Insensitive Traffic Models for Communication Networks

Thomas Bonald

To cite this version:

Thomas Bonald. Insensitive Traffic Models for Communication Networks. Discrete Event Dynamic Systems, Springer Verlag, 2007, 10.1007/s10626-007-0012-5 . hal-01275545

Insensitive Traffic Models for Communication Networks

T. Bonald
France Telecom
Issy-les-Moulineaux, France
thomas.bonald@orange-ftgroup.com

June 11, 2007

Abstract: We present a survey of traffic models for communication networks whose key performance indicators like blocking probability and mean delay are independent of all traffic characteristics beyond the traffic intensity. This insensitivity property, which follows from that of the underlying queuing networks, is key to the derivation of simple and robust engineering rules like the Erlang formula in telephone networks.

Keywords: Traffic modeling, communication networks, bandwidth sharing, insensitivity, Kelly-Whittle queuing networks.

1 Introduction

Since its publication in 1917, the Erlang formula has provided an essential tool for sizing telephone networks [15]. 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 [27]. 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.

Contrary to common belief, it is not even necessary to assume that calls arrive as a Poisson process. Each user typically generates several calls during the same activity period, which may produce call bursts. Referring to a *session* as the sequence of calls generated by the same user, it is in fact sufficient to assume that *sessions* arrive as a Poisson process [4]. This assumption is reasonable for a large user population. For a small user population, sessions may simply be considered as permanent and this corresponds to the Engset model [12, 13]. The Erlang model is just a limiting case of the Engset model when the user population grows to infinity. Both are insensitive to all traffic characteristics beyond the traffic intensity. Call durations and idle period durations may have arbitrary distributions and be correlated. The blocking probability is still given by the corresponding Erlang or Engset formula, which is a function of traffic intensity only [4]. This insensitivity property extends to circuit-switched networks where users require circuits of different bandwidth. 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 [14, 16, 19, 22, 25].

Similar insensitive models have recently been applied to packet-switched networks like the Internet. Most traffic in today's Internet is generated by the transfer of digital documents like web pages, audio and video files. This traffic is *elastic* in that the duration of each transfer depends on network congestion. Each document is split into a sequence of packets, referred to as a *flow*, whose sending rate is adapted in response to congestion indications such as packet losses, typically under the control of TCP[1]. The quality of the transfer then depends on the time required to transfer all the packets of the flow. In this sense, network performance for elastic traffic is mainly manifested at flow level and can be gauged by measures like the mean flow duration.

An appropriate abstraction in this context entails entirely disregarding the complex packet-level phenomena and considering each flow content as a fluid which is transmitted as a continuous stream through the network. The rate of this fluid data stream depends on the set of flows that compete for the same network resources. The basic model is a processor-sharing queue that represents a single bottleneck link whose bandwidth is equally shared by the ongoing flows [24]. Both the distribution of the number of ongoing flows and the mean flow duration are insensitive to all traffic characteristics beyond the traffic intensity [1]. Like circuit traffic, referring to a *session* as the sequence of flows generated by the same user, it is sufficient to assume that *sessions* arrive as a Poisson process. For a finite user population, the analogue of the Engset model with a fixed number of permanent sessions applies [2, 17].

The same insensitivity property holds for more general network models whose resources are shared according to balanced fairness [6]. The mean duration of each flow then depends only on its characteristics (resource requirements, rate limit) and on the traffic intensity of each type of flow. Again, this insensitivity property is key to the derivation of simple and robust engineering rules that do not require knowledge of fine traffic statistics.

In this paper, we present a survey of these insensitive traffic models in the unified framework of Kelly-Whittle queuing networks. Circuit traffic and elastic traffic are considered in Sections 2 and 3, respectively. The underlying queuing networks and the corresponding insensitivity results are presented in Section 4. Section 5 concludes the paper.

[1] Transmission Control Protocol, see [28].

2 Circuit traffic

We start with networks where each communication necessitates the prior establishment of a physical or virtual *circuit*, corresponding to the set of required resources. A typical example is the switched telephone network. By analogy, we shall refer to any communication as a *call*. We are interested in the evaluation of the call *blocking* probability, that is the probability that the resources required by a new call are not available.

Unless otherwise specified, we assume that calls arrive as a Poisson process and have independent, exponentially distributed durations. The insensitivity property of the subsequent traffic models with respect to these traffic characteristics will be shown in Section 4.

2.1 Erlang model

We first consider a single link consisting of C circuits. We denote by λ the call arrival rate and by $1/\mu$ the mean call duration. Let $\alpha = \lambda/\mu$ be the traffic intensity in Erlangs. By Little's law, this corresponds to the mean number of calls in the absence of blocking. The system is an $M/M/C/C$ queue. The number of ongoing calls has the stationary distribution:

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

By the PASTA[2] property, the call blocking probability B is equal to $\pi(C)$, the stationary probability that all lines are occupied:

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

This is the well-known Erlang formula [15]. Note that B is decreasing in C for fixed load $\rho = \alpha/C$. This "economy of scale" property easily follows from the integral representation of B:

$$\frac{1}{B} = \int_0^\infty e^{-t}\left(1 + \frac{t}{\alpha}\right)^C dt.$$

2.2 Engset model

Now consider a finite number of sources, K. Each source is either active, i.e., with an ongoing call, or idle. Call durations are independent, exponentially distributed with mean $1/\mu$. Idle period durations are independent, exponentially distributed with mean $1/\nu$. We refer to $\alpha = \nu/\mu$, the ratio of the mean call duration to the mean idle period duration, as the traffic intensity per idle source.

[2] Poisson Arrivals See Time Averages, see [26].

We assume that $K > C$ so that some calls may be blocked. A source whose call is blocked starts an new idle period as if the call were accepted and completed instantaneously. The system corresponds to a closed network of two queues, an $./M/\infty$ queue and an $./M/C/C$ queue, with K customers that alternately visit both queues and jump over the $./M/C/C$ queue in case of blocking. The number of ongoing calls has the stationary distribution [9]:

$$\pi(x) = \pi(0)\binom{K}{x}\alpha^x, \quad x = 0, 1, \dots, C.$$

Note that the PASTA property does not hold in this case. The steady state distribution as seen by new calls is in fact equal to the steady state distribution of a system with $K-1$ sources, yielding the following expression for the blocking probability:

$$B = \frac{\binom{K-1}{C}\alpha^C}{1 + (K-1)\alpha + \frac{(K-1)(K-2)}{2}\alpha^2 + \dots + \binom{K-1}{C}\alpha^C}. \tag{2}$$

This is the Engset formula [13]. Note that B is decreasing in C and increasing in K for fixed load $\rho = K\alpha/C$, the limiting case $K \to \infty$ corresponding to the Erlang formula. These properties easily follow from the expression:

$$\begin{aligned}\frac{1}{B} &= 1 + \alpha^{-1}\frac{C}{K-C} + \frac{\alpha^{-2}}{2}\frac{C(C-1)}{(K-C)(K-C+1)} \\ &+ \dots + \frac{\alpha^{-C}}{C!}\frac{C!}{(K-C)(K-C+1)\dots K}.\end{aligned}$$

2.3 Multi-rate model

The multi-rate model consists of a link of C bit/s shared by N types of calls. Type-i calls require a circuit of constant bit rate $r_i \le C$ bit/s. They arrive as a Poisson process of intensity λ_i and have independent, exponentially distributed durations with mean $1/\mu_i$. We denote by $\alpha_i = \lambda_i/\mu_i$ the corresponding traffic intensity in Erlangs. The system state is described by the line vector $x = (x_1, \dots, x_N)$ of the number of ongoing calls of each type. Denoting by r the column vector $(r_1, \dots, r_N)'$, this system has the stationary distribution:

$$\pi(x) = \pi(0)\frac{\alpha_1^{x_1}}{x_1!}\dots\frac{\alpha_N^{x_N}}{x_N!}, \quad xr \le C.$$

By the PASTA property, the blocking probability of type-i calls is equal to the probability that the link occupancy is higher than $C - r_i$.

When the rates $r_1, \dots, r_N$ and the link capacity C have integer values, the stationary distribution of link occupancy can be evaluated through the following recursion known as the Kaufman-Roberts algorithm [19, 25]. Let:

$$p(n) = \sum_{x:xr=n}\frac{\alpha_1^{x_1}}{x_1!}\dots\frac{\alpha_N^{x_N}}{x_N!}, \quad n = 0, 1, \dots, C.$$

This is, up to a normalization constant, the probability that the link occupancy is equal to n. In particular, the blocking probability of type-i calls is given by:

$$B_i = \frac{\sum_{n=C-r_i+1}^{C} p(n)}{\sum_{n=0}^{C} p(n)}. \tag{3}$$

For all $n = 1, \ldots, C$, we have:

$$p(n) = \sum_{i=1}^{N} \frac{\alpha_i r_i}{n} p(n - r_i),$$

with $p(0) = 1$ and $p(n) = 0$ for all $n < 0$. Note that the computational cost of this recursive formula is linear in the number of classes N, unlike the direct calculation based on the stationary distribution which is exponential in N. Similar results exist for non-Poisson arrivals, when calls are generated by a finite number of sources like in the Engset model [10].

2.4 Network model

Finally, we consider a finite set of facilities that may represent any transmission resource of a communication network (e.g. bandwidth, code, timeslot, power). We consider J such resources. Resource j has a capacity of C_j units. There are N types of calls. We denote by λ_i the arrival rate of type-i calls, by $1/\mu_i$ their mean duration and by $\alpha_i = \lambda_i/\mu_i$ the corresponding traffic intensity in Erlangs. Type-i calls require a circuit with a_{ij} resource-j units, for all $j = 1, \ldots, J$. We may have $a_{ij} = 0$ in which case type-i calls require no resource-j unit. Denote by A the matrix with i, j-entry a_{ij} and by C the line vector $(C_1, \ldots, C_J)$. As above, the system state is described by the line vector $x = (x_1, \ldots, x_N)$. The state space is the set of states x such that $xA \leq C$ component-wise. The system state has the stationary distribution:

$$\pi(x) = \pi(0) \frac{\alpha_1^{x_1}}{x_1!} \ldots \frac{\alpha_N^{x_N}}{x_N!}, \quad xA \leq C.$$

By the PASTA property, the blocking probability of type-i calls, B_i, is equal to the probability that the consumption of at least one resource j is higher than $C_j - a_{ij}$. Equivalently, $1 - B_i$ is the steady-state probability that the consumption of each resource j is less than or equal to $C_j - a_{ij}$. Denoting by a_i the i-th line of matrix A, we get:

$$B_i = 1 - \frac{\sum_{x: xA \leq C - a_i} \frac{\alpha_1^{x_1}}{x_1!} \cdots \frac{\alpha_N^{x_N}}{x_N!}}{\sum_{x: xA \leq C} \frac{\alpha_1^{x_1}}{x_1!} \cdots \frac{\alpha_N^{x_N}}{x_N!}}.$$

The exact evaluation of this expression turns out to be computationally expensive, even using the generalization of the Kaufman-Roberts algorithm [11]. The following bound proved by Whitt [29] allows one to decouple the system and to evaluate the contribution of each resource independently. We have:

$$B_i \leq 1 - \prod_{j=1}^{J} (1 - B_{ij}),$$

where B_{ij} denotes the blocking probability of type-i calls when the network reduces to resource j (i.e. there is no other resource contraint). This corresponds to the multi-rate model described in §2.3 for which the Kaufman-Roberts algorithm applies. In practice, one may use the looser bound:

$$B_i \leq \sum_{j=1}^{J} B_{ij}. \tag{4}$$

The model extends to a finite number of sources, like the Engset model described in §2.2. Denote by K_i the number of type-i sources and by α_i the traffic intensity per type-i idle source (i.e. the ratio of the mean call duration to the mean idle period duration of type-i sources, cf. §2.2). The stationary distribution of the system state becomes:

$$\pi(x) = \pi(0) \binom{K_1}{x_1} \alpha_1^{x_1} \dots \binom{K_N}{x_N} \alpha_N^{x_N}, \quad xA \leq C,\ x_1 \leq K_1, \dots, x_N \leq K_N.$$

We deduce the blocking probability as for the Engset model. The previous model with Poisson call arrivals corresponds to the limiting case $K_1 \to \infty, \dots, K_N \to \infty$ for fixed total traffic intensities $K_1\alpha_1, \dots, K_N\alpha_N$.

3 Elastic traffic

We now consider fluid models of packet-switched networks, as described in Section 1. Traffic is *elastic* in that each flow has a fixed size (in bits) but a variable duration depending on its throughput. This is a key difference with the models of circuit traffic considered so far, where the duration of each call is independent of the network state. We are interested in the evaluation of the mean flow duration, as well as in the flow blocking probability in the presence of admission control.

Unless otherwise specified, we assume that flows arrive as a Poisson process and have independent, exponentially distributed sizes. The insensitivity property of the subsequent traffic models with respect to these traffic characteristics will be shown in Section 4.

3.1 Processor sharing model

Consider a single link of C bit/s. We denote by λ the flow arrival rate and by σ the mean flow size (in bits). The traffic intensity is given by $\beta = \lambda\sigma$ (in bit/s) and the link load by $\rho = \beta/C$. Let x be the number of ongoing flows. We assume that flows equally share the link capacity so that each flow has throughput C/x in the presence of x flows, for all $x \geq 1$. The system is an $M/M/1$ processor-sharing queue. In particular, the number of ongoing flows has the stationary distribution:

$$\pi(x) = \pi(0)\rho^x, \quad x = 0, 1, 2, \ldots,$$

under the stability condition $\rho < 1$. We deduce the mean number of flows:

$$\bar{x} = \frac{\rho}{1-\rho}.$$

The mean per-bit delay τ, defined as the ratio of the mean flow duration to the mean flow size, follows from Little's law:

$$\tau = \frac{\bar{x}}{\lambda\sigma} = \frac{1}{C(1-\rho)}. \tag{5}$$

In the presence of admission control that limits the number of ongoing flows to some constant $M \geq 1$, the stationary distribution of the number of flows is the restriction of π to the state space $\{0, 1, \ldots, M\}$. The mean per-bit delay becomes:

$$\tau = \frac{\bar{x}}{\lambda(1-B)\sigma} = \frac{1-(M+1)B}{1-B}\frac{1}{C(1-\rho)},$$

where B is the blocking probability, given by:

$$B = \frac{\rho^M}{1+\rho+\rho^2+\ldots+\rho^M}.$$

We verify that τ tends to M/C and B tends to 1 when $\rho \to \infty$.

3.2 Finite source model

Now consider a finite number of sources, K. Each source is either active, i.e., with an ongoing flow, or idle. Flow sizes are independent, exponentially distributed with mean σ. Idle period durations are independent, exponentially distributed with mean $1/\nu$. We refer to $\beta = \nu\sigma$, the ratio of the mean flow size to the mean idle period duration, as the traffic intensity per idle source (in bit/s). We denote by $\alpha = \beta/C$ the load per idle source and by $\rho = K\alpha$ the total load. The system corresponds to a closed network of two queues, an $./M/1$ processor-sharing queue and an $./M/\infty$ queue, with K customers that alternately visit both queues. The number of ongoing flows has the stationary distribution:

$$\pi(x) = \pi(0)\frac{K!}{(K-x)!}\alpha^x, \quad x = 0, 1, \ldots, K.$$

Denote by λ the flow arrival rate. We have:

$$\lambda = \frac{C}{\sigma}(1-\pi(0)).$$

The mean per-bit delay then follows from Little's law:

$$\tau = \frac{\bar{x}}{\lambda\sigma} = \frac{1}{C} \frac{\sum_{x=1}^{K} x \frac{K!}{(K-x)!}\alpha^x}{\sum_{x=1}^{K} \frac{K!}{(K-x)!}\alpha^x}. \tag{6}$$

It may be verified that τ increases in K for fixed load $\rho = K\alpha$. The limiting case $K \to \infty$ corresponds to the processor sharing model.

In the presence of admission control that limits the number of ongoing flows to some constant $M < K$, the stationary distribution of the number of flows is the restriction of π to the state space $\{0, 1, \ldots, M\}$. The mean per-bit delay becomes:

$$\tau = \frac{1}{C} \frac{\sum_{x=1}^{M} x \frac{K!}{(K-x)!}\alpha^x}{\sum_{x=1}^{M} \frac{K!}{(K-x)!}\alpha^x}.$$

The steady state distribution as seen by new flows is equal to the steady state distribution of a system with $K - 1$ sources, yielding the following expression for the blocking probability:

$$B = \frac{\frac{(K-1)!}{(K-1-M)!}\alpha^M}{1 + (K-1)\alpha + \ldots + \frac{(K-1)!}{(K-1-M)!}\alpha^M}.$$

We verify that τ tends to M/C and B tends to 1 when $\rho \to \infty$.

3.3 A common rate limit

Assume flows are additionally constrained by some fixed bit rate $r \le C$. Flows arrive as a Poisson process of intensity λ and have independent, exponentially distributed sizes with mean σ. We denote by $\beta = \lambda\sigma$ the traffic intensity in bit/s and by $\alpha = \beta/r$ the equivalent traffic intensity in Erlangs for virtual circuits of r bit/s. This corresponds to the mean number of ongoing flows in the absence of link capacity constraint, that is for $C = \infty$. The link load is $\rho = \beta/C$. In the presence of x flows, each flow has throughput r if $xr \le C$ and C/x otherwise. The number of ongoing flows has the stationary distribution:

$$\pi(x) = \pi(0)\frac{\alpha^x}{x!} \quad \text{if } xr \le C,$$

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

under the stability condition $\rho < 1$. The mean per-bit delay τ follows from Little's law. When C/r is an integer, say m, we get:

$$\tau = \frac{1}{r} + \frac{S/\rho}{C(1-\rho)}, \tag{7}$$

where S denotes the probability that the link is saturated:

$$S = \sum_{x>m} \pi(x) = \frac{\frac{\alpha^m}{m!}\frac{\rho}{1-\rho}}{1 + \alpha + \ldots + \frac{\alpha^{m-1}}{(m-1)!} + \frac{\alpha^m}{m!}\frac{1}{1-\rho}}.$$

In the presence of admission control that limits the number of ongoing flows to some constant $M \geq m$, the stationary distribution of the number of flows is the restriction of π to the state space $\{0, 1, \ldots, M\}$. We deduce the mean per-bit delay as above. The flow blocking probability, given by

$$B = \frac{\dfrac{\alpha^m}{m!}\rho^{M-m}}{1 + \alpha + \ldots + \dfrac{\alpha^{m-1}}{(m-1)!} + \dfrac{\alpha^m}{m!}(1 + \rho + \ldots + \rho^{M-m})},$$

coincides with the Erlang formula when $M = m$, in which case there is no elastic bandwidth sharing. Similar results can be derived for the finite source model of §3.2. In the presence of admission control with a maximum of M flows, the flow blocking probability coincides with the Engset formula when $M = m$.

3.4 Multi-rate model

Consider the extension of the above model to N types of flows. Type-i flows have the rate limit $r_i \leq C$. We denote by λ_i their arrival rate, by σ_i their mean size in bits, by $\beta_i = \lambda_i \sigma_i$ their traffic intensity in bit/s, by $\alpha_i = \beta_i / r_i$ their traffic intensity in Erlangs for virtual circuits of r_i bit/s and by $\rho_i = \beta_i / C$ their contribution to the link load. We denote by $\beta = \sum_{i=1}^N \beta_i$ the total traffic intensity (in bit/s) and by $\rho = \sum_{i=1}^N \rho_i = \beta / C$ the link load.

The system state is described by the line vector $x = (x_1, \ldots, x_N)$ of the number of ongoing flows of each type. Let r be the column vector $(r_1, \ldots, r_N)'$. Each ongoing type-i flow has throughput r_i when $xr \leq C$ and must share the link capacity C with the other flows in progress when $xr > C$. Denoting by $\phi_i(x)$ the total throughput of type-i flows in state x, the throughput constraints are the following:

$$\forall x \in \mathbb{N}^N, \quad \sum_{i=1}^N \phi_i(x) \leq C \quad \text{and} \quad \phi_i(x) \leq x_i r_i, \quad i = 1, \ldots, N. \tag{8}$$

The stationary distribution of the system state depends on the way flows share link capacity when $xr > C$. We consider balanced fair sharing [6], which is defined in such a way that the underlying queuing network is a Kelly-Whittle network, as explained in Section 4.

In the following, we denote by e_i the N-dimensional line vector whose i-th component is equal to 1 and other components are equal to 0. Under balanced fairness, the total throughput of type-i flows in state x is given by:

$$\phi_i(x) = \frac{\Phi(x - e_i)}{\Phi(x)},$$

where Φ is the associated balance function, recursively defined by:

$$\Phi(x) = \prod_{i=1}^N \frac{1}{r_i^{x_i} x_i!} \quad \text{if} \quad xr \leq C, \tag{9}$$

$$\Phi(x) = \frac{1}{C} \sum_{i=1}^{N} \Phi(x - e_i) \quad \text{otherwise,} \tag{10}$$

with $\Phi(x) = 0$ if $x \notin \mathbb{N}^N$. Note that the capacity constraints (8) are satisfied: the throughput of each flow is equal to its rate limit when $xr \leq C$ and the total throughput is equal to C otherwise. For a unit capacity link with two rate limits, $r_1 = 1$ and $r_2 = 1/2$, for instance, we verify that $\phi_1(x) = 2/3$ and $\phi_2(x) = 1/3$ for $x = (1,1)$. Max-min fair sharing, on the other hand, that tends to allocate resources as equally as possible, would give the same throughput to both flows in state $x = (1,1)$ [3].

A key property of balanced fairness is that the associated system state has an explicit stationary distribution, given by:

$$\pi(x) = \pi(0)\Phi(x) \prod_{i=1}^{N} \beta_i^{x_i}, \quad x \in \mathbb{N}^N, \tag{11}$$

under the stability condition $\rho < 1$. In view of (9) and (10), the stationary distribution satisfies the recursion:

$$\pi(x) = \pi(0) \prod_{i=1}^{N} \frac{\alpha_i^{x_i}}{x_i!} \quad \text{if} \quad xr \leq C, \tag{12}$$

$$\pi(x) = \sum_{i=1}^{N} \rho_i \pi(x - e_i) \quad \text{otherwise,} \tag{13}$$

which generalizes the expression obtained for a common rate limit, cf. §3.3. The mean duration of each type of flow then follows by Little's law.

We have the analogue of the Kaufman-Roberts algorithm described in §2.3. Assume the rates $r_1, \ldots, r_N$ and the link capacity C have integer values. Let:

$$p(n) = \sum_{x:xr=n} \Phi(x) \prod_{i=1}^{N} \beta_i^{x_i}, \quad n \in \mathbb{N}.$$

For all $n < C$, this is, up to a normalization constant, the probability that the link occupancy is equal to n. Defining:

$$\bar{p} = \sum_{n>C} p(n),$$

the probability of link saturation is given by:

$$S = \sum_{x:xr>C} \pi(x) = \frac{\bar{p}}{1 + p(1) + \ldots + p(C) + \bar{p}}.$$

For all $n = 1, \ldots, C$, we have:

$$p(n) = \sum_{i=1}^{N} \frac{\alpha_i r_i}{n} p(n - r_i),$$

with $p(0) = 1$ and $p(n) = 0$ for all $n < 0$, and

$$\bar{p} = \sum_{i=1}^{N} \frac{\rho_i \bar{p}_i}{1-\rho} \quad \text{with} \quad \bar{p}_i = \sum_{C-r_i<n\leq C} p(n).$$

The former equality is the Kaufman-Roberts algorithm and follows from (12). The latter follows from (13):

$$\bar{p} = \sum_{x:xr>C} \pi(x) = \sum_{x:xr>C} \sum_{i=1}^{N} \rho_i \pi(x-e_i) = \sum_{i=1}^{N} \rho_i(\bar{p}_i + \bar{p}).$$

A similar recursive algorithm is described in [8] for the mean per-bit delay.

3.5 Multi-need model

Now consider a single resource of C units shared by flows having different resource requirements. Specifically, there are N types of flows and type-i flows require a_i resource units per bit/s. Like for the multi-rate model, type-i flows have a specific rate limit, here equal to C/a_i. But now a type-i flow needs *all* the resource units, C, to achieve this rate limit. We denote by λ_i the arrival rate of type-i flows, by σ_i their mean size in bits, by $\beta_i = \lambda_i\sigma_i$ their traffic intensity in bit/s and by $\rho_i = \beta_i a_i/C$ their contribution to the system load. The total system load is given by $\rho = \sum_{i=1}^{N} \rho_i$.

The system state is described by the line vector $x = (x_1, \dots, x_N)$ of the number of ongoing flows of each type. Denoting by $\phi_i(x)$ the total throughput of type-i flows in state x, the throughput constraint is the following:

$$\forall x \in \mathbb{N}^N, \quad \sum_{i=1}^{N} \phi_i(x) a_i \leq C. \tag{14}$$

Again, the stationary distribution of the system state depends on the way flows share the resource. Balanced fair sharing is defined by

$$\phi_i(x) = \frac{\Phi(x-e_i)}{\Phi(x)},$$

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

$$\Phi(x) = \frac{1}{C} \sum_{i=1}^{N} \Phi(x-e_i) a_i,$$

with $\Phi(x) = 0$ if $x \notin \mathbb{N}^N$. Note that the capacity constraint (14) is attained in all states $x \neq 0$. We get:

$$\Phi(x) = \binom{x_1 + \ldots + x_N}{x_1, \ldots, x_N} \prod_{i=1}^{N} \left(\frac{a_i}{C}\right)^{x_i}$$

and

$$\phi_i(x) = \frac{x_i}{x_1 + \ldots + x_N} \frac{C}{a_i}. \tag{15}$$

Note that $\phi_i(x)a_i$ corresponds to the total amount of resource units allocated to type-i flows in state x. In view of (15), the available resource units C are equally shared under balanced fairness, which results in throughputs $\phi_1(x), \ldots, \phi_N(x)$ inversely proportional to the resources requirements $a_1, \ldots, a_N$.

In view of (11), the system state has the stationary distribution:

$$\pi(x) = \pi(0) \binom{x_1 + \ldots + x_N}{x_1, \ldots, x_N} \prod_{i=1}^{N} \rho_i^{x_i}, \quad x \in \mathbb{N}^N,$$

under the stability condition $\rho < 1$. We deduce the mean number of type-i flows:

$$\bar{x}_i = \frac{\rho_i}{1 - \rho},$$

and by Little's law, the mean per-bit delay of type-i flows:

$$\tau_i = \frac{\bar{x}_i}{\lambda_i \sigma_i} = \frac{a_i}{C(1 - \rho)}. \tag{16}$$

Thus the mean per-bit delay is proportional to the resource requirement.

3.6 Network model

Finally, we consider a general model that combines the previous two models and extends them to the case of several resources. Various examples of data networks covered by this model are described in [6]. We consider J resources indexed by j. Resource j has a capacity of C_j units. There are N types of flows. Type-i flows consume a_{ij} resource-j units per bit/s, for all $j = 1, \ldots, J$. We may have $a_{ij} = 0$ in which case type-i flows require no resource-j unit, but we assume that $a_{ij} > 0$ for at least one resource j. Type-i flows have the rate limit r_i bit/s. We may have $r_i = \infty$ in which case type-i flows have no rate limit. We denote by λ_i the arrival rate of type-i flows, by σ_i their mean size in bits, by $\beta_i = \lambda_i \sigma_i$ their traffic intensity in bit/s and by $\rho_{ij} = \beta_i a_{ij}/C$ their contribution to the resource-j load. The total resource-j load is given by $\rho_j = \sum_{i=1}^{N} \rho_{ij}$.

The system state is described by the line vector $x = (x_1, \ldots, x_N)$ of the number of ongoing flows of each type. Let $\phi_i(x)$ be the total throughput of type-i flows in state x. Denoting by $\phi(x)$ the line vector $(\phi_1(x), \ldots, \phi_N(x))$, by A the matrix with i,j-entry a_{ij} and by C the line vector $(C_1, \ldots, C_J)$, the throughput constraints are:

$$\phi(x)A \leq C \quad \text{and} \quad \phi_i(x) \leq x_i r_i, \quad i = 1, \ldots, N, \tag{17}$$

where the first inequality is component-wise. Note that for a single resource, that is $J = 1$, the multi-rate model corresponds to the matrix $A = (1, \ldots, 1)'$ while the multi-need model corresponds to the case $r_i = \infty$ for all $i = 1, \ldots, N$.

Balanced fair sharing is defined by

$$\phi_i(x) = \frac{\Phi(x - e_i)}{\Phi(x)},$$

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

$$\Phi(x) = \max\left\{\max_j \frac{1}{C_j}\sum_{i=1}^N a_{ij}\Phi(x - e_i), \max_{i:x_i>0} \frac{\Phi(x - e_i)}{x_i r_i}\right\}, \tag{18}$$

with $\Phi(x) = 0$ if $x \notin \mathbb{N}^N$. The capacity constraints (17) are satisfied. Moreover, at least one of these capacity constraints is attained in all states x.

The system state has the stationary distribution:

$$\pi(x) = \pi(0)\Phi(x)\prod_{i=1}^N \beta_i^{x_i},$$

under the stability condition $\rho_j < 1$ for all $j = 1, \dots, J$. The mean per-bit delay follows from Little's law. Its evaluation turns out to be computationally expensive, however. Like in the case of circuit traffic, the following bound allows one to decouple the system and evaluate the contribution of each resource independently [5]:

$$\tau_i \le \delta_i + \sum_{j=1}^J \frac{a_{ij}}{C_j}\frac{\rho_j}{1 - \rho_j},$$

where $\delta_i = \max\{\max_j \frac{a_{ij}}{C_j}, \frac{1}{r_i}\}$ is the minimum per-bit delay of type-i flows. In practice, one may use the looser bound:

$$\tau_i \le \frac{1}{r_i} + \sum_{j=1}^J \frac{a_{ij}}{C_j(1 - \rho_j)}.$$

This is the analogue of the bound (4) derived for the blocking probability in circuit-switched networks: in view of (16), the j-th term of the sum is equal to the mean per-bit delay of type-i flows when the network reduces to resource j (i.e. there is no other resource constraint).

Like in the case of circuit traffic, the model extends to a finite number of sources. Denote by K_i the number of type-i sources and by β_i the traffic intensity per type-i idle source (i.e. the ratio of the mean flow size to the mean idle period duration of type-i sources, cf. §3.2). The stationary distribution of the system state becomes:

$$\pi(x) = \pi(0)\Phi(x)\prod_{i=1}^N \frac{K_i!}{(K_i - x_i)!}\beta_i^{x_i}, \quad x_1 \le K_1, \dots, x_N \le K_N.$$

The previous model with Poisson flow arrivals corresponds to the limiting case $K_1 \to \infty, \dots, K_N \to \infty$ for fixed total traffic intensities $K_1\beta_1, \dots, K_N\beta_N$.

For both finite and infinite source models, one may apply admission control to guarantee a minimum throughput u_i to each type-i flow in progress. In the specific case $u_i = r_i = 1$ for all $i = 1, \dots, N$, traffic becomes inelastic and the model reduces to that described in §2.4 for circuit traffic, with traffic intensity $\alpha_i = \beta_i$ for type-i calls. The corresponding balance function is simply given by:

$$\Phi(x) = \frac{1}{x_1!} \cdots \frac{1}{x_N!}.$$

4 Insensitivity results

In this section, we show that all traffic models considered so far, that are covered by the general model of §3.6, correspond to Kelly-Whittle networks and have the insensitivity property.

4.1 State-dependent service rates

Consider a network of N processor-sharing queues with state-dependent service rates. Customers arrive in queue i according to a Poisson process of intensity λ_i, require independent, exponentially distributed services with mean σ_i (in service units, say bits) and leave the network once served. We denote by $\beta_i = \lambda_i \sigma_i$ the traffic intensity at queue i (in bit/s). The network state is described by the line vector $x = (x_1, \dots, x_N)$ of the number of customers in each queue. We denote by $\phi_i(x)$ the service rate of queue i in state x (in bit/s), with $\phi_i(x) = 0$ if $x_i = 0$. Note that this service rate does not only depend on x_i, the number of customers in queue i, but on the whole network state x: the N queues are coupled through their service rates.

The network state is a Markov process with transition rates λ_i from state x to state $x + e_i$ and $\phi_i(x)/\sigma_i$ from state x to state $x - e_i$. Assume that:

$$\forall i = 1, \dots, N, \quad \phi_i(x) = \frac{\Phi(x - e_i)}{\Phi(x)}, \tag{19}$$

for some positive function Φ on $\mathbb{N}^N$ such that $\Phi(0) = 1$ and $\Phi(x) = 0$ if $x \notin \mathbb{N}^N$. The network state is then a *reversible* Markov process whose stationary distribution is given by:

$$\pi(x) = \pi(0)\Phi(x) \prod_{i=1}^{N} \beta_i^{x_i}, \tag{20}$$

under the stability condition:

$$\sum_{x \in \mathbb{N}^N} \Phi(x) \prod_{i=1}^{N} \beta_i^{x_i} < \infty.$$

The local balance equations follow from (19):

$$\pi(x - e_i)\lambda_i = \pi(x)\frac{\phi_i(x)}{\sigma_i}.$$

The queuing network is a Kelly-Whittle network, whose stationary distribution is known to be independent of the distribution of service requirements beyond the mean for the processor-sharing service discipline [7, 18, 20, 26, 30]. In particular, the elastic traffic models of Section 3 with Poisson flow arrivals are insensitive to the flow size distribution beyond the mean.

As mentioned in Section 1, it is in fact not necessary that flows arrive as a Poisson process. In practice, each user typically generates a sequence of flows during the same activity period. We refer to a *session* as the sequence of flows generated by the same user, with an assumed idle period of random duration between the end of a flow and the beginning of the following flow. It is then sufficient to assume that *sessions* arrive as a Poisson process: the number of flows of a session, the successive flow sizes and idle period durations of this session may be arbitrarily distributed and correlated. The stationary distribution of the system state and the mean flow duration are insensitive to all traffic characteristics beyond the traffic intensity.

This insensitivity property comes again from that of Kelly-Whittle networks. Each session may be represented as the alternating visits of a customer to one of the N queues and to an additional infinite-server queue representing the idle periods. The associated queuing network is still a Kelly-Whittle network, whose stationary distribution is independent of the customer routes beyond the arrival rate at each queue [20]. There may be an arbitrary, predefined set of fixed routes, with arbitrary distributions and correlation of successive service requirements, allowing one to represent virtually any traffic characteristics.

Similarly, traffic models with a finite number of sources correspond to *closed* queuing networks. Assume K_i customers visit alternately queue i and an infinite-server queue with exponentially distributed service times with mean $1/\nu_i$. We refer to $\beta_i = \nu_i \sigma_i$ as the traffic intensity per idle source at queue i (in bit/s). The network state is again a *reversible* Markov process whose stationary distribution is given by:

$$\pi(x) = \pi(0)\Phi(x)\prod_{i=1}^{N}\frac{K_i!}{(K_i - x_i)!}\beta_i^{x_i}, \quad x_1 \le K_1, \ldots, x_N \le K_N.$$

The local balance equations follow from (19):

$$\pi(x - e_i)\nu_i(K_i - x_i + 1) = \pi(x)\frac{\phi_i(x)}{\sigma_i}.$$

The queuing network is a closed Kelly-Whittle network, whose stationary distribution is independent of the distribution of service requirements beyond the mean [7, 18, 20, 26]. The successive service requirements of a customer may even be correlated. Thus finite-source models are insensitive to all traffic characteristics beyond the traffic intensity of each source.

It is worth noting that Kelly-Whittle networks are the only networks of processor-sharing queues that satisfy the insensitivity property [7]. Thus it is essential that network resources are shared according to balanced fairness.

For other common resource allocations like max-min fairness and proportional fairness, the balance property (19) is violated in most cases and performance is sensitive to traffic characteristics like the flow size distribution [6]. Some structural properties of proportional fairness make its performance close to that of balanced fairness, however, and thus approximately insensitive [23].

4.2 State-dependent arrival rates

Previous results apply to traffic models without admission control. In the presence of admission control, the arrival rate at each queue is a function of the network state. Assume customers arrive in queue i according to a Poisson process of intensity $\lambda_i(x)$ in state x, require independent, exponentially distributed services with mean σ_i and leave the network once served. Denote by $\mathcal{X}$ the set of admissible states and assume $\lambda_i(x)$ is equal to some constant λ_i if $x + e_i \in \mathcal{X}$ and to 0 otherwise. The stationary distribution of the network state is still given by (20) on the state space $\mathcal{X}$ and is independent of the distribution of service requirements beyond the mean [7, 18, 20, 26]. In particular, the circuit traffic models of Section 2 with Poisson call arrivals are insensitive to the call duration distribution beyond the mean.

Again, the insensitivity property extends to the flow arrival process provided *sessions* arrive as a Poisson process. It is sufficient to assume that the session goes on in case of blocking, as if the blocked flow were completed instantaneously, to preserve the stationary distribution of the network state [9, 18]. Other models that allow one to represent random retrials in case of blocking are described in [4]. Key performance indicators like blocking probability and mean delay are the same for all flows of the same type, independently of the session they belong to and of their position in this session (e.g. first, second or last flow of the session). In particular, the circuit traffic models of Section 2 with Poisson call arrivals are insensitive to all traffic characteristics beyond the traffic intensity of each type of call.

Finally, these insensitivity properties extend to finite-source models as in the absence of admission control. The corresponding queuing network is a closed Kelly-Whittle network with state-dependent routing probabilities. We conclude that all traffic models of Sections 2 and 3 are insensitive.

5 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 average demand only, and not on the distribution of call holding times that has been evolving over the years. We believe the insensitive traffic models described in the present paper are useful for sizing communication networks and could serve as guidelines for the design of new traffic control mechanisms in next generation networks.

References

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

[2] A.W. Berger, Y. Kogan, Dimensioning bandwidth for elastic traffic in high-speed data networks, IEEE/ACM Trans. on Networking 8-5 (2000) 643–654.

[3] D. Bertsekas and R. Gallager, *Data Networks*, Prentice Hall, 1987.

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

[5] T. Bonald, Throughput performance in networks with linear capacity constraints, in: *Proc. of CISS*, 2006.

[6] T. Bonald, L. Massoulié, A. Proutière, J. Virtamo, A queueing analysis of max-min fairness, proportional fairness and balanced fairness, Queueing Systems 53 (2006) 65–84.

[7] T. Bonald and A. Proutière, Insensitivity in processor-sharing networks, Performance Evaluation 49 (2002) 193–209.

[8] T. Bonald and J. Virtamo, A recursive formula for multi-rate systems with elastic traffic, IEEE Communications Letters 9 (2005) 753–755.

[9] N.M. van Dijk, On Jackson's product form with "jump-over" blocking, Operations Research Letters 7 (1988) 233–235.

[10] L.E.N. Delbrouck, On the steady state distribution in a service facility with different peakedness factors and capacity requirements, IEEE Transactions on Communications 11 (1983) 1209–1211.

[11] Z. Dziong, J.W. Roberts, Congestion probabilities in a circuit-switched integrated services network, Performance Evaluation 7-4 (1987) 267–284.

[12] J.W. Cohen, The Generalized Engset Formula, Phillips Telecommunications Review 18 (1957) 158–170.

[13] 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.

[14] O. Enomoto, H. Miyamoto, An Analysis of mixtures of multiple bandwidth traffic on time division switching networks, in: *Proc. of the 7th International Teletraffic Congress*, 1973.

[15] A.K. Erlang, Solution of some problems in the theory of probabilities of significance 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.

[16] L.A. Gimpelson, Analysis of mixtures of wide and narrow-band traffic, IEEE Trans. Comm. Technology 13-3 (1965) 258–266.

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

[18] A. Hordijk and N. van Dijk, Adjoint processes, job local balance and insensitivity of stochastic networks, Bull:44 session Int. Stat. Inst. 50 (1982) 776–788.

[19] J.S. Kaufman, Blocking in a shared resource environment, IEEE Trans. Commun. 29 (1981) 1474–1481.

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

[21] F.P. Kelly, A. Maulloo and D. Tan, Rate control for communication networks: Shadow prices, proportional fairness and stability, Journal of the Operat. Res. Society 49 (1998).

[22] F.P. Kelly, Loss networks, Annals of Applied Probability 1 (1991) 319–378.

[23] L. Massoulié, Structural properties of proportional fairness: Stability and insensitivity, Annals of Applied Probability 17-3 (2007) 809–839.

[24] L. Massoulié and J.W. Roberts, Bandwidth sharing and admission control for elastic traffic, Telecommunication Systems 15 (2000) 185–201.

[25] 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.

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

[27] B.A. Sevastyanov, An ergodic theorem for Markov processes and its application to telephone systems with refusals, Theor. Probability Appl. 2 (1957) 104–112.

[28] W.R. Stevens, *TCP/IP Illustrated, Volume 1: The Protocols*, Addison-Wesley, 1994.

[29] W. Whitt, Blocking when service is required from several facilities simultaneously, AT&T Technical Journal 64-8 (1985) 1807–1856.

[30] P. Whittle, Partial balance and insensitivity, Journal of Applied Probability 22 (1985) 168–176.