Cloud Networking: Framework and VPN Applicability

draft-bitar-datacenter-vpn-applicability-01.txt

Nabil Bitar (Verizon)
Florin Balus, Marc Lasserre, and Wim Henderickx (Alcatel-Lucent)
Ali Sajassi and Luyuan Fang (Cisco)
Yuichi Ikejiri (NTT Communications)
Mircea Pisica (BT)

November 2011
IETF-82, Taipei, Taiwan

Scope

- **Requirements for large scale multi-tenant data centers and cloud-networks**

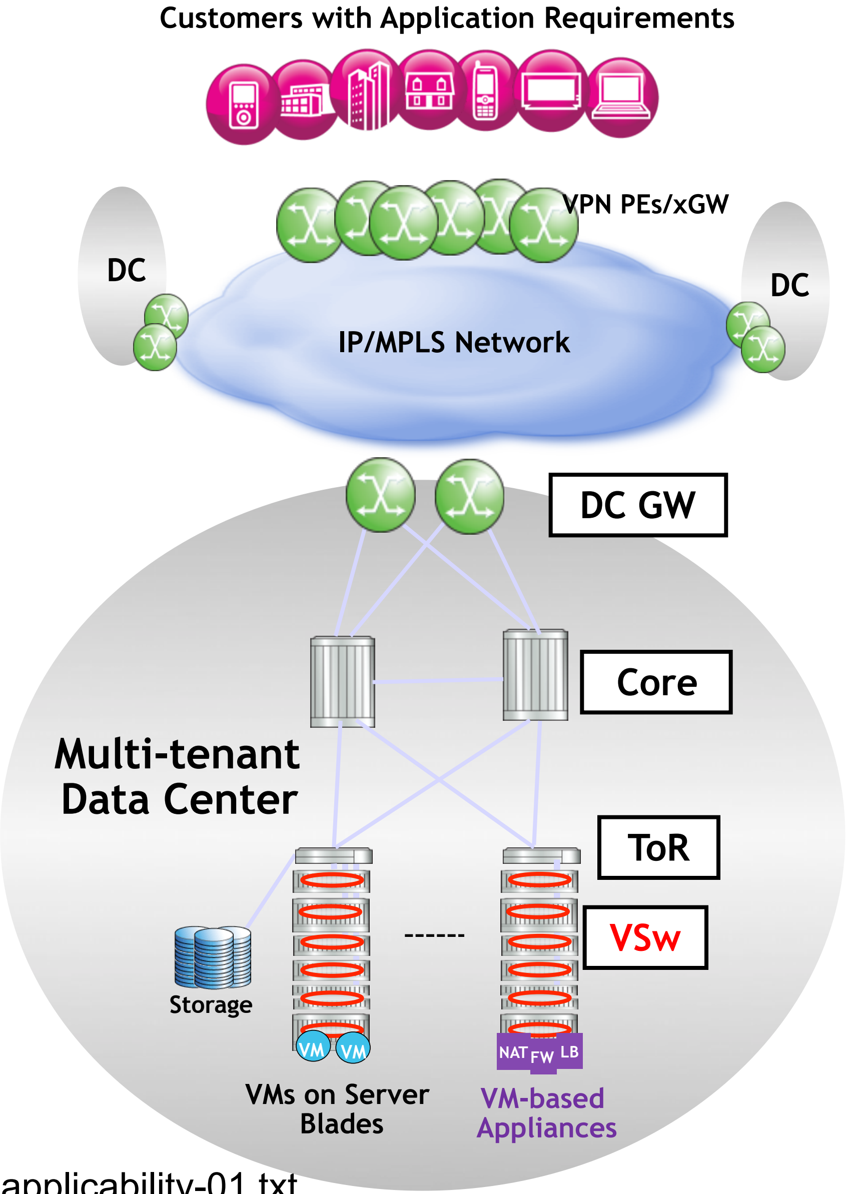
- **Applicability of existing and evolving Ethernet, L2VPN, and L3VPN technologies to multi-tenant cloud networking and tradedoffs:**
  - Intra-Data Center networks
  - Inter-data center connectivity
    - Data centers can belong to the same data center service provider, different data center providers, the tenant, and any hybrid
  - Tenant and public access to data centers

- **Scenarios – cloud networks**

- **Challenges/Gaps that still require work**

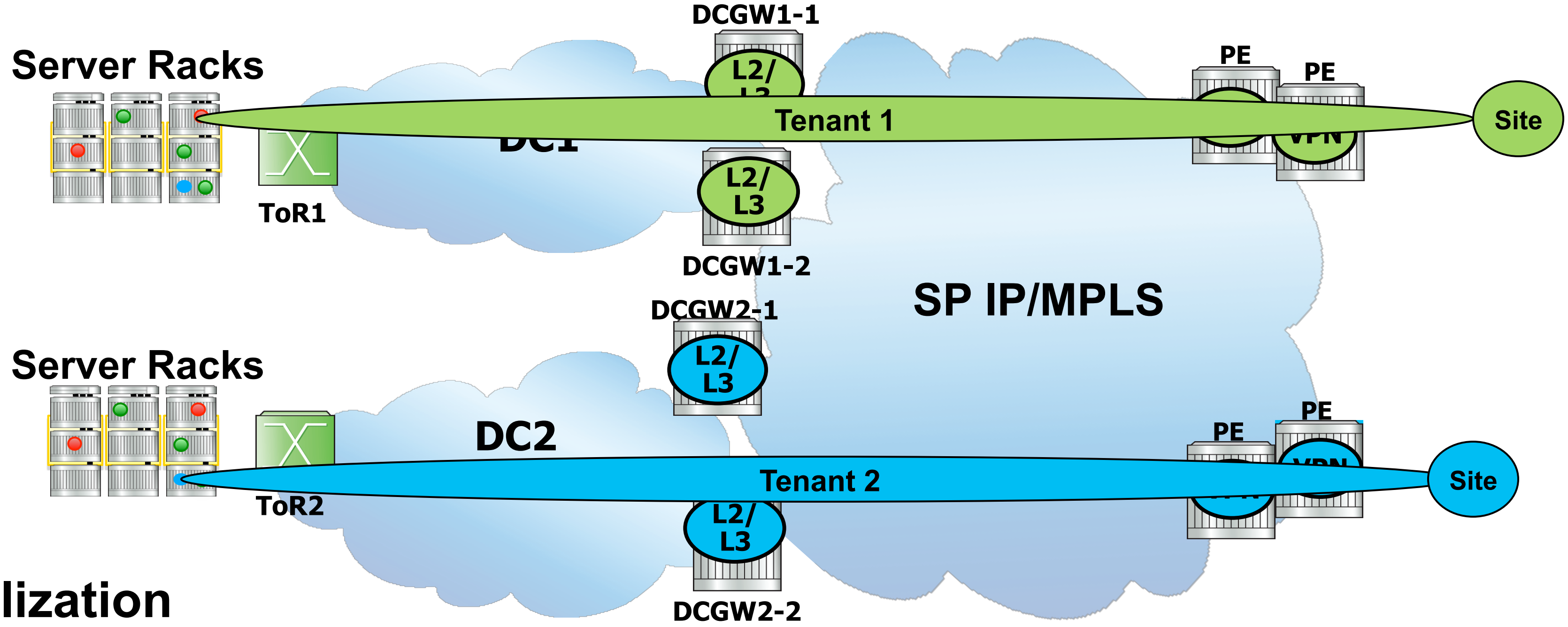
Cloud networking framework

- **DC GW – gateway to the outside world providing DC Interconnect and connectivity to Internet and VPN customers.**
- **Core Switch/Router – high capacity core node, usually a cost effective Ethernet switch; may support routing capabilities.**
- **ToR or Top of Rack – hardware-based Ethernet switch; may perform IP routing.**
- **VSw or virtual switch – software based Ethernet switch running inside the server blades**



draft-bitar-datacenter-vpn-applicability-01.txt

Multi-Tenant Data Center and Data Center-Interconnect Requirements



Virtualization

- Provide for network virtualization among tenants with overlapping addresses on the same data center network infrastructure – layer2 and layer3, and integrated routing and bridging
- Provide for compute and storage resources allocated to a tenant an attachment to the tenant virtual private network
- Provide connectivity between a tenant DC virtual infrastructure and the tenant sites, including tenant operated DCs
- Provide for dynamic stretching and shrinking of a tenant virtual infrastructure flexibly within a DC and across DCs
- Provide for DC operator virtual network management

Multi-Tenant Data Center and Data Center-Interconnect Requirements

Support large Scale DCs :

- Large number of tenants – a tenant identified by a service ID in data plane and/or control plane.(e.g., >> 4K VLAN IDs)
- Large number of VMs and multiple per-VM virtual NICs → large number of Ethernet MACs, IP addresses and ARP entries that need to be accommodated in the data center network infrastructure
- Multicast and broadcast containment per tenant virtual domain to conserve bandwidth resources
- VM movement and network rapid convergence in the presence of a large number of tenants and VMs

Optimize network resource utilization

- Bandwidth utilization within data center, on the DC connection to the WAN, and across the WAN
- FIB utilization at routers and switches
- Control plane resource utilization on routers and switches

Multi-Tenant Data Center and Data Center-Interconnect Requirements

Path Optimization

- Provide for optimized forwarding – shortest path between any two communicating endpoints in a virtual network to improve latency and network utilization efficiency
- Eliminate or reduce traffic black-holing when a VM is moved from one location to another during network transition – traffic redirection until convergence to shortest path

Resiliency: Fast recovery around failure

VM Mobility

- Maintain the existing client sessions upon VM move: VM keeps the same IP and MAC address
- Expand/shrink L2/L3 domains within a DC and across DCs
- Optimal traffic forwarding: shortest path, avoid triangular routing in steady state and provide for traffic redirection during transition
- Rewrite the MAC FIBs to redirect traffic to new location
- Have a VM IP route where needed to direct traffic to the VM

Multi-Tenant Data Center and Data Center-Interconnect Requirements

- **Auto-discovery by the network of a VM location with minimal network configuration touches – cater to ease of management**
- **Support for OAM to troubleshoot connectivity problems and provide for SLAs at the service layer (layer2 or layer3)**
- **Ease of introduction of new DC networking technologies in existing DC environments**
- **Allow for the following networking models**
  - DC service provider and the WAN network service provider providing access to a tenant site are two different entities.
  - DC service provider and the WAN network service provider providing access to a tenant site are same entities
  - DC can have its own private network for its own data center connectivity or can use another network service provider

VPN applicability to Cloud Networking

- **Layer 3 option**
  - e.g. RFC4364

- **Layer 2 options**
  - VLANs and L2VPN toolset
  - PBB and L2VPN toolset
  - TRILL and L2VPN toolset

  - In current draft version, PBB with L2VPN options have been detailed

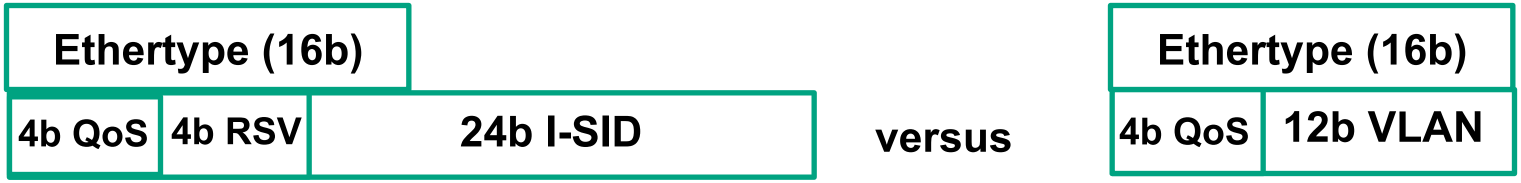
Addressing L3 virtualization with IP VPNs

- **Use full fledge IP VPN for L3 Virtualization inside a DC**
- **IP VPN advantages**
  - Interoperates with existing WAN VPN technology
  - Deployment tested, provides a full networking toolset
  - Scalable core routing – only one BGP-MP routing instance is required compared with one per customer/tenant in the Virtual Routing case
  - Service Auto-discovery - automatic discovery and route distribution between related service instances
  - Well defined and deployed Inter-Provider/Inter-AS models
  - Supports a variety of VRF-to-VRF tunneling options accommodating different operational models: MPLS [RFC4364], IP or GRE [RFC4797]
- **Connectivity models for customer IP VPN instances located in the WAN**
  - DC GW may participate directly in the WAN IP VPN
  - Inter-AS Options A, B or C - applicability to both Intra and Inter-Provider use cases

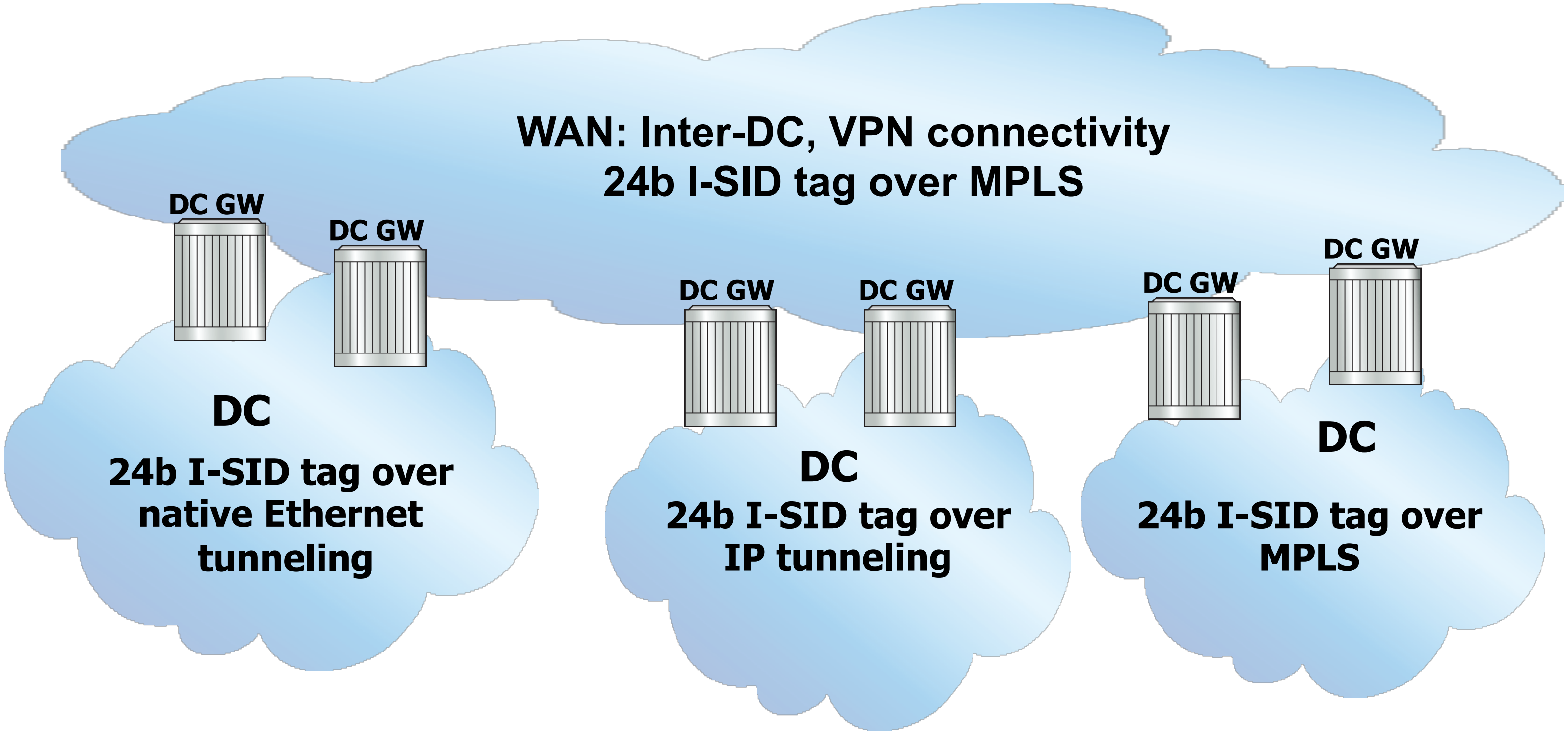
PBB + L2VPN applicability to Cloud Networking

- **24b ISID tag vs. 12b VLAN tag used for Tenant identification**
  - Expands L2 domains from 4K VLANs to 16M ISIDs
  - Standardized in 2008 by IEEE – inherits current and future IEEE specs (QoS, OAM, control plane etc…)
  - Supported in merchant silicon, proven vendor interoperability
  - Deployed in a number of large service provider networks

- **ISID tag follows the VLAN tag format**
  - I-Tag code point implies the presence of (VM) MAC DA, SA right after I-SID

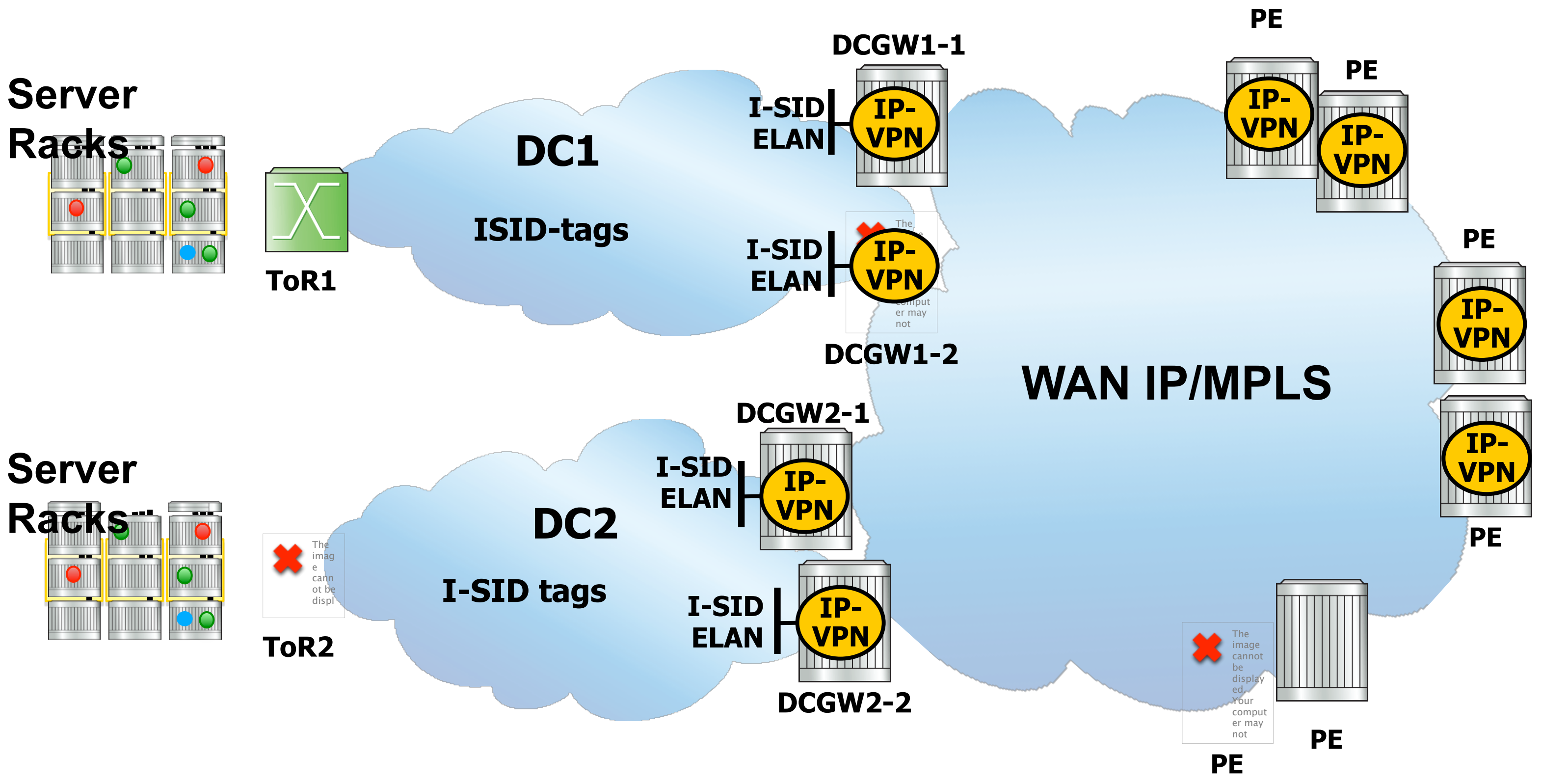


Supported tunneling options for 24b ISID Tag


WAN: Inter-DC, VPN connectivity
24b I-SID tag over MPLS
DC GW
DC GW
DC GW
DC GW
DC GW
DC GW
DC
24b I-SID tag over native Ethernet tunneling
DC
24b I-SID tag over IP tunneling
DC
24b I-SID tag over MPLS

- Native Ethernet – IEEE 802.1ah-2008
- Ethernet over IP (L2TPv3) or MPLS tunneling - PBB-VPLS
- Other more optimized IP tunneling options could be explored

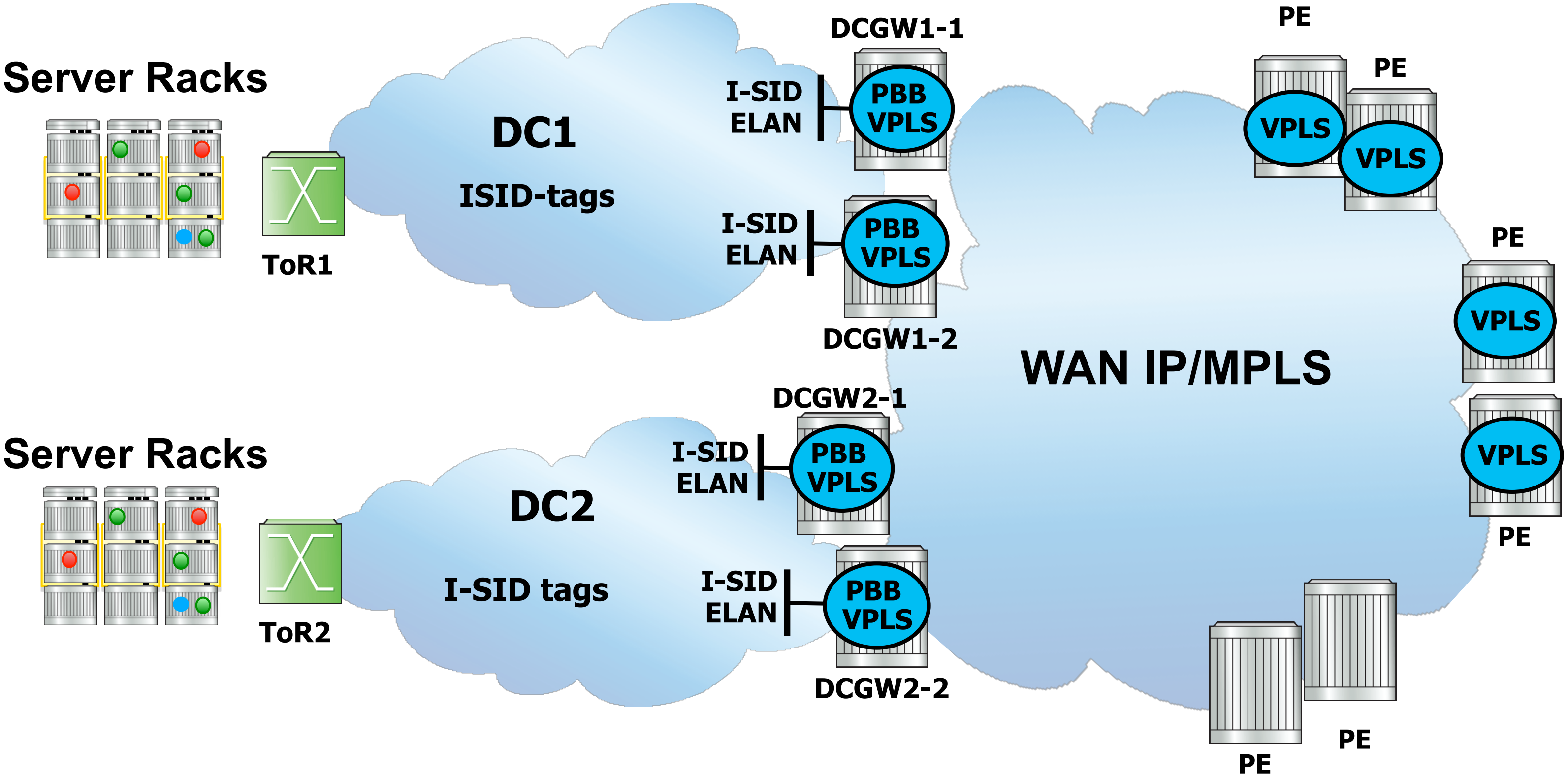
VPN interoperability w/ PBB+L2VPN IP VPN Example



PBB I-SID tag termination into IP VPN VRFs: from IP over VLAN to IP over I-SID interfaces

- **Same tunneling options: Native Ethernet, IP or MPLS or a mix**

VPN interoperability w/ PBB+L2VPN Example



Option1: PBB I-SID termination into PBB-VPLS

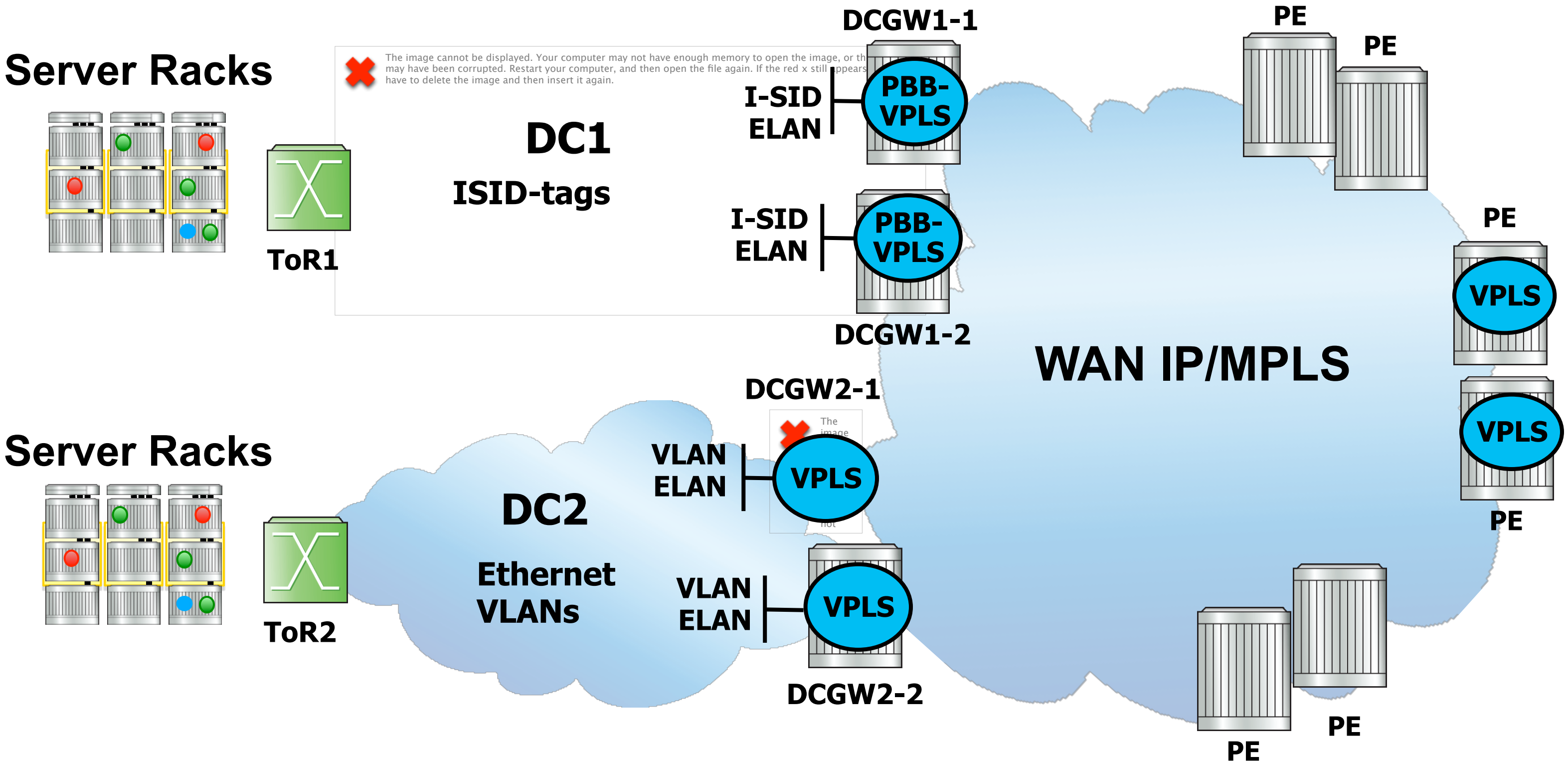
- DCGW translates back to regular VPLS

Option2: PBB I-SID transparently transported over PBB-VPLS

- DCGW acts as a Backbone Core Bridge: no ISID provisioning, no VM MAC awareness

Same tunneling options available: Ethernet or IP or MPLS or a mix

VLAN interoperability w/ PBB+L2VPN

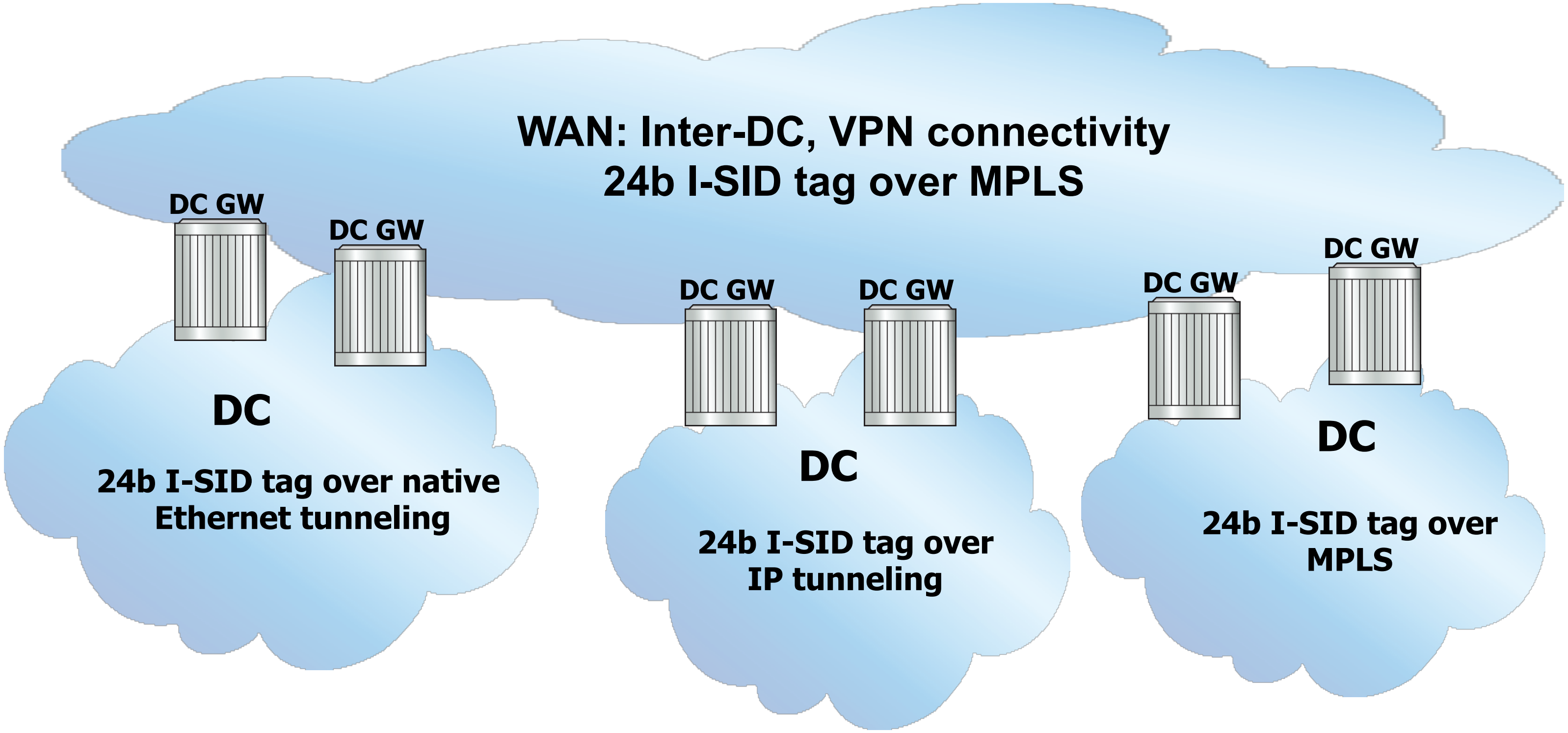

Server Racks
ToR1
The image cannot be displayed. Your computer may not have enough memory to open the image, or th
may have been corrupted. Restart your computer, and then open the file again. If the red x still appears
have to delete the image and then insert it again.
DC1
ISID-tags
DCGW1-1
I-SID
ELAN
PBB-
VPLS
I-SID
ELAN
PBB-
VPLS
DCGW1-2
PE
PE
PE
VPLS
VPLS
PE
WAN IP/MPLS
Server Racks
ToR2
DC2
Ethernet
VLANs
DCGW2-1
VLAN
ELAN
VPLS
VLAN
ELAN
VPLS
DCGW2-2
PE
PE

Similarly with VPLS interop, DCGWs in DC1 translate PBB I-SIDs to VPLS

- **Alternatively DCGWs in DC2 may run PBB-VPLS and translate I-SIDs to VLANs**

PBB and L2VPN - control plane options

Legacy: PW Mesh with split horizon
Nextgen: BGP (PBB-EVPN)



WAN: Inter-DC, VPN connectivity
24b I-SID tag over MPLS

24b I-SID tag over native Ethernet tunneling

24b I-SID tag over IP tunneling

24b I-SID tag over MPLS

Legacy: STP, MC-LAG
Nextgen: IS-IS (SPB)

Regular IP Routing

Regular IP Routing,
MPLS toolset

PBB and L2VPN - control plane options

- **Re-use of IP Routing toolset: IS-IS, BGP based control plane choices**
- **Service Auto-discovery, minimize operator provisioning**
  - Hypervisor to ToR VM discovery methods: VDP (IEEE 802.1Qbg), IGMP, SDN, others
- **Supports L2 multipathing and Active/Active Multihoming**
- **Fast convergence, Traffic Steering**
- **Inter-AS expansion with BGP**

Other work in progress

- Discussion on VM Mobility, Optimal traffic forwarding – see draft-raggarwa-data-center-mobility-01.txt

- ARP suppression discussed in PBB-EVPN (draft-sajassi-l2vpn-pbb-evpn-02.txt) and EVPN (draft-raggarwa-sajassi-l2vpn-evpn-04.txt)
- ARP Broadcast Reduction for Large Data Centers (draft-shah-armd-arp-reduction-02.txt)

PBB+L2VPN Solution Summary

Component	PBB+L2VPN toolset
Tenant ID	24b tag
Tag format	IEEE 802.1ah I-SID
VM MAC hiding	Yes
Tunneling options	IP, MPLS, Ethernet
IP tunnel format	PW/L2TPv3
IP core routing	Yes

pathing

PBB+L2VPN and DC Challenges

Draft Requirements	VPN Applicability
Service Scale	Yes (16M)
MAC scale	Yes (overlay)
Flood containment	Yes (Ethernet, MPLS) TBD for IP overlay
Multi-	Yes (IS-IS, BGP)
Multicast efficiency	P2MP LSPs, TBD (IP)
Interop	Yes
VM Mobility	Work in progress

Next steps

- IP tunneling optimization for I-SID tag transport
- Network auto-provisioning and flood containment through the auto-discovery of VM and VM groups: agree on mechanism(s)
- Broadcast, Multicast handling over IP Core requires work
- Tunnel and Service Address Translation between Cloud Provider and Tenant/Network Service Provider