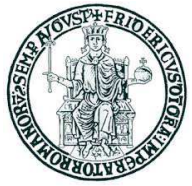
# Cloud e Datacenter Networking

Università degli Studi di Napoli Federico II
Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione DIETI
Laurea Magistrale in Ingegneria Informatica

Prof. Roberto Canonico

## Inter datacenter networking



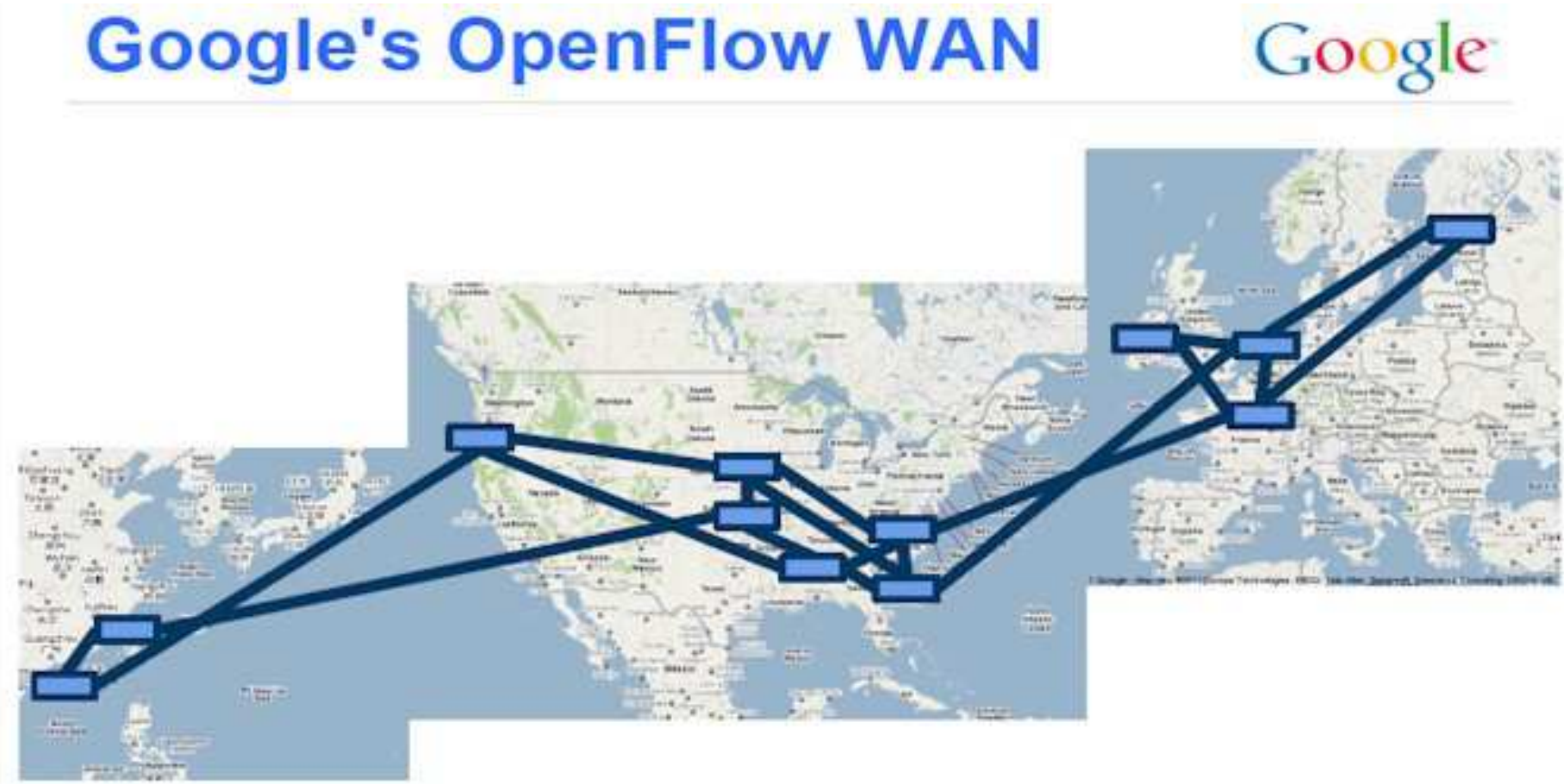
# Lesson outline

- Connection of geographicaly dispersed datacenters
- Characteristics of inter-datacenter traffic
- Role of MPLS and Metro Ethernet in datacenter connectivity

# WAN connectivity for datacenters

- Private companies build multiple datacenters for several reasons
  - Data availability
  - Workload partitioning / Load balancing
  - Legal reasons (sensitive data constrained to remain within national boundaries)
  - Disaster response (disaster recovery)
- These private datacenters need to be efficiently interconnected so that data and applications may be migrated from one datacenter to another
- Public cloud providers also build multiple datacenters
  - for the above reasons
  - plus the need to keep their servers as close as possible to end users
    - Low latency is a QoS requirement for several applications
- Two possible approaches for geographic inter-datacenter connectivity:
  - Rely on the public Internet
  - Build a dedicated WAN infrastructure

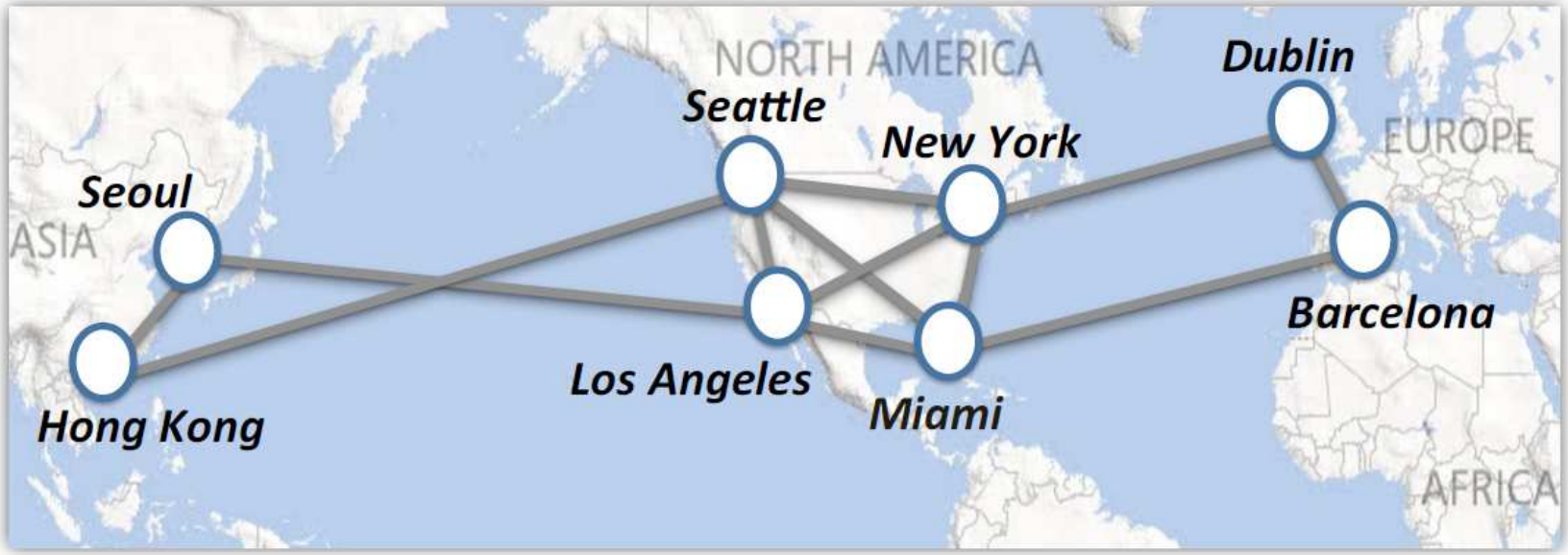
# An example of a private WAN: Google




- Google's B4 private Software-Defined WAN
  - Picture shows B4 topology as of 2011

**Sushant Jain, Alok Kumar, Subhasree Mandal, Joon Ong, Leon Poutievski, Arjun Singh, Subbaiah Venkata, Jim Wanderer, Junlan Zhou, Min Zhu, Jon Zolla, Urs Hölzle, Stephen Stuart, and Amin Vahdat.**
***B4: Experience with a Globally-Deployed Software Defined WAN*. In Proceedings of ACM SIGCOMM 2013.**

# An example of a private WAN: Microsoft

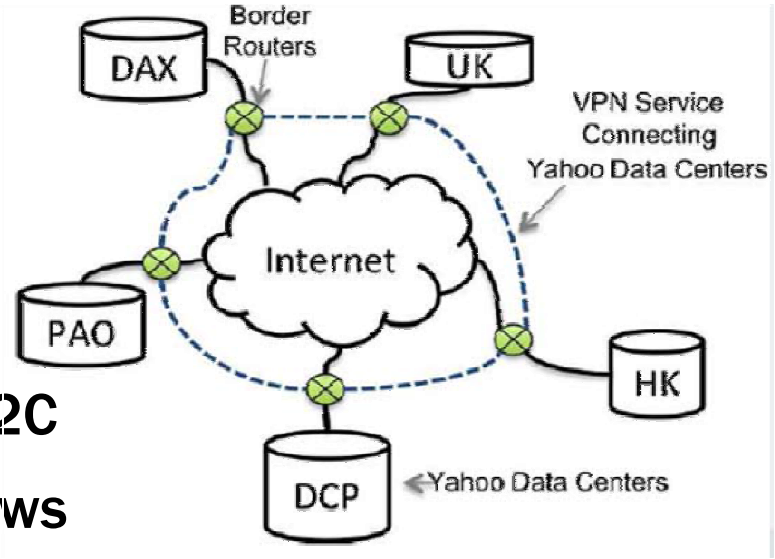


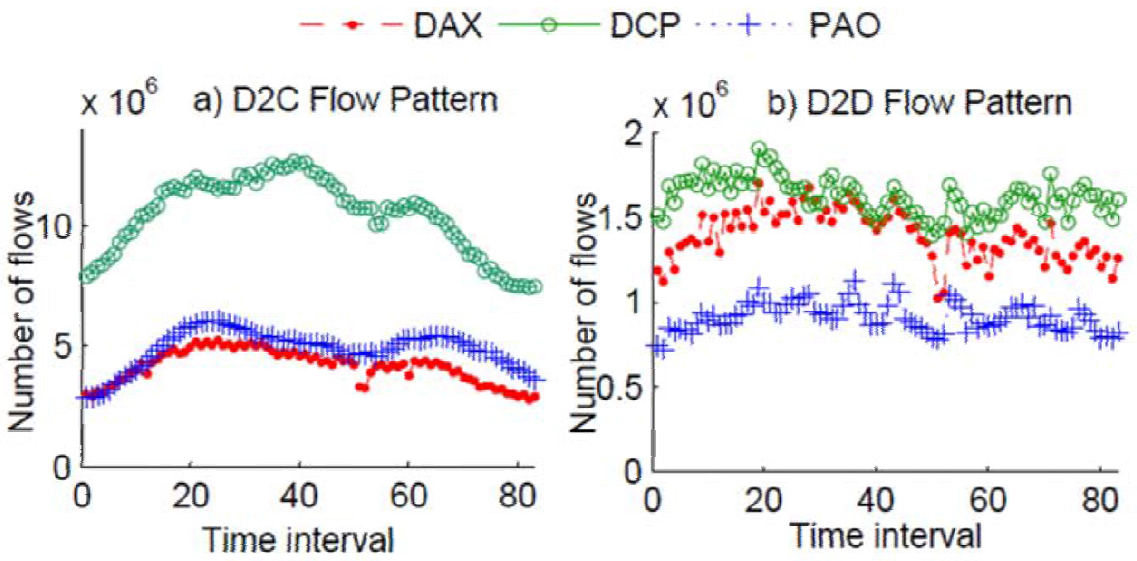

- Microsoft inter-DC WAN
  - average utilization over time of busy links is only 30-50%

**Chi-Yao Hong, Srikanth Kandula, Ratul Mahajan, Ming Zhang, Vijay Gill, Mohan Nanduri, and Roger Wattenhofer. *Achieving high utilization with software-driven WAN*. Proceedings of ACM SIGCOMM 2013.**

# Inter-datacenter traffic characteristics

- Analysis conducted in 2011 on the WAN links connecting 5 Yahoo datacenters
- D2C means Datacenter to Clients
- D2D means Datacenter to Datacenter
- In terms of number of flows: D2D is 10÷20% of D2C
- But D2D traffic mostly made of long-lived heavy flows
- In terms of traffic volume, D2D traffic has been underestimated
- According to Google, B4 now carries more traffic than the public facing WAN connecting its datacenters with the rest of Internet




Y. Chen, S. Jain, V. K. Adhikari, Z. L. Zhang and K. Xu. *A first look at inter-data center traffic characteristics via Yahoo! datasets.* Proceedings of IEEE INFOCOM 2011.

# Advantages of building a private WAN

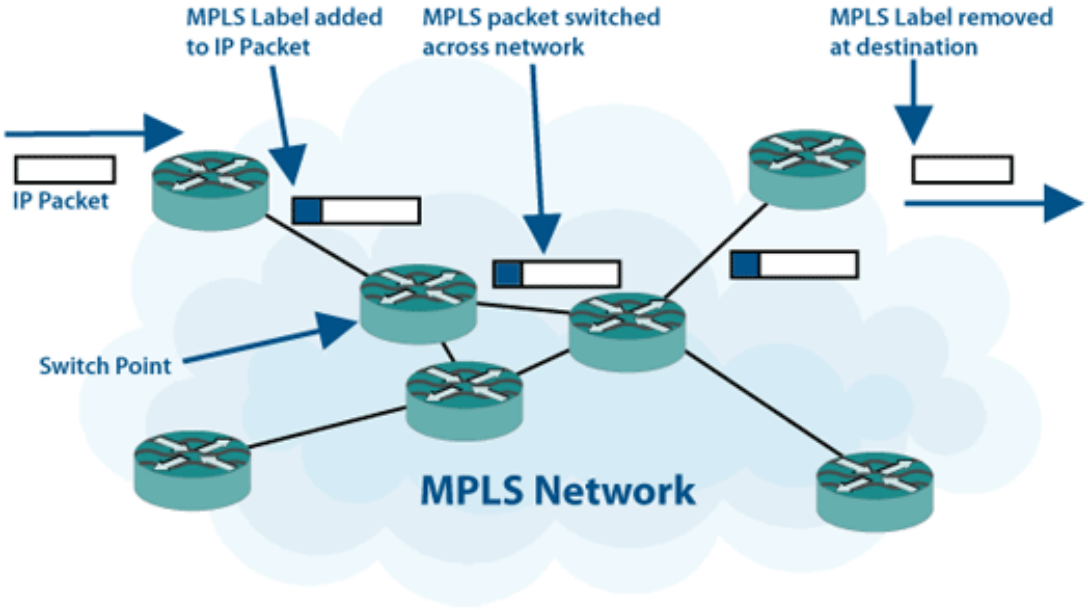
- Ownership of the WAN infrastructure gives the maximum degree of flexibility in designing and managing inter-datacenter connectivity
- A private WAN is usually created by establishing point-to-point links either by
  - Buying dark fibers or
  - Leasing a portion of a fiber link
- from a telecommunications operator
- On the other hand, if one has to rely on intermediate transit ISPs, the infrastructure needs to participate to interdomain routing (BGP)
  - In this case, traffic engineering between datacenters is possible if sites are multi homed, i.e. they are connected to multiple transit ISPs

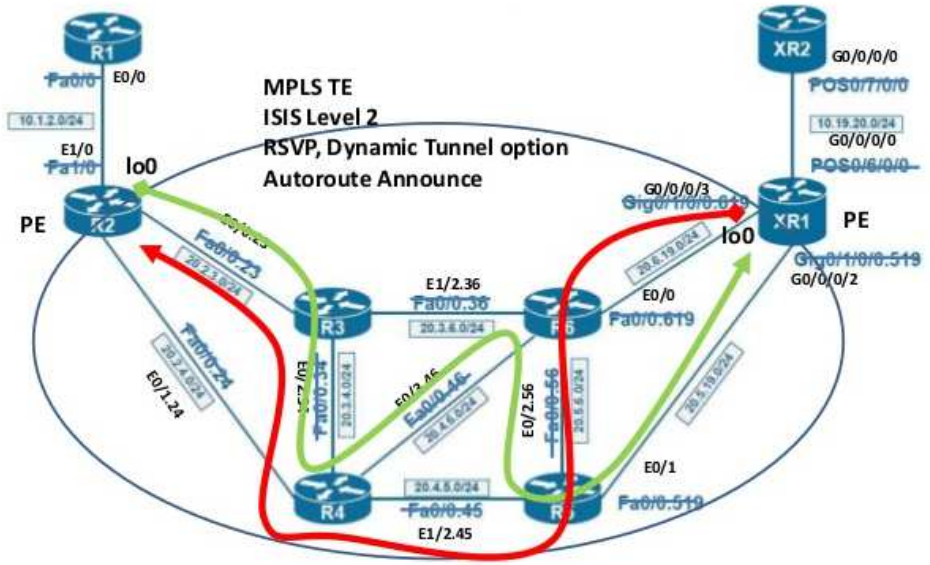
# Inter-datacenter point-to-point link types

- Dark Fiber
  - Dark fiber is a viable method for extending VLANs over data center or campus distances. The maximum attainable distance is a function of the optical characteristics (transmit power and receive sensitivity) of the LED or laser that resides in a Small Form-Factor Pluggable (SFP) or Gigabit Interface Converter (GBIC) transponder, combined with the number of fiber joins, and the attenuation of the fiber.
- Coarse Wavelength Division Multiplexing (CWDM)
  - CWDM offers a simple solution to carry up to eight channels (1 Gbps or 2 Gbps) on the same fiber. These channels can carry Ethernet or fiber channel. CWDM does not offer protected lambdas, but client protection allows rerouting of the traffic on the functioning links when a failure occurs. CWDM lambdas can be added and dropped, allowing the creation of hub-and-spoke, ring, and meshed topologies. The maximum achievable distance is approximately 100 km with a point-to-point physical topology and approximately 40 km with a physical ring topology.
- Dense Wavelength Division Multiplexing (DWDM)
  - DWDM enables up to 32 channels (lambdas). Each of these channels can operate at up to 10 Gbps. DWDM networks can be designed either as multiplexing networks that are similar to CWDM or with a variety of protection schemes. DWDM also offers the possibility to amplify the channels to reach greater distances.

# Traffic engineering with MPLS

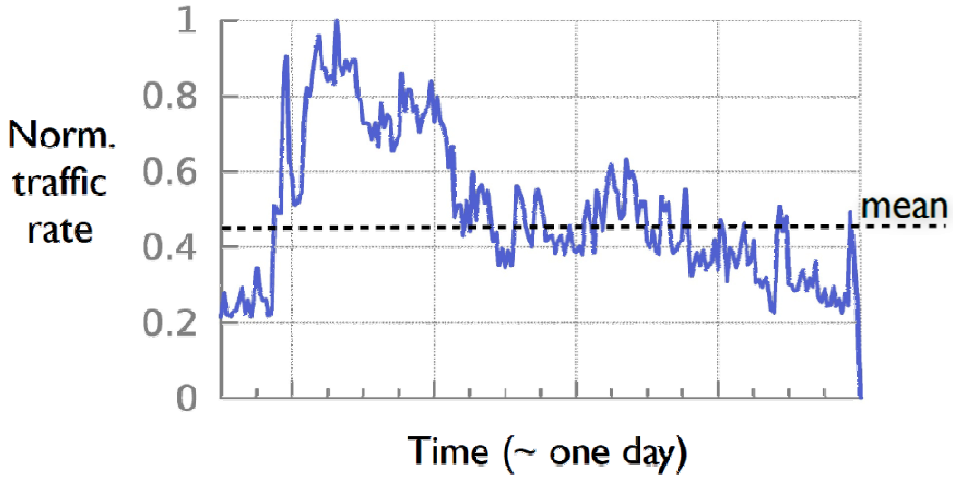
- With MultiProcotol Label Switching (MPLS) it is possible to establish end-to-end paths and control their routes
- An intra-domain link state protocol (such as OSPF or IS-IS) is used to spread topology information across all network nodes
- For traffic engineering, it is also necessary to spread available bandwidth info for each link among nodes
- When a new path is requested to be setup, the shortest path is computed that has enough bandwidth available
- A signalling protocol is used to inform network nodes of the newly established path

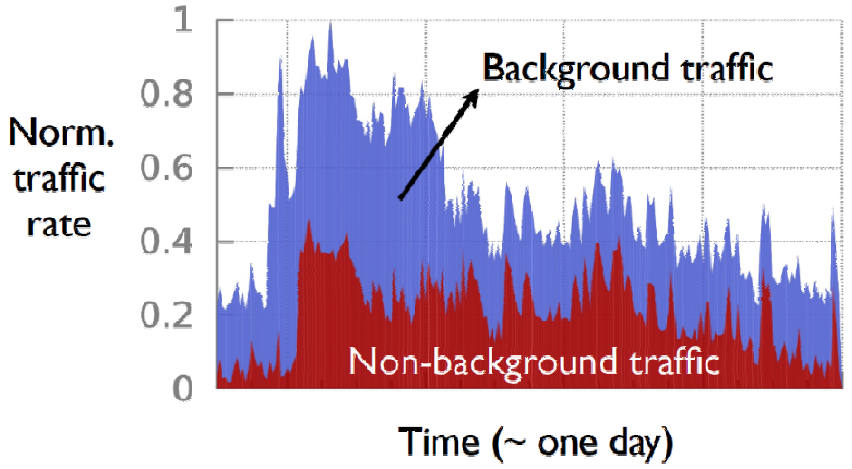


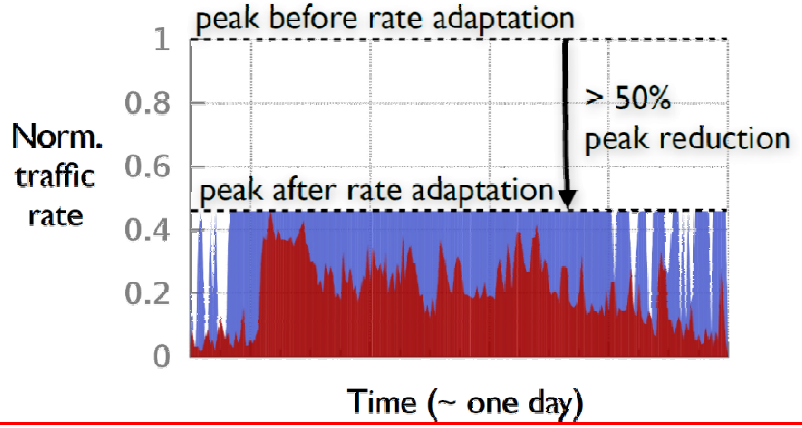
# WAN link utilization

- Since WAN links are dimensione according to traffic peaks, they are usually quite poorly utilized



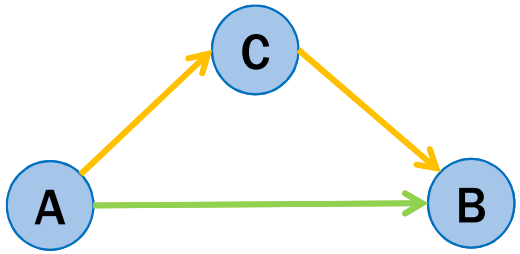




- However, a large amount of traffic is actually background traffic that could be delayed without affecting applications
- Hence links might be provisioned according to the peak of delay-sensitive traffic and background traffic could be transmitted using spare capacity



# Site-to-site delay

- Internet routing often violates the triangle inequality rule




Triangle inequality rule

Delay(A,B) < Delay(A,C) + Delay(C,B)

Triangle inequality rule is violated if:

Delay(A,B) > Delay(A,C) + Delay(C,B)

- In order to connect A to B, it may be convenient to force packets via C
- If Internet routing forces the A to B route, one may build an overlay network that transmits A to B packets in two tunnels: A to C and C to B
  - The C relay node might be deployed in a public cloud provider site

# Metro Ethernet

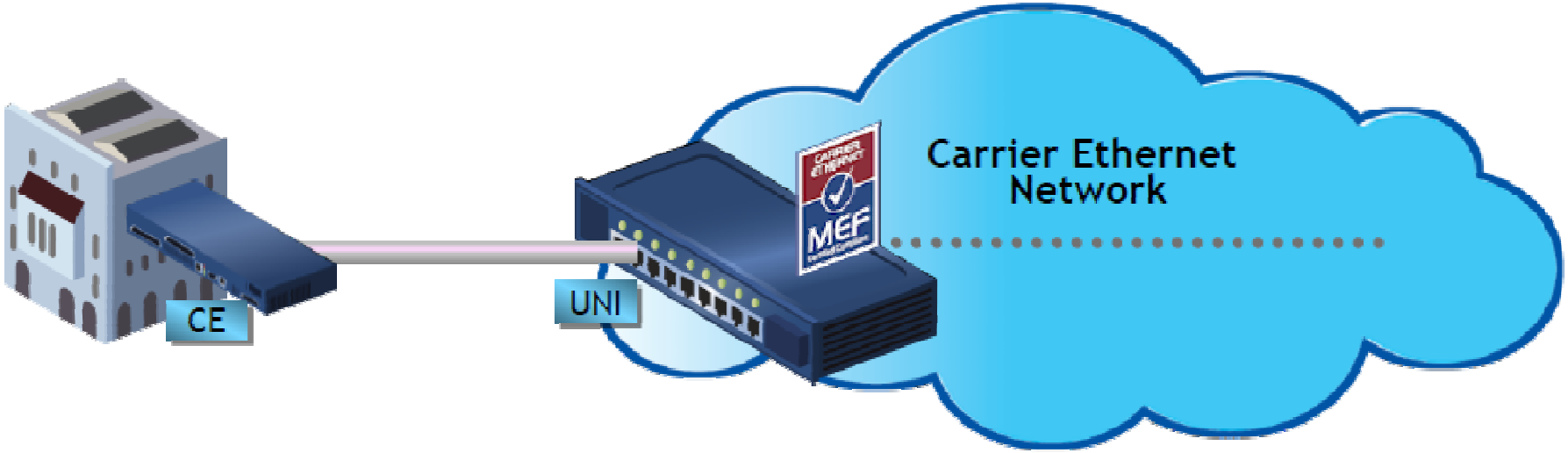
- What is Metro Ethernet ?

  > "... generally defined as the network that bridges or connects geographically separated enterprise LANs while also connecting across the WAN or backbone networks that are generally owned by service providers. The Metro Ethernet Networks provide connectivity services across Metro geography utilising Ethernet as the core protocol and enabling broadband applications"
  >
  > from "Metro Ethernet Networks - A Technical Overview" from the Metro Ethernet Forum

- Why Metro Ethernet ?
  - Benefits both providers and customers in numerous ways ...
  - Packet traffic has now overtaken all other traffic types
  - Need for rapid provisioning
  - Reduced CAPEX/OPEX
  - Increased and flexible bandwidth options
  - Well-known interfaces and technology

# Metro Ethernet UNI and CE

- The User Network Interface (UNI) is the physical interface or port that is the demarcation between the customer and the Service Provider/Carrier
- The UNI is always provided by the Service Provider
- The UNI in a Carrier Ethernet Network is a standard physical Ethernet Interface at operating speeds 10Mbs, 100Mbps, 1Gbps or 10Gbps




- Customer Equipment (CE) attaches to the Metro Ethernet Network at the UNI
  - Using standard Ethernet frames
  - CE can be a Router or bridge/switch - IEEE 802.1 bridge

# Metro Ethernet services

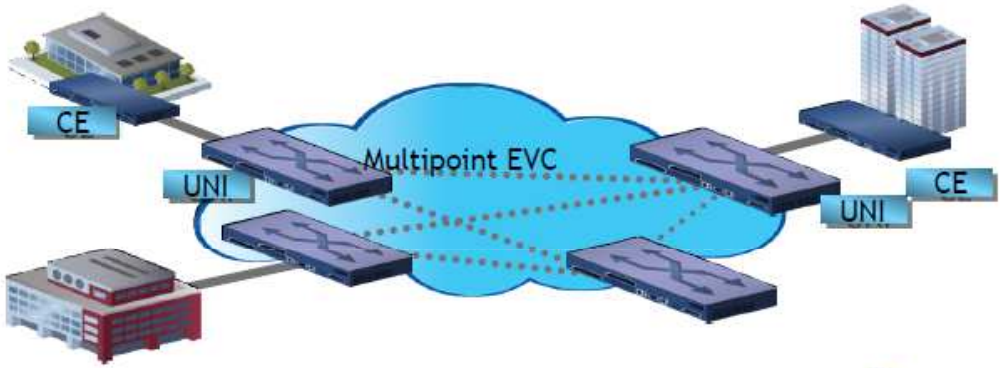
E-LINE


Point to Point

Service Type used to create

- Ethernet Private Lines
- Virtual Private Lines
- Ethernet Internet Access

E-LAN

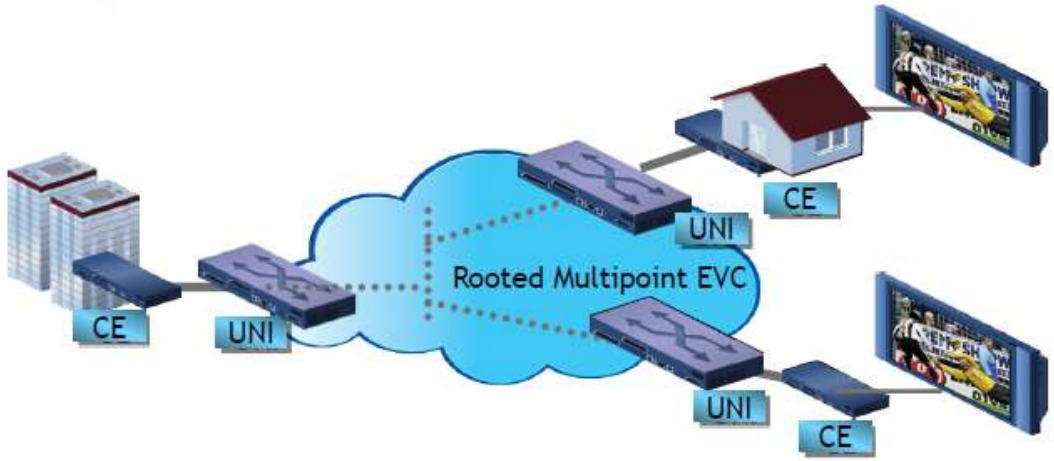



Multi-Point to Multi-Point

Service Type used to create

- Multipoint Layer 2 VPNs
- Transparent LAN Service

E-TREE




Point to Multi-Point

- Efficient use of Service Provider ports
- Foundation for Multicast networks e.g. IPTV

# MPLS-based Extended Layer-2 networks

- MPLS has been viewed as an IP traffic engineering technology
  - Allows a carrier to increase operational efficiency, but service remains the same
- Layer 2 transport is a new application of MPLS
  - MPLS becomes a forwarding infrastructure for a number of services
    - IP services
    - Private Data (e.g. Ethernet)
- MPLS has also been used to create MPLS-based Layer 2 VPNs connecting several sites logically forming one single LAN

# Ethernet over MPLS

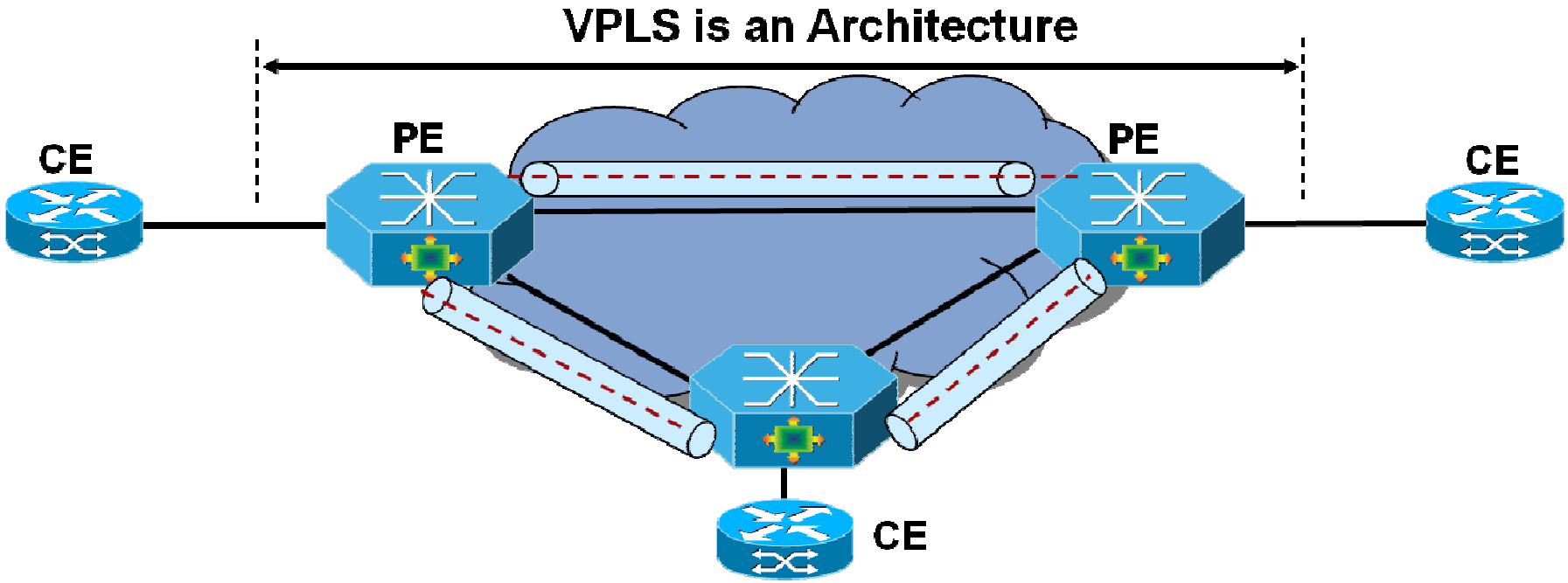
- Ingress device strips the Ethernet preamble and CRC, but transports the entire header
- 802.1q VLAN ID may be overwritten at egress


Ethernet frame

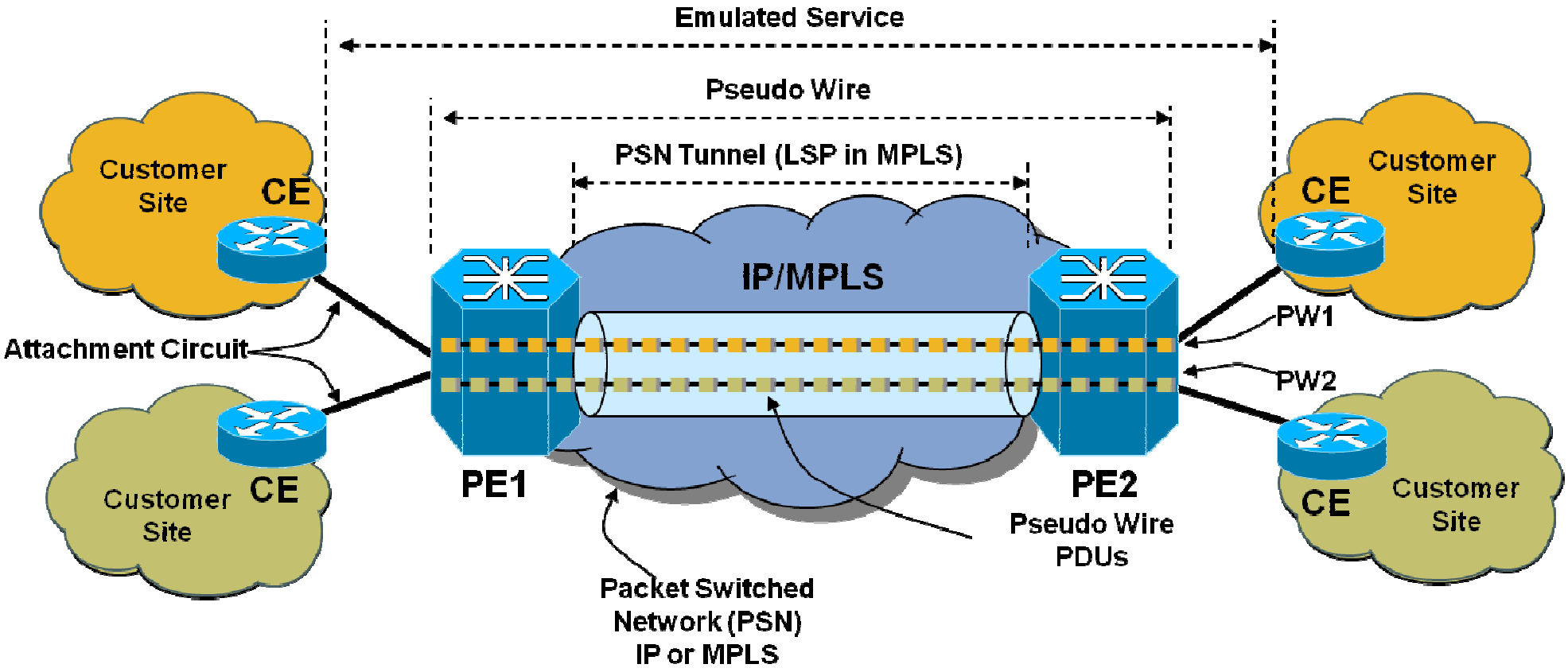

Ethernet over MPLS

# Virtual Private LAN Service (VPLS)

- Virtual Private LAN Service (VPLS) is an architecture that provides multipoint Ethernet LAN services, often referred to as *Transparent LAN Services* (TLS) across geographically dispersed locations using MPLS as transport
- SP emulates an IEEE Ethernet bridge network (virtual)
- A VPLS is based on a full mesh of MPLS Pseudo Wires
  - Data Plane used is same as EoMPLS (point-to-point)



# Pseudo Wire Reference Model (RFC 3916)

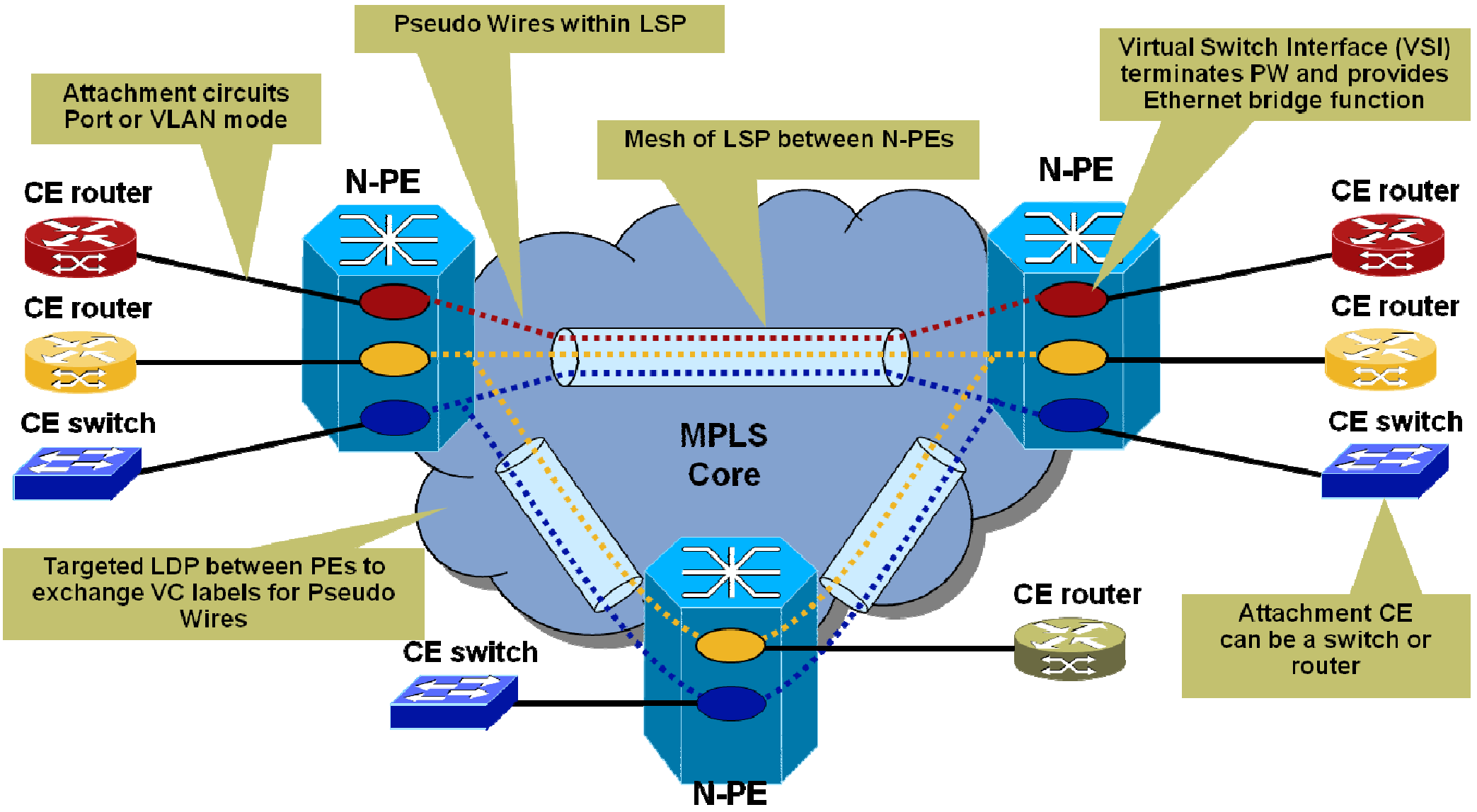



- A Pseudo Wire (PW) is a connection between two provider edge devices connecting two attachment circuits (ACs)
- In an MPLS core a Pseudo Wire uses two MPLS labels
  - Tunnel Label (LSP) identifying remote PE router
  - VC Label identifying Pseudo Wire circuit within tunnel

# Reasons for chosing a VPLS service from a telco

- Customer have full operational control over their routing neighbours
- Privacy of addressing space - they do not have to be shared with the carrier network
- Customer has a choice of using any routing protocol including non IP based (IPX, AppleTalk)
- Customers could use an Ethernet switch instead of a router as the CPE
- A single connection could reach all other edge points emulating an Ethernet LAN (VPLS)

# VPLS components



# Virtual Switch Interface functions

- Flooding / Forwarding
  - MAC table instances per customer (port/vlan) for each PE
  - VFI will participate in learning and forwarding process
  - Associate ports to MAC, flood unknowns to all other ports
- Address Learning / Aging
  - LDP enhanced with additional MAC List TLV (label withdrawal)
  - MAC timers refreshed with incoming frames
- Loop Prevention
  - Create full-mesh of Pseudo Wire VCs (EoMPLS)
  - Unidirectional LSP carries VCs between pair of N-PE Per
  - A VPLS use “split horizon” concepts to prevent loops