# Insensitive Load Balancing

Thomas Bonald, Matthieu Jonckheere, Alexandre Proutière

# Insensitive Load Balancing

T. Bonald, M. Jonckheere and A. Proutière
France Telecom R&D
38-40 rue du Général Leclerc
92794 Issy-les-Moulineaux, France
{thomas.bonald,matthieu.jonckheere,alexandre.proutiere}@francetelecom.com

## ABSTRACT

A large variety of communication systems, including telephone and data networks, can be represented by so-called Whittle networks. The stationary distribution of these networks is insensitive, depending on the service requirements at each node through their mean only. These models are of considerable practical interest as derived engineering rules are robust to the evolution of traffic characteristics. In this paper we relax the usual assumption of static routing and address the issue of dynamic load balancing. Specifically, we identify the class of load balancing policies which preserve insensitivity and characterize optimal strategies in some specific cases. Analytical results are illustrated numerically on a number of toy network examples.

## Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics

## General Terms

Algorithms, Performance.

## Keywords

Load balancing, Insensitivity, Whittle networks.

## 1. INTRODUCTION

Load balancing is a key component of computer systems and communication networks. It consists in routing new demands (e.g., jobs, database requests, telephone calls, data transfers) to service entities (e.g., processors, servers, routes in telephone and data networks) so as to ensure the efficient utilization of available resources. One may distinguish different types of load balancing, depending on the level of information used in the routing decision:

**Static load balancing.** In this case, the routing decision is "blind" in the sense that it does not depend on the system state. This is a basic scheme which is widely used in practice but whose performance suffers from the lack of information on the current distribution of system load.

**Semi-static load balancing.** The routing decision is still blind but may depend on the period of the day [18]. Such a scheme may be useful when traffic has a well-know dayly profile for instance. Like the static scheme, it is unable to react to sudden traffic surges for some service entities.

**Dynamic load balancing.** In this case, routing depends on the system state. Such a scheme is much more efficient as it instantaneously adapts the routing decisions to the current load distribution among service entities.

Designing "optimal" load balancing schemes is a key issue. While this reduces to an optimization problem for static schemes [6, 10, 11, 24], the issue is much more complex for dynamic schemes [23]. The basic model consists of a set of parallel servers fed by a single stream of customers. Most existing results are based on the assumption of i.i.d. exponential service requirements and equal service rates. For infinite buffer sizes, Winston proved that joining the shortest queue is optimal in terms of the transient number of completed jobs, provided customers arrive as a Poisson process [27]. This result was extended by Ephremides et. al. [7] to non-Poisson arrivals. For finite buffer sizes, Hordjik and Koole [9] and Towsley et. al. [25] proved that joining the shortest (non-full) queue minimizes the blocking rate. Extending these results to non-exponential service requirements or unequal service rates proves extremely difficult. Whitt gave a number of counter-examples showing that joining the shortest queue is not optimal in certain cases [26]. Harchol-Balter et. al. studied the impact of the job size distribution for those load balancing schemes where the routing decision may be based on the amount of service required by each customer [8].

Identifying the optimal load balancing becomes even more complex in the presence of several customer classes. The simplest multiclass system consists of two parallel servers, each receiving a background customer stream. As a particular case of so-called "V2-symmetric" networks, Towsley

*SIGMETRICS/Performance'04,* June 12–16, 2004, New York, NY, USA.
.

et. al. proved that joining the shortest queue is optimal provided service times are i.i.d. exponential with the same mean at each server and all arrival streams have the same intensity [25]. Alanyali and Hajek considered this system for telephone traffic, i.e., the servers have a given number of available circuits and each service corresponds to the occupation of a circuit during the telephone call [1]. They proved under very general assumptions that in heavy traffic, joining the server with the highest number of available circuits minimizes the call blocking rate. Extending this result to any traffic load seems impossible in view of the results by van Leeuwaarden et. al. [16]. They gave an algorithm for evaluating the optimal policy with i.i.d. exponential call durations. Again, the optimal load balancing can be characterized in symmetric conditions only, in which case it boils down to joining the server with the highest number of available circuits. For more complex multiclass systems like communications networks, it is even more difficult to characterize the optimal solution. For circuit-switched networks, Kelly highlighted the potential inefficiency of certain dynamic load balancing at high loads, due to the fact that most calls may take overflow routes and therefore consume much more resources than necessary [14, 11, 12]. So-called "trunk reservation" was proposed as a means for overcoming this problem [12, 17]. More recently, similar phenomena have been observed in the context of data networks [18, 20].

Thus it seems extremely difficult if not impossible to characterize optimal load balancing schemes. In addition, the resulting performance cannot be explicitly evaluated in general. Instead of restricting the analysis to a specific distribution of service requirements (e.g., exponential), we here consider the class of *insensitive* load balancings, whose performance depends on this distribution through its mean only. Specifically, we consider the class of Whittle queueing networks, which can represent a large variety of computer systems and communication networks and whose stationary distribution is known to be insensitive under the usual assumption of static routing [21]. We identify the class of load balancing policies which preserve insensitivity and characterize optimal "decentralized" strategies, i.e., which depend on local information only. The resulting performance can be explicitly evaluated.

The model is described in the next section. In the following two sections, optimal load balancing schemes are characterized in the case of a single class and several customer classes, respectively. Section 5 is devoted to examples. Section 6 concludes the paper.

# 2. MODEL

We consider a network of $N$ processor sharing nodes. The service rate of node $i$ is a function $\phi_i$ of the network state $x = (x_1, \ldots, x_N)$, where $x_i$ denotes the number of customers present at node $i$. Required services at each node are i.i.d. exponential of unit mean. As shown in §2.5 below, this queueing system can represent a rich class of communication networks. We first present the notion of load balancing in this context, the key relation between the balance property and the insensitivity property, and the performance objectives.

*Notation.* Let $\mathcal{N} \equiv \mathbb{N}^N$. For $i = 1, \ldots, N$, we denote by $e_i \in \mathcal{N}$ the unit vector with 1 in component $i$ and 0 elsewhere. For $x, y \in \mathcal{N}$, we write $x \leq y$ if $x_i \leq y_i$ for all $i$. We use the notation:

$$|x| \equiv \sum_{i=1}^{N} x_i \quad \text{and} \quad \binom{|x|}{x} \equiv \frac{|x|!}{x_1! \ldots x_N!}.$$

We denote by $\mathcal{F}$ the set of $\mathbb{R}_+$-valued functions on $\mathcal{N}$.

## 2.1 Load balancing

We consider $K$ customer classes. Class-$k$ customers arrive as a Poisson process of intensity $\nu_k$ and require a service at one node $i \in \mathcal{I}_k$ before leaving the network, where $\mathcal{I}_1, \ldots, \mathcal{I}_K$ form a partition of the set of nodes $\{1, \ldots, N\}$. We denote by $\nu = \sum_{k=1}^{K} \nu_k$ the overall arrival intensity. A class-$k$ customer is routed to node $i \in \mathcal{I}_k$ with probability $p_i(x)$ in state $x$, and "blocked" with probability $1 - \sum_{i \in \mathcal{I}_k} p_i(x)$, in which case she/he leaves the network without being served. Let $\lambda_i(x) = p_i(x)\nu_k$ be the arrival rate at node $i$ in state $x$. We have:

$$\sum_{i \in \mathcal{I}_k} \lambda_i(x) \leq \nu_k. \tag{1}$$

We assume that the network capacity is finite in the sense that there exists a finite set $\mathcal{Y} \subset \mathcal{N}$ such that if $x \in \mathcal{Y}$ then $y \in \mathcal{Y}$ for all $y \leq x$ and:

$$\lambda_i(x) = 0, \quad \forall x \in \mathcal{Y},\ x + e_i \notin \mathcal{Y}. \tag{2}$$

Thus $\mathcal{Y}$ defines the set of attainable states and we let:

$$\lambda_i(x) = 0, \quad \forall x \notin \mathcal{Y}. \tag{3}$$

Any state-dependent arrival rates that satisfy (1), (2) and (3) determine an "admissible" load balancing.

## 2.2 Balance property

*Service rates.* We say that the service rates are balanced if for all $i, j$:

$$\phi_i(x)\phi_j(x - e_i) = \phi_j(x)\phi_i(x - e_j), \quad \forall x :\ x_i > 0,\ x_j > 0.$$

This property defines the class of so-called Whittle networks, an extension of Jackson networks where the service rate of a node may depend on the overall network state [21]. For Poisson arrivals at each node, the balance property is equivalent to the reversibility of the underlying Markov process. We assume that $\phi_i(x) > 0$ in all $x$ such that $x_i > 0$. Let $\Phi$ be the function recursively defined by $\Phi(0) = 1$ and:

$$\Phi(x) = \frac{\Phi(x - e_i)}{\phi_i(x)}, \quad x_i > 0. \tag{4}$$

The balance property implies that $\Phi$ is uniquely defined. We refer to $\Phi$ as the balance function of the service rates. Note that if there is a function $\Phi$ such that the service rates satisfy (4), these service rates are necessarily balanced.

For any $x \in \mathcal{N}$, $\Phi(x)$ may be viewed as the weight of any direct path from state $x$ to state 0, where a direct path is a set of consecutive states $x(0) \equiv x, x(1), x(2), \ldots, x(n) \equiv 0$ such that $x(m) = x(m-1) - e_{i(m)}$ for some $i(m)$, $m = 1, \ldots, n$, with $n = |x|$, and the weight of such a path is

the inverse of the product of $\phi_{i(m)}(x(m))$ for $m = 1, \ldots, n$ (refer to Figure 1). As will become clear in §2.3 below, the balance function plays a key role in the study of Whittle networks.

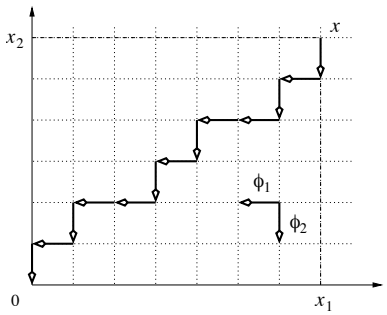


**Figure 1: The balance function $\Phi(x)$ is equal to the weight of any direct path from state $x$ to state 0.**

*Arrival rates.* A similar property may be defined for the arrival rates. We say that the arrival rates are balanced if for all $i, j$:

$$\lambda_i(x)\lambda_j(x+e_i) = \lambda_j(x)\lambda_i(x+e_j), \quad \forall x \in \mathcal{N}.$$

Let $\Lambda$ be the function recursively defined by $\Lambda(0) = 1$ and:

$$\Lambda(x+e_i) = \lambda_i(x)\Lambda(x). \tag{5}$$

The balance property implies that $\Lambda$ is uniquely defined. We refer to $\Lambda$ as the balance function of the arrival rates. Again, if there is a function $\Lambda$ such that the arrival rates satisfy (5), these arrival rates are necessarily balanced. For any $x \in \mathcal{N}$, $\Lambda(x)$ may be viewed as the weight of any direct path $x(n), x(n-1), \ldots, x(0)$ from state 0 to state $x$, defined as the product of $\lambda_{i(m)}(x(m-1))$ for $m = 1, \ldots, n$. In particular, the fact that $\Lambda(x) > 0$ implies that $\Lambda(y) > 0$ for all $y \leq x$. We define:

$$\mathcal{X} = \{x \in \mathcal{N} : \ \Lambda(x) > 0\}. \tag{6}$$

In view of (1), (2), (3) and (5), we deduce that:

$$\mathcal{X} \subset \mathcal{Y} \tag{7}$$

and

$$\sum_{i \in \mathcal{I}_k} \Lambda(x+e_i) \leq \nu_k \Lambda(x), \quad x \in \mathcal{N}. \tag{8}$$

We refer to $\mathcal{A}$ as the set of "admissible" functions $\Lambda \in \mathcal{F}$ for which properties (7) and (8) are satisfied.

## 2.3 Insensitivity property

*Static load balancing.* Consider a static load balancing where the arrival rates $\lambda_i(x)$ do not depend on the network state $x$ (within the network capacity region defined by $\mathcal{Y}$). If the service rates are balanced, the stochastic process that describes the evolution of the network state $x$ is an irreducible Markov process on the state space $\mathcal{Y}$, with stationary distribution:

$$\pi(x) = \pi(0) \times \Phi(x) \prod_{i=1}^{N} \lambda_i^{x_i}, \quad x \in \mathcal{Y}, \tag{9}$$

where $\pi(0)$ is given by the usual normalizing condition and $\Phi$ is the balance function defined by (4). In addition, the system is insensitive in the sense that the stationary distribution $\pi$ depends on the distribution of required services at any node through the mean only [21]. It has recently been shown that the balance property is in fact necessary for the system to be insensitive [2]. In the rest of the paper, we always assume that the service rates are balanced.

*Dynamic load balancing.* We now consider a dynamic load balancing where the arrival rates $\lambda_i(x)$ depend on the network state $x$. A sufficient condition for the insensitivity property to be retained is that the arrival rates are balanced, in which case the stochastic process that describes the evolution of the network state $x$ is an irreducible Markov process on the state space $\mathcal{X}$ defined by (6), with stationary distribution:

$$\pi(x) = \pi(0) \times \Phi(x)\Lambda(x), \quad x \in \mathcal{X}, \tag{10}$$

where $\pi(0)$ is given by the usual normalizing condition. Again, the balance property is in fact necessary for the system to be insensitive [2]. The class of insensitive load balancings is thus simply characterized by the set of admissible balance functions $\mathcal{A}$ defined above.

## 2.4 Performance objectives

Our aim is to characterize insensitive load balancings that are optimal in terms of blocking probability. For a given class $k$, the blocking probability is given by:

$$p_k = \sum_{x \in \mathcal{X}} \pi(x) \left(1 - \sum_{i \in \mathcal{I}_k} \frac{\lambda_i(x)}{\nu_k}\right).$$

The objective is to minimize either the maximum per-class blocking probability $\max_k p_k$ or the overall blocking probability, given by:

$$p = \sum_{k=1}^{K} \frac{\nu_k}{\nu} p_k = \sum_{x \in \mathcal{X}} \pi(x) \left(1 - \sum_{i=1}^{N} \frac{\lambda_i(x)}{\nu}\right).$$

It is worth noting that the balance function $\Lambda$, which gives the routing probability $p_i(x)$ in each state $x$, also determines the state space $\mathcal{X}$. In general, the state of actually attainable states $\mathcal{X}$ associated with the optimal solution is strictly included in the set of potentially attainable states $\mathcal{Y}$ defining the network capacity region.

## 2.5 Application to communication networks

The considered model is sufficiently generic to represent a rich class of computer systems and communication networks. The basic example is a set of parallel servers as mentioned in Section 1. We use this toy example as a reference system in Section 5 to assess the performance of insensitive load balancing strategies. The model can be used to represent much more complex systems, however.

*Circuit switched networks.* Consider for instance a circuit switched network composed of $L$ links with respective capacities $C_1, \ldots, C_L$ and shared by $K$ user classes. Class-$k$ users arrive at rate $\nu_k$ and require a circuit of capacity $a_k$ for a random duration of mean $1/\mu_k$ through one of the routes $r_i$, $i \in \mathcal{I}_k$, where each route $r_i$ consists of a subset of the links $\{1, \ldots, L\}$. This defines $N$ types of users with $\mathcal{I}_1 \cup \ldots \cup \mathcal{I}_K = \{1, \ldots, N\}$. Such a circuit switched network can be represented by the above queueing system where each node $i$ corresponds to type-$i$ users. If $i \in \mathcal{I}_k$, this corresponds to users that occupy a circuit of capacity $a_k$ along route $r_i$ during a random duration of mean $1/\mu_k$. Specifically, the service rate of node $i$ is given by:

$$\phi_i(x) = x_i a_k \mu_k, \quad \text{for } i \in \mathcal{I}_k.$$

Thus the service rates are balanced with corresponding balance function:

$$\Phi(x) = \prod_{k=1}^{K} \prod_{i \in \mathcal{I}_k} \frac{1}{x_i!\ a_k^{x_i} \mu_k^{x_i}}.$$

The network capacity is determined by the link capacities:

$$\mathcal{Y} = \left\{ x \in \mathcal{N} :\ \forall l,\ \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k :\ l \in r_i} x_i a_k \le C_l \right\}.$$

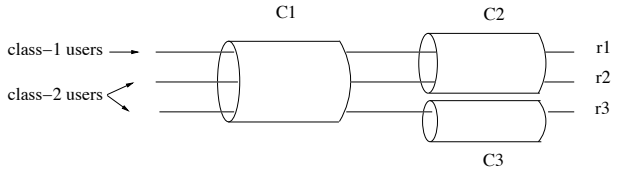


**Figure 2: A network with 2 user classes**

A typical example is a telephone network where $a_k = 1$ for all $k$ and the capacity $C_l$ corresponds to the number of available circuits on link $l$. The example depicted in Figure 2 consists of $K = 2$ user classes where class-1 users can take route $r_1 = \{1, 2\}$ only while class-2 users can take either route $r_2 = \{1, 2\}$ or route $r_3 = \{1, 3\}$. The network capacity is given by:

$$\mathcal{Y} = \{x \in \mathcal{N} :\ x_1 + x_2 + x_3 \le C_1,\ x_1 + x_2 \le C_2, x_3 \le C_3\}.$$

*Data networks.* We now consider a data network composed of $L$ links with respective capacities $C_1, \ldots, C_L$ and shared by $K$ user classes. Class-$k$ users arrive at rate $\nu_k$ and require the transfer of a document of random size of mean $1/\mu_k$ through one of the routes $r_i$, $i \in \mathcal{I}_k$. Again, this defines $N$ types of users with $\mathcal{I}_1 \cup \ldots \cup \mathcal{I}_K = \{1, \ldots, N\}$. The duration of a data transfer depends on its bit rate. We assume that the overall bit rate $\gamma_i(x)$ of type-$i$ users depends on the network state $x$ only and is equally shared between these users. Such a data network can be represented by the above queueing system where each node $i$ corresponds to type-$i$ users, that transfer a document through route $r_i$. The service rate of node $i$ is:

$$\phi_i(x) = \mu_k \gamma_i(x), \quad i \in \mathcal{I}_k.$$

The allocation must satisfy the capacity constraints:

$$\forall l, \quad \sum_{i:\ l \in r_i} \gamma_i(x) \le C_l.$$

The balanced allocation for which at least one capacity constraint is reached in any state is known as "balanced fairness" [3]. We have:

$$\gamma_i(x) = \frac{\Gamma(x - e_i)}{\Gamma(x)}, \quad \forall x : x_i > 0,$$

where $\Gamma$ is the positive function recursively defined by $\Gamma(0) = 1$ and:

$$\Gamma(x) = \max_l \frac{1}{C_l} \sum_{i:\ l \in r_i, x_i > 0} \Gamma(x - e_i).$$

The balance function $\Phi$ which characterizes the service rates of the corresponding queueing network is then given by:

$$\Phi(x) = \Gamma(x) \prod_{k=1}^{K} \prod_{i \in \mathcal{I}_k} \frac{1}{\mu_k^{x_i}}.$$

The network capacity can be determined so as to guarantee a minimum data rate $\gamma$ for instance, in which case:

$$\mathcal{Y} = \left\{ x \in \mathcal{N} :\ \forall i,\ \frac{\gamma_i(x)}{x_i} \ge \gamma \right\}.$$

# 3. A SINGLE CLASS

We first consider the case $K = 1$. There is a single stream of incoming customers, that can be routed to any of the $N$ network nodes. We first characterize the class of admissible load balancings and then use this characterization to identify optimal strategies in terms of blocking probability.

## 3.1 Characterization

Denote by $\mathcal{S} \subset \mathcal{A}$ the set of balance functions that correspond to "simple" load balancings in the sense that customers can be blocked in a single state. We denote by $\Lambda_y$ the element of $\mathcal{S}$ such that customers are blocked in state $y \in \mathcal{Y}$ only.

PROPOSITION 1. *We have:*

$$\Lambda_y(x) = \Lambda_y(y) \binom{|y - x|}{y - x} \frac{1}{\nu^{|y-x|}} \quad \text{if } x \le y,$$

*and $\Lambda_y(x) = 0$ otherwise. The constant $\Lambda_y(y)$ is determined by the normalizing condition $\Lambda_y(0) = 1$.*

PROOF. For any state $x \le y$, $x \ne y$, we have:

$$\Lambda_y(x) = \frac{1}{\nu} \sum_{i=1}^{N} \Lambda_y(x + e_i).$$

In particular, $\Lambda_y(x)$ is equal to the product of $\Lambda_y(y)/\nu^{|y-x|}$ by the number of direct paths from $x$ to $y$. □

The blocking probability associated with a simple load balancing can be easily evaluated using the following recursive formula:

PROPOSITION 2. *Let $1/\delta(y)$ be the blocking probability associated with the balance function $\Lambda_y \in \mathcal{S}$. We have:*

$$\delta(y) = 1 + \sum_{i=1}^{N} \frac{\phi_i(y)}{\nu} \delta(y - e_i),$$

*with $\delta(0) = 1$ and $\delta(y) = 0$ for any $y \notin \mathcal{N}$.*

PROOF. Using the identity:

$$\binom{|y-x|}{y-x} = \sum_{i:y_i>x_i} \binom{|y-x-e_i|}{y-x-e_i}, \quad x \leq y,$$

we deduce from (10) and Proposition 1 that:

$$\begin{aligned}
\delta(y) &= \frac{\sum_{x\leq y} \pi(x)}{\pi(y)} \\
&= \frac{1}{\Phi(y)} \sum_{x\leq y} \Phi(x) \binom{|y-x|}{y-x} \frac{1}{\nu^{|y-x|}} \\
&= 1 + \frac{1}{\Phi(y)} \sum_{x\leq y, x\neq y} \sum_{i:y_i>x_i} \Phi(x) \binom{|y-x-e_i|}{y-x-e_i} \frac{1}{\nu^{|y-x|}} \\
&= 1 + \sum_{i=1}^{N} \frac{\phi_i(y)}{\nu} \times \\
&\quad \frac{1}{\Phi(y-e_i)} \sum_{x\leq y-e_i} \Phi(x) \binom{|y-x-e_i|}{y-x-e_i} \frac{1}{\nu^{|y-x-e_i|}} \\
&= 1 + \sum_{i=1}^{N} \frac{\phi_i(y)}{\nu} \delta(y - e_i).
\end{aligned}$$

□

The following result characterizes the set of admissible balance functions $\mathcal{A}$ as linear combinations of elements of $\mathcal{S}$:

THEOREM 1. *For any balance function $\Lambda \in \mathcal{A}$, we have:*

$$\Lambda = \sum_{y\in\mathcal{Y}} \alpha(y) \Lambda_y, \tag{11}$$

*where for all $y \in \mathcal{Y}$,*

$$\alpha(y) = \frac{\beta(y)}{\Lambda_y(y)} \quad \textit{with} \quad \beta(y) = \Lambda(y) - \frac{1}{\nu} \sum_{i=1}^{N} \Lambda(y + e_i).$$

*Conversely, for any $\alpha \in \mathcal{F}$ such that $\sum_{y\in\mathcal{Y}} \alpha(y) = 1$, the balance function $\Lambda$ defined by (11) lies in $\mathcal{A}$.*

PROOF. In view of (8), there exists a function $\beta \in \mathcal{F}$ such that for any state $x$:

$$\Lambda(x) = \beta(x) + \frac{1}{\nu} \sum_{i=1}^{N} \Lambda(x + e_i). \tag{12}$$

As $\Lambda(x) = 0$ for any state $x \notin \mathcal{Y}$, we deduce that $\Lambda$ is in fact entirely determined by the function $\beta$ through (12). The proof of equality (11) then follows from the fact that the function $\sum_{y\in\mathcal{Y}} \alpha(y)\Lambda_y$ satisfies (12) in any state $x \in \mathcal{Y}$:

$$\begin{aligned}
\sum_{y\in\mathcal{Y}} \alpha(y)\Lambda_y(x) &= \alpha(x)\Lambda_x(x) + \sum_{\substack{y\in\mathcal{Y} \\ y\geq x, y\neq x}} \alpha(y)\Lambda_y(x), \\
&= \beta(x) + \sum_{\substack{y\in\mathcal{Y} \\ y\geq x, y\neq x}} \alpha(y) \frac{1}{\nu} \sum_{i=1}^{N} \Lambda_y(x + e_i), \\
&= \beta(x) + \frac{1}{\nu} \sum_{i=1}^{N} \sum_{y\in\mathcal{Y}} \alpha(y)\Lambda_y(x + e_i).
\end{aligned}$$

Conversely, any linear combination $\Lambda = \sum_{y\in\mathcal{Y}} \alpha(y)\Lambda_y$ of elements of $\mathcal{S}$ with $\sum_{y\in\mathcal{Y}} \alpha(y) = 1$ satisfies $\Lambda(0) = 1$, $\mathcal{X} \subset \mathcal{Y}$ as well as inequalities (8). □

## 3.2 Optimal load balancing

We deduce from Theorem 1 that there exists an optimal admissible load balancing which is simple. In particular, the state of actually attainable states $\mathcal{X}$ associated with this optimal solution is of the form $\{x \in \mathcal{N} : x \leq y\}$ and therefore generally smaller than the set of potentially attainable states $\mathcal{Y}$.

COROLLARY 1. *There is a balance function $\Lambda \in \mathcal{S}$ which minimizes the blocking probability over the set $\mathcal{A}$.*

PROOF. The blocking probability is given by:

$$p = \sum_{x\in\mathcal{X}} \pi(x) \left(1 - \sum_{i=1}^{N} \frac{\lambda_i(x)}{\nu}\right).$$

In view of (5) and (10), we deduce:

$$\begin{aligned}
p &= \frac{\sum_{x\in\mathcal{Y}} (1 - \frac{1}{\nu}\sum_{i=1}^{N} \lambda_i(x))\Lambda(x)\Phi(x)}{\sum_{x\in\mathcal{Y}} \Lambda(x)\Phi(x)} \\
&= \frac{\sum_{x\in\mathcal{Y}} (\Lambda(x) - \frac{1}{\nu}\sum_{i=1}^{N} \Lambda(x + e_i))\Phi(x)}{\sum_{x\in\mathcal{Y}} \Lambda(x)\Phi(x)}.
\end{aligned}$$

It then follows from Theorem 1 that

$$p = \frac{\sum_{y\in\mathcal{Y}} \beta(y)\Phi(y)}{\sum_{y\in\mathcal{Y}} \beta(y)\Psi(y)}$$

with

$$\Psi(y) = \sum_{x\in\mathcal{Y}} \frac{\Lambda_y(x)}{\Lambda_y(y)} \Phi(x), \quad y \in \mathcal{Y}.$$

In particular,

$$p \geq \min_{y\in\mathcal{Y}} \frac{\Phi(y)}{\Psi(y)}.$$

Let $y^\star$ be a state where the function $y \mapsto \Phi(y)/\Psi(y)$ is minimal. The blocking probability $p$ is minimal if $\beta(y) = 0$ for all $y \in \mathcal{Y}$ except $y^\star$. The corresponding balance function is $\Lambda_{y^\star} \in \mathcal{S}$. □

REMARK 1. *In view of Corollary 1 and Proposition 2, finding the optimal admissible load balancing requires $O(|\mathcal{Y}|)$ operations only, where $|\mathcal{Y}|$ denotes the number of elements in the set $\mathcal{Y}$.*

It is possible to further characterize the optimal load balancing when the network is "monotonic" in the sense that:

$$\phi_i(x) \geq \phi_i(x - e_j), \quad \forall i, j, \ \forall x : x_j > 0. \tag{13}$$

We say that a state $y \in \mathcal{Y}$ is extremal if $y + e_i \notin \mathcal{Y}$ for all $i = 1, \ldots, N$. The following result is a consequence of Proposition 2:

PROPOSITION 3. *Let $\Lambda_y \in \mathcal{S}$ be a balance function which minimizes the blocking probability over the set $\mathcal{A}$. If the network is monotonic, $y$ is an extremal state of $\mathcal{Y}$.*

PROOF. Let $1/\delta(y)$ be the blocking probability associated with the balance function $\Lambda_y \in \mathcal{S}$ if $y \in \mathcal{Y}$, and $\delta(y) = 0$ otherwise. We prove by induction on $|y|$ that:

$$\forall y \in \mathcal{Y}, \ \forall j, \quad \delta(y) \geq \delta(y - e_j).$$

The property holds for $y = 0$. Now assume that the property holds for all $y \in \mathcal{Y}$ such that $|y| = n$, for some integer $n$. Let $y \in \mathcal{Y}$ such that $|y| = n+1$. It follows from Proposition 2 and the monotonicity property that:

$$\begin{aligned} \forall j, \quad \delta(y) &= 1 + \sum_{i=1}^{N} \frac{\phi_i(y)}{\nu} \delta(y - e_i), \\ &\geq 1 + \sum_{i=1}^{N} \frac{\phi_i(y - e_j)}{\nu} \delta(y - e_i - e_j), \\ &= \delta(y - e_j). \end{aligned}$$

Thus $\delta(y)$ is maximal for an extremal state of $\mathcal{Y}$. □

# 4. SEVERAL CLASSES

When $K \geq 2$, admissible load balancings can still be written as linear combinations of simple load balancings as in Theorem 1 but with additional constraints that cannot be simply characterized. Thus we restrict the analysis to the class of so-called "decentralized" load balancings where the routing decision for a class-$k$ customer does not depend on the number of customers of other classes. This class presents the practical interest of requiring local information only, unlike the general class of admissible load balancings where the routing decisions are based on the overall network state.

## 4.1 Characterization

For any state $x \in \mathcal{N}$, we define the restricted state $x^{(k)} \equiv \sum_{i \in \mathcal{I}_k} x_i e_i$ giving the number of class-$k$ customers in each node $i \in \mathcal{I}_k$. Let $\mathcal{N}^{(k)} \equiv \{x^{(k)}, x \in \mathcal{N}\}$ be the corresponding state space. We define the class of decentralized load balancings as those for which the corresponding balance function has the following product-form:

$$\Lambda(x) = \prod_{k=1}^{K} \Lambda^{(k)}(x^{(k)}) \quad \text{if } x \in \mathcal{Y}, \quad \Lambda(x) = 0 \quad \text{otherwise},$$

where $\Lambda^{(k)}$ is the restriction of $\Lambda$ to the set $\mathcal{N}^{(k)}$. We define $\mathcal{D}$ as the set of balance functions $\Lambda \in \mathcal{A}$ having such a product-form. Note that the load balancing is decentralized in the sense that the routing probability of a class-$k$ customer to each node $i \in \mathcal{I}_k$ is independent of the number of customers of other classes:

$$\lambda_i(x - e_i) = \frac{\Lambda(x)}{\Lambda(x - e_i)} = \frac{\Lambda^{(k)}(x^{(k)})}{\Lambda^{(k)}(x^{(k)} - e_i)}, \quad \forall x \in \mathcal{Y} : \ x_i > 0.$$

We now extend the notion of "simple" load balancing defined in §3.1. Let $\mathcal{Y}^{(k)} = \{x^{(k)}, x \in \mathcal{Y}\}$ and:

$$\mathcal{Y}' = \{x \in \mathcal{N} : \ \forall k, \ x^{(k)} \in \mathcal{Y}^{(k)}\}.$$

Thus $\mathcal{Y}'$ is the set of states $x$ such that any restricted state $x^{(k)}$ belongs to $\mathcal{Y}$. Note that $\mathcal{Y}'$ contains $\mathcal{Y}$ and is equal to $\mathcal{Y}$ if $\mathcal{Y} = \{x \in \mathcal{N} : \ x \leq y\}$ for some state $y$. We now refer to $\mathcal{S}$ as the set of balance functions $\Lambda_y, y \in \mathcal{Y}'$, defined by:

$$\Lambda_y(x) = \prod_{k=1}^{K} \Lambda_{y^{(k)}}(x^{(k)}),$$

where $\Lambda_{y^{(k)}}(x^{(k)})$ is the simple load balancing for class-$k$ customers as defined in §3.1. The following result is the analog of Theorem 1.

THEOREM 2. *For any balance function $\Lambda \in \mathcal{D}$, we have:*

$$\Lambda^{(k)} = \sum_{y \in \mathcal{Y}^{(k)}} \alpha^{(k)}(y) \Lambda_y,$$

*where for all $y \in \mathcal{Y}^{(k)}$,*

$$\alpha^{(k)}(y) = \frac{\beta^{(k)}(y)}{\Lambda_y(y)} \quad \textit{with} \quad \beta^{(k)}(y) = \Lambda(y) - \frac{1}{\nu_k} \sum_{i \in \mathcal{I}_k} \Lambda(y + e_i).$$

*In particular, we have for any $x \in \mathcal{Y}$:*

$$\Lambda(x) = \sum_{y \in \mathcal{Y}'} \alpha(y) \Lambda_y(x),$$

*where for all $y \in \mathcal{Y}'$,*

$$\alpha(y) = \frac{\beta(y)}{\Lambda_y(y)} \quad \textit{with} \quad \beta(y) = \prod_{k=1}^{K} \beta^{(k)}(y^{(k)}).$$

PROOF. We obtain the expressions for $\Lambda^{(k)}$ as in the proof of Theorem 1. The proof then follows from the fact that for all $x \in \mathcal{Y}$:

$$\begin{aligned} \Lambda(x) &= \prod_{k=1}^{K} \Lambda^{(k)}(x^{(k)}) \\ &= \prod_{k=1}^{K} \sum_{y \in \mathcal{Y}^{(k)}} \frac{\beta^{(k)}(y)}{\Lambda_y(y)} \Lambda_y(x) \\ &= \sum_{y \in \mathcal{Y}'} \prod_{k=1}^{K} \frac{\beta^{(k)}(y^{(k)})}{\Lambda_{y^{(k)}}(y^{(k)})} \Lambda_{y^{(k)}}(x^{(k)}) \\ &= \sum_{y \in \mathcal{Y}'} \alpha(y) \Lambda_y(x). \end{aligned}$$

□

## 4.2 Optimal load balancing

As in the case of a single class, we deduce from Theorem 2 that there exists an optimal admissible load balancing which is simple, for both the overall blocking probability and the maximum per-class blocking probability objectives.

COROLLARY 2. *There is a balance function $\Lambda \in \mathcal{S}$ which minimizes the overall blocking probability over the set $\mathcal{D}$.*

PROOF. The overall blocking probability is given by:

$$p = \sum_{x \in \mathcal{X}} \pi(x) \left(1 - \sum_{i=1}^{N} \frac{\lambda_i(x)}{\nu}\right).$$

In view of (5) and (10), we deduce:

$$p = \left( \sum_{x \in \mathcal{Y}} \sum_{k=1}^{K} \frac{\nu_k}{\nu} (\Lambda^{(k)}(x^{(k)}) - \frac{1}{\nu_k} \sum_{i \in \mathcal{I}_k} \Lambda^{(k)}(x^{(k)} + e_i)) \right.$$
$$\left. \times \quad \prod_{l \neq k} \Lambda^{(l)}(x^{(l)}) \Phi(x) \right) / \left( \sum_{x \in \mathcal{Y}} \Lambda(x) \Phi(x) \right).$$

It then follows from Theorem 2 that

$$p = \frac{\sum_{y \in \mathcal{Y}'} \beta(y) \Phi'(y)}{\sum_{y \in \mathcal{Y}'} \beta(y) \Psi(y)},$$

with

$$\Phi'(y) = \sum_{x \in \mathcal{Y}} \sum_{k=1}^{K} \frac{\nu_k}{\nu} \times \prod_{l \neq k} \frac{\Lambda_{y^{(l)}}(x^{(l)})}{\Lambda_{y^{(l)}}(y^{(l)})} \times \Phi(y^{(k)} + \sum_{l \neq k} x^{(l)})$$

and

$$\Psi(y) = \sum_{x \in \mathcal{Y}} \frac{\Lambda_y(x)}{\Lambda_y(y)} \Phi(x), \quad y \in \mathcal{Y}'.$$

In particular,

$$p \geq \min_{y \in \mathcal{Y}'} \frac{\Phi'(y)}{\Psi(y)}.$$

We conclude that the blocking probability $p$ is minimal if $\beta(y) = 0$ for all $y \in \mathcal{Y}$ except for one state $y^\star$ where the function $y \mapsto \Phi'(y)/\Psi(y)$ is minimal. The corresponding balance function is $\Lambda_{y^\star} \in \mathcal{S}$. □

COROLLARY 3. *There is a balance function $\Lambda \in \mathcal{S}$ which minimizes the maximum per-class blocking probability over the set $\mathcal{D}$.*

PROOF. The blocking probability of class-$k$ customers is given by

$$p_k = \sum_{x \in \mathcal{X}} \pi(x) \left(1 - \sum_{i \in \mathcal{I}_k} \frac{\lambda_i(x)}{\nu_k}\right).$$

It then follows as in the proof of Corollary 2 that

$$\max_{k=1,\ldots,K} p_k = \frac{\sum_{y \in \mathcal{Y}'} \beta(y) \Phi''(y)}{\sum_{y \in \mathcal{Y}'} \beta(y) \Psi(y)},$$

with

$$\Phi''(y) = \max_{k=1,\ldots,K} \sum_{x \in \mathcal{Y}} \frac{1}{\nu_k} \times \prod_{l \neq k} \frac{\Lambda_{y^{(l)}}(x^{(l)})}{\Lambda_{y^{(l)}}(y^{(l)})} \times \Phi(y^{(k)} + \sum_{l \neq k} x^{(l)})$$

and

$$\Psi(y) = \sum_{x \in \mathcal{Y}} \frac{\Lambda_y(x)}{\Lambda_y(y)} \Phi(x), \quad y \in \mathcal{Y}'.$$

We conclude that the maximum per-class blocking probability is minimal if $\beta(y) = 0$ for all $y \in \mathcal{Y}$ except for one state $y^\star$ where the function $y \mapsto \Phi''(y)/\Psi(y)$ is minimal. The corresponding balance function is $\Lambda_{y^\star} \in \mathcal{S}$. □

# 5. EXAMPLES

We apply previous theoretical results to a reference system consisting of a set of parallel servers, with and without background traffic.

## 5.1 Absence of background traffic

We first consider $N$ parallel servers fed by a single stream of customers, as illustrated in Figure 3. Such a system may represent a supercomputer center, for instance, or any other distributed server system. For communication networks, it might correspond to a logical link split over several physical links. In this case, we consider data traffic only as, for telephone traffic, any policy which blocks a call only when all circuits are occupied is obviously optimal.

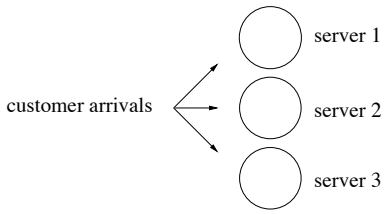


**Figure 3: The reference system: a set of parallel servers.**

The model is the same as that considered in [2]. Denote by $x = (x_1, \ldots, x_N)$ the network state, $C_1, \ldots, C_N$ the server capacities and $1/\mu$ the mean service requirement. We have:

$$\phi_i(x) = \mu C_i,$$

corresponding to the balance function:

$$\Phi(x) = \frac{1}{\prod_{i=1}^{N} (\mu C_i)^{x_i}}$$

The network capacity region, defined so as to guarantee a minimum service rate $\gamma$, may be written:

$$\mathcal{Y} = \{x : \forall i,\ x_i \leq C_i/\gamma\}.$$

The overall system load is given by:

$$\varrho = \frac{\nu}{\mu \sum_{l=1}^{N} C_i}.$$

The monotonicity property (13) trivially holds and it follows from Corollary 1 and Proposition 3 that the optimal insensitive load balancing is characterized by the balance function:

$$\Lambda(x) = \binom{|y-x|}{y-x} / \binom{|y|}{y} \times \nu^{|x|} \quad \text{if } x \leq y,$$

and $\Lambda(x) = 0$ otherwise, where $y$ is the vector such that $y_i$ is the largest integer smaller than $C_i/\gamma$. In view of (5), the corresponding arrival rates are:

$$\lambda_i(x) = \frac{y_i - x_i}{|y-x|}\nu, \quad x \leq y.$$

Viewing each server $i$ as a resource of $y_i$ potential "circuits" of rate $\gamma$, this boils down to routing new demands in proportion to the number of available circuits at each server.

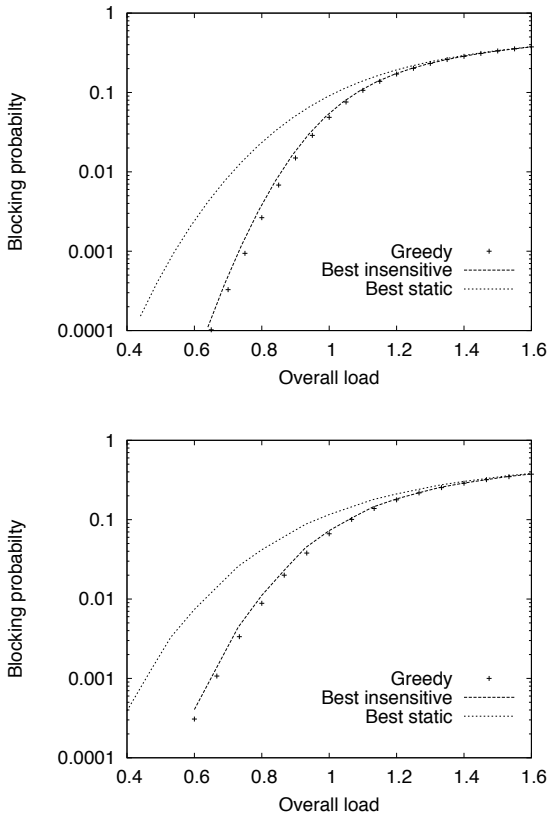


**Figure 4: Two parallel links with symmetric capacities (upper graph) and asymmetric capacities (lower graph).**

We compare the resulting blocking probability with that obtained for the best static load balancing and the greedy load balancing, respectively. We refer to the greedy strategy as that where users are routed at time $t$ to server $i(t)$ with the highest potential service rate :

$$i(t) = \arg \max_{i=1,\ldots,N} \frac{C_i}{x_i(t^-)+1}. \tag{14}$$

Note that for symmetric capacities, this is equivalent to joining the shortest queue. The greedy strategy is sensitive. Results are derived by simulation for i.i.d. exponential services.

Figure 4 gives the blocking probability for two servers of symmetric capacities ($C_1 = C_2 = 1$) and asymmetric capacities ($C_1 = 1$, $C_2 = 0.5$) with $\gamma = 0.1$. We observe that the best insensitive load balancing provides a tight approximation for the greedy strategy, which is known to be optimal for i.i.d. exponential services with the same mean (cf. Section 1). We verified by simulation that, for both symmetric and asymmetric scenarios, the performance of the greedy strategy is in fact only slighty sensitive to the service requirement distribution and therefore that the above approximation remains accurate under general traffic assumptions.

## 5.2 Presence of background traffic

We now consider a set of $N$ parallel servers as depicted in Figure 5, with a "flexible" stream corresponding to users that can be routed to either server, and one "background" stream per server corresponding to users whose route is fixed. This system with $N = 2$ servers is used in [16] for modeling a wireless telephone network with two overlapping cells: the flexible stream corresponds to those users in the overlapping area that can be served by either cell while each background stream corresponds to users that can be served by one cell only.

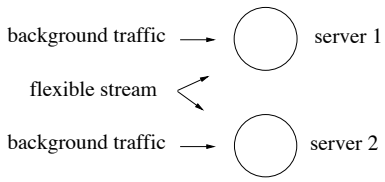


**Figure 5: Parallel servers with background traffic.**

Let $(x, x')$ be the network state, where $x = (x_1, \ldots, x_N)$ and $x' = (x'_1, \ldots, x'_N)$ describe the number of users of the flexible and background streams, respectively. We denote by $C_i$ the capacity of server $i$ and by $\phi_i(x, x')$ and $\phi'_i(x, x')$ the service rates of flexible and background users at server $i$, respectively. We here consider the two types of traffic described in §2.5:

- *Telephone traffic*, where each user requires a circuit of unit capacity. Denoting by $1/\mu$ the mean call duration for a user of the flexible stream, $1/\mu_i$ the mean call duration for a user of the background stream of server $i$, we get:

  $$\phi_i(x, x') = x_i\mu \text{ and } \phi'_i(x, x') = x'_i\mu_i, \text{ for all } i.$$

  This corresponds to the balance function:

  $$\Phi(x, x') = \frac{1}{\prod_{i=1}^N x_i!\ x'_i!\mu^{x_i}\mu_l^{x'_i}}.$$

  The network capacity region is defined by:

  $$\mathcal{Y} = \{(x, x') : \forall i,\ x_i + x'_i \leq C_i\}.$$

- *Data traffic*, where the capacity of each link is fairly shared between active users. Denoting by $1/\mu$ the mean document size for a user of the flexible stream, $1/\mu_i$ the mean document size for a user of the background stream of server $i$, we get for all $i$:

$$\phi_i(x, x') = \frac{x_i}{x_i + x'_i} C_i \mu$$

and

$$\phi'_i(x, x') = \frac{x'_i}{x_i + x'_i} C_i \mu_i.$$

This corresponds to the balance function:

$$\Phi(x, x') = \prod_{i=1}^{N} \binom{x_i + x'_i}{x_i} \frac{1}{C_i^{x_i + x'_i}} \frac{1}{\mu^{x_i} \mu_i^{x'_i}}.$$

The network capacity region is characterized by a minimum data rate $\gamma$ so that:

$$\mathcal{Y} = \{(x, x') : \forall i,\ x_i + x'_i \leq C_i/\gamma\}.$$

*Traffic parameters.* Let $\nu$ be the arrival rate of the flexible stream, $\nu_i$ the arrival rate of the background stream of server $i$. The overall traffic intensity is given by:

$$\rho = \frac{\nu}{\mu} + \sum_{i=1}^{N} \frac{\nu_i}{\mu_i}.$$

For telephone traffic, this is expressed in Erlangs and the capacity of each link corresponds to the number of available circuits. For data traffic, $\rho$ and the link capacities are expressed in bits/s. The overall system load is defined as:

$$\varrho = \frac{\rho}{\sum_{i=1}^{N} C_i}.$$

In the presence of background traffic, the optimal insensitive strategy depends on the traffic intensity. As shown in the following examples, it remains approximately the same for a large range of offered loads, however, indicating that a fixed strategy could be chosen in practice.

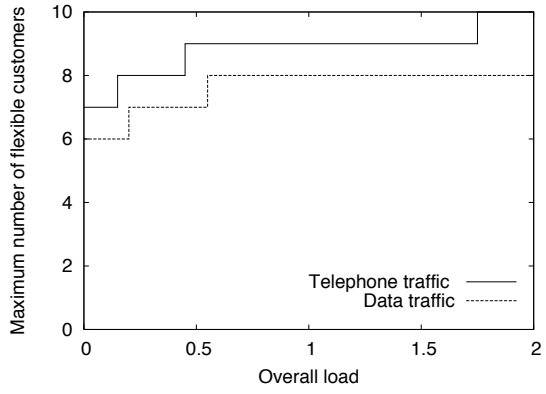


**Figure 6: The best decentralized load balancing with homogeneous service requirements.**

We consider $N = 2$ servers of the same capacity, $C_1 = C_2 = 10$. Each background stream represents 10% of the overall traffic intensity. The minimum data rate is $\gamma = 1$.

*Homogeneous service requirements.* We first consider the case of homogeneous service requirements, $\mu = \mu_1 = \mu_2$. In view of Corollary 2, there is a simple load balancing, characterized by a state $(y, y') \in \mathcal{Y}'$, that minimizes the overall blocking rate over all decentralized insensitive load balancings. It turns out that $y' = (10, 10)$ for all traffic loads, i.e., the corresponding balance function is:

$$\Lambda(x, x') = \binom{|y - x|}{y - x} / \binom{|y|}{y} \times \nu^{|x|} \nu_1^{x'_1} \nu_2^{x'_2} \quad \text{if } x \leq y,$$

and $\Lambda(x, x') = 0$ otherwise. Let $n$ be the maximum number of flexible users per link, i.e., $y = (n, n)$. Figure 6 shows how $n$ varies with respect to the system load $\varrho$ for both telephone and data traffic.

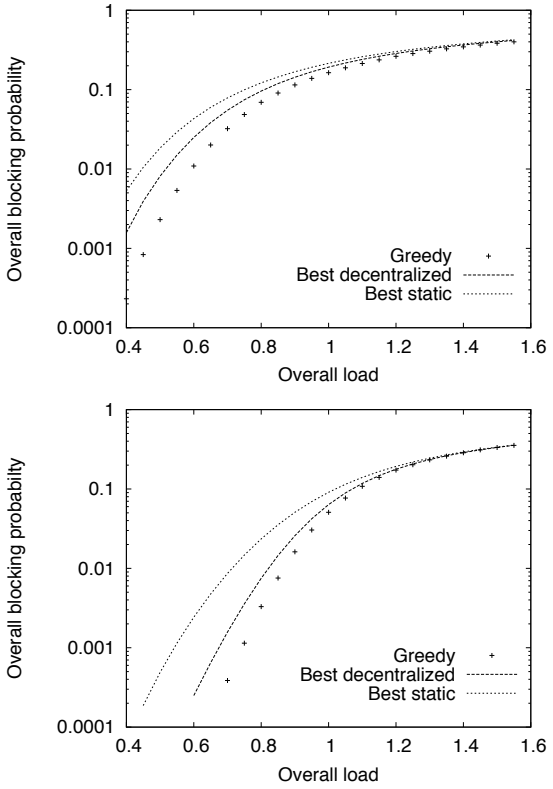


**Figure 7: Overall blocking rate for telephone traffic (upper graph) and data traffic (lower graph) with homogeneous service requirements.**

Figure 7 gives the resulting overall blocking probability. Results are compared with those of the best static load balancing and the greedy load balancing where a user is routed to that server where most resources are available.

Specifically, a user of the flexible stream arriving at time $t$ is routed to that server $i(t)$ with the largest number of available circuits for telephone traffic:

$$i(t) = \arg\max_{i=1,2} C_i - x_i(t^-) - x'_i(t^-),$$

with the highest potential rate for data traffic:

$$i(t) = \arg\max_{i=1,2} \frac{C_i}{x_i(t^-) + x'_i(t^-) + 1}.$$

Again, the greedy policy is sensitive and results are derived by simulation with i.i.d. exponential services in this case. We observe that the greedy load balancing outperforms any other strategy for both telephone and data traffic. This strategy requires a complete knowledge of the network state, however, unlike the other two strategies. Note that the static load balancing is decentralized and insensitive, therefore leads to a higher blocking rate than the best decentralized insensitive strategy.

*Heterogeneous service requirements*. We now consider the case of heterogeneous service requirements, namely $\mu = \mu_1/100 = \mu_2/100$. A strategy that minimizes the overall blocking probability would tend to block the flexible stream in view of the higher service requirements. Thus we rather consider the best decentralized load balancing in terms of the maximum per-class blocking rate. In view of Corollary 3, there is a simple load balancing that minimizes the maximum per-class blocking rate, characterized by a state $(y, y') \in \mathcal{Y}'$. Again, we have $y = (n, n)$ and $y' = (10, 10)$ for all traffic loads. Figure 8 shows how $n$ varies with respect to the system load $\varrho$ for both telephone and data traffic.

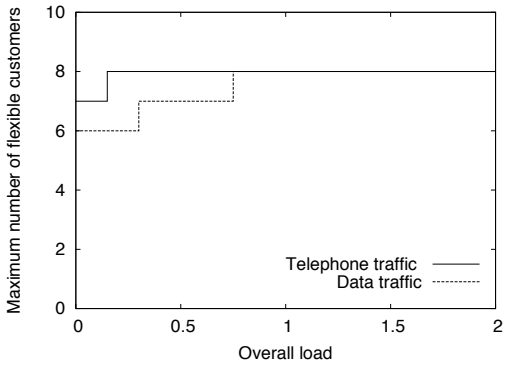


**Figure 8: The best decentralized load balancing with heterogeneous service requirements.**

Figure 9 gives the resulting maximum per-class blocking probability compared with that of the best static load balancing and the greedy load balancing. We observe that the greedy load balancing in no longer the best strategy, especially at high loads. For data traffic at load $\varrho = 1$ for instance, both the greedy strategy and the best static strategy lead to a blocking rate approximately equal to 10%, while the best insensitive strategy gives a blocking rate of 5%.

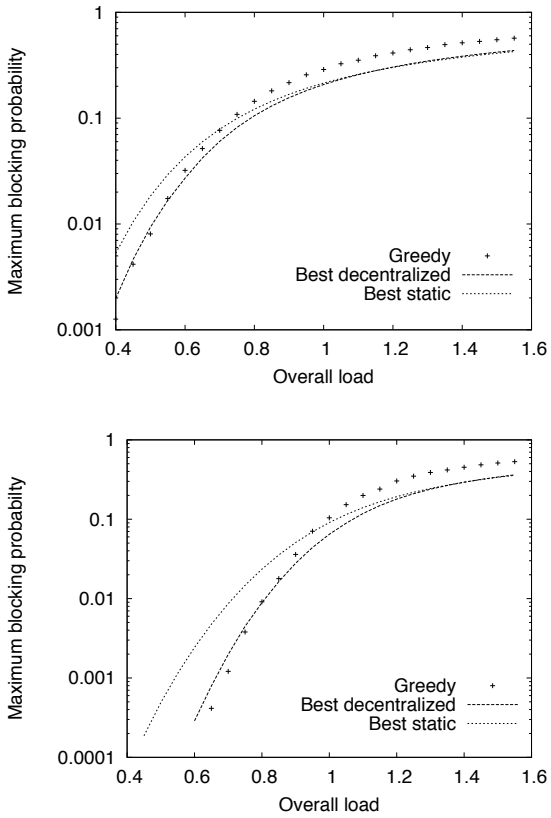


**Figure 9: Maximum per-class blocking rate for telephone traffic (upper graph) and data traffic (lower graph) with heterogeneous service requirements.**

# 6. CONCLUSION

While load balancing is a key component of computer systems and communication networks, most existing optimality and performance results are derived for specific topologies and traffic characteristics. In this paper, we have focused on those strategies that are *insensitive* to the distribution of service requirements, in the general setting of Whittle networks. We have characterized any insensitive load balancing as a linear combination of so-called "simple" strategies, for both a single customer class and several customer classes with decentralized routing decisions. This characterization leads to simple optimality and performance results that were illustrated on toy examples.

While we focused on *external* routing decisions only, it would be interesting to extend these results to *internal* routing decisions, where the successive nodes visited by a customer depend on the load of these nodes. This is the subject of future research.

## REFERENCES


[1] M. Alanyali, B. Hajek, Analysis of simple algorithms for dynamic load balancing, Mathematics of Operations Research 22-4 (1997) 840–871.

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

[3] T. Bonald, A. Proutière, Insensitive bandwidth sharing in data networks, Queueing Systems 44-1 (2003) 69–100.

[4] X. Chao, M. Miyazawa, R. F. Serfozo and H. Takada, Markov network processes with product form stationary distributions, Queueing Systems 28 (1998) 377–401.

[5] J.W. Cohen, The multiple phase service network with generalized processor sharing, Acta Informatica 12 (1979) 245–284.

[6] M.B. Combe, O.J. Boxma, Optimization of static traffic allocation policies, Theoretical Computer Science 125 (1994) 17–43.

[7] A. Ephremides, P. Varaiya, J. Walrand, A simple dynamic routing problem, IEEE Transactions on Automatic control 25 (1980) 690–693.

[8] M. Harchol-Balter, M. Crovella, C. Murta, On choosing a task assignment policy for a distributed server system, IEEE journal of parallel and distributed computing 59 (1999) 204–228.

[9] A. Hordijk, G. Koole, On the assignment of customers to parallel queues, Probability in the Engineering and Informational Sciences 6 (1992) 495–511.

[10] F.P. Kelly, Blocking Probabilities in Large Circuit-switched Networks, Adv. Applied Probability 18 (1986) 473–505.

[11] F.P. Kelly, Routing and capacities allocations in networks with trunk reservations. Mathematics of operations research 15 (1990) 771–793.

[12] F.P. Kelly, Network routing, Philosophical Transactions of the Royal Society A337 (1991) 343–367.

[13] F.P. Kelly, Loss Networks, Annals of Applied Probabilities 1-3 (1991) 319–378.

[14] F.P. Kelly, Bounds on the performance of dynamic routing schemes for highly connected networks, Mathematics of Operations Research 19 (1994) 1–20.

[15] F.P. Kelly, Mathematical modelling of the Internet, in: *Mathematics Unlimited - 2001 and Beyond* (Editors B. Engquist and W. Schmid), Springer-Verlag, Berlin (2001) 685–702.

[16] J. van Leeuwaarden, S. Aalto, J. Virtamo, Load balancing in cellular networks using first policy iteration, Technical Report, Networking Laboratory, Helsinki University of Technology, 2001.

[17] D. Mitra, R.J. Gibbens, B.D. Huang, State dependent routing on symmetric loss networks with trunk reservations, IEEE Transactions on communications 41-2 (1993) 400–411.

[18] S. Nelakuditi, Adaptive proportional routing: a localized QoS routing approach, in: *Proc. of IEEE Infocom*, 2000.

[19] R. Nelson, T. Philips, An approximation for the mean response time for shortest queue routing with general interarrival and service times, Performance evaluation 17 (1993) 123–139.

[20] S. Oueslati-Boulahia, E. Oubagha, An approach to routing elastic flows, in: *Proc. of ITC 16*, 1999.

[21] R. Serfozo, *Introduction to stochastic networks*, Springer, 1999.

[22] P. Sparaggis, C. Cassandras, D. Towley, Optimal control of multiclass parallel service systems with and without state information, in: *proc. of the 32nd Conference on Decision Control*, San Antonio, 1993.

[23] S. Stidham, Optimal control of admission to a queueing system, IEEE Transactions on Automatic Control 30-8 (1985) 705–713.

[24] A.N. Tantawi and D. Towsley, Optimal static load balancing in distributed computer systems, Journal of the ACM 32-2 (1985) 445–465.

[25] D. Towsley, D. Panayotis, P. Sparaggis, C. Cassandras, Optimal routing and buffer allocation for a class of finite capacity queueing systems, IEEE Trans. on Automatic Control 37-9 (1992) 1446–1451.

[26] W. Whitt, Deciding which queue to join: some counterexamples, Operations research 34-1 (1986) 226–244.

[27] W. Winston, Optimality of the shortest line discipline, Journal of Applied Probability 14 (1977) 181–189.