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

Zhen LIU and Philippe NAIN
{liu,nain}@sophia.inria.fr
INRIA
2004, Route des Lucioles
B.P. 93, 06902 Sophia Antipolis Cedex
France

November 1991

Abstract

General formulas are proposed to quantify the effects of changing the model parameters in the so-called BCMP network [6]. These formulas relate the derivative of the expectation of any function of both the state and the paramaters of the network with respect to any model parameter (i.e., arrival rate, mean service demand, service rate, visit ratio, traffic intensity) to known functions of the state variables. Applications of our results to sensitivity analysis and optimization problems are given.

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 [21] at the beginning of the century, it became widespread only after the pioneering work of Jackson [30], Gordon and Newell [25], Baskett, Chandy, Muntz, and Palacios [6], Kelly [33], on a special class of queueing networks known as *product form queueing networks*. A 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 extended the product form property to state-dependent routing and/or blocking phenomena (Towsley [65], Hordijk and van Dijk [26]), service disciplines that may depend on the class of the customer (Chandy and Martin [12]), non-differentiable service time distribution functions (Samelson and Bulgren [53]), stationary dependent service times (Jansen and König [31]), and concurrent classes of customers (Chiola et al. [14], Le Boudec [38]). A fairly complete survey on product form queueing networks can be found in Disney and König [20].

Besides these theoretical results, efficient computational algorithms for computing the primary 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 [8], Reiser and Kobayashi [49], Chandy and Sauer [13], Reiser and Lavenberg [50], McKenna and Mitra [39, 41], Strelen [59], Conway and Georganas [16] for BCMP networks, and Akyildiz and von Brand [2] for product form networks with blocking mechanisms.

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 [58] 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 [55] to the case where the service rates are nondecreasing functions of the queue lengths. For the same network, Shanthikumar and Yao [56] 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 [42] and Ross and Yao [51]. Monotonicity results have also been derived lately for non-Markovian queueing networks by Adan and Van der Wal [1], Shanthikumar and Yao [57], Tsoucas and Walrand [68].

These properties have been obtained using stochastic comparison techniques involving different stochastic orderings, coupling and pathwise arguments. However, these probabilistic methods do not provide *gradient estimates* allowing 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.).

Simulations can also be used to derive gradient estimates, as it is used in classical performance analysis of discrete event systems (e.g., computation of the buffer occupancy in queueing networks). To do so, methods avoiding lengthy, biased and inaccurate computations have been proposed in the last decade. Of particular interest are the methods based on *pertubation analysis* (Ho [27, 28, 29], Cao [9, 10, 11], Suri [62, 64]) and on *likelihood ratios* (Glynn [23], Glynn and Sanders [24]) which enable to compute gradient estimates by observing only one sample path (i.e., by running only one simulation). A comprehensive survey on pertubation analysis can be found in Suri [63] and an overview on likelihood ratio gradient estimation can be found in Glynn [22]. In particular, likelihood ratios have been successfully applied to light traffic analysis of open queueing networks (Reiman and Simon [47], Reiman and Weiss [48]). A new proof of the result in [47] based on rare perturbation analysis and Campbell's formula for point processes has recently been proposed by Baccelli and Brémaud [4].

There also exists a perturbation theory for general Markov reward processes based on Markov reward equations. Pioneering results have been obtained by Schweitzer [54] and Meyer [40] for stationary probabilities of finite Markov chains. Lately, error bounds of reward functions have been derived by van Dijk and Puterman [72] and van Dijk [69] for various Markov reward models. The Markov reward method has also been used by van Dijk [70] to obtain monotonicity results for non-product form queueing networks.

Our objective is to derive *explicit formulas* for the derivatives of various performance measures in the network studied by Baskett, Chandy, Muntz, and Palacios [6] (BCMP network). More precisely, we show that the derivative of the expectation of any function of the state of the system (defined to be the state of the network and of the model parameters) with respect to any model parameter (i.e., arrival rate, mean service demand, service rate, visit ratio, traffic intensity), simply expresses in terms of known functions of the state variables (queue lengths, typically). Then, the derivative can be evaluated with the aid of the aforementioned computational algorithms.

Observe that partial derivatives of queueing measures in product form queueing networks have been used in many previous works. In [66, 67] Trivedi et al. considered the problem of the optimal selection of CPU speed and device capacities in a computer system so as to maximize the

throughput. Kobayashi, Gerla and de Souza e Silva in [36, 18] considered routing and load balancing problems with the aim of maximizing system throughput. Partial derivatives were also used by Kenevan and von Mayrhauser [34] to establish convexity results for a single class closed product form queueing network. Suri [60, 61] obtained expressions in terms of the difference of probability distributions for the partial derivatives of specific performance measures such as throughput in closed queueing networks with homogeneous service times. Jordan and Varaiya [32] carried out similar analysis to get sensitivity results in a generalized Erlang loss system. This approach was also employed to devise various algorithms for computing queue length moments using MVA-like recursions (cf. Conway et al. [17], de Souza e Silva and Muntz [19] and Strelen[59]).

Our results can be used to analyze both quantitative and qualitative effects due to any changes in the model parameters of a BCMP network. In particular, first order properties (monotonicity/nonmonotonicity) and second order properties (concavity, convexity) of the system performance measures can be brought to light via our formulas.

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 whereas Section 4 presents applications including monotonicity and optimization issues, as well as some numerial results. Section 5 contains a discussion on implementation issues.

2 The Model

The network considered in this paper is similar to that analyzed in [6], 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 $\mathbf{P} = [p_{i,r;j,s}]$ 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 will be said to be 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 $N \times R$ external sources according to independent Poisson processes. To be more specific, define $M_l(\mathbf{S})$ as the number of customers of type E_l when the state of the network is $\mathbf{S}$. Then, the exogeneous arrival rate of customers of type $(i, r) \in E_l$ when the

state of the network is $\mathbf{S}$ is $\lambda_{ir}\gamma_l(M_l(\mathbf{S}))$, where γ_l is an arbitrary mapping $\mathbb{N} \to [0, +\infty)$ and $\lambda_{ir} \geq 0$ (throughout this paper $\mathbb{N} := \{0, 1, 2, \ldots\}$).

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 nontrivial partition $(L_{\mathcal{A}}, L_{\mathcal{B}})$ of the set $\{1, 2, \ldots, L\}$ such that

$$\forall (i,r) \in \bigcup_{l \in L_{\mathcal{A}}} E_l, \quad \lambda_{ir} = 0,$$

and

$$\forall l \in L_{\mathcal{B}} \;\; \exists (i,r) \in E_l, \quad \lambda_{ir} > 0,$$

then the network is *mixed*. If for all l, $1 \leq l \leq L$ 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 $(i,r) \in E_l$ (in which case $M_l(\mathbf{S})$ is constant) 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 type $(i,r) \in E_l$ leaves the system upon its service completion at station i, if E_l is open.

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 demand distribution which is a negative exponential with mean $\tau_i > 0$ if station i is a station of type 1.

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 α_i be any mapping $\mathbb{N} \to [0, +\infty)$ such that $\alpha_i(0) = 0$ with the interpretation that $\alpha_i(j)$ is the service rate at station i when there are $j > 0$ customers at this station. In particular, if $\alpha_i(j) = \min(c_i, j)$ then station i has c_i servers in parallel working at unit speed. For

stations of type 2, 3 or 4, each class of customers may have a distinct and arbitrary service demand distribution (GI-servers). Let $\tau_{ir} > 0$ denote the mean service demand of a customer of type (i, r) for $i \in \{\text{PS, IS, LCFS}\}$.

Let $\mathcal{S}$ be the set of feasible states for the network under consideration. Let X_{ir} denote the number of customers of class r at station i. The state of the network is $\mathbf{S} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N)$ where $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iR}) \in \mathbb{N}^{\mathrm{R}}$ for $i = 1, 2, \ldots, N$. Let $\mathbf{n}_i \in (n_{i1}, n_{i2}, \ldots, n_{iR}) \in \mathbb{N}^{\mathrm{R}}$. For any (possible random) vector $\mathbf{a} = (a_1, a_2, \ldots, a_R)$ in $\mathbb{N}^{\mathrm{R}}$ the notation $|\mathbf{a}| = \sum_{r=1}^{R} a_r$ will be used.

The joint equilibrium distribution of queue sizes in the network is (cf. [6]):

$$\mathrm{P}\left(\mathbf{S} = (\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_N)\right) = C\, d(\mathbf{S})\, g_1(\mathbf{n}_1)\, g_2(\mathbf{n}_2) \cdots g_N(\mathbf{n}_N), \tag{2.2}$$

for all $\mathbf{S} \in \mathcal{S}$, where C is a normalizing constant and $d(\mathbf{S})$ is a function of the number of customers in the system. More precisely, if the network is closed then $d(\cdot) \equiv 1$, otherwise

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

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

Each $g_i(\mathbf{n}_i)$ in (2.2) is a function that depends on the type of station i:

- if $i \in \text{FCFS}$, then

$$g_i(\mathbf{n}_i) = |\mathbf{n}_i|! \left(\prod_{j=1}^{|\mathbf{n}_i|} \frac{1}{\alpha_i(j)} \right) \left(\prod_{r=1}^{R} \frac{\rho_{ir}^{n_{ir}}}{n_{ir}!} \right); \tag{2.4}$$

- if $i \in \{\text{PS, LCFS}\}$, then

$$g_i(\mathbf{n}_i) = |\mathbf{n}_i|! \prod_{r=1}^{R} \frac{\rho_{ir}^{n_{ir}}}{n_{ir}!}; \tag{2.5}$$

- if $i \in \text{IS}$, then

$$g_i(\mathbf{n}_i) = \prod_{r=1}^{R} \frac{\rho_{ir}^{n_{ir}}}{n_{ir}!}, \tag{2.6}$$

where $\rho_{ir} = \tau_i\, \theta_{ir}$ if $i \in \text{FCFS}$ and $\rho_{ir} = \tau_{ir}\, \theta_{ir}$ for $i \in \{\text{PS, IS, LCFS}\}$. The θ_{ir}'s satisfy the set of linear equations (cf. [6])

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

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

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

The parameter ρ_{ir} is usually referred to as the traffic intensity at station i due to customers of class r, whereas in the case of closed networks θ_{ir} is called the visit ratio of customers of class r to station i (see e.g., [18, 50]).

3 Sensitivity Results

From now on we assume that the network is in equilibrium.

Two types of first order sensitivity results are derived in this section: results with respect to workload intensities (mean service demands and service rates in Theorem 1, visit ratios and traffic intensities in Theorem 3) and results with respect to exogeneous arrival rates (Theorem 2). We conclude this section by showing through an example that these results also enable us to infer higher order properties of the system performance measures (Remark 4).

Let us first introduce some notation. Unless otherwise mentioned (see (3.1) in Theorem 1), we assume that $\partial x/\partial y = 0$ for all $x, y \in \mathcal{A}$, $x \neq y$, where $\mathcal{A} := \{\lambda_{ir}, \tau_i, \tau_{ir}, \alpha_i(j), \gamma_l(j)\}_{i,j,r,l}$, with the exception that $\partial\lambda_{ir}/\partial\lambda_{js}$ may be non zero for $(i,r) \neq (j,s)$ (this assumption is needed to cover the modeling of the exogeneous arrivals in the standard BCMP formulation, since in that case $\sum_{(i,r)\in E_l} \lambda_{ir} = 1$ for all $l \in L_{\mathcal{O}}$, see [6]). For sake of simplicity, we also assume that the elements of the routing matrix $\mathbf{P}$ do not depend on x for all $x \in \mathcal{A}$ (i.e., $\partial p(i,r;j;s)/\partial x = 0$).

Sensitivity results will be obtained with respect to any model parameter lying in the set $\mathcal{X} := (\tau_i, \tau_{ir}, \rho_{ir}, \lambda_{ir}, \theta_{ir}, \alpha_i(j), \alpha_i(j)/\tau_i, \gamma_l(j))_{i,j,r,l}$.

Let Φ be a mapping $\mathcal{S} \times \mathcal{X} \to \mathbb{R}$. We say that Φ satisfies assumption **A1** if

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

2. $$\frac{\partial \mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right]}{\partial x} = \sum_{\mathbf{n}\in\mathcal{S}} \frac{\partial\{\Phi(\mathbf{n}, \mathbf{x})\mathrm{P}\left(\mathbf{S} = \mathbf{n}\right)\}}{\partial x},$$

for all $x \in \{\tau_i\}_i$, $\mathbf{n} \in \mathcal{S}$, $\mathbf{x} \in \mathcal{X}$. Similarly, we say that Φ satisfies assumption **A2** (resp. **A3**,

A4, **A5**, **A6**, **A7**, **A8**) if conditions 1 and 2 hold simultaneously for all $x \in \{\alpha_i(j)\}_{i,j}$ (resp. $x \in \{\alpha_i(j)/\tau_i\}_{i,j}$, $x \in \{\tau_{ir}\}_{i,r}$, $x \in \{\lambda_{ir}\}_{i,r}$, $x \in \{\gamma_l(j)\}_{l,j}$, $x \in \{\theta_{ir}\}_{i,r}$, $x \in \{\rho_{ir}\}_{i,r}$).

It is easily seen that conditions 1 and 2 are satisfied when $\Phi(\mathbf{n}, \mathbf{x})$ is continuous as a function of x, and when it is bounded by a polynomial in $\mathbf{n}$. In particular, condition 2 is satisfied for closed networks.

We now establish sensitivity results with respect to the mean service demands (resp. service rates) in the case of an arbitrary network (i.e., open, closed or mixed).

Theorem 1 *If Φ satisfies assumption* **A1***, then for $i \in$ FCFS (resp. for $i \in$ {PS, IS, LCFS} such that $\tau_{ir} = \tau_i$ for $r = 1, 2, \ldots, R$),*

$$\frac{\partial \mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right]}{\partial \tau_i} = \mathrm{E}\left[\frac{\partial\, \Phi(\mathbf{S}, \mathbf{x})}{\partial \tau_i}\right] + \frac{\mathrm{Cov}\left(\Phi(\mathbf{S}, \mathbf{x}), |\mathbf{X}_i|\right)}{\tau_i}. \tag{3.1}$$

If Φ satisfies assumption **A2** *(resp.* **A3***), then for $i \in$ FCFS, $j \geq 1$,*

$$\frac{\partial \mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right]}{\partial x} = \mathrm{E}\left[\frac{\partial\, \Phi(\mathbf{S}, \mathbf{x})}{\partial x}\right] - \frac{\mathrm{Cov}\left(\Phi(\mathbf{S}, \mathbf{x}), \mathbf{1}(|\mathbf{X}_i| \geq j)\right)}{x}, \tag{3.2}$$

for $x \in \{\alpha_i(j)\}_{i,j}$ (resp. $x \in \{\alpha_i(j)/\tau_i\}_{i,j}$).

If Φ satisfies assumption **A4***, then for $i \in$ {PS, IS, LCFS}, $r = 1, 2, \ldots, R$,*

$$\frac{\partial \mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right]}{\partial \tau_{ir}} = \mathrm{E}\left[\frac{\partial\, \Phi(\mathbf{S}, \mathbf{x})}{\partial \tau_{ir}}\right] + \frac{\mathrm{Cov}\left(\Phi(\mathbf{S}, \mathbf{x}), X_{ir}\right)}{\tau_{ir}}. \tag{3.3}$$

Proof. We first prove (3.1). It follows from (2.2) and assumption **A1** that

$$\begin{aligned}
\frac{\partial \mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right]}{\partial \tau_i} &= \sum_{\mathbf{n} \in \mathcal{S}} \frac{\partial}{\partial \tau_i} \left\{ C\, d(\mathbf{n})\, \Phi(\mathbf{n}, \mathbf{x}) \prod_{k=1}^{N} g_k(\mathbf{n}_k) \right\}, \\
&= C \sum_{\mathbf{n} \in \mathcal{S}} d(\mathbf{n}) \left(\frac{\partial \Phi(\mathbf{n}, \mathbf{x})}{\partial \tau_i} \right) \prod_{k=1}^{N} g_k(\mathbf{n}_k) + \left(\frac{\partial C}{\partial \tau_i} \right) \sum_{\mathbf{n} \in \mathcal{S}} d(\mathbf{n})\, \Phi(\mathbf{n}, \mathbf{x}) \prod_{k=1}^{N} g_k(\mathbf{n}_k) \\
&+ C \sum_{\mathbf{n} \in \mathcal{S}} d(\mathbf{n})\, \Phi(\mathbf{n}, \mathbf{x}) \left(\frac{\partial}{\partial \tau_i} \prod_{k=1}^{N} g_k(\mathbf{n}_k) \right).
\end{aligned} \tag{3.4}$$

One can easily check from (2.4)-(2.6) that

$$\frac{\partial}{\partial \tau_i} \prod_{k=1}^{N} g_k(\mathbf{n}_k) = \frac{|\mathbf{n}_i|}{\tau_i} \prod_{k=1}^{N} g_k(\mathbf{n}_k). \tag{3.5}$$

Now differentiating the identity

$$C = \frac{1}{\sum_{\mathbf{n}\in\mathcal{S}} d(\mathbf{n}) \prod_{k=1}^{N} g_k(\mathbf{n}_k)}, \tag{3.6}$$

we obtain

$$\frac{\partial C}{\partial \tau_i} = -C^2 \left(\frac{\partial}{\partial \tau_i} \sum_{\mathbf{n}\in\mathcal{S}} d(\mathbf{n}) \prod_{k=1}^{N} g_k(\mathbf{n}_k) \right),$$

which, together with (3.5), yields

$$\frac{\partial C}{\partial \tau_i} = -\frac{C}{\tau_i} \mathrm{E}\left[|\mathbf{X}_i|\right]. \tag{3.7}$$

Consequently, cf. (3.4), (3.5), and (3.7),

$$\begin{aligned} \frac{\partial \mathrm{E}\left[\Phi(\mathbf{S},\mathbf{x})\right]}{\partial \tau_i} &= \mathrm{E}\left[\frac{\partial \Phi(\mathbf{S},\mathbf{x})}{\partial \tau_i}\right] - \frac{\mathrm{E}\left[|\mathbf{X}_i|\right]\mathrm{E}\left[\Phi(\mathbf{S},\mathbf{x})\right]}{\tau_i} + \frac{\mathrm{E}\left[\Phi(\mathbf{S},\mathbf{x})\,|\mathbf{X}_i|\right]}{\tau_i}, \\ &= \mathrm{E}\left[\frac{\partial \Phi(\mathbf{S},\mathbf{x})}{\partial \tau_i}\right] + \frac{\mathrm{Cov}\left(\Phi(\mathbf{S},\mathbf{x}), |\mathbf{X}_i|\right)}{\tau_i}. \end{aligned}$$

The proofs of (3.2) and (3.3) follow similarly and are therefore omitted. ∎

Remark 1 A relation similar to (3.7) was first derived by Kobayashi and Gerla [36] in a less general context.

Remark 2 Theorem 3.1 of Strelen [59] is a simple corollary of relation (3.1). Set $\Phi(\mathbf{n},\mathbf{x}) = n_{i_1}^{j_1-1} n_{i_2}^{j_2} \cdots n_{i_k}^{j_k}$, with $k \geq 1$, $1 \leq l \leq k$, $j_l \geq 1$, $1 \leq i_l \leq N$. Then, using (3.1), we see that

$$\begin{aligned} &\mathrm{E}\left[|X_{i_1}|^{j_1}|X_{i_2}|^{j_2}\cdots|X_{i_k}|^{j_k}\right] = \\ &\quad \mathrm{E}\left[|X_{i_1}|\right]\mathrm{E}\left[|X_{i_1}|^{j_1-1}|X_{i_2}|^{j_2}\cdots|X_{i_k}|^{j_k}\right] + \tau_{i_1}\frac{\partial \mathrm{E}\left[|X_{i_1}|^{j_1-1}|X_{i_2}|^{j_2}\cdots|X_{i_k}|^{j_k}\right]}{\partial \tau_{i_1}}. \end{aligned}$$

The following theorem addresses sensitivity results with respect to exogeneous arrival rates when the network is either open or mixed.

Theorem 2 *Assume that Φ satisfies assumption* **A5**. *If there exists an external source of customers of type $(i,r) \in E_l$ (i.e., $\lambda_{ir} > 0$), then*

$$\frac{\partial \mathrm{E}\left[\Phi(\mathbf{S},\mathbf{x})\right]}{\partial \lambda_{ir}} = \mathrm{E}\left[\frac{\partial\, \Phi(\mathbf{S},\mathbf{x})}{\partial \lambda_{ir}}\right] + \mathrm{Cov}\left(\Phi(\mathbf{S},\mathbf{x}), \sum_{l'\in L_{\mathcal{O}}} \frac{M_{l'}(\mathbf{S})}{\Lambda_{l'}} \sum_{(j,s)\in E_{l'}} \frac{\partial \lambda_{js}}{\partial \lambda_{ir}}\right)$$

$$+ \quad \mathrm{Cov}\left(\Phi(\mathbf{S},\mathbf{x}), \sum_{l'\in L_{\mathcal{O}}} \sum_{(j,s)\in E_{l'}} \left(\frac{X_{js}}{\theta_{js}}\right)\left(\frac{\partial\theta_{js}}{\partial\lambda_{ir}}\right)\right). \tag{3.8}$$

In particular, if $E_l = \{(i,r)\}$ *and if* $\partial\lambda_{js}/\partial\lambda_{j's'} = 0$ *for all* j, j', s, s' *such that* $(j,s) \neq (j',s')$, *then*

$$\frac{\partial \mathrm{E}\left[\Phi(\mathbf{S},\mathbf{x})\right]}{\partial\lambda_{ir}} = \mathrm{E}\left[\frac{\partial\,\Phi(\mathbf{S},\mathbf{x})}{\partial\lambda_{ir}}\right] + \frac{\mathrm{Cov}\left(\Phi(\mathbf{S},\mathbf{x}), M_l(\mathbf{S})\right)}{\Lambda_l}. \tag{3.9}$$

Assume that Φ *satisfies assumption* **A6**. *If* E_l *is open, then for* $j \geq 0$,

$$\frac{\partial \mathrm{E}\left[\Phi(\mathbf{S},\mathbf{x})\right]}{\partial\gamma_l(j)} = \mathrm{E}\left[\frac{\partial\,\Phi(\mathbf{S},\mathbf{x})}{\partial\gamma_l(j)}\right] + \frac{\mathrm{Cov}\left(\Phi(\mathbf{S},\mathbf{x}), \mathbf{1}(M_l(\mathbf{S}) \geq j)\right)}{\gamma_l(j)}. \tag{3.10}$$

Proof. Assume that $(i,r) \in E_l$. We have, cf. (2.2) and assumption **A5**,

$$\begin{aligned}
\frac{\partial \mathrm{E}\left[\Phi(\mathbf{S},\mathbf{x})\right]}{\partial\lambda_{ir}} &= C\sum_{\mathbf{n}\in\mathcal{S}}\left(\frac{\partial\Phi(\mathbf{n},\mathbf{x})}{\partial\lambda_{ir}}\right) d(\mathbf{n}) \prod_{k=1}^{N} g_k(\mathbf{n}_k) + \left(\frac{\partial C}{\partial\lambda_{ir}}\right)\sum_{\mathbf{n}\in\mathcal{S}}\Phi(\mathbf{n},\mathbf{x})\, d(\mathbf{n}) \prod_{k=1}^{N} g_k(\mathbf{n}_k) \\
&+ C\sum_{\mathbf{n}\in\mathcal{S}}\Phi(\mathbf{n},\mathbf{x})\left(\frac{\partial d(\mathbf{n})}{\partial\lambda_{ir}}\right)\prod_{k=1}^{N} g_k(\mathbf{n}_k) \\
&+ C\sum_{\mathbf{n}\in\mathcal{S}}\Phi(\mathbf{n},\mathbf{x})\, d(\mathbf{n})\left(\frac{\partial}{\partial\lambda_{ir}}\prod_{k=1}^{N} g_k(\mathbf{n}_k)\right).
\end{aligned} \tag{3.11}$$

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

$$\frac{\partial}{\partial\lambda_{ir}}\prod_{k=1}^{N} g_k(\mathbf{n}_k) = \prod_{k=1}^{N} g_k(\mathbf{n}_k) \sum_{l'\in L_{\mathcal{O}}}\sum_{(j,s)\in E_{l'}}\left(\frac{n_{js}}{\theta_{js}}\right)\left(\frac{\partial\theta_{js}}{\partial\lambda_{ir}}\right). \tag{3.12}$$

It follows from (2.3) that

$$\frac{\partial d(\mathbf{n})}{\partial\lambda_{ir}} = d(\mathbf{n}) \sum_{l'\in L_{\mathcal{O}}} \frac{M_{l'}(\mathbf{n})}{\Lambda_{l'}} \sum_{(j,s)\in E_{l'}} \frac{\partial\lambda_{js}}{\partial\lambda_{ir}}. \tag{3.13}$$

By differentiating (3.6) with respect to λ_{ir} and by using (3.12) and (3.13), we obtain

$$\frac{\partial C}{\partial\lambda_{ir}} = -\left(\sum_{l'\in L_{\mathcal{O}}} \frac{M_{l'}(\mathbf{S})}{\Lambda_{l'}} \sum_{(j,s)\in E_{l'}} \frac{\partial\lambda_{js}}{\partial\lambda_{ir}} + \sum_{l'\in L_{\mathcal{O}}}\sum_{(j,s)\in E_{l'}}\left(\frac{\mathrm{E}\left[X_{js}\right]}{\theta_{js}}\right)\left(\frac{\partial\theta_{js}}{\partial\lambda_{ir}}\right)\right) C. \tag{3.14}$$

Substituting (3.12), (3.13), and (3.14) into (3.11) yields (3.8). The same line of arguments gives us (3.10).

Finally, (3.9) is readily derived from (3.8) by observing that

$$\frac{\partial \theta_{ir}}{\partial \lambda_{ir}} = 0,$$

whenever $E_l = \{(i,r)\}$. Indeed, in this case $q_{ir}(l) = 1$ (see (2.8)), so that $\theta_{ir} = 1/(1 - p_{i,r;i,r})$ from (2.7), and therefore $\partial\theta_{ir}/\partial\lambda_{ir} = 0$. ∎

Remark 3 The partial derivatives $\{\partial\theta_{js}/\partial\lambda_{ir}\}_{i,r;j,s}$ involved in Theorem 2 are uniquely determined by the equations, cf. (2.7),

$$\frac{\partial \theta_{js}}{\partial \lambda_{ir}} = \frac{\partial q_{js}(l)}{\partial \lambda_{ir}} + \sum_{(j',s') \in E_l} p_{j',s';j,s} \frac{\partial \theta_{j's'}}{\partial \lambda_{ir}}, \qquad \forall (j,s) \in E_l,\ l \in L_{\mathcal{O}}, \tag{3.15}$$

where the first term in the right-hand side of (3.15) is obtained from (2.8).

Theorem 3 *Assume that* Φ *satisfies assumptions* **A7** *(resp.* **A8***).*

Then, for all $i \in \{$*FCFS, PS, IS, LCFS*$\}$, $r = 1, 2, \ldots, R$,

$$\frac{\partial \mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right]}{\partial x} = \mathrm{E}\left[\frac{\partial\, \Phi(\mathbf{S}, \mathbf{x})}{\partial x}\right] + \frac{\mathrm{Cov}\left(\Phi(\mathbf{S}, \mathbf{x}), X_{ir}\right)}{x}, \tag{3.16}$$

for $x = \theta_{ir}$ *(resp.* $x = \rho_{ir}$*).*

Proof. The proof is analogous to those of Theorems 1 and 2, and is therefore omitted. ∎

It is worthwhile noting that (3.16) analyses the effect of increasing θ_{ir} while keeping the θ_{js}'s unchanged for all $(j,s) \neq (i,r)$. Therefore, it is assumed that $\partial\theta_{js}/\partial\theta_{ir} = 0$ for all $(j,s) \neq (i,r)$. This can be achieved by modifying the routing matrix $\mathbf{P}$. Indeed, for any $h > 0$, it is always possible to find a routing matrix $[p^h_{j,s;j',s'}]$ such that

$$\theta^h_{js} = q_{js}(l) + \sum_{(j',s') \in E_l} \theta^h_{j',s'}\, p^h_{j',s';j,s},$$

with $\theta^h_{js} := \theta_{js}$ for $(j,s) \neq (i,r)$ and $\theta^h_{ir} := \theta_{ir} + h$, where $\{\theta_{js}\}_{(j,s) \in E_l}$ is a solution of (2.7) (choose, for instance, $p^h_{j,s;j',s'} = 1 - \theta_{ir}(1 - p_{i,r;i,r})/(\theta_{ir} + h)$ for $(j,s) = (j',s') = (i,r)$, $p^h_{j,s;j',s'} = \theta_{ir}\, p_{i,r;j',s'}/(\theta_{ir} + h)$ for $(j,s) = (i,r)$ and $(j',s') \neq (i,r)$, and $p^h_{j,s;j',s'} = p_{j,s;j',s'}$ for $(j,s) \neq (i,r)$). The same comment applies to the partial derivatives $\{\partial\rho_{js}/\partial\rho_{ir}\}_{(j,s) \neq (i,r)}$.

Remark 4 Higher order derivatives of the sample function Φ can also be obtained from Theorems 1-3. For instance, an iterative use of formula (3.1) yields

$$\frac{\partial^2 \mathrm{E}\left[\Phi(\mathbf{S},\mathbf{x})\right]}{\partial \tau_i^2} = \mathrm{E}\left[\frac{\partial^2 \Phi(\mathbf{S},\mathbf{x})}{\partial \tau_i^2}\right] + 2\,\frac{\mathrm{Cov}\left(\partial\Phi(\mathbf{S},\mathbf{x})/\partial\tau_i, |\mathbf{X}_i|\right)}{\tau_i} - (1+2\,\mathrm{E}\left[|X_i|\right])\,\frac{\mathrm{Cov}\left(\Phi(\mathbf{S},\mathbf{x}), |\mathbf{X}_i|\right)}{\tau_i^2} + \frac{\mathrm{Cov}\left(\Phi(\mathbf{S},\mathbf{x}), |\mathbf{X}_i|^2\right)}{\tau_i^2}, \quad (3.17)$$

for $i \in \mathrm{FCFS}$, provided that $\Phi(\mathbf{n},\mathbf{x})$ and $\partial\Phi(\mathbf{n},\mathbf{x})/\partial\tau_i$ both satisfy assumption **A1**.

Remark 5 Formula (3.16) generalizes Theorem 1 in de Souza e Silva and Muntz [19]. Indeed, letting $\Phi(\mathbf{n}) = n_{ir}$ we readily get from (3.3) that $\mathrm{Var}\left(X_{ir}\right) = x\,\partial \mathrm{E}\left[X_{ir}\right]/\partial x$, for $x = \theta_{ir}$ or $x = \rho_{ir}$.

Remark 6 As mentioned in [6], 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. However, Theorems 1-3 remain valid provided the introduced service dependencies do not involve the parameters in $\mathcal{A}$.

Remark 7 Similarly, the product form (2.2) is preserved under various forms of state-dependent arrival rates. In this case, Theorems 1-3 still hold if the introduced arrival dependencies do not involve the parameters in $\mathcal{A}$.

As a consequence of Remark 7, Theorem 1 is applicable to the product form queueing network with population size constraints as introduced by Lam [37]. Note, however, that Theorems 2 or 3 are not directly applicable in that case.

Remark 8 Formula (3.1) shows that the derivative and the expectation in $\partial \mathrm{E}\left[\Phi(\mathbf{S},\mathbf{x})\right]/\partial\tau_i$ can be interchanged if and only if $\mathrm{Cov}\left(\Phi(\mathbf{S},\mathbf{x}), |X_i|\right) = 0$. Similar results can also be derived from (3.2), (3.3), (3.8)-(3.10), (3.16). Such interchange properties are very useful in perturbation analysis (cf. [10, 11]).

4 Applications

Many results of practical interest can be derived from Theorems 1-3. We point out some of them below. Recall that any vector $\mathbf{n} \in \mathcal{S}$ can be written as $\mathbf{n} = (\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_N)$ with $\mathbf{n}_i = (n_{i1}, n_{i2}, \ldots, n_{iR})$, where $n_{ir} \in \mathbb{N}$. Recall also that $\mathbf{x} \in \mathcal{X}$ denotes the vector of system parameters (see Section 3).

4.1 Sensitivity of queue lengths

Let f be any nondecreasing mapping $\mathbb{N} \to \mathbb{R}$.

1. Let $\Phi(\mathbf{n}, \mathbf{x}) = f\left(\sum_{k \in N_{\mathcal{I}}} |\mathbf{n}_k|\right)$, $N_{\mathcal{I}} \subset \{1, 2, \ldots, N\}$.

   Fix $i \in$ FCFS. We assume that Φ satisfies assumptions **A1** **A2** and **A3** (this holds, in particular, if f is bounded by a polynomial). Then, cf. (3.1), (3.2),

   $$\frac{\partial \mathrm{E}\left[f\left(\sum_{k \in N_{\mathcal{I}}} |\mathbf{X}_k|\right)\right]}{\partial \tau_i} = \mathrm{Cov}\left(f\left(\sum_{k \in N_{\mathcal{I}}} |\mathbf{X}_k|\right), |\mathbf{X}_i|\right) / \tau_i; \tag{4.1}$$

   $$\frac{\partial \mathrm{E}\left[f\left(\sum_{k \in N_{\mathcal{I}}} |\mathbf{X}_k|\right)\right]}{\partial x} = -\mathrm{Cov}\left(f\left(\sum_{k \in N_{\mathcal{I}}} |\mathbf{X}_k|\right), \mathbf{1}(|\mathbf{X}_i| \geq j)\right) / x, \tag{4.2}$$

   for all $x \in \{\alpha_i(j),\, \alpha_i(j)/\tau_i;\, j \geq 1\}$.

   Assume that $N_{\mathcal{I}} = \{i\}$. We know from Barlow and Proschan ([5], Theorem 4.7, p. 146) that $\mathrm{Cov}\,(g(X), h(X)) \geq 0$ for any random variable X and for any nondecreasing mappings g and h. Applying this result to the right-hand side of (4.1) (resp. (4.2)) yields:

   - $|\mathbf{X}_i|$ is increasing in τ_i (resp. decreasing in $\alpha_i(j)$, $\alpha_i(j)/\tau_i$, $j = 1, 2, \ldots$) in the sense of stochastic ordering ([52], pp. 251-252);
   - $\sum_{k \neq i} |\mathbf{X}_k|$ is decreasing in τ_i (resp. increasing in $\alpha_i(j)$, $\alpha_i(j)/\tau_i$, $j = 1, 2, \ldots$) in the sense of stochastic ordering when the network is closed.

   The above results extend earlier results of Shanthikumar and Yao (see [55], Corollary 3.1) to *multiclass* closed/*open*/*mixed* BCMP networks (note however that our results are only established for the stochastic ordering, whereas the results in [55] hold for the (stronger) likelihood ratio ordering).

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

   Assume that Φ satisfies assumption **A4**. Then, cf. (3.3),

   $$\frac{\partial \mathrm{E}\left[f(X_{js})\right]}{\partial \tau_{ir}} = \mathrm{Cov}\left(f(X_{js}), X_{ir}\right) / \tau_{ir}, \tag{4.3}$$

   for $i \in \{\text{PS, IS, LCFS}\}$,

   Assume that $E_l = \{(i, r)\}$. Then, similar to case 1 above, we deduce from (4.3) that

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

Assume now that E_l is open and that $E_l = \{(j,s)\}$. Then, cf. (3.9),

$$\frac{\partial \mathrm{E}\left[f(X_{js})\right]}{\partial \lambda_{js}} = \mathrm{Cov}\left(f(X_{js}), X_{js}\right)/\lambda_{js}, \tag{4.4}$$

provided that f does not depend on λ_{js}, which shows that

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

3. Let $\Phi(\mathbf{n}, \mathbf{x}) = |\mathbf{n}_i|$

   Then, for $i \in$ FCFS, cf. (3.17),

$$\frac{\partial^2 \mathrm{E}\left[|X_i|\right]}{\partial \tau_i^2} = \frac{\mathrm{Cov}\left(|X_i|, |X_i|^2\right)}{\tau_i^2} - \left(1 + 2\,\mathrm{E}\left[|X_i|\right]\right)\frac{\mathrm{Var}\left(|X_i|\right)}{\tau_i^2}. \tag{4.5}$$

In particular, formulas can also be derived for $\partial \mathrm{E}\left[f(X_{js})\right]/\,\partial \tau_i$ for $i \in$ FCFS and for $\partial \mathrm{E}\left[f(|\mathbf{X}_j|)\right]/\,\partial \tau_{ir}$ for $i \in \{$PS, IS, LCFS$\}$.

4.2 Sensitivity of throughputs

1. Let $\Phi(\mathbf{n}, \mathbf{x}) = \dfrac{n_{js}}{\tau_{js}\, n_j}\, \mathbf{1}_{\{n_{js}>0\}}$.

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

   For $i \in$ PS, we have, cf. (3.3),

$$\frac{\partial \mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right]}{\partial \tau_{ir}} = -\mathbf{1}_{\{(i,r)=(j,s)\}}\frac{\mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right]}{\tau_{ir}} + \frac{1}{\tau_{ir}\,\tau_{js}}\,\mathrm{Cov}\left(\frac{X_{js}}{|X_j|}\,\mathbf{1}_{\{X_{js}>0\}}; X_{ir}\right). \tag{4.6}$$

   If only customers of class r may visit station i (i.e., $|X_i| = X_{ir}$) and if $(j,s) = (i,r)$, then (4.6) reduces to

$$\tau_{ir}^2\,\frac{\partial \mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right]}{\partial \tau_{ir}} = -\mathrm{P}\left(X_{ir} > 0\right) + \mathrm{Cov}\left(\mathbf{1}(X_{ir} > 0; X_{ir}\right).$$

2. Let $\Phi(\mathbf{n}, \mathbf{x}) = n_{js}/\tau_{js}$.

   Then $\mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right] = \mathrm{E}\left[X_{js}\right]/\tau_{js}$ is the throughput of customers of class s in station j if $j \in \mathrm{IS}$.

   We have, for $i \in \mathrm{IS}$, cf. (3.3),

$$\frac{\partial \mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right]}{\partial \tau_{ir}} = -\mathbf{1}_{\{(i,r)=(j,s)\}} \frac{\mathrm{E}\left[X_{js}\right]}{\tau_{js}^2} + \frac{\mathrm{Cov}\left(X_{js}, X_{ir}\right)}{\tau_{ir}\tau_{js}}. \tag{4.7}$$

   If $(i, r) = (j, s)$, then (4.7) reduces to

$$\tau_{ir}^2 \frac{\partial \mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right]}{\partial \tau_{ir}} = \mathrm{Var}\left(X_{ir}\right) - \mathrm{E}\left[X_{ir}\right].$$

3. Let $\Phi(\mathbf{n}, \mathbf{x}) = \alpha_j(|\mathbf{n}_j|)/\tau_j$.

   Recall that $\alpha_j(0) = 0$. Then $\mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right] = \mathrm{E}\left[\alpha_j(|\mathbf{X}_j|)\right]/\tau_j$, and $\mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right]$ is the throughput of station j if $j \in \mathrm{FCFS}$.

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

$$\frac{\partial \mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right]}{\partial \tau_i} = -\mathbf{1}_{\{i=j\}} \frac{\mathrm{E}\left[\alpha_j(|\mathbf{X}_j|)\right]}{\tau_i^2} + \frac{\mathrm{Cov}\left(\alpha_j(|\mathbf{X}_j|), |\mathbf{X}_i|\right)}{\tau_i\tau_j}. \tag{4.8}$$

   If $i = j$, then (4.8) simply becomes

$$\tau_i^2 \frac{\partial \mathrm{E}\left[\Phi(\mathbf{S}, \mathbf{x})\right]}{\partial \tau_i} = \mathrm{Cov}\left(\alpha_i(|\mathbf{X}_i|), |\mathbf{X}_i|\right) - \mathrm{E}\left[\alpha_i(|\mathbf{X}_i|)\right]. \tag{4.9}$$

   Corollary 3.1 in Shanthikumar and Yao [55] implies that the left-hand side of (4.8) is nonpositive, for any single class closed Jackson network. We have not been able to derive this result by showing that the right-hand side of (4.8) is nonpositive.

A numerical illustration of the results obtained in this section is now given. The computations have been performed using the queueing modeling software QNAP 2 [46]. We consider the mixed queueing network depicted in Figure 1. There are two subchains: customers of class 1 belong to the open chain and (ten) customers of class 2 belong to the closed chain. All stations are FCFS stations but station 3 that is an IS station with class-dependent (exponential) service demands. Further, station 2 has nonmonotone state-dependent service rates with $\alpha_2(1) = 1$, $\alpha_2(2) = 2$, $\alpha_2(3) = 7$, and $\alpha_2(j) = 1$ for $j \geq 4$. The nominal mean service demands are $\tau_1 = 0.5$, $\tau_2 = 1.0$, $\tau_{31} = 1.5$, $\tau_3 := \tau_{32} = 2.4$, $\tau_4 = 1.1$ and $\tau_5 = 1.7$.

Let T_i (resp. T_{ir}) be the throughput of station i (resp. the throughput of customers of class r at station i). Figure 2 displays the mapping $\tau_2 \to \partial T_2/\partial \tau_2$ (resp. $\tau_{32} \to \partial T_{32}/\partial \tau_{32}$, $\tau_4 \to \partial T_4/\partial \tau_4$)

for τ_2 (resp. τ_{32}, τ_4) lying in $[0.5, 10]$, when the other mean service demands are kept fixed to their nominal values. We may observe from Figure 2 that (i) the throughput of station 2 is monotone (nonincreasing) and nonconvex as a function τ_2, (ii) the throughput of customers of class 2 at station 3 is monotone (nonincreasing) and convex as a function of τ_{32}, (iii) the throughput of station 4 is neither monotone nor convex as a function of τ_4.

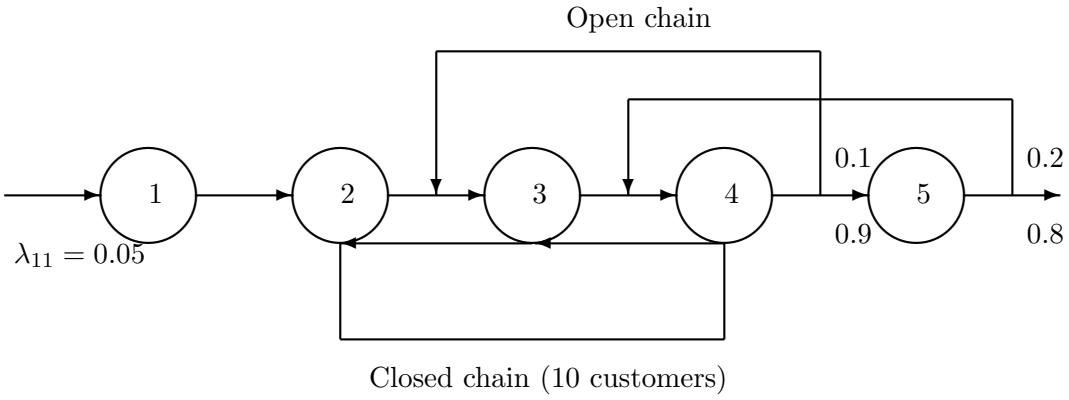

Figure 1: Mixed BCMP queueing network.

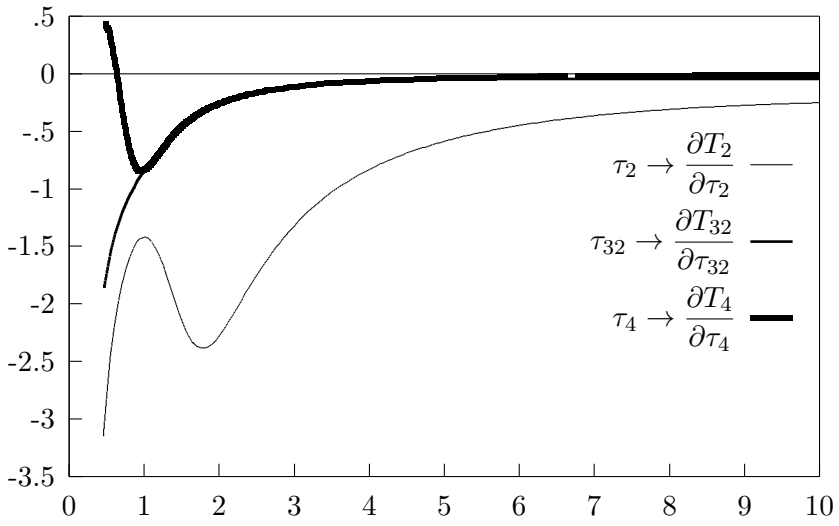

Figure 2: Throughput derivatives vs mean service demands.

4.3 Parameter 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 many cases, 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 problems 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 not feasible because the derivative of the cost function is hard to obtain.

In product form queueing networks, however, owing to Theorems 1-3, the determination of the derivative is reduced to computing the covariance between some state variables. This provides a new approach, at least from a numerical point of view, to the solution of some optimization problems.

As an easy example, let the cost function be $\mathrm{E}\left[a\left|\mathbf{X}_i\right| + b/\tau_i\right]$, where a and b are nonnegative real coefficients representing holding and service costs, respectively. Recall that $|\mathbf{X}_i|$ denotes the queue length of an FCFS station i in a BCMP network and that τ_i is the mean service demand of the customers visiting that station. One might want to find a value of τ_i that minimizes this cost function. Using Theorem 1, one immediately obtains

$$\frac{\partial \mathrm{E}\left[a|\mathbf{X}_i| + b/\tau_i\right]}{\partial \tau_i} = -\frac{b}{\tau_i^2} + \frac{a}{\tau_i}\,\mathrm{Var}\left(|\mathbf{X}_i|\right). \tag{4.10}$$

The problem is then reduced to the computation of the roots of the right-hand side of equation (4.10). Further applications to parameter optimizations can be found in [18, 36, 44, 45, 66, 67, 73].

4.4 Correlation between state variables

Let us choose $\Phi(\mathbf{n})$ as in application 1 of Section 4.1. Assume that the network is closed and that there is only one class of customers. Then, for $j \geq 1$,

$$\mathrm{Cov}\left(f(|\mathbf{X}_k|), |\mathbf{X}_i|\right) \;\leq\; 0; \tag{4.11}$$

$$\mathrm{Cov}\left(f(|\mathbf{X}_k|), \mathbf{1}(|\mathbf{X}_i| \geq j)\right) \;\geq\; 0, \tag{4.12}$$

for all $i \neq k$, and further

$$\mathrm{Cov}\left(f\left(\sum_{k \in N_{\mathcal{I}}} |\mathbf{X}_k|\right), |\mathbf{X}_i|\right) \;\geq\; 0; \tag{4.13}$$

$$\text{Cov}\left(f\left(\sum_{k\in N_{\mathcal{I}}}|\mathbf{X}_k|\right),\mathbf{1}(|\mathbf{X}_i|\geq j)\right) \quad \leq \quad 0, \tag{4.14}$$

for all $i \in$ FCFS. Formulas (4.12) and (4.14) hold under the additional assumptions that the mappings $\{\alpha_i,\, i \in N_{\mathcal{I}}\}$ are nondecreasing when the set $N_{\mathcal{I}}$ contains at least two elements.

To prove (4.11)-(4.14), let us recall the following result due to Shanthikumar and Yao (use Corollary 3.1 in [55] with $\mu_i(j) := \alpha_i(j)/\tau_i$). In a closed network with a single class of customers, let $(N_{\mathcal{A}}, N_{\mathcal{B}})$ denote a nontrivial partition of $\{1, 2, \ldots, N\}$. Then,

1. $|\mathbf{X}_j|$ $(j \in N_{\mathcal{A}})$ is decreasing in $\mu_i(j)$ in the sense of likelihood ratio ordering for all $i \in N_{\mathcal{B}}$;
2. $\sum_{j\in N_{\mathcal{B}}}|\mathbf{X}_j|$ is increasing in $\mu_i(j)$ in the sense of likelihood ratio ordering for all $i \in N_{\mathcal{B}}$,

provided (i) the service rate at station i is nondecreasing in $|\mathbf{X}_i|$ for all $i \in N_{\mathcal{A}}$, (ii) $|N_{\mathcal{B}}| = 1$ or $|N_{\mathcal{B}}| \geq 2$ and the service rate at station i is nondecreasing in $|\mathbf{X}_i|$ for all $i \in N_{\mathcal{B}}$ ($|\mathcal{G}|$ denotes the cardinality of any set $\mathcal{G}$).

The proof of (4.11)-(4.14) now follows from the above results and formulas (4.1), (4.2), and from the property that "likelihood ratio ordering $\Rightarrow$ stochastic ordering" (see Ross [52], Chap. 8).

Note that the negative correlation (4.11) implies the negative dependence (introduced by Block et al. [7]) between the numbers of customers in FCFS stations and that the positive correlation (4.12) has been obtained by Ott and Shanthikumar [43] for queueing networks with identical visit ratios and service rates.

5 Implementation issues and computational algorithms

As already mentioned in the introduction, there exists several efficient computational algorithms for computing the primary performance measures of BCMP-like networks [8], [13], [16], [39], [41], [50], [49], [59]. We show below that some of these algorithms can be used to compute the sensitivity results obtained in this paper.

Two cases need to be distinguished. In the first case, the sample function $\Phi(\mathbf{n}, \mathbf{x})$ is a polynomial with respect to the state variables n_{ir}, $1 \leq i \leq N$, $1 \leq r \leq R$, for any vector of system parameters $\mathbf{x} \in \mathcal{X}$ (e.g., $\Phi(\mathbf{n}, \mathbf{x}) = |\mathbf{n}_i|$).

Then, it is seen from Theorems 1-3 that the computation of the derivative of $\mathrm{E}\,[\Phi(\mathbf{S}, \mathbf{x})]$ with

respect to some system parameter reduces to the computation of covariances between the state variables X_{ir}, $1 \leq i \leq N$, $1 \leq r \leq R$, once the derivative of $\Phi(\mathbf{n}, \mathbf{x})$ is (formally) obtained. These covariances can be numerically computed using the approximate algorithm of Strelen [59].

In the second case, the function $\Phi(\mathbf{n}, \mathbf{x})$ is arbitrary (for optimization purposes we may have to deal for instance with non-polynomial cost functions of the state variables). In this case, a good candidate for the computation of our sensitivity results is the well-known *convolution algorithm* of Reiser and Kobayashi [49]. This algorithm initially aims to compute detailed performance measures of BCMP networks (e.g., marginal probabilities of the number of customers in the stations). However, since the convolution algorithm relies on the computation of a multidimensional generating function of a detailed state of the network, it can also provide covariances between state variables.

6 Concluding remarks

In this paper, various formulas have been established which relate the derivative of the expectation of any function of the state of the system with respect to any model parameter (i.e., arrival rate, mean service demand, service rate, visit ratio, traffic intensity) to covariances between state variables (Section 3). Applications of these results including monotonicity results have been given (Section 4). In particular, the results in Section 4.2 show that, in general, the throughputs in mixed/closed BCMP networks are not monotonic functions of the system parameters. We have also shown in Section 4.4 how monotonicity properties could be used to determine correlations between certain state variables.

It is worthwhile to note that the results in this paper provide an approach to the numerical computation of the derivative of the expectation of any function of the state of the system with respect to any model parameter, and is thus of particular interest for numerical solutions of various optimization problems arising in queueing systems. Efficient computational algorithms enabling such computations have been identified in Section 5.

Although the general relations in Theorems 1-3 have been obtained within the context of BCMP networks, similar results can be derived for other queueing networks such as, for instance, networks with blocking phenomena and/or state dependent routing [3, 26, 65, 71], networks with failures [15], provided the product form is preserved.

Acknowledgment

The authors are grateful to the referees for their useful comments on the first version of this paper.

References

[1] I. Adan and J. Van der Wal, "Monotonicity of the throughput of a closed queueing network in the number of jobs." *Oper. Res.* **37**, 6 (1989), 953-957.

[2] I. F. Akyildiz and H. von Brand, "Computational algorithms for networks of queues with rejection blocking." *Acta Informatica* **26** (1989), 559-576.

[3] I. F. Akyildiz and H. von Brand, "Exact solution for open, closed and mixed queueing networks with rejection blocking." *Theoretical Comp. Sci.* **64** (1989), 203-219.

[4] F. Baccelli and P. Brémaud, "Virtual customers in sensitivity and light traffic analysis via Campbell's formula for point processes." INRIA Report No. 1294 (Oct. 1990). To appear in *J. Appl. Prob.*

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

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

[7] H. W. Block, T. H. Savits and M. Shaked, "A concept of negative dependence using stochastic ordering." *Statist. Prob. Let.* **3** (1985), 81-86.

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

[9] X. R. Cao, "Realization probability in closed Jackson queueing networks and its application." *Adv. Appl. Prob.* **19** (1987) 708-738.

[10] X. R. Cao, "A sample performance function of closed Jackson queueing networks." *Oper. Res.* **36**, 1 (Jan. 1988) 128-136.

[11] X. R. Cao, "The convergence property of sample derivatives in closed Jackson queueing networks." *Stoch. Proc. and their Applications* **33** (1989) 105-122.

[12] K. M. Chandy and A. J. Martin, "A characterization of product-form queueing networks." *J. ACM* **34** (1983) 286-299.

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

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

[15] A. E. Conway, "Product-form and insensitivity in circuit-switched networks with failing links." *Perf. Eval.* **9**, 3 (Jun. 1989), 209-215.

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

[17] A. E. Conway, E. de Souza e Silva and S. S. Lavenberg, "Mean Value Analysis by chain of product form queueing networks." *IEEE Trans. Comp.* **38**, 3 (Mar. 1989), 432-441.

[18] E. de Souza e Silva and M. Gerla, "Load balancing in distributed systems with multiple classes and site constraints." *Proc. Performance '84*, E. Gelenbe (Ed.) (1984), 17-33.

[19] E. de Souza e Silva and R. R. Muntz, "Simple relationships among moments of queue lengths in product form queueing networks." *IEEE Trans. Comput.* **37**, 8 (Sep. 1988), 1125-1129.

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

[21] A. K. Erlang, 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.

[22] P. Glynn, "Likelihood ratio gradient estimation: an overview." *Proc. of the 1987 Winter Simulation Conference*, 366-375, 1987.

[23] P. Glynn, "A GSMP formalism for discrete event systems." *IEEE Proc.* Special Issue on Dynamics Discrete Events Systems **77**, 1 (Jan. 89), 14-23.

[24] P. Glynn and L. Sanders, "Monte Carlo optimization in manufacturing systems: two new approaches." *Proc. of ASME Computer in Engineering Conf.*, Chicago, IL, 1986.

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

[26] A. Hordijk and N. M. van Dijk, "Networks of queues." *Lectures Notes in Control and Information Sciences* **60** (Springer Verlag, 1983), 158-205.

[27] Y. C. Ho and X. R. Cao, "Perturbation analysis and optimization of queueing networks." *J. Opt. Theory and Appl.* **40**, 4 (1983), 559-582.

[28] Y. C. Ho, X. R. Cao and C. Cassandras, "Infinitesimal and finite perturbation analysis for queueing networks." *Automatica* **19** (1983), 439-445.

[29] Y. C. Ho and S. Li, "Extensions of infinitesimal perturbation." *IEEE Trans. Aut. Cont.* AC-33, 5 (1988), 427-438.

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

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

[32] S. Jordan and P. Varaiya, "Throughput in multiple service, multiple resource communication networks." *Proc. of the IEEE Conf. on Decision and Control*, Tampa, FL (Dec. 1989), 236-241.

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

[34] J. R. Kenevan and A. K. von Mayrhauser, "Convexity and concavity properties of analytic queueing models for computer systems." *Proc. Performance '84*, E. Gelenbe (Ed.) (1984), 361-375.

[35] L. Kleinrock, *Queueing Systems*, Vol. I. Wiley-Interscience, New York, 1975.

[36] H. Kobayashi and M. Gerla, "Optimal routing in closed queueing networks." *ACM Trans. Comp. Syst.* **1** (1983), 294-310.

[37] S. S. Lam, "Queueing networks with population size constraints." *IBM J. Res. Dev.* **21** (1977), 370-378.

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

[39] J. McKenna and D. Mitra, "Asymptotic expansions and integral representations of queue lengths in closed Markovian networks." *J. ACM* **31**, 2 (Apr. 1984), 346-360.

[40] C. D. Jr. Meyer, "The condition of a finite Markov chain and perturbation bounds for the limiting probabilities." *SIAM J. Alg. Disc. Math.* **1** (1980), 273-283.

[41] D. Mitra and J. McKenna, "Asymptotic expansions for closed Markovian networks with state-dependent service rates." *J. ACM* **33**, 3 (Jul. 1986), 598-592.

[42] P. Nain, "Qualitative properties of the Erlang blocking model with heterogeneous user requirements." *Queueing Systems* **6** (1990), 189-206.

[43] T. J. Ott and J. G. Shanthikumar, "On maxima and minima of partial sums of strongly interchangeable random variables." *Prob. Eng. Inf. Sci.* **4** (1990), 319-332.

[44] K. R. Pattipati, J. Wolf and S. Deb, "A calculus of variations approach to file allocation problems in computer systems." IBM Research Report RC 14948, Sep. 1989.

[45] K. R. Pattipati, J. Wolf and S. Deb, "A calculus of variations approach to parameter optimization in product-form queueing networks." IBM Research Report RC 15355, Jan. 1990.

[46] M. Véran and D. Potier, "QNAP 2: A portable environment for queueing systems modelling." *Proc. Int. Conf. on Modelling Techniques and Tools for Performance Analysis*, Paris, May 16-18, 1984. Also INRIA Report No. 314, Jun. 1984 (extensive version).

[47] M. I. Reiman and B. Simon, "Open queueing systems in light traffic." *Math. of Oper. Res.* **14**, 1 (1989), 26-59.

[48] M. I. Reiman and A. Weiss, "Light traffic derivatives via likelihood ratios." *IEEE Trans. Inf. Theory* IT-35, 3 (May 89), 648-654.

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

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

[51] K. W. Ross and D. D. Yao, "Monotonicity properties for the stochastic knapsack." *IEEE Trans. Inf. Theory* IT-36, 5 (1990), 1173-1179.

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

[53] C. L. Samelson 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), 830-840.

[54] P. J. Schweitzer, "Perturbation theory and finite Markov chains." *J. Appl. Prob.* **5** (1968), 401-413.

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

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

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

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

[59] J. C. Strelen, "A generalization of mean value analysis to higher moments: moment analysis." *Perf. Eval. Review* **14**, 1 (1986), 78-91.

[60] R. Suri, "Robustness of queueing network formulas." *J. ACM* **30**, 3 (Jul. 1983), 564-594.

[61] R. Suri, "A Concept of monotonicity and its characterization for closed queueing networks." *Oper. Res.* **33**, 3 (May-Jun. 1985), 606-624.

[62] Suri, R., "Infinitesimal perturbation analysis for general discrete event systems." *J. ACM* **34**, 3 (1987), 686-717.

[63] R. Suri, "Perturbation analysis: the state of the art and research issues explained via the $GI/G/1$ queue." *IEEE Proc.* Special Issue on Dynamics Discrete Events Systems **77**, 1 (Jan. 89), 114-137.

[64] Suri, R. and M. A. Zazanis, "Perturbation analysis gives strongly consistent sensitivity estimates for the $M/G/1$ queue." *Oper. Res.* **34**, 1 (Jan. 1988), 39-64.

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

[66] K. S. Trivedi, R. A. Wagner and T. M. Sigmon, "Optimal selection of CPU speed, device capacities, and file assignments." *J. ACM* **27**, 3 (Jul. 1980), 457-473.

[67] K. S. Trivedi and R. E. Kinicki, "A model for computer configuration design." *Computer* **13** (Apr. 1980), 47-54.

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

[69] N. M. van Dijk, "Perturbation theory for unbounded Markov reward processes with applications to queueing." *Adv. Appl. Prob.* **20** (1988), 99-111.

[70] N. M. van Dijk, "On the importance of bias-terms for error bound and comparison results." *Proc. First Int. Work. on Numerical Solutions of Markov Chains*, Raleigh, NC (Jan. 1989).

[71] N. M. van Dijk, "On 'stop=repeat' servicing for non-exponential queueing networks with blocking." *J. Appl. Prob.* **28** (1991), 159-173.

[72] N. M. van Dijk and M. L. Puterman, "Perturbation theory for Markov reward processes with applications to queueing systems." *Adv. Appl. Prob.* **20** (1988), 79-89.

[73] A. K. von Mayrhauser and K. S. Trivedi, "Computer configuration design to minimize response time." *Computer Performance* **3**, 1 (Mar. 1982), 32-39.