Lecture-22: Resource allocation in networks

1 Resource allocation as utility maximization

The Internet is a shared resource, shared by many millions of users, who are connected by a huge network consisting of many, many routers and links. The capacity of the links must be split in some *fair* manner among the users. Economists solve such problems by associating a so-called utility function with each individual, and then finding an allocation that maximizes the net utility of the individuals. We now formally describe and model the resource allocation problem in the Internet.

Definition 1.1 (Links and sources). Consider a network consisting of a set of links $\mathcal{L}$ with capacity vector $c \in \mathbb{R}_+^{\mathcal{L}}$ such that $c_\ell \in \mathbb{R}_+$ is the capacity of link $\ell \in \mathcal{L}$. These links are accessed by a set of sources $\mathcal{S}$ with rate vector $x \in \mathbb{R}_+^{\mathcal{S}}$ such that $x_r \in \mathbb{R}_+$ is the rate of source $r \in \mathcal{S}$. We will use the terms source and user interchangeably.

Definition 1.2 (Routes). Each source is associated with a route, where a route is simply a collection of links. We denote the routing matrix by $R \in \{0,1\}^{\mathcal{L}\times\mathcal{S}}$ such that $R_{\ell,r}$ is the indicator that route $r \in \mathcal{S}$ includes link $\ell \in \mathcal{L}$. That is,

$$R_{\ell,r} \triangleq \begin{cases} 1, & \text{if route } r \text{ uses link } \ell, \\ 0, & \text{otherwise.} \end{cases}$$

Remark 1. Thus, we assume that the route used by a source to convey packets to their destination is fixed. Since the route is fixed for a source, we use the same index (typically r or s) to denote both a source and its route. We allow multiple sources to share exactly the same route. Thus, two routes can consist of exactly the same set of links.

Definition 1.3 (Utility). Each user derives a certain utility $U_r(x_r)$ when transmitting at rate x_r, where $U_r : \mathbb{R}_+ \to \mathbb{R}$ is an increasing, strictly concave, and continuously differentiable function.

Remark 2. The utility can be interpreted as the level of satisfaction that a user derives when its transmission rate is x_r. It is also usually the case that the rate at which the utility increases is larger at smaller rates than at larger rates. For example, a user's level of satisfaction will increase by a larger amount when the rate allocated to him or her increases from 0 Mbps to 1 Mbps than when the rate increases from 1 Mbps to 2 Mbps. This justifies the assumption that $U_r(x_r)$ is a strictly concave function.

Definition 1.4 (Capacity constraint). The total traffic on link $\ell \in \mathcal{L}$ from all sources $r \in \mathcal{S}$ is denoted by $y_\ell \triangleq \sum_{r\in\mathcal{S}} R_{\ell,r}x_r = (Rx)_\ell$. Since capacity of link $\ell \in \mathcal{L}$ is c_ℓ, the link capacity constraint is $Rx \leqslant c$.

Definition 1.5 (Feasible allocation). Given a routing matrix R, we call a source rate allocation $x \in \mathbb{R}_+^{\mathcal{S}}$ *feasible* if it is non-negative for all users and satisfies the link capacity constraints at each link $\ell \in \mathcal{L}$. The set of feasible allocations is denoted by

$$\mathcal{D} \triangleq \left\{ x \in \mathbb{R}_+^{\mathcal{S}} : Rx \leqslant c \right\}. \tag{1}$$

Lemma 1.6. *The set of feasible allocations $\mathcal{D}$ is a convex set.*

Proof. Let $x, y \in \mathcal{D}$ and $\lambda \in [0,1]$, then we observe that $Rx \leqslant c$ and $Ry \leqslant c$. Further, from the linearity of matrix multiplication, we get $R(\lambda x + \bar{\lambda} y) = \lambda Rx + \bar{\lambda} y \leqslant c$. Hence, $\lambda x + \bar{\lambda} y \in \mathcal{D}$, and the result holds since the choice of x, y, λ were arbitrary. □

1.1 Network utility maximization

Definition 1.7 (Network utility maximization). The goal of resource allocation is to solve the following optimization problem, called *network utility maximization (NUM)* that finds the feasible non-negative source rate allocation that satisfies the link capacity constraint and maximizes the sum utility of all users. That is,

$$x^* \triangleq \arg\max \left\{ \sum_{r \in \mathcal{S}} U_r(x_r) : x \in \mathcal{D} \right\}. \tag{2}$$

Remark 3. Recall that U_r is a strictly concave function and hence the sum network utility $\sum_{r \in \mathcal{S}} U_r$ is also strictly concave. Further, we showed that $\mathcal{D}$ is convex set and hence NUM is a convex optimization problem with a unique optimal allocation $x^* \in \mathcal{D}$.

Example 1.8 (Line network). Consider a network with L links numbered 1 through L each with unit capacity, and $L+1$ sources numbered 0 through L. All link capacities are assume to be equal to 1. Source 0's route includes all links, while source r uses only link r. The routing matrix for this example is $R_{r\ell} = \mathbb{1}_{\{r=\ell\}} \mathbb{1}_{\{r \in [L]\}} + \mathbb{1}_{\{r=0\}}$. The capacity constraint for each link $\ell \in \mathcal{L}$ is

$$\sum_{r=0}^{L} R_{\ell,r} x_r = x_0 + x_\ell \leqslant 1.$$

We take the logarithmic utility function, such that $U_r(x_r) \triangleq \ln x_r$, then we can write the NUM as

$$\begin{aligned} &\max_x \sum_{r=0}^{L} \ln x_r, \\ &\text{such that } x_0 + x_\ell \leqslant 1,\ \ell \in \{0, \ldots, L\}, \\ &\text{and } x \geqslant 0. \end{aligned}$$

Since $\lim_{x \downarrow 0} \ln x = -\infty$, the optimal solution will assign a strictly positive rate to each user, and so the last constraint can be ignored. Let p_ℓ be the Lagrange multiplier associated with the capacity constraint at link ℓ and let $p \in \mathbb{R}_+^L$ denote the vector of Lagrange multipliers. Then, the Lagrangian is given by

$$L(x, p) = \sum_{r=0}^{L} \ln x_r - \sum_{\ell=1}^{L} p_\ell (x_0 + x_\ell - 1).$$

Setting $\frac{\partial L}{\partial x_r} = 0$ for each $r \in \mathcal{S}$ gives $x_0^* = \frac{1}{\sum_{\ell=1}^{L} p_\ell^*}$ and $x_r^* = \frac{1}{p_r^*}$ for each source $r \in [L]$. Further, the KKT conditions require that $p_\ell^*(x_0^* + x_\ell^* - 1) = 0$ and $p_\ell^* \geqslant 0$ for all links $\ell \in [L]$. If $p_\ell = 0$ for some link $\ell \in [L]$, then $x_\ell = \infty$ and $x_0 + x_\ell > 1$. Hence, we observe that $p_\ell \neq 0$ for any link $\ell \in [L]$, and therefore $p_\ell^* = \frac{1}{x_\ell^*} = \frac{1}{1 - x_0^*}$ for all $\ell \in [L]$ and $x_0^* = \frac{1 - x_0^*}{L}$. It follows that optimal values of Lagrange multipliers and source rates are

$$p_\ell^* = \frac{L+1}{L} \text{ for all } \ell \in [L], \qquad x_0^* = \frac{1}{L+1}, \qquad x_r^* = \frac{L}{L+1}, r \in [L].$$

We note an important feature of the solution. The optimal rate of each source explicitly depends on the sum of the Lagrange multipliers on its route. Thus, if a simple algorithm exists to compute the Lagrange multipliers on each link and feed back the sum of the Lagrange multipliers on its route to each source, then the source rates can also be computed easily. This feature of the optimal solution will be exploited later to derive a distributed algorithm to solve the resource allocation problem.

1.2 Utility function and fairness

In our network utility maximization framework, we have associated a utility function with each user. The utility function can be interpreted in one of two different ways.

1. There is an inherent utility function associated with each user.
2. A utility function is assigned to each user by the network.

In the latter case, the choice of utility function determines the resource allocation to the users. Thus, the utility function can be viewed as imposing different notions of fair resource allocation. Of course, there is no notion of fair allocation that is universally accepted. Here, we will discuss some commonly used notions of fairness.

1.2.1 Proportional fairness

Definition 1.9 (Proportional fairness). An allocation $x^* \in \mathcal{D} \subseteq \mathbb{R}_+^{\mathcal{S}}$ is called *proportionally fair* if it satisfies the following property. For any other allocation $x \in \mathcal{D}$, we have

$$\sum_{r \in \mathcal{S}} \frac{x_r - x_r^*}{x_r^*} \leqslant 0. \tag{3}$$

Remark 4. The reason for this terminology is as follows. If one of the source rates is increased by a certain amount, the sum of the fractions (also called proportions) by which the different users' rates change is non-positive. A consequence of this observation is that, if the proportion by which one user's rate changes is positive, there will be at least one other user whose proportional change will be negative.

Lemma 1.10. *If $f : \mathcal{D} \to \mathbb{R}$ is a concave continuously differentiable function with maximizer x^*, then for all $x \in \mathcal{D}$*

$$\langle \nabla f(x^*), (x - x^*) \rangle \leqslant 0.$$

Proof. Let $x, x^* \in \mathcal{D}$. Since f is concave, it follows that for any $h \in [0,1]$, we have $f(x^* + h(x - x^*)) \geqslant (1-h)f(x^*) + hf(x)$. We can rewrite this as

$$\frac{f(x^* + h(x - x^*)) - f(x^*)}{h} \leqslant f(x) - f(x^*).$$

Taking limit $h \to 0$ for the continuously differentiable function f and using the fact that x^* is a maximizer for f, we get $\langle \nabla f(x^*), (x - x^*) \rangle \leqslant f(x) - f(x^*) \leqslant 0$. □

Lemma 1.11. *Proportionally fair resource allocation is achieved by associating a logarithmic utility function with each user, i.e., $U(x_r) = \ln x_r$ for all users $r \in \mathcal{S}$.*

Proof. Let x^* be the maximizer of $\sum_{r \in \mathcal{S}} U_r(x_r)$ in the constraint set $\mathcal{D}$. From Lemma 1.10, we observe that the set of optimal rates $x^* \in \mathcal{D}$ for logarithmic utility function $U_r(x_r) = \ln x_r$ satisfies (3) for any feasible allocation $x \in \mathcal{D}$. □

Definition 1.12. For each user $r \in \mathcal{S}$, if the utility functions are of the form $w_r \ln x_r$ for some weight $w_r > 0$, then the resulting allocation is called *weighted proportionally fair.*

1.2.2 Max-min fairness

Definition 1.13 (Max-min fairness). An allocation $x^* \in \mathcal{D}$ is called *max-min fair* if it satisfies the following property. If there is any other allocation $x \in \mathcal{D}$ such that a user s's rate increases, i.e., $x_s > x_s^*$, then there has to be another user u with the property $x_u < x_u^*$ and $x_u^* \leqslant x_s^*$.

Remark 5. In other words, if we attempt to increase the rate for one user, the rate for a less-fortunate user will suffer.

Lemma 1.14. *Let $x, x^* \in \mathcal{D}$ be arbitrary and optimal max-min fair allocations. Then $\min_r x_r^* \geqslant \min_r x_r$.*

Proof. To see why this is true, suppose that there exists an allocation $x \in \mathcal{D}$ such that $\min_r x_r^* < \min_r x_r$. This implies that, for any s such that $\min_r x_r^* = x_s^*$, we have $x_s^* < \min_r x_r \leqslant x_s$. However, this implies that if we switch the allocation from x^* to x, we have increased the allocation for s without affecting a less-fortunate user, since there is no less-fortunate user than s under x^*. Thus, the max-min fair resource allocation attempts first to satisfy the needs of the user who gets the least amount of resources from the network. □

1.2.3 Minimum potential delay fairness

Definition 1.15 (Minimum potential delay fairness). If each user $r \in \mathcal{S}$ is associated with the utility function $U_r(x_r) \triangleq -\frac{1}{x_r}$, then the optimal solution $x^* \in \mathcal{D}$ to NUM problem is *minimum potential delay fair*.

Remark 6. The goal of maximizing the sum of the user utilities is equivalent to minimizing the sum $\sum_{r\in\mathcal{S}} \frac{1}{x_r}$. The term $\frac{1}{x_r}$ can be interpreted as follows. If user r with rate r needs to transfer a file of unit size, then the delay associated with completing this file transfer is $\frac{1}{x_r}$. Hence, the name minimum potential delay fairness.

1.2.4 α fairness

Definition 1.16 (α-fair). Let $\alpha > 0$. Resource allocation is called *α-fair* if the utility function of each user $r \in \mathcal{S}$ is

$$U_r(x_r) \triangleq \frac{x_r^{1-\alpha}}{1-\alpha}. \tag{4}$$

Lemma 1.17. *α-fair utility function are concave, and consequently so is the α-fair network utility.*

Proof. We observe that α-fair utility functions are concave for $\alpha > 0$, since $U_r''(x_r) = -\alpha x_r^{-1-\alpha} < 0$. Therefore, the sum of concave functions $\sum_{r\in\mathcal{S}} U_r(x_r)$ is also concave. □

Remark 7. All of the different notions of fairness discussed above can be unified by considering α-fair utility functions for some $\alpha > 0$. Different values of α yield different ideas of fairness.

Theorem 1.18. *Consider α-fair network utility maximization problem. It reduces to*
(a) minimum potential delay fairness for $\alpha = 2$,
(b) proportional fairness for $\alpha = 1$, and
(c) max-min fairness for $\alpha = \infty$.

Proof. Let $\mathcal{D} \subseteq \mathbb{R}_+^{\mathcal{S}}$ be the set of feasible allocations.

(a) For $\alpha = 2$, we have $U_r(x_r) = -\frac{1}{x_r}$ and the result follows.

(b) We observe that the optimal allocation remains unchanged for a constant shift in all utility functions. In particular, we take $U_r(x_r) \triangleq \frac{x_r^{1-\alpha}-1}{1-\alpha}$. For $\alpha = 1$, we have $U_r(x_r) = \lim_{\alpha\to 1} \frac{x_r^{1-\alpha}-1}{1-\alpha} = \ln x_r$, and the result follows.

(c) Let $x^*(\alpha) \in \mathcal{D}$ be the optimal α-fair allocation. We assume that $\lim_{\alpha\to\infty} x_r^*(\alpha) = x_r^*$ exists for all sources $r \in \mathcal{S}$ and $x_1^* < x_2^* < \cdots < x_n^*$ for $n = |\mathcal{S}|$ sources. We define the minimum consecutive difference for allocation x^* as $\epsilon \triangleq \min_r(x_{r+1}^* - x_r^*)$. For this ϵ, we can choose α sufficiently large such that $|x_r^*(\alpha) - x_r^*| \leqslant \frac{\epsilon}{4}$. This implies that $x_1^*(\alpha) < x_2^*(\alpha) < \cdots < x_n^*(\alpha)$.
Applying Lemma 1.10 to the concave aggregate of α-utility functions, we obtain for any arbitrary feasible allocation $x \in \mathcal{D}$ and optimal α-fair allocation $x^*(\alpha) \in \mathcal{D}$

$$\sum_{r\in\mathcal{S}} U_r'(x_r^*(\alpha))(x_r - x_r^*(\alpha)) = \sum_{r\in\mathcal{S}} (x_r^*(\alpha))^{-\alpha}(x_r - x_r^*(\alpha)) \leqslant 0.$$

Considering an arbitrary flow $s \in \mathcal{S}$ and multiplying both sides by $(x_s^*(\alpha))^\alpha$, the above expression can be rewritten as

$$\sum_{r=1}^{s-1}(x_r - x_r^*(\alpha))\frac{(x_s^*(\alpha))^\alpha}{(x_r^*(\alpha))^\alpha} + (x_s - x_s^*(\alpha)) + \sum_{i=s+1}^{n}(x_i - x_i^*(\alpha))\frac{(x_s^*(\alpha))^\alpha}{(x_i^*(\alpha))^\alpha} \leqslant 0.$$

Since $|x_r^*(\alpha) - x_r^*| \leqslant \frac{\epsilon}{4}$, we further have

$$\sum_{r=1}^{s-1}(x_r - x_r^*(\alpha))\frac{(x_s^*(\alpha))^\alpha}{(x_r^*(\alpha))^\alpha} + (x_s - x_s^*(\alpha)) - \sum_{i=s+1}^{n}|x_i - x_i^*(\alpha)|\frac{(x_s^* + \frac{\epsilon}{4})^\alpha}{(x_i^* - \frac{\epsilon}{4})^\alpha} \leqslant 0.$$

Note that $x_i^* - \frac{\epsilon}{4} - (x_s^* + \frac{\epsilon}{4}) \geqslant \frac{\epsilon}{2}$ for any $i > s$. So, by increasing α, the third term in the above expression will become negligible. Thus, if $x_s > x_s^*(\alpha)$, the allocation for at least one user whose rate satisfies $x_r^*(\alpha) < x_s^*(\alpha)$ will decrease. The argument can be made rigorous and extended to the case $x_r^* = x_s^*$ for some r and s. Therefore, as $\alpha \to \infty$, the α-fair allocation approaches max-min fairness.

□