Introduction to

# BGP-MPLS Ethernet VPN

Emil Gągała

PLNOG, 16.03.2011

Slides thanks to Rahul Aggarwal

# Agenda

Data Center Interconnect requirements

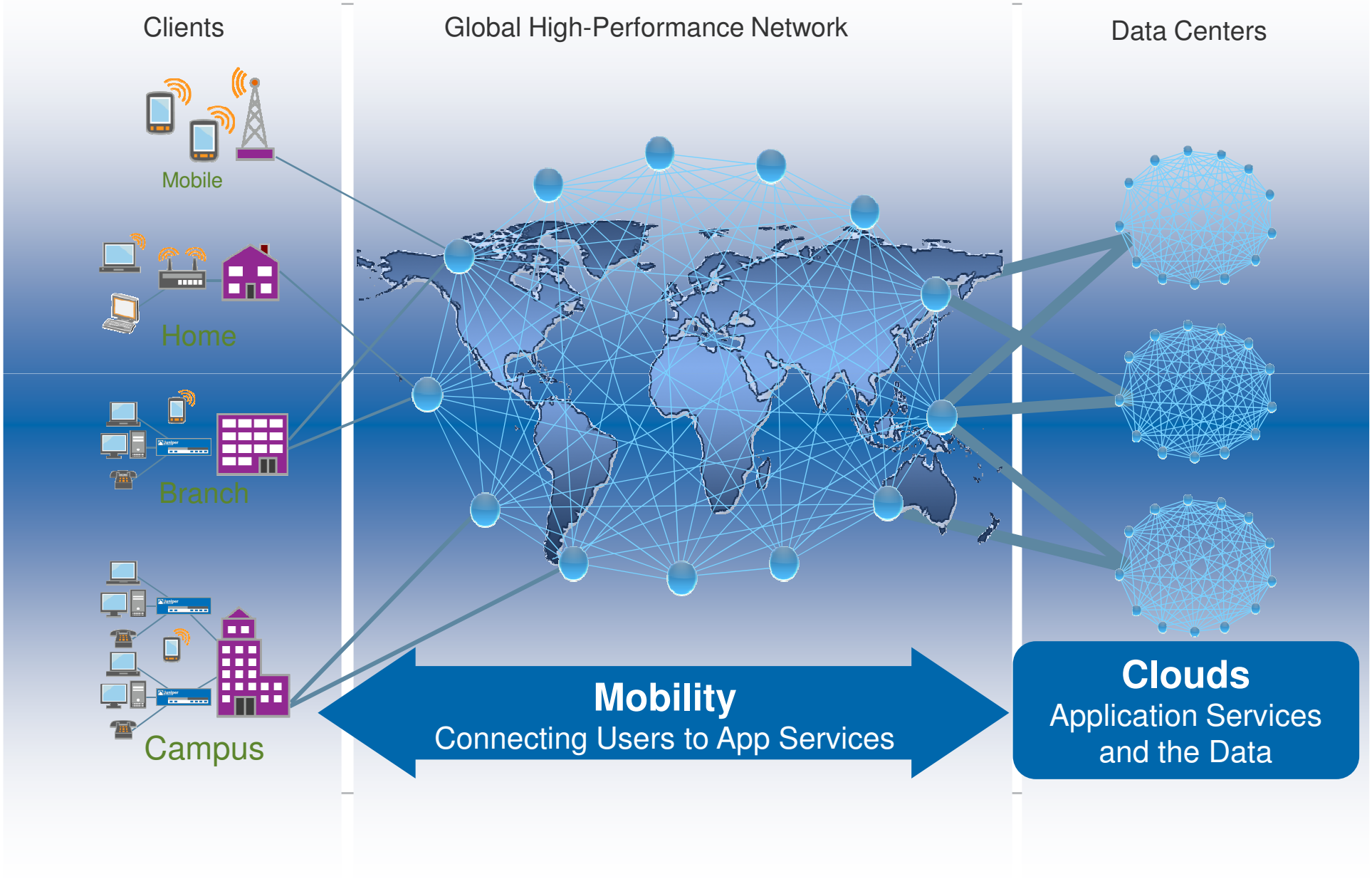
VPLS Status Quo and Areas of Improvements

Ethernet VPN (BGP/MPLS MAC VPN) overview

Ethernet VPNs for layer 2 extension in data centers

Comparison of Ethernet VPN and VPLS service

# THE 2 ROLES OF IT INFRASTRUCTURE



# Data Center Interconnect

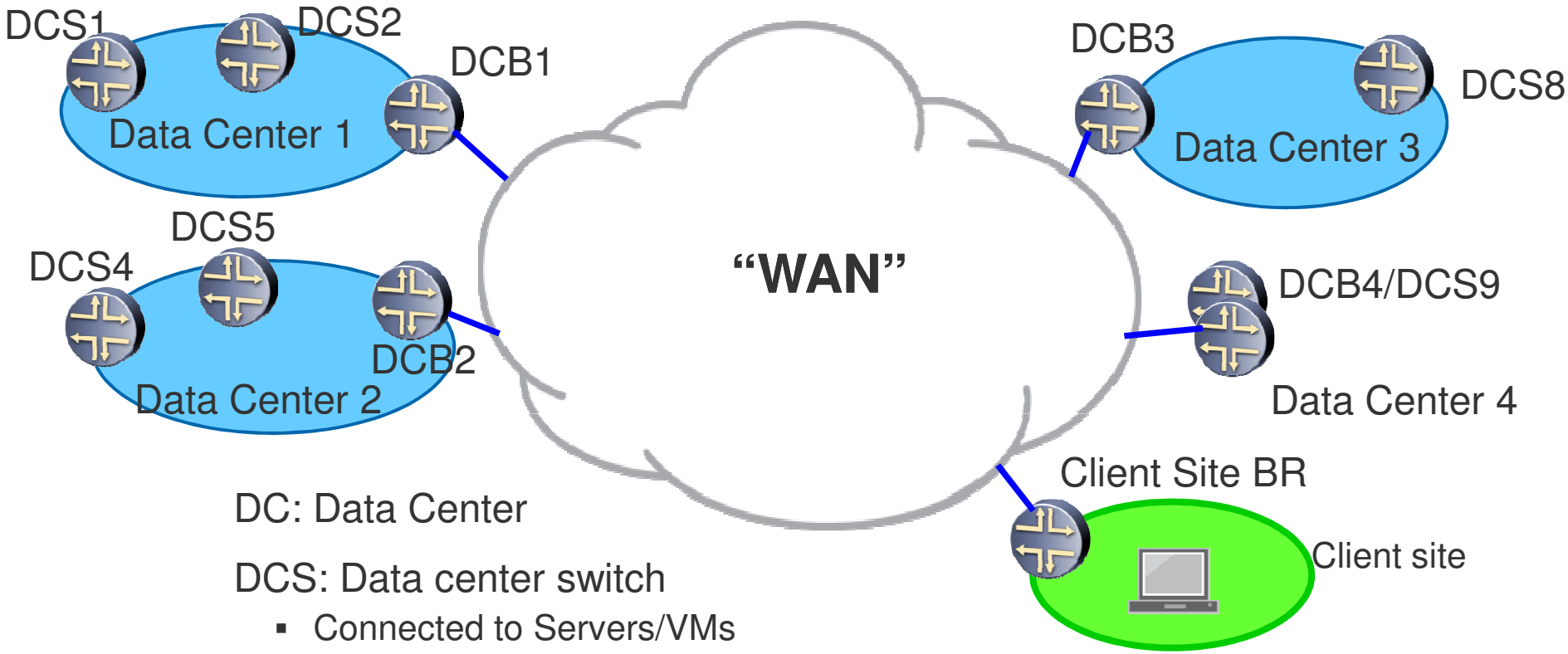
Data Center Interconnect refers to the network or networks that provide both layer 2 and layer 3 interconnect between

- Servers/Virtual Machines (VMs) in data centers and
- Clients to servers/VMs in data centers

The servers/VMs connected by the data center interconnect may be

- In the same data center
- In different data centers in close geographical proximity
  - E.g., the same metro or the same region such
- In different data centers which are not in close geographical proximity
  - E.g., the data centers are reachable via the Wide Area Network (WAN)

# Reference Model and Terminology




DC: Data Center

DCS: Data center switch

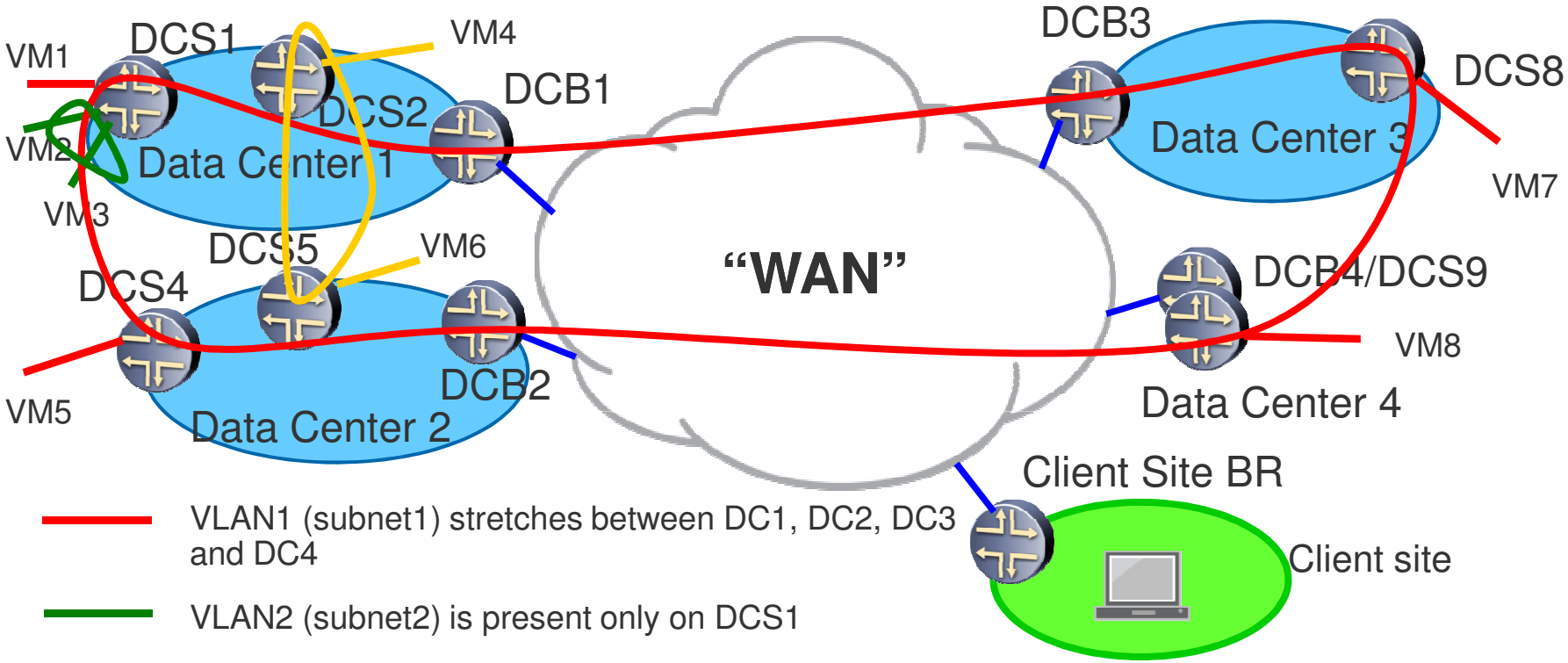
- Connected to Servers/VMs

DCB: Data center border router

- Could be co-located with DCS

"WAN" provides interconnect among DCs, and between DCs and Client Site BR

# Data Center Interconnect Layer 2 Extension and Layer 3 Routing Example




VLAN1 (subnet1) stretches between DC1, DC2, DC3 and DC4

VLAN2 (subnet2) is present only on DCS1

VLAN3 (subnet3) stretches between DC1 and DC2

VLAN stretch is required for cloud computing "resource fungibility", redundancy etc.

Communication between VMs on different VLANs/subnets and between clients and the VMs requires layer 3 routing

# Data Center Interconnect Requirements (1)

Allow VMs to move within the same subnet/VLAN without requiring re-numbering

- E.g. VM migration in the same VLAN for load optimization or failure reasons
- Retain the MAC address and the IP address of a VM for seamless application experience
- The subnet i.e., VLAN may span multiple data centers which may not be in close geographical proximity
  - In practice VM mobility across data centers is limited by bandwidth and distance between data centers

Provide “optimal” forwarding between servers/VMs and between clients and servers/VMs in the presence of seamless VM mobility

# Data Center Interconnect Requirements (2)

Scalability

- Thousands of VLANs supported by data center interconnect
- Hundreds of thousands to Millions of MAC addresses

Minimize or eliminate flooding of unknown unicast traffic.

- While performing layer 2 forwarding within a VLAN

Support of multiple active points of attachment for servers/VMs (e.g., VMs on multi-homed servers)

- Active-active load-balancing among multiple links to and from a server

Minimize latency

- Both in the data center interconnect within a data center and in the data center interconnect across data centers

# Data Center Interconnect Requirements (3 )

Fast service restoration

- Recovery based on local repair from link and node failures in the entire infrastructure including “edge” links and “edge” nodes

Virtualization

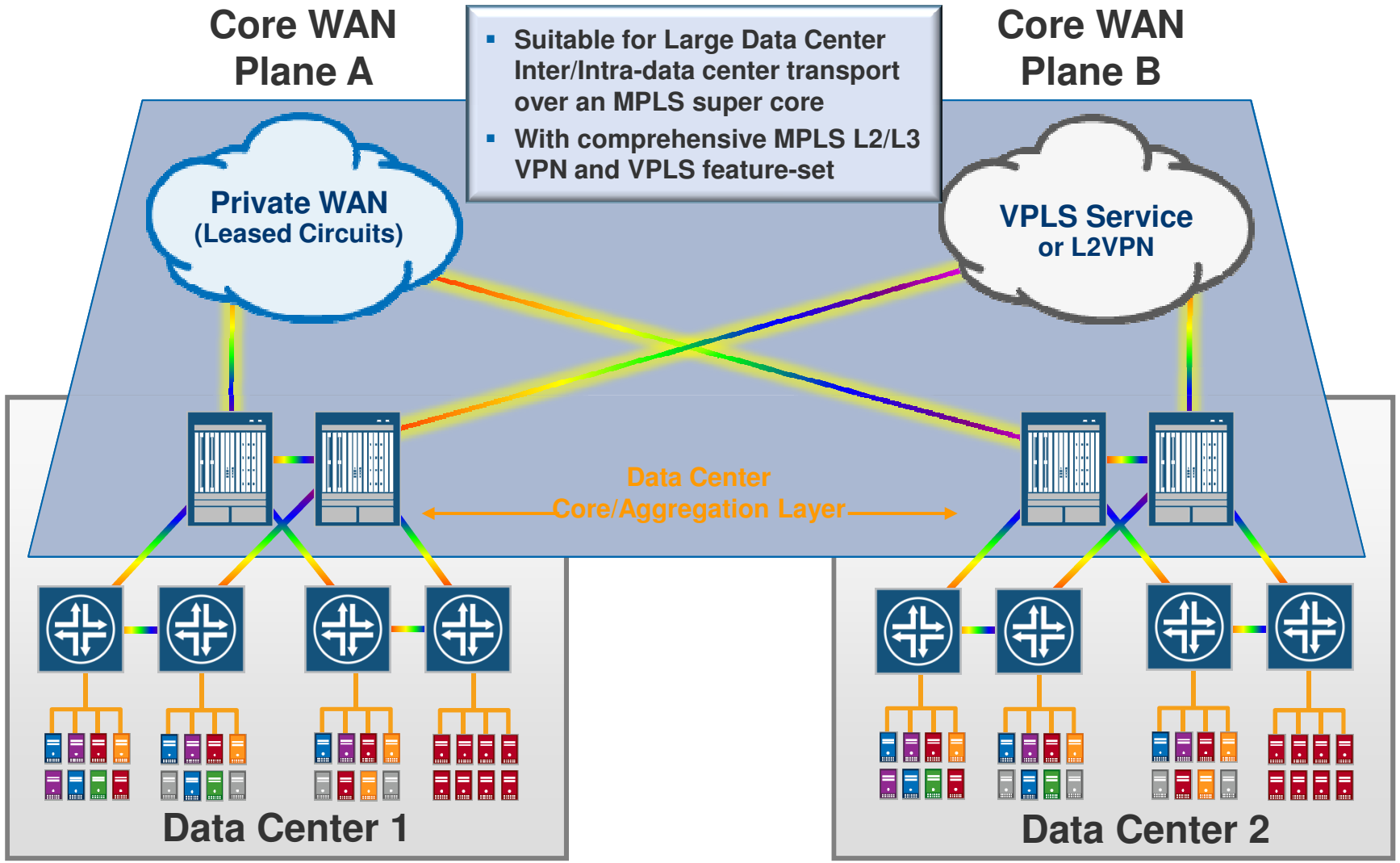
- Segmentation of different business units, hosted customers, services etc.

Simplified provisioning

Common technology for inter-data center and intra-data center interconnect

- At the same time cater to low cost and low feature set hardware requirements on DCS
- In practice inter-data center interconnect characteristics are limited by bandwidth and distance between data centers

# INTER AND INTRA DATA CENTER L2 AND L3 CONNECTIVITY



# Data Center Interconnect: Layer 2 Extension Technology

BGP-VPLS or LDP-VPLS

- The best available option in shipping code
- Does not meet some of the data center interconnect requirements

BGP-MPLS MAC VPN

- New technology to meet all of the data center interconnect requirements
- Draft-raggarwa-sajassi-l2vpn-evpn-01.txt
- See following slides for an overview

# VPLS – CHOICE TO CONNECT SEPARATE L2 DOMAINS

VPLS - Virtual Private LAN Service – connects geographically separate Ethernet broadcast domains over WAN with MPLS pseudo-wires

VPLS allows P2MP Ethernet connections over MPLS networks

One of the two protocols used for peer discovery and signaling between VPLS peers

- BGP VPLS (RFC.4561) and LDP VPLS (RFC.4562)

VPLS provides any to any connectivity – requiring full-mesh connectivity

- BGP solution provides auto-discovery function while LDP requires each peer to be configured

VPLS leverages the benefits of MPLS

- Traffic Engineering
- Multiplexing
- Auto-recovery, rerouting, node/path protection capabilities
- High scale
- Co-existence with other MPLS based services; L3VPN, NG-MVPN etc

# VPLS OVERVIEW

VPLS uses 2-labels in a stack – inner label represents the associated VLAN / IFL

No mechanism to exchange MAC information between PEs exist

Learning by SMAC address

- On physical ports *and* on logical ports (LSPs)
- Learned via packets received from LAN *and* WAN

Forwarding by DMAC address

- Unknown unicast, multicast, broadcast – flooded

Loop prevention by Split Horizon

- CE to CE forwarding needs Spanning Tree

Multi-homing

- Multiple exit points allowed
- But no active-active connection ports are allowed

# IMPROVING ON VPLS

Areas of opportunities for VPLS:

- Active-active multi homing, load balancing
- Relies on flooding to propagate MAC reach-ability information
- Host mobility support
- Ability to make policy based distribution, forwarding decisions
- Fast convergence from edge failures using local repair
- Ability to multiplex VLANs on a single VPLS instance
- Necessity to perform a MPLS lookup AND a MAC lookup on an egress PE

Juniper's current plans:

- MAC-VPN protocol to be implemented improve on VPLS
- A natural upgrade from VPLS - leveraging many of the same elements; BGP, MPLS, RSVP/LDP, P2MP LSPs

BGP MAC-VPN currently an IETF draft:

- http://tools.ietf.org/html/draft-raggarwa-mac-vpn-00
- With non-Juniper co-authors – pushed for industry-wide standardization

# BGP-MPLS MAC VPNs for Data Center Interconnect

BGP-MPLS based technology, one application of which is data center interconnect between data center switches for ***intra-VLAN forwarding i.e., layer 2 extension***

Why?

- Not all data center interconnect layer 2 extension requirements are satisfied by existing MPLS technology such as VPLS
  - E.g., minimizing flooding, active-active points of attachment, fast edge protection, scale, etc.

How?

- Reuses several building blocks from existing BGP-MPLS technologies
- Requires extensions to existing BGP-MPLS technologies…

# BGP MAC-VPN OVERVIEW

Shares many of the characteristics with BGP VPLS and provides following enhancements in addition

Discovery and NLRI exchange by BGP

- MAC address and MPLS label exchanged between MAC-VPN peers

Learning by SMAC address *and* Control Plane

- SMAC: On physical ports and on logical ports (LSPs)
- SMAC: Learned via packets received from LAN and WAN
- CP: Learning from the MAC-VPN peer
- CP: Better convergence times as VMs move from one DC to the other

Learning control

- Ability to apply policies, restrictions
- Ability to provide virtual groups of MAC addresses

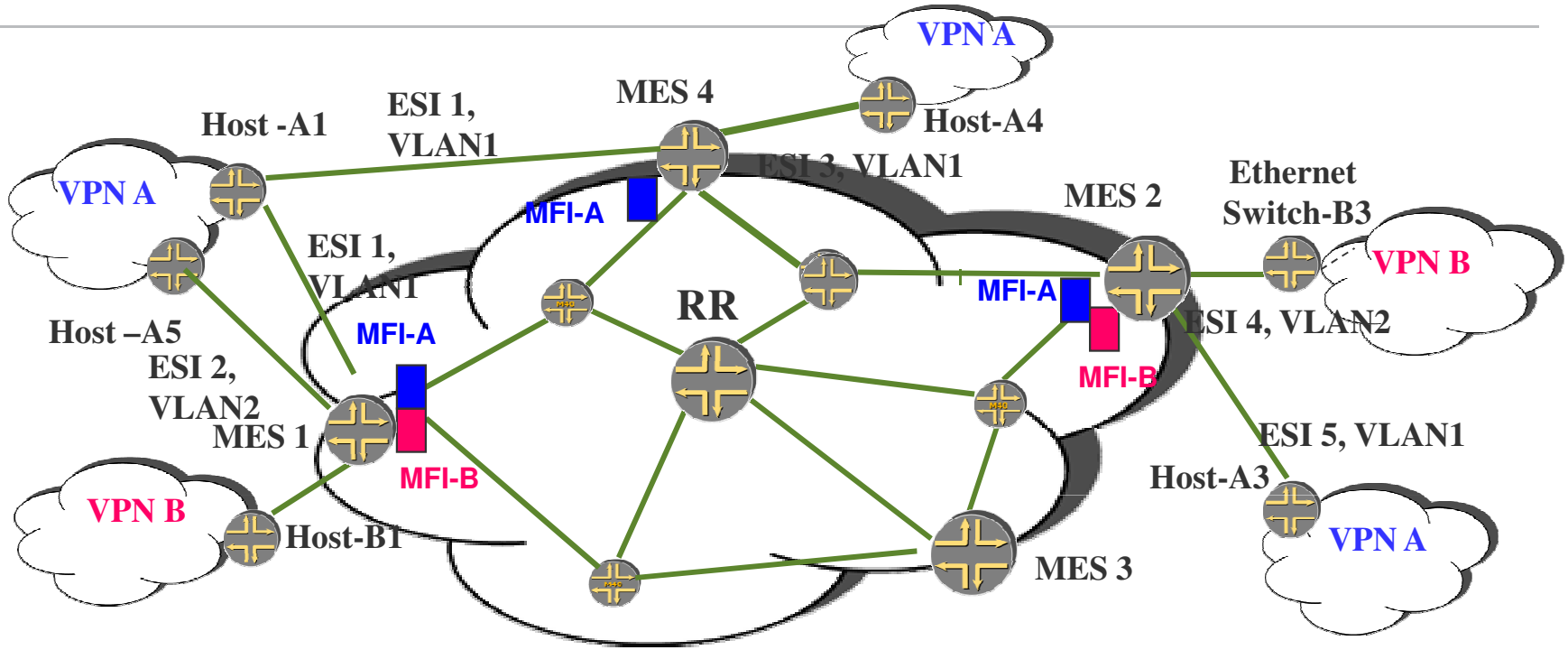
Multi-homing

- Active-active multi-homing and load balancing supported

Forwarding by DMAC address

- Unknown unicast, multicast, broadcast – flooded
- Split-Horizon implemented to prevent loops

# MAC VPN Reference Model




MES - MPLS Edge Switch; EVI – Ethernet VPN Forwarding Instance; ESI – Ethernet Segment Identifier (e.g., LAG identifier)

MESes are connected by an IP/MPLS infrastructure

Transport may be provided by MPLS P2P or MP2P LSPs and optionally P2MP/MP2MP LSPs for “multicast”

Transport may be also be provided by IP/GRE Tunnels

# MAC VPN
# Local MAC Address Learning

A MES must support local data plane learning using vanilla ethernet learning procedures

- When a CE generates a data plane packet such as an ARP request

MESes may learn the MAC addresses of hosts in the control plane using extensions to protocols such as LLDP that run between the MES and the hosts

MESes may learn the MAC addresses of hosts in the management plane

- E.g., Between VM and DCS

# MAC VPN
# Remote MAC Address Learning

MAC VPN introduces the ability for an MES to advertise locally learned MAC addresses in BGP to other MESes, using principles borrowed from IP VPNs

MAC VPN requires an MES to learn the MAC addresses of CEs connected to other MESes in the *control plane using BGP*

- *Remote MAC addresses are not learned in the data plane*

# Remote MAC Address Learning in the BGP Control Plane Architectural Benefits (1)

Increases the scale of MAC addresses and VLANs supported

- BGP capabilities such as constrained distribution, Route Reflectors, inter-AS etc., are reused

Allows hosts to connect to multiple active points of attachment

Improves convergence in the event of certain network failures

- Fast convergence based on local repair on MES-CE link failures

Allow hosts to relocate within the same subnet without requiring renumbering

# Remote MAC Address Learning in the BGP Control Plane Architectural Benefits (2)

Minimizes flooding of unknown unicast packets

- Particularly if local MAC address learning can be performed in the control or management plane

Control over which MAC addresses are learned by which devices

- Simplifies operations; enables flexible topologies etc.

Minimizes flooding of ARP

# MAC VPN Policy Attributes

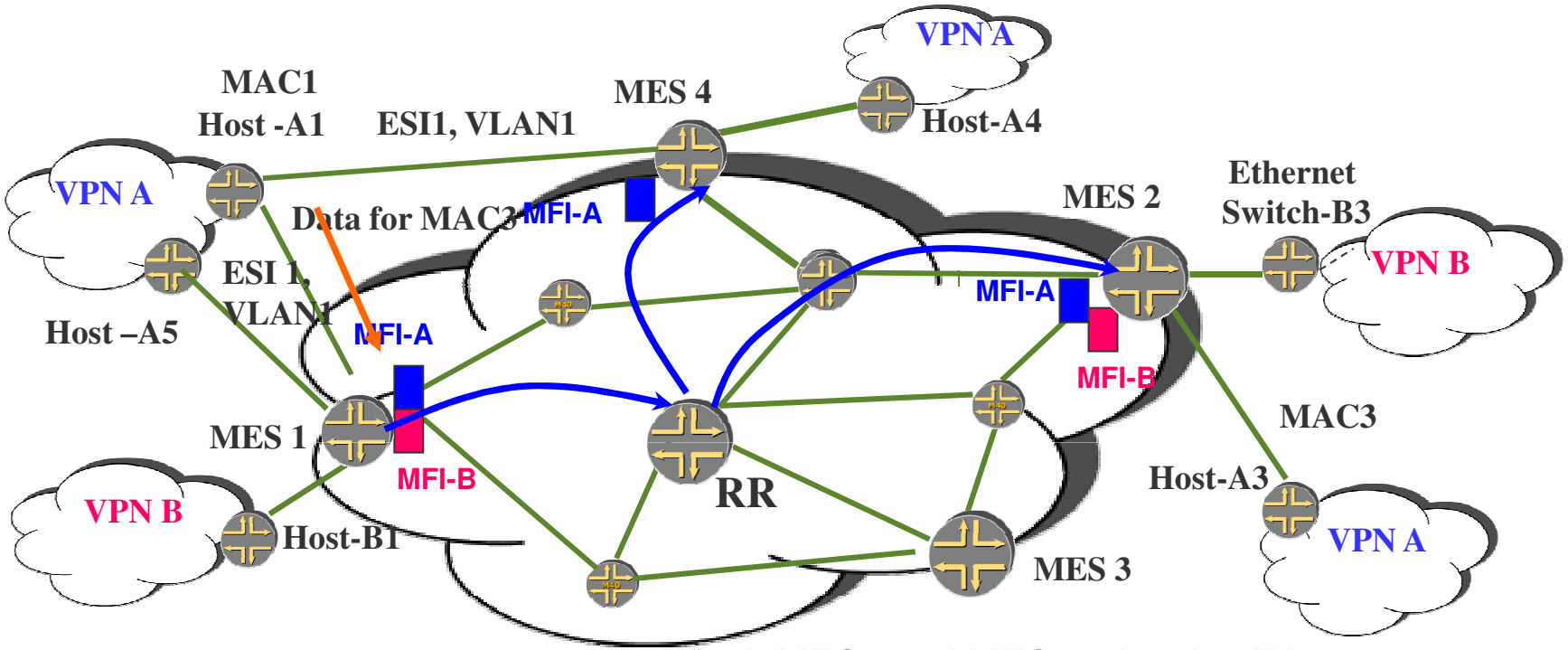
Route Targets (RT) to define the membership of a MAC VPN

- The MESes, ethernet interfaces/VLANs connecting CEs (e.g., hosts, switches, routers) to DCSes

RTs can be auto-derived from a VLAN ID

- Particularly applicable if there is a one to one mapping between a MAC VPN and a VLAN
- Simplifies configuration (by removing the need to configure RTs)

# MAC MVPN Functionality
# MAC Address Learning using BGP




MES1 advertises MAC Route in BGP.
<RD-1, MAC1, A1-VLAN, A1-ESI ID,
MAC lbl L1, VPN A RT>

Both MES1 and MES4 advertise Ethernet
Tag auto-discovery routes for
<ESI1, VLAN1> along with MPLS label and VPNA RT

MES2 learns via BGP that MAC1 is dual homed to MES1 and MES4

- Even if MES4 doesn't advertise MAC1 (as MES2 knows via the Ethernet Tag A-D routes that MES4 is connected to ESI1, VLAN1)

# Auto-Discovery of Ethernet Segments

Each MES in a given MAC VPN learns the Ethernet Segment membership of all the other MESes in the MAC VPN

Enables designated forwarder election

- To ensure that multicast, broadcast and unknown unicast packets are sent to a multi-homed CE by a single MES

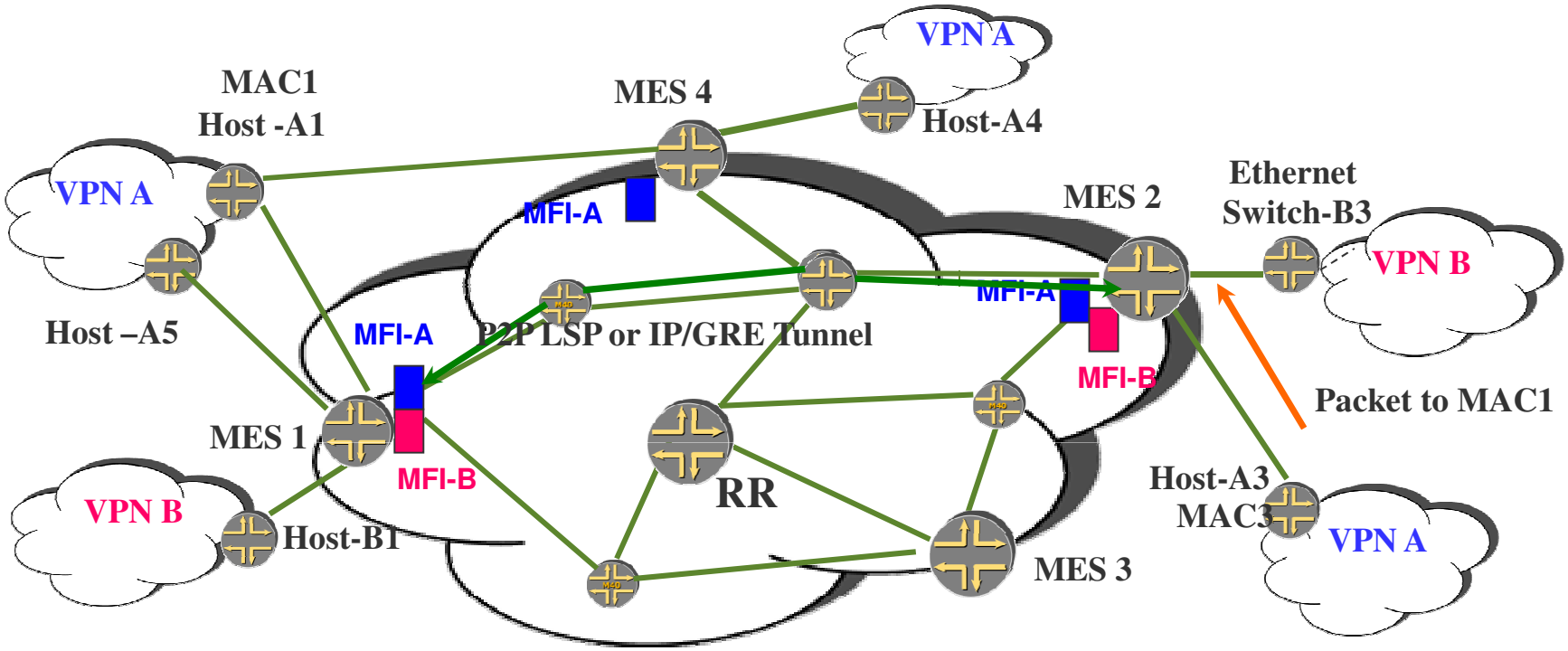
Enables “split horizon”

- To ensure that a multicast, broadcast or unknown unicast packet that is sent by Host-A1 to MES1, and then sent by MES1 to all other MESes in the MAC VPN, isn’t sent back by MES2 to Host-A1 – when host-A1 is dual homed to MES1 and MES2

Auto-discovery of Inclusive Trees for multicast, broadcast and unknown unicast traffic

One of the building blocks for load balancing i.e. active-active points of attachment

# MAC MVPN Functionality
# Unicast Data Plane



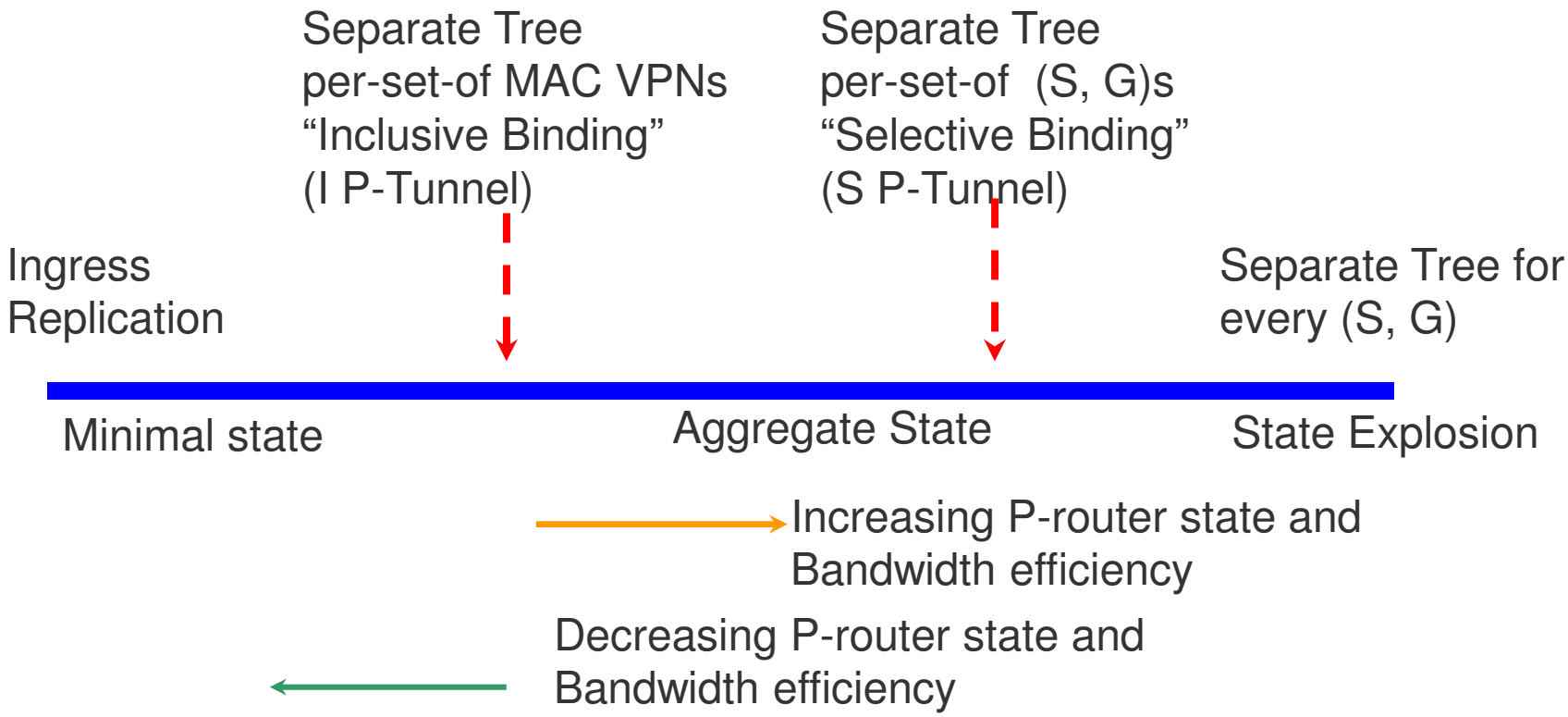

If MES2 decides to send the packet to MES1 it will use inner label <MAC VPN label advertised by MES1 for MAC1> and outer encaps as <MPLS Label for LSP to MES1> or <IP/GRE header for IP/GRE tunnel to MES1>

MES1 can load balance the traffic To MAC1 between MES1 and MES4

Granularity of label advertised by MES1 and MES4 for MAC1 is a local matter

# MAC VPN Architecture
# Multicast Data Plane – Flexible Toolkit

Tree = RSVP-TE P2MP LSPs, Receiver Initiated P2MP LSPs, MP2MP LSPs, PIM Based IP/GRE Tunnels
Ingress Replication = P2P RSVP-TE LSPs, MP2P LDP LSPs



# MAC VPN Scaling
# Control Plane

Propagation scope of MAC-VPN routes for a given VLAN/subnet is constrained by the scope of MESes that span that VLAN/subnet

- Controlled by Route Target of the MAC-VPN

Hub and spoke MAC-VPNs can help improve both control plane and data plane scale

- Spokes may have only a default route to the hub

Route Reflector infrastructure can scale using existing techniques

MAC-VPN routes may be optionally dampened

# MAC VPN Scaling
# Data Plane State

The presence of a MAC address in the control plane does not imply that it must be installed in the forwarding plane

The forwarding plane may perform only label switching

- E.g., WAN border routers
- See inter-region data center interconnect later

The forwarding plane may not require the MAC address

- If there are no active flows destined to that MAC address
- The data plane may be populated with only routes to “active” MACs

# Reducing ARP Flooding (1)

MESes perform Proxy ARP

An MES responds to an ARP request, for an IP address, with the MAC address bound to the IP address

- When the destination is in the same subnet as the sender of the ARP request
- The ARP request is not forwarded to other MESes

# Reducing ARP Flooding (2)

How does the MES learn the IP address bound to the MAC address when the MAC address is remote?

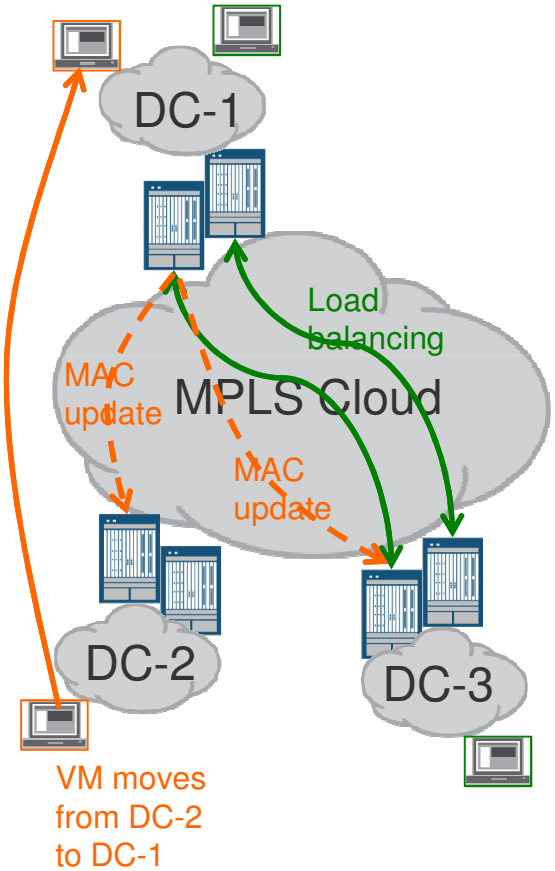
- BGP MAC routes carry the IP address bound to the MAC address

How does an MES learn the IP to MAC binding when the MAC address is local?

- Control or management plane between MES and CEs or data plane snooping

An MES advertises the local IP to MAC bindings in the MAC routes

# ENHANCING DC CONNECTIVITY FURTHER: UPCOMING MAC VPN




BGP-MAC-VPN a new standards based protocol to inter-connect L2 domains over MPLS

- Enhancing industry standard VPLS further
- Multi-vendor / open initiative led by Juniper – non-proprietary
- MPLS investment protection - builds easily over VPLS, L2/L3VPN environments

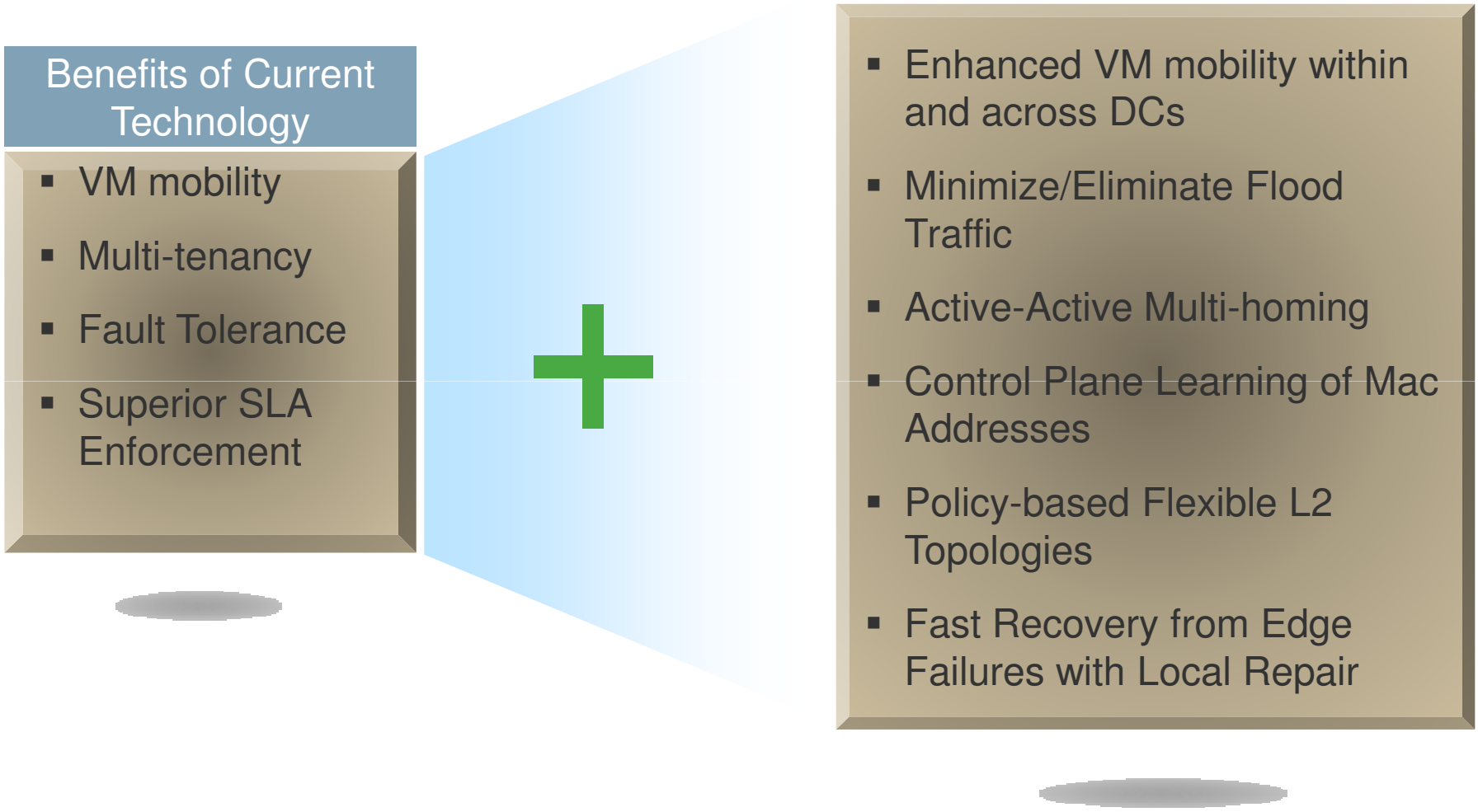
Enhancements delivered by MAC-VPN:

- Active multi-homed
- Extended control plane (MAC address) scaling
- Faster convergence from edge failures using local repair
- Flooding AND Control Plane learning
- Increased granularity on MAC address reach-ability distribution – increased support for host mobility – policy based decisions

# VPLS/MAC-VPN – SUMMARY COMPARISON

| Desirable L2 extension attributes | VPLS | MAC VPN |
|---|---|---|
| VM Mobility without renumbering L2 and L3 addresses | ✔ | ✔ |
| Ability to span VLANs across racks in different locations | ✔ | ✔ |
| Scalability (number of end hosts) in an L2 domain | ✔✔ | ✔✔✔ |
| Policy-based flexible L2 topologies similar to L3 VPNs | | ✔ |
| Multiple points of attachment with ability to load-balance vlans across them | ✔ | ✔ |
| Active-Active points of attachment with ability to load-balance flows within a single VLAN | | ✔ |
| Multi-tenant support (secure isolation, overlapping MAC, IP addresses) | ✔ | ✔ |
| Control-Plane Based Learning | | ✔ |
| Minimize or eliminate flooding of unknown unicast | | ✔ |
| Fast convergence from edge failures based on local repair | | ✔ |
| Not reliant on limited scale technology (IP Multicast) in the core for optimal performance | ✔ | ✔ |

# ETHERNET-VPN : THE "REAL" NEXT-GENERATION L2-EXTENSION SOLUTION



# For Further Reading

"BGP MPLS Based Ethernet VPN" Draft-raggarwa-sajassi-l2vpn-evpn-01.txt

- R. Aggarwal (Juniper), A. Sajassi (Cisco), W. Henderickx (Alcatel), A. Isaac (Bloomberg), J. Uttaro (AT&T), N. Bitar (Verizon), R. Shekhar (Juniper), F. Balus (Alcatel), K. Patel (Cisco), S. Boutros (Cisco)

"Requirements for Ethernet VPN (E-VPN)" draft-sajassi-raggarwa-l2vpn-evpn-req-00.txt

- A. Sajassi (Cisco), R. Aggarwal (Juniper), W. Henderickx (Alcatel), A. Isaac (Bloomberg), J. Uttaro (AT&T), N. Bitar (Verizon), S. Salem (Cisco), C. Filsfils (Cisco)


everywhere