Multiprotocol Label Switching (MPLS) and Applications

How Flows and a Centralized Control Plane Augment Distributed IP Routing

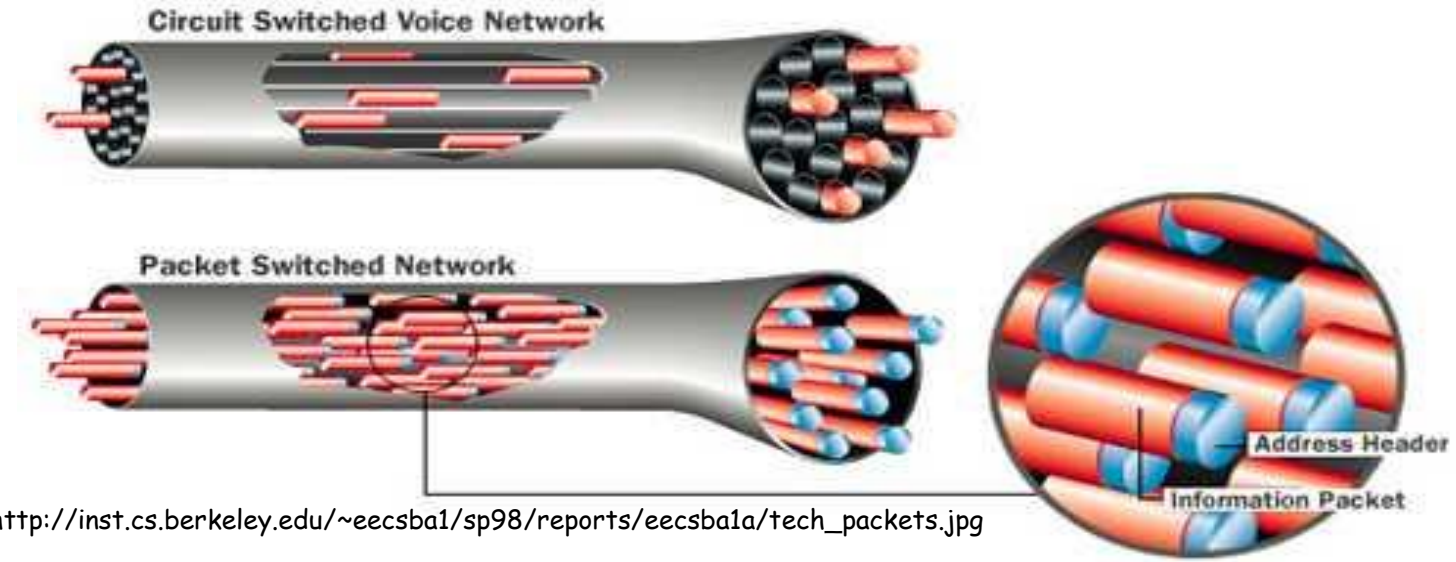
Outline

- Review of Circuit Switching vs. Packet Switching
- Multi Protocol Label Switching (MPLS) Protocol
- Traffic Engineering with MPLS
- Path Computation Engine Architecture and Protocol

REVIEW OF CIRCUIT SWITCHING VS. PACKET SWITCHING

Packet switching vs. circuit switching

- Packet switching
  - Data traffic divided into packets
    - Each packet contains its own header (with address)
    - Packets sent separately through the network
  - Router performs longest prefix matching at each hop
  - Destination reconstructs the message
  - Example: sending a letter through postal system
- Circuit switching
  - Source first establishes a connection to the destination
    - Each router on the path may reserve bandwidth
  - Switches send data on a predetermined path
  - Source sends data over the connection
    - No destination address, since routers know the path
  - Source tears down the connection when done
  - Example: voice conversation on telephone network circa 1970



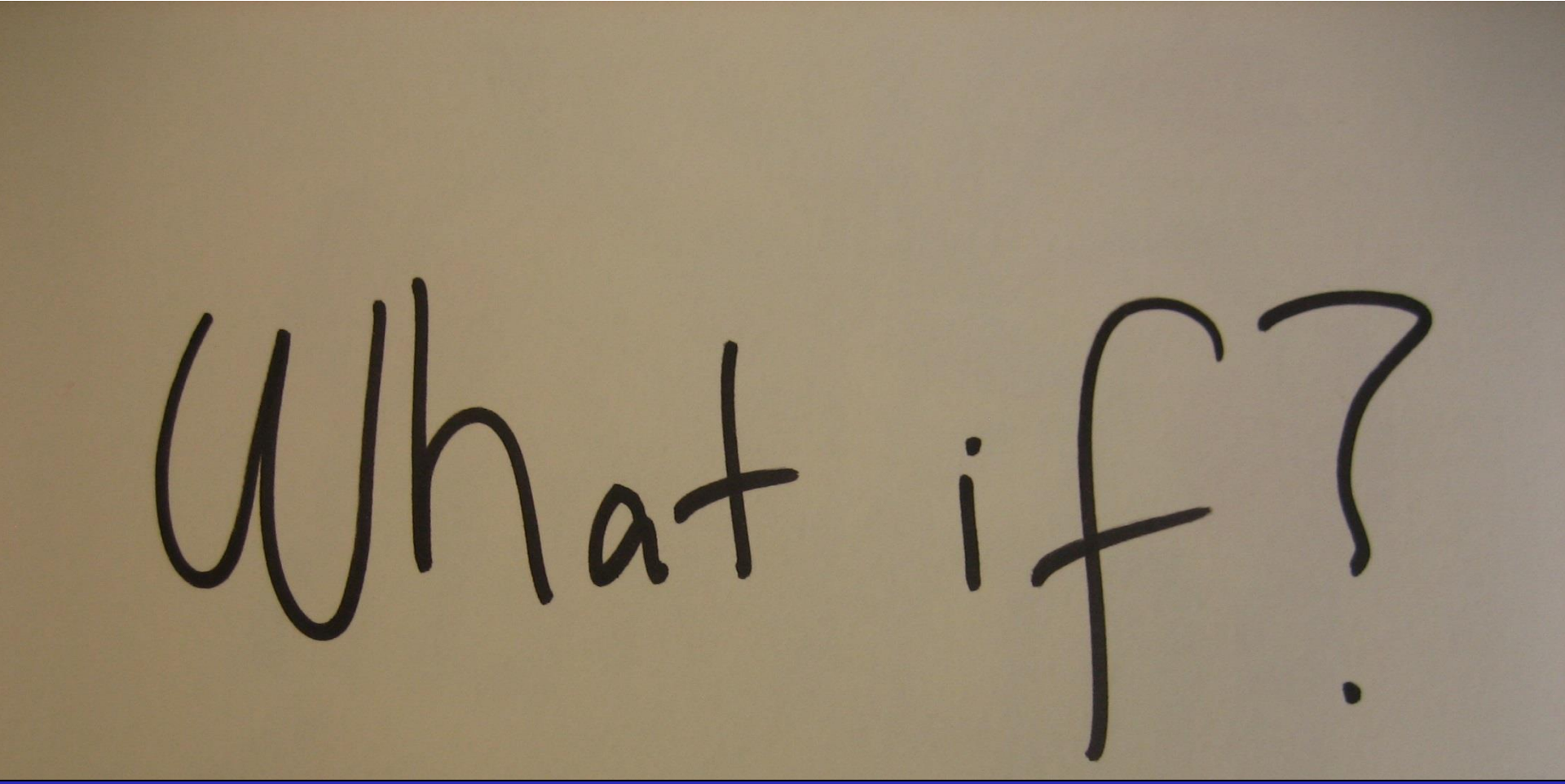
Source:http://inst.cs.berkeley.edu/~eecsba1/sp98/reports/eecsba1a/tech_packets.jpg

Advantages of circuit switching

- Guaranteed bandwidth
  - Predictable communication performance
  - Not "best-effort" delivery with no real guarantees
- Simple abstraction
  - Reliable communication channel between hosts
  - No worries about lost or out-of-order packets
- Simple forwarding means cheaper hardware
  - Forwarding based on time slot or frequency
  - No "longest prefix match" on each packet
- Low overhead
  - Only data sent, control plane context kept in switches
  - No IP, TCP, UDP headers on packets

Disadvantages of circuit switching

- Wasted bandwidth
  - Bursty traffic leads to idle connection during silent period
  - Unable to achieve gains from statistical multiplexing
- Blocked connections
  - Connection refused when bandwidth is not sufficient
  - Unable to offer "okay" service to everybody
- Connection set-up delay
  - No communication until the connection is set up
  - Unable to avoid extra latency for small data transfers
- Network state
  - Switches must store per-connection information
  - Unable to avoid per-connection storage and state failover



We could have all the advantages of circuit-switching without any of the disadvantages?

Source:http://www.danielfusco.com/wp-content/uploads/what-if.jpg

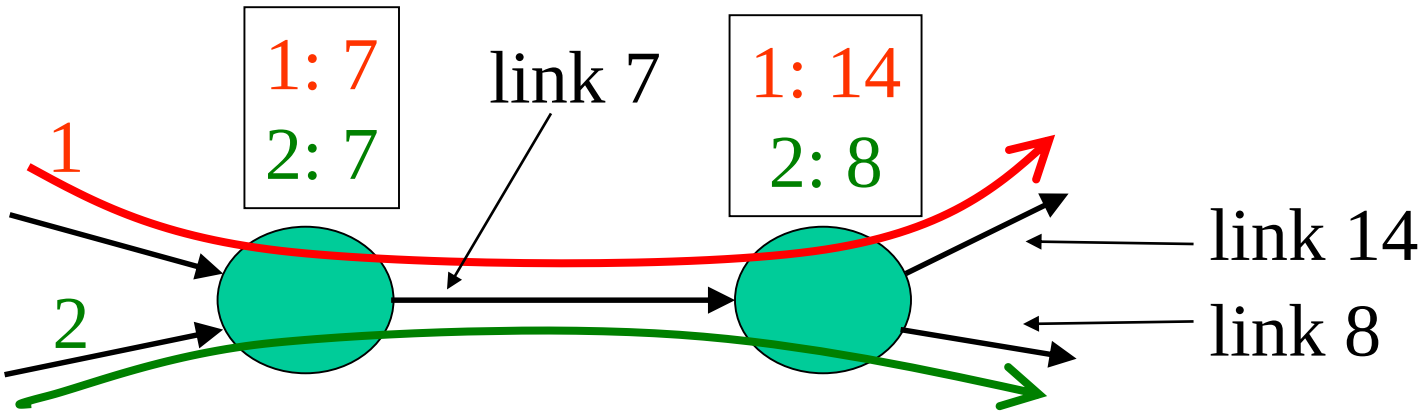
Asychronous Transfer Mode (ATM)

- Layer 2 standard developed in the late 1980s and early 1990s
  - Alternative to Ethernet
- Packets divided into equal sized cells
  - ATM: 53 bytes per cell
  - Ethernet: packet sizes are variable
- Virtual circuits set up between sender and receiver
  - Could be long lasting
- Asynchronous time division multiplexing used at the physical layer
- Widely deployed in operator networks in mid to late 1990s
- ATM superseded by optical switching for transport networks

Source:http://penta2.ufrgs.br/rc952/trab2/atm_forum.gif

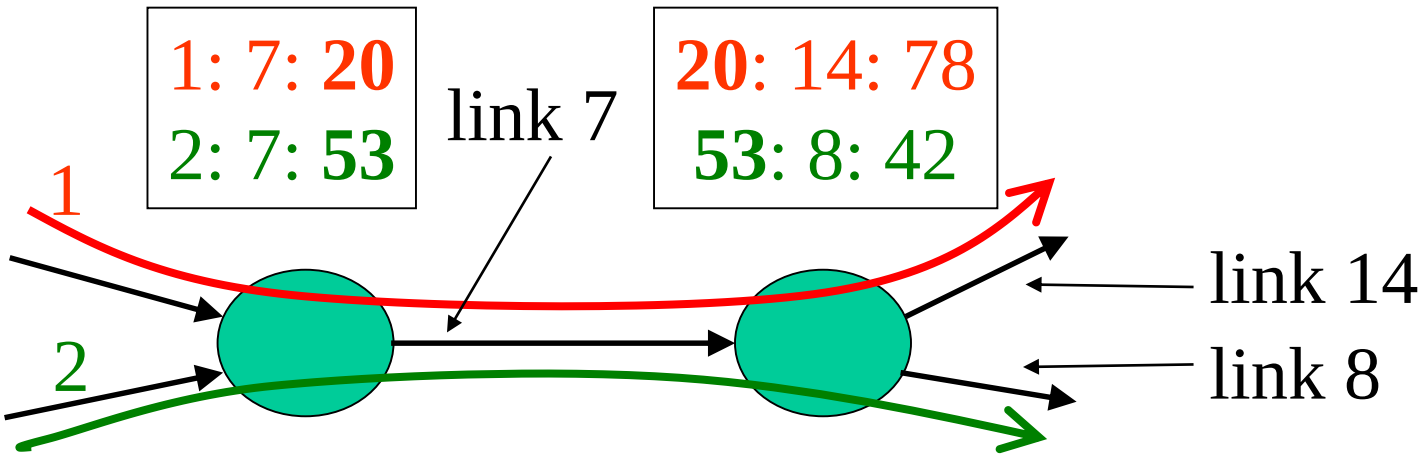
Virtual circuits

- Hybrid of packet and circuit switching
  - Logical circuit between a source and destination
  - Packets from different VCs multiplex on a link
- Virtual Circuit Identifier (VC ID)
  - Source set-up: establish path for the VC
  - Switch: mapping VC ID to an outgoing link
  - Packet: fixed length label in the header



Swapping the label at each hop

- Problem: using VC ID along the whole path
  - Each virtual circuit consumes a unique ID
  - Starts to use up all of the ID space in the network
- Label swapping
  - Map the VC ID to a new value at each hop
    - Table has old ID, next link, and new ID
  - Allows reuse of the IDs at different links



Virtual Circuit Data Format

- Similar to IP datagrams
  - Sender divides data into packets
- Packet has an address
  - IP address for IP
  - VC ID for virtual circuit
- Store-and-forward transmission
  - Multiple packets may arrive at once
  - Buffer packets that can't be immediately forwarded
- Multiplexing on a link
  - No reservations: statistical multiplexing
    - Packets are interleaved without a fixed pattern
  - Reservations: resources for group of packets
    - Guarantees to get a certain number of "slots"

How Virtual Circuits Differ from IP

- Forwarding look-up
  - Virtual circuits: small fixed-length connection id
  - IP: destination IP address 4 or 16 bytes
- Initiating data transmission
  - Virtual circuits: must signal along the path
  - IP: just start sending packets
- Router state
  - Virtual circuits: routers know about connections
  - IP: no state, easier failure recovery
- Quality of service
  - Virtual circuits: resources and scheduling per VC
  - IP datagrams: difficult to provide QoS

MULTI PROTOCOL LABEL SWITCHING (MPLS) PROTOCOL

Multiprotocol Label Switching

- Apply the Virtual Circuit idea from ATM to IP forwarding
- Why Multi Protocol?
  - Originally designed to handle more than just IP
    - IPX (ancient Xerox L3 protocol used by Novell Networks in the 1990's)
    - Appletalk (still used today in some cases)
    - …
    - Some future network protocol?
- Wildly successful
  - Widely adopted by vendors (especially Cisco)
  - Most carriers run an MPLS core
  - Many also run MPLS access/aggregation networks
    - Alternative for access/aggregation is carrier Ethernet (802.1aq)

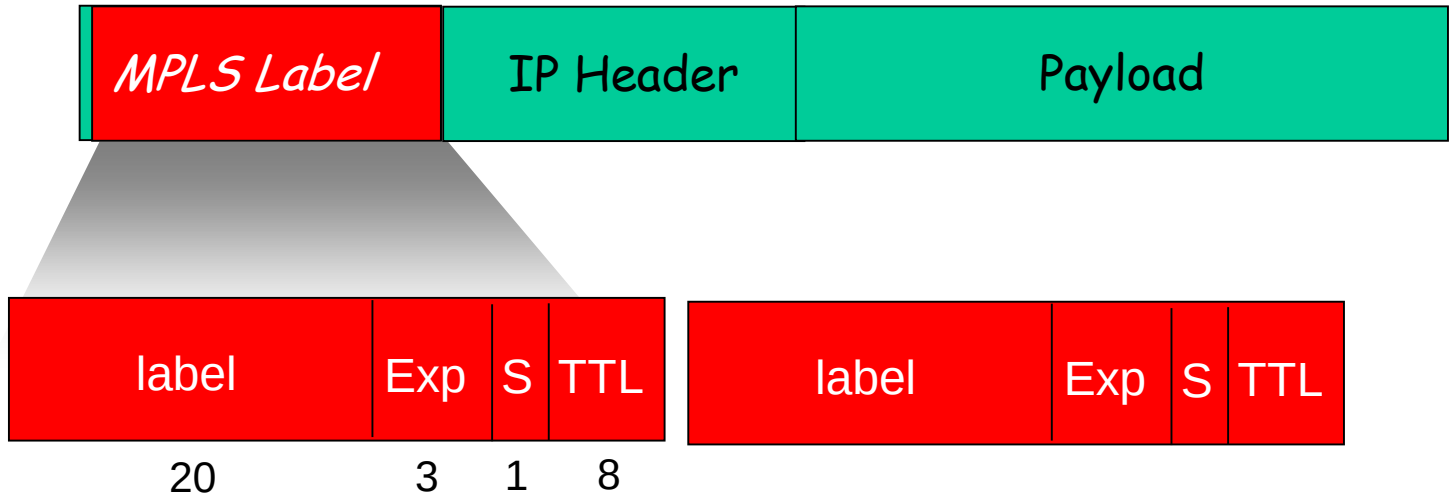
IETF Standardization

- First MPLS Working Group in IETF formed in 1997
- First standards track RFC published in 2001
- WG still going strong today
  - 140+ RFCs
  - 16 WG drafts
- New applications keep popping up
  - Most recent is segment routing

Source:https://www.ietf.org/logo/ietf-logo.gif

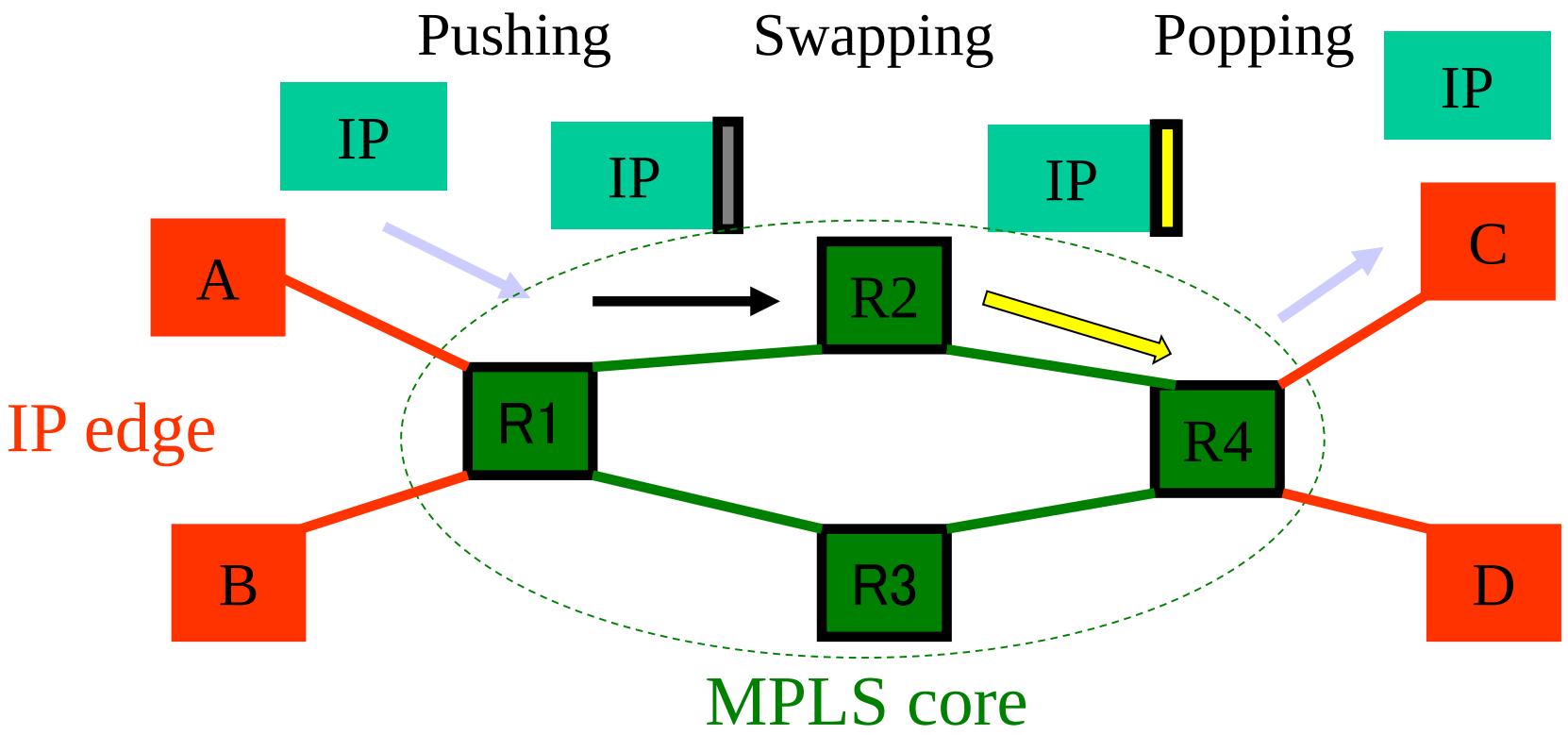
Layer 2.5 Protocol

- Insert label between Layer 2 and Layer 3 header
- Fields
  - 20 bit routing label
  - 3 bit "Exp" field carries packet queuing priority for Class of Service
  - 1 bit "Bottom of Stack" field
  - 8 bit Time To Live field
- Labels can be stacked



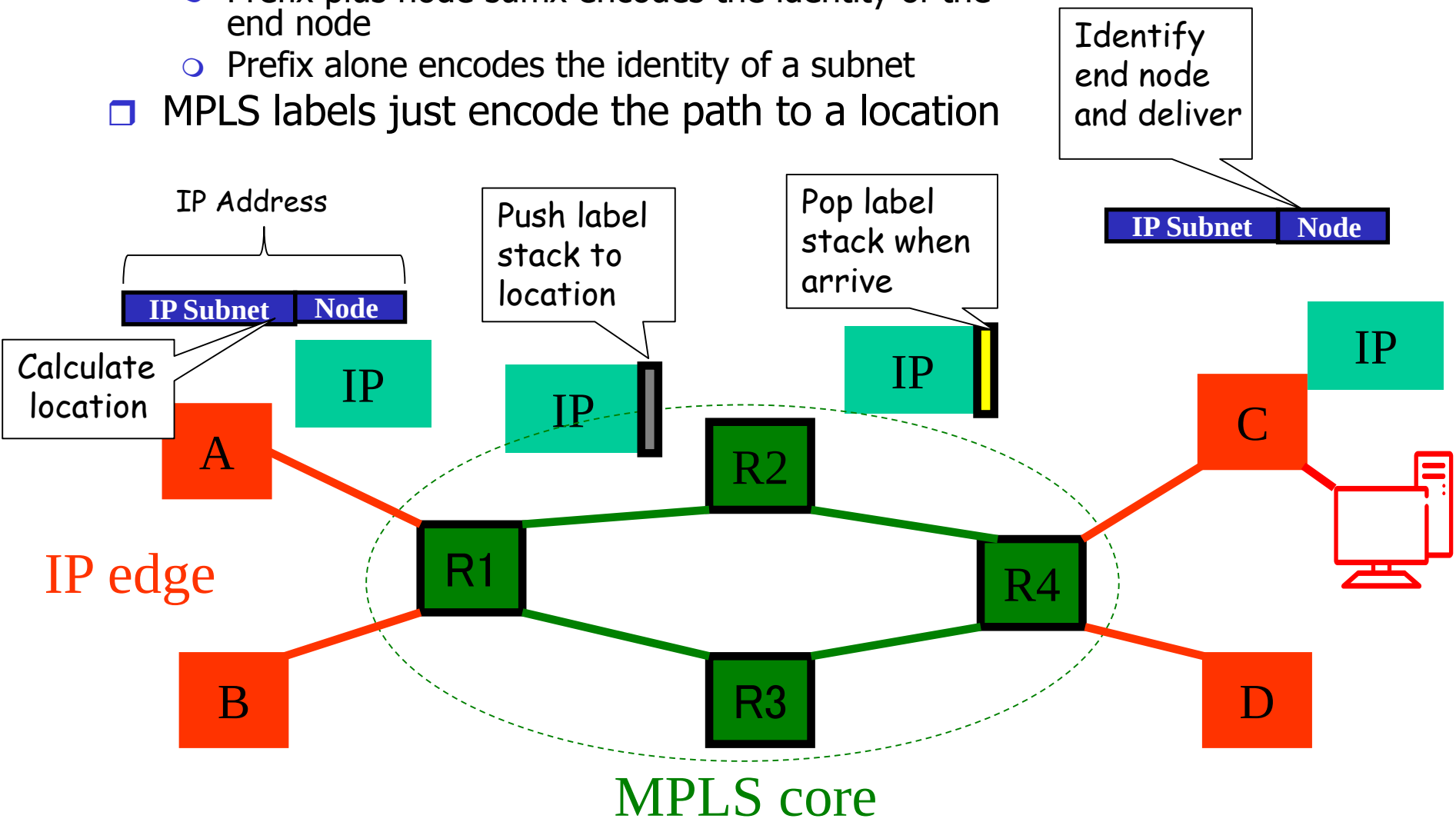
Simple Data Plane Operations

- Pushing: add the initial "in" label (stack)
- Swapping: map "in" label to "out" label (stack)
- Popping: remove the "out" label (stack)



Identity and Location

- An IP address encodes both the location of a node and the node's identity
  - Subnet prefix encodes location in topology
  - Prefix plus node suffix encodes the identity of the end node
  - Prefix alone encodes the identity of a subnet
- MPLS labels just encode the path to a location



Forwarding Equivalence Class (FEC)

- FEC:
  - A rule for grouping packets according to their destination location and forwarding treatment
  - All packets in a FEC are labelled the same way
  - FEC is calculated at the entry point to the MPLS network
- Example FECs
  - Destination prefix
    - Rule: Longest-prefix match in forwarding table to determine route
    - Useful for: Conventional destination-based forwarding
  - Src/dest address, src/dest port, and protocol
    - Rule: Five-tuple match
    - Useful for: Quality of service treatment of the traffic
  - Sent by a particular customer site
    - Rule: Incoming interface
    - Useful for: Virtual private networks
- A label is a locally significant identifier for a FEC

Forwarding Equivalence Class is just another name for an aggregated flow!

Terminology

- A MPLS labelled path is called a *label-switched path* (LSP)
  - One way only
  - For two way, add another label in opposite direction
- An MPLS router at the start/end of an LSP is called a *label-edge router* (LER)
  - Start: classifies packets, inserts label (stack) before fowarding onto LSP
  - End: Pops label (stack), forwards according to IP longest prefix matching
- An MPLS router in the middle of an LSP is called a *label-switched router* (LSR)
  - Forwards according to the top of stack label

Operation of an MPLS Router

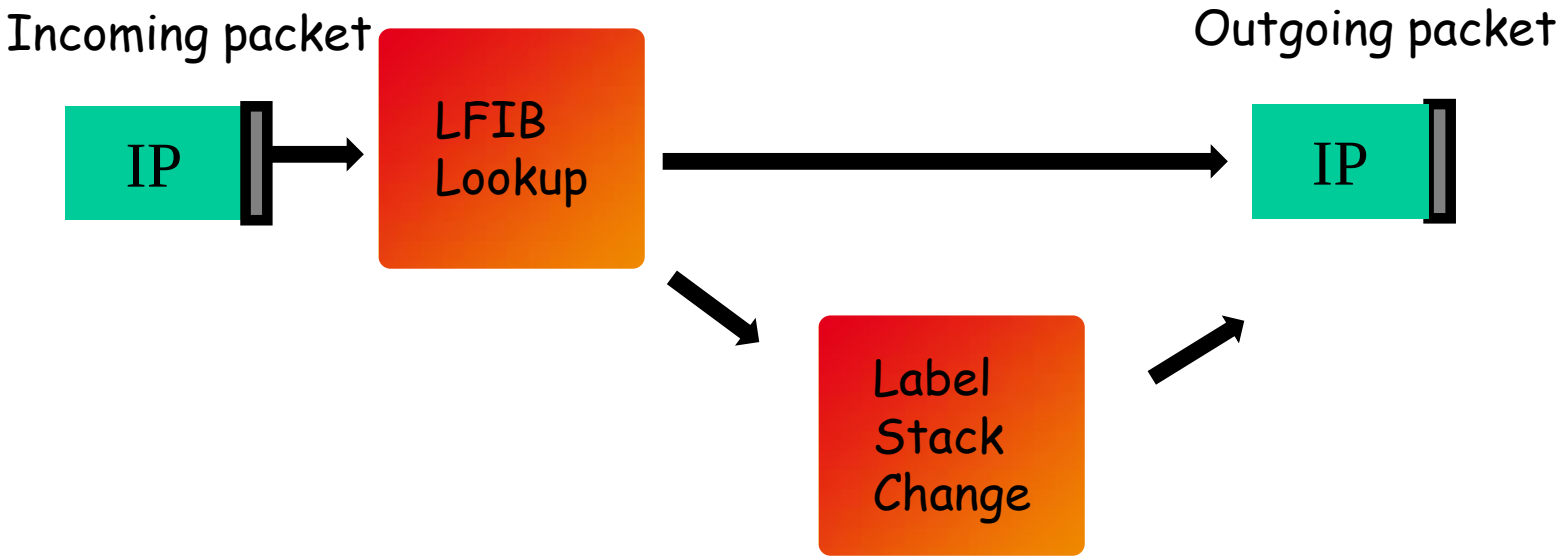
- MPLS control plane sets label to forwarding treatment mapping from outside
  - Control plane is very complex
- An LSR/LER maintains a Label Forwarding Information Base (LFIB)
  - Like FIB, maintained in the line card
  - Soft state, so must be refreshed periodically
- LFIB contains MPLS stack operation for each label value
  - Push, pop, swap
  - Outgoing interface for each label value

Operation of a Start LER

- FEC classifier examines IP header and determines which label (stack) to insert
- Label (stack) insertion puts label (stack) between IP and Ethernet header
- LFIB lookup determines what outgoing interface to use
- Labelled packet sent through router switch fabric to outgoing interface

Operation of an LSR

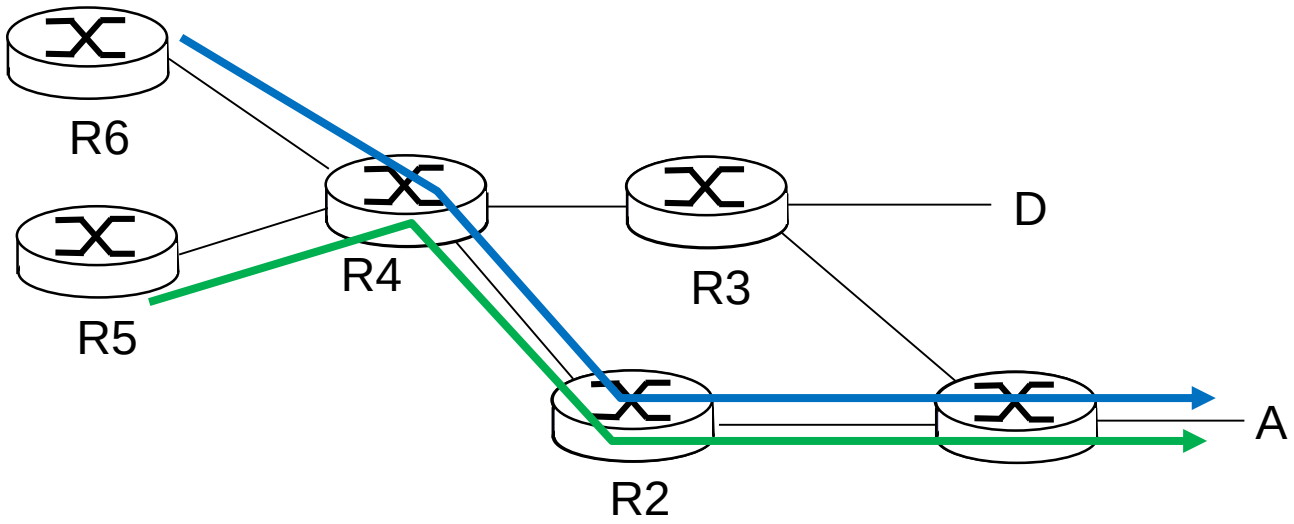
- Look up Top of Stack label in the LFIB
- Does label stack need changing?
  - If so push, pop or swap
- Labelled packet sent through the router switch fabric to outgoing interface



Operation of an End LER

- Examine top of stack label
  - Is "Bottom of Stack" bit set?
  - Is label value "Explicit NULL Label"?
    - 0 for IPv4
    - 2 for IPv6
- Pop label and send to appropriate IP forwarding engine

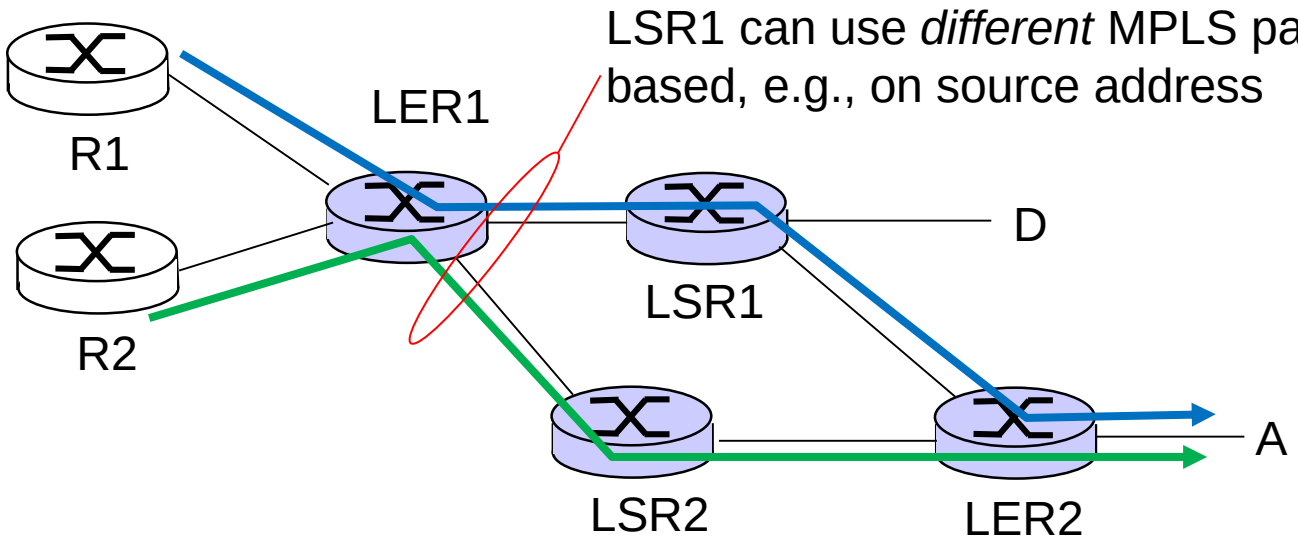
MPLS versus IP paths



- *IP routing: path to destination determined by destination address alone*

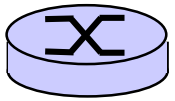
IP router

MPLS versus IP paths



- *IP routing: path to destination determined by destination address alone*
- *MPLS routing:* path to destination can be based on source *and* dest. address

IP-only router



MPLS and IP router

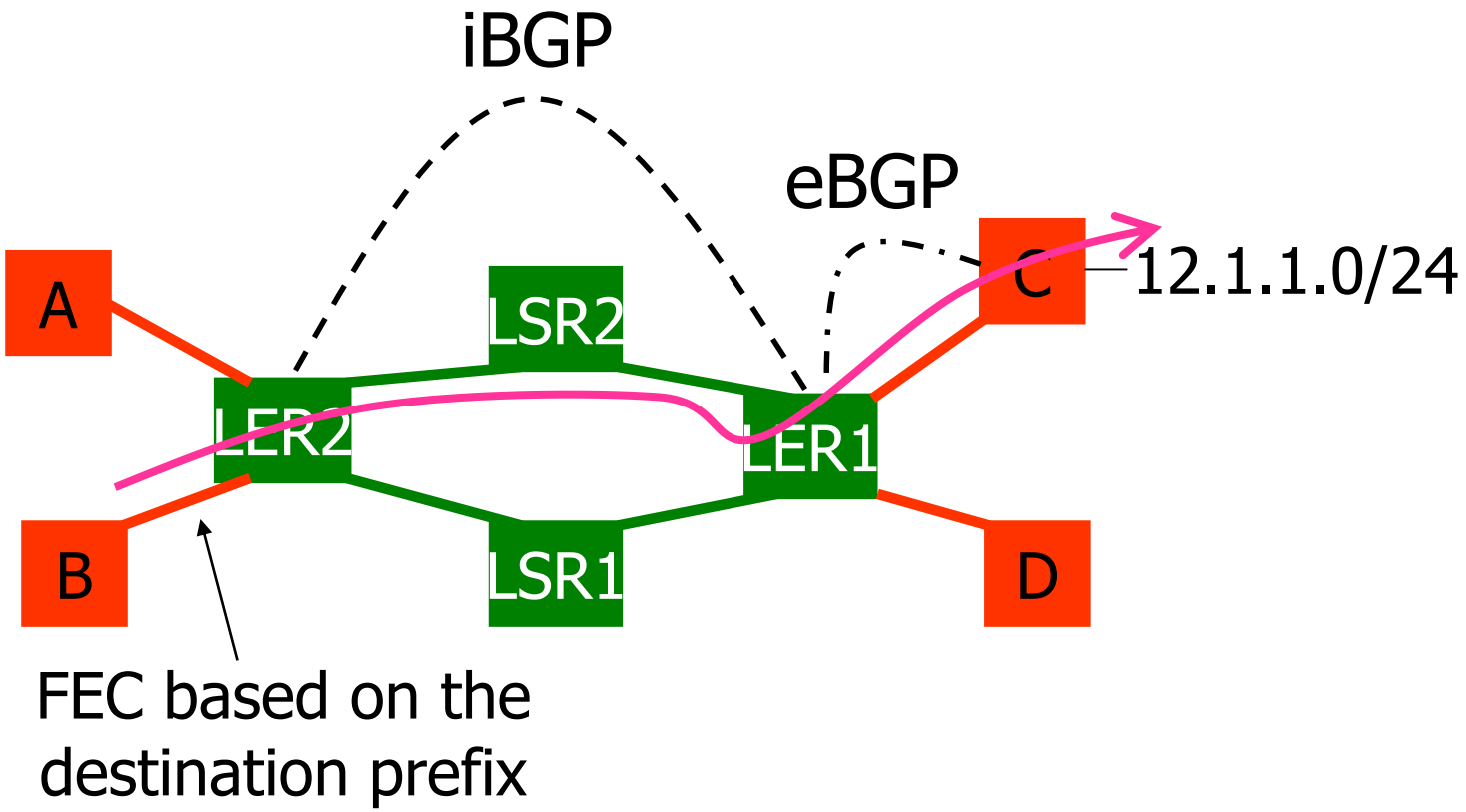
MPLS forwarding decisions can differ from IP!

Complex Control Plane

- MPLS Control Plane is very complex
- A new control plane for each new application
  - E.g. TE, VPN, etc.
- Adapted from existing protocols
  - Extensions added to routing protocols
  - Some protocols repurposed for MPLS applications
    - E.g. RSVP
- MPLS specific protocols
  - Label Distribution Protocol (LDP)

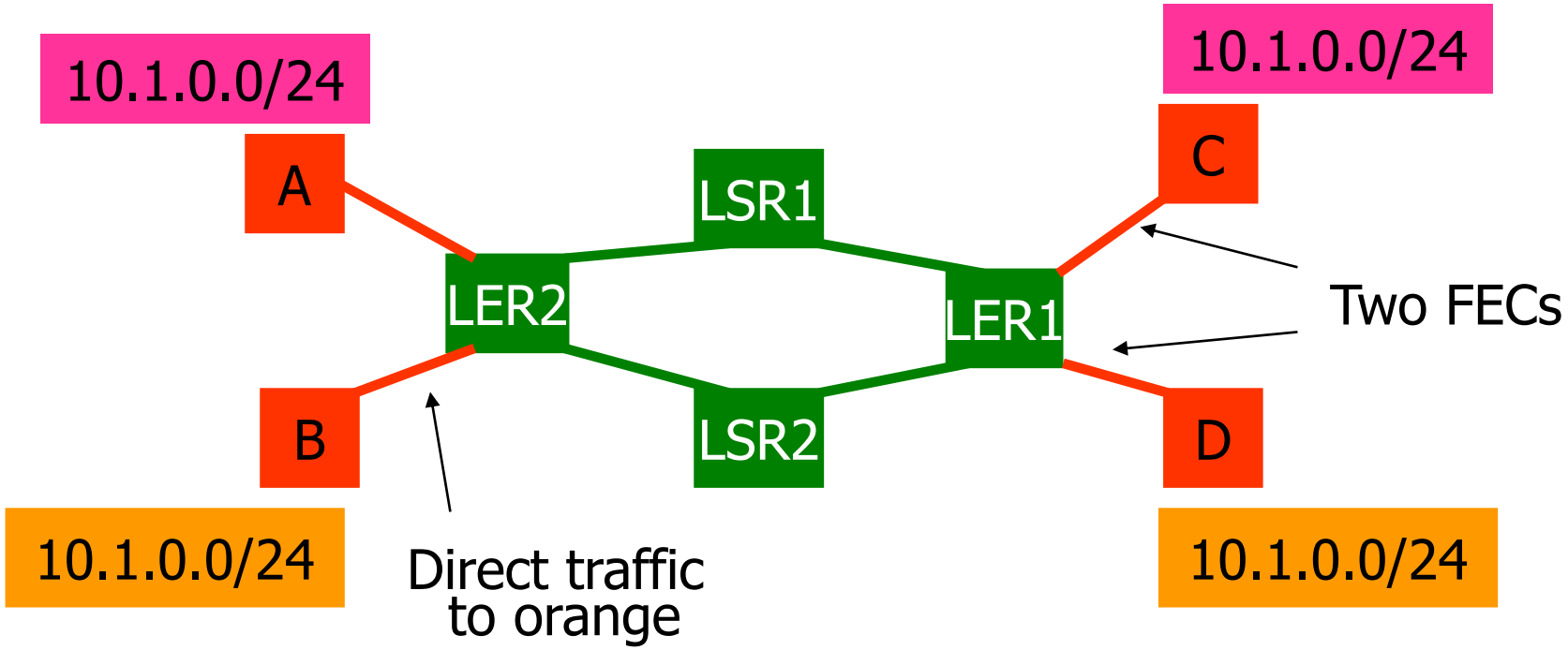
Source: http://www.savvycomsoftware.com

Example 1: BGP-Free core



- Routers LSR1 and LSR2 don't need to speak BGP
  - Complexity reduction because iBGP peers need to be fully meshed
- Control Plane Protocol
  - IGP (OSPF, IS-IS) extended to distribute topology and traffic information
  - Label Distribution Protocol (LDP) used to set up LSPs

Example 2: VPNs with private addresses



- Different LSPs used for green VPN and orange VPN
- Control Plane depends on type of VPN
  - L3VPN: Multiprotocol BGP
  - L2VPN
    - VPLS: Label Distribution Protocol (LDP) or network management system
    - Ethernet VPN (eVPN): Multiprotocol BGP

TRAFFIC ENGINEERING WITH MPLS

What is Traffic Engineering?

- Traffic Engineering:
  - Control and optimization of routing, to steer traffic through the network in the most cost effective way
- Costs:
  - Cost of congestion
  - Cost of violating customer Service Level Agreements (SLAs)
  - Transit costs
- Decision variables:
  - Bandwidth
  - Latency
    - For voice and real time video traffic
- Why not just use routing?
  - Link state IGPs only use one equal cost path at a time
  - May want to send some traffic over a higher cost link

Source: wikipedia.org

Traffic Engineering (TE) with Constraint-Based Routing

- Path calculation uses constrained shortest-path first
  - Compute shortest path based on weights
  - Exclude paths that do not satisfy constraints
    - Constraint example: insufficient bandwidth
- Constraint information dissemination
  - Extend OSPF/IS-IS to carry the constraint information
    - Link-state attributes for available bandwidth
- LSRs calculate Traffic Engineering Database (TED)
- Path signaling
  - Establish label-switched path on explicit route with RSVP-TE
- Forwarding
  - MPLS labels

Traffic Engineering Database (TED)

- With distributed control plane, every router in the Traffic Engineered network calculates a TED
- For every link between router$_i$ and router$_j$ in the TE domain:
  - Administrative group (color)
  - Traffic engineering metric for TE calculation
  - Topology info
    - Local interface IP address
    - Remote interface IP address
  - Bandwidth
    - Maximum link bandwidth
    - Maximum reservable bandwidth
    - Available bandwidth

TED Example

```
NodeID: R5.00(10.0.0.5)
  Type: Rtr , Age: 103 secs, LinkIn: 3, LinkOut: 3
  Protocol: IS-IS(2)
   To: R1.00(10.0.0.1), Local: 10.1.15.2, Remote: 10.1.15.1
    Color: 0x100 red
    Metric: 10
    Static BW: 155.52Mbps
    Reservable BW: 155.52Mbps
    Available BW [priority] bps:
       [0] 155.52Mbps   [1] 155.52Mbps  [2] 155.52Mbps  [3] 155.52Mbps
       [4] 155.52Mbps   [5] 155.52Mbps  [6] 155.52Mbps  [7] 155.52Mbps
    Interface Switching Capability Descriptor(1):
      Switching type: Packet
      Encoding type: Packet
      Maximum LSP BW [priority] bps:
       [0] 155.52Mbps   [1] 155.52Mbps  [2] 155.52Mbps  [3] 155.52Mbps
       [4] 155.52Mbps   [5] 155.52Mbps  [6] 155.52Mbps  [7] 155.52Mbps
   To: R4.00(10.0.0.4) , Local: 10.1.45.2, Remote: 10.1.45.1
    Color: 0 <none>
    Metric: 10
    Static BW: 155.52Mbps
    Reservable BW: 155.52Mbps
    Available BW [priority] bps:
       [0] 155.52Mbps   [1] 155.52Mbps  [2] 155.52Mbps  [3] 155.52Mbps
       [4] 155.52Mbps   [5] 155.52Mbps  [6] 155.52Mbps  [7] 155.52Mbps
    Interface Switching Capability Descriptor(1):
      Switching type: Packet
      Encoding type: Packet
      Maximum LSP BW [priority] bps:
       [0] 155.52Mbps   [1] 155.52Mbps  [2] 155.52Mbps  [3] 155.52Mbps
       [4] 155.52Mbps   [5] 155.52Mbps  [6] 155.52Mbps  [7] 155.52Mbps
  [...Output truncated...]
```

- This is Node 5
- 3 input links and 3 output links
- Type is router (could also be network)
- Softstate age is 103 seconds
  Routing protocol is IS-IS Level 2

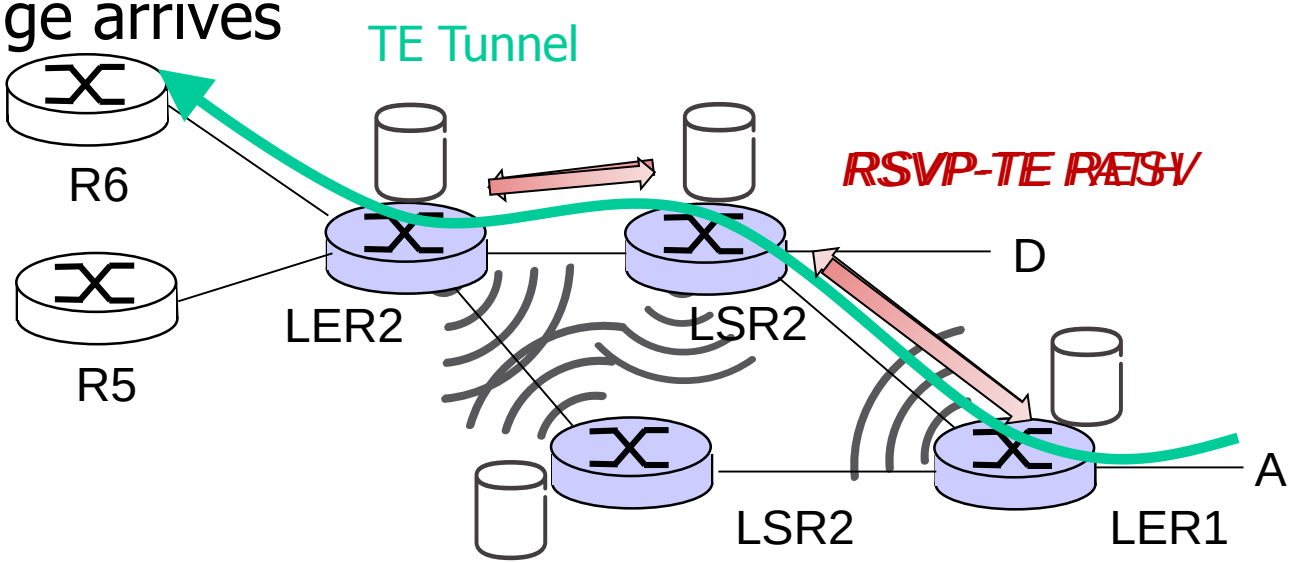
- Link to Node 1 has local IP 10.1.15.2 and remote IP 10.1.15.1

- Administrative group is red
- LS-TE Metric is 10

- Static Bandwidth is 155.52 Mbps
- Reservable Bandwidth is 155.52 Mbps
- Available bandwidth by Priority (3 bitToS) level

Control Plane for MPLS TE

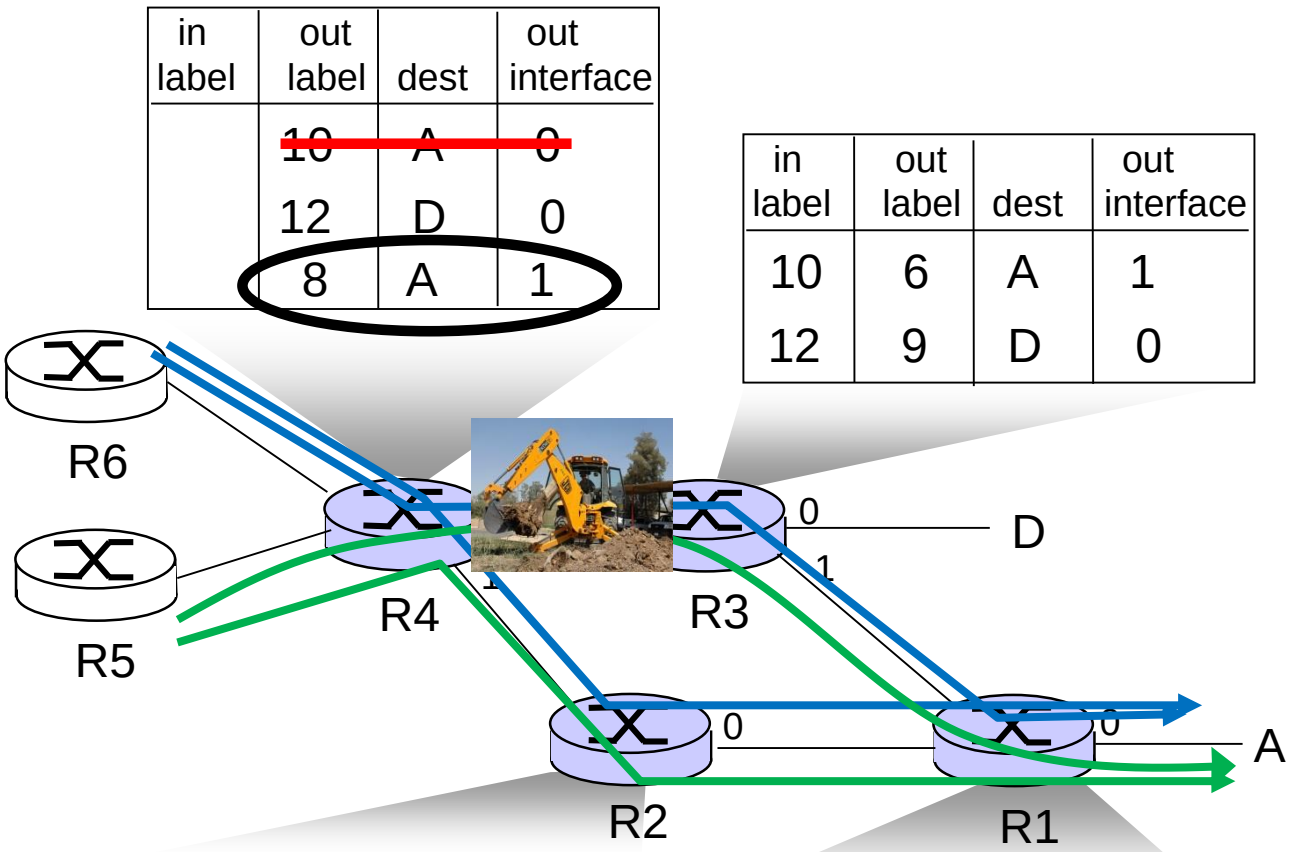
- LS Routers in core floods TE info using, e.g., OSPF, IS-IS
  - Link bandwidth, amount of "reserved" link bandwidth
- Calculate TED
- Entry (head) end LER signals *forward* along path to request a tunnel with RSVP-TE PATH message
- Exit (tail) end LER signals *backward* along path to reserve bandwidth with RSVP-TE RESV message
- Entry router knows it can use the tunnel when the RESV message arrives



Path Protection and Fast Reroute

- Realtime media (voice, real time video) requires maximum failover time
  - 60 ms for VoIP
- Reserve bandwidth on an alternate route
  - Protect a label-switched path by having a stand-by
- Precise control over where the traffic will go
  - Stand-by path can be chosen to be physically disjoint
- Ensure fast recovery from a link failure
  - LFIB has forwarding rule for backup path at a lower priority
- How upstream router detects path failure
  - Upstream router sends heartbeat packets every 10 ms
  - Downstream router detecting failure sends failure message to upstream router
- Upstream router fails over to backup path

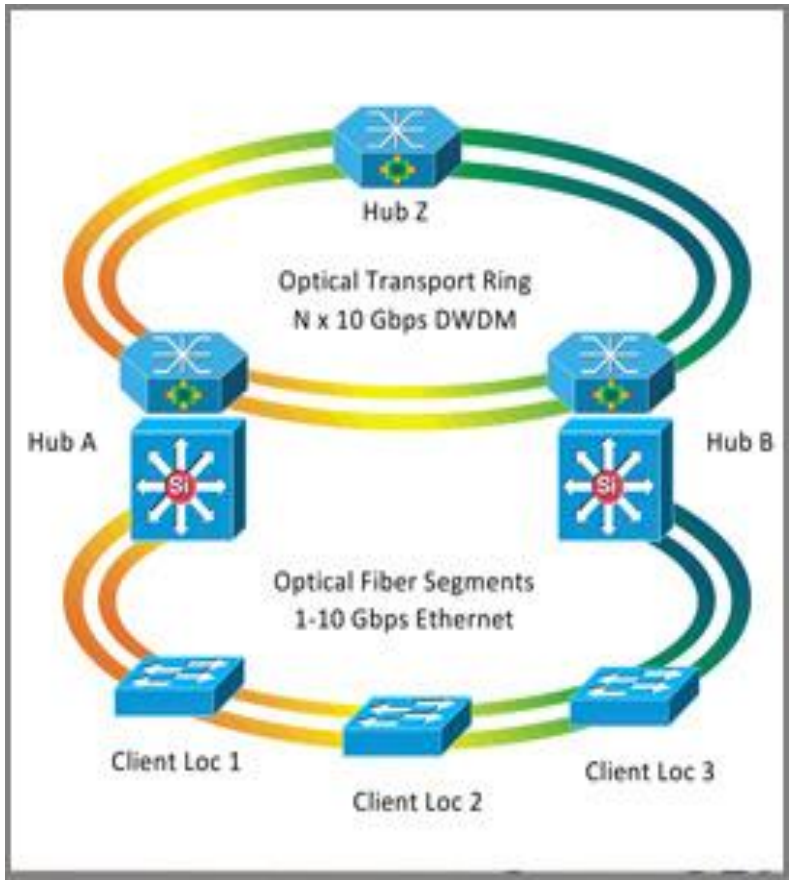
MPLS Fast Reroute



PATH COMPUTATION ENGINE ARCHITECTURE AND PROTOCOL

Use Case #1: MPLS Extension to Optical Transport Networks

- In the mid-2000's MPLS extended to optical transport networks
  - Full circle:
    - 1980/early 90's: ATM
    - late 1990's:MPLS for routers
    - 2000's: MPLS for optical switches
- Generalized MPLS (GMPLS)
  - Labels can be mapped to optical transport network constructs
    - Circuits
    - Wavelengths
    - Time-slots
- MPLS now used for "multi-layer" networks
  - L3 routing
  - L2 switching
  - L1 WDM/TDM



Source: metrons.com

Use Case #2: InterAS VoIP

- In the mid-2000's incumbent network operators started sending all voice traffic over IP networks
  - "Multi-service" networks
  - IP Multimedia Subsystem control plane
  - Today only the last hop from the analog telephone uses old analog technology
- Problems:
  - Maintaining real time traffic classification across transit domains and a peering/transit points
  - Optimal routing between operators
  - Optimal routing within an operator
    - Different administrative domains (ex. wireless, WAN)
    - Different ASes in Tier 1s (ex. NTT, NTT International)

Why a Centralized Path Computation Engine/Element (PCE)?

- Multilayer networks
  - No visibility from routers with TED when calculating paths
    - Routers handle L3 for IP routing
    - Optical switches handle L2/L1 for transport
- Constraint-based path computation in a large, multi-domain network takes too much CPU time for control plane processor
  - Optical switches might not even have computation capacity
- TED database may contain too much information for individual forwarding elements to handle
  - Some optical switches don't have much control plane memory
- Establishing paths between different IGP routing areas may be inhibited from limited visibility
- Operator may have particular policy rules that need enforcement

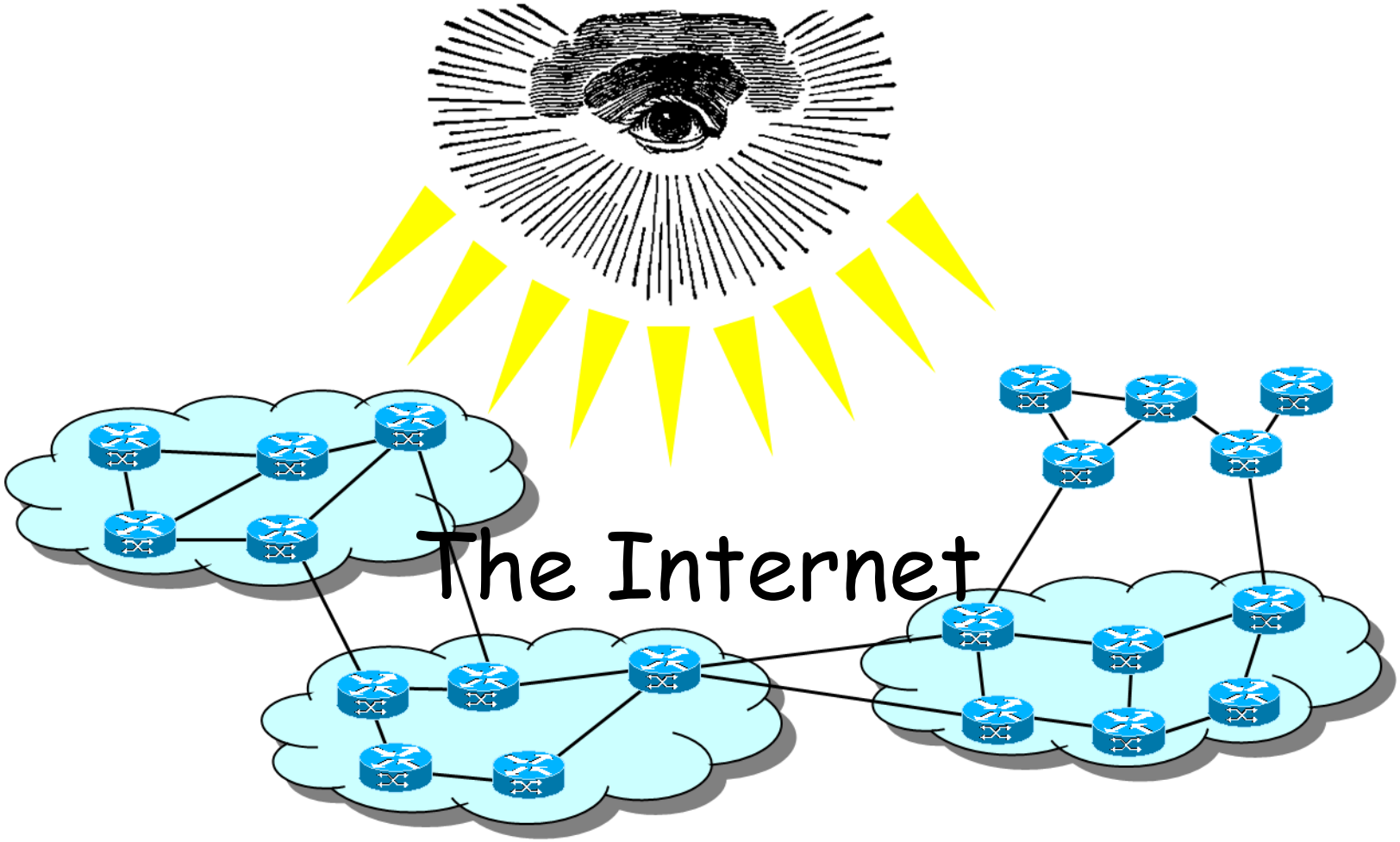
Solution: Centralized Path Computation Engine

PCE Separates Concerns

- PCE control function
  - Collect topology and traffic engineering information from inside a particular administrative domain
  - Respond to requests for LSPs through the domain
  - Respond to transit requests
    - Requests from PCEs external to the domain about optimal paths through the internal domain
- Routing/switching control plane function
  - Set up and maintain LSPs through L1/L2 and L3 devices within an administrative domain
  - Ensure that LSPs connect up with optimal transit points where traffic enters/exits
- Data plane function
  - Route traffic as congestion free and fast as possible in accordance with their traffic classification

The PCE allows the application of appropriate computational power where it is needed

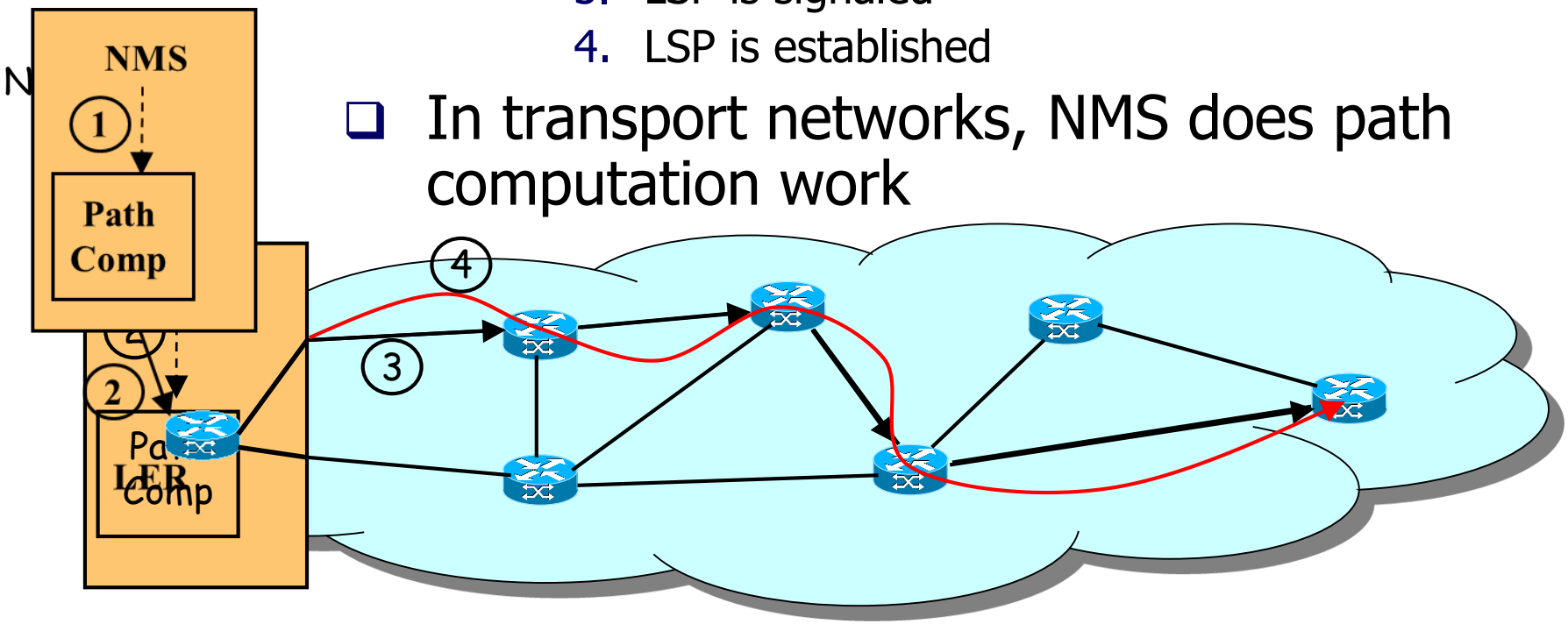
What the PCE is Not!



Source: A. Farrell, "Path Computation Element Tutorial", MPLS 2008

Path Computation In Existing Systems

- For MPLS LERs:
  1. NMS sends request to the LER asking for an LSP
  2. LER performs a path computation
  3. LSP is signaled
  4. LSP is established
- In transport networks, NMS does path computation work



Source: A. Farrell, "Path Computation Element Tutorial", MPLS 2008

Path Computation Element Protocol

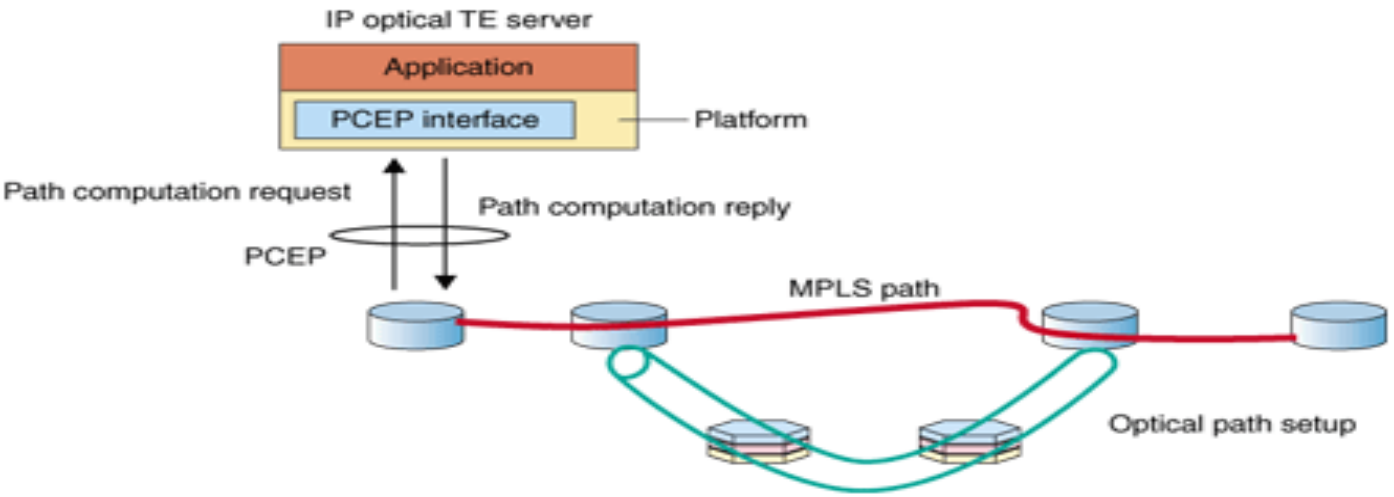
- PCEP allows a data plane element to request a path from the PCE
  - Data plane element is called a Path Computation Client (PCC)
- Operates over TCP
  - Reliable and in-order delivery
  - Security delegated to TCP (TLS, etc.)
- Session-based protocol
  - PCE and PCC open a session
  - Negotiate parameters and learn capabilities
  - All message exchanges within the scope of the session

PCEP Messages

- Open
  - Set up session, negotiate capabilities
- Keepalive
  - Heartbeat
- Request
  - Ask for a path
- Response
  - Respond with a path
- Notify
  - PCE notifies of various conditions, e.g. currently overloaded
- Error
  - Protocol error (e.g. malformed packet)
- Close
  - Close session

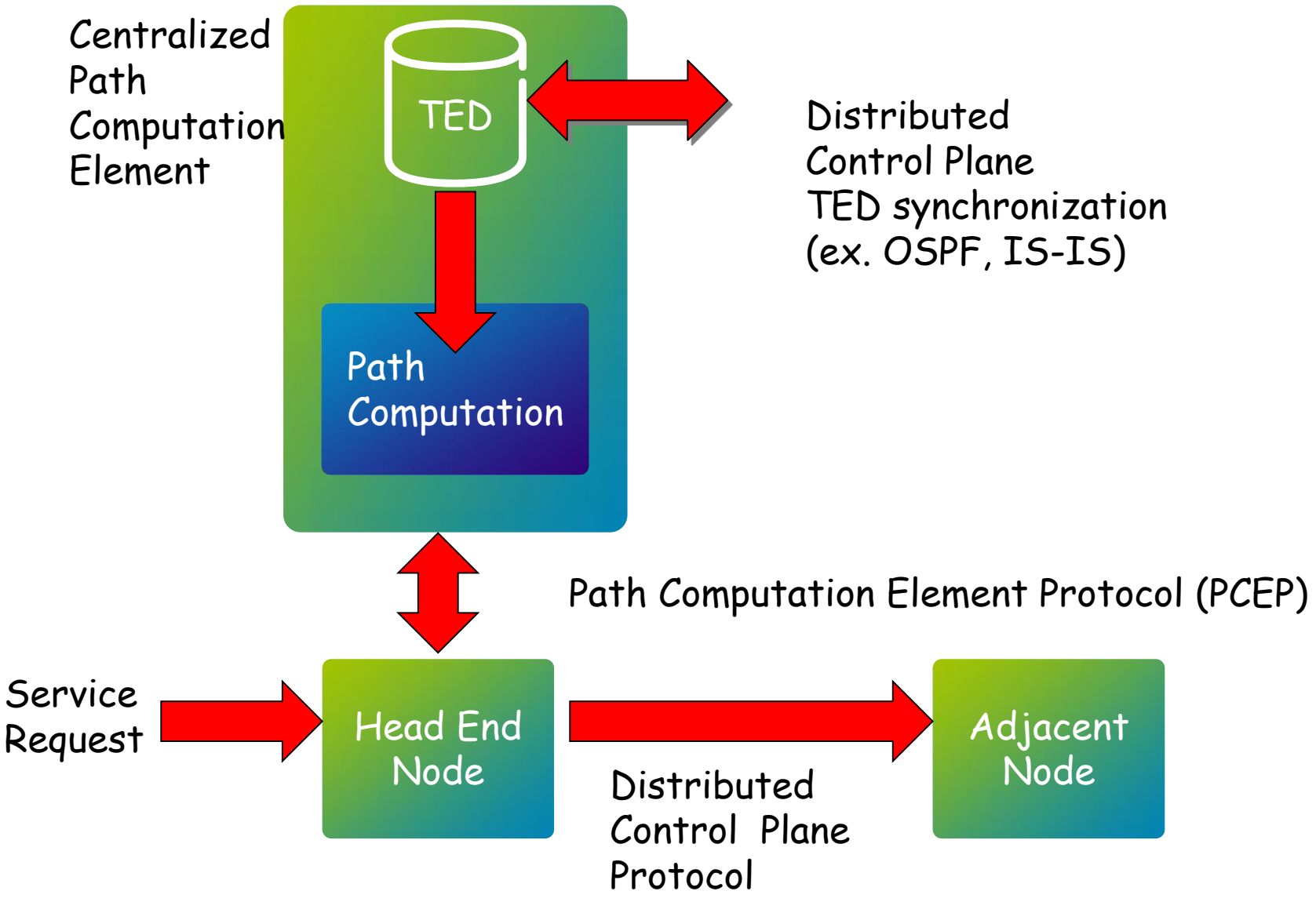
Request / Response Message

- Request message provides:
  - Start and end points
  - Basic constraints
    - Bandwidth
    - LSP attributes
    - Setup/holding priorities
    - Path inclusions
  - Metric to optimise
    - IGP metric
    - TE metric
    - Hop count
  - Associated paths
- Response reports the computed path:
  - Explicit route
  - Actual path metrics
  - Or the failure to find a path

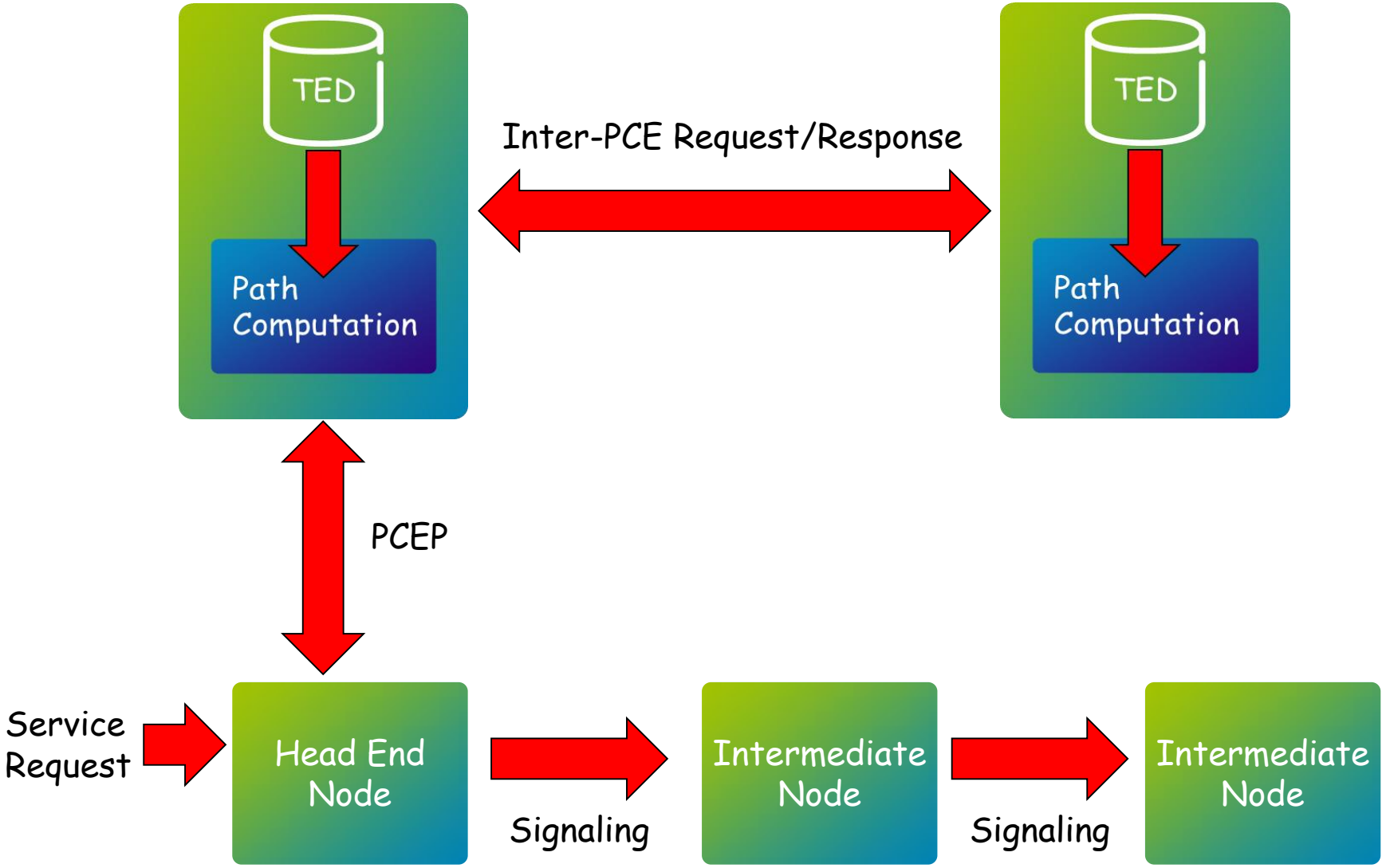


Source: ntt-review.jp

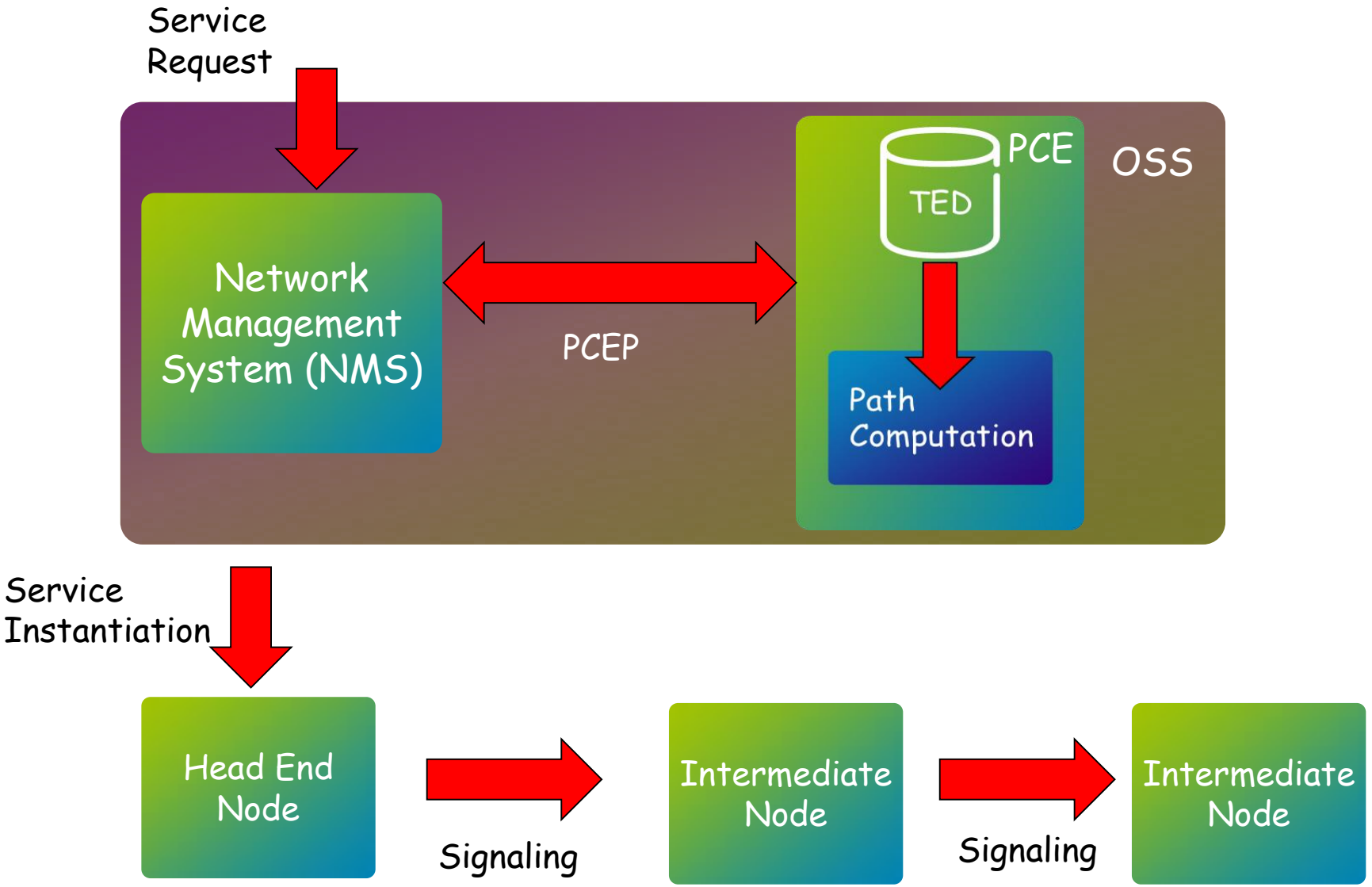
Centralized PCE Architecture



Inter-PCE Communication



Integration with Operations Support System (OSS)



Summary

- Virtual circuit has many attractions for transport networks
  - Development of MPLS made virtual circuits available to IP networks
- MPLS labels packets with a 32 bit label in a label stack
  - Simple data plane: push, pop, swap
  - 20 bit routing label, 3 ToS bits for quality of service
- Packet flows are labelled according to their forwarding equivalence class
  - A FEC is just an aggregated flow
- MPLS control plane differs depending on applications
  - Traffic engineering is a major use of MPLS
  - Operator sets up label switched paths across their network to spread load
- Application of MPLS to optical transport networks and other use cases required centralized path computation
  - Path Computation Engine provides server-computed path information to routers and optical switches

Acknowledgements

- Adrian Farrell, Old Dog Consulting

EXTRA SLIDES

Multiple PCE