IEEE INFOCOM 2001

Tutorial T3

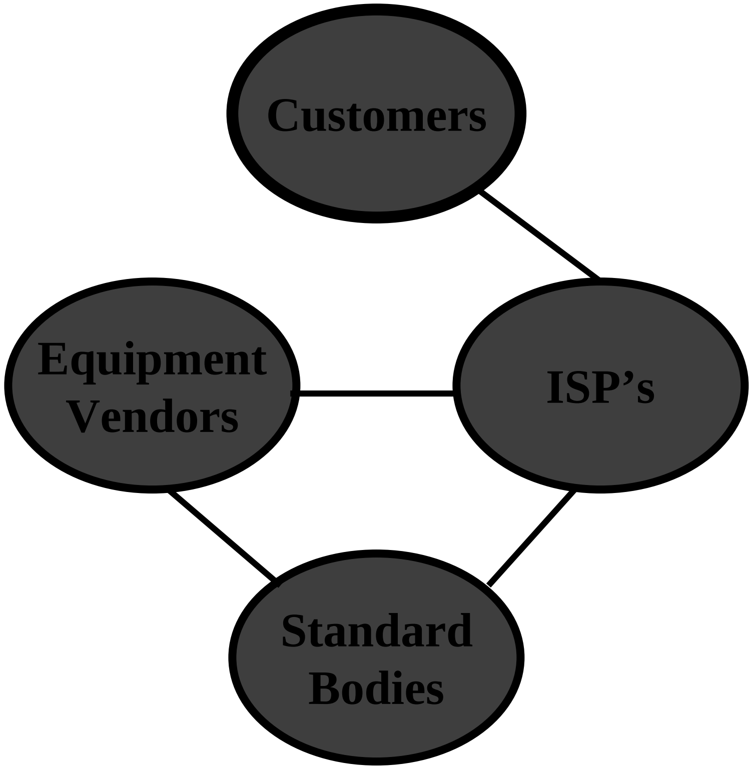
Traffic Engineering in IP/MPLS Networks

Anwar Elwalid (Bell Labs, Lucent Technologies)

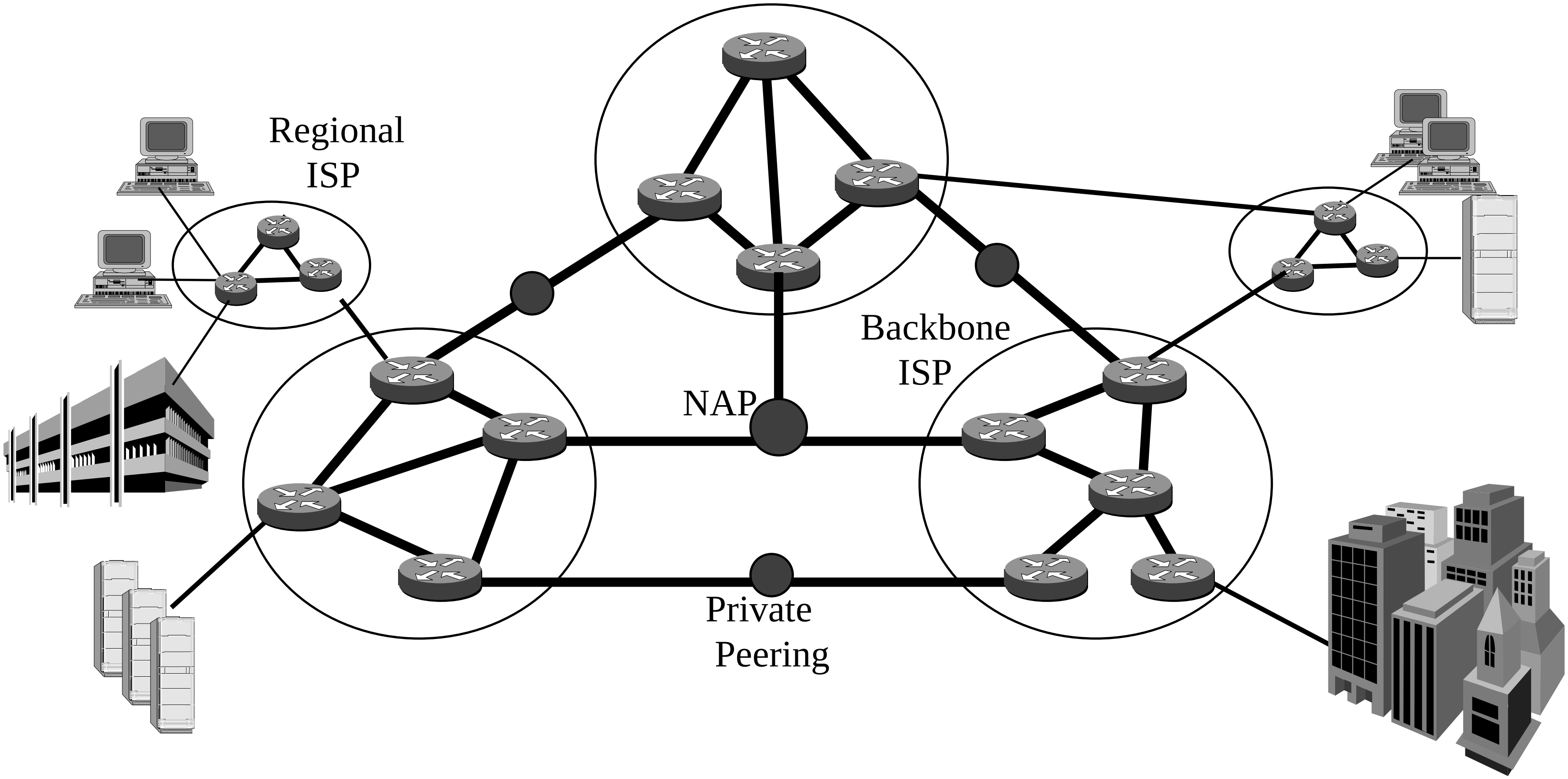
Sunday, 22 April, 2001 - Afternoon

Driving Forces for next generation Internet

- Customer
  - Demand for fast, reliable and differentiated service
- Internet Service Providers (ISPs):
  - Competition
  - Service Level Agreements (SLA’s)
  - Traffic Engineering (TE)
- Equipment Vendors:
  - New Technologies
- Standard Bodies:
  - Internet Engineering Task Force (IETF)



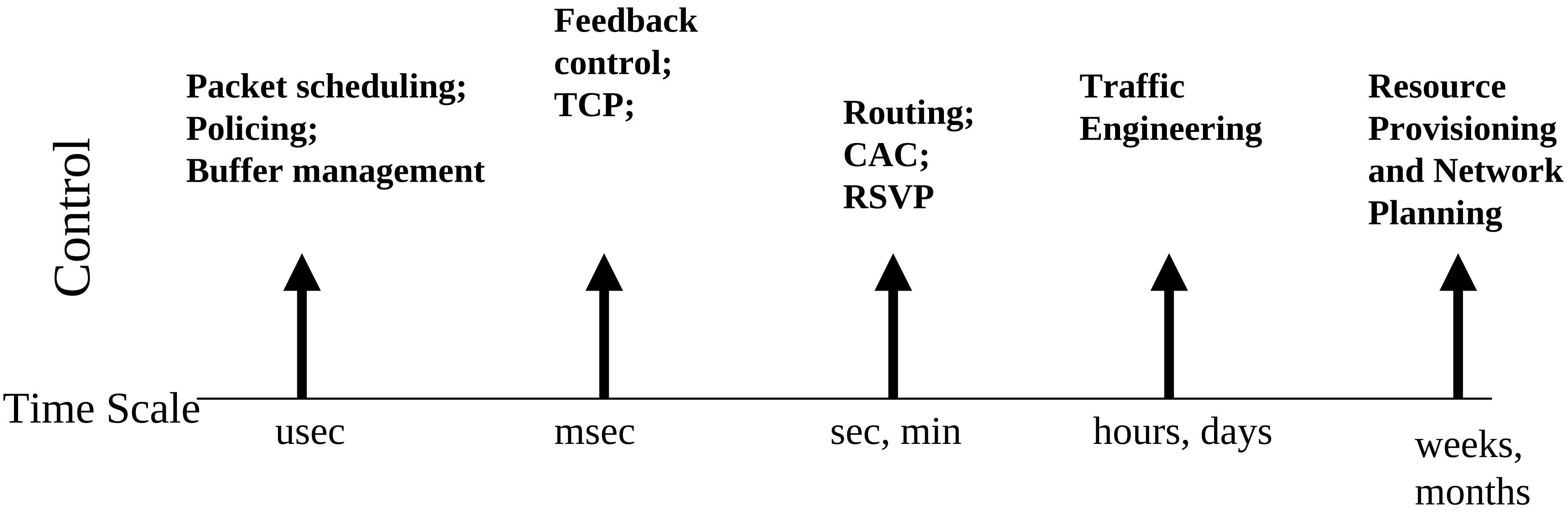
Internet Physical Architecture



Service Level Agreements (SLA’s)

- A contract between an ISP and a customer: the ISP provides service with certain QoS at specified cost
- Customer: dial-up user, enterprise or another ISP
- Service:
  - Spatial characteristics:
    - “Pipe” - point-to-point
    - “Hose” - point-to-multipoint
    - “Cloud” - Multipoint-to-multipoint
  - Traffic characteristics:
    - e.g. specified by leak bucket; voice with certain coding rate
  - QoS characteristics:
    - Bandwidth: fixed, burstable
    - End-to-end delay, loss
    - Service availability

Time-Scale Classification of TE & QoS Managment



Control

Time Scale

Packet scheduling;
Policing;
Buffer management

usec

Feedback
control;
TCP;

msec

Routing;
CAC;
RSVP

sec, min

Traffic
Engineering

hours, days

Resource
Provisioning
and Network
Planning

weeks,
months

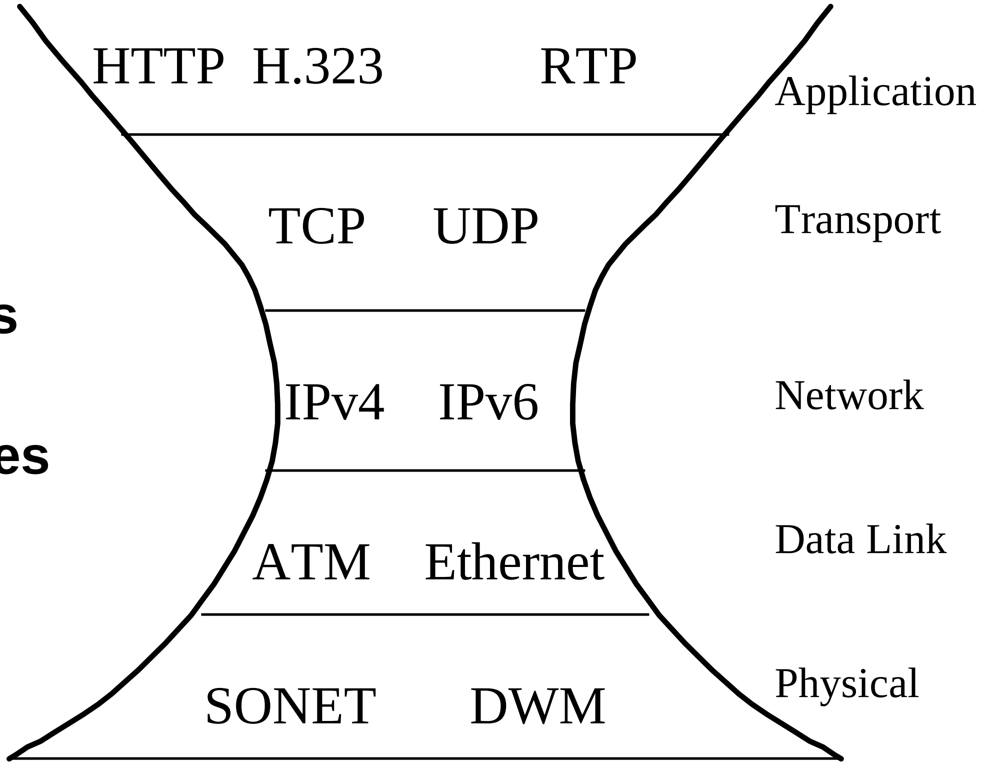
Road Map

- **TE and QoS Management in IP (datagram) Networks:**
  - **Routing**
  - **Congestion Control**
  - **Integrated Services (IntServ)**
  - **Reservation Protocol (RSVP)**
  - **Differentiated Services (DiffServ)**
  - **Scheduling and Admission Control in Routers**

- **TE and QoS Management in MPLS Networks**
  - **MPLS Architecture and Mechanisms**
  - **MPLS &DiffServ**
  - **Traffic Engineering Solutions**
  - **Virtual Private Networks**
  - **IP over Optical, MPλS, GMPLS**

Internet Protocol Architecture

- **Layered hourglass architecture**
- **IP supports different applications (e.g. VoIP, HTTP)**
- **IP runs over different technologies (IP over ATM, SONET, anything!)**
- **Connectionless network**
- **Stateless**



Internet Routing Protocols

- Internet is divided into autonomous systems (AS’s)
  - AS: set of routers administered by a single organization (e.g. ISP)
- Routing protocols divided into two types of protocols:
  - Interior Gateway Protocols (IGP’s) provide routing within an AS
    - Routing Information Protocol (RIP): distance vector protocol using Bellman-Ford algorithm
    - ISIS
    - Open-Shortest Path First (OSPF): link-state protocol using Dijkstra algorithm
  - Exterior Gateway Protocols (EGP’s) provide routing among AS’s
    - Border Gateway Protocol (BGP): path-vector protocol

OSPF Basics

- Each router floods link-state advertisements (LSAs) to the network
- From link-state information, each router knows the topology of the network
- Each router computes the shortest path to each destination using Dijkstra’s Algorithm
- Each router maintains routing table of next hop for each destination
- Scalability is achieved by using sub-areas

Review of Routing/TE Approaches in IP networks

- **Shortest Path First (SPF**):
  - Static link metric: proportional to the inverse of link capacity
    - Issue: routing oblivious to traffic demands and QoS ==> “super aggregation problem”
  - Dynamic link metric: function of some congestion measure
    - Issue: instability and service disruption
- **Equal Cost MultiPath (ECMP):** Unlike SPF, distributes traffic equally among equal cost paths
    - Issue: routing oblivious to traffic demands and QoS and paths with comparable costs are not considered
- **TOS Routing:** Different routes selected based on the packet Type of Service (TOS) field
  - issue: oblivious to traffic demands and not universally implemented
- **Overlay Model (e.g. IP over ATM):**
  - Traffic engineering on ATM using virtual circuits
  - Complex logical topology for routers, and other drawbacks*(will be revisited later)*

Traffic Engineering by Optimizing OSPF Weights

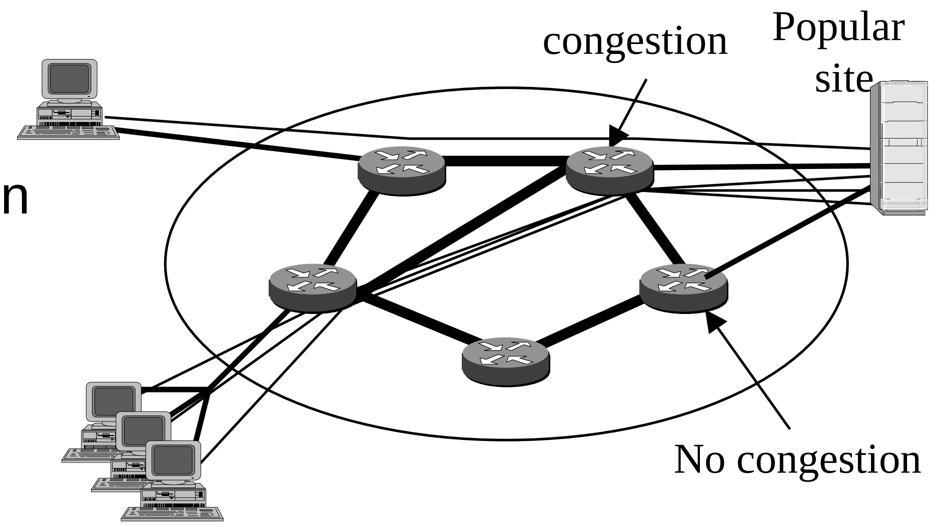
- **Given:**
  - a network with link capacities $\{c_l\}$
  - Source-destination demand matrix $\{\lambda_i^k\}$
- **Find:** the link weights $\{w_l\}$ for which OSPF (single path)routing leads to minimization of a global cost function, for example:

$$D = \sum_l \frac{f_l}{(c_l - f_l)}$$

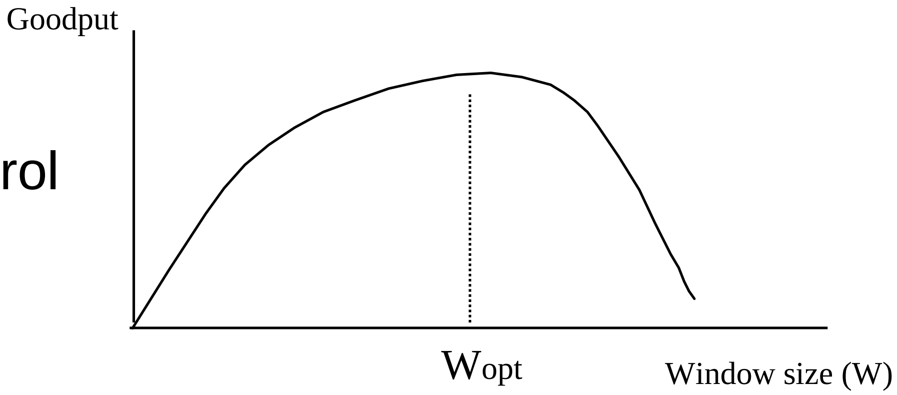
Where f_l is the resulting total flow on link *l*

- Heuristic Combinatorial algorithm: (Rodrigues and Ramakrishnan)
- **Variations:**
  - **Equal-cost multi-path**
  - **Inverse-shortest path:** find explicit paths - harder problem
- Solution could be far from optimal (general) routing
- **Implementation issue**:could lead to global service disruption

Congestion Control in the Internet



- Routing oblivious to congestion
- Intelligence at end points
- Stateless Network
- Best Effort Service

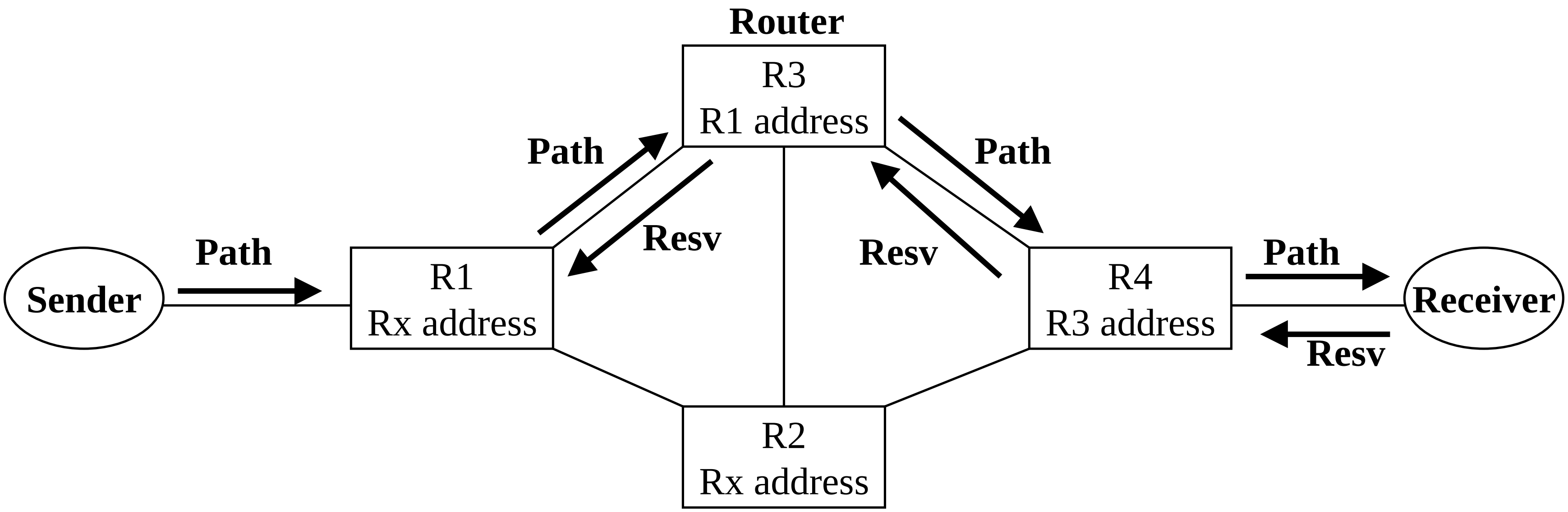


- Objective:
  - Dynamic Window Flow Control
  - Implicit feedback
  - End-to-End control
  - Fairness to users

Integrated Services (IntServ)

- Provides New IP service models beyond the best-effort service:
  - **Controlled-load service** : the QoS that a flow gets when the network is overloaded is comparable to the QoS it gets when the network is under-loaded
  - **Guaranteed service**: the flow is guaranteed bounded delay
- Requires new router mechanisms:
  - Admission Control
  - Traffic Control
    - Packet classifier: assign packet to service classes
    - Packet scheduler: forward packets according to class requirements
- Requires signaling Protocol:
  - **RSVP**: protocol to set-up and tear down resource reservations in the routers

RSVP



- Signaling protocol for establishing “*soft state*” per flow
- Path messages from sender to receiver carry flow traffic characteristics (Tspec) and path information
- Resv messages (Flowspec) from receiver to routers along the path carry Tspec and resource reservation requests Rspec - Receiver-initiated reservations

- Flow: generally specified by:
    - 5-tuple: source and destination addresses, source and destination port numbers and protocol ID
    - Bounded inter-packet times
- “Soft State”:
    - Path and Resv state information at routers are cleared after a timer expiration
    - State refresh messages are sent periodically
    - Advantages:
        - Robust adaptation to route changes and lost signaling messages
- Admission Control:
    - Routers decide if a reservation can be accommodated and assign appropriate packet classification and scheduling.

RSVP Reservation Styles

- *Wildcard-Filter (WF) style* creates a single reservation for all flows from upstream senders
- *Fixed-Filter (FF) style* – creates a distinct reservation for selected senders
- *Shared Explicit (SE) style* – creates a shared reservation for selected senders

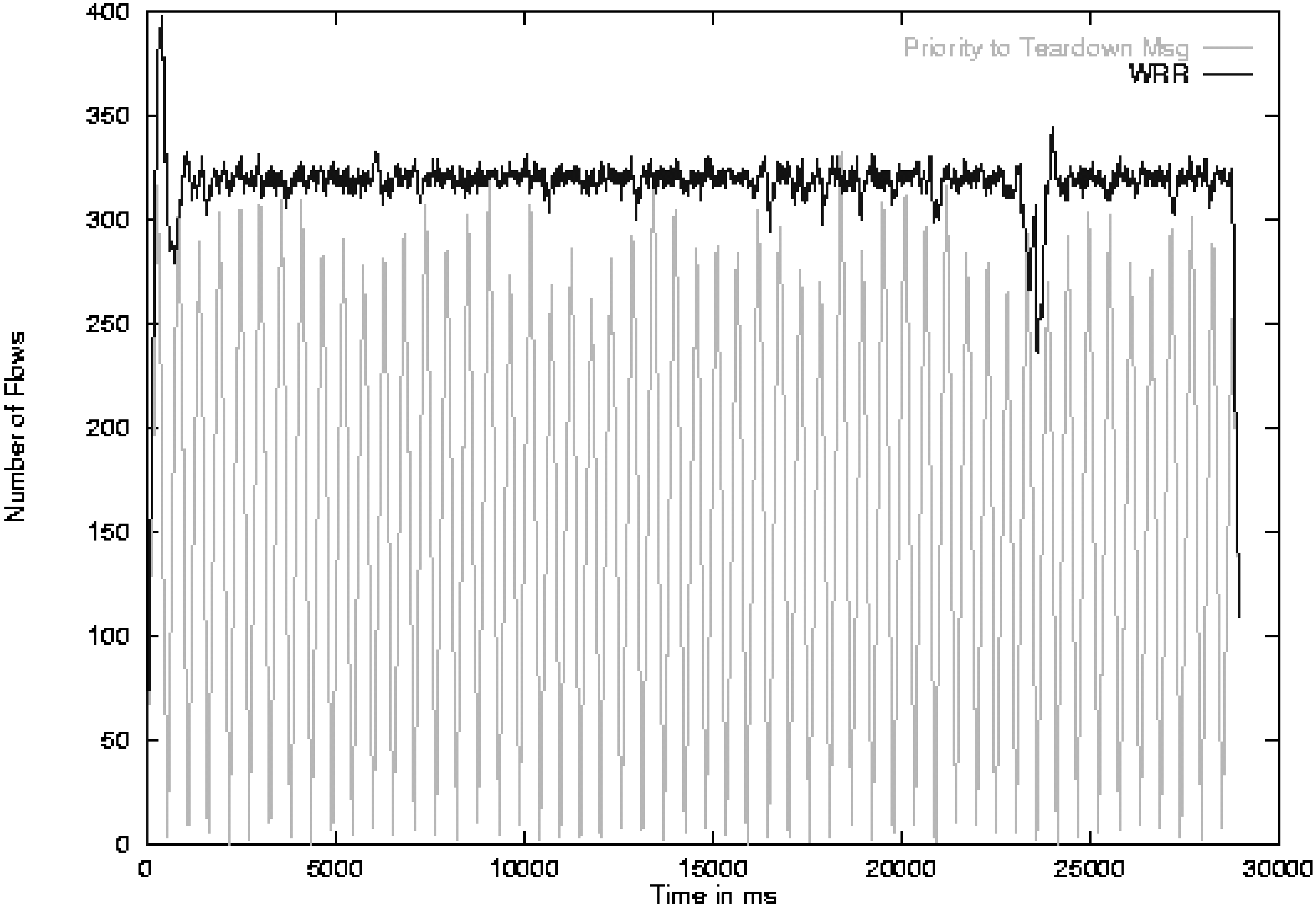
Issues with IntServ

- Scalability issues:
  - In large network with many flows:
    - Per-flow state maintenance in routers is expensive
    - Large overhead associated with processing of signaling packets which could affect routing function
- Scaling RSVP
  - Message Bundling
  - RSVP Summary Refresh
  - RSVP Message Acknowledgement
- We now examine and give approaches for:
  - RSVP message processing
  - Connection admission control and packet scheduling

Study of RSVP Message Processing Mechanisms

- One central CPU per interface which processes routing and control messages
- Control messages: Path, Resv, , Update, Tear-Down
- Observation: FIFO processing of control messages could lead to:
  - Reservation blocking when link bandwidth is available
  - Oscillating link utilization
- Proposed Solution:
  - Adaptive Weighted Round Robin scheduling
  - Weight for update messages is function of established flows
  - Weights for Path/Resv and Tear-Down messages are functions of link utilization and request sizes
- (M. May, T.V. Lakshman and A. Elwalid) NOSSDAV ‘98, Cambridge.

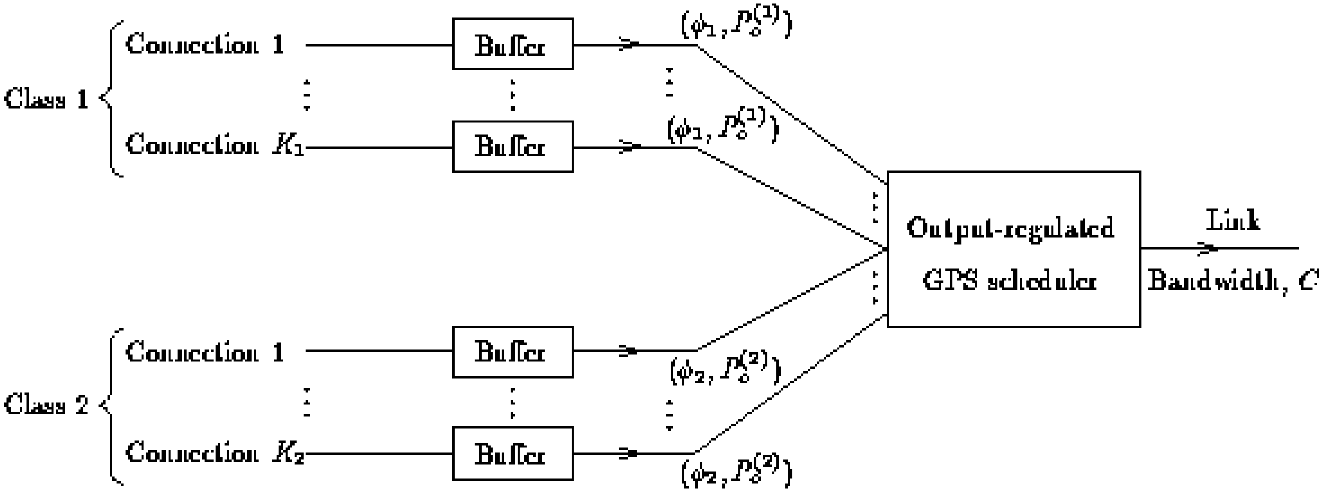
Comparison of FIFO Scheduing with Adaptive WRR



Design of Admission Control and Generalized Processor Sharing (GPS) Schedulers with Statistical Multiplexing

- Consider 3 service classes Gold, Silver, Bronze (best effort):
- Design of admission control and GPS scheduling: *coupled* problem
- Objective:
  - design the GPS weights to maximize resource utilization (flow-carrying capacity)
  - Simple admission control rules
- Features of the solution
  - No reliance on statistical source models
  - Exploits statistical multiplexing
  - Simple: No more than two sets of GPS weights are needed
- (Elwalid and Mitra INFOCOM’99)

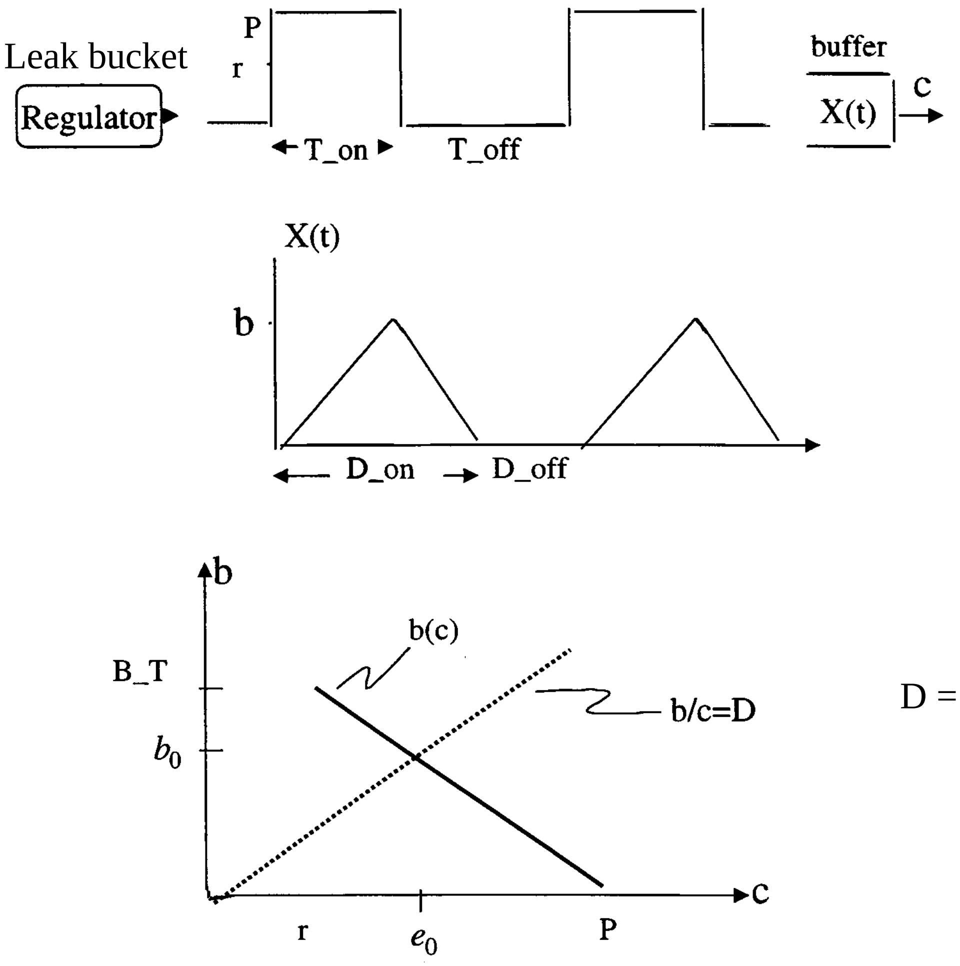
GPS Schedulers



- Output-regulation: allocates only the bandwidth needed for QoS; excess to best-effort
- QoS parameters for class *j :* delay bound and violation probability
- Design parameters: set of weights $\{\phi_j\}$ and set of peak rates $\{P_0^{(j)}\}$

Deterministic Analysis

- Consider a single connection



D = max delay

e_0 is the lossless effective bandwidth

Deterministic QoS Guarantees

Consider K_1 sources of class 1 and K_2 sources of class 2 in the system. The requirements of a source of class j (j=1,2) is satisfied if

$$e_0^{(j)} \leq \frac{\phi_j}{\phi_1 \sum_{i=1}^{K_1} \xi_i^{(1)} + \phi_2 \sum_{i=1}^{K_2} \xi_i^{(2)}} C,$$

where $\xi_i^{(j)}$ is an indicator variable for the activity of source i of class j.

- For lossless multiplexing, we have

$$\phi_1 = e_0^{(1)}, \quad \phi_2 = e_0^{(2)}$$

which give the realizable admissible set:

$$K_1 e_0^{(1)} + K_2 e_0^{(2)} \leq C$$

- For peak-rate regulation, we set $P_o^{(j)} = e_0^{(j)}$.

Statistical QoS Guarantees

The demand of source i of class j for bandwidth is an on-off process taking values $e_0^{(j)}$ and 0, independently from other sources. Hence,

$$\Pr(\xi_i^{(j)} = 1) = 1 - \Pr(\xi_i^{(j)} = 0) = \frac{D_{on}^{(j)}}{D_{on}^{(j)} + D_{off}^{(j)}}$$

- Let L_j (small) be the violation probability for class j.

$$\Pr\left\{ e_0^{(1)} > \frac{\phi_1}{\phi_1 \sum_{i=1}^{K_1} \xi_i^{(1)} + \phi_2 \sum_{i=1}^{K_2} \xi_i^{(2)}} C \right\} \leq L_1$$

$$\Pr\left\{ e_0^{(2)} > \frac{\phi_2}{\phi_1 \sum_{i=1}^{K_1} \xi_i^{(1)} + \phi_2 \sum_{i=1}^{K_2} \xi_i^{(2)}} C \right\} \leq L_2$$

$$\mathsf{Pr}\left[\phi_1 \sum_{i=1}^{K_1} \xi_i^{(1)} + \phi_2 \sum_{i=1}^{K_2} \xi_i^{(2)} > \phi_1 (C/e_0^{(1)})\right] \leq L_1$$

$$\mathsf{Pr}\left[\phi_1 \sum_{i=1}^{K_1} \xi_i^{(1)} + \phi_2 \sum_{i=1}^{K_2} \xi_i^{(2)} > \phi_2 (C/e_0^{(2)})\right] \leq L_2$$

- Use Chernoff Bound to approximate the probabilities

Let $\mathcal{A}^{(1)}(\phi) \triangleq \{(K_1, K_2)$: QoS for class 1 is satisfied given $\phi\}$ $\qquad \phi = \dfrac{\phi_1}{\phi_2}$

Let $\mathcal{A}^{(2)}(\phi) \triangleq \{(K_1, K_2)$: QoS for class 2 is satisfied given $\phi\}$

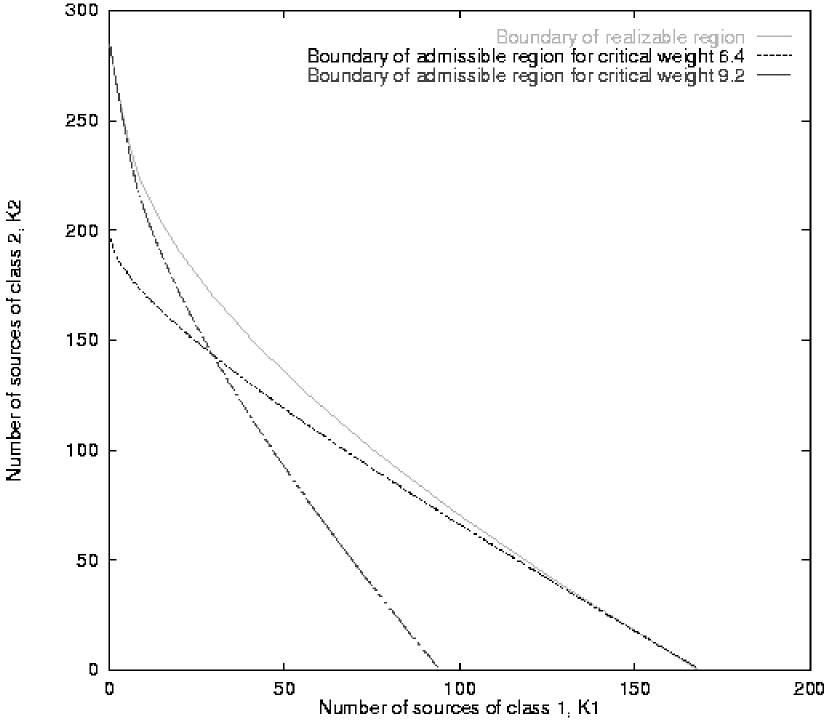
- Admissible Set: $\mathcal{A}(\phi) \triangleq \mathcal{A}^{(1)}(\phi) \cap \mathcal{A}^{(2)}(\phi)$
- Realizable Set: $\mathcal{R} \triangleq \bigcup_{\phi} \mathcal{A}(\phi)$
- Two Critical weights

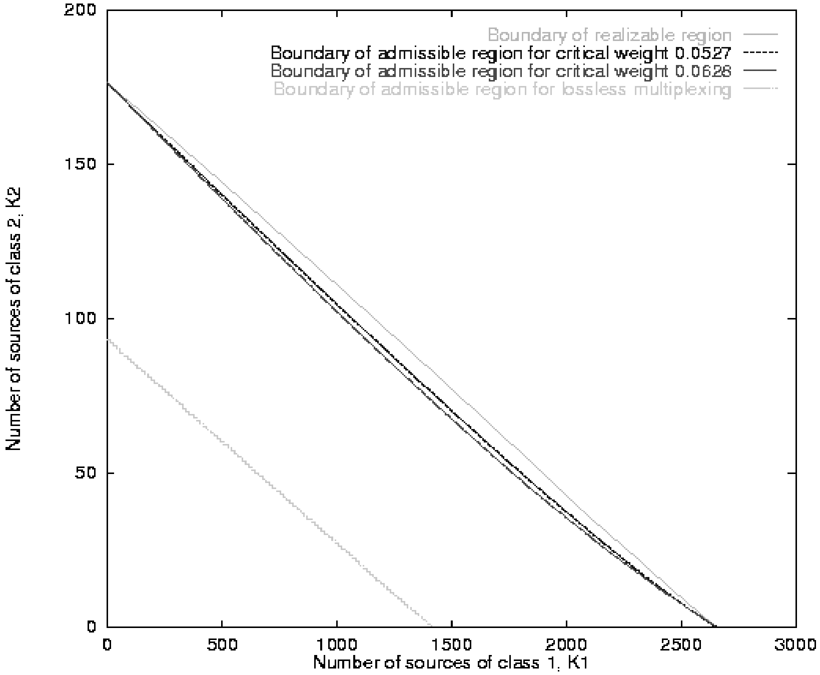
Numerical Examples

	Case 1		Case 2		Case 3	
	class 1	class 2	class 1	class 2	class 1	class 2
Q (cells)	250	250	25	250	100	1000
P (Mbps)	1.5	1.5	1.5	1.5	0.032	1.5
r (Mbps)	0.15	0.15	0.15	0.15	0.016	0.15
L	10^{-3}	10^{-9}	10^{-3}	10^{-9}	10^{-3}	10^{-9}
D (sec)	0.001	0.6	0.01	0.01	0.01	0.6
$e_0^{(\cdot)}$ (Mbps)	1.48	0.158	0.62	1.31	0.0318	0.48

class1 in case 3 corresponds to voice traffic

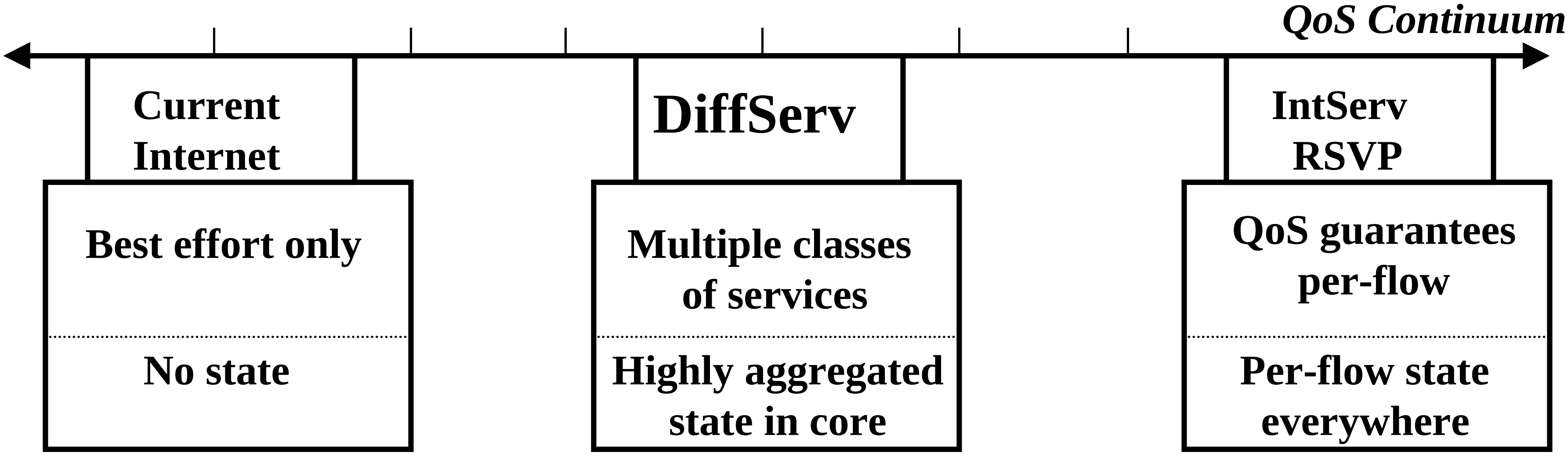
$$\text{Weight} = \frac{\phi_1}{\phi_2}$$





Differentiated Services (DiffServ)

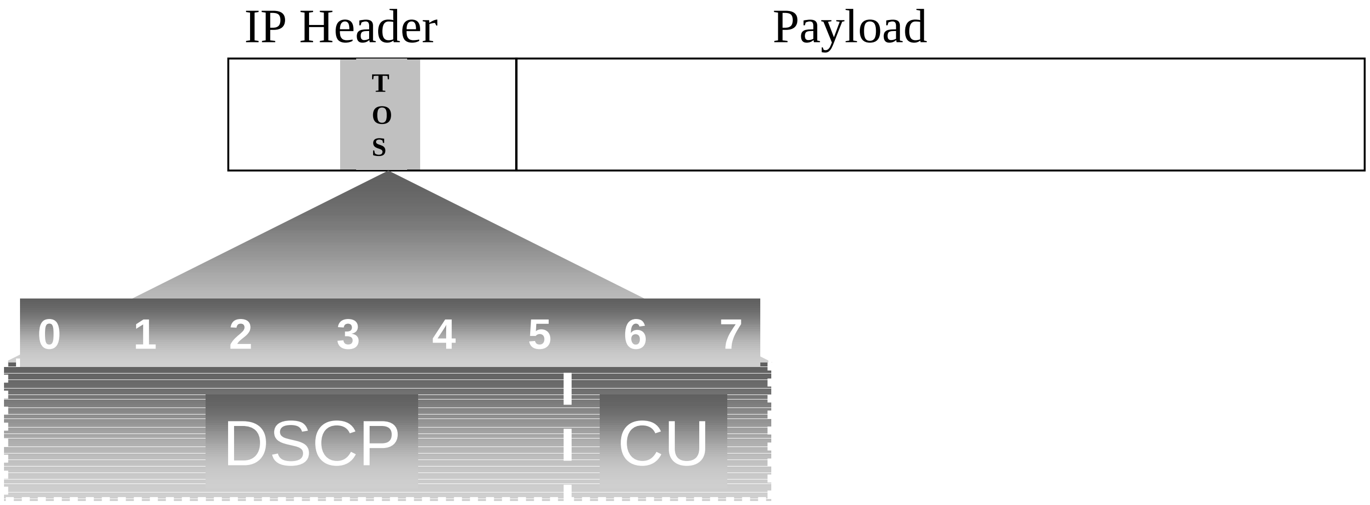
- Objective: Provide scalable service differentiation in the Internet without the need for per-flow state and signaling at every hop.
- Simplicity and scalability:
  - Maintain state for small number of traffic *aggregates*
  - Push complex processing to network edges
  - No hop-by-hop application signaling
- Decouples traffic conditioning and service provisioning from forwarding functions implemented within the core network.



DiffServ Architecture

- Service level agreement (SLAs) between a customer and an ISP which includes traffic conditioning agreement (TCA)
- Static ,dynamic
- A Packet is assigned a “ DS code point” which corresponds to a particular packet-forwarding treatment at each router - per hop behavior (PHB)
- PHBs are implemented in nodes using buffer management and packet scheduling mechanisms. PHBs are defined in terms of behavior characteristics relevant to service provisioning policies, and not in terms of particular implementation mechanisms.
- ISP enforces TCA at its network ingress points

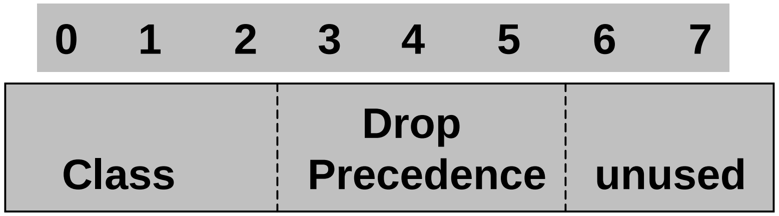
DiffServ : DSCP



- DSCP : Differentiated Service Code Point = 6 bits
- DSCP field defines Per-Hop Behavior (PHB), i.e., encodes which treatment the packet should receive
- CU: Currently Unused = 2 bits (potentially for ECN)

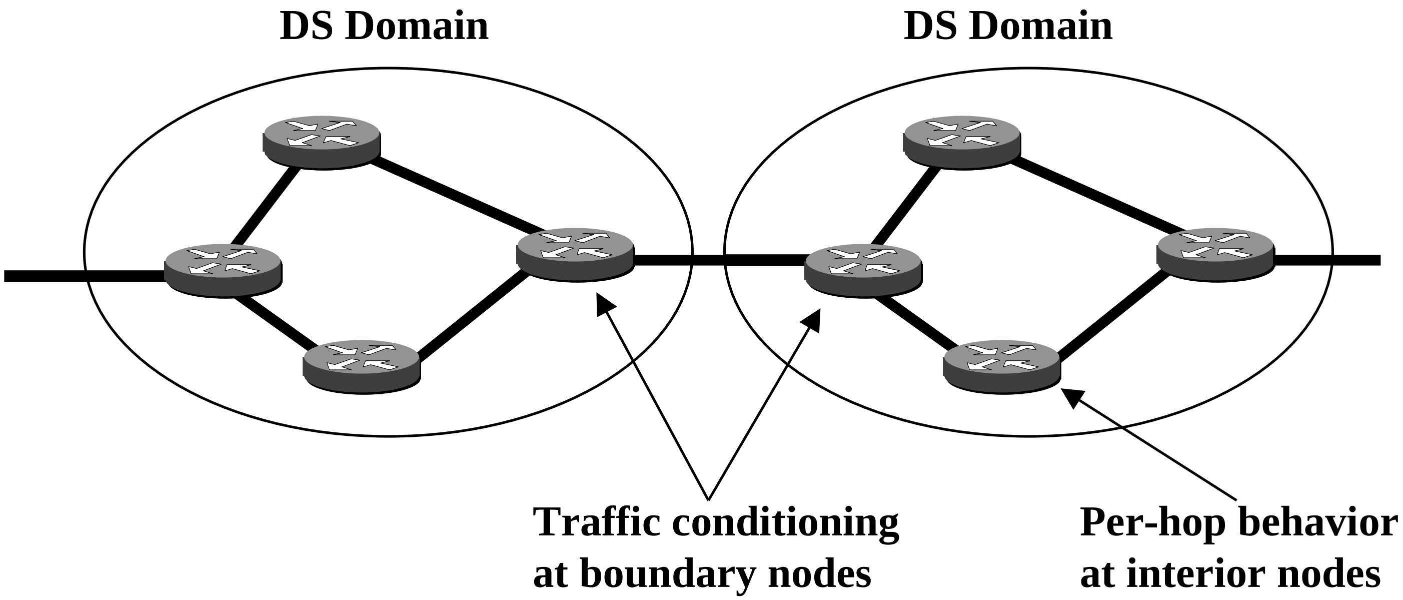
Per-Hop Behaviors

- Default PHB: best-effort
- Expedited Forwarding (EF PHB)
  - Provides service equivalent to “virtual leased line” - assured bandwidth, low jitter
  - packets must be policed/shaped at ingress; non conforming packets are discarded
- Assured Forwarding (AF PHB)
  - Provides in-profile traffic a high probability of delivery
  - Four levels of forwarding assurances
  - Within each level packets are assigned one of three possible drop-precedence values (priorities)



0	1	2	3	4	5	6	7
Class			Drop Precedence			unused	

DiffServ Model

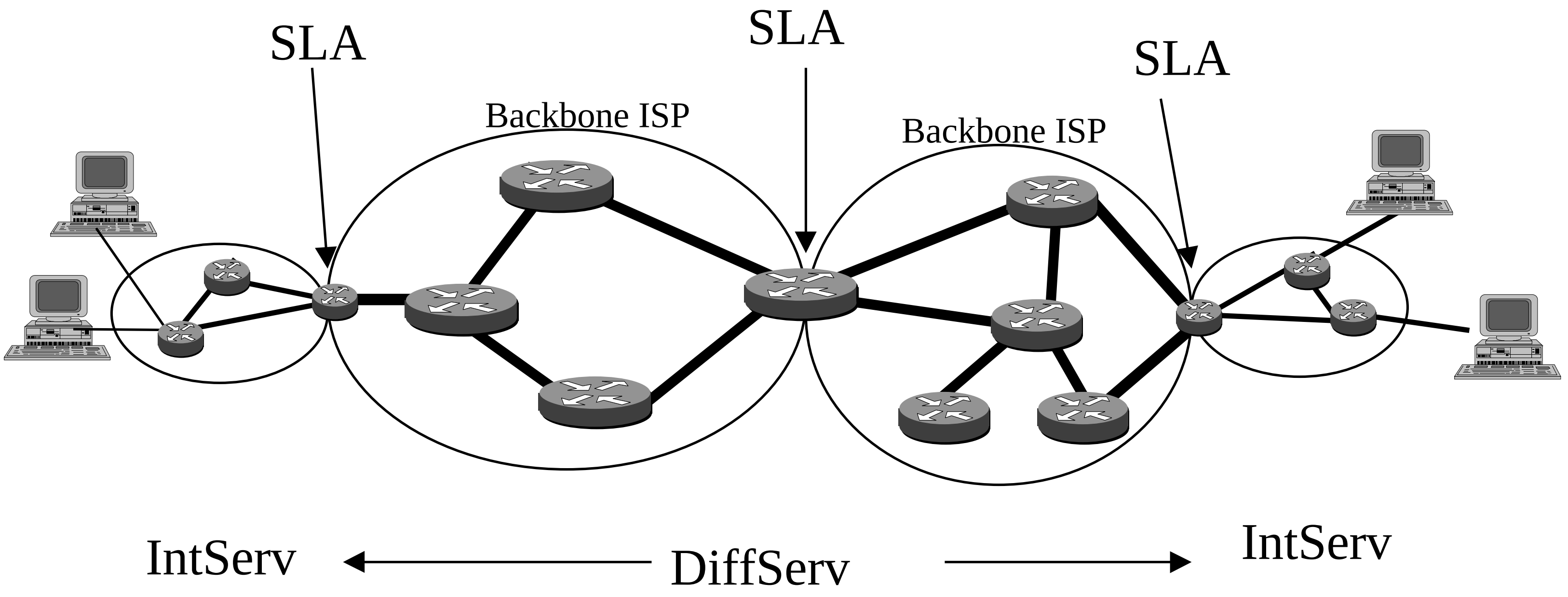


•DS Administrators set up DS-capable routers within their domain for conditioning and PHB per service class

•Service = Conditioning + Behaviors

•Service for a given DiffServ category (eg. Gold) in one domain is not necessarily the same as in another domain. Policy-driven approach is seen as a good mechanism to achieve end to end consistency

Integration of IntServ and DiffServ



SLA

SLA

SLA

Backbone ISP

Backbone ISP

IntServ

DiffServ

IntServ

- Edge Network: Per customer/flow state
- Core Network: Per class state

Key issue remains: Traffic path is based on destination forwarding

Traffic Engineering

- Mapping of traffic demands to network topology efficiently while meeting SLA’s.

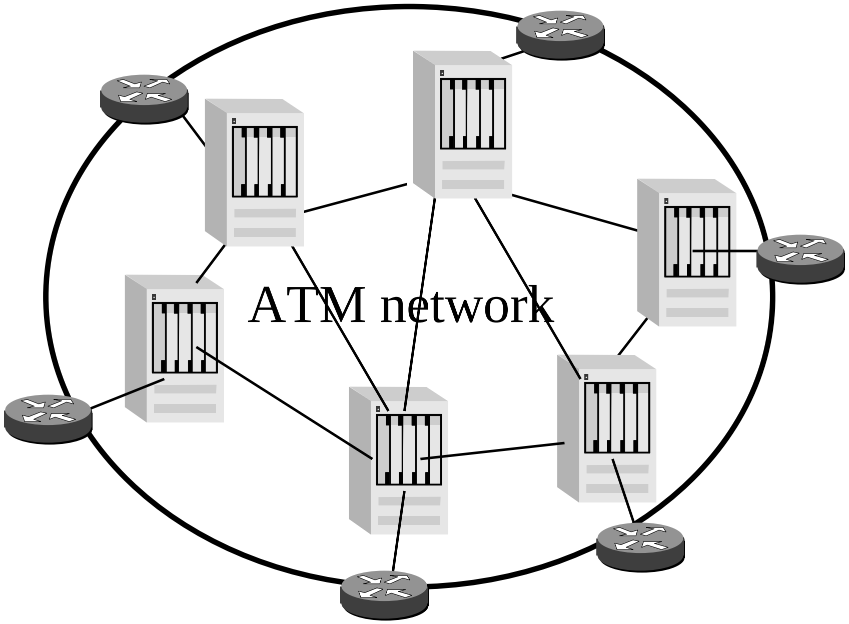
- Macroscopic and longer term view of the network

- Mechanisms at shorter time scale (TCP, IntServ, DiffServ, …) deal with congestion when it occurs by blocking, shaping and discriminating among flows

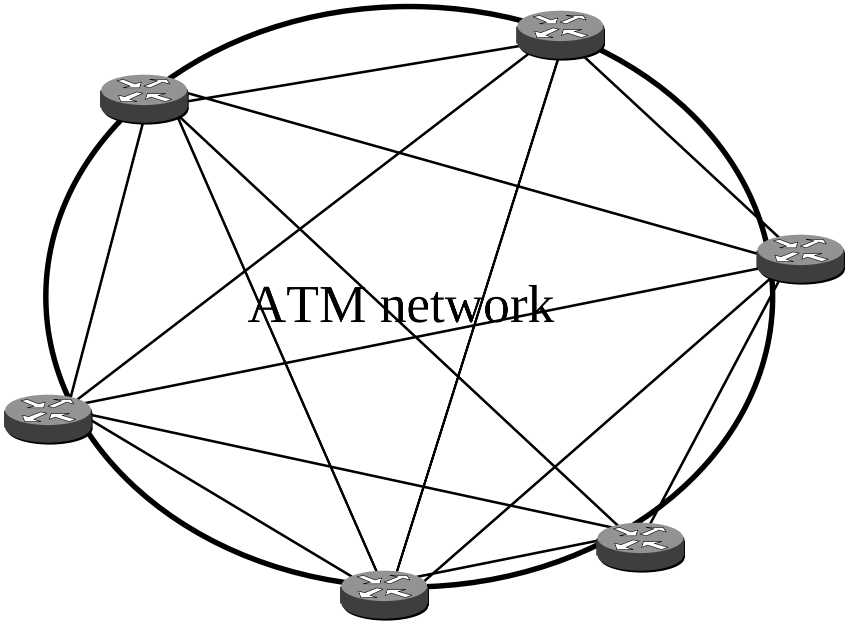
- Classifications:
  - Time-dependent vs. State-dependent
  - Off-network vs. On-network
  - Centralized vs Distributed

IP over ATM

Traffic Engineering in IP over ATM (overlay) Networks



Physical Topology



Logical Topology

- PVC are setup between POPs
- IP packets are encapsulated using ATM Adaptation Layer (AAL) 5
- Dense PVC meshes
- ATM network is engineered

IP/ATM Overlay Model:
Disadvantages

- Management complexity- (two networks to manage)
- ATM SAR interfaces are not available for OC-48
- Bandwidth overhead (cell tax)
- Scalability (router adjacency problem):
  - Number of PVCs and number of IGP adjacencies grows by n^2
  - flooding information for link and node failures grow by higher order of n

MPLS

What is MPLS?

- Highly flexible technology for bringing new services to IP networks
- Integrates network layer routing and label switching in a Label Switching Router (LSR)
- LSR runs normal IP routing protocols just like a router
- LSR forwards packets by layer-2 label switching
- Layer 2 can be ATM, frame relay, ...
- Layer 3 can potentially support other protocols (AppleTalk, IPX, IPv6, …)
- LSR uses a label distribution protocol (RSVP/LDP) to map IP routing information to layer-2 labels

Multi-Protocol Label Switching

Benefits:

- Simplified and improved forwarding via label switching
- Simplified ISP backbone architecture (one protocol)
- Traffic engineering via explicit routes using efficient tunneling mechanism which doesn’t require the explicit route in the packet header
- Facilitates differentiated grades of service
- Facilitates VPN design
- Facilitates fast-rerouting via backup paths (aka protection paths)
- Solves the N^2 VC mesh network scalability problem by VC merge
- IETF Standards
- Supported by major vendors and network operators
- Being extended to support dynamic optical bandwidth provisioning

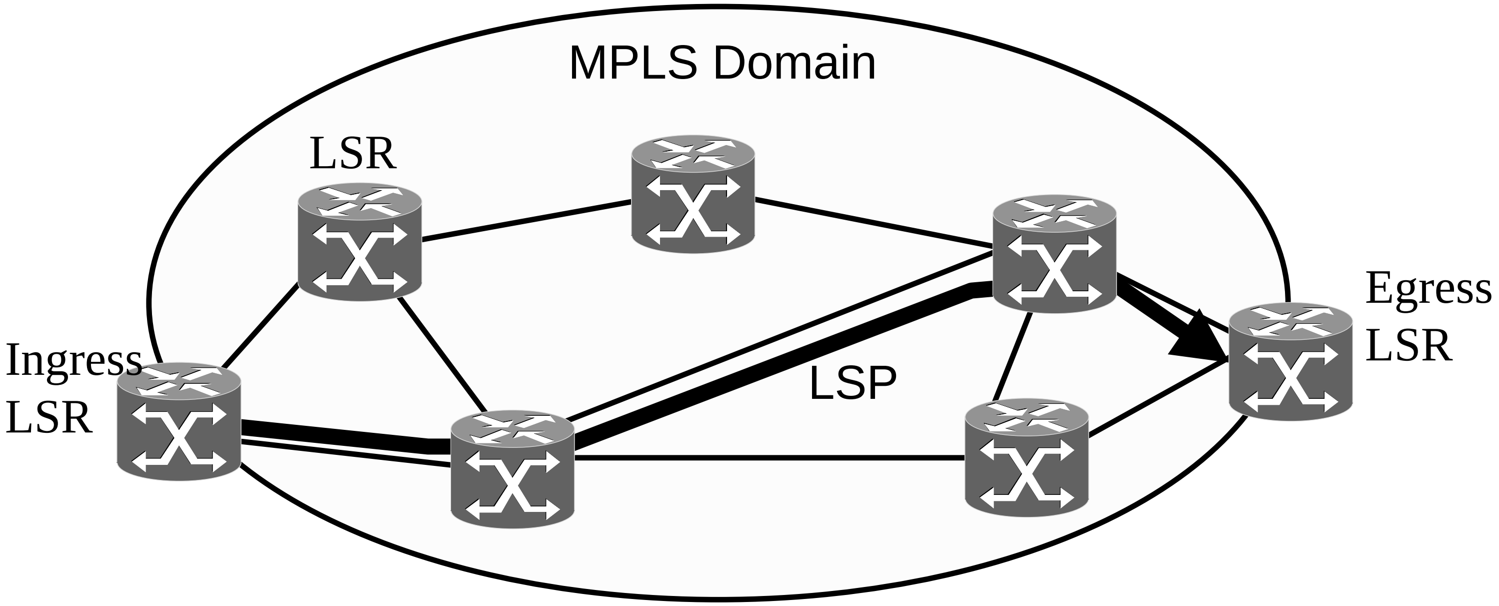
MPLS Architecture Principles

- **Forwarding Plane**
  - **Simple label swapping to forward packets along a Label Switched Path (LSP)**
  - **Map traffic to LSP based on “Forwarding Equivalence Class” (FEC)**
  - **MPLS forwarding Can be used over any packet link; when ATM label switching is used, label value is copied to the VCI field**
- **Control Plane**
  - **Multiple control functions (Traffic engineering, DiffServ, VPNs) influence label assignements**

Terminology

- **Label:** a short fixed length local identifier (e.g., Shim header,VPI/VCI)
- **Label Switching Router (LSR):** a node capable of forwarding packets based on labels, and aware of MPLS control protocols
- **Label Switched Path (LSP):** the path through one or more LSRs on which forwarding is done using labels (i.e., virtual circuit)
- **Label Stack:** an ordered set of labels within a packet
- **Forwarding Equivalence Class (FEC):** the set of packets which may be mapped to the same LSP
  - Examples of FEC
    - Application Flow: the finest level of granularity, best suited for local or campus networks
    - IP Prefix: middle of the road granularity, best suited for enterprise networks
    - Egress Router: the coarsest level of granularity, best suited for the core of the Internet (best scaling properties)
  - A FEC may be a function of destination address and QoS
- **Label Distribution Protocol (LDP):** a protocol that enables LSRs to establish LSPs by mapping FECs to labels

MPLS Forwarding Model



- (Label-Switching Router) LSR:
  - **Forwards MPLS packets using label-switching**
  - **Executes IP routing protocols and participates in MPLS control protocols**
- **Ingress LSR**:

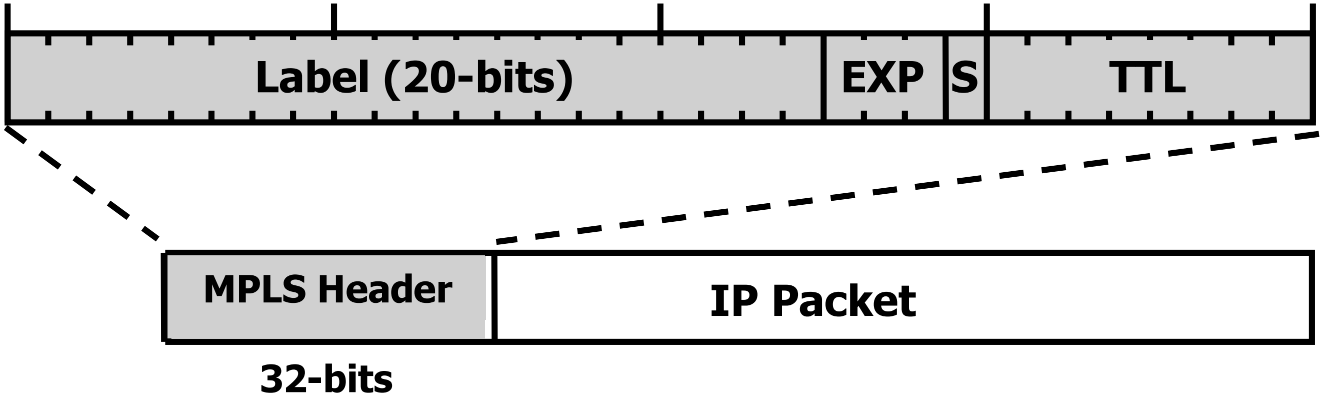
  - Assign incoming IP packets to a FEC

  - Generates MPLS header and assigns (binds) initial label to FEC
- **Egress LSR**:

  - Removes the MPLS header

  - Forwards packets based on IP destination address

MPLS (Shim) Header



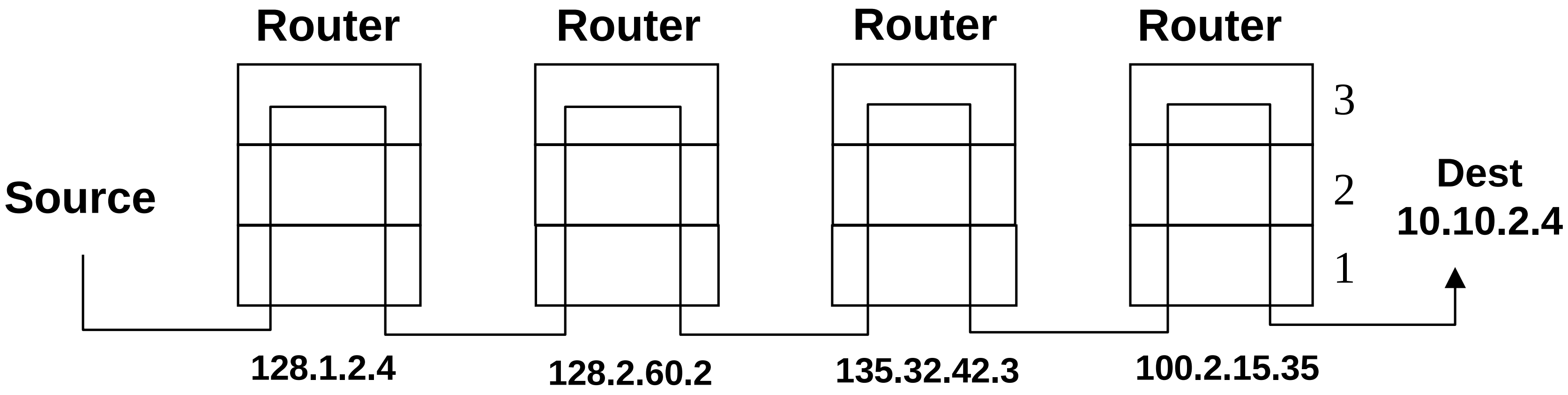
- **Fields**
  - **Label**
  - **Experimental bits (can be used for CoS mapping)**
  - **Stacking bit: indicates bottom of label stack when set**
  - **Time to live (copied from IP TTL)**

More on labels

- Actually a stack of labels
- Successful look up of top label determines:
  - The next forwarding hop
  - The CoS treatment, if any
  - the operation to be performed on the label stack before forwarding
    - swap label,
    - pop label off the stack, or
    - swab label and then push one or more additional labels on the stack.
- Label has Local significance, and is unique within a given space (e.g., the space of incoming (outgoing) packets on a given incoming (outgoing) interface
- Values 0-14 are reserved. If Value is 0 (NULL), label is popped and packet is forwarded based on IP header.

Packet Forwarding in Internet

- **At each router, each packet is assigned to a FEC**
- **Each FEC is mapped to a next hop.**



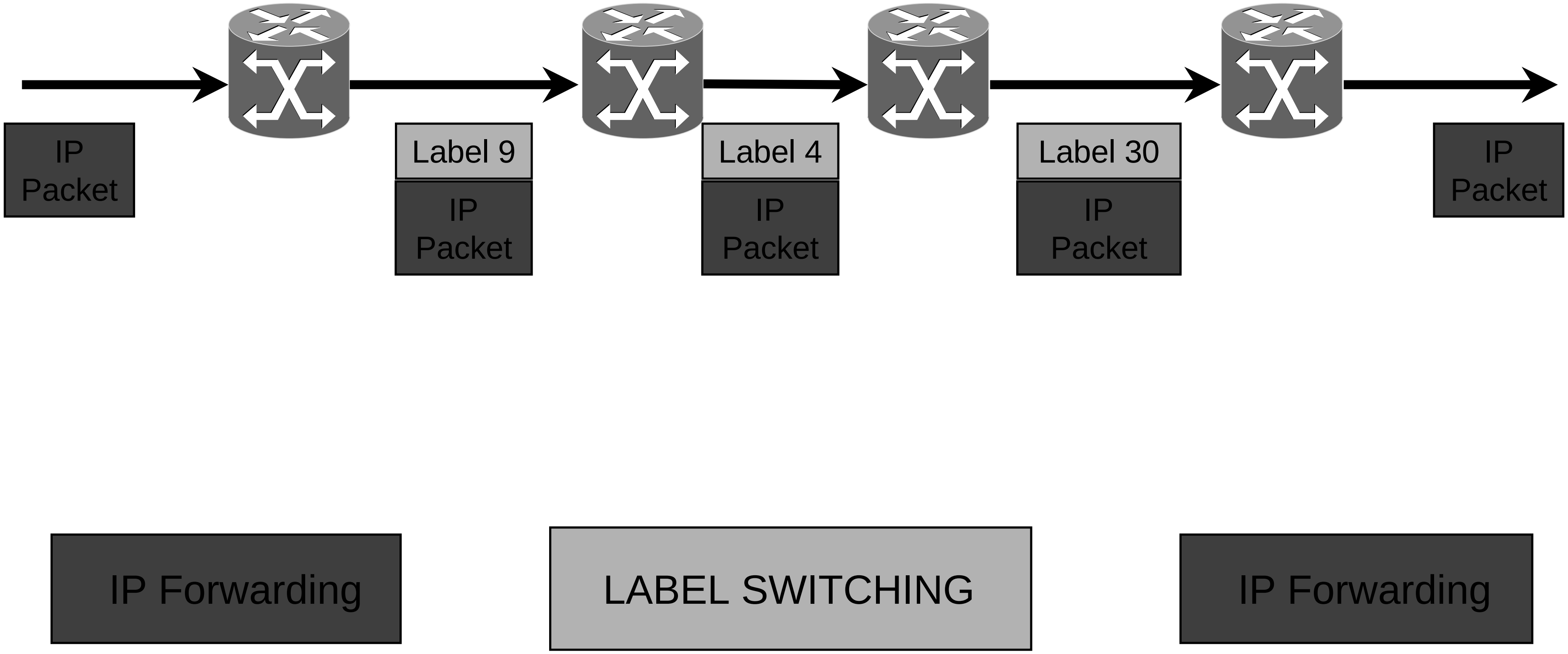
Dest	Next hop
10.10.2.4	128.2.60.2

Dest	Next hop
10.10.2.4	135.32.42.3

Dest	Next hop
10.10.2.4	100.2.15.35

Dest	Next hop
10.10.2.4	10.10.2.4

MPLS Packet Forwarding Example
MPLS in the Backbone



IP Packet

Label 9
IP Packet

Label 4
IP Packet

Label 30
IP Packet

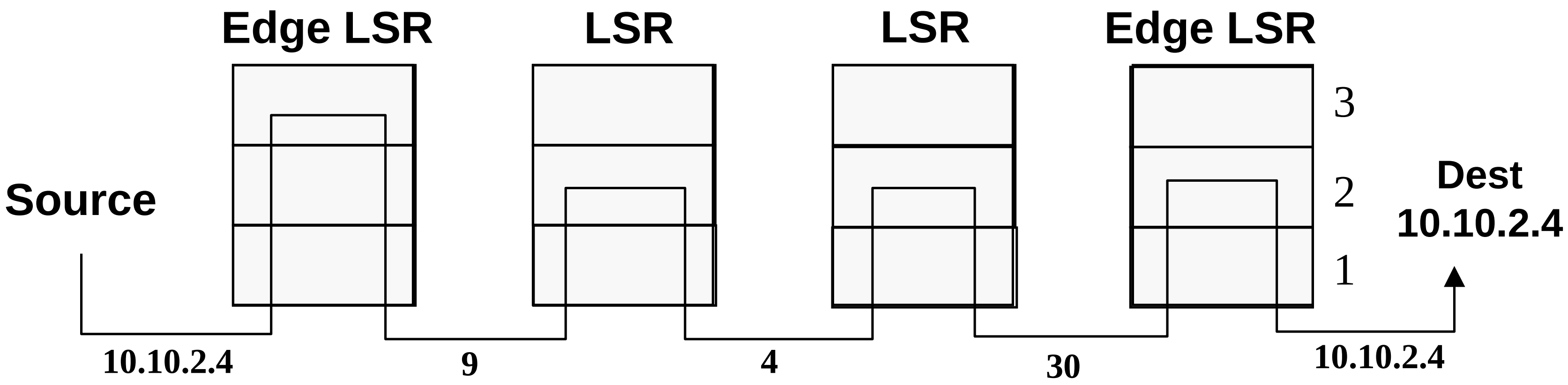
IP Packet

IP Forwarding

LABEL SWITCHING

IP Forwarding

Packet Forwarding in MPLS



Forwarding Table:

Dest	Out label
10.10.2.4	9

FTN

In Label	Out label
9	4

ILM

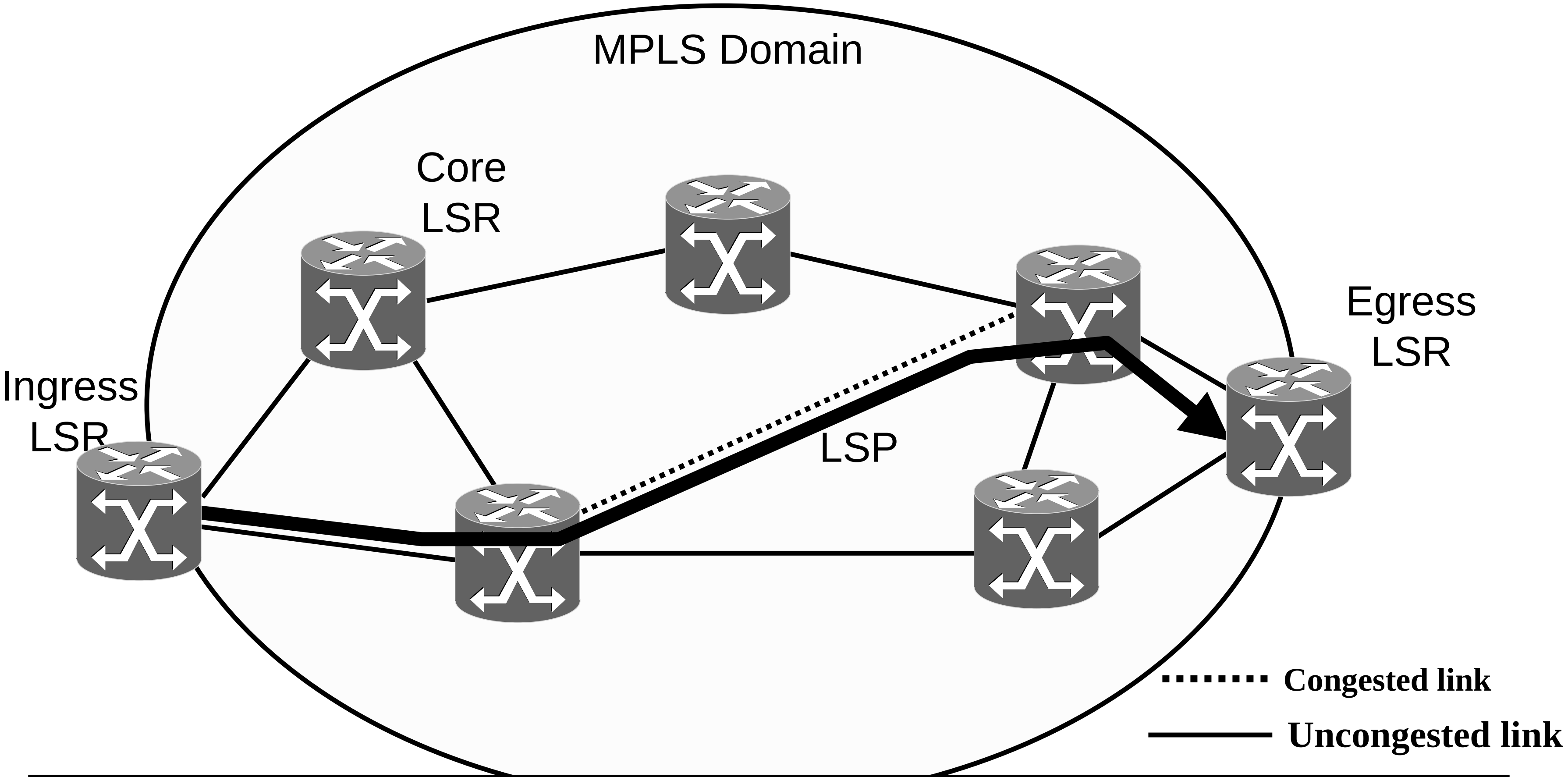
In Label	Out label
4	30

ILM

In Label	Next hop
30	10.10.2.4

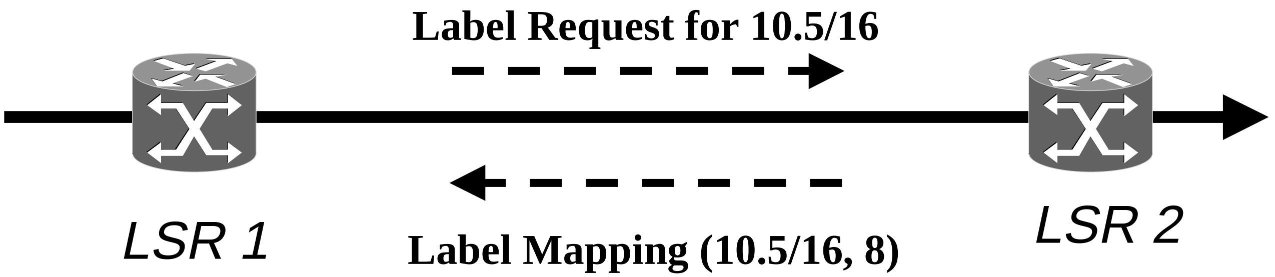
ILM

Label Switched Paths That Follow Routing



LDP is used to build the LSPs, using IP forwarding tables to follow the paths used by hop-by-hop routing

Label Distribution Protocol (LDP)



- Basic flow of LSP set-up using LDP

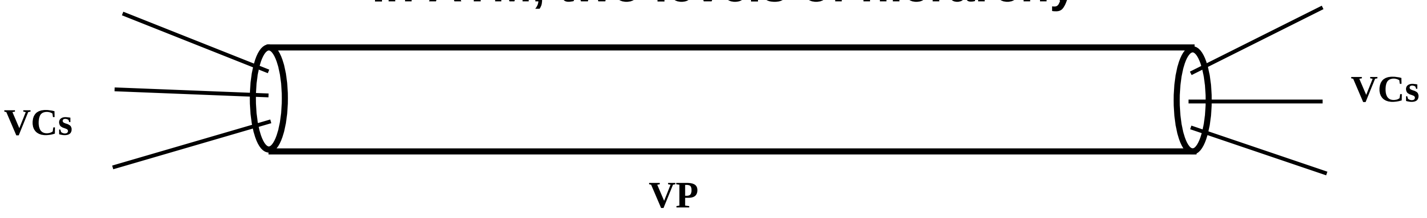
1) LSR1 detects that LSR2 is its next hop for FEC=10.5/16

2) LSR1 sends a Label Request message to LSR2 for FEC=10.5/16

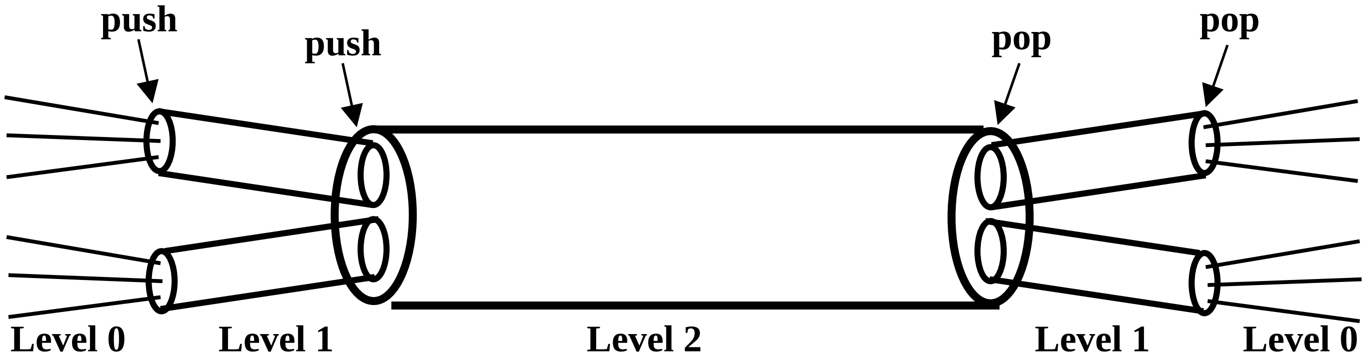
3) LSR2 responds with a Label Binding message that specifies FEC-label binding

Tunnel Hierarchy

In ATM, two levels of hierarchy



Label stack in MPLS allows for arbitrary levels of hierarchy



Label 0	packet

Label 2	Label 1	Label 0	packet

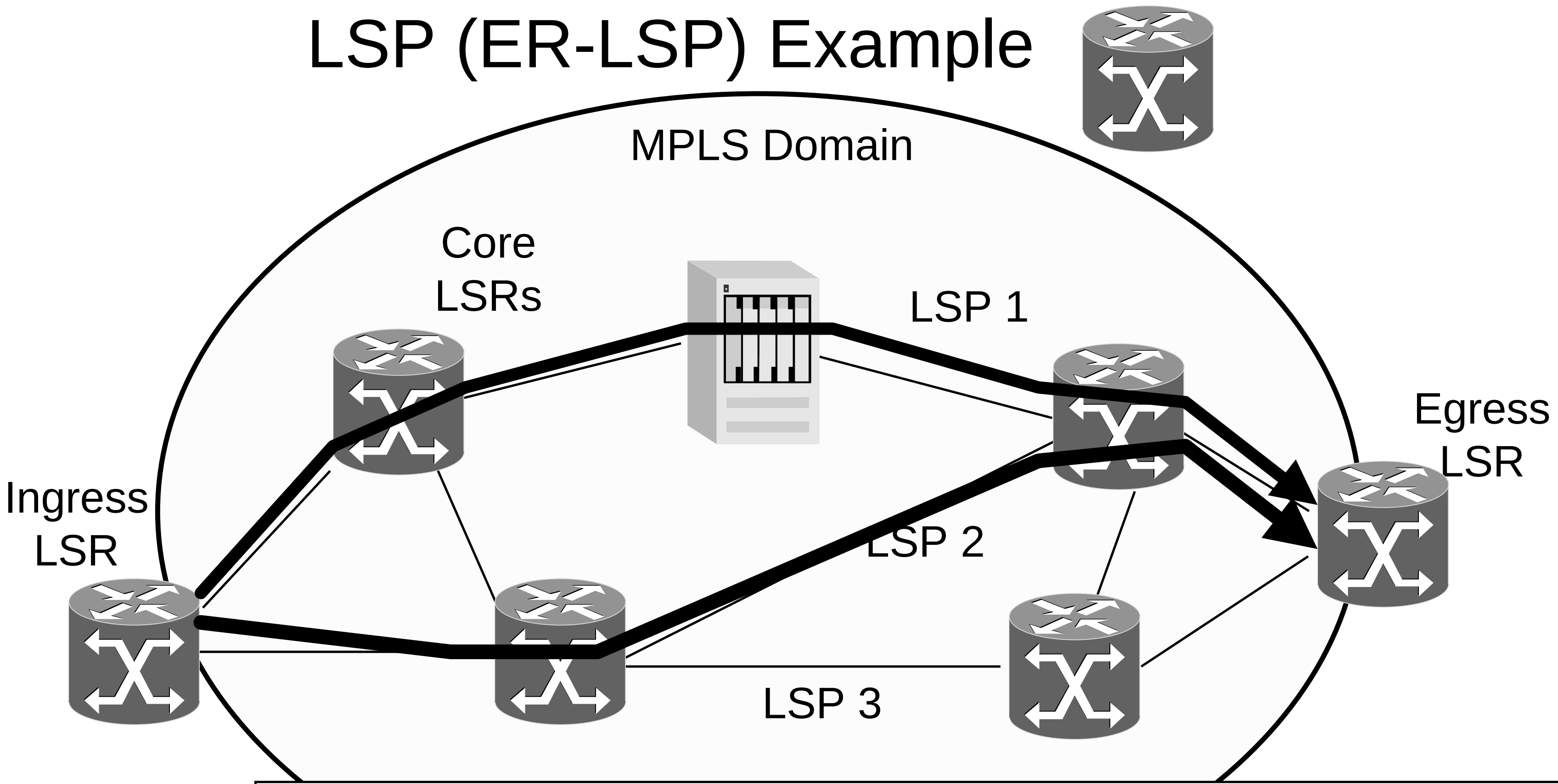
Label 0	packet

- Label stacks are used for multiplexing multiple LSPs into an aggregate LSP
- Reduce the number of LSPs through the core
- Application to VPNs

MPLS Traffic Engineering Using Explicitly Routed LSPs (ER-LSPs)

- LSP setup using source routing
- Builds a path from source to destination
- Policy, QoS, and/or other constraints may be used to determine LSP routing
- Explicit route is an example of constrained route where the constraint is the order (“strict” or “loose”)in which LSRs are visited
- Backup LSPs may be setup for rerouting traffic in case of failure of primary paths.

Explicitly Routed LSP (ER-LSP) Example



- Blue - path followed by routing, produced by LDP
- Red - an explicitly routed LSP from the Ingress LSR to the Egress LSR
- ER-LSP follows the path chosen by the source

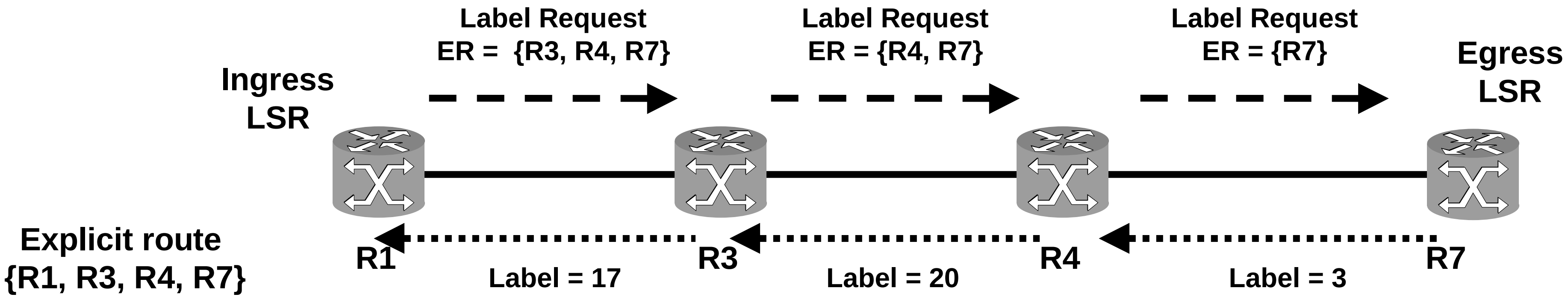
Two requirements

- Signaling protocols:
  - two signaling protocols defined by the IETF
    - CR-LDP: Constraint-Based Routing - Label Distribution Protocol
    - RSVP-TE: Extensions to RSVP for Traffic Engineering
  - Market will decide the success of each
- Enhanced routing protocols to facilitate traffic engineering (TE) capabilities and populate TE database with link attributes and topology information.
  - ISIS-TE
  - OSPF-TE:
    - Opaque TE LSA

CR – LDP

- Extensions to LDP to signal user requests for resource reservations and other constraints
- LSRs exchange LDP messages using TCP
- A mechanism for establishing explicitly routed LSPs
- Label Request Message
  - Includes:
    - Explicit Route TLV (optional)
    - Traffic Parameters TLV (optional)
      - Peak Data Rate
      - Peak Burst Size
      - Committed Data Rate
      - Committed Burst Size
      - Excess Burst Size
    - Pinning TLV (optional)

CR-LDP



Label Request
ER = {R3, R4, R7}

Label Request
ER = {R4, R7}

Label Request
ER = {R7}

Ingress
LSR

Egress
LSR

Explicit route
{R1, R3, R4, R7}

R1

R3

R4

R7

Label = 17

Label = 20

Label = 3

In	Out
IP Route	17

In	Out
17	20

In	Out
20	3

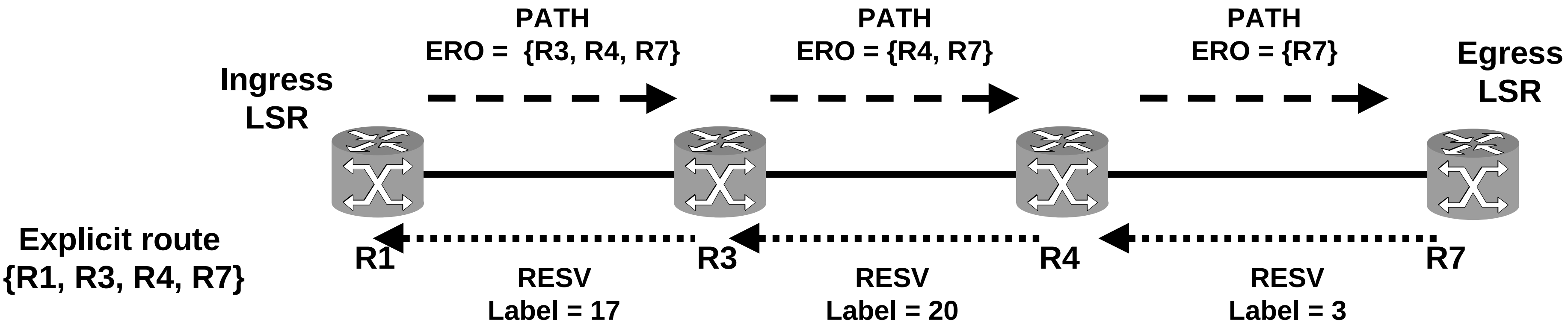
- Ingress LSR R1 obtains explicit route to egress LSR R7
- R1 transmits a Label Request Message addressed to R7
- Route list modified at each hop

- R7 transmits a Label Mapping message to R4 with Label = 3
- Intermediate LSRs R4 and R3:
  - Store “outbound” label, allocate an “inbound” label
  - Transmit Label Mapping with inbound label to upstream LSR
- R1 binds label to FEC

RSVP –TE

- RSVP-TE is traditional RSVP with explicit routing and scalability improvements
- LSRs exchange messages using raw IP
- supports downstream-on-demand label allocation only
- RSVP extensions (new objects):
  - PATH Message:
    - LABEL_REQUEST Object -- mandatory
    - EXPLICIT_ROUTE Object (ERO)
    - RECORD_ROUTE Object (RRO)
    - SESSION_ATTRIBUTE Object
  - RESERV Message:
    - LABEL Object -- mandatory
    - RECORD_ROUTE Object

RSVP-TE



PATH
ERO = {R3, R4, R7}

PATH
ERO = {R4, R7}

PATH
ERO = {R7}

Ingress
LSR

Egress
LSR

Explicit route
{R1, R3, R4, R7}

R1

R3

R4

R7

RESV
Label = 17

RESV
Label = 20

RESV
Label = 3

In	Out
IP Route	17

In	Out
17	20

In	Out
20	3

- **Establishing state and requesting label assignment**
  - Ingress LSR transmits a PATH message addressed to Egress router LSR
    ERO = {strict R3, strict R4, strict R7}

- **Distributing labels and reserving resources**
  - Egress LSR sends a RESV message to R4
    Label = 3
    Session object to identify the LSP
  - Intermediate LSR (R4 and R3)
    - Stores “outbound” label, allocate an “inbound” label
    - Transmits RESV with inbound label to upstream LSR
  - Ingress LSR binds label to FEC

CR-LDP vs RSVP-TE

- Underlying Protocol
  - CR-LDP : TCP
  - RSVP-TE: raw IP
- Protocol State
  - CR-LDP: hard
  - RSVP-TE: soft
- Resource Reservation
  - CR-LDP: forward path
  - RSVP-TE: reverse path
- Other functional differences arise from the above

MPLS & ATM

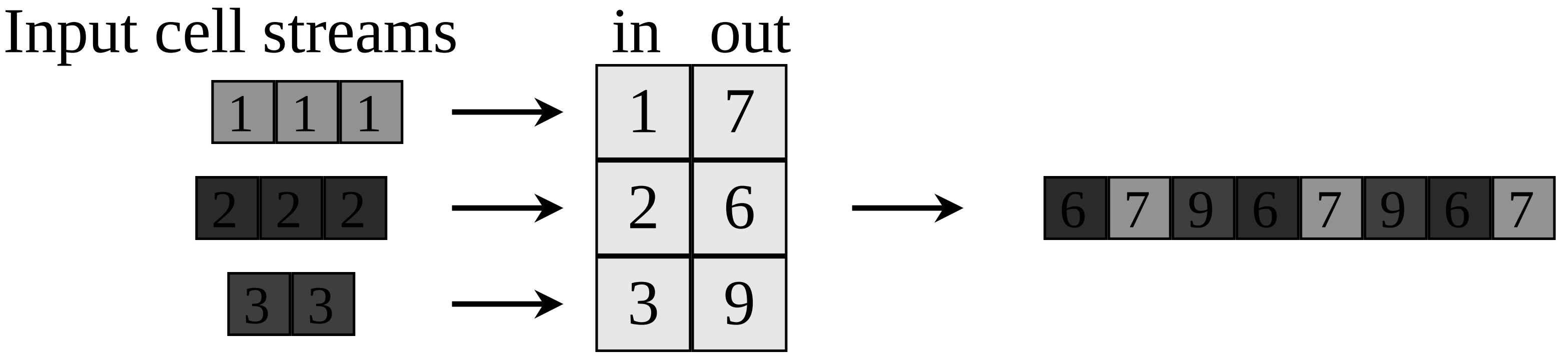
- Label-Controlled ATM
  - LSRs use ATM hardware
- Ships in the night
  - ATM and MPLS control planes independently run on the same hardware
  - VPI/VCI label space is partitioned
  - Intermediate solution

- Scalability with label merging
  - VC Merging
  - VP Merging

ATM-based LSRs

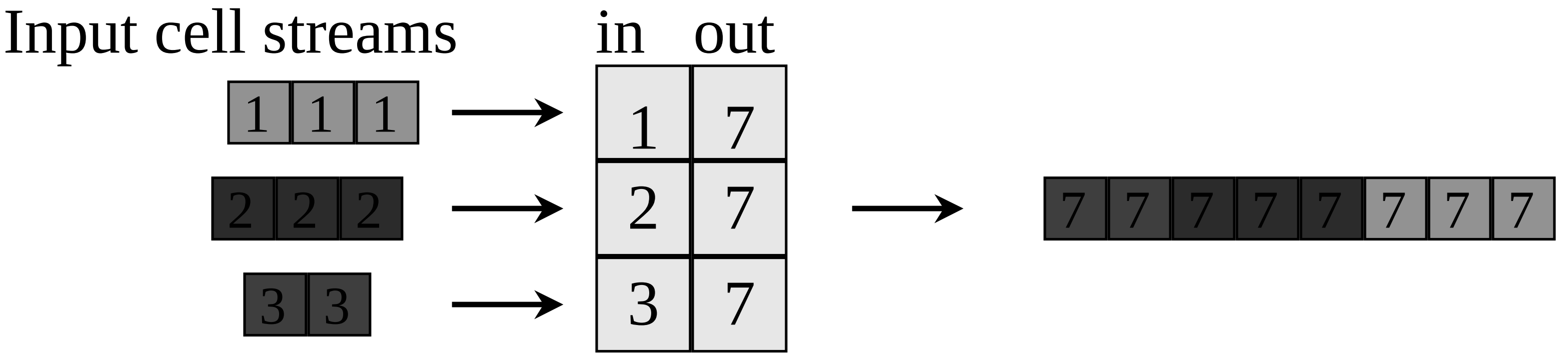
- How to Map Route Information to VC Labels?
- Non-VC merging: Map each source-destination pair to a unique VC value at a switch.
  - Problem: Need O(n^2) VC labels, where n is the number of destinations.
- VP merging: Map incoming VP labels for the same destination to the same outgoing VP label. For each VP, a unique VC value is used to identify the sender.
  - Problem: VP space exhaustion.
- VC merging: Map incoming VC labels for the same destination to the same outgoing VC label.
  - Need only O(en) VC labels, where e is the number of switch ports (typically small).
  - Potential Problem: Cells belonging to different packets for the same destination cannot interleave with each other.

VC Merging versus Non-VC Merging



in	out
1	7
2	6
3	9

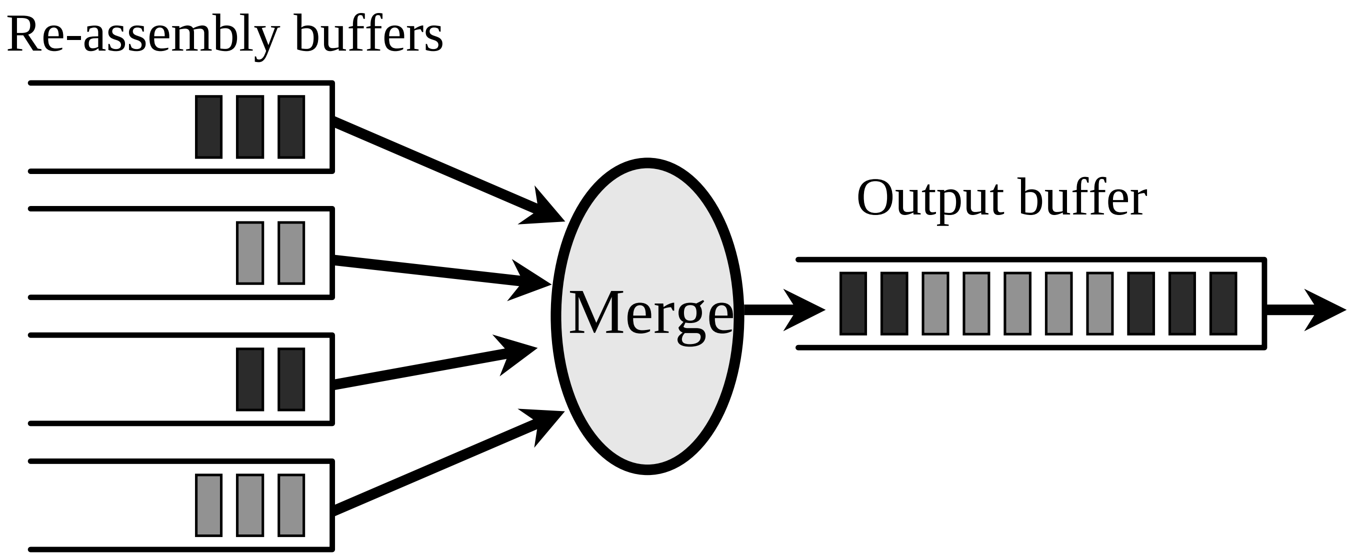
Non-VC merging



in	out
1	7
2	7
3	7

VC merging

Structure of Switch Output Module

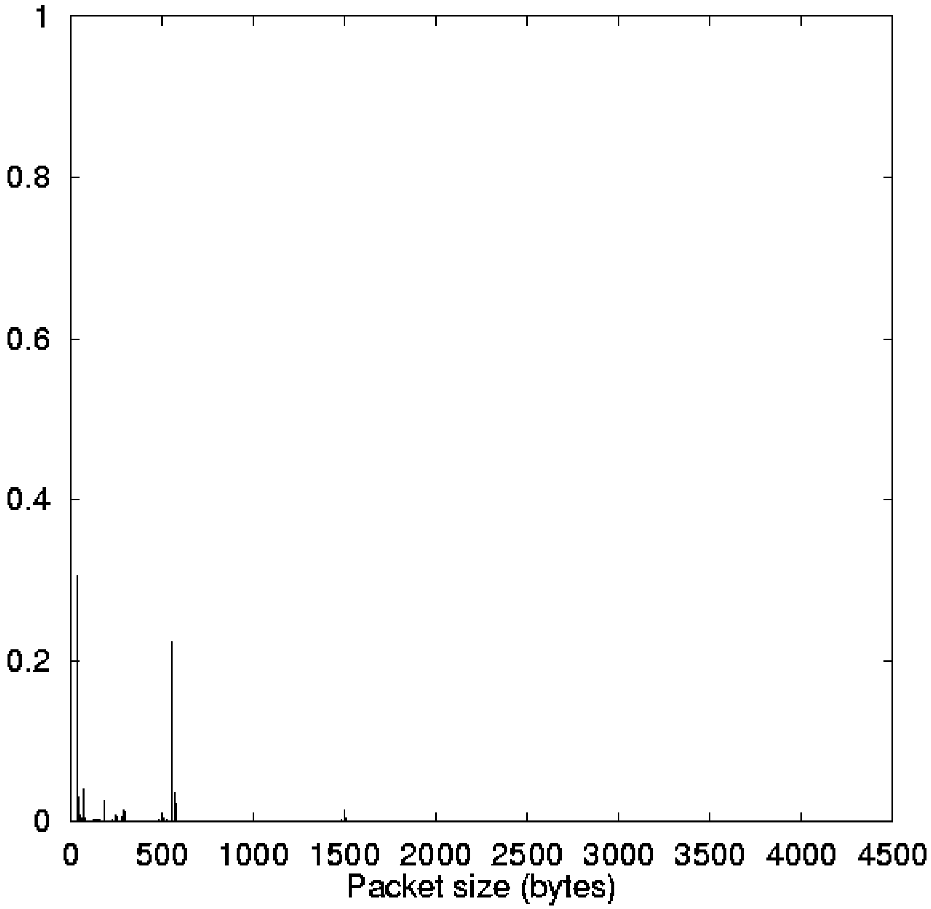


- For each packet, incoming cells are stored in a re-assembly buffer until the last cell arrives

- Issue: **Additional** buffers required

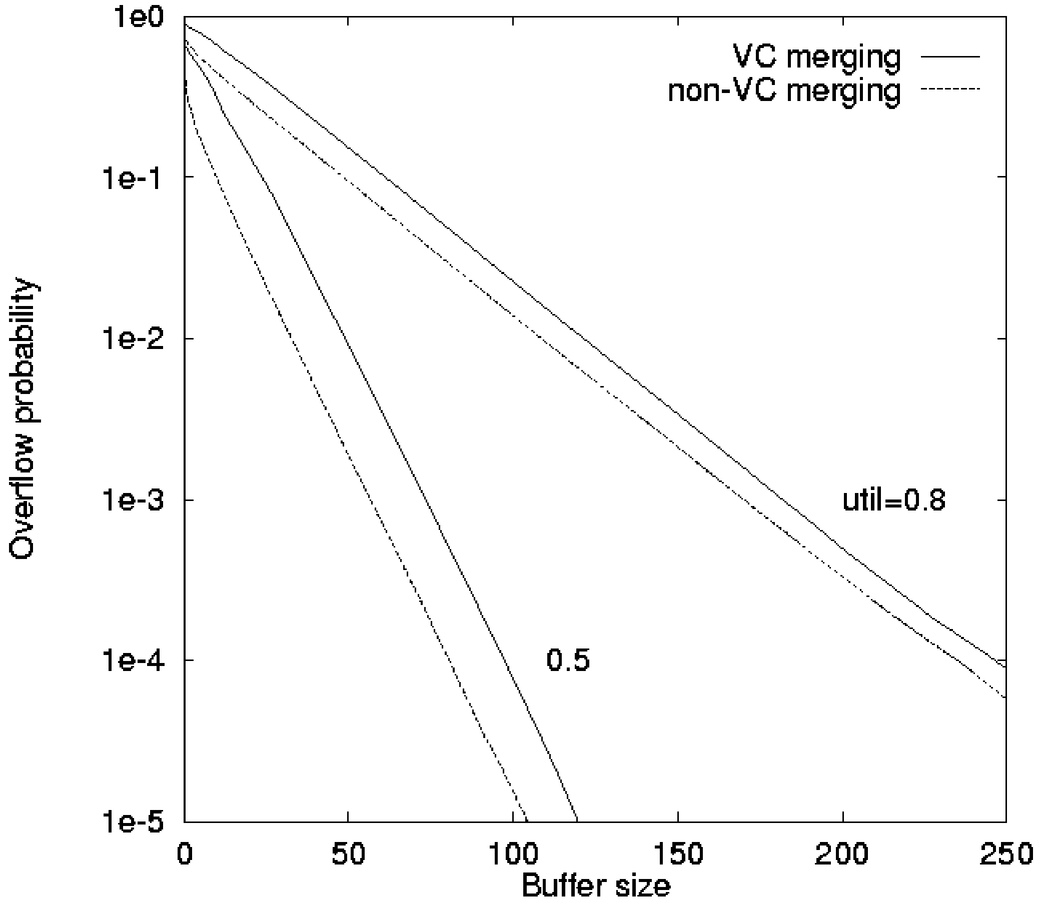
- Analysis using D-BMAP/D/1 (discrete-time batch Markovian arrival process) and simulation (IETF RFC 2682 -- I. Widjaja, A. Elwalid)

Internet Packet Size Distribution



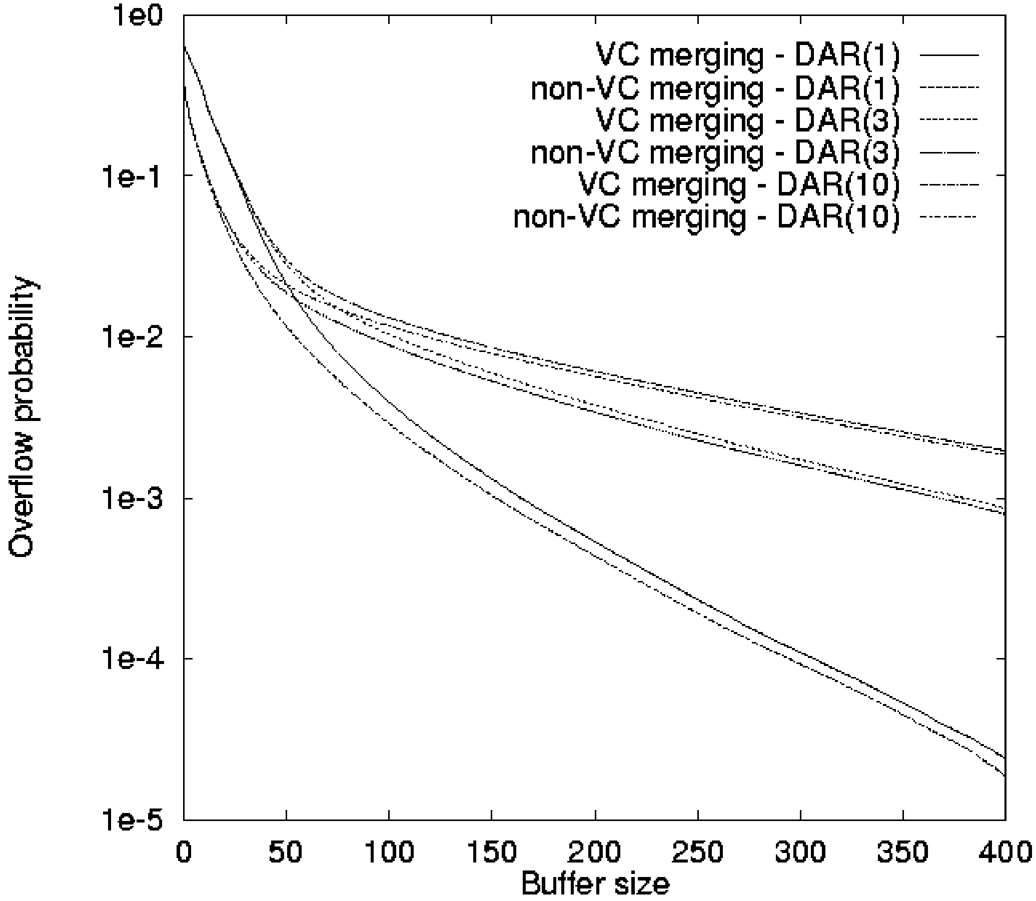
Bi-modal distribution
Mean packet size = 257 bytes

Comparison of VC merging versus non-VC merging for Internet Packets (size=6.2)



Additional buffer requirements for VC merging decreases as the traffic utilization increases

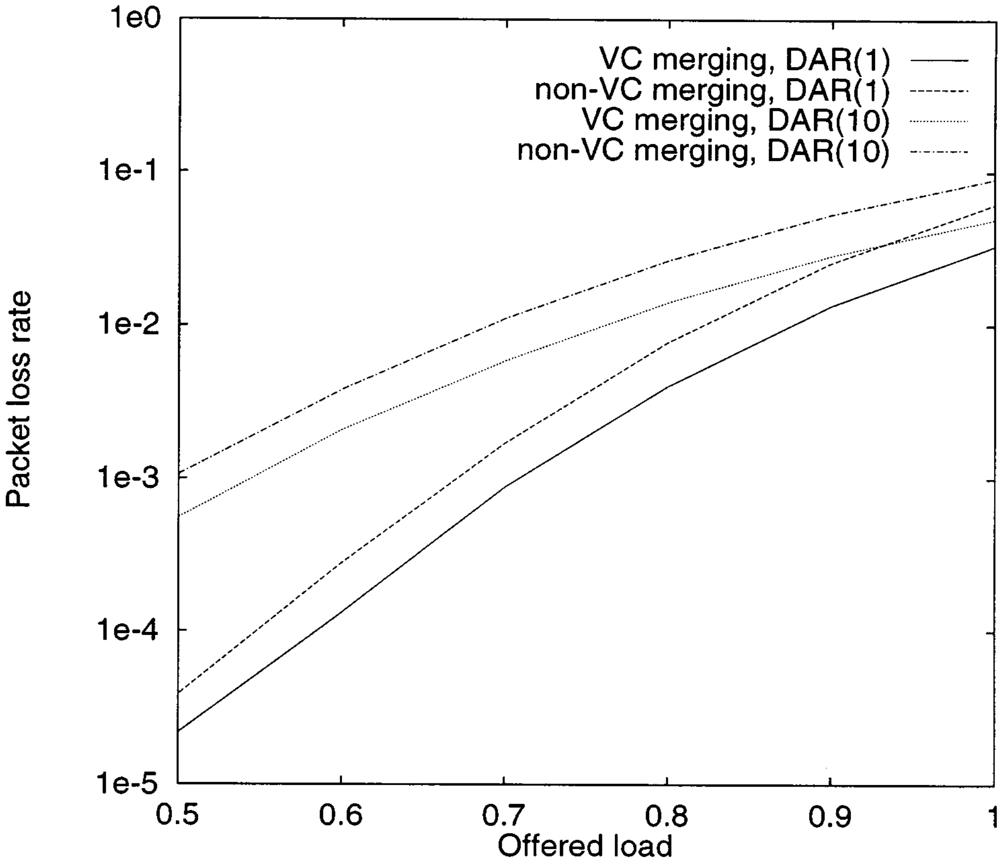
Effect of Correlated Interarrival Times on Additional Buffer Requirement



DAR(*p*) = Discrete-Auto-regressive process of order *p*

With VC merging, higher correlation of packet arrivals leads to smaller additional buffer

Packet Loss Comparison



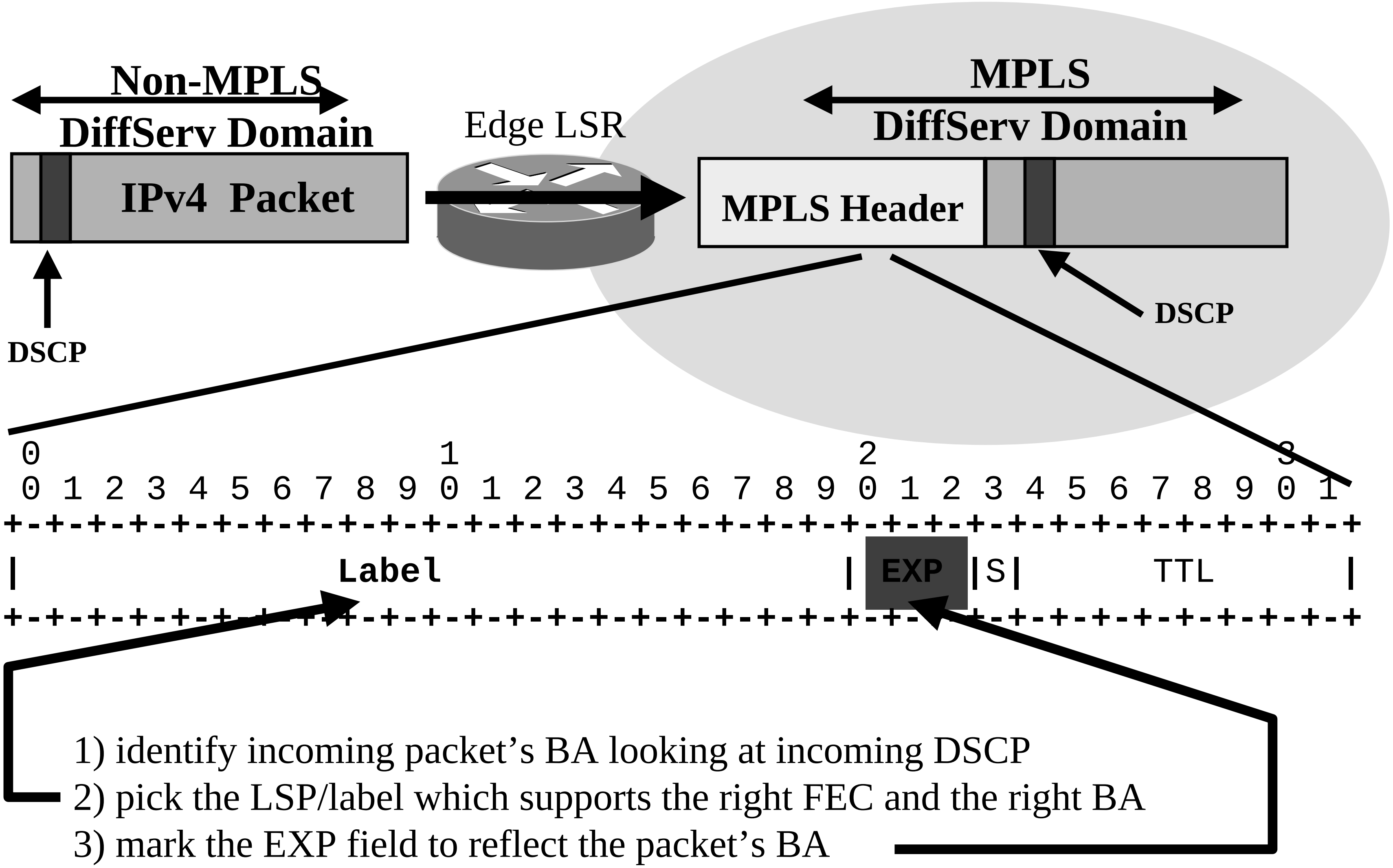
For fair comparison,
EPD (Early packet Discard)
is used with non-VC merging

VC merging leads to higher
packet goodput

Conclusion

- VC merging is scalable solution for MPLS
- The overhead for VC merging in terms of additional buffer requirement is minimal
- As utilization increases and/or traffic becomes more bursty, the additional buffer requirement for VC merging decreases
- VC merging achieves higher packet goodput than non-VC merging

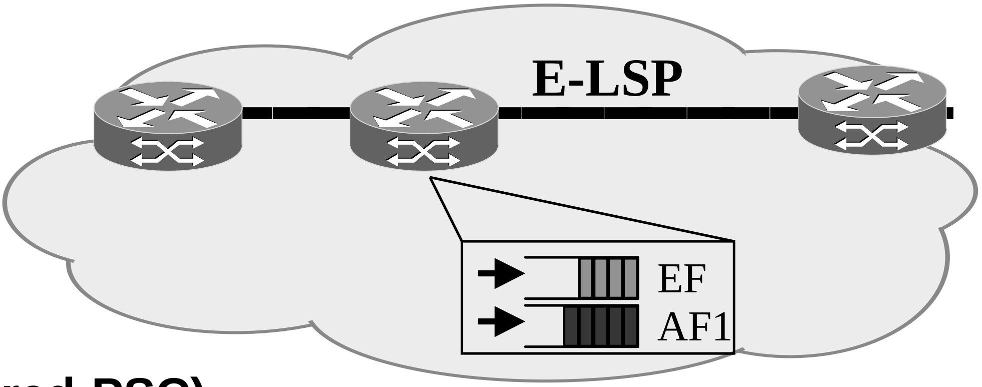
DiffServ over MPLS



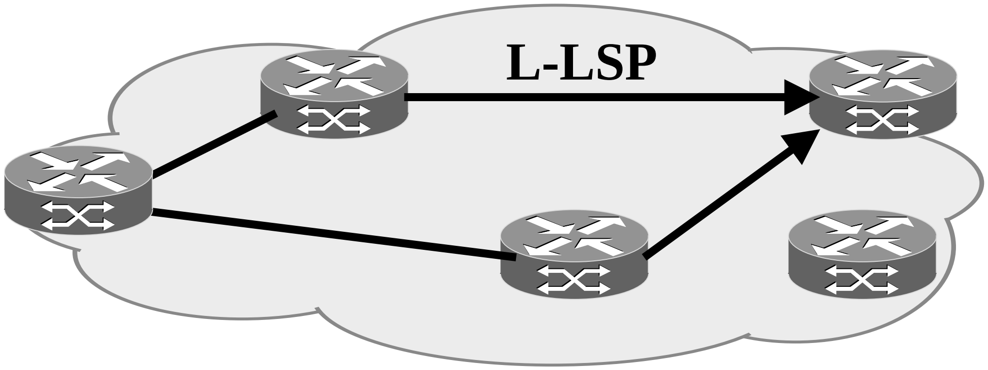
```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                Label                  | EXP |S|       TTL     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

1) identify incoming packet’s BA looking at incoming DSCP
2) pick the LSP/label which supports the right FEC and the right BA
3) mark the EXP field to reflect the packet’s BA

- **E-LSP (EXP-Inferred-PSC):**
  - EXP field of MPLS header determines the PHB to be applied to the packet. This includes both the PSC and the drop preference.
  - A single LSP can support up to eight BAs of a given FEC



- **L-LSP (Label-Only-Inferred-PSC)**
  - A separate LSP for a single FEC / BA pair
  - Label maps LSP using DSCP (6-bits)
  - requires signaling extension to bind “queue” to a label



Traffic Engineering Solutions for MPLS Networks

- LSP Design
  - Local optimization: constrained-based routing
  - Global optimization:
    - Offline (hard)
    - Time-dependent (uses historical data, traffic forecasts, SLA’s)

- Adaptive traffic engineering:
  - online based on real-time measurements
  - load balancing of traffic among existing LSP’s

- VPN Design
  - Global optimization (even harder !)

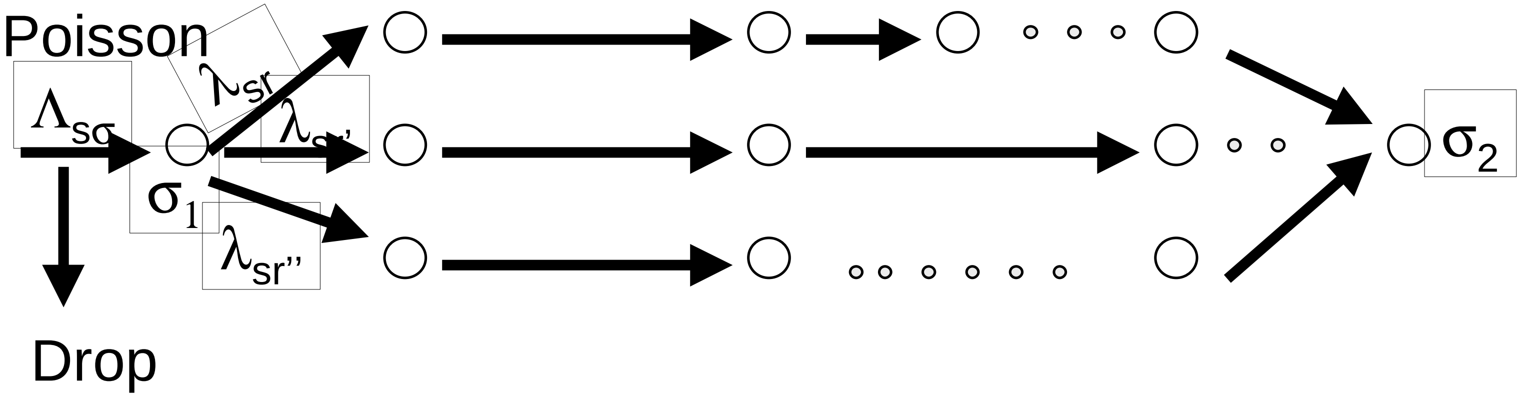
Constrained Shortest Path Design

- Problem: Given a directed graph, where with each link associated two parameters: length (administrative cost) and delay (or other QoS parameter). Find the shortest length path from a source node to a destination node, such that the total path delay does not exceed a given threshold (constraint).
- Exact solution is NP-hard
- Approximate solution for on-line implementations
  - Delay-scaling:
    - Allow the delay to be within ε of the maximum value, where ε is user-given error tolerance
    - Polynomial-time algorithm
    - (Kataria, Goel, Ramakrishnan)

Global LSP Design

(Mitra, Morrison and Ramakrishnan)

- MODEL: *N* **NODES; *L* LINKS, AND *S* SERVICES*.* *SERVICE* s REQUIRES BANDWIDTH d_{sl} ON LINK *l* OF CAPACITY C_l**



$$\sigma = (\sigma_1 , \sigma_2) = (o , d)$$

GIVEN ARRIVAL RATE $\Lambda_{s\sigma}$ AND THE SET OF ADMISSIBLE ROUTES R(s, σ) FOR STREAM (s, σ)

OPTMIZATION PROBLEM

Maximize Network Revenue W = $\sum_{s,\sigma} \sum_{r\in R(s,\sigma)} e_{sr}\rho_{sr}(1-L_{sr})$

e_{sr} EARNINGS PER UNIT TIME IF A CALL OF TYPE s IS CARRIED ON ROUTE r

ρ_{sr} IS THE OFFERED TRAFFIC OF TYPE s ON ROUTE r

L_{sr} IS THE LOSS PROBABILITY OF CALLS OF TYPE s IN ROUTE r

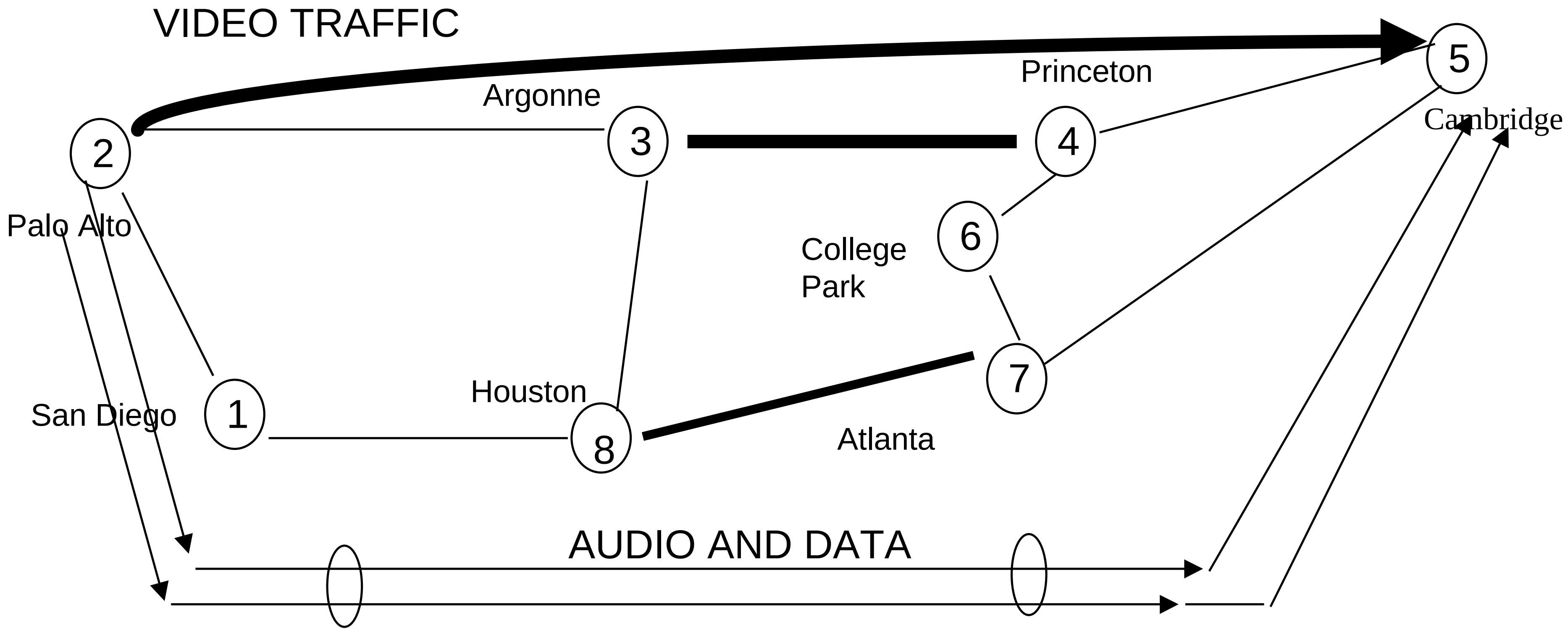
SUBJECT TO CONSTRAINTS

$$\sum_{r\in R\ (s,\sigma)} \rho_{sr} \le \Lambda_{s\sigma} / \mu_{s\sigma}\ \forall s, \forall \sigma$$

$$\rho_{sr} \ge 0$$

- NONLINEAR OPTIMIZATION PROBLEM

Features of Optimal Solution



- •TRAFFIC FROM PALO ALTO TO CAMBRIDGE
  - •VIDEO TRAFFIC SENT THE NORTHERN ROUTE
  - •AUDIO AND DATA TRAFFIC SENT ON THE SOUTHERN ROUTE

Joint design of LSPs and OSPF Weights

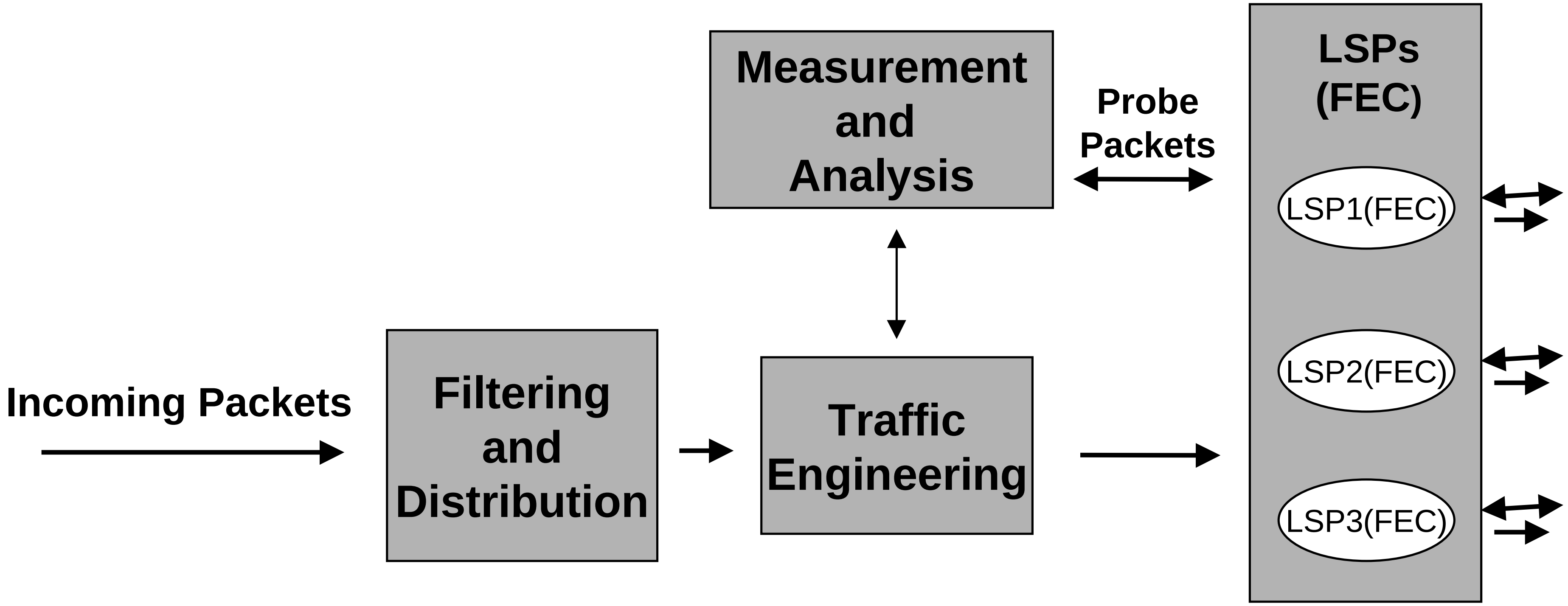
- **IGP (OSPF)**
  - Advantages: routing is simple, resilient, distributed, automated
  - Disadvantages: limitation on choice of routes;
  - Optimizing OSPF weights to meet objective may not be feasible
- **MPLS ER-LSP:**
  - Advantages: control of routes and performance
  - Disadvantages: cost of path setup and maintenance
- **Objective:**
  - Enhance OSPF by setting up a small number of ER-LSPs, so that OSPF weight optimization is possible
- ER-LSP is used as **Forwarding Adjacency**
  - Improves scalability
  - Advertised and used as a link in path computation in OSPF

MATE: MPLS Adaptive Traffic Engineering

- Internet Draft: <draft-widjaja-mpls-mate-02.txt>
- (A. Elwalid, C. Jin, S. Low, I. Widjaja)

- Features of MATE:
  - Assumes multiple LSP's are setup between ingress/egress pairs
  - Adaptive traffic mapping onto LSPs to minimize congestion
  - End-to-end control with no new hardware or protocol requirements at intermediate LSRs
  - No knowledge of a priori traffic distribution is required, and no particular scheduling or buffering schemes are assumed
  - Minimal packet re-sequencing

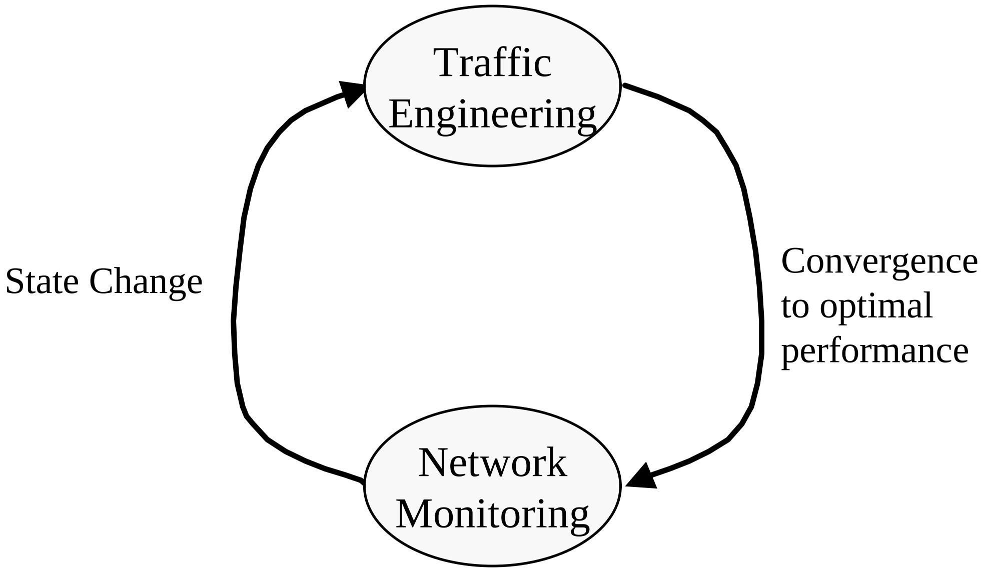
MATE Functions in Ingress LSRs



Probe packets are sent to estimate the *relative* one-way mean packet delay and packet loss rate along the LSP

Adaptive Traffic Engineering

- Alternate between Two Phases:
  - Traffic Engineering Phase
  - Network State Monitoring Phase

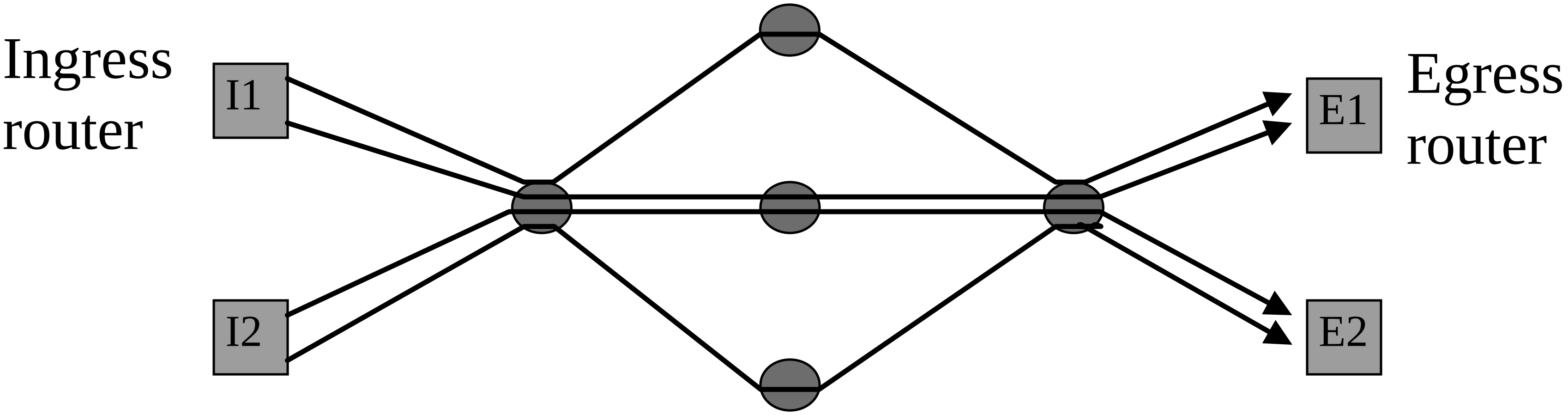


- Different Time Scales

Model

Ingress-egress node pair *s:*

- Input traffic rate a_s
- Set of paths P_s
- Assigns fraction λ_{sp} to path p in P_s



Objective

Split traffic to minimize total cost:

$$\min \quad C(\lambda) = \sum_{p} C_p(\lambda)$$

$$\text{s.t.} \quad \sum_{p \in P_s} \lambda_{sp} = a_s$$

$$\lambda_{sp} \geq 0$$

where λ = vector of global traffic splits

Cost is a function of mean packet delay and loss probability

Optimality:

A split λ is optimal if and only if, for each OD pair, all paths with positive flow have minimum (& equal) cost derivatives

Gradient Projection Algorithm:

Each pair s *individually* adjusts its traffic split $\lambda_s(t)$:

$$\lambda_s(t+1) = [\lambda_s(t) - \gamma \nabla C^s(t)]^+$$

- ◆ $\lambda_s(t)$: vector of traffic splits
- ◆ $\nabla C^s(t)$: vector of (measured) path cost derivatives
- ◆ γ :gain parameter

Asynchronous Environment

- Feedback delays:
  - substantial
  - different
  - time-varying
- IE pairs update
  - at different times
  - with different frequencies
- Network state probed asynchronously at different rates

Convergence

Theorem

Starting from any initial rate vector $\lambda(0)$, any accumulation point of the sequence $\{\lambda(t)\}$ is optimal, provided stepsize is sufficiently small

Stability

- Stepsize = how fast traffics are changed
- Tradeoff
  - Small stepsize : converges, but slowly
  - Large stepsize : rapid convergence, but may diverge
- **Theorem**

  Convergence if gain parameter $< \dfrac{1}{L(1+a\,(2T+1))}$

  - T : degree of asynchronism
  - L : steepness of cost function
  - a : size of network

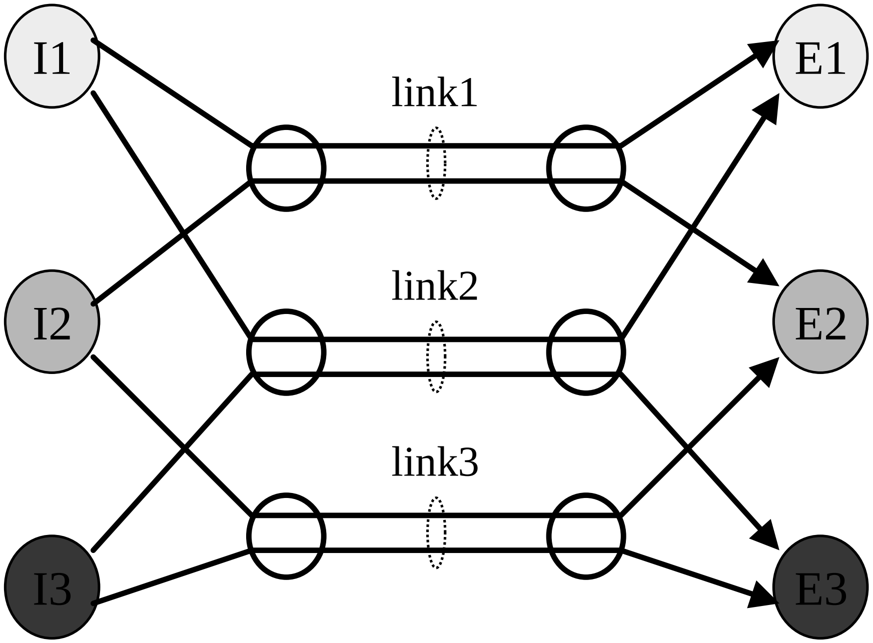
Traffic Measurement and Statistical Analysis

- Probe packets are sent to estimate the ***relative*** one-way mean packet delay and packet loss rate along the LSP

- Probing:
  - ICMP (Internet Control Message Protocol) Extension for one-way performance metrics
    - Internet draft <draft-elwalid-icmp-ext-02.txt>
  - UDP-based protocol

- Statistical techniques to obtain reliable estimates of congestion measures - Bootstrap Resampling technique

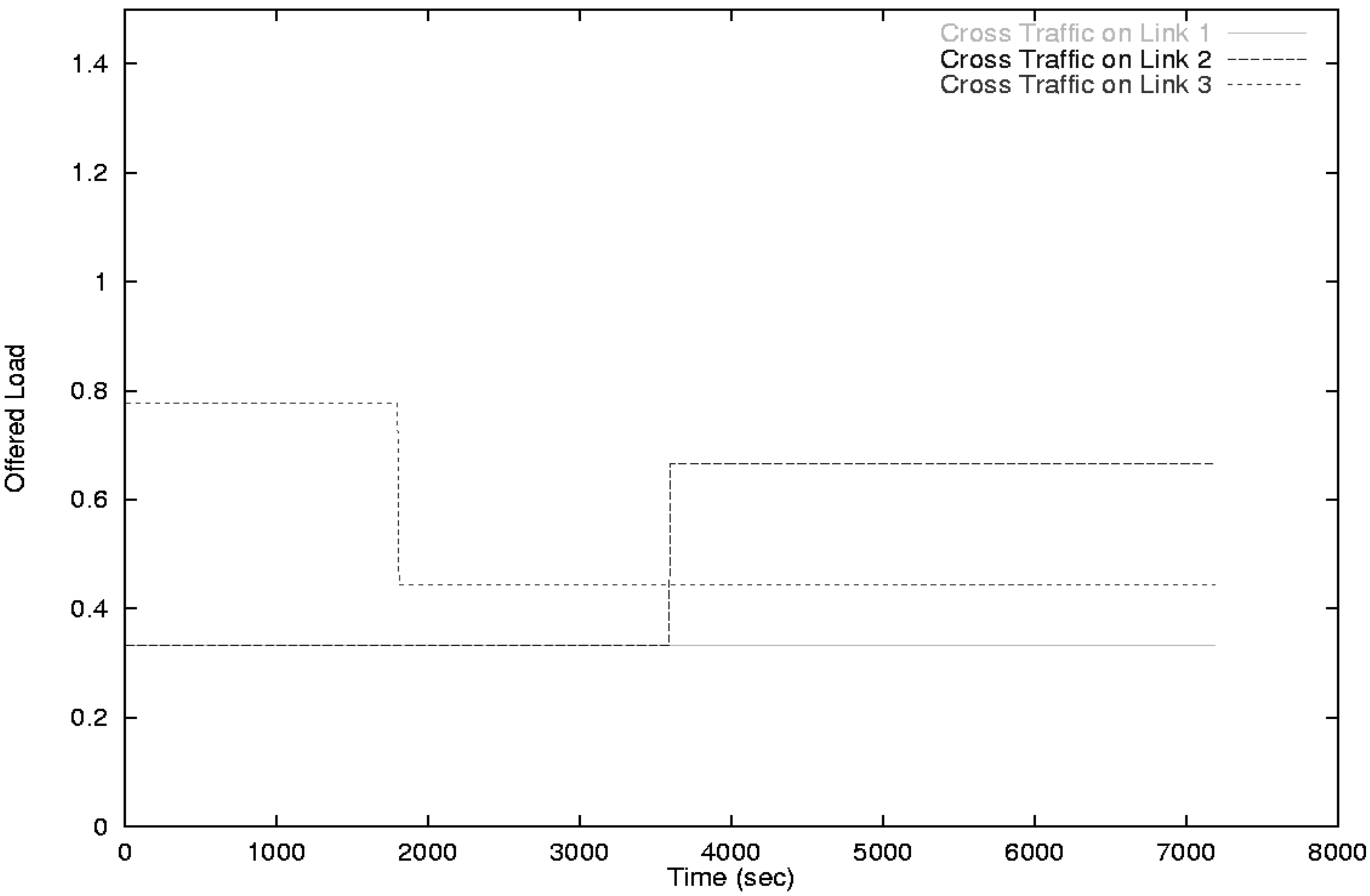
Experiment



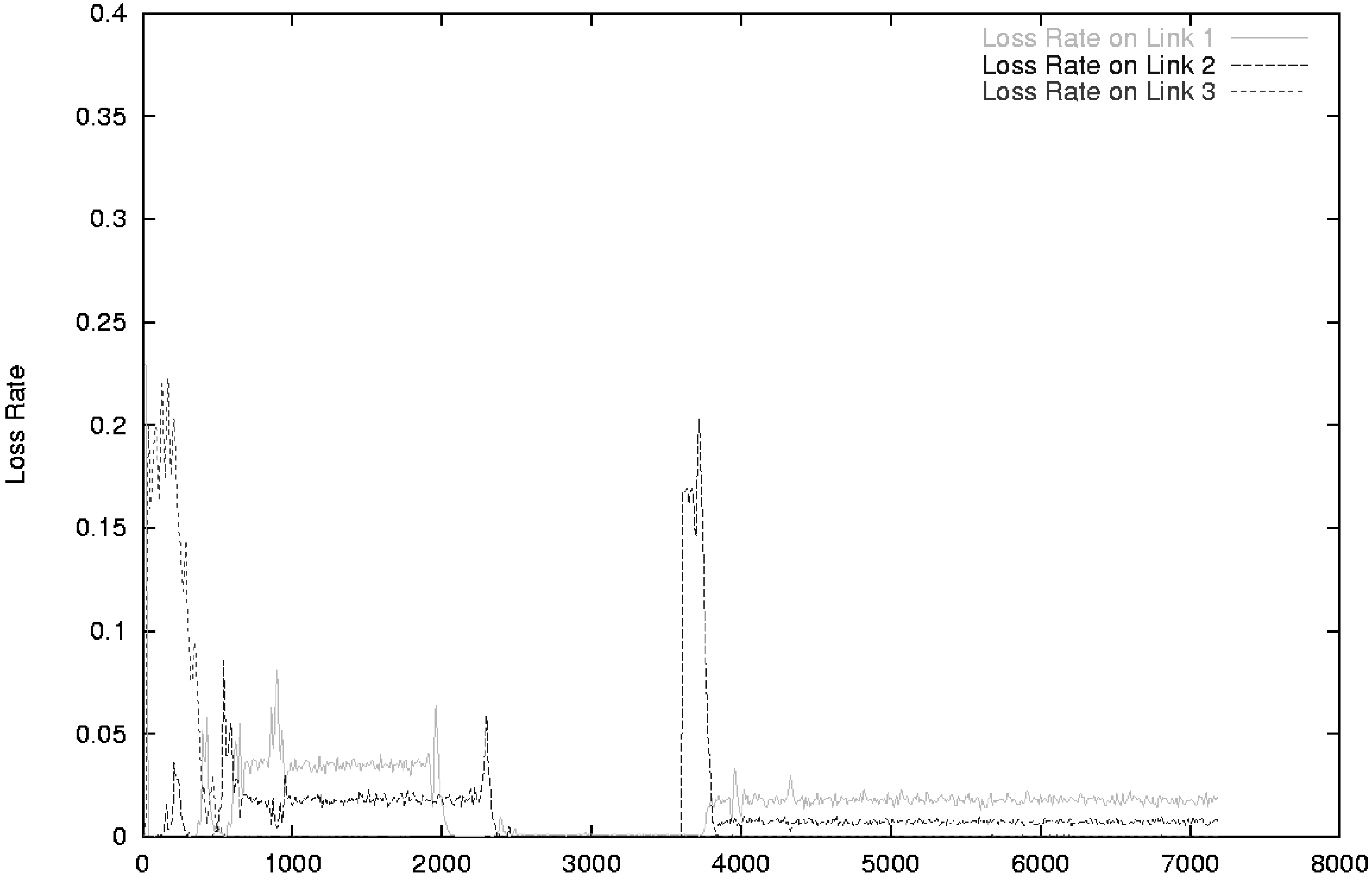
Topology Example

- 3 Ingress/Egress pairs
- 2 LSPs per pair
- Link1, Link2, Link3 each support 2 LSPs from different pairs plus additional “cross traffic”

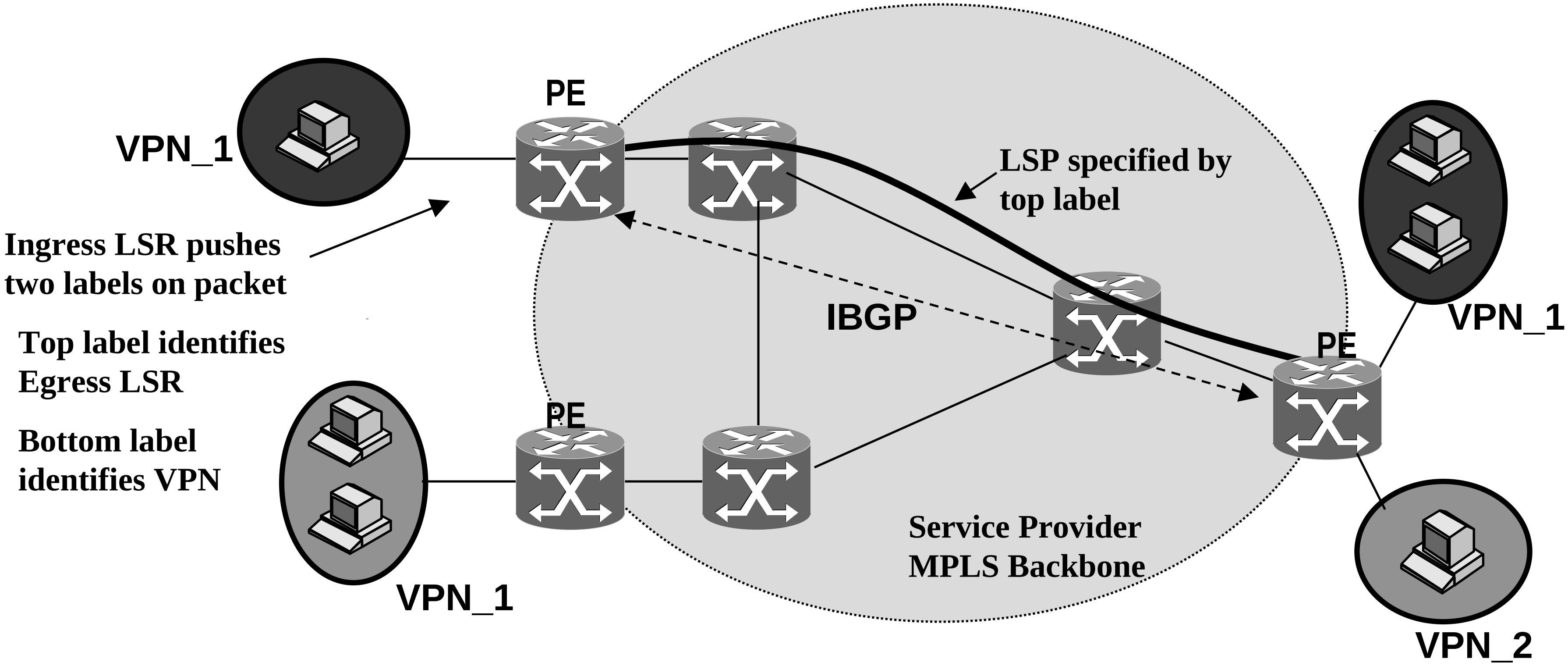
Cross Traffic



Packet Loss Rate



MPLS VPN architecture



LSP specified by top label

Ingress LSR pushes two labels on packet

Top label identifies Egress LSR

Bottom label identifies VPN

Service Provider MPLS Backbone

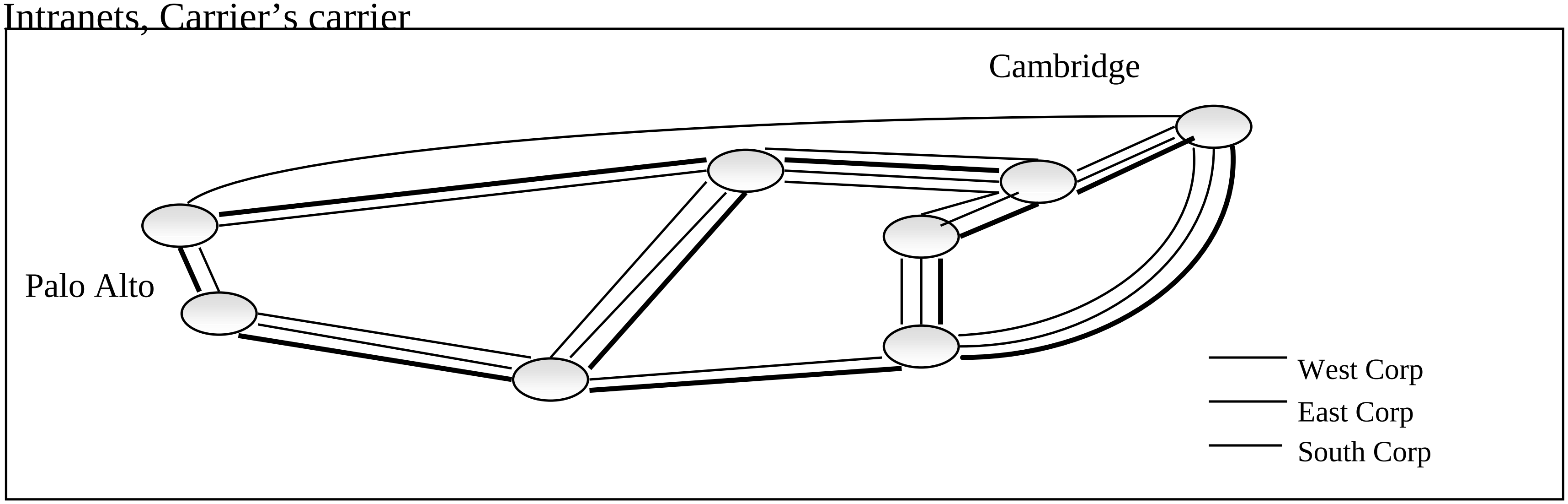
- •A PE (Provider Edge) LSR maintains VPN routes for those VPNs to which it is directly attached.

- •BGP used to distribute VPN route information

- •Scalability: by using MPLS two-level label stack, intermediate LSRs do not maintain any VPN routes

VIRTUAL PRIVATE NETWORK DESIGN

Intranets, Carrier’s carrier



Is there bandwidth & resources available for all SLAs?

How to craft a SLA?

- joint design for traffic and resource management
- large scale optimization with QoS constraints

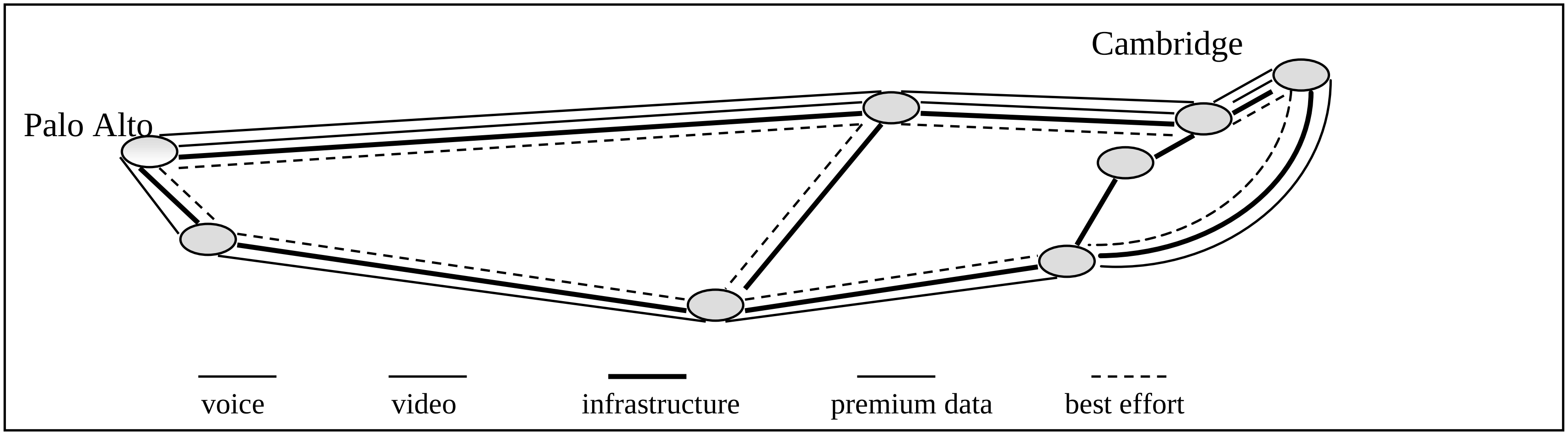
QoS requirements in SLA

packet-level (delay, loss)

flow level

class dependent

OPTIMUM ROUTING AND RESOURCE MANAGEMENT



- Given infrastructure
- 4 QoS classes: voice, premium data, video, best effort
- Point-to-point traffic for each QoS class
- “Effective Bandwidth” concept to encapsulate packet behavior, QoS
- Voice & Video: for low delay require routes to have few hops
- Premium data: routes may have more hops
- Best Effort data: many more hops allowed

OPTIMIZATION PROBLEM

High level goal: network-wide load balancing

Max “Network Revenue”
with respect to traffic management,
subject to above constraints.

VIRTUAL PRIVATE NETWORKS

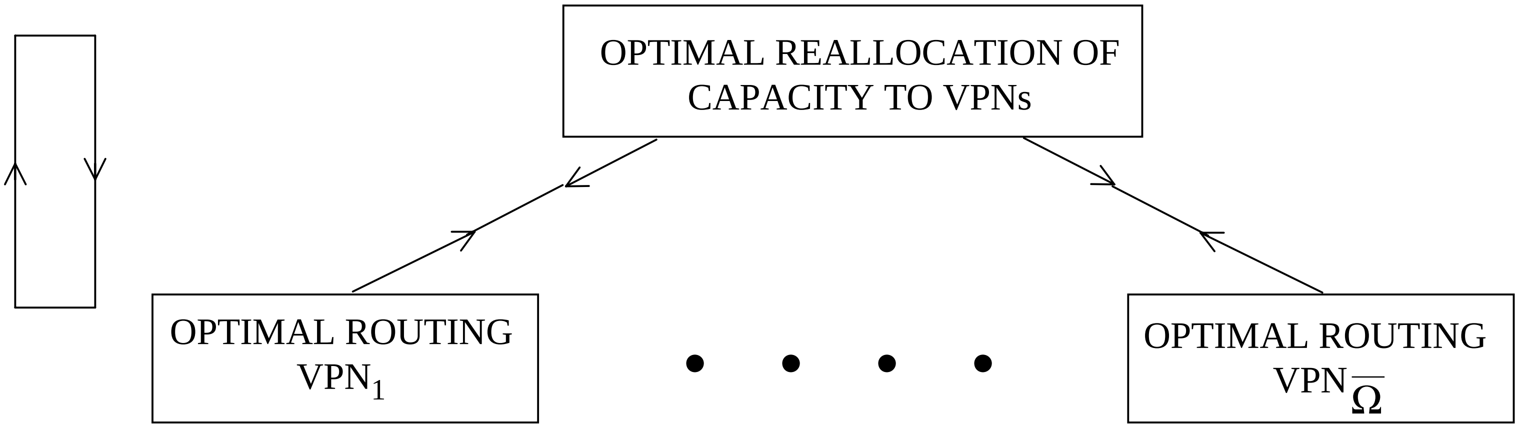
Allocate bandwidth on each link of infrastructure to VPNs such that, when each VPN's multi-QoS-class traffic is optimally routed over its allocated resources, a weighted aggregate measure of carried bandwidth over the service infrastructure ("network revenue") is maximized, subject to constraints that each VPN carries a specified minimum

Joint resource allocation and routing design

Multiplexing is across services and routes within each VPN, but not across VPNs in the interest of QoS protection

Alternately, Hierarchical Virtual Partitioning may be used to share resources across VPNs. The design here is used to select the algorithm parameters.

(Mitra, Morrison and Ramakrishnan)


OPTIMAL REALLOCATION OF CAPACITY TO VPNs
OPTIMAL ROUTING VPN$_1$
• • • •
OPTIMAL ROUTING VPN$_{\overline{\Omega}}$

DESIGN OF VIRTUAL PRIVATE NETWORKS (1)

INFRASTRUCTURE: L links, N nodes

$C_\ell = \text{capacity of link } \ell$

VPNs VPN Ω is allocated cap. $C_\ell^{(\Omega)}$ on link ℓ

$\left(\sum_{\Omega} C_\ell^{(\Omega)} \le C_\ell\right)$ decision variable

$(\mathrm{s},(\sigma_1,\sigma_2)) =$ “stream”

$R^{(\Omega)}_{(\mathrm{s},\sigma)} =$ set of admissible routes for (s,σ) on Ω

$\bar{\rho}^{(\Omega)}_{\mathrm{s},\sigma} =$ offered traffic on stream (s,σ) for Ω

$\rho^{(\Omega)}_{\mathrm{s,r}} =$ traffic offered to route $r \in R^{(\Omega)}_{(\mathrm{s},\sigma)}$

decision variable

REVENUE Revenue for subnet Ω

$$W^{(\Omega)} = \sum_{(\mathrm{s},\sigma)^{(\Omega)}} \; \sum_{r \in R\,(\mathrm{s},\sigma)^{(\Omega)}} e^{(\Omega)}_{\mathrm{sr}} \rho^{(\Omega)}_{\mathrm{sr}} \left(1 - \mathrm{L}^{(\Omega)}_{\mathrm{sr}}\right)$$

Total network revenue, $W = \sum_{\Omega} W^{(\Omega)}$

DESIGN OF VIRTUAL PRIVATE NETWORKS (2)

Revenue for subnet Ω, $W^{(\Omega)}$

Revenue for network $W = \sum W^{(\Omega)}$

GENERAL VPN DESIGN PROBLEM

$$\max_{\{C_\ell^{(\Omega)}\},\{\rho_{\text{sr}}^{(\Omega)}\}} \quad W$$

st $\quad W^{(\Omega)} \geq W_{\min}^{(\Omega)} \qquad \forall\Omega$

routing

$$\sum_{r \in R(s,\sigma)^{(\Omega)}} \rho_{\text{sr}}^{(\Omega)} \leq \overline{\rho}_{\text{s}\sigma}^{(\Omega)} \qquad \forall(\text{s},\sigma)^{(\Omega)} \quad \forall\Omega$$

$$\rho_{\text{sr}}^{(\Omega)} \geq 0$$

resource allocation

$$\sum_{\Omega} C_\ell^{(\Omega)} \leq C_\ell \qquad \forall\ell$$

$$C_\ell^{(\Omega)} \geq 0$$

SPECIAL VPN DESIGN PROBLEM: $W_{\min}^{(\Omega)} = 0 \qquad \forall\Omega$

Feasibility is issue in GENERAL problem

Special problem has simpler solution

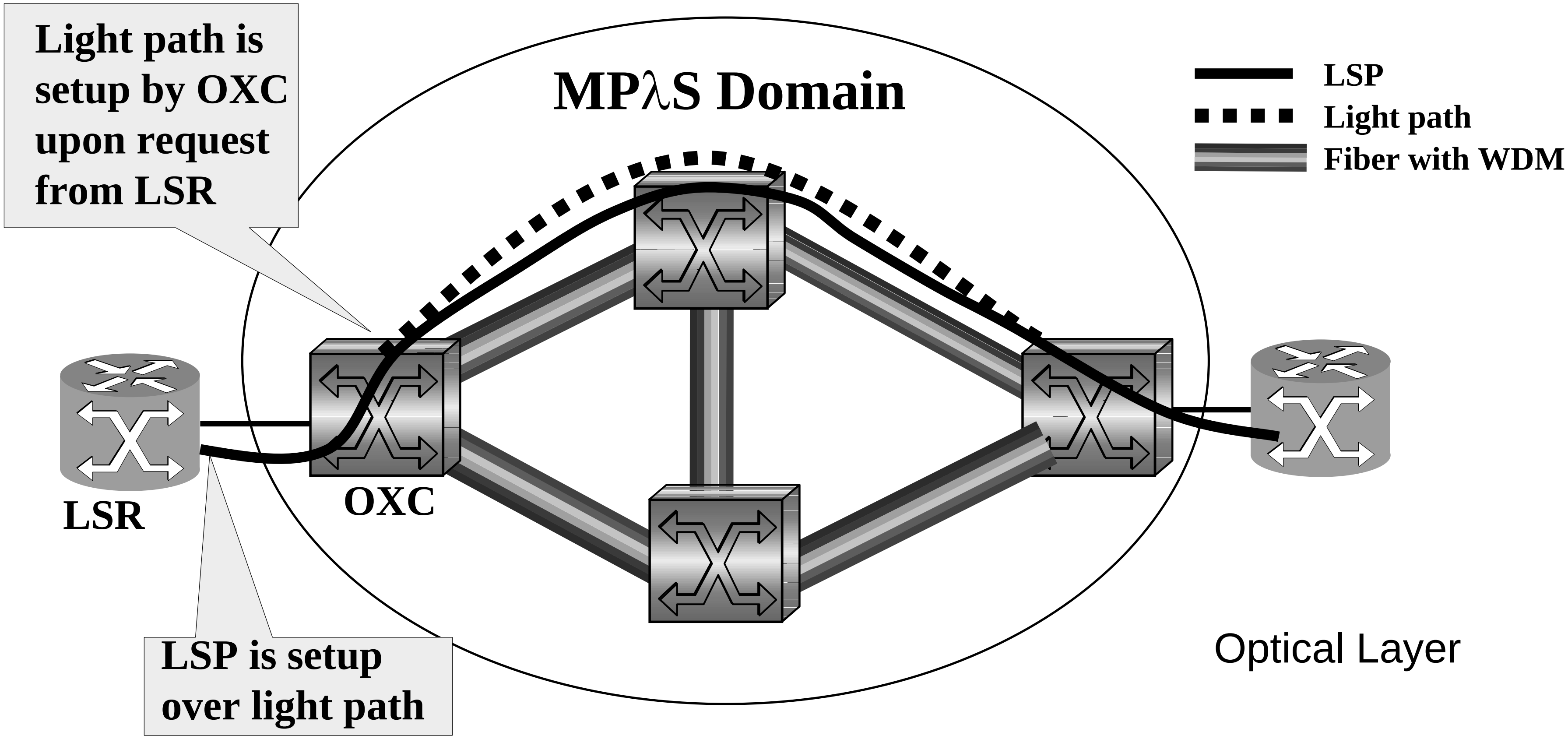
Protection and Restoration in MPLS

- Faster than layer3 rerouting
- More granular than layer1 mechanisms
- Link/Node protection:
  - Fast detour around local failure
  - Effective when applied to the most unreliable path components
- Path protection
  - protection path is disjoint from working path
  - More efficient (but slower than) link/node protection
  - Protection options:
    - 1:1: one working LSP is protected/restored by one protection LSP;
    - n:1: one working LSP is protected/restored by n protection LSPs with configurable load splitting ratio;
    - 1:n: one protection LSP is used to protect/restore n working LSP;
    - 1+1: traffic is sent on both the working LSP as well as the protection LSP, and the egress LSR selects one of the two copies.
- Link resources may be shared among protection LSP’s associated with different working LSPs
- Protection requirements may be included in optimization of traffic engineering

Optical Transport Requirements

- Real-time establishment of optical channels
- Dynamic reconfiguration / rearrangement
- Support of traffic engineering, protection and restoration
- Interoperability among diverse devices (routers, OXCs, etc.)

Dynamic Light (Lambda) Path Setup



Light path is setup by OXC upon request from LSR

LSP is setup over light path

- Direct LSP and light path establishment/reconfiguration to meet defined constraints via common signaling protocols
- **MPλS** for signaling in the optical transport network

MPλS and GMPLS (Generalized MPLS)

- **MPλS**:
  - A control plane of OXCs to facilitate dynamic light path (λ**)** setup and optical layer bandwidth management
  - Based on MPLS control protocols
  - Facilitates protection and restoration in optical links
- **GMPLS**: Extensions to MPLS to cover
  - TDM switching
  - λ switching (hence MPλS is a subset of GMPLS)
  - Port switching

MPλS

- Data plane driven by a switching matrix
  - LSR: (ingress label) ⇒ (egress label)
  - OXC: (ingress λ) ⇒ (egress λ)
- Based on Extensions to IGP and RSVP-TE/CR-LDP
- Extensions to OSPF:
  - Opaque LSA to carry optical TE and protection parameters
  - Link bundling
- Extension to RSVP-TE
  - Label Request Object Modification
    - Added Link Media Type
    - Label Object identifies λ/fiber requested

Service Models

- Overlay (optical UNI) model:
  - A client-network model, where IP (client) networks request connectivity over the OTN (optical transport network) via an UNI signal. Clients are unaware of the OTN architecture
  - Light paths across the OTN appear as links to IP devices outside the optical domain
  - Independent OTN and Client control planes
  - analogous to classical IP over ATM
- Peer Model:
  - LSRs and OXCs are peers
  - Integrated control of IP and optical networks
  - Reduced number of routing adjacencies
  - Full visibility of topology at L3
  - LSPs may span OXCs and LSRs
- Interesting combined traffic engineering and network design problems

Thank you