Pricing models for network services

Costas Courcoubetis

Dept. of Computer Science
University of Crete
and
Telecommunications and Networks Group
Institute of Computer Science
FORTH
Greece

Email: *courcou@ics.forth.gr*

ITC 99

Prelude

- Major sources of inspiration and collaboration
  - ACTS project **CA$h*MAN*** (Charging and Accounting for Multiservice Networks)
  - INDEX project (UC Berkeley)
  - F. Kelly, R. Weber, P. Varaiya, G. Stamoulis, V. Siris

Some thoughts...

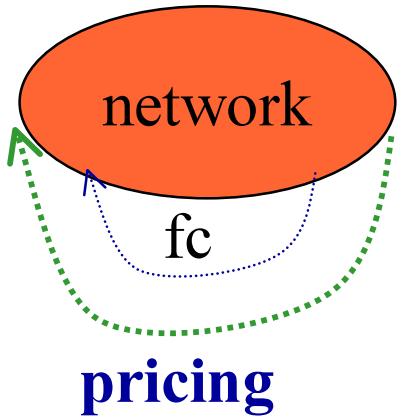
- What is different in pricing network services?
  - Network externalities, special cost structures, large monopolies
- Things are getting more complex…
  - New technologies (from application layer to physical layer)
  - Demand grows extremely fast, unpredictable
  - Costs decrease, many unpredictable aspects (interconnection, bottleneck services, technology evolution)
  - Demand for new killer applications is related to pricing

Some thoughts… (cont.)

- **New issues**
  - Regulatory actions to increase competition (unbundling, incentives for alternative technologies, reduce risk, etc.)
  - New business models, richer competitive services, E-commerce
  - Interconnection services are key
  - Resilience to new service technologies, issue of scalability
  - Congestion due to bad charging practices
  - Bottleneck technology is SW

Correct view on charging

- **Charging is not only for making profits, but for**
  - improving **value** of services to users
  - providing **stability and robustness**
  - improving **scalability** of network control
- **Charging should provide**
  - ***the right incentives*** *to users*
  - ***important information*** *to network control*
- **Charging should be**
  - simple but not simplistic
  - understandable
  - implementable
  - competitive



Outline

- **Network services**
  - **guaranteed, elastic, traffic contracts, network control, multiplexing, effective bandwidths**
- **Economic concepts**
  - **basic economic models**
  - **finite resource sharing models, congestion pricing**
  - **regulation and competition**
  - **flat rate pricing**
- **Charging schemes for elastic services**
  - **congestion price implementations, proportional fairness proposal**
- **Charging schemes for guaranteed services**
  - **constructing incentive compatible tariffs from effective bandwidths**
  - **properties of a simple time-volume charging scheme, extensions**

Network services

Contents

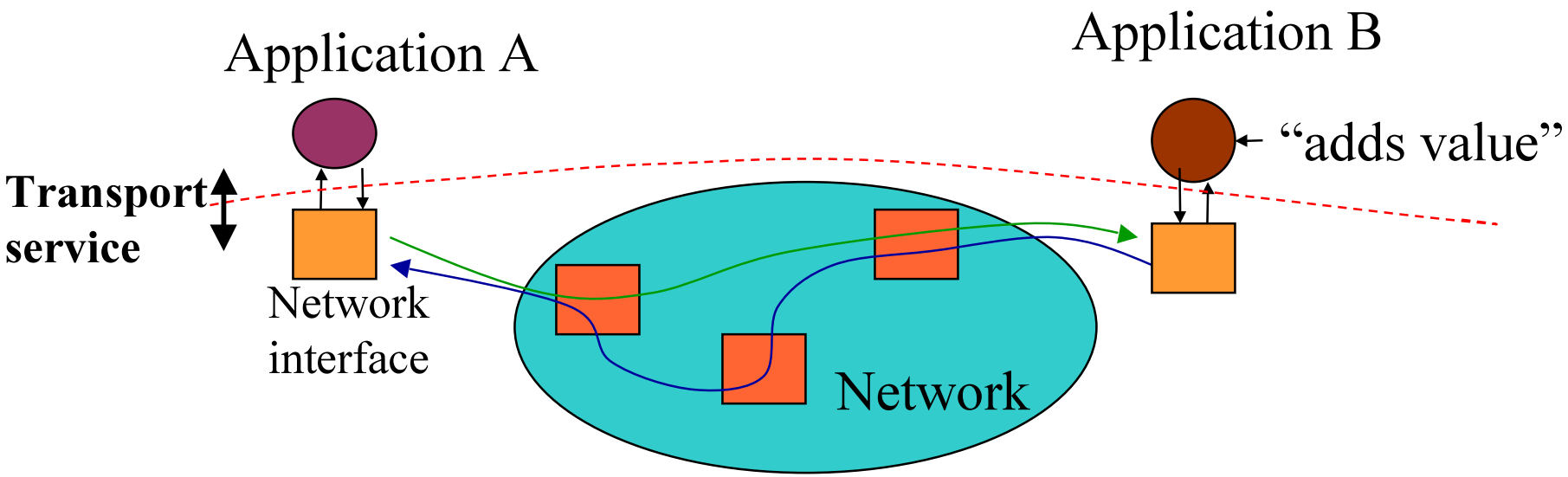
- **Service contracts**
- **Fulfilling service contracts**
- **Network control**
- **Connections with technology**
  - **ATM services**
  - **Internet services**
- **Conclusions**

General concepts:

- service contracts
- guaranteed and elastic services
- service control architectures

Network Services

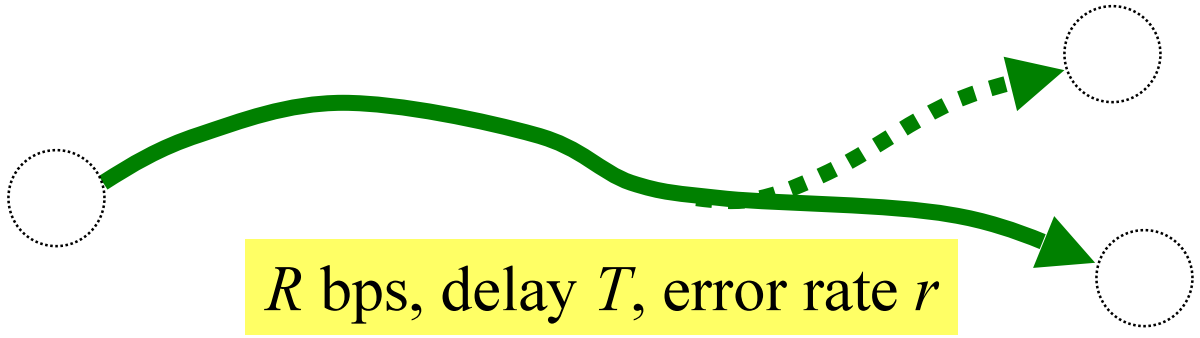
- Transport services
- Value added services



Service = transport + value-added

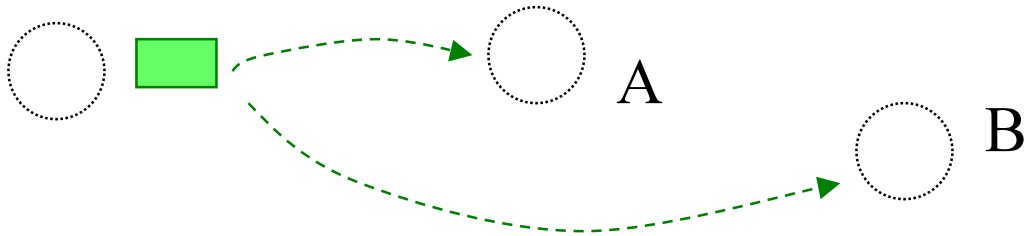
Network Transport Services

- Connection-oriented services
  - Semantics = directed virtual bit pipe (tree)



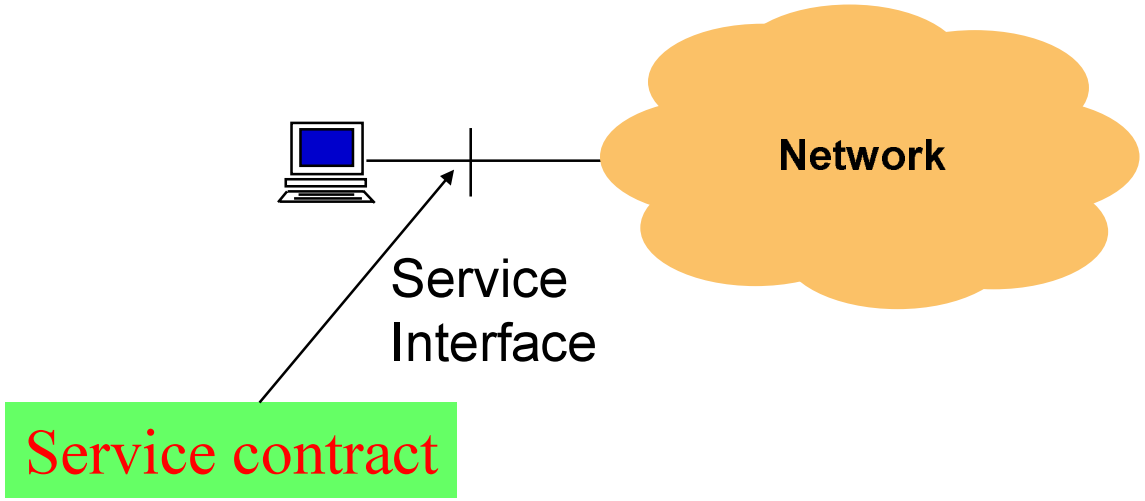
- Connectionless services
  - Semantics = datgram service (to multiple destinations)

Deliver message of size M to A,B with delay *T* and ber = *r*



Service Contracts

- Services = packet/cell transport service (1->1, 1->M)
- Traffic contract = connection's (or flow's) performance + traffic profile user must conform to
- Unicast: usually sender initiates service establishment
- Multicast: might be receiver initiated, more flavours



Network Contract Types

- **Guaranteed services (contracts)**:
  - network provides some form of performance guarantees in terms of loss, delay, and delay jitter
  - users request some amount of resources
  - subject to admission control
- **Elastic services (contracts)**:
  - no specific performance guarantees
  - performance deteriorates during overload periods
  - no specific bandwidth request; user's are able to use all available bandwidth
  - intended for applications that can adapt their sending rate

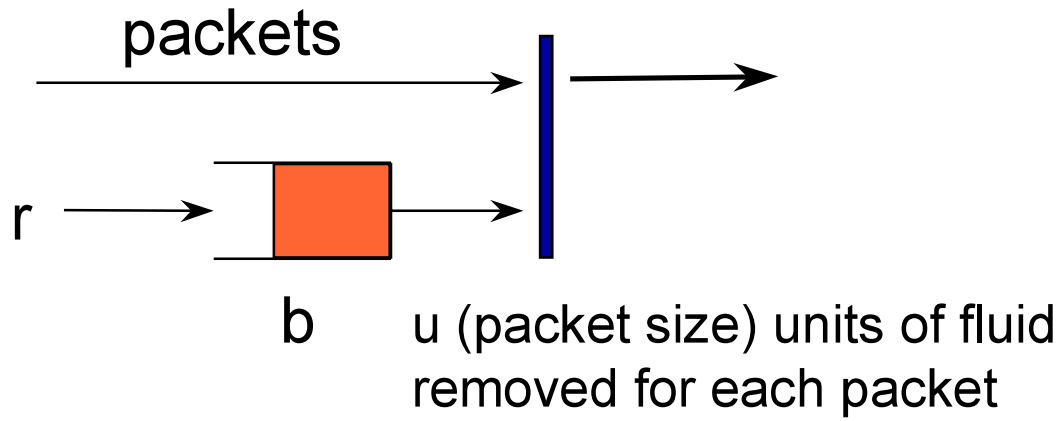
Guaranteed Services

- Performance guarantees
  - Quality of Service (QoS): loss, delay, and delay jitter
  - statistical (e.g., loss < 10^{-7}) or deterministic (delay < 30 ms)
- Required mechanisms:
  - Connection Admission Control (CAC)
  - Policing
- **User-network traffic contract**: connection's QoS and traffic description:

Network promises to support the specified QoS, provided the user's traffic is within his traffic contract

Guaranteed Services (cont.)

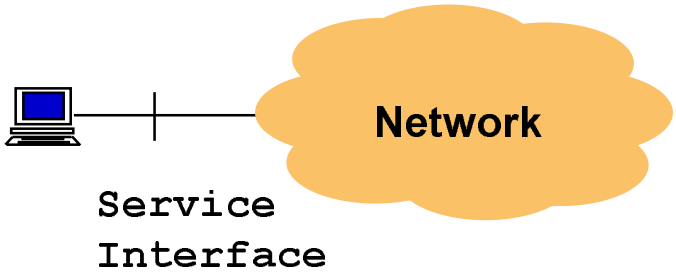
- Both ATM Forum and IETF use **leaky bucket descriptor**:
  - ATM Forum: Generic Cell Rate Algorithm (GCRA)
  - IETF: token bucket filter
- Leaky bucket: two parameters r,b
  - r: leak rate
  - b: bucket size



Elastic Services

- No specific performance guarantees, but can provide some form of fair treatment to different connections
- Feedback mechanisms inform source of congestion
  - Explicit (binary, rate), implicit (packet loss)
- Mechanisms in routers/switches to share bandwidth, enforce fairness, etc.
- Source behaviour
  - **increase** (additive) when there is no congestion
  - **decrease** (multiplicative) when there is congestion
- Examples:
  - ABR: rate-based flow control (EFCI, Explicit Rate)
  - Internet: TCP flow control

Network Control

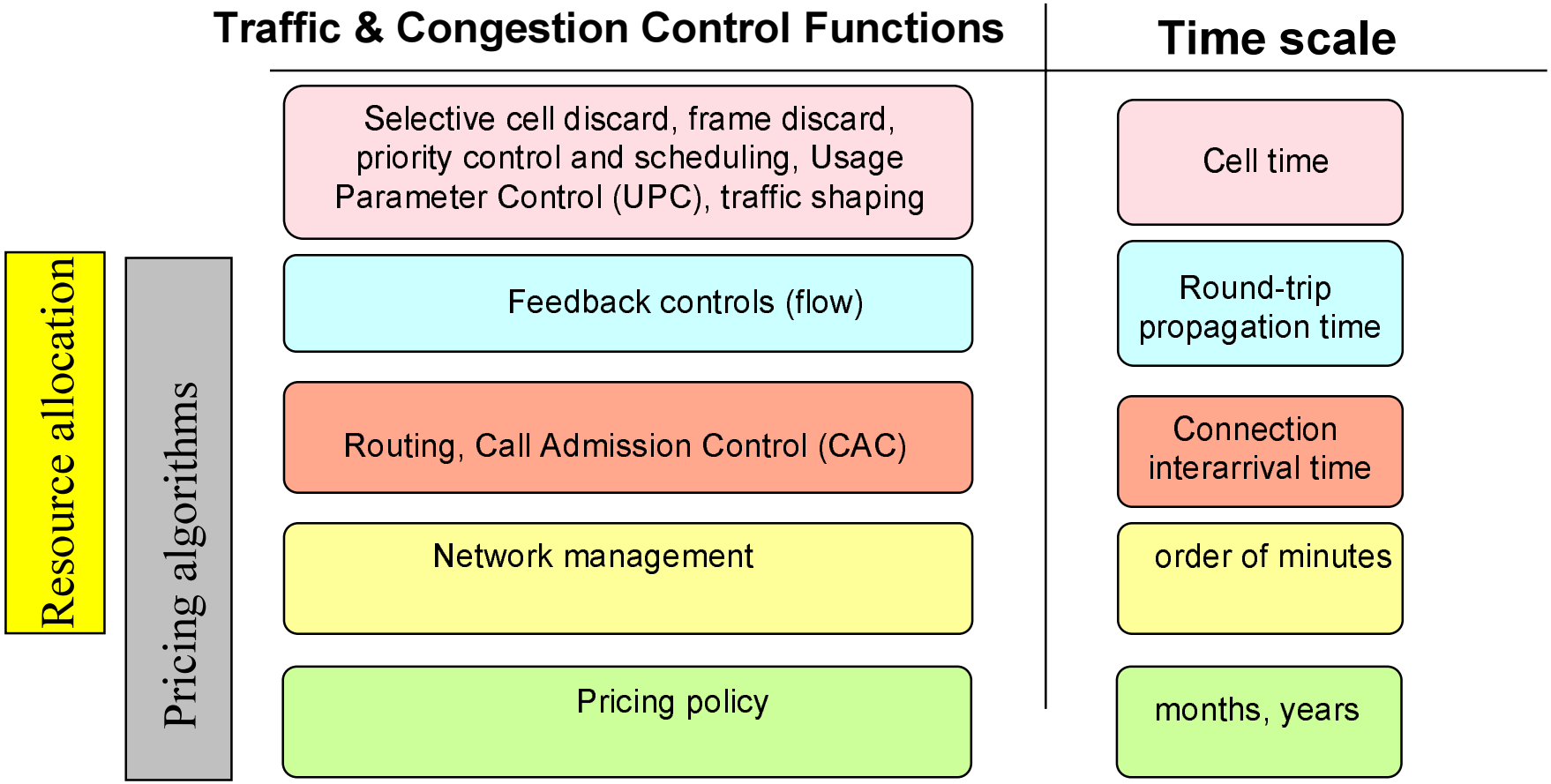


- Network control is the set of internal mechanisms used by the network in order to comply to its part of the service contracts
- Finer control capabilities -> larger set of services
- Layers of control:
  - policing and shaping
  - switching, scheduling
  - routing
  - admission control
  - multicasting
  - congestion and flow control
  - resource management
  - pricing policy
- Service architecture: control blocks needed to support a particular class of services

Network Control for Various Contract Types

- **Guaranteed services:**
  - Call Admission Control - CAC
  - **no flow control**
    - **Open loop control**
- **Elastic services:**
  - flow control
  - **no CAC** (except for MCR in ABR)
    - **Closed loop control**

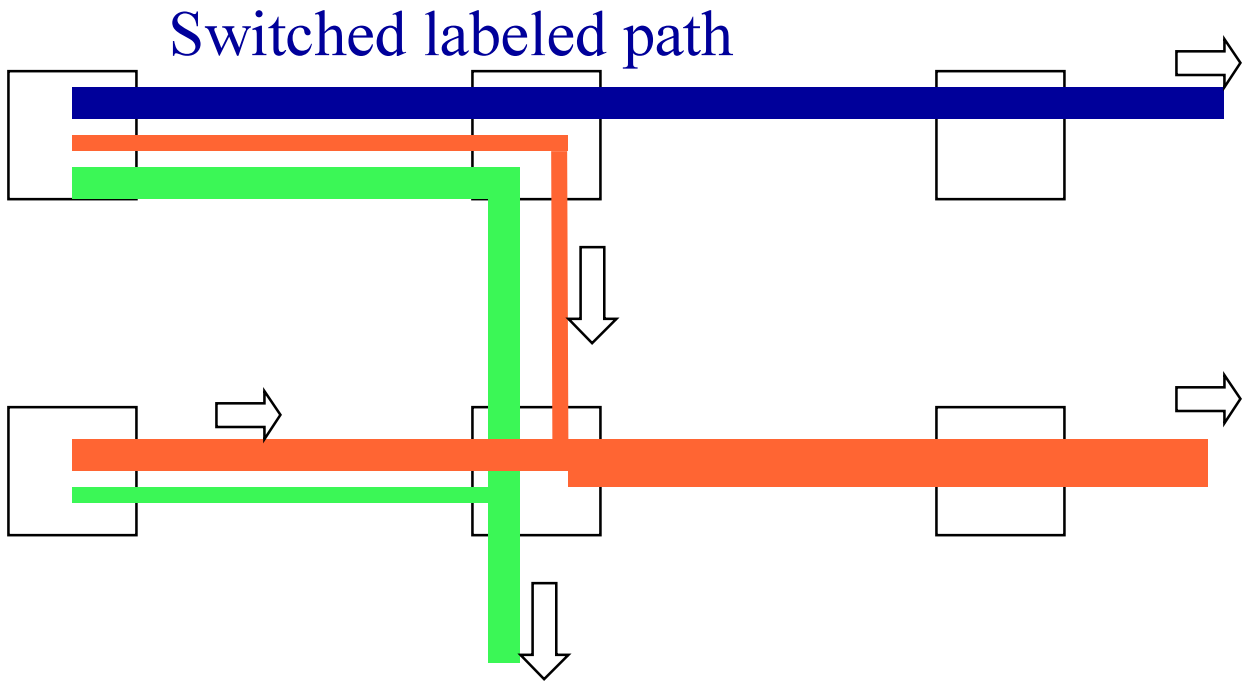
Time scales of network control



Switching

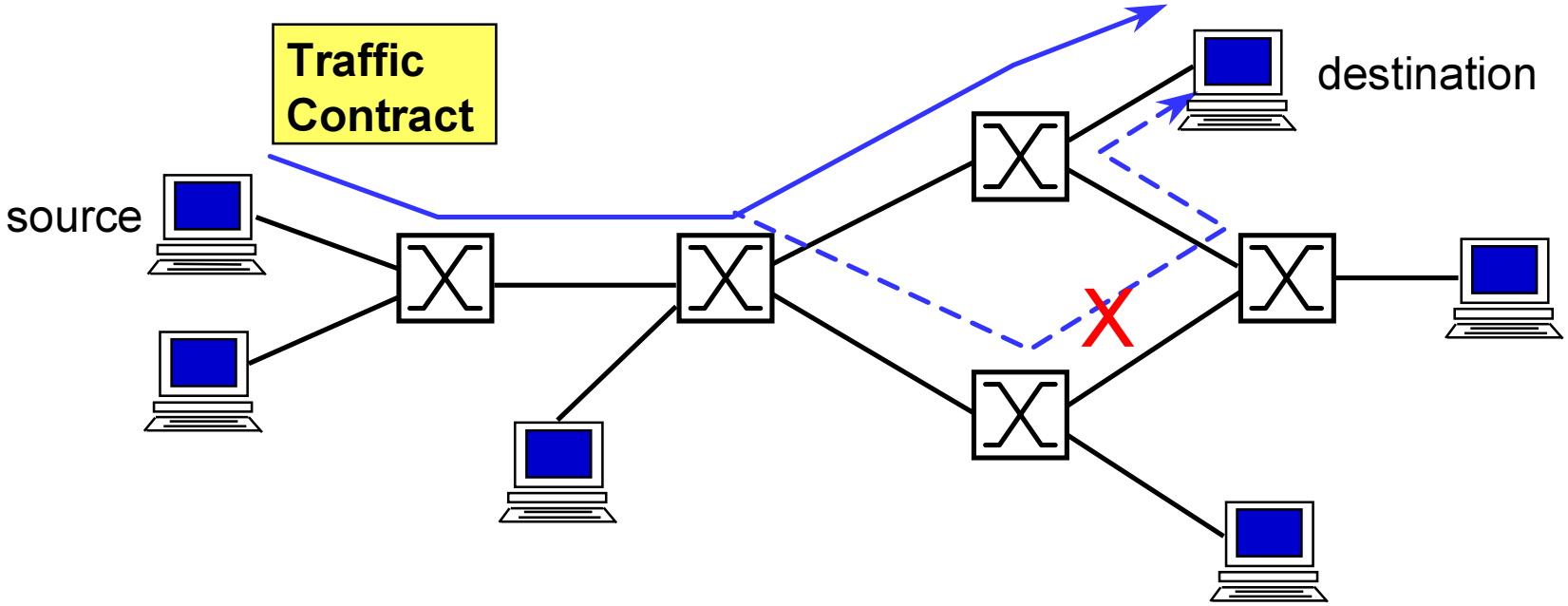
- End-to-end services are provided by switching information from node to node
  - synchronous = circuit switching (PSTN, ISDN)
  - asynchronous = packet switching (ATM, Frame Relay, IP sw)
- Packet switches
  - datagram switching
    - based on source-destination
  - label switching (virtual circuit switching)
    - based on incoming link + label
    - label is being changed at each switch

Label Switching



We can allocate resources per information pipe
How does this compares to the capabilities of datagram networks?

Control for connection-oriented services



- **Call Admission Control** (CAC): performed at every switch, determines whether there are enough resources to accept a call
- **Routing**: find path from source to destination that fulfils user requirements (bandwidth, QoS)
- **Connection set-up**: uses signalling mechanisms (labels + resource reservation)
- **Flow control**: controls flow in the circuit once it is established
- **Issues**: minimize blocking

Congestion and flow control

- Congestion = network state where resources are not utilized as desired leading to unacceptable performance
  - bandwidth not available to flows that need it the most
  - buffers are not available -> packet loss
- Congestion control
  - long time scales: pricing (tariffs), admission control
  - short time scales: priorities, flow control, dynamic pricing
- Flow control = mechanisms for controlling congestion by adjusting sending rates of applications
  - **goal: efficiency, fairness**
  - **mechanisms: rate flow control, window flow control**

Network Management

- Uses global network state to control resource allocation
- Works on slow time scales (order of minutes)
- Stirs the network operating point to a global optimum
  - where faster controls (CAC, routing, signaling, flow control) do local optimization
- Example:
  - management allocates large bandwidth pipes (virtual paths) to traffic aggregates based on historical information
  - faster control mechanisms fill the above predefined pipes as effectively as possible
  - then, management corrects pipe sizes according to actual demand

Service semantics:

- ATM
- Internet
  - integrated services
  - differentiated services

ATM Forum Service Categories

	Service Category	Typical Application
Real-Time :	Constant Bit Rate (CBR)	Circuit emulation, videoconferencing,
	Real-Time Variable Bit Rate (rt-VBR)	Compressed video/audio
	Non-Real-Time Variable Bit Rate (nrt-VBR)	Critical data
Non-Real-Time:	Available Bit Rate (ABR)	LAN interconnection,
	Unspecified Bit Rate (UBR)	File transfer, message transfer

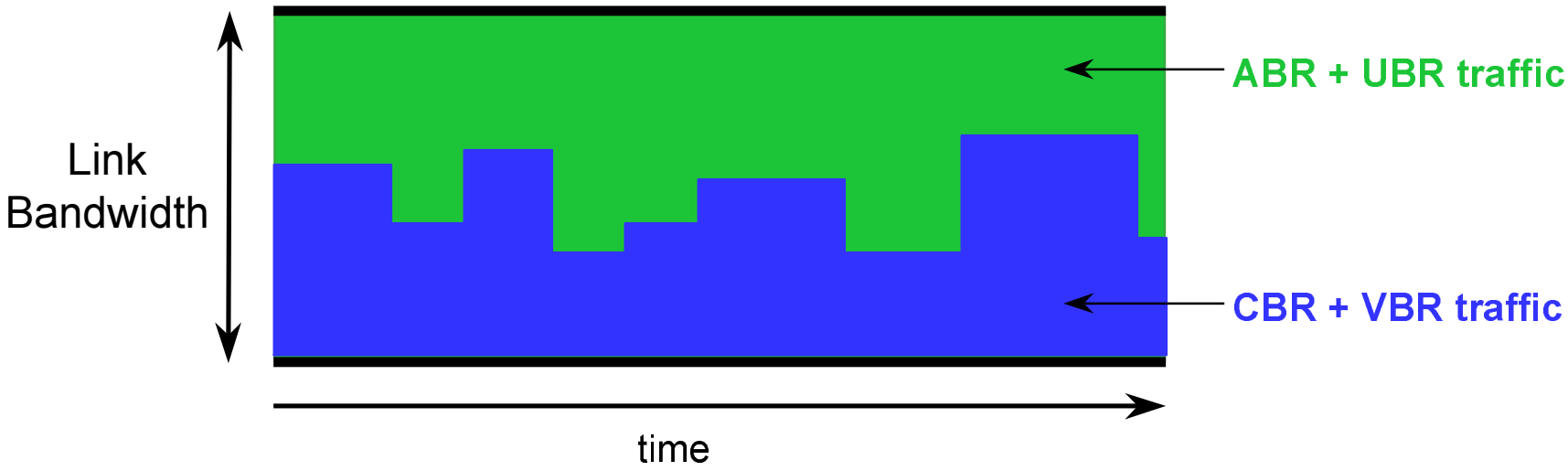
ATM Forum Real-Time Service categories

- **Constant Bit Rate (CBR):**
  - real-time applications requiring a static amount of bandwidth
  - Quality of Service (QoS) in terms of delay, delay variation, cell loss
- **Real-Time Variable Bit Rate (rt-VBR):**
  - real time applications with “bursty” traffic
  - Quality of Service (QoS) in terms of delay, delay variation, cell loss

ATM Forum Non-Real-Time Service categ.

- **Non-Real-Time Variable Bit Rate (nrt-VBR):**
  - non-real-time applications with bursty traffic
  - cell loss bound but no delay bounds
- **Available Bit Rate (ABR):**
  - “elastic” applications which can adapt their traffic rate
  - closed loop flow control supported
- **Unspecified Bit Rate (UBR):**
  - non-real-time applications, no service guarantees

ABR and UBR



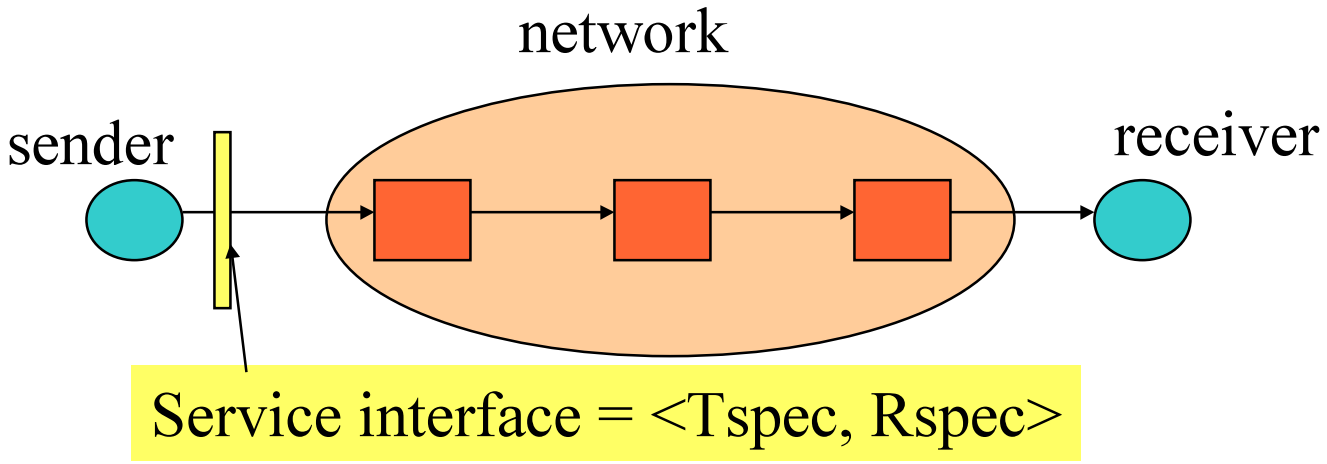
Available Bit Rate (ABR) Services

- Intended for **elastic sources** (i.e., sources which can increase-decrease their traffic rate)
- For each ABR connection:
  - PCR (Peak Cell Rate)
  - MCR (Minimum Cell Rate) - subject to admission control
- **No specific QoS parameters**
  - CLR (Cell Loss Ratio): expected low for compliant sources
  - fair share of available bandwidth
- **Rate-based flow control**
  - binary feedback (Explicit Forward Congestion Indication - EFCI)
  - rate based (Explicit Rate - ER), Resource Management cell

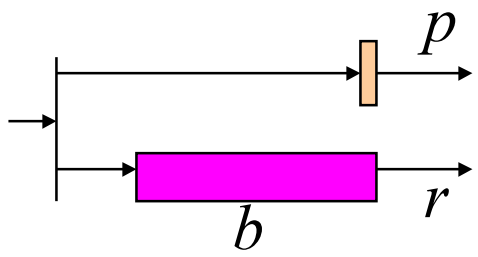
IETF Integrated Services Architecture

- **Guaranteed Service:**
  - deterministic delay guarantee
  - token bucket used to specify traffic and QoS
- **Controlled-Load Service:**
  - network provides service close to that provided by a best-effort network under lightly loaded conditions
  - token bucket used to specify traffic
- **Best-Effort Service:**
  - no guarantees

IETF Integrated services (cont.)



Tspec = $(r, b), p, M, m$



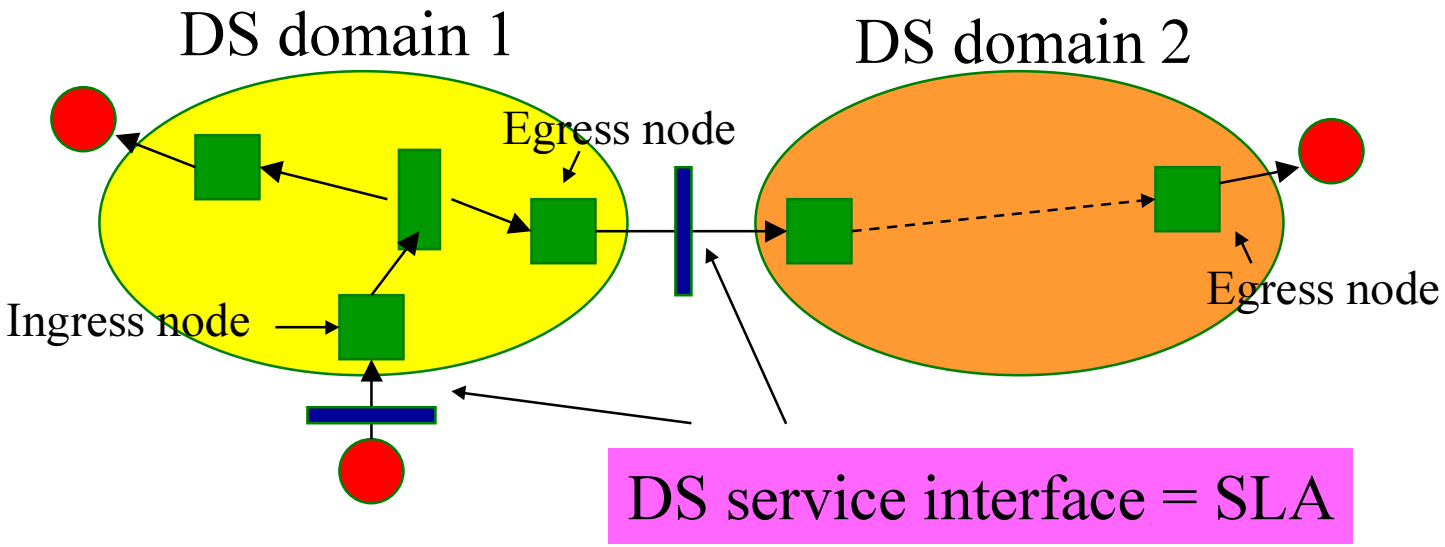
QoS = guaranteed upper bound on delay (**Guaranteed Services**)
= "as in an uncongested best effort network" (**Controlled Load**)

Rspec = implicit QoS specification
= minimum reserved capacity along the path = (R, S)

IETF Differentiated Services Architecture

- Goal: offer a range of network services (levels of performance)
  - improve revenues (premium pricing)
  - competitive differentiation
- Key concepts:
  - scalability
  - simple model:
    - traffic that enters the network is classified into a small number of classes and conditioned at the boundaries of the network
    - a class ("behavior aggregate") is characterized by a tag ("DS codepoint")
    - a router services packets according to the tags

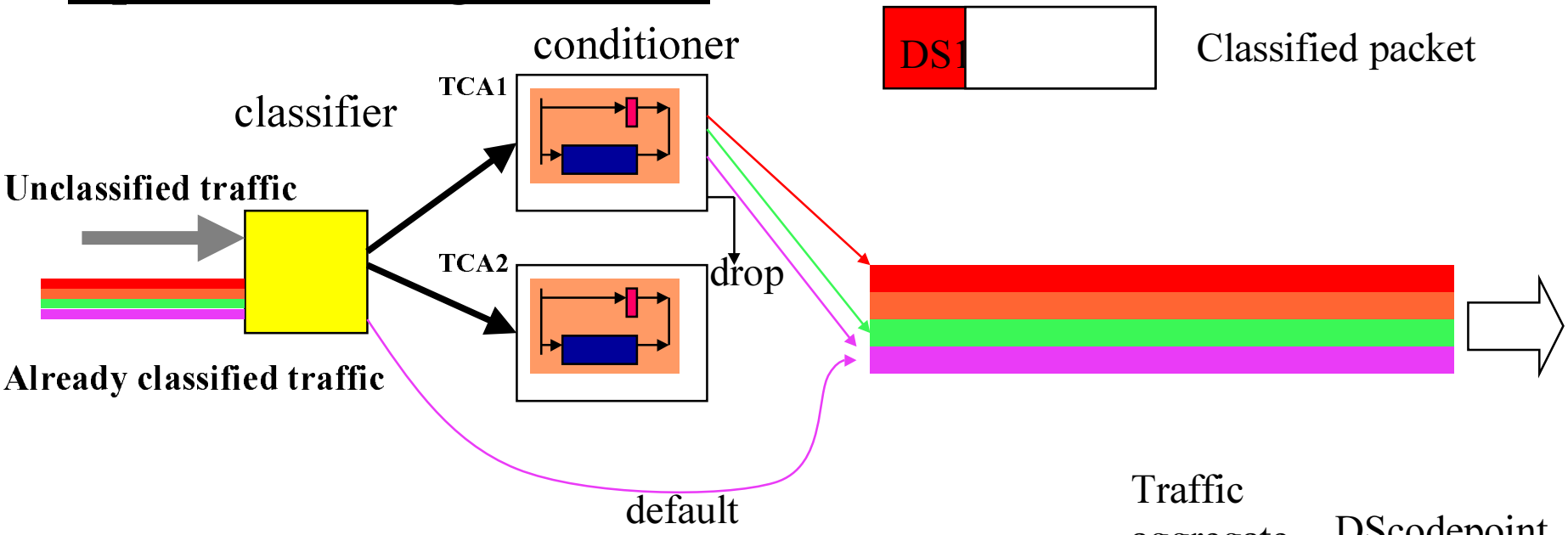
IETF DS Architecture (cont.)



- DS region: one or more DS domains
- Scope of service: one-directional traffic, point-to-multipoint,
- across domains
- QoS: quantitative, qualitative
- Dynamic and static SLAs

IETF DS Architecture (cont.)

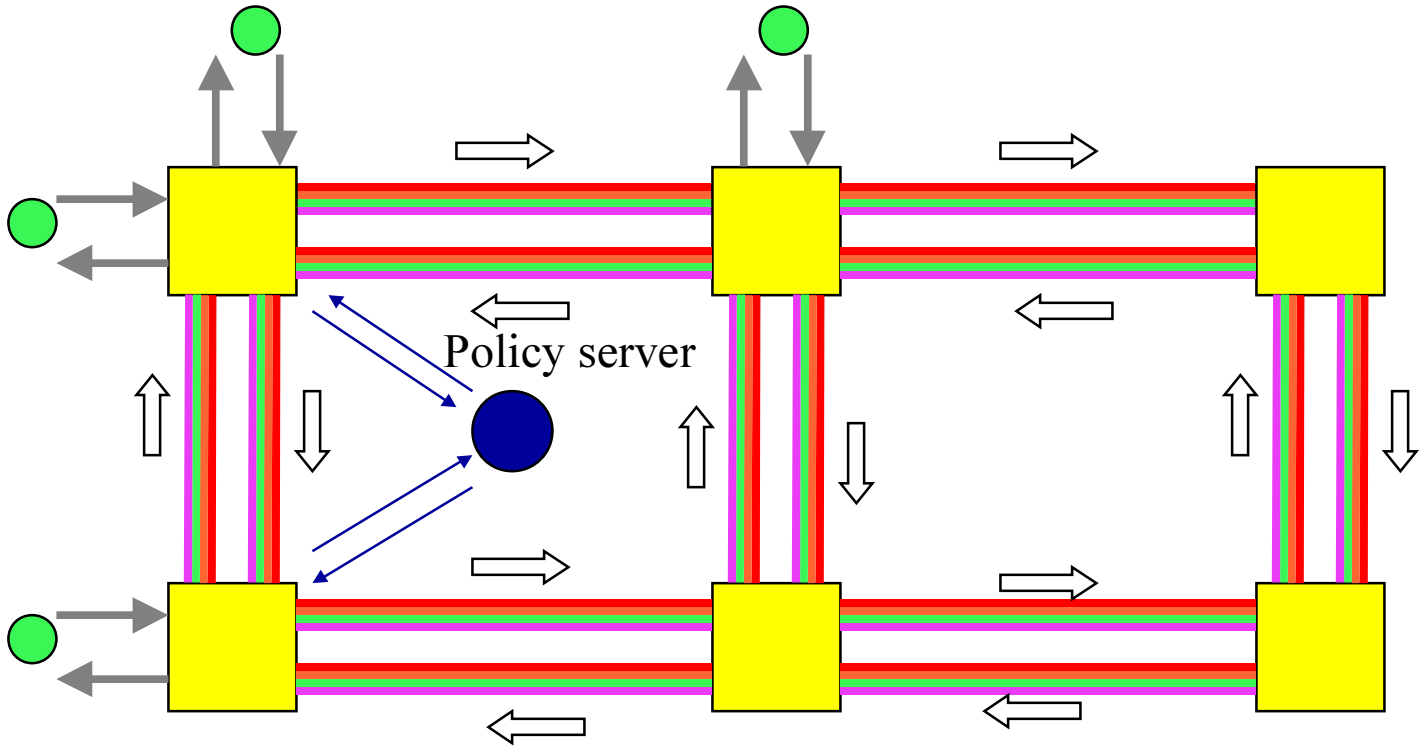
Operations at ingress nodes



Classification: based on mark, flow id
Conditioning: enforces TCA
- marks unmarked packets
- possible remarking, discarding

	Traffic aggregate	DScodepoint
	1	DS1
	2	DS2
	3	DS3
Best-effort	4	DS4

IETF DS Architecture (cont.)



Important issues:

- how to allocate resources to PHBs
- how to define implementable PHBs

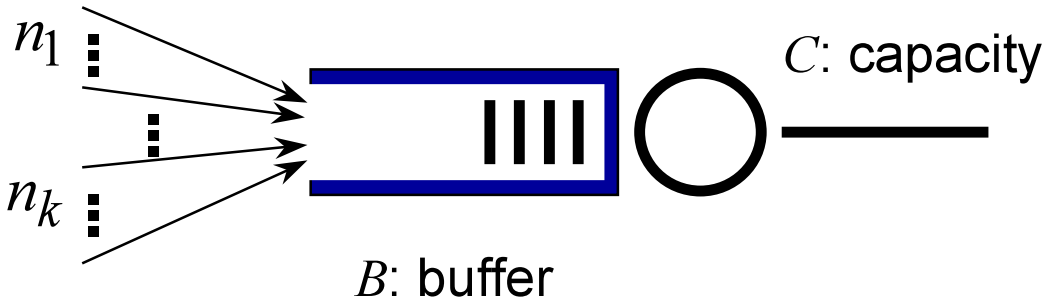
Multiplexing of guaranteed services:

- call acceptance control
- effective bandwidths

Call Admission Control (CAC)

- k traffic classes (actual or contract types)
- class i contributes

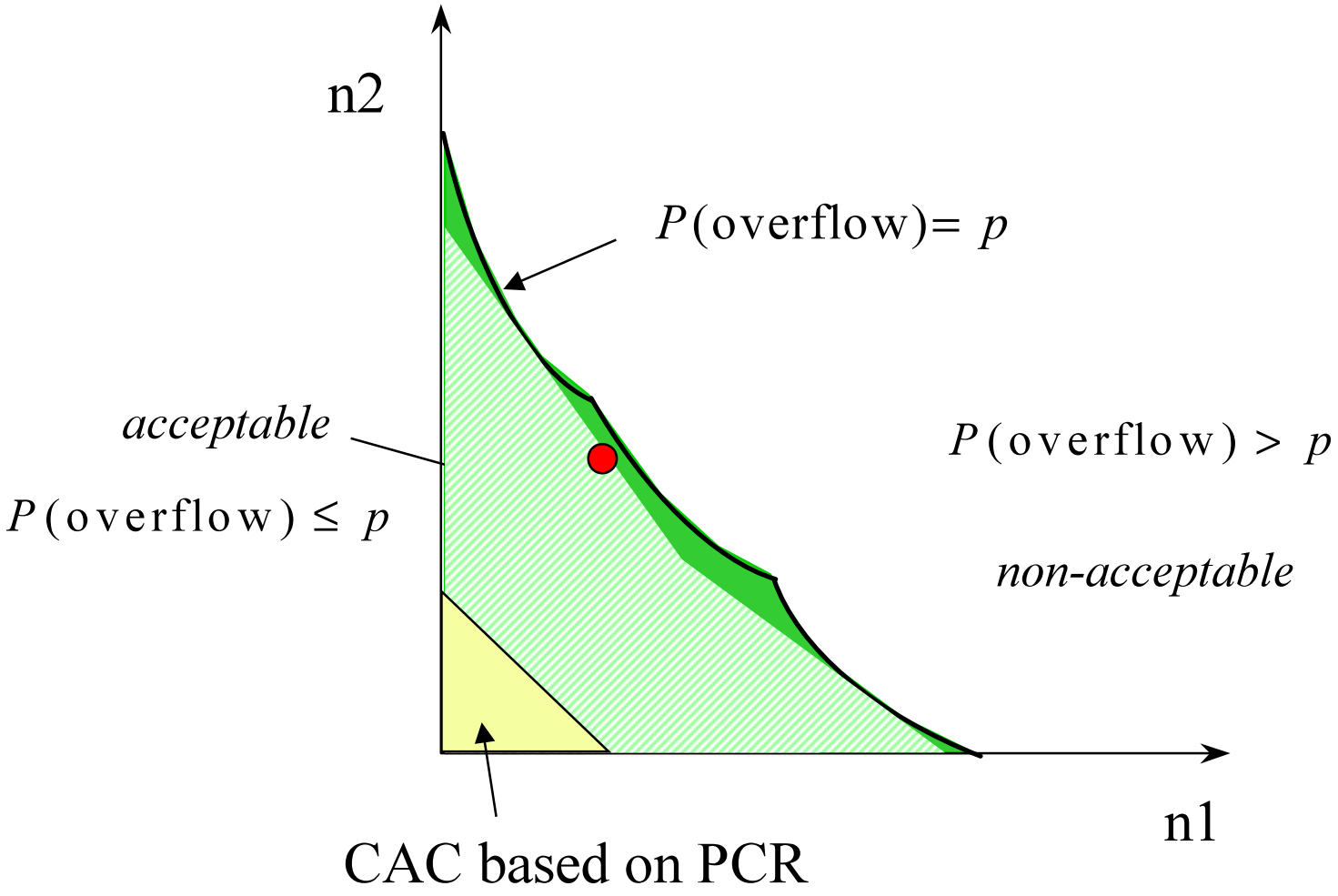
  n_i sources



- QoS constraint (contract obligation): CLP $\leq$ p (e.g. p=10^{-8})

- What $(n_1,\ldots,n_k)$ do not violate QoS constraints ?
- Approaches to CAC:
  - Non-dynamic: based only on traffic contract parameters
  - Dynamic: includes on-line measurements and contract parameters

Acceptance region



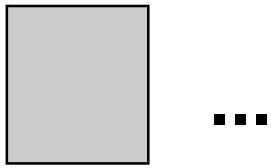
Simplifying the problem of CAC

Use *Effective Bandwidths:*

acceptance condition: $n_1 \cdot \alpha_1 + \ldots + n_k \cdot \alpha_k \leq C^*$

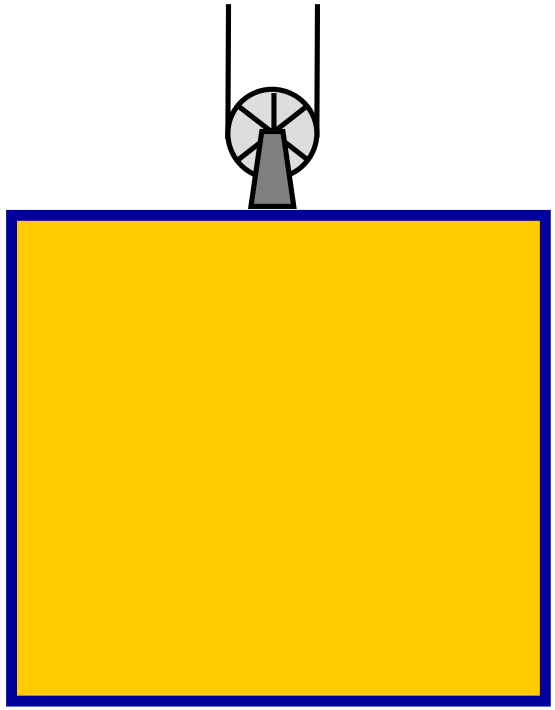
- *Can we define* $\alpha_1, \ldots \alpha_k, C^*$ *such that*
- α_i depends on **source traffic statistics**, *as well as* **traffic mix, capacity, buffer, QoS**
- C^* depends on **traffic mix, capacity, buffer, and QoS**
- Calculation of α_i can be done off-line
- The **true** acceptance region is well approximated
- **YES !**

Loading an elevator with boxes



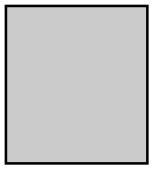
...

w_i, v_i



W_{max}, V_{max}

- What is the **relative effective usage** of a box ?
- Equivalently, in what sense



= k x 

or $\alpha_1 = k \times \alpha_2$

w_1, v_1 w_2, v_2

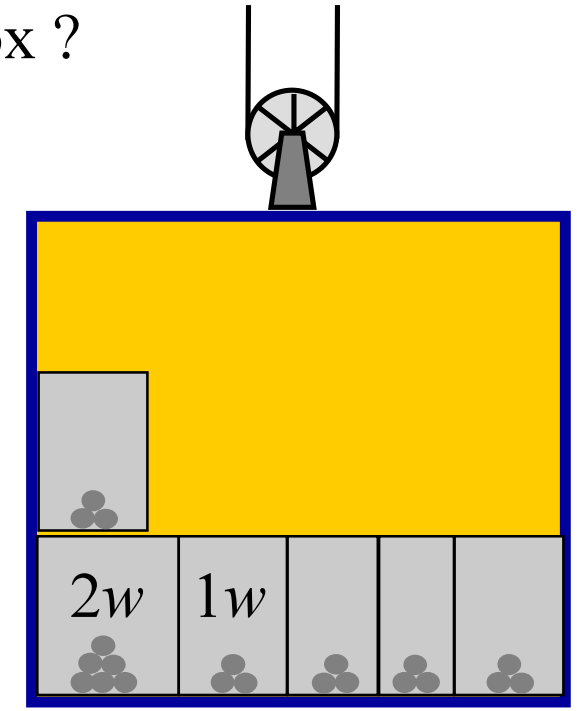
Key notion: substitution

Loading an elevator (cont.)

- What is the relative effective usage of a box ?
  - Depends on which constraint is active: *max. weight* or *max. volume*
  - Determined by operating point
- If *max. weight* is active, then effective usage equals box's *weight*

$$\sum_i w_i = W_{\max}$$

$$\sum_i v_i < V_{\max}$$



$W_{\max}$, $V_{\max}$

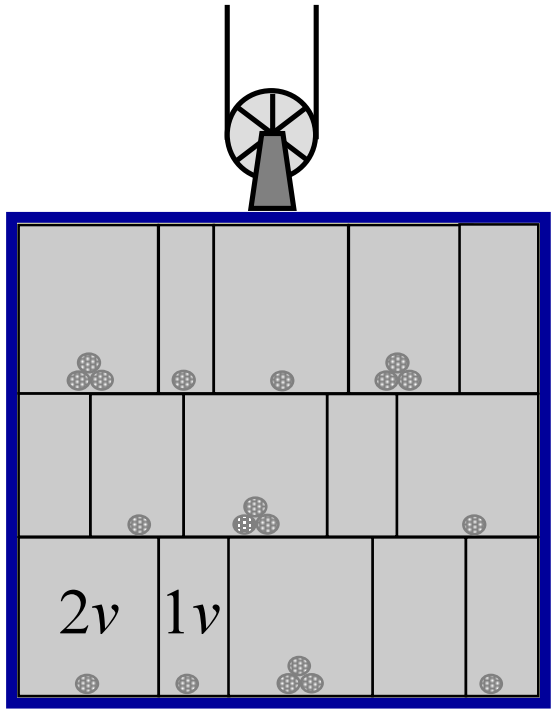
- **Effective bandwidth = weight**

Loading an elevator (cont.)

- If *max. volume* is active, then effective usage equals box's *volume*

$$\sum_i v_i = V_{\max}$$

$$\sum_i w_i < W_{\max}$$

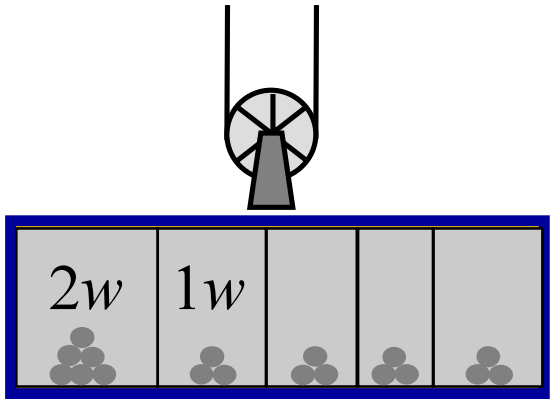


$W_{\max}$, $V_{\max}$

- **Effective bandwidth = volume**

Loading an elevator (cont.)

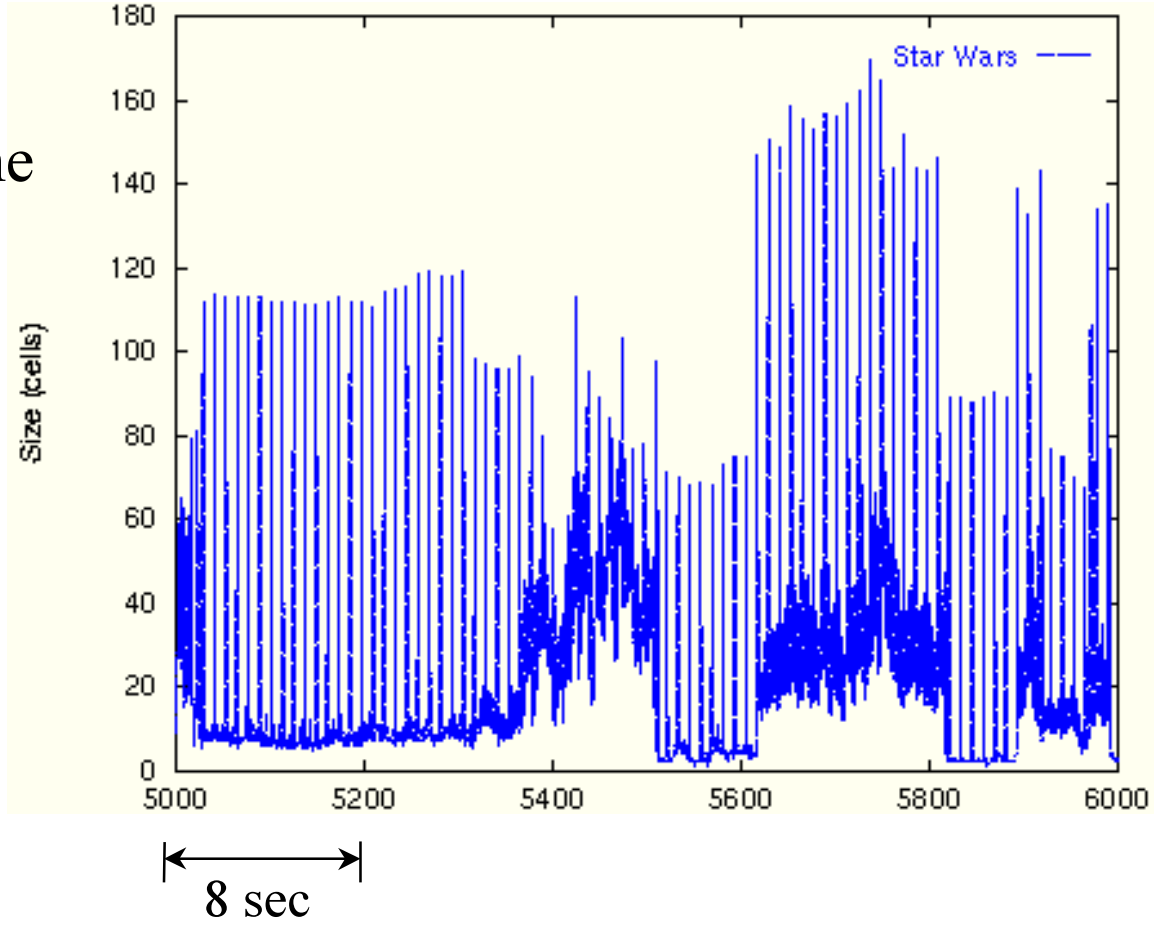
- What is the relative effective usage of a box ?
  - Depends on which constraint is active: *max. weight* or *max. volume*
  - Determined by operating point



- **Effective bandwidth = ~~weight~~ = volume**

Effective bandwidth of traffic streams

- Broadband traffic has burstiness in different time scales
- Effective bandwidth (resource usage) depends on time scales which are important for buffer overflow
- ➡ How can we identify which time scales are important for overflow?
- Dependence on context



***Star Wars* MPEG-1 trace**

An effective bandwidth formula

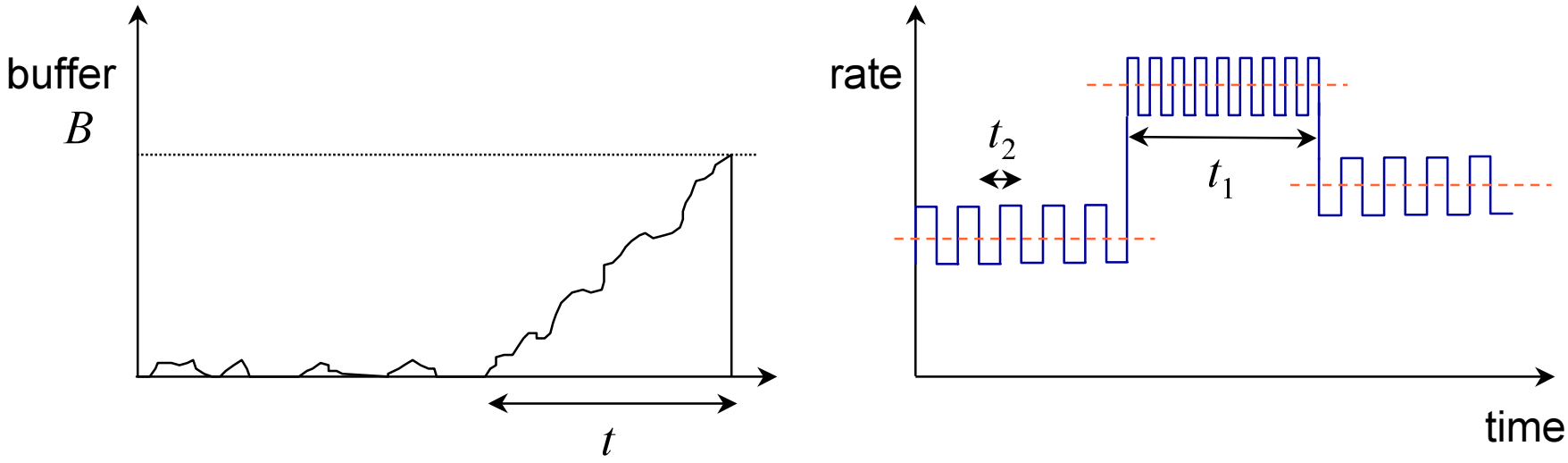
- Effective bandwidth of a source of type j

$$\alpha_j(s,t) = \frac{1}{st}\log E\left[e^{sX_j[0,t]}\right]$$

  $X_j[0,t]$: load produced by source of type j in window t

- (s,t) = **operating point of the link**
  - depends on the link param. (C,B), traffic mix, and CLP $(= e^{-\gamma})$
  - t: *time* parameter, related to time for buffer overflow
  - s: *space* parameter, depends on link's multiplexing capability, exponential tilt parameter of distributions
  - $s = \frac{\bar{\sigma}\gamma}{\partial B}$, $st = \frac{\bar{\sigma}\gamma}{\partial C}$ where $\gamma = -\log \mathrm{CLP}$

Operating point parameters *s,t*



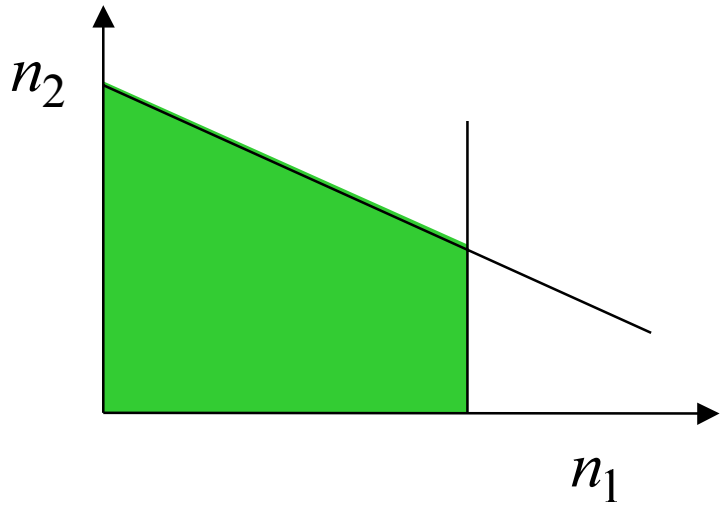
- During the overflow, the inputs have a different distribution with higher means: exponentially tilted distribution with parameter *s* (= distribution of most probable behaviour)
- Overflow period has duration *t* => we care for contribution of input sources in window *t*
  - time scale of relevant burstiness = t_1 , not t_2

Multiple QoS constraints

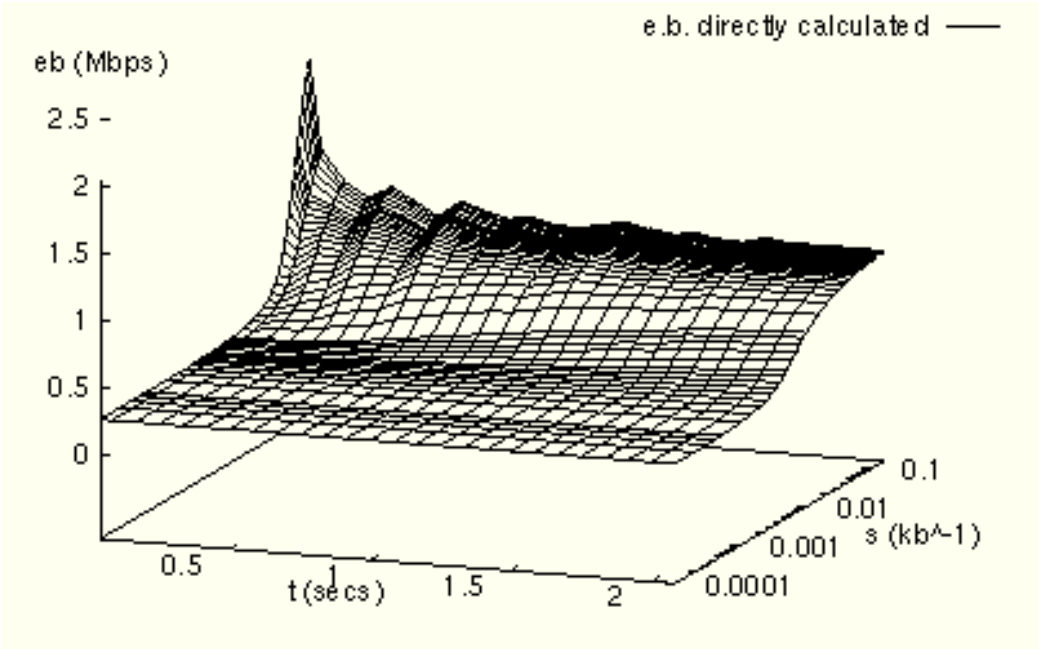
- Acceptance region described by multiple constraints
- Example: Priority queuing
  - two classes: $J_1 > J_2$
  - for J_1 : $P(delay > B_1 / C) \le e^{-\gamma_1}$
  - for $J_1 \bigcup J_2$: $P(buffer\ overflow) \le e^{-\gamma_2}$
- Two constraints:

$$n_1 \alpha_1(s_1, t_1) \le K_1$$

$$n_1 \alpha_1(s_2, t_2) + n_2 \alpha_2(s_2, t_2) \le K_2$$

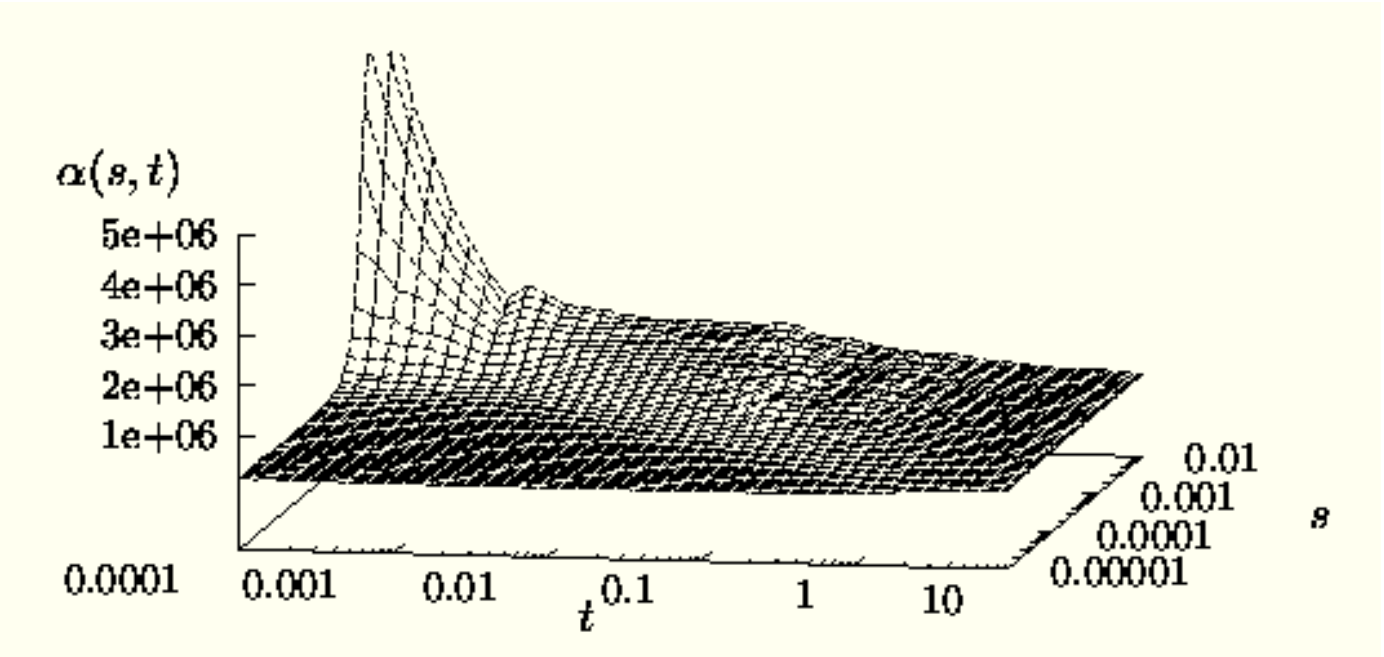


Effective bandwidth for MPEG-1 traffic



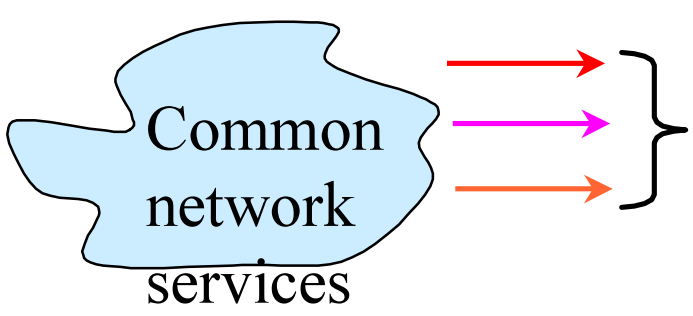
- *Star Wars* MPEG-1 trace

Effective bandwidth for Ethernet traffic

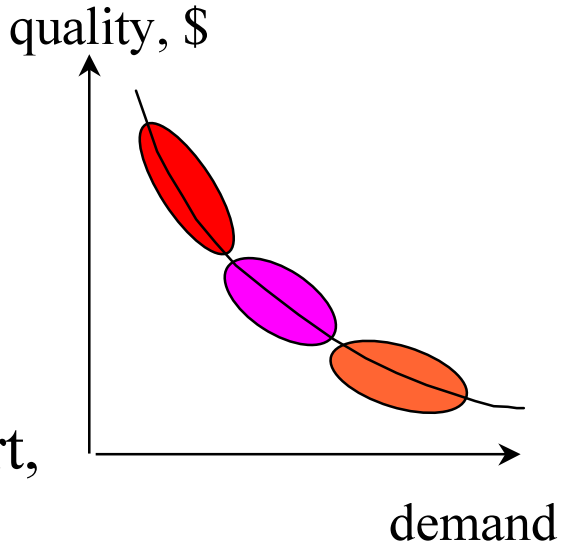


- Bellcore Ethernet trace

A proposal for pricing



Quality differentiated services



- **Quality differentiation**: guaranteed, best-effort,
  - low-high delay, blocking, reliability, access
- Prices differentiate **quality** of service, **not content**
- Between classes:
  - Prices depend on **demand**, driven towards cost by competition
- within a class:
  - Price relation defined by **substitution**; proportional to
    - -> *effective bandwidths* for guaranteed services
    - -> *throughput* for best effort services

Conclusions

- Important similarities in modern service architectures
- Convergence of basic concepts for QoS
- Generic service contract concepts (for IP and ATM)
- Effective bandwidths provide a mathematically rigorous approximation of the acceptance region
  - can be approximated by a set of **linear constraints**
- The effective bandwidth of a stream is a function of the **operating point** of the link
  - defined by network resources and consistency of traffic mix

Basic economic concepts

Contents

- Basic concepts:
  - user utility
  - demand
  - producers
- The surplus-based models:
  - social welfare maximization
  - monopoly
  - perfect competition
- Sharing finite capacity:
  - network expansion
  - congestion models
  - effective bandwidth charges
- Regulation:
  - information models
  - price regulation
  - competition
  - unbundling
- Flat rate pricing:
  - waste
  - stability
  - quality differentiation

The context

- Communication services are **economic commodities**
- **Demand factors:** amounts of services purchased by users
  - utility of using a service, demand elasticity
- **Supply factors:** amounts of services produced
  - technology of network elements, service control architecture, cost of production
- **Market model:** models interaction and competition
- **Prices:** control mechanism
  - control demand and production, deter new entry
  - provide income to cover costs
  - structure and value depends on underlying model

Basic concepts:

- user utility
- demand curve
- producer profit

Terminology

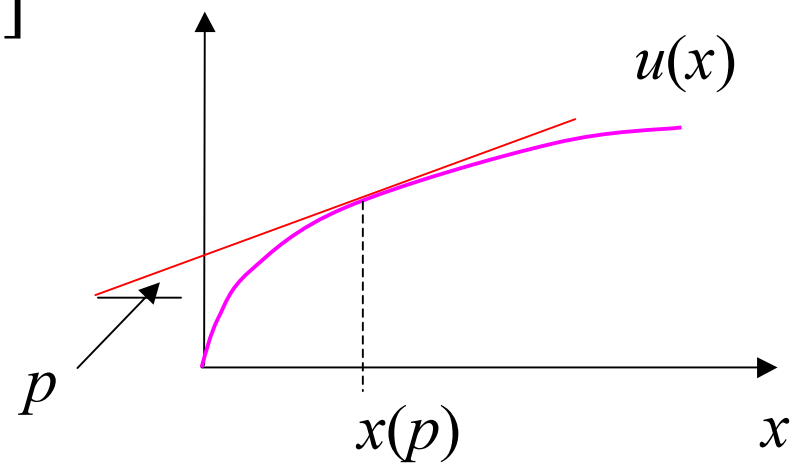
- **Terminology**:
  - **price**: associated with unit of service
  - **tariff**: price structure
    - general form of price (e.g., $a+px$)
    - *instrument for pursuing control objectives*
  - **charge**: amount to be paid (bill)

The basic model: the consumer

- Model = < consumers, producers, market mechanism >
- **Consumers:**
  - utility function $u(x)$ increasing, concave
  - **consumer surplus** (net benefit): $u(x) -$ charge for x
  - solve optimisation problem (linear prices):

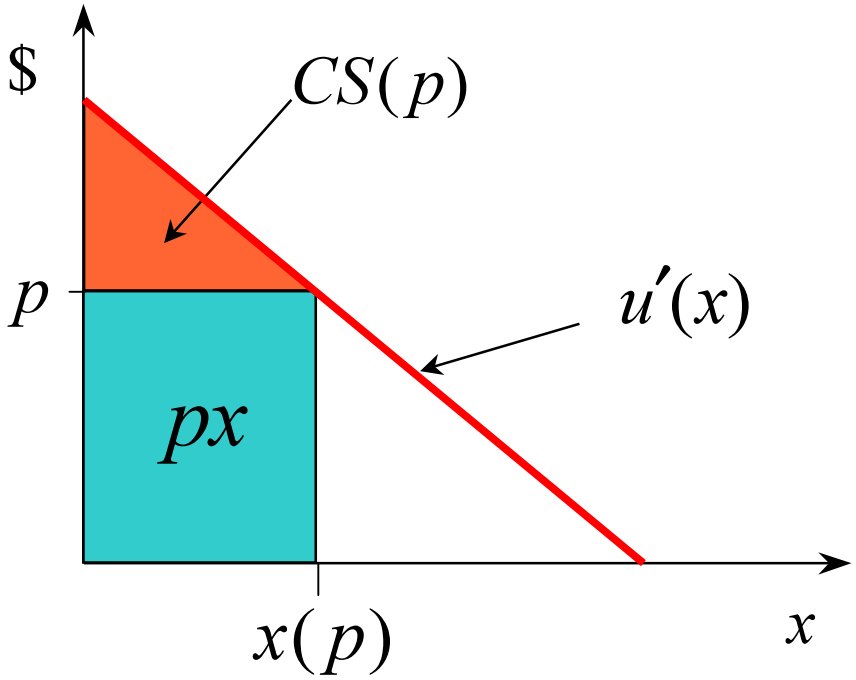
$$x(p) = \arg\max_x [u(x) - p^T x]$$

  - at optimum $\dfrac{\partial u(x)}{\partial x_i} = p_i$



The demand curve

The demand curve:



Solving the user problem:

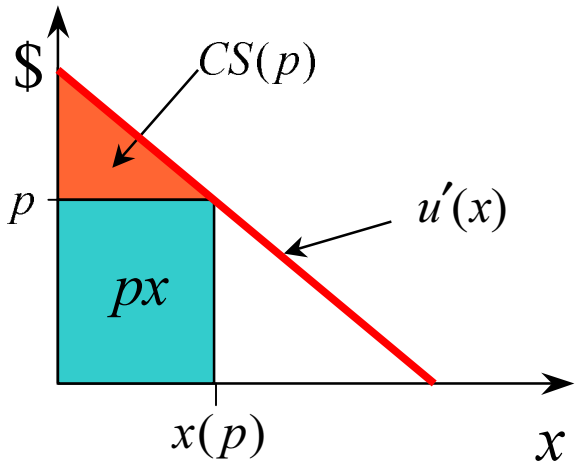
$$\max_x \{u(x) - px\}$$

$$\Rightarrow u'(x) = p$$

CS = consumer surplus

user utility $u(x) = CS(p) + px$

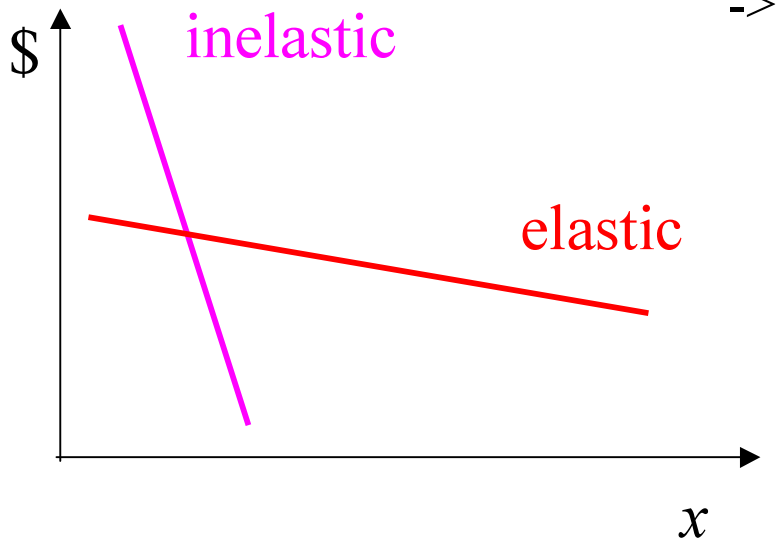
The demand curve (cont.)



Elasticity of demand: $\varepsilon_i = \dfrac{\partial x_i / x_i}{\partial p_i / p_i}$

Cross-elasticity: $\varepsilon_{ij} = \dfrac{\partial x_i / x_i}{\partial p_j / p_j}$

-> Complements, substitutes



Economic models and tariffs

- Prices result from the solution of economic models
- Three major contexts for deriving optimal prices
  - **surplus maximization**: standard market models with ***actual*** competition: monopoly, oligopoly, perfect competition
  - **stability under competition and fairness**: sustainability against ***potential*** entry, recovering costs, fairness w.r.t. cost causation, no subsidization
  - **Asymmetric information models**: principal-agent models, hidden action and hidden information

The surplus-based models

- social welfare maximization
- monopoly
- perfect competition
- oligopoly

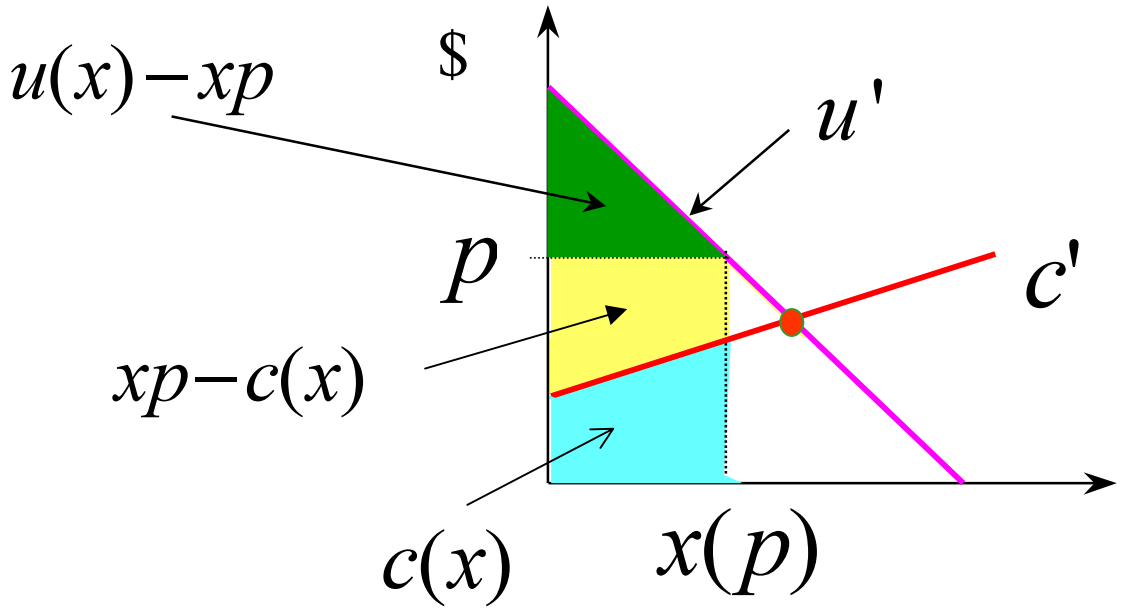
The unifying social surplus formulation

The general surplus equation:

$$\max_x \; [U(x) - xp(x)] + \lambda [xp(x) - c(x)]$$

Consumer surplus **Producer surplus**

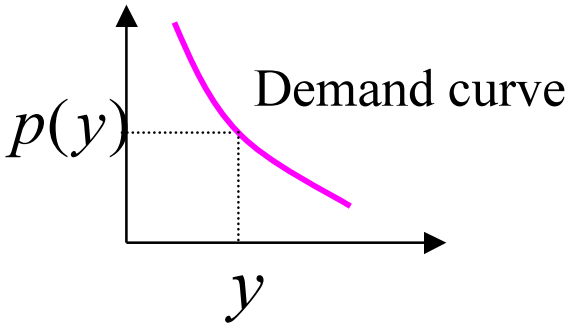
Monopoly: $\lambda = \infty$
Oligopoly: $\lambda > 1$
Perfect comp.: $\lambda = 1$
Ramsey prices: $\lambda > 1$



The supplier

- **Producer:** profit function (**producer surplus**):

$$\pi(y) = yp(y) - c(y),\ y \in Y$$



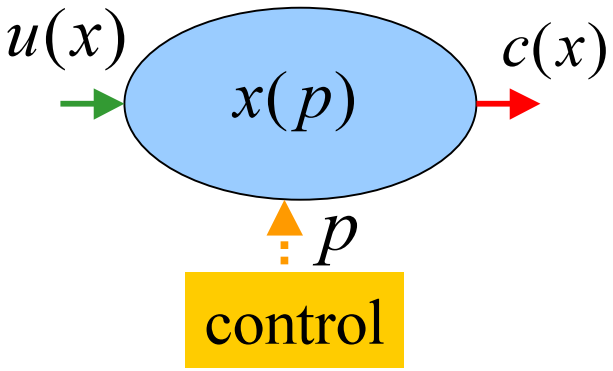
Monopoly: $\max_{y \in Y} [\, p(y)\, y - c(y)]$

Perfect competition: $\max_{y \in Y} [\, py - c(y)]$, for given p

Oligopoly: $\max_{y \in Y} [\, p(y + z)\, y - c(y)]$

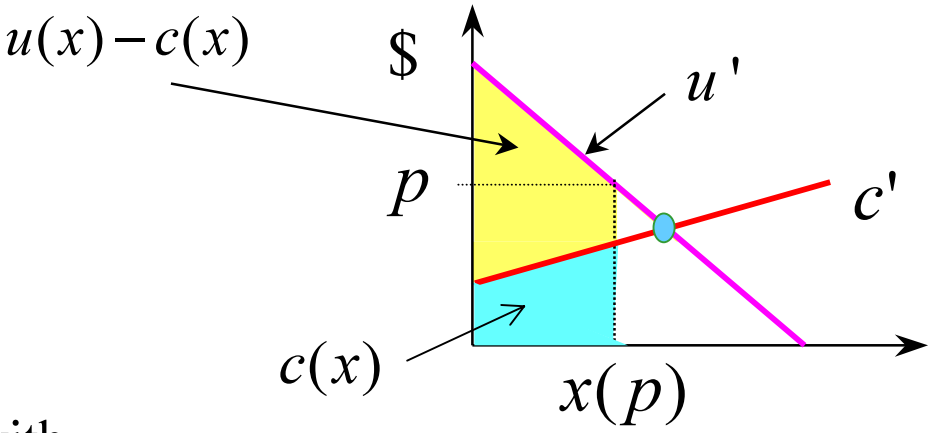
Regulation: fix p, produce $y = y(p)$

The social planner's problem



$$\max_{x} \; u(x) - c(x) \Leftrightarrow$$

$$\frac{\partial u(x^*)}{\partial x_i} = \frac{\partial c(x^*)}{\partial x_i} \quad \textbf{= MC}$$



Note that this is equivalent with

$$\max_{p}[\{u(x(p)) - x(p)\,p\} + \{x(p)\,p - c(x(p))\}] = \max[CS + \pi]$$

Marginal cost prices

- **Strong points:**
  - welfare maximisation under appropriate conditions
  - firmly based on costs
  - easy to understand
- **Weak points:**
  - **do not cover total cost** (need for subsidisation)
  - must be defined w.r.t. time frame of output expansion
    - **short run marginal cost = 0 or ∞**
    - use long-run marginal cost (planned permanent expansion)
  - difficult to predict demand and to dimension the network
  - difficult to relate cost changes to marginal output changes

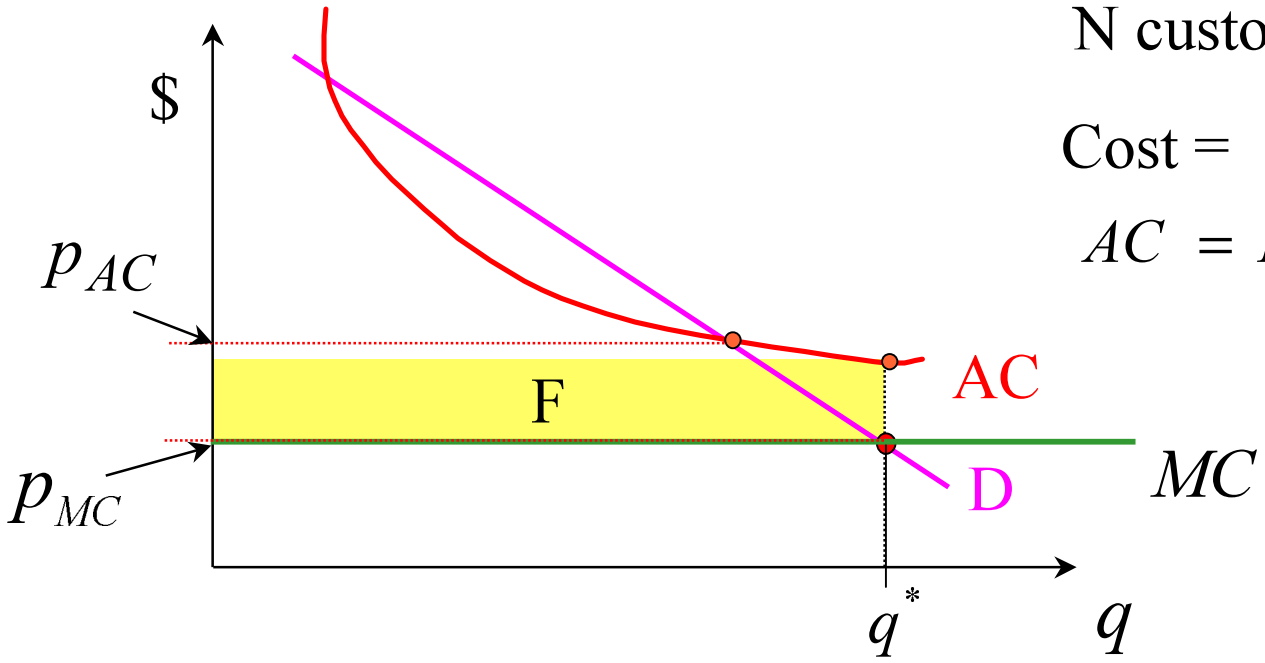
Recovering network cost

- **Pricing at marginal cost maximises efficiency but does not necessarily recover network cost**
  - example: assume $r(q) = \alpha + \beta q$

  Then under marginal cost pricing, $p = \beta$

  and the network revenue is βq, hence we are short of α
- **Ways out:**
  - add fixed fee (two-part tariffs)
  - Ramsey prices
  - general non-linear tariffs

Two-part tariffs



N customers

Cost = $F + MCq$

$AC = F / q + MC$

Under *MC*, network needs to recover an additional amount *F*

Use tariff $F / N + MCq$

Customer benefit = $u(q^*) - F / N - MCq^*$ < 0 ?

Ramsey pricing

- **Problem**: **Maximise overall efficiency given that network revenue covers network cost**

$$\max_{\mathbf{p}} \; u(x(p)) - c(x(p)) \qquad s.t. \quad c(x(p)) = x(p)\,p$$

$$\max_{x} \; u(x) - c(x) + \lambda(x^T p(x) - c(x))$$

$$\Leftrightarrow p_i - c' + \lambda(p_i + x_i \frac{\partial p_i}{\partial x_i} - c') = 0$$

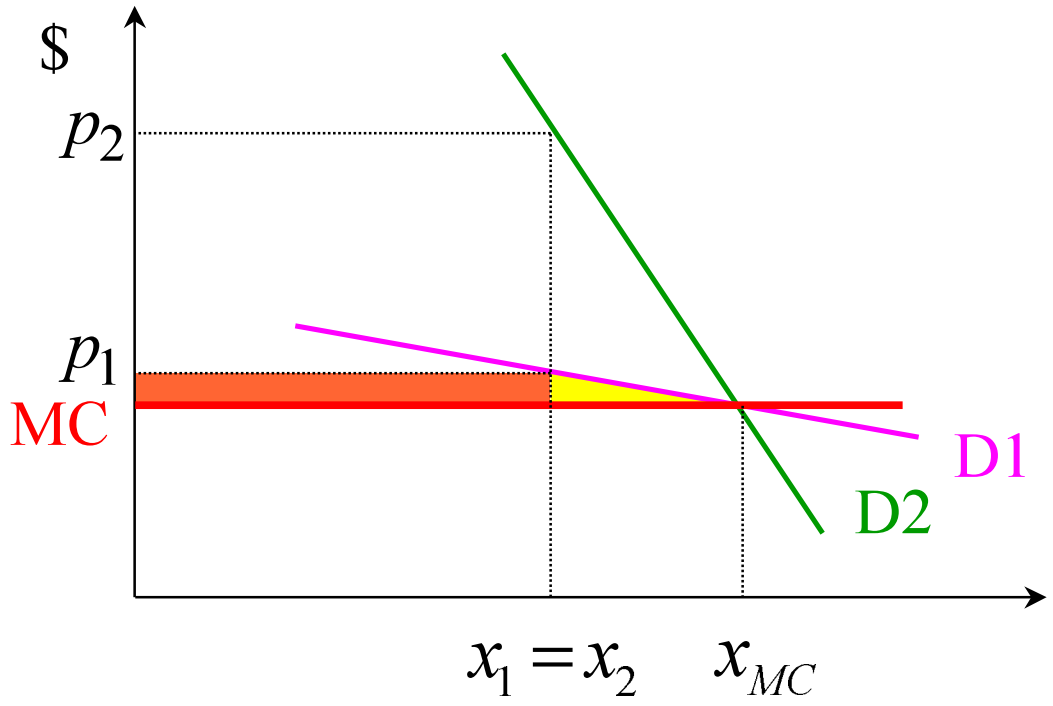
$$\Leftrightarrow p_i(1 + \gamma \frac{1}{\varepsilon_i}) = c', \quad \varepsilon_i = \frac{\partial x_i / x_i}{\partial p_i / p_i}$$

$$\Leftrightarrow \frac{p_i - c'}{p_i} = -\gamma / \varepsilon_i$$

Ramsey pricing (cont.)

- **Property of Ramsey pricing:**

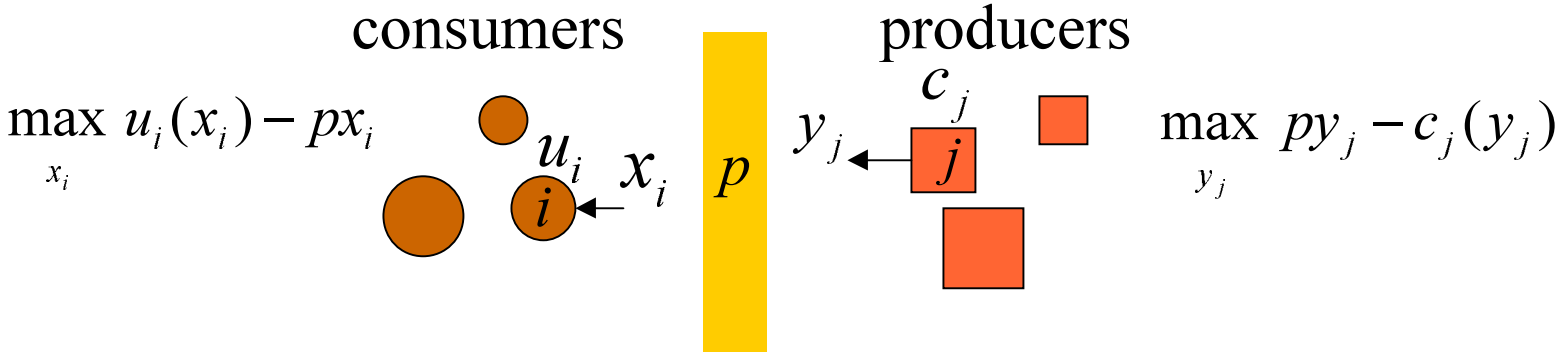
Quantities deviate almost proportionally from the ones under marginal cost pricing



Same increase in profit per unit of reduction of social surplus

Perfect competition

- Every participant in the market is small, can not affect prices
- Equilibrium: stable point where production = demand, price p



Market clearance: $\sum_i x_i(p) = \sum_j y_j(p)$

$$\frac{\partial u_i}{\partial x_i} = \frac{\partial c_j}{\partial y_j} = p \Leftrightarrow \max_{\{x_i, y_j\}} \sum_i u_i(x_i) - \sum_j c_j(y_j)$$

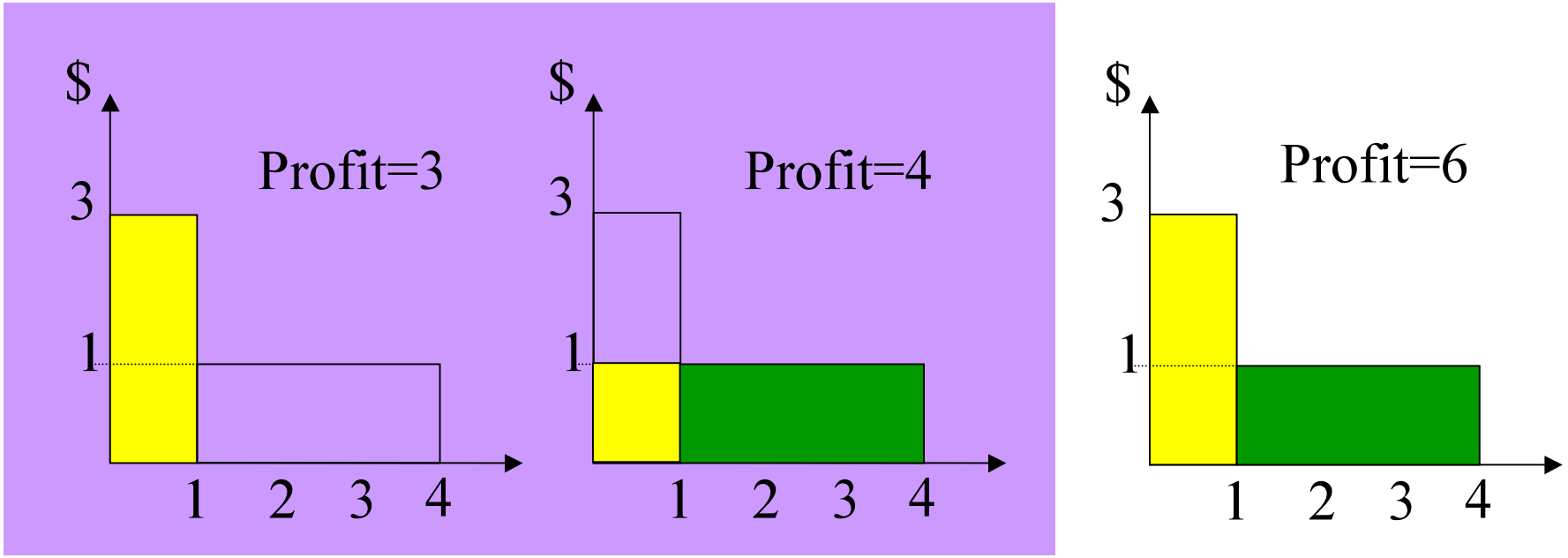
$$s.t. \quad \sum_i x_i \geq \sum_j y_j$$

=> **Social welfare optimum!**

=> **Tatonnement**

Monopoly: an example

Sell a product to different customer types



Price discrimination: personalized pricing, versioning, group pricing

Monopoly

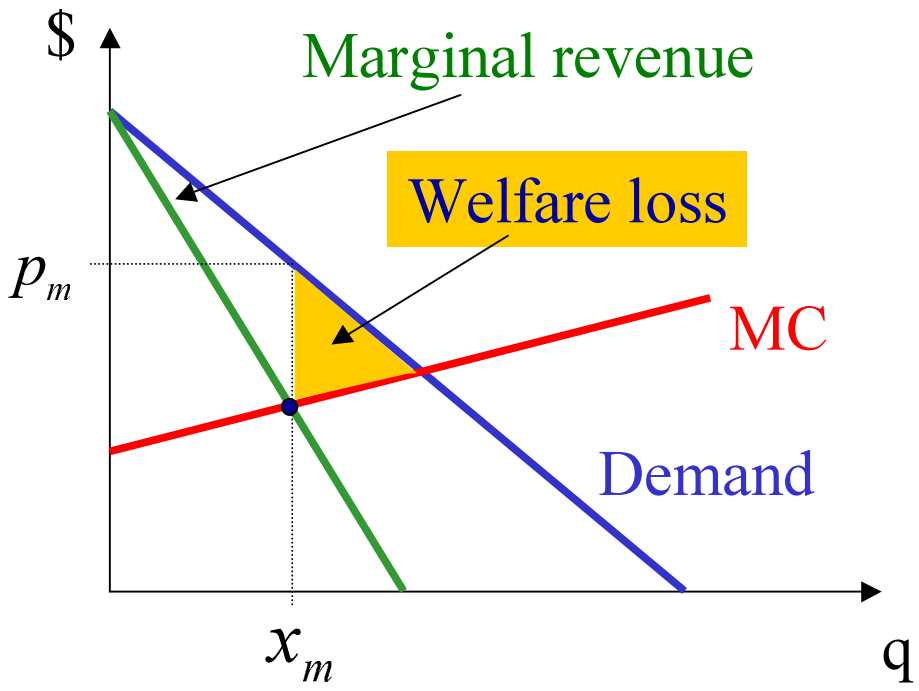
- **Goal**: maximize profits
- **Advantage:** economies of scale (small MC)
- **Disadvantage:** inefficiency, small consumer surplus

$\Rightarrow$ Combine with regulation

$$\max_x p(x)^T x - c(x) \Leftrightarrow$$

$$p_i(x) + \frac{\partial p_i}{\partial x_i} x_i = c' \Leftrightarrow$$

$$p_i(x)[1 + \frac{1}{\varepsilon_i}] = c'$$



Price discrimination

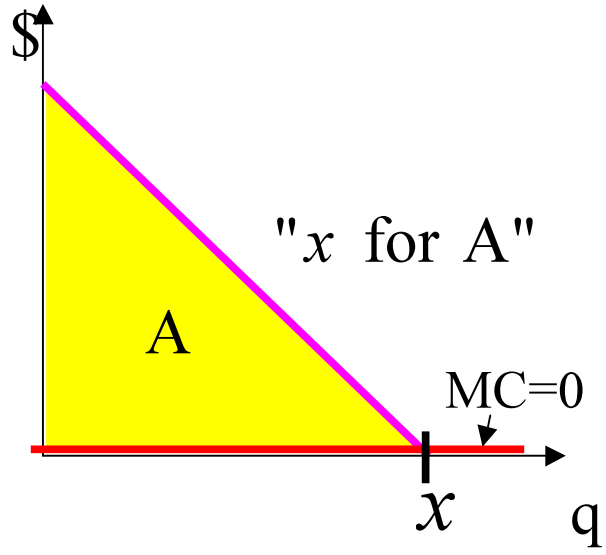
First-degree price discrimination:

- extracts maximum profit from customer
- addresses each customer separately
- "take it or leave it" offer "amount x for m dollars"
- Pareto efficient operation

$$\max_{x,m} \; m - c(x) \quad s.t. \quad u(x) - m \geq 0 \Leftrightarrow$$

$$\max_{x} \; u(x) - c(x) \Leftrightarrow$$

$$u'(x) = c'(x)$$



Sharing finite capacity:

- network expansion
- congestion models
- effective bandwidth charges

Sharing finite resources

- Network resource management occurs in various time scales
  - ***short time scales***: amount of resources is **fixed**, and control deals with **optimal sharing**
  - ***long time scales***: assuming the optimal operation of the network during the short time scales, **resources are expanded** in order to improve average performance and accommodate increased demand

The short time-scale problem:

•prices are used to control the way resources are shared

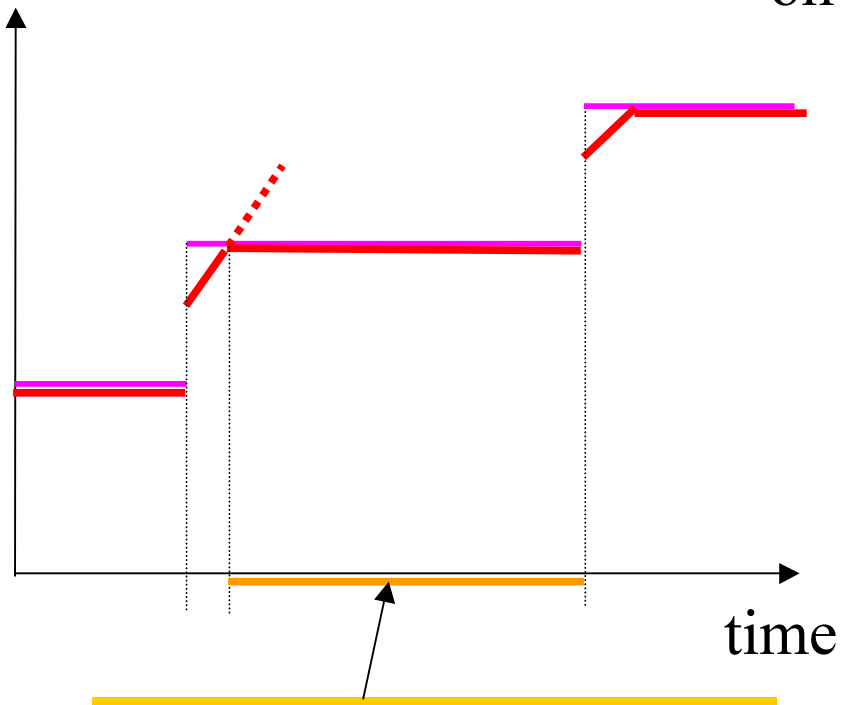
•can be used as input for deciding capacity expansion

Sharing finite resources (cont.)

capacity
demand

Tariff = $A+px$

Constant part A: recovers network cost (expansion+operational) on top of usage part px



Maximizing SW with congestion cost

The general form of the social surplus problem is

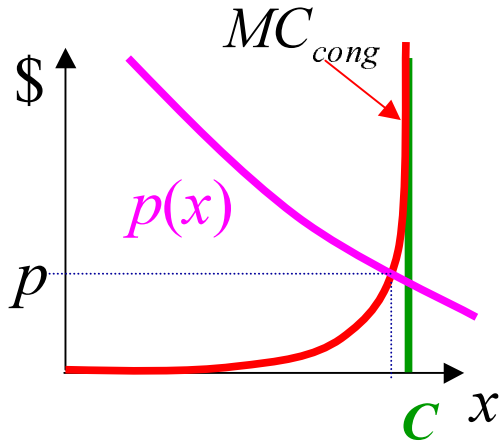
$$\max_x U(x) - c(x) \; s.t. \; x \le C$$

where $c(x) = c_{op}(x) + c_{cong}(x)$

$c_{cong}(x)$ = **congestion cost** = cost due to performance degradation when load = x

Assume $c_{op}(x) = F, c_{cong}(x) \underset{x->C}{\longrightarrow} \infty$

Then $\frac{\partial U(x^*)}{\partial x} = \frac{\partial c_{cong}(x^*)}{\partial x} = p_{cong}$



Maximising SW with no congestion cost

- M service types, N users, finite capacity C
- a_i = effective bandwidth of service i

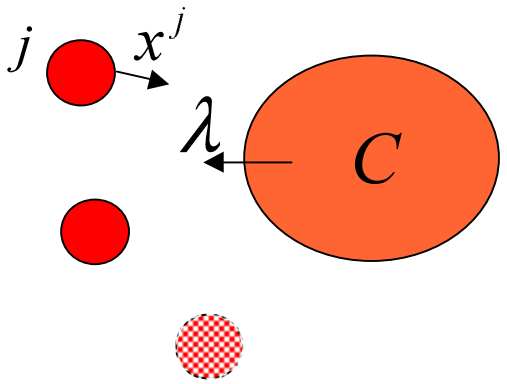
$$\max_{\{x_i^j\}} \sum_{j=1,\ldots,N} u^j(x_1^j,\ldots,x_M^j) \quad s.t. \sum_j \sum_i a_i x_i^j \le C \Leftrightarrow$$

$$\max_{\{x_i^j\}} \sum_{j=1,\ldots,N} u^j(x_1^j,\ldots,x_M^j) - \lambda[\sum_j \sum_i a_i x_i^j - C] \Leftrightarrow$$

$$\frac{\partial u^j}{x_i^j} = \lambda a_i, \forall j$$

=> price per unit of service i = λa_i

λ can be computed by tatonnement



Price discrimination and usage charging

Interesting case: monopolist with complete control on prices, and no operating cost

Optimal strategy: use congestion prices to maximize consumer surplus, then take it all back using subscription fees

Example: two customers, $u_i(x_i), i = 1,2, \quad a_1 x_1 + a_2 x_2 = C$

If λ is the optimal congestion price in SW maximization with corresponding shares x_1^*, x_2^*, then use tariffs

$$[u_i(x_i^*) - \lambda a_i x_i^*] + \lambda a_i x_i$$

subscription fee usage charge

Time-of-day pricing

- Single best-effort service differentiated by the time of day
- Two periods, $t = 1,2$ (peak, off-peak), I = set of users
- Utility functions $u^i(x_1^i, x_2^i), i \in I$
  - x_t^i is *amount* of data sent during period t
- C_t is the bandwidth available at period t (duration $= T$)
- **Global planner**: $\max_{\{x_t^i\}} \sum_I u^i(x_1^i, x_2^i)\ s.t. \sum_I x_t^i \le C_t T$
- Optimum $\{\hat{x}_t^i\}$ characterised by prices p_1, p_2 s.t.
- **User *i* solves**: $\max_{x,y} u^i(x, y) - p_1 x - p_2 y$

Congestion pricing for delay

- Delay-sensitive traffic: delay ⇨benefit reduction
  - when total traffic approaches link capacity
  - interesting for Internet
- Benefit of user *i* sending at rate x^i : $u^i(x^i) - \sigma^i d(x) x^i$
- Social planner solves:

$$\max_{\{x^i\}} \{\sum_i u^i(x^i) - d(C, \sum_i x^i) \sum_i \sigma^i x^i\}$$

- Optimum achieved for congestion price $p = \frac{\partial d}{\partial x} \sum_i \sigma^i x^i$
- User solves $\max_x \{u^i(x) - \sigma^i d^i x - px\}$

Summary of results (finite capacity)

Price differentiation based on service types

- monopoly: $\gamma = 1$
- Ramsey prices: $\gamma \neq 1$
- perfect competition: $\gamma = 0$

$$p_i\left(1+\gamma\frac{1}{\varepsilon_i}\right)=\lambda a_i$$

Time-of-day price differentiation

$$p_i^t(1+\gamma Y_i)=\lambda^t a_i$$

Consistency issues in pricing services

Price discrimination based on type of services might generate

- **arbitrage** (customers make profits by buying services of certain types, and then repackage and resale them as a different service, priced cheaper than the market value of the new service)
- **splitting** (split service into different sub-services, with a total cost being less than the cost of the original service)

If prices are proportional to effective bandwidths, then

- arbitrage is not possible
- splitting is encouraged since sub-services are assumed independent
- splitting can be avoided by adding a fixed cost per connection
- Social welfare is maximized

Cost-based pricing

- Tariffs must cover some notion of cost related to service provisioning
- Basic criterion is **feasibility** (not optimality)
  - prices are not unique
- Three independent criteria for characterizing feasibility
  1. Stability under entry and bypass
  2. Meeting a set of axioms for relating prices to costs
  3. Satisfying certain accounting principles (FDC, LRIC, etc.)

What is cost?

- Cost of a service = **value** of economic means used in order to provide the service => **Cost is a relative notion!**
  - Associate the use of equipment to services
  - definition of cost of equipment (historical vs current, net replacement cost, modern equivalent asset with abatements,etc.)
- Cost definition => different incentives
  - replacement of equipment, introduction of new technologies, encourage or deter entry, invest in sunk costs

Regulation

- information models
- price regulation
- competition
- unbundling

Regulation and information models

- Economic efficiency of an economy has many aspects
  - **Allocative efficiency**: SW maximization
  - **Productive efficiency:** non-static cost, needs effort by firms
  - Competitive markets achieve efficiency
- Market power (MP): reduces efficiency, possible market failure
- **Regulation mechanisms:**
  - incentives to firms with MP to adjust prices -> econ. efficiency
  - direct control on prices
  - indirect control (increase competition, incentives)
  - negative effects: drive suppl. surplus to zero, deter new entry
- Main difficulty for regulation: **asymetry of information**
  - private information increases profits

Principal-agent model

The **principal** (regulator) wants to induce the **agent** (firm) to take an action that is costly to him through **incentive payments**.

- the output is directly observed instead of the action
- the action defines the operating regime (cost structure)

$$\max_{s(.),a} \; x(a) - s(x(a)) \quad s.t.$$

participation

$$s(x(a)) - c(a) \geq 0, \quad \text{and}$$

Incentive compatibility

$$s(x(a)) - c(a) \geq s(x(b)) - c(b) \; \forall b \in A - a$$

A solution: offer $s(x) = x - F$ ← participation fee

Agent solves: $\max_a x(a) - F - c(a) \;\; s.t. \;\; x(a) - F - c(a) \geq 0$

=> economic efficiency, $F \leq x(a^*) - c(a^*) = F^*$

Price regulation

- Direct control of monopoly prices
- Regulator: specifies a set of constraints on prices = **price caps**
  - firm: free to choose any price in the set
  - social surplus increases
- Examples

$$\{p^t : \sum_i p_i^t q_i^{t-1} \le \sum_i p_i^{t-1} q_i^{t-1}\}$$

$$\{p^t : \sum_i p_i^t q_i^{t-1} \le c(q^{t-1})\}$$

$$\{p^t : \sum_i p_i^t q_i^{t-1} \le (1-X)\overline{p}\sum_i q_i^{t-1}\}$$

- the RPI-X mechanism

Price regulation (cont.)

- Dynamic regulation can be modeled as a game between the regulator and the firm
  - anticipation of future regulatory decisions influence policy and decisions of firms for current interval (investments, etc.)
  - a "confused" regulatory policy might have very negative effects
- Frequency of policy updates:
  - *low frequency*: not adapting to technology improvements, provides stability for optimal adaptation: **benefit from regulatory lags**
  - *fast frequency*: hard for firms to adapt, better incentives for introducing new technologies

Regulation and competition

- Advantages and disadvantages of monopoly:
  - best when large sunk costs are required, economies of scale
  - reduced allocative and distributive efficiency
- Is competition always profitable to the society?
  - **Cream skimming**: new entrants can target at most profitable part of the market, produce inefficient entry, make monopolist collapse
  - Excessive entry: increasing the number of firms may drive producer surplus to zero (even <0)
  - no economies of scale (increased marginal cost)
- Is regulation always needed?
  - **Contestable markets**: threat of ***hit-and-run*** entry keeps prices near marginal cost

Regulation and competition (cont.)

- Entry occurs when the competitor can produce services at lower prices than the incumbent, and still be profitable
  - **Efficient entry**: entrant provides service at lower total cost than the incumbent
  - **Inefficient entry**: service is priced cheaper, but costs more than when produced by the incumbent.
- Regulation might
  - bias for competition => encourage inefficient entry, negative PS
  - barrier to competition => prohibit efficient entry, reduce SW

Entry deterrence and unbundling

- Monopolist has strategic advantages due to his position
  - has the “first move” in the game
  - is in position to make **viable threats** to deter entry
  - can subsidize from other parts of the market
  - can prohibit access to bottleneck distribution networks
  - can bundle bottleneck services with other services
- Regulatory “medicine”: **unbundling**
  - force monopolist to offer services in stand-alone fashion
  - price unbundled services near actual cost
  - low prices might impede innovation by prohibiting deployment of new alternative technologies

Flat rate pricing

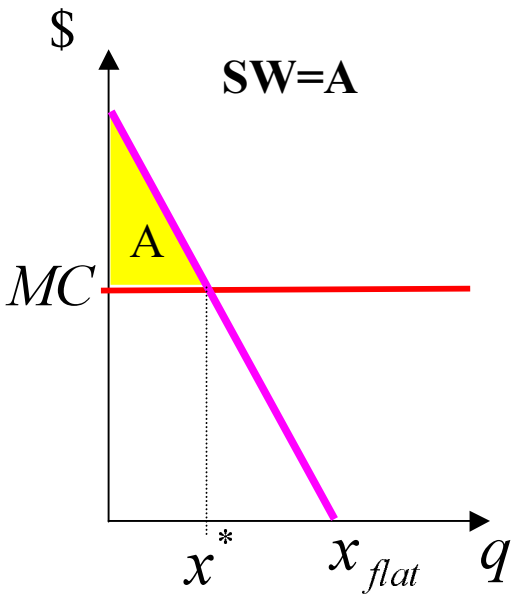
- waste
- stability
- quality differentiation

Flat rate pricing

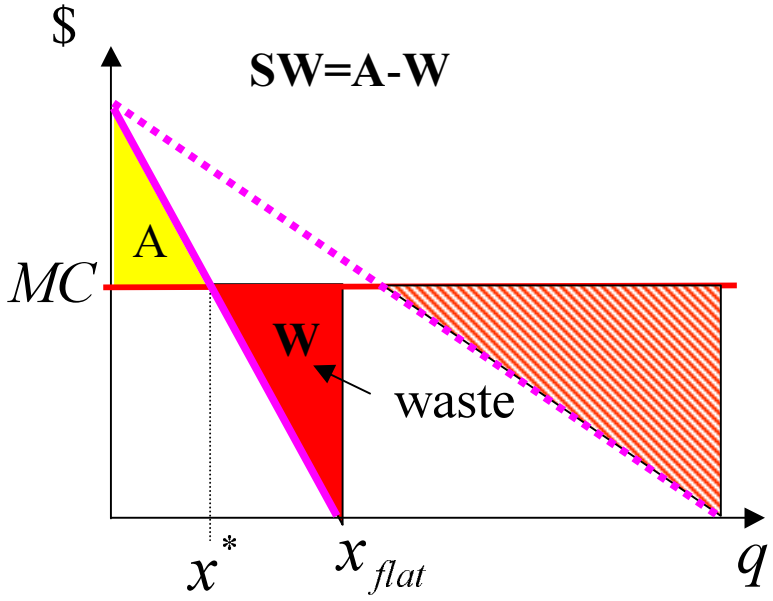
- Flat rate pricing is widely used because
  - easy to implement, **some** users like it
- Problems with flat rate:
  - high social cost (produces waste)
  - light users subsidize heavy users
  - unstable under competition
  - inefficient market segmentation
  - generates lower income for providers
  - lower benefit for most users (except the heavy ones)
  - recent experimental results for Internet pricing in INDEX experiment (UC Berkeley)

Flat rate pricing (cont.)

Assume network cost = $x\ MC$



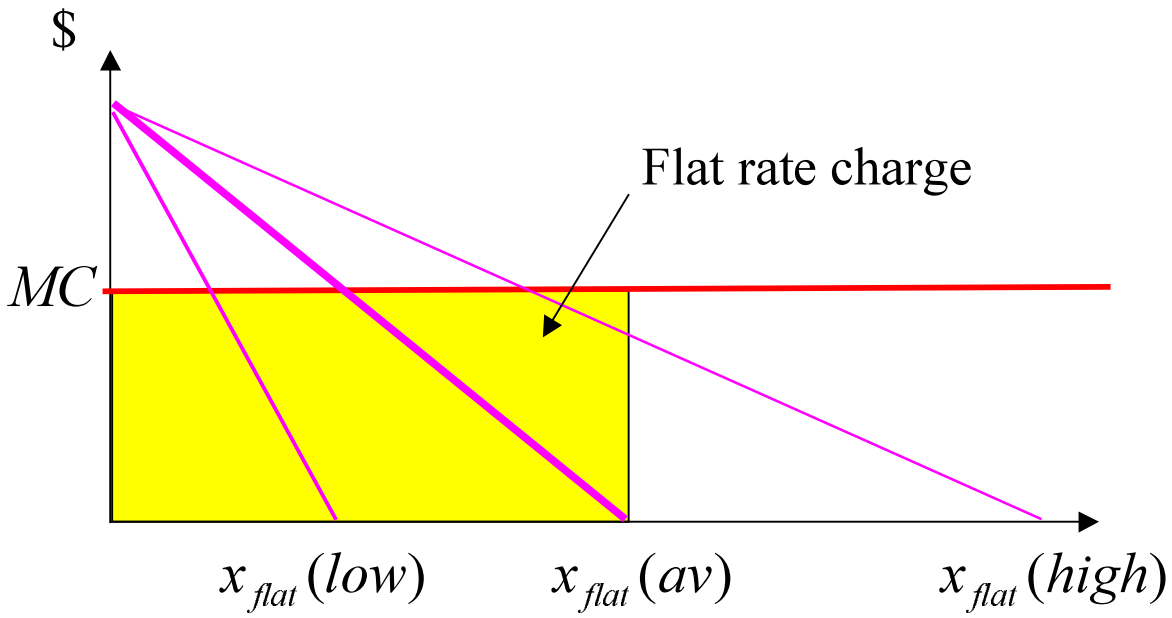
$p=MC$



Flat price ($p=0$)

Under flat pricing, social efficiency decreases

Cross-subsidization

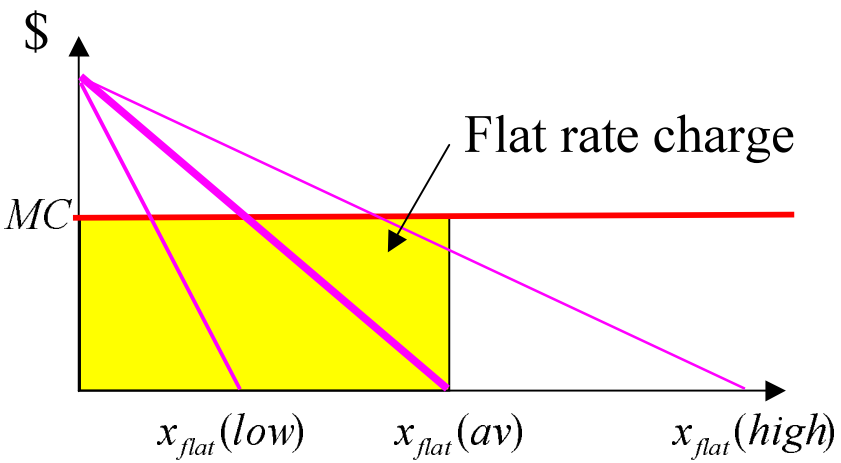


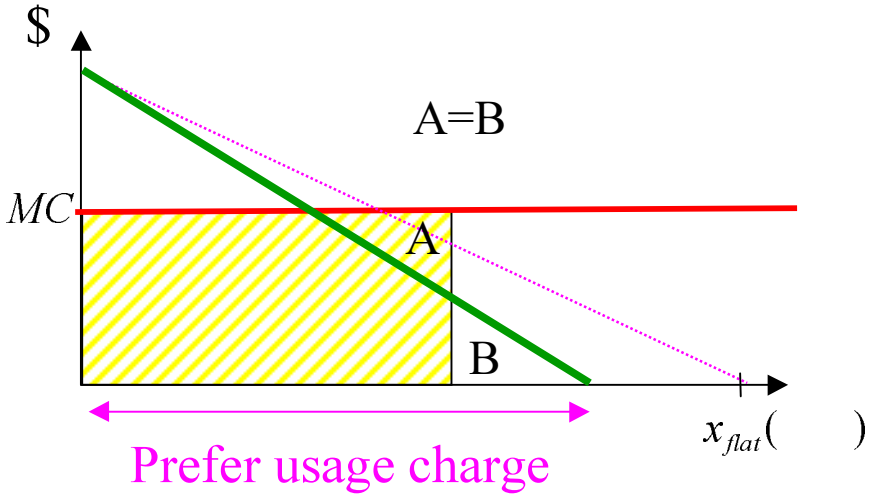
Low users will not participate => revenue + SW loss

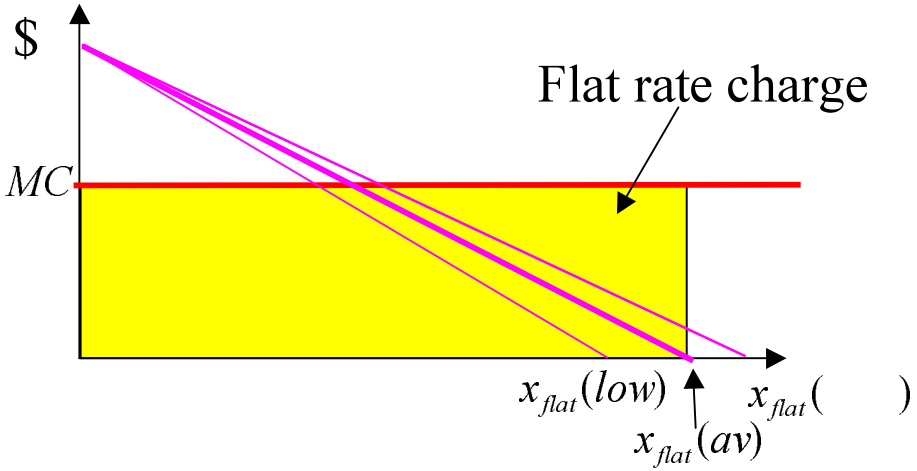
- decrease flat fee (=> losses or constrain usage)

Cross-subsidization (cont.)

Game: competitive provider with usage charge = MC

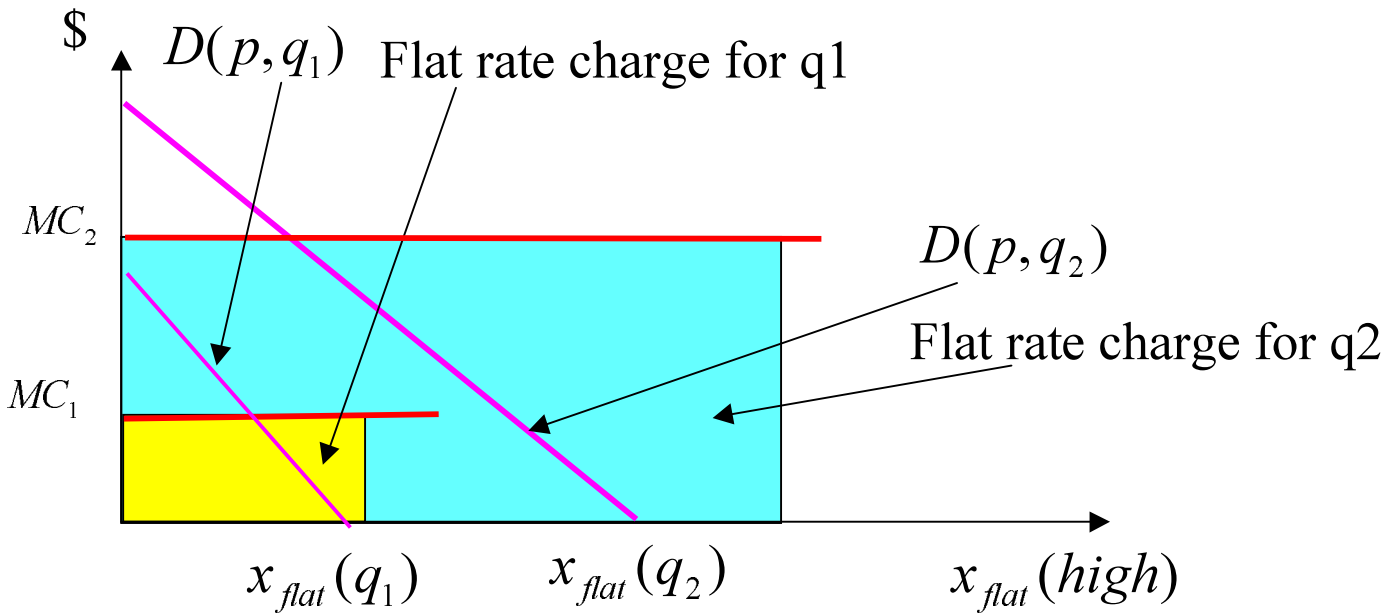






=> all customers prefer usage charge

Flat rate and tiered quality



Case of a user that can use both service qualities, is “low user” for q2:

- will subscribe to lower quality

=> high quality becomes even more expensive

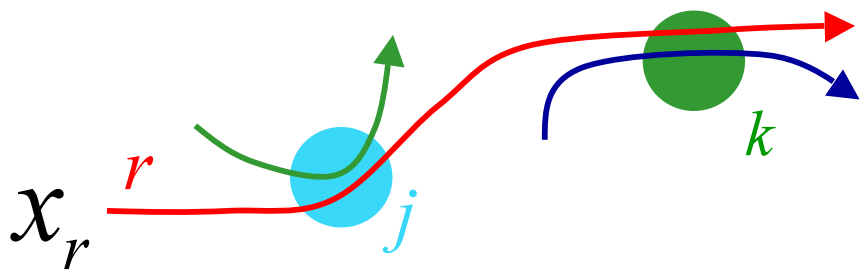
=> lose SW: could subscribe in both (Pareto improvement)

Charging elastic and best-effort services

Contents

- **The problem**
- **Congestion pricing**
- **Implementation issues**
- **Proportional fairness**
- **Relation to current Internet**
- **Conclusions**

Basic concepts



- **Elastic traffic**: flexible contract with network
  - no guarantees on delay, throughput, CLP
  - examples: TCP/IP, ABR
  - Sources of randomness:
    - number of users + amount of data
    - amount of available resources

Basic concepts (cont.)

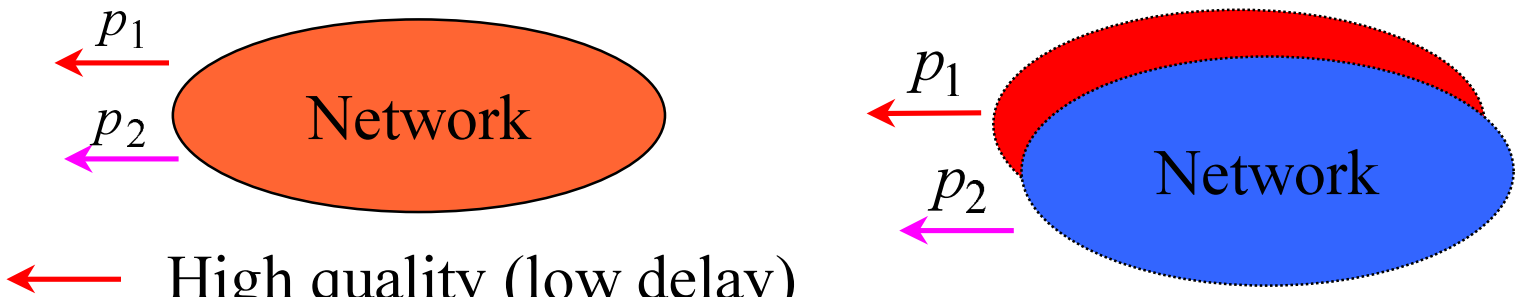
- Need for flow control
- Various notions of fairness (max-min, etc.)
  - **not economically efficient !(no consideration of demand)**
- The goal: **Provide optimal economic sharing of resources based on demand**
- **Two time scales:**
  - *fast time scale*: flow control, adjustment of prices
    - maintains feasibility of flows
    - congestion prices define sharing
  - *slow time scale*: adjustment of demand
    - provides optimal sharing

Basic concepts (cont.)

- Common approach: **congestion pricing**
  - 1. **reduce excess demand**
  - 2. **account for congestion costs**
- Prices can be
  - *computed dynamically*
  - *approximated from historical data* (time-of-day prices)
- Important issues
  - cost of computing prices
  - cost of exchanging info with users
  - stability of price mechanism
  - expression of user preferences

Approaches

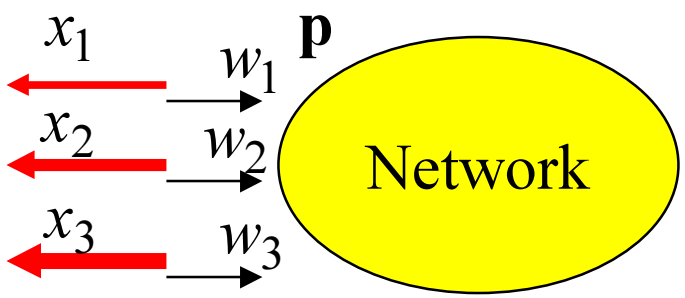
Quality differentiation of services



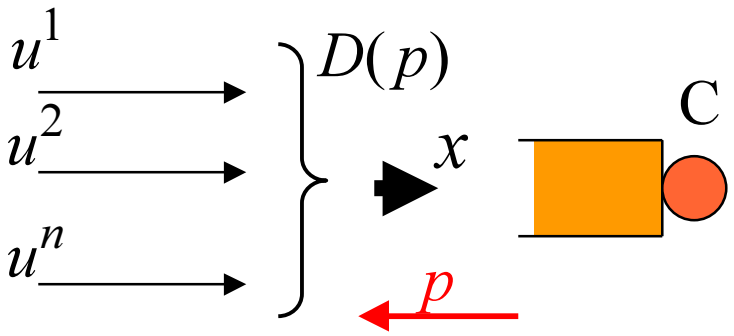
High quality (low delay)

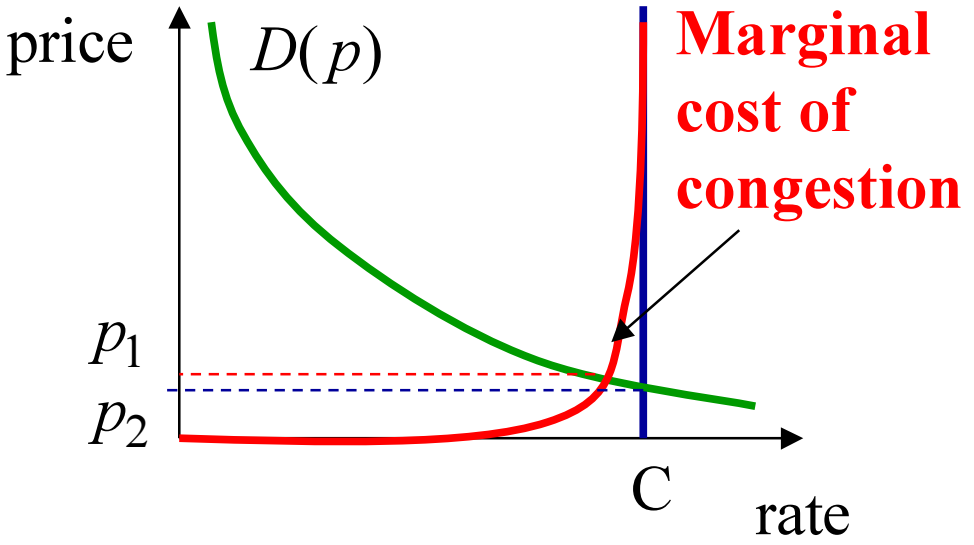
Low quality (higher delay)

Quantity differentiation



Congestion pricing



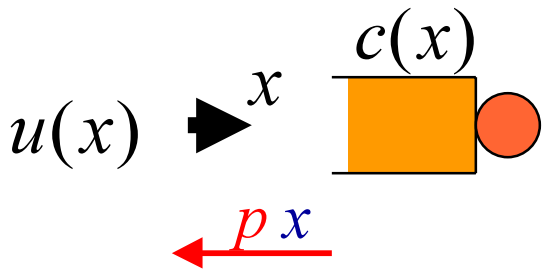


$$\max_x \; u(x) - c(x) \; s.t. \; x \le C$$

Congestion price: marginal utility = marginal congestion cost
Note: ***almost*** **demand = capacity**

Computing congestion prices

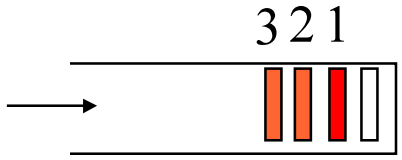
1. Congestion charge is computed on an average basis



$$p = \frac{\partial c}{\partial x}$$

x = average flow
c = average cost

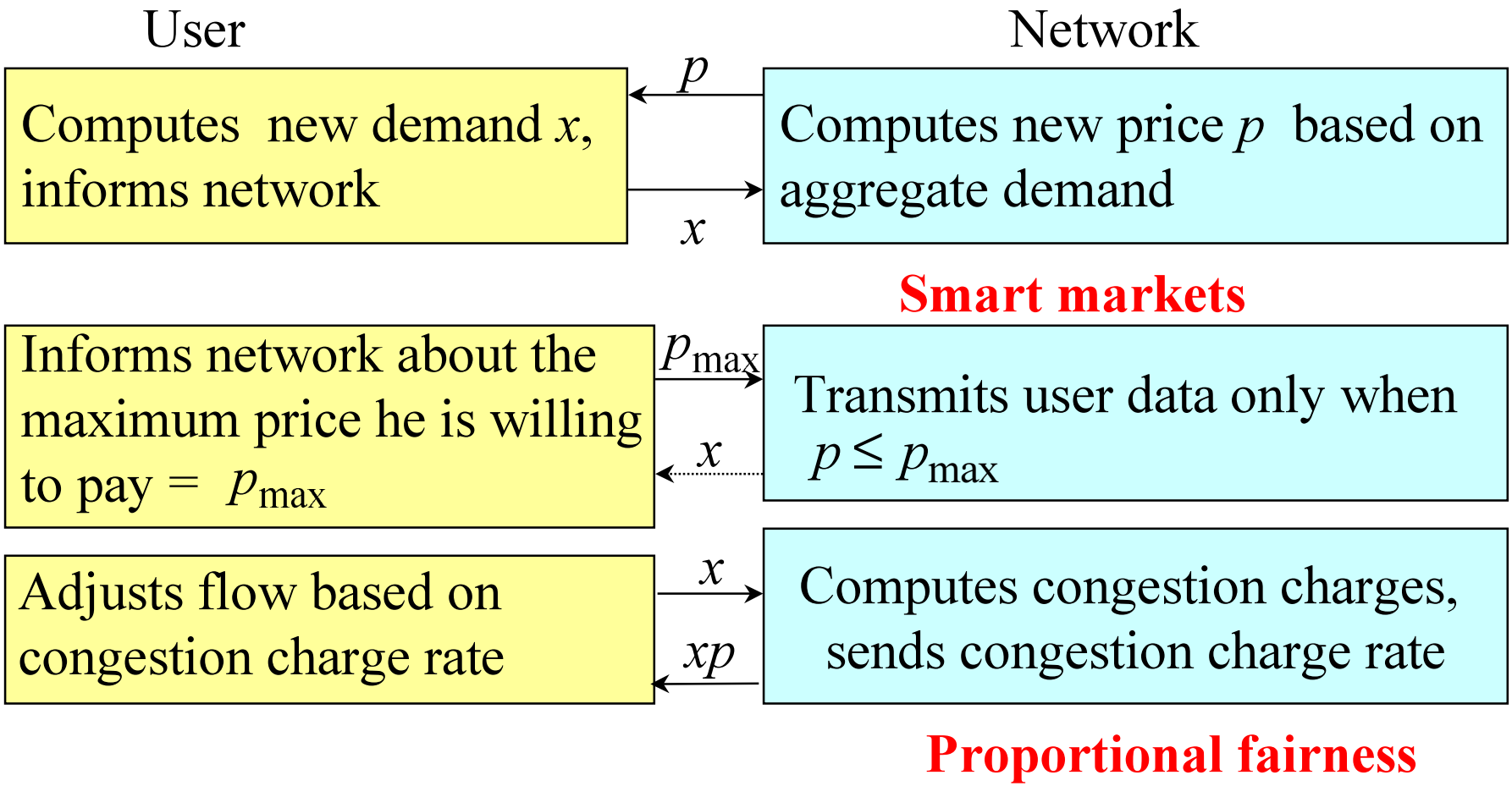
2. Congestion charge is computed per sample path:
Each packet is charged the cost increment that it causes



Packet 1 is charged the **extra** waiting cost it causes to 2,3

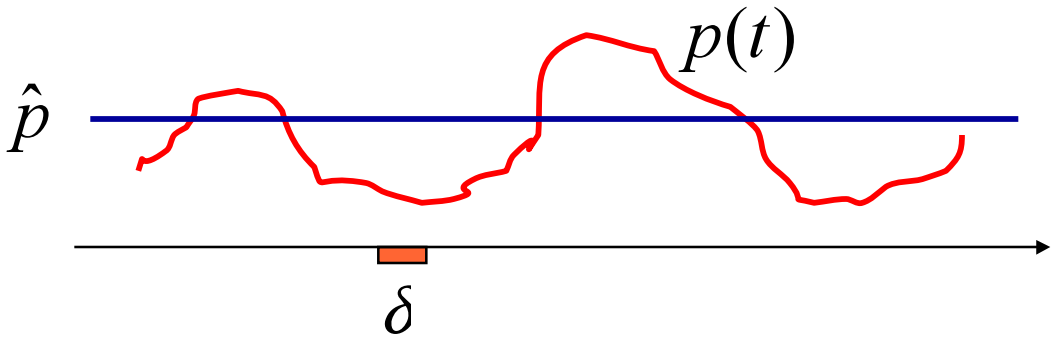
The rate of charge *px* is averaged on the particular sample path

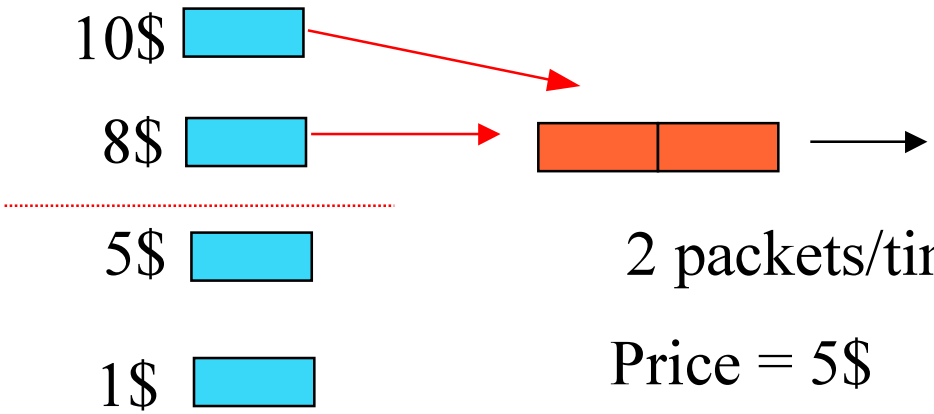
Implementation approaches



Smart markets

Smart markets:
bandwidth auctions



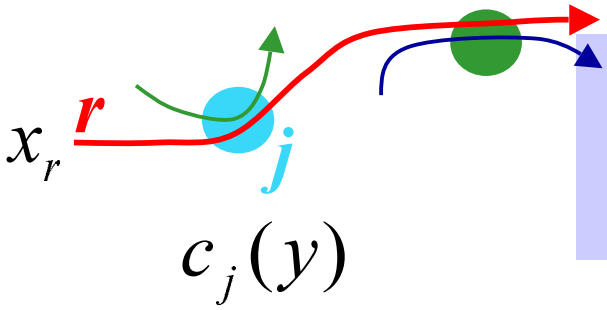


2 packets/time slot

Price = 5$

Congestion price = shadow cost of capacity constraint

Proportional fairness



$$\max_x U(x) = \sum_{r \in R} u_r(x_r) - \sum_j c_j\left(\sum_{s:\, j \in r} x_s\right)$$

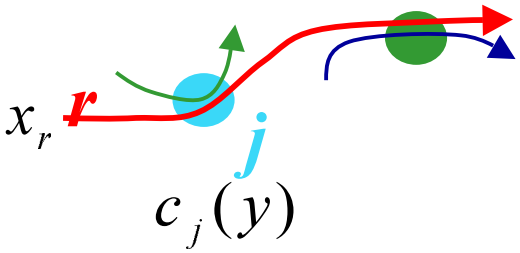
Congestion price at resource j : $p_j(y) = \frac{d}{dy} c_j(y)$

Total congestion price for route r : $\mu_r(t) = \sum_{j \in r} p_j(y_j),\ y_j = \sum_{s:\, j \in s} x_s(t)$

Theorem1 (Proportional fairness):

If $\frac{d}{dt} x_r(t) = k\left(w_r - x_r(t)\mu_r(t)\right)$, then there is a unique stable point $\{x_r\}$ to which all trajectories converge, and $x_r = \frac{w_r}{\mu_r}$

Proportional fairness (cont.)



$$\max_x U(x) = \sum_{r \in R} u_r(x_r) - \sum_j c_j\Big(\sum_{s:\, j \in r} x_s\Big)$$

$$p_j(y) = \frac{d}{dy} c_j(y), \quad \mu_r(t) = \sum_{j \in r} p_j(y_j), \quad y_j = \sum_{s:\, j \in s} x_s(t)$$

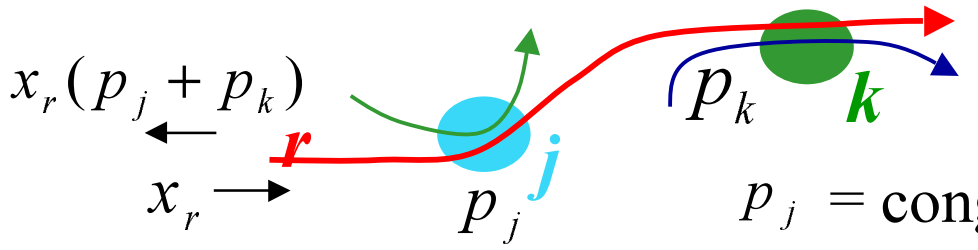
Fast time scales t : Proportional fairness

$$\frac{d}{dt} x_r(t) = k\left(w_r(\tau) - x_r(t)\mu_r(t)\right) \quad \rightarrow x_r(\tau) = \frac{w_r(\tau)}{\mu_r}$$

Theorem2 (social welfare optimality):

If $w_r(\tau+1) = x_r(\tau)u'_r(x_r(\tau))$, then $\{x_r(\tau)\}$ converges to the social welfare optimum

Proportional fairness: implementation



p_j = congestion price of resource j

$$\dot{x}_r = k[w_r - x_r(p_j + p_k)]$$

willingness to pay \$/s

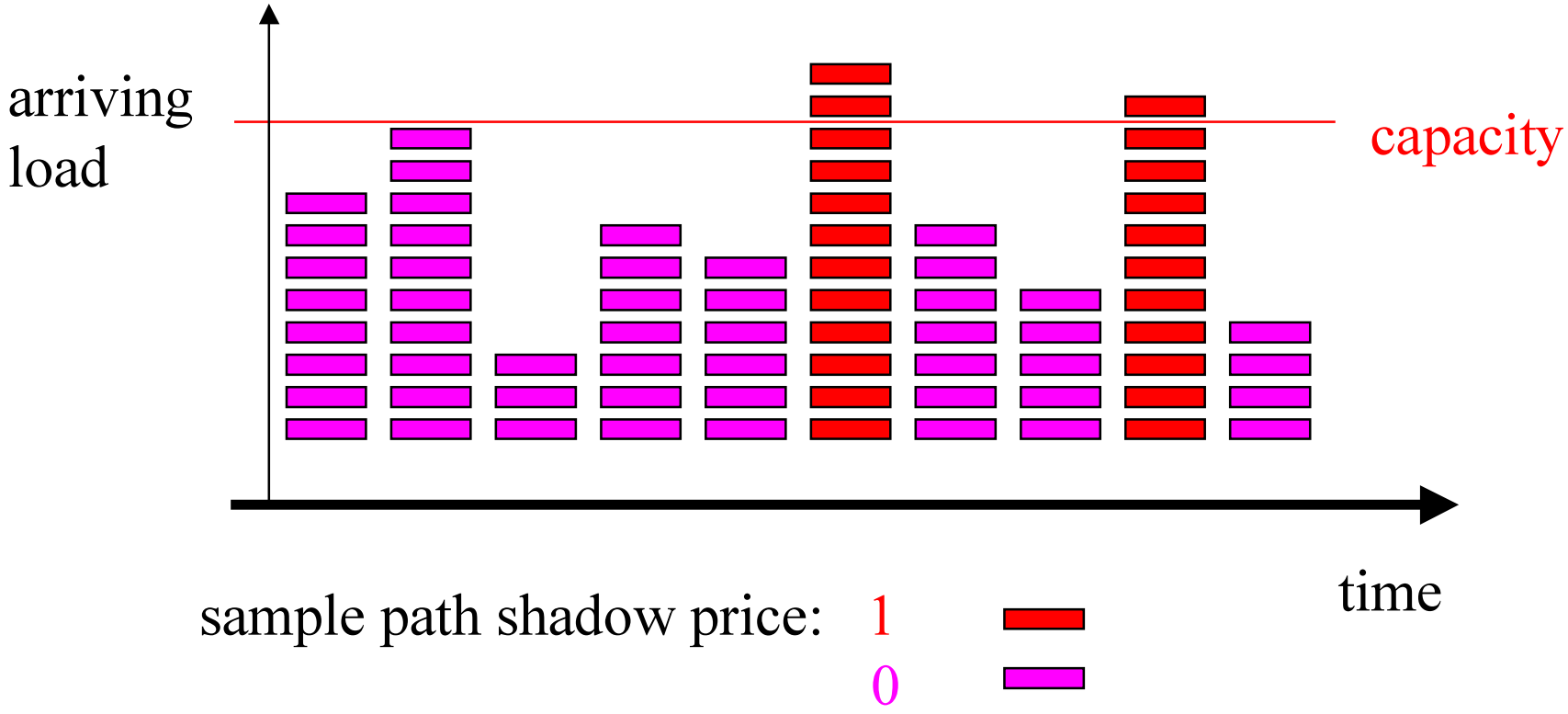
current rate of congestion charge \$/s

Question: is there a technology-sensible way to implement this?

Proposal: use as $c_j(y)$ the **rate of losing packets** at the resource j

Construction of the charge: charge one unit for each packet that if removed, one less packet would be lost.

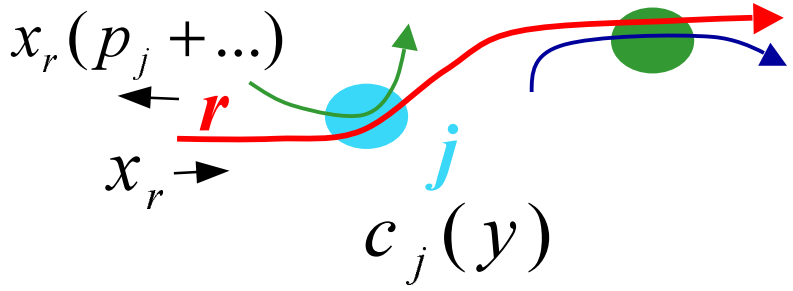
Sample path shadow prices



Proportional fairness: implementation (cont.)

Implementation: "bad" packet = packets that contribute to losses

- for each "bad" packet, send a charging mark back (ECN bit)
- translate charging mark into money depending on desired QoS



$$c_j(y) = d \times \text{loss rate}$$

d = cost of loosing a packet
= monetary value of a mark

$$p_j = d \times \text{proportion of lost packets}$$

Note: in general $x_r p > x_r p_{loss}$
(for $d = 1$)

=> losses are not good signals for congestion cost!

$$\rho = 1 - 1/\sqrt{N}$$

$$p \approx 1/2 \quad p_{loss} \approx a/\sqrt{N}$$

Comparing with current Internet

$$\dot{x}_r = k[w_r - x_r(p_j + p_k)]$$

Proportional fairness: additive increase, multiplicative decrease

$x \approx 1/p$

Internet semantics: p = packet loss probability

Internet (TCP): additive increase, (multiplicative decrease)^2
- rate of congestion signals + halving the rate

$x \approx 1/\sqrt{p}$

Important issue: use existing router technology to implement marking

Evolution of proportional fairness

- compatible with TCP, RED and ECN
- allows marking and flow control strategies to evolve
- able to support arbitrarily differentiated services, defined by users
- no need for large buffers or multiple priorities within network, or for CAC at edge

Possible user control algorithms

If proportional fairness marking is implemented, the intelligence is at the user end: its up to the user to respond
Some characteristic user-response algorithms:

1. Elastic user: has fixed w, transmits $X(t) = \lfloor x(t) + z(t) \rfloor^+$ where

$$z(t+1) = x(t) + z(t) - X(t)$$
$$x(t+1) = x(t) + k(w - f(t))$$

$f(t) =$ # of marks received in slot (t,t+1)

2. File transfer: given initial budget W, file size F

$$w(t+1) = \max\{x(t)W(t)/F(t), 0.01\}$$

Conclusions

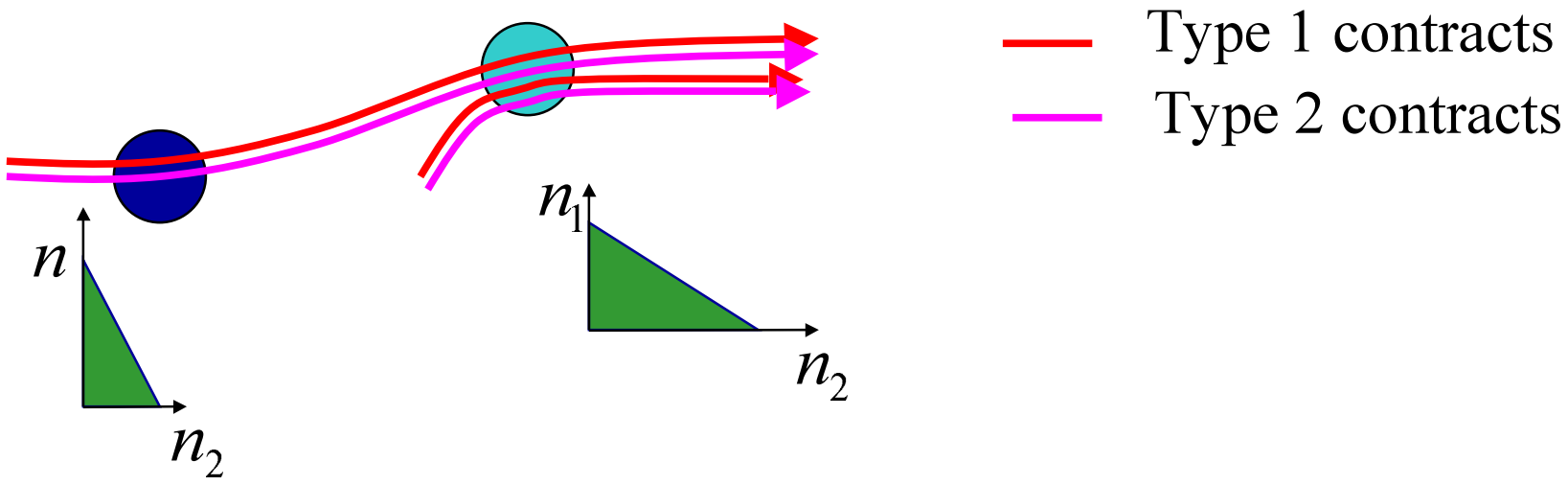
- Large users can anticipate their effect on congestion prices
  - many possible game formulation
  - potential decrease of social optimality
- Congestion pricing sensible also for monopolists
- Exploit existing router technologies for implementing marking
- Combine with existing Internet QoS architectures (diff serv)
- Integrate with traffic management in MPLS

Charging Guaranteed Services

Contents

- Effective Bandwidths and charging
- Time- and Volume-based charging
  - Simple Charging Scheme
  - Properties and Incentives
  - Examples
  - Simplifications
- CBR Charging

The problem



Constraints: CAC at each node,
uses some notion of effective bandwidth

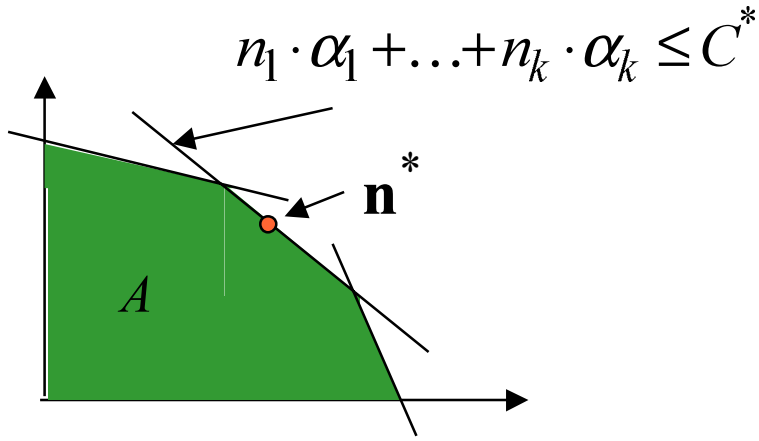
Problem: how to price different traffic contract types
Economic context: maximize social welfare, perfect competition

Economic theory reminder

- Social welfare maximisation:

$$\max_{\mathbf{n}} U(\mathbf{n})$$

$$s.t.\ \mathbf{n} \in A$$



- **Prices defined by shadow costs of effective bandwidth constraints**

$$p_i = \lambda eb_i, \qquad \frac{p_i}{p_j} = \frac{eb_i}{eb_j}$$

- **Which is the right effective bandwidth definition?**

Which is the right effective bandwidth?

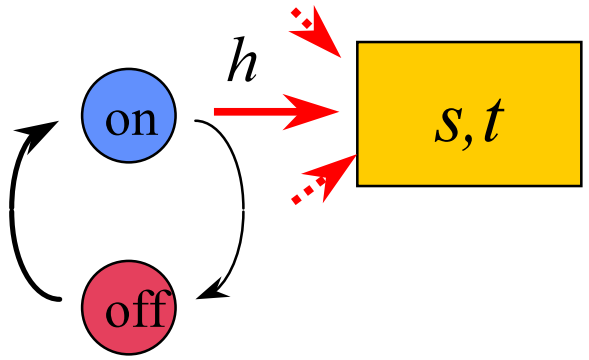
- **Which is the appropriate effective bandwidth definition?**
- **Criteria:**
  - **consistent with CAC**
  - **incentive compatibility (fairness, accuracy)**
- **the actual eb?**
  - **can not be used for CAC, good incentives**
- **the typical eb?**
  - **can not be used for CAC, bad incentives**
- **the worst possible eb?**
  - **good for CAC, bad incentives, unfair**
- **the worst possible eb given the measurements?**
  - **can not be used for CAC, good incentives**
  - **can we make the user reveal the measurement info before the actual measurements?**

The time-volume scheme

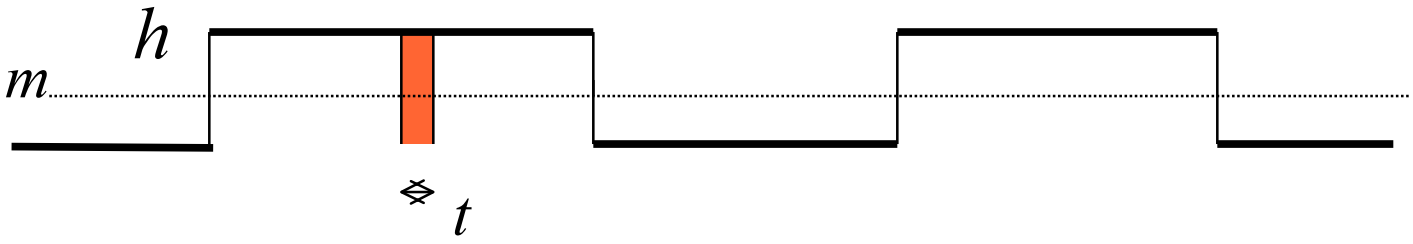
- Network post a set of tariffs of the form $a(...)T + b(...)V + c(...)$
  - T = duration of call (e.g. seconds)
  - V = volume of call (e.g. Mbits or Mcells)
  - $a(...)$, $b(...)$, $c(...)$ capture SLA choices (peak rate, QoS, etc)
- User chooses particular $< a, b, c >$
- Total charge for a call is $aT + bV + c = T(a + bm) + c$
- **Can we use such a scheme to charge for effective bandwidths?**
- **Can we make the user reveal his mean rate *m* ?**

The worst eb given time and volume

- Traffic contract = h
- Measurements = m
- Operating point of multiplexer = s,t



Worst case traffic: slow on-off



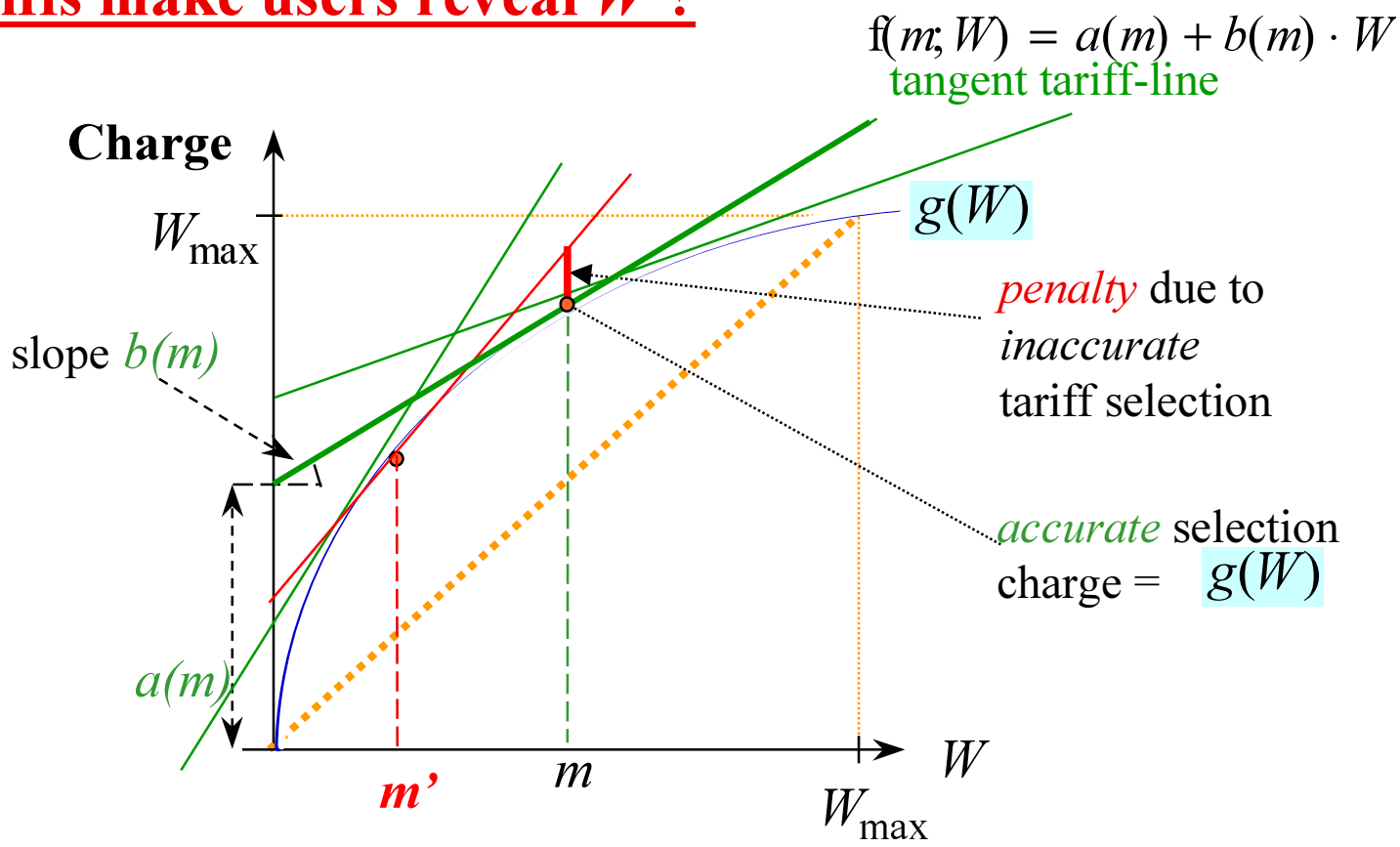
Effective bandwidth = $$\alpha_{on/off}(s,t) = \frac{1}{st}\log\left[1 + \frac{m}{h}\left(e^{sth} - 1\right)\right]$$

Can we make the user reveal his information about m at connection setup?

An interesting structural property for tariffs

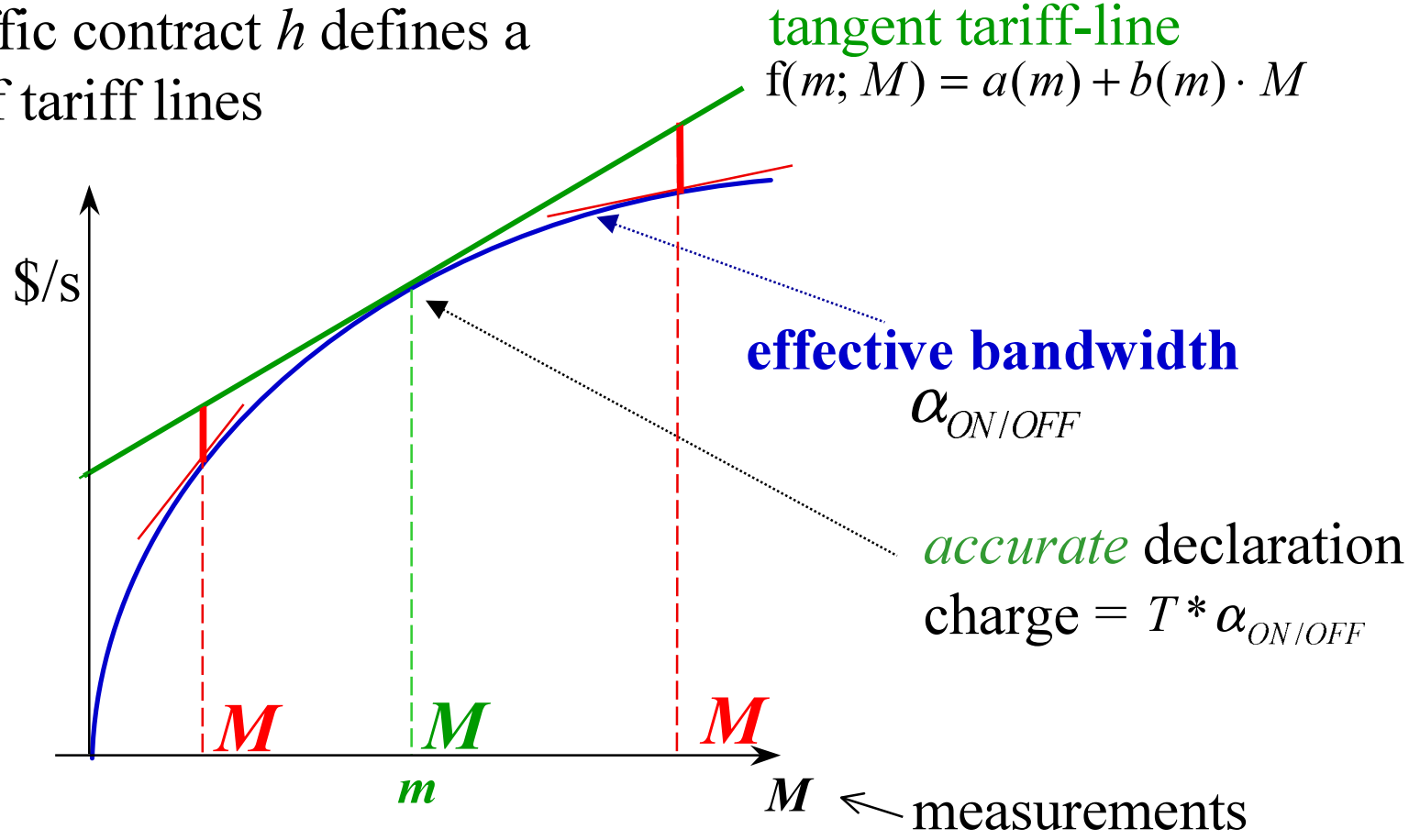
- Customers arrive to eat
  - a priori info: $W_{\max}$
  - a posterior info: W
- **Can tariffs make users reveal W ?**

charge $g(W)$ = concave function of amount eaten



Simple Charging Scheme for VBR

- Each traffic contract *h* defines a **family** of tariff lines



1. User chooses tariff <=> **declares *m***
2. Final charge = $T[a(m)+b(m)M]$

Properties of Simple Charging Scheme

- Total charge = $T \cdot [a(m) + b(m) \cdot M]$

  $= a(m) \cdot T + b(m) \cdot V$

- Accounts both for
  - resource reservation => time-component
  - actual usage => volume-component
- Simple Accounting
  - Requires only *simple* measurements: T and V
- **Flexibility added to traffic contracts**
- Rational users pay in proportion to their effective use
  - Tariff coefficients depend on traffic contract parameters

Examples of Tariffs

h = 3 Mbps	st = 1 sec/Mbit	
M	*a(m)*	*b(m)*
0.20 Mbps	0.26	2.80
0.75 Mbps	0.93	1.10
1.50 Mbps	1.46	0.60
2.25 Mbps	1.81	0.41
2.80 Mbps	1.98	0.34

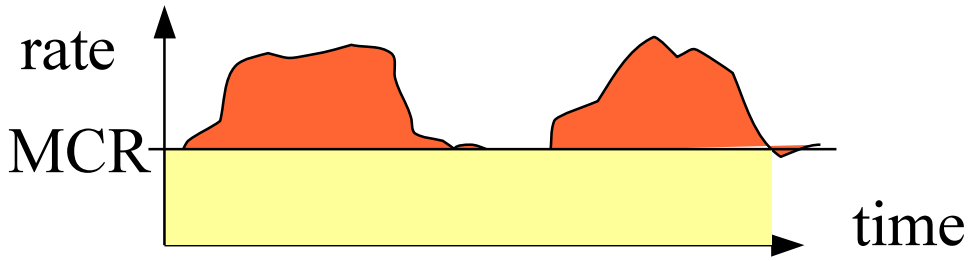
h = 1.5 Mbps	st = 1 sec/Mbit	
M	*a(m)*	*b(m)*
0.20 Mbps	0.06	1.59
0.75 Mbps	0.37	0.85
1.50 Mbps	0.72	0.52

***a(m)* => $/sec**

***b(m)* => $/Mbit**

h = 3 Mbps	st = 2 sec/Mbit	
M	*a(m)*	*b(m)*
0.20 Mbps	1.18	2.41
0.75 Mbps	1.82	0.66
1.50 Mbps	2.16	0.33
2.25 Mbps	2.36	0.22
2.80 Mbps	2.46	0.18

Simple Time-Volume scheme for ABR



- $aT + bV + c$ can be used for ABR
- User buys an amount m of MCR at posted price p_{MCR}
- Network charges for a period of usage T
  - $p_{MCR} \times m \times T$ for the data sent within MCR
  - $p_{UBR} \times V$ where V is the volume sent **on top** of MCR
  - c (for signalling congestion, discourage splitting)
- No incentive for splitting connections if excess capacity allocated proportionally to the amounts of MCR

Conclusions

- Charging for effective usage can be made simple
- Charging for time and volume is adequate
- Incentive compatibility is an important issue
- Interesting relation with CAC
- Can be further simplified

Final remarks

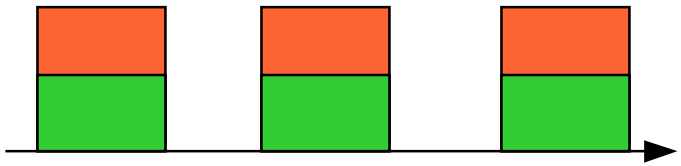
- There is important theory for constructing charges
- Charging can be a mechanism of control
- Competition will motivate the use of sophisticated charging
- Simple charging rules can result from sophisticated models
- There is no single best choice in charging
- Charging the Internet: bad (or the absence of) charging can impede the deployment of services (same for ATM)
- Usage based charging is definitely feasible and it will gain acceptance in the near future

More topics

Splitting of traffic

peak rate = h mean rate = m

effective bandwidth = a_{total}

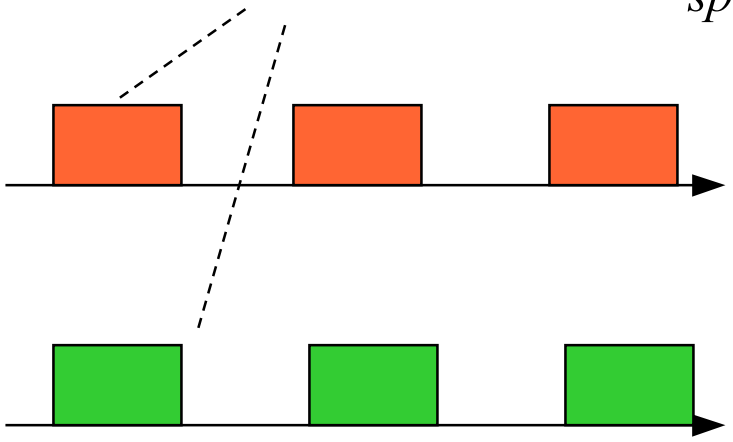


VC 1

VC 2

peak rate = $h/2$, mean rate = $m/2$

effective bandwidth = a_{split}



- Splitting can be *beneficial* to the user => possibly less total charge, because

$$2a_{split} < a_{total}$$

- correlated traffic streams are erroneously charged as independent ones

Discouraging splitting - fixed charge

- Traffic splitting is undesirable to provider, because:
  - may lead to reduced revenue
  - set of available VPI/VCI may be exhausted
  - increased signalling overhead for setting more VCs

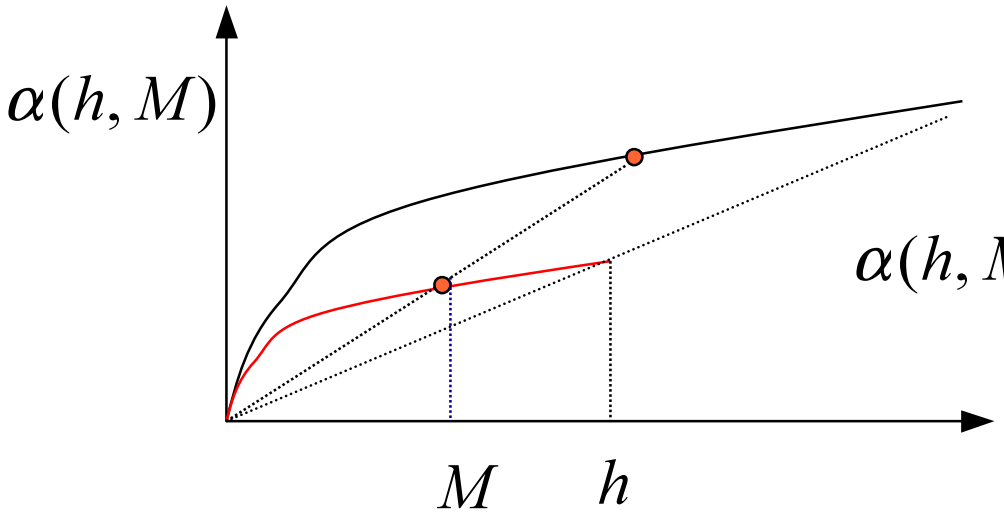
➡ **Splitting should be discouraged => add a *fixed* charge per VC**

➡ Total Charge = $a(m) \cdot T + b(m) \cdot V + c(m)$

- However, traffic splitting could be beneficial to provider, if substreams can only be accommodated through *different routes*

Discouraging splitting of traffic (cont.)

- Use **homothetic** tariffs

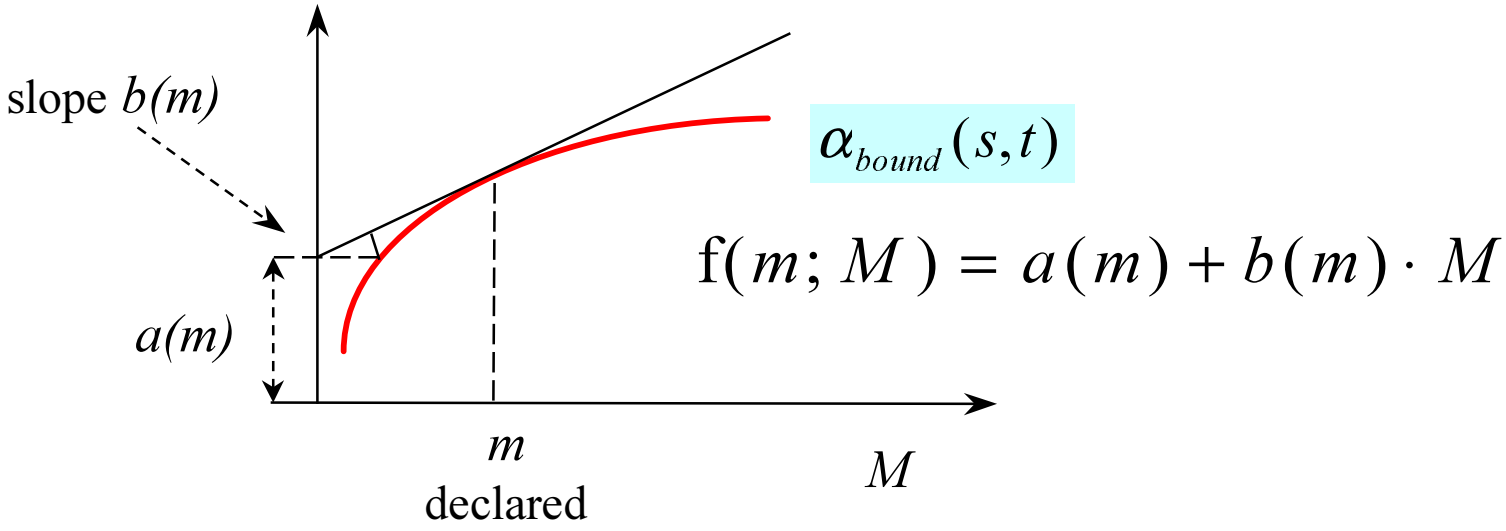


$$\alpha(h, M) = k\alpha(h/k, M/k)$$

- **Pros:** convexity makes users reveal their mean rates, no incentive to split
- **Cons:** charge not proportional to eb (but close!)

Improving accuracy of Simple Charging Scheme

- The simple charging scheme bounds the effective bandwidth according to the ON/OFF bound
  - does not capture general traffic contracts for VBR
- *Other* bounds can also be used
  - **functions of mean rate *and the LBs* of the traffic contract**
  - *Same* approach: charge per unit time derived according to the *tangent* selected by the user



Taking into account leaky bucket constraints

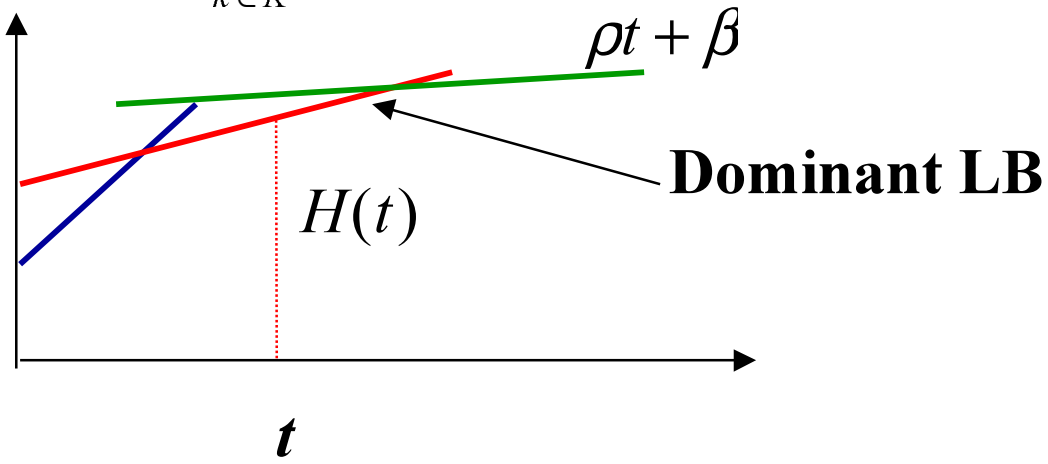
- ON/OFF bound corresponds to a single leaky bucket, constraining **only** the peak rate

$$\alpha_{on/off}(s,t) = \frac{1}{st}\log\left[1+\frac{m}{h}\left(e^{sth}-1\right)\right]$$

- For traffic contracts involving multiple leaky buckets, we can use the *tighter* bound

$$\alpha_{lb}(s,t) = \frac{1}{st}\log\left[1+\frac{tm}{H(t)}\left(e^{sH(t)}-1\right)\right]$$

where $H(t) := \min_{k\in K}\{\rho_k t + \beta_k\}$



More general charging schemes

- Simple scheme can not distinguish users having the same mean

- Need for more detailed traffic measurements
- Consider the general linear tariff

$$f(X) = a_0 + a_1 g_1(X) + \ldots + a_L g_L(X)$$

  - $X = X_1, \ldots, X_T$, $g_i(X) =$ measurement function$(= \frac{1}{T} \sum_{j=1,T} X_j)$

- Can we construct such functions that charge for **effective usage?**
  - Evaluate implementation cost vs accuracy gain

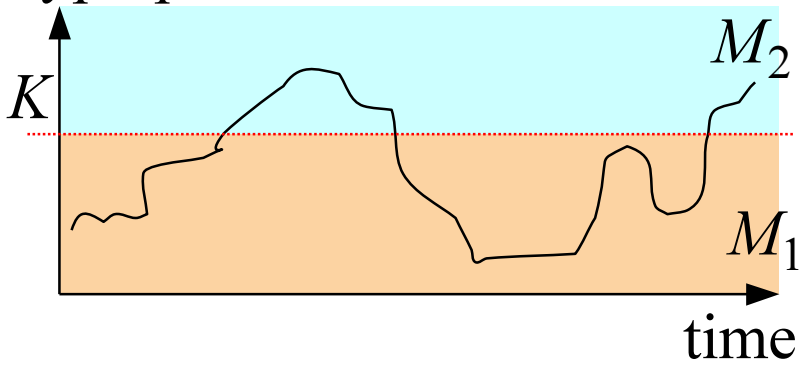
More general charging schemes (cont.)

- Approach used in Simple Charging Scheme can be extended
  - Define the effective bandwidth to be the function

$$\alpha(h, M) = \sup_{X_t}\left\{\frac{1}{st}\log Ee^{sX[0,t]}\right\}$$

$$s.t.\ Eg(X) = M,\ X_t \in \Xi(h)$$

  - concave in M
- Construct linear tariffs = tangent hyperplanes to $\alpha(M)$
- Example: the *2-tax band* scheme



Simpler Charging: Dispensing with Duration

- **The time-component of charge can be eliminated**
  - total charge = $b \cdot V + c$
  - tariff will be simpler
  - dependence of usage-charge on QoS will be clearer
- **Reasoning:**
  - c can be set to account for *typical time-charge*, or
  - we can assume a typical value for m and infer $T \approx V / m$, hence

$$a(m) \cdot T + b(m) \cdot V + c(m) \approx a(m) \cdot (V / m) + b(m) \cdot V + c(m) = b'(m) \cdot V + c(m)$$

- **However, users will have *no* incentive to close connections**
  - set of available VPI/VCI may be exhausted
  - provider can limit the maximum number of VPI/VCIs permissible per user

Charging CBR Services

- **Simple charging scheme can also be applied to CBR services**
  - users should declare $m = h$
  - Total Charge = $a(h) \cdot T + b(h) \cdot V + c(h)$
  - Volume-charge does *not* vanish, because $b(h) \neq 0$

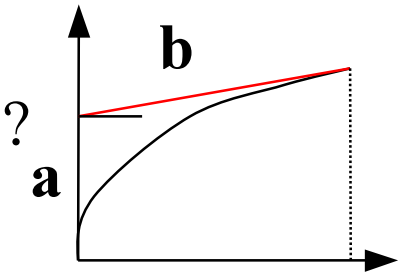
- CBR services should be charged *only* on the basis of time, if their peak rate is really *reserved*, and CBR is *not* multiplexed statistically
  - simpler scheme
  - already adopted in practice

Charging PVCs

- So far have only dealt with *Switched* VCs for VBR services (SVCs)
- Simple charging scheme can also be applied to *Permanent* VCs (PVCs) for VBR services
- However, PVCs can also be charged *only* on the basis of time, if they are *not* multiplexed statistically, due to their long duration
  - simpler scheme
  - already adopted in practice

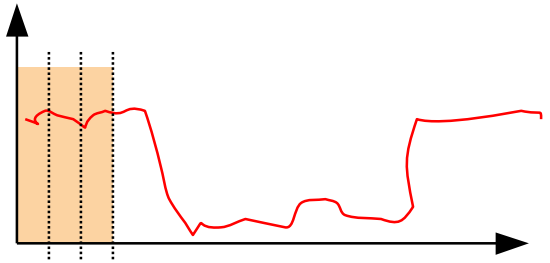
Charging and CAC

- **Consistency of CAC and charging function:**
  - Natural to charge with the eb used in CAC
  - Suppose CAC according to PCR. How to charge?
    - Not a competitive CAC
    - Better provide incentives to reduce volume



$$aT + bV + c, \quad a >> b$$

- Suppose "perfect" **dynamic** CAC is used
  - call arrival and departures occur every T
  - control mechanism (by blocking calls) achieves QoS at all times
  - ⇨ effective bandwidth of a call



= average of **actual** effective bandwidth in each period T

= *almost the mean rate*!!

URLs

- CA$hMAN: Charging and Accounting Schemes in Mutliservice ATM Networks. ACTS Project AC-039. URL: http://www.isoft.intranet.gr/cashman/
- IETF's Differential Service for the Internet working group. URL: http://diffserv.lcs.mit.edu/
- INDEX: The Internet Demand Experiment. Department of EECS, University of California, Berkeley. URL: http://www.INDEX.berkeley.edu
- Frank Kelly's Proportional Fairness page. URL: http://www.statslab.cam.ac.uk/~frank/pf/
- Hal Varian's The Information Economy. URL: http://www.sims.berkeley.edu/resources/infoecon/index.html
- Telecommunications and Networks Division, Institute of Computer Science (ICS), Foundation for Research and Technology (FORTH). URL: http://www.ics.forth.gr/netgroup/

Closely related references

- D. Botvich and N. Duffield. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. Queueing Systems, 20, 1995.
- C. Courcoubetis, F.P. Kelly, V.A. Siris, and R. Weber. A study of simple usage-based charging schemes for broadband networks. Accepted under revisions by Telecommunication Systems. URL: http://www.ics.forth.gr/netgroup/publications/
- C. Courcoubetis, F.P. Kelly, and R. Weber. Measurement-based charging in communications networks. Technical report 1997-19, Statistical Laboratory, University of Cambridge, 1997. To appear in Operations Research. URL: http://www.statslab.cam.ac.uk/Reports/1997/1997-19.html
- C. Courcoubetis, V.A. Siris, and G.D. Stamoulis. Application of the many sources asymptotic and effective bandwidths to traffic engineering. To appear in Telecommunication Systems. A shorter version appeared in ACM SIGMETRICS'98. URL: http://www.ics.forth.gr/netgroup/publications/
- C. Courcoubetis, V.A. Siris. Managing and pricing service level agreements for differentiated services. In 7th IEEE/IFIP IWQoS'99, UCL, London, UK. URL: http://www.ics.forth.gr/netgroup/publications/

Closely related references

- C. Courcoubetis, V.A. Siris. An approach to pricing and resource sharing for Available Bit Rate (ABR) services. In IEEE Globecom'98, Sydney, Australia, November 1998. URL: http://www.ics.forth.gr/netgroup/publications/
- C. Courcoubetis, V.A. Siris, and G.D. Stamoulis. Integration of pricing and flow control for Available Bit Rate services in ATM networks. In Proc. of IEEE GLOBECOM'96, London, UK, 1996. URL: http://www.ics.forth.gr/netgroup/publications/
- C. Courcoubetis and R. Weber. Buffer overflow asymptotics for a switch handling many traffic sources. Journal of Applied Prob., 33, 1996. URL: http://www.ics.forth.gr/netgroup/publications/
- G. de Veciana, C. Courcoubetis, and J. Walrand. Decoupling bandwidths for networks: a decomposition approach to resource management for networks. In Proc. IEEE INFOCOM'94.
- R.J. Edell, N. Mckeown, and P.P. Varaiya. Billing users and pricing for TCP. IEEE J. Select. Areas in Commun., September 1995.

Closely related references (cont.)

- A. Elwalid and D. Mitra. Effective bandwidths of general Markovian traffic sources and admission control of high speed networks. IEEE/ACM Trans. on Networking, October 1993.
- R.J. Gibbens. Traffic characterization and effective bandwidths for broadband network traces. In F.P. Kelly, S. Zachary, and I. Zeidins editors, Stochastic Networks: Theory and Applications, Oxford University Press, 1996.
- F.P. Kelly. On tariffs, policing and admission control for multiservice networks. Operations Research Letters, 15, 1994
- F.P. Kelly. Notes on effective bandwidths. In F.P. Kelly, S. Zachary, and I. Zeidins editors, Stochastic Networks: Theory and Applications, Oxford University Press, 1996.
- R.J. Gibbens and F.P. Kelly. Distributed connection acceptance control for a connectionless network. In ITC16, Edinburgh, UK, 1999.
- R.J. Gibbens and F.P. Kelly. Resource pricing and the evolution of congestion control. Automatica (1999).
- F.P. Kelly. Charging and accounting for bursty connections. In L.W. McKnight and J.P. Bailey, editors, Internet Economics. MIT Press, 1996.

Closely related references (cont.)

- F.P. Kelly. Charging and rate control for elastic traffic. European Transactions on Telecommunications, January 1997.
- F.P. Kelly, A. Maulloo, and D. Tan. Rate control in communication networks: shadow prices, proportional fairness and stability. Journal of the Operational Research Society, 49, 1998.
- N. Likhanov and R.R. Mazumdar. Cell loss asymptotics for buffers fed with a large number of independent stationary sources. In IEEE INFOCOM'98.
- S.H. Low and P.P. Varaiya. A new approach to service provisioning in ATM networks. IEEE/ACM Trans. on Networking, October 1993.
- S.H. Low. Equilibrium bandwidth and buffer allocation for elastic traffics. May 1997, Partial and preliminary results were presented at the International Conference on Telecommunications (ICT'97), Melbourne, Australia, April 1997.
- S.H. Low. Equilibrium Allocation and Pricing of Network Resources Among User-Suppliers. Preprint, November 1997.
- J.K. Mackie-Mason an H.R. Varian. Pricing the Internet. Available at Hal Varian's The Information Economy. URL: http://www.sims.berkeley.edu/resources/infoecon/index.html

Closely related references (cont.)

- B.M. Mitchell and I. Vogelsang. Telecommunications pricing: Theory and Practice. Cambridge University Press, 1991
- M. Montgomery and G. de Veciana. On the relevance of time scales in performance oriented traffic characterizations. In Proc. of IEEE INFOCOM'96, San Fransisco, CA, March 1996.
- S. Shenker. Fundamental design issues for the future Internet. IEEE J. Select. Areas Commun., September 1995.
- A. Simonian and J. Guibert. Large deviations approximations for fluid queues fed by a large number of on-off sources. IEEE J. Select. Areas Commun., August 1995.
- V. A. Siris, D.J. Songhurst, G.D. Stamoulis, and M. Stoer. Usage-based charging using effective bandwidths: studies and reality. In ITC-16, Edinburgh, UK, June 1999.
- D. Walker, F.P. Kelly and J. Solomon. Tariffing in the new IP/ATM environment. Telecommunications Policy, 21, 1997.
- J. Walrand and P.P. Varaiya. High Performance Communication Networks. Morgan Kaufmann Publishers, Inc., 1996.

Other references

- R. Cocchi, D. Estrin, S. Shenker, and L. Zhang. A study of priority pricing in multiple service class networks. In Proc. of ACM SIGCOMM'91.
- D.D. Clark. Adding service discrimination to the Internet. Telecommunications Policy, 20, 1996.
- D.D. Clark. A model for cost allocation and pricing in the Internet. In L.W. McKnight and J.P. Bailey, editors, Internet Economics. MIT Press, 1996.
- R. Cocchi, S. Shenker, D. Estrin, and L. Zhang. Pricing in computer networks: Motivation, formulation, and examples. IEEE/ACM Trans. on Networking, November 1993.
- A. Gupta, D.O. Stahl, and A.B. Whinston. An economic approach to networked computing with priority classes. Technical report, University of Texas, Austin, December 1994.
- A. Gupta, D.O. Stahl, and A.B. Whinston. Managing the Internet as an economical system. Technical report, University of Texas, Austin, July 1994.
- M.L. Honig and K. Steiglitz. Usage-based pricing of packet data generated by a heterogeneous user population. In Proc. fo IEEE INFOCOM'95, Boston, MA.

Other references (cont.)

- S. Herzog, S. Shenker, and D. Estrin. Sharing the "cost" of multicast trees: an axiomatic analysis. In Proc. of ACM SIGCOMM'95, Cambridge, MA.
- J. Murphy, L. Murphy, and E.C. Posner. Distributed pricing for embedded ATM networks. In Proc. of the 14th International Teletraffic Congress (ITC-14).
- J.K. Mackie-Mason an H.R. Varian. Some FAQs about usage-based pricing, September 1994. Available at Hal Varian's The Information Economy. URL: http://www.sims.berkeley.edu/resources/infoecon/index.html
- J.K. Mackie-Mason an H.R. Varian. Economic FAQs about the Internet, June 1995. Available at Hal Varian's The Information Economy.
- J.K. Mackie-Mason an H.R. Varian. Pricing congestible network resources. IEEE J. Select Areas in Commun., September 1995.
- A. Odlysko. A modest proposal for preventing Internet congestion. Preprint, September 1997. AT&T Labs-Research.
- A. Orda and N. Shimkin. Incentive pricing in multi-class communication networks. In Proc. IEEE INFOCOM'97, Kobe, Japan, April 1997.

Other references (cont.)

- C. Parris and D. Ferrari. A resource based pricing policy for real-time channels in a packet-switching network. Technical report, International Computer Science Institute, Berkeley, CA, 1992.
- C. Parris, S Keshav, and D. Ferrari. A framework for the study of pricing in integrated networks. Technical report TR-92-016, International Computer Science Institute, Berkeley, CA, March 1993.
- S. Shenker, D. Clark, D. Estrin, and S. Herzog. Pricing in computer networks: Reshaping the research agenda. ACM Computer Communications Review, 26, 1996.
- J. Sairameshm D.F. Ferguson, adn Y. Yemini. An approach to pricing, optimal allocation, and quality of service provisioning in high-speed packet networks. In Proc of IEEE INFOCOM'95, Boston, MA, April 1995.
- S. Shenker. Service models and pricing policies for an integrated services Internet. In B. Kahin adn J. Keller, editors, Public Access to the Internet. Prentice Hall, 1995.

Other references (cont.)

- A. Shwartz and A. Weiss. Large Deviations for Performance Analysis. Chapman and Hall, 1995.
- H. R. Varian. Microeconomic Analysis. W.W. Norton & Company Inc., 1992.
- Q. Wang, J.M. Peha, and M. A. Sirbu. The design of an optimal pricing scheme for ATM integrated services networks. In L.W. McKnight and J.P. Bailey, editors, Internet Economics. MIT Press, 1996.

Technology background

- The ATM Forum. Traffic Management Specification Version 4.0. April 1996.
- International Telecommunication Union Telecommunications Sector (ITU-T) I.371. Traffic control and congestion control in B-ISDN.
- R. Braden, D. Clark, and S. Shenker. Integrated services in the Internet architecture: an overview. RFC1633, July 1994.
- M.W. Garrett. A service architecture for ATM: From applications to scheduling. IEEE Network Magazine, May 1996.
- S. Shenker, C. Partridge, and R. Guerin. Specification of guaranteed quality of service. RFC1212, Integrated Services Working Group, September 1997.
- J. Wroclawski. Specification of the controlled load network element service. RFC2211, Integrated Services Working Group, September 1997.

Some thoughts...

- **There is no unique view on charging for network services**
  - Disparate models, contradicting proposals
- **There is no need for pricing network services!**
  - No congestion in the future
  - price only content
- **There is nothing new! (Economists did everything already)**
  - yes and no!
  - new issues:
    - complex service semantics, not obvious charging structures
    - congestion and stability depends on charging
    - scalability issues, interconnection
    - dynamic control structures