TRILL and VXLAN

Emerging fundamental protocols
for virtual networking

Presented by *Baohua Yang*

December 28, 2011

Content

- TRILL
  - From STP, to scale the L2 networks
- VXLAN
  - From VLAN, to decouple the virtual and physical networks

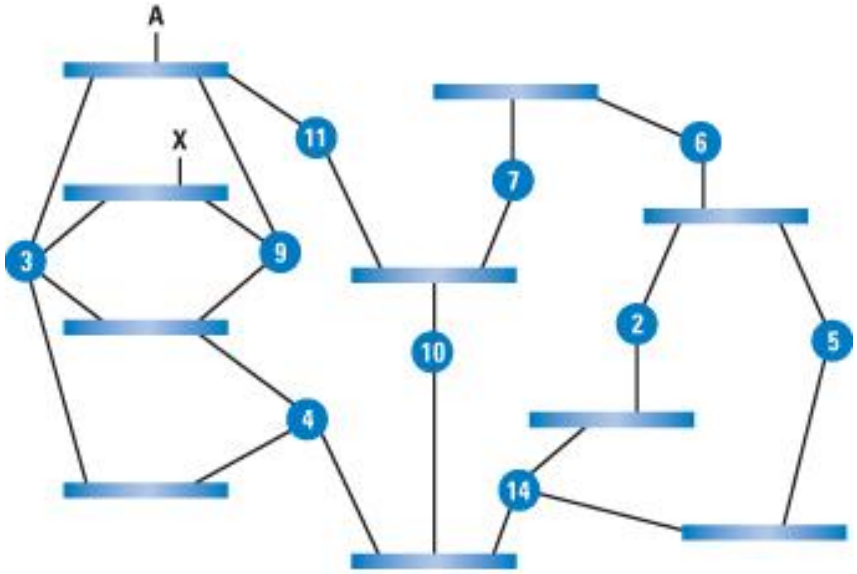
TRILL

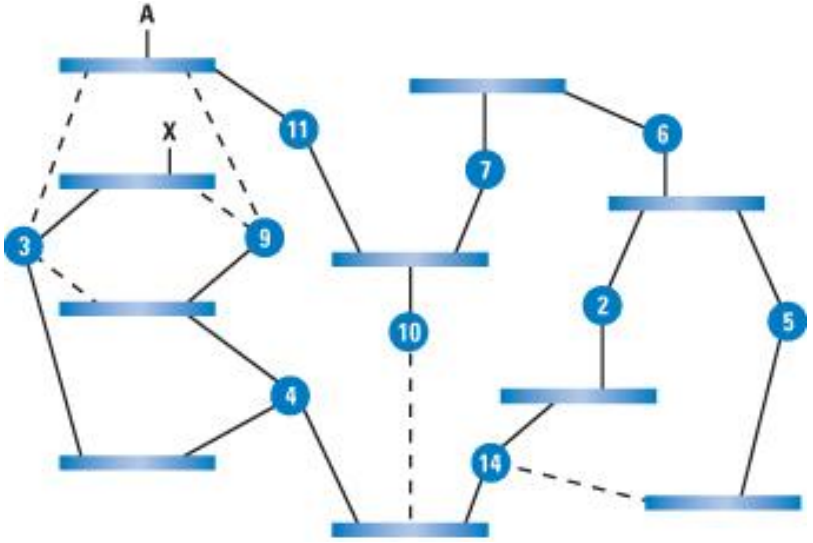
- Transparent Interconnection of Lots of Links (TRILL):Problem and Applicability Statement
  - RFC 5556, 2009.5
- Routing Bridges (RBridges): Base Protocol Specification
  - RFC 6325, 2011.7
- Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS
  - RFC 6326, 2011.7
- Routing Bridges (RBridges): Adjacency
  - RFC 6327, 2011.7
- PPP Transparent Interconnection of Lots of Links (TRILL) Protocol Control Protocol
  - RFC 6361, 2011.8
- Routing Bridges (RBridges): Appointed Forwarders
  - RFC 6439, 2011.11
- Core technique in Cisco FabricPath: An improved TRILL

L2/Ethernet Networks

- Relocate without requiring renumbering
- Automatic configuration
- The basis is the Spanning Tree Protocol
  - Automatically find a tree (loop-free)
  - Bandwidth across the subnet is limited (only single path)
  - Re-calculation/Converge slowly

L2/Ethernet Networks





L3/IP Networks

- Route to links, not nodes (IP address clusters)
- Use link-state routing protocols
- Higher capacity and more resistance to link failures
- Complicated, require renumbering when relocating, cause network interruption

TRILL

- Combines the features of these two existing solutions (multi-path, simplicity, etc.)
- Capable of using network-style routing, while still providing Ethernet service
- Allows reuse of well-understood network Routing protocols to benefit the link layer
- Layer 2.5

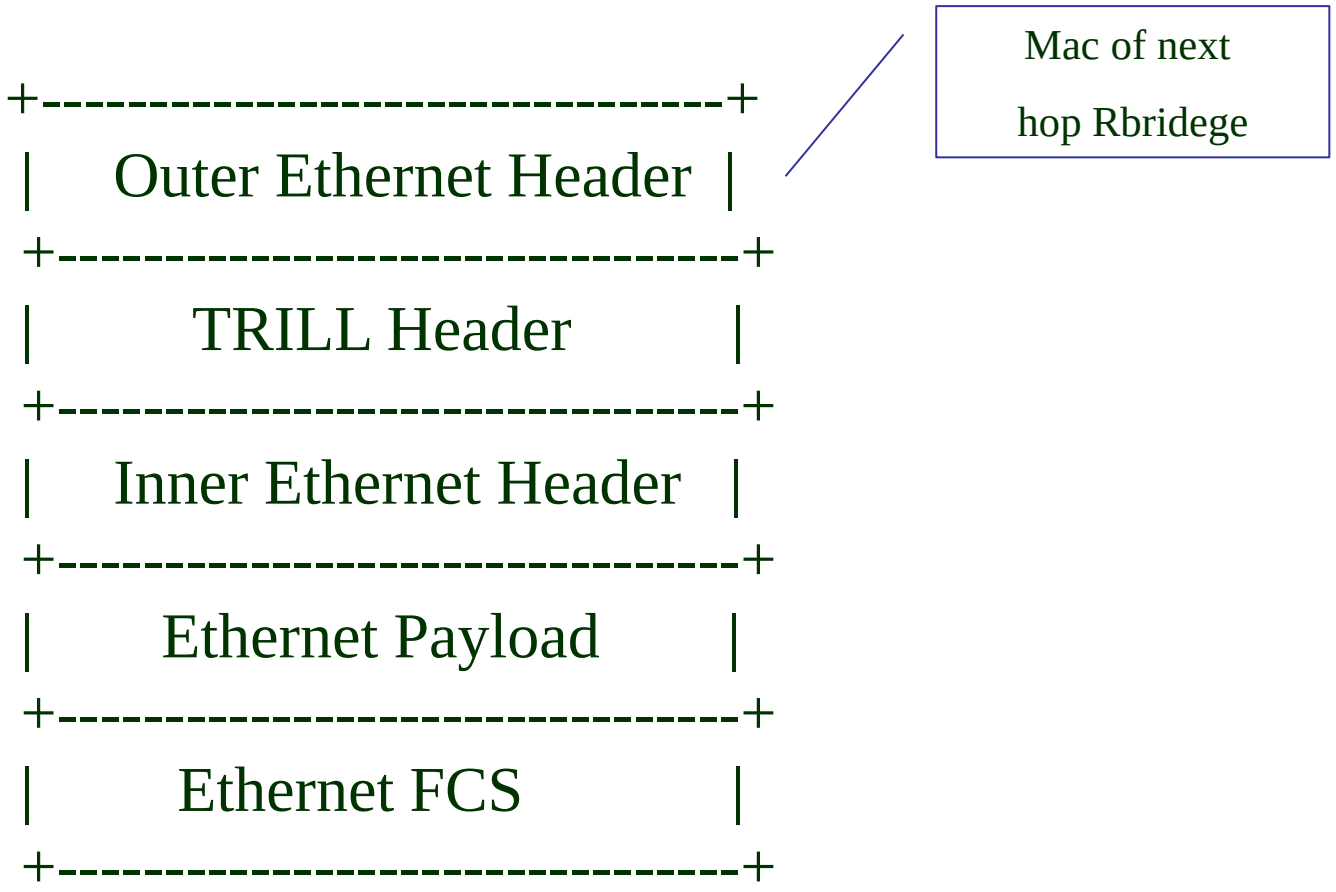
TRILL properties – rfc5556

- No change to link capabilities
  - Unicast, broadcast, multicast, auto-learning
- Zero Configuration and Zero Assumption
  - Bridges, hubs
- Forwarding Loop Mitigation
- Spanning Tree Management
- Multiple Attachments
- VLAN Issues
- Operational Equivalence
- Optimizations
- Internet Architecture Issues (routing, etc)

Rbridge – rfc6325

- Forward based on a header with a hop count
- The src-Rbridge encapsulate the frames with a TRILL header, specifying the dst-Rbridge
- The nickname of Rbridge are selected with 2-octet, by a dynamic protocol
- The dst-Rbridge removes the encapsulated header, and fwd the native frames
- A Designated Rbridge (DRB) is selected by link protocol

Packet Format

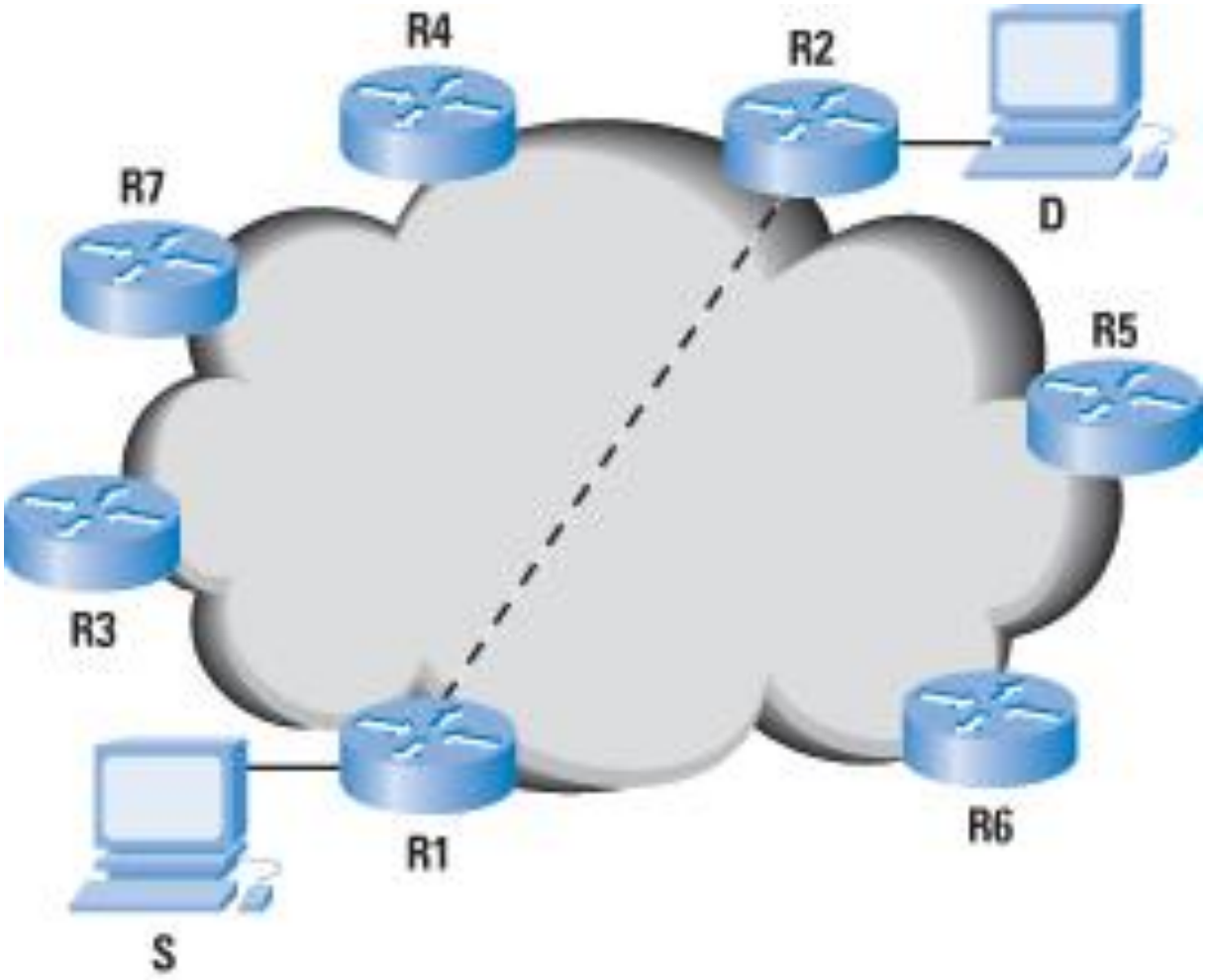


TRILL header

	V, R, M, OP-Length, Hop count
Egress RB Nickname	Ingress RB Nickname
Options…	

V: version, 2 bits
R: reserved, 2 bits
M: Multi-destination, 1 bit
Op-Length: option length, 5 bits
Hop count: 6-bit
Egress/Ingress RB Nickname: 16-bit
Options: present if OP-Length is non-zero

TRILL example



TRILL routing

- An extension of IS-IS
  - It runs directly over Layer 2, so therefore it may be run without configuration (no IP addresses need to be assigned).
  - It is easy to extend by defining new TLV (type-length-value) data elements and sub-elements for carrying TRILL information.
- TRILL-IS-IS frames

Summary

- Creates a cloud with a flat Ethernet address
- Can use all the Layer 3 techniques, including shortest paths, Equal Cost Multipath (ECMP), and traffic engineering
- Supports VLANs and multicast
- Compatible with existing Ethernet bridges (switches)

Algorhyme - Radia Perlman

- I think that I shall never see
- A graph more lovely than a tree.

- A tree whose crucial property
- Is loop-free connectivity.

- A tree that must be sure to span
- So packets can reach every LAN.

- First, the root must be selected.
- By ID, it is elected.

- Least-cost paths from root are traced.
- In the tree, these paths are placed.

- A mesh is made by folks like me,
- Then bridges find a spanning tree.

Algorhyme v2

- I hope that we shall one day see
- A graph more lovely than a tree.

- A graph to boost efficiency
- While still configuration-free.

- A network where RBridges can
- Route packets to their target LAN.

- The paths they find, to our elation,
- Are least cost paths to destination!

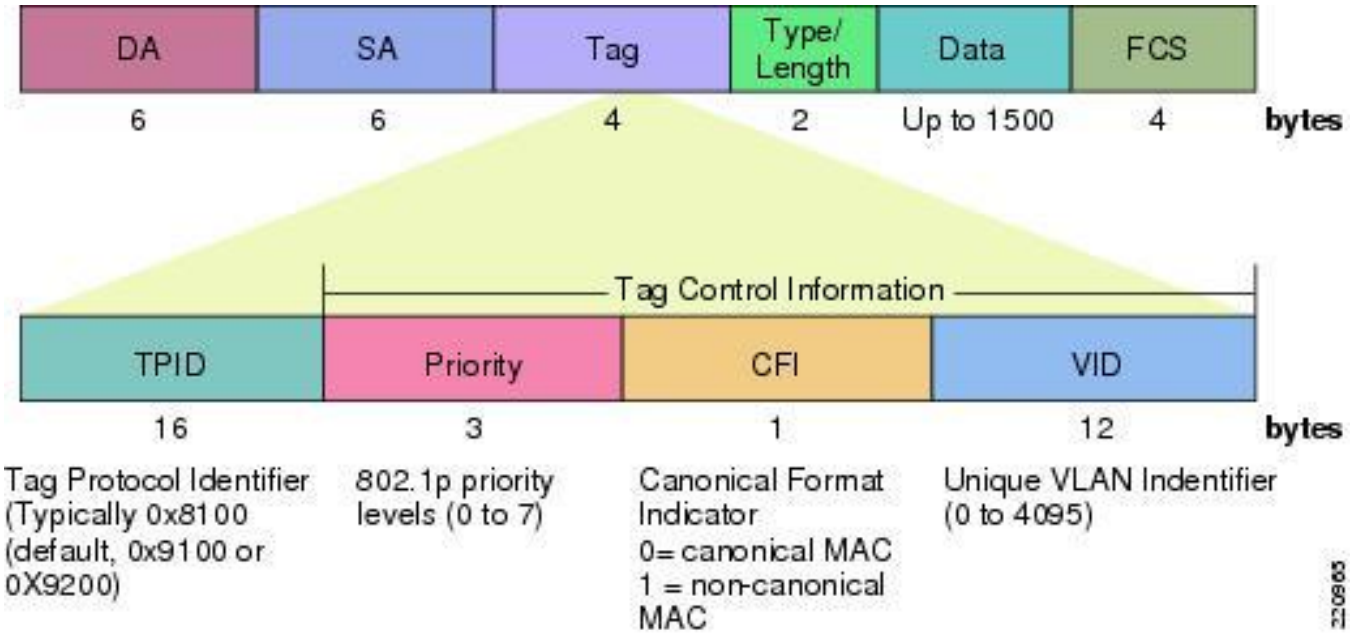
- With packet hop counts we now see,
- The network need not be loop-free!

- RBridges work transparently,
- Without a common spanning tree.

VXLAN: Decouple virtual networking from the physical world!

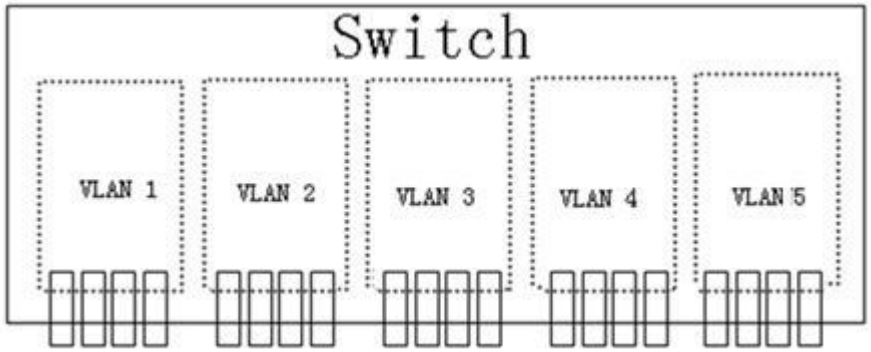
Vlan Introduction

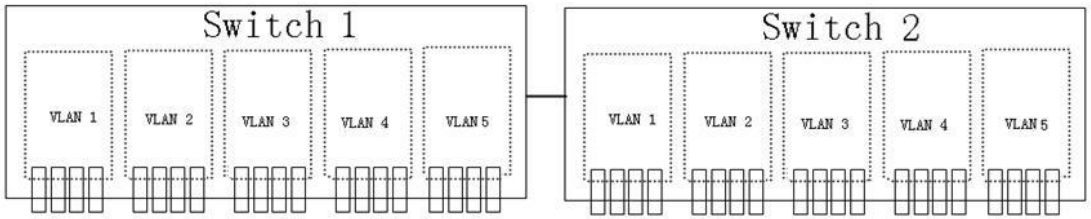
- IETF 1999, 802.1Q
  - Map the physical broadcast domains (LAN) into virtual ones.
  - Add Vlan header into the Ethernet header



Vlan

- Least scalable
- Simplest VN





Existing solutions of virtualization

- VMware vSwitch
  - Based on vlan
- vCDNI
  - MAC-in-MAC encapsulation
  - Lots of flooding
- PBB/VPLS

VXLAN Draft

- VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks
  - 2011.8.26
  - Internet-Draft, by IETF, 6 months

Virtual eXtensible LAN

- Address the need for overlay networks within virtualized data centers accommodating multiple tenants
  - cloud service provider
  - enterprise data center networks

Requirement

- Current vlan limit of 4094 is inadequate
  - Too many tenants
  - Over 100,000 VMs
- STP is difficult to work (shutdown some links for loop prevention)
- Each tenant may independently assign MAC addresses and VLAN IDs
- Arp table pressure for TOR
- IP for infrastructure interconnection, while L2 for inter-vm communication

Why IP doesn’t work

- Two tenants might use the same set of Layer 3 addresses within their networks
- Customers relying on direct Layer 2 or non-IP Layer 3 protocols for inter-VM communication

The answer is an overlay network.

Carry the MAC traffic from the individual VMs, in an encapsulated format over a logical "tunnel“.

Basic Idea

- Layer 2 overlay scheme over a Layer 3 network
  - Each overlay is termed a VXLAN segment
  - Only VMs within the same segment can communicate with each other
  - 24 bit segment ID (16,777,216 VNI)
  - VNI/VXLAN are known only to the end point of the tunnel (VTEP)
- Also a tunneling (encapsulation) scheme

VXLAN packet

- Outer Header
  - Ethernet + IP + UDP
- VXLAN Header
- Inner Ethernet Header
- Payload

VXLAN header

- Flags (8 bits)
  - Bit 4 should be set to 1, others are 0.
- VNI (24 bits)
- Reserved fields (24+8 bits)
  - Set to 0

Unicast VM-VM

- src-vm
  - Unaware of VXLAN
  - Send mac frame to the dst-vm
  - VTEP check the VNI of the src-vm and the dst-vm, to see if on the same segment
  - If on the same segment, add an outer header (outer mac, outer IP, VXLAN header), send to the remote VTEP

Unicast VM-VM cont.

- dst-vm
  - Remote VTEP receives the pkt and check its VNI if valid.
  - If so, strip the outer header and remember the inner-mac to outer src IP mapping
  - Dst-vm receives the pkt.

Broadcast and Multicast

- The src-vm attemps to communicate with the dst-vm with IP
  - Map VNI and the IP multicast group, by the management layer
  - The broadcast pkt is sent to the multicast group, requiring multicast routing protocol (PIM-SM)
  - The dst-vm sends an arp-response using unicast. The frame will be encapsulated back to the source VTEP.

Discussion

- Virtual L2 networks
  - Over UDP/GRE
  - MAC-to-VTEP mapping problem: NVP, VXLAN
- VXLAN
  - Control plane?
  - Scalability?

Thanks!
Q&A