



# Quality of Service on the Internet: New Protocols and Traffic Engineering

Part B of Seminar on Current Issues and Technologies for the Internet

**Dr. Junaid Ahmed Zubairi**

Visiting Associate Professor

CIT, Arid Agriculture University

Aug 15<sup>th</sup>, 2002 8:30PM



# Overview of Presentation

- ◆ Motivation
- ◆ IETF's Intserv Model
- ◆ IETF's DiffServ Model
- ◆ Challenges in hybrid approaches
- ◆ MPLS
- ◆ Traffic Engineering
- ◆ GMPLS



# References

- ◆ <http://www.ietf.org>
- ◆ Internet QoS Architectures and Mechanisms for Quality of Service Z. Wang, Morgan Kaufmann 2001
- ◆ T3: Traffic Engineering in IP/MPLS Networks by Anwar ElWalid, IEEE Infocom 2001
- ◆ Data and Computer Communications by W. Stallings, Prentice Hall 2000



# Who is IETF?

- ◆ IETF is Internet Engineering Task Force
- ◆ IETF has several special interest focus groups
- ◆ These groups have members from the industry and academia working on developing protocol standards (RFC's) and proposed ideas (ID's)
- ◆ **<http://www.ietf.org>** has the complete list of special interest groups and their RFC's and current ID's



# The Motivation

- ◆ Many new applications have different requirements from those for which the Internet was designed
- ◆ New applications need performance and resource assurance
- ◆ Service differentiation is also needed so that the traffic from different applications is treated in service-appropriate way
- ◆ Resource assurance and service differentiation means QoS (Quality of Service)



# IETF's Models

- ◆ It was felt that instead of focusing on coping with congestion, Internet should be run in a way that there is no congestion
- ◆ Applications should be able to reserve or obtain network resources at a given QoS
- ◆ IETF has been working on developing new models and protocols for the Internet
- ◆ During the last decade, Intserv and Diffserv models have been developed



# Integrated Services

- ◆ Intserv stands for “Integrated Services”
- ◆ IntServ provides quantitative guarantees to each flow and requires all intermediate routers to keep track of flows through “soft state”
- ◆ To receive resource reservation, an application describes its requirements
- ◆ The network determines a path based on the request



# Intserv

- ◆ A reservation protocol is used to install the reservation state along the selected path
- ◆ Reservation is enforced by packet classification and scheduling on routers along the path
- ◆ The reservation setup protocol in the Intserv model is the RSVP (Resource ReSerVation Protocol)





# The Control Plane

- ◆ The control plane of Intserv contains following components:
  - ***QoS Routing Agent***: To allow the determination of the next hop for the current request
  - ***Admission Control***: To decide if sufficient resources are available to meet the request
  - ***Reservation Setup Agent***: To install the reservation
  - ***Resource Reservation Table***: To record the soft state for the reserved flow
- ◆ When the packets arrive, the data plane identifies the flow and schedules packets as per reservations



# RSVP's Services

- ◆ RSVP offers two types of services
- ◆ CONTROLLED LOAD service means that the service offered to a flow in an overloaded network is the same as it would get in a lightly loaded network
- ◆ GUARANTEED SERVICE is when a flow gets hard guarantees on the delay it will suffer

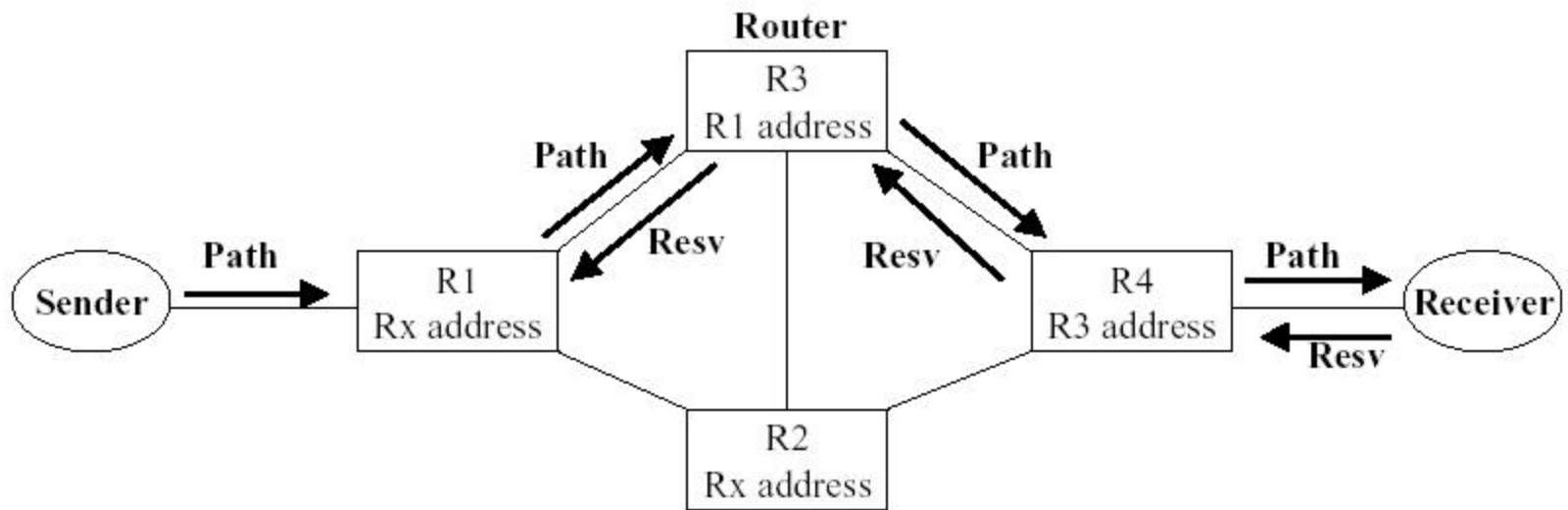


# RSVP Features

- ◆ RSVP makes simplex reservations
- ◆ Support of multicasting is provided by making RSVP receiver oriented
- ◆ It is independent of the routing and policy
- ◆ RSVP installs soft state that may be timed out if not refreshed periodically
- ◆ In RSVP, PATH and RESV messages are sent for installing reservations



# RSVP





# PATH and RESV

- ◆ PATH messages are sent from sources to receivers
- ◆ PATH messages carry source information and path features to the receivers
- ◆ PATH messages also install the necessary state for RESV messages to get back to the sources
- ◆ Receivers can request reservations by sending RESV messages along the exact reverse path of the PATH messages



# RSVP Signaling

- ◆ RSVP relies on extensive signaling for obtaining flow reservations along a path. It also entails soft state overhead and therefore does not scale well to the Internet
- ◆ Most of the Internet traffic consists of short-lived web transactions
- ◆ It will be unwise to go through reservations for such traffic
- ◆ All reservations must be authenticated and accounted; something not developed yet for the Internet
- ◆ RSVP may be successfully deployed in a campus network



# IETF's DiffServ Model

- ◆ Intserv's problems prevented its deployment
- ◆ IETF started developing a new model in 1997 to provide differing levels of service to different applications without the overhead of signaling and state maintenance
- ◆ The DiffServ model uses the TOS field in IPv4 header to affix labels on packets belonging to different service levels
- ◆ DiffServ has the potential to offer QoS on the Internet, *at last!!*



# IETF's DiffServ Model

- ◆ Consider a petrol station, you can buy regular, super or premium gasoline from the same pump
- ◆ DiffServ offers various service levels to the customer from the same network with SLA
- ◆ DiffServ adopts techniques used in ATM for traffic management, in a simplified way





# Diffserv Outline

- ◆ Diffserv works on the basis of dividing the traffic into a small number of forwarding classes
- ◆ For each FEC, the amount of traffic entering the network is controlled at the edge of the Diffserv network
- ◆ FEC's are prioritized, with each one coded into the IP header's TOS byte. Core routers offer priority treatment based on the coding



# How does it differ from Intserv?

- ◆ Diffserv provides resource allocation to aggregated traffic instead of individual flows
- ◆ Diffserv enforces policing at the edge and class based forwarding in the core. Intserv requires all nodes to classify packets and use per flow queuing
- ◆ Diffserv does provisioning instead of reservations
- ◆ Diffserv deployment is incremental instead of end-to-end
- ◆ Diffserv emphasizes long term SLA's instead of per-flow signaling



# IETF's DiffServ Model

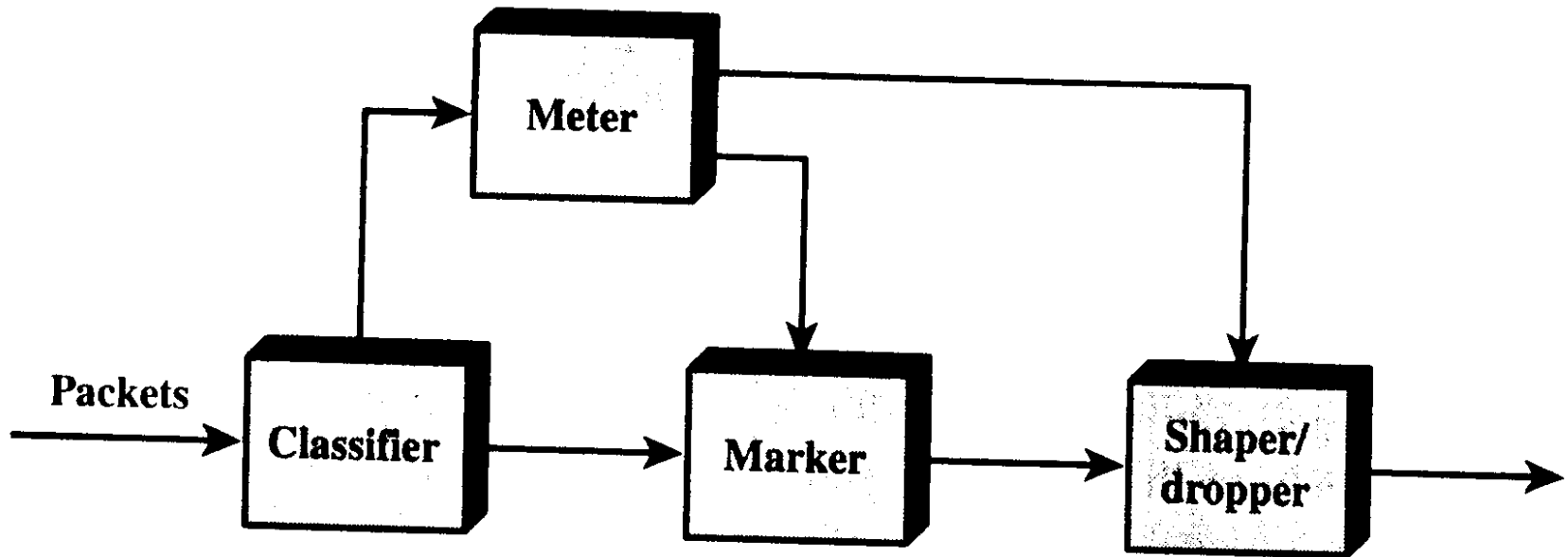
- ◆ DiffServ levels of service are implemented in a DiffServ domain
- ◆ The customer connects to the “edge router” at the edge of the DiffServ domain
- ◆ The edge router performs traffic classification (using DS codepoint marked by customer in TOS to separate the packets)
- ◆ It then measures submitted traffic for conformance to the agreed profile



# IETF's DiffServ Model

- ◆ The edge router then changes the DS code byte of offending packets
- ◆ It may also do traffic shaping by delaying the packets as necessary and dropping the offending packets
- ◆ Diffserv tries to follow the Internet example of keeping the complexity at the edges
- ◆ Refer to the diagram in the next slide to see the edge router function

# Diffserv Traffic Conditioner





# Traffic Treatment

- ◆ Users should agree to a profile of their traffic to avoid unforeseen congestion
- ◆ Some of the important parameters of agreed profile include the committed rate and allowed peak rate
- ◆ Some flows may be violating agreed profile
- ◆ It is important to enforce the policing (metering and marking) mechanisms at the ingress node.
- ◆ Marking is a way to ensure that the user does not violate the agreed profile



# Traffic Treatment

- ◆ The purpose of marking is to indicate if the current packet violates the profile or not
- ◆ Three color marking is considered sufficient with green indicating a good packet, yellow showing a packet that exceeds committed profile but falls within the peak rate and red showing a violation
- ◆ Colors are coded using the drop precedence of the AF (assured forwarding) class



# Single-Rate Three Color Marker

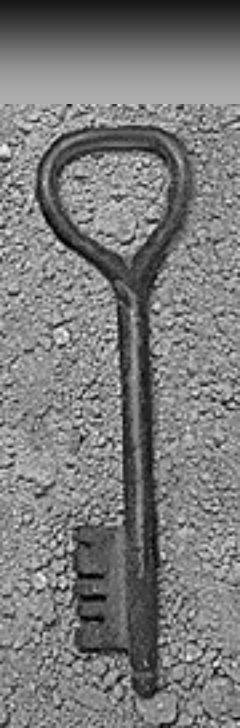
- ◆ srTCM marks the packets according to their length and an agreed rate known as Committed Information Rate (CIR)
- ◆ CIR is applied as token generation rate for two token buckets C and E
- ◆ If no traffic arrives, buckets C and E get filled in this order until CBS or EBS
- ◆ If a packet arrives and it is found less than or equal to the size of C, it is colored green





# Single-Rate Three Color Marker

- ◆ If the packet size exceeds the size of C but does not exceed the size of E, it is marked yellow else it is colored red
- ◆ Marking a packet green results in removing enough tokens from C
- ◆ Marking a packet yellow removes enough tokens from E but marking a packet red does not remove any tokens from C or E
- ◆ Packet length thus determines its color



# Two-Rate Three Color Marker

- ◆ trTCM operates with two token buckets P and C
- ◆ C gets filled with rate CIR to a maximum size of CBS (Committed Burst Size)
- ◆ P gets filled with a rate of PIR (Peak Information Rate) to a maximum size of PBS (Peak Burst Size)



# Two-Rate Three Color Marker

- ◆ Each packet arriving is first checked against the current size of  $P$ . If packet size exceeds the size of  $P$ , it is colored red
- ◆ If packet does not exceed  $P$ , it is checked against  $C$ . If it is larger, it is colored yellow else it is colored green



# Two-Rate Three Color Marker

- ◆ For red packets, neither bucket is modified
- ◆ For yellow packets, only P is decremented
- ◆ For green packets, both P and C are decremented
- ◆ Thus trTCM is useful when a peak rate is also agreed upon besides committed rate



# Time Sliding Window Three Color Marker

- ◆ tswTCM does not use token buckets rather it uses a rate estimator that computes the average rate of the offered traffic
- ◆ The estimated rate is used by the marker in comparing it to CTR (Committed Target Rate)
- ◆ Packets that conform to CTR are marked green



# Time Sliding Window Three Color Marker

- ◆ Packets that exceed CTR but do not exceed PTR are marked yellow
- ◆ Packets contributing to the portion of the rate above PTR are marked red
- ◆ tswTCM is useful for AF class traffic in Diffserv domain
- ◆ Its rate estimator uses a time window that is based on RTT (round trip time) of TCP or 1 second for UDP



# Traffic Shaping

- ◆ Congestion occurs due to the bursty traffic
- ◆ Offered load to the network may include bursts of data followed by relatively inactive time slices
- ◆ If the traffic peaks are smoothed over time, the magnitude of congestion can be controlled



# Traffic Shaping

- ◆ After shaping, the traffic may be more evenly spread over time, thus offering a well behaved load to the network
- ◆ Thus deploying a shaper for the traffic entering the network may result in improved throughput and efficient utilization of resources





# Leaky and Token Bucket Shapers

- ◆ Leaky bucket with a counter and a buffer can convert unregulated flow into a regulated smooth flow. However, a burst larger than the buffer gets discarded
- ◆ Token Bucket allows limited bursts to pass through by accumulating tokens at a fixed rate and letting a burst pass if enough tokens have accumulated



# Per-Hop Behaviors

- ◆ IETF has defined two DS services that are visible as PHB (per-hop-behavior) of an intermediate router for the marked packet
- ◆ EF (Expedited Forwarding)
  - EF is the premium service offered. It can appear as a virtual leased line for the customer. It offers low loss/latency and assured bandwidth
- ◆ <http://www.ietf.org/rfc/rfc2598.txt>



# Per-Hop Behaviors

- ◆ AF (Assured Forwarding)

- The AF PHB group provides delivery of IP packets in four independently forwarded AF classes. Within each AF class, an IP packet can be assigned one of three different levels of drop precedence. A DS node does not reorder IP packets of the same microflow if they belong to the same AF class.

- ◆ **<http://www.ietf.org/rfc/rfc2597.txt>**

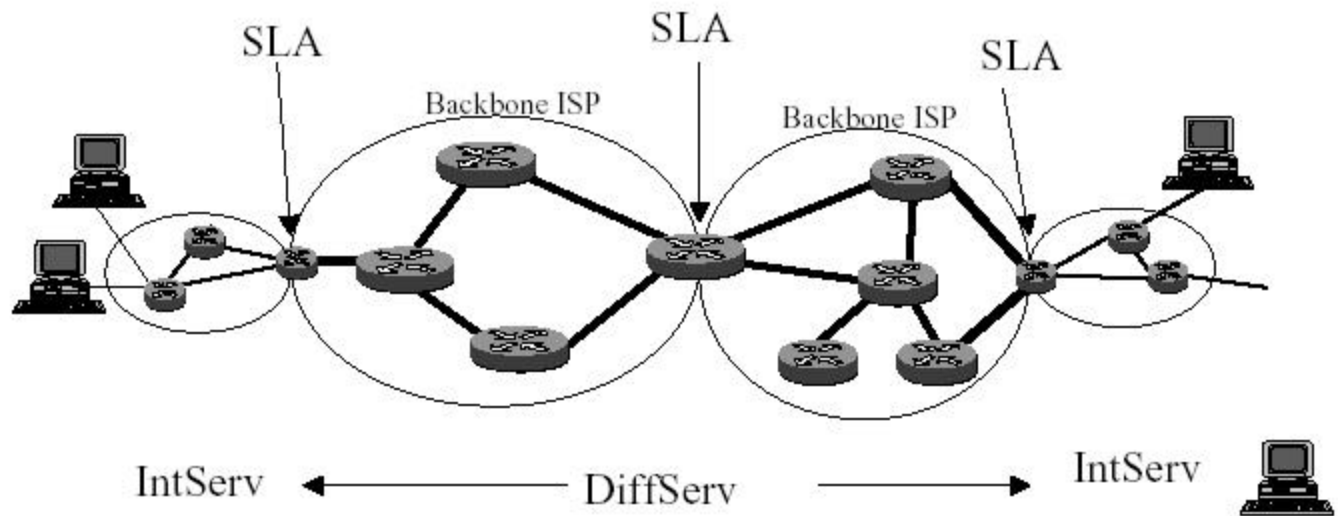


# Hybrid Approach

- ◆ Integrated services model may be applied end to end across a network containing one or more Diffserv regions
- ◆ For example, Diffserv may allocate aggregate bandwidth for a company in the core network. The access router of that company may allocate individual user flows using Intserv
- ◆ Challenges include
  - Selecting an appropriate PHB for Intserv flow (mapping aggregates and flows)
  - Performing appropriate policing
  - Exporting Intserv parameters from Diffserv domains

# Hybrid Approach


## Integration of IntServ and DiffServ





# Motivation for MPLS

- ◆ MPLS (Multi Protocol Label Switching) is a very interesting recent development
- ◆ Let us see why MPLS was developed
- ◆ ATM switches are deployed in the Internet backbones due to their extremely fast switching and provisioning
- ◆ All Internet traffic is based on IP. So IP must be carried over the ATM

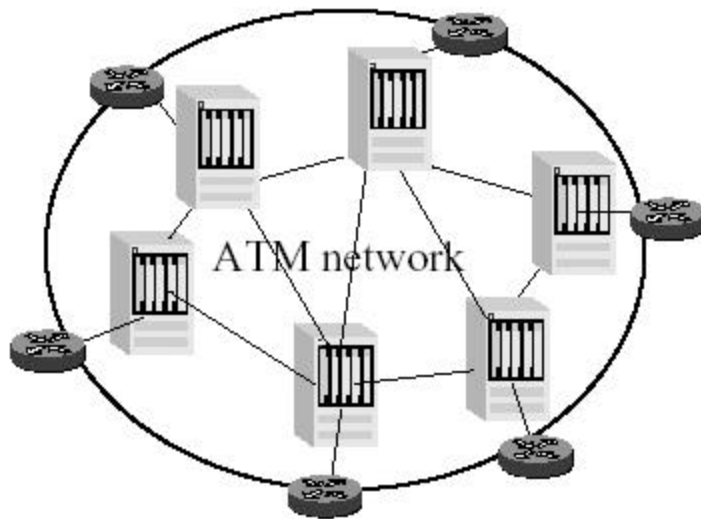


# IP/ATM → MPLS

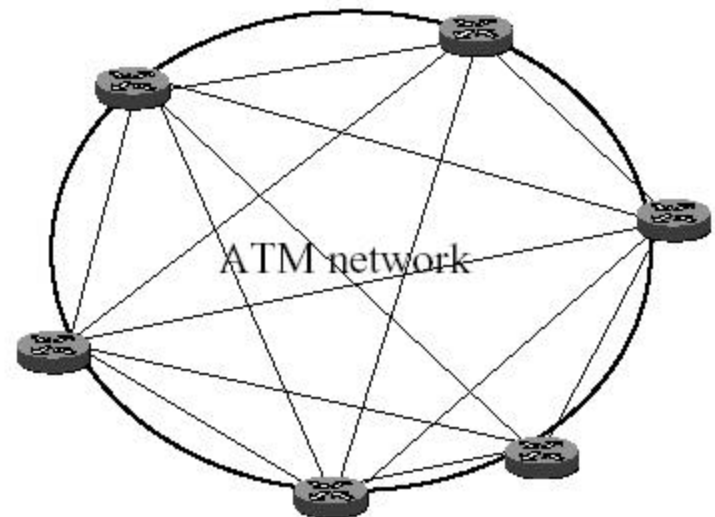
- ◆ Classical IP over ATM (Overlay model) suffers from several problems
- ◆ First, all ATM switches are connected in a mesh. A small increase in the number of switches can drastically increase the number of virtual circuits
- ◆ The QoS features of ATM are not exploited and all connections are best effort
- ◆ IP and ATM have incompatible addressing and control protocols so overlaying is expensive

# IP Over ATM

Traffic Engineering in IP over ATM (overlay) Networks



Physical Topology



Logical Topology





# MPLS

- ◆ The industry developed tag switching and label switching to solve the above problems
- ◆ In label switching, a short fixed length label is encoded into the packet
- ◆ The intermediate LSR (Label Switched Router) finds the next hop from a table, using the label as an index
- ◆ If the LSR is an ATM switch, label is just the VPI/VCI identifier
- ◆ If the LSR is an IP router, the label helps eliminate the destination based routing and reduces the job of the router to label switching

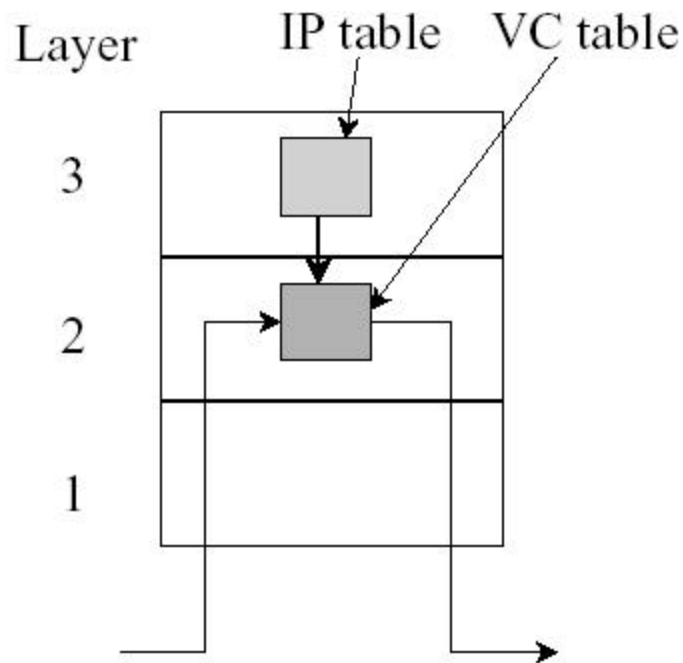


# MPLS

- ◆ A label switched path (LSP) must be set up prior to the start of transmission
- ◆ IP and ATM are tightly integrated with label switching
- ◆ IP takes over the control path and ATM switches are used only for data transmission
- ◆ IP can use the ATM switches as label switched nodes (or IP routers)

# IP Over ATM With MPLS

## Structure of a Label Switching Router



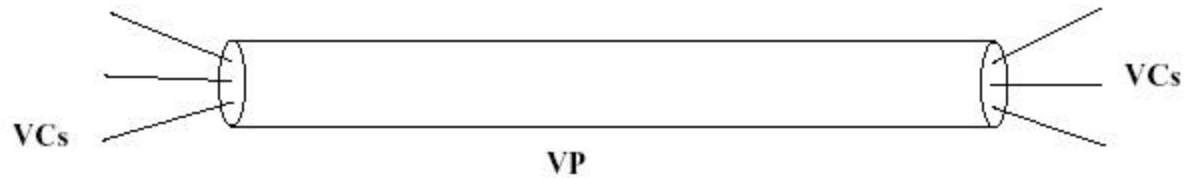
- Layer 3 (IP) performs routing and label distribution
- Layer 2 (ATM) performs fast forwarding



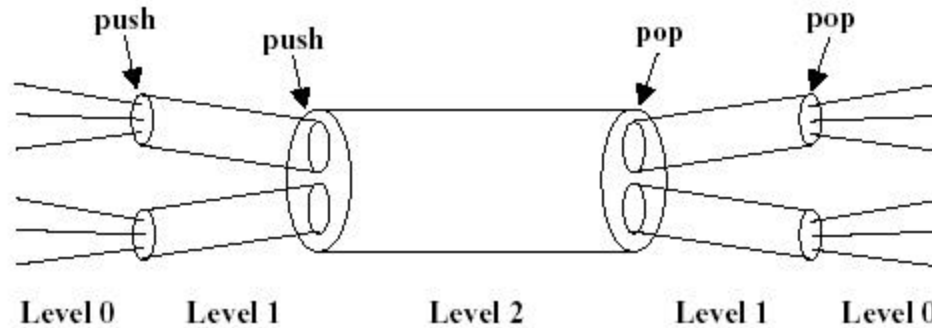
# LSP Hierarchy

## Tunnel Hierarchy and Label stacks

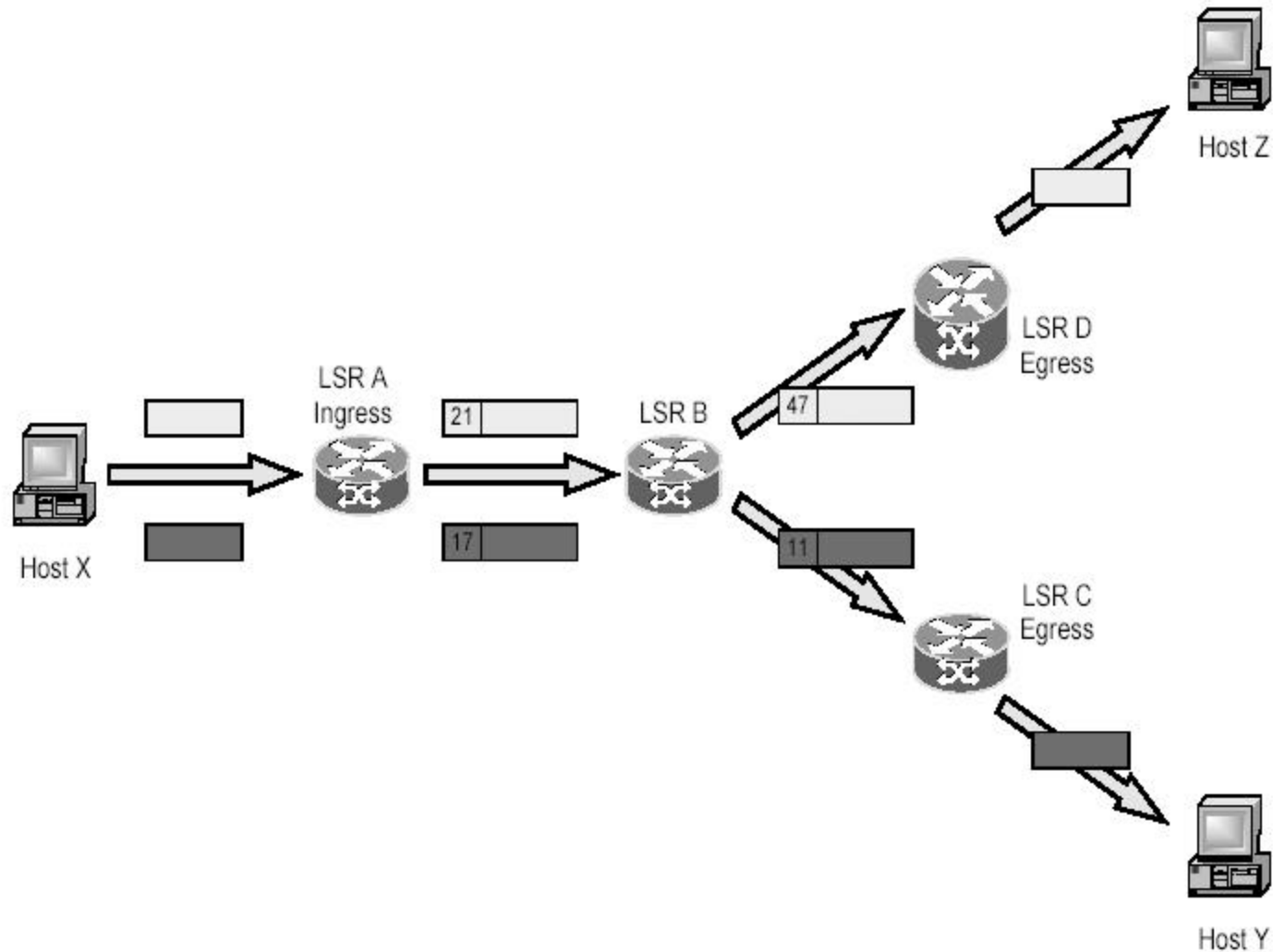
In ATM, two levels of hierarchy



Label stack allows for arbitrary levels of hierarchy



# LSP's in an MPLS Network





# MPLS

- ◆ MPLS simplifies the routing problem in an all IP subnet
- ◆ An MPLS domain has an ingress node that nails down paths through the maze of core routers for every requesting flow until the exit door (egress node)
- ◆ Thus every router does not have to decide about the path of each packet
- ◆ In MPLS, the connectionless network is converted into connection oriented network



# MPLS

- ◆ Intermediate routers use a “shim header” or a layer 2.5 header to find out the next hop of a packet
- ◆ This shim header is inserted between the frame header and packet header
- ◆ It is used by the router to consult a table that tells what path is to be taken for this packet



# MPLS

- ◆ Instead of routing, now the routers do label switching
- ◆ Since the path is pre-determined, routers can speed up the processing of packets
- ◆ Also, the network manager can decide LSP's (label switched paths) based on load distribution and other administrative goals
- ◆ Directing traffic on paths not determined by traditional IGP's provides flexibility and load balancing





# Traffic Engineering

- ◆ TE optimizes the network efficiency with the control of the
  - Mapping
  - Distribution
- ◆ Of the traffic across the network
- ◆ TE tries to balance the load across the network and addresses fault tolerance and congestion avoidance



# Traffic Engineering

- ◆ Earlier, the routing protocols favored shortest or least cost paths, building up congestion on some paths
- ◆ TE was not practiced, leaving the network overloaded in some parts and underutilized in others



# MPLS AND TE

- ◆ MPLS runs constrained routing to determine an LSP within an MPLS domain.
- ◆ This LSP will run from an ingress node to an egress node of the domain
- ◆ LSP may have some QoS features, based on the algorithm used
- ◆ The path could be strictly specified or loosely outlined and backup paths may be specified for handling link failures



# TE

- ◆ The LSP setup may follow TE principles thus solving the chronic inefficient utilization problem of the networks
- ◆ For example, constrained routing may prefer longer and lightly loaded paths over shortest paths
- ◆ MPLS + TE → Balanced and well utilized network



# Automated Provisioning

- ◆ The networks are growing bigger!!
- ◆ The protocols are becoming more complex
- ◆ With Diffserv, MPLS, RSVP-TE, CR-LDP, COPS and associated protocols, it is impossible to allow manual provisioning
- ◆ Therefore, there is a need for automated TE-based path selection algorithms



# Constrained Routing

- ◆ Constrained routing applies extended IGP parameters to the tree to find a suitable path
- ◆  $BW_{avail}$  and hop count may be used to determine paths
  - Shortest widest path
  - Widest shortest path
  - Shortest distance path ( $dist = 1/BW_{avail}$ )



# QoS Traffic Considerations

- ◆ If only the available bandwidth is considered, the class of service may not be taken into consideration
- ◆ Thus, the best effort traffic may intersect the QoS traffic at several points within the domain
- ◆ In Diffserv, this may be a recipe for disaster!!



# TELIC

- ◆ An efficient dynamic traffic engineering algorithm is developed for selecting paths across an MPLS-Diffserv domain
- ◆ TELIC (Traffic Engineering with Link Coloring) works with a set of traffic requests present at an ingress router of such a domain
- ◆ It allocates paths to an egress node using Dijkstra's shortest path algorithm





# TELIC

- ◆ Each request specifies the amount of bandwidth requested followed by the Diffserv class of service (EF,AF,DF)
- ◆ While processing a request, TELIC partitions the network into several monochromatic subgraphs and makes an effort to match the request with an appropriate subgraph



# TELIC

- ◆ In case a subgraph has no path to the egress node, TELIC merges it with another subgraph as per rules carefully built-in and starts the search all over again
- ◆ In case a search is exhausted, rules are available to deallocate a best effort class LSP and start the search again
- ◆ TELIC is written as a flexible tool in C++



# Software Operations

- ◆ Traffic requests are read in and placed in a FIFO queue
- ◆ The program will then:
  - Look at the type of request
  - create sub-graphs based on color and available bandwidth to find the best match for a request



# Software Operations

- ◆ If a path is found, the links on the path are updated to reflect the increase in usage
  - Higher cost, less bandwidth, different color
- ◆ Otherwise, the request is not allocated, and the next traffic request in the queue is processed



# Software Features

- ◆ Variable bandwidth requests
- ◆ Domains and traffic requests can be placed in files so multiple configurations may be tested



# Software Features

- ◆ Results displayed:
  - Bandwidth utilized
  - LSP Table (Traffic allocated and its path)
  - Overall condition of the domain
- ◆ Object-Oriented design promotes easy adaptability
  - Visual C++
- ◆ Questions? Email [Track605@aol.com](mailto:Track605@aol.com)
- ◆ TELIC results were presented in Applied Telecommunications Symposium (part of ASTC'02)



# GMPLS

- ◆ Recently the industry has gravitated towards GMPLS (Generalized MPLS) as the control plane solution for automatic lightpath setup and teardown in optical networks
- ◆ GMPLS is an extension of MPLS
- ◆ The Internet backbone must use optical switching instead of electronic switching to handle the projected huge bandwidth
- ◆ MPLS cannot handle non-packet routing
- ◆ GMPLS allows control and provisioning of non-packet devices



# GMPLS

- ◆ Using GMPLS, it is possible to perform switching based on:
  - Wavelengths
  - Wavebands
  - Timeslots
  - Ports
  - And Labels





# GMPLS

- ◆ For example, in an all-optical switch, there may be thousands of tiny mirrors that can be moved by miniature motors
- ◆ Switching can be done by adjusting a mirror so that light entering from one fiber can be reflected (switched) to the desired path forward
- ◆ Optical switching is thus entirely different from packet switching

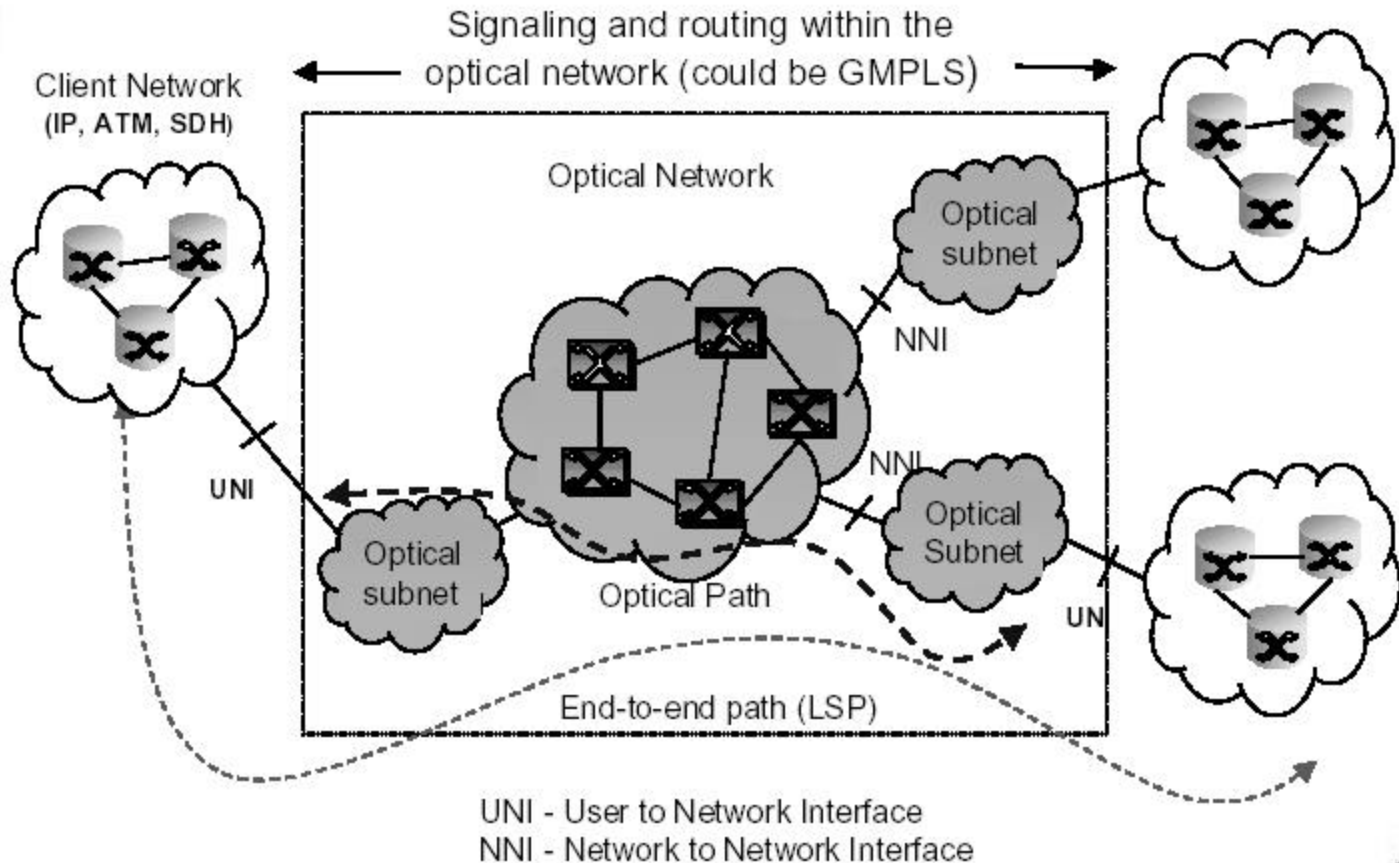


# LMP

- ◆ A link management protocol has been developed for GMPLS. It provides link provisioning, fault isolation and link aggregation
- ◆ Selection of label in MPLS → Selection of wavelength and OXC port in GMPLS
- ◆ MPLS LSP → GMPLS lightpath
- ◆ Before GMPLS, control and provisioning of optical network could take weeks!!
- ◆ Vendors were also reluctant to de-provision due to any changes

# End to End Provisioning

## Lightpath Provisioning: Architecture





# Summary

- ◆ We have taken a detailed look at the Internet and how it is changing
- ◆ MPLS and Diffserv are being combined to provide EF paths to certain flows such as IP telephony, AF paths to multimedia streaming and DF paths to ftp, email etc
- ◆ In future, Internet may be able to provide the QoS that is only enjoyed by telephone and Radio/TV broadcasting