Utility, Fairness and Rate Allocation

Laila Daniel and Krishnan Narayanan

11th March 2013
Outline of the talk

- A rate allocation example
- Fairness criteria and their formulation as utilities
- Convex optimization - some background
- Network Utility Maximization (NUM) principle
A rate allocation example

(x0, x1, x2) = (0.5, 0.5, 1.5) is the Max–min fair allocation

- Links $L_1$ and $L_2$ in series with their respective capacities $C_1 = 1$ and $C_2 = 2$ shared by 3 flows $x_0$, $x_1$ and $x_2$ as follows. $x_0$ threads $L_1$ and $L_2$, $x_1$ is confined to $L_1$ and similarly $x_2$ is confined to $L_2$.
- A 'reasonable allocation' of rates for the flows is the following: $x_0 = 0.5$, $x_1 = 0.5$ and $x_2 = 1.5$
Max-min fair (MMF) allocation

- The idea behind the allocation is ’to divide the link capacity equally among the flows sharing the links and if there is any excess capacity share it equally between flows that require it’
- This allocation principle is ’max-min fair allocation’
- Definition: A vector of rates $\mathbf{x} = (x_s)_s \in S$ is max-min fair if it is feasible and no individual rate $x_s$ can be increased without decreasing any other rate equal or smaller
- Max-min fair allocation
  - maximizes the minimum rate
  - can be viewed as giving the maximum protection to the minimum of the allotted rates (absolute property)
  - All unsatisfied sources get the same rate which means that there is no incentive for a source to benefit from inflating its required rate
Max-min Fair (MMF) allocation

- A link $l$ is a bottleneck for a source $s$ if the link is saturated and the source has the largest rate on that link of all the flows sharing that link.
- In our example link $L_2$ is a bottleneck for flow $x_2$, and $L_1$ is a bottleneck for flows $x_0$ and $x_1$.

**Theorem**

A feasible allocation of rates is max-min fair iff every source has a bottleneck link.

**Theorem**

There is a unique max-min fair allocation which can be obtained by progressive filling (algorithm)

- Max-min fair allocation can be adapted using weights
- How to specify max-min fair allocation ?
- Is it the only reasonable allocation ?
- If not, what about other allocation schemes?
The need for utility function

- Scenario: Suppose the flow $x_1$ needs minimum rate 0.75 and flow $x_0$ has little 'worth' for any rate greater than 0.25
- (flow $x_1$ corresponds to real-time traffic and flow $x_0$ is a delay-insensitive (elastic) traffic)
- The notion of a utility function quantifies the worth of a given rate to a flow.
- Allocation of network resources based on the utility that sources specify for their rates
- Examples of utility functions are $\log x$ and $\frac{1}{x}$
- Utility function in general is a smooth concave function.
- Utility function can model fairness requirements.
An Example of rate allocation involving Utility

- For the same network scenario as above, but with logarithmic utility for the sources
- Now the resource allocation problem becomes

\[ \max (\log x_0 + \log x_1 + \log x_2) \text{ such that } \]

\[ x_0 + x_1 \leq 1 \]

\[ x_0 + x_2 \leq 2 \]

\[ x_0 \geq 0, x_1 \geq 0, x_2 \geq 0 \]

- Note that the constraints are linear inequalities and the objective function is a concave function.
Convex optimization problem

- If a function $f$ is convex if and only if its negation $-f$ is concave. E.g. The exponential function is convex and the log function is concave.

- Geometrically, the set of linear constraints (which in general can include both weak inequalities as well as equations defines a convex set.

- The optimization problem of minimizing a convex function (objective) over a convex set called a convex optimization problem.

- Maximizing a concave function is equivalent to minimizing a convex function obtained by its negation.

- So our rate allocation example is a convex optimization problem.
Using Lagrangian to solve the rate allocation example

- Let $\lambda_1$ and $\lambda_2$ be the Lagrangian multipliers correspondingly to the capacity constraints on the links $L_1$ and $L_2$ respectively.
- The Lagrangian for our problem is given by
  \[ L(x, \lambda) = \log x_0 + \log x_1 + \log x_2 - \lambda_1 (x_0 + x_1 - 1) - \lambda_2 (x_0 + x_2 - 2) \]
- Here $x$ is the data rate allocation vector and $\lambda$ is a vector of Lagrangian multipliers (non-negative real numbers).
- To solve this, set $\frac{\partial L}{\partial x_r} = 0$ for each $r \in 0, 1, 2$.
- This gives us
  \[ x_0 = \frac{1}{\lambda_1 + \lambda_2}, \quad x_1 = \frac{1}{\lambda_1}, \quad x_2 = \frac{1}{\lambda_2} \]
- Using $x_0 + x_1 = 1$ and $x_0 + x_2 = 2$ and solving we get
  \[ \lambda_1 = \sqrt{3}, \quad \lambda_2 = \frac{\sqrt{3}}{\sqrt{3} + 1} \]
- Substituting the $\lambda$ values we get the optimal allocation rates:
  \[ x_0 = 0.42, \quad x_1 = 0.58, \quad x_2 = 1.58 \]
Observations about Log utility

- Notice in comparison with MMF-allocation $x_0$ is smaller though in both cases both the links are saturated.
- Note that the non-zero Lagrangian multipliers $\lambda_1$ and $\lambda_2$ correspond to the case where the capacity constraints are active (i.e. equality) and the Lagrangian multipliers corresponding to $x_0$, $x_1$ and $x_2$ are all 0 as these constraints are slack, meaning $x_0 > 0$, $x_1 > 0$ and $x_2 > 0$ (CS principle)
- The log utility seems 'natural' in that it pulls up the smaller rates thereby giving them protection as in MMF allocation but not as strongly as MMF in an attempt to maximize the global system utility.
Fairness: Max-min vs Proportional Fairness

- In the case of a SINGLE bottleneck link, max-min allocation and proportional fairness allocation coincide.
- The sum of the rates received from all links is the same for all the users under proportional fairness criteria.
- Engenders the view that a resource does useful 'work' for a flow, so under proportional fairness allocation ALL FLOWS receive the SAME amount of work from the network.

\[(x_0, x_1, x_2) = (0.5, 0.5, 0.5)\] is the Max−min fair allocation.

\[(x_0, x_1, x_2) = (1/3, 2/3, 2/3)\] is the Proportional fair allocation.

\[L_1 \quad C_1 = 1 \quad L_2 \quad C_2 = 1\]
Log utility and proportional fairness

- A rate vector $\mathbf{x}^* = (x_s^*)_{s \in S}$ is proportionally fair if for any other rate vector $\mathbf{x} = (x_s)_{s \in S}$ the aggregate of the proportional changes is non-positive i.e,

$$\sum_{s \in S} \frac{x_s - x_s^*}{x_s^*} \leq 0$$

- The log utility function $U_s(x_s) = w_s \log x_s$ has this property ($w_s$ is the weight)

- The resource allocation scheme corresponding to $U_s(x_s) = w_s \log x_s$ is called weighted proportionately fair (WPF)

- If $w_s = 1 \ \forall s \in S$, then it is called proportionally fair

- Kelly has shown that AIMD principle is roughly corresponds to proportional fairness

- TCP fairness has been shown to be close to proportional fairness
Minimum potential delay fairness (MPD)

- Here the utility function $U(x) = -\frac{w}{x}$ where $w$ is the weight (usually a constant $\geq 0$) associated with the rate $x$.
- Here the motivation for the choice of the utility function is to minimize the time taken to complete a transfer, i.e., the higher the allocated rate, the smaller the transfer time.
- So the optimization problem can be regarded as minimizing the total time to complete all the transfers.
- Corresponding allocation is MPD allocation.
- For MPD allocation, for our rate allocation example we get the following equations in the same manner as before:

\[
\frac{1}{x_0^2} = \lambda_1 + \lambda_2, \quad \frac{1}{x_1^2} = \lambda_1 \quad \text{and} \quad \frac{1}{x_2^2} = \lambda_2
\]

- Solution: $x_0 = 0.49, x_1 = 0.51$ and $x_2 = 1.51$
- This solution was computed using MAPLE, a symbolic computing package.
- MPD fairness is a compromise between MMF and WPF.
Some Observations

- The goal of the congestion control algorithm is to provide the rate allocation of the SYSTEM OPTIMUM RATES (obtained by solving the network optimization problem) as the EQUILIBRIUM RATES when all the flows have enough data to send.
- The mechanism should be STABLE for this equilibrium rates and should converge to this equilibrium no matter what the initial state is.
- The network equilibrium is characterized by some fairness criterion.
- How quickly the algorithm can converge to this equilibrium without large oscillations of the network allocated rates is also an important question.
- These are some considerations in the design of a congestion control algorithm.
Optimal bandwidth sharing

- Optimal bandwidth sharing is got by solving the following utility maximization problem:

\[
\max \sum_{s \in S} U(x_s)
\]

\[
\sum_{s \in S} x_s \leq C_l, \forall l \in L
\]

\[
x_s \geq 0, \forall s \in S
\]

- A unique maximizer called primal optimal solution exists as the objective function is strictly concave and the feasible region is a compact convex set.

- A rate vector \(x\) will be optimal iff KKT holds at this rate vector.

- Now use KKT to get the following relationships between the optimal rates \(x_s\) and the dual variables \(p_l, l \in L\).
Summary of the talk

- Utility functions can be used to specify different fairness properties
- Log utility function is associated with proportional fairness
- The optimization problem of congestion control can be decomposed into subproblems solved by each SOURCE and each LINK in the network
- Pricing interpretation of the Lagrangian multiplier (dual variable)
- The operating point of the network is given by the equilibrium between the users willingness to pay (in price per unit time) and the system allotted rates computed (packets/unit price) are in equilibrium
- The equilibrium point is described using the some fairness criterion