2011-9-9

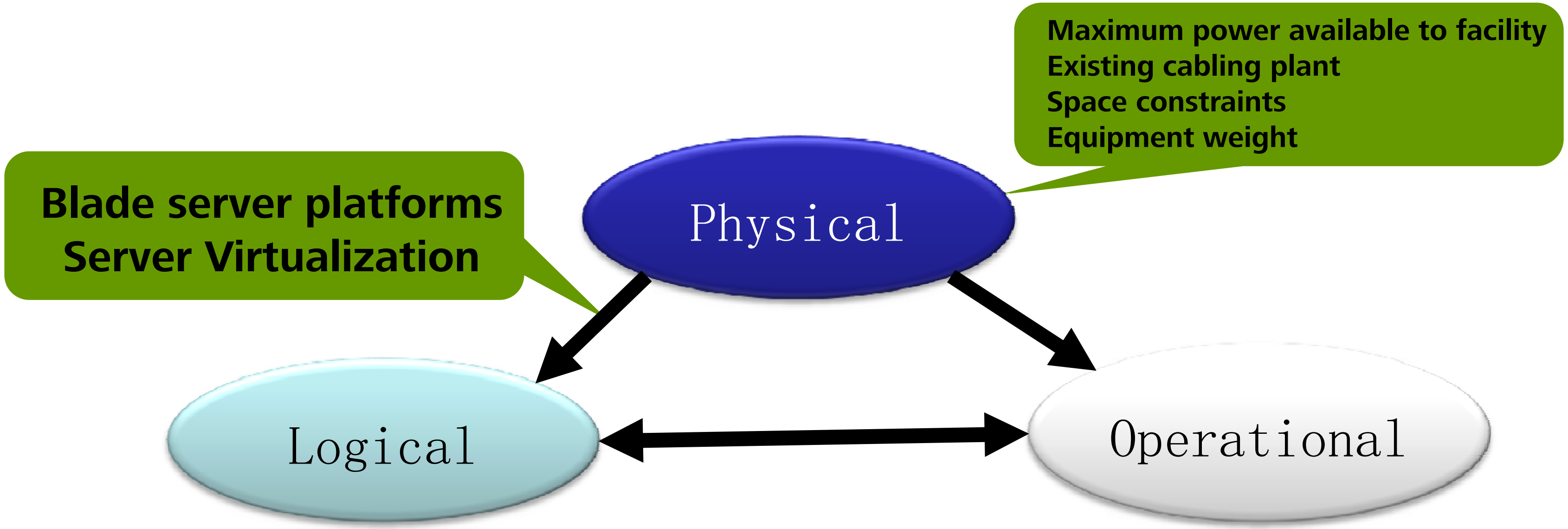
Address Resolution Scalability for VPN oriented Data Center Services

www.huawei.com

Linda Dunbar (ldunbar@huawei.com)

Mega Data Centers

Data Center Challenges



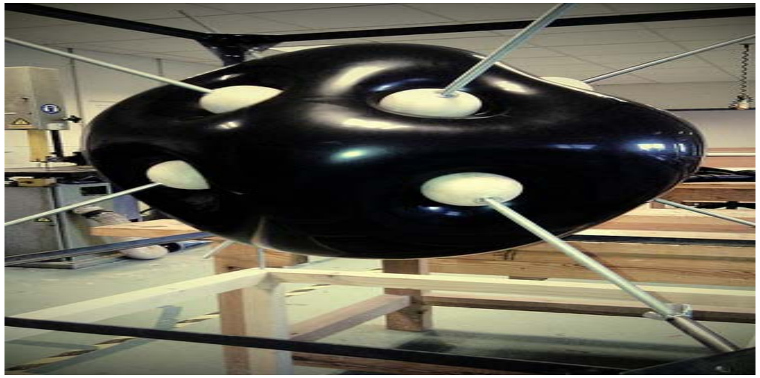
Maximum power available to facility
Existing cabling plant
Space constraints
Equipment weight

Blade server platforms
Server Virtualization

Physical

Logical

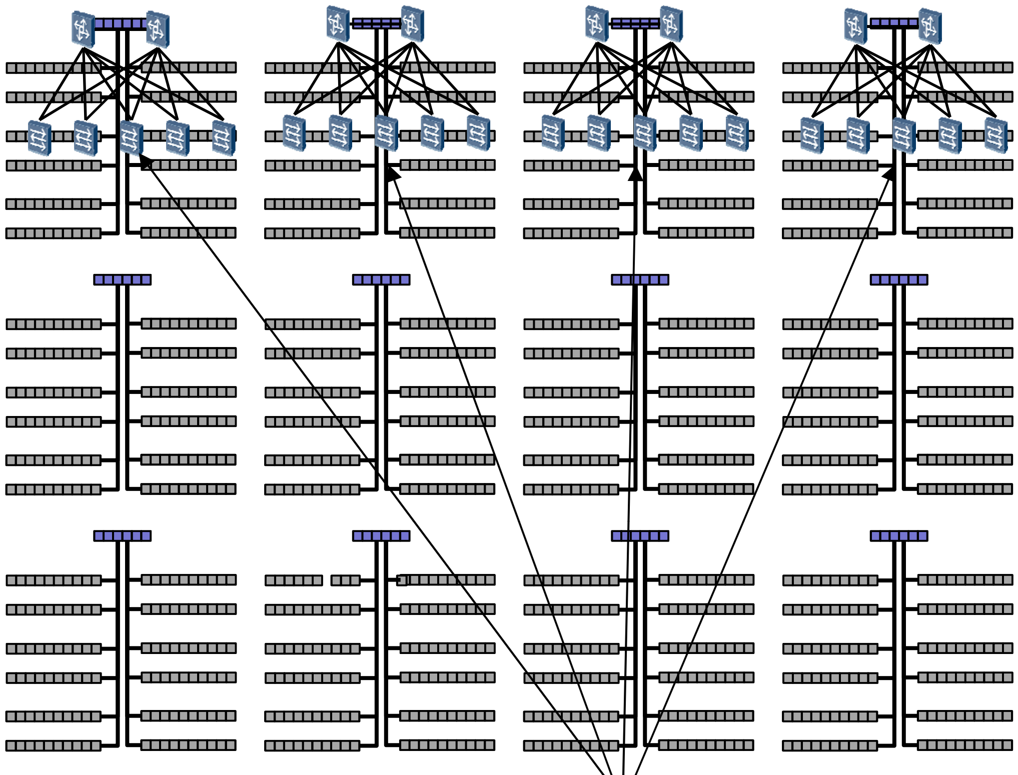
Operational



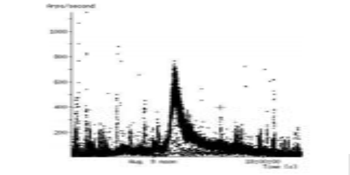
IETF ARMD: Address Resolution for Massive number of hosts in Data center

When subnets are not confined to one location:

- ARP/ND broadcast/multicast messages are no longer confined to smaller number of ports.
- Gateway routers can't handle all the ARP/ND requests from all hosts
- Some hosts might be temporarily out of service during VM migration.
- Gratuitous ARP broadcast from new location flood to all TOR switches



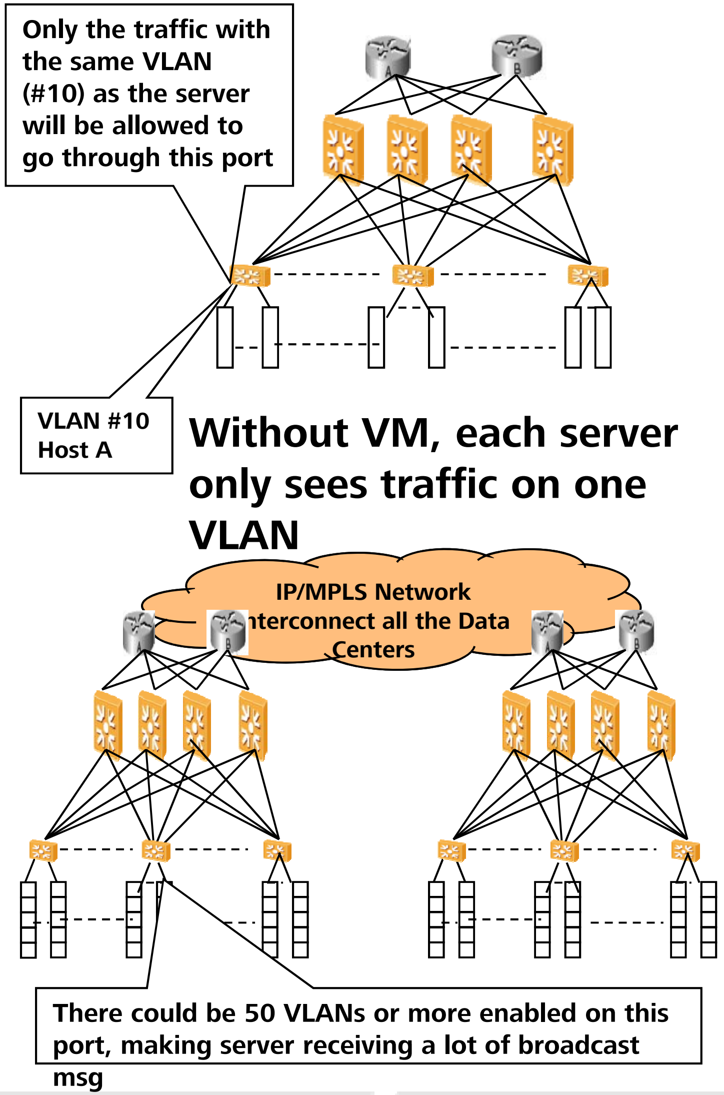
Mixed L3 domains caused by VM mobility



ARPs received per second over time on a LAN of 2456 hosts.

Why VLAN partition (smaller subnet) alone is not enough

- VLAN works well when all hosts belonging to one VLAN are confined to one location .
- When hosts belonging to one VLAN are placed at different racks and one rack has multiple VLAN enabled, all broadcast messages are no longer confined anymore.
  - The effect is same as one large VLAN.



Only the traffic with the same VLAN (#10) as the server will be allowed to go through this port

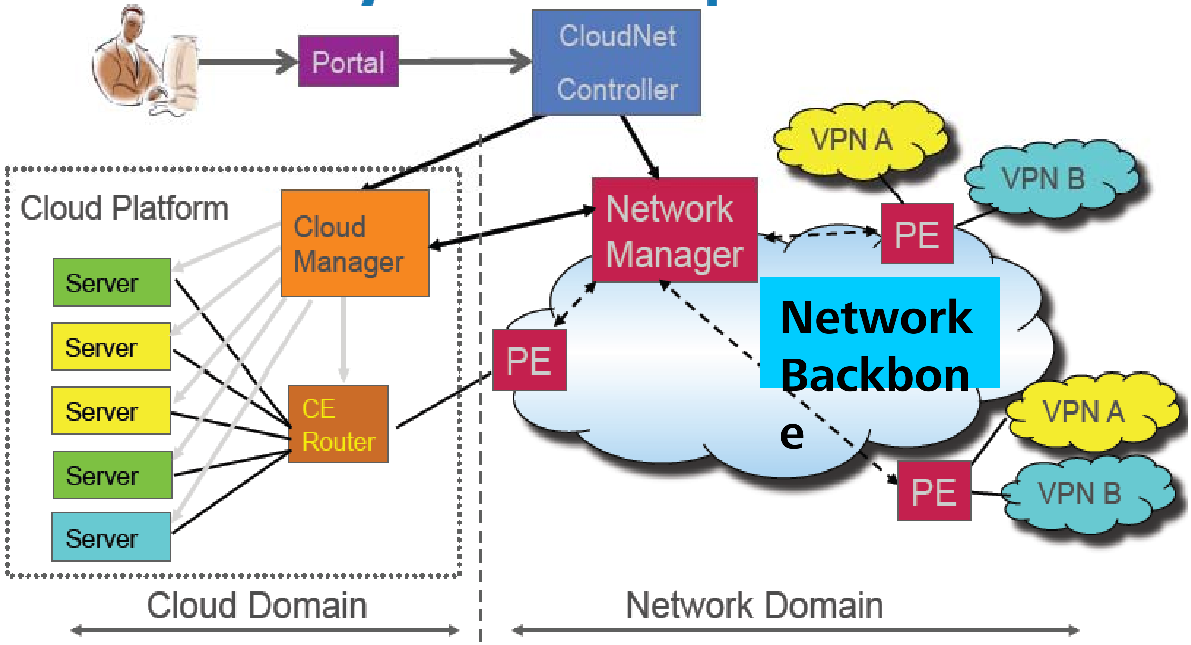
VLAN #10
Host A

Without VM, each server only sees traffic on one VLAN

There could be 50 VLANs or more enabled on this port, making server receiving a lot of broadcast msg

VPN Oriented Data Center Services & CloudNet (http://www2.research.att.com/~kobus/doc)s/cloudnet.pdf)

CloudNet System Components



High level abstraction:

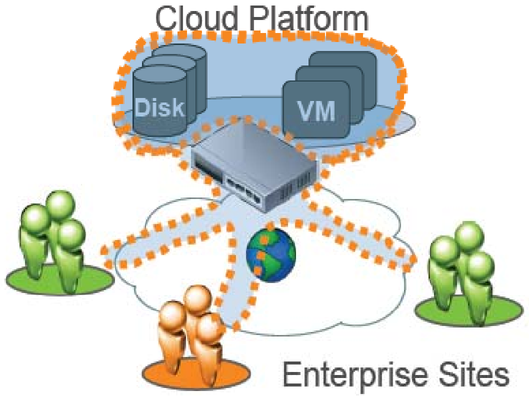
- Create compute resources
- Map into VPN
- Cross domain interaction

Cloud Manager:

- Create compute resources
- Map into VPN (cloud side)

Network Manager (IRSCP):

- VPN management (network side)
- Dynamic VPN mapping/stitching

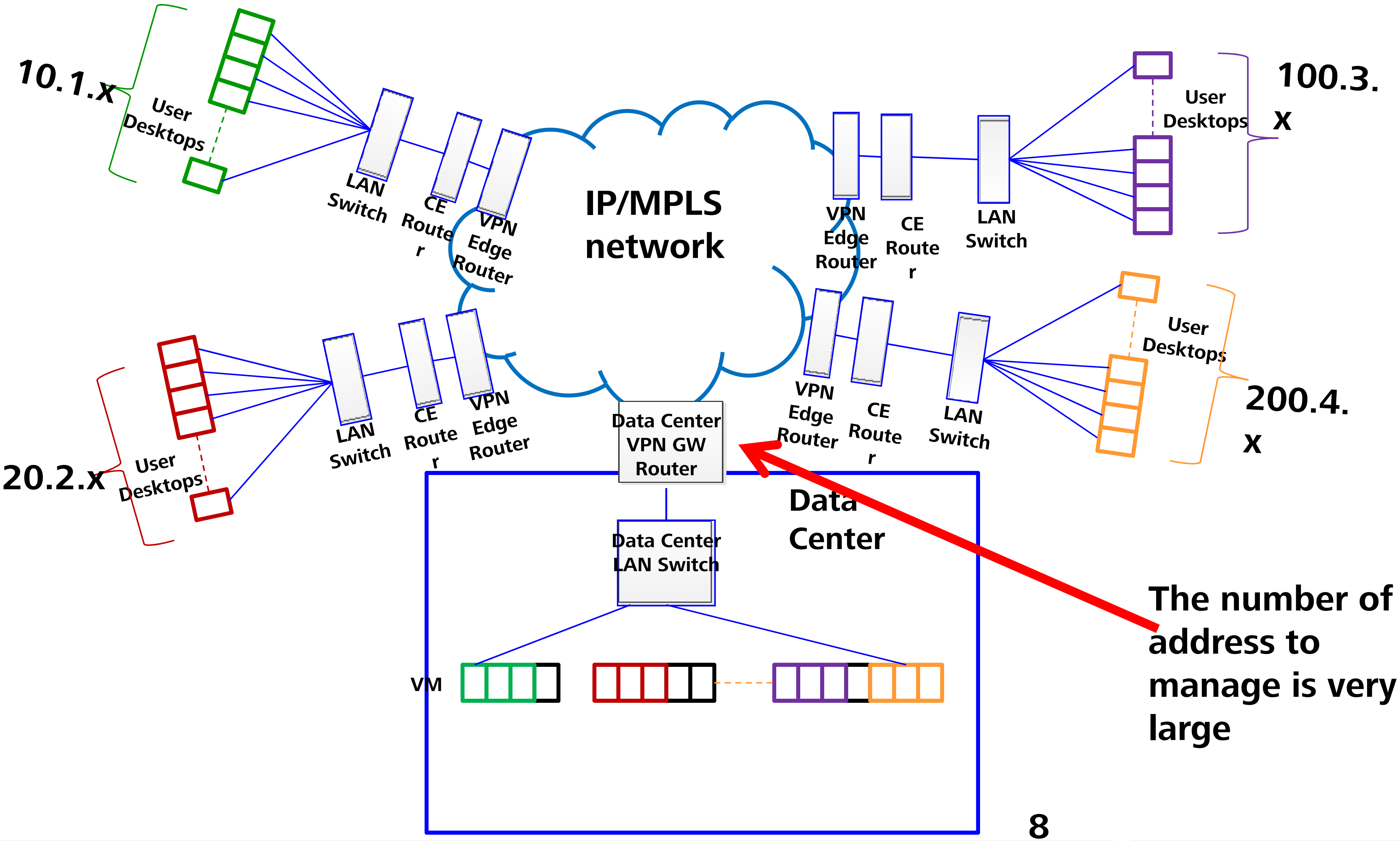


Network Operator's VPC: Span across multiple data centers. Focus on VPNs provided by a network operator, as opposed to IPSec VPNs that create software tunnels between each end host.
-- Different from Amazon's VPC

Address Issues Induced by VPN-o-DS

- **Address scalability**
- **Address conflict**

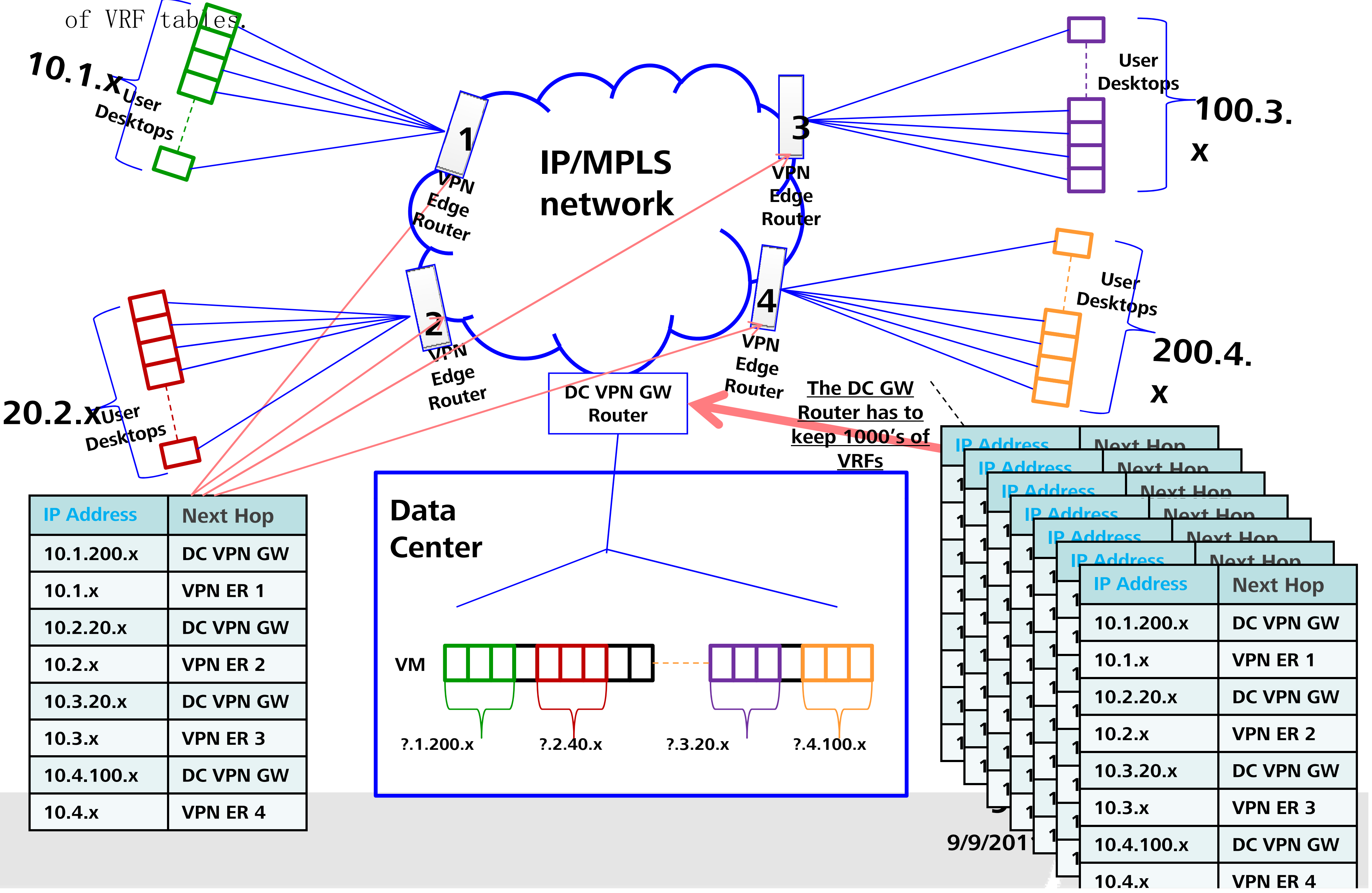
Scalability



The number of address to manage is very large

But the routing tables at PEs become more complicated

- If Data Center hosts VMs for thousands of VPNs, gateway router has to maintain thousands of VRF tables.

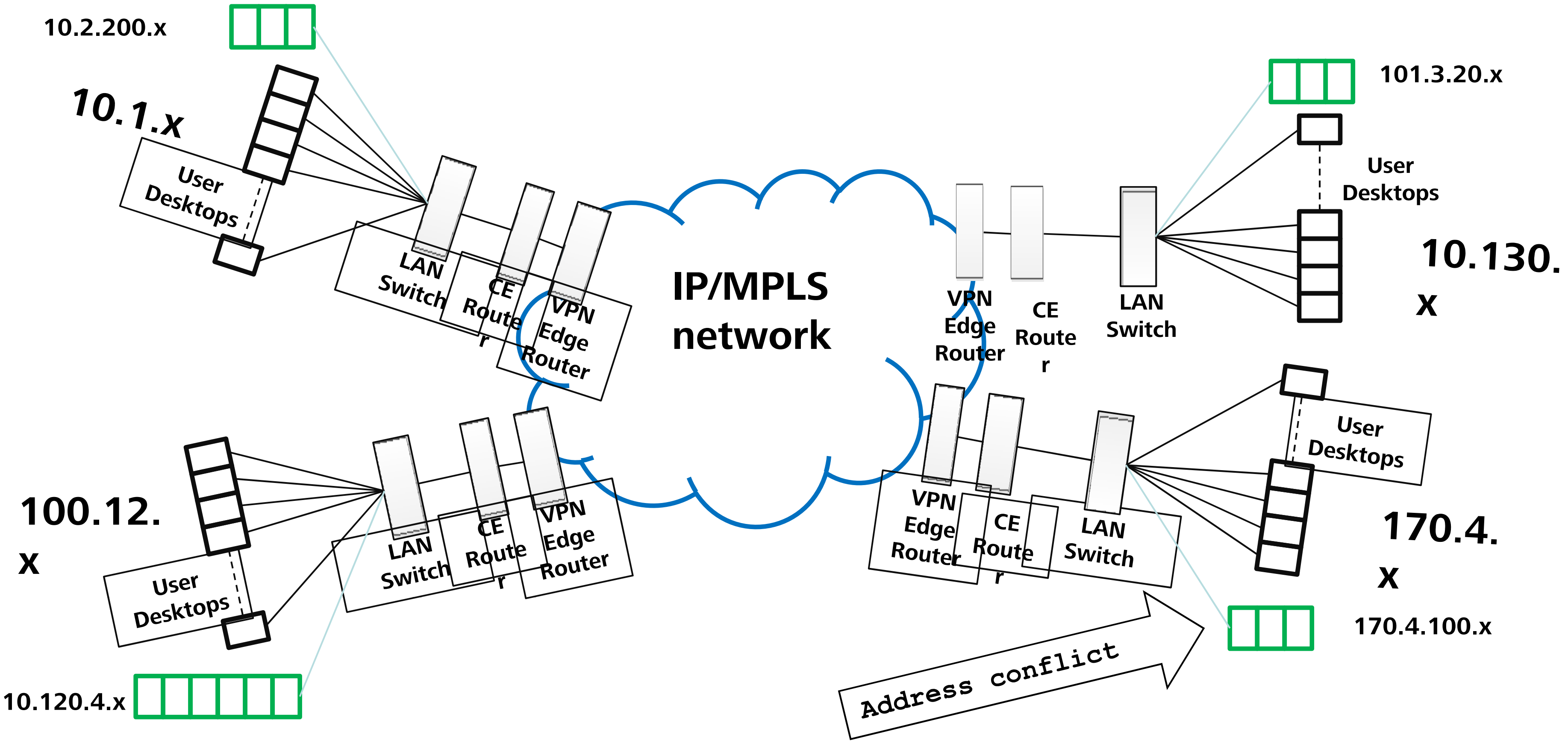


The DC GW Router has to keep 1000's of VRFs

IP Address	Next Hop
10.1.200.x	DC VPN GW
10.1.x	VPN ER 1
10.2.20.x	DC VPN GW
10.2.x	VPN ER 2
10.3.20.x	DC VPN GW
10.3.x	VPN ER 3
10.4.100.x	DC VPN GW
10.4.x	VPN ER 4

IP Address	Next Hop
10.1.200.x	DC VPN GW
10.1.x	VPN ER 1
10.2.20.x	DC VPN GW
10.2.x	VPN ER 2
10.3.20.x	DC VPN GW
10.3.x	VPN ER 3
10.4.100.x	DC VPN GW
10.4.x	VPN ER 4

Address conflict in L3VDCS



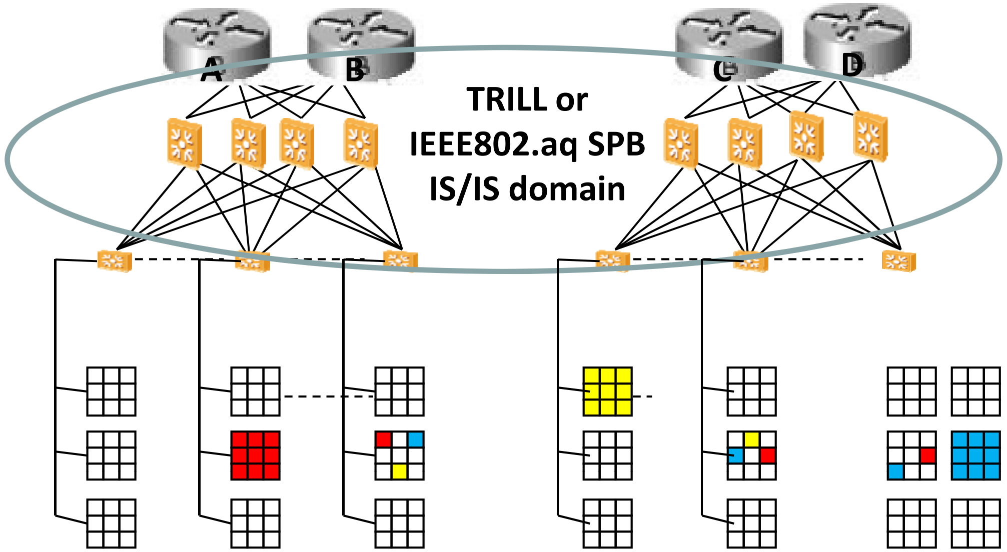
- **Potential Solutions**

Key Components of Data Center Network: Managed vs. Dynamic

- **Managed information:**
  - **Association between hosts and racks/rows, which is orchestrated by Server/VM management system**
- **Dynamic Information: (Multiple attachment scenarios)**
  - **When a server have multiple ports connected to network,**
    - **the server is aware of connectivity to its immediately connected switches (ToR or Access Modules) and can choose any of the valid links to forward traffic to the network.**
    - **But network is not aware which switches can reach the server (because the link between a server and a switch is out of Routing domain)**
  - **When backbone IS/IS routing domain is bounded by the Aggregation switches and a ToR (or Access Module) is connected to multiple Aggregation switches:**
    - **A ToR (or an Access Module) is aware of connectivity to its immediately connected Aggregation (backbone edge) switch and can choose any of the valid links to forward traffic to the network.**
    - **But backbone network is not aware which Edge nodes can reach the ToR**

Directory Assistance is Simple when there is single path from hosts to backbone network

- Just one ToR switch (Access Module) per rack & IS/IS domain to ToR



Target host	Access Module
MAC&VLAN/IP	Switch ID

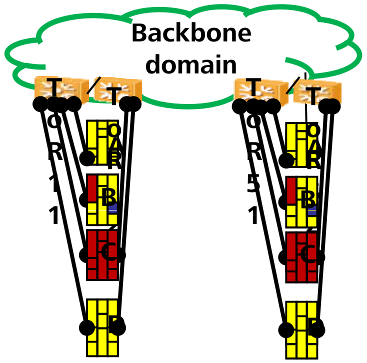
Table 1: Directory for Access Switch

- Managed information:
- Dynamic information: IS/IS routing

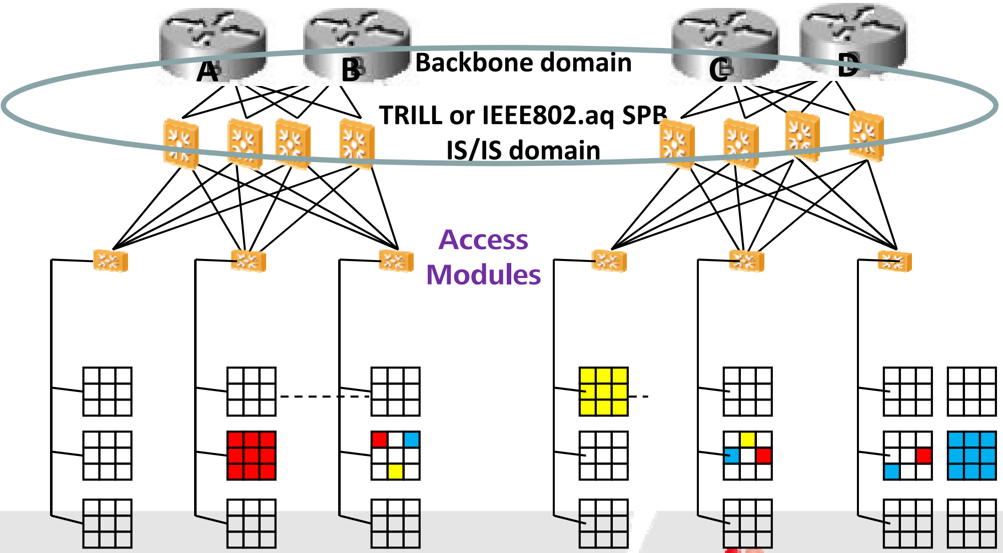
Directory assistance becomes tricky when one node can be accessed by multiple switches

– Multiple Attachment Scenario

- **Scenario 1: Servers with multiple ports connecting to different switches into network**



- **Scenario 2: Access Modules (or ToR) are connected to multiple Aggr switches which are the boundary of backbone network**



Directory Information

- **Managed information:**

Target host	Access Module
MAC&VLAN/IP	**ToR Switch ID**

Scenario #1

Node ID	Edge switch to network
ToR ID (or Server ID)	Aggr ID (or Edge Switch ID)
	Aggr ID (or Edge Switch ID)
	Aggr ID (or Edge Switch ID)
	Aggr ID (or Edge Switch ID)

Scenario #2

- **Dynamic information:**
  - **ToR to Aggr local link status**
  - **Remote Aggr to Target ToR link status**

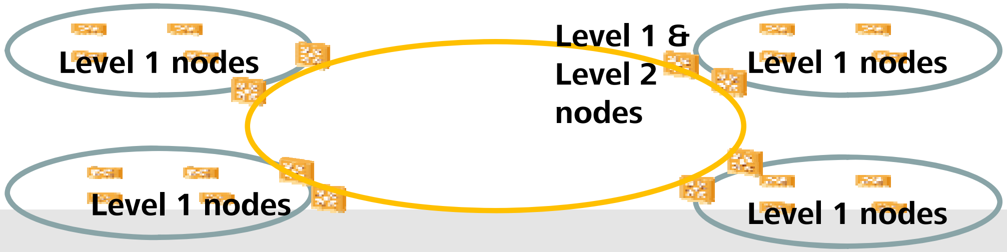
Node ID	Edge switch to network	Status
ToR ID (or Server ID)	**Aggr ID (or Edge Switch ID)**	**Up/Down**
	Aggr ID (or Edge Switch ID)	**Up/ Down**
	Aggr ID (or Edge Switch ID)	**Up/ Down**
	Aggr ID (or Edge Switch ID)	**Up/ Down**

Scenario #2

Table 2: Directory for Access Switch (not in IS/IS domain)

Why not using Level 1 & Level 2 hierarchical IS/IS Routing for network with very large number of nodes?

- **Level 1 and Level 2 Hierarchical routing (both ISIS and OSPF) works well when the addresses can be aggregated:**
  - **One aggregated IP address can represent a very large number of individual addresses.**
    - each area can have its distinct prefix.
  - **Area ID encoded in the Address (OSI).**
- **Hierarchical Routing blows up when addresses are flat…**
  - **Can’t have one aggregated address to represent a set of nodes (hosts)**
  - **Route insertion time for large number of non-aggregaable addresses will be too long**
  - **Too many entries for each switch, especially boundary switches.**
    - Any switches in any Area need to know if a target address “b” should be sent to boundary node (i.e. Level 1&Level 2 node).
    - Boundary nodes need to know how to forward a target “b”.



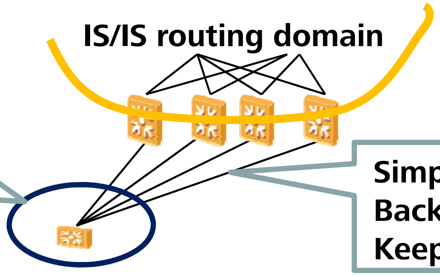
Solution to achieve hierarchical routing for large network with flat addresses:

- **Only the backbone addresses, i.e. the addresses for the nodes in IS/IS routing domain, are visible in the network**
- **With Directory assistance, source node encapsulates data packet with Backbone (i.e. boundary node) Addresses:**

Behavior of Simplified backbone Edge

(Simplified Rbridge in TRILL or Simplified Provider Edge in MAC-in-MAC)

- Goal: to achieve hierarchy in hiding addresses for large network with flat addresses
- Basic behavior:
  - Does not participate in IS/IS routing.
  - But perform header encapsulation (TRILL or backbone MAC header with SA = Ingress Edge and DA=Egress Edge)
- Mechanism in selecting Source Address for the header of backbone encapsulation:



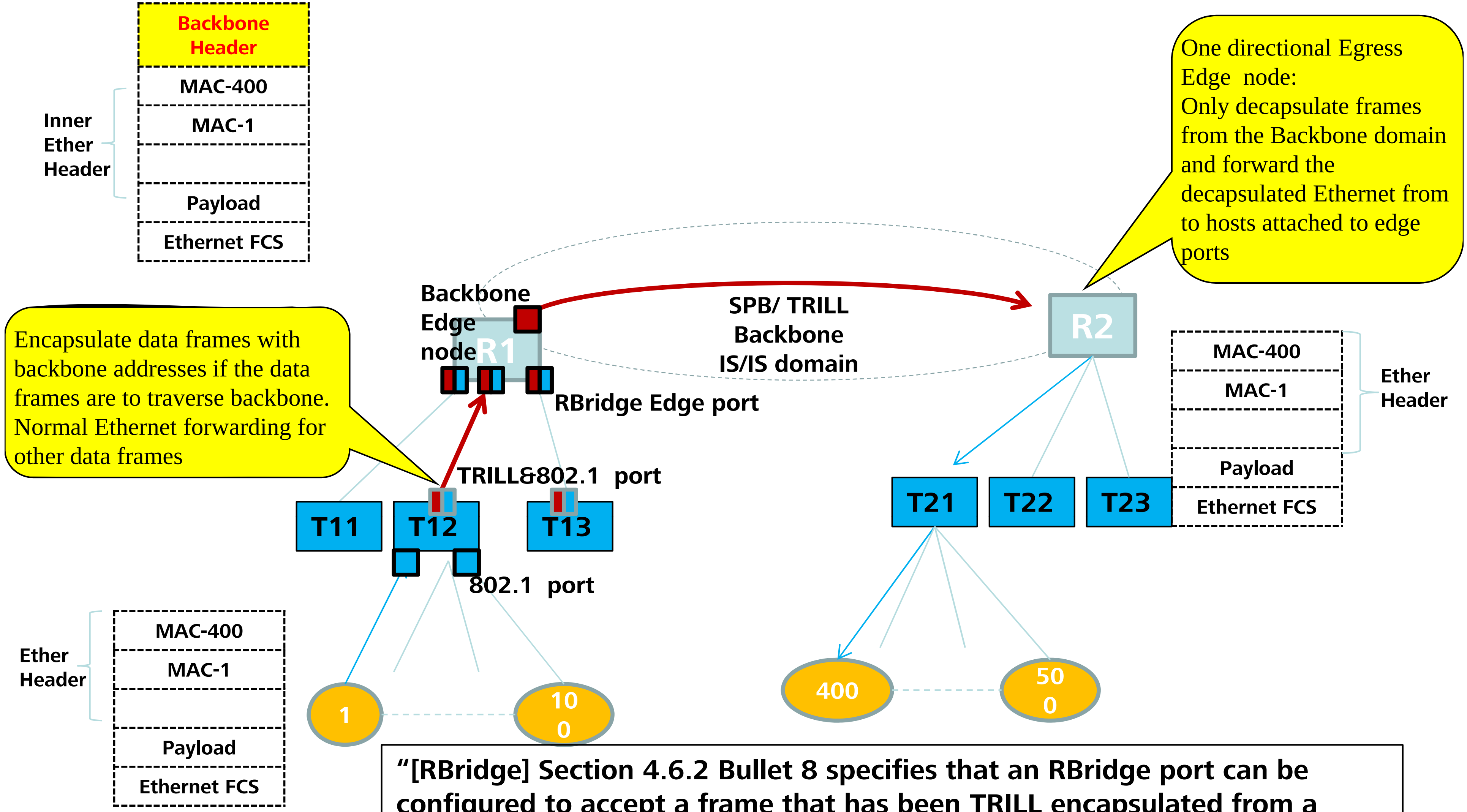
Choose any of directly attached Edge nodes as Ingress SA (local decision, same as ECMP)

Simplified edge node can monitor the Backbone Edge nodes status by a simple KeepAlive (Hello) exchange.

- Mechanism in selecting Destination Address for the header of backbone encapsulation:
  - Directory assistance
  - Distributed routing discovery

Simplified Backbone Edge node

-Does not participate in IS/IS routing. But encapsulate backbone header with SA = Ingress Edge and DA=Egress Edge



Encapsulate data frames with backbone addresses if the data frames are to traverse backbone. Normal Ethernet forwarding for other data frames

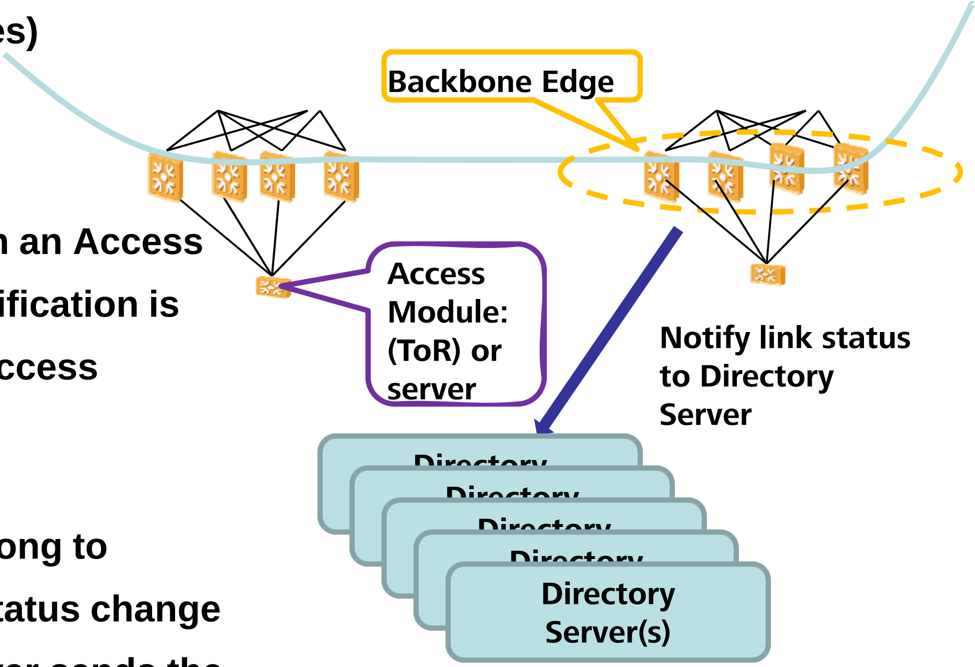
One directional Egress Edge node:
Only decapsulate frames from the Backbone domain and forward the decapsulated Ethernet from to hosts attached to edge ports

"[RBridge] Section 4.6.2 Bullet 8 specifies that an RBridge port can be configured to accept a frame that has been TRILL encapsulated from a neighbor that is not an RBridge..

Mechanisms to learn which egress can reach target:

– Directory maintains the link status between Access Module to its Backbone Edge nodes.

- Directory contents (with some dynamic attributes)
  - {Access Module, {Backbone Edge, Link Status}}
    - The link status is dynamic.
  - {Access Module, {host-addr&VLAN}}
- Whenever there is status change in link between an Access Module and a Backbone Edge, the following notification is sent to the corresponding directory (either by Access Switch or the Backbone Edge):
  - (Access Switch X, BackBone Edge Y, LinkStatus).
- Assume hosts attached to Access Module X belong to VLAN#1, …, or VLAN#n. Upon receiving a link status change notification for Access Module X, Directory Server sends the following notification to all Access Modules which have hosts belonging to any of the VLAN#1, … or VLAN#n :
  - (Access Module X, BackBone Edge Y, LinkStatus).
- All Backbone Edge nodes maintain the reachability to each other via SPB or TRILL's IS/IS


Backbone Edge
Access Module: (ToR) or server
Notify link status to Directory Server
Directory
Directory
Directory
Directory
Directory Server(s)

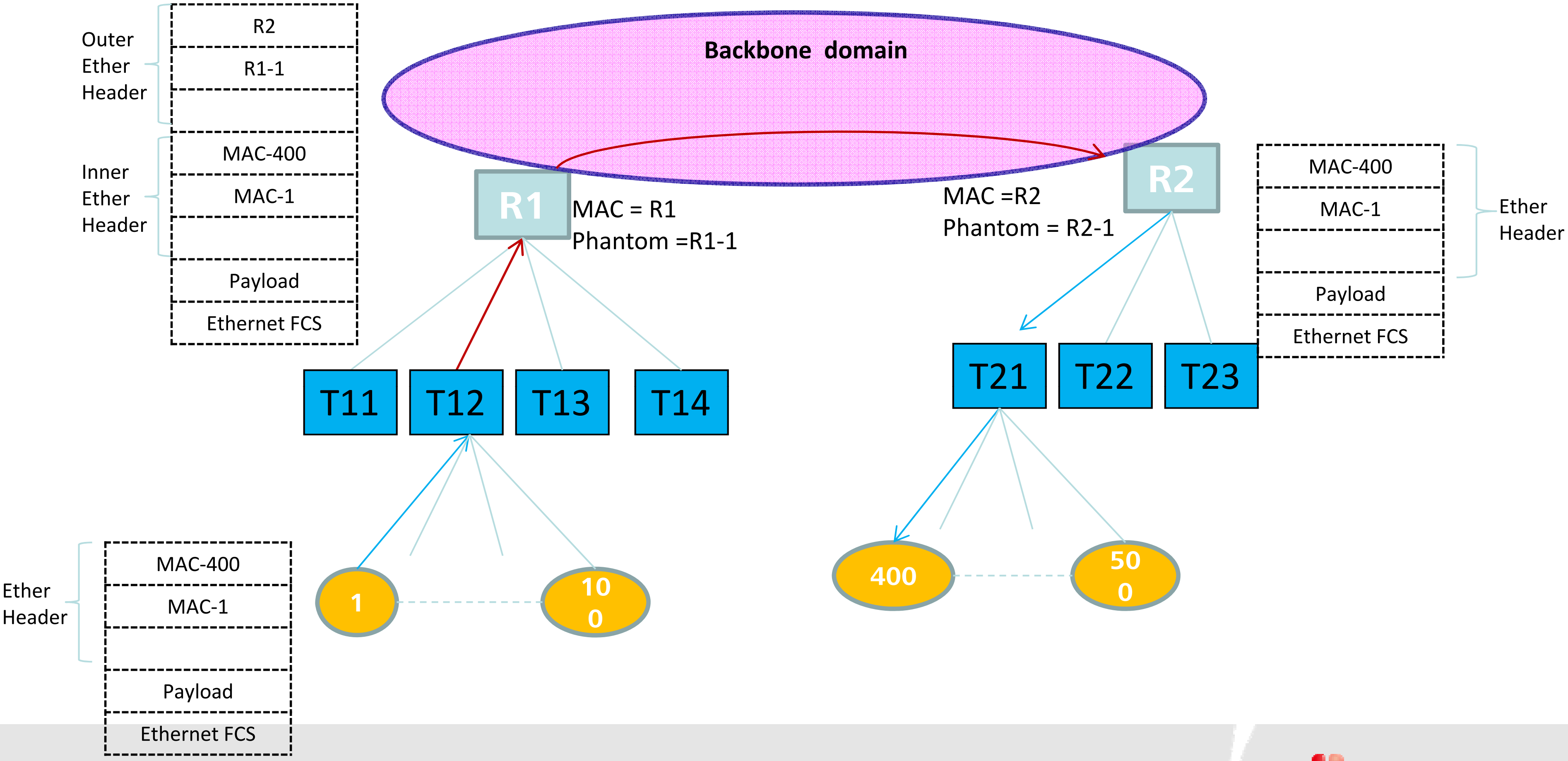
Phantom Backbone Source Address (Optional)

to avoid R1 receiving a data frame with its own address as SA.

Phantom B-Addr represents a group of non SPB nodes attached to the edge ports of a SPB edge node

which encapsulates Backbone MAC addresses.

- Phantom name is only useful when Edge node (e.g. R1) filter out data frames with its own address in the SA field.

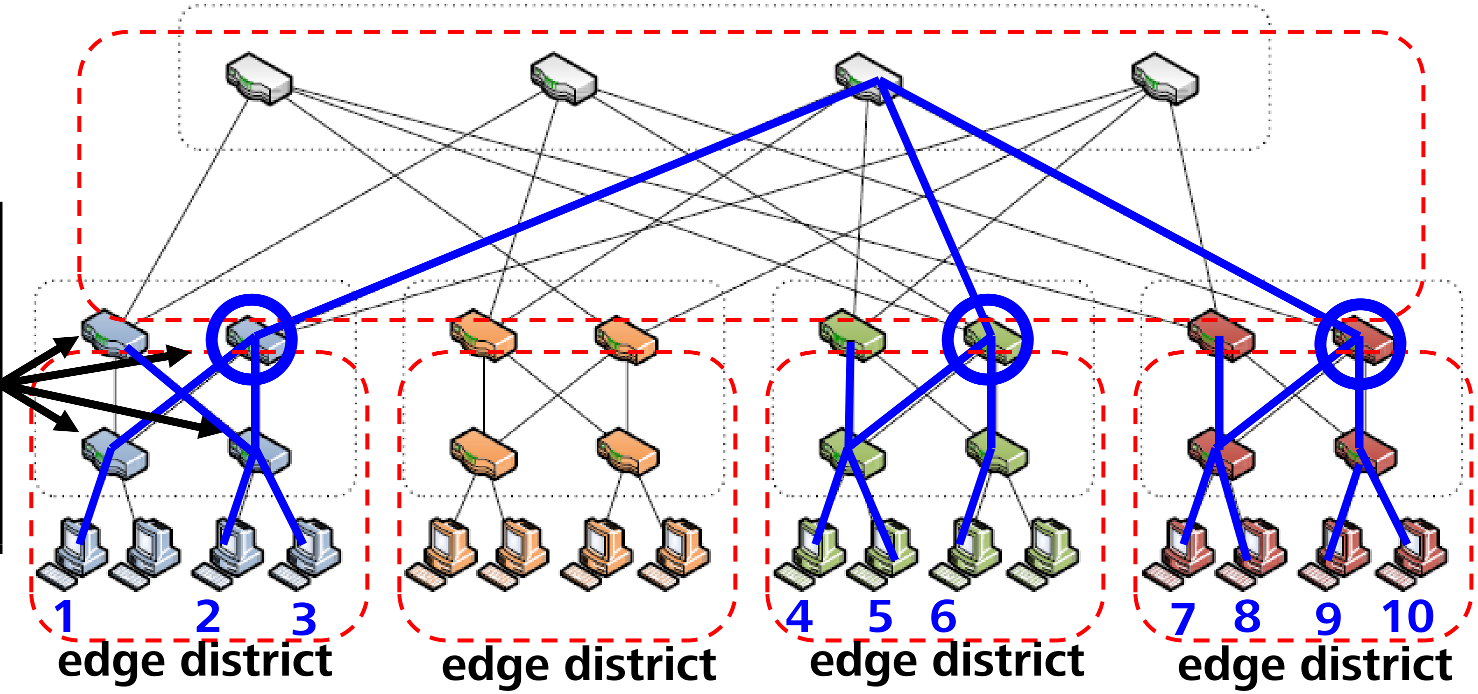


- Switch behavior

Dynamic information, e.g. FDB’s Port Map, is learned

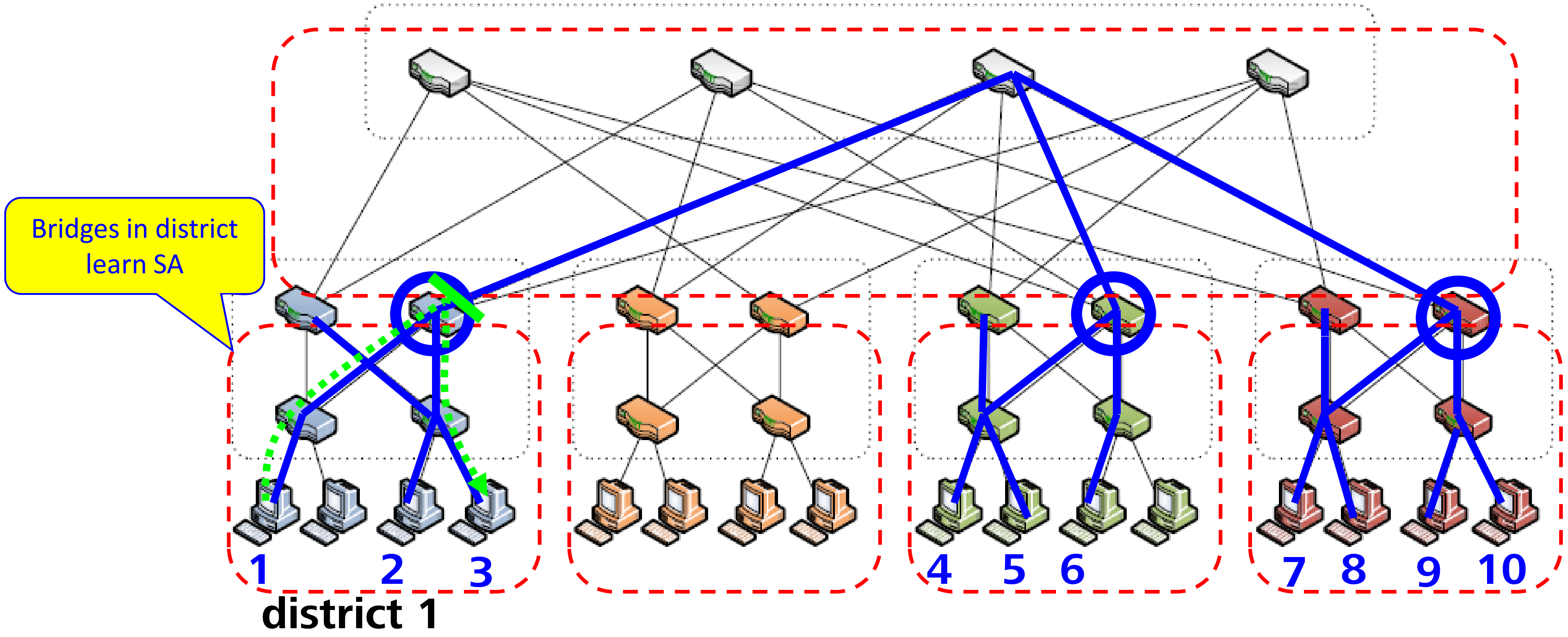
Partial FDB

DA	VID	Port Map
1	blue	NULL
2	blue	NULL
3	blue	NULL



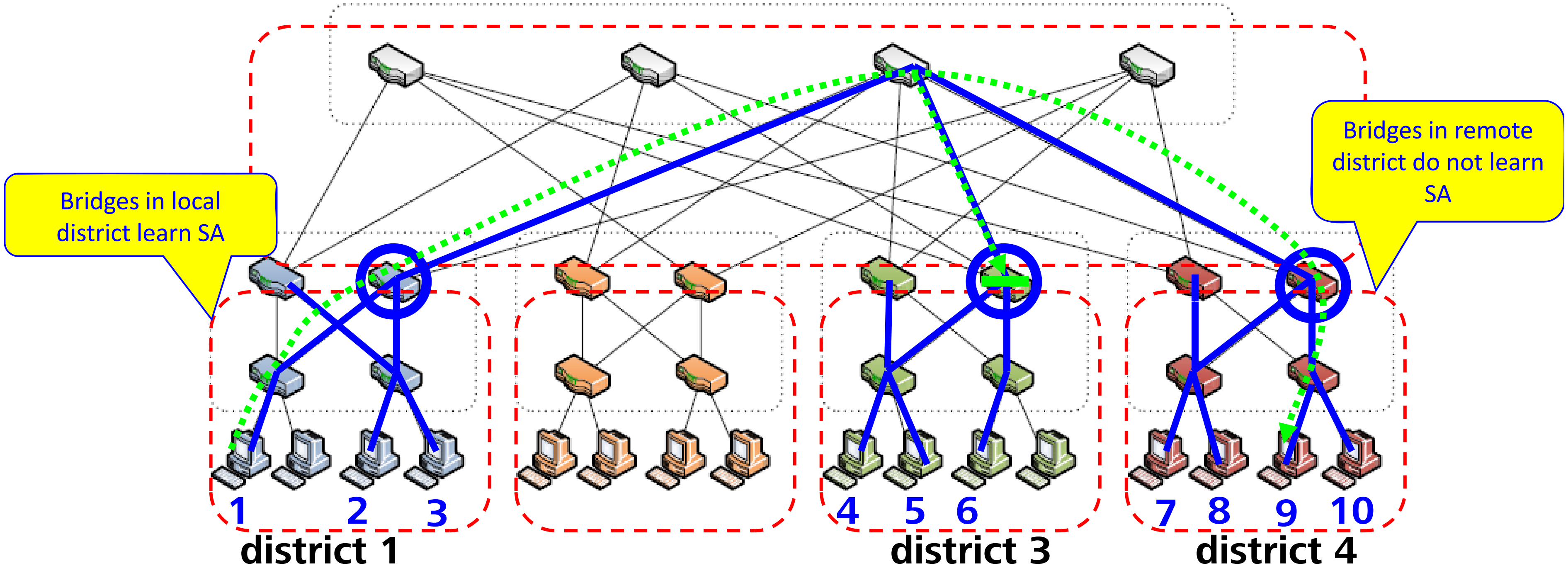
- **Each bridge has a partial FDB, with DA & VID for all hosts assigned to the Edge District (Rack or Row)**
  - **The partial FDB can be populated by the newly defined IEEE802.1Qbg’s VDP, or being pushed down from Directory Server.**
- **If the port for a <MAC, VID> has not been learned, the port value associated with that FDB entry is NULL;**
- **An address in a VLAN is determined to be local or remote depending on whether or not there is an FDB entry for that address.**

SA and DA are in same edge-district



- Host 1 sends frame to host 3;
- Bridge receiving frame in district 1 determines that DA *is* in district 1 by observing that an FDB entry for district 1 exists;
- Bridge floods the frame if outbound port is NULL;
  - or forwards the frame on specified outbound port;
- Bridge learns the SA (overwriting the NULL value if the SA was not previously learned);
- District 1 DBB recognizes that DA is in district 1 and does not allow the frame to flood into the core district;

SA and DA in different edge-districts


Bridges in local district learn SA
Bridges in remote district do not learn SA
1
2
3
4
5
6
7
8
9
10
district 1
district 3
district 4

- Host 1 sends frame to host 9;
- Bridges in district 1 do not have an FDB entry for host 9’s DA and forward frame to default port (towards SPB Gateway);
- SPB Edge node finds the outer MAC DA by MAC&VLAN <-> SPB Edge Mapping table, encapsulate the data frame, and forward it to the SPB core.
- The target SPB Edge will decapsulate the data frame and forward (or flood) the frame into its district since the DA is in its district;