Distributed Network Resource Allocation for Multi-Tiered Multimedia Applications

Georgios Tychogiorgos, Athanasios Gkelias and Kin K. Leung
Electrical and Electronic Engineering
Imperial College
London SW7 2AZ, UK
{*g.tychogiorgos, a.gkelias, kin.leung*}*@imperial.ac.uk*

***Abstract*—The continuously growing number of multimedia applications in current communication networks highlights the necessity for an efficient resource allocation mechanism to capture the unique characteristics of multi-tiered multimedia applications and allocate network capacity in an efficient way. This paper examines the problem of sharing the network throughput under the existence of inelastic traffic flows that follow a multi-tiered utility function. First, the concept of multi-sigmoidal utilities is introduced in order to describe user satisfaction, then, the implications of the use of such utilities are discussed for two different allocation policies; the *bandwidth-proportional* and the *utility-proportional* fairness allocation policies. In the former case, the intrinsic reasons of possible network oscillations are analyzed in detail and a heuristic to overcome such situations is proposed. In the latter one, where such oscillations are not possible, efficient ways to calculate a closed form solution for the optimal rate allocation are described. Moreover, a novel mathematical representation of such a multi-sigmoidal utility is presented and closed form solutions for a number of application types are calculated. Finally, the efficiency and robustness of the proposed algorithms is evaluated by simulations for different network topologies and compared against other work in literature.**

I. Introduction

The end-to-end communication and resource allocation services in current communication networks are provided by Transport layer protocols such as TCP, whose various extensions have been shown to implicitly solve a resource allocation optimization problem [1] where all applications have been modelled using concave utility functions. Although this was a valid assumption in the past, the traffic generated by current applications has such *Quality of Service (QoS)* requirements that must be modelled by non-concave functions. Therefore, existing resource allocation schemes provide suboptimal solutions that may significantly affect both network performance and user experience.

More specifically, network traffic can be classified into two categories: *elastic* and *inelastic* [2]. Elastic applications include file transfer (FTP), email, network management (SNMP) and Web access (HTTP), where user satisfaction is modelled using logarithmic and other concave utility functions [1] (e.g. $U(x) = \log x$). Inelasticity usually characterizes *real-time* applications such as Video Streaming, Teleconferencing, Voice over IP (VoIP), Stock Trading etc. where non-concave utilities of sigmoidal shape are typically used [3][4][5]. The sole use of concave functions had little effect in the resource allocation in previous decades, since elastic applications were responsible for almost all the traffic. In current networks though, where the majority of the traffic is generated by inelastic real-time applications [6], such an assumption may lead to significant misuse of resources which can prove the use of TCP impractical [7].

Network Utility Maximization (NUM) [8], contrary to the resource allocation algorithm in TCP, can distinguish between elastic and inelastic applications by choosing different utility functions for each one. Since the seminal work of Kelly et al. [8], there have been several pieces of work that cultivated a deep understanding in the ways that optimization theory can be utilized in solving various convex resource allocation formulations in a distributed way. Interested readers are referred to [9] and the references therein for a complete overview of convex network resource allocation methods and references [7][10][11] and [12] that introduce the use of single sigmoidal utilities and examine the effects of non-convexity of the resulting optimization problem on the development of a distributed algorithm to solve it.

The single sigmoidal utility functions were introduced to model multimedia applications but as technology advances they may not be suitable to model many state of the art multimedia applications. Several video streaming applications used nowadays offer services at different quality levels with each level having different bit-rate requirements and offering different *Quality of Experience (QoE)* for the user. For example, an online video content provider offers four distinct levels of video quality (e.g. low, medium, high, ultra high) based on the video resolution and bit rate. Each quality option represents a different level of user satisfaction. Moreover, for a specific video resolution the allocated bit rate affects user satisfaction. For example, if low resolution is chosen, the increase of bit rate above a certain level will not result in significantly better visual results since the resolution is too low for a visible improvement. Therefore, user satisfaction at this quality level is saturated and further increase can only be a result of the transition to a higher resolution profile. Such multi-tiered multimedia applications can not be modelled satisfactorily by single sigmoidal utilities.

Prior research has shown [8] that the resulting bandwidth allocations for traditional NUM approaches follow the so-called *bandwidth-proportional* fairness (BPF). While this type of fairness seems to perform well when all users follow the same utility, this approach is responsible for some contradictory behavior in cases that users have different QoS needs, i.e. when they follow different utilities. More specifically, a bandwidth-

proportional fair optimization algorithm favors users with low demand, i.e. those with rapidly increasing utility function since this leads to a larger increase in the aggregate utility than when allocating to users with high demand, i.e. with small value of utility derivative. To resolve this contradictory behavior, authors in [13] define a new type of fairness, called *utility-proportional* fairness (UPF). According to that, a feasible bandwidth allocation vector $\boldsymbol{x^*} = [x_1^*, x_2^*, \ldots, x_R^*]^T$ is utility proportional fair, if for any other feasible vector $\boldsymbol{x}$ the following condition holds:

$$\sum_{r \in R} \frac{x_r - x_r^*}{U_r\left(x_r^*\right)} \leq 0. \tag{1}$$

The *utility-proportional* fairness can be achieved if the utility function of each user is transformed according to:

$$\mathcal{U}_r\left(x_r\right) = \int_{m_r}^{x_r} \frac{1}{U_r\left(y\right)} dy, \quad m_r \leq x_r \leq M_r, \tag{2}$$

where m_r and M_r are the minimum and maximum transmission rates for user r respectively, and the objective function of the NUM problem is changed to $\sum_{r \in R} \mathcal{U}_r\left(x_r\right)$.

The most intuitive, yet very challenging, solution to resolve the inefficiency of single-sigmoidal utilities is the use of multi-sigmoidal functions. Multi-sigmoidal utilities, such as the one shown in black at the top plot in Figure 1, are capable of capturing the step-like behavior of user satisfaction with respect to the various quality levels of modern video applications. **The development of appropriate multi-sigmoidal utilities that can capture the QoS/QoE characteristics of the underlying applications and the extension of NUM to incorporate such utilities are the main motivation behind our work**.

The organization of the paper in terms of the contribution of each section is as follows. Section II introduces the multi-sigmoidal utility to express user satisfaction in multi-tiered multimedia applications and describes how this affects the traditional NUM framework. Section III examines the incorporation of such function while trying to allocate resources according to BPF. This includes the impact of such a choice on the continuity properties of the optimal rate allocation function, describes a detailed procedure to determine all these discontinuity points and proposes an efficient heuristic algorithm in order to resolve network oscillations, when they occur. Section IV discusses the changes necessary to the traditional NUM framework in order to be able to allocate resources according to the UPF policy. Section V proposes a novel tangent-based mathematical representation of a multi-sigmoidal utility function and analyzes how this family of utility functions can lead to the approximation and calculation of closed form solutions in BFP and UPF, respectively. Section VI presents extended simulation results of the proposed algorithms in various topologies and Section VII concludes our work.

II. NUM with Multi-sigmoidal Utilities

A. Properties of a Multi-sigmoidal Utility Function

Before introducing multi-sigmoidal utilities, it is necessary to define a set of properties that a function should possess in order to be able to model user satisfaction. Such functions should:

P1) take positive values in the range $[0, 1]$;
P2) be increasing functions of the transmission rate,
P3) be zero when no rate is allocated to a particular user;
P4) have value 1 for rates above the maximum rate, r_{max};
P5) be continuous in the range $(0, r_{max})$.

One could argue that a potentially sixth property could be added as well. This describes the need that all quality levels, i.e. all concave parts, of the utility to be reachable by a NUM algorithm. In other words, a multi-sigmoidal utility can indeed model multi-tiered applications only if all distinct utility levels can be optimal selections under some conditions. While this will be explained in more detail later, in Section III, it is not considered a requirement for a multi-sigmoidal utility since the exact shape of a utility function is determined depending on each user's appreciation of the allocated bitrate not affected by the operational characteristics of NUM.

B. Network Resource Allocation with Multi-sigmoidal Utilities

This subsection extends the initial NUM framework [8] by allowing the utilities to be multi-sigmoidal and discusses the research challenges that this imposes to the distributed algorithm, which will be answered later in this paper.

Consider a multi-hop network where M nodes act as sources sending streams of traffic to a set of destination nodes using a set of J links. A single node can operate as source, destination or even as relay node that just forwards traffic to its neighbours. When a source node i sends traffic at a rate r_i, it enjoys utility $U_i(r_i)$. It is assumed that all links in the network are wired, vector $\boldsymbol{C} = [C_1, C_2, \cdots, C_J]^T$ contains the capacity of each link and $\boldsymbol{r} = [r_1, r_2, \cdots, r_M]^T$ includes the transmission rates of all sources. The optimization problem describing the *Network Resource Allocation (NRA)* problem is:

Problem $\mathbf{\Pi^P_{NRA}}$: Find the optimal rate vector $\boldsymbol{r}$

$$\max_{\boldsymbol{r}} \quad \sum_{i=1}^{M} \mathcal{U}_i(r_i) \qquad \text{s. t.} \quad \sum_{i=1}^{M} \alpha_{i,j} r_i \leq C_j, \quad \forall \text{ links } j$$

where routing coefficient $\alpha_{i,j}$ is 1 if user i sends traffic through link j and 0 otherwise. We assume that the routing matrix $\boldsymbol{A}$, containing all routing coefficients $\alpha_{i,j}$, is known a priori and considered fixed throughout the optimization process. The rates $r_i \geq 0$, $i \in [1, M]$, represent the transmission rates of the respective source nodes. Note that $\mathcal{U}_i(r_i)$ represents a transformation of U_i. As mentioned later, $\mathcal{U}_i(r_i) = U_i(r_i)$ for *bandwidth-proportional* fairness but not for *utility-proportional* one.

Problem $\mathbf{\Pi^P_{NRA}}$ can be solved using Duality Theory by constructing its dual problem and trying to solve the primal-dual pair of problems in a distributed way [14]. Dual problem variables are the so-called "*lagrange* multipliers" and represent the "price" that user i has to pay to send each of the r_i units of traffic through link j. Following the analysis in [14][15], it is evident that each user is trying to maximize their *Net Utility* and thus the optimal resource allocation for user i is:

$$r_i^*\left(\boldsymbol{\lambda}\right) = \text{argmax } \left\{NU_i\left(r_i\right) = \mathcal{U}_i\left(r_i\right) - r_i \cdot \lambda^i\right\}, \tag{3}$$

where $\lambda^i = \sum_{j=1}^{J} \alpha_{i,j}\lambda_j$. Equation (3) can be used to calculate the optimal rate of user i for a given price vector $\boldsymbol{\lambda}$. The optimal value of the dual variables λ_j, $j \in [1, J]$, can be calculated iteratively using a gradient method, such as the *Gradient Descent* [14],

$$\lambda_j(t+1) = \lambda_j(t) - s_\lambda(t)\left(C_j - \sum_{i=1}^{M} \alpha_{i,j} r_i\right). \quad (4)$$

$s_\lambda(t)$ is the step size of the method at time t and affects the convergence speed and distance from the true optimum [14].

Equations (3) and (4) constitute a joint primal-dual distributed algorithm of NUM, which can converge to an optimal solution, even in the case of non-concave utilities (such as single-sigmoidal), as long as (3) is continuous around the optimal price vector $\boldsymbol{\lambda}^*$ [7]-[10]. The properties of equations (3) and (4) in the case of multi-sigmoidal utilities will be discussed in the remainder of this paper.

The optimal solution of (3) for a particular user i is also the optimal rate for this user for Problem $\mathbf{\Pi}^{\mathbf{p}}_{\mathbf{NRA}}$. This rate is at a point where the derivative of the objective function diminishes [14], which leads to:

$$r_i^*(\boldsymbol{\lambda}) = \mathcal{U}_i'\left(\lambda^i\right)^{-1}, \quad (5)$$

where $\mathcal{U}_i'(\cdot)^{-1}$ is the inverse first derivative function.

The two resource allocation policies examined in this paper differ in the calculation of the inverse of the first derivative. In BPF $\mathcal{U}_i(\cdot) = U_i(r_i)$ and therefore the calculation of the inverse of the first derivative is possible only for utilities whose derivative is a $1-1$ function. In any other case, there might be multiple optimal rates for a single aggregate price λ^i. This is the inherent reason for the existence of oscillations in the rate allocation process, as discussed later in this paper. On the other hand, the utility function (2) in UPF can be inverted allowing the calculation of closed form solutions for (5).

III. Multi-sigmoidal Utilities in BPF

This section provides a detailed analysis of the effect of the use of multi-sigmoidal utilities in the resource allocation process while preserving *bandwidth-proportional* fairness (BPF). For simplicity, we will use notation $U_i(r_i)$ instead of $\mathcal{U}_i(r_i)$ in the remainder of this section since $\mathcal{U}_i(r_i) = U_i(r_i)$.

A. Discontinuity

In the case of *bandwidth-proportional* fairness, $r_i^*(\boldsymbol{\lambda})$ is continuous for all price vectors if the utility is either concave or convex function of rates while it is discontinuous at only one point for single-sigmoidal utilities [4][7]. Equation (5) shows that $r_i^*(\boldsymbol{\lambda})$ is in essence a function of the aggregate price per unit of traffic and does not depend on the individual values of λ_j, $j \in [1, J]$. Therefore, we will also refer to $r_i^*(\boldsymbol{\lambda})$ as $r_i^*\left(\lambda^i\right)$, where λ^i is the aggregate price for user i. In addition, it turns out that the shape of a utility function determines the discontinuity points of the rate allocation function and that the discontinuity points correspond to jumps from one concave region to another or from one concave region to zero. Moreover, there is a number of candidate discontinuity points that may or may not appear as discontinuities of $r_i^*\left(\lambda^i\right)$.

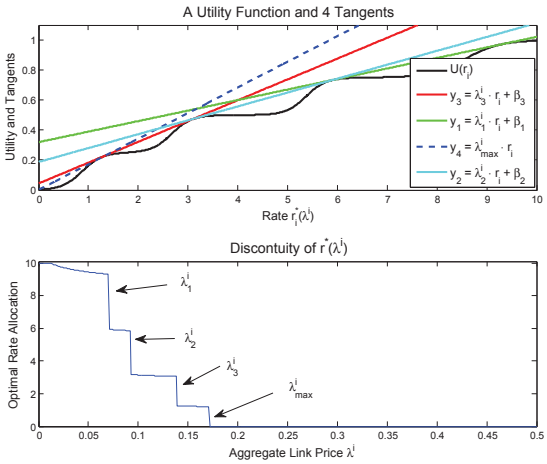

Fig. 1: A multi-sigmoidal utility with 4 discontinuity points

The methodology to identify these points involves the use of lines that are tangent to the utility function $U_i(r_i)$. Initially, we draw a tangent line $y = \alpha r_i + \beta$ that osculates the utility function at two or more points. Let r_i^n, with $n = 1, 2, ..., N$ and $r_i^1 < r_i^2 < \cdots < r_i^N$, be the rates at which the tangent line y osculates the utility function, such that the tangent line is graphically always above the utility function. In multi-sigmoidal utilities, a tangent line such as y can osculate the utility function at most at $N = K$ points, where K is the number of inflection points in the utility shape, and there can be at most $\frac{K(K-1)}{2}$ distinct tangents, in the case where each one of them osculates the utility function at exactly two points. Using the example of tangent y we can prove that the candidate discontinuity points are aggregate prices equal to the slopes of these tangent lines.

Theorem III.1. *If $\lambda^i = \alpha$, the rates r_i^n, $n = 1, 2, \ldots, N$ are all globally optimal rates for user i and aggregate price λ^i.*

Theorem III.1 shows that function $r_i^*\left(\lambda^i\right)$ has multiple values for aggregate price λ^i equal to the slope of the tangent y and the multiplicity of the function at that point is equal to the number of points N that the slope osculates the utility function. Regarding the discontinuity and monotonicity properties of $r_i^*\left(\lambda^i\right)$ it is possible to prove the following theorems. The proofs of Theorems III.1-III.5 are omitted for brevity but can be found in [16].

Theorem III.2. *If $\lambda^i = \alpha + \delta$, where δ is a very small positive constant, then the globally optimal rate $r_i^*\left(\lambda^i\right)$ is smaller than the smallest optimal rate for $\lambda^i = \alpha$, i.e $r_i^*\left(\lambda^i\right) < r_i^1$.*

Theorem III.3. *If $\lambda_i = \alpha - \delta$, where δ is a very small positive constant, then the globally optimal rate r_i^* is larger than the largest optimal rate for $\lambda_i = \alpha$, i.e $r_i^* > r_i^N$.*

Theorem III.4. *The optimal rate function of user i, $r_i^*\left(\lambda^i\right)$, is a decreasing function of λ^i.*

Apart from the discontinuity of $r_i^*\left(\lambda^i\right)$ around the points determined by the tangents at the utility function, these theorems imply that rates in the range $\left(r_i^1, r_i^N\right)$ can never be globally optimal rates and therefore $r_i^*\left(\lambda^i\right)$ will "jump" from

r_i^N to r_i^1. Moreover, there will be a maximum value for λ^i, let λ_{max}^i, above which the optimal rate will be zero. In other words, $r_i^*\left(\lambda^i\right)$ has a positive value for $0 \leq \lambda^i \leq \lambda_{max}^i$ and is zero for aggregate prices $\lambda^i \geq \lambda_{max}^i$. This maximum non-zero aggregate price λ_{max}^i is called *maximum willingness to pay* for user i and is a discontinuity point of $r_i^*\left(\lambda^i\right)$ for all sigmoidal utilities. To calculate λ_{max}^i we can use the same procedure as for single-sigmoidal utilities [10].

It is evident from the above that every tangent at two or more points of the utility function represents a candidate discontinuity point of function $r_i\left(\boldsymbol{\lambda}\right)$. Each one of these points represents a "jump" from one hyperbolic tangent component to another, while the discontinuity point around λ_{max}^i represents a "jump" from a hyperbolic tangent component to zero rate. The latter point will always appear in the rate function but the rest depend on their relative value compared to λ_{max}^i. For example, if λ_{max}^i is smaller than all the other candidate discontinuity aggregate prices, then none of them will appear and there will be only one discontinuity point, λ_{max}^i. The maximum number of discontinuity points are K, as many as the inflection points of the utility. This can happen if there are $K-1$ distinct tangent lines, each one touching the utility at two points that belong to two consecutive hyperbolic tangent components, with the K^{th}, corresponding to $\lambda^i = \lambda_{max}^i$, being graphically represented by a tangent line that passes from point $(0,0)$ and osculates the utility function at its first hyperbolic tangent component.

To provide an example of the above, the top sub-figure of Figure 1 shows a utility function with four discontinuity points, along with the four tangent lines responsible for these discontinuity points, while the bottom one shows the optimal rate $r_i^*\left(\boldsymbol{\lambda}\right)$ with the discontinuity points clearly shown. This figure illustrates the connection between the shape of the utility function and the discontinuity points of the price-based rate function $r_i^*\left(\boldsymbol{\lambda}\right)$. Moreover, it illustrates that $r_i^*\left(\boldsymbol{\lambda}\right)$ consists of decreasing continuous parts and decreasing jump discontinuity points. Commenting on the feasibility of all K sigmoidal components to be selected as optimal choices, it is evident that this is possible only under the existence of K distinct discontinuity points, i.e. the utility function is fully reachable. In any other case, there will be at least one sigmoidal component that is unreachable during NUM. Based on this observation, we can prove the following:

Theorem III.5. *A multi-sigmoidal utility will have all levels reachable, and hence will have the maximum number of discontinuity points, iff the following conditions hold:*

(1) $\lambda_{k,k-1}^i < \lambda_{k,j}^i, \quad \forall j \in [1,\ldots,k-2], \quad k \in [3,\ldots,K]$
(2) $\lambda_{k,k-1}^i < \lambda_{max}^i, \quad \forall k \in [2,\ldots,K],$

where $\lambda_{k,l}^i$ is the slope of the tangent that osculates the utility of user i at the k^{th} and l^{th} sigmoidal component.

For an arbitrary multi-sigmoidal utility function we can calculate all the aggregate prices for which $r_i^*\left(\boldsymbol{\lambda}\right)$ is discontinuous. To this purpose, we assume the existence of a tangent that osculates the utility function at exactly two points, let p_1 and p_2. After calculating all such possible tangents and their touching points, a candidate discontinuity point is the slope of this line:

$$\lambda_c^i = \frac{U_i\left(p_1\right) - U_i\left(p_2\right)}{p_1 - p_2}. \tag{6}$$

Using these candidate points, we create the symmetric matrix $\boldsymbol{S}^i$ of size $K \times K$, where $S^i\left(s_1, s_2\right)$ represents the slope of the tangent that osculates the ${s_1}^{th}$ and ${s_2}^{th}$ concave region of the utility. By convention, we assume that the elements of the main diagonal of matrix $\boldsymbol{S}^i$ contain some very large positive value. Algorithm 1 can be used to determine which of these candidate discontinuity points will actually appear in $r_i\left(\boldsymbol{\lambda}\right)$. Note that $S^i\left(ctr_1, 1:ctr_1\right)$ denotes the first ctr_1 elements of the ${ctr_1}^{th}$ row of matrix $\boldsymbol{S}^i$. The resulting vector $\boldsymbol{disc}$ contains the discontinuity points of $r_i^*\left(\boldsymbol{\lambda}\right)$. Algorithm 1 is an iterative algorithm that can be run independently by each user in order to determine the discontinuity points of its rate allocation function. Note that in case that one of the tangents osculates the utility function at more than two points, then two or more elements of matrix $\boldsymbol{S}^i$ will be equal.

Algorithm 1 – Calculation of discontinuity points of $r_i^*\left(\boldsymbol{\lambda}\right)$

```
ctr_1 = K; ctr_2 = 1; Calculate λ^i_max, S^i;
while true do
    index = argmin{S^i(ctr_1, 1:ctr_1)};
    λ^i_tmp = min{S^i(ctr_1, 1:ctr_1)};
    if λ^i_max < λ^i_tmp then
        break;
    else
        disc(ctr_2) = λ^i_tmp; ctr_1 = index;
    end if
    ctr_2 = ctr_2 + 1;
end while
disc(ctr_2) = λ^i_max;
```

B. Oscillations

The discontinuity points calculated by Algorithm 1 play an important role in the convergence of the algorithm comprised of eq. (3) and (4) in the case of multi-sigmoidal utilities.

When the condition in [12] is not satisfied, there can be oscillations in the network. The phenomenon of oscillation occurs when a user transmits at an excessive data rate (compared to the available capacity) in one iteration, and then, in the next iteration, the user transmits at an exceedingly low rate. An oscillation is formed as the repetition of these two events continues indefinitely, prevents the user from converging to the optimal transmission rate and leads to a wider network oscillation. The oscillation rates of user i are in fact very close to the optimal rates for λ^i. Based on this observation, we propose the *Oscillation Resolving Heuristic (ORH)* to assure the convergence of the gradient based distributed algorithm.

The *Oscillation Resolving Heuristic (ORH)* allocates a constant non-zero rate to oscillating users and removes them from the rest of the optimization process, which continues for the remaining users in the network. The allocated rate r_i^{osc} to oscillating users i is equal to the smallest touching point of the tangent with slope equal to the aggregate price λ^i for which the oscillation happens. This approach assures that no

users are restricted from accessing network resources, contrary to other approaches in literature. This allocated rate satisfies the necessary conditions for optimality and by selecting the smallest of all the optimal rates for price λ^i we assure that there will be more resources for the rest of the users in the network, thus leading to higher network utility.

The implementation of ORH is very simple, requires a simple oscillation detection mechanism with no need for any centralized coordination. Note also that, the ORH does not represent a complete solution for solving Problem $\mathbf{\Pi^{P}_{NRA}}$ and does not affect the convergence properties of the algorithm. In fact, (3) and (4) are responsible for solving Problem $\mathbf{\Pi^{P}_{NRA}}$ iteratively, while the ORH is merely part of the process for resolving an oscillation that might occur during this process. The ORH leads towards more fair resource allocations compared to mechanisms such as the heuristic proposed in [4].

IV. Multi-sigmoidal Utilities in UPF

By considering the *utility proportional fairness* transformation of (2), the problem becomes convex and (3) always satisfies the condition in [12]. More importantly this allows us to calculate a closed form solution for (5). This stems from the fact that the first derivative of the utility function can be easily calculated as $\mathcal{U}_i^{'}(r_i) = \frac{1}{U_i(r_i)}$, which is invertible as long as the utility is continuous and monotonic, which are both true for any concave utility and the multi-sigmoidal utility considered in this paper. In this case, the optimal rate is given by:

$$r_i^*\left(\lambda^i\right) = U_i^{-1}\left(\frac{1}{\lambda^i}\right). \tag{7}$$

As explained in the next section, based on (7), we can calculate a closed form solution for utilities that satisfy these two properties. This is a significant advantage of the *utility proportional fairness approach* which leads to the development of algorithms that calculate the optimal solution even for non-concave utilities and converge significantly faster than the traditional iterative approach.

V. A Novel Multi-sigmoidal Function

A. A Hyperbolic Tangent Based Utility Function

Based on the desired properties of a multi-sigmoidal utility, presented in Section II, we propose the use of the following family of multi-sigmoidal functions:

$$U(r) = \frac{1}{2K}\left\{\sum_{k=1}^{K}\tanh\left(\frac{r-c_k}{b_k}\right) + K\right\}, \tag{8}$$

where r is the transmission rate, c_k is the k^{th} inflection point, with $c_1 > c_2 > \cdots > c_K$, and b_k is a positive design parameter that determines the steepness of the k^{th} component of the multi-sigmoidal function. K is the number of single sigmoidal components comprising the multi-sigmoidal function, each one of them having a single inflection point. For example, the multi-sigmoidal function in black in the top plot of Figure 1 consists of four hyperbolic tangent components.

Hyperbolic tangent functions have been extensively used in neural networks research area [17] but their convenient properties make them also applicable within the context of multi-tiered multimedia applications for the following reasons:

- They possess the five properties described in Section II.
- They can be combined to create multi-sigmoidal shapes of arbitrary number of rate levels.
- They can be calibrated using the inflection vector $\boldsymbol{c}$ and the steepness vector $\boldsymbol{b}$ to achieve the desired shape.
- Their first derivative can be easily inverted to calculate the optimal rate allocation for a specific price vector.

The hyperbolic tangent function, $tanh\,(x)$, is a symmetric, continuous (property P5), differentiable and increasing (property P2) function, which is centered around its inflection point at $r = 0$ and has two horizontal asymptotes, the lines $y = -1$ and $y = 1$. Each tangent component can be scaled and shifted appropriately so that the resulting utility has values within the range $[0, 1]$. The resulting multi-sigmoidal function has horizontal asymptotes the lines $y = 0$ and $y = 1$ (property P1). Note that inflection points c_k can be used as design parameters to create the step-like behaviour of the utility around the rate values of each application quality level.

Parameters b_k, $k = 1, \ldots, K$, can be used to calibrate the steepness of the respective tangent components. In general, larger values for b_k lead to smoother shapes. In particular, they can be used to bring $U(0)$ and $U(r_{max})$ as close to the bounds (0 and 1 respectively) as necessary, where r_{max} is the maximum rate above which the utility is equal to 1. Specifically, for $r_i = 0$ equation (8) becomes $\tanh\left(-\frac{c_k}{b_k}\right) \approx -1$, for $k = 1, 2, \ldots, K$. Given that $y = -1$ is an asymptote, the equation will never be satisfied in the equality but we can select variables b_k, $k \in \{1, 2, \ldots, K\}$, so that the maximum error ϵ_k of the k^{th} tangent component is bounded. More specifically, it is possible to calculate an upper bound for each b_k in order to meet property P3 according to

$$\tanh\left(-\frac{c_k}{b_k}\right) \leq -1 + \epsilon_k \Rightarrow b_k \leq -\frac{c_k}{\tanh^{-1}(\epsilon_k - 1)} \tag{9}$$

and since $\tanh^{-1}(\cdot)$ is negative around $r = -1$,

$$b_k \leq \frac{c_k}{|\tanh^{-1}(\epsilon_k - 1)|}. \tag{10}$$

By selecting the component bounds appropriately, it is possible to bound the total error $\epsilon = \sum_{k=1}^{K}\epsilon_k$ below a maximum threshold. In addition, it can be shown that the effect of parameter b_1, i.e. the sigmoidal component that is closer to the point $r = 0$, is dominant over the rest and therefore the calculated bound for b_1 is expected to be much tighter for the same error. Working in the same way, it is possible to calculate the upper bounds for parameters b_k to assure that property P4 is also satisfied and, as seen later, to minimize the approximation error of the optimal rate.

B. Approximating the Optimal Rate in BPF

The family of multi-sigmoidal utilities described in (8) is a non-concave function with multiple concave and convex regions. Its first derivative is given by

$$V(r) = \frac{1}{2K}\left\{\sum_{k=1}^{K}\frac{1}{b_k}sech^2\left(\frac{r-c_k}{b_k}\right)\right\}, \tag{11}$$

which is not a 1-1 function since the same value $V(\cdot)$ corresponds to more than one rates and therefore is not invertible. Figure 2 shows the utility derivative for the multi-sigmoidal example in black of Figure 1, which illustrates that a single value of utility derivative corresponds to at most $2 \times K$ distinct rates.

It is possible however to approximate these rates efficiently. The approximation methodology relies on the fact that $V(\cdot)$ in (11) is a summation of a number of independent hyperbolic secant components. Moreover, they are symmetric, they can be inversed separately, and by taking into account that the rate that maximizes Problem $\mathbf{\Pi^i_{NU}}(\boldsymbol{\lambda})$ can only be in a concave region or at zero rate, it is possible to calculate a single rate for each component by

$$r_i^c(\boldsymbol{\lambda}, k) = b_k^i \cdot sech^{-1}\left(\sqrt{2 \cdot K \cdot b_k^i \cdot \lambda^i}\right) + c_k^i, \qquad (12)$$

where $sech^{-1}(\cdot)$ is the inverse hyperbolic secant, b_k^i, $k = 1, 2, \ldots, K$, form steepness vector $\boldsymbol{b^i}$ and inflection points c_k^i, $k = 1, 2, \ldots, K$, form inflection vector $\boldsymbol{c^i}$ of user i[1]. An additional candidate solution is at $r_i^c(\boldsymbol{\lambda}, K+1) = 0$, which must be also taken into account. Consequently, the optimal rate of user i for vector $\boldsymbol{\lambda}$ will be the one that yields the maximum *net utility*, i.e.

$$r_i^*(\boldsymbol{\lambda}) = \text{argmax}\left\{NU_i\left(r_i^c(\boldsymbol{\lambda}, k)\right) | k = 1, 2, \ldots, K+1\right\}. \qquad (13)$$

The use of equation (13) to approximate the optimal rate for any price vector $\boldsymbol{\lambda}$ transforms the distributed algorithm described in Section II to use (13) instead of (3). The resulting algorithm comprised of (13) and (4) is an extension of the standard gradient algorithm [14] and any oscillations that are likely to appear due to discontinuities can be resolved using the heuristic presented in Section III-B.

The procedure described above has transformed (3), which involves the solution of a non-convex optimization problem, into a simple selection (out of $K+1$ choices) of the rate that maximizes the net utility using (13). However, since this is an approximation method, it is necessary to determine the approximation error and propose methods to minimize it. It is easy to verify from Figure 2 that the approximation error depends on the degree of overlap[2] of the hyperbolic secant components and, moreovoer, it has its maximum values at the intersection points x_k of two consecutive components.

The effects of this overlapping can be restricted efficiently. The inflection points of the utility's sigmoidal components are determined by the technology used at the source node and they are assumed that can not be changed. However, there is often more freedom in selecting the steepness parameters of a multi-sigmoidal utility. In such cases, the steepness parameters b_k, $k = 1, 2, \ldots, K$, can be used as design parameters to assure that the approximation error is small. In this way, it is possible to calculate some additional bounds for the values of the parameters b_k of the utility function so that the

[1]The mathematical derivation of (12) is presented in detail in [16].

[2]We assume that two hyperbolic secant components c_1 and c_2 are not overlapping if $f_{c1}(x_c) = f_{c2}(x_c) \approx 0$ at their intersection point x_c.

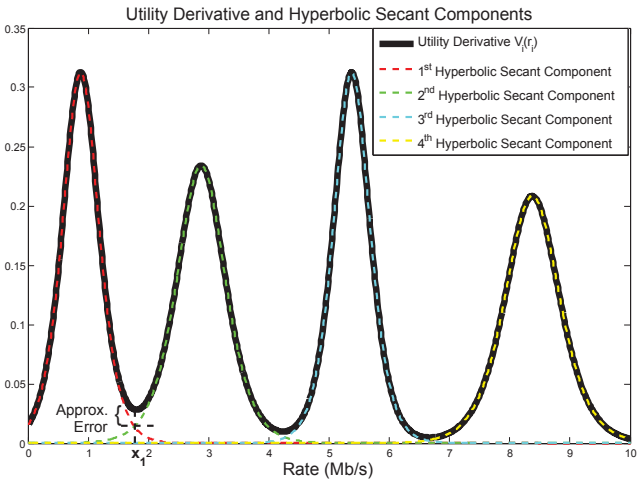

Fig. 2: Example of a multi-sigmoidal utility derivative and its 4 Hyperbolic Secant Components

hyperbolic secant components of the utility derivative are non-overlapping. In general, the smaller the values of b_k are, the more concentrated the respective hyperbolic secant component is around the inflection point. Clearly, the choice of b_k for component k affects the range of choices at the neighboring ones and therefore it is not possible to determine analytically a single steepness vector $\boldsymbol{b}$ to assure low approximation error. However, it is possible to formulate optimization problems that calculate the optimal steepness vector $\boldsymbol{b}$ according to various criteria, such as the maximum tolerable approximation error. Such an optimization problem is formulated in [16].

C. Calculating Optimal Rate in UPF

The existence of various types of user applications with diverse QoS requirements complicates the process of calculating a generic closed form solution for the optimal rate. It is however possible to derive application-specific analytical solutions for (3) in the case of *Utility Proportional fairness*.

Based on the analysis above, it is possible to derive the optimal rate allocation for browsing, file transfer and video streaming applications using the utility functions suggested in [4], [5] and [16]. These optimal rate allocation functions are demonstrated in Table I. r^{min} and r^{max} represent the minimum and maximum transmission rates of a user, and parameters α and β are calibration parameters of the single-sigmoidal utility. The calculation of analytical solutions for concave and single-sigmoidal utilities is relatively easy and will be omitted for brevity. However, the calculation for multi-sigmoidal utilities such as those described in (8) is more complicated and, therefore, will be described in detail in the remainder of this section.

Eq. (8) consists of K hyperbolic tangent components that have been scaled and shifted so that the resulting utility has values in the range $[0, 1]$. Therefore, each of the scaled components is restricted in a different non overlapping region. For example, values in the range $(0.5, 0.75)$ correspond to the third hyperbolic tangent component of the utility in the top plot of Figure 1. This implies that a value of utility belongs to only one of the hyperbolic tangent components, while the rest

TABLE I: The Optimal Resource Allocation Function for Widely Used Types of Applications

Application Type	User Utility Function	Optimal Rate Allocation Function
HTTP	$U_i\left(r_i\right)=\frac{\log\left(\frac{r_i}{r^{min}}\right)}{\log\left(\frac{r^{max}}{r^{min}}\right)}$	$r_i^*\left(\boldsymbol{\lambda}\right)=r^{min}\cdot\left(\frac{r^{max}}{r^{min}}\right)^{\frac{1}{\lambda^i}}$
FTP	$U_i\left(r_i\right)=\frac{\log(r_i+1)}{\log(r^{max}+1)}$	$r_i^*\left(\boldsymbol{\lambda}\right)=\left(r^{max}+1\right)^{\frac{1}{\lambda^i}}-1$
Single-tiered Video Application	$U_i\left(r_i\right)=\frac{1}{1+\exp(-\alpha(r_i-\beta))}$	$r_i^*\left(\boldsymbol{\lambda}\right)=\frac{\alpha\cdot\beta-\log\left(\lambda^i-1\right)}{\alpha}$
Multi-tiered Video Application	$U_i\left(r_i\right)=\frac{1}{2K}\left\{\sum_{k=1}^{K}\tanh\left(\frac{x_r-c_k}{b_k}\right)+K\right\}$	$r_i^*\left(\lambda^i\right)=b_j\cdot atanh\left(2\left(K\frac{1}{\lambda^i}-j\right)+1\right)+c_j$

of the components have value either 1 or -1. To calculate the inverse of (8), we write:

$$y=\frac{1}{2K}\left\{\sum_{k=1}^{K}\tanh\left(\frac{r_i-c_k}{b_k}\right)+K\right\}\Leftrightarrow$$

$$2Ky-K=\sum_{k=1}^{K}\tanh\left(\frac{r_i-c_k}{b_k}\right)\Rightarrow$$

$$2Ky-K=\mu+\tanh\left(\frac{r_i-c_j}{b_j}\right)-\varphi. \tag{14}$$

Index j represents the index of the hyperbolic tangent component that corresponds to the requested point. Term μ represents the components before j that have value 1, i.e. $\mu=j-1$, and term φ represents the components after j that have value -1, i.e. $\varphi=K-j$. Based on these, (14) becomes:

$$2\left(Ky-j\right)+1=\tanh\left(\frac{r_i-c_k}{b_k}\right), \tag{15}$$

and by solving with respect to r_i, we find that:

$$r_i^*\left(y\right)=b_j\cdot\operatorname{arctanh}\left(2\left(\mathrm{Ky}-\mathrm{j}\right)+1\right)+\mathrm{c_j}. \tag{16}$$

Moreover, by combining (7) and (16) we calculate the optimal rate allocation of user i with respect to the aggregate network price for i as

$$r_i^*\left(\lambda^i\right)=b_j\cdot\operatorname{arctanh}\left(2\left(\mathrm{K}\frac{1}{\lambda^{\mathrm{i}}}-\mathrm{j}\right)+1\right)+\mathrm{c_j}. \tag{17}$$

Eq. (17) is a closed form of the optimal rate allocation for a specific aggregate price λ^i when the utility function has multi-sigmoidal shape, i.e. when it models multi-tiered multimedia applications. In order to evaluate (17), it is necessary to determine the hyperbolic tangent component that the specific aggregate price λ^i corresponds to, i.e. determine the value of j. According to the *first order necessary condition for optimality* [14], at the optimal solution $\mathcal{U}_i^{'}\left(r_i^*\right)=\lambda^i$, which leads to $U_i\left(r_i^*\right)=\frac{1}{\lambda^i}$, which implies that the regions of utility values can be easily mapped to regions of aggregate price values. Specifically, for a multi-sigmoidal utility with K inflection points of the form described in (8), the hyperbolic component j is within region $\left[\frac{j-1}{K},\frac{j}{K}\right]$, with $j=1,2,\ldots,K$, of the utility values and corresponds to prices in the region $\left(\frac{K}{j},\frac{K}{j-1}\right)$, with $\frac{K}{0}\rightarrow\infty$. In other words, depending on the value of the aggregate price λ^i, we can determine the component that the optimal rate belongs to and specify j. For example, Table II shows the utility value regions and their respective aggregate price regions for a multi-sigmoidal utility given by (8) for $K=4$. Note, that aggregate prices within $[0,1)$ correspond to $U_i=1$ and therefore to component $j=K$.

By splitting the summation of hyperbolic tangent components and calculating the inverse of only one of them, we create some discontinuities on the boundaries of the aggregate price regions. These discontinuities are caused by the fact that $\operatorname{arctanh}(\mathrm{x})\rightarrow\pm\infty$ when $x\rightarrow\pm1$ respectively. Specifically for (17) the discontinuities appear on the intermediate boundaries since, by definition of the utility function, $r_i^*(0)=r_i^{max}$ and $r_i^*(\infty)=0$. For example, in the case of a multi-sigmoidal utility with $K=4$, the discontinuities exist for $\lambda^i=\frac{4}{3}$, $\lambda^i=2$ and $\lambda^i=4$. In order to handle these discontinuities and assure continuity of the rate allocation function, one could assign an approximation of the optimal rate for these boundary cases based on neighboring rate values. In other words, the optimal rate $r_i^*\left(\lambda^i\right)$ for the boundary aggregate prices can be calculated by a transformation of the form:

$$r_i^*\left(\lambda^i\right)=f\left(r_i^*\left(\lambda_-^i\right),r_i^*\left(\lambda_+^i\right)\right), \tag{18}$$

where $\lambda_-^i=\lambda^i-\epsilon$, $\lambda_+^i=\lambda^i+\epsilon$ and ϵ is a very small positive constant. A potential approach could be a weighted average of the rates for prices λ_-^i and λ_+^i according to:

$$r_i^*(\lambda^i)=\frac{w_1\cdot r_i^*(\lambda_-^i)+w_2\cdot r_i^*(\lambda_+^i)}{w_1+w_2}, \tag{19}$$

where w_1 and w_2 are weighting parameters with $w_k>0$, $k\in\{1,2\}$. The relative values of the parameters w_1 and w_2 can be used to select a rate value that is closer to one or the other discontinuity end. For example, $w_1>w_2$ implies that $r_i^*(\lambda^i)$ will be closer to $r_i^*(\lambda_-^i)$ than to $r_i^*(\lambda_+^i)$. For the numerical results later, we will use $w_1=w_2=\frac{1}{2}$ and $\epsilon=10^{-8}$ to calculate the optimal rate for boundary aggregate prices.

This weighted averaging of neighboring points for the estimation of the optimal rate is a way to make (17) a continuous function of the aggregate price. This continuity for all aggregate prices also implies that when using *utility proportional fairness* all rates within the range $\left[r^{min},r^{max}\right]$

Component	Utility Value Region	Aggregate Price Region
1	$\left[0,\frac{1}{4}\right]$	$(4,\infty)$
2	$\left[\frac{1}{K},\frac{2}{K}\right]$	$(2,4)$
3	$\left[\frac{2}{K},\frac{3}{K}\right]$	$\left(\frac{4}{3},2\right)$
4	$\left[\frac{3}{K},1\right]$	$\left[0,\frac{4}{3}\right)$

TABLE II: Tangent Components and the Respective Utility and Aggregate Price Value Regions for a Utility with $K=4$

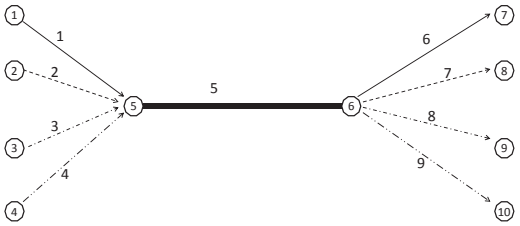

Fig. 3: Example of a network topology with a single bottleneck link

are feasible contrary to the bandwidth proportional fairness case, where only a small part of the total rate range is feasible (see bottom sub-plot in Figure 1). This shows that the rate allocation function has the robustness and elasticity to adjust to any changes in the link prices and take advantage of the full range of the available rate region in order to maximize user satisfaction in the network.

VI. Simulation Results

The optimization framework presented in the previous sections was simulated in MATLAB to study its performance. Several examples where network oscillations occurred were examined to evaluate the efficiency of ORH to stabilize the network in the case of BPF and illustrate the ability of UPF to provide stability and lead to fair allocation of resources when heterogeneous applications compete.

The simulation results are organized in two sections; a single bottleneck network case and a multiple bottleneck network one. The simulation setup included a variety of types of applications, including FTP, HTTP and multimedia applications. This dictated the use of different utility functions, concave or multi-sigmoidal, according to the type of application. All multimedia applications were modelled using multi-sigmoidal utilities according to (8) for different inflection and steepness vectors. Furthermore, the calculation of the steepness parameter vector $\boldsymbol{b}^i$ for each multi-sigmoidal utility was done by solving the optimization problem in [16] for a maximum approximation error $\sigma = 10^{-4}$ using the Global Optimization Toolbox in MATLAB, and all utilities where designed so that their maximum transmitted rate r_{max} is $10Mb/s$ and $U_i(r_{max}) = 1$ for all source nodes.

A. Single bottleneck link

Figure 3 shows an example topology of a network that has a single bottleneck link. The traffic flows are designated by a different line style. The capacities of links $1-4$ and $6-9$ where selected to be larger than r_{max}, while the capacity of link 5 was chosen so that it is inadequate for all sources to transmit at their maximum rate r_{max}, thus creating a bottleneck.

Source nodes 1 and 4 have multi-sigmoidal utilities of four hyperbolic tangent components while sources 1 and 3 represent HTTP and FTP traffic respectively and are modelled using logarithmic utility functions [5]. Several different values for the capacity of the bottleneck link were used in order to examine cases of network oscillation or stability. In essence, by increasing the bottleneck link capacity, one can decrease

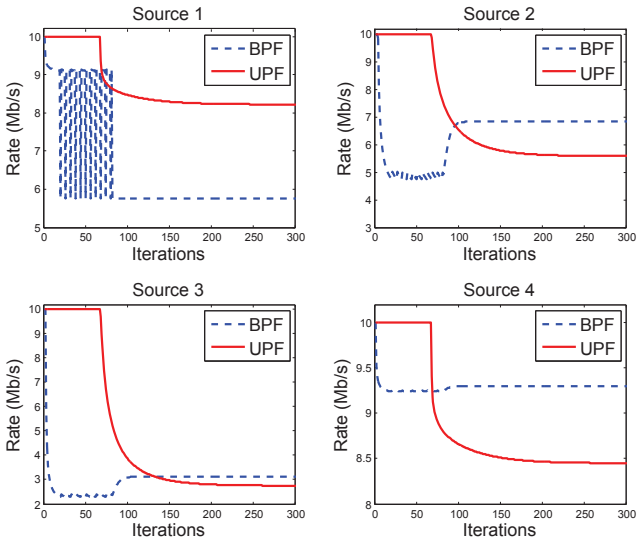

Fig. 4: Convergence of Rates

the optimal link price due to the availability of more resources and the weakening of the competition among users.

Figure 4 shows the convergence of the rates of the source nodes for bottleneck capacity $C_5 = 25Mb/s$. All links apart from link 5 have zero price and $\lambda_5 = 0.072485$. This happens as link 5 saturates and its link price increases. In BPF, user 1 starts oscillating and the ORH heuristic is invoked. The algorithm allocates some rate to User 1 and removes him from the rest of the optimization process in order to resolve oscillations and assure convergence of the distributed algorithm. As shown in blue, the oscillation of user 1 also causes other users to oscillate but at a smaller extend. On the other hand, when UPF is applied all 4 users are shown to converge smoothly without any oscillations.

The top plot in Figure 5 shows the convergence of the aggregate utility in the network, i.e. the summation of the individual utilities illustrating the effect of the oscillation of User 1 in the objective function of the optimization problem. This instability is resolved successfully by ORH. The bottom plot in Figure 5 shows that when the oscillating user is removed, the remaining users compete for the rest of the network resources which leads to higher individual utilities for these users. On the other hand, there are no oscillations when UPF is used and the resulting rate allocation leads to exactly the same degree of satisfaction for all sources. Therefore, the utility of all users is depicted using a single line (in red).

B. Multiple bottleneck links

Figure 6 illustrates a topology with three bottleneck links where eight flows are competing for network resources. The different traffic flows are distinguished by a different line style and colour combination. Links 5, 8 and 13 are the bottlenecks while the rest are sufficiently large to accommodate traffic even at the maximum rate r_{max}. Nodes 2, 3 and 6 measure user satisfaction using concave utilities, while the remaining five flows model multi-tiered multimedia applications. Figure 7 shows the convergence of the aggregate objective function when BPF and UPF are applied. As the blue line shows,

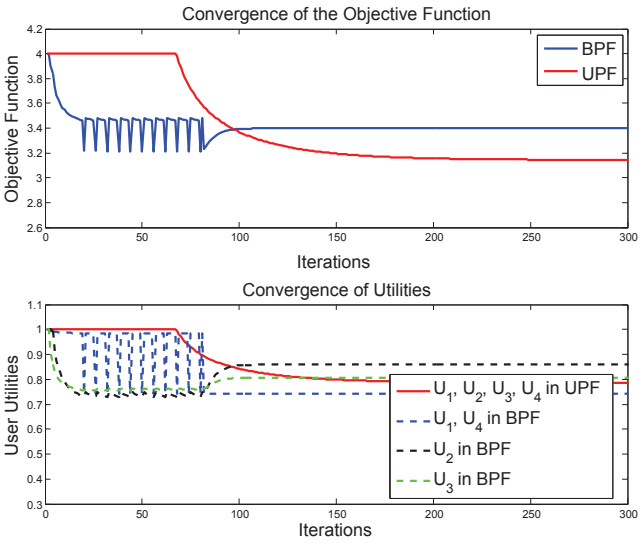

Fig. 5: Convergence of Objective Function and Individual User Utilities

the ORH can successfully resolve user oscillations that occur during the optimization process while the UPF manages to prevent any oscillations and allow the distributed optimization algorithm to converge smoothly to the optimal rates even for this more complex network scenario. The convergence of rates was omitted in this case as it resembles the behaviour in the previous simulation example, however, it is worth noting that, in general, UPF gives priority to users with higher rate requirements while BPF allocates more rate to users that are satisfied easier in an attempt to achieve higher aggregate utility in the network. This behavior can be verified in all the aforementioned figures.

VII. Concluding Remarks

This paper studied the resource allocation problem motivated by the fast growing number of multimedia applications in current communication networks. We introduced the concept of multi-sigmoidal utilities and proposed efficient methods to overcome any challenges that their use imposes for two different allocation policies, BPF and UPF. We proposed a novel mathematical representation of such utility functions and a distributed algorithm to optimize the allocation of bandwidth by exploiting the special structure of the utility function. Finally, the performance and robustness of the proposed framework were evaluated through extensive simulations for various network topologies.

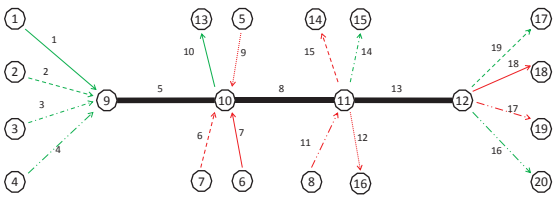

Fig. 6: Example of a network topology with multiple bottleneck links

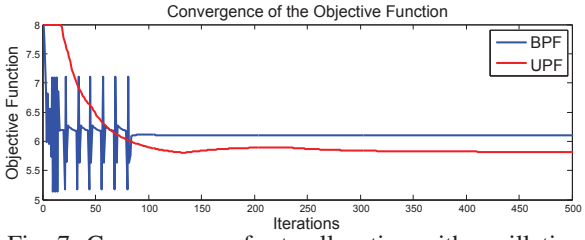

Fig. 7: Convergence of rate allocation with oscillation

References

[1] S. Low, "A duality model of tcp and queue management algorithms," *IEEE/ACM Transactions on Networking*, vol. 11, no. 4, pp. 525 – 536, Aug. 2003.

[2] W. Stallings, *Data and Computer Communications*, 9th ed. Pearson Custom Publishing, 2010.

[3] S. Shenker, "Fundamental design issues for the future internet," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1176–1188, September 1995.

[4] J.-W. Lee, R. R. Mazumdar, and N. B. Shroff, "Non-convex optimization and rate control for multi-class services in the internet," *IEEE J. on Selected Areas in Commun.*, vol. 13, no. 4, pp. 827–840, August 2005.

[5] C. Liu, L. Shi, and B. Liu, "Utility-based bandwidth allocation for triple-play services," in *Universal Multiservice Networks, 2007. ECUMN '07. Fourth European Conference on*, feb. 2007, pp. 327 –336.

[6] Cisco, "Visual networking index: Global mobile data traffic forecast update, 20102015," http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf, Tech. Rep., February 2011.

[7] G. Tychogiorgos, A. Gkelias, and K. Leung, "A non-convex distributed optimization framework and its application to wireless ad-hoc networks," *Wireless Communications, IEEE Transactions on*, vol. 12, no. 9, pp. 4286–4296, September 2013.

[8] F. P. Kelly, A. Maulloo, and D. Tan, "Rate control in communication networks: Shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, pp. 237–252, 1998.

[9] M. Chiang, S. Low, A. Calderbank, and J. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255 –312, jan. 2007.

[10] J.-W. Lee, R. Mazumdar, and N. Shroff, "Nonconvexity issues for internet rate control with multiclass services: stability and optimality," in *INFOCOM 2004*, vol. 1, March 2004, pp. 24–34.

[11] P. Hande, S. Zhang, and M. Chiang, "Distributed rate allocation for inelastic flows," *Networking, IEEE/ACM Transactions on*, vol. 15, no. 6, pp. 1240 –1253, December 2007.

[12] G. Tychogiogos, A. Gkelias, and K. K. Leung, "A new distributed optimization framework for hybrid ad-hoc networks," in *IEEE GlobeCom 2011 – Workshop on Heterogeneous, Multi-Hop, Wireless and Mobile Networks*, Houston, USA, December 2011.

[13] W.-H. Wang, M. Palaniswami, and S. H. Low, "Application-oriented flow control: Fundamentals, algorithms and fairness," *IEEE/ACM Transactions on Networking*, vol. 14, no. 6, pp. 1282 –1291, December 2006.

[14] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.

[15] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[16] "Optimizing resource allocation for multi-tiered multimedia applications," https://dl.dropboxusercontent.com/u/16464897/Report.pdf, Online, Tech. Rep., July 2014.

[17] J. Drakapoulos, "Multi-sigmoidal neural networks and back-propagation," in *Artificial Neural Networks, Fourth International Conference on*, June 1995, pp. 154 –159.