Resource Negotiation and Pricing in DiffServ for Adaptive Multimedia Applications

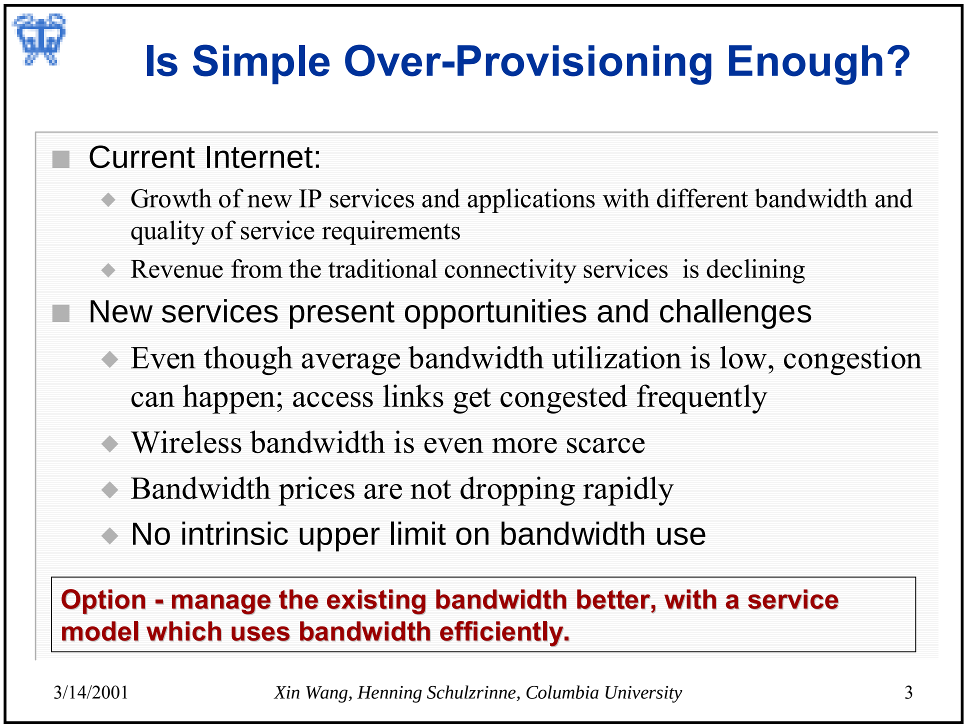
Xin Wang and Henning Schulzrinne
Internet Real -Time Laboratory
Columbia University
http://www.cs.columbia.edu/~xinwang

Outline

- Background
- RNAP: architecture and messaging
- Pricing models
- User adaptation

Resource Negotiation Framework

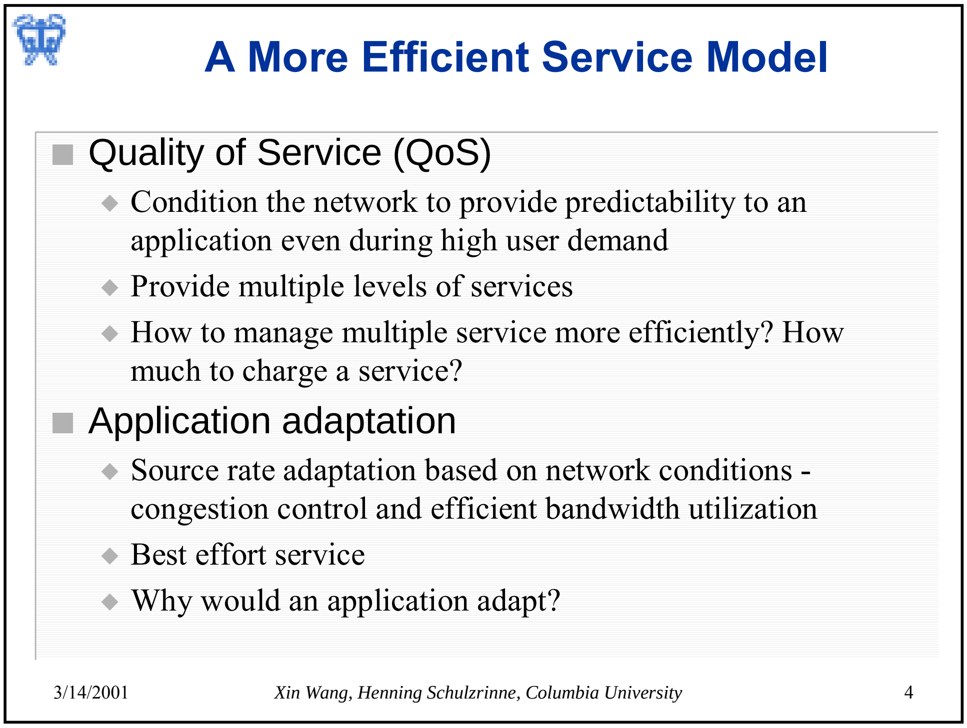
- Testbed demonstration of Resource Negotiation Framework
- Simulation and discussion of Resource Negotiation Framework



Is Simple Over-Provisioning Enough?

- Current Internet:
  - Growth of new IP services and applications with different bandwidth and quality of service requirements
  - Revenue from the traditional connectivity services is declining
- New services present opportunities and challenges
  - Even though average bandwidth utilization is low, congestion can happen; access links get congested frequently
  - Wireless bandwidth is even more scarce
  - Bandwidth prices are not dropping rapidly
  - No intrinsic upper limit on bandwidth use

Option - manage the existing bandwidth better, with a service model which uses bandwidth efficiently.



A More Efficient Service Model

- Quality of Service (QoS)
  - Condition the network to provide predictability to an application even during high user demand
  - Provide multiple levels of services
  - How to manage multiple service more efficiently? How much to charge a service?
- Application adaptation
  - Source rate adaptation based on network conditions - congestion control and efficient bandwidth utilization
  - Best effort service
  - Why would an application adapt?

A More Efficient Service Model (cont'd)

- Requirements of QoS/adaptive model:
  - mechanism to select and negotiate services
  - adaptive applications
  - short-term resource configuration for better response to user demand and network conditions, for more efficient resource usage
  - price network services based on QoS (resources consumed), allocate resources based on user willingness-to-pay
  - provide signal / incentive for user adaptation through pricing

Allow dynamic resource negotiation during ongoing service

↓

- A dynamic service selection and resource negotiation mechanism
- Usage-,QoS-,demand-sensitive pricing

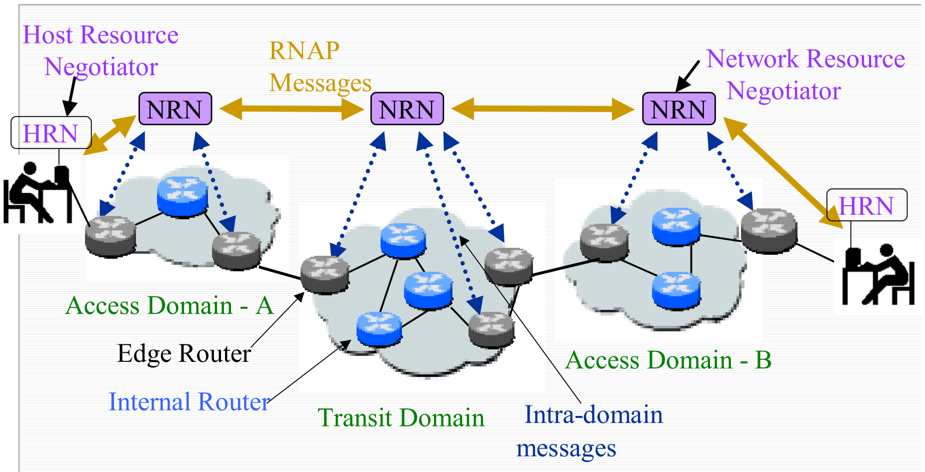
What We Add to Enable This Model

- A dynamic resource negotiation protocol: **RNAP**
  - An abstract **Resource Negotiation And Pricing** protocol
  - Enables user and network (or two network domains) to dynamically negotiate multiple services
  - Enables network to formulate and communicate prices and charges
  - Service predictability: commit service and price for an interval
  - Multi-party negotiation: senders, receivers, or both
  - Reliable and scalable
  - Lightweight and flexible: embedded in other protocols, e.g., RSVP, or implemented independently
- A demand-sensitive pricing model
  - Enables differential charging for supporting multiple levels of services; services priced to reflect the cost and long-term user demand
  - Allows for congestion pricing to motivate user adaptation

What We Add... (cont'd)

- Demonstrate a complete resource negotiation framework (RNAP, pricing model, user adaptation) on test-bed network
- Show significant advantages relative to static resource allocation and fixed pricing using simulations:
  - Much lower service blocking rate under resource contention
  - Service assurances under large or bursty offered loads, without highly conservative provisioning
  - Higher perceived user benefit and higher network revenue

Protocol Architectures: Centralized
(RNAP-C)



Host Resource Negotiator

RNAP Messages

Network Resource Negotiator

NRN

NRN

NRN

HRN

HRN

Access Domain - A

Edge Router

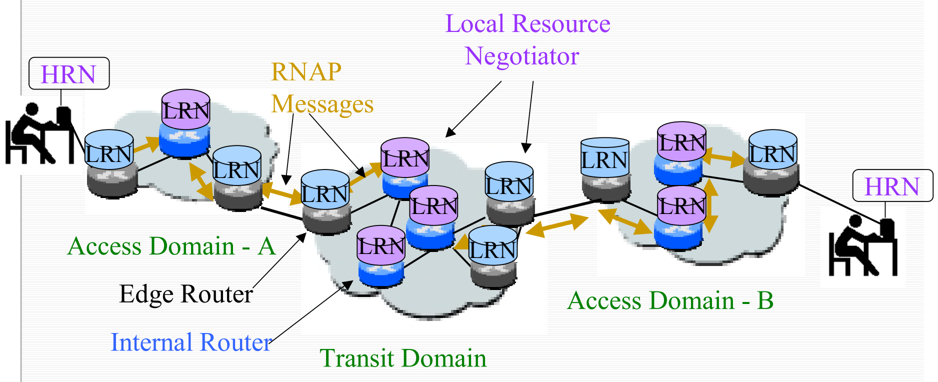
Internal Router

Transit Domain

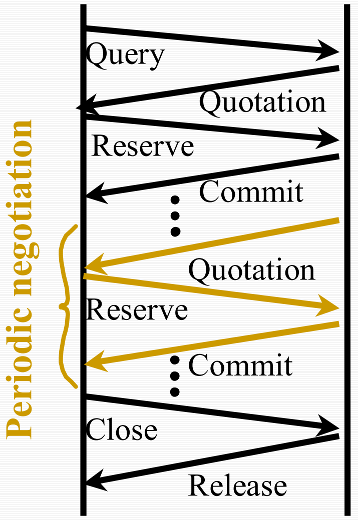
Intra-domain messages

Access Domain - B

Protocol Architectures: Distributed
(RNAP-D)



RNAP Messages



Query: Inquires about available services, prices

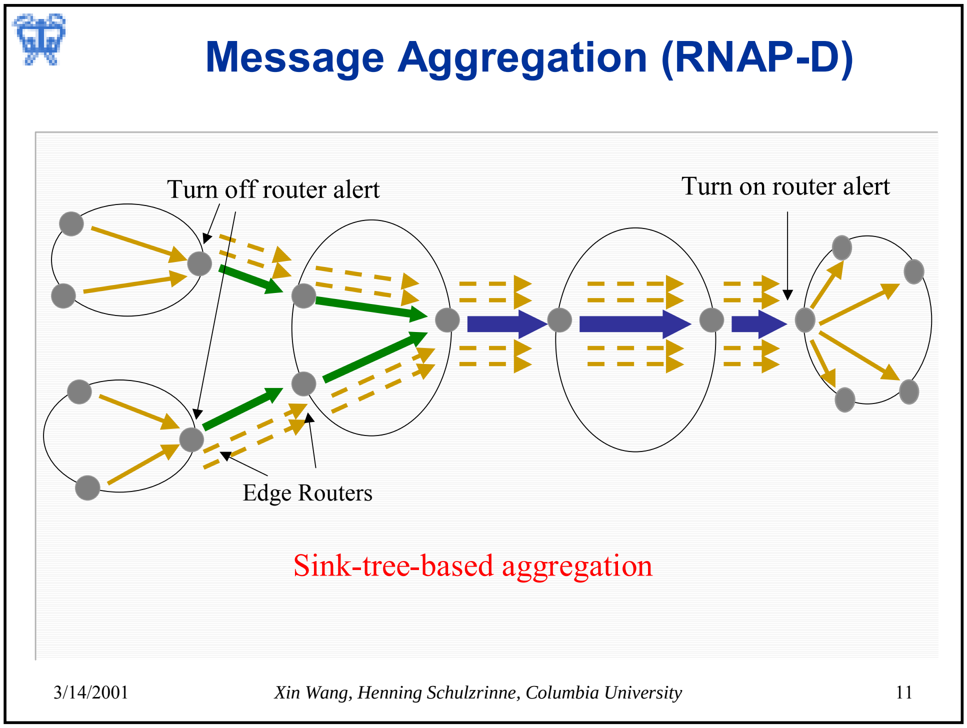
Quotation: Specifies service availability, accumulates service statistics, prices

Reserve: Requests services and resources, Modifies earlier requests

Commit: Confirms the service request at a specific price or denies it.

Close: Tears down negotiation session

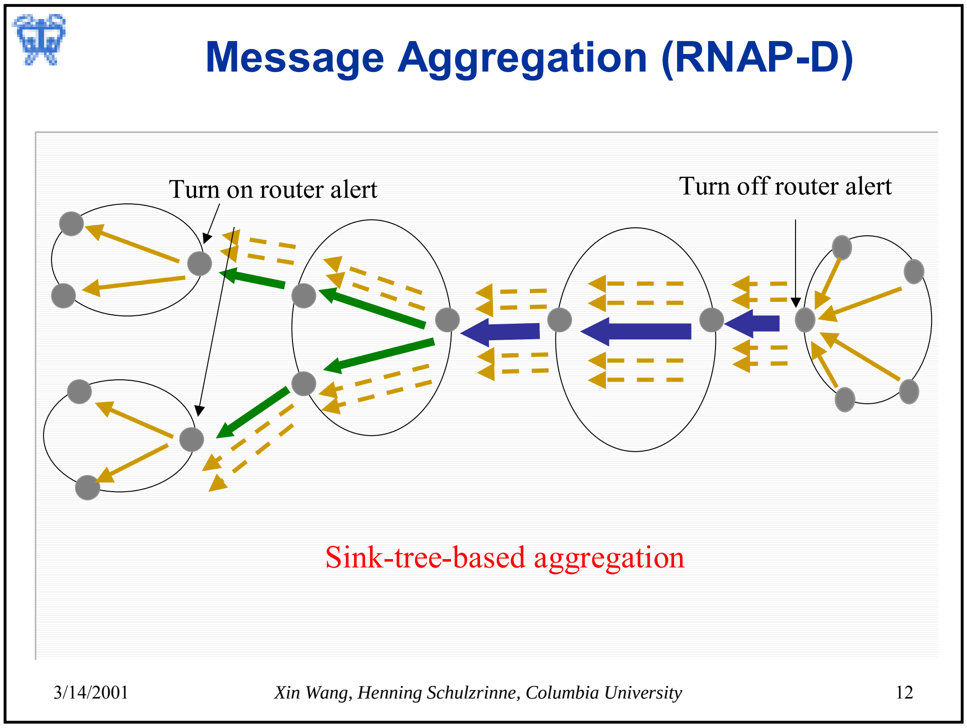
Release: Releases the resources



Message Aggregation (RNAP-D)

Sink-tree-based aggregation

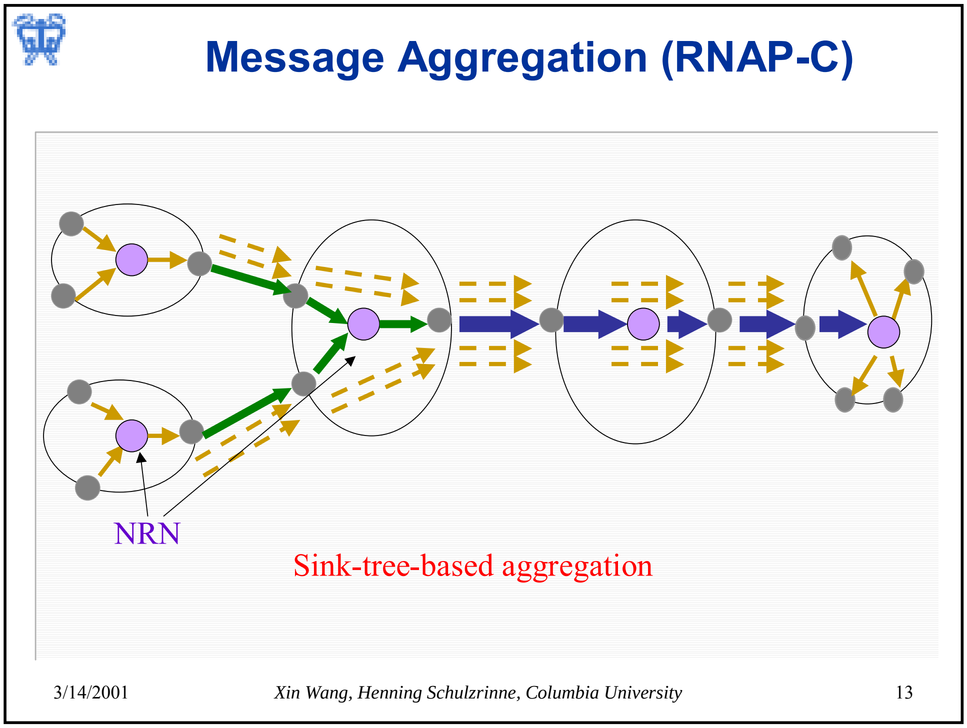
3/14/2001 *Xin Wang, Henning Schulzrinne, Columbia University* 11



Message Aggregation (RNAP-D)

Sink-tree-based aggregation

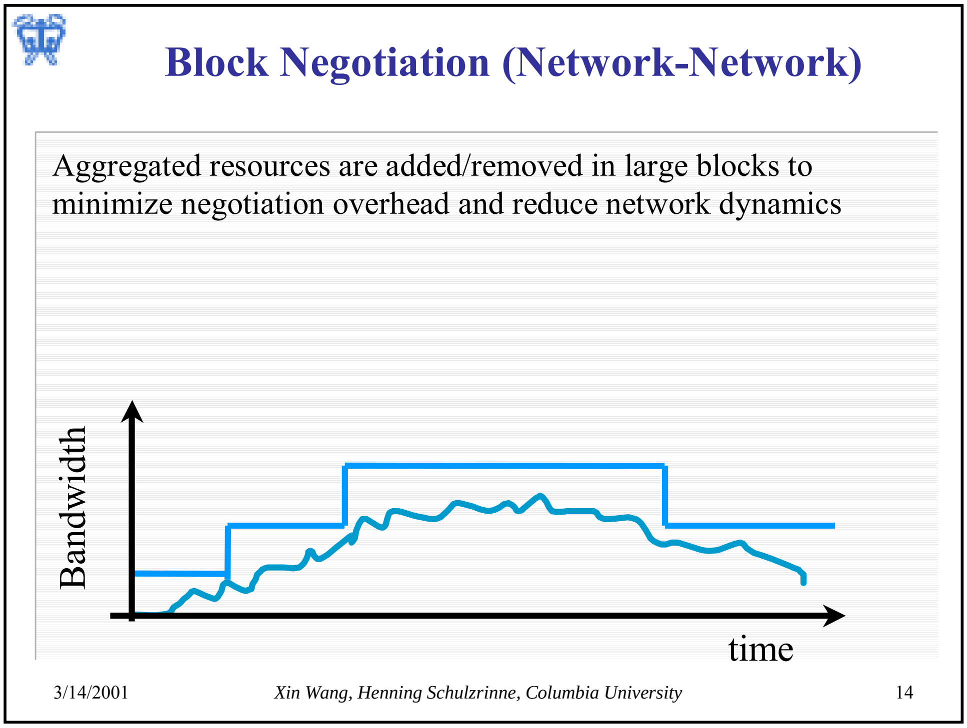
3/14/2001 *Xin Wang, Henning Schulzrinne, Columbia University* 12



Message Aggregation (RNAP-C)

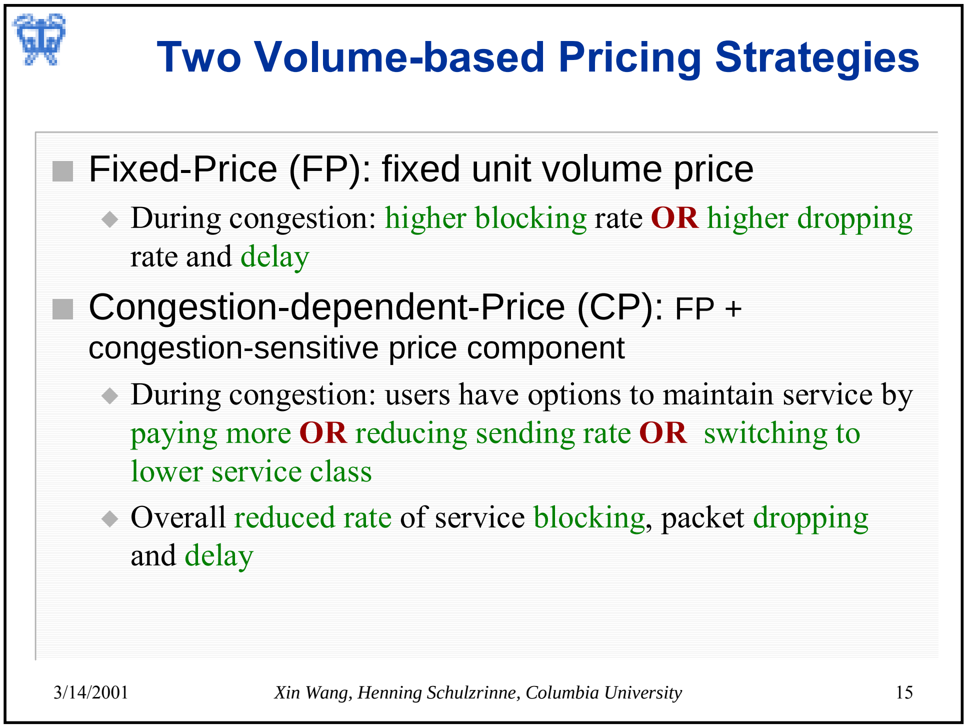
NRN

Sink-tree-based aggregation



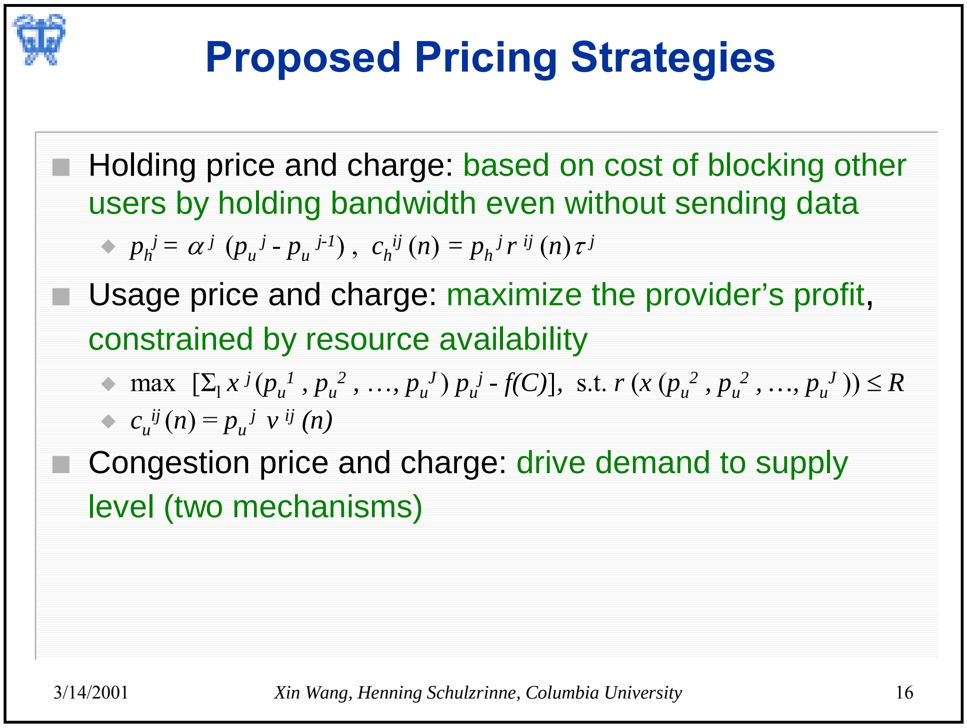
Block Negotiation (Network-Network)

Aggregated resources are added/removed in large blocks to minimize negotiation overhead and reduce network dynamics



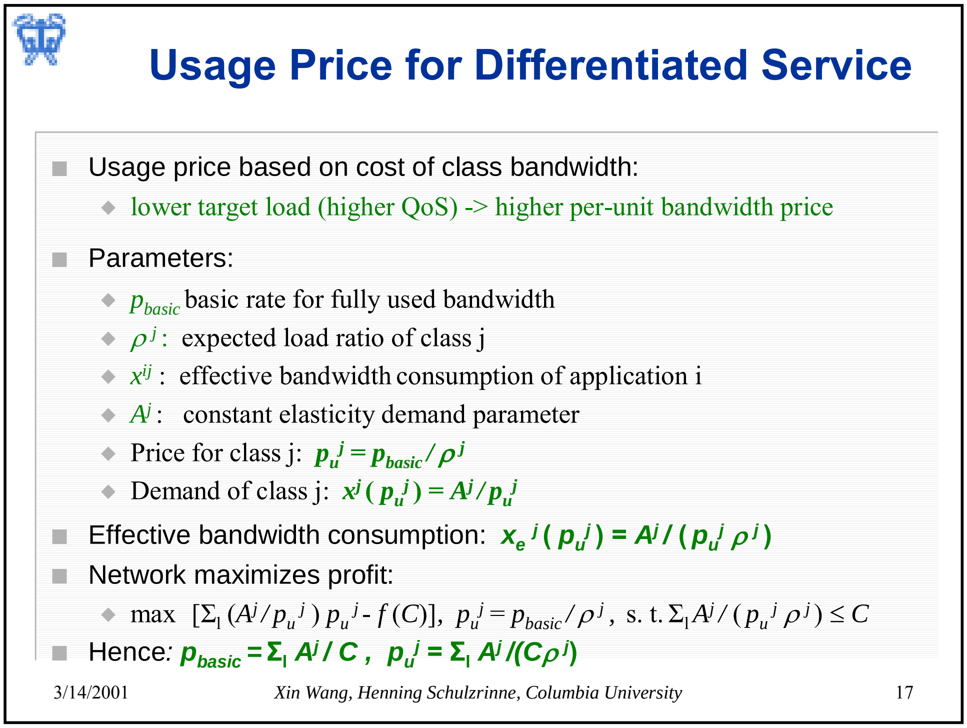
Two Volume-based Pricing Strategies

- Fixed-Price (FP): fixed unit volume price
  - During congestion: higher blocking rate **OR** higher dropping rate and delay
- Congestion-dependent-Price (CP): FP + congestion-sensitive price component
  - During congestion: users have options to maintain service by paying more **OR** reducing sending rate **OR** switching to lower service class
  - Overall reduced rate of service blocking, packet dropping and delay



Proposed Pricing Strategies

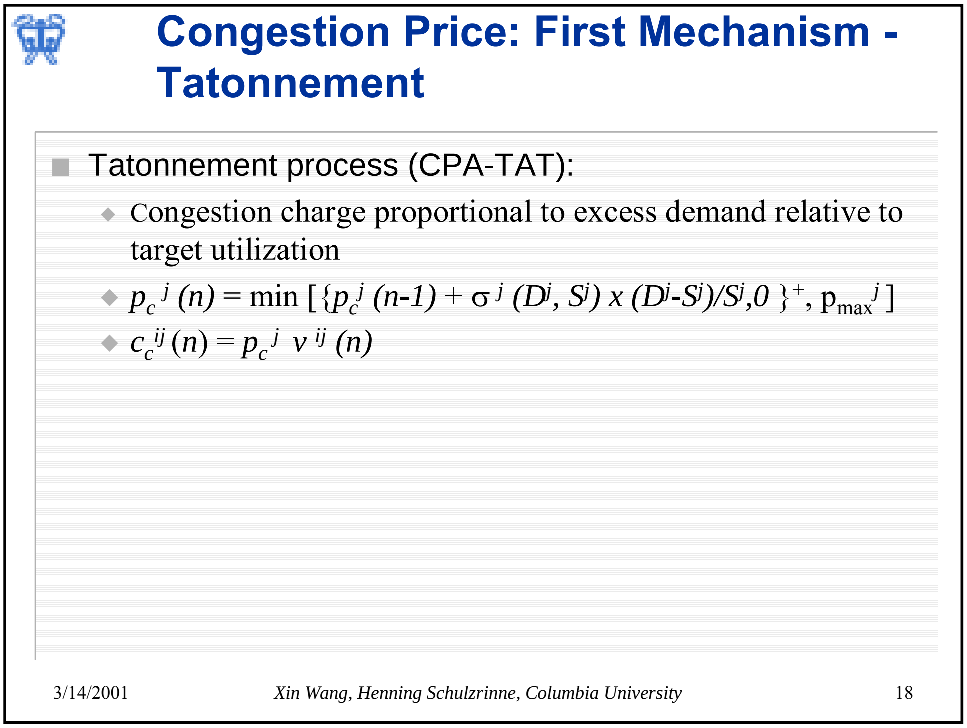
- Holding price and charge: based on cost of blocking other users by holding bandwidth even without sending data
  - $p_h^j = \alpha^j \, (p_u^j - p_u^{j-1})\,,\ c_h^{ij}(n) = p_h^j r^{ij}(n) \tau^j$
- Usage price and charge: maximize the provider’s profit, constrained by resource availability
  - $\max\ [\Sigma_l\, x^j (p_u^1, p_u^2, \ldots, p_u^J)\, p_u^j - f(C)]$, s.t. $r\,(x\,(p_u^2, p_u^2, \ldots, p_u^J)) \leq R$
  - $c_u^{ij}(n) = p_u^j\ v^{ij}(n)$
- Congestion price and charge: drive demand to supply level (two mechanisms)



Usage Price for Differentiated Service

- Usage price based on cost of class bandwidth:
  - lower target load (higher QoS) -> higher per-unit bandwidth price
- Parameters:
  - p_{basic} basic rate for fully used bandwidth
  - ρ^j: expected load ratio of class j
  - x^{ij}: effective bandwidth consumption of application i
  - A^j: constant elasticity demand parameter
  - Price for class j: $p_u^j = p_{basic} / \rho^j$
  - Demand of class j: $x^j(p_u^j) = A^j / p_u^j$
- Effective bandwidth consumption: $x_e^j(p_u^j) = A^j / (p_u^j \rho^j)$
- Network maximizes profit:
  - max $[\Sigma_l (A^j / p_u^j) p_u^j - f(C)]$, $p_u^j = p_{basic} / \rho^j$, s. t. $\Sigma_l A^j / (p_u^j \rho^j) \le C$
- Hence: $p_{basic} = \Sigma_l A^j / C$, $p_u^j = \Sigma_l A^j / (C \rho^j)$

3/14/2001 *Xin Wang, Henning Schulzrinne, Columbia University* 17



Congestion Price: First Mechanism - Tatonnement

- Tatonnement process (CPA-TAT):
  - Congestion charge proportional to excess demand relative to target utilization
  - $p_c^j(n) = \min[\{p_c^j(n-1) + \sigma^j(D^j, S^j) x (D^j - S^j)/S^j, 0\}^+, p_{max}^j]$
  - $c_c^{ij}(n) = p_c^j v^{ij}(n)$

3/14/2001 *Xin Wang, Henning Schulzrinne, Columbia University* 18

Congestion Price: Second Mechanism - *M*-bid Second-price Auction

- Auction models in literature:
  - Assume unique bandwidth/price preference, one bid
  - Service uncertainty: user does not know about high demand until rejected
  - Other issues: setup delay, signaling burst, user response to auction results
- *M*-bid auction Model
  - User bids (bandwidth, price) for a number of bandwidths, bids obtained by sampling utility function.
  - Reduce uncertainty
  - Network selects highest bids, charges highest rejected bid price
  - During high demand: lower bandwidth (higher price per unit bandwidth) bids get selected; more users served
  - Periodic auctions - support congestion control
  - Inter-auction admission to reduce setup delay

Example of *M*-bid Auction

- Total capacity 70, congestion price is 2

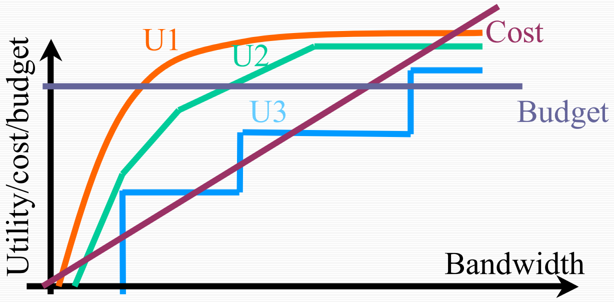
Bid Price	Bid Bandwidth	Bidder	Bid Selection
5	10	1	
4	10	2	
4	15	1	←
3.5	20	3	←
3	25	2	←
$\boxed{2}$	30	**3**	

Cutoff

Congestion Price

Rate Adaptation of Multimedia System

- Gain optimal perceptual value of the system based on the network conditions and user profile
- Utility function: users' preference or willingness to pay



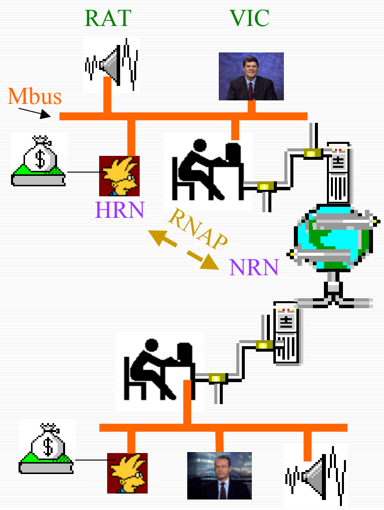
Example Utility Function

- Utility is a function of bandwidth at fixed QoS
  - An example utility function: $U(x) = U_0 + \omega \log(x / x_m)$
  - U_0 : perceived (opportunity) value at minimum bandwidth
  - ω : sensitivity of the utility to bandwidth
- Function of both bandwidth and QoS
  - $U(x) = U_0 + \omega \log(x / x_m) - k_d d - k_l l$, for $x \geq x_m$
  - k_d : sensitivity to delay
  - k_l : sensitivity to loss

Two Rate-Adaptation Models

- Model1: User adaptation under CPA-TAT (tatonnement-based pricing)
  - Optimize perceived surplus of the multimedia system subject to budget and application requirements
  - With the example utility functions, resource request of application *i*:
    - Without budget constraint: $x^i = \omega^i / p^i$
    - With budget constraint: $x^i = b^i / p^i$, with $b^i = b\,(\omega^i / \Sigma_l\, \omega^k)$
- Model2: User adaptation under CPA-AUC (second-price auction)
  - Submit *M*-bid derived by sampling utility function; adapt rate based on allocated bandwidth/QoS

Testbed Architecture

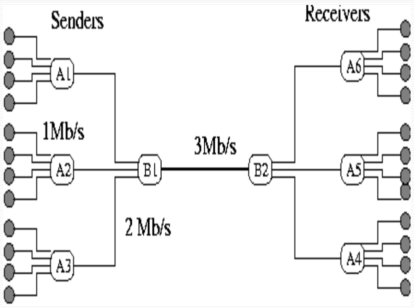


- Demonstrate functionality and performance improvement:
  - blocking rate, loss, delay, price stability, perceived media quality
- Host
  - HRN negotiates for a system
  - Host processes (HRN, VIC, RAT) communicate through Mbus
- Network
  - Router: FreeBSD 3.4 + ALTQ 2.2, CBQ extended for DiffServ
  - NRN: (1) Process RNAP messages; (2) Admission control, monitor statistics, compute price; (3) At edge, dynamically configure the conditioners and form charge
- Inter-entity signaling: RNAP

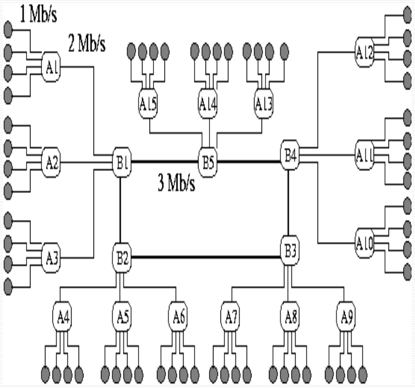
Simulation Design

- Performance comparison:
  - **Fixed price policy (FP)** (usage price + holding price) versus **congestion price based adaptive service (CPA)** (usage price + holding price + congestion price)
- Four groups of experiments: effect of traffic load, admission control, traffic burstiness, and load balance between classes
- Weighted Round Robin (WRR) scheduler
- Three classes: EF, AF, BE
  - EF: load threshold 40%, delay bound 2 ms, loss bound 10^{-6}
  - AF: load threshold 60%, delay bound 5 ms, loss bound 10^{-4}
  - BE: load threshold 90%,delay bound 100 ms,loss bound 10^{-2}
- Sources: mix of on-off traffic and Pareto on-off traffic

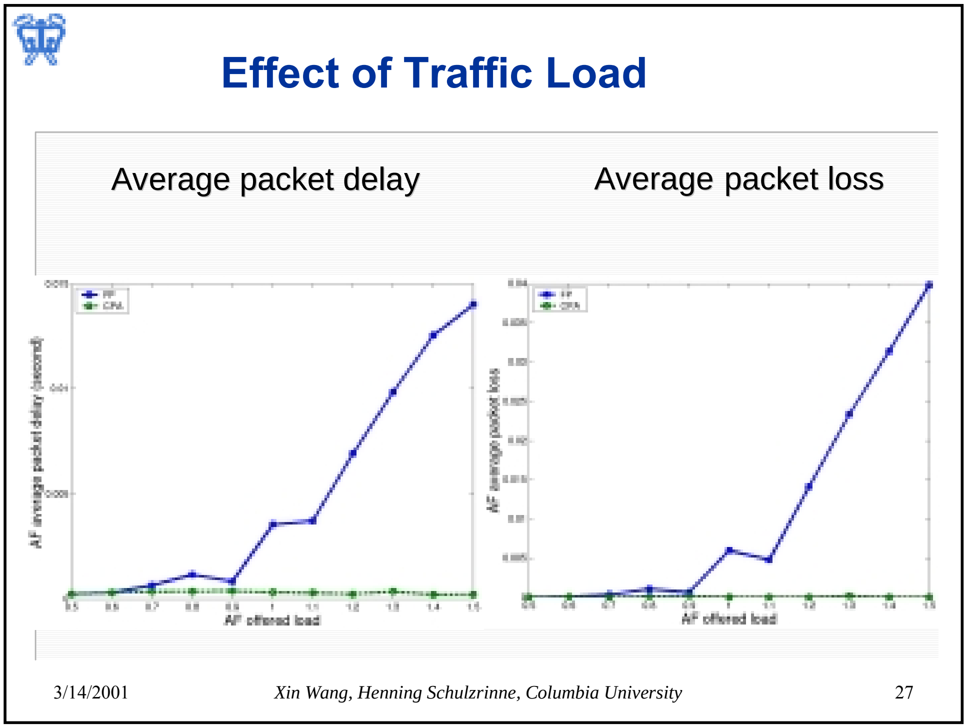
Simulation Architecture



Topology 1 (60 users)



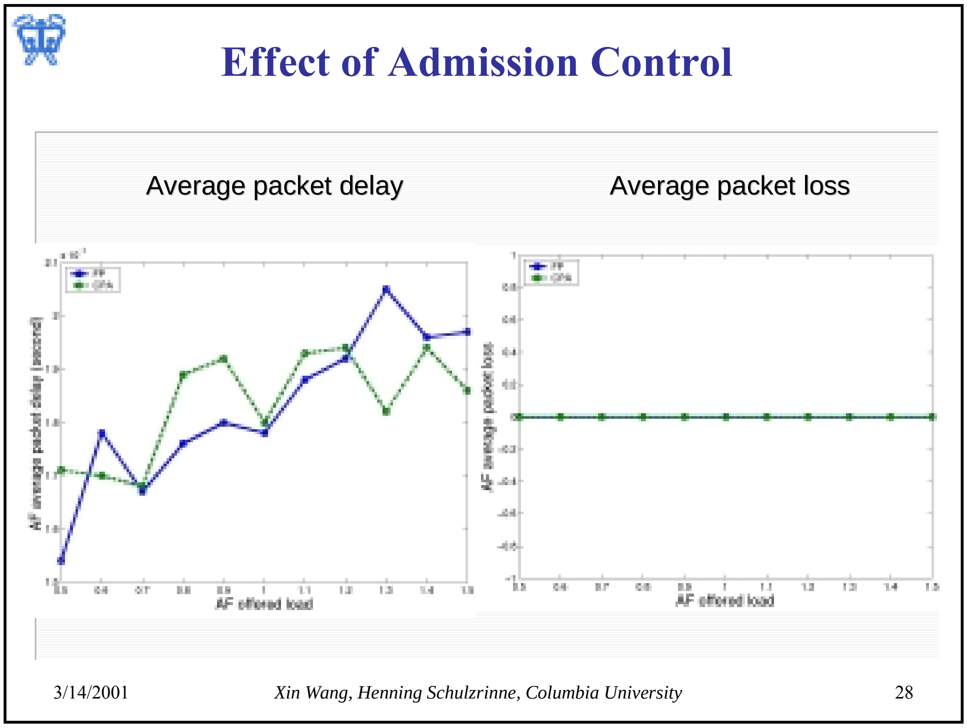
Topology 2 (360 users)



Effect of Traffic Load

Average packet delay

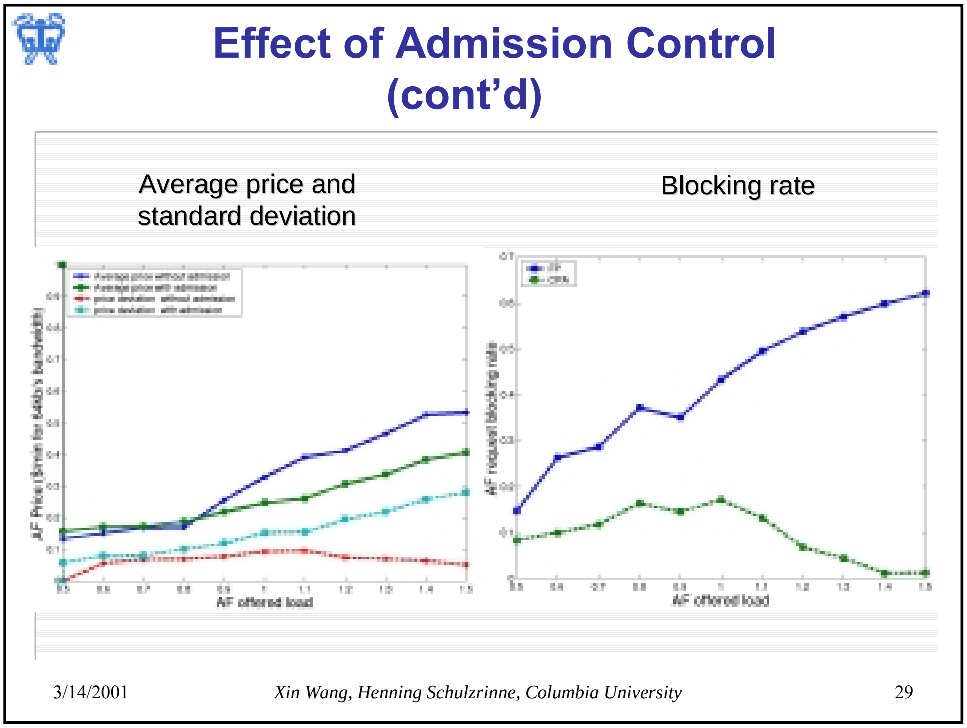
Average packet loss



Effect of Admission Control

Average packet delay

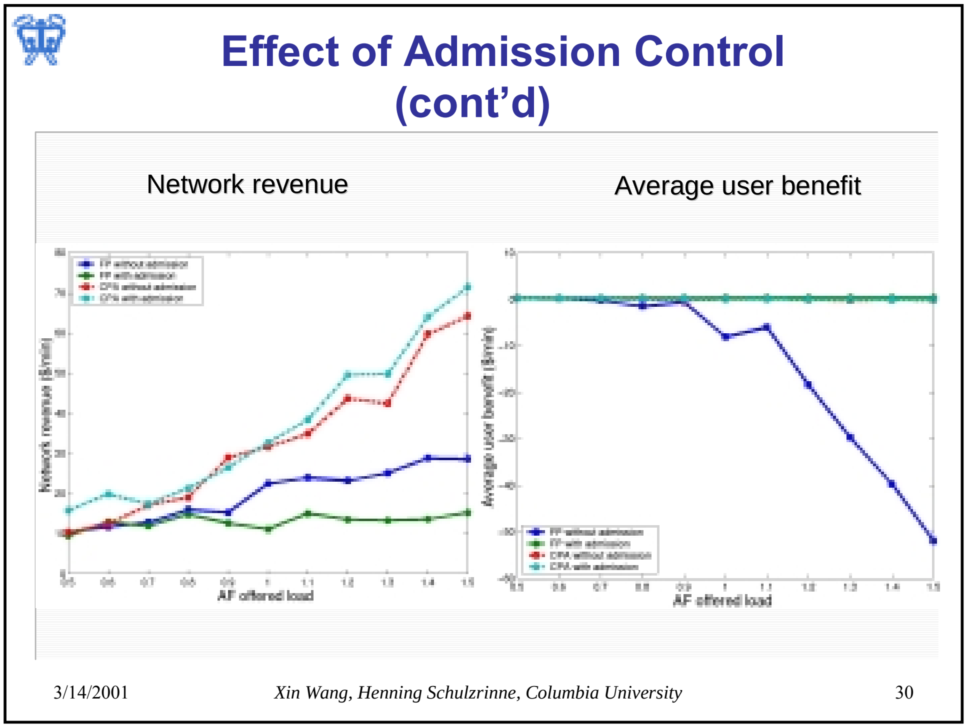
Average packet loss



Effect of Admission Control (cont'd)

Average price and standard deviation

Blocking rate



Effect of Admission Control (cont'd)

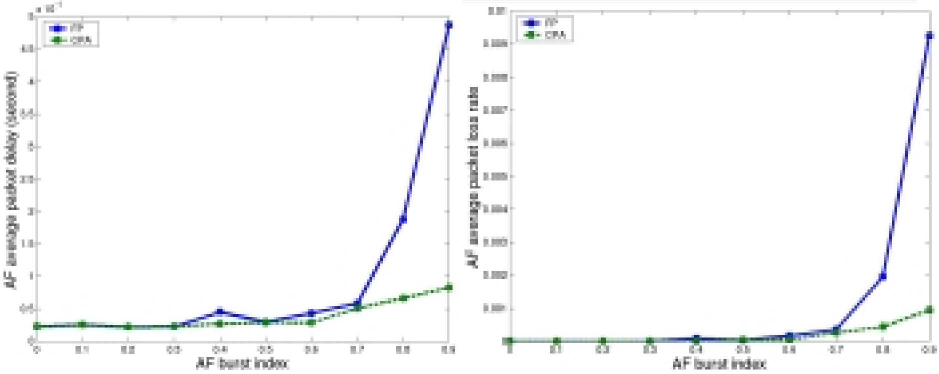
Network revenue

Average user benefit

Effect of Traffic Burstiness

Average packet delay

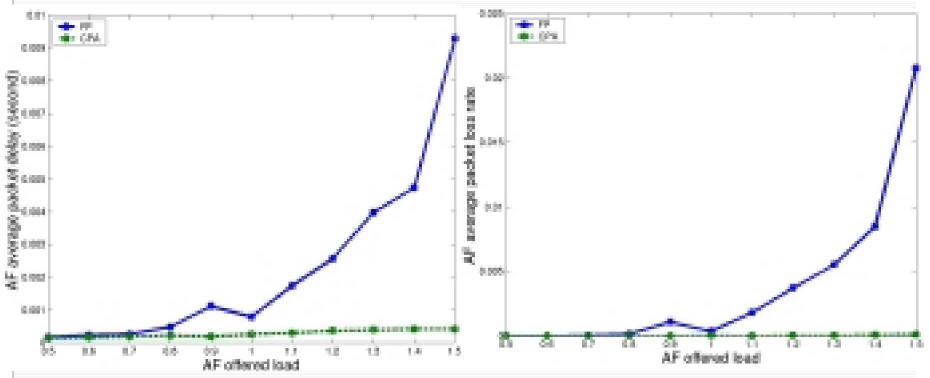
Average packet loss



Load Balance Between Classes

Average packet delay

Average packet loss

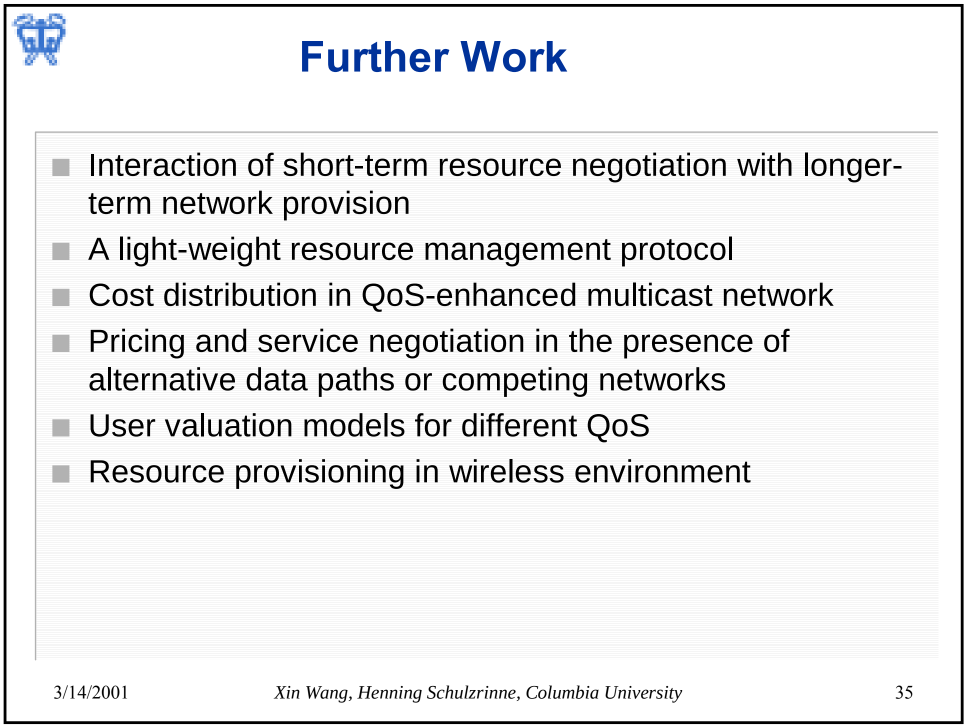


Simulation Results

- Congestion-price-based policy (CPA) + user adaptation vs Fixed price policy (FP) + no adaptation:
  - limit congestion
  - lower request blocking rate,
  - higher user satisfaction
  - higher network revenue
- Differentiated service requires different target loads in each class
- Even without admission control, CPA policy restricts load to targeted level, can meet service assurance
- With admission control, blocking rate and price dynamics further reduced
- Allowing service class migration allows for service assurance at predicted level and further stabilizes price

Conclusions

- Proposed a dynamic resource negotiation framework: A Resource Negotiation And Pricing protocol (**RNAP**) , a rate and QoS adaptation model, and a pricing model
- RNAP: Supports dynamic service negotiation between network and users, and between peer networks
- Pricing models
  - Based on resources consumed by service class and long-term user demand, including congestion-sensitive component to motivate user demand adaptation during resource contention
  - *M*-bid Auction Model serves more users than comparable auction schemes, and reduces uncertainty of service availability
- User adaptation: maximize perceived user satisfaction



Further Work

- Interaction of short-term resource negotiation with longer-term network provision
- A light-weight resource management protocol
- Cost distribution in QoS-enhanced multicast network
- Pricing and service negotiation in the presence of alternative data paths or competing networks
- User valuation models for different QoS
- Resource provisioning in wireless environment