Charging and rate control for elastic traffic *

Frank Kelly
University of Cambridge

Abstract

This paper addresses the issues of charging, rate control and routing for a communication network carrying elastic traffic, such as an ATM network offering an available bit rate service. A model is described from which max–min fairness of rates emerges as a limiting special case; more generally, the charges users are prepared to pay influence their allocated rates. In the preferred version of the model, a user chooses the charge per unit time that the user will pay; thereafter the user's rate is determined by the network according to a proportional fairness criterion applied to the rate per unit charge. A system optimum is achieved when users' choices of charges and the network's choice of allocated rates are in equilibrium.

1 Introduction

This paper describes a model designed to shed light on the issues of charging, rate control and routing. Its main purpose is to support ongoing work on charging schemes for broadband multiservice networks, described in [3] and [6]. A subsidiary aim is to investigate the relationship between various fairness criteria and 'smart market' approaches to dynamic pricing [7],[9],[10].

The organization of the paper is as follows. Section 2 presents a system model of charging, routing and flow control, where the system comprises both users with utility functions and a network with capacity constraints. Standard results from the theory of convex optimization show that the optimization of the system may be decomposed into subsidiary optimization problems, one for each user and one for the network, by using price per unit flow as a Lagrange multiplier that mediates between the subsidiary problems.

*This is a (corrected) version of a paper that appeared in *European Transactions on Telecommunications*, volume 8 (1997) pages 33-37.

Low and Varaiya [7] and Murphy *et al.* [9] describe how such results may be used as the basis for distributed pricing algorithms, and MacKie-Mason and Varian [10] describe a 'smart market' based on a per-packet charge when the network is congested.

In Section 3 we use a simple example to explore how various fairness criteria are associated with particular choices of utility function. We note that max–min fairness [1] emerges as a limiting special case, and describe a proportional fairness criterion associated with the logarithmic utility function.

In the system decomposition of Section 2, price per unit flow is the mediating variable. This may cause a particular difficulty for elastic traffic. In an implementation of an ATM available bit rate service, for example, users would be subject to *two* sources of uncertainty about the service offered: both the allocated rate *and* the price charged per unit flow would be allowed to fluctuate at the network's discretion. In Section 4 we describe an alternative system decomposition where price *per unit share* is the mediating variable. Under this decomposition the user chooses the charge per unit time that it pays, and the network determines allocated rates by a proportional fairness criterion, but applied to the rate per unit charge, rather than just the rate. It is shown that a system optimum is achieved when users' choices of charges and the network's choice of allocated rates are in equilibrium.

2 The model

Consider a network with a set J of *resources*, and let C_j be the finite capacity of resource j, for $j \in J$. Let a *route* r be a non-empty subset of J, and write R for the set of possible routes. Set $A_{jr} = 1$ if $j \in r$, so that resource j lies on route r, and set $A_{jr} = 0$ otherwise. This defines a $0 - 1$ matrix $A = (A_{jr}, j \in J, r \in R)$. Suppose that several routes through the network may substitute for one another: formally, suppose that a *source-sink* s is a subset of R and write S for the set of possible source-sinks. Set $H_{sr} = 1$ if $r \in s$, so that route r serves the source-sink s, and set $H_{sr} = 0$ otherwise. This defines a $0 - 1$ matrix $H = (H_{sr}, s \in S, r \in R)$. For each $r \in R$ let $s(r)$ identify a value $s \in S$ such that $H_{sr} = 1$, and suppose this value is unique; we view $s(r)$ as the source-sink served by route r.

We associate a source-sink s with a user, and suppose that if a rate x_s is allocated to the source-sink s then this has utility $U_s(x_s)$ to the user. We assume that the utility $U_s(x_s)$ is an increasing, strictly concave and continuously differentiable function of x_s over the range $x_s \geq 0$ (following Shenker [12], we call traffic that leads to such a utility function *elastic* traffic). Assume further that utilities are additive, so that the aggregate utility of rates

$x = (x_s, s \in S)$ is $\sum_{s \in S} U_s(x_s)$.

A flow pattern $y = (y_r, r \in R)$ supports the rates $x = (x_s, s \in S)$ if $Hy = x$, so that the flows y_r over routes r serving the source-sink s sum to the rate x_s. A flow pattern $y = (y_r, r \in R)$ is feasible if $y \geq 0$ and $Ay \leq C$, where $C = (C_j, j \in J)$, so that the flows over routes through resource j sum to not more than the capacity C_j of resource j. Let $U = (U_s(\cdot), s \in S)$ and let $U'(x) = (U'_s(x_s), s \in S)$.

To find the system optimal rates and flows we need to consider the following optimization problem.

SYSTEM(U, H, A, C):

$$\begin{aligned} \text{maximize} \quad & \sum_{s \in S} U_s(x_s) && (1) \\ \text{subject to} \quad & Hy = x, Ay \leq C && (2) \\ \text{over} \quad & x, y \geq 0. && (3) \end{aligned}$$

The objective function (1) is differentiable and strictly concave and the feasible region (2),(3) is compact; hence a maximizing value of (x, y) exists and can be found by Lagrangian methods. There is a unique optimum for the rate vector x, since the objective function (1) is a strictly concave function of x, but there may be many corresponding values of the flow rate y satisfying the relations (2) and (3). Say that x *solves SYSTEM*(U, H, A, C) if there exists y such that (x, y) solves the optimization problem (1)–(3).

Consider the Lagrangian form

$$\begin{aligned} L(x, y, z; \lambda, \mu) &= \sum_{s \in S} U_s(x_s) - \lambda^T(x - Hy) + \mu^T(C - Ay - z), \\ &= \sum_{s \in S} \big(U_s(x_s) - \lambda_s x_s\big) + \sum_{r \in R} y_r \big(\lambda_{s(r)} - \sum_{j \in r} \mu_j\big) \\ &\qquad - \sum_{j \in J} \mu_j z_j + \sum_{j \in J} \mu_j C_j \end{aligned}$$

where $\lambda = (\lambda_s, s \in S), \mu = (\mu_j, j \in J)$ are vectors of Lagrange multipliers and $(z_j, j \in J)$ is a vector of slack variables. Then

$$\begin{aligned} \frac{\partial L}{\partial x_s} &= U'_s(x_s) - \lambda_s \\ \frac{\partial L}{\partial y_r} &= \lambda_{s(r)} - \sum_{j \in r} \mu_j \\ \frac{\partial L}{\partial z_j} &= -\mu_j. \end{aligned}$$

Hence, at a maximum of L over the orthant $x, y, z \geq 0$, the following conditions hold:

$$\begin{aligned} \lambda_s &= U_s'(x_s) \quad \text{if } x_s > 0 \\ &\geq U_s'(x_s) \quad \text{if } x_s = 0 \end{aligned} \tag{4}$$

$$\begin{aligned} \lambda_{s(r)} &= \sum_{j \in r} \mu_j \quad \text{if } y_r > 0 \\ &\leq \sum_{j \in r} \mu_j \quad \text{if } y_r = 0 \end{aligned} \tag{5}$$

$$\begin{aligned} \mu_j &= 0 \qquad \text{if } z_j > 0 \\ &\geq 0 \qquad \text{if } z_j = 0 \end{aligned} \tag{6}$$

From the general theory of constrained convex optimization ([8], chapter 5; [13], chapter 3) it follows that there exists a quadruple (λ, μ, x, y) which satisfies

$$\lambda \geq U'(x), \quad Hy = x, \quad \big(\lambda - U'(x)\big)^T x = 0 \tag{7}$$

$$\mu \geq 0, \quad Ax \leq C, \quad \mu^T(C - Ax) = 0 \tag{8}$$

$$\lambda^T H \leq \mu^T A, \quad y \geq 0, \quad (\mu^T A - \lambda^T H)y = 0 \tag{9}$$

and that, further, the vector x then solves $SYSTEM(U, H, A, C)$.

The Lagrange multipliers λ, μ have several simple interpretations. For example, if route r has positive flow on it, $y_r > 0$, then necessarily $\sum_{j \in r} \mu_j \leq \sum_{j \in r^*} \mu_j$ for any other route r^* which serves the same source-sink. We may view μ_j as the *implied cost* of unit flow through link j. Alternatively μ_j is the *shadow price* of additional capacity at link j.

If user s is charged a price λ_s per unit flow, and is allowed to freely vary the flow x_s, then the utility maximization problem for user s is as follows.

$USER_s(U_s; \lambda_s)$

$$\begin{aligned} \text{maximize} \quad & U_s(x_s) - \lambda_s x_s \\ \text{over} \quad & x_s \geq 0. \end{aligned} \tag{10}$$

If the network receives a revenue λ_s per unit flow from user s, and is allowed to freely vary the flows x, then the revenue optimization problem for the network is as follows.

NETWORK$(H, A, C; \lambda)$

$$\begin{aligned} \text{maximize} \quad & \sum_s \lambda_s x_s && (11) \\ \text{subject to} \quad & Hy = x, Ay \leq C && (12) \\ \text{over} \quad & x, y \geq 0. && (13) \end{aligned}$$

Say that x *solves NETWORK*$(H, A, C; \lambda)$ if there exist y such that (x, y) solves the optimization problem (11) – (13).

Theorem 1 *There exists a price vector $\lambda = (\lambda_s, s \in S)$ such that the vector $x = (x_s, s \in S)$, formed from the unique solution x_s to USER$_s(U_s; \lambda_s)$ for each $s \in S$, solves NETWORK$(H, A, C; \lambda)$. The vector x then also solves SYSTEM(U, H, A, C).*

Proof. First note that *USER*$_s(U_s; \lambda_s)$ has a unique solution x_s, by the strict concavity of U_s, and that x_s is determined by $U_s'(x_s) \leq \lambda_s, x_s \geq 0$ and $(\lambda_s - U_s'(x_s))x_s = 0$. Next observe that the Lagrangian form for the optimization problem (11)–(13) is

$$\begin{aligned} L(x, y, z; p, q) &= \sum_s \lambda_s x_s - p^T(x - Hy) + q^T(C - Ay - z) \\ &= \sum_s x_s(\lambda_s - p_s) + \sum_r y_r(p_{s(r)} - \sum_{j \in r} q_j) \\ &\quad - \sum_{j \in J} q_j z_j + \sum_{j \in J} q_j C_j. \end{aligned}$$

Hence any quadruple (λ, μ, x, y), which satifies conditions (7), (8) and (9), identifies Lagrange multipliers $p = \lambda, q = \mu$, which establish that (x, y) solves *NETWORK*$(H, A, C; \lambda)$, as well as *SYSTEM*(U, H, A, C).

Conversely, for any solution x to *NETWORK*$(H, A, C; \lambda)$ there exist Lagrange multipliers p, q, where if $x_s > 0$ then $p_s = \lambda_s$ and if $x_s = 0$ then $p_s \geq \lambda_s$. Thus if x_s solves *USER*$_s(U_s; \lambda_s)$ then it will also solve *USER*$_s(U_s; p_s)$, and so we may construct a quadruple satisfying conditions (7), (8) and (9) by replacing λ and μ by p and q respectively. This establishes that x solves *SYSTEM*(U, H, A, C), and hence the final part of the theorem.

3 An example: fairness criteria

Suppose that each source-sink s is served by a single route r, and abbreviate notation by writing $s = r$, rather than $s = \{r\}$; thus $H = I$, the identity matrix. Suppose also that $U_s(x_s) = m_s \log x_s$. (Formally, we define $U_s(\cdot)$ over the range $[0, \infty)$, with $U_s(0) = -\infty$ and $U'_s(0) = \infty$.) Then at the optimum x_s is necessarily positive, and conditions (7) and (9) become simply

$$\lambda_s = \frac{m_s}{x_s}, \qquad \lambda_s = \sum_{j \in s} \mu_j.$$

Thus the optimal rate x_s is

$$x_s = \frac{m_s}{\sum_{j \in s} \mu_j} \tag{14}$$

where $(x_s, s \in S), (\mu_j, j \in S)$ solve

$$\mu \geq 0, \quad Ax \leq C, \quad \mu^T(C - Ax) = 0 \tag{15}$$

and relation (14).

Next we investigate the relationship between the solution to relations (14), (15) and concepts of fairness. The most common fairness criterion is that of max–min fairness: a vector of rates $x = (x_s, s \in S)$ is *max–min fair* if it is feasible (that is $x \geq 0$ and $Ax \leq C$), and if for each $s \in S$, x_s cannot be increased (while maintaining feasibility) without decreasing x_{s^*} for some s^* for which $x_{s^*} \leq x_s$ [1]. (The compactness and convexity of the feasible region for x imply that such a vector exists, and is unique.) The max–min fairness criterion gives an absolute priority to the smaller flows, in the sense that if $x_{s^*} < x_s$ then no increase in x_s, no matter how large, can compensate for any decrease in x_{s^*}, no matter how small. An alternative fairness criterion, which favours smaller flows less emphatically, is proportional fairness, defined as follows. A vector of rates $x = (x_s, s \in S)$ is *proportionally fair* if it is feasible (that is $x \geq 0$ and $Ax \leq C$) and if for any other feasible vector x^*, the aggregate of proportional changes is zero or negative[1]:

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

Say that resource j is a *bottleneck* if the solution x to relations (14), (15) has $(Ax)_j = C_j$. If $m_s = 1$ for $s \in S$, and if each flow x_s passes through a single bottleneck, then the solution x is necessarily max–min fair. This conclusion does not, however, apply when flows pass through multiple

[1] The printed version of the paper erroneously had strict inequality in this relation. I'm grateful to Martin Biddiscombe, Nortel, Harlow, for the correction - see http://www.ucl.ac.uk/∼uceemdb/phd.html

bottlenecks. To investigate this situation further, consider a small feasible perturbation $(x_s, s \in S) \to (x_s + \delta x_s, s \in S)$. This increases the objective function (1) provided

$$\sum_{s \in S} U'_s(x_s).\, \delta x_s > 0,$$

which condition becomes, with $U_s(x_s) = \log x_s$, the condition

$$\sum_{s \in S} \frac{\delta x_s}{x_s} > 0.$$

From the convexity of the feasible region for x and the strict concavity of the logarithm function, it follows that, when $m_s = 1$ for $s \in S$, the solution x to relations (14), (15) is the unique vector of rates that is proportionally fair.

We note that the definition of proportional fairness directly extends to the case where each source-sink s may be served by multiple routes: the definition of feasibility simply becomes that there exists $y \geq 0$ such that $x = Hy$ and $Ay \leq C$. Once again the solution x to $SYSTEM(U, H, A, C)$ with $U_s(x_s) = \log x_s$, $s \in S$, is the unique vector of rates that is proportionally fair.

The logarithmic utility function is thus intimately associated with the concept of proportional fairness. Is there a utility function that plays a similar role for the concept of max–min fairness? To explore this question further, let us suppose that any feasible flow satisfies $x_s < 1$. (This assumption loses no generality, since we can clearly rescale capacity units so that $\sum_{j \in J} C_j < 1$.) Next, let $U_s(x_s) = U_{(\alpha)}(x_s)$ for $s \in S$, where

$$U_{(\alpha)}(x) = -(-\log x)^{\alpha} \qquad 0 < x < 1, \alpha \geq 1.$$

The case $\alpha = 1$ is just the logarithmic utility function associated with a proportionally fair allocation of rates. If $0 < x_{s^*} < x_s < 1$,

$$\frac{U'_{(\alpha)}(x_{s^*})}{U'_{(\alpha)}(x_s)} = \frac{x_s}{x_{s^*}} \left(\frac{\log x_{s^*}}{\log x_s} \right)^{\alpha - 1} \to \infty \qquad \text{as } \alpha \to \infty.$$

Thus the collection of utility functions $U = U_{(\alpha)}$ provides a priority to smaller flows which increases as α increases and becomes absolute as $\alpha \to \infty$. The max–min fair allocation of rates is the limit of the solution to $SYSTEM(U_{(\alpha)}, H, A, C)$ as $\alpha \to \infty$.

4 An alternative decomposition

The decomposition described by Theorem 1 uses a vector λ giving prices per unit flow. In this Section we describe an alternative decomposition expressed in terms of prices per unit share.

If user s is charged an amount m_s per unit time, and receives in return a flow x_s proportional to m_s, then the utility maximization problem for user s is as follows.

$USER_s[U_s; \lambda_s]$

$$\begin{aligned} \text{maximize} \quad & U_s\left(\frac{m_s}{\lambda_s}\right) - m_s && (16) \\ \text{over} \quad & m_s \geq 0. && (17) \end{aligned}$$

Let $m = (m_s, s \in S)$, and define the following optimization problem.

$NETWORK[H, A, C; m]$

$$\begin{aligned} \text{maximize} \quad & \sum_s m_s \log x_s && (18) \\ \text{subject to} \quad & Hy = x \quad Ay \leq C && (19) \\ \text{over} \quad & x, y \geq 0. && (20) \end{aligned}$$

Say that x *solves* $NETWORK[H, A, C; m]$ if there exists y such that (x, y) solve the optimization problem (18)–(20).

Theorem 2 *There exist vectors* $\lambda = (\lambda_s, s \in S)$, $m = (m_s, s \in S)$ *and* $x = (x_s, s \in S)$ *such that*

(i) m_s *solves* $USER_s[U_s; \lambda_s]$, *for* $s \in S$*;*

(ii) x *solves* $NETWORK[H, A, C; m]$*;*

(iii) $m_s = \lambda_s x_s$ *for* $s \in S$.

The vector x *then also solves* $SYSTEM(U, H, A, C)$.

Proof. The derivative of the objective function (16) is

$$\frac{\partial}{\partial m_s}\left\{U_s\left(\frac{m_s}{\lambda_s}\right) - m_s\right\} = \frac{1}{\lambda_s}U_s'\left(\frac{m_s}{\lambda_s}\right) - 1.$$

Thus the conditions

$$\begin{aligned} U_s'\left(\frac{m_s}{\lambda_s}\right) &= \lambda_s \quad \text{if} \quad m_s > 0 \\ &\leq \lambda_s \quad \text{if} \quad m_s = 0 \end{aligned} \tag{21}$$

identify a solution m_s to $USER_s[U_s; \lambda_s]$.

The Lagrangian for the optimization problem (18)–(20) is

$$\begin{aligned} L(x, y, z; p, q) &= \sum_s m_s \log x_s - p^T(x - Hy) + q^T(C - Ay - z) \\ &= \sum_s (m_s \log x_s - p_s x_s) + \sum_r y_r(p_{s(r)} - \sum_{j \in r} q_j) \\ &\qquad - \sum_{j \in J} q_j z_j + \sum_{j \in J} q_j C_j. \end{aligned}$$

Then

$$\begin{aligned} \frac{\partial L}{\partial x_s} &= \frac{m_s}{x_s} - p_s \\ \frac{\partial L}{\partial y_r} &= p_{s(r)} - \sum_{j \in r} q_j \\ \frac{\partial L}{\partial z_j} &= -q_j. \end{aligned}$$

Hence, at a maximum of L over the orthant $x, y, z \geq 0$, the following conditions hold:

$$\frac{m_s}{x_s} = p_s \tag{22}$$

$$\begin{aligned} p_{s(r)} &= \sum_{j \in r} q_j \quad \text{if} \quad y_r > 0 \\ &\leq \sum_{j \in r} q_j \quad \text{if} \quad y_r = 0 \end{aligned} \tag{23}$$

$$\begin{aligned} q_j &= 0 \qquad \text{if} \quad z_j > 0 \\ &\geq 0 \qquad \text{if} \quad z_j = 0. \end{aligned} \tag{24}$$

But the quadruple (λ, μ, x, y) which satisfies conditions (7), (8) and (9) identifies a solution to (22), (23) and (24), with $p = \lambda$, $q = \mu$, and $m_s = \lambda_s x_s$, $s \in S$. Moreover, this solution satisfies the feasibility constraints (12) and (13), and the relation (21). This establishes the existence of the claimed vectors λ, m and x.

Conversely, for any solution x to $NETWORK[H, A, C; m]$ there exist Lagrange multipliers p, q, where if $x_s > 0$ then $m_s = x_s p_s$, and if $x_s = 0$ then $m_s = 0$. Thus if $m_s = \lambda_s x_s$ and m_s solves $USER_s[U_s; \lambda_s]$ then m_s will also solve $USER_s[U_s; p_s]$ [2], and so we may construct a quadruple satisfying conditions (7), (8) and (9) by replacing λ and μ by p and q respectively. Hence conditions (i), (ii) and (iii) of the Theorem imply that x solves $SYSTEM(U, H, A, C)$.

Note that if $m_s = 1$ for $s \in S$ then the solution to $NETWORK[H, A, C; m]$ is the proportionally fair allocation of rates. If m_s, $s \in S$, are all integral then the solution to $NETWORK[H, A, C; m]$ may be constructed as follows. For each $s \in S$ replace the single user s by m_s identical sub-users, calculate the proportionally fair allocation over the resulting $\sum_{s \in S} m_s$ rates, and then provide to user s the aggregate rate allocated to its m_s associated sub-users. Then the rates *per unit charge* are proportionally fair.

5 Concluding remarks

We have shown that if each user is able to choose a charge per unit time that it is prepared to pay, and if the network determines allocated rates so that the rates per unit charge are proportionally fair, then a system optimum is achieved when users' choices of charges and the network's choice of allocated rates are in equilibrium. We have not discussed convergence to equilibrium and an interesting and challenging question concerns whether rate control algorithms such as those described in [2], [3], and [4] may be adapted to implement the proportional fairness criterion described in this paper.

A further challenging question concerns how the choice of parameter m_s might be implemented in an ATM network. One possibility would be to use the Minimum Cell Rate of ATM standards [5] to buy a share of spare capacity, as well as to provide a lower bound on the rate. In [6] some of the consequential influences on user behaviour are discussed.

[2] This deduction is incorrect: it may fail when $m_s = x_s = 0$. A counterexample to the Theorem is provided by the example $m_s = x_s = 0$, $\lambda_s > U_s'(0)$, for all s: then statements (i), (ii) and (iii) of the Theorem hold, but x does not solve $SYSTEM(U, H, A, C)$. One way to sidestep the difficulty is to assume that $U_s'(x_s) \to \infty$ as $x_s \downarrow 0$, a technical assumption that turns out to have some other advantages in the dynamic context. But within the current context a much more satisfactory resolution is appended as a revised Section 4. I'm grateful to Ramesh Johari for pointing out the difficulty and developing its resolution.

Acknowledgements

The partial support of the Commission of the European Communities ACTS project AC039, entitled Charging and Accounting Schemes in Multiservice ATM Networks (CA$hMAN), is acknowledged. For comments influencing this paper, the author is particularly grateful to Costas Courcoubetis, Jon Crowcroft, Nicky van Foreest, Richard Gibbens, Aman Maulloo, Marion Raffali, David Tan, Kees van der Wal and Richard Weber.

References

[1] D. Bertsekas and R. Gallager, *Data Networks*. Prentice-Hall, 1987.

[2] A. Charny, K.K. Ramakrishnan and A. Lauck, Time scale analysis and scalability issues for explicit rate allocation in ATM networks. *IEEE/ACM Transactions on Networking* **4**, 569–581, 1996.

[3] C. Courcoubetis, V.A. Siris and G.D. Stamoulis, Integration of pricing and flow control for Available Bit Rate services in ATM networks. *Proc. IEEE Globecom '96*, London, 1996.

[4] E.J. Hernandez-Valencia, L. Benmohamed, S. Chong and R. Nagarajan, Rate control algorithms for the ATM ABR service. *European Transactions on Telecommunications* **8**, 7–20, 1997.

[5] ITU Recommendation I371, Traffic control and congestion control in B–ISDN. Geneva, Switzerland, 1995.

[6] F.P. Kelly, Charging and accounting for bursty connections. In *Internet Economics*, ed. L.W. McKnight and J.P. Bailey (eds). MIT Press, 1996.

[7] S.H. Low and P.P. Varaiya, A new approach to service provisioning in ATM networks. *IEEE Transactions on Networking* **1**, 547–553, 1993.

[8] M. Minoux, *Mathematical Programming: Theory and Algorithms*. Wiley, Chichester, 1986.

[9] J. Murphy, L. Murphy and E.C. Posner, Distributed pricing for embedded ATM networks. In *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*, ed. J. Labetoulle and J.W. Roberts. Elsevier, Amsterdam, 1053–1063, 1994.

[10] J.K. MacKie-Mason and H.R. Varian, Pricing the Internet. In *Public Access to the Internet*, ed. B. Kahin and J. Keller. Prentice-Hall, Englewood Cliffs, New Jersey, 1994.

[11] J.K. MacKie-Mason and H.R. Varian, Pricing congestible network resources. *IEEE Journal Selected Areas Communication* **13**, 1141–1149, 1995.

[12] S. Shenker, Fundamental design issues for the future Internet. *IEEE Journal Selected Areas Communication* **13**, 1176–1188, 1995.

[13] P. Whittle, *Optimization Under Constraints*. Wiley, Chichester, 1971.

Charging and rate control for elastic traffic
Correction to published version

Ramesh Johari and Frank Kelly
University of Cambridge

This document contains a revised Section 4 of: Kelly, F.P. (1997). "Charging and rate control for elastic traffic." *European Transactions on Telecommunications* **8**: 33-37.

4 An alternative decomposition

The decomposition described by Theorem 1 uses a vector λ giving prices per unit flow. In this Section we describe an alternative decomposition expressed in terms of prices per unit share.

If user s is charged an amount m_s per unit time, and receives in return a flow x_s proportional to m_s, then the utility maximization problem for user s is as follows.

$USER_s[U_s; \lambda_s]$

$$\begin{aligned} \text{maximize} \quad & U_s\left(\frac{m_s}{\lambda_s}\right) - m_s && (16) \\ \text{over} \quad & m_s \geq 0. && (17) \end{aligned}$$

Let $m = (m_s, s \in S)$, $\mathcal{B}(m) = \{s \in S : m_s > 0\}$, and define the following optimization problem.

$NETWORK[H, A, C; m]$

$$\begin{aligned} \text{maximize} \quad & \sum_{s \in \mathcal{B}(m)} m_s \log x_s && (18) \\ \text{subject to} \quad & Hy = x, \;\; Ay \leq C && (19) \\ \text{over} \quad & x, y \geq 0. && (20) \end{aligned}$$

Note that if $m_s = 1$ for $s \in S$ then the solution to $NETWORK[H, A, C; m]$ is the proportionally fair allocation of rates. If m_s, $s \in S$, are all integral, then the solution to $NETWORK[H, A, C; m]$ may be constructed as follows. For each $s \in S$ replace the single user s by m_s identical sub-users, calculate the proportionally fair allocation over the resulting $\sum_{s \in S} m_s$ rates, and then provide to user s the aggregate rate allocated to its m_s associated sub-users. Then the rates *per unit charge* are proportionally fair.

Say that x *solves* $NETWORK[H, A, C; m]$ if there exists y such that (x, y) solve the optimization problem (18)-(20). The corresponding Lagrangian form is

$$\begin{aligned} L_{NET}(x, y, z; \lambda, \mu) &= \sum_{s \in \mathcal{B}(m)} m_s \log x_s - \lambda^T(x - Hy) + \mu^T(C - Ay - z) \\ &= \sum_{s \in \mathcal{B}(m)} m_s \log x_s - \sum_{s \in S} \lambda_s x_s + \sum_{r \in R} y_r \left(\lambda_{s(r)} - \sum_{j \in r} \mu_j \right) \\ &\quad - \sum_{j \in J} \mu_j z_j + \sum_{j \in J} \mu_j C_j . \end{aligned}$$

Applying the stationarity conditions, we conclude that at an optimum of L_{NET} over $x, y, z \geq 0$:

$$\begin{aligned} \frac{\partial L_{NET}}{\partial x_s} = \frac{m_s}{x_s} - \lambda_s &= 0, \quad \text{if } s \in \mathcal{B}(m) \\ = -\lambda_s &= 0, \quad \text{if } x_s > 0, s \notin \mathcal{B}(m) \\ &\leq 0, \quad \text{if } x_s = 0, s \notin \mathcal{B}(m) \\ \frac{\partial L_{NET}}{\partial y_r} = \lambda_{s(r)} - \sum_{j \in r} \mu_j &= 0, \quad \text{if } y_r > 0 \\ &\leq 0, \quad \text{if } y_r = 0 \\ \frac{\partial L_{NET}}{\partial z_j} = -\mu_j &= 0, \quad \text{if } z_j > 0 \\ &\leq 0, \quad \text{if } z_j = 0. \end{aligned}$$

We may express these conditions more compactly: (x, y) solve $NETWORK[H, A, C; m]$ if and only if there exist multipliers (λ, μ) such that:

$$Hy = x, \quad Ay \leq C, \quad x, y \geq 0 \tag{21}$$

$$\lambda^T H \leq \mu^T A, \quad \lambda, \mu \geq 0 \tag{22}$$

$$\mu^T(C - Ay) = 0, \quad (\mu^T A - \lambda^T H)y = 0, \quad m_s = \lambda_s x_s, \ s \in S. \tag{23}$$

The first row of conditions is *primal feasibility*; the second row is *dual feasibility*; and the third row comprises *complementary slackness*.

We may use these conditions to construct the dual of $NETWORK[H, A, C; m]$. Given vectors λ and μ, the global maximum of L_{NET} occurs when $x_s = m_s/\lambda_s$

for $s \in \mathcal{B}(m)$. After elision of terms independent of λ and μ, the dual optimization problem is as follows.

$DUAL[H, A, C; m]$

$$\begin{aligned} \text{maximize} \quad & \sum_{s \in \mathcal{B}(m)} m_s \log \lambda_s - \sum_{j \in J} \mu_j C_j & (24) \\ \text{subject to} \quad & \lambda^T H \leq \mu^T A & (25) \\ \text{over} \quad & \lambda, \mu \geq 0. & (26) \end{aligned}$$

Say that λ *solves* $DUAL[H, A, C; m]$ if there exists μ such that (λ, μ) solve the optimization problem (24)-(26).

Theorem 2 *There exist vectors $m = (m_s, s \in S)$, $\lambda = (\lambda_s, s \in S)$, and $x = (x_s, s \in S)$ such that*

1. *m_s solves $USER_s[U_s; \lambda_s]$, for $s \in S$;*
2. *λ solves $DUAL[H, A, C; m]$, and $\lambda_s > 0$ for $s \in S$;*
3. *x solves $NETWORK[H, A, C; m]$;*
4. *$m_s = \lambda_s x_s$ for $s \in S$.*

Further, given any such triple (m, λ, x), the vectors m and x are uniquely determined, and x solves $SYSTEM(U, H, A, C)$.

Proof. As $\lambda_s > 0$ for all $s \in S$, $USER_s[U_s; \lambda_s]$ is well defined for all $s \in S$. The derivative of the objective function (16) is

$$\frac{\partial}{\partial m_s} \left\{ U_s \left(\frac{m_s}{\lambda_s} \right) - m_s \right\} = \frac{1}{\lambda_s} U_s' \left(\frac{m_s}{\lambda_s} \right) - 1.$$

Thus the conditions

$$\begin{aligned} U_s' \left(\frac{m_s}{\lambda_s} \right) &= \lambda_s, \text{if } m_s > 0 \\ &\leq \lambda_s, \text{if } m_s = 0 \end{aligned}$$

identify a solution m_s to $USER_s[U_s; \lambda_s]$. We may write these conditions more compactly as

$$m_s \geq 0, \quad \lambda_s \geq U_s' \left(\frac{m_s}{\lambda_s} \right), \quad \left(\lambda_s - U_s' \left(\frac{m_s}{\lambda_s} \right) \right) m_s = 0. \qquad (27)$$

By Lagrangian duality, (λ, μ) solve $DUAL[H, A, C; m]$ if and only if there exists a pair (x, y) such that (x, y, λ, μ) satisfy (21)-(23). But the quadruple (x, y, λ, μ) which satisfies conditions (7)-(9) identifies a solution to (21)-(23), by defining $m_s = \lambda_s x_s$, $s \in S$. Thus, x solves $NETWORK[H, A, C; m]$, and λ solves $DUAL[H, A, C; m]$. Finally, the conditions (7) together with the definition of m_s imply the conditions (27) are satisfied. This establishes the existence of the claimed vectors, m, λ, and x.

Conversely, suppose we are given m, λ, and x satisfying conditions (i)-(iv) of the theorem. Then, again by Lagrangian duality, we know there exists a pair $(\hat{x}, \hat{y})$ such that $(\hat{x}, \hat{y}, \lambda, \mu)$ satisfy (21)-(23). We now claim that in fact, $x = \hat{x}$, by the following reasoning. Given $s \in S$, $m_s = \lambda_s x_s$ by (iv); and $m_s = \lambda_s \hat{x}_s$ by (23). But since $\lambda_s > 0$, it must be true that $\hat{x}_s = x_s$ for all $s \in S$. Thus, the quadruple $(x, \hat{y}, \lambda, \mu)$ satisfies (21)-(23). But, by (i), m_s and λ_s satisfy (27) for all $s \in S$, so $(x, \hat{y}, \lambda, \mu)$ satisfies (7)-(9). We conclude x solves $SYSTEM(U, H, A, C)$, and therefore x is uniquely determined. Since $\lambda_s > 0$ for all $s \in S$, m is uniquely determined as well. □

We have shown that if each user is able to choose a charge per unit time that it is prepared to pay, and if the network determines allocated rates so that the rates per unit charge are proportionally fair, then a system optimum is achieved when users' choices of charges and the network's choice of allocated rates and prices per unit share are in equilibrium. In economic terms, equilibrium is achieved when demand (m_s) equals supply, or price times quantity $(\lambda_s x_s)$; and further, in this case, aggregate utility is maximized.