for all $(i,r) \in E_l$ and that E_l is open if for some $(i,r) \in E_l$, $\lambda_{ir} > 0$.

In case E_l is open, then we assume that there exists at least one state $(i,r) \in E_l$ such that

$$0 \leq \sum_{(j,s) \in E_l} p_{i,r;j,s} < 1. \tag{2.1}$$

Thus, $1 - \sum_{(j,s) \in E_l} p_{i,r;j,s}$ is the probability that a customer of class r leaves the system upon its service completion at station i.

Four distinct types of service stations are considered:

Type 1. The service discipline is First-Come-First-Served (FCFS) and multiple servers are allowed. All customers have the same service time distribution which is a negative exponential.

Type 2. There is a single server and the service discipline is Processor Sharing (PS).

Type 3. There is an Infinite number of Servers (IS).

Type 4. There is a single server and the service discipline is preemptive resume Last-Come-First-Served (LCFS).

When station i is of type 1, we write $i \in$ FCFS. The notation $i \in$ PS, $i \in$ IS and $i \in$ LCFS will have the obvious meaning.

For $i \in$ FCFS, let $\mu_i\,\alpha_i(n)$ be the service rate at station i when there are $n > 0$ customers at this station. For stations of type 2, 3 or 4, each class of customer may have a distinct and arbitrary service time distribution (GI-servers). For $i \in \{$PS, IS, LCFS$\}$, let $1/\mu_{ir}$ denote the mean value of the service time of a customer of class r at station i. We further assume that α_i is an

arbitrary mapping $\mathbb{N} \to [0, +\infty)$ with $\alpha_i(0) = 0$, which does not depend on the model parameters $\{\lambda_{ir}, \mu_i, \mu_{ir}\}_{ir}$.

Let X_{ir} denote the number of customers of class r at station i. The state of the network is $S = (X_1, X_2, \ldots, X_N)$ where $X_i = (X_{i1}, X_{i2}, \ldots, X_{iR})$ for $i = 1, 2, \ldots, N$. Set $|X_i| = \sum_{r=1}^{R} X_{ir}$.

The joint equilibrium distribution of queue sizes in the network is [3], [14],

$$\mathrm{P}\,(S = (x_1, x_2, \ldots, x_N)) = C\, d(S)\, g_1(x_1)\, g_2(x_2) \cdots g_N(x_n), \tag{2.2}$$

where C is a normalizing constant and $d(S)$ is a function of the number of customers in the system. If the network is closed then $d(S) \equiv 1$, otherwise

$$d(S) = \prod_{l \in L_{\mathcal{O}}} \left\{ \Lambda_l^{M_l(S)} \prod_{m=0}^{M_l(S)-1} \gamma_l(m) \right\}, \tag{2.3}$$

where $L_{\mathcal{O}} := \{l \mid 1 \le l \le L,\ E_l \text{ is open}\}$ and where $\Lambda_l := \sum_{(i,r) \in E_l} \lambda_{ir}$ for all $l \in L_{\mathcal{O}}$.

Let $x_i = (n_{i1}, n_{i2}, \ldots, n_{iR})$ and $n_i = \sum_{r=1}^{R} n_{ir}$. Each $g_i(x_i)$ in (2.2) is a function that depends on the type of station i:

- if station i is of type 1, then

$$g_i(x_i) = n_i! \left(\prod_{m=1}^{n_i} \frac{1}{\mu_i\, \alpha_i(m)} \right) \left(\prod_{r=1}^{R} \frac{e_{ir}^{n_{ir}}}{n_{ir}!} \right); \tag{2.4}$$

- if station i is of type 2 or 4, then

$$g_i(x_i) = n_i! \prod_{r=1}^{R} \left\{ \left(\frac{e_{ir}}{\mu_{ir}} \right)^{n_{ir}} \left(\frac{1}{n_{ir}!} \right) \right\}; \tag{2.5}$$

- if station i is of type 3, then

$$g_i(x_i) = \prod_{r=1}^{R} \left\{ \left(\frac{e_{ir}}{\mu_{ir}} \right)^{n_{ir}} \left(\frac{1}{n_{ir}!} \right) \right\}. \tag{2.6}$$

The e_{ir}'s satisfy the following set of linear equations [3]:

$$e_{ir} = q_{ir}(l) + \sum_{(j,s) \in E_l} e_{js}\, p_{j,s;i,r}, \tag{2.7}$$

for all $(i,r) \in E_l$, $l = 1, 2, \ldots, L$, where

$$q_{ir}(l) := \begin{cases} \dfrac{\lambda_{ir}}{\Lambda_l}, & \text{if } \lambda_{ir} > 0; \\ 0, & \text{if } \lambda_{ir} = 0. \end{cases} \tag{2.8}$$

3 Sensitivity Results

From now on we assume that the network is in equilibrium. Two types of sensitivity results are derived in this section: sensitivity results with respect to service rates (Theorem 3.1) and those with respect to exogeneous arrival rates (Theorem 3.2).

Let us introduce some notation. Let Φ be a mapping $\left(\mathbb{N}^{\mathrm{R}}\right)^N \to [0, +\infty)$. We say that Φ satisfies assumption A1 (resp. A2, A3) if

1. $\dfrac{\partial \Phi(x)}{\partial z}$, $\mathrm{E}\left[\dfrac{\partial \Phi(S)}{\partial z}\right]$ and $\dfrac{\partial \mathrm{E}[\Phi(S)]}{\partial z}$ exist;

2. $\dfrac{\partial \mathrm{E}[\Phi(S)]}{\partial z} = \sum\limits_{y \in \mathcal{F}} \dfrac{\partial \{\Phi(y) \mathrm{P}(S = y)\}}{\partial z}$,

for $z \in \{\mu_i\}_i$ (resp. $z \in \{\mu_{ir}\}_{ir}$, $z \in \{\lambda_{ir}\}_{ir}$), where $\mathcal{F}$ denotes the set of all feasible states for the network under consideration (i.e., open, closed, or mixed).

We now state the first result of this section.

Theorem 3.1 *Assume that Φ satisfies assumption A1. Then, for $i \in$ FCFS,*

$$\frac{\partial \mathrm{E}[\Phi(S)]}{\partial \mu_i} = \mathrm{E}\left[\frac{\partial \Phi(S)}{\partial \mu_i}\right] - \frac{\mathrm{cov}\left(\Phi(S), |X_i|\right)}{\mu_i}. \tag{3.1}$$

Assume that Φ satisfies assumption A2. Then, for $i \in \{\mathrm{PS}, \mathrm{IS}, \mathrm{LCFS}\}$, $1 \le r \le R$,

$$\frac{\partial \mathrm{E}[\Phi(S)]}{\partial \mu_{ir}} = \mathrm{E}\left[\frac{\partial \Phi(S)}{\partial \mu_{ir}}\right] - \frac{\mathrm{cov}\left(\Phi(S), X_{ir}\right)}{\mu_{ir}}. \tag{3.2}$$

Proof. Assume first that $i \in$ FCFS. It follows from (2.2) and assumption A1 that

$$\frac{\partial \mathrm{E}[\Phi(S)]}{\partial \mu_i} = \sum_{y \in \mathcal{F}} \frac{\partial}{\partial \mu_i} \left\{ C\, d(y)\, \Phi(y) \prod_{k=1}^{N} g_k(x_k) \right\},$$

$$= C \sum_{y \in \mathcal{F}} d(y) \left(\frac{\partial \Phi(y)}{\partial \mu_i} \right) \prod_{k=1}^{N} g_k(x_k) + \left(\frac{\partial C}{\partial \mu_i} \right) \sum_{y \in \mathcal{F}} d(y)\, \Phi(y) \prod_{k=1}^{N} g_k(x_k)$$

$$+ \; C \sum_{y \in \mathcal{F}} d(y)\, \Phi(y) \left(\frac{\partial}{\partial \mu_i} \prod_{k=1}^{N} g_k(x_k) \right). \tag{3.3}$$

Using (2.4) it is easily checked that

$$\frac{\partial}{\partial \mu_i} \prod_{k=1}^{N} g_k(x_k) = -\frac{n_i}{\mu_i} \prod_{k=1}^{N} g_k(n_k). \tag{3.4}$$

Now differentiating the identity

$$C = \frac{1}{\sum_{y \in \mathcal{F}} d(y) \prod_{k=1}^{N} g_k(x_k)}, \tag{3.5}$$

we obtain

$$\frac{\partial C}{\partial \mu_i} = -C^2 \left(\frac{\partial}{\partial \mu_i} \sum_{y \in \mathcal{F}} d(y) \prod_{k=1}^{N} g_k(x_k) \right),$$

which, together with (3.4), yields

$$\frac{\partial C}{\partial \mu_i} = \frac{C}{\mu_i} \mathrm{E}\left[|X_i| \right]. \tag{3.6}$$

Consequently, cf. (3.3), (3.4), and (3.6),

$$\begin{aligned} \frac{\partial \mathrm{E}[\Phi(S)]}{\partial \mu_i} &= \mathrm{E}\left[\frac{\partial \Phi(S)}{\partial \mu_i} \right] + \frac{\mathrm{E}[|X_i|]\, \mathrm{E}[\Phi(S)]}{\mu_i} - \frac{\mathrm{E}[\Phi(S)\, |X_i|]}{\mu_i}, \\ &= \mathrm{E}\left[\frac{\partial \Phi(S)}{\partial \mu_i} \right] - \frac{\mathrm{cov}\left(\Phi(S), |X_i| \right)}{\mu_i}. \end{aligned}$$

The proof of (3.2) follows similarly by observing that from (2.5) and (2.6),

$$\frac{\partial}{\partial \mu_{ir}} \prod_{k=1}^{N} g_k(x_k) = -\frac{n_{ir}}{\mu_{ir}} \prod_{k=1}^{N} g_k(x_k).$$

This yields

$$\frac{\partial C}{\partial \mu_{ir}} = \frac{C}{\mu_{ir}} \mathrm{E}\left[X_{ir} \right],$$

for $i \in \{\mathrm{PS}, \mathrm{IS}, \mathrm{LCFS}\}$. ∎

Let us assume that the network is either open or mixed. We now state the second result of this section that establishes sensitivity results with respect to exogeneous arrival rates.

Theorem 3.2 *Assume that Φ satisfies assumption A3. If there exists an external source of customers of type $(i,r) \in E_l$, then*

$$\frac{\partial \mathrm{E}[\Phi(S)]}{\partial \lambda_{ir}} = \mathrm{E}\left[\frac{\partial \Phi(S)}{\partial \lambda_{ir}}\right] + \frac{\mathrm{cov}(\Phi(S), M_l(S))}{\Lambda_l} + \left(\frac{\partial e_{ir}}{\partial \lambda_{ir}}\right) \frac{\mathrm{cov}(\Phi(S), X_{ir})}{e_{ir}}. \tag{3.7}$$

In particular, if $E_l = \{(i,r)\}$, then

$$\frac{\partial \mathrm{E}[\Phi(S)]}{\partial \lambda_{ir}} = \mathrm{E}\left[\frac{\partial \Phi(S)}{\partial \lambda_{ir}}\right] + \frac{\mathrm{cov}(\Phi(S), M_l(S))}{\Lambda_l}. \tag{3.8}$$

Proof. Assume that $(i,r) \in E_l$. We have, cf. (2.2) and assumption A3,

$$\begin{aligned} \frac{\partial \mathrm{E}[\Phi(S)]}{\partial \lambda_{ir}} &= C \sum_{y \in \mathcal{F}} \left(\frac{\partial \Phi(y)}{\partial \lambda_{ir}}\right) d(y) \prod_{k=1}^{N} g_k(x_k) + \left(\frac{\partial C}{\partial \lambda_{ir}}\right) \sum_{y \in \mathcal{F}} \Phi(y)\, d(y) \prod_{k=1}^{N} g_k(x_k) \\ &+ C \sum_{y \in \mathcal{F}} \Phi(y) \left(\frac{\partial d(y)}{\partial \lambda_{ir}}\right) \prod_{k=1}^{N} g_k(x_k) + C \sum_{y \in \mathcal{F}} \Phi(y)\, d(y) \left(\frac{\partial}{\partial \lambda_{ir}} \prod_{k=1}^{N} g_k(x_k)\right). \end{aligned} \tag{3.9}$$

From (2.2), (2.4)-(2.6), it is easily seen that

$$\frac{\partial}{\partial \lambda_{ir}} \prod_{k=1}^{N} g_k(x_k) = \frac{n_{ir}}{e_{ir}} \left(\frac{\partial e_{ir}}{\partial \lambda_{ir}}\right) \prod_{k=1}^{N} g_k(x_k). \tag{3.10}$$

It follows from (2.3) that

$$\frac{\partial d(y)}{\partial \lambda_{ir}} = \frac{M_l(y)}{\Lambda_l}\, d(y). \tag{3.11}$$

By differentiating (3.5) with respect to λ_{ir} and by using (3.10) and (3.11), we obtain

$$\frac{\partial C}{\partial \lambda_{ir}} = -\left(\frac{\mathrm{E}[M_l(S)]}{\Lambda_l} + \left(\frac{\partial e_{ir}}{\partial \lambda_{ir}}\right) \frac{\mathrm{E}[X_{ir}]}{e_{ir}}\right) C. \tag{3.12}$$

Substituting (3.10), (3.11), and (3.12) into (3.9) yields (3.7).

It remains to check that

$$\frac{\partial e_{ir}}{\partial \lambda_{ir}} = 0$$

in (3.7) whenever $E_l = \{(i,r)\}$. In this case, $q_{ir} = 1$ from (2.8), so $e_{ir} = 1/(1 - p_{i,r;i,r})$ from (2.7), and therefore $\partial e_{ir}/\partial \lambda_{ir} = 0$, which completes the proof. ∎

Remark 3.1 The quantity $\partial e_{ir}/\partial\lambda_{ir}$ involved in (3.7) can be computed by solving the linear equations, cf. (2.7),

$$\frac{\partial e_{js}}{\partial\lambda_{ir}} = \frac{\partial q_{js}(l)}{\partial\lambda_{ir}} + \sum_{(j',s')\in E_l} p_{j',s';j,s}\,\frac{\partial e_{j's'}}{\partial\lambda_{ir}}, \qquad \forall (j,s)\in E_l, \tag{3.13}$$

where (cf. (2.8))

$$\frac{\partial q_{js}(l)}{\partial\lambda_{ir}} = \begin{cases} \dfrac{\Lambda_l - \lambda_{js}}{\Lambda_l^2}, & \text{if } (j,s) = (i,r); \\ -\dfrac{1}{\Lambda_l^2}, & \text{if } (j,s) \neq (i,r). \end{cases}$$

It follows from assumption (2.1) that $\{\partial e_{js}/\partial\lambda_{ir}\}_{(j,s)\in E_l}$ is the unique solution of (3.13).

Remark 3.2 As mentioned in [3], pp. 256-257, the product-form (2.2) is preserved when various forms of state-dependent service rates are incorporated in stations PS, IS and LCFS. If so, formulas (2.4)-(2.6) have to be modified accordingly but Theorems 3.1 and 3.2 remain valid provided the introduced service dependencies do not involve the model parameters $\{\lambda_{ir}, \mu_i, \mu_{ir}\}_{ir}$.

4 Applications

Many results of practical interest can be derived from Theorems 3.1 and 3.2. We point out some of them below. Recall that any point x in the domain of definition of the function Φ can be written as $x = (x_1, x_2, \ldots, x_N)$ with $x_i = (n_{i1}, n_{i2}, \ldots, n_{iR})$, where $n_{ir} \in \mathbb{N}$. Also recall that $n_i = \sum_{r=1}^{R} n_{ir}$.

4.1 Sensitivity of queue lengths

Let f be any nondecreasing mapping $\mathbb{N} \to [0, +\infty)$.

1. $\Phi(x) = f\left(\sum_{j\in N_{\mathcal{I}}} n_j\right)$, $N_{\mathcal{I}} \subset \{1, 2, \ldots, N\}$.

   Then, for $i \in$ FCFS, cf. (3.1),

$$\frac{\partial \mathrm{E}\left[f\left(\sum_{j\in N_{\mathcal{I}}} |X_j|\right)\right]}{\partial\mu_i} = -\frac{\mathrm{cov}\left(f\left(\sum_{j\in N_{\mathcal{I}}} |X_j|\right), |X_i|\right)}{\mu_i}. \tag{4.1}$$

   Assume that $N_{\mathcal{I}} = \{i\}$. Then, the right-hand side of (4.1) is nonpositive since the random variable $f(|X_i|)$ is *stochastically increasing* in $|X_i|$, which implies that $\mathrm{cov}\,(f(|X_i|), |X_i|) \geq 0$ (see [2], Theorem 4.7, p. 146). Therefore,

- $|X_i|$ is decreasing in μ_i in the sense of *stochastic ordering* ([20], pp. 251-252) for all $i \in$ FCFS.
- As a consequence, $\sum_{j \neq i} |X_j|$ is increasing in μ_i in the sense of stochastic ordering when the network is closed, for all $i \in$ FCFS.

2. $\Phi(x) = f(n_{js})$ with $(j,s) \in E_l$.

   Then, for $i \in \{$PS, IS, LCFS$\}$, cf. (3.2),

$$\frac{\partial \mathrm{E}[f(X_{js})]}{\partial \mu_{ir}} = -\frac{\mathrm{cov}(f(X_{js}), X_{ir})}{\mu_{ir}}. \tag{4.2}$$

   Similar to case 1 above, we deduce from (4.2) that

   - X_{ir} is decreasing in μ_{ir} in the sense of stochastic ordering;
   - $\sum_{(j,s) \neq (i,r)} |X_{js}|$ is increasing in μ_{ir} in the sense of stochastic ordering when the network is closed,

   for all $i \in \{$PS, IS, LCFS$\}$, $1 \leq r \leq R$.

3. Assume now that E_l is open and that $E_l = \{(j,s)\}$. Then, cf. (3.8),

$$\frac{\partial \mathrm{E}[f(X_{js})]}{\partial \lambda_{js}} = \frac{\mathrm{cov}(f(X_{js}), X_{js})}{\lambda_{js}}, \tag{4.3}$$

   which shows that

   - X_{js} is increasing in λ_{js} in the sense of stochastic ordering, for all $1 \leq j \leq N$, $1 \leq s \leq R$.

The interested reader can also derive formulas for $\partial \mathrm{E}[f(X_{js})]/\partial \mu_i$ for $i \in$ FCFS, and for $\partial \mathrm{E}[f(X_j)]/\partial \mu_{ir}$ for $i \in \{$PS, IS, LCFS$\}$.

Note that the results in applications 1 and 2 generalize earlier results of Shanthikumar and Yao (see [22], Corollary 3.1) to arbitrary state-dependent service rates (see Remark 3.2) and multi-class closed/open/mixed queueing networks.

4.2 Sensitivity of throughputs

1. $\Phi(x) = \mu_{js} \frac{n_{js}}{n_j} 1_{\{n_{js}>0\}}$ with $(j,s) \in E_l$.

Then $\mathrm{E}[\Phi(S)] = \mu_{js}\mathrm{E}\left[(X_{js}/|X_j|)\,\mathbf{1}_{\{X_{js}>0\}}\right]$ is the throughput of customers of class s in station j if $j \in \mathrm{PS}$ and if E_l is closed.

For $i \in \mathrm{PS}$, we have, cf. (3.2),

$$\frac{\partial \mathrm{E}[\Phi(S)]}{\partial \mu_{ir}} = \mathbf{1}_{\{(i,r)=(j,s)\}}\frac{\mathrm{E}[\Phi(S)]}{\mu_{ir}} - \frac{\mu_{js}}{\mu_{ir}}\,\mathrm{cov}\left(\frac{X_{js}}{|X_j|}\,\mathbf{1}_{\{X_{js}>0\}}; X_{ir}\right). \tag{4.4}$$

If only customers of class r may visit station i and if $(j,s) = (i,r)$, then (4.4) becomes

$$\frac{\partial \mathrm{E}[\Phi(S)]}{\partial \mu_{ir}} = \mathrm{P}(X_{ir} > 0) - \mathrm{E}[X_{ir}]\,\mathrm{P}(X_{ir} = 0).$$

2. $\Phi(x) = \mu_{js}\, n_{js}$ with $(j,s) \in E_l$.

Then $\mathrm{E}[\Phi(S)] = \mu_{js}\mathrm{E}[X_{js}]$ is the throughput of customers of class s in station j if $j \in \mathrm{IS}$ and if E_l is closed.

We have, for $i \in \mathrm{IS}$, cf. (3.2),

$$\frac{\partial \mathrm{E}[\Phi(S)]}{\partial \mu_{ir}} = \mathbf{1}_{\{(i,r)=(j,s)\}}\mathrm{E}[X_{js}] - \frac{\mu_{js}}{\mu_{ir}}\,\mathrm{cov}(X_{js}, X_{ir}). \tag{4.5}$$

If $(i,r) = (j,s)$, then (4.5) reduces to

$$\frac{\partial \mathrm{E}[\Phi(S)]}{\partial \mu_{ir}} = \mathrm{E}[X_{ir}] - \mathrm{var}(X_{ir}).$$

3. $\Phi(x) = \mu_j \alpha_j(n_j)$ with $(j,s) \in E_l$.

Recall that $\alpha_j(0) = 0$. Then $\mathrm{E}[\Phi(S)] = \mu_j\mathrm{E}[\alpha_j(|X_j|)]$, and $\mathrm{E}[\Phi(S)]$ is the throughput of station j if $j \in \mathrm{FCFS}$ and if E_l is closed.

We have, for $i \in \mathrm{FCFS}$, cf. (3.1),

$$\frac{\partial \mathrm{E}[\Phi(S)]}{\partial \mu_i} = \mathbf{1}_{\{i=j\}}\mathrm{E}[\alpha_j(|X_j|)] - \frac{\mu_j}{\mu_i}\,\mathrm{cov}(\alpha_j(|X_j|), |X_i|). \tag{4.6}$$

If $i = j$, then (4.6) simply becomes

$$\frac{\partial \mathrm{E}[\Phi(S)]}{\partial \mu_i} = \mathrm{E}[\alpha_i(|X_i|)] - \mathrm{cov}(\alpha_i(|X_i|), |X_i|).$$

The reader is encouraged to establish cross term derivatives, e.g., the derivative of an FCFS station with respect to the service rate of an IS station.

4.3 Numerical approach to optimization problems

Many optimization problems in queueing networks are formulated in terms of the search for an optimal value of a system parameter so that some cost function is minimized/maximized. In general, the cost function is expressed as the mathematical expectation of a function of the system state.

One of the most direct ways of solving this kind of problem is to compute the derivative of the cost function with respect to the system parameter and to find the minima/maxima of the cost function. Unfortunately, this approach is often unfeasible because the derivative of the cost function is hard to obtain.

In product-form networks, however, owing to the Theorems 3.1 and 3.2, the determination of the derivative is reduced to the computation of the covariance of certain state variables. This provides a new approach, at least from a numerical point of view, to the solution of certain optimization problems.

As an example, let the cost function be $\mathrm{E}[a|X_i| + b\mu_i]$, where a and b are nonnegative real coefficients representing holding and service costs, respectively, and where $|X_i|$ is the queue length of an FCFS station i in a BCMP network and μ_i is the service rate. One might want to find a value of μ_i that minimizes this cost function. Using Theorem 3.1, one obtains

$$\frac{\partial \mathrm{E}[aX_i + b\mu_i]}{\partial \mu_i} = b - \frac{a}{\mu_i}\mathrm{var}(X_i). \tag{4.7}$$

Thus, the problem reduces to the computation of the roots of the right-hand side of equation (4.7).

4.4 Correlation between state variables

Let us choose $\Phi(x)$ as in application 1 of section 4.1. Assume that the network is closed and that there is only one class of customers. Let T_i denote the throughput of station i, $1 \leq i \leq N$ (cf. section 4.2). Then,

$$\mathrm{cov}(f(|X_j|), |X_i|) \leq 0; \tag{4.8}$$

$$\mathrm{cov}(T_j, |X_i|) \leq 0, \tag{4.9}$$

for all $i \neq j$, and

$$\mathrm{cov}\left(f\left(\sum_{j \in N_I} |X_j|\right), |X_i|\right) \geq 0, \tag{4.10}$$

for all $i \in N_{\mathcal{I}}$, provided the service rate at each station is a nondecreasing function of the total number of customers at that station.

The proof of (4.8) and (4.10) (resp. (4.9)) follows directly from (4.1) (resp. (4.4)-(4.6)) together with the following result due to Shanthikumar and Yao ([22], Corollary 3.1). In a closed network with a single class of customers, let $(N_{\mathcal{A}}, N_{\mathcal{B}})$ denote any nontrivial partition of $\{1,2,\ldots,N\}$; then

1. $|X_j|$ $(j \in N_{\mathcal{A}})$ is increasing in μ_i in the sense of stochastic ordering for all $i \in N_{\mathcal{B}}$;
2. $\sum_{j \in N_{\mathcal{B}}} |X_j|$ is decreasing in μ_i in the sense of stochastic ordering for all $i \in N_{\mathcal{B}}$,
3. T_j is nondecreasing in μ_i for all $i \in N_{\mathcal{B}}$, $1 \leq j \leq N$,

provided (i) the service rate at station i is nondecreasing in $|X_i|$ for all $i \in N_{\mathcal{A}}$, (ii) card $N_{\mathcal{B}} = 1$ or card $N_{\mathcal{B}} \geq 2$ and the service rate at station i is nondecreasing in $|X_i|$ for all $i \in N_{\mathcal{B}}$.

5 Conclusions

In this paper, we have revisited quantitive and qualitive analysis of the BCMP network. Various formulas have been established that relate the derivative of the expectation of any function Φ of the state of the network with respect to any arrival/service rate in the network, to known functions of the state of the network.

As an application of these results, we have shown that monotonicity/nonmonotonicity properties of the throughputs and of the queue length moments (some of which are known in the literature) can easily be derived by appropriately choosing the function Φ. As formulas in section 4.2 suggest, in general the throughputs in mixed/closed BCMP networks are nonmonotonic in the system parameters.

It is worthwhile to note that the results obtained in the paper provide an approach to the numerical computation of the derivatives of the expectation of an arbitrary function of the system state with respect to arrival and service rates in the queueing network, and is thus of particular interest for numerical solutions of various optimization problems arising in queueing systems.

References

[1] Adan, I. and Van der Wal, J. Monotonicity of the throughput in single server production and assembly networks with respect to the buffer sizes. Proc. Queueing Networks with Blocking,

H. G. Perros and T. Altiok, Eds, Elsevier Science Publishers B. V. (North Holland, 1989), 147-171.

[2] Barlow, R. E. and Proschan, F. *Statistical Theory of Reliability and Life Testing Probability Models*, Holt, Rinehart & Winston, New York, 1975.

[3] Baskett, F., Chandy, K. M., Muntz, R. R., and Palacios, F. G. Open, closed and mixed network of queues with different classes of customers. *J. ACM* 22, 2 (Apr. 1975), 248-260.

[4] Buzen, J. P. A computational algorithm for closed queueing netwoks with exponential servers. *Comm. ACM* 14, 9 (Sep. 1973), 527-531.

[5] Chandy, K. M. and Sauer, C. H. Computational algorithms for product form queueing networks. *Comm. ACM* 23, 10 (Oct. 1980), 573-583.

[6] Chiola, G., Marsan, M. A., and Balbo, G. Product-form solution techniques for the performance analysis of multiple-bus multiprocessor systems with nonuniform memory references. *IEEE Trans. Comptr.* C-37, 5 (May 1988), 532-540.

[7] Conway, A. E. and Georganas, N. D. RECAL: A new efficient algorithm for the exact analysis of multiple-chain closed queueing networks. *J. ACM* 33, 4 (Oct. 1984), 768-791.

[8] Disney, R. L. and König, D. Queueing networks: A survey of their random processes. *SIAM Review* 27, 3 (Sep. 1985), 335-403.

[9] Erlang, A. K. On the rational determination of the number of circuits in *The Life and Works of A. K. Erlang*, E. Brockmeyer, H. L. Halstrøm and A. Jensen. *Trans. Danish Academy of Tech. Sci.* (1948), 216-221.

[10] Gordon, W. J. and Newell, G. F. Closed queueing systems with exponential servers. *Oper. Res.* 15, 2 (Apr. 1967), 252-265.

[11] Jackson, J. R. Jobshop-like queueing systems. *Manage. Sci.* 10, 1 (1963), 131-142.

[12] Jansen, U. and König, D. Insentivity and steady state probabilities in product form for queueing networks. *Elektron. Inform. Kybernet.* 16 (1980), 385-397.

[13] Jordan, S. and Varaiya, P. Throughput in multiple service, multiple resource communication networks. Preprint (Apr. 1989).

[14] Kelly, F. *Reversibility and Stochastic Networks.* John Wiley & Sons, New York, 1979.

[15] Le Boudec, J. Y. A BCMP extension to multiserver stations with concurrent classes of customers. *Perf. Eval. Review* 14, 1 (1986), 78-91.

[16] Nain, P. Qualitative properties of the Erlang blocking model with heterogeneous user requirements. INRIA Report No. 1018 (Apr. 1989). To appear in *QUESTA*.

[17] Reiser, M. and Lavenberg, S. S. Mean-value analysis of closed multichain queueing networks. *J. ACM* 27, 2 (Apr. 1980), 313-322.

[18] Reiser, M. and Kobayashi, H. Queueing networks with multiple closed chains: theory and computational algorithms. *IBM J. Res. and Develop.* 19 (May 1975), 283-294.

[19] Ross, K. W. and Yao, D. D. Monotonicity properties for the stochastic knapsack. Preprint. Submitted for publication to *IEEE Trans. Inform. Theory*.

[20] Ross, S. M. *Stochastic Processes*. John Wiley & Sons, New York, 1983.

[21] Samelson, C. L. and W. G. Bulgren. A note on product-form solution for queueing networks with Poisson arrivals and general service-time distributions with finite means. *J. ACM* 29, 3 (Jul. 1982), pp. 830-840.

[22] Shanthikumar, J. G. and Yao, D. D. The effect of increasing service rates in a closed queueing network. *J. Appl. Prob.* 23 (1986), 474-483.

[23] Shanthikumar, J. G. and Yao, D. D. Stochastic monotonicity of the queue lengths in closed queueing networks. *Oper. Res.* 35, 4 (1987), 583-588.

[24] Shanthikumar, J. G. and Yao, D. D. Stochastic monotonicity in general queueing networks. *J. Appl. Prob.* 26 (1989), 413-417.

[25] Stewart, W. J. and Stohs, W. P. Some equivalence results for load-independent exponential queueing networks. *IEEE Trans. Soft. Eng.* SE-10, 4 (1984), 414-422.

[26] Towsley, D. Queueing network models with state-dependent routing. *J. ACM* 27, 2 (Apr. 1980), 323-337.

[27] Tsoucas, P. and Walrand, J. Monotonicity of throughput in non-Markovian networks. *J. Appl. Prob.* 26, 1 (Mar. 1989), 134-141.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique