# IP End-To-End Quality of Service: Recent Solutions and Issues

**Raj Jain**

Raj Jain is now at
Washington University in Saint Louis
Jain@cse.wustl.edu
http://www.cse.wustl.edu/~jain/

These slides are available at
http://www.cis.ohio-state.edu/~jain/talks/ipqos2.htm


- ATM QoS and Issues
- Integrated services/RSVP and Issues
- Differentiated Services and Issues
- QoS using MPLS
- End-to-end QoS
- This is an update to the May'98 talk
  http://www.cis.ohio-state.edu/~jain/talks/ipqos.htm

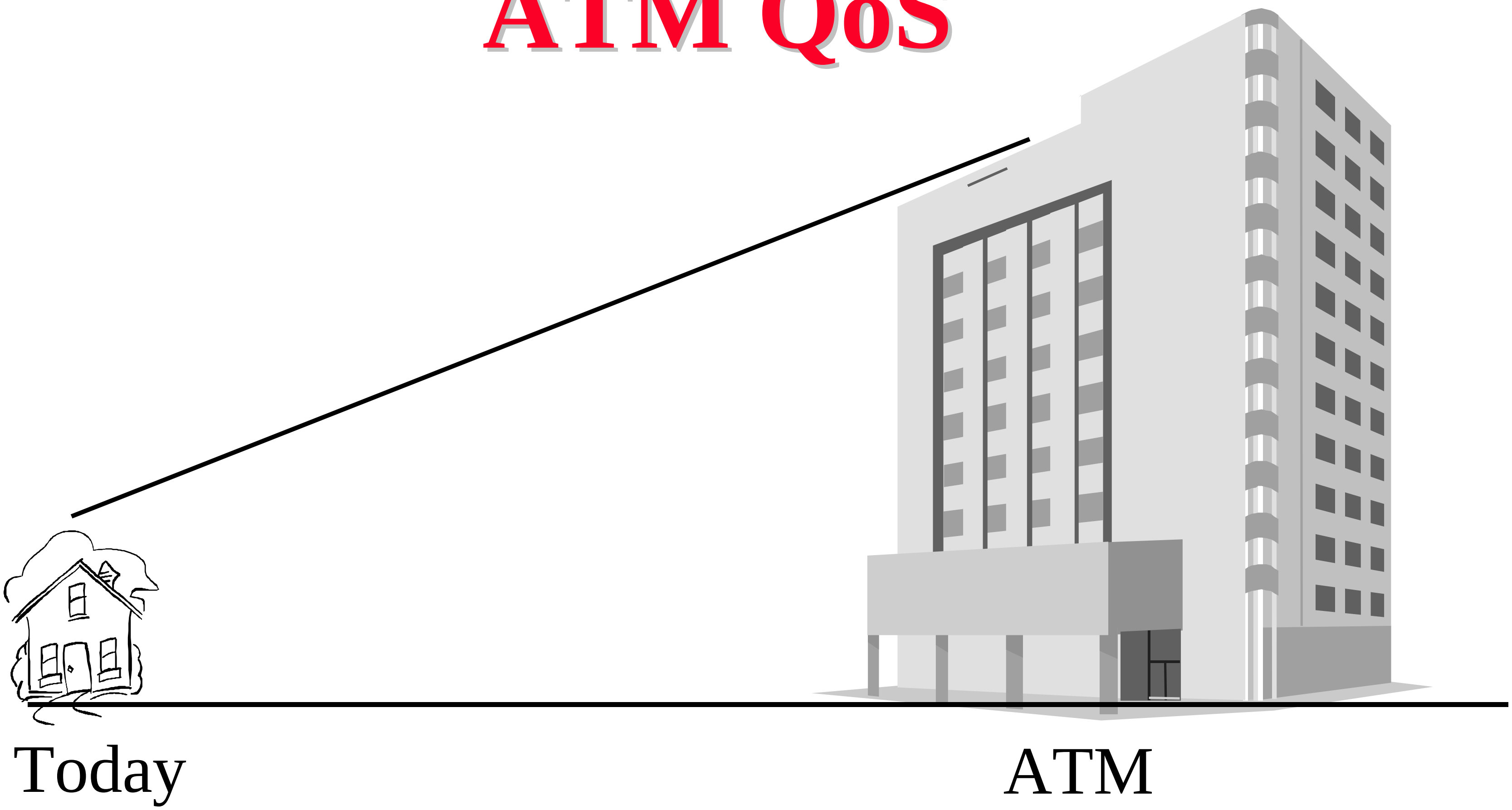
# What is QoS?

- "Unequal" allocation of resources
- Predictable Quality: Throughput, Delay, Loss, Delay jitter, Error rate
- Mechanisms: Routing, Classifiers, Scheduling, Queueing, Buffer Management, Admission Control, Shaping, Policing, capacity planning

# ATM Service Categories

- **CBR**: Throughput, delay, delay variation
- **rt-VBR**: Throughput, delay, delay variation
- **nrt-VBR**: Throughput
- **UBR**: No Guarantees
- **GFR**: Minimum Throughput
- **ABR**: Minimum Throughput. Very low loss. Feedback.
- ATM also has QoS-based routing (PNNI)

# ATM QoS




Too much too soon

# ATM QoS: Issues

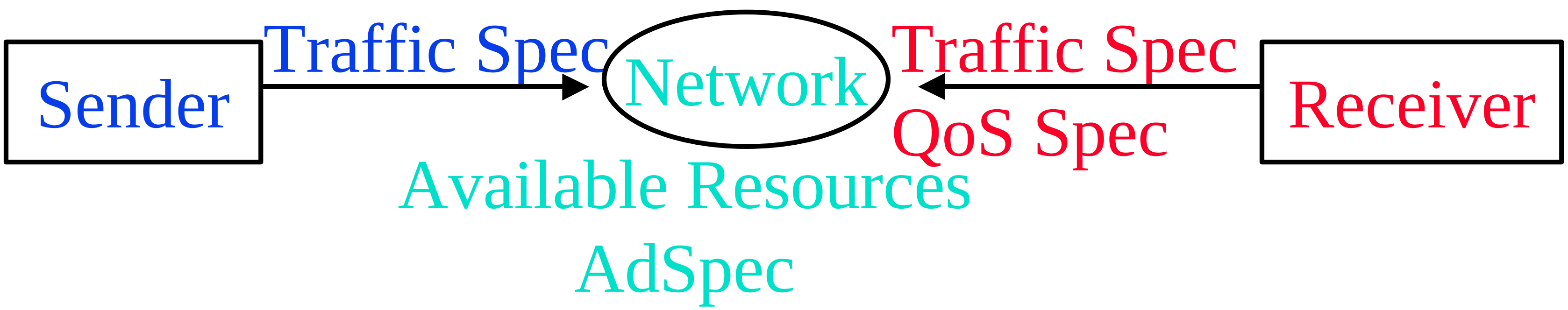
- Can't easily aggregate QoS: VP = Σ VCs
- Can't easily specifiy QoS: What is the CDV required for a movie?
- Signaling too complex ⇒ Need Lightweight Signaling
- Need Heterogeneous Point-to-Multipoint: Variegated VCs
- Need QoS Renegotiation
- Need Group Address
- Need priority or weight among VCs to map DiffServ and 802.1D

# Integrated Services

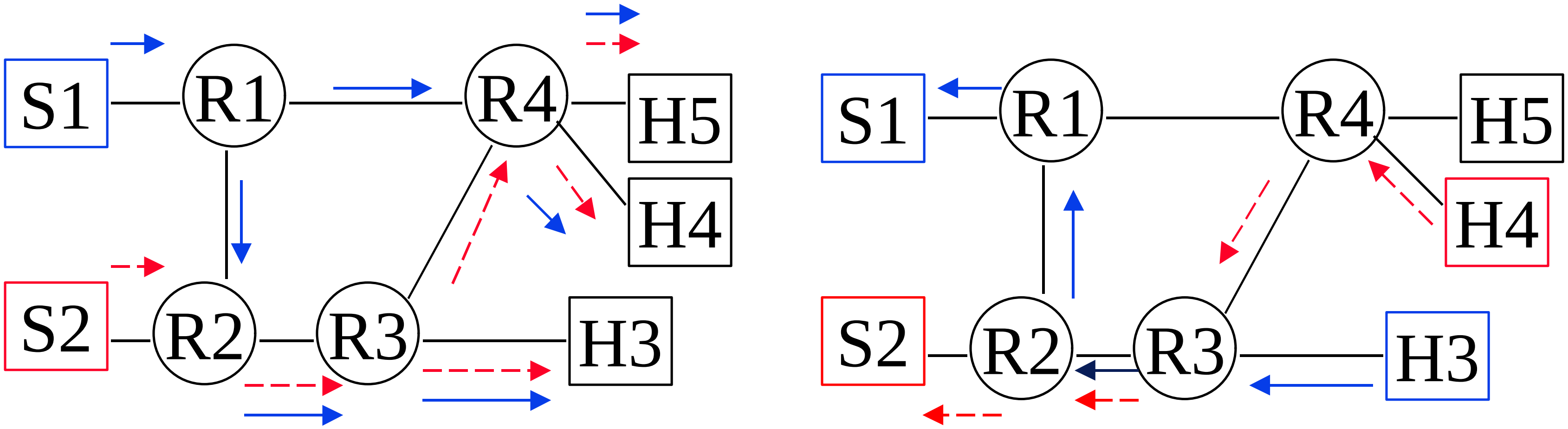
- Best Effort Service: Like UBR.
- Controlled-Load Service: Performance as good as in an unloaded datagram network. No quantitative assurances. Like nrt-VBR or UBR w MCR
- Guaranteed Service: rt-VBR
  - Firm bound on data throughput and delay.
  - Delay jitter or average delay not guaranteed or minimized.
  - Every element along the path must provide delay bound.
  - Is not always implementable, e.g., Shared Ethernet.
  - Like CBR or rt-VBR

# RSVP

- Resource ReSerVation Protocol
- Internet signaling protocol
- Carries resource reservation requests through the network including traffic specs, QoS specs, network resource availability
- Sets up reservations at each hop



# RSVP Messages




- Sources send PATH messages to the multicast address. Contain traffic spec and has place for network to indicate available resources.
- Receivers send ResV messages in the reverse direction. Contain QoS spec.
- Similar requests from multiple receivers are merged.

# Problems with RSVP and Integrated Services

- Complexity in routers: packet classification, scheduling
- Scalable in number of receivers per flow but Per-Flow State: O(n) $\Rightarrow$ Not scalable with # of flows. Number of flows in the backbone may be large. $\Rightarrow$ Suitable for small private networks
- Need a concept of "Virtual Paths" or aggregated flow groups for the backbone
- Need policy controls: Who can make reservations? Support for accounting and security. $\Rightarrow$ RSVP admission policy (rap) working group.

# Problems (Cont)

- Receiver Based:
  Need sender control/notifications in some cases.
  Which receiver pays for shared part of the tree?
- Soft State: Need route/path pinning (stability).
  Limit number of changes during a session.
- RSVP does not have negotiation and backtracking
- Throughput and delay guarantees require support of lower layers. Shared Ethernet $\Rightarrow$ IP can't do GS or CLS. Need switched full-duplex LANs.
- Can't easily do RSVP on ATM either
- Most of these arguments also apply to integrated services.

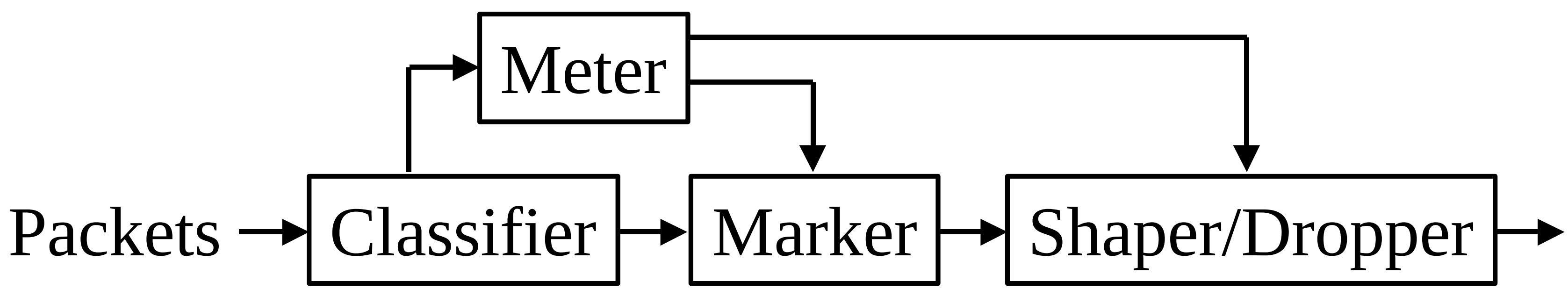
# Differentiated Services

| Ver | Hdr Len | Precedence | ToS | Unused | Tot Len |
|---|---|---|---|---|---|
| 4b | 4b | 3b | 4b | 1b | 16b |

- IPv4: 3-bit precedence + 4-bit ToS
- OSPF and integrated IS-IS can compute paths for each ToS
- Many vendors use IP precedence bits but the service varies $\Rightarrow$ Need a standard $\Rightarrow$ Differentiated Services
- DS working group formed February 1998
- Charter: Define ds byte (IPv4 ToS field)
- Mail Archive: http://www-nrg.ee.lbl.gov/diff-serv-arch/

# DiffServ Concepts

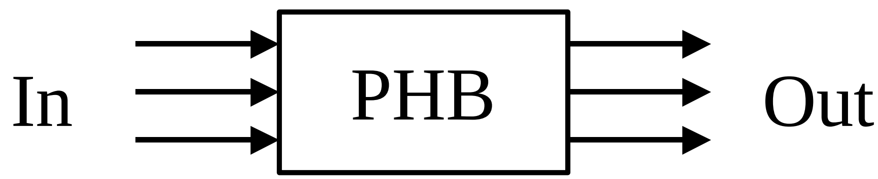
- Micro-flow = A single application-to-application flow
- Traffic Conditioners: Meters (token bucket), Markers (tag), Shapers (delay), Droppers (drop)
- Behavior Aggregate (BA) Classifier:
  Based on DS byte only
- Multi-field (MF) Classifiers:
  Based on IP addresses, ports, DS-byte, etc..



# Diff-Serv Concepts (Cont)

- Service: Offered by the protocol layer
  - Application: Mail, FTP, WWW, Video,...
  - Transport: Delivery, Express Delivery,...
    Best effort, controlled load, guaranteed service
  - DS group will not develop services
    They will standardize "Per-Hop Behaviors"

# Per-hop Behaviors




- Externally Observable Forwarding Behavior
- x% of link bandwidth
- Minimum x% and fair share of excess bandwidth
- Priority relative to other PHBs
- PHB Groups: Related PHBs. PHBs in the group share common constraints, e.g., loss priority, relative delay

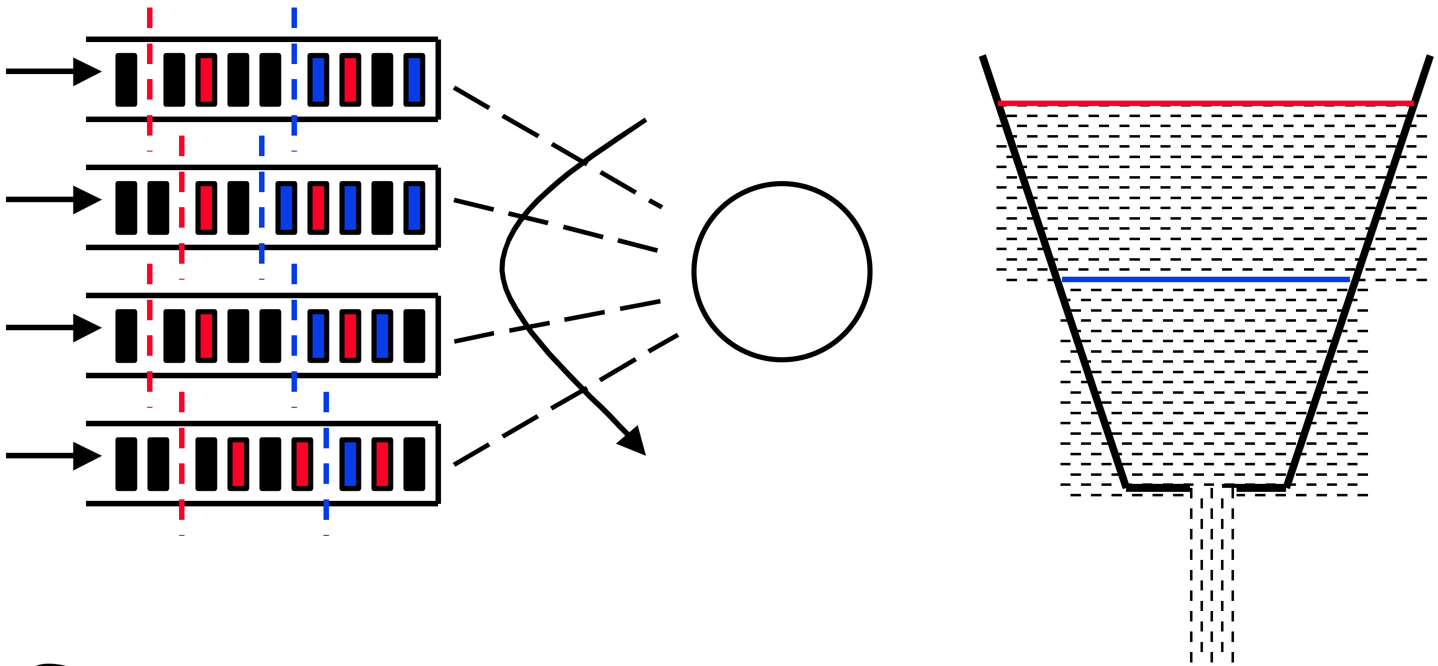
# Code Points

- Three Subsets:
  - xxxxx0 Standard
  - xxxx11 Experimental/Local Use
  - xxxx01 Currently Experimental/Local Use. May be used for future standards.
- xxx000 = Class Selectors
  Should follow current IP precedence rules
  - Larger ⇒ Relatively better performance
  - 11x000 must be better than 000000
    (110 000 and 111 000 used for network control)
- Two proposals: Expedited and Assured

# Expedited Forwarding

- Also known as "Premium Service"
- Virtual leased line
- Similar to CBR
- Guaranteed minimum service rate
- Policed: Arrival rate < Minimum Service Rate
- Not affected by other data PHBs
  $\Rightarrow$ Highest data priority (if priority queueing)
- Code point: 101 110

# Assured Forwarding



- PHB Group
- Four Classes: Decreasing weights in WFR/WFQ
- Three drop preference per class
  (one rate and two bucket sizes)

# Assured Forwarding (Cont)

- DS nodes SHOULD implement all 4 classes and MUST accept all 3 drop preferences
- Lower delay for lower classes
- Similar to nrt-VBR/ABR/GFR
- Code Points:

| Drop Prec. | Class 1 | Class 2 | Class 3 | Class 4 |
|---|---|---|---|---|
| Low | 010 000 | 011 000 | 100 000 | 101 000 |
| Medium | 010 010 | 011 010 | 100 010 | 101 010 |
| High | 010 100 | 011 100 | 100 100 | 101 100 |

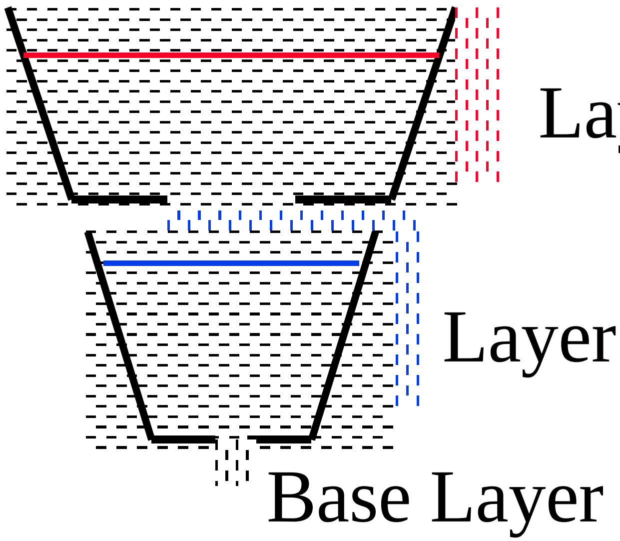
- Avoids 11x000 (used for network control)

# Assured Forwarding: Issues

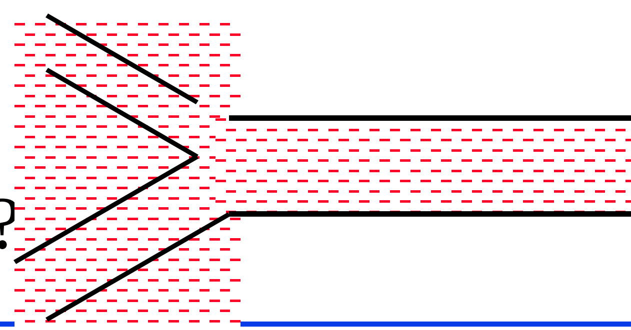
- Lower delay = Average, min, or max? (not specified)
- TCP slow/start does not distinguish between multiple drop preferences ⇒ Not useful for TCP

# AF Issues (Cont)

- Layered Video would like multiple rates (not thresholds) $\Rightarrow$ Multiple leaky buckets




- Merging of AF-flows
  AF(R1, DT11, DT12)
  + AF(R2, DT21, DT22) = ?



# AF Simulation Results

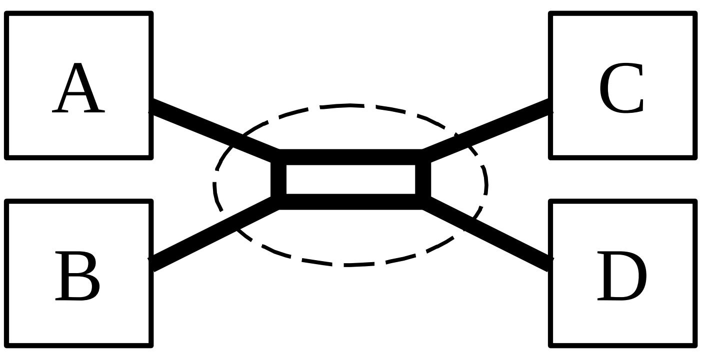
- TCP throughput is not close to target rates.
- Larger RTT $\Rightarrow$ Smaller throughput
- Larger target rates $\Rightarrow$ Smaller target/allocated ratio
- Non-TCP (non-adaptive) sources can degrade AF-TCP connections
- Performance of the aggregate changes when its composition changes.
- Token bucket marking is better than average queue marking for RIO (RED with In/Out a.k.a. WRED).
- Ref: J. Ibanez, "Preliminary Simulation Studies of the Assured Service," Bay Networks, BALTR98-023, July 1998.

# Problems with DiffServ

- per-hop $\Rightarrow$ Need at every hop
  One non-DiffServ hop can spoil all QoS
- End-to-end $\neq \Sigma$ per-Hop
  Designing end-to-end services with weighted guarantees at individual hops is difficult.
  Only EF will work.
- Designed for static Service Level Agreements (SLAs)
  Both the network topology and traffic are highly dynamic.
- Multicast $\Rightarrow$ Difficult to provision
  Dynamic multicast membership $\Rightarrow$ Dynamic SLAs?

# DiffServ Problems (Cont)

- DiffServ is unidirectional ⇒ No receiver control
- Modified DS field ⇒ Theft and Denial of service. Ingress node should ensure.
- How to ensure resource availability inside the network?
- QoS is for the aggregate not per-destination. Multi-campus enterprises need inter-campus QoS.



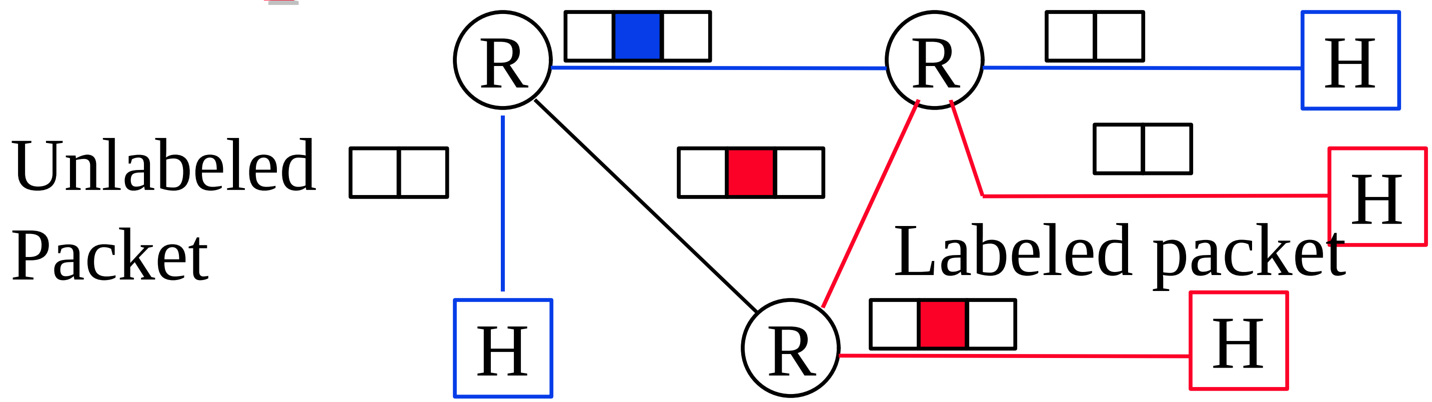
# DiffServ Problems (Cont)

- QoS is for the aggregate not micro-flows.
  Not intended/useful for end users. Only ISPs.
  - Large number of short flows are better handled by aggregates.
  - Long flows (voice and video sessions) need per-flow guarantees.
  - High-bandwidth flows (1 Mbps video) need per-flow guarantees.
- All IETF approaches are open loop control $\Rightarrow$ Drop
  Closed loop control $\Rightarrow$ Wait at source
  Data prefers waiting $\Rightarrow$ Feedback

# DiffServ Problems (Cont)

- Guarantees $\Rightarrow$ Stability of paths
  $\Rightarrow$ Connections (hard or soft)
  Need route pinning or connections.

# Multiprotocol Label Switching




- Entry "label switch router (LSR)" attaches a label to the packet based on the route
- Other LSRs switch packets based on labels. Do not need to look inside ⇒ Fast.
- Labels have local significance ⇒ Different label at each hop (similar to VC #)
- Exit LSR strips off the label

# MPLS

- Initially focused on IPv4 and IPv6.
  Technology extendible to other L3 protocols.
- Works on all LANs, ATM, Frame Relay, ...
- Not specific to a routing protocol (OSPF, RIP, ...)
- Optimization only. Labels do not affect the path.
  Only speed. Networks continue to work w/o labels
- Initially, MPLS was being designed for fast routing.
  Hardware based fast routers
  $\Rightarrow$ Switching not required for performance
- Now the group focus has been changed to "Traffic Engineering"

# Traffic Engineering Using MPLS

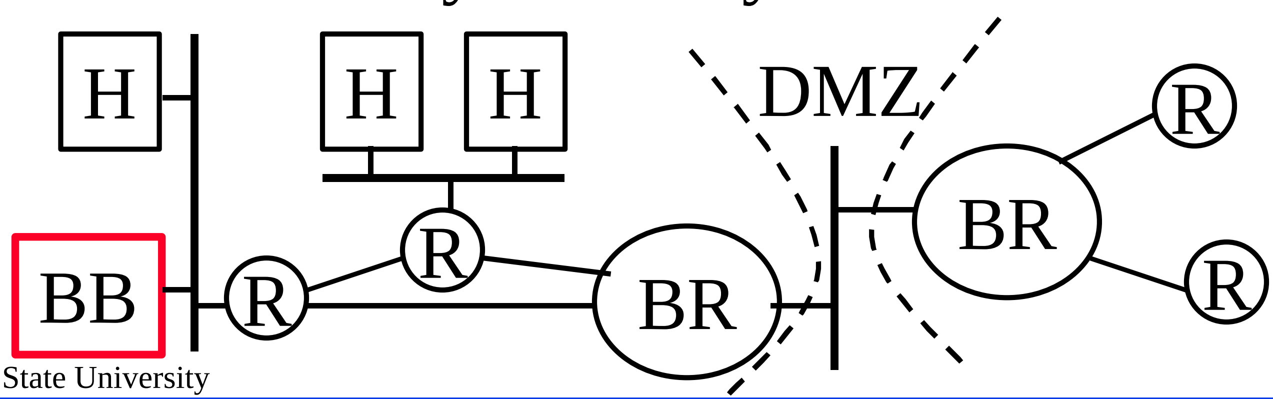
- Traffic Engineering = Performance Optimization
  ⇒ Maximum throughput, Min delay, min loss
  ⇒ Quality of service
- Traffic Engineering = Efficient resource allocation
  Minimize congestion, Path splitting
- In MPLS networks: "Traffic Trunks" = SVCs
  Traffic trunks are routable entities like VCs
- Each traffic trunk can have a set of associated characteristics, e.g., priority, preemption, policing
- Characteristics of packets assigned to that trunk determine its "Forwarding Equivalence Class (FEC)"

# Traffic Engineering (Cont)

- Trunk paths are setup based on policies or specified resource availability.
- A traffic trunk can have alternate sets of paths in case of failure of the main path. Trunks can be rerouted.
- Multiple trunks can be used in parallel to the same egress.
- Some trunks may preempt other trunks. A trunk can be preemptor, non-preemptor, preemptable, or non-preemptable.
- Each trunk can have its own overbooking rate

# Bandwidth Broker

- Repository of policy database. Includes authentication
- Users request bandwidth from BB
- BB sends authorizations to leaf/border routers
  Tells what to mark.
- Ideally, need to account for bandwidth usage along the path
- BB allocates only boundary or bottleneck



# IEEE 802.1D Model

| Dest Addr | Src Addr | Tag Prot ID | Pri | CFI | VLAN ID |
|---|---|---|---|---|---|

802.1Q header

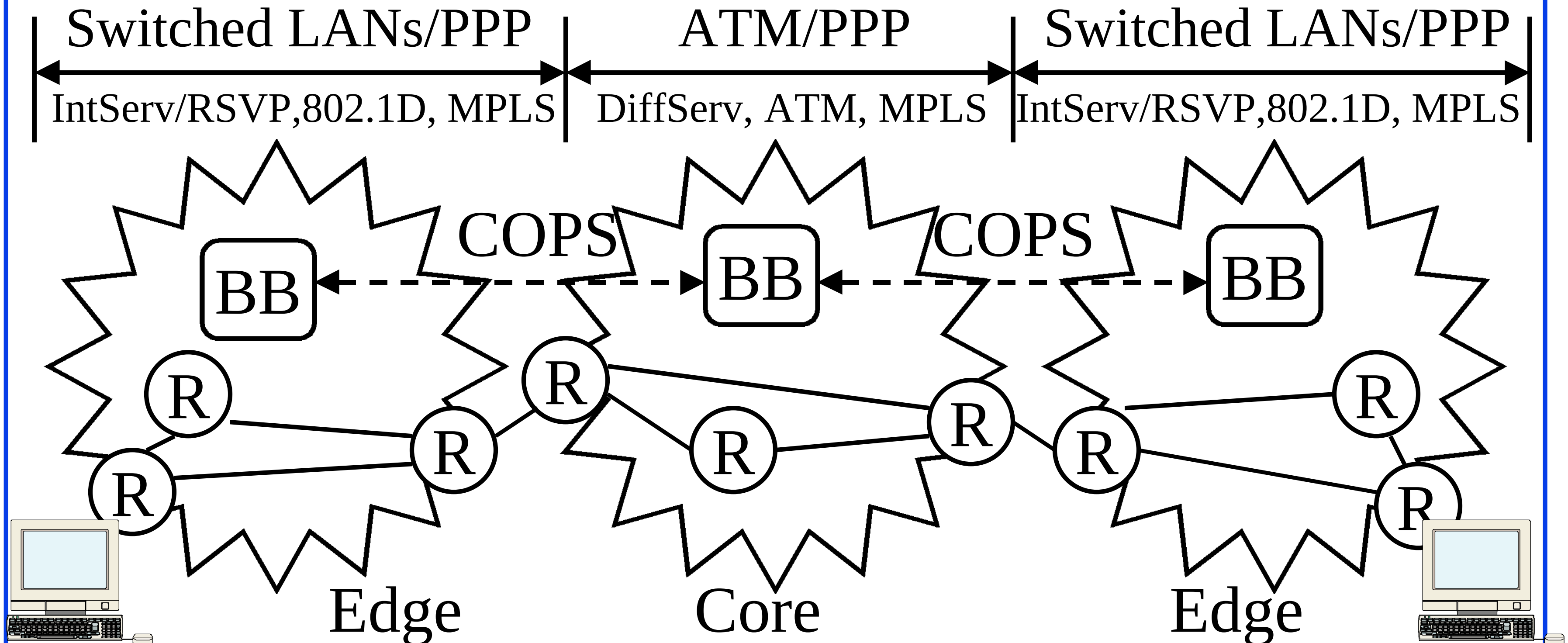
| Prot Type | Payload | FCS |
|---|---|---|

CFI = Canonical Format Indicator (Source Routing)

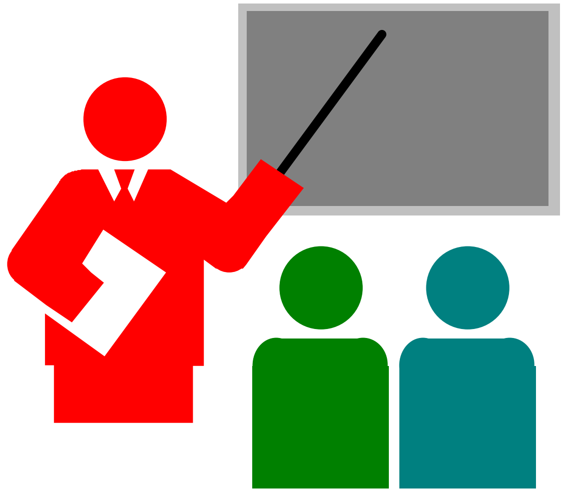
- **Up to eight priorities:** Strict.
  1 Background
  2 Spare
  0 Best Effort
  3 Excellent Effort
  4 Control load
  5 Video (Less than 100 ms latency and jitter)
  6 Voice (Less than 10 ms latency and jitter)
  7 Network Control

# End-to-end View

- ATM/PPP backbone, Switched LANs/PPP in Stub
- IntServ/RSVP, 802.1D, MPLS in Stub networks
- DiffServ, ATM, MPLS in the core



# Summary



- Integrated Services: GS = rtVBR, CLS = nrt-VBR
- Signaling protocol: RSVP
- Differentiated Services will use the DS byte
- MPLS allows traffic engineering

# Conclusions

- Multiple drop preferences does not help TCP (it does not care which packet is lost) or data in general. Will need to send probe packets with different drop preferences to sense the level of congestion and act accordingly.
- Multiple drop preferences does not help voice/video. Need multiple leaky bucket rates for layered/scalable coding.
- QoS = Weakest link in the chain
  $\Rightarrow$ All layers and all systems along the path need QoS support.

# Conclusions (Cont)

- Need additivity or mathematical aggregatability $\Rightarrow$ Simple deterministic guarantees (CBR) is easier to understand and aggregate.
  CBR was first step in telecommunications networks.
  CBR (EF) should be the first step for IP.
- Start with throughput guarantees.
  Fair allocation of excess throughput should be next.
  Delay is automatic with isolation.
- Coarse levels (factor of 2 to 10) of throughput guarantees will do.
- Two priorities will take a long way.

# Conclusions (Cont)

- Excess allocation is useful with closed loop (e.g., ABR). Dropping on the way (open loop) is not the right way.
- Network/application dynamics
  $\Rightarrow$ Need closed loop and active bandwidth management

# QoS Support By Vendors

| **Vendor** | **Criteria** | **Mechanism** |
|---|---|---|
| Aponet | L3 addresses, L4 Port, time | Priority |
| Bay networks | MAC, L3 addresses, Subnet, Protocol type, Switch port, VLAN | ATM, Priority, RSVP, 802.1 |
| Cabletron | MAC, protocol, L3 addresses, time, L4 port, VLAN | 802.1, IFMP, RSVP |
| Checkpoint | MAC, L3 addresses, protocol, L4 port, URLs | WFQ |
| Cisco | MAC, L3 addresses, protocol, switch or router port, L4 port | 802.1, IP Precedence, RSVP, Tag/MPLS, RED, WFQ |
| Class data | Application name, file name, L3 addresses, time, L4 port, URL, user name | 802.1, IP Precedence, RSVP |
| Digital | MAC, L3 addresses, switch port | Priority, 802.1 |
| Extreme networks | MAC, L3 addresses, subnet, switch port, L4 port, VLAN | 802.1, RSVP, QFW |
| Flowwise | MAC, L3 addresses, subnet, switch port, L4 port, VLAN | Priority, 802.1, RSVP |
| Fore | ATM, MAC, L3 addresses, subnet, Protocol type, switch port, L4 port, VLAN | ATM (per-VC Q), IP Precedence, RSVP |

| Vendor | Criteria | Mechanism |
|---|---|---|
| Foundry net | MAC, L3 addresses, subnet, L4 port, VLAN | Priority, 802.1,RSVP |
| IBM | ATM, MAC, L3 addresses, subnet, Protocol, switch port, VLAN | 802.1, ATM, ARIS/MPLS, RSVP |
| Ipsilon | L3 addresses, subnet, switch port, L4 port, VLAN | IFMP, RSVP, WRR, Priority, ATM (per-VC Q) |
| Newbridge | L3 addresses, Protocol, L4 port, VLAN | 802.1, ATM, RSVP |
| New oak | L3 addresses, L4 Port | RSVP, IFMP, Tag, WFQ, RED |
| Packeteer | L3 addresses, Protocol, L4 port, time, URLs | TCP/IP flow control |
| Prominet | MAC, L3 addresses, switch port, L4 port | 802.1, priority, RSVP |
| The Structure | L3 addresses, time, L4 port | TCP/IP flow control |
| 3Com | ATM, MAC, L3 addresses, switch port, L4 port, time | ATM, 802.1, priority, RSVP, QFQ, PACE |
| Torrent net | MAC, L3 addresses, switch port, L4 port | 802.1, priority, RSVP |
| Xedia | L3 addresses, device port, time, L4 port | CBQ, RSVP |
| Xylan | ATM, MAC, L3 addresses, switch port | 802.1, priority, RSVP |
| Yago | MAC, L3 addresses, switch port, L4 port | 802.1, priority, RSVP |

Ref: E. Roberts, "The New Class Systems," Data Communications, October 1997

# References

- For a detailed list of references see: http://www.cis.ohio-state.edu/~jain/refs/ipqs_ref.htm

# List of Acronyms

| | |
|---|---|
| ABR | Available Bit Rate |
| ATM | Asynchronous Transfer Mode |
| BA | Behavior Aggregate |
| BGP | Border Gateway Protocol |
| BOF | Birds of a Feather |
| CBR | Constant Bit Rate |
| CDV | Cell Delay Variation |
| CFI | Canonical Format Indicator |
| CLP | Cell Loss Priority |
| CLS | Controlled Load Service |
| COPS | Common Open Policy Service Protocol |

# Acronyms (Cont)

| | |
|---|---|
| CoS | Class of Service |
| DA | Destination Address |
| DQDB | Distributed Queue Dual Bus |
| DSBM | Designated Subnet Bandwidth Manager |
| DVMRP | Distance Vector Routing Multicast Protocol |
| FCS | Frame Check Sequence |
| FDDI | Fiber Distributed Data Interface |
| FIFO | First in First out |
| FTP | File Transfer Protocol |
| GS | Guaranteed Service |
| ICMP | Internet Control Message Protocol |

# Acronyms (Cont)

| | |
|---|---|
| IEEE | Institution of Electrical and Electronic Engineers |
| IETF | Internet Engineering Task Force |
| IGMP | Internet Group Management Protocol |
| IP | Internet Protocol |
| IPv4 | Internet Protocol Version 4 |
| IPv6 | Internet Protocol Version 6 |
| IS | Internal System |
| IntServ | Integrated Services |
| LANs | Local Area Networks |
| LLC | Logical Link Control |
| LU | Local Use |

# Acronyms (Cont)

| | |
|---|---|
| MAC | Media Access Control |
| MBONE | Multicast Backbone |
| MBS | Maximum Burst Size |
| MF | Multi-field |
| MPLS | Multiprotocol Label Switching |
| MTU | Maximum Transmission Unit |
| NHRP | Next Hop Resolution Protocol |
| OOPS | Open Outsourcing Policy Service |
| OSPF | Open Shortest Path First |
| PASTE | Provider Architecture for Differentiated Services and Traffic Engineering |

# Acronyms (Cont)

| | |
|---|---|
| PCR | Peak Cell Rate |
| PHB | Per-Hop Behavior |
| PIM | Protocol Independent Multicast |
| PT | Protocol Type |
| QOSPF | QoS-OSPF |
| QoS | Quality of Service |
| RED | Random Early Discard |
| ResV | Reservation Request |
| RFC | Request for Comment |
| RIF | Routing Information Field |
| RSVP | Resource Reservation Protocol |

# Acronyms (Cont)

| | |
|---|---|
| RSpec | QoS Specification |
| RTP | Real-time Transport Protocol |
| SBM | Subnet Bandwidth Manager |
| SONET | Synchronous Optical Network |
| TCP | Transmission Control Protocol |
| TPID | Tag Protocol ID |
| TR | Token Ring |
| TSpec | Traffic Specification |
| ToS | Type of Service |
| UBR | Unspecified Bit Rate |
| UDP | User Datagram Protocol |

# Acronyms (Cont)

| | |
|---|---|
| UNI | User-Network Interface |
| VBR | Variable Bit Rate |
| VC | Virtual Circuit |
| VLAN | Virtual Local Area Network |
| WAN | Wide Area Network |
| WFQ | Weighted Fair Queueing |