A Smart Market for Resource Reservation in a Multiple QoS Network

Jeffrey K. MacKie-Mason*
Dept. of Economics and School of Information
University of Michigan

March 1997

Allocating scarce network resources

- information networks offer heterogeneous mix of service qualities (bandwidth, mean delay, max delay, delay jitter, packet loss, &c.)
  - packet nets ill-suited for inelastic applications ("best efforts")
  - circuit nets ill-suited for widely heterogeneous service requirements, highly bursty traffic (no sharing)
- critical need for efficient and feasible mechanisms to allocate various scarce resources in a multiple QoS net

Network model: shared media network planning problem

- multiple nodes and fixed capacity links
- packet or cell switching technology
- protocol for directed path bandwidth reservation in advance (ATM, RSVP)
- single user at each node i would like to send traffic to users at other nodes j
- traffic characterized in two dimensions: mean delay and minimum bandwidth

Network management objective

Maximize sum of user utilities by specifying:

1. admission rule (which flows?)

2. routing plan (which resources do they get?)

The problem with multiple QoS

- Bandwidth constraint for only one service class, for source h on link l: $\sum_h b_{hl} < B_l$

- with multiple QoS the constraint is: $B(\vec{b}_{11}, \vec{b}_{2l}, \ldots, \vec{b}_{Hl}) < B_l$

  where $\vec{b}_{hl}$ is a $Kx1$ vector of traffic of types $1, \ldots, k$ from source h on link l, and $B(\cdot)$ is a nonlinear function of all traffic of all types from all sources on link l

- nonlinearity matters: (4×5Mbps, mean delay μ_1) $\neq$ (2 × 5 Mbps, delay μ_1) + (10 Mbps, delay μ_2)

Re-simplifying the problem to bandwidth sharing

- Near real-time computation aided by *effective bandwidth* formulation (e.g., Kelly 1991):

$$\sum_{hk} b_{hl}^{k}(b_{hl}, \phi^{k}) \leq B_{l}$$

  - b_{hl}^{k} is effective bandwidth required by source h for traffic of type k
  - ϕ^{k} are the service parameters for traffic of type k

The model: technology

- multicommodity flow in a capacitated graph, $G(N, L, B)$ (each flow is separate commodity, no mixing)

- origin/destination pairs $p = (i, j)$; service types $k \in K$; traffic delivered is d_{pk}

- cost of flow on link (i, j): $c_{ij,pk}$

The model: routing

- Each flow must be nonnegative ($f_{ij,kp} \geq 0$), start at its origin, end at its destination, and leave every intermediate node that it enters (MCF constraints):

$$\sum_{j \neq i} f_{ij,kp} - \sum_{j \neq i} f_{ji,kp} = \begin{cases} d_{pk} & \text{if } i \text{ is } p\text{'s origin} \\ -d_{pk} & \text{if } i \text{ is } p\text{'s dest} \\ 0 & \text{otherwise} \end{cases}$$

- Aggregate flows on each link must satisfy the effective bandwidth constraint:

$$\sum_k b_{ijk} \left(\sum_p f_{ij,kp} \right) \leq B_{ij} \qquad \forall i \neq j$$

The model: user and network valuations

- user i's valuation:

$$U_i = \sum_{p \in P_i, k \in K} u_{pk}(d_{pk})$$

- network planner's objective (if user utility functions observable):

$$\max_{d_{pk}, f_{ij,kp}} \sum_{p \in P, k \in K} u_{pk}(d_{pk}) - c_{ij,pk} f_{ij,pk}$$

Characterizing the solution

- Consider one of the complementary slackness conditions for the routing problem:

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

  where the w_{ij} are the shadow values on the capacity constraints, and the $\pi_{i,pk}$ are the shadow values on the routing constraints

- Think of traffic as abstract commodities that can be inserted at i and extracted at j (Temporarily ignore the flow nature of traffic: that we want to receive at D what we inserted at O)

- Then π_{ipk} is the "spot price" for inserting traffic of type (p, k) at node i

- Marginal system cost of getting traffic "from" i to j is cost of inserting at i less cost of inserting (value of extracting) at i: $\pi_{ipk} - \pi_{jpk}$

- True when (i, j) equal (O, D)

Usefulness of nodal spot prices

- Only links, not nodes, have physical costs and constraints in this system
  - Could report link costs to users (physical plus shadow cost of congestion)
  - But then users need to know optimal route, and price for each link
- Instead, only need to know two numbers: nodal prices π_{ipk} and π_{jpk} for O/D pair (i, j)
- Users need not know routing information: all transport and capacity costs *along an optimal route* are embedded

Where does nodal spot price interpretation come from?

- Define π_j to be the minimum cost of reaching an intermediate node along optimal path for flow (p, k); assume no capacity constraints ($w_{ij} = 0$)

- Then, if want to obtain (extract) packet at j, could either pay $-\pi_j$, or pay to obtain at i ($-\pi_i$) plus pay the cost of transporting from i to j (c_{ij})

- So, for a link (i, j) on the optimal path, must have $-\pi_i + c_{ij} = -\pi_j$

- This yields the complementary slackness condition on "excess cost":

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

Nodal prices for non-adjacent nodes

- Users don't want packets at arbitrary j: they want to insert at O and extract at D. Ignoring capacity constraints, total resource cost along optimal path is $\sum_{(i,j)\in L^*} c_{ij}$

- But consider the sum of excess costs on optimal path:

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

  - where sum is zero because excess cost of each link is zero on optimal path

  - and intermediate nodal prices drop out because a packet both enters and departs each transit node

Summary: nodal spot prices

- Thus, the resource cost of the optimal path is equal to the net nodal spot price $\pi_{OD} \equiv \pi_O - \pi_D$ from paying to insert traffic at origin and receiving to extract the traffic at the destination:

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

- Complete the problem by reintroducing capacity constraints
  - If constraint binding, then cost of sending traffic along a link is the resource cost, plus the cost of reducing some other traffic, or of paying for it on a more costly link

– I.e., the amount by which network value would in be increased if a bit more capacity were added to the constrained link — this is precisely the capacity shadow value, w_{ij}

Decentralizing the solution

- Solution characterized in terms of "spot prices" to hint that there might be a decentralized pricing scheme that could obtain efficient network usage

- Returning to network planning problem, there is also a complementary slackness conditions for the optimal traffic deliveries, d_{pk}:

$$\left(u'_{pk}(d_{pk}) + (\pi_{O,pk} - \pi_{D,pk})\right) d_{pk} = 0 \quad \forall p \in P, k \in K$$

- If users charged a price $\widehat{\pi}_{pk}$ for flows of type k between O/D pair p, then the necessary and sufficient conditions for optimal purchase are

$$\left(u'_{pk}(d_{pk}) - (\widehat{\pi}_{i,pk} - \widehat{\pi}_{j,pk})\right) d_{pk} = 0 \quad \forall p \in P, k \in K$$

- Thus, if decentralized prices ($\widehat{\pi}_{ipk}$) are chosen same as nodal prices (π_{ipk}) from planning problem, get same solution

- Problem: To find optimal π_{ipk} requires solving the full central planning problem, with knowledge of the users' true benefit functions, $u_{pk}(\cdot)$.

  - So, *nothing gained with "decentralized" pricing*

  - congestion *externality* is the root problem: when capacity binding, need to know the (user) value of the next best use of the capacity

Generalized Vickrey Auction (MacKie-Mason and Varian 1994)

- Vickrey: Award good to highest bidder, at second highest price

- truthtelling is a dominant strategy

- bid doesn't affect price paid; truthful bid wins whenever true value $>$ price

- Generalized Vickrey:

  1. multiple units

  2. multiple goods

  3. goods with externalities (agents care how much others consume)

Intuition for Generalized Vickrey Auction

- To induce truthtelling, give each participant her utility plus all social surplus of the other agents measured using reported utility functions

- If agent reports truthfully, payoff is total social welfare – so truthtelling is dominant strategy

- But, giving *each* agent *total surplus* is very costly

- Can require a payment if it depends only on reports made by *other* agents (the "second price" in Vickrey)

Smart Market for Network Reservation

- Social objective function:

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

1. each user p reports utility fcn, $\sum_k r_{pk}(d_{pk})$
2. network planner computes optimal routing and usages taking $r(\cdot)$ as true
3. User p receives payoff

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

   where W_{-p} is reported surplus from other users
4. A convenient $G_p(\cdot)$ is

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

(subject to the routing, capacity and nonnegativity constraints for users $q \neq p$)

- Thus, the payoff is p's utility plus increase (decrease) in welfare of others caused by p's participation

- If participation creates congestion and thus lowers value of network to others, p pays precisely the incremental congestion cost as a congestion tax or premium

Smart Market Summary

- GVA elicits truthful preference revelation
- scarce bandwidth is allocated to highest valued uses, subject to technology constraints
- auction computation yields nodal spot prices, which might be used for spot pricing between reservation periods
- computation also yields congestion costs for each link, w_{ij}, which send correct signal to network planner about efficient capacity investments
- Note: Computationally intensive. The GVA mechanism requires solution of the network optimization problem once, but the payment function requires another solution for each participant (peripheral node), p