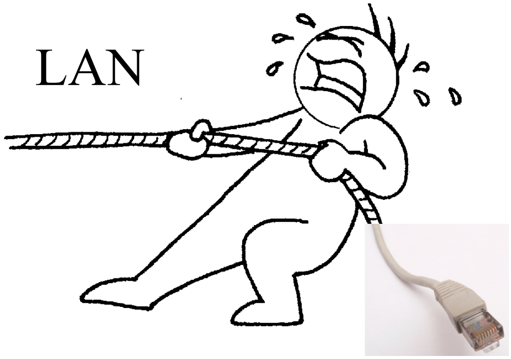
LAN Extension and Virtualization using Layer 3 Protocols



Raj Jain
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@cse.wustl.edu

These slides and audio/video recordings of this class lecture are at:

http://www.cse.wustl.edu/~jain/cse570-13/

Overview

1. Data Center Interconnection and LAN extension
2. TRILL
3. LISP

Note: Data Center partitioning techniques for multi-tenancy are discussed in another module that covers NVO3, VXLAN, NVGRE, and STT.

Network Virtualization Techniques

Entity	Partitioning	Aggregation/Extension/Interconnection**
NIC	SR-IOV	MR-IOV
Switch	VEB, VEPA	VSS, VBE, DVS, FEX
L2 Link	VLANs	LACP, Virtual PortChannels
L2 Network using L2	VLAN	PB (Q-in-Q), PBB (MAC-in-MAC), PBB-TE, Access-EPL, EVPL, EVP-Tree, EVPLAN
L2 Network using L3	NVO3, VXLAN, NVGRE, STT	MPLS, VPLS, A-VPLS, H-VPLS, PWoMPLS, PWoGRE, OTV, **TRILL, LISP**, L2TPv3, EVPN, PBB-EVPN
Router	VDCs, VRF	VRRP, HSRP
L3 Network using L1		GMPLS, SONET
L3 Network using L3*	MPLS, GRE, PW, IPSec	MPLS, T-MPLS, MPLS-TP, GRE, PW, IPSec
Application	ADCs	Load Balancers

*All L2/L3 technologies for L2 Network partitioning and aggregation can also be used for L3 network partitioning and aggregation, respectively, by simply putting L3 packets in L2 payloads.

**The aggregation technologies can also be seen as partitioning technologies from the provider point of view.

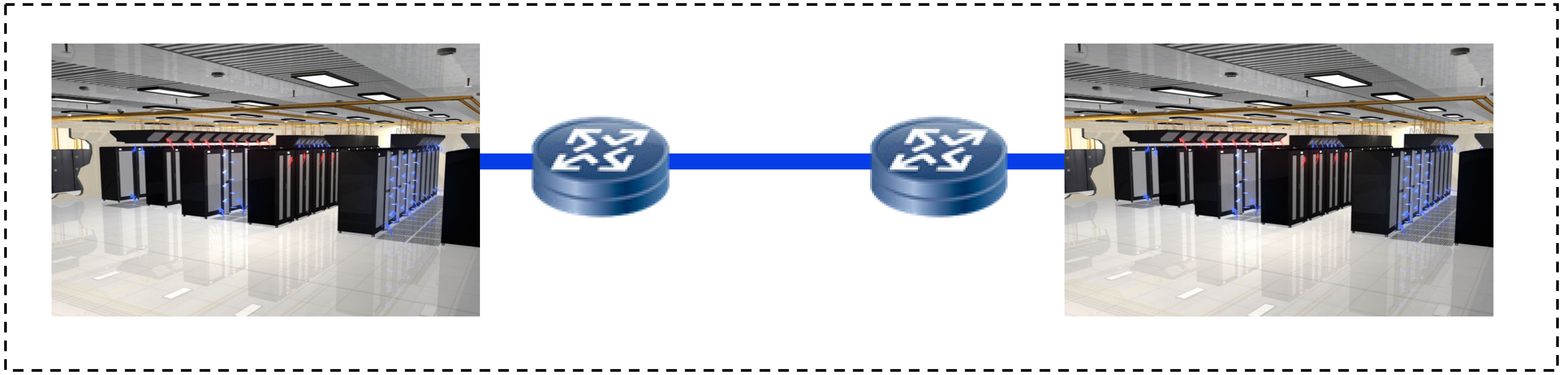
Geographic Clusters of Data Centers

- Multiple data centers are used to improve availability
- Cold-Standby: Data is backed up on tapes and stored off-site. In case of disaster, application and data are loaded in standby. Manual switchover ⇒ Significant downtime. (1970-1990)
- Hot-Standby: Two servers in different geographically close data centers exchange state and data continuously. Synchronous or Asynchronous data replication to standby. On a failure, the application automatically switches to standby. Automatic switchover ⇒ Reduced downtime (1990-2005) Only 50% of resources are used under normal operation.
- Active-Active: All resources are used. Virtual machines and data can be quickly moved between sites, when needed.

Ref: G. Santana, "Datacenter Virtualization Fundamentals," Cisco Press, 2014, ISBN: 1587143240

Data Center Interconnection (DCI)

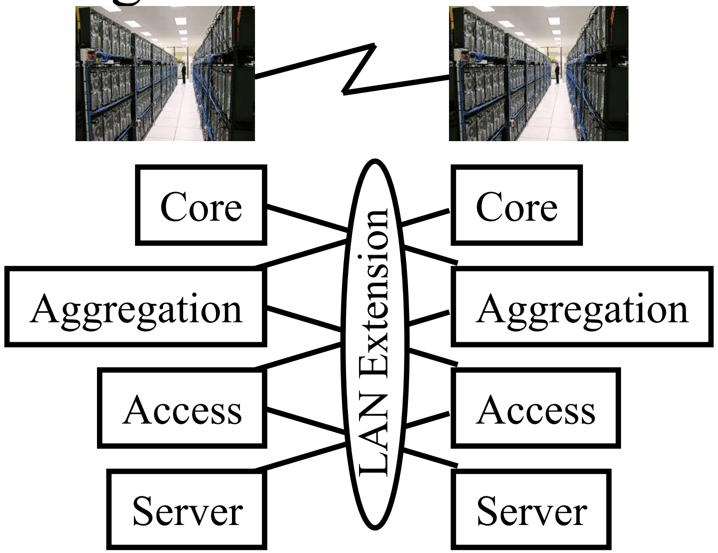
- Allows distant data centers to be connected in one L2 domain
  - Distributed applications
  - Disaster recovery
  - Maintenance/Migration
  - High-Availability
  - Consolidation
- Active and standby can share the same virtual IP for switchover.
- Multicast can be used to send state to multiple destinations.



Challenges of LAN Extension

- **Broadcast storms**: Unknown and broadcast frames may create excessive flood
- **Loops**: Easy to form loops in a large network.

- **STP Issues**:
  - High spanning tree diameter (leaf-to-leaf): More than 7.
  - Root can become bottleneck and a single point of failure
  - Multiple paths remain unused
- **Tromboning**: Dual attached servers and switches generate excessive cross traffic
- **Security**: Data on LAN extension must be encrypted



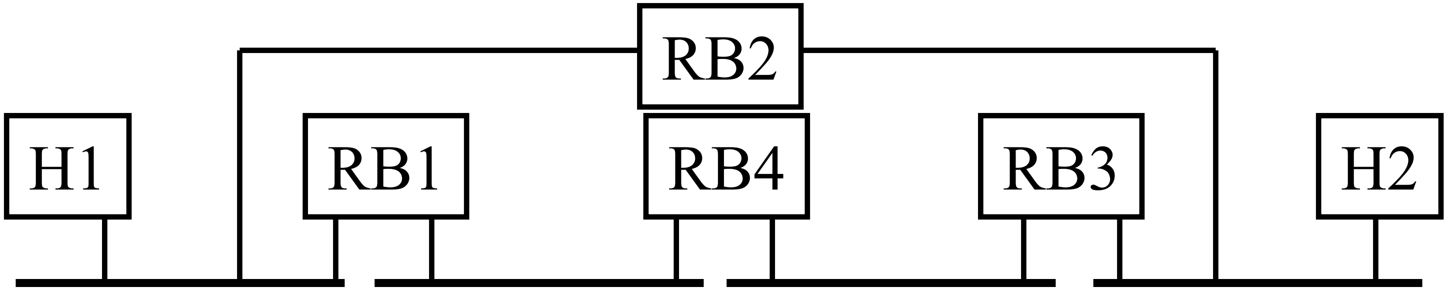
TRILL

- Transparent Interconnection of Lots of Links
- Allows a large campus to be a single extended LAN
- LANs allow free mobility inside the LAN but:
  - Inefficient paths using Spanning tree
  - Inefficient link utilization since many links are disabled
  - Inefficient link utilization since multipath is not allowed.
  - Unstable: small changes in network ⇒ large changes in spanning tree
- IP subnets are not good for mobility because IP addresses change as nodes move and break transport connections, but:
  - IP routing is efficient, optimal, and stable
- Solution: Take the best of both worlds ⇒ Use MAC addresses and IP routing

Ref: RFCs 5556, 6325, 6326, 6327, 6361, 6439

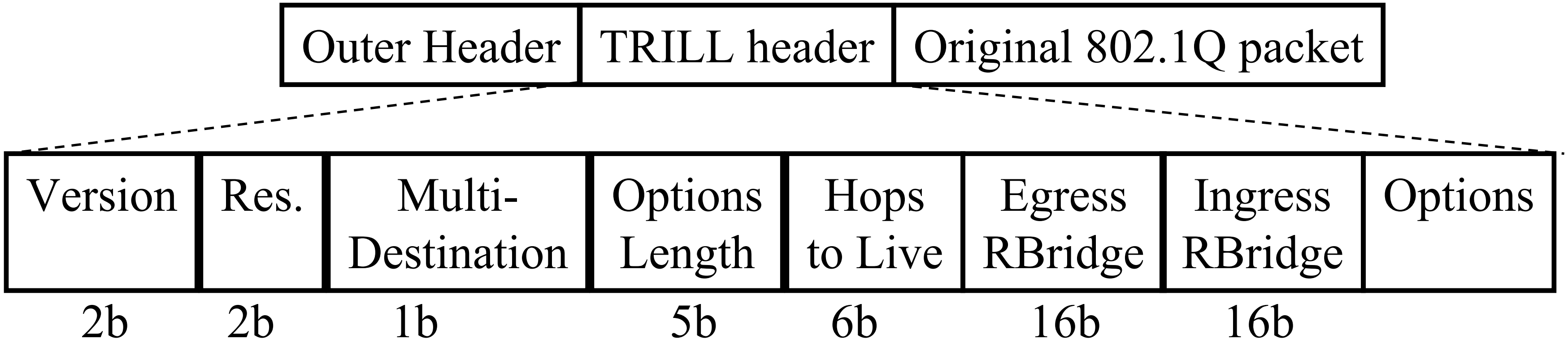
TRILL Architecture

- Routing Bridges (RBridges) encapsulate L2 frames and route them to destination RBridges which decapsulate and forward
- Header contains a hop-limit to avoid looping
- RBridges run IS-IS to compute pair-wise optimal paths for unicast and distribution trees for multicast
- RBridge learn MAC addresses by source learning and by exchanging their MAC tables with other RBridges
- Each VLAN on the link has one (and only one) designated RBridge using IS-IS election protocol



Ref: R. Perlman, "RBridges: Transparent Routing," Infocom 2004

TRILL Encapsulation Format



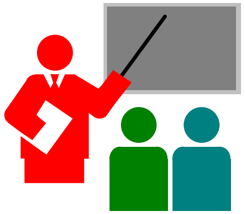
Outer Header	TRILL header	Original 802.1Q packet

Version	Res.	Multi-Destination	Options Length	Hops to Live	Egress RBridge	Ingress RBridge	Options
2b	2b	1b	5b	6b	16b	16b	

- ❑ For outer headers both PPP and Ethernet headers are allowed. PPP for long haul.
- ❑ Outer Ethernet header can have a VLAN ID corresponding to the VLAN used for TRILL.
- ❑ Priority bits in outer headers are copied from inner VLAN

TRILL Features

- Transparent: No change to capabilities.
  Broadcast, Unknown, Multicast (**BUM**) support. Auto-learning.
- Zero Configuration: RBridges discover their connectivity and learn MAC addresses automatically
- Hosts can be multi-homed
- VLANs are supported
- Optimized route
- No loops
- Legacy bridges with spanning tree in the same extended LAN



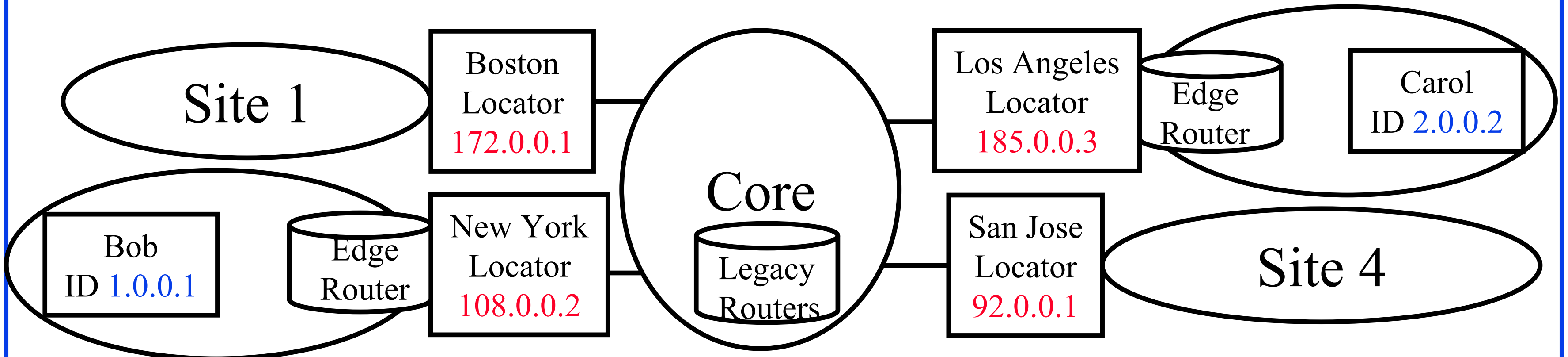
TRILL: Summary

- TRILL allows a large campus to be a single Extended LAN
- Packets are encapsulated and routed using IS-IS routing

Locator ID Separation Protocol (LISP)

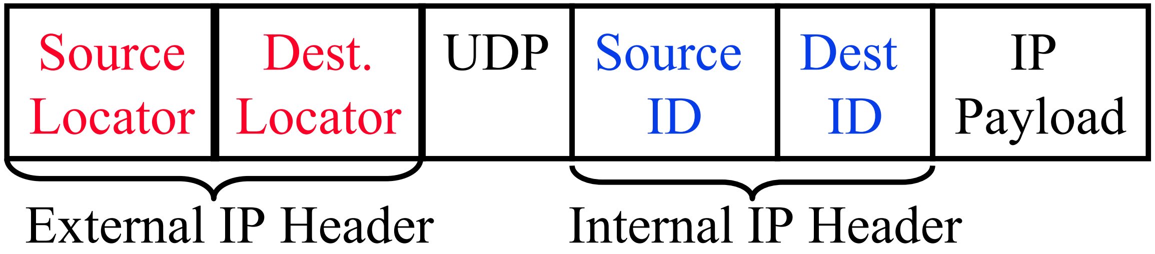
- Each host has an ID and a *locator*
  e.g., Raj Jain (EID) at WashU (RLOC)
- IPv6: $\underbrace{\text{2001:0034:0000:0000}}_{\text{Locator}}\text{:}\underbrace{\text{0001:0002:0003:0004}}_{\text{ID}}$

  IPv4: $\underbrace{\text{128.72.45.65}}_{\text{Locator}}\text{.}\underbrace{\text{192.168.0.1}}_{\text{ID}}$
- Inside a site, the routing is based on ID.
  Between sites, the routing is based on locators
- Edge routers encapsulate packets with locator on outer header.



LISP (Cont)

- IDs look like IP addresses ⇒ No changes to hosts
- Locators look like IP addresses ⇒ No changes to core routers between sites
- Changes are required only in routers at the edge of the sites.
- Trick: Edge routers use IP-in-IP tunneling to send packets between sites.
- A "*map server*" keeps track of ID to locator mapping



Source Locator	Dest. Locator	UDP	Source ID	Dest ID	IP Payload

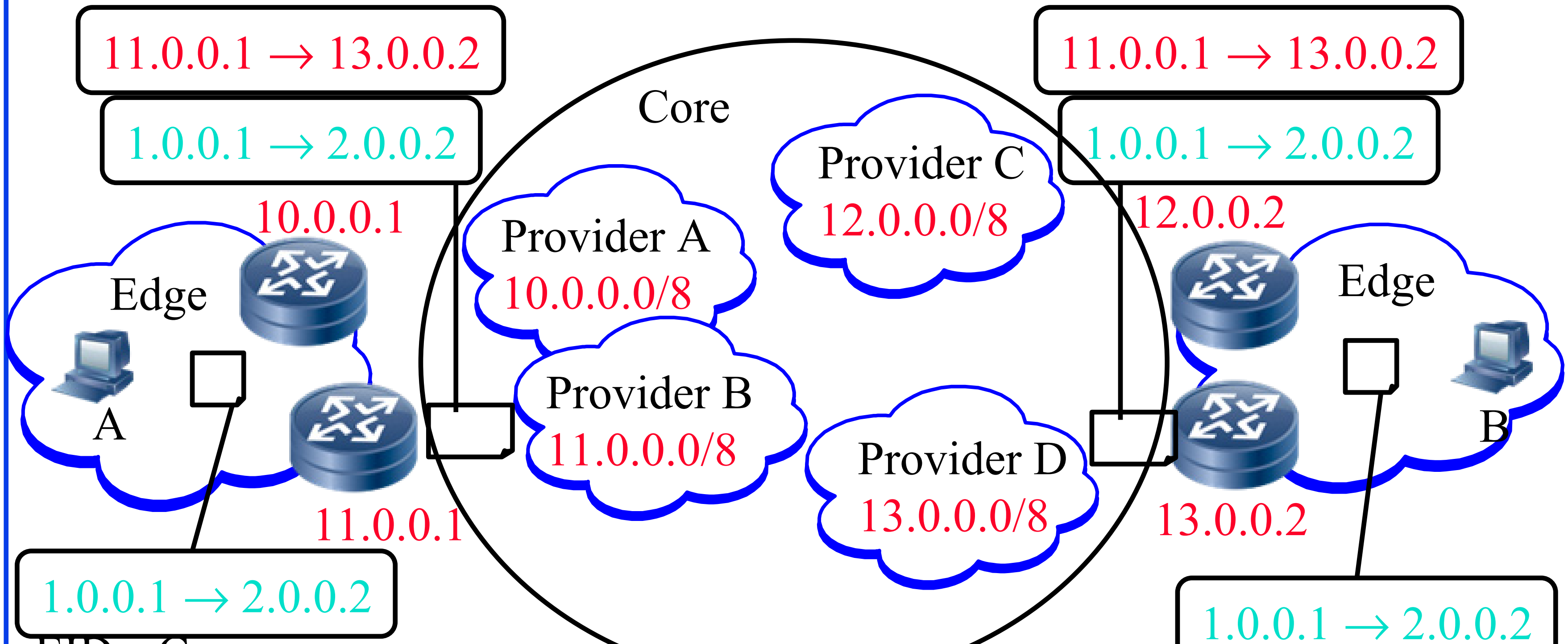
$\underbrace{\text{Source Locator} \quad \text{Dest. Locator}}_{\text{External IP Header}}$ UDP $\underbrace{\text{Source ID} \quad \text{Dest ID}}_{\text{Internal IP Header}}$ IP Payload

Ref: LISP – Routing in the Cloud, Sep 2012, http://lisp.cisco.com/LISP_Update.pdf

LISP Terminology

- Endpoint Identifier (EID): ID from different name space.
  Not routable on global Internet. Registered in DNS.
- Routing Locators (RLOC): Existing name space.
  Globally routable. Assigned to routers.
  Hosts do not know about them.
- Ingress Tunnel Router (ITR): Encapsulates and transmits
- Egress Tunnel Router (ETR): Receives and decapsulates
- xTR: Both ITR and ETR functions (common)
- Map-server: ETRs register their EID prefix-to-RLOC mappings
  Receives map requests via mapping system and forwards them
  to ETRs. ETR is "authoritative" for its EIDs.
- Map-Resolver: Receives map requests from ITR. Forwards
  them to mapping system.

LISP Example



EIDs: Green
Locators: Red

DNS: B→2.0.0.2

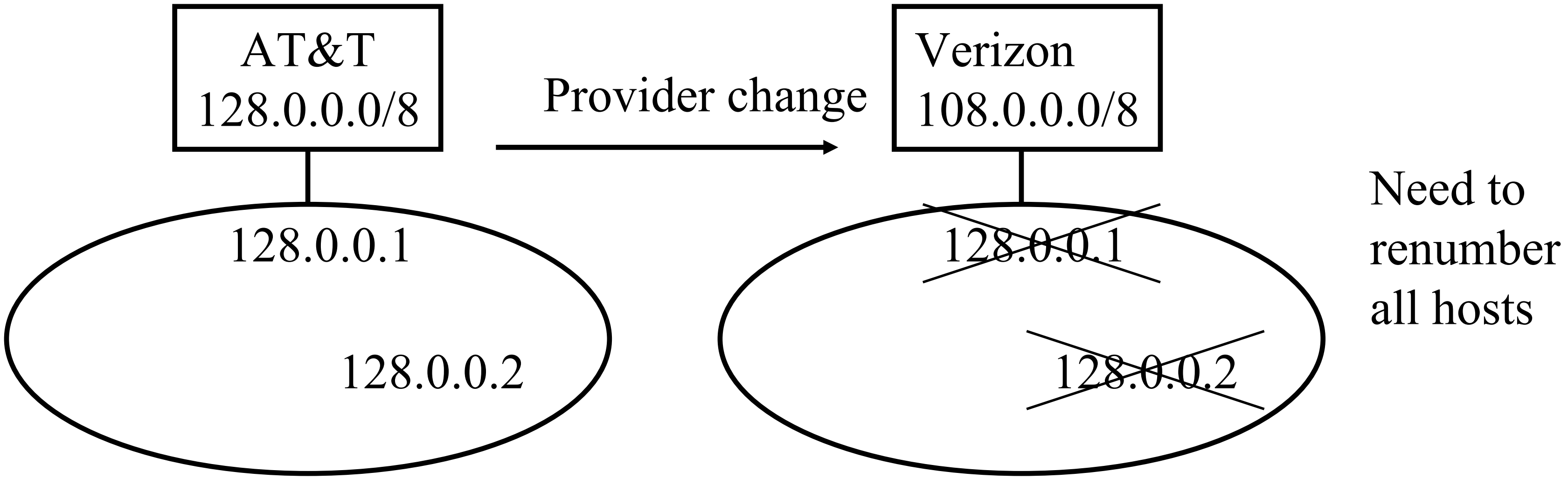
Map Server Entry:

EID-Prefix: 2.0.0.0/8
Locator Set:
12.0.0.2, priority 1, weight 50
13.0.0.2, priority 1, weight 50

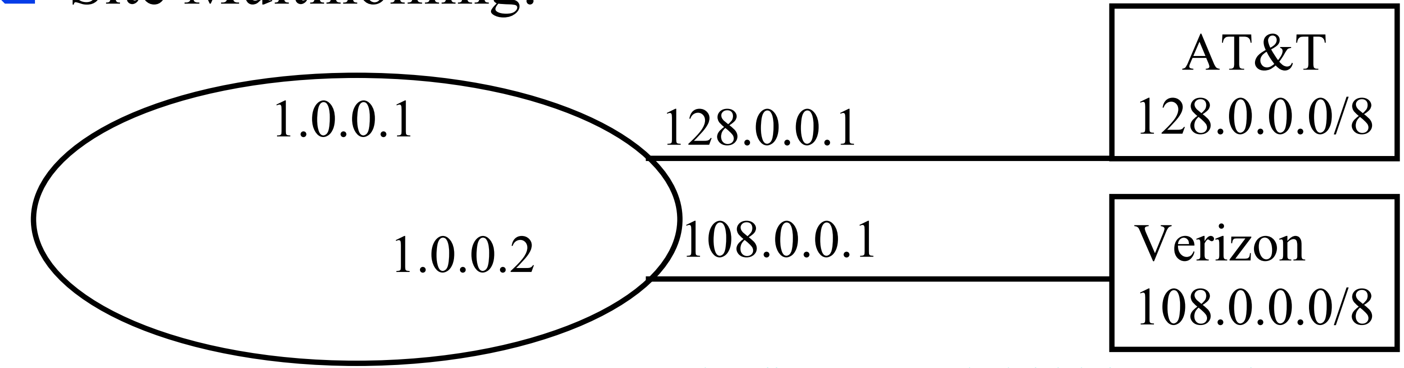
Ref: http://www.nanog.org/meetings/nanog41/presentations/lisp-nanog-abq.pdf

LISP Applications

- No renumbering if carrier changes

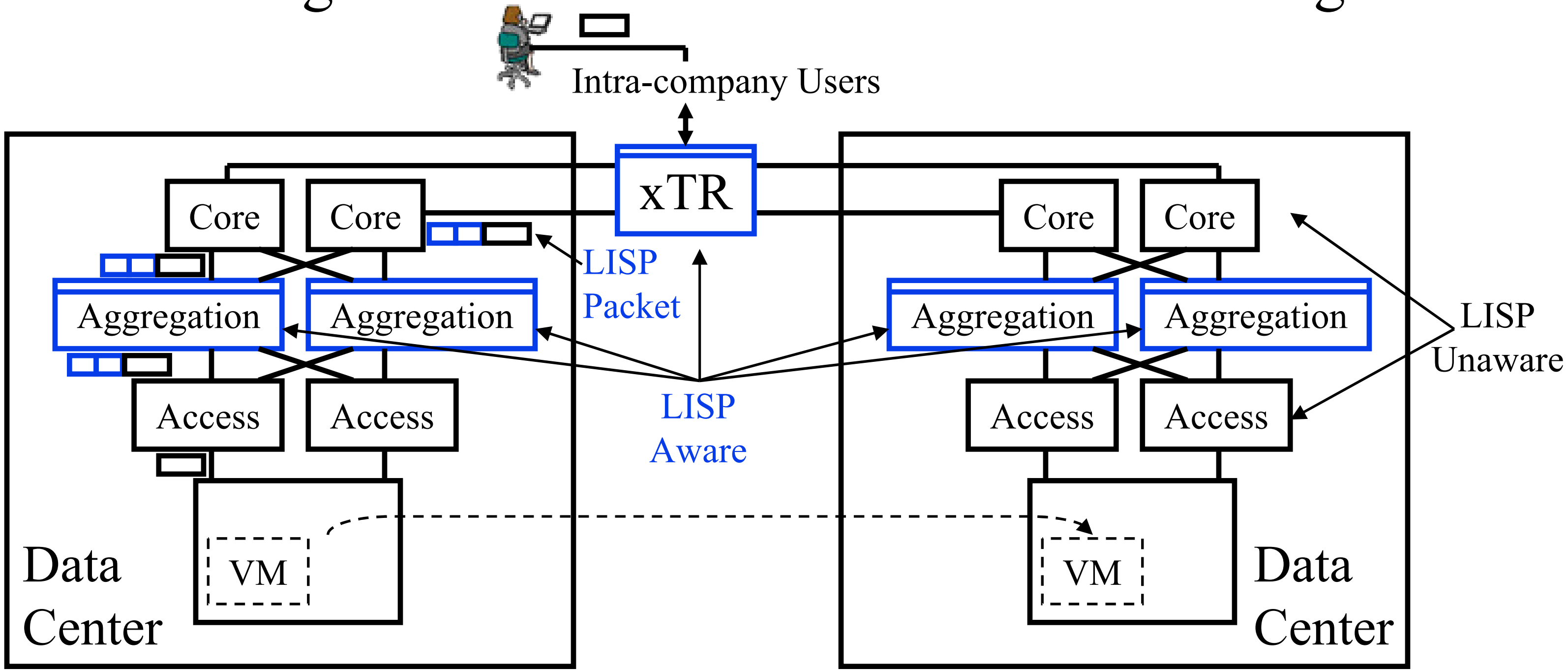


- Site Multihoming:



VM Migration Using LISP

- When an aggregator switch receives an IP packet from a VM, it notes its EID and registers its RLOC with map-server
- Map-server deletes the old entry (if any)
- Push or pull models for resolution
- Note: No triangulation after move. No IP address change.



Ref: G. Santana, "Datacenter Virtualization Fundamentals," Cisco Press, 2014, ISBN: 1587143240

LISP Summary

- Separates IDs from Locators
- Legacy IP needs locators ⇒ Use it on the outside
- Mobility requires IDs ⇒ Use it on the inside
- Uses IP-in-IP tunneling.

Summary

1. Ethernet is being extended to cover multiple data centers and large campuses. Networks are being "flattened" (L2 end-to-end)
2. Most of these efforts encapsulate Ethernet frames and transport them using layer 3 protocols
3. TRILL allows a single LAN to cover a large campus by using Rbridges that act as bridge for address learning and as router for forwarding. They exchange learnt MAC addresses using IS-IS.
4. LISP allows a network to span multiple sites. IDs are used inside while locators are used between sites. UDP encapsulation is used for inter-site communication.

Reading List

- Cisco, "Enhance Business Continuance with Application Mobility Across Data Centers," http://www.cisco.com/en/US/prod/collateral/switches/ps9441/ps9402/white_paper_c11-591960.pdf
- G. Santana, "Datacenter Virtualization Fundamentals," Cisco Press, 2014, ISBN: 1587143240 (Safari book)
- V. Fuller, et al., "LISP: A level of Indirection for Routing," http://www.nanog.org/meetings/nanog41/presentations/lisp-nanog-abq.pdf
- LISP - Routing in the Cloud, Sep 2012, http://lisp.cisco.com/LISP_Update.pdf
- R. Perlman, "RBridges: Transparent Routing," Infocom 2004
- V. Josyula, M. Orr, and G. Page, "Cloud Computing: Automating the Virtualized Data Center," Cisco Press, 2012, 392 pp., ISBN: 1587204347 (Safari book)

Wikipedia Links

- http://en.wikipedia.org/wiki/TRILL_(computing)
- http://en.wikipedia.org/wiki/Locator/Identifier_Separation_Protocol

Acronyms

A-VPLS	Advanced Virtual Private LAN Service
ASM	Across Subnet Mode
BFD	Bidirectional Forwarding Detection
BGP	Border Gateway Protocol
BUM	Broadcast, Unicast, Multicast
CRC	Cyclic Redundancy Check
DCI	Data Center Interconnection
DNS	Domain Name System
DWDM	Dense Wavelength Division Multiplexing
EID	Endpoint Identifier
EoMPLS	Ethernet over MPLS
EoMPLSoGRE	Ethernet over MPLS over GRE
ESM	Extended Subnet Mode
ETR	Egress Tunnel Router
EVPN	Ethernet Virtual Private Network
GRE	Generic Routing Encapsulation

Acronyms (Cont)

H-VPLS	Hierarchical Virtual Private LAN Service
ID	Identifier
IP	Internet Protocol
IPv4	Internet Protocol version 4
IPv6	Internet Protocol version 6
IS-IS	Intermediate System to Intermediate System
ITR	Ingress Tunnel Router
LAN	Local Area Network
LISP	Locator ID Separation Protocol
MAC	Media Access Control
MPLS	Multiprotocol Label Switching
NVGRE	Network Virtualization Using GRE
NVO3	Network Virtualization using L3
OAM	Operations, Administration, and Maintenance
OTV	Overlay Transport Virtualization
PB	Provider bridging

Acronyms (Cont)

PBB	Provider Backbone Briding
PPP	Point to Point Protocol
RBridge	Routing Bridges
RFC	Request for Comments
RLOC	Routing Locators
STP	Spanning Tree Protocol
STT	Stateless Transport Tunneling
TE	Traffic Engineering
TR	Tunnel Router
TRILL	Transparent Interconnection of Lots of Link
UDP	User Datagram Protocol
VLAN	Virtual Local Area Network
VM	Virtual Machine
vPC	Virtual PortChannel
VPLS	Virtual Private LAN Service
VPLSoGRE	VPLS over GRE

Acronyms (Cont)

❑ VPN	Virtual Private Network
❑ VSS	Virtual Switching System
❑ VXLAN	Virtual Extensible Local Area Network
❑ xTR	Ingress/Egress Tunnel Router

References

- "TRILL: Problem and Applicability Statement," RFC 5556, May 2009, https://datatracker.ietf.org/doc/rfc5556/
- "RBridges: Base Protocol Specification," RFC 6325, Jul 2011, https://datatracker.ietf.org/doc/rfc6325/
- "RBridges: Adjacency," RFC 6327, July 2011, https://datatracker.ietf.org/doc/rfc6327/
- "PPP TRILL Protocol Control Protocol," RFC 6361, Nov 2011, https://datatracker.ietf.org/doc/rfc6361/
- " RBridges: Appointed Forwarders," RFC 6439, Nov 2011, https://datatracker.ietf.org/doc/rfc6439/
- "Definitions of Managed Objects for RBridges," RFC 6850, Jan 2013, https://datatracker.ietf.org/doc/rfc6850/
- "Requirements for OAM in TRILL," RFC 6905, Mar 2013, https://datatracker.ietf.org/doc/rfc6905/