Insensitive Queueing Models for Communication Networks

T. Bonald†
France Telecom R&D
38-40 rue du general Leclerc
92794 Issy-les-Moulineaux, France
thomas.bonald@francetelecom.com

***Abstract*— A rich class of communication networks can be represented as queueing networks with state-dependent arrival rates and service rates. We provide necessary and sufficient conditions for such queueing networks to be insensitive in the sense that the steady-state distribution depends on the service time distribution at each queue through the mean only. This insensitivity property is key to the development of simple engineering rules that do not require the knowledge of fine traffic statistics.**

***Index Terms*— Queueing theory, partial reversibility, insensitivity, communication networks.**

I. Introduction

Since its publication in 1917, the Erlang formula has proved instrumental in sizing telephone networks [13]. It determines the required number of telephone lines given a prediction of expected demand and a target blocking probability. A key property of the Erlang formula is its insensitivity: the blocking probability does not depend on the holding time distribution beyond the mean [28]. Traffic is in fact characterized by a unique parameter, the *traffic intensity*, which is defined as the product of the call arrival rate and the mean holding time. This makes the Erlang formula both simple to apply and robust to changes in fine traffic characteristics, and explains its enduring success.

The only assumption required by the Erlang model is that calls arrive as a Poisson process. It may in fact be shown that the Erlang formula holds for non-Poisson call arrivals provided call *sessions* arrive as a Poisson process, where a session corresponds to the sequence of calls generated by the same user [4]. This assumption is reasonable for a large user population. For a small user population, sessions do not arrive as a Poisson process (the higher the number of ongoing sessions, the less likely the arrival of new sessions). All sessions are then considered as permanent, the traffic intensity being determined by the ratio of mean call duration to mean idle duration. This is the well-known Engset model [10], [11]. For equal traffic intensities, the Engset formula gives a lower blocking probability than the Erlang formula and tends to the latter when the user population grows to infinity.

Both the Erlang model and the Engset model are insensitive to all traffic characteristics beyond the traffic intensity. Call durations and idle durations may have arbitrary distributions. There may be arbitrary correlation between these random variables within the same session. The blocking probability is given by the corresponding Erlang or Engset formula, which is a function of traffic intensity and the number of telephone lines only [4].

These strong insensitivity results extend to reservation-based communication networks like circuit-switched networks. The corresponding models are known as loss networks [21] and include the extension of the Erlang model to a link shared by users having different bandwidth requirements [12], [15], [19], [23]. The blocking probability of a call depends only on its resource requirement (bandwidth, path in the network) and on the traffic intensity of each type of call.

It has recently been shown that similar insensitivity results hold for connection-less data networks like the Internet. Data transfers are represented as fluid flows whose bit rate changes at each flow arrival and flow departure. The basic model is a processor-sharing queue that represents a single, evenly shared bottleneck link [22]. Both the distribution of the number of active flows and the throughput of each flow is insensitive to all traffic characteristics beyond the traffic intensity [2]. It is again sufficient that *sessions* arrive as a Poisson process, each session consisting of a random sequence of flows separated by idle periods. For a finite user population, the analogue of the Engset model with a fixed number of permanent sessions applies [3], [16].

The same insensitivity property is satisfied by more complex models that consist of several links and where flows do not have a full access to the network but are constrained by some access line. It is only required that resources are shared according to balanced fairness [6], [8]. The throughput of each flow then depends only on its characteristics (access bit rate, path in the network) and on the traffic intensity of each type of flow. Various traffic control schemes like admission control and load balancing can be considered as well [5].

We shall see that all these traffic models belong in fact to a class of queueing systems we refer to as *partially reversible* networks. Specifically, consider a queueing network with Poisson external arrivals and independent, exponentially distributed service requirements. The service discipline is processor-sharing at each queue. We refer to the network state as the vector of the number of customers at each queue. External arrival rates, routing probabilities and service rates depend on the network state. The network may be open, closed or mixed depending on its state.

† Also affiliated with École Normale Supérieure, France.

Unformally stated, we say that the network is partially reversible if:

(i) in each network state, the traffic equations have a positive solution, referred to as the arrival rates;
(ii) the corresponding network with the same arrival rates and without routing (cf. figure 1) is described by a reversible Markov process.

We prove that the invariant measure of a partially reversible queueing network is explicit and insensitive to the service requirement distribution at each queue. Conversely, *any* queueing network whose invariant measure is insensitive to the service requirement distribution at each queue must satisfy conditions (i) and (ii). Partially reversible networks are therefore the only queueing systems leading to insensitive results.

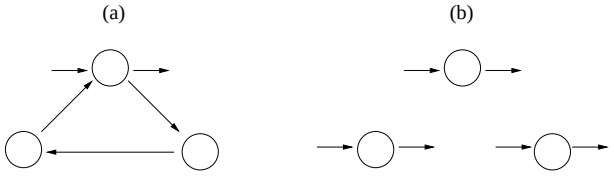


Fig. 1. A queueing network (a) and the corresponding network with the same arrival rates and without routing (b).

We believe these insensitive queueing models are key to the derivation of simple and robust engineering rules that do not require the knowledge of fine traffic statistics. Moreover, they may be useful for the design of traffic control schemes like congestion control, admission control and load balancing that make the performance of communication networks both simple to predict and robust to changes in traffic characteristics.

We present a brief overview of related work in the next section. The notion of partial reversibility is introduced in section III. We prove in section IV that partial reversibility is a necessary and sufficient condition for insensitivity. Section V is devoted to examples of communication networks that can be represented as partially reversible queueing networks. Section VI concludes the paper.

II. Related work

The study of queueing networks started with the seminal work of Jackson [18]. We first introduce Jackson networks, then present various extensions including state-dependent arrival rates and service rates. For convenience, we describe open networks only though all these networks have their closed counterpart (cf. section III). We refer the reader to the book of Serfozo [27] for a detailed bibliography on queueing networks. In all the paper, we say that a Markov process is reversible if the local balance property holds. In particular, we do not impose ergodicity.

A. Jackson networks

Consider a Jackson network of N queues with external arrival rates ν_i, routing probabilities p_{ij} and service rates μ_i. Let $X(t)$ be the N-dimensional vector whose i-th component gives the number of customers in queue i at time t. It is well-known that the invariant measure of the Markov process $\{X(t)\}_{t\geq 0}$ has the product form:

$$\pi(x) = \left(\frac{\lambda_1}{\mu_1}\right)^{x_1} \dots \left(\frac{\lambda_N}{\mu_N}\right)^{x_N},$$

where the arrival rates λ_i are the solution of the traffic equations:

$$\lambda_i = \nu_i + \sum_{j=1}^{N} \lambda_j p_{ji}, \quad i = 1, \dots, N. \tag{1}$$

This solution exists and is unique provided all customers eventually leave the network.

B. State-dependent service rates

Assume that the service rates now depend on the network state. Let e_i be the N-dimensional unit vector with 1 in component i and 0 elsewhere. If there is a positive function Φ such that for all i and all states x,

$$\Phi(x) = \Phi(x + e_i)\mu_i(x + e_i), \tag{2}$$

the invariant measure of the Markov process $\{X(t)\}_{t\geq 0}$ becomes:

$$\pi(x) = \lambda_1^{x_1} \dots \lambda_N^{x_N} \Phi(x).$$

We refer to such networks as Whittle networks. Usual product-form networks like BCMP networks [1] and Kelly networks [20] may be seen as Whittle networks where each customer class is represented as a separate queue [27]. It is worth noting that condition (2) is equivalent to the reversibility of the *service* process, which describes the evolution of the state of a virtual network with unit external arrival rates and no routing. The function Φ corresponds to the invariant measure of this Markov process.

C. State-dependent arrival rates

Assume now that, in addition to the service rates, both the external arrival rates and the routing probabilities depend on the network state. Let $\lambda_i(x)$ be the arrival rate at queue i in state x. The arrival rates are the solution of the traffic equations:

$$\lambda_i(x) = \nu_i(x) + \sum_{j=1}^{N} \lambda_j(x) p_{ji}(x + e_j), \quad i = 1, \dots, N. \tag{3}$$

We assume that this solution exists and is unique for all states x. If there is a positive function Λ such that for all i and all states x:

$$\Lambda(x)\lambda_i(x) = \Lambda(x + e_i), \tag{4}$$

the invariant measure of the Markov process $\{X(t)\}_{t\geq 0}$ becomes:

$$\pi(x) = \Lambda(x)\Phi(x).$$

Similarly, condition (4) is equivalent to the reversibility of the *arrival* process, which describes the evolution of the state in a virtual network with external arrival rates equal to the arrival rates, no routing and unit service rates. The function Λ corresponds to the invariant measure of this Markov process. Thus

the invariant measure π of the Markov process $\{X(t)\}_{t\geq 0}$ may be seen as the product of the invariant measures of the arrival process and the service process. This is often considered as the most general class of queueing networks whose invariant measure is explicit.

It is in fact not necessary to assume that both the arrival process *and* the service process are reversible [7]. Provided the traffic equations (3) have a positive solution in all states x, it is sufficient that the Markov process that describes the evolution of the state of a virtual network with external arrival rates equal to the arrival rates and no routing is reversible (see condition (ii) above). This is the notion of partial reversibility we present in section III in its most general form, that is for an arbitrary state space, a state-dependent network structure (open, closed or mixed), and possibly null arrival rates and service rates.

D. Insensitivity

Partially reversible networks are actually covered by the results of Hordijk and van Dijk [17], where the so-called *adjoint process* plays the role of the Markov process of condition (ii). But the network state is described in terms of the location of each *customer* in the network instead of the number of customers in each *queue*. As a consequence, the insensitivity property is expressed in terms of the service distribution of each customer. This is a much stronger requirement than the insensitivity to the service distribution at each queue we are interested in. Necessary conditions for the latter insensitivity property cannot be deduced from the results of Hordijk and van Dijk. The more general results derived by Whittle [29] for Markov processes and by Schassberger [24] for generalised semi-Markov processes also apply to the insensitivity to the service distribution of each customer only [25], [26].

Regarding the insensitivity to the service distribution at each queue, we proved in [7] that condition (ii) is necessary and sufficient in networks that satisfy condition (i). We here prove the much stronger result that both conditions (i) and (ii) are necessary and sufficient for insensitivity. In particular, the traffic equations must have a positive solution in all states x for a network to be insensitive to the service distribution at each queue. This solution is unique for open networks and defined up to a multiplicative constant per closed subnetwork for closed or mixed networks.

E. Other queueing networks

Following the seminal work of Kelly [20], many authors have studied so-called *symmetric* service disciplines, that lead to insensitive results. We here restrict the analysis to the processor-sharing service discipline, which is sufficient to represent most usual communication networks (cf. section V). Other extensions include networks with batch arrivals and batch services, introduced by Boucherie and van Dijk [9], and networks with negative customers introduced by Gelenbe [14]. We do not consider such extensions in the present paper.

III. Partial reversibility

We first introduce the notion of partial reversibility in its most general form, then consider the specific cases of open, closed and mixed networks. Since the service requirements are assumed to be independent, exponentially distributed at each queue, we do not specify the service discipline in this section.

A. General framework

Consider a network of N queues. Some customers are generated by a source, follow a random path in the network and eventually leave the network. Some other customers stay forever in the network. We refer to arrivals from the source as *external arrivals*, to other arrivals as *internal arrivals*. Let $x = (x_1, \ldots, x_N)$ be the network state, where x_i denotes the number of customers in queue i. External arrivals at queue i form a Poisson process of intensity $\nu_i(x)$ in state x. We may have $\nu_i(x) = 0$, in which case there are no external arrivals at queue i in state x. The overall external arrival rate in state x is denoted by:

$$\nu(x) \stackrel{\text{def}}{=} \sum_{i=1}^{N} \nu_i(x).$$

After service completion at queue i in state x, a customer is routed to queue j with probability $p_{ij}(x)$ and leaves the network with probability:

$$p_i(x) \stackrel{\text{def}}{=} 1 - \sum_{j=1}^{N} p_{ij}(x).$$

Service requirements are independent, exponentially distributed of unit mean at each queue. The service rate of queue i is $\mu_i(x)$ in state x. We may have $\mu_i(x) = 0$, in which case no service is provided by queue i in state x. By convention, we let $p_{ii}(x) = 1$ in this case. We denote by e_i the N-dimensional unit vector with 1 in component i and 0 elsewhere.

Network dynamics: We are interested in the evolution of the network state. We denote by $X(t)$ the network state at time t. Under the above assumptions, the stochastic process $\{X(t)\}_{t\geq 0}$ is a Markov process with the following transition rates:

$$q(x, x + e_i) = \nu_i(x),$$
$$q(x + e_i, x + e_j) = \mu_i(x + e_i) p_{ij}(x + e_i),$$
$$q(x + e_i, x) = \mu_i(x + e_i) p_i(x + e_i).$$

We assume that this Markov process is *irreducible* and denote by $\mathcal{S}$ its state space, $\mathcal{S} \subset \mathbb{N}^N$. Without any loss of generality, we let $\nu_i(x) = \mu_i(x) = 0$ for all $x \notin \mathcal{S}$.

Customer path: We now describe the random path followed by an arbitrary customer in the network when the other customers are frozen in some state $x \in \mathbb{N}^N$. Specifically, let $\{R_n(x)\}_{n\geq 0}$ be the Markov chain on $\{0, 1, \ldots, N\}$ with transition matrix $P(x)$ defined by $P_{00}(x) = 1$ if $\nu(x) = 0$,

$$P_{00}(x) = 0, \quad P_{0i}(x) = \frac{\nu_i(x)}{\nu(x)} \quad \forall i \neq 0 \quad \text{otherwise},$$

and

$$P_{i0}(x) = p_i(x + e_i), \quad P_{ij}(x) = p_{ij}(x + e_i) \quad \forall i, j \neq 0.$$

Viewing state 0 as the source of external arrivals, the sequence of states visited by this Markov chain between two consecutive visits of state 0 corresponds to the path followed by an arbitrary customer arriving in the network when the other customers are frozen in state x. We assume that the Markov chain $\{R_n(x)\}_{n\geq 0}$ has *closed* communication classes. Its invariant measure $\eta(x)$, defined by the balance equation:

$$P(x)\eta(x) = \eta(x) \tag{5}$$

with $\eta_i(x) > 0$ for all $i \in \{0, 1, \ldots, N\}$, is then unique up to a multiplicative constant per communication class. We denote by $\mathcal{C}_i(x)$ the communication class of state i. For all $i \in \mathcal{C}_0(x)$, the ratio $\eta_i(x)/\eta_0(x)$ corresponds to the mean number of visits to state i between two consecutive visits to state 0.

Traffic equations: We refer to the *arrival rate* to queue i in state x as:

$$\lambda_i(x) = \nu(x)\frac{\eta_i(x)}{\eta_0(x)}, \quad i = 1, \ldots, N.$$

In view of the balance equation (5), the arrival rates satisfy the traffic equations:

$$\nu(x) = \sum_{i=1}^{N} \lambda_i(x) p_i(x + e_i) \tag{6}$$

and

$$\lambda_i(x) = \nu_i(x) + \sum_{j=1}^{N} \lambda_j(x) p_{ji}(x + e_j). \tag{7}$$

The positive solution to these equations is unique up to a multiplicative constant per communication class. Equation (6) states that the departure rate from the source is equal to the arrival rate to the source. Equations (7) state that the arrival rate at each queue i is the sum of the external arrival rate and the internal arrival rate. Note that (6) follows from (7) by summation.

Partial reversibility: The network is said to be partially reversible if the traffic equations have a positive solution in all states $x \in \mathbb{N}^N$ and if the Markov process $\{\bar{X}(t)\}_{t\geq 0}$, defined by the transition rates:

$$\bar{q}(x, x + e_i) = \lambda_i(x), \ \bar{q}(x + e_i, x) = \mu_i(x + e_i), \quad i \in \mathcal{C}_0(x),$$

$$\bar{q}(x + e_i, x + e_j) = \mu_i(x + e_i)\lambda_j(x), \quad i \notin \mathcal{C}_0(x), \ j \in \mathcal{C}_i(x),$$

is reversible. This describes the evolution of the state of a network with arrival rate $\lambda_i(x)$ at queue i in state x and without routing in the corresponding open component (cf. figure 1), with homogeneous routing in the corresponding closed components (cf. figure 3 below). Following Hordijk and van Dijk [17], we refer to the Markov process $\{\bar{X}(t)\}_{t\geq 0}$ as the *adjoint* process. We denote by $\mathcal{G}$ the corresponding transition graph.

We have the following key result:

Theorem 1: For a partially reversible network, the Markov process $\{X(t)\}_{t\geq 0}$ has the same invariant measure π as the adjoint process, given by $\pi(y) = 1$ for some reference state $y \in \mathcal{S}$ and for all states $x \in \mathcal{S}$, $x \neq y$, by

$$\pi(x) = \prod_{k=1}^{n} \frac{\bar{q}(x(k-1), x(k))}{\bar{q}(x(k), x(k-1))},$$

where $x(0) \equiv y, x(1), \ldots, x(n) \equiv x$ denotes any path from state y to state x in the transition graph $\mathcal{G}$.

Proof: It follows from the traffic equations (6) and (7) that the measure π satisfies:

$$\pi(x)\nu(x) = \sum_{i=1}^{N} \pi(x + e_i)\mu_i(x + e_i)p_i(x + e_i),$$

$$\pi(x + e_i)\mu_i(x + e_i) = \pi(x)\nu_i(x) + \sum_{j=1}^{N} \pi(x + e_j)\mu_j(x + e_j)p_{ji}(x + e_j).$$

These are equations of partial balance, which imply the global balance by summation:

$$\pi(x)\left(\nu(x) + \sum_{i=1}^{N} \mu_i(x)\right) = \sum_{i=1}^{N} \pi(x - e_i)\nu_i(x - e_i) + \sum_{i,j=1}^{N} \pi(x + e_j - e_i)\mu_j(x + e_j - e_i)p_{ji}(x + e_j - e_i) + \sum_{i=1}^{N} \pi(x + e_i)\mu_i(x + e_i)p_i(x + e_i),$$

where we use the convention that $\pi(x) = 0$ if $x \notin \mathbb{N}^N$. ■

It is worth noting that the Markov process $\{X(t)\}_{t\geq 0}$ associated with a partially reversible network is generally not reversible. We verify that the Markov process $\{X(t)\}_{t\geq 0}$ is reversible if and only if for all states $x \in \mathbb{N}^N$, the Markov chain $\{R_n(x)\}_{n\geq 0}$ is reversible:

Corollary 1: For a partially reversible network, the Markov process $\{X(t)\}_{t\geq 0}$ is reversible if and only if for all states $x \in \mathbb{N}^N$, the Markov chain $\{R_n(x)\}_{n\geq 0}$ describing the path of an arbitrary customer when the other customers are frozen in state x is reversible, that is:

$$\nu_i(x) = \lambda_i(x)p_i(x + e_i), \quad \forall i,$$

$$\lambda_i(x)p_{ij}(x + e_i) = \lambda_j(x)p_{ji}(x + e_j), \quad \forall i, j.$$

Proof: The Markov process $\{X(t)\}_{t\geq 0}$ is reversible if and only if for all states $x \in \mathbb{N}^N$,

$$\pi(x)\nu_i(x) = \pi(x + e_i)\mu_i(x + e_i)p_i(x + e_i), \quad \forall i,$$

$$\pi(x + e_i)\mu_i(x + e_i)p_{ij}(x + e_i) = \pi(x + e_j)\mu_j(x + e_j)p_{ji}(x + e_j), \quad \forall i, j.$$

The result is then a direct consequence of Theorem 1. ■

B. Open networks

By definition, the network is said to be open if:

$$\forall x \in \mathbb{N}^N, \quad \mathcal{C}_0(x) = \{0\} \cup \{i : x + e_i \in \mathcal{S}\}.$$

Thus all other communication classes correspond to non-admissible states. An example of open network is given in figure 1.

Partial reversibility implies the existence of a positive solution to the traffic equations. Figure 2 below gives examples of non-partially reversible queueing networks. In both cases, there is no positive solution to the traffic equations. Note that it is sufficient that the traffic equations do not have a positive solution in at least *one* state x for the insensitivity property to be violated.



Fig. 2. Examples of non-partially reversible queueing networks: the traffic equations do not have a positive solution.

C. Closed networks

By definition, the network is said to be closed if:

$$\forall x \in \mathbb{N}^N, \quad \mathcal{C}_0(x) = \{0\}.$$

In particular, there is no external arrivals and the network contains a fixed number of customers. The network may consist in several closed subnetworks, as illustrated in Figure 3 (a). Note that these subnetworks may interact through the state-dependent service rates. The set of closed subnetworks may also change depending on the state x. Partial reversibility implies the existence of a positive solution to the traffic equations in all states x. In addition, the Markov process that describes the evolution of the network with the same arrival rates and homogeneous routing probabilities per subnetwork, as illustrated in Figure 3 (b), must be reversible.

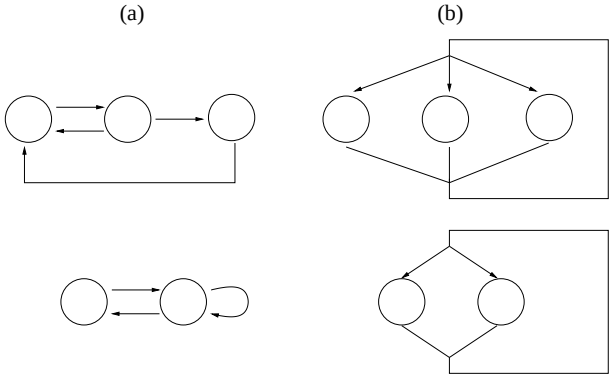


Fig. 3. A closed queueing network consisting of two subnetworks (a) and the corresponding network with the same arrival rates and homogeneous routing probabilities per subnetwork (b).

Figure 4 below gives an example of a non-partially reversible closed network: the traffic equations do not have a positive solution. Again, it is sufficient that the traffic equations do not have a positive solution in at least *one* state x for the insensitivity property to be violated.

D. Mixed networks

Mixed networks may have open and closed components, as illustrated in Figure 5 (a). Partial reversibility implies the existence of a positive solution to the traffic equations in all states x. In addition, the Markov process that describes the evolution of the network with the same arrival rates, no routing for the open subnetwork and homogeneous routing probabilities per closed subnetwork, as illustrated in Figure 5 (b), must be reversible.

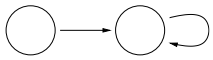


Fig. 4. Example of a non-partially reversible closed queueing network: the traffic equations do not have a positive solution.

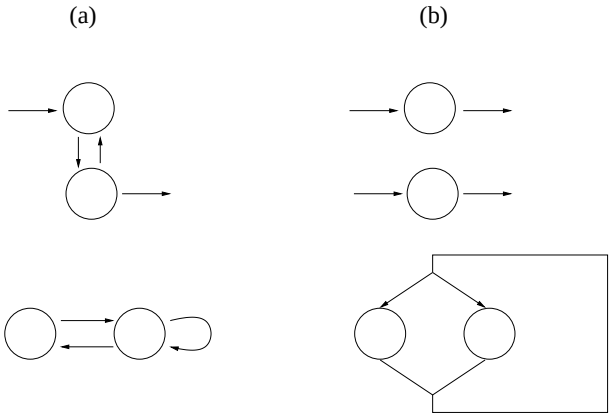


Fig. 5. A mixed queueing network consisting of an open subnetwork and a closed subnetwork (a) and the corresponding network with the same arrival rates, no routing for the open subnetwork and homogeneous routing probabilities for the closed subnetwork (b).

IV. Insensitivity

In this section, we assume that the service discipline is processor-sharing at each queue and prove that partial reversibility is a necessary and sufficient condition for insensitivity. We restrict the analysis to phase-type distributions (also known as *Cox* distributions) which are known to form a dense subset within the set of all distributions with nonnegative support [27].

Specifically, we say that the network is insensitive if the *invariant measure* of the number of customers at each queue remains unchanged when the distribution of service requirements at each queue is replaced by any phase-type distribution with the same mean. Note that we do not assume that the Markov process $\{X(t)\}_{t\geq 0}$ is ergodic. The following result is proved in the appendix:

Theorem 2: Partial reversibility is a necessary and sufficient condition for insensitivity.

In particular, the insensitivity property *implies* the existence of a positive solution to the traffic equations in all states x.

V. Application to Communication Networks

We now show that various communication networks can be represented as partially reversible queueing networks.

A. Reservation-based networks

Erlang model: We first consider the Erlang model with C telephone lines, Poisson call arrivals of intensity ν and independent, exponentially distributed call durations of mean $1/\mu$. This is represented as a single queue[1] with state-dependent arrival rate $\nu(x) = \nu$ if $x < C$, $\nu(x) = 0$ otherwise, and state-dependent service rate $\mu(x) = \mu x$. This is trivially a partially reversible network, with invariant measure:

$$\pi(x) = \frac{\rho^x}{x!}, \quad x = 0, 1, \ldots, C,$$

where $\rho \stackrel{\text{def}}{=} \nu/\mu$ corresponds to the traffic intensity in Erlangs. By the PASTA property, the call blocking probability B is equal to $\pi^t(C)$, the stationary probability that all lines are occupied:

$$B = \frac{\dfrac{\rho^C}{C!}}{1 + \rho + \dfrac{\rho^2}{2} + \ldots + \dfrac{\rho^C}{C!}}. \tag{8}$$

This is the Erlang formula, which in view of Theorem 2 does not depend on distribution of call durations beyond the mean.

Now assume users generate calls within sessions. We consider the simple case where sessions arrive as a Poisson process of intensity ν, start with a call of exponential duration of mean $1/\mu_1$, followed after an idle period of exponential duration of mean $1/\mu_2$ by a second call of exponential duration of mean $1/\mu_3$. In case of call blocking, the session goes on as if the call were accepted and terminated instantaneously. This is represented as a network of $N = 3$ queues on the state space:

$$\mathcal{S} = \{x : x_1 + x_3 \leq C\},$$

with null external arrival rates except for

$$\nu_1(x) = \nu \quad \text{if } x_1 + x_3 < C,$$

$$\nu_2(x) = \nu \quad \text{if } x_1 + x_3 = C,$$

null routing probabilities except for

$$p_{12}(x) = 1 \quad \text{if } x_1 + x_3 \leq C,$$

$$p_{23}(x) = 1 \quad \text{if } x_1 + x_3 < C,$$

and service rates:

$$\mu_1(x) = \mu_1 x_1, \quad \mu_2(x) = \mu_2 x_2, \quad \mu_3(x) = \mu_3 x_3, \quad \forall x \in \mathcal{S}.$$

This is an open queueing network with arrival rates:

$$\lambda_1(x) = \nu \quad \text{if } x_1 + x_3 < C, \quad \lambda_1(x) = 0 \quad \text{otherwise,}$$

$$\lambda_2(x) = \nu \quad \forall x \in \mathcal{S},$$

$$\lambda_3(x) = \nu \quad \text{if } x_1 + x_3 < C, \quad \lambda_3(x) = 0 \quad \text{otherwise.}$$

The network is partially reversible with invariant measure:

$$\pi(x) = \frac{\rho_1^{x_1} \rho_2^{x_2} \rho_3^{x_3}}{x_1! x_2! x_3!}, \quad x \in \mathcal{S},$$

[1]For convenience, we use the same notation for the arrival rate to the queue ν (resp. the per-customer service rate μ) and the state-dependent arrival rate $\nu(x)$ (resp. the state-dependent service rate $\mu(x)$).

where $\rho_i \stackrel{\text{def}}{=} \nu/\mu_i$. The blocking probability is again given by the stationary probability $\sum_{x:x_1+x_3=C} \pi^t(x)$ that all lines are occupied, which coincides with the Erlang formula (8) for the traffic intensity $\rho = \rho_1 + \rho_3$. In view of Theorem 2, the result is independent of the distributions of call and idle durations except for the traffic intensity. More generally, the call blocking probability is given by (8) for any type of session, with a random number of calls and possible correlation between the duration of successive calls and idle periods of the same session [4].

Engset model: The Engset model with a fixed number M of permanent sessions leads to similar insensitive results. Assume call durations are exponential of mean $1/\mu_1$, idle durations are exponential of mean $1/\mu_2$. This is represented as a closed queueing network of $N = 2$ queues on the state space:

$$\mathcal{S} = \{x : x_1 \leq C,\ x_1 + x_2 = M\},$$

with routing probabilities:

$$p_{12}(x) = 1 \quad \forall x \in \mathcal{S},$$

$$p_{21}(x) = 1 \quad \text{if } x_1 < C, \quad p_{21}(x) = 0 \quad \text{otherwise,}$$

and service rates:

$$\mu_1(x) = \mu_1 x_1, \quad \mu_2(x) = \mu_2 x_2, \quad \forall x \in \mathcal{S}.$$

The network is a partially reversible with invariant measure:

$$\pi(x) = \binom{M}{x_1} \varrho^{x_1}, \quad x \in \mathcal{S},$$

where $\varrho \stackrel{\text{def}}{=} \mu_2/\mu_1$ denotes the ratio of mean call duration to mean idle duration. Note that the PASTA property does not hold in this case. It is indeed well known that the network state distribution seen by new calls is in fact equal to the steady state distribution with $M - 1$ permanent sessions, yielding the following expression for the blocking probability when $C < M$:

$$B = \frac{\binom{M-1}{C} \varrho^C}{1 + (M-1)\varrho + \ldots + \binom{M-1}{C} \varrho^C}.$$

This is the Engset formula, which is insensitive to the distributions of call durations and idle durations beyond the ratio of their mean ϱ.

Multirate systems: We now consider a multirate system with N types of calls. Calls of type i arrive as a Poisson process of intensity ν_i and require a circuit of c_i bit/s during a random period of exponential duration of mean $1/\mu_i$. The link capacity is equal to C bit/s. New type-i calls are blocked when the link occupancy is higher than $C - c_i$. This is represented as a network of N queues on the state space:

$$\mathcal{S} = \{x : x.c \leq C\}, \quad x.c \stackrel{\text{def}}{=} \sum_{i=1}^{N} x_i c_i,$$

with null external arrival rates except for

$$\nu_i(x) = \nu_i \quad \text{if } x.c \leq C - c_i,$$

null routing probabilities and service rates:

$$\mu_i(x) = \mu_i x_i, \quad \forall x \in \mathcal{S}.$$

This is a partially reversible network with invariant measure:

$$\pi(x) = \prod_{i=1}^{N} \frac{\rho_i^{x_i}}{x_i!}, \quad x \in \mathcal{S}.$$

where $\rho_i \stackrel{\text{def}}{=} \nu_i/\mu_i$ corresponds to the traffic intensity of type-i calls in Erlangs. The blocking probability of type-i calls is equal to the steady-state probability that the link occupancy is higher than $C - c_i$ and, like the Erlang formula, is independent of all traffic characteristics beyond the traffic intensities $\rho_1, \ldots, \rho_N$.

Loss networks: Finally, we consider a network of L links. Link l has a capacity of C_l bit/s. There are N types of calls. Calls of type i arrive as a Poisson process of intensity ν_i and require a circuit of c_i bit/s through links $r_i \subset \{1, \ldots, L\}$ during a random period of exponential duration of mean $1/\mu_i$. Similarly, this is represented as a partially reversible network of N queues with invariant measure

$$\pi(x) = \prod_{i=1}^{N} \frac{\rho_i^{x_i}}{x_i!}, \quad x \in \mathcal{S},$$

on the state space:

$$\mathcal{S} = \left\{ x : \quad \forall l, \ \sum_{i: l \in r_i} x_i c_i \leq C_l \right\},$$

where $\rho_i \stackrel{\text{def}}{=} \nu_i/\mu_i$ corresponds to the traffic intensity of type-i calls in Erlangs. The blocking probability of type-i calls is equal to the steady-state probability that for some $l \in r_i$, link-l occupancy is higher than $C_l - c_i$, and is independent of all traffic characteristics beyond the traffic intensities $\rho_1, \ldots, \rho_N$.

B. Connection-less networks

Single bottleneck: Consider a single link of C bit/s. Data flows arrive as a Poisson process of intensity ν and have independent, exponential sizes of mean σ bits. This is represented as a single queue of arrival rate ν and service rate C/σ, which is trivially a partially reversible network with invariant measure:

$$\pi(x) = \rho^x, \quad x \in \mathbb{N},$$

where $\rho \stackrel{\text{def}}{=} \nu\sigma/C$ corresponds to the link load, which is assumed to be less than 1.

We measure user performance in terms of flow throughput, defined as the ratio γ of the mean flow size to the mean flow duration. By Little's law, we get:

$$\gamma = C - A,$$

where $A \stackrel{\text{def}}{=} \nu\sigma$ corresponds to the traffic intensity in bit/s. In view of Theorem 2, the flow throughput γ is insensitive to the flow size distribution beyond the mean. It may be easily verified as for the Erlang model that it is also insensitive to the flow arrival process provided flows are generated within sessions and sessions arrive as a Poisson process [2].

Access rates: Now assume flows are additionnally constrained by some fixed bit rate $c \leq C$. This is represented as a single queue of arrival rate ν and state-dependent service rate $\mu(x) = \min(xc, C)/\sigma$. The invariant measure becomes:

$$\pi(x) = \frac{\left(\frac{C}{c}\rho\right)^x}{x!}, \quad \text{if } xc \leq C,$$

$$\pi(x) = \rho\pi(x-1), \quad \text{otherwise.}$$

The system is stable if and only if the link load ρ is less than 1. The corresponding flow throughput can then be easily derived and is insensitive to all traffic characteristics beyond the traffic intensity.

Multirate systems: We now consider a multirate system with N types of data flows. Data flows of type i arrive as a Poisson process of intensity ν_i, have an exponential size of mean σ_i bits and are constrained by some fixed bit rate $c_i \leq C$. We assume that the link capacity C is shared according to balanced fairness [8]. The corresponding model is a partially reversible network of N queues with invariant measure:

$$\pi(x) = \prod_{i=1}^{N} \frac{\left(\frac{C}{c_i}\rho_i\right)^{x_i}}{x_i!}, \quad \text{if } x.c \leq C,$$

$$\pi(x) = \sum_{i=1}^{N} \rho_i \pi(x - e_i) \quad \text{otherwise,}$$

where $\rho_i \stackrel{\text{def}}{=} \nu_i\sigma_i/C$ corresponds to the link load due to the type-i flows. The system is stable if and only if the total link load $\sum_{i=1}^{N} \rho_i$ is less than 1. Again, the corresponding throughput of each type of flow can be easily derived and is insensitive to all traffic characteristics beyond the traffic intensities $A_1, \ldots, A_N$, with $A_i \stackrel{\text{def}}{=} \nu_i\sigma_i$.

Networks: These results can be generalized to networks with several links. Consider a network of L links. Link l has a capacity of C_l bit/s. There are N types of flows. Flows of type i arrive as a Poisson process of intensity ν_i, have an exponential size of mean σ_i bits and go through links $r_i \subset \{1, \ldots, L\}$. We do not consider per-flow rate limits for the sake of simplicity. Under balanced fairness, the traffic model is a partially reversible network of N queues with invariant measure:

$$\pi(x) = \Phi(x) \prod_{i=1}^{N} A_i^{x_i},$$

where Φ is the function recursively defined by $\Phi(0) = 1$,

$$\Phi(x) = \max_{l=1,\ldots,L} \frac{1}{C_l} \sum_{i: l \in r_i} \Phi(x - e_i),$$

with $\Phi(x) = 0$ if $x \notin \mathbb{N}^N$, and $A_i = \nu_i\sigma_i$ corresponds to the traffic intensity of type-i flows. The network is stable if and only if the load of each link is less than 1:

$$\forall l = 1, \ldots, L, \quad \sum_{i: l \in r_i} \frac{A_i}{C_l} < 1.$$

The results are insensitive to all traffic characteristics beyond the traffic intensities $A_1, \ldots, A_N$.

Admission control: Flows may be subject to admission control if some minimum throughput must be guaranteed to ongoing flows. The corresponding invariant measure is then the restriction of the invariant measure without admission control to the set of admissible states. The results are insensitive to all traffic characteristics beyond the traffic intensity provided the session goes on in case of flow blocking, as for the above considered reservation-based networks.

Load balancing: Flows may additionnally be routed to less loaded links. Consider N parallel links of respective capacities $C_1, \ldots, C_N$. Flows arrive as a Poisson process of intensity ν and have independent, exponential sizes of mean σ bits. We denote by x the network state, where x_i is the number of ongoing flows on the i-th link. We assume the number of flows on the i-th link cannot exceed some fixed value M_i. New flows are routed to link i with probability $q_i(x)$, with $q_i(x) = 0$ if $x_i = M_i$, and are blocked with probability:

$$p(x) = 1 - \sum_{i=1}^{N} q_i(x).$$

This is represented as a network of N queues on the state space

$$\mathcal{S} = \{x : x_1 \leq M_1, \ldots, x_N \leq M_N\}$$

with external arrival rates $\nu_i(x) = \nu q_i(x)$, null routing probabilities and service rates $\mu_i = C_i/\sigma$. Assume that flows are blocked if and only if all links are fully occupied in the sense that $x_i = M_i$ for all i. Partial reversibility imposes to route flows to link i in proportion to $M_i - x_i$ in non-blocking states [5]:

$$q_i(x) = \frac{M_i - x_i}{\sum_{j=1}^{N}(M_j - x_j)}.$$

The corresponding invariant measure is given by:

$$\pi(x) = \binom{\sum_{j=1}^{N}(M_j - x_j)}{M_1 - x_1, \ldots, M_N - x_N} \prod_{i=1}^{N} \rho_i^{x_i}, \quad x \in \mathcal{S},$$

where $\rho_i \stackrel{\text{def}}{=} \nu\sigma/C_i$ corresponds to the relative load of link i. User performance can then be evaluated in terms of flow blocking probability and flow throughput. It is insensitive to all traffic characteristics beyond the traffic intensity.

VI. Conclusion

The insensitivity property is key to the derivation of simple and robust engineering rules that do not require the knowledge of fine traffic statistics. Since Erlang's pioneer work, telephone networks have been sized based on the prediction of the demand only, and not on the distribution of holding times that changes over the years. We believe the insensitive queueing models described in the present paper are useful for sizing current communication networks and could serve as guidelines for the design of future traffic control schemes like congestion control, admission control and load balancing.

The following appendices are devoted to the proof of Theorem 2: partial reversibility is a necessary and sufficient condition for insensitivity.

Appendix I
Sufficient condition

A partially reversible network is insensitive: We prove the result for i.i.d. service requirements consisting of M exponential phases. Denote by α_{im} the mean service requirement of the m-th phase at queue i, $m = 1, \ldots, M$, and by q_{im} the probability that a customer enters the $(m+1)$-th phase after the completion of the m-th phase at queue i, $m = 1, \ldots, M-1$. We assume that $\alpha_{im} > 0$ and $q_{im} > 0$ for all i, m and define:

$$\beta_{i1} = \alpha_{i1}, \quad \beta_{im} = \alpha_{im} \prod_{n=1}^{m-1} q_{in} \quad \text{for all } m = 2, \ldots, M.$$

Service requirements are kept of unit mean so that:

$$\forall i = 1, \ldots, N, \quad \sum_{m=1}^{M} \beta_{im} = 1. \tag{9}$$

Let y_{im} be the number of customers in m-th phase of service at queue i. We denote by y the vector $(y_1, \ldots, y_M)$, where $y_m \stackrel{\text{def}}{=} (y_{1m}, \ldots, y_{Nm})$ gives the number of customers in m-th phase of service in each queue. This describes the state of a network of $N \times M$ queues, indexed by im with $i = 1, \ldots, N$, $m = 1, \ldots, M$, with external arrival rates in state y:

$$\tilde{\nu}_{i1}(y) = \nu_i(y_1 + \ldots + y_M),$$
$$\tilde{\nu}_{im}(y) = 0 \quad \text{for } m = 2, \ldots, M,$$

routing probabilities:

$$\tilde{p}_{im,j1}(y) = p_{ij}(y_1 + \ldots + y_M)(1 - q_{im}),$$
$$\tilde{p}_{im,jn}(y) = 0 \quad \text{for } n = 2, \ldots, M,$$

with $q_{iM} \stackrel{\text{def}}{=} 0$, and service rates:

$$\tilde{\mu}_{im}(y) = \mu_i(y_1 + \ldots + y_M) \times \frac{1}{\alpha_{im}} \frac{y_{im}}{y_{i1} + \ldots + y_{iM}}$$
$$\text{if } y_{i1} + \ldots + y_{iM} > 0, \quad \tilde{\mu}_{im}(y) = 0 \quad \text{otherwise.}$$

The probability that a customer leaves the network in state y after service completion at queue im is given by:

$$\tilde{p}_{im}(y) = 1 - \sum_{j=1}^{N} \tilde{p}_{im,j1}(y),$$

that is

$$\tilde{p}_{im}(y) = p_i(y_1 + \ldots + y_M)(1 - q_{im}).$$

In view of the corresponding traffic equations, the arrival rates are given by:

$$\tilde{\lambda}_{i1}(y) = \lambda_i(y_1 + \ldots + y_M),$$
$$\tilde{\lambda}_{im}(y) = \lambda_i(y_1 + \ldots + y_M) \prod_{n=1}^{m-1} q_{in}, \quad m = 2, \ldots, M.$$

Since the original network of N queues is partially reversible, the new network of $N \times M$ queues is partially reversible with invariant measure $\tilde{\pi}$ given by:

$$\begin{aligned} \tilde{\pi}(y) &= \pi(y_1 + \ldots + y_M) \\ &\times \prod_{i=1}^{N} \frac{(y_{i1} + \ldots + y_{iM})!}{y_{i1}! \ldots y_{iM}!} \beta_{i1}^{y_{i1}} \ldots \beta_{iM}^{y_{iM}} \end{aligned}$$

on the state space:

$$\tilde{\mathcal{S}} = \{y : y_1 + \ldots + y_M \in \mathcal{S}\}.$$

Denoting by f_{im} the $N \times M$-dimensional unit vector with 1 in component im and 0 elsewhere, we indeed have that for all $y \in \mathbb{N}^{N\times M}$:

$$\tilde{\pi}(y)\tilde{\lambda}_{im}(y) = \tilde{\pi}(y + f_{im})\tilde{\mu}_{im}(y + f_{im})\tilde{p}_{im}(y + f_{im}), \quad \forall i, m,$$

and

$$\tilde{\pi}(y + f_{im})\tilde{\mu}_{im}(y + f_{im})\tilde{p}_{im,jn}(y + f_{im}) = \tilde{\pi}(y + f_{jn})\tilde{\mu}_{jn}(y + f_{jn})\tilde{p}_{jn,im}(y + f_{jn}), \quad \forall i, m, j, n.$$

The insensitivity property then follows from the fact that:

$$\sum_{y:y_1+\ldots+y_M=x} \tilde{\pi}(y) = \pi(x) \sum_{y:y_1+\ldots+y_M=x} \prod_{i=1}^{N} \frac{(y_{i1} + \ldots + y_{iM})!}{y_{i1}! \ldots y_{iM}!} \beta_{i1}^{y_{i1}} \ldots \beta_{iM}^{y_{iM}},$$

which is equal to $\pi(x)$ for all states $x \in \mathcal{S}$ in view of (9).

Appendix II
Necessary condition

An insensitive network is partially reversible: We now consider a queueing network as described in §III-A, whose associated Markov process $\{X(t)\}_{t\geq 0}$ is irreducible on some state space $\mathcal{S}$. We do not assume that the communication classes of the Markov chain $\{R_n(x)\}_{n\geq 0}$ are closed in all states x. In particular, the traffic equations may have no solution. We prove that the insensitivity property *implies* the existence of a solution to the traffic equations. Moreover, the adjoint process is reversible.

The proof is by induction on N. The property holds for $N = 1$. Since $\mathcal{S} \subset \mathbb{N}$, the Markov process $\{X(t)\}_{t\geq 0}$ is indeed reversible and coincides with the adjoint process. Now assume that it holds for any network of $N-1$ queues, for some $N \geq 2$, and consider an insensitive network of N queues. Assume that the corresponding Markov process $\{X(t)\}_{t\geq 0}$ is irreducible on some state space $\mathcal{S}$. The balance equations in any state $x \in \mathcal{S}$ are:

$$\pi(x)\left(\nu(x) + \sum_{i=1}^{N} \mu_i(x)\right) = \sum_{i=1}^{N} \pi(x - e_i)\nu_i(x - e_i) + \sum_{i,j=1}^{N} \pi(x + e_j - e_i)\mu_j(x + e_j - e_i)p_{ji}(x + e_j - e_i),$$

where we use the convention that $\pi(x) = 0$ if $x \notin \mathbb{N}^N$.

We now consider i.i.d. service requirements at queue N equal to 0 with probability $1-\alpha$ and exponentially distributed of mean $1/\alpha$ with probability α, for some constant α such that $0 < \alpha < 1$. Note that the mean service requirement remains equal to 1. This corresponds to a new network with i.i.d. exponential service requirements of unit mean but external arrival rates:

$$\nu_i^{(\alpha)}(x) = \nu_i(x) + (1-\alpha)\nu_N(x)\frac{p_{Ni}(x + e_N)}{1 - (1-\alpha)p_{NN}(x + e_N)}$$

for all $i = 1, \ldots, N-1$,

$$\nu_N^{(\alpha)}(x) = \frac{\alpha\nu_N(x)}{1 - (1-\alpha)p_{NN}(x + e_N)},$$

routing probabilities:

$$p_{ij}^{(\alpha)}(x) = p_{ij}(x) + (1-\alpha)p_{iN}(x)\frac{p_{Ni}(x + e_N - e_i)}{1 - (1-\alpha)p_{NN}(x + e_N - e_i)}$$

for all $i = 1, \ldots, N$, $j = 1, \ldots, N-1$,

$$p_{iN}^{(\alpha)}(x) = \alpha p_{iN}(x) + (1-\alpha)p_{iN}(x)\frac{\alpha p_{NN}(x + e_N - e_i)}{1 - (1-\alpha)p_{NN}(x + e_N - e_i)}$$

for all $i = 1, \ldots, N$, and service rates:

$$\mu_i^{(\alpha)}(x) = \mu_i(x), \quad i = 1, \ldots, N-1, \quad \mu_N^{(\alpha)}(x) = \alpha\mu_N(x).$$

By assumption, π is an invariant measure for that network for all α, $0 < \alpha < 1$. We let α tend to 0. If $p_{NN}(x+e_N) < 1$, the limiting external arrival rates are given by:

$$\tilde{\nu}_i(x) = \nu_i(x) + \nu_N(x)\frac{p_{Ni}(x + e_N)}{1 - p_{NN}(x + e_N)}$$

for all $i = 1, \ldots, N-1$,

$$\tilde{\nu}_N(x) = 0.$$

Now if $p_{NN}(x + e_N) = 1$, we have $p_{Ni}(x + e_N) = 0$ for all $i = 1, \ldots, N-1$ so that the limiting external arrival rates are given by:

$$\tilde{\nu}_i(x) = \nu_i(x), \quad i = 1, \ldots, N.$$

Similarly, consider the routing probabilities from queue i. If $p_{NN}(x + e_N - e_i) < 1$, the limiting routing probabilities are given by:

$$\tilde{p}_{ij}(x) = p_{ij}(x) + p_{iN}(x)\frac{p_{Nj}(x + e_N - e_i)}{1 - p_{NN}(x + e_N - e_i)}$$

for all $i = 1, \ldots, N$, $j = 1, \ldots, N-1$,

$$\tilde{p}_{iN}(x) = 0, \quad i = 1, \ldots, N.$$

Now if $p_{NN}(x+e_N-e_i) = 1$, we have $p_{Nj}(x+e_N-e_i) = 0$ for $j = 1, \ldots, N-1$ so that the limiting routing probabilities are given by:

$$\tilde{p}_{ij}(x) = p_{ij}(x), \quad j = 1, \ldots, N.$$

Finally, the limiting service rates are given by:

$$\tilde{\mu}_i(x) = \mu_i(x), \quad i = 1, \ldots, N-1, \quad \tilde{\mu}_N(x) = 0.$$

Now assume that for some state $x \in \mathcal{S}$, $\tilde{\nu}_N(x) > 0$ or $\tilde{p}_{iN}(x) > 0$ for some $i = 1, \ldots, N-1$. In the latter case, we must have $\tilde{\mu}_i(x) > 0$ in view of the convention that $\mu_i(x) = 0$ implies $p_{ii}(x) = 1$ (see §III-A). Since $\tilde{\mu}_N(x) = 0$ in all states x, π is the invariant measure of a network with a positive probability flow *into* queue N and a null probability flow *out of* queue N, which is a contradiction. Thus for all $x \in \mathcal{S}$, $\tilde{\nu}_N(x) = 0$ and $\tilde{p}_{iN}(x) = 0$ for all $i = 1, \ldots, N-1$. In particular, if $p_{NN}(x + e_N) = 1$ for some $x \in \mathbb{N}^N$, then $\nu_N(x) = 0$ and $p_{iN}(x + e_i) = 0$ for all $i = 1, \ldots, N-1$.

The limiting balance equations in state x are:

$$\pi(x)\sum_{i=1}^{N-1}(\tilde{\nu}_i(x)+\mu_i(x)) = \sum_{i=1}^{N-1}\pi(x-e_i)\tilde{\nu}_i(x-e_i)$$
$$+ \sum_{i,j=1}^{N-1}\pi(x+e_i-e_j)\mu_i(x+e_i-e_j)\tilde{p}_{ij}(x+e_i-e_j)$$
$$+ \sum_{i=1}^{N-1}\pi(x+e_i)\mu_i(x+e_i)\tilde{p}_i(x+e_i).$$

with

$$\tilde{p}_i(x) \stackrel{\text{def}}{=} 1-\sum_{j=1}^{N-1}\tilde{p}_{ij}(x), \quad i=1,\ldots,N-1.$$

These equations are the balance equations of a Markov process $\{\tilde{X}(t)\}_{t\geq 0}$ describing the state of a network of $N-1$ queues. We apply the inductive assumption to this insensitive network. If the Markov process $\{\tilde{X}(t)\}_{t\geq 0}$ is reducible, we consider its restriction to each of its communication classes. Since $\pi(x) > 0$ for all $x \in \mathcal{S}$, the Markov process $\{\tilde{X}(t)\}_{t\geq 0}$ has no transient state so that the restriction to each of its communication classes is irreducible.

By induction, the Markov chain $\{\tilde{R}_n(y)\}_{n\geq 0}$ describing the path of an arbitrary customer in this network of $N-1$ queues when the other customers are frozen in state y has closed communicating classes for all $y \in \mathbb{N}^{N-1}$ and the adjoint Markov process is reversible. Since $p_{NN}(x+e_N)=1$ implies $\nu_N(x)=0$ and $p_{iN}(x+e_i)=0$ for all $i=1,\ldots,N-1$, the Markov chain $\{R_n(x)\}_{n\geq 0}$ describing the path of an arbitrary customer in the original network of N queues when the other customers are frozen in state x also has closed communicating classes for all $x \in \mathbb{N}^N$. We denote by $\lambda_1(x),\ldots,\lambda_N(x)$ the corresponding arrival rates. We verify from the traffic equations (7) that $\lambda_1(x),\ldots,\lambda_{N-1}(x)$ are solutions of the traffic equations associated with the network of $N-1$ queues:

$$\lambda_i(x) = \tilde{\nu}_i(x) + \sum_{j=1}^{N-1}\lambda_j(x)\tilde{p}_{ji}(x+e_j).$$

Thus the transition rates of the adjoint Markov process associated with the network of $N-1$ queues are equal to the corresponding transition rates $\bar{q}$ of the adjoint Markov process associated with the original network of N queues. Since the former is reversible, we deduce that for all states $x,y \in \mathcal{S}$, $x \neq y$, such that $x_N = y_N$,

$$\pi(x)\bar{q}(x,y) = \pi(y)\bar{q}(y,x). \tag{10}$$

Since node N does not play any particular role, this equality is satisfied for all states $x,y \in \mathcal{S}$, $x \neq y$, such that $x_i = y_i$ for some $i=1,\ldots,N$. If $N \geq 3$, we have $\bar{q}(x,y)=0$ for all states $x,y \in \mathcal{S}$ such that $x_i \neq y_i$ for all $i=1,\ldots,N$. This is also true for $N=2$ except if $p_{12}(x+e_1)=p_{21}(x+e_2)=1$ for some $x \in \mathbb{N}^2$. But the state space $\mathcal{S}$ then reduces to the two states $x+e_1$, $x+e_2$ and the Markov process $\{X(t)\}_{t\geq 0}$ is reversible and coincides with the adjoint process. In all cases, equation (10) is satisfied for all states $x,y \in \mathcal{S}$, $x \neq y$. The adjoint Markov process associated with the original network of N queues is reversible.

REFERENCES

[1] F. Baskett, K.M. Chandy, R.R. Muntz and F.G. Palacios, "Open, closed and mixed networks of queues with different classes of customers," *J. Assoc. Comput. Mach.*, vol. 22, pp. 248–260, 1975.

[2] S. Ben Fredj, T. Bonald, A. Proutière, G. Régnié and J.W. Roberts, "Statistical bandwidth sharing: A study of congestion at fbw level," in *Proc. ACM SIGCOMM*, 2001.

[3] A.W. Berger, Y. Kogan, "Dimensioning bandwidth for elastic traffi c in high-speed data networks," *IEEE/ACM Trans. on Networking*, vol. 8-5, pp. 643–654, 2000.

[4] T. Bonald, "The Erlang model with non-Poisson call arrivals," to appear in *Proc. ACM Sigmetrics / IFIP Performance*, 2006.

[5] T. Bonald, M. Jonckheere, A. Proutière, "Insensitive load balancing," in *Proc. ACM Sigmetrics / IFIP Performance*, 2004.

[6] T. Bonald, L. Massoulié, A. Proutière, J. Virtamo, "A queueing analysis of max-min fairness, proportional fairness and balanced fairness," to appear in *Queueing Systems*, 2006.

[7] T. Bonald and A. Proutière, "Insensitivity in processor-sharing networks," *Performance Evaluation*, vol. 49, pp. 193–209, 2002.

[8] T. Bonald and A. Proutière, "Insensitive bandwidth sharing in data networks," *Queueing Systems*, vol. 44-1, pp. 69–100, 2003.

[9] R. Boucherie and N. van Dijk, "Product forms for queueing networks with state dependent multiple job transitions," *Adv. App. Prob.*, vol. 23, pp. 152–187, 1991.

[10] J.W. Cohen, "The Generalized Engset Formula," *Phillips Telecommunications Review*, vol. 18, pp. 158–170, 1957.

[11] T.O. Engset, "On the calculation of switches in an automatic telephone system," in: *Tore Olaus Engset: The man behind the formula*, Eds: A. Myskja, O. Espvik, 1998. First appeared as an unpublished report in Norwegian, 1915.

[12] O. Enomoto, H. Miyamoto, "An Analysis of mixtures of multiple bandwidth traffi c on time division switching networks," in *Proc. of the 7th International Teletraffi c Congress*, 1973.

[13] A.K. Erlang, "Solution of some problems in the theory of probabilities of signifi cance in automatic telephone exchanges," in: *The life and works of A.K. Erlang*, Eds: E. Brockmeyer, H.L. Halstrom, A. Jensen, 1948. First published in Danish, 1917.

[14] E. Gelenbe, "Queuing networks with negative and positive customers," *J. App. Prob.*, vol. 28, pp. 656–663, 1991.

[15] L.A. Gimpelson, "Analysis of mixtures of wide and narrow-band traffi c," *IEEE Trans. Comm. Technology*, vol. 13-3, pp. 258–266, 1965.

[16] D.P. Heyman, T.V. Lakshman, A.L. Neidhardt, "A new method for analysing feedback-based protocols with applications to engineering Web traffi c over the Internet," in *Proc. ACM SIGMETRICS*, 1997.

[17] A. Hordijk and N. van Dijk, "Adjoint processes, job local balance and insensitivity of stochastic networks," *Bull:44 session Int. Stat. Inst.*, vol. 50, pp. 776–788, 1982.

[18] J.R. Jackson, "Networks of waiting lines," *Operat. Res.*, vol. 5, pp. 518–521, 1957.

[19] J. S. Kaufman, "Blocking in a shared resource environment," *IEEE Trans. Commun.*, vol. 29, pp. 1474–1481, 1981.

[20] F.P. Kelly, *Reversibility and Stochastic Networks*, Wiley, 1979.

[21] F.P. Kelly, "Loss networks," *Annals of Applied Probability*, vol. 1, pp. 319–378, 1991.

[22] L. Massoulié and J.W. Roberts, "Bandwidth sharing and admission control for elastic traffi c," *Telecommunication Systems*, vol. 15, pp. 185–201, 2000.

[23] J. W. Roberts, "A service system with heterogeneous user requirement," in *Performance of Data Communications Systems and Their Applications*, G. Pujolle, Ed. Amsterdam, The Netherlands: North-Holland, 1981, pp. 423–431.

[24] R. Schassberger, "Insensitivity of steady state distributions of generalized semi-Markov processes with speeds," *Advances in Applied Probability*, vol. 10, pp. 836–851, 1978.

[25] R. Schassberger, "The insensitivity of stationary distributions in networks of queues," *Advances in Applied Probability*, vol.10, pp.906–912, 1978.

[26] R. Schassberger, "Two remarks on insensitive stochastic models," *Advances in Applied Probability*, vol. 18, pp. 791–814, 1986.

[27] R.F. Serfozo, *Introduction to Stochastic Networks*, Springer Verlag, 1999.

[28] B.A. Sevastyanov, "An ergodic theorem for Markov processes and its application to telephone systems with refusals," *Theor. Probability Appl.*, vol. 2, pp. 104–112, 1957.

[29] P. Whittle, "Partial balance and insensitivity," *Journal of Applied Probability*, vol. 22, pp. 168–176, 1985.