

Fair resource allocation schemes and network dimensioning problems

Włodzimierz Ogryczak, Tomasz Śliwiński, and Adam Wierzbicki

Abstract — Resource allocation problems are concerned with the allocation of limited resources among competing activities so as to achieve the best overall performances of the system but providing fair treatment of all the competitors. Telecommunication networks are facing the increasing demand for Internet services. Therefore, a problem of network dimensioning with elastic traffic arises which requires to allocate bandwidth to maximize service flows with fair treatment of all the services. In such applications, the so-called max-min fairness (MMF) solution concept is widely used to formulate the resource allocation scheme. This guarantees the fairness but may lead to significant losses in the overall throughput of the network. In this paper we show how multiple criteria optimization concepts can be used to generate various fair resource allocation schemes. The solution concepts are tested on the network dimensioning problem and their abilities to model various preferences are demonstrated.

Keywords — telecommunication networks, network dimensioning, resource allocation, fairness.

1. Introduction

Resource allocation decisions are concerned with the allocation of limited resources so as to achieve the best system performances. In this paper, we focus on approaches that, while allocating resources, attempt to provide a fair (equal) treatment of all the competing activities [8, 13]. The problems of efficient and fair resource allocation arise in various systems which serve many users, like in telecommunication systems among others [8].

The development of the Internet has led to an increased role of the traffic carried by the IP protocol in telecommunication networks. Due to the use of packet switching, the IP protocol can provide greater network utilization (the so-called multiplexing gain). For these reasons, network management can be interested in designing networks which have a high throughput for the IP protocol.

At the same time, data traffic carried by the TCP protocol (which is the most frequently used transport protocol in IP networks) has a unique characteristic. The TCP protocol will adapt its throughput to the amount of available bandwidth. It is therefore capable to use the entire available bandwidth, but it will also be able to reduce its throughput in the presence of contending traffic. This type of network traffic has been called *elastic traffic*.

Network design today often considers the problem of designing networks that carry elastic traffic. If the network is also used for other types of communication that require guaranteed quality of service, the network design problem

can be decomposed into two parts: first, design the network to carry non-elastic traffic in such a way that all demands for that communication are satisfied. Next, use the spare capacity to carry elastic traffic of the IP protocol. Resource allocation models may be used to help to solve such network design problems.

Within a telecommunication network the data traffic is generated by a huge number of nodes exchanging data. In such a network, a relatively small subset of nodes are chosen to serve as hubs which can be used as intermediate switching points [2, 6]. Given a set of hubs, data traffic generated by a service is sent from the source node to a hub first. It can be then sent along communications link between hubs, and finally reach the destination node along a link from a hub. The hub-based network organization allows the data traffic to be consolidated on the inter-hub links. The problem of network dimensioning with elastic traffic arises when there is a need to design the (inter-hub) link capacities to carry as much traffic as possible between a set of network nodes. This can occur in the case described above, when the network capacity available after considering all non-elastic demands has to be used for elastic traffic, or in another case: when the network capacity is insufficient to carry all non-elastic demands. In such a case, the problem is to determine how much traffic of the non-elastic demands can be admitted into the network. To do so, the demands can be treated as elastic traffic. The outcome of network design will also specify the limits of traffic to be admitted into the network for each demand [16].

Network management must stay within a budget of expenses for purchasing link bandwidth. Network management will want to have a high throughput of the IP network, to increase the multiplexing gains. This traffic is offered only a best-effort service, and therefore network management is not concerned with offering guaranteed levels of bandwidth to the traffic. Network dimensioning with elastic traffic can therefore be thought of as a search for such network flows that will maximize the network throughput (the sum of all flows in the network) while staying within a budget constraint for the costs of link bandwidth. However, such a problem formulation would lead to the starvation of flows between certain network nodes.

Looking at the problem from the user perspective, the network flows between different nodes should be treated as fairly as possible. The users may be interested in high available bandwidth between any two nodes of the network, or in high available bandwidth from all other network nodes to the user's node, or in high available bandwidth from the user's node to all other nodes. Whatever the user prefer-

ence, it would be expressed in terms of fairness for a certain set of criteria which depend on the individual flows. Let us first consider providing fairness for all flows between any two network nodes. Such a goal would clearly lead to lower levels of throughput, since resources must be allocated to distant nodes, which is more expensive than using the entire budget to purchase a high capacity for close nodes.

Therefore, network management must consider two goals: increasing throughput and providing fairness. These two goals are clearly conflicting, if the budget constraint has to be satisfied. Network management could therefore be interested in finding compromise solutions that do not starve network flows, and give satisfying levels of throughput.

The search for such compromise solutions has led to the development of a method that finds solutions which are fair with respect to flows in certain categories. These categories can depend on the distance between the source and destination of a flow. The details of this method will be given below; it is referred to as proportional fairness (PF) [5]. However, this method gives only one possible compromise solution. The purpose of this work is to show that there exists a methodology that allows the decision maker to explore a set of solutions that could satisfy his preferences with respect to throughput and fairness, and choose the solution which the decision maker finds best. This interactive approach to decision making is superior to a black box approach, when the decision maker has only one solution and cannot express his preferences [18].

The paper is organized as follows. In the next section we recall the network dimensioning problem. In Section 3, basic fair solution concepts for resource allocation are formally introduced. In the next section, the ordered outcomes are used to introduce LP implementable solution concepts allowing to model various fair allocation schemes. Finally, in Section 5, we report some results of our initial computational experience with this new approach.

2. The network dimensioning problem

The generic resource allocation problem may be stated as follows. There is given a set I of m services. There is also given a set Q of allocation patterns (allocation decisions). For each service $i \in I$ a function $f_i(\mathbf{x})$ of the allocation pattern \mathbf{x} has been defined. This function, called the individual objective function, measures the outcome (effect) $y_i = f_i(\mathbf{x})$ of the allocation pattern for service i . In applications, we consider, an outcome usually expresses the service flow. However, outcomes can be measured (modeled) as service time, service costs, service delays as well as in a more subjective way. In typical formulations a larger value of the outcome means a better effect (higher service quality or client satisfaction). Otherwise, the outcomes can be replaced with their complements to some large number. Therefore, without loss of generality, we can assume that each individual outcome y_i is to be maximized which results in a multiple criteria maximization model.

The problem of network dimensioning with elastic traffic can be formulated as a linear programming (LP) resource allocation problem as follows. Given a network routing topology $G = \langle V, E \rangle$, consider a set of pairs of nodes as the set I of services. For each service $i \in I$, the elastic flow from source u_i^s to destination u_i^d will be denoted by y_i , which is a variable representing the model outcome. For each service, we have given the information about the routing path in the network from the source to the destination. This information can be in the form of a matrix $\mathbf{A} = (a_{ei})$, which satisfies the relation: $a_{ei} = 1$ if link e belongs to the routing path connecting u_i^s with u_i^d . Further, for each link $e \in E$, marginal costs c_e of link bandwidths is given. Hence, the cost of the entire path for service i can be expressed as:

$$\kappa_i = \sum_{e \in E} c_e a_{ei}.$$

The network dimensioning problem depends on allocating the bandwidth to several links in order to maximize flows of all the services while remaining within available budget B for all link bandwidths. The decisions are usually modeled with (decision) variables: x_e – representing the bandwidth allocated to link $e \in E$. They have to fulfill the following constraints:

$$\sum_{e \in E} c_e x_e = B \quad (1)$$

$$\sum_i a_{ei} y_i = x_e \quad \forall e \in E, \quad (2)$$

where Eq. (1) represents the budget limit while Eqs. (2) establish the relation between service flows and links bandwidth (the quantity $\sum_{i \in I} a_{ei} y_i$ is the load of link e). Certainly, all the decision and outcome variables must be non-negative: $x_e \geq 0$ for all $e \in E$ and $y_i \geq 0$ for all $i \in I$. Alternatively, one may eliminate variables x_e formulating the problem as a simplified resource allocation model with only one constraint:

$$\sum_{i=1}^m \kappa_i y_i = B \quad (3)$$

and variables y_i representing directly decisions.

The model could have various objective functions, depending on the chosen approach. One may consider two extreme approaches. The first extreme approach is the maximization of the throughput (the sum of flows) $\sum_{i \in I} y_i$. Due to possible alternative formulation (3), it is apparent that this approach would choose one variable y_{i^0} which has the smallest marginal cost $\kappa_{i^0} = \min_{i \in I} \kappa_i$ and make that flow maximal within the budget limit ($y_{i^0} = B/\kappa_{i^0}$), while limiting all other flows to zero. A slightly more fair optimal solution would give equal values to all flows which have marginal costs equal to the minimal marginal cost. However, all flows that have marginal costs larger than the minimum would have to be zero in a solution that maximizes throughput.

The so-called max-min fairness solution concept is widely used to formulate fair resource allocation schemes [1, 8]. The worst performance (minimum flow) is there maximized

and additionally regularized, if necessary, with the lexicographic (sequential) maximization of the second worst performance, the third worst etc. The MMF concept is consistent with Rawlsian [15] theory of justice.

Actually, due to possible alternative formulation (3), the MMF concept would lead us to a solution that has equal values for all the flows [12]:

$$y_i^{MMF} = B / \sum_{i \in I} \kappa_i \quad \text{for } i = 1, \dots, m.$$

Allocating the resources to optimize the worst performances may cause a large worsening of the overall (mean) performances. In such a solution the throughput ($mB / \sum_{i=1}^m \kappa_i$) could be considerably smaller than the maximal throughput ($B / \min_{i=1, \dots, m} \kappa_i$). In an example analyzed further, we shall show that the throughput in a perfectly fair solution can be less than 50% of the maximal throughput.

Network management can be interested in seeking a compromise between the two extreme approaches discussed above. The approach called *proportional fairness* proposed in [5] maximizes the sum of logarithms of the flows y_i . The use of the logarithmic function makes it impossible to choose zero flows for any pair of nodes, and, on the other hand, makes it not profitable to assign too much flow to any individual demand. The optimization model of the PF method takes the following form:

$$\max \sum_{i=1}^m \log(y_i). \quad (4)$$

For the problem of network dimensioning with elastic traffic and unbounded flows, the solution found by the PF method has an interesting property [11]. The optimal flows y_i^{PF} are given by the expression:

$$y_i^{PF} = B / \kappa_i \quad \text{for } i = 1, \dots, m. \quad (5)$$

This property implies that the optimal flow in the PF model is inversely proportional to the cost of the path that the flow travels in the network. Due to this property, it is not necessary to solve nonlinear models in order to find the PF optimal solution. Also, the solution provides fairness to the flows which have the same path costs. Arguably, the PF solution is a good compromise solution to the problem, since it provides a higher throughput than the perfectly fair solution. However, network management could be interested in choosing among a larger set of compromise solutions in order to satisfy their preferences. In the following sections, we shall describe an approach that allows to search for such compromise solutions.

3. Basic fair allocation schemes

Consider a generic resource allocation problem defined as an optimization problem with m objective functions $f_i(\mathbf{x})$:

$$\max \{ \mathbf{f}(\mathbf{x}) : \mathbf{x} \in Q \}, \quad (6)$$

where $\mathbf{f}(\mathbf{x})$ is a vector-function that maps the decision space $X = R^n$ into the criterion space $Y = R^m$, $Q \subset X$ denotes

the feasible set, and $\mathbf{x} \in X$ denotes the vector of decision variables.

Model (6) only specifies that we are interested in maximization of all objective functions f_i for $i \in I = \{1, 2, \dots, m\}$. In order to make it operational, one needs to assume some solution concept specifying what it means to maximize multiple objective functions.

Typical solution concepts for multiple criteria problems are defined by aggregation functions $g : Y \rightarrow R$ to be maximized. Thus the multiple criteria problem (6) is replaced with the maximization problem

$$\max \{ g(\mathbf{f}(\mathbf{x})) : \mathbf{x} \in Q \}. \quad (7)$$

In order to guarantee the consistency of the aggregated problem (7) with the maximization of all individual objective functions in the original multiple criteria problem, the aggregation function must be strictly increasing with respect to every coordinate, i.e., for all $i \in I$,

$$g(y_1, \dots, y_{i-1}, y'_i, y_{i+1}, \dots, y_m) < g(y_1, y_2, \dots, y_m) \quad (8)$$

whenever $y'_i < y_i$.

In order to guarantee fairness (equitability) of the solution concept, the aggregation function must be additionally symmetric (impartial), i.e. for any permutation τ of I ,

$$g(y_{\tau(1)}, y_{\tau(2)}, \dots, y_{\tau(m)}) = g(y_1, y_2, \dots, y_m) \quad (9)$$

as well as be equitable (to satisfy the principle of transfers)

$$g(y_1, \dots, y_p - \varepsilon, \dots, y_p + \varepsilon, \dots, y_m) > g(y_1, y_2, \dots, y_m) \quad (10)$$

for any $0 < \varepsilon < y_p - y_{p'}$. In the case of an aggregation function satisfying all the requirements (8), (9) and (10), we call the corresponding problem (7) a *fair (equitable) aggregation* of problem (6). Every optimal solution to the fair aggregation (7) of a resource allocation problem (6) defines some fair allocation scheme.

Note that symmetric functions satisfying the requirement

$$g(y_1, \dots, y_p - \varepsilon, \dots, y_p + \varepsilon, \dots, y_m) \geq g(y_1, y_2, \dots, y_m) \quad (11)$$

for $0 < \varepsilon < y_p - y_{p'}$ are called (weakly) Schur-concave [10] while the stronger requirement of equitability (10), we consider, is related to strictly Schur-concave functions. In other words, an aggregation (7) is fair if it is defined by a strictly increasing and strictly Schur-concave function g .

The simplest aggregation functions commonly used for the multiple criteria problem (6) are defined as the sum of outcomes

$$g(\mathbf{y}) = \sum_{i=1}^m y_i \quad (12)$$

or the worst outcome

$$g(\mathbf{y}) = \min_{i=1, \dots, m} y_i. \quad (13)$$

In the network dimensioning problem, the former represents throughput maximization while the latter corresponds

to the MMF model. The sum (12) is a strictly increasing function while the minimum (13) is only non-decreasing. Therefore, the aggregation (7) using the sum of outcomes always generates a Pareto-optimal solution while the maximization of the worst outcome may need some additional refinement. Both the functions are symmetric and satisfy the requirement (11), although they do not satisfy the equitability requirement (10). Hence, they are Schur-concave but not strictly Schur-concave. To guarantee the fairness of solutions, some enforcement of concave properties is required.

For any strictly concave, increasing function $s : R \rightarrow R$, the function

$$g(\mathbf{y}) = \sum_{i=1}^m s(y_i) \quad (14)$$

is a strictly monotonic and strictly Schur-concave function [10]. This defines a family of the fair aggregations according to the following corollary [7].

Corollary 1. For any strictly convex, increasing function $s : R \rightarrow R$, the optimal solution of the problem

$$\max \left\{ \sum_{i=1}^m s(f_i(\mathbf{x})) : \mathbf{x} \in Q \right\} \quad (15)$$

is a fair solution for resource allocation problem (6).

In the case of the outcomes restricted to positive values, one may use logarithmic function thus resulting in the proportional fairness model (4). Various other concave functions s can be used to define fair aggregations (15) and the resulting resource allocation schemes. However, the problem of network dimensioning, we consider, is originally an LP model. Therefore, it is important if various fair allocation schemes can be generated with LP tools. We will show such LP models in the next section.

The standard maximin approach (13) may be lexicographically extended to the full MMF model where, in addition to the smallest outcome, one maximizes also the second smallest outcome (provided that the smallest one remains as large as possible), maximizes the third smallest (provided that the two smallest remain as large as possible), etc. Note that the lexicographic maximization is not applied to any specific order of the original criteria. Nevertheless, in the case of LP problems, there exists a dominating objective function which is constant on the entire optimal set of the maximin problem [9]. Hence, having solved the maximin problem, one may try to identify the dominating objective and eliminate it to formulate a restricted maximin problem on the former optimal set. Therefore, the lexicographic maximin solution to LP problems can be found by sequential maximin optimization with elimination of the dominating functions. Although, the LP models, we will present in the next section, provide us with a direct formulation for the MMF model.

4. Ordered outcomes

Multiple criteria optimization defines the dominance relation by the standard vector inequality. The theory of majorization [10] includes the results which allow us to express the relation of fair (equitable) dominance as a vector inequality on the cumulative ordered outcomes [7]. This can be mathematically formalized as follows. First, introduce the ordering map $\Theta : R^m \rightarrow R^m$ such that $\Theta(\mathbf{y}) = (\theta_1(\mathbf{y}), \theta_2(\mathbf{y}), \dots, \theta_m(\mathbf{y}))$, where $\theta_1(\mathbf{y}) \leq \theta_2(\mathbf{y}) \leq \dots \leq \theta_m(\mathbf{y})$ and there exists a permutation τ of set I such that $\theta_i(\mathbf{y}) = y_{\tau(i)}$ for $i = 1, \dots, m$. Next, apply to ordered outcomes $\Theta(\mathbf{y})$, a linear cumulative map thus resulting in the *cumulative ordering map* $\bar{\Theta}(\mathbf{y}) = (\bar{\theta}_1(\mathbf{y}), \bar{\theta}_2(\mathbf{y}), \dots, \bar{\theta}_m(\mathbf{y}))$ defined as

$$\bar{\theta}_i(\mathbf{y}) = \sum_{j=1}^i \theta_j(\mathbf{y}) \quad \text{for } i = 1, \dots, m. \quad (16)$$

The coefficients of vector $\bar{\Theta}(\mathbf{y})$ express, respectively: the smallest outcome, the total of the two smallest outcomes, the total of the three smallest outcomes, etc.

Vector $\bar{\Theta}(\mathbf{y})$ can be viewed graphically with a piecewise linear curve connecting point (0,0) and points $(i/m, \bar{\theta}_i(\mathbf{y})/m)$ for $i = 1, \dots, m$. Such a curve represents the absolute Lorenz curve which can be mathematically formalized as follows. First, we introduce the right-continuous cumulative distribution function:

$$F_{\mathbf{y}}(d) = \sum_{i=1}^m \frac{1}{m} \delta_i(d), \quad \text{where } \delta_i(d) = \begin{cases} 1 & \text{if } y_i \leq d \\ 0 & \text{otherwise} \end{cases}$$

which for any real value d provides the measure of outcomes smaller or equal to d . Next, we introduce the quantile function $F_{\mathbf{y}}^{(-1)}$ as the left-continuous inverse of the cumulative distribution function $F_{\mathbf{y}}$:

$$F_{\mathbf{y}}^{(-1)}(\eta) = \inf \{ d : F_{\mathbf{y}}(d) \geq \eta \} \quad \text{for } 0 < \eta \leq 1.$$

By integrating $F_{\mathbf{y}}^{(-1)}$ one gets $F_{\mathbf{y}}^{(-2)}(0) = 0$ and

$$F_{\mathbf{y}}^{(-2)}(\eta) = \int_0^{\eta} F_{\mathbf{y}}^{(-1)}(\alpha) d\alpha \quad \text{for } 0 < \eta \leq 1.$$

Graphs of functions $F_{\mathbf{y}}^{(-2)}(\eta)$ (with respect to η) take the form of concave curves (Fig. 1), the *(upper) absolute Lorenz curves*. In our case of m outcomes, the absolute Lorenz curve is completely defined by the values $F_{\mathbf{y}}^{(-2)}(i/m) = \frac{1}{m} \bar{\theta}_i(\mathbf{y})$ for $i = 1, \dots, m$, where $F_{\mathbf{y}}^{(-2)}(1/m) = \bar{\theta}_1(\mathbf{y}) = \theta_1(\mathbf{y})$ represent the worst outcome and $F_{\mathbf{y}}^{(-2)}(1) = \frac{1}{m} \bar{\theta}_m(\mathbf{y}) = \frac{1}{m} \sum_{i=1}^m \theta_i(\mathbf{y})$.

In income economics the Lorenz curve is a cumulative population versus income curve [10]. A perfectly equal distribution of income has the diagonal line as the Lorenz curve and no outcome vector can be better. The absolute Lorenz curves, we consider, are unnormalized taking into account also values of outcomes. Vectors of equal outcomes are distinguished according to the value of outcomes. They are

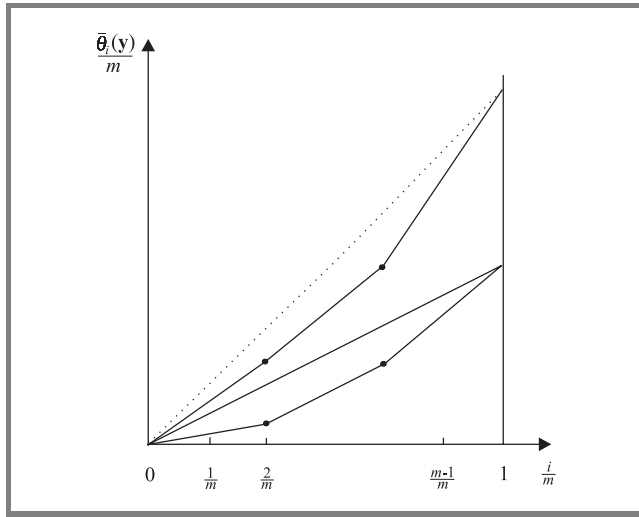


Fig. 1. Vectors $\bar{\Theta}(\mathbf{y})$ as the absolute Lorenz curves.

graphically represented with various ascent lines in Fig. 1. Hence, with the relation of fair dominance an outcome vector of large unequal outcomes may be preferred to an outcome vector with small equal outcomes.

Note that fair solutions to problem (6) can be expressed as Pareto-optimal solutions for the multiple criteria problem with objectives $\bar{\Theta}(\mathbf{f}(\mathbf{x}))$

$$\max \left\{ (\bar{\theta}_1(\mathbf{f}(\mathbf{x})), \bar{\theta}_2(\mathbf{f}(\mathbf{x})), \dots, \bar{\theta}_m(\mathbf{f}(\mathbf{x}))) : \mathbf{x} \in Q \right\}. \quad (17)$$

Corollary 2. A feasible solution $\mathbf{x} \in Q$ is a fair solution of the resource allocation problem (6), iff it is a Pareto-optimal solution of the multiple criteria problem (17).

Corollary 2 provides the relationship between fair allocation schemes and Pareto-optimality. Moreover, the multiple criteria problem (17) may serve as a source of fair allocation schemes.

Although the definitions of quantities $\bar{\theta}_k(\mathbf{y})$, used as criteria in (17), are very complicated, the quantities themselves can be modeled with simple auxiliary variables and constraints. It is commonly known that the worst (largest) outcome may be defined by the following optimization: $\theta_1(\mathbf{y}) = \max\{t : t \leq y_i \text{ for } i = 1, \dots, m\}$, where t is an unrestricted variable. It turns out that this approach can be generalized to provide an effective modeling technique for quantities $\bar{\theta}_k(\mathbf{y})$ with arbitrary k [14]. Namely, for a given outcome vector \mathbf{y} the quantity $\bar{\theta}_k(\mathbf{y})$ may be found by solving the following linear program:

$$\begin{aligned} \bar{\theta}_k(\mathbf{y}) = \max & \quad kt - \sum_{i=1}^m d_i \\ \text{s.t. } & \quad t - y_i \leq d_i, \quad d_i \geq 0 \quad \text{for } i = 1, \dots, m, \end{aligned} \quad (18)$$

where t is an unrestricted variable while nonnegative variables d_i represent, for several outcome values y_i , their downside deviations from the value of t . Independently from the formal proof [14], this formula can be justified as follows. It is obvious that $\max(kt - \sum_{i=1}^m d_i) = \bar{\theta}_k(\mathbf{y})$

whenever no more than $k - 1$ deviations d_i are strictly positive. On the other hand, for any t and d_i feasible to (18) one can define an alternative feasible values: $\tilde{t} = t - \Delta$ and $\tilde{d}_i = d_i - \Delta$ for $d_i > 0$, where Δ is an arbitrary small positive number. For at least k positive values one gets $k\tilde{t} - \sum_{i=1}^m \tilde{d}_i \geq kt - \sum_{i=1}^m d_i$, which justifies (18).

Formula (18) provides us with a computational formulation for the worst conditional mean $M_{\frac{k}{m}}(\mathbf{y})$ defined as the mean outcome for the k worst-off services, i.e.:

$$M_{\frac{k}{m}}(\mathbf{y}) = \frac{1}{k} \bar{\theta}_k(\mathbf{y}), \quad \text{for } k = 1, \dots, m. \quad (19)$$

Note that for $k = 1$, $M_{\frac{1}{m}}(\mathbf{y}) = \bar{\theta}_1(\mathbf{y}) = \theta_1(\mathbf{y}) = M(\mathbf{y})$ thus representing the minimum outcome, and for $k = m$, $M_{\frac{m}{m}}(\mathbf{y}) = \frac{1}{m} \bar{\theta}_m(\mathbf{y}) = \frac{1}{m} \sum_{i=1}^m \theta_i(\mathbf{y}) = \frac{1}{m} \sum_{i=1}^m y_i = \mu(\mathbf{y})$ which is the mean outcome. Formula (18) allows us to maximize effectively the worst conditional means for various intermediate values k [13].

Note that Corollary 2 allows one to generate equitably efficient solutions of (6) as efficient solutions of problem (17). The aggregation maximizing the sum of outcomes, corresponds to maximization of the last (m th) objective in problem (17). Similar, the maximin scalarization corresponds to maximization of the first objective in (17). For modeling various fair preferences one may use some combinations of the cumulative ordered outcomes $\bar{\theta}_i(\mathbf{y})$. In particular, for the weighted sum on gets

$$\sum_{i=1}^m w_i \bar{\theta}_i(\mathbf{y}). \quad (20)$$

Note that, due to the definition of map $\bar{\Theta}$ with (16), the above function can be expressed in the form with weights $v_i = \sum_{j=i}^m w_j$ ($i = 1, \dots, m$) allocated to coordinates of the ordered outcome vector. Such an approach to aggregation of outcomes was introduced by Yager [19] as the so-called ordered weighted averaging (OWA). When applying OWA to problem (6) we get

$$\max \left\{ \sum_{i=1}^m v_i \theta_i(\mathbf{f}(\mathbf{x})) : \mathbf{x} \in Q \right\}. \quad (21)$$

The OWA aggregation is obviously a piece wise linear function since it remains linear within every area of the fixed order of arguments.

If weights v_i are strictly decreasing and positive, i.e. $v_1 > v_2 > \dots > v_{m-1} > v_m > 0$, then each optimal solution of the OWA problem (21) is a fair solution of (6). Moreover, in the case of LP models, as the network dimensioning one, every fair allocation scheme can be identified as an optimal solution to some OWA problem with appropriate monotonic weights [7].

While equal weights define the linear aggregation, several decreasing sequences of weights lead to various strictly Schur-concave and strictly monotonic aggregation functions. Thus, the monotonic OWA aggregations provide a family of piece wise linear aggregations filling out the

space between the piece wise linear aggregation functions (12) and (13) as shown in Fig. 2. Actually, formulas (20) and (18) allow us to formulate any monotonic (not necessarily strictly) OWA problem (21) as the following LP extension of the original multiple criteria problem:

$$\max \sum_{k=1}^m w_k z_k \tag{22}$$

subject to $\mathbf{x} \in Q$

$$z_k = kt_k - \sum_{i=1}^m d_{ik} \quad \text{for } k = 1, \dots, m \tag{23}$$

$$t_k - d_{ik} \leq f_i(\mathbf{x}), \quad d_{ik} \geq 0 \quad \text{for } i, k = 1, \dots, m, \tag{24}$$

where $w_m = v_m$ and $w_k = v_k - v_{k+1}$ for $k = 1, \dots, m - 1$.

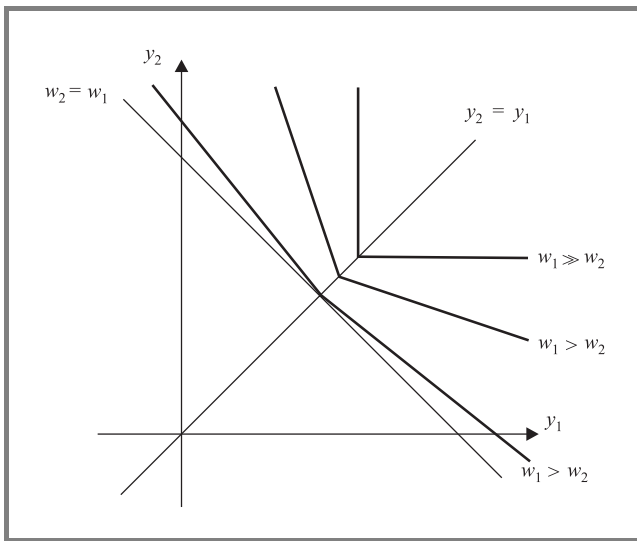


Fig. 2. Isoline contours for equitable OWA.

When differences among weights tend to infinity, the OWA aggregation approximates the lexicographic ranking of the ordered outcome vectors [20]. That means, as the limiting case of the OWA problem (21), we get the lexicographic problem:

$$\text{lexmax } \{ \Theta(\mathbf{f}(\mathbf{x})) : \mathbf{x} \in Q \} \tag{25}$$

which represents the MMF (lexicographic maximin) approach to the original resource allocation problem (6). Problem (25) is a regularization of the standard maximin optimization (13), but in the former, in addition to the worst outcome, we maximize also the second worst outcome (provided that the smallest one remains as large as possible), maximize the third worst (provided that the two smallest remain as large as possible), and so on. Due to (16), the MMF problem (25) is equivalent to the problem:

$$\text{lexmax } \{ \tilde{\Theta}(\mathbf{f}(\mathbf{x})) : \mathbf{x} \in Q \}$$

which leads us to a standard lexicographic optimization with predefined linear criteria defined according to (18).

5. Computational results

First we have tested the OWA computational models (22)–(24) when applied to a generic LP resource allocation problem. We tested solution times for different size parameters. For each number of decision variables n and number of criteria (services) m we solved 20 randomly generated problems (Table 1). All computations were performed on a PC with the Pentium 200 MHz processor employing the CPLEX 6.0 package [4].

Table 1
Computation times for randomly generated problems

Services m	Allocations – n					
	5	10	20	40	60	100
10	0.05	0.10	0.10	0.15	0.15	0.20
20	0.30	0.35	0.40	0.60	0.75	1.00
30	0.80	1.00	1.55	2.15	2.65	3.35
40	1.95	2.35	3.20	5.25	6.75	9.50
60	7.30	8.80	10.95	20.75	31.30	44.95
100	49.05	54.60	65.40	104.15	173.10	278.80

Further we have analyzed sample network dimensioning problem with elastic traffic. For this purpose we have considered a network of the topology is patterned after the backbone network of a Polish ISP (Fig. 3). The network has 12 nodes, and we consider flows between any pair of these nodes (therefore, there are $144 - 12 = 132$ flows). All links have marginal costs equal to one, and the budget for link bandwidth is $B = 1000$. Since all links have equal costs of one, path cost will be equal to the link length, which is 1, 2, 3 or 4 in the example topology. All flows are unbounded. However, it is clear that due to the budget constraint no flow can exceed B .

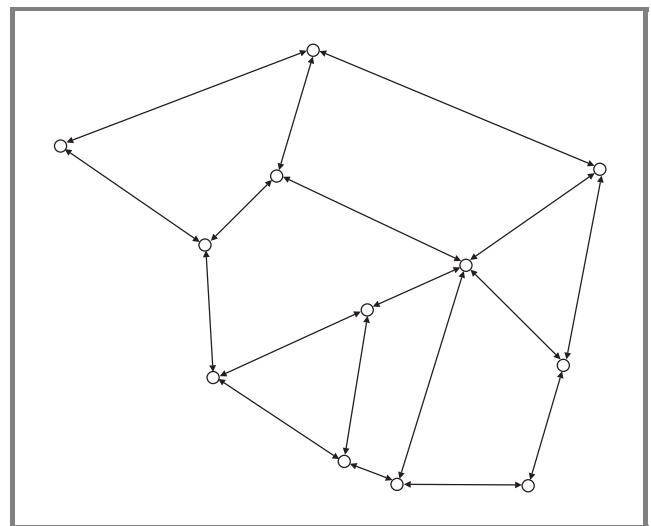


Fig. 3. Sample network topology.

To have control over the solution that will be found by the model, we decided to scale the outcomes (flows). Following the concepts of reference point methodology [17], we as-

sume that the decision maker (DM) specifies requirements in terms of aspiration and reservation levels, i.e., by introducing desired (acceptable) and required values for several outcomes. Depending on the specified aspiration and reservation levels, y_i^a and y_i^r , respectively, a special achievement function is built which can be interpreted as a measure of the DM's satisfaction with the current value of outcome the i th outcome. It is a strictly increasing function of outcome y_i with value 1 if $y_i = y_i^a$, and value 0 for $y_i = y_i^r$. Thus the partial achievement functions map the outcomes values onto a normalized scale of the DM's satisfaction. Various functions can be built meeting those requirements [18]. We use the piece wise linear function:

$$\sigma_i(y_i) = \begin{cases} \gamma(y_i - y_i^r)/(y_i^a - y_i^r), & \text{for } y_i \leq y_i^r \\ (y_i - y_i^r)/(y_i^a - y_i^r), & \text{for } y_i^r < y_i < y_i^a \\ \beta(y_i - y_i^a)/(y_i^a - y_i^r) + 1, & \text{for } y_i \geq y_i^a \end{cases}$$

where β and γ are arbitrarily defined parameters satisfying $0 < \beta < 1 < \gamma$. Parameter β represents additional increase of the DM's satisfaction over level 1 when a criterion generates outcomes better than the corresponding aspiration level. On the other hand, parameter $\gamma > 1$ represents dissatisfaction connected with outcomes worse than the reservation level. The achievement function σ_i can be viewed as an extension of the fuzzy membership function to a strictly monotonic and concave utility (Fig. 4).

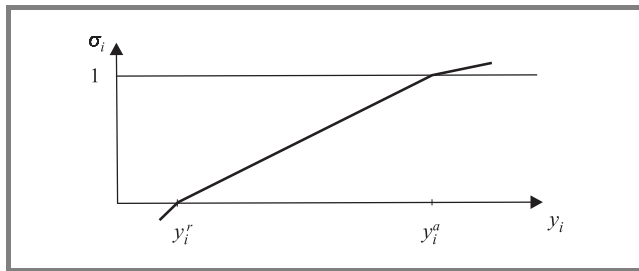


Fig. 4. Outcomes scaled with the achievement function.

The scaled flows are combined into an objective function using the OWA model. The linear program formulation of the OWA approach uses weights w_i , which are first-order differences of the weights v_i which are coefficients of the ordered outcome vector in the OWA model. In the approach used here, the weights $w_i = 1$ for all i . Thus, the OWA model has linearly decreasing weights. In the next section, we shall apply the outlined approach to search for compromise solutions of the network dimensioning problem with elastic flows using the sample topology given in Fig. 3.

The first application of the outlined approach used the same reservation and aspiration levels for all flows. Predictably, the result was a perfectly fair solution with each flow equal to 3.546, and a throughput of 468.1. This solution has a throughput which is less than 50% of the optimum throughput (equal to the budget constraint, 1000).

Next, the aspiration and reservation levels were chosen close to the values of the flows predicted by the property of the PF approach. Indeed, we got the optimal solution of the PF model, which has a throughput of 573.3. While the throughput of this solution is larger than in the perfectly fair solution, it is still not large when compared to the optimum.

Finally, the aspiration levels were set to 999 (close to the maximal flow), and the reservation levels were chosen for flows that had identical path costs in the following way: the flows with path cost equal to 1 had a reservation level of 15; flows with path cost equal to 2 had a reservation level of 2.0; flows with path cost equal to 3 had a reservation level of 1.0, and flows with path cost equal to 4 had a reservation level of 0.5. This approach resulted in a solution that had a throughput of 732.7, yet the smallest flow was larger than 1.0, and flows with equal path costs were treated fairly, like in the proportionally fair solution. The Lorenz curves of all the described solutions are shown in Fig. 5. Note that none of the solutions dominates any other.

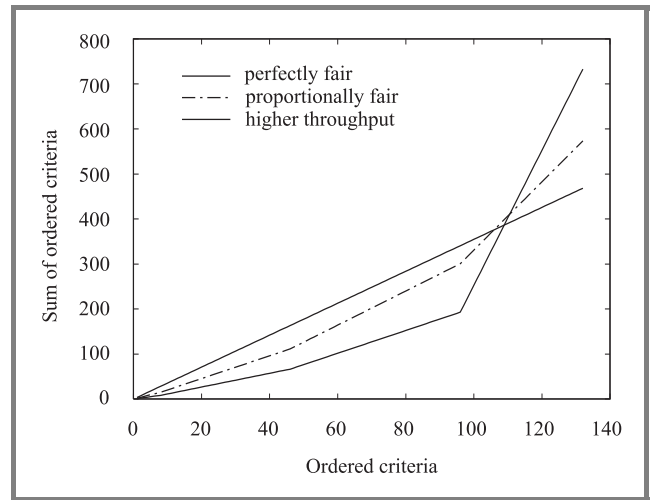


Fig. 5. Solutions obtained for the sample topology.

As was indicated in the introduction, the users of a network could be interested in fair treatment of flows between any pair of nodes, or in some other form of fairness. For example, the users could be interested in having fair amounts of available throughput from all other nodes to the user's node. This form of preferences could be expressed by the criteria:

$$n_v = \sum_{p_i=(u,v)} x_i \quad \forall v \in V. \quad (26)$$

In this case, the number of criteria is reduced. Also, note that in approaches which make the value of a flow dependent on the distance between the origin and destination (like proportional fairness), nodes which are distant from all other nodes will be treated unfairly. The three solutions described above will be shown in Fig. 6. The figure plots the Lorenz curves for the 12 criteria n_v for each of the three

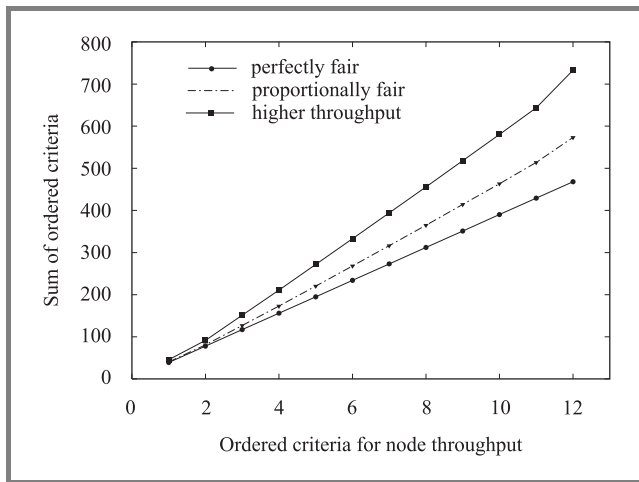


Fig. 6. The three solutions with respect to node criteria n_v .

solutions. It can be seen that the solution which increases throughput dominates the other two. This is a consequence of the design of the network topology, which is such that increasing network throughput improves the throughputs toward all the nodes. Another consequence of the topology is that all nodes have close values of criteria n_v , which is why the curves on the figure are almost straight; in more detail one could notice that the curves for proportional fairness and the OWA method have each 6 changes of slope. The perfectly fair solution predictably remains perfectly fair for the criteria n_v .

6. Concluding remarks

In various systems which serve many users, like in telecommunications systems, there is a need to respect the fairness rules, i.e. to allocate resources equitably among the competing services. Allocating the resources to optimize the worst performances may cause a large worsening of the overall (mean) performances. Therefore, several other fair allocation schemes are searched and analyzed.

The conditional mean is based on averaging restricted to the group of the worst performances defined by the tolerance level. Our earlier computational experiments with the conditional mean criterion applied to a traffic engineering model (a single ring bidirectional loading) were very promising [13]. The OWA aggregation further enriches modeling capacity offered by the conditional mean. In the case of LP models all equitable preferences may be modeled by selection of weights in the OWA aggregation.

Initial experiments with application of the OWA criterion (together with the reference point methodology) to the problem of network dimensioning with elastic traffic have confirmed the theoretical properties of the approach. We were able to generate easily allocations representing classical fairness models as well as to find new compromise solutions.

Maximization of the OWA aggregation, similar to the standard minimax approach, can be defined by optimization of a linear objective and a number of auxiliary linear inequalities. Many specific large-scale allocation models (especially discrete ones) may need some specialized exact or approximate algorithms. Thus, further research on computational aspects is necessary.

The problem of network dimensioning with elastic traffic could be extended with constraints on the individual flows. For example, network management could obtain traffic statistics that indicate the maximum throughputs which will be required between a pair of nodes. On the other hand, network statistics could also determine how much of the IP traffic requires guaranteed throughput (for example, from voice over IP applications). From this, minimal throughputs between a pair of nodes could be derived. In this work, we have analyzed in details the network design with elastic traffic without flow constraints. However, our approach allows to express such constraints in the objective function.

References

- [1] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs: Prentice-Hall, 1987.
- [2] J. F. Campbell, "Hub location and the p -hub median problem", *Oper. Res.*, vol. 44, pp. 923–935, 1996.
- [3] E. Gourdin, "Optimizing Internet networks", *ORMS Today*, vol. 28, pp. 46–49, 2001.
- [4] ILOG Inc., *Using the CPLEX Callable Library*. Incline Village: ILOG Inc., CPLEX Division, 1997.
- [5] F. Kelly, A. Mauloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability", *J. Oper. Res. Soc.*, vol. 49, pp. 206–217, 1997.
- [6] J. G. Klincewicz, "Hub location in backbone/tributary network design: a review", *Loc. Sci.*, vol. 6, pp. 307–335, 1998.
- [7] M. M. Kostreva and W. Ogryczak, "Linear optimization with multiple equitable criteria", *RAIRO Oper. Res.*, vol. 33, pp. 275–297, 1999.
- [8] H. Luss, "On equitable resource allocation problems: a lexicographic minimax approach", *Oper. Res.*, vol. 47, pp. 361–378, 1999.
- [9] E. Marchi and J. A. Oviedo, "Lexicographic optimality in the multiple objective linear programming: the nucleolar solution", *Eur. J. Oper. Res.*, vol. 57, pp. 355–359, 1992.
- [10] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and its Applications*. New York: Academic Press, 1979.
- [11] P. Nilsson and M. Pióro, "Solving dimensioning problems for proportionally fair networks carrying elastic traffic", Lund Institute of Technology at Lund University, 2002.
- [12] W. Ogryczak, "Comments on properties of the minimax solutions in goal programming", *Eur. J. Oper. Res.*, vol. 132, pp. 17–21, 2001.
- [13] W. Ogryczak and T. Śliwiński, "On equitable approaches to resource allocation problems: the conditional minimax solution", *J. Telecommun. Inform. Technol.*, no. 3, pp. 40–48, 2002.
- [14] W. Ogryczak and A. Tamir, "Minimizing the sum of the k largest functions in linear time", *Inform. Proc. Lett.*, vol. 85, pp. 117–122, 2003.
- [15] J. Rawls, *The Theory of Justice*. Cambridge: Harvard University Press, 1971.

- [16] A. Tomaszewski, "A polynomial algorithm for solving a general max-min fairness problem" in *Proc. 2nd Pol.-Germ. Teletraffic Symp. PGTS*, 2002, pp. 253-258.
- [17] A. P. Wierzbicki, "A mathematical basis for satisficing decision making", *Math. Modell.*, vol. 3, pp. 391-405, 1982.
- [18] A. P. Wierzbicki, M. Makowski, and J. Wessels, Eds., *Model Based Decision Support Methodology with Environmental Applications*. Dordrecht: Kluwer, 2000.
- [19] R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decision making", *IEEE Trans. Syst., Man Cybern.*, vol. 18, pp. 183-190, 1988.
- [20] R. R. Yager, "On the analytic representation of the Leximin ordering and its application to flexible constraint propagation", *Eur. J. Oper. Res.*, vol. 102, pp. 176-192, 1997.



Włodzimierz Ogryczak is a Professor of Optimization and Decision Support in the Institute of Control and Computation Engineering (ICCE) at Warsaw University of Technology, Poland. He received both his M.Sc. (1973) and Ph.D. (1983) in mathematics from Warsaw University, and D.Sc. (1997) in computer science

from Polish Academy of Sciences. His research interests are focused on theoretical research, computer solutions and interdisciplinary applications in the area of optimization and decision making with the main stress on: multiple criteria optimization and decision support, decision making under risk, location and distribution problems. He has published three books and numerous research articles in

international journals. Since 2000 with ICCE, arlier with Warsaw University (Institute of Informatics).
e-mail: wogrycza@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Adam Wierzbicki received his B.S. in mathematics and M.S. in informatics from the University of Warsaw, in 1997 and 1998. In June 2003, he received a Ph.D. degree from the Institute of Telecommunications of Warsaw University of Technology. His Ph.D. thesis titled "Content distribution and streaming media communication on the Internet" concerned the design of content delivery networks, particularly for improving quality and performance of streaming media communication. His current research interests include Peer-to-Peer computing, application layer multicast and network design. He is employed as a research assistant at the Polish-Japanese Institute of Information Technology in Warsaw, Poland.

e-mail: adamw@icm.edu.pl
Polish-Japanese Institute of Information Technology
Koszykowa st 86
02-008 Warsaw, Poland

Tomasz Śliwiński
e-mail: tswiwns@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland