

On Light and Heavy Traffic Approximations of Balanced Fairness

Thomas Bonald, Aleksi Penttinen, Jorma Virtamo

▶ **To cite this version:**

Thomas Bonald, Aleksi Penttinen, Jorma Virtamo. On Light and Heavy Traffic Approximations of Balanced Fairness. ACM Sigmetrics / Performance, 2006, Saint Malo, France. hal-01277612

HAL Id: hal-01277612
https://inria.hal.science/hal-01277612

Submitted on 22 Feb 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On Light and Heavy Traffic Approximations of Balanced Fairness

Thomas Bonald
France Telecom R&D
France

Aleksi Penttinen
Helsinki Univ. of Technology
Finland

Jorma Virtamo
Helsinki Univ. of Technology
Finland

ABSTRACT

Flow level analysis of communication networks with multiple shared resources is generally difficult. A recently introduced sharing scheme called balanced fairness has brought these systems within the realm of tractability. While straightforward in principle, the numerical evaluation of practically interesting performance metrics like per-flow throughput is feasible for limited state spaces only, besides some specific networks where the results are explicit. In the present paper, we study the behaviour of balanced fairness in light and heavy traffic regimes and show how the corresponding performance results can be used to approximate the flow throughput over the whole load range. The results apply to any network, with a state space of arbitrary dimension. A few examples are explicitly worked out to illustrate the concepts.

1. INTRODUCTION

One of the current trends in telecommunications is leading towards all-IP networks in which most of the traffic consists of elastic flows. This means that flows, typically file transfers, adapt their transmission speed to the available capacity. Flow level analysis of such networks is usually difficult due to the dynamic nature of traffic. Flows arrive at random and leave the network once the corresponding file transfer is completed. The bandwidth shares between concurrent flows are adjusted at each flow arrival and departure. Such networks prove very difficult to analyse in practice even for the simplest topologies [1].

In this context, a sharing scheme called balanced fairness has proved to be a very interesting concept, making flow level analysis tractable [2, 3, 4, 5]. Moreover, the performance of any network whose resources are shared according to balanced fairness depends on traffic characteristics through the traffic intensity only.

While balanced fairness was originally introduced for wired networks with single-path routing, the basic principle can be extended to any network, including multi-hop wireless networks [6], multi-path routing [7] and cellular networks with soft handover [8]. A survey with a collection of various examples is presented in [8].

Besides of being an interesting concept in its own, balanced fairness can also be viewed as an approximation tool. Namely, it turns out that the performance of many other resource sharing schemes like max-min fairness [9] and proportional fairness [10] is often similar to that of balanced fairness [2]. The performance under balanced fairness is much easier to analyse than under other sharing schemes. This is due to the reversibility of the underlying Markov process, which allows one to derive the equilibrium distribution by a simple recursion.

While balanced fairness leads to explicit performance results for some network topologies, including the practically interesting case of trees [5], one must turn to numerical analysis in most cases. Obtaining accurate results is then difficult for heavily loaded systems as the number of states influencing the results grows quickly. Even at moderate load, the recursion easily becomes prohibitive when the dimension of the state space increases.

In the present paper, we study the behaviour of balanced fairness in light and heavy traffic regimes for a network with an arbitrary state space. We derive some simple, easy-to-calculate results for a key performance metric referred to as the *flow throughput*. Specifically, we obtain the derivative of the flow throughput with respect to the network load both at zero load and at maximum load, when the capacity limit is reached.

With the above information at both ends of the load range, the overall behaviour of the throughput curve becomes more or less defined. We use a simple rational interpolation function, the form of which is suggested by the known solvable cases, to join the extreme load regimes, resulting in an overall approximation for the whole range. We emphasize that this approximation can be obtained with little effort for basically any system, and thus believe it is a useful tool for flow level performance evaluation in practice.

The light traffic analysis simply follows from the truncation of the recursion to states where the total number of flows is less than or equal to two. The heavy traffic analysis is more complex. We need to distinguish between two cases,

depending on whether the set of feasible allocation vectors, the so-called capacity set, is bounded by linear constraints or whether its boundary is a smooth surface. We use the store-and-forward bound [3] in the former case, an approximation scheme based on the so-called L^α-ellipsoid in the latter case. Several examples are worked out to demonstrate the accuracy of the approximation scheme.

While the main body of the paper deals with networks of finite dimension, we extend the concept of balanced fairness and the corresponding light traffic approximation to continuous systems where the number of classes is infinite. As an example we consider a two-base-station cellular system where mobiles are located anywhere on the line segment between the two base stations. Such a system defies exact analysis, thus again demonstrating the usefulness of the approximation scheme.

The rest of the paper is organized as follows. In Section 2 we present the network model and summarize the basics of balanced fairness. Sections 3 and 4 are devoted to the light and heavy traffic analysis, respectively. We show in the following section how these results can be combined to approximate the flow throughput over the whole load range. In Section 6 we extend the results to continuous systems. We conclude in Section 7.

2. BALANCED FAIRNESS

The notion of balanced fairness was introduced by Bonald and Proutière as a means to approximately evaluate the performance of fair allocations like max-min fairness and proportional fairness in wired networks [2]. We here define balanced fairness in the usual form for a finite set of discrete flow classes $i = 1, \ldots, N$. In Section 6, the concept is extended to the setting of continuous class indices.

2.1 Network model

A flow class represents a set of similar flows in terms of network resource usage. In a wired network for instance, a flow class is characterized by a set of links representing a path in the network. Let x_i be the number of class-i flows in progress and denote by $x = (x_1, \ldots, x_N)$ the network state. We use the notation $|x| = \sum_i x_i$.

Class-i flows arrive stochastically and have finite, random sizes. We denote by ρ_i the traffic intensity of class-i flows, defined as the product of the flow arrival rate and the mean flow size (in bit/s). We denote by $\rho = (\rho_1, \ldots, \rho_N)$ the vector of traffic intensities.

In state x class-i flows share evenly a total rate of $\phi_i(x)$ bit/s, with the convention $\phi_i(x) = 0$ if $x_i = 0$. We denote by $\phi(x) = (\phi_1(x), \ldots, \phi_N(x))$ the vector of allocated rates. In all states x, this vector belongs to some coordinate convex capacity set $\mathcal{C}$ that represents the physical constraints of the network. In many practically interesting examples, the capacity set is a polytope, that is

$$\mathcal{C} = \{\phi\,, \phi A \leq C\}, \tag{1}$$

for some $N \times L$-dimensional non-negative matrix A and L-dimensional positive vector C [8]. For wired networks with single-path routing for instance, the vector C gives the capacity of the L network links and the matrix A is the incidence matrix, that is the i, l-entry of A is equal to 1 if class-i flows go through link l and equal to zero otherwise.

2.2 Resource allocation

An allocation is said to be balanced if for all states x and all classes $i, j = 1, \ldots, N$:

$$\phi_i(x)\phi_j(x - e_i) = \phi_i(x - e_j)\phi_j(x), \tag{2}$$

where e_i represents the N-dimensional unit vector whose components are equal to 0 except for component i which is equal to 1. For Poisson flow arrivals and i.i.d. exponential flow sizes, this balance property is equivalent to the reversibility of the Markov process describing the evolution of the network state. The invariant measure of this Markov process is then given by:

$$\pi(x) = \Phi(x)\rho_1^{x_1} \ldots \rho_N^{x_N}, \tag{3}$$

where $\Phi(x)$ is the inverse of the product of allocated rates along any direct path from state x to state 0, which in view of (2) does not depend on the considered path. This so-called balance function is recursively defined by $\Phi(0) = 1$ and for all states x such that $x_i > 0$,

$$\Phi(x) = \frac{\Phi(x - e_i)}{\phi_i(x)}. \tag{4}$$

It should be emphasized that all results are valid under non-Markovian assumptions. The invariant measure (3) is insensitive to all traffic characteristics beyond the vector of traffic intensities ρ provided flows are generated within sessions. A session consists of a finite, random number of flows separated by intervals of inactivity referred to as 'think times'. The number of flows per session, flow sizes and think times may have arbitrary distributions. There may be correlation between these random variables. The only required assumption is that sessions arrive as a Poisson process, see [2] for details.

There is a continuum of allocations that satisfy the balance property (2). There is a single allocation such that $\phi(x)$ belongs to the boundary of the capacity set $\mathcal{C}$ in any state $x \neq 0$. This is balanced fairness. Under balanced fairness, the invariant measure (3) has a finite sum provided ρ belongs to the interior of the capacity set $\mathcal{C}$ [8, 2]. When the capacity set $\mathcal{C}$ is a polytope (1), this is equivalent to the strict component-wise inequality $\rho A < C$.

In the rest of the paper, we assume that this stability condition is satisfied and denote by π the equilibrium distribution of the network state:

$$\pi(x) = \frac{1}{G(\rho)}\Phi(x)\rho_1^{x_1} \ldots \rho_N^{x_N}, \tag{5}$$

where $G(\rho)$ is the normalization constant,

$$G(\rho) = \sum_x \Phi(x)\rho_1^{x_1} \ldots \rho_N^{x_N}. \tag{6}$$

Rewriting (4) as:

$$\phi_i(x) = \frac{\Phi(x - e_i)}{\Phi(x)}, \tag{7}$$

we deduce that the balance function associated with balanced fairness is recursively defined by

$$\Phi(x) = \min\{\alpha : \frac{(\Phi(x-e_1), \ldots, \Phi(x-e_N))}{\alpha} \in \mathcal{C}\}, \quad (8)$$

with the convention $\Phi(0) = 1$ and $\Phi(x) = 0$ for all $x \notin \mathbb{Z}_+^N$. Note that this recursion defines a unique balance function, which in turn defines a unique allocation in view of (7). For a polytope capacity set, there are L capacity constraints and the recursion takes the following simple form:

$$\Phi(x) = \min_{l=1,\ldots,L} \frac{1}{C_l} \sum_{i=1}^{N} \Phi(x-e_i) A_{il}.$$

2.3 Scaling transformation

For any capacity set $\mathcal{C}$, let:

$$a_i = \max\{\alpha : \alpha e_i \in \mathcal{C}\}, \quad i = 1, \ldots, N. \quad (9)$$

By the convexity of $\mathcal{C}$, this is the maximum rate allocated to class-i flows. Since $\phi(x)$ belongs to the border of the capacity set, this is also the rate allocated to class-i flows when there are no other flows in the network.

Now consider the capacity set obtained by scaling the i-axis by the factor $1/a_i$. The associated balance function $\varphi(x)$, referred to as the scale-free balance function, is given by:

$$\varphi(x) = \Phi(x) a_1^{x_1} \cdots a_N^{x_N}, \quad (10)$$

where $\Phi(x)$ denotes the original balance function. The scale-free balance function satisfies:

$$\forall n \geq 0, \quad \varphi(ne_i) = 1, \quad i = 1, \ldots, N.$$

In view of (7) and (10), we have in all states $x \neq 0$:

$$\phi_i(x) = a_i \times \frac{\varphi(x-e_i)}{\varphi(x)}.$$

2.4 Flow throughput

A key performance metric is the flow throughput, defined as the ratio of the mean flow size to the mean flow duration. This may be viewed as the *equivalent bandwidth* as perceived by users. By Little's result, the flow throughput γ_i of class-i flows is given by

$$\gamma_i = \frac{\rho_i}{\mathrm{E}[x_i]}.$$

In view of (5) and (6), we have:

$$\mathrm{E}[x_i] = \sum_x x_i \pi(x) = \frac{\rho_i}{G(\rho)} \frac{\partial}{\partial \rho_i} G(\rho).$$

We deduce:

$$\gamma_i = \frac{G(\rho)}{\frac{\partial}{\partial \rho_i} G(\rho)} \quad . \quad (11)$$

Recalling that a_i is the maximum bit rate allocated to class-i flows, we have $\gamma_i \leq a_i$ for all $i = 1, \ldots, N$.

3. LIGHT TRAFFIC REGIME

As noted in Section 1, the direct evaluation of the flow throughput from the recursion (8) may be computationally prohibitive for networks of high dimension N. This section is devoted to the light traffic analysis.

3.1 Low load expansion

We use the recursion (8) to calculate the normalization constant (6) up to the second order (i.e., only the states up to occupation $|x| = 2$ are taken into account). Using the scale-free balance function $\varphi(x)$ discussed above, recalling that $\varphi(0) = \varphi(e_i) = 1$, and introducing the shorthand notation $\varphi_{ij} = \varphi(e_i + e_j)$, we have

$$G(\rho) = 1 + \sum_i \frac{\rho_i}{a_i} + \sum_{i \geq j} \varphi_{ij} \frac{\rho_i \rho_j}{a_i a_j} + \ldots, \quad (12)$$

where a_i is defined by (9), $\varphi_{ii} = 1$ and, in view of (8),

$$\varphi_{ij} = \min\{\alpha : \frac{a_i e_i + a_j e_j}{\alpha} \in \mathcal{C}\} \quad \forall i \neq j. \quad (13)$$

The parameter φ_{ij} can be described as the contraction factor needed to bring an $a_i \times a_j$ rectangle inside $\mathcal{C}$, as depicted in Figure 1. Since $\mathcal{C}$ is convex, we have $1 \leq \varphi_{ij} \leq 2$ for all i, j.

In the case of a polytope capacity set (1), expressions (9) and (13) take the forms

$$\begin{cases} a_i &= \min\limits_{l=1,\ldots,L} \dfrac{C_l}{A_{il}}, \\ \varphi_{ij} &= \max\limits_{l=1,\ldots,L} \dfrac{a_i A_{il} + a_j A_{jl}}{C_l} \quad \forall i \neq j. \end{cases} \quad (14)$$

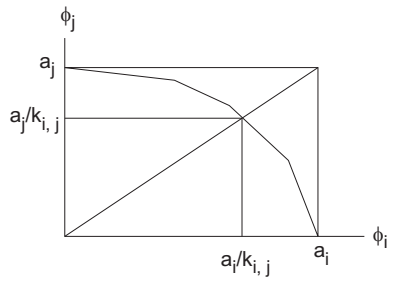


Figure 1: The contraction factor φ_{ij}.

In order to study the behaviour of the flow throughput we define a load line with a given traffic profile $p = (p_1, \ldots, p_N)$, with $|p| = 1$. By this we mean the set of all $\rho \in \mathcal{C}$ such that the proportions of the loads in different classes are given by p. Let $\hat{\rho}$ be the end point of the load line on the boundary of $\mathcal{C}$. Then the load line consists of points $r\hat{\rho}$, where $r \in [0, 1]$. We wish to characterize the function $\gamma_i(r)$ on a given load line. Using (11) and (12) we easily find

$$\begin{cases} \gamma_i(0) &= a_i, \\ \gamma_i'(0) &= -\Big(\hat{\rho}_i + a_i \sum\limits_{j \neq i} \hat{\rho}_j \dfrac{\varphi_{ij} - 1}{a_j}\Big). \end{cases} \quad (15)$$

It is worth noting that in order to calculate this derivative what is essentially only needed are the values of the φ_{ij} and these can easily be found by the pairwise consideration of (13) as illustrated in Figure 1.

3.2 Examples

Star network

Consider the star network of Figure 2, which is topologically equivalent to a so-called hypercycle [2]. Its capacity set, also

illustrated in the figure, is a polytope with

$$C = (1, 1, 1), \qquad A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

While no exact solution for this system is known, the directional derivative (15) is easily calculated. Eq. (14) gives $a_i = 1$ for all i and $\varphi_{ij} = 2$ for $i \neq j$. Then, for instance with equal loads, $\hat{\rho}_i = 1/2$, we obtain $\gamma_i'(0) = -3/2$ for all i. This example is further worked out in §5.2.

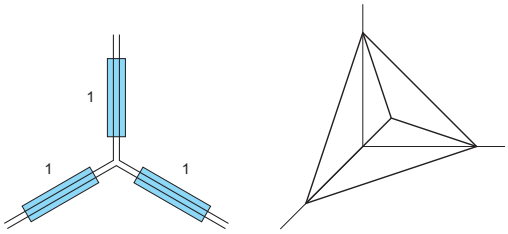


Figure 2: Star network and its capacity set.

Parking lot network

As another example consider the parking lot network consisting of two links with capacities $C_1 = 2$ and $C_2 = 1$ and two flow classes. Class-1 flows go through link 1 only while class-2 flows go through both links. The topology and the capacity set of the system are illustrated in Figure 3. Six load lines representing different traffic profiles $(1 - p_2, p_2)$, with $p_2 = 0,\ 0.1, \ldots, 0.5$, are also shown. Again, the capacity set is a polytope with

$$C = (2, 1), \qquad A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

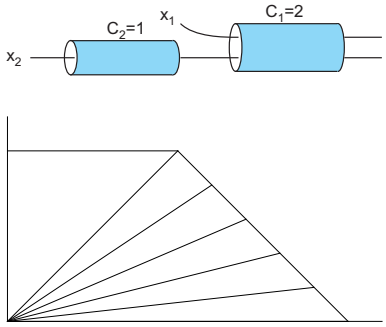


Figure 3: Parking lot network and its capacity set with radial load lines.

Eq. (14) gives $a_1 = 2$, $a_2 = 1$ and $\varphi_{12} = 3/2$. It then follows from (15) that $\gamma_1'(0) = -2$ and $\gamma_2'(0) = -(1 + 3p_2)/2$ for all $p_2 \in (0, 1/2)$. In fact, a general closed-form expression for the flow throughput is known for this type of network [4]. In particular, the class-2 flow throughput is given by:

$$\begin{aligned} \gamma_2 &= \left(\tfrac{1}{(1-\rho_2)(2-\rho_2)} + \tfrac{1}{2-(\rho_1+\rho_2)} \right)^{-1} \\ &= \tfrac{(2-(\rho_1+\rho_2))(2-3\rho_2-\rho_2^2)}{(2-\rho_2)^2-\rho_1}. \end{aligned} \tag{16}$$

It is easy to verify that the directional derivatives at zero load agree with the above simple expressions. The complete throughput curves along the load lines of Figure 3 are plotted in Figure 4 (the lowest curve corresponds to the upper load line where $p_2 = 1/2$). The derivatives as given by $\gamma_2'(0) = -(1 + 3p_2)/2$ are (from top to bottom) -0.5, -0.65, -0.8, -0.95, -1.1, -1.25, in agreement with the figure. Note that this directional derivative formula is much easier to find than the full expression (16).

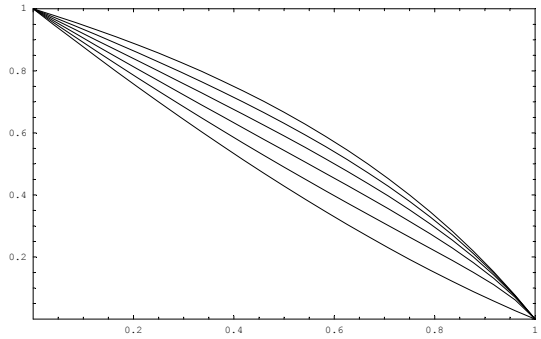


Figure 4: Throughput along the load lines of Fig. 3.

Tree network

As a third example consider a more complex access network of tree topology depicted in Figure 5. The networks consists of 10 links and 9 flow classes (numbered according to the corresponding access link; class 3 does not exist). The capacity set is again a polytope; the corresponding capacity vector C and incidence matrix A follow similarly as in previous examples.

We wish to study the class-10 throughput along the load line with equal traffic loads. The link that first becomes saturated is link 3. This happens at $\hat{\rho}_i = \frac{12}{7}$ for all i. A mechanical calculation using Eqs. (14) and (15) then yields $\gamma_{10}(0) = 3$ and $\gamma_{10}'(0) = -\frac{141}{64}$. We will also return to this example in §5.2.

4. HEAVY TRAFFIC REGIME

We now turn to the heavy traffic analysis. Our aim here is to derive the derivative of the flow throughput along a load line at the capacity limit $\hat{\rho}$. We have to distinguish between two cases depending on whether the boundary of the capacity set is a polytope or a smooth surface.

4.1 A polytope capacity set

A polytope capacity set (1) arises when flows are contending for L resources with respective capacities $C_1, \ldots, C_L$, and each unit of bandwidth allocated to class-i flows requires A_{il} units of capacity from resource l. As shown in [8], this model describes a rich variety of communication networks.

Let $\hat{\rho}$ be a load vector on the boundary of the capacity set and denote by $\hat{\varrho} = \hat{\rho}A$ the corresponding resource demand vector. Note that $\hat{\varrho}_l = C_l$ for all l in some non-empty set $\mathcal{L}$ of resource indices (otherwise, $\hat{\rho}$ wouldn't be on the boundary of the capacity set). As in the light traffic analysis, we denote by $\gamma_i(r)$ the class-i flow throughput on the load line $r\hat{\rho}$, where $r \in [0, 1]$. We are interested in the derivative $\gamma_i'(1)$ for those classes i such that $\gamma_i(1) = 0$.

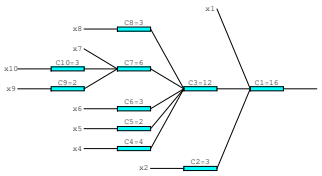


Figure 5: An access network with tree topology.

A single saturated resource

If $\mathcal{L}$ reduces to a single resource l, we have $\gamma_i(1) = 0$ if and only if class-i flows use resource l, that is if $A_{il} > 0$. For these classes, the heavy traffic regime is then determined by that resource only, in the sense that:

$$\gamma_i(r) \sim \frac{C_l}{A_{il}}(1-r), \tag{17}$$

when $r \to 1$, from which we deduce:

$$\gamma_i'(1) = -\frac{C_l}{A_{il}}.$$

This is the result one would obtain if there were no other resource constraint than l. Note that the derivative depends on the direction $\hat{\rho}$ through the saturated constraint l only.

The asymptotic result (17) is a consequence of the following general bounds:

$$\max_{k=1,\ldots,L} \frac{A_{ik}}{C_k - \varrho_k} \le \gamma_i^{-1} \le \sum_{k=1}^{L} \frac{A_{ik}}{C_k - \varrho_k}. \tag{18}$$

Recall that γ_i^{-1} corresponds to the ratio of the mean duration to the mean size for class-i flows. The left-hand side inequality means that the mean flow duration is higher than if there were no other resource constraint than k, for all $k = 1, \ldots, L$; the presence of other resources make data transfers longer. The right-hand side inequality means that the mean flow duration is less than the sum of the mean flow durations due to each individual resource constraint. We refer to this inequality as the store-and-forward bound since it was originally derived in [3] in the context of wired networks with single-path routing: the mean flow duration increases when flows are successively transmitted, in a store-and-forward way, on each link of their path in the network. While the bounds (18) are proved in [3] in the particular case where the elements of the matrix A are equal to 0 or 1, the same inequalities are satisfied for a general matrix A. The proof is essentially the same as in [3].

The resource demand vector is equal to $r\hat{\varrho}$ as a function of r, where $r \in [0, 1]$. If $\mathcal{L}$ reduces to a single resource l, it follows from (18) that

$$\frac{A_{il}}{C_l(1-r)} \le \gamma_i^{-1} \le \frac{A_{il}}{C_l(1-r)} + \sum_{k \ne l} \frac{A_{ik}}{C_k - r\hat{\varrho}_k}.$$

where we used the fact that $\hat{\varrho}_l = C_l$. The asymptotic result (17) then follows from the fact that $\hat{\varrho}_k \ne C_k$ for all $k \ne l$.

Several saturated resources

When the set $\mathcal{L}$ consists of more than one resource, we have $\gamma_i(1) = 0$ if and only if class-i flows use at least one resource in $\mathcal{L}$, that is if $A_{il} > 0$ for some $l \in \mathcal{L}$. While the asymptotic values of the bounds (18) do not coincide, it is conjectured that the heavy traffic regime is given by the store-and-forward bound, namely

$$\gamma_i(r) \sim \frac{1-r}{\sum_{l \in \mathcal{L}} A_{il}/C_l},$$

from which we deduce:

$$\gamma_i'(1) = -\frac{1}{\sum_{l \in \mathcal{L}} A_{il}/C_l}. \tag{19}$$

This is suggested by the result of Schweitzer [13] who proved that the throughputs in multiclass closed networks with a large number of customers, which correspond to the store-and-forward model in heavy traffic, are given by the proportional fair allocation and the recent results of Massoulié [14] showing the asymptotic equivalence of proportional fairness and balanced fairness. A check with the examples presented in [5, 8] confirms that the heavy traffic regime is indeed given by expression (19).

Again, the derivative (19) depends on the direction $\hat{\rho}$ through the set of saturated constraints $\mathcal{L}$ only, i.e. the derivatives have constant values on any given facet of the polytope, similarly on any given edge between two facets, etc. In particular, the flow throughput is not a continuous function of the load vector $\hat{\rho}$: in approaching the boundary of the capacity set, the flow throughput behaves differently depending on how many constraints are saturated at $\hat{\rho}$. Thus the directional derivative of the flow throughput has discontinuities at edges of the boundary where two or more facets of the polytope meet.

This behaviour is demonstrated by the parking lot example considered in §3.2. From Figure 4 we see that indeed the directional derivative at the capacity limit is -2 for all the load lines that end at the line representing the capacity constraint of link 1, whereas the directional derivative at the capacity limit is -2/3 for the load line ending at the corner point where both links are saturated, i.e. precisely the values predicted by (19). Similarly, one could find that for load lines ending at the horizontal line representing the

capacity constraint of link 2, have the directional derivative -1 at the end point. No matter how close the load line is to the diagonal one, at the capacity limit, i.e. at the very end of the line, the system only 'sees' a single constraint.

4.2 A capacity set with smooth boundary

When the capacity set is constrained by a smooth surface instead of a polytope, virtually no exact result is known. Since on a facet of a polytope the directional derivative depends only on the facet plane itself and any smooth surface can locally be approximated by its tangent plane, it is tempting to think that the directional derivative of the throughput at any point $\hat{\rho}$ on the boundary is determined as soon as the tangent plane at $\hat{\rho}$ is known. This thinking, however, is fallacious. The difficulty becomes obvious if one tries to approximate the boundary by a polytope and consider the limit when the polytope becomes denser. Depending on whether $\hat{\rho}$ is on a facet of the approximating polytope or at an edge of two or more facets, one gets a different result. There seems to be no simple way to resolve this ambiguity in the limit process.

There is, however, a special system with curved boundary that yields to exact analysis, viz. the case where the boundary of the capacity set is a N-dimensional L^α-ellipsoid. In Appendix A we derive the directional derivative of throughput at the capacity limit for the corresponding scale-free L^α-sphere. Here we present the result in the general case of an ellipsoid, obtained by the scaling transformation of §2.3. The result is very useful as any smooth surface at a given point can be approximated up to second order by an L^α-ellipsoid, so that not only the tangent plane but also the curvature (tensor) of the approximating surface matches that of the original surface. This is utilized in our approximation scheme in Section 5.

The L^α-ellipsoid is defined by the surface

$$\left(\frac{\phi_1}{a_1}\right)^\alpha + \cdots + \left(\frac{\phi_N}{a_N}\right)^\alpha = 1, \quad \alpha > 1,$$

where the a_i, $i = 1, \ldots, N$, are positive scaling constants and α parameterizes the curvature. The asymptotic derivative is obtained by rescaling from (31),

$$\gamma_i = \frac{2\,a_i}{(N+1)\alpha - (N-1)} \, \frac{1 - ((\frac{\rho_1}{a_1})^\alpha + \cdots + (\frac{\rho_N}{a_N})^\alpha)}{(\frac{\rho_i}{a_i})^{\alpha-1}}. \quad (20)$$

Now consider the directional derivative of γ_i along the load line defined by the traffic profile $(p_1, \ldots, p_N)$, i.e.,

$$\hat{\rho}_i = \left(\frac{p_i^\alpha}{(\frac{p_1}{a_1})^\alpha + \cdots + (\frac{p_N}{a_N})^\alpha}\right)^{1/\alpha}.$$

Substituting $\rho = r\hat{\rho}$ into expression (20) and forming the derivative with respect to r yields

$$\gamma_i'(1) = -\frac{2\,\alpha\,a_i}{(N+1)\alpha - (N-1)} \left(\frac{(\frac{p_1}{a_1})^\alpha + \cdots + (\frac{p_N}{a_N})^\alpha}{(\frac{p_i}{a_i})^\alpha}\right)^{\frac{\alpha-1}{\alpha}}. \quad (21)$$

5. JOINING THE LIGHT AND HEAVY TRAFFIC REGIMES

In this section we demonstrate how the above light and heavy traffic results can be used to find approximate throughput function along given load line, $\rho = r\hat{\rho}$, $r \in (0, 1)$, for a system with a capacity set bounded by either a polytope or an arbitrary smooth convex curve. As examples we consider the star and tree networks discussed in §3.2, as well as a system where a single sender transmits simultaneously to two recipients, assuming the information theoretic capacity set of the broadcast channel [11].

5.1 Interpolation scheme

There are two steps in our approach. First, we find the low-load throughput $\gamma_i(0)$ and its derivative $\gamma_i'(0)$ using (15) and the derivative at capacity limit $\gamma_i'(1)$ a) either using (19), when the capacity set is a polytope, or b) in the case of a smooth boundary, by fitting the parameters $a_1, \ldots, a_N$ and α of the L^α-ellipsoid so that $\hat{\rho}$ is on the ellipsoid and that the first and second derivatives of the ellipsoid surface match those of the boundary of the original capacity set (see Figure 8), and then using (21) for the throughput derivative.

In the second step, having found $\gamma_i(0)$, $\gamma_i'(0)$ and $\gamma_i'(1)$, we fit a rational expression of the form

$$\tilde{\gamma}_i(r) = \gamma_i(0)(1-r)\,\frac{1 + c\,r}{1 + d\,r}, \quad (22)$$

where c and d are free parameters, to give the right derivatives $\gamma_i'(0)$ and $\gamma_i'(1)$. The form of this function is motivated by the observation that in all explicitly solvable cases with a polytope capacity set, e.g. trees [5], the throughput is given by a rational expression where the degree of the numerator is higher by one than that of the denominator.

5.2 Examples

Star network

Returning to the star network example of Figure 2 we consider the throughput along the load line with the traffic profile $p = \frac{1}{3}(1, 1, 1)$. In §3.2 we already found $\gamma_i(0) = 1$ and $\gamma_i'(0) = -3/2$ for all i. As the capacity set is a polytope, the derivative at the capacity limit (all three constraints saturated at $\hat{\rho}$) is determined by (19) giving $\gamma_i'(0) = -\frac{1}{2}$ for all i. Fitting the parameters c and d in (22) yields the simple parabola $\tilde{\gamma}_i(r) = (1-r)(1-r/2)$. This is shown in Figure 6 along with values obtained from the numerical solution of the recursion (8). We find that the approximation obtained with virtually no work fits very well with the actual data.

Tree network

Next we come back to the tree network of Figure 5 assuming a uniform traffic profile. In §3.2 we already calculated $\gamma_{10}(0) = 3$ and $\gamma_{10}'(0) = -\frac{141}{64}$. At the capacity limit with $\hat{\rho}_i = \frac{12}{7}$ for all i, link 3 solely is saturated. Then (19) immediately gives $\gamma_{10}'(1) = -12$. The same fitting procedure as before yields the rational approximation shown in Figure 7 (upper curve). The lower curve represents the exact result that is available for all tree networks [5] but with a heavier computational procedure. The approximation is again good, though there is a slight deviation in the heavy traffic regime.

The reason for this deviation is basically the fact that even though link 3 only is saturated at the capacity limit, also

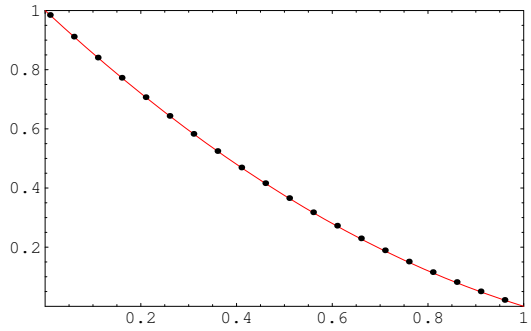


Figure 6: Approximate throughput of the star network with equal loads (solid line) and corresponding numerically evaluated values (dots).

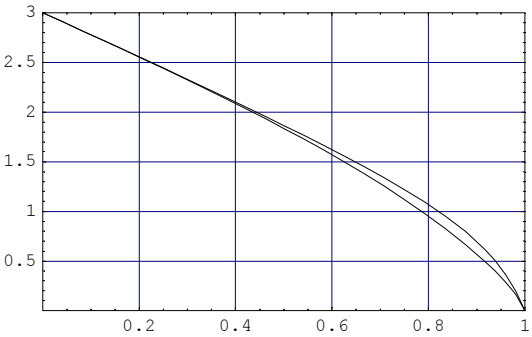


Figure 7: Approximate and exact class-10 throughput of the tree network with equal loads (upper and lower curves, respectively).

link 1 is close to saturation: its load is $9 \cdot \frac{12}{7} \approx 15.4$ while its capacity is $C_1 = 16$. In other words, the considered load line ends on a facet of the polytope at a point that is close to an edge. In such cases the directional derivative at the capacity limit is not a very good descriptor of the overall behaviour of the throughput curve. This is obvious also in the throughput curves of the parking lot configuration in Figure 4. Though the derivative given by (19) is correct at the very end of the curve, the curve bends rather sharply and soon, with a load slightly below the capacity limit, the derivative is closer a value predicted by (19) assuming all the constraints active at the edge to be saturated. In the present example, if we do assume that also link 1 was saturated, then (19) gives $\gamma'_{10}(1) = -1/(\frac{1}{12} + \frac{1}{16}) = -\frac{48}{7}$. Using this value in the fitting renders the match with the exact curve almost perfect (not shown in the figure).

Broadcast channel

As a third example consider the broadcast channel. A base station transmits files simultaneously to two mobiles, 1 and 2. We consider Gaussian radio channels and assume mobile 1 has a better radio channel than mobile 2 in the sense that the corresponding noise powers N_1, N_2 satisfy $N_1 < N_2$. In theory, data may be coded in such a way that mobile 1 can cancel the interference caused by the transmission to mobile 2 [11]. This leads to the ideal transmission rates:

$$\begin{cases} \phi_1 &= W \log_2 \left(1 + \frac{P_1}{N_1}\right), \\ \phi_2 &= W \log_2 \left(1 + \frac{P_2}{N_2 + P_1}\right), \end{cases}$$

where W is the bandwidth of the channel (in Hz) and P_1, P_2 are the transmission powers to mobiles 1 and 2, whose sum is constrained by the total transmission power P of the base station. The corresponding capacity set $\mathcal{C}$ is coordinate convex. We take $W = 0.5$, $P/N_1 = 30$ dB and $P/N_2 = 0$ dB in the following.

The corresponding capacity set $\mathcal{C}$ is illustrated in Figure 8. We approximate the flow throughputs along the load line with the traffic profile $p = (\frac{10}{11}, \frac{1}{11})$, i.e. the load of mobile 1 is 10 times as large as that of mobile 2. In this case the load vector at the capacity limit is $\hat{\rho} = (3.77, 0.377)$.

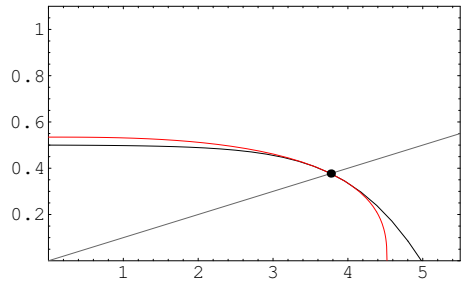


Figure 8: Capacity set (the lower curve at zero class-1 capacity), load line with $\rho_1/\rho_2 = 10$ and the L^α-ellipsoid approximation at the capacity limit.

From the capacity set we get the values $\gamma_1(0) = 4.984$, $\gamma_2(0) = 0.5$. The light-load derivatives are again obtained from (15), yielding $\gamma'_1(0) = -4.986$ and $\gamma'_2(0) = -0.499$. For the heavy traffic regime, we use the fitted L^α-ellipsoid with parameters $a_1 = 4.522$, $a_2 = 0.535$ and $\alpha = 2.712$ shown in Figure 8 and get the asymptotic derivatives by (21): $\gamma'_1(1) = -4.689$, $\gamma'_2(1) = -0.738$. Finally, by fitting the parameters c and d in (22) for each class we obtain the approximate flow throughput curves shown in Figure 9, in good accordance with the results obtained by numerical evaluation of the recursion (8).

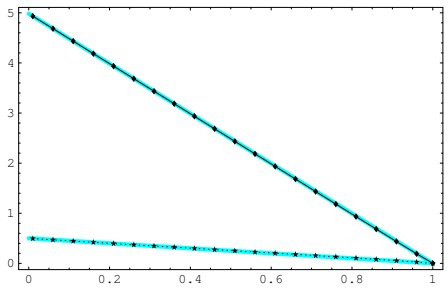


Figure 9: Approximated (lines) and numerically evaluated (dots) throughputs of classes 1 and 2 with load ratio $\rho_1/\rho_2 = 10$ as a function of the total load.

6. CONTINUOUS SETTING

In this section we formulate the concept of balanced fairness for networks with an infinite number of flow classes. We then give the low load expansion for the flow throughput, allowing us to calculate derivatives at zero load to any desired order. By using the fitting procedure of Section 5 we work out an example to get an approximation for the flow throughputs as a (continuous) function of the flow class.

6.1 General formulation

We consider here the balanced fairness concept in systems where the flow classes are indexed by a continuous variable $x \in \mathcal{A} \subseteq \mathbb{R}^d$. Flows are assumed to arrive according to a Poisson process with the arrival density $\lambda(x)$ and to have the mean size $1/\mu(x)$; the load density is denoted $\rho(x) = \lambda(x)/\mu(x)$. Whereas in the discrete case the state of the system is specified by a vector where an entry for each class gives the number of flows in that class, in the continuous index case we have to go to 'sparse matrix notation' and instead describe the locations of the flows present in the system. So the system state is specified by set of indices of the active flows $\xi = \{x_1, \ldots, x_n\}$, where n is any non-negative integer. The set of all possible states is denoted by Ξ. Note that ξ is an unordered set of indices, i.e. any permutation of the indices in the set ξ gives just another, equivalent label for the same state. We also assume that $\lambda(x)$ has no atoms, whence the probability that any two arriving flows have exactly the same index is zero, so we can restrict ourselves to states where no two indices are equal.

As in the discrete case, the balance requirement is satisfied if the service rates are of the form

$$\phi(x, \xi) = \frac{\Phi(\xi \backslash \{x\})}{\Phi(\xi)}, \quad \forall x \in \mathbb{R}^d, \ x \in \xi, \ \xi \in \Xi,$$

in which $\Phi(\cdot)$ is an arbitrary function. In this case the state probability density function reads

$$f(\xi) = \frac{1}{G} \Phi(\xi) \prod_{x_i \in \xi} \rho(x_i),$$

where G is the normalization constant.

Balanced fairness refers to the balanced allocation where the resource usage is as efficient as possible. Let $\mathcal{C}^\xi$ be the capacity set for the constellation ξ. Analogously with the discrete case, cf. (8), the balance function is defined recursively,

$$\Phi(\xi) = \min\{\alpha : \frac{(\Phi(\xi \setminus \{x_1\}), \ldots, \Phi(\xi \setminus \{x_n\}))}{\alpha} \in \mathcal{C}^\xi\},$$

that is, we remove the active flows in any order until we reach $\Phi(\emptyset)$, which can be fixed arbitrarily, e.g., $\Phi(\emptyset) = 1$. The system is stable if and only if $(\rho(x_i), \ldots, \rho(x_n)) \in \text{int}(\mathcal{C}^\xi)$ for all $\xi \in \Xi$.

If the capacity set is a polytope, cf. (1), for all ξ, we may write

$$\Phi(\xi) = \min_{i=1,\ldots,L^\xi} \frac{1}{C^\xi_l} \sum_{i=1}^{n} \Phi(\xi \setminus \{x_i\}) A^\xi_{il}, \tag{23}$$

where $C^\xi = (c_{x_1}, \ldots, c_{x_n})$ and A^ξ is an $n \times L^\xi$ matrix, both defined for the constellation ξ. Typically, determination of A^ξ for all ξ is cumbersome and in numerical evaluation of the recursion (23) one uses an alternative formulation, cf. Appendix B.

6.2 Low load expansion

The normalization constant can be written as a series in terms of multiple integrals with progressively larger number of active flows,

$$\begin{aligned} G = 1 + \int_{x \in \mathcal{A}} \Phi(x) \rho(x) dx + \\ + \frac{1}{2!} \int_{x \in \mathcal{A}} \int_{y \in \mathcal{A}} \Phi(x, y) \rho(x) \rho(y) dx dy + \ldots. \end{aligned} \tag{24}$$

Here and hereafter, we write for simplicity $\Phi(x_1, \ldots, x_n)$ instead of $\Phi(\{x_1, \ldots, x_n\})$. The factorial in (24) compensates for the fact that in an n-fold integral over the full range each state (unordered set of indices) appears $n!$ times. Hereafter, we suppress the integration limits with the understanding that all the integrals are over the full range.

The throughput $\gamma(x)$ is defined in the usual way,

$$\gamma(x) = \frac{\rho(x) dx}{\mathrm{E}\left[X(x, x + dx)\right]},$$

where $X(x, x + dx)$ is the number of flows in the interval $(x, x + dx)$. The expectation of this occupancy can be identified from the normalization constant by noting that the multiplier of dx in an n-fold integral gives the unnormalized probability density of having n flows in the system, one of them about point x, and that a given dx can be found at n places in an n-fold integral. Thus we get,

$$\gamma(x) = \frac{1 + \int \Phi(y) \rho(y) dy + \frac{1}{2!} \iint \Phi(y, z) \rho(y) \rho(z) dy dz + \ldots}{\Phi(x) + \int \Phi(x, y) \rho(y) dy + \frac{1}{2!} \iint \Phi(x, y, z) \rho(y) \rho(z) dy dz + \ldots}. \tag{25}$$

In general, calculating the throughput exactly from (25) is difficult. However, it provides a means to expand the throughput in terms of the load parameter. As in the discrete case (cf. §2.3) we define the scale-free balance function $\varphi(x_1, \ldots, x_n) = a(x_1) \cdots a(x_n) \Phi(x_1, \ldots, x_n)$, where $a(x)$ is the allocation for a sole flow with index x. Denoting $\varrho(x) = \rho(x)/a(x)$ we expand the expression to the second order, which is sufficient for, e.g., computation of the two first derivatives at $\rho(x) \to 0$,

$$\begin{aligned} \tfrac{\gamma(x)}{a(x)} = 1 - \textstyle\int (\varphi(x, y) - 1) \varrho(y) dy \\ - \tfrac{1}{2} \textstyle\iint (\varphi(x, y, z) - \varphi(y, z)) \varrho(y) \varrho(z) dy dz \\ + \textstyle\int \varphi(x, y) \varrho(y) dy \int (\varphi(x, y) - 1) \varrho(y) dy + \ldots. \end{aligned} \tag{26}$$

The throughput analysis of systems with continuous class index is generally more difficult than in the discrete class case. The light load behaviour is determined by the above equations and the throughput can be evaluated from (25) to any desired order by applying, e.g., the Monte Carlo method to the multiple integrals. This is itself an important achievement, since for instance a straight-forward process simulation (in time) of the system does not easily work for determining $\gamma(x)$, because in the simulation no samples are obtained with the flow index exactly equalling x. In the heavy load regime, however, less is known. We can determine the stability limit where the throughput goes to zero but the directional derivative of the throughput at this

limit is not known (this problem is left for future work). Nonetheless, even with this information we can approximate the throughput curve by an interpolating function, as will be demonstrated in the next section.

6.3 Example: a two-cell network

As an example we study a two-cell network in linear configuration illustrated in Figure 10. Mobiles are located on the segment between the two base stations. The total load ρ is assumed to be distributed uniformly along the unit distance between the base stations.

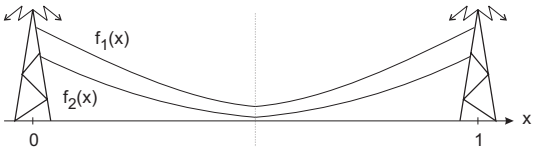


Figure 10: Linear two-cell system.

The system has three active transmission modes: either base station 1 is active, base station 2 is active, or both of them are active. Assuming a continuous link adaptation we have an infinity of flow classes indexed by the location $x \in (0,1)$. Each flow is served by the closest base station. Now $a(x)$, the maximum feasible rate of a mobile at point x, is achieved when only the closest base station is active. When both the base stations are active we denote the rate by $b(x)$ with $b(x) < a(x)$. For later use we give a special notation x^* for the point in $(0, 1/2)$ where $a(x^*) = 2\,b(x^*)$. If $a(1/2) < 2\,b(1/2)$, i.e. the equality is satisfied nowhere in the interval, we define $x^* = 1/2$.

We use the Shannon capacity formula with the standard signal attenuation behaviour

$$\begin{aligned} a(x) &= \log_2\Big(1 + \tfrac{x^{-\alpha}}{\nu}\Big), \\ b(x) &= \log_2\Big(1 + \tfrac{x^{-\alpha}}{\nu + (1-x)^{-\alpha}}\Big), \end{aligned}$$

where α is the attenuation exponent and ν is the normalized noise power (normalized by the signal power at distance 1). The above formulae hold for $x \in [0, 1/2]$. Values for $x \in [1/2, 1]$ are obtained by symmetry, $a(x) = a(1-x)$ and $b(x) = b(1-x)$.

By the results of [12] we know that the capacity of the system is:

$$C = \Big(\int_0^{x^*} \frac{1}{b(x)}\, dx + 2 \int_{x^*}^{1/2} \frac{1}{a(x)}\, dx \Big)^{-1},$$

in the sense that the network is stable if and only if $\rho < C$. To further analyze the system we consider the derivative at x when $\rho \to 0$. We denote it by $\gamma'(x)$, though the derivative is taken with respect to ρ (and not x). It follows from (26) that:

$$\gamma'(x) = -a(x) \int_0^1 \frac{\varphi(x,y) - 1}{a(y)}\, dy. \qquad (27)$$

To compute $\varphi(x,y)$ we need to construct the capacity set $\mathcal{C}^\xi$ separately for all flow pairs (x,y). Without loss of generality we may assume that $x \in (0, 1/2)$. Then the numerator of the integrand in (27) can be split into three different cases (cf. Figure 11):

$$\varphi(x,y) - 1 = \begin{cases} 1, & y \le 1/2, \\ \min\Big(1, \dfrac{a(x)}{b(x)}\,(1 - \dfrac{b(y)}{a(y)})\Big), & 1/2 < y \le 1 - x, \\ \min\Big(1, \dfrac{a(y)}{b(y)}\,(1 - \dfrac{b(x)}{a(x)})\Big), & 1 - x < y \le 1. \end{cases}$$

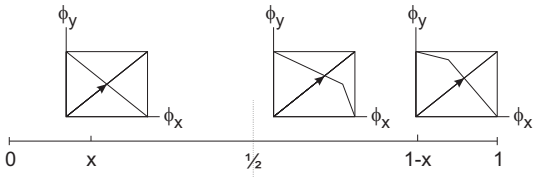


Figure 11: Balanced fair allocation for two flows at x (fixed) and y (three different regions).

As mentioned above, the derivative of the throughput at the capacity limit is a difficult problem, but we may still fix the two parameters of an interpolating function of the form (22) if we match also the second order derivative at load $\rho \to 0$. The second derivative can be found analogously to the first order case starting from (26). For brevity, the details are omitted. Figure 12 shows the resulting approximative throughput curves for the whole load range for flows located at points $x = 0.1,\ 0.3,\ 0.5$, with the assumed system parameters $\alpha = 3$ and $\nu = 1$. The figure presents also the numerical evaluation of (25) when the integrals both in the numerator and denominator were calculated up to 3rd, 5th and 8th order by Monte Carlo integration (numerical evaluation of the balance function is presented in Appendix B). This shows that a direct calculation of the throughput from the low load expansion is possible up to about one half of the maximum load, and it also suggests that the interpolation approximation is accurate.

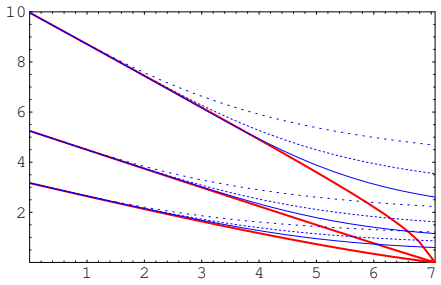


Figure 12: Flow throughput at points $x = 0.1,\ 0.3,\ 0.5$. Fitted rational functions are represented by thick lines. The results of numerical evaluation of (25) with integrals up to 3rd, 5th and 8th order are also shown.

7. CONCLUSIONS

Exact flow level performance analysis of networks for elastic data traffic is usually difficult. The main contribution of the

present paper is to demonstrate that under balanced fairness the performance of any system, measured in terms of flow throughput, can easily be evaluated in an approximate way.

Our starting point was to study the system under both light and heavy traffic conditions and, in particular, to derive simple expressions for the derivatives of the flow throughput with respect to the load. Using the information available for the extreme load conditions, we interpolate the throughput function on a given load line, representing a fixed traffic profile, for the whole range of load values. A rational expression, suggested by exactly solvable systems, is used for the interpolation. The approximation scheme is illustrated by a few examples. In the case of a star network for instance, the approximation is immediate and gives a perfect match with the numerical results.

For the derivative of the throughput at the heavy traffic limit we presented a conjecture based on an earlier store-and-forward bound [3] in the case of capacity set with linear constraints. When the capacity set has a smooth boundary, we approximate the derivative at the capacity limit by that of a so-called L^α-ellipsoid, for which we were able to find the derivative in closed form.

We emphasize that whereas the analysis of balanced fairness is limited to networks of low dimension or with a special structure, the proposed scheme is feasible for virtually any system for which the capacity set can be determined. It is worth noting, however, that the determination of the capacity set can occasionally be computationally difficult, for instance calling for finding the convex hull of a set vectors in a multi-dimensional space. This kind of situations arise, e.g. in multi-station or multi-hop wireless networks where the feasible combinations of link rates are determined by different transmission schedules [8].

A separate contribution of the paper is the introduction of balanced fairness for continuous class index systems. The basic formulation was given along with a low load expansion for the throughput. This, together with the capacity of the system, was again used to get an approximation for the throughput curve as a function of the load. An otherwise intractable system of two base stations with users anywhere between the base stations was worked out to demonstrate the usefulness of the method.

Further work is required for a formal proof of the conjecture on the asymptotic behaviour matching to the store-and-forward bound. This is closely related to the plausible conjecture on asymptotic equivalence of balanced fairness, proportional fairness and store-and-forward. In the continuous class index case the heavy load behaviour is also an open question.

8. REFERENCES

[1] G. Fayolle, A. de La Fortelle, J.M. Lasgouttes, L. Massoulié, J.W. Roberts, *Best-effort networks: modeling and performance analysis via large networks*, Proc. of IEEE Infocom, 2001.

[2] T. Bonald and A. Proutière, *Insensitive bandwidth sharing in data networks*, Queueing Systems, Vol. 44(1), 2003.

[3] T. Bonald, A. Proutière, *On performance bounds for balanced fairness*, Performance Evaluation, Vol. 55, 2004.

[4] T. Bonald, A. Proutière, J. Roberts, J. Virtamo, *Computational aspects of balanced fairness*, Proc. of the 18th International Teletraffic Congress (ITC-18), 2003.

[5] T. Bonald, J. Virtamo, *Calculating the flow level performance of balanced fairness in tree networks*, Performance Evaluation, Vol. 58, 2004.

[6] A. Penttinen, J. Virtamo and R. Jäntti, *Performance analysis in multi-hop radio networks with balanced fair resource sharing*, Telecommunications Systems, to appear, 2005.

[7] J. Leino and J. Virtamo, *Insensitive Traffic Splitting in Data Networks*, Proc. of 19th International Teletraffic Congress (ITC-19), 2005.

[8] T. Bonald, L. Massoulié, A. Proutière, J. Virtamo, *A Queueing Analysis of Max-Min Fairness, Proportional Fairness and Balanced Fairness*, submitted, 2005.

[9] D. Bertsekas and R. Gallager, *Data Networks*, Prentice Hall, 2:nd edition, 1992.

[10] F. Kelly, *Charging and rate control for elastic traffic*, European Transactions on Telecommunications, Vol. 8, 1997.

[11] T. Cover, J. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1991.

[12] T. Bonald, S. Borst, A. Proutière, *Inter-cell scheduling in wireless data networks*, Proc. of European Wireless, 2005.

[13] P.J. Schweitzer, *Approximate analysis of multiclass closed networks of queues*, Proc. International Conf. on Stochastic Control and Opt., Amsterdam, Netherlands, 1979.

[14] L. Massoulié, *Structural properties of proportional fairness: stability and insensitivity*, submitted to Annals of App. Probab., 2005.

APPENDIX

A. ASYMPTOTIC DERIVATIVE IN AN L^α-SPHERE

We wish to find the directional derivative of the flow throughput at the capacity limit when the capacity set is a N-dimensional L^α-sphere, defined as:

$$\mathcal{C} = \{\phi : \phi_1^\alpha + \cdots + \phi_N^\alpha \le 1\},$$

for some $\alpha \ge 1$. When $\alpha = 1$ the system is the ordinary multi-class processor-sharing queue, whose balance function is given by the multinomial coefficient. It is easy to see that for a general α the result is obtained by a simple power-raising transformation and the invariant measure is

$$\pi(x) = \left(\binom{|x|}{x_1,\ldots,x_N} \varrho_1^{x_1} \cdots \varrho_N^{x_N} \right)^{1/\alpha},$$

where $\varrho_i = \rho_i^\alpha$, $i = 1,\ldots,N$. We also use the notation $\varrho = \sum_i \varrho_i$ and $q_i = \varrho_i/\varrho$. The stability condition of the system is $\varrho < 1$.

The normalization constant is partitioned according to the total occupancy $|x|$,

$$G = \sum_x \pi(x) = \sum_{n=0}^{\infty} G_n,$$

where

$$G_n = \sum_{|x|=n} \pi(x) = \varrho^{n/\alpha} \sum_{|x|=n} \left(\binom{|x|}{x_1,\ldots,x_N} q_1^{x_1} \cdots q_N^{x_N} \right)^{1/\alpha}. \tag{28}$$

The expression in the parentheses is a multinomial distribution. It is well known that samples $x_1,\ldots,x_N$ from independent Poisson distributions with means $q_i a$, conditioned on the total number of samples $|x| = n$, obey the multinomial distribution with the parameters n and $q_1,\ldots,q_N$, irrespective of the value a. This follows from the identity:

$$\binom{|x|}{x_1,\ldots,x_N} q_1^{x_1} \cdots q_N^{x_N} = \frac{\frac{(q_1 a)^{x_1}}{x_1!}\, e^{-q_1 a} \cdots \frac{(q_N a)^{x_N}}{x_N!}\, e^{-q_N a}}{\frac{a^n}{n!}\, e^{-a}},$$

where the latter form admits the above interpretation as a conditional (Bayesian) probability. Now let $a = n$ whence the total sample size n represents a typical value. For large n, the Poisson distributions can be approximated by Gaussian distributions, so that

$$\frac{\frac{(q_1 n)^{x_1}}{x_1!}\, e^{-q_1 n} \cdots \frac{(q_N n)^{x_N}}{x_N!}\, e^{-q_N n}}{\frac{n^n}{n!}\, e^{-n}} \approx \frac{\frac{1}{\sqrt{2\pi q_1 n}}\, e^{-\frac{1}{2}\frac{(x_1 - q_1 n)^2}{q_1 n}} \cdots \frac{1}{\sqrt{2\pi q_N n}}\, e^{-\frac{1}{2}\frac{(x_N - q_N n)^2}{q_N n}}}{\frac{1}{\sqrt{2\pi n}}}.$$

Raising this to the power $1/\alpha$ yields

$$\left(\binom{|x|}{x_1,\ldots,x_N} q_1^{x_1} \cdots q_N^{x_N} \right)^{1/\alpha} \approx \theta\, n^{N/2-(N-1)/2\alpha}\, \frac{e^{-\frac{1}{2}\frac{(x_1 - q_1 n)^2}{\alpha q_1 n}}}{\sqrt{2\pi\alpha q_1 n}} \cdots \frac{e^{-\frac{1}{2}\frac{(x_N - q_N n)^2}{\alpha q_N n}}}{\sqrt{2\pi\alpha q_N n}}.$$

for some constant θ. The product of the fractions represents the probability distribution function of a multidimensional Gaussian random variable. The sum of the corresponding components is also Gaussian, with probability distribution function:

$$f(t) = \frac{e^{-\frac{1}{2}\frac{(t-n)^2}{\alpha n}}}{\sqrt{2\pi\alpha n}}.$$

On the other hand, this marginal probability distribution function is obtained by integrating the probability distribution function of the full multivariate distribution over the hyperplane $|x| = t$. In particular, the integral is $1/\sqrt{2\pi\alpha n}$ over the hyperplane $|x| = n$. But this integral is exactly the counterpart of the sum in (28) when the sum is approximated by an integral. Thus, we finally conclude that asymptotically

$$G \approx \theta \sum_{n=0}^{\infty} \varrho^{n/\alpha} n^{(N-1)(\alpha-1)/2\alpha}, \tag{29}$$

Similarly, we can write an expression for $\mathrm{E}[x_i]$. To this end, note that for a fixed, large n, the distribution of x_i is concentrated (in relative terms) around the value $q_i n$. Thus we have approximately

$$\mathrm{E}[x_i] \approx \frac{C}{G} \sum_{n=0}^{\infty} q_i n\, \varrho^{n/\alpha} n^{(N-1)(\alpha-1)/2\alpha}. \tag{30}$$

Approximating the sums in (29) and (30) by integrals and by making use of the result

$$\int_0^\infty a^x x^b dx = \frac{\Gamma(1+b)}{(\log\frac{1}{a})^{1+b}}, \qquad a < 1,\ b > -1,$$

we finally obtain for $\varrho \to 1$ the mean occupancy

$$\mathrm{E}[x_i] \approx \frac{(N+1)\alpha - (N-1)}{2}\, \frac{\varrho_i}{1-\varrho},$$

and the throughput

$$\gamma_i \approx \frac{2}{(N+1)\alpha - (N-1)}\, \frac{1 - (\rho_1^\alpha + \cdots + \rho_N^\alpha)}{\rho_i^{\alpha-1}}. \tag{31}$$

Consider now the behaviour of γ_i along a line when the loads ρ_i are changed in given proportions p_i, with $|p| = 1$, as a function of a load parameter r,

$$\rho_i = \frac{p_i}{(p_1^\alpha + \cdots + p_N^\alpha)^{1/\alpha}}\, r = \left(\frac{p_i^\alpha}{p_1^\alpha + \cdots + p_N^\alpha} \right)^{1/\alpha} r.$$

Parameter r varies from 0 to 1 with the value $r = 1$ corresponding to the maximum load with $\rho_1^\alpha + \cdots + \rho_N^\alpha = 1$. Inserting this into (31) yields

$$\gamma_i(r) \approx \frac{2}{(N+1)\alpha - (N-1)} \left(\frac{p_1^\alpha + \cdots + p_N^\alpha}{p_i^\alpha} \right)^{\frac{\alpha-1}{\alpha}} \frac{1 - r^\alpha}{r^{\alpha-1}}.$$

From this we can calculate the derivative of the throughput with respect to the load parameter r at the capacity limit $r = 1$. Since the above expressions are asymptotically correct, we get an exact result,

$$\gamma_i'(1) = -\frac{2\alpha}{(N+1)\alpha - (N-1)} \left(\frac{p_1^\alpha + \cdots + p_N^\alpha}{p_i^\alpha} \right)^{\frac{\alpha-1}{\alpha}}.$$

B. BALANCE FUNCTION COMPUTATION IN A CONTINUOUS SETTING

If the capacity set $\mathcal{C}^\xi$ is a polytope for all $\xi \in \Xi$, we may alternatively describe it as the convex hull of its corner points. These corner points, referred to as the spanning rate vectors, are usually much easier to derive from the system than the A and C required for the polytope form (1). Applying this alternative representation, the balance function $\Phi(\xi)$ is evaluated recursively as follows (cf. the discrete class case [6]):

1. Enumerate all m possible rate vectors for the constellation ξ. These vectors form the rows of the $m \times n$ rate matrix R^ξ. Only the spanning rate vectors need to be included, but it is not necessary to eliminate any non-spanning vectors.
2. Denote $\Theta = (\Phi(\xi \backslash \{x_1\}), \ldots, \Phi(\xi \backslash \{x_n\}))$, $e = (1, \ldots, 1)$ and $y = (y_1, \ldots, y_m)$. Now the value of the balance function is obtained from the LP-problem

$$\Phi(\xi) = \min_y \; e \cdot y^{\mathrm{T}},$$
$$yR^\xi \geq \Theta,$$
$$y \geq 0.$$

In the example §6.3 the rate matrix R^ξ contains the rate vectors in which each flow $x \in \xi$ is served alone. In addition, one includes rate vectors for all pairs of flows $\{x, y\}$, $x, y \in \xi$, which can be served simultaneously, i.e. for which $\min(x, y) \leq 0.5$ and $\max(x, y) > 0.5$. For example, let $\xi = (0.2, 0.3, 0.7)$, then

$$R^\xi = \begin{pmatrix} a(0.2) & 0 & 0 \\ 0 & a(0.3) & 0 \\ 0 & 0 & a(0.7) \\ b(0.2) & 0 & b(0.7) \\ 0 & b(0.3) & b(0.7) \end{pmatrix}.$$

For a small number of flows it is generally not efficient to separately remove non-spanning rate vectors from R^ξ.