On performance bounds for balanced fairness

Thomas Bonald, Alexandre Proutière

On performance bounds for balanced fairness

T. Bonald and A. Proutière
France Telecom R&D
38-40, rue du Général Leclerc
92794 Issy-les-Moulineaux Cedex 9, France
{Thomas.Bonald,Alexandre.Proutiere}@francetelecom.com

May 7, 2003

Abstract

While Erlang's formula has helped engineers to dimension telephone networks for over eighty years, such a three-way "performance - demand - capacity" relationship is still lacking for data networks. It may be argued that the enduring success of Erlang's formula is essentially due to its simplicity: the call blocking rate does not depend on the distribution of call duration but on overall demand only. In this paper, we consider data networks and characterize those capacity allocations which have the same insensitivity property, in the sense that performance of data transfers does not depend on precise traffic characteristics such as the distribution of data volume but on overall demand only. We introduce the notion of "balanced fairness" and prove some key properties satisfied by this insensitive allocation. It is shown notably that the performance of balanced fairness is always better than that obtained if flows are transmitted in a "store and forward" fashion, allowing simple formula applying to the latter to be used as a conservative evaluation for network design and provisioning purposes.

1 Introduction

Erlang's formula has helped engineers to dimension telephone networks for more than eighty years [10]. It gives the proportion of calls that are blocked, B, as a function of demand (call arrival rate $\times$ mean call duration, A) and capacity (number of circuits, C) only:

$$B = \frac{A^C/C!}{1 + A + \ldots + A^C/C!}.$$

In particular, Erlang's formula is *insensitive* in the sense that the blocking probability does not depend on the distribution of call durations. The only required assumption is that calls arrive as a Poisson process, which is verified in practice as calls are generated by a large number of users with mutually independent behaviors. This insensitivity property notably explains why Erlang's formula is still used for dimensioning current telephone networks, despite the fact that telephone traffic characteristics have changed considerably since Erlang's publication in 1917.

1.1 Data network provisioning

Such a simple three-way "performance - demand - capacity" relationship is not yet available for data networks. First, what is meant by performance depends on whether the considered application is time-critical (e.g., voice, audio and video streaming, interactive games) or not (e.g., file transfers, Web browsing). In

this paper, we consider the latter only, that is, those applications that require a succession of document transfers and for which performance does not depend on the delay of individual packets but on the transfer time of an entire document. These applications constitute the majority of traffic in current data networks, the document transfer being typically controlled by TCP [12]. The corresponding flows are often referred to as "elastic" as their rate varies with respect to the level of congestion in the network. Thus the duration of an elastic flow does not only depend on its size (the volume of the data transfer) but also on its rate which varies in a random way as new flows arrive and existing flows cease.

Like telephone traffic, demand in data networks (in bit/s) can be defined simply as the product of the flow arrival rate by the mean flow size. Note, however, that the flow arrival process is not Poisson. Flows are usually generated within sessions, each session being composed of a succession of flows separated by an interval of inactivity generally referred to as "think-time". A typical example is the succession of Web pages downloaded by a user in a period of continuous activity. This may result in a strongly correlated flow arrival process, depending on the number of flows in a session, the distribution of flow sizes and think-time durations and their possible correlation [7, 22]. However, assuming that sessions are generated independently by a large population of users, the *session* arrival process is well approximated by a Poisson process. This has been recognized as one of the rare invariants of Internet traffic [22].

The way capacity is shared is not as simple as in telephone networks where a circuit is occupied throughout the call holding time. Capacity allocation in data networks results from the complex interaction of a number of packet-level mechanisms, including congestion control, scheduling and buffer management, and has been the subject of considerable recent research [13, 15, 16, 17, 18, 20, 21]. These studies generally consider a fixed number of long-lived flows and thus implicitly make a *time-scale separation* assumption: the time-scale of packet-level dynamics (the time to attain the equilibrium capacity allocation given a fixed number of flows) is so small compared to the time-scale of flow-level dynamics (the time between successive flow arrivals) that flows can be assumed to last indefinitely. In this paper, we also use the time-scale separation assumption, but to study the impact on performance of the random nature of traffic at flow level. Specifically, we neglect packet-level dynamics in that the equilibrium capacity allocation is assumed to be immediately attained on each flow arrival or flow departure. Under this assumption, we study those capacity allocations which are insensitive for data traffic in the same way Erlang's formula is insensitive for telephone traffic: performance depends on demand and capacity only, and not on precise traffic characteristics such as the flow size distribution or the structure of sessions.

The main motivation for studying insensitive allocations is to derive simple performance results for data networks, depending on demand and capacity only and therefore robust with respect to evolutions in the nature of user applications (Web, peer to peer,...). These allocations could then be used as design objectives for future packet-level mechanisms. Simulations suggest, however, that the performance of these allocations is generally close to that of well-known allocations such as max-min fairness [5, 6], which has long been stated as an ideal objective for congestion control algorithms [2]. It is thus expected that the results obtained in the present paper constitute a sufficiently accurate approximation of the performance realized by existing packet-level mechanisms and that derived engineering guidelines can be used for current data networks.

1.2 Related work

The first insensitivity result for elastic traffic was given in [19] for the case of a single bottleneck whose capacity is fairly shared between flows in progress. The corresponding model is the processor sharing queue. The mean duration $T(s)$ of a flow of size s is then a very simple function of demand, A, and capacity, C:

$$T(s) = \frac{s}{C - A}.$$

Like Erlang's formula, the only required assumption is that *sessions* arrive as a Poisson process [1, 7]. While this result is extremely simple and useful, it is not sufficient to evaluate the impact on performance of multiple bottlenecks, given that most flows go through several links in data networks. Assuming flows sharing the same links have the same rate, each *route* can then be represented as a processor sharing queue, with a state-dependent service speed given by the capacity allocated to those flows on this route. Using key properties of Whittle queueing networks [4], we proved in [5, 6] that for any network topology, the capacity allocations that lead to insensitive performance are those for which the following *balance* property holds:

> For all pairs of flows f, g, the relative change in the capacity allocated to f when g is removed is the same as the relative change in the capacity allocated to g when f is removed.

There is a continuum of allocations satisfying this property, each characterized by a so-called "balance function". Among these there is just one allocation for which in any state, the capacity of at least one link is fully allocated. We refer to this unique insensitive allocation as "balanced fairness". Balanced fairness differs from well-known allocations such as max-min fairness [2] or proportional fairness [15] except for very specific network topologies referred to as homogeneous "hypercubes", in which case it coincides with proportional fairness. These networks are the generalization of so-called homogeneous lines and grids for which proportional fairness was indeed shown to be insensitive in the case of Poisson flow arrivals [3]. For any other network topology, max-min fairness, proportional fairness and, more generally, any capacity allocation based on the maximization of some utility function, are sensitive [6]. This means that the performance of these allocations cannot be evaluated without specific assumptions on traffic characteristics, and explains why such performance results are so difficult to derive, even for the simplest network topologies [11].

1.3 Contribution

While the performance of balanced fairness does not depend on detailed traffic characteristics, it is still a complex function of demand on all routes and of the capacity of all links. This renders the exact evaluation difficult to apply for network engineering purposes. It is necessary in this case to apply an appropriate decomposition allowing provisioning decisions to be made locally. The main contribution of the present paper is to demonstrate that such a decomposition is possible. Specifically, we prove that the performance of balanced fairness on any given route is worse than that obtained if flows were transmitted on a single isolated link of this route, and better than that obtained if flows were handled link by link on this route in a "store and forward" fashion. Thus the mean duration $T_r(s)$ of a flow of size s on route r satisfies the inequalities:

$$\max_{l \in r} \frac{s}{C_l - A_l} \leq T_r(s) \leq \sum_{l \in r} \frac{s}{C_l - A_l}, \tag{1}$$

where C_l and A_l denote the capacity and the overall demand of link l, respectively. These simple performance bounds require per-link information only and coincide for single-link routes.

After describing the model in the next section, we introduce in Section 3 the notions of balance and insensitivity. We notably show that the balance property is in fact equivalent to three milder forms of insensitivity: insensitivity to the distribution of successive flow sizes and think-time durations, insensitivity to the flow arrival process and time-scale insensitivity. We also prove the lower bound (1), which holds for any insensitive allocation. In the following two sections, we define and give key properties of the insensitive allocations referred to as "store and forward" and "balanced fairness", and prove the upper bound (1). Finally, we illustrate these results on some toy network topologies and conclude the paper.

2 Flow-level modeling of data networks

We first introduce a generic flow-level model of data networks. We then show how this model can be represented by a processor sharing queueing network with state-dependent service speeds. For sake of clarity, we show step by step the generality of the model, from the simplest case where flow arrivals are Poisson and flow sizes exponential i.i.d. to the most general case where flows arrive in sessions with an arbitrary distribution and correlation of successive flow sizes and think-time durations.

2.1 Data network model

We represent a data network as a set of L links where each link l has a capacity $C_l > 0$, $l = 1, \ldots, L$. A random number of flows compete for access to these links. Each flow is characterized by a volume of information to be transferred (referred to as the flow size) on a route consisting of a set of links. The flows are "elastic" in the sense that their duration depends on their rate which varies as new flows begin and other cease. Specifically, a flow of size s arriving at time t_{start} on route r is completed at time t_{end} given by:

$$s = \int_{t_{\text{start}}}^{t_{\text{end}}} c(t)dt,$$

where $c(t)$ denotes the flow rate at time t, that is the capacity allocated to this flow on *each* link of route r at time t, $t_{\text{start}} \leq t \leq t_{\text{end}}$. This rate is limited by the capacity C_l of each link $l \in r$ that is shared between all flows in progress on route r and on other routes containing link l. It may additionally be constrained by a fixed maximum limit representing external constraints such as the user's access line.

Capacity allocation. We consider K classes of flow in this data network. Each class k is characterized by a route r_k consisting of a non-empty set of links and a per-flow rate limit $a_k > 0$ we refer to as the "access rate". We adopt the convention that either $a_k < \min_{l \in r_k} C_l$, in which case the access rate is limiting, or $a_k = \infty$. We denote by $x = (x_1, \ldots, x_K)$ the network state, where x_k is the number of flows of class k in progress. It is assumed that the total capacity ϕ_k allocated to flows of class k is equally shared between these flows and depends on the network state x only. The allocation must satisfy the capacity constraints:

$$\sum_{k: l \in r_k} \phi_k(x) \leq C_l, \quad l = 1, \ldots, L \quad \text{and} \quad \phi_k(x) \leq x_k a_k, \quad k = 1, \ldots, K. \tag{2}$$

Traffic conditions. The evolution of the network state x does not only depend on the way capacity is allocated between flows in progress but on traffic characteristics, i.e., on the way new flows are generated and on the distribution of their size. The traffic characteristics considered in this paper are quite general and described in detail in §2.3-2.4. It is sufficient at this stage to assume that the marked point process of flow arrivals of each class, with marks corresponding to the flow sizes, is stationary and ergodic. Denote by ρ_k the traffic intensity of class k. This corresponds to the mean volume of information offered by flows of class k per unit of time. We denote by $A_l = \sum_{k: l \in r_k} \rho_k$ the overall traffic intensity at link l, and refer to the usual traffic conditions as the inequalities:

$$A_l < C_l, \quad l = 1, \ldots, L. \tag{3}$$

2.2 A processor sharing queueing network

We now introduce a queueing network of processor sharing nodes with state-dependent service speeds. We show in §2.3-2.4 that this queueing network can represent the data network described in §2.1 with virtually any traffic characteristics (arbitrary flow size distribution, correlated arrivals of flows within sessions, etc).

Definition. Consider an open queueing network of N processor sharing nodes with state-dependent speeds, that is, the service speed ψ_i of node i depends on the state $y = (y_1, \ldots, y_N)$, where y_i is the number of customers in node i. We only assume that $\psi_i(y) > 0$ if and only if $y_i > 0$. Exogenous arrivals at node i form a Poisson process of rate ν_i. The services required at node i are exponential i.i.d. of mean $1/\mu_i$. After service completion at node i, a customer moves to node j with probability p_{ij} and leaves the network with probability $p_i \equiv 1 - \sum_j p_{ij}$. The routing matrix (p_{ij}) is assumed to be transient, so that the effective arrival rate λ_i at node i is uniquely defined by the equations:

$$\lambda_i = \nu_i + \sum_j \lambda_j p_{ji}, \quad i = 1, \ldots, N.$$

We denote by $\varrho_i = \lambda_i/\mu_i$ the traffic intensity at node i.

Balance equations. The stochastic process $Y = \{Y_t, t \geq 0\}$ that describes the evolution of the number of customers at each node is an irreducible Markov process. Let f_i be the unit vector with 1 in component i and 0 elsewhere, for $i = 1, \ldots, N$. When the network is in state y, the possible transitions are triggered by the movement of a customer from node i to node j, in which case the next state is $T_i^j y \equiv y - f_i + f_j$, a departure from node i, in which case the next state is $T_i y \equiv y - f_i$, and an exogenous arrival at node j, in which case the next state is $T^j y \equiv y + f_j$. The balance equations that an invariant measure χ must satisfy are thus:

$$\chi(y) \sum_i (\nu_i + \psi_i(y)\mu_i) = \sum_i \chi(T_i y)\nu_i + \sum_{i,j} \chi(T_i^j y)\psi_j(T_i^j y)\mu_j p_{ji} + \sum_i \chi(T^i y)\psi_i(T^i y)\mu_i p_i. \qquad (4)$$

This measure may be of infinite sum, in which case the Markov process Y is transient or null recurrent. Otherwise, the Markov process Y is positive recurrent with stationary distribution:

$$\lim_{t\to\infty} \Pr(Y_t = y) = \frac{\chi(y)}{\sum_{y'} \chi(y')}.$$

2.3 Poisson flow arrivals

Consider the data network of §2.1 where flows of each class arrive as an independent Poisson process. This may be represented by the above considered processor sharing queueing network, where each customer corresponds to an ongoing flow in case of exponential flow size distributions, to a phase of an ongoing flow in case of phase-type flow size distributions.

Exponential flow size distribution. If flows have exponential i.i.d. sizes, the corresponding processor sharing queueing network has $N = K$ nodes and no routing, i.e., $p_{ij} = 0$ for all i, j and $\varrho_i = \rho_i$ for all i. The service speed ψ_i of node i represents the total capacity ϕ_i allocated to flows of class i, which is equally shared between these flows. A simple example is given in Figure 1.

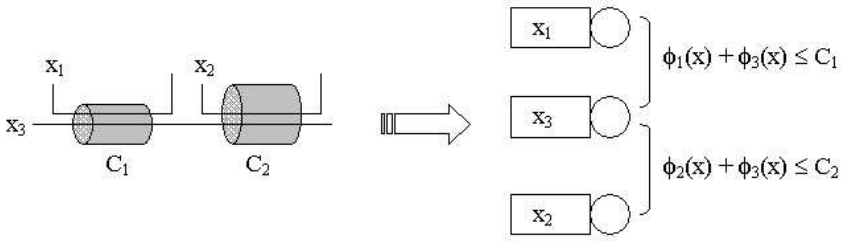

Figure 1: A data network represented as a processor sharing queueing network

Phase-type flow size distribution. Measurements of the size of flows in real data networks show that their distribution is not exponential but typically much more variable [9]. The considered queueing network allows phase-type distributions, which are known to form a dense subset within the set of all distributions of nonnegative random variables. A phase-type distribution for flows of class k can be represented simply by a set of consecutive nodes $S_k \subset \{1, \ldots, N\}$ such that $\nu_i > 0$ for the first node $i \in S_k$ and for any node $i \in S_k$, $p_{ij} = 0$ for all nodes j except for $j = i + 1$, if $i + 1 \in S_k$ (refer to Figure 2). As each node $i \in S_k$ represents a phase of flows of class k, we have:

$$\psi_i(y) = \frac{y_i}{x_k} \phi_k(x), \quad \text{with} \quad x_k = \sum_{i \in S_k} y_i.$$

The traffic intensity of flows of class k is given by:

$$\rho_k = \sum_{i \in S_k} \varrho_i.$$

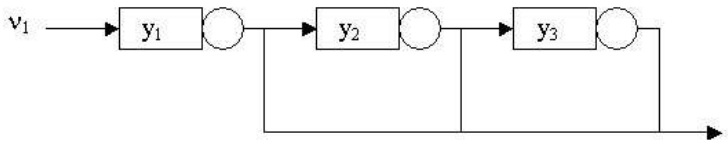

Figure 2: A 3-phase distribution of flow sizes

2.4 Poisson session arrivals

As mentioned in Section 1, flows do not arrive as independent Poisson processes in data networks. They are typically generated within sessions, each session being composed of a succession of flows separated by an interval of inactivity which we call "think-time". Again, the considered processor sharing network is sufficiently general to account for this complex structure of traffic, provided sessions arrive as a Poisson process and think-time durations do not depend on the network state (unlike flow durations).

Exponential flow size and think-time duration distributions. We first consider the case where successive flow sizes and think-time durations are all exponentially distributed. Think-times can simply be represented by infinite server nodes, i.e., those nodes i in the set $S_0 \subset \{1, \ldots, N\}$ such that:

$$\psi_i(y) = y_i. \tag{5}$$

We still denote by $S_k \subset \{1, \ldots, N\}$ the set of nodes representing flows of class k, i.e., such that:

$$\psi_i(y) = \frac{y_i}{x_k}\phi_k(x), \quad x_k = \sum_{i \in S_k} y_i. \tag{6}$$

A session can then be represented as a random walk of a customer in an alterning series of nodes in the sets S_k, $k \neq 0$, and in the set S_0. That is, for any node $i \notin S_0$, we have $p_{ij} = 0$ for all nodes $j \notin S_0$, and for any node $i \in S_0$, we have $p_{ij} = 0$ for all nodes $j \in S_0$. We assume without loss of generality that $\nu_i = 0$ and $p_i = 0$ for all nodes $i \in S_0$, which means that a session necessarily starts and ends with a flow (and not a think-time). We say that a session is *multi-class* if its successive flows may belong to different classes, and *single-class* otherwise. Again, the traffic intensity of flows of class k is simply given by:

$$\rho_k = \sum_{i \in S_k} \varrho_i. \tag{7}$$

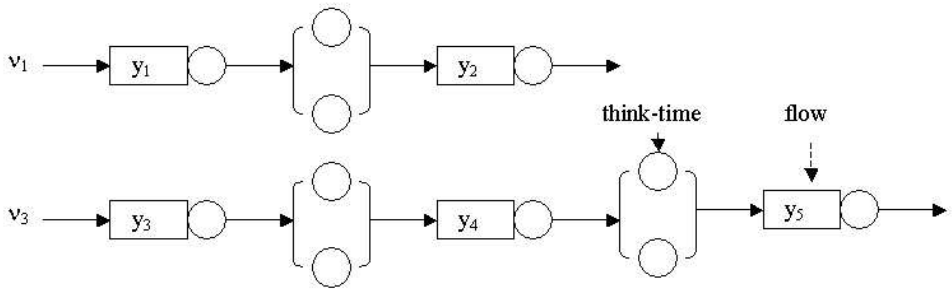

Figure 3: Example of two types of session, composed of two and three flows, respectively

It is worth noting that the distribution of the number of flows per session may be perfectly general. Successive flow sizes and think-time durations may also be correlated. Figure 3 gives an example of two types of session, composed of two and three flows, respectively. The mean flow sizes may well be higher for the first type of session for instance. In fact, arbitrary correlations between successive flow sizes and think-time durations may be represented by considering as many session types as necessary and introducing phase-type distributions.

Phase-type flow size and think-time duration distributions. As in §2.3, assume now that each node represents a phase of a flow or a think-time (and not the flow or the think-time itself). We still denote by S_0 the set of nodes representing think-times, satisfying (5), and S_k the set of nodes representing flows of class k, satisfying (6). Assume without loss of generality that successive phases of the same flow or think-time consist of consecutive nodes. A session with phase-type distributions of flow sizes and think-time durations can be represented as a random walk such that any visit of a customer to a node

$i \in S_k$, $k \neq 0$, can be followed by a visit to the node $i+1 \in S_k$ if this node corresponds to a new phase of the same flow, or a visit to a node $j \in S_0$ representing the first phase of a think-time. Similarly, any visit to a node $i \in S_0$ can be followed by a visit to the node $i+1 \in S_0$ if this node corresponds to a new phase of the same think-time, or a visit to a node $j \notin S_0$ representing the first phase of a flow. New sessions are represented by those nodes $i \notin S_0$ such that $\nu_i > 0$, corresponding to the first phase of a flow. The traffic intensity of flows of class k is still given by (7).

3 Insensitive allocations

We now characterize those capacity allocations for which performance is insensitive to the above described traffic characteristics. Specifically, we prove in Theorems 1 and 2 that the insensitivity property is equivalent to three milder forms of insensitivity, which all imply the balance property. We then give key properties of these allocations, including the lower bound (1) in Proposition 1.

3.1 Balance property

Let e_k be the unit vector with 1 in component k and 0 elsewhere, for $k = 1, \ldots, K$.

Definition 1 (Balance property) *The capacities $\phi_1, \ldots, \phi_K$ are said to be balanced if:*

$$\phi_k(x)\phi_{k'}(x - e_k) = \phi_{k'}(x)\phi_k(x - e_{k'}), \quad \forall k, k' \; \forall x \text{ such that } x_k > 0 \text{ and } x_{k'} > 0.$$

Let $\langle x, x - e_{k_1}, x - e_{k_1} - e_{k_2} \ldots, x - e_{k_1} - \ldots - e_{k_{n-1}}, 0\rangle$ be a direct path from state x to state 0, i.e., a path of length n where $n \equiv |x|$ is the number of flows in state x. The balance property implies that for any $x \neq 0$, the expression

$$\Phi(x) = \frac{1}{\phi_{k_1}(x)\phi_{k_2}(x - e_{k_1})\phi_{k_3}(x - e_{k_1} - e_{k_2}) \ldots \phi_{k_n}(x - e_{k_1} - \ldots - e_{k_{n-1}})} \tag{8}$$

is independent of the considered direct path. Defining $\Phi(0) = 1$, the capacities are uniquely characterized by the function Φ, referred to as the balance function:

$$\phi_k(x) = \frac{\Phi(x - e_k)}{\Phi(x)}, \quad \forall k, \; x_k > 0. \tag{9}$$

Conversely, if there exists a positive function Φ such that the capacities satisfy (9), it can be easily verified that these capacities are balanced. We say that the capacities are balanced by Φ.

3.2 Sufficient condition for insensitivity

Consider an allocation for which the balance property holds. The processor sharing queueing network introduced in §2.2 can represent virtually any traffic characteristics, provided session arrivals form a Poisson process. In view of (5) and (6), it may be readily verified that the corresponding service speeds $\psi_1, \ldots, \psi_N$ are balanced by the function Ψ defined by:

$$\Psi(y) = \prod_{i \in S_0} \frac{1}{y_i!} \times \prod_{k=1}^{K} \binom{x_k}{y_i, i \in S_k} \times \Phi(x),$$

where we use the notation:

$$\binom{x_k}{y_i, i \in S_k} \equiv \frac{x_k!}{\prod_{i \in S_k} y_i!}.$$

The processor sharing queueing network is then a so-called Whittle network [23], for which an invariant measure χ is simply given by:

$$\chi(y) = \Psi(y) \prod_{i=1}^{N} \varrho_i^{y_i}. \tag{10}$$

Summing this expression over all states corresponding to x_k flows of class k, we obtain in view of (7):

$$\sum_{y:\sum_{i\in S_k} y_i = x_k} \chi(y) = \prod_{i\in S_0} e^{\varrho_i} \times \Phi(x) \prod_{k=1}^{K} \rho_k^{x_k}. \tag{11}$$

Thus the invariant measures of the number of flows of each class are insensitive to *any* traffic characteristics (flow size distribution, distribution of the number of flows per session, correlation between successive flow sizes and think-time durations, etc) except the traffic intensities $\rho_1, \ldots, \rho_K$. This is actually a direct consequence of the well-known insensitivity of Whittle networks [6]. We conclude that the balance property indeed implies insensitivity.

3.3 Necessary condition for insensitivity

A key result is that the converse is also true: an allocation for which the invariant measures of the number of flows of each class are insensitive to any traffic characteristics except the traffic intensities $\rho_1, \ldots, \rho_K$ is balanced [6]. In fact, each of the following milder forms of insensitivity implies the balance property:

(I1) Insensitivity to the flow size distribution: For Poisson flow arrivals, the invariant measures of the number of flows of each class remain unchanged when for any class, the exponential distribution of flow sizes is replaced by any phase-type distribution of same mean.

(I2) Insensitivity to the flow arrival process: For exponential i.i.d. flow sizes, the invariant measures of the number of flows of each class remain unchanged when for any class, the Poisson flow arrivals are replaced by Poisson *session* arrivals with the same flow arrival rate.

(I3) Time-scale insensitivity: For Poisson flow arrivals and exponential i.i.d. flow sizes, the invariant measures of the number of flows of each class remain unchanged when for any class, flow inter-arrival times and flow sizes are multiplied by the same constant.

Theorem 1 *Any allocation that satifies one of the properties (I1), (I2), (I3) is balanced.*

The proof of Theorem 1, given in Appendix A, directly follows from the necessary condition for insensitivity in processor sharing networks proved in [4]. Theorem 2 below, also proved in Appendix A, is a stronger result than Theorem 1 since the properties (I1), (I2), (I3) correspond to respective particular cases of the following properties:

(I1') Insensitivity to the distribution of successive flow sizes and think-time durations: For Poisson session arrivals, the invariant measures of the number of flows of each class remain unchanged when for any class, the exponential distributions of successive flow sizes and think-time durations are replaced by any phase-type distributions of same respective means.

(I2') Insensitivity to the flow arrival process: For i.i.d. flow sizes, the invariant measures of the number of flows of each class remain unchanged when for any class, the Poisson flow arrivals are replaced by Poisson *session* arrivals with the same flow arrival rate.

(I3') Time-scale insensitivity: For Poisson *single-class* session arrivals, the invariant measures of the number of flows of each class remain unchanged when for any class, session inter-arrival times and successive flow sizes and think-time durations are multiplied by the same constant.

Theorem 2 *Any allocation that satifies one of the properties (I1'), (I2'), (I3') is balanced.*

In view of Theorems 1 and 2 and §3.2, all six insensitivity properties above are equivalent.

3.4 Properties of insensitive allocations

In view of previous results, there exists a continuum of insensitive allocations, each characterized by a positive function Φ according to (9). In the rest of the paper, we use the convention $\Phi(x) = 0$ for any x such that $x_k < 0$ for some k. In view of the capacity constraints (2), Φ must satisfy the following inequalities in any state x:

$$\sum_{k:l\in r_k} \frac{\Phi(x-e_k)}{\Phi(x)} \leq C_l,\ l = 1,\ldots,L, \quad \text{and} \quad \frac{\Phi(x-e_k)}{\Phi(x)} \leq x_k a_k,\ k = 1,\ldots,K. \tag{12}$$

Given a function Φ which satisfies these inequalities, it follows from (11) that the stability condition for the corresponding allocation is:

$$\sum_x \Phi(x) \prod_{k=1}^K \rho_k^{x_k} < \infty, \tag{13}$$

in which case the stationary distribution of the number of flows of each class is given by:

$$\pi(x) = \pi(0) \times \Phi(x) \prod_{k=1}^K \rho_k^{x_k}, \tag{14}$$

with

$$\pi(0) = \left(\sum_x \Phi(x) \prod_{k=1}^K \rho_k^{x_k} \right)^{-1}. \tag{15}$$

This corresponds to the stationary distribution of the Markov process describing the evolution of the number of flows of each class for Poisson flow arrivals and exponential i.i.d. flow sizes.

Using the fact that mean sojourn time of a customer in any node of a Whittle network is proportional to its quantity of service [4], we deduce that the mean duration $T_k(s)$ of a class-k flow of size s is proportional to s. Applying Little's formula, we get:

$$T_k(s) = s \times \frac{E[x_k]}{\rho_k}. \tag{16}$$

The following performance bound proved in Appendix A holds for any insensitive allocation.

Proposition 1 *For any class k, the mean duration of a class-k flow of size s satisfies:*

$$T_k(s) \geq \frac{s}{a_k} \quad \textit{and} \quad T_k(s) \geq \frac{s}{C_l - A_l} \quad \forall l \in r_k.$$

4 Store and forward

In this section, we introduce an insensitive allocation which has the property that the stationary distribution of the number of flows of each class is the same *as if* flows were successively transferred on each link of their route, in a "store and forward" fashion. In particular, the mean flow duration of each class has a simple and explicit expression.

The insensitivity of the "store and forward" allocation actually follows from that of Kelly queueing networks [14]. Thus, before defining the corresponding balance function $\Phi^{\rm SF}$, we first successively introduce an open and a closed Kelly queueing network. We shall deduce from the analysis of the latter the capacity constraints (12), from the analysis of the former the stability condition and the mean flow duration of each class.

4.1 An open Kelly queueing network

Consider a data network model different to that described in §2.1 in that flows of each class k are *successively* transmitted on an access link of capacity a_k and on network links $l \in r_k$ instead of *simultaneously* consuming capacity on each of these links. This model can then be represented by the following open queueing network.

Definition. The network consists of processor sharing nodes $1, \ldots, L$ of respective capacities $C_1, \ldots, C_L$ and infinite server nodes $1, \ldots, K$ with respective per-server capacities $a_1, \ldots, a_K$. Note that the per-server capacity of some of these infinite server nodes could be infinite in which case they contain no customer with probability 1. Services at each node are exponential i.i.d. of unit mean. There are K classes of customer. Customers of class k arrive as a Poisson process of rate ρ_k, visit the infinite server node k and the processor sharing nodes $l \in r_k$, in a fixed but arbitrary order, then leave the network.

Stationary distribution. This is an open Kelly queueing network, stable under the usual traffic conditions (3). Let z_k be the number of customers of class k visiting the infinite server node k and z_{kl} the number of customers of class k visiting the processor sharing node l, $l \in r_k$. The stationary distribution η of the Markov process $Z = \{Z_t, t \geq 0\}$ that describes the evolution of the number of customers of each class at each node (processor sharing nodes and infinite server nodes) is the same *as if* customers of each class arrive as an independent Poisson process at each node:

$$\eta(z) = \eta(0) \times \prod_{k=1}^{K} \frac{1}{z_k!} \left(\frac{\rho_k}{a_k}\right)^{z_k} \times \prod_{l=1}^{L} \binom{\sum_{k:l\in r_k} z_{kl}}{z_{kl}, k : l \in r_k} \prod_{k:l\in r_k} \left(\frac{\rho_k}{C_l}\right)^{z_{kl}}. \tag{17}$$

It follows from the insensitivity of Kelly networks that this stationary distribution does not depend on the service distribution of each class of customer at each node, nor on possible correlations between successive services required by the same customer, including the case where each customer requires the *same* service at each node [14]. This represents the above considered data network where each flow is successively transferred from link to link, in a "store and forward" fashion (see Figure 4). Denote by $\mathcal{Z}(x)$ the set of states z corresponding to x_k flows of class k, for each class k:

$$\mathcal{Z}(x) = \left\{ z : \ \forall k, \ z_k + \sum_{l\in r_k} z_{kl} = x_k \right\} \tag{18}$$

The stationary distribution of the number of flows of each class is given by:

$$\pi^{\mathrm{SF}}(x) = \sum_{z \in \mathcal{Z}(x)} \eta(z) = \pi^{\mathrm{SF}}(0) \times \Phi^{\mathrm{SF}}(x) \prod_{k=1}^{K} \rho_k^{x_k}, \tag{19}$$

where

$$\Phi^{\mathrm{SF}}(x) = \sum_{z \in \mathcal{Z}(x)} \prod_{k=1}^{K} \frac{1}{z_k!} \left(\frac{1}{a_k}\right)^{z_k} \times \prod_{l=1}^{L} \binom{\sum_{k:l \in r_k} z_{kl}}{z_{kl}, k : l \in r_k} \prod_{k:l \in r_k} \left(\frac{1}{C_l}\right)^{z_{kl}} \tag{20}$$

and

$$\pi^{\mathrm{SF}}(0) = \left(\sum_{x} \Phi^{\mathrm{SF}}(x) \prod_{k=1}^{K} \rho_k^{x_k} \right)^{-1}.$$

The mean number of flows of class k is:

$$E[x_k] = E[z_k] + \sum_{l \in r_k} E[z_{kl}] = \frac{\rho_k}{a_k} + \sum_{l \in r_k} \frac{\rho_k}{C_l - A_l}. \tag{21}$$

As in Section 2, any traffic characteristics may actually be represented: the stationary distribution of the number of flows of each class is still given by (19). Now note that flows may additionally be divided into an arbitrary number of "blocks" exactly as sessions are divided into an arbitrary number of flows, with the constraint that each block must leave the network before the next block of the same flow can enter: the stationary distribution of the number of blocks of each class at each link is still given by (17), *as if* blocks of each class arrive as an independent Poisson process at each node.

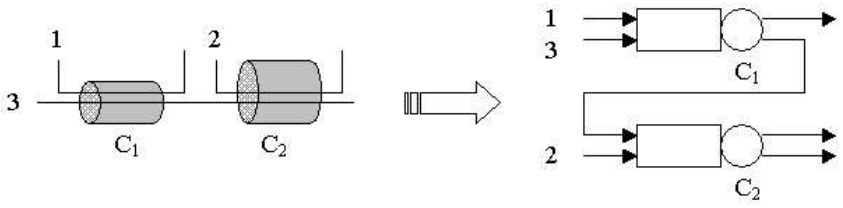

Figure 4: A data network with "store and forward" transfers represented as a Kelly queueing network

4.2 A closed Kelly queueing network

Results of §4.1 suggest that an insensitive allocation can be derived from the above considered network model with block transfers by letting the size of a block tend to zero so that flows are in fact emitted continuously. Given a fixed number of ongoing flows of each class, this network behaves like a closed queueing network, the transfer of successive blocks of a class-k flow being represented by cyclic visits of the same customer to the infinite server node k and the processor sharing nodes $l \in r_k$. The continuous transfer rate of a flow is then determined by the steady state of this closed queueing network through the rate at which the corresponding customer visits a particular node. This is a direct consequence of (9) and expression (22) below.

Definition. Consider the same network as that of §4.1 except that there is a fixed number x_k of customers of class k, $k = 1, \ldots, K$. The customers of class k visit the infinite server node k and the processor sharing nodes $l \in r_k$ in a cyclic way, in a fixed but arbitrary order (each of these nodes is visited exactly once in a cycle).

Stationary distribution. This is a closed Kelly network [14]. The stochastic process $\tilde{Z} = \{\tilde{Z}_t, t \geq 0\}$ that describes the evolution of the number of customers of each class at each node is an irreducible Markov process on the state space $\mathcal{Z}(x)$, given by (18). The stationary distribution of $\tilde{Z}$ is:

$$\tilde{\eta}(z) = \tilde{\eta}(0) \times \prod_{k=1}^{K} \frac{1}{z_k!} \left(\frac{1}{a_k}\right)^{z_k} \times \prod_{l=1}^{L} \binom{\sum_{k:l\in r_k} z_{kl}}{z_{kl}, k : l \in r_k} \prod_{k:l\in r_k} \left(\frac{1}{C_l}\right)^{z_{kl}},$$

where

$$\tilde{\eta}(0) = \left(\sum_{z\in\mathcal{Z}(x)} \prod_{k=1}^{K} \frac{1}{z_k!} \left(\frac{1}{a_k}\right)^{z_k} \times \prod_{l=1}^{L} \binom{\sum_{k:l\in r_k} z_{kl}}{z_{kl}, k : l \in r_k} \prod_{k:l\in r_k} \left(\frac{1}{C_l}\right)^{z_{kl}} \right)^{-1} \equiv \frac{1}{\Phi^{\mathrm{SF}}(x)}.$$

For any class k and any x such that $x_k > 0$, the rate (number of visits per unit of time) at which the x_k customers of class k visit any node $l \in r_k$ is given by:

$$\sum_{z\in\mathcal{Z}(x):z_{kl}>0} \tilde{\eta}(z) \times \frac{z_{kl}}{\sum_{k':l\in r_{k'}} z_{k'l}} \times C_l = \frac{\Phi^{\mathrm{SF}}(x - e_k)}{\Phi^{\mathrm{SF}}(x)}. \tag{22}$$

As the rate at which customers visit a server cannot exceed the speed of this server, we deduce:

$$\sum_{k:l\in r_k} \frac{\Phi^{\mathrm{SF}}(x - e_k)}{\Phi^{\mathrm{SF}}(x)} \leq C_l, \ l = 1, \ldots, L, \quad \text{and} \quad \frac{\Phi^{\mathrm{SF}}(x - e_k)}{\Phi^{\mathrm{SF}}(x)} \leq x_k a_k, \ k = 1, \ldots, K. \tag{23}$$

As in §4.1, it follows from the insensitivity of Kelly networks that the stationary distribution of $\tilde{Z}$ does not depend on the service distribution of each class of customer, nor on possible correlations between successive services required by the same customer. In particular, it remains the same if each customer requires the *same* service at each node in a cycle, which represents the transfer of the same block from link to link.

4.3 Definition and properties

In the rest of the paper, we refer to *store and forward* as the insensitive allocation characterized by the balance function Φ^{SF} given by (20), which in view of (23) satisfies the capacity constraints (12). The stability condition (13), which corresponds to that of the open queueing network considered in §4.1, is satisfied under the usual traffic conditions (3). In this case, the stationary distribution, given by (19), coincides with that one would obtain if flows were transferred in a "store and forward" way. In view of (16) and (21), the mean duration of class-k flows of size s is simply given by:

$$T_k^{\mathrm{SF}}(s) = \frac{s}{a_k} + \sum_{l\in r_k} \frac{s}{C_l - A_l}. \tag{24}$$

5 Balanced fairness

In this section, we define and give key properties of an allocation we refer to as "balanced fairness" [5]. This is the most efficient insensitive allocation in the following two senses. First, this is the only insensitive allocation such that in any state, a network link is saturated or a flow rate limit constraint is attained. Second, we prove in Proposition 3 below that this is the insensitive allocation for which the data network is empty with the highest probability. The main result of this paper is given in Theorem 4: the performance of balanced fairness is better than that of store and forward.

5.1 Definition

Consider the balance function Φ^{BF} recursively defined by $\Phi^{\mathrm{BF}}(0) = 1$ and:

$$\forall x \neq 0, \quad \Phi^{\mathrm{BF}}(x) = \max\left(\max_l \left\{\frac{1}{C_l}\sum_{k:l\in r_k}\Phi^{\mathrm{BF}}(x-e_k)\right\}, \max_{k:x_k>0}\left\{\frac{1}{a_k x_k}\Phi^{\mathrm{BF}}(x-e_k)\right\}\right). \tag{25}$$

This function clearly satisfies the inequalities (12). The corresponding allocation will be referred to as *balanced fairness*. Observe that in any state $x \neq 0$, at least one of the inequalities (12) is an equality, which means that a network link is saturated or a flow rate limit constraint is attained. This property characterizes balanced fairness among insensitive allocations. The following result, which is a direct consequence of definition (25), shows that balanced fairness is also the insensitive allocation with the minimum balance function Φ such that $\Phi(0) = 1$.

Proposition 2 *Let Φ be any positive function such that $\Phi(0) = 1$ and the inequalities (12) are satisfied. We have:*

$$\forall x, \quad \Phi(x) \geq \Phi^{\mathrm{BF}}(x).$$

Proof. The proof is by induction on the total number of flows $n = \sum_{k=1}^{K} x_k$. As $\Phi(0) = \Phi^{\mathrm{BF}}(0) = 1$, the inequality is satisfied for $n = 0$. Now assume it is satisfied for $n = m$, $m \geq 0$. Let x be any state with $n = m + 1$ total number of flows. From (12) and (25), we get:

$$\begin{aligned}\Phi(x) &\geq \max\left(\max_l \left\{\frac{1}{C_l}\sum_{k:l\in r_k}\Phi(x-e_k)\right\}, \max_{k:x_k>0}\left\{\frac{1}{a_k x_k}\Phi(x-e_k)\right\}\right)\\ &\geq \max\left(\max_l \left\{\frac{1}{C_l}\sum_{k:l\in r_k}\Phi^{\mathrm{BF}}(x-e_k)\right\}, \max_{k:x_k>0}\left\{\frac{1}{a_k x_k}\Phi^{\mathrm{BF}}(x-e_k)\right\}\right) = \Phi^{\mathrm{BF}}(x).\end{aligned}$$

□

5.2 Properties

We first characterize the stability region, then show that balanced fairness is the insensitive allocation for which the data network is empty with the highest probability.

Theorem 3 *The stability condition (13) holds for balanced fairness under the traffic conditions (3).*

Proof. From Proposition 2,

$$\Phi^{\mathrm{SF}}(x)\prod_{k=1}^{K}\rho_k^{x_k} \geq \Phi^{\mathrm{BF}}(x)\prod_{k=1}^{K}\rho_k^{x_k}.$$

The proof then follows from the fact that the stability condition (13) holds for store and forward under the traffic conditions (3). □

Proposition 3 *Consider any balanced allocation which does not coincide with balanced fairness and for which the stability condition (13) holds. The probability that the network is empty for this allocation is lower than for balanced fairness, i.e.,* $\pi(0) < \pi^{\mathrm{BF}}(0)$.

Proof. As the considered allocation does not coincide with balanced fairness, it follows from Proposition 2 that the corresponding balance function Φ satisfies $\Phi(x) \geq \Phi^{\mathrm{BF}}(x)$ for all states x, and $\Phi(x) > \Phi^{\mathrm{BF}}(x)$ for at least one state x. The proof then follows from (15). □

Finally, we give in Theorem 4 the main result of this paper: the mean flow duration is always smaller for balanced fairness than for store and forward. Lemmas 1 and 2 as well as Theorem 4 are proved in Appendix B.

Lemma 1 *Consider a class of flow k whose route consists of a single link l and is not subject to a rate limit, i.e., such that $r_k = \{l\}$ and $a_k = \infty$. Then in any state x such that $x_k > 0$, balanced fairness saturates link l, that is:*

$$\forall x,\ x_k > 0, \quad \Phi^{\mathrm{BF}}(x) = \frac{1}{C_l}\sum_{k':l\in r_{k'}} \Phi^{\mathrm{BF}}(x - e_{k'}).$$

Lemma 2 *Consider a class of flow k whose route consists of a single link l and is not subject to a rate limit, i.e., such that $r_k = \{l\}$ and $a_k = \infty$. Then, the mean number of class-k flows is:*

$$E[x_k] = \frac{\rho_k}{C_l - A_l}.$$

Theorem 4 *The mean duration of a class-k flow of size s satisfies:*

$$T_k^{\mathrm{BF}}(s) \leq T_k^{\mathrm{SF}}(s) = \frac{s}{a_k} + \sum_{l\in r_k}\frac{s}{C_l - A_l}.$$

Theorem 4 allows a simple conservative evaluation of the performance of balanced fairness, requiring per-link information only. In view of Proposition 1, this approximation will typically be accurate for a given class k when there is a clearly identified bottleneck on route r_k, i.e., a link $l \in r_k$ such that the so-called "residual capacity" of this link, $C_l - A_l$, is much smaller than that of any other link on route r_k, or when the flow rate limit a_k is much smaller than the residual capacity of each link $l \in r_k$.

6 Application to specific network topologies

In this section, we apply the previous results to specific network topologies. Specifically, we use (25) and (14) to evaluate the performance of balanced fairness and compare it to that of store and forward. In all graphs below, we plot the so-called *flow throughput*, defined as the ratio of the size s of a flow to its mean duration, which is independent of s in view of (16). We consider toy network topologies like lines and trees without flow rate limit where explicit expressions can be derived, and choose traffic conditions

where the performance of balanced fairness significantly differs from that of store and forward. In any more realistic scenario with a large number of routes and a large spectrum of link capacities and flow rate limits, explicit expressions can hardly be derived for balanced fairness. Numerical evaluations are always possible, however, and the difference between the performance of both allocations is typically much less significant (refer to [8] where an efficient recursive algorithm was developed for such complex topologies). This highlights the interest of the simple performance bounds derived in this paper.

6.1 Lines

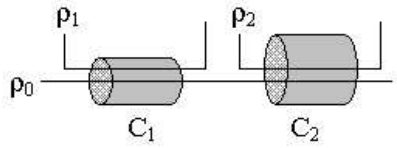

Figure 5: A 2-link line

We refer to a line as a network composed of L links of respective capacities $C_1, \ldots, C_L$, with one L-link route and L single-link routes. For simplicity, we assume without loss of generality that the minimum link capacity is equal to one. Denote by ρ_0 the traffic intensity on the L-link route. Under the stability condition $\rho_0 + \rho_l < C_l$ for all links l, the mean duration of a flow of size s on the L-link route is given by:

$$T_0^{\mathrm{BF}}(s) = \frac{s}{1-\rho_0} + \sum_{l=1}^{L} \left(\frac{s}{C_l - \rho_l - \rho_0} - \frac{s}{C_l - \rho_0} \right),$$

while for store and forward:

$$T_0^{\mathrm{SF}}(s) = \sum_{l=1}^{L} \frac{s}{C_l - \rho_l - \rho_0}.$$

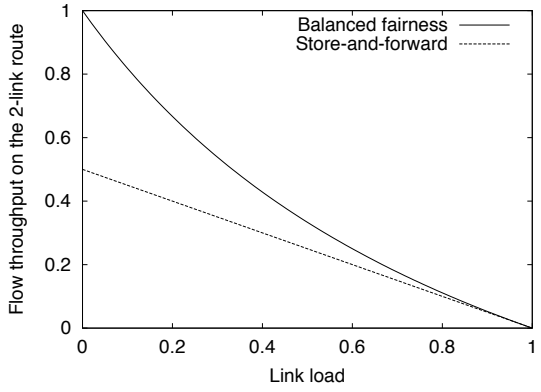

Figure 6: Performance of balanced fairness and store and forward in a homogeneous 2-link line

Note that, in view of Lemma 2, the flow throughput on single-link routes are the same. Figure 6 below compares for a line of two unit capacity links the flow throughput on the 2-link route obtained with balanced fairness and store and forward, when $\rho_0 \to 0$ and $\rho_1 = \rho_2$. We observe that while store and forward gives a good approximation of balanced fairness at high load, the difference is significant at low load. This can be explained simply by the fact that the rate of a flow on the 2-link route in the absence of any other flow is equal to 1 for balanced fairness, 1/2 for store and forward.

6.2 Trees

We refer to a tree as a network of $L = K + 1$ links: a trunk of unit capacity and K branches $1, \ldots, K$ of respective capacities $C_1, \ldots, C_K \leq 1$, with $\sum_k C_k > 1$. Route r_k contains the trunk and branch k, as illustrated in Figure 7. Tree networks may represent metropolitan area networks for instance, that consist of several multiplexing stages before access to backbone networks.

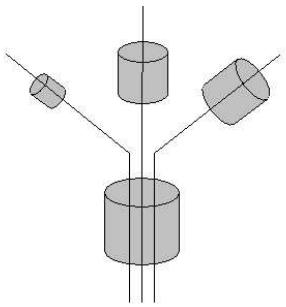

Figure 7: Tree networks

For a 2-branch tree, the mean duration of a flow of size s on branch 1 is given by:

$$T_1^{\mathrm{BF}}(s) = \frac{s}{1-\rho_1-\rho_2} + \frac{s}{C_1-\rho_1} \times \frac{C_1(1-C_1)(C_2-\rho_2)}{\rho_1(1-C_1)(C_2-\rho_2) + C_2(C_1-\rho_1)(1-\rho_2)},$$

while for store and forward:

$$T_1^{\mathrm{SF}}(s) = \frac{s}{1-\rho_1-\rho_2} + \frac{s}{C_1-\rho_1}.$$

In Figure 8, we compare balanced fairness and store and forward on a 2-branch tree with branches of capacities $C_1 = 0.1$ and $C_2 = 1$, in the case $\rho_1 \to 0$. Note that, in view of Lemma 2, the flow throughput on route 2 is the same for both allocations. The difference in the flow throughput on route 1 decreases with the capacity of branch 1.

6.3 A single link with different flow rate limits

Finally, we consider a single unit capacity link with different flow rate limits $a_1, \ldots, a_K < 1$. It proves difficult to derive explicit performance results for balanced fairness. For store and forward, the mean duration of a class-k flow of size s is simply:

$$T_k^{\mathrm{SF}}(s) = \frac{s}{a_k} + \frac{s}{1-\rho},$$

where $\rho = \sum_{k=1}^{K} \rho_k$ denotes the link load. Figure 9 compares the the performance of balanced fairness and store and forward for two flow rate limits, $a_1 = 0.01$ and $a_2 = 0.02$, in the case $\rho_1 = \rho_2$. We observe that store and forward provides a good conservative approximation of balanced fairness.

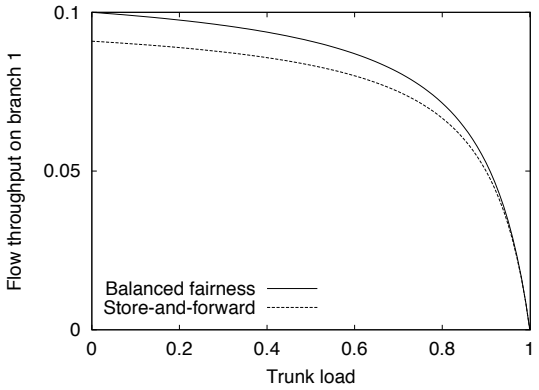

Figure 8: Performance of balanced fairness and store and forward in a 2-branch tree ($C_1 = 0.1,\ C_2 = 1$)

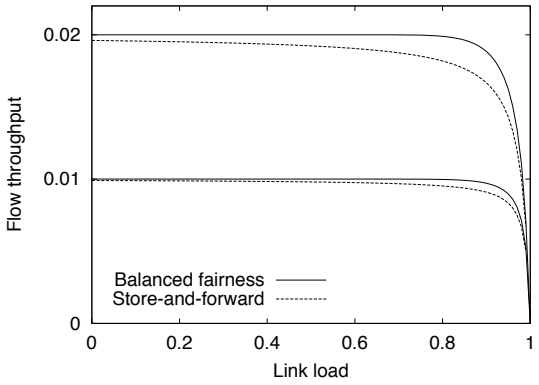

Figure 9: Performance of balanced fairness and store and forward for a single link with two flow rate limits ($a_1 = 0.01,\ a_2 = 0.02$)

7 Conclusion

Insensitivity is key to the development of simple and robust engineering rules for data networks. We have characterized in Theorems 1 and 2 the class of allocations which are insensitive. Balanced fairness refers to the most efficient allocation in this class. While the performance of balanced fairness does not depend on detailed traffic characteristics, it is still a complex function of demand on all routes and of the capacity of all links. This renders the exact evaluation difficult to apply for pratical purposes. The main result of the paper, given in Theorem 4, provides a simple conservative evaluation of the performance of balanced fairness, requiring per-link information only. In particular, links can be dimensioned independently to meet a partial response time target. The response time in a network realizing balanced fairness is guaranteed to be less than the sum of the partial targets on the links of a given route.

An important question that has not been addressed in this paper is how to realize a balanced fair allocation with a distributed congestion control algorithm. Similarly, it remains to evaluate the extent to which the performance of balanced fairness constitutes a good approximation to that realized by existing packet-level mechanisms. Preliminary results from work in progress suggest that accuracy is good and that the store and forward bound is a useful practical tool for dimensioning current data networks.

Appendix

A Insensitive allocations

Proof of Theorem 1. Consider the processor sharing network introduced in §2.3 representing the data network with Poisson flow arrivals and exponential i.i.d. flow sizes, i.e., with $N = K$ nodes and $\nu_i/\mu_i = \rho_i$ for $i = 1, \ldots, N$. We refer to this processor sharing network as the *initial* network. From [4, Theorem 2], the following insensitivity property (P) implies the balance property:

(P) The invariant measures of the Markov process describing the number of customers at each node of the initial network remain unchanged when for any node i and for any α_i, $0 < \alpha_i < 1$, the exponential i.i.d. services at node i are replaced by i.i.d. services, exponentially distributed of mean $1/\alpha_i \times 1/\mu_i$ with probability α_i, null with probability $1 - \alpha_i$.

The proof then follows from the fact that each property (I1), (I2), (I3) implies property (P):

(I1)⇒(P) Consider the initial network with 2-phase services, i.e., with K additional nodes j such that $S_k = \{i, j\}$ for any class k, with modified routing probability $\tilde{p}_{ij} = \alpha_i$ and modified service rates $\tilde{\mu}_i = m \times \mu_i$ and $\tilde{\mu}_j = m/(m-1) \times \alpha_i \mu_i$, for some integer $m > 1$. Letting m tend to infinity, this corresponds to the initial network where the services at any node i are replaced by exponentially distributed services of mean $1/\alpha_i \times 1/\mu_i$ with probability α_i, null services with probability $1 - \alpha_i$.

(I2)⇒(I3) Consider the initial network with K additional infinite server nodes S_0 representing think-times and for any node $i \notin S_0$, modified exogenous arrival rates $\tilde{\nu}_i = \alpha_i \nu_i$ and routing probabilities $\tilde{p}_{ij} = 1 - \alpha_i$ and $\tilde{p}_{ji} = 1$ for some node $j \in S_0$. Letting μ_j tend to infinity for all $j \in S_0$, this corresponds to the initial network where the arrival rates and service rates at any node i are multiplied by the same constant α_i.

(I3)⇒(P) Consider the initial network where the arrival rates and service rates at any node i are multiplied by the same constant α_i, $0 < \alpha_i < 1$. This also corresponds to the initial network with the same arrival rates but where the services at node i are replaced by exponentially distributed services of mean $1/\alpha_i \times 1/\mu_i$ with probability α_i, null services with probability $1 - \alpha_i$.

□

Proof of Theorem 2. As Theorem 2 of [4] holds for any routing probabilities, we conclude as in the proof of Theorem 1 that (I1') implies the balance property. In the following, we prove that (I2') implies the balance property. It can be shown in a very similar way that (I3') implies the balance property.

Consider an allocation for which (I2') holds. When flow arrivals are Poisson, the data network can be modeled as in §2.3 by a processor sharing network where phases of flows of any class k are represented by consecutive nodes $S_k \subset \{1, \ldots, N\}$. Any invariant measure χ of the number of customers at each node of this network satisfies the balance equations (4). Now consider the new processor sharing network obtained by replacing the Poisson arrivals of flows of class 1 by Poisson *session* arrivals with the same flow arrival rate, where each session consists of a geometrically distributed number of flows of mean $1/\alpha_1$, for some parameter α_1, $0 < \alpha_1 < 1$. Letting the tink-time durations tend to zero, this simply corresponds to the initial network with modified arrival rate $\tilde{\nu}_1 = \alpha_1 \nu_1$ at node 1 and routing probabilities $\tilde{p}_{i1} = \alpha_1$, $\tilde{p}_i = (1 - \alpha_1) p_i$ for all nodes $i \in S_1$. From the insensitivity property (I2'), χ satisfies the corresponding

balance equations (4):

$$\chi(y)\left(\alpha_1\nu_1+\sum_{i\notin S_1}\nu_i+\sum_i\psi_i(y)\mu_i\right)=\chi(T_1y)\alpha_1\nu_1+\sum_{i\notin S_1}\chi(T_iy)\nu_i$$
$$+\sum_{i\in S_1}(\chi(T^i_{i+1}y)\psi_i(T^i_{i+1}y)\mu_ip_{i,i+1}+\chi(T^i_1y)\psi_i(T^i_1y)\mu_i\alpha_1)+\sum_{i,j\notin S_1}\chi(T^j_iy)\psi_j(T^j_iy)\mu_jp_{ji}$$
$$+\sum_{i\in S_1}\chi(T^iy)\psi_i(T^iy)\mu_i(1-\alpha_1)p_i+\sum_{i\notin S_1}\chi(T^iy)\psi_i(T^iy)\mu_ip_i.$$

Letting α_1 tend to zero in these equations, observing that for any $i\in S_1$ and any x_1,

$$\sum_{y:\sum_{i'\in S_1}y_{i'}=x_1}\chi(y)\psi_i(y)=\sum_{y:\sum_{i'\in S_1}y_{i'}=x_1}(\chi(T^i_{i+1}y)\psi_i(T^i_{i+1}y)p_{i,i+1}+\chi(T^iy)\psi_i(T^iy)p_i),$$

we obtain:

$$\sum_{y:\sum_{i'\in S_1}y_{i'}=x_1}\chi(y)\sum_{i\notin S_1}(\nu_i+\psi_i(y)\mu_i)=\sum_{i\notin S_1}\sum_{y:\sum_{i'\in S_1}y_{i'}=x_1}\chi(T_iy)\nu_i$$
$$+\sum_{i,j\notin S_1}\sum_{y:\sum_{i'\in S_1}y_{i'}=x_1}\chi(T^j_iy)\psi_j(T^j_iy)\mu_jp_{ji}+\sum_{i\notin S_1}\sum_{y:\sum_{i'\in S_1}y_{i'}=x_1}\chi(T^iy)\psi_i(T^iy)\mu_ip_i.$$

Applying successively the same reasoning to flows of class $2,3,...,K-1$, we prove that, for any fixed $x_1,\ldots,x_{K-1}$, the function $\sum_{y:\sum_{i\in S_k}y_i=x_k,k\neq K}\chi(y)$ is an invariant measure for the number of customers at nodes S_K, with service capacities given by:

$$\forall i\in S_k,\quad \psi_i(y)=\frac{y_i}{x_K}\phi_K(x),\quad x_K=\sum_{i\in S_K}y_i.$$

For any fixed $x_1,\ldots,x_{K-1}$, these service capacities are balanced by the function:

$$\binom{x_K}{y_i,i\in S_K}\prod_{n=0}^{x_K-1}\frac{1}{\phi_K(x-ne_K)}.$$

The corresponding queueing network is a Whittle network, so that:

$$\sum_{y:\sum_{i\in S_k}y_i=x_k,k\neq K}\chi(y)\propto\binom{x_K}{y_i,i\in S_K}\prod_{i\in S_K}\varrho_i^{y_i}\times\prod_{n=0}^{x_K-1}\frac{1}{\phi_K(x-ne_K)}.$$

Summing this expression over all states y such that $\sum_{i\in S_K}y_i=x_K$, we get:

$$\sum_{y:\sum_{i\in S_k}y_i=x_k}\chi(y)\propto\rho_K^{x_K}\prod_{n=0}^{x_K-1}\frac{1}{\phi_K(x-ne_K)}.$$

In particular, the service capacity ϕ_K satisfies (9) for the balance function Φ defined by:

$$\Phi(x)=\frac{\sum_{y:\sum_{i\in S_k}y_i=x_k}\chi(y)}{\prod_{k=1}^K\rho_k^{x_k}}.$$

By symmetry, this property holds for any class k and the allocation is balanced. □

Proof of Proposition 1. From (12), we have:

$$\forall x,\ x_k > 0, \quad \Phi(x) \geq \frac{\Phi(x - e_k)}{x_k a_k}.$$

Using (14), we get:

$$E[x_k] \geq \frac{\rho_k}{a_k}.$$

Similarly, we know from (12) that for any link $l \in r_k$:

$$\forall x, \quad \Phi(x) \geq \frac{1}{C_l} \sum_{k':l \in r_{k'}} \Phi(x - e_{k'}).$$

Using (14), we get:

$$E[x_k] \geq \frac{\rho_k}{C_l} + \frac{1}{C_l} \sum_{k':l \in r_{k'}} \rho_{k'} E[x_k],$$

so that

$$E[x_k] \geq \frac{\rho_k}{C_l - A_l}.$$

The proof follows from (16). □

B Balanced fairness

Proof of Lemma 1. The proof is by induction on the number of flows not in class k, $n' = \sum_{k' \neq k} x_{k'}$. The equality holds for $n' = 0$ as the capacity constaints reduce to that of link l. Now assume it holds for $n' = m$, $m \geq 0$. Let x be any state with $n' = m + 1$ flows not in class k. From (12), we get for any link $l' \neq l$:

$$\frac{1}{C_l} \sum_{k':l \in r_{k'}} \Phi^{\mathrm{BF}}(x - e_{k'}) \geq \frac{1}{C_l} \sum_{k':l \in r_{k'}} \frac{1}{C_{l'}} \sum_{k'':l' \in r_{k''}} \Phi^{\mathrm{BF}}(x - e_{k'} - e_{k''}) = \frac{1}{C_{l'}} \sum_{k'':l' \in r_{k''}} \Phi^{\mathrm{BF}}(x - e_{k''}).$$

From (25), it remains to prove that for any class $k'' \neq k$ such that $x_{k''} > 0$:

$$\frac{1}{C_l} \sum_{k':l \in r_{k'}} \Phi^{\mathrm{BF}}(x - e_{k'}) \geq \frac{1}{x_{k''} a_{k''}} \Phi^{\mathrm{BF}}(x - e_{k''}).$$

If $l \notin r_{k''}$, we have:

$$\frac{1}{C_l} \sum_{k':l \in r_{k'}} \Phi^{\mathrm{BF}}(x - e_{k'}) \geq \frac{1}{C_l} \sum_{k':l \in r_{k'}} \frac{1}{x_{k''} a_{k''}} \Phi^{\mathrm{BF}}(x - e_{k'} - e_{k''}) = \frac{1}{x_{k''} a_{k''}} \Phi^{\mathrm{BF}}(x - e_{k''}).$$

Otherwise, we first consider the case where $x_{k''} > 1$:

$$\begin{aligned} \frac{1}{C_l} \sum_{k':l \in r_{k'}} \Phi^{\mathrm{BF}}(x - e_{k'}) &\geq \frac{1}{C_l} \frac{1}{(x_{k''} - 1) a_{k''}} \Phi^{\mathrm{BF}}(x - 2e_{k''}) + \frac{1}{C_l} \sum_{k' \neq k'':l \in r_{k'}} \frac{1}{x_{k''} a_{k''}} \Phi^{\mathrm{BF}}(x - e_{k'} - e_{k''}) \\ &\geq \frac{1}{C_l} \sum_{k':l \in r_{k'}} \frac{1}{x_{k''} a_{k''}} \Phi^{\mathrm{BF}}(x - e_{k'} - e_{k''}) = \frac{1}{x_{k''} a_{k''}} \Phi^{\mathrm{BF}}(x - e_{k''}). \end{aligned}$$

Now if $x_{k''} = 1$:

$$\frac{1}{C_l} \sum_{k':l \in r_{k'}} \Phi^{\mathrm{BF}}(x - e_{k'}) \geq \frac{1}{C_l} \sum_{k' \neq k'':l \in r_{k'}} \frac{1}{a_{k''}} \Phi^{\mathrm{BF}}(x - e_{k'} - e_{k''}) = \frac{1}{a_{k''}} \Phi^{\mathrm{BF}}(x - e_{k''}).$$

□

Proof of Lemma 2. From Lemma 1,

$$\forall x,\ x_k > 0, \quad \Phi^{\mathrm{BF}}(x) = \frac{1}{C_l} \sum_{k':l \in r_{k'}} \Phi^{\mathrm{BF}}(x - e_{k'}).$$

As in the proof of Proposition 1, we get from (14):

$$E[x_k] = \frac{\rho_k}{C_l} + \frac{1}{C_l} \sum_{k':l \in r_{k'}} \rho_{k'} E[x_k].$$

□

Proof of Theorem 4. From the insensitivity property, we can assume without loss of generality that flows have exponential i.i.d. sizes of unit mean and arrive as independent Poisson processes of intensities $\rho_1, \ldots, \rho_K$. We introduce a new class 0 that shares the same resources as class k, i.e., such that $r_0 = r_k$ and $a_0 = a_k$. The Poisson arrival process of flows of class k in the original network is splitted into two Poisson processes, one of intensity $\varepsilon\rho_k$ for arrivals of flows of class 0 and another of intensity $(1-\varepsilon)\rho_k$ for arrivals of flows of class k, where ε is a fixed parameter, $0 < \varepsilon < 1$. We denote by x_0 the number of flows of class 0 in progress in this new network. As balanced fairness equally shares capacity between flows sharing the same resources, the corresponding balance function is given by:

$$\tilde{\Phi}^{\mathrm{BF}}(x_0, x) = \binom{x_0 + x_k}{x_k} \Phi^{\mathrm{BF}}(x + x_0 e_k).$$

In view of Theorem 3 and expression (14), the corresponding stationary distributions satisfy:

$$\tilde{\pi}^{\mathrm{BF}}(x_0, x) = \binom{x_0 + x_k}{x_k} \varepsilon^{x_0} (1-\varepsilon)^{x_k} \pi^{\mathrm{BF}}(x + x_0 e_k).$$

Thus the steady-state probability that an ongoing flow of class 0 or k is a flow of class 0 is equal to ε and, in view of (16), $\tilde{T}_0^{\mathrm{BF}}(s) = T_k^{\mathrm{BF}}(s)$, i.e., the mean duration of a class-0 flow of size s in the modified network is equal to that of a class-k flow of size s in the original network, independently of ε.

We now consider another balanced allocation where the capacity allocated to flows of class 0 differs from that allocated to flows of class k. We first introduce another modified network where those classes $l \in r_k$ with routes $\hat{r}_l = \{l\}$ and without rate limit are added to existing classes $1, \ldots, K$. We also add a class 0 which is constrained by the rate limit a_k only, i.e., such that $\hat{r}_0 = \emptyset$ and $\hat{a}_0 = a_k$. We denote by $\hat{\Phi}^{\mathrm{BF}}$ the balance function associated with balanced fairness in this new modified network, defined for any state $(\hat{x}, x)$, where $\hat{x} = (\hat{x}_0, \hat{x}_l, l \in r_k)$ gives the number of flows of each new class. If flows of any new class have exponential i.i.d. sizes of unit mean and arrive as independent Poisson processes of same intensity $\varepsilon\rho_k$, it follows from Theorem 3 that the stability condition (13) holds in this new network under the usual traffic conditions (3) and from Lemma 2 that:

$$E[\hat{x}_l] = \frac{\varepsilon\rho_k}{C_l - A_l}, \quad l \in r_k.$$

Now we consider the allocation balanced by the function $\tilde{\Phi}$ defined by:

$$\tilde{\Phi}(x_0, x) = \sum_{|\hat{x}| = x_0} \hat{\Phi}^{\mathrm{BF}}(\hat{x}, x). \tag{26}$$

This function satisfies the capacity constraints (12) for the first modified network where class k is splitted into classes 0 and k. This is immediate for any link $l \notin r_k$. For any link $l \in r_k$, this follows from the inequality:

$$\hat{\Phi}^{\mathrm{BF}}(\hat{x}, x) \geq \frac{1}{C_l}\hat{\Phi}^{\mathrm{BF}}(\hat{x} - \hat{e}_l, x) + \frac{1}{C_l}\sum_{k':l\in r_{k'}} \hat{\Phi}^{\mathrm{BF}}(\hat{x}, x - e_{k'}),$$

where $\hat{e}_l$ denotes the unit vector with 1 in the component corresponding to new class l and 0 elsewhere, and the equality:

$$\sum_{|\hat{x}|=x_0} \hat{\Phi}^{\mathrm{BF}}(\hat{x} - \hat{e}_l, x) = \sum_{|\hat{x}|=x_0-1} \hat{\Phi}^{\mathrm{BF}}(\hat{x}, x).$$

Similarly, the rate limit constraint on class 0 follows from the fact that for any $x_0 > 0$:

$$\hat{\Phi}^{\mathrm{BF}}(\hat{x}, x) \geq \frac{1}{\hat{x}_0 a_k}\hat{\Phi}^{\mathrm{BF}}(\hat{x} - \hat{e}_0, x) \geq \frac{1}{x_0 a_k}\hat{\Phi}^{\mathrm{BF}}(\hat{x} - \hat{e}_0, x), \quad 0 < \hat{x}_0 \leq x_0,$$

and

$$\sum_{|\hat{x}|=x_0} \hat{\Phi}^{\mathrm{BF}}(\hat{x} - \hat{e}_0, x) = \sum_{|\hat{x}|=x_0-1} \hat{\Phi}^{\mathrm{BF}}(\hat{x}, x).$$

We deduce from the above properties that the stability condition (13) holds for this allocation under the usual traffic conditions (3) and:

$$E[x_0] = E[|\hat{x}|] = \frac{\varepsilon\rho_k}{a_k} + \sum_{l\in r_k} \frac{\varepsilon\rho_k}{C_l - A_l}.$$

In view of (16) and (24), $\tilde{T}_0(s) = T_k^{\mathrm{SF}}(s)$, i.e., the mean duration of a class-0 flow in the first modified network for the allocation balanced by $\tilde{\Phi}$ is equal to the mean duration of a class-k flow of size s in the original network for the store and forward allocation, independently of ε.

To conclude the proof, we use Proposition 2:

$$\forall x_0, x, \quad \tilde{\Phi}(x_0, x) \geq \tilde{\Phi}^{\mathrm{BF}}(x_0, x).$$

It follows from (14) that the corresponding stationary distributions satisfy:

$$\forall x_0, x, \quad \frac{\tilde{\pi}(x_0, x)}{\tilde{\pi}_\varepsilon(0)} \geq \frac{\tilde{\pi}^{\mathrm{BF}}(x_0, x)}{\tilde{\pi}^{\mathrm{BF}}(0)},$$

where

$$\tilde{\pi}_\varepsilon(0) = \left(\sum_{x_0,x} (\varepsilon\rho_k)^{x_0}(1-\varepsilon)^{x_k}\tilde{\Phi}(x_0, x)\prod_{k'=1}^{K}\rho_{k'}^{x_{k'}}\right)^{-1} \quad \text{and} \quad \tilde{\pi}^{\mathrm{BF}}(0) = \left(\sum_{x} \Phi^{\mathrm{BF}}(x)\prod_{k'=1}^{K}\rho_{k'}^{x_{k'}}\right)^{-1}.$$

We deduce from (16) that

$$\frac{\tilde{T}_0(s)}{\tilde{\pi}_\varepsilon(0)} \geq \frac{\tilde{T}_0^{\mathrm{BF}}(s)}{\tilde{\pi}^{\mathrm{BF}}(0)}.$$

The proof then follows from the fact that $\lim_{\varepsilon\to 0}\tilde{\pi}_\varepsilon(0) = \tilde{\pi}^{\mathrm{BF}}(0)$. □

References

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

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

[3] T. Bonald and L. Massoulié, Impact of fairness on Internet performance, in: *Proc. of ACM SIGMETRICS*, 2001.

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

[5] T. Bonald and A. Proutière, Insensitive bandwidth sharing, in: *Proc. of IEEE GLOBECOM*, 2002.

[6] T. Bonald and A. Proutière, Insensitive bandwidth sharing in data networks, to appear in Queueing Systems, 2003.

[7] T. Bonald, A. Proutière, G. Régnié, J.W. Roberts, Insensitivity results in statistical bandwidth sharing, in: *Proc. of ITC 17th*, 2001.

[8] T. Bonald, A. Proutière, J. Roberts, J. Virtamo, Computational aspects of balanced fairness, submitted, 2003.

[9] M. Crovella and A. Bestravos, Self-similarity in WWW traffic: Evidence and possible cause, in: *Proc. of ACM SIGMETRICS*, 1996.

[10] A.K. Erlang, Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges, Elektroteknikeren 13, 1917.

[11] G. Fayolle, A. de la Fortelle, J-M. Lasgouttes, L. Massoulié, and J.W. Roberts, Best-effort networks: Modeling and performance analysis via large networks asymptotics, in: *Proc. of IEEE INFOCOM*, 2001.

[12] V. Jacobson, Congestion avoidance and control, in: *Proc. of ACM SIGCOMM*, 1988.

[13] R. Johari and D. Tan, End-to-end congestion control for the Internet: delays and stability, ACM/IEEE Transactions of Networking 9-6 (2001) 818–832.

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

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

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

[17] R.J. La and V. Anantharam, Charge-sensitive TCP and rate control in the Internet, in: *Proc. of IEEE INFOCOM*, 2000.

[18] L. Massoulié, Stability of distributed congestion control with heterogeneous feedback delays, IEEE Transactions on Automatic Control 47-6 (2002) 895–902.

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

[20] L. Massoulié and J.W. Roberts, Bandwidth sharing: Objectives and algorithms, IEEE/ACM Transactions on Networking 10-3 (2002) 320–328.

[21] J. Mo and J. Walrand, Fair end-to-end window-based congestion control, IEEE/ACM Transactions on Networking 8-5 (2000) 556–567.

[22] V. Paxson and S. Floyd, Difficulties in Simulating the Internet, IEEE/ACM Transactions on Networking 9-4 (2001) 392–403.

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