

# Potential Roles of the sub-IP technologies in Future large-scale Network Architecture

**NetXworking Workshop, Berlin  
April 19-20, 2007**

Dimitri Papadimitriou  
<dpapadimitriou@psg.com>  
<dimitri.papadimitriou@alcatel-lucent.be>

# Outline

- Landscape
  - Advances
  - Control plane
  - Packet networks
- Some lessons
- Paradigm/concepts ?
- Approach
- Conclusion

# Landscape (1)

Several perspectives in sub-IP techno's

- Optical
  - Circuit-switching in the optical domain → OWS
  - ATM → OBS
  - MPLS → OLSP
  - Others: OPS, TDM → OTDM, etc.
- Control plane paradigms
- Packet
  - MPLS → MPLS-TE → ?
  - Ethernet from LAN → MAN → ?

# Landscape (2)

- ATM → Optical Burst Switching (OBS)
  - Design objectives
    - statistical multiplexing in optical domain
    - buffer-less / buffer-limited network
    - transparent data (optical signal with high rate cut through)
  - Data driven, connectionless (with temporary one- or two-way reservation)
    - (control) header sent to temporary reserve wavelength resource
    - optical burst (block of packets) sent through the network
    - note: burst in optical domain and header in the electronic domain
  - Advantages
    - adaptation to traffic variation (variable burst size)
    - burstification (relaxes processing time, no re-sequencing)
  - Problems
    - synchronization difficult between header and data burst
    - burst loss
    - contention resolution (note: edge node shaping (bursty traffic profile) decreases contention)

# Landscape (3)

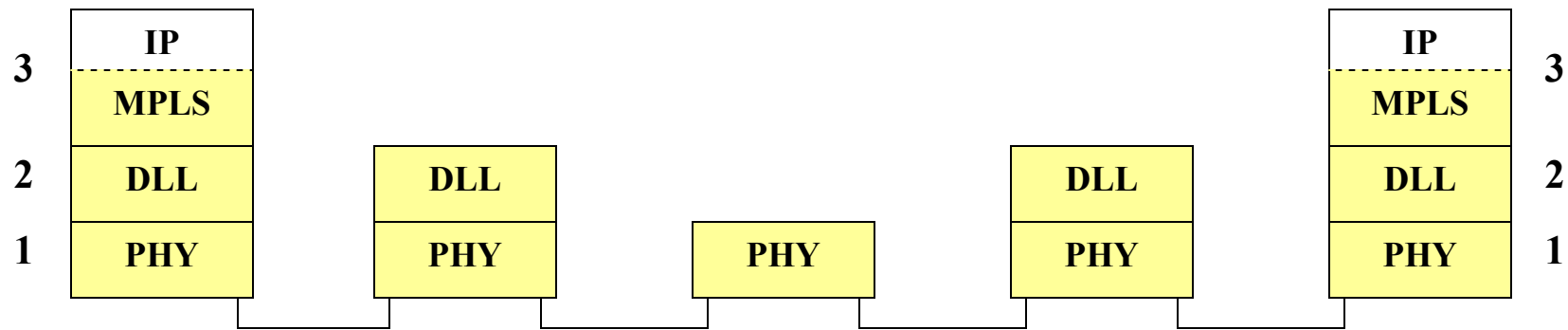
- Optical Packet Switching (OPS)
  - Control driven, Connection-oriented
    - signaling for connection setup (one-way reservation)
    - optical packets (fixed duration/length) created at ingress and sent in the network after the acknowledgement of the signaling process
    - control plane in electronic domain and data path in optical domain
  - Advantages
    - less contention (burstiness reduced)
    - limit packet jitter
    - fixed duration packet => control independent from the line bit rate (data plane)
  - Problems
    - lack of optical buffering
    - filling ratio at the payload level can be difficult to achieve (timeout, dummy packets)
    - need for fast packet switching and header processing
    - fill gaps between useful packets by dummy packets (to ease packet rhythm locking and monitoring)
    - large packet jitter => reassembly mechanisms at network egress difficult to handle

# Landscape (4)

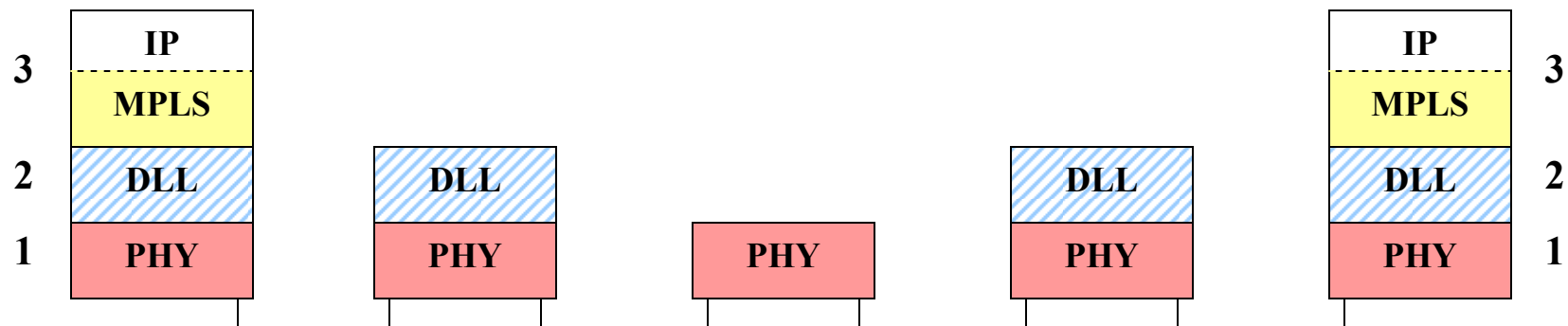
- Optical Wavelength Switching (OWS)
  - Bandwidth/resource allocation
    - fixed, limited, and discrete bandwidth granularity (x10 Gbps, STM-x, OC-x)
    - no flexible adaptation (packet over circuit)
    - permanent or semi-permanent
    - circuit-oriented, hard-state “cross-connection”
  - Associated to a control-driven control-plane paradigm
    - edge-to-edge (SPVC mode) or end-to-end (SVC mode)
    - unified or overlay control plane inter-connection model
  - Deployment limitations
    - Cost of interfaces on routers/operating end-to-end SVC without real gain compared to more flexible packet processing at routers ... raise of IP over ETH (Gb)
    - limited aggregation/multiplexing capability
- At the end its main advantages are
  - Robustness/Stability (if not all-optical)

# Control Plane Paradigms (1)

- Unified IP control plane



- Overlay control plane - major issue: unknown adjacency problem => full mesh (n square), no trigger



# Control Plane Paradigms (2)

- Idea: “control plane” = preliminary form of a "packetisation" of the sub-IP environment
  - Objective: make sub-IP “infrastructure” responsive and adaptive to traffic demands by transposing the control plane of IP/MPLS devices to sub-IP nodes
- Limitations (to both models)
  - complexity of protocol implementation (side effect of MPLS-TE protocol suite generalization)
  - impossibility for pure IP device to setup data path without support of “switching capable interface”
  - limitation induced by control-driven paradigm itself

# Packet

- MPLS: multiple lives
  - Step 1: forwarding paradigm (following L3 paradigm)
  - Step 2: traffic engineering
  - Step 3: “service” delivery (PW: L2 traffic adaptation/multiplexing, L3VPN, etc.)
- Evolution
  - From 90’s: IP enhanced/complemented with X (with X = MPLS)
  - Questions:
    - X = Ethernet ?
    - But Ethernet = LAN technology => X = Ethernet’

# Ethernet IEEE 802.1 shortcomings

## 1. Lack of Traffic Engineering capabilities

multiple VLAN per spanning tree (MSTP, 802.1s) instance leads to an inefficient allocation of resources, particularly bandwidth

=> source/constraint-based routing

## 2. Slow convergence

spanning tree protocol (STP, 802.1d) rapid spanning tree protocol (RSTP, 802.1w) subject to tens of second lasting loops in several cases (count-to-infinity) e.g. root bridge failure

=> fast re-routing

## 3. Lack of fast recovery capabilities

during convergence, network capacity drops significantly as the Ethernet bridges fall back to flooding (MAC learning)

hence, a local event (e.g. a link failure) has global impact

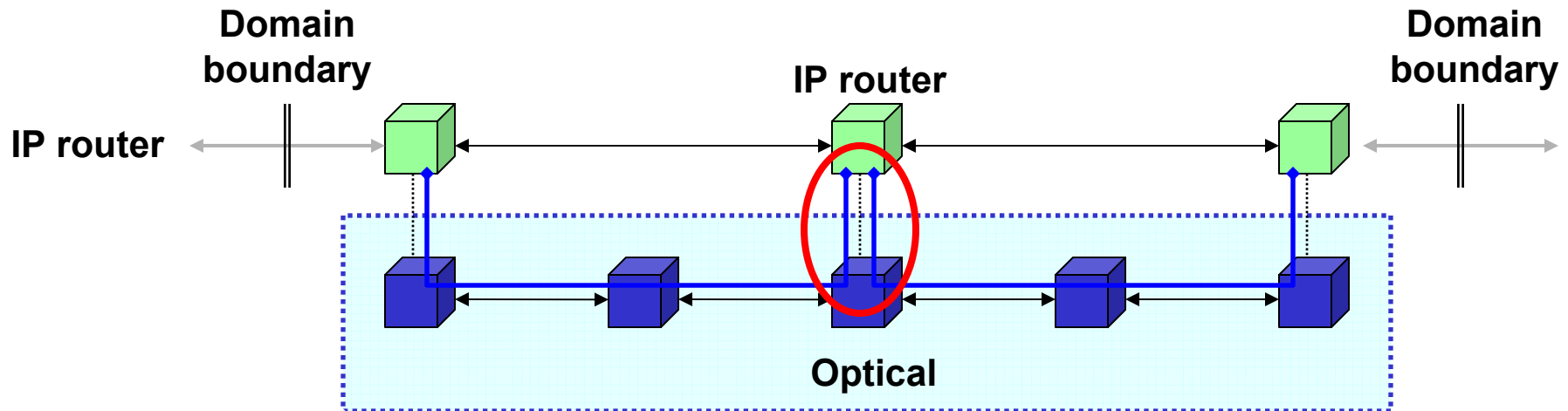
=> forwarding independent of MAC learning

# Packet

- Ethernet evolves as (intra-domain) aggregation technology for core networks (... by better adapting Ethernet - as MPLS is adapted to IP)
- Implications
  - Ethernet control plane:
    - From distance vector routing protocol (spanning tree protocol) to link state routing protocol
    - As IP routing evolved from RIP (distance vector) to OSPF (link state)
  - Ethernet forwarding plane:
    - Ethernet switching without specific mechanisms suitable/dedicated for LAN (campus, enterprise, etc.) environments
    - Mechanisms fitting specific needs of aggregation

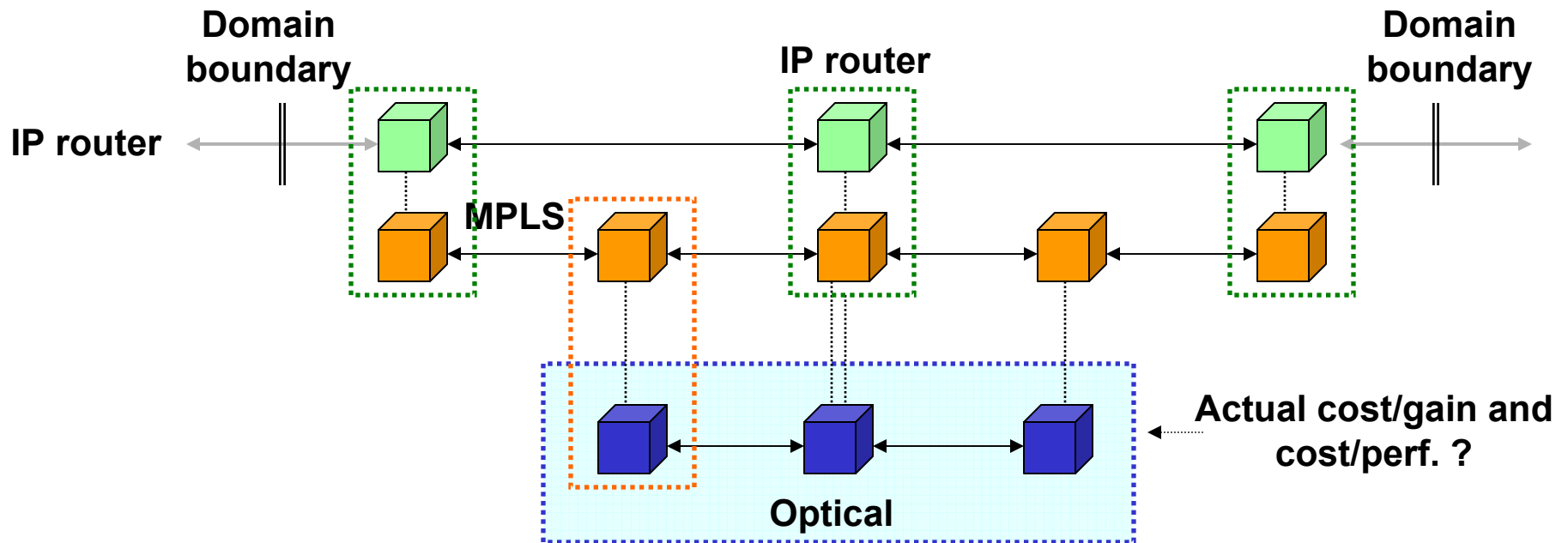
# A Priori

- End 90's expectations
  - Core routers interconnected by flexible optical/DWDM infrastructure
  - Easy/fast provisioning (dump router-to-router optical pipes)
  - Drawbacks (at that time) cost of optical interfaces on IP routers compensated by “revenues”
- Result: OWS not fulfilling promises (cost, not adapted to granularities of incoming traffic demands, etc.)



# A Posteriori

- Result in mid'00s: aggregation/multiplexing on routers via MPLS-TE (tunnels)
  - IP Router: networking (single peering point)
  - “LSR”: adaptation, multiplexing and aggregation
  - **Optical equipment: (internal) connectivity only ?**

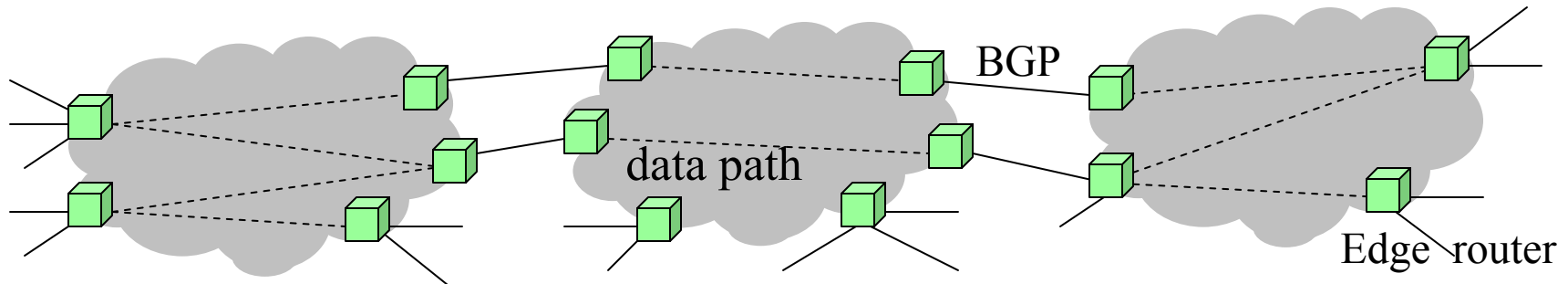


# Problem

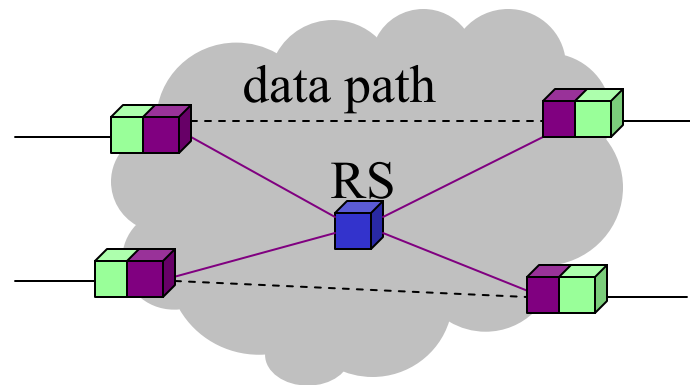
- In core / large-scale architectures for packet networks
  - if control plane / traffic is aggregated, then it is aggregated on the same platform that aggregates data plane / traffic
  - imposes set of two-dimensional requirements on that platform
  - => platform must scale in terms of both bandwidth and throughput
  - + protocol message processing
- Implication
  - core platform must include state-of-the-art capabilities for both dimensions
  - cost and complexity of platform
- Problem: how to reduce the two-dimensional nature of the core scaling problem

# Approach (1)

- Inter-domain routing



- Intra-domain routing



RS = Route Server acting as IP routing information (distributed/centralized) proxy server

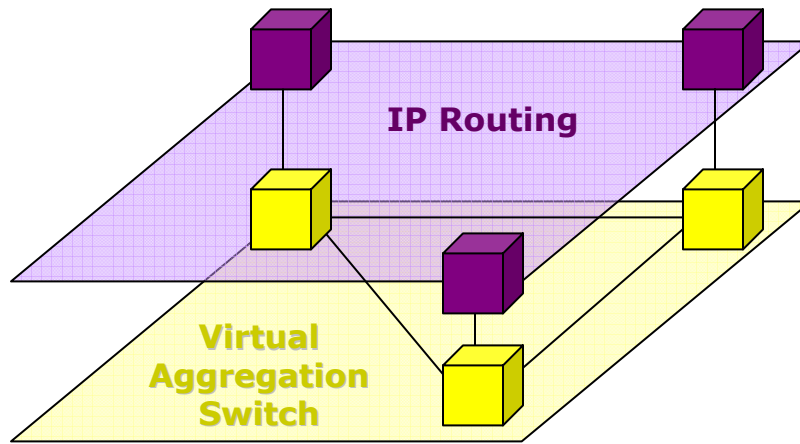
# Approach (2)

- Retain benefits of both IP traffic engineering and original control plane “separation” of overlay network (control plane separation between IP and transit network)
- Advantages
  - Increased stability and reduced transit control plane scaling requirements
  - Ability to instantiate multiple IP networks relying on the same transit network, without fate-sharing their control planes
  - TE of router-to-router flow in the network, whilst providing the “advantages” of an IP network
  - Within core:
    - IP signaling and control plane protocols (new paradigm possible)
    - commoditized interfaces to the IP routers connected to it

# Concept

- Decouple control from data plane aggregation functions
- Advantages
  - Technical complexity associated to each aggregation problem can be addressed separately
  - Each aggregation problem can be addressed with a specific, rather than generalized platform (possible cost reduction)
  - Differences in expansion rates in logical and physical space are no longer dependent
    - As traffic is increasing ~ 100% per year, and routing table growth is closer ~ 20%, this approach would not require upgrading both the physical and logical scaling platforms at the same time, as they are no longer linked

# Overall Architecture



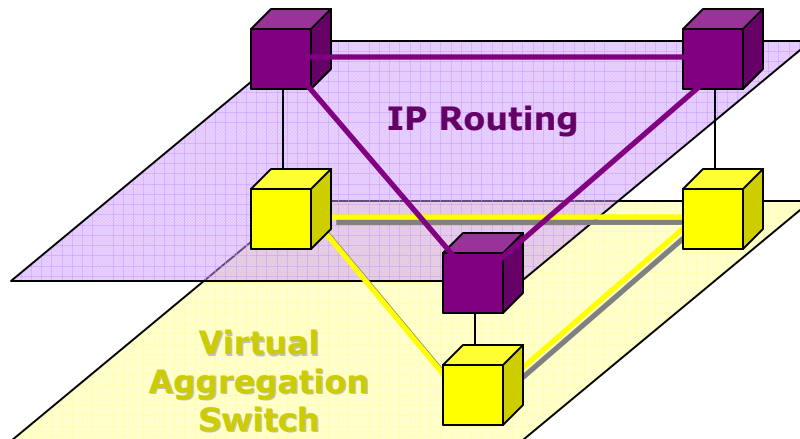
IP Router



Component Switch



Data Plane



IP Router

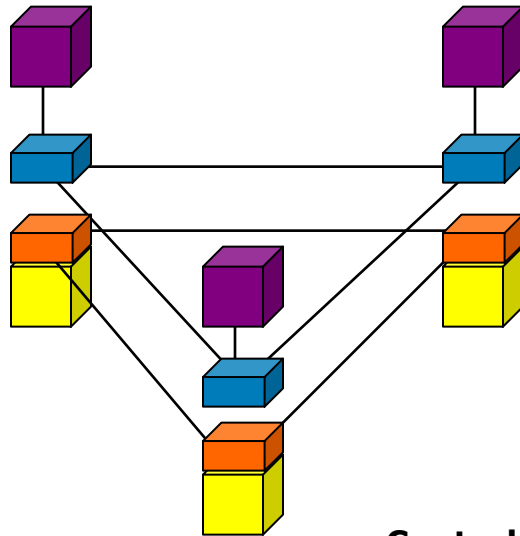


Component Switch



Data Plane (Link) Adjacencies

# Control Plane and Adjacencies



IP Router



Component Switch



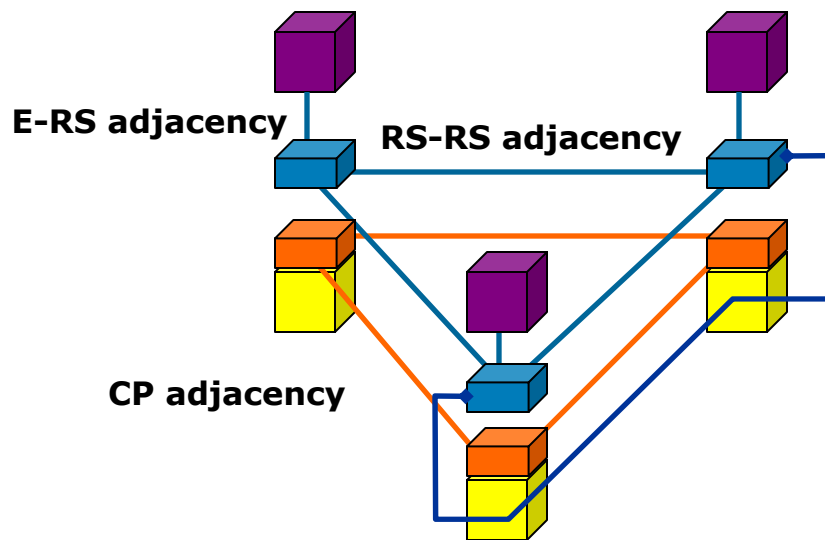
Router Server



Controller



Control Plane

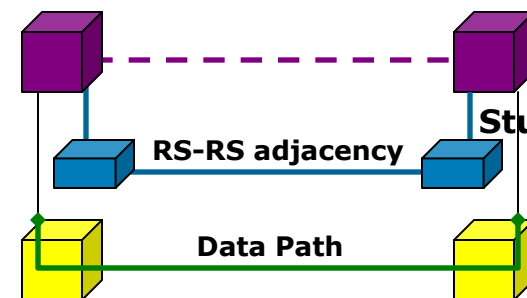
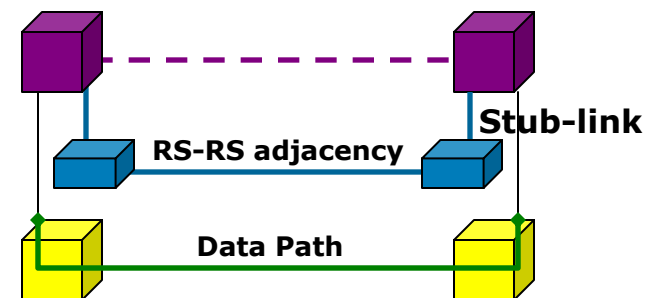


E-RS adjacency

RS-RS adjacency

CP adjacency

Control Plane Adjacencies

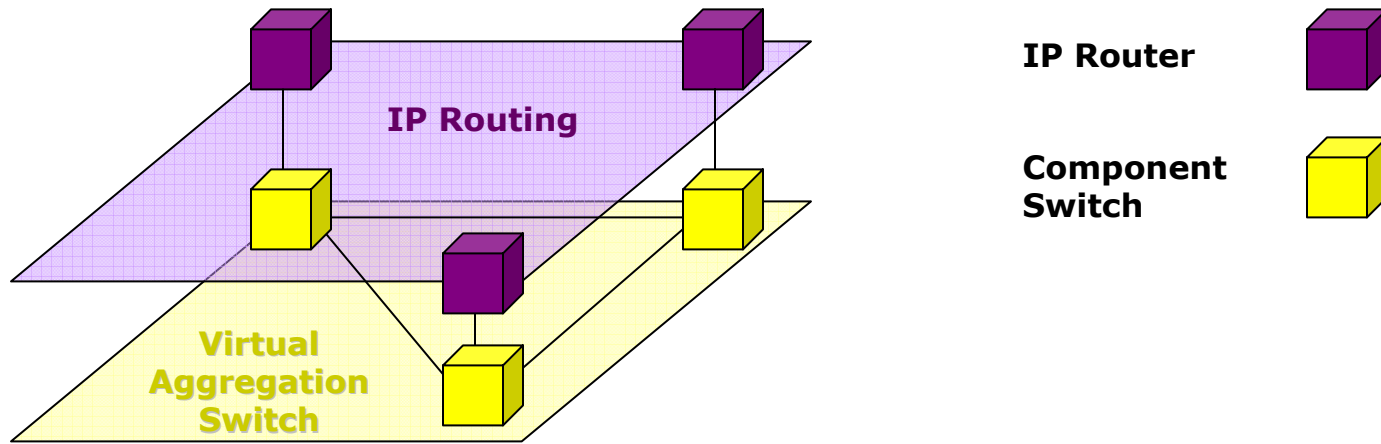


Data Path

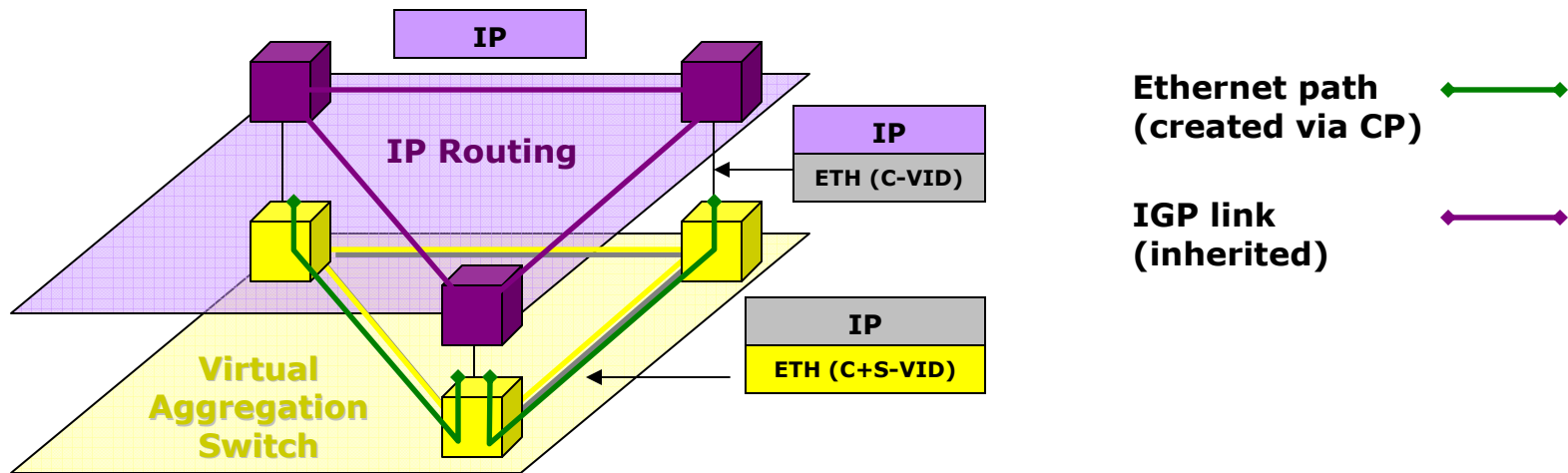
RS-RS adjacency

Stub-link

# Data Plane and Adjacencies



Data Plane



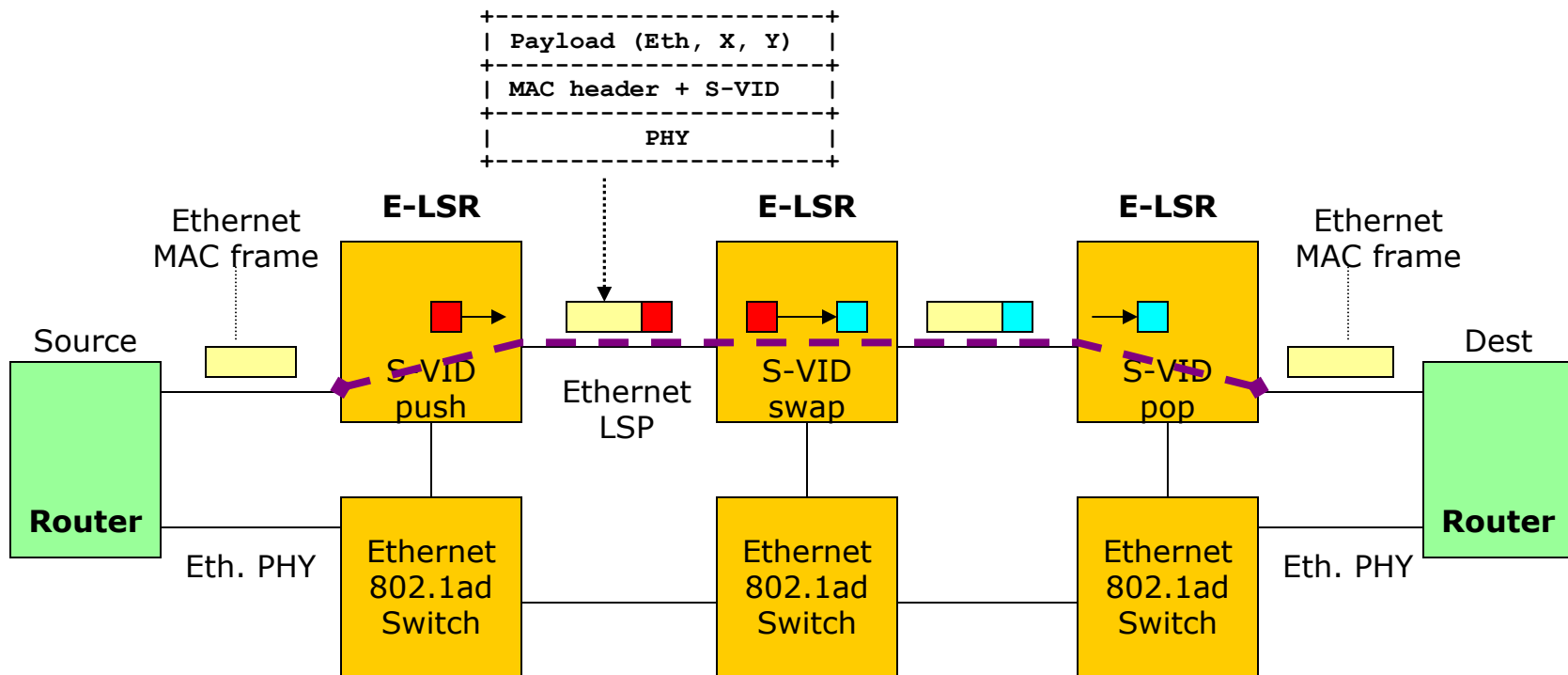
Data Plane Adjacencies

# Several issues

1. RS acting as IP routing information (distributed) proxy server => data vs control path separation
2. Address resolution for data path (between IP routers)
3. IP multicast traffic (IP mcast routing as overlay to IP unicast routing)
4. Model: pipe vs hose (→ adaptive edge connectivity)

# Ethernet Aggregation: VLAN-label Switching

- Ethernet LER (E-LER) function: take an incoming Ethernet MAC frame, add or remove the label (encoded in the TAG field)
- Ethernet LSR (E-LSR): take incoming labelled Ethernet MAC frame and perform label swap (VID in  $\rightarrow$  VID out)  $\Rightarrow$  forwarding independent of dest. MAC address
- Ethernet: point-to-point and point-to-multipoint data paths



# Several Issues

- **Data Plane**
  - Ethernet Label space and scalability (→ Label/LSP merging)
  - Ethernet QoS mechanisms (DSCP to Ethernet PCP mapping → DCP)
  - Ethernet multicast traffic (connectivity and mapping)
- **Control Plane**
  - Unified traffic engineering (TE) much simpler than existing implementations (including fast re-routing)
  - Adaptive TE including Bandwidth Constraint Models (BCM)
  - Lightweight measurement/monitoring capabilities including performance

# Conclusion

- IP-Optical paradigm => IP-Ethernet paradigm
- To reduce the two-dimensional nature of the core scaling problem => decouple control from data plane aggregation
- Core routing without core router: approach applicable to larger scale IP networks
  - Ethernet as (intra-domain) aggregation technology