Preliminary Draft

A Smart Market for Resource Reservation in a Multiple Quality of Service Information Network

by

Jeffrey K. MacKie-Mason

University of Michigan and NBER

July 1995
Revised: September 8, 1997

A Smart Market for Resource Reservation in a Multiple Quality of Service Information Network

Jeffrey K. MacKie-Mason

1. Allocating Scarce Resources in a Multiple Quality of Service Network

The technology is nearly available to offer remarkably powerful new communications services: multiple streams, from multiple users, composed of different applications that require different qualities of service (QoS), all travelling over a single interconnected physical infrastructure. Society will benefit from integrated applications (video conferencing with interactive demos and shared whiteboards; computer-integrated telephony, &c.), and from increased access to information resources: access by more people, more of the time, from more places. However, as long as the laws of thermodynamics hold, the resources on which these systems are built will not be free. Efficient use of advanced networks requires a rational mechanism for allocating the scarce resources to the rapidly growing number of users and service types. Allocation in a multiple quality of service network may be the single greatest barrier to communications “anytime, anywhere”. In this paper I present a fairly general model of the problem, and, after showing that a decentralized open market will fail, I propose a mechanism for solving the problem.

Historical information networks have been based on separate physical networks for each major class of service. Different wires or segments of the spectrum were used for telephony, cable and broadcast TV, telegraph, paging, data. We now appear to be in an era of dramatic change in the information transport model, viz., convergence. For example, one of the less well-known marvels of the Internet is its new engineering model: multiple heterogeneous applications are supported simultaneously on a single shared physical network. "Telephony" circuits also support fax and Internet sessions, and are being tested for video programming. Cable networks are adding Internet transport to their video programming service, and are testing telephony.

Heterogeneous, shared networks offer a cost-saving opportunity. Mixing more, and more diverse, service flows on a single network improves the opportunity for *statistical multiplexing*, which then improves capacity utilization. Further, fewer physical, redundant physical infrastructures need to be built. Perhaps most important is that shared transport networks open possibilities for interaction and sharing between different communications applications. For example, on a network that provided video conferencing, and shared applications software, remote colleagues could work in real time using multiple tools: voice, facial and hand expressions, a shared whiteboard, software demos, and so forth.

Internet bandwidth sharing is supported by two features: packet switching and layering. MacKie-Mason and Varian (1994). By breaking flows into discrete, small packets it is possible to more efficiently share limited bandwidth: since each packet contains its own identification information, channels can be shared by the packets from multiple flows or sessions. Layering supports sharing by multiple *different* applications. The transport and application layers are separate,

This work has been supported by NSF grant SBR-9320481, NSF/ARPA/NASA cooperative agreement IRI-94-11287, and DARPA grant XXX.

and any application can use transport as long as it follows the rules for wrapping packet envelopes around the data to be transported.

As remarkable as the Internet is, it lacks a third crucial feature for simultaneously supporting multiple, heterogeneous applications: multiple qualities of service. Certain services — especially those that are inelastic (Braden, Clark, and Shenker (1994)) — require some service guarantees. For example, real-time, synchronous conversation is very intolerant of delay. Inelastic applications are poorly suited for a packet-switched network using best efforts service. Other applications or users differently value packet loss, maximal packet delay, average delay, delay variance (*jitter*), and other quality factors. However, at least two recent protocol developments offer the possibility that shared multi-application networks may start to see multiple qualities of service: ATM (asynchronous transfer mode) and RSVP. ATM creates a *virtual circuit* for each flow, and assignment of circuits can be restricted (*admission control*) in order to guarantee certain qualities of service. RSVP is a new protocol being implemented on top of the Internet's TCP/IP service that permits end-to-end reservation of bandwidth, based on which some service guarantees can be offered.

Thus, we are moving toward more flexible network services, with increasing convergence: capacity shared across multiple sessions, from multiple users, with multiple guaranteed qualities of service. A critical need, however, is to design efficient and feasible mechanisms for allocating the various resources in an integrated services network (Shenker (1993), MacKie-Mason and Varian (1995)). The mere fact that it is technically possible to share a network across many uses with different service requirements does not mean that the network will be provide valuable service. Already on the Internet with only a single quality of service (best efforts) the problem of congestion is well known. When resources *are* limited — when not everyone on the planet cannot simultaneously watch live video transmissions from Mars — how should they be allocated to maximize the value of the network to its user community overall? For this paper I pursue the objective of efficiency: ensuring that fixed resources (bandwidth, delay bounds, packet loss guarnatees, etc.) are allocated to the highest value uses.[1]

A good allocation mechanism is absolutely essential in a multiple quality of service network: with so many diverse competing demands for network resources, network performance could be quite poor if resources are not directed to their best uses. Suppose, for example, that users could obtain guaranteed low-delay service on a first-come, first-served basis (the current Internet allocation mechanism): nothing would stop all of the email users from requesting guaranteed service (even though email easily tolerates moderate delays), leaving none for users who need, and highly value guaranteed low-delay service for video conferencing. (Imagine such an allocation mechanism for first-class seats on airlines!) The social value realized from modern integrated services communications networks will be greatly constrained until efficient, feasible allocation mechanisms are developed.

In this paper, I characterize a rather general solution to the problem of efficient resource allocation in a multiple QoS network. Unfortunately, as I show, standard distributed or market mechanisms are unlikely to do a very good job of approximating the efficient allocation. I then describe a relevation mechanism that does obtain efficient allocation.

The generality of the model follows from envision a network that can guarantee multiple QoS as doing so by scheduling resources in advance, after users request reservations for the resources.

[1] The mechanism proposed at the end of this paper could be applied using any well-defined social objective function, including one that values some measure of fairness rather than just efficiency.

This architecture allows us to design an allocation mechanism that is responsive to the state of demand. State-responsive allocation allows the “rationing rule” during congested periods to assign the available resources to the highest value users. I am silent on how long in advance one must reserve resources, and on how long the reservation interval lasts. By suitably shrinking both the advance notice period and the length of the reservation interval, it is possible to approach arbitrarily close to continuous real-time pricing. How close an actual system can approach to real-time depends on the costs of implementing the mechanism, and the necessary propagation lags required to communicate information between the users and the network pricing system.

I draw on results from several literatures in order to develop a reasonable and general multiple QoS network model, and for it an efficient allocation mechanism. In particular, the solution of the network allocation problem leans on multicommodity flow results from the transportation engineering literature. Congestion pricing has been studied in the network economics literature; see, e.g., MacKie-Mason and Varian (1995) for a recent treatment relevant for information networks. The *smart market* proposed emerged from the mechanism design literature. I also borrow the theory of effective bandwidth from computer science. The effective bandwidth results are not necessary for the general results, but they greatly aid the exposition, and may be essential for *feasible* implementation of a smart market with reasonably frequent reservation intervals.

2. The Basic Problem

I consider a network with multiple nodes and links. Each link has a fixed capacity measured in bits per second. The network transport technology permits some form of directed path bandwidth reservation in advance, sufficient to guarantee various qualities of service (e.g., ATM, or RSVP over TCP/IP).

Users located at each node wish to send data to other nodes. For expository ease I aggregate all users at a particular node into a single, representative user. The user at node i would like to send traffic to each of the other nodes, j. There are different types of data traffic, which require different network service characteristics. For simplicity, I limit the model to two traffic characteristics: mean delay and minimum bandwidth. For example, users may value a video broadcast only if they can be guaranteed a bandwidth of 1.5 Mbps and a maximum delay of 100 ms. However, all of the results generalize straightforwardly to a finite-dimensional vector of service quality attributes. The number of dimensions that can be managed will largely depend on computational and other transactions costs.

An individual’s utility from using the network depends on the amount of traffic admitted from the user, of each service type, for each of the possible destinations. I simplify by assuming that the user’s benefit received from traffic between nodes i and j is independent of traffic between i and any other node h not equal to j. (That is, the value of my email to mom doesn’t depend on whether my email to dad got through — not a perfect assumption.) I make the usual assumption that the marginal utility of traffic between any two nodes is decreasing. There is “free disposal” of traffic received at a node j, so that absent rationing or usage fees each user would send an infinite amount of traffic (or at least far more than the available capacity could manage).

The network planner’s objective is to maximize the sum of user utilities by specifying an admission rule and specifying a network routing plan. By the assumption of free disposal there is an excess demand for capacity, thus the need for an admission rule that specifies how much traffic each user may deliver to each destination. The network has a mesh topology and uses a protocol that makes it possible to share the links among traffic from various users, of various QoS, going to various destinations, and to route traffic along multiple paths. Therefore, since the values that users place on their traffic differ by destination and quantity delivered, the network manager needs to solve a routing problem in order to deliver the combination of traffic that maximizes total user benefits.

The network planning problem is complicated due to the quality of service requirements imposed by users. If all traffic had the same delay requirements and differed only in the bandwidth required, then a routing plan would be constrained by the requirement that $\sum_h b_{hl} < B_l$, where b_{hl} is the bandwidth required by source h on link l, and B_l is the capacity on link l. However, when sources differ in their service requirements, the aggregation of traffic is typically not linear across the different types of flows. For example, a link that can accomodate four 5 Mbps flows with mean delay requirement μ_1 may not be able to accomdate two 5 Mbps flows of type μ_1 and one 10Mbps flow of type μ_2. Therefore, in general the network planner needs to optimize the admission and routing of flows across a mesh network with complex nonlinear interactions among the different types of flows that share links.

Recently remarkably powerful results on “effective bandwidth” have emerged that may greatly simplify the network allocation problem. These results indicate that for a broad range of source types an aggregation may be performed that permits a multiple-dimension service quality guarantee to be made based solely on a one-dimensional bandwidth reservation without efficiency loss.[2] If an effective bandwidth formulation is possible for a given traffic source and network technology, the constraint on link utilization can again be expressed as linear and one-dimensional: $\sum_{hk} b^k_{hl}(\phi) \leq B_l$, where b^k_{hl} is the effective bandwidth required by user h for traffic of type k, which is a function of the quality of service parameters for traffic type k, denoted by ϕ.

As an example, consider the result from Kelly (1991) (see also de Veciana and Walrand (1993). He models a system with n_j independent sources of type j. The distribution of bursts from each source is Poisson with mean arrival time λ_j; the length of each burst has mean μ_j and variance σ^2_j. The network offers first-come, first-served delivery. Then the mean delay is given by

$$\mathrm{ED} = \frac{\sum_\mathrm{j} \mathrm{n_j}\lambda_\mathrm{j}(\mu^2_\mathrm{j} + \sigma^2_\mathrm{j})}{2(1 - \sum_\mathrm{j} \mathrm{n_j}\lambda_\mathrm{j}\mu_\mathrm{j})}.$$

If the service constraint is that $\mathrm{ED} < \mathrm{d}$, then

$$\sum_j n_j b_j(d) \leq 1$$

where $b_j(d) = \lambda_j[\mu_j + \frac{1}{2d}(\mu^2_j + \sigma^2_j)]$.

For expository I shall assume that the traffic sources in the network are susceptible to an effective bandwidth characterization. This allows us to aggregate the traffic of each type on a given link into the effective bandwidth required for that traffic, and then sum across traffic types to find the total

[2] For the original work, see, e.g., de Veciana, Olivier, and Walrand (1993), de Veciana and Walrand (1993), Kelly (1991), Elwalid and Mitra (1992), Chang (1992), Gibbens and Hunt (1991), (Kesidis92). Of course, subject to absolute network performance constraints, it is always possible to guarantee a service quality by assigning enough bandwidth. The effective bandwidth results show how the efficient lower bound on necessary bandwidth can be found in some cases.

required bandwidth and constrain this requirement to be no more than the bandwidth available. This reduces the complexity of an already messy problem. This convenience may extend beyond exposition: feasible implementations of a smart market may also rely on dimension-reduction to limit the complexity of the bidding space and the magnitude of required computation for each allocation. In that case, effective bandwidth theory is an efficient, theoretically correct approach to such simplification, in contrast to arbitrary elimination of dimensions (e.g., bidding only on raw bandwidth and ignoring user preferences over mean delay).

3. The Model

I now specify the planning problem. The network is given by a capacitated graph, $G(N, L, B)$, specifying a set of nodes, $i \in N$, connected by links, $(i, j) \in L$, with each link (i, j) constrained to offer a maximum bandwidth of B_l. P is the set of all origin–destination pairs, with element $p = (i, j) \in P$. In general, not all O/D pairs p will have a direct link; that is, there exist pairs $(i, j) \in P$ such that $(i, j) \notin L$. K is the set of different traffic types (characterized by different service requirements), with element k. The unit cost of sending a flow along link (i, j) is given by $c_{ij,pk}$.

There is a single representative user located at each node i. Thus every unique origin/destination pair has a single source of traffic of each type k; the traffic actually delivered is denoted d_{pk}. Let P_i denote the set of origin/destination pairs that originate at node i. A user located at node i derives utility from her traffic d_{pk}, $p \in P_i$ equal to $u_{pk}(d_{pk})$; utility from traffic of different types and to different destinations is additively separable, so the user's total utility is given by $\sum_{p \in P_i, k \in K} u_{pk}(d_{pk})$. I assume that the $u_{pk}(\cdot)$ are concave.

The network described has multiple commodities, indexed in two dimensions: p and k. A flow of type k is different from a flow of type k': users value them differently and they may impose different costs on the network. Likewise, a flow between O/D pair p is different from a flow between pair p'. Thus, a unique commodity is d_{pk}, and I will sometimes refer to a flow of type (p, k).[3]

Suppose (part of) the flow of type k between node pair p transits a link from node i to j. I denote this flow on line (i, j) by $f_{ij,kp}$. Assume that there is an effective bandwidth function that aggregates all traffic of type k along link (i, j) into a required bandwidth to support that traffic and its service requirements,

$$b_{ijk} = g\left(\sum_p f_{ij,kp}\right).$$

With a full-duplex network, there is independent and equal capacity in both directions, so effective bandwidth calculations will be required for each link in each direction, that is for all links (i,j), $i \neq j$.

[3] This type of multicommodity network differs from, say, an electricity transmission network. In an electric grid, there is only one commodity type. Electrons are electrons, so the set K has only one element. Likewise, electric grid O/D pairs are generally irrelevant: buyers and sellers want to deliver or receive electrons at a given node, but don't care where those particular electrons came from, or where they are going. Indeed, due to Kirchoff's physical laws it is generally impossible to know or guarantee the specific path followed, or the destination reached by any given electrons placed on the grid at a particular node. See MacKie-Mason (1994) for application of a smart market mechanism to an electric transmission network problem.

Suppose the network planner can directly observe the users' utility functions. Then the planning objective is

$$\max_{d_{pk}, f_{ij,kp}} \sum_{pk} u_{pk}(d_{pk}) - c_{ij,pk} f_{ij,pk}. \tag{1}$$

The admitted flows must be routed through the network. Routing requires that each flow start at its origin, end at its destination, and that it leave every intermediate node that it enters. Therefore, the assignment of flows to links must satisfy the multi-commodity flow constraints (Ahuja, Magnanti, and Orlin (1993)),

$$\sum_{j \neq i} f_{ij,kp} - \sum_{j \neq i} f_{ji,kp} = D(d_{pk}, i) = \begin{cases} d_{pk} & \text{if } i \text{ is the origin for } p \\ -d_{pk} & \text{if } i \text{ is the destination for } p \\ 0 & \text{otherwise} \end{cases} \quad \forall i \in N. \tag{2}$$

This set of constraints can be compactly represented in matrix notation as $\mathcal{N} f_{pk} = D(d_{pk}, i)$, where $\mathcal{N}$ is the "node-link incidence matrix." Each column ij corresponds to the variable d_p, with a $+1$ in the ith row and a -1 in the jth row; the rest of the entries are zero.

Feasible routing is constrained by the capacity of each link. To enforce this constraint, flows of a given type on a given link are aggregated, and the resulting effective bandwidth is determined:

$$b_{ijk} = g\left(\sum_p f_{ij,kp}\right) \qquad \forall i \neq j. \tag{3}$$

Then the sum of the effective bandwidths on a link is constrained to be less than the available capacity:

$$\sum_k b_{ijk} \leq B_{ij}. \tag{4}$$

Finally, all of the flows must be nonnegative, so

$$f_{ij,kp} \geq 0. \tag{5}$$

4. Solving the Network Planning Problem

The problem in (1), subject to the constraints in (2), (4), and (5), has a special structure that makes it easier to interpret the optimality conditions. To see this, consider a simple decomposition of the problem: for any vector of usages, find the minimal cost of routing; then choose the usages the maximize total benefits net of costs. I can use this decomposition into a pair of problems — routing and usage — because the choice vectors in the two problems are independent in the objective function. That is, the only link between usages and link flows (d_{pk} and $f_{ij,pk}$) occurs through the technological routing constraints, (2).

Formally, an optimal solution of (1) and its constraints must satisfy the subproblem

$$C(\hat{d}_{pk}) = \min_{f_{ij,kp}} \sum_{p \in P, k \in K, (i,j) \in L} c_{ij,pk} f_{ij,pk} \tag{6}$$

subject to

$$\begin{aligned}
\mathcal{N} f_{pk} &= D(\hat{d}_{pk}, i) \\
\sum_k g\left(\sum_p f_{ij,pk}\right) &\leq B_{ij} \\
f_{ij,pk} &\geq 0 \\
d_{pk} &= \hat{d}_{pk} \qquad \text{for all } p \in P, k \in K.
\end{aligned}$$

That is, the cost should be minimized whatever the vector of usages $\hat{d}_{pk}$. Then the optimal usages are chosen to

$$\max_{d_{pk}} \sum_{p \in P, k \in K} u_{pk}(d_{pk}) - C(d_{pk}).$$

The problem in (6) is a slight generalization of the standard multicommodity flow (MCF) problem.[4] Although I could write down the Kuhn-Tucker sufficient conditions for an optimum directly, it turns out to be easier to interpret the result if I use the sufficient complementary slackeness conditions from the programming dual to the MCF problem. These are:

$$w_{ij}\left(\sum_k g\left(\sum_p f^*_{ij,pk}\right) - c_{ij}\right) = 0 \qquad \forall (i,j) \in L \qquad (7)a$$

$$c^{\pi}_{ij,pk} \equiv c_{ij,pk} + w_{ij} - \pi_{i,pk} + \pi_{j,pk} \geq 0 \qquad \forall (i,j) \in L, p \in P, k \in K \qquad (7b)$$

$$c^{\pi}_{ij,pk} f^*_{ij,pk} = 0 \qquad \forall (i,j) \in L, k \in K, p \in P \qquad (7c)$$

where the w_{ij} are the Lagrangian multipliers (shadow values or dual variables) on the capacity constraints, and the $\pi_{i,pk}$ are the multipliers on the routing constraints.

I shall call $c^{\pi}_{ij,pk}$ the *excess cost* associated with traffic of type (p,k) travelling over link (i,j).[5] The excess cost is central to the economic interpretation of the network planning problem; I discuss it in detail below.

To solve the full problem, (1), I also need the complementary slackness conditions corresponding to the optimal usages, d^*_{pk}, which are easily obtained from the Kuhn-Tucker conditions:

$$\left(u'_{pk}(d_{pk}) + \sum_{i \in N} \pi_{i,pk} \frac{\partial D(d_{pk}, i)}{\partial d_{pk}}\right) d_{pk} = 0 \forall p \in P, k \in K. \qquad (8)$$

[4] The difference is that one of the constraints, (4), involves a nonlinear function of the control variables.

[5] The network programming literature usually refers to this variable as the "reduced cost", but for "excess cost" is more natural for the economic interpretation I give it.

5. Characterizing the Solution

The optimal solution $(d^*_{pk}, f^*_{ij,pk})$ has a rather useful interpretation. I shall first provide the interpretation, then return to show how it is supported by the optimality conditions. Neglect for a moment that flows are defined by a specific origin/destination pair, and think of them instead as a variety of abstract commodities indexed by type (p, k). Suppose that a network user could place a flow of type (p, k) on the network at node i, or could extract a flow at node j. We can interpret the π_{ipk} as the *spot prices* or marginal system costs for inserting traffic of type (p, k) at node i. Then the marginal system cost of getting traffic from node i to node j is the cost of inserting the flow at i less the cost of inserting it at j (*i.e.*, plus the value of extracting it at j). For any two nodes that actually lie on an optimal route from origin to destination the net cost between nodes i and j is $\pi_{ipk} - \pi_{jpk}$, where this relationship holds when i and j are themselves the origin and destination nodes.

How does this interpretation of the dual variables π_{ipk} help us? Only links, not nodes, have physical costs and constraints in this network model. Therefore, I could find the marginal system cost for a flow of type (p, k) by summing the marginal link costs (including both transport costs and the shadow or opportunity costs due to the capacity constraints) along all of the links that constitute an optimal route for p. However, that requires knowing the list of links in an optimal route, and the transport and capacity costs for each of those links. This information is available to the network planner, of course, but communicating the information to users would be a heavy burden (suppose a typical route involved 15 links?).

Instead, the system cost information can be abstracted to only two numbers for each flow: the nodal prices π_{ipk} and π_{jpk} for O/D pair (i, j). All of the transport and capacity costs for the links along an optimal route are embedded in the nodal prices, and need not be reported. Users could be informed of system costs without knowing the route their traffic follows.

I shall now show how the nodal spot price interpretation of the π_{ipk} emerges from the optimality conditions. To begin, consider a simplified mesh network in which there is only one type of traffic, no capacity constraints, and only one origin-destination pair (but possibly multiple feasible routes). Imagine a physical model of this network in which the links are lengths of string, with the length of each link equal to the cost of traversal, c_{ij}.

Optimality requires that for any usage load, d, the routing cost must be minimized. To find the minimum cost path I can use the string model to solve a minimum distance problem (since distance equals cost): I do that by holding the mesh at the origin and destination node and pulling tight. The links that are taut form the minimum cost path. Along this path, designate the minimum cost of reaching an intermediate node, j, by π_j. Optimality requires

$$\pi_j \leq \pi_i + c_{ij}$$

for all $(i, j) \in L$ since if this were false, one could get to j more cheaply than π_j by going first to i at cost π_i and then going from i to j at cost c_{ij}. Further, for any link (i, j) that is part of the optimal path, $\pi_j - \pi_i + c_{ij} = 0$. If I recognize that the shadow values are zero, $w_{ij} = 0$ when there are no capacity constraints, I have obtained the excess cost condition in (7b):

$$c^{\pi}_{ij} f^*_{ij} = (c_{ij} - \pi_i + \pi_j) f^*_{ij} = 0 \qquad \forall (i, j) \in L$$

That is, if there is a flow on link (i, j), the excess cost on that link must be zero. If the excess cost is positive, the flow must be zero. In the string model, a link with a positive excess cost will be slack (it has excess “length”).

Now I can explain the interpretation of the π_i as spot prices for inserting traffic at node i. On a link that is in use, $-\pi_i + c_{ij} = -\pi_j$. To “purchase” or obtain a packet at j, one can pay the spot price at j, ($-\pi_j$), or pay the spot price at node i and pay the transport costs from i to j, for a total of $-\pi_i + c_{ij} = -\pi_j$.

In fact, users don’t have a demand for packets at node j, but for inserting packets at an origin node, O, and receiving them at a destination node, D. However, getting a packet from O to D can be thought of as a sequence of transactions along the minimum cost path. Indeed, it is helpful to imagine a set of artificial brokers or arbitrage agents stationed at each node who transact in packet insertions and extractions. The user pays a broker π_O to take a packet at the originating node, O. The broker hires transport to carry the packet to intermediate node i_1 at a cost c_{Oi_1}. The broker then extracts the packet from the network, and pays the user π_{i_1} to take it off his hands at i_1. This sequence is repeated for each hop along the optimal route. Clearly, the total cost for this route is $\sum_{(i,j)\in L^*} c_{ij}$, where L^* is the set of links in the optimal path. To interpret this in terms of nodal spot prices, consider the sum of the excess costs along the optimal path:

$$\sum_{(i,j)\in L^*} c^{\pi}_{ij} = \sum_{(i,j)\in L^*} c_{ij} - \pi_O + \pi_D = 0$$

where the sum is zero because the excess cost of each link is zero on the optimal path, and all of the nodal prices for intermediate nodes drop out because a packet both enters and departs each of those nodes. Thus,

$$\pi_{OD} \equiv \pi_O - \pi_D = \sum_{(i,j)\in L^*} c_{ij},$$

and I can see that paying the cost of the optimal path is equal to paying the net nodal spot price π_{OD} which is found by paying π_O to insert traffic at the origin, and receiving π_D for extracting the traffic at the destination.

Suppose the network manager were going to charge each user for the marginal cost that user placed on the network. The manager could post at node i a list of $\pi_p \equiv \pi_{ip} - \pi_{jp}$ for all $j \neq i$. Then users at i could consult the list to find the cost of their traffic to each destination, without needing any knowledge about the routing of their traffic through the network or the costs of each network link.

Let us now return to the complete problem, in which traffic can originate at any node, and there may be more than one type of flow (each with different effective bandwidth requirements). In this general network problem, a flow of type (p, k) may compete with the single flow in our simple network for the use of some links. I can visualize again what happens with our string model. Suppose I have stretched the mesh taught to find the minimum cost route for the first traffic flow of type k serving pair p. Now add a second flow that serves a different origin and different destination, p'. Using a second pair of hands, grasp the new O/D pair and attempt to pull taut. If the minimum cost path for the new flow runs along an entirely disjoint set of links from the original flow, I will succeed in finding an optimal taut path without having any effect on the optimal route for the first flow.[6]

[6] For a simple example, think of an X-shaped network with four outer nodes and one central node, and links from each outer node to the central node. I can pull taught between nodes 1 and 2, and between nodes 3 and 4, without either pair having an effect on the other.

There is a second possibility: the minimum cost path for the second flow will want to share some links with the first flow. This is possible in our simple string model: the two taut paths will have some overlapping links. However, the string model does not have any capacity constraints. Suppose that the combined flows are greater than the capacity along some of the shared links. I can represent the demand for bandwidth by the force with which I pull between the O/D pairs. If I now allow the strings to have some elasticity, then I can say that if a taut path is found without stretching any links I have not exceeded the capacity, but if the combined flows exceed capacity along some links, those links will be stretched by the excess pull (demand for capacity). The result is that those links will become longer: since length measures link cost in this model, I have effectively found that the cost of sending flows along the congested links has increased.

At this point I may be stretching the string metaphor too far. What in fact happens when link capacity is binding is that some of the traffic either must not be delivered, or it must be routed along a more costly route. But the idea that I have increased the cost (stretched the length) along the congested link is quite correct. When all of the traffic cannot flow along a least cost link, the cost of using that link for a flow becomes the sum of the physical transport cost, plus the opportunity cost (congestion cost) incurred by having to either reduce some traffic or displace it to a more costly link. Indeed, if the excess traffic can be rerouted, the additional cost is precisely the difference in transport cost between the preferred path and the best available alternative. More generally, the opportunity cost for a given link will be either the incremental cost incurred by rerouting, or the marginal user benefit foregone by reducing delivered traffic. That is, it is the marginal increase in net system benefits (expression (1)) that would be obtained if capacity along the link were increased.

Once all of the demanded flows are optimally routed, each link has an associated congestion cost, $w_{ij} \geq 0$, which represents the next best use of that link. Therefore, when I calculate the excess cost of a link, optimality requires that the cost of getting traffic to node j must be less than or equal to the nodal cost at i plus both the transport and congestion costs from i to j:

$$c^{\pi}_{ij,pk} = c_{ij,pk} + w_{ij} - \pi_{ipk} + \pi_{jpk}.$$

The congestion cost on link (i,j) is independent of the type of traffic (p,k) because it represents the *next* best use of the link, regardless of the type of traffic that is currently on the link.

In the full network, the marginal system cost for traffic of type (p,k) is

$$\pi_{pk} \equiv \pi_{ipk} - \pi_{jpk} = \sum_{(i,j)\in L^*} \left(c_{ij,pk} + w_{ij}\right).$$

The nodal prices π_{ipk} are different for each type of traffic k because each traffic type puts a different load on scarce capacity, as shown by the $g(\cdot)$ functions in (4). Nodal prices differ for each O/D pair p because flows between different pairs follow different routes, and the nodal prices embody the cost of the links that are followed to get to that node.

6. Decentralizing the optimal reservation of network resources

In the previous section I provide the necessary and sufficient conditions for a solution, and an interpretation of the result for a network resource reservation problem. I described the values of the the dual variables as *spot prices*, and remarked that users could be assessed the cost of their incremental contribution to system cost by charging π_{pk} for traffic of type (p, k). I intended this description to suggest there might be a pricing scheme lurking that could obtain efficient network usage through decentralized decisionmaking, rather than through a centralized solution to the problem (1).

In fact, only an imperfect decentralized solution is possible. Suppose that user i maximizes a quasi-linear utility function

$$\max_{\{y_i, d_{pk}\}} \sum_{p \in P_i, k \in K} u_{pk}(d_{pk}) + y_i$$

where y_i is income spent on other goods, subject to a budget constraint

$$\sum_{p \in P_i, k \in K} \pi_{pk} d_{pk} + y_i = M_i \qquad \forall i \in N$$

where π_{pk} is the unit price charged for flows of type k between O/D pair p, and M is the user's total income. The necessary and sufficient (assuming $u_{pk}(\cdot)$ are concave) conditions for the user's optimal purchases are

$$\left(u'_{pk}(d_{pk}) - \lambda(\pi_{\hat{i}pk} - \pi_{\hat{j}pk})\right) d_{pk} = 0 \qquad \forall p \in P, k \in K \tag{9}$$

and

$$\lambda = 1$$

where $(\hat{i}, \hat{j})$ are the origin and destination nodes in pair p. It is immediately obvious that after substituting for λ, the conditions in (9) are identical in form to the necessary conditions (8) in the network planner's problem. Thus, if the prices in the decentralized problem (π_{ipk}) are chosen to be the same as the nodal prices that result from the planner's problem, the user's demands in the decentralized network will be identical to their usage vectors in the centrally planned network, and the same optimal usage and routing solution will obtain.

There is a serious problem with this decentralization result, however. I have shown that if the optimal prices π_{pk} are announced, users will choose to demand the efficient traffic levels. However, finding the optimal π_{pk} required solving the full central planning problem for the network Once the central planning problem is already solved (which, recall, requires that the network manager know the users' true benefit functions, $u_{pk}(\cdot)$), there is nothing gained by announcing prices and soliciting traffic, since the efficient traffic flows have already been found.

If user demands are reasonably predictable (perhaps by day and time of day), the decentralized solution might provide a reasonably efficient method to approximate the optimum. The network manager could solve the planning problem for the expected "normal" traffic demand, announce the resulting nodal prices, and let users choose their actual traffic demands in light of the posted (time-of-day) prices. The outcome would be approximately efficient. One problem is that the network manager would have to also announce a rationing scheme that would deny resources to some users when the decentralized set of demands for the network occasionally exceeded the available capacity.

7. An efficient, decentralized auction mechanism

Posting prices in advance and allowing users to choose their desired level of network usage at those prices can at best yield an approximately efficient allocation. Unfortunately, even that approximate solution has a serious drawback: for the central planner to calculate the efficient prices for "normal" demand requires that the planner know the users' benefit functions, $u_{pk}(\cdot)$.

In many industries, firms seem to do reasonably well at estimating the shape of customers' preferences for their products, and setting prices accordingly. It may be unrealistic to think that similar success will be possible for services in a multiple QoS network, at least for the next many years. The services are new and constantly changing. Users have so far had little or no experience in specifying their demands in a multiple QoS environment, so survey data are unlikely to be much help. And dynamic congestion pricing is rare, so there are almost no useful data on how consumers shift their usage over time to respond to it.

I propose a method for inducing users to truthfully reveal their preferences (as best they know them) so that an efficient allocation can be calculated. This mechanism is known as a "smart market." A smart market combines modern computing power with the theory of a revelation game to elicit truthfully reported utility functions and then solve for the optimal allocation. The allocation is implemented by charging the nodal spot prices derived above. Thus, the smart market has the same desirable efficiency properties as the solution to the omniscient network planner's problem. In particular, the congestion cost component of prices correctly signals where and how much additional capacity should be added.

The smart market is a "generalized Vickrey auction" (GVA) (Varian and MacKie-Mason (1994)). The Vickrey auction is a well-known scheme for assigning a good to the agent who places the highest value on it, when individual valuations are private information. The idea is to solicit bids and award the good to the highest bidder, but charge the second highest bid as the price. Bidding one's true valuation is a dominant strategy for each agent. The generalized Vickrey auction extends the idea to allocate multiple units of a good, multiple goods, and goods with externalities (so that agents care about how much others are consuming).

To understand the intuition for a generalized Vickrey auction, consider first the standard Vickrey auction. Suppose there is one unit of one good, and two agents. Efficiency requires that the agent with the higher valuation, v_i receive the good. If agents announce a bid b_i, then agent 1's expected payoff is

$$\Pr[b_1 > b_2][v_1 - b_2].$$

Suppose that agent 1 announces the truth, $b_1 = v_1$. Then agent 1 always gets the good when the payoff is positive ($v_1 - b_2 > 0$), and never when it is negative. Clearly this is a dominant strategy. Since both agents will tell the truth, the good always goes to the agent with the higher true valuation.

The intuition is straightforward. First, the probability that an agent wins the auction depends on her bid, but the payoff she receives does not. Second, since the agent's bid cannot affect the payoff, the payoff should be structured so that truthtelling always wins the auction when the payoff is positive, and always loses when the payoff is negative. Truthtelling then dominates, and the resulting allocation is Pareto efficient.

The payoff structure that induces truthtelling in the one good, one unit Vickrey auction is second-pricing. More generally, the payoff structure that induces truthtelling is to give each participant his or her utility plus all of the social surplus of the other agents measured using reported utility functions. If the agent reports truthfully, then the agent's payoff is precisely equal to total social welfare from the auction, which is the desired maximand. Since a truthtelling agent receives precisely what the auction is maximizing, there can be no better strategy than truthtelling.

Of course, giving each agent all of the social surplus clearly is very costly! Therefore, the auction design usually requires a payment by each agent. If that payment depends only on the reports made by *other* agents it will not change truthtelling incentives. This is the role that the "second-price" plays in the Vickrey auction: it takes away the value of the good to the highest value user other than the agent in question. Thus in the two agent example, the winning agent gets her surplus less the second agent's surplus; the losing agent gets the winning agent's surplus less the winning agent's surplus, or zero.

The generalized Vickrey auction (GVA) follows this intuition, but applies to problems in which agents can have preferences over more than one good, more than one unit of the goods, and over the quantities of the goods that are consumed by other agents (externalities). Assume that each agent consumes a vector x_a, and that the total matrix of consumptions by all agents is given by x. Assume each agent has concave preferences over all consumptions, $u_a(x)$. Then the basic auction is implemented as follows (Varian and MacKie-Mason (1994)):

1. Each agent reports a utility function $r_a(\cdot)$.
2. The planner computes

   $$x^* = \arg\max \sum_a r_a(x)$$

   subject to

   $$F(x) = 0$$

   and assigns action x_a^* to agent $a = 1, \ldots, A$. Then compute

   $$W_{-a}(x^*) = \sum_{b \neq a} r_b(x^*)$$

   which is the total valuation of all agents other than a according to their reported utility functions.
3. Agent a receives payoff

   $$u_a(x^*) + [W_{-a}(x^*) - G_a(r_{-a)}],$$

   where $G_a(r_{-a})$ is an arbitrary function of other agents' reported utilities.

Truthtelling is the dominant strategy for this auction. The "second-price" analogue for a payment $G_a(\cdot)$ by each agent is to charge each agent the total social surplus that would be possible if that agent did not participate in the auction at all. The result, then, is that the net payoff received by agent a is the net increment in total surplus that his participation creates. This payment would be

$$G_a(r_{-a}) = \max_x \sum_{b \neq a} r_b(0, x_{-a})$$

subject to

$$F(0, x_{-a}) = 0.$$

As a simple example, consider the original Vickrey problem: one unit of one good, winner-takes-all; assume there are only two bidders, a and b. In that case, the person with the highest reported utility gets $u_a + [0 - u_b]$. u_b is the surplus that the second bidder would get if a *did not participate*, because b would then get the good. This is the standard Vickrey rule: bidder a gets the good, pays bidder b's bid (which truthfully revealed b's valuation), and keeps the net, $u_a - u_b$. The loser, of course, gets $0 + [u_a - u_a] = 0$.

If the only externalities are negative, not positive, then the auction raises non-negative revenue. With negative externalities, $G_i > W_{-i}$, for all agents i, because the other agents are better off with a *not* participating.[7] That means that each agent pays a non-negative amount for her allocation. If there are costs of production, these can be simply recovered by adding them to G_a without changing the problem, since the production costs do not depend on the bid values.

I now apply the GVA to the network pricing problem. Recall that the objective function is

$$\max_{d_{pk}, f_{ij,kp}} \sum_{pk} u_{pk}(d_{pk}) - c_{ij,pk} f_{ij,pk}. \tag{1}$$

The GVA for this problem is:

1. Each user p reports a utility function, $\sum_k r_{pk}(d_{pk})$.
2. The network planner computes

   $$\{d^*_{pk}, f^*_{ij,pk}\} = \arg\max \sum_{pk} r_{pk}(d_{pk}) - c_{ij,pk} f_{ij,pk}$$

   for each p, subject to constraints (2), (4), and (5) given earlier. The planner then assigns the vector d^*_p to each agent $p \in P$, and computes

   $$W_{-p}(d^*) = \sum_{q \neq p, k \in K} r_{qk}(d^k_q) - c_{ij,qk} f_{ij,qk}.$$

3. User p receives a payoff

   $$\sum_k \left(u_{pk}(d^*_{pk}) - c_{ij,pk} f_{ij,pk}\right) + [W_{-p}(d^*) - G_p(r_{-p})]. \tag{10}$$

A convenient form for $G_p(\cdot)$ is

$$G_p(r_{-p}) = max \sum_{q \in P \neq p, k \in K} r_{qk}(d^k_q) - c_{ij,qk} f_{ij,qk},$$

where the maximization is subject to the same routing, capacity and nonnegative constraints as above, except that they are applied only to the users $q \neq p$.

[7] Though not usually phrased this way, the *second price* in the traditional Vickrey auction is actually the correct congestion price: when a gets the good, that causes congestion and crowds out b's consumption. The social cost of that congestion is the utility foregone by b.

With this definition of the payment $G_p(\cdot)$, user p's payoff in (10) has two nice interpretations. First, the expression in square brackets is the increase (decrease) in the welfare of all other users that is created by user p's participation. Thus, for participating in the network, p receives his own direct utility (net of transport costs) plus the net increase in the social welfare of others that he creates by his participation. If his traffic creates incremental congestion, and thus has a negative effect on the value of the network to others, he pays precisely the incremental reduction in network value to others that his congestion causes.

Alternatively, if I regroup the expression (10),

$$\left[\sum_k \big(u_{pk}(d^*_{pk}) - c_{ij,pk}f_{ij,pk}\big) + W_{-p}(d^*)\right] - G_p(r_{-p})],$$

it becomes clear that p is receiving the net increment in total social surplus that his participation creates. Since he gets all of the incremental surplus, he wants to truthfully reveal his preferences. If his participation creates a negative congestion externality, then the net increment in total surplus is less than his direct utility gain, and he pays a congestion tax.

The GVA for the network resource reservation problem has very nice properties. It elicits truthful revelation of user valuations, and thus efficiently allocates the scarce bandwidth to the most valued uses (subject to the constraints imposed by the flow routing problem). The solution of the optimization in step (1) yields the nodal spot prices that can be used to communicate the cost of different traffic types to users, and also to implement an approximate solution by applying the spot prices from one auction to other, similar demand periods. The solution also yields the marginal congestion costs for each link, w_{ij}, which send the correct signal to the network planner about efficient capacity investments.

One problem with the GVA is that it requires a substantial amount of computation. If the network planner simply knew the correct utility functions, as I assumed in section 6, the problem would not be very computationally intense. If the utility functions u_{pk} were linear, then the problem would be equivalent to a linear multicommodity flow problem, which has a special structure that can be exploited in solution algorithms. Good MCF algorithms are two to five times as fast as a general simplex linear programming code (Ahuja et al. (1993)). With nonlinear, but strictly concave utility functions, the problem becomes somewhat more costly, but a solution is guaranteed in finite time. The separability of the problem into an MCF conditional on a choice of optimal demands that I exploited in the interpretation of the optimality conditions suggests that solution would not be too costly.

However, when I turn to the GVA, with unknown user utility functions, a solution becomes considerably more costly. The GVA mechanism itself requires only a single solution of the network optimization, plus some ancillary linear computations. However, the GVA with a payment in the suggested form of $G_p(\cdot)$ requires that the network problem be resolved once for every p (with a minor reduction in dimensionality by the elimination of one user). Since some price $G_p(\cdot)$ would certainly be necessary for any practical GVA, it appears that the computation may be on the order of p network optimization problems. It is clear that the granularity of the allocation problem is a crucial decision variable: the more disaggregated are users, the larger the computational burden becomes in two dimensions: the solution time for the network optimization is likely to increase polynomially in p, and the number of such optimizations increases linearlly in p. The computational cost will also affect the desirable interval over which resources can be reserved (e.g., blocks of an hour, a minute, a second, etc.).

References

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Englewood Cliffs, NJ.

Braden, R., Clark, D., and Shenker, S. (1994). Integrated services in the Internet architecture: an overview. Tech. rep., ISI, MIT and Xerox PARC. Network Working Group, RFC 1633.

Chang, C. S. (1992). Stability, queue length and delay of deterministic and stochastic queueing networks. Tech. rep., XXX. Submitted to *IEEE AC*.

de Veciana, G., Olivier, C., and Walrand, J. (1993). Large deviations of birth death Markov fluids. *Probability in the Engineering and Informational Sciences*, *7*, 237–255.

de Veciana, G., and Walrand, J. (1993). Effective bandwidths: Call admission, traffic policing & filtering for ATM networks. Tech. rep., Electronics Research Laboratory, UC Berkeley. Memorandum No. UCB/ERL M39/47.

Elwalid, A. I., and Mitra, D. (1992). Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. Tech. rep., XXX. Submitted to *IEEE Networks*.

Gibbens, R. J., and Hunt, P. J. (1991). Effective bandwidths for the multi-type UAS channel. *Queuing Systems*, *9*, 17–28.

Kelly, F. P. (1991). Effective bandwidths at multi-class queues. *Queuing Systems*, *9*, 5–16.

MacKie-Mason, J. K. (1994). A spatial smart market for wholesale electricity exchanges. Tech. rep., University of Michigan, Department of Economics.

MacKie-Mason, J. K., and Varian, H. (1995). Pricing the Internet. In Kahin, B., and Keller, J. (Eds.), *Public Access to the Internet*. Prentice-Hall, Englewood Cliffs, New Jersey. Available from URL:

ftp: //gopher. econ. lsa.umich.edu/pub/Papers/Pricing_the_Internet.ps.Z.

MacKie-Mason, J. K., and Varian, H. R. (1994). Economic FAQs about the Internet. *Journal of Economic Perspectives*, *8*(3). Available from URL:

ftp: //gopher.econ.lsa.umich.edu/pub/Papers/FAQs.ps.Z.

MacKie-Mason, J. K., and Varian, H. R. (1995). Pricing congestible network resources. *IEEE Journal of Selected Areas in Communications*, *13*(7). Available at URL:

ftp: //gopher. econ. lsa.umich.edu/pub/Papers/pricing-congestible.ps.Z.

Shenker, S. (1993). Service models and pricing policies for an integrated services Internet. Tech. rep., Palo Alto Research Center, Xerox Corporation.

Varian, H. R., and MacKie-Mason, J. K. (1994). Generalized Vickrey auctions. Tech. rep., Dept. of Economics, Univ. of Michigan.