TRILL History

Transparent Interconnection of Lots of Links

Donald E. Eastlake 3rd

Principal Engineer, Huawei

d3e3e3@gmail.com

What is TRILL?

- Basically a simple idea:
  - Encapsulate native frames in a transport header providing a hop count
  - Route the encapsulated frames using IS-IS
  - Decapsulate native frames before delivery

- Provides
  - Least cost paths with zero/minimal configuration
  - Equal Cost Multi-Pathing of unicast
  - Multi-pathing of multi-destination traffic

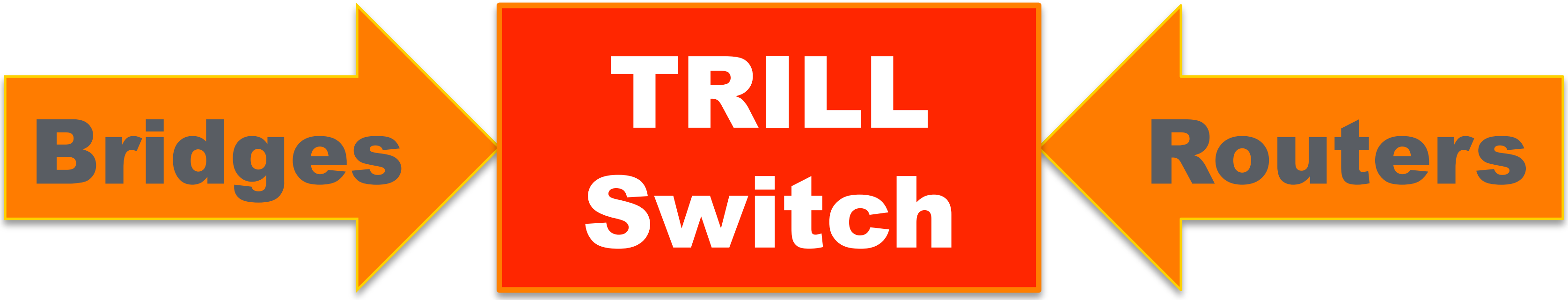
Original Bridging Constraints

- When transparent bridging was initially designed, it was subject to severe constraints:
  - No modifications could be made to the frames. There was not a single spare bit that could safely be used.
  - There was a hard limit to the size. Thus, encapsulation would not have been possible.
  - End nodes must not have to do anything differently than they would on a single CSMA/CD LAN.
- Spanning Tree / Bridging was a brilliant solution, given these constraints, but it is brittle.
  - Failure to properly process control messages leads to melt down.

Routing versus Bridging

- Routing only sends data out a port when it receives control messages on that port indicating this is safe and routing has a TTL for safety.
  - If control messages are not received or not processed, it "fails safe" and does not forward data.
- Bridging (Spanning Tree Protocol) forwards data out all ports (except the one where the data was received) unless it receives control messages on that port indicate this is unsafe. There is no TTL.
  - If control messages are not received or not processed, it "fails unsafe", forwards data, and can melt down due to data loops.

TRILL Features


Bridges

TRILL Switch

Routers

- Transparency
- Plug & Play
- Virtual LANs
  - Multi-tenant support
- Frame Priorities
- Data Center Bridging
- Virtualization Support

- Multi-pathing
- Optimal Paths
- Rapid Fail Over
- The safety of a TTL
  - Implemented in data plane
- Multi-Topology
- Extensions

More TRILL Features

- Breaks up and minimizes spanning tree for greater stability.
- Unicast forwarding tables at transit RBridges scale with the number of RBridges, not the number of end stations.
- Transit RBridges do not learn end station addresses.
- Compatible with existing IP Routers. TRILL switches are as transparent to IP routers as bridges are.
- Support for VLANs, frame priorities, and 24-bit data labels (“16 million VLANs”).

More TRILL Features

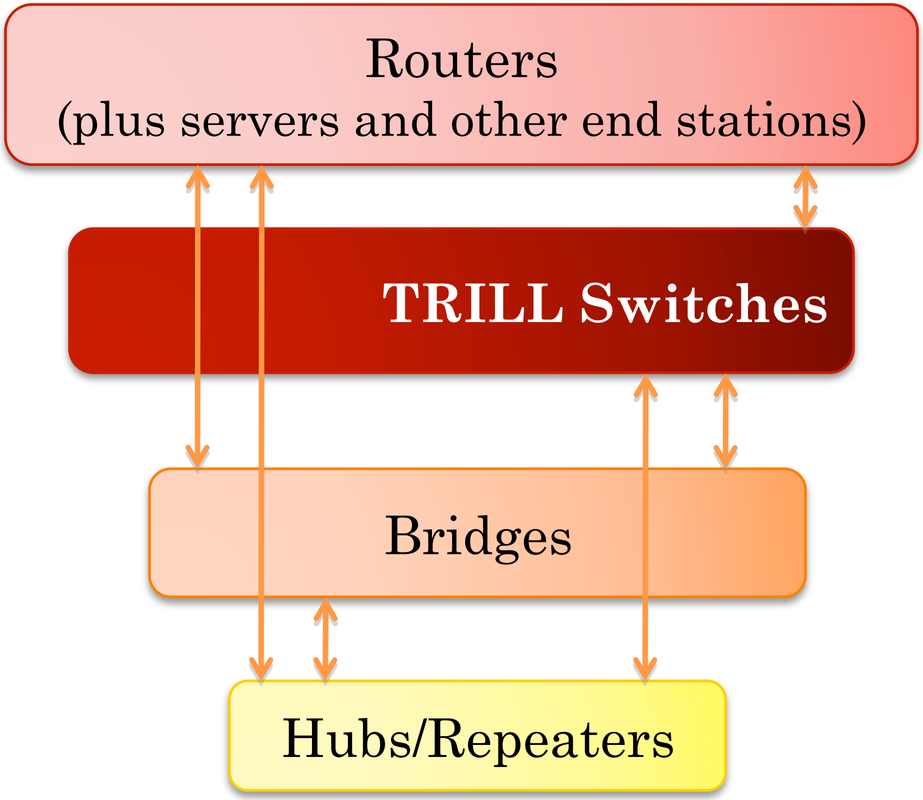
- MTU feature and jumbo frame support including jumbo routing frames.
- Active-active connection at the edge
- Standardized push and pull directory features
- Multi-topology support
- Has a poem.
  - The only other bridging or routing protocol with a poem is Spanning Tree (see Algorhyme).

Algorhyme V2 (TRILL and RBridges)

- 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.
- - By Ray Perlner
  (Radia Perlman’s son)

Peering: Are TRILL Switches Bridges or Routers?

- Really, they are a new species, between bridges and routers:



Inspired by a Real Life Incident

- In November 2002, Beth Israel Deaconess Hospital in Boston, Massachusetts, had a total network meltdown:
  - Their network took four days of heroic efforts to be restored to an operational state! In the mean time the staff was reduced to using paper and pencil.
  - Beth Israel Deaconess had grown by acquiring various clinics and just plugged all those bridged networks together.
  - The article in Boston's primary newspaper specifically mentioned "Spanning Tree Protocol" as the problem!
  - Radia Perlman, who invented spanning tree over 25 years ago, decided it was time to come up with a better way.

TRILL History up to 2004

- 1964: Packet switching/routing invented by Paul Baran.
- 1973: Ethernet invented by Robert Metcalfe
- 1979: Link State Routing invented by John McQuillan.
- 1985: Radia Perlman invents the Spanning Tree Protocol.
- 1987: DECnet Phase V / IS-IS designed by Radia Perlman.

- **2002: Beth Israel Deaconess Hospital network in Boston melts down due to deficiencies in the Spanning Tree Protocol.**

- 2004: TRILL/RBridges presented by inventor Radia Perlman at Infocom.

TRILL IN 2005—2007

- 2005: TRILL presented to IEEE 802 by Radia Perlman, rejected.
- **2005: TRILL presented to IETF which Charters the TRILL Working Group in the Internet Area.**
- **2005 TRILL Chair: Eric Nordmark**
- **2005 TRILL INT AD: Margaret (Cullen) Wasserman**
- **2005 TRILL INT AD: Mark Townsley**
- **2006: TRILL co-Chairs:**
  Eric Nordmark, Donald Eastlake

TRILL IN 2008—2010

- 2008: MTU problem delays protocol while fix is incorporated.
- 2008 July: TRILL did not meet
- 2009: RFC 5556: TRILL: Problem and Applicability Statement
- **2009: TRILL INT AD: Ralph Droms**
- 2009 July: TRILL did not meet
- 2009: TRILL Protocol passed up to IESG for Approval.
- **2010: TRILL approved IETF Standard (2010-03-15)**
  - Ethertypes, Multicast addresses & NLPID assigned
- 2010: Successful TRILL control plane interop at UNH IOL

TRILL in 2009—2011

- 2011: TRILL Protocol base document set:
  - **RFC 6325: RBridges: TRILL Base Protocol Specification**
  - RFC 6326: TRILL Use of IS-IS
  - RFC 6327: RBridges: Adjacency
  - RFC 6361: TRILL over PPP
  - RFC 6439: RBridges: Appointed Forwarders
- 2011: TRILL Working Group Re-Chartered to do further development of the TRILL protocol

TRILL IN 2012—2013

- 2012: Second Successful TRILL control plane interop at UNH IOL
- **2012: TRILL WG Secretary: Jon Hudson**
- 2013: Additional TRILL documents published:
  - RFC 6847: FCoE (Fibre Channel over Ethernet) over TRILL
  - RFC 6850: RBridge MIB
  - RFC 6905: TRILL OAM Requirements
  - RFC 7067: TRILL Directory Assistance Problem and High-Level Design Proposal
- **2013: TRILL INT AD: Ted Lemon**
- 2013: Third TRILL interop for control and data plane at UNH IOL week of May 20th
- 2013: TRILL Working Group Re-Chartered to do further development of the TRILL protocol

TRILL IN 2014

- **2014: March**
- **2014: TRILL moved to Routing Area**
- **2014: TRILL RTG AD: Alia Atlas**
- **2014: TRILL WG Co-Chairs:**
  Donald Eastlake, Jon Hudson
- **2014: TRILL WG Secretary: Sue Hares**

- (continued next slide)

TRILL IN 2014 (CONT.)

- 2014: Additional TRILL documents published:
  - RFC 7172: TRILL Fine Grained Labeling
  - RFC 7173: TRILL Transport using pseudo-wired
  - RFC 7174: TRILL OAM Framework
  - RFC 7175: TRILL BFD Support
  - RFC 7176: TRILL use of IS-IS
  - RFC 7177: TRILL Adjacency
  - RFC 7178: TRILL RBridge Channel Support
  - RFC 7179: TRILL Header Extension
  - RFC 7180: TRILL Clarifications, Corrections, and Updates
  - RFC 7357: TRILL ESADI
  - RFC 7379: Problem Statement and Goals for Active-Active TRILL Edge

TRILL in 2014—2015

- **2014: July**
- **2014: TRILL WG Co-Chairs: Sue Hares, Jon Hudson**
- **2014: TRILL WG Secretary: Donald Eastlake**

- 2015: Additional TRILL documents published:
  - RFC 7455: TRILL Fault Management

TRILL IN 2016

- 2016: Additional TRILL document published
  - RFC 7780: TRILL Clarifications, Corrections, and Updates
  - RFC 7781: TRILL Pseudo-Nickname for Active-Active Access
  - RFC 7782: TRILL Active-Active Edge Using Multiple MAC Attachments
  - RFC 7783: Coordinated Multicast Trees for TRILL
  - RFC 7784: TRILL OAM MIB

  - RFC 7956: TRILL Distributed Layer 3 Gateway
  - RFC 7961: TRILL Interface Addresses APPsub-TLV
  - RFC 7978: TRILL RBridge Channel Header Extension

TRILL IN 2017—2018

- 2017: Additional TRILL documents published:
  - RFC 8139: TRILL Appointed Forwarders
  - RFC 8171: TRILL Edge Directory Assistance Mechanisms
  - RFC 8249: TRILL MTU Negotiation
  - RFC 8243: Alternatives for Multilevel TRILL

- 2018: Additional TRILL document published so far:
  - RFC 8302: ARP and Neighbor Discovery (ND) Optimization

TRILL Plugfests

- Cumulative participants at the UNH IOL TRILL interoperability events:
  - Broadcom
  - Extreme Networks
  - HP/H3C Networking
  - Huawei Technologies
  - Ixia
  - JDSU
  - Oracle
  - Spirent

Two TRILL Non-RFC References

- TRILL Introductory Internet Protocol Journal Article:
  - http://www.cisco.com/web/about/ac123/ac147/archived_issues/ipj_14-3/143_trill.html

- The first paper: Perlman, Radia. “Rbridges: Transparent Routing”, Proceeding Infocom 2004, March 2004.
  - http://www.ieee-infocom.org/2004/Papers/26_1.PDF

END

Donald E. Eastlake 3rd

Co-Chair, TRILL Working Group

Principal Engineer, Huawei

d3e3e3@gmail.com

Backup Slides

Donald E. Eastlake 3rd
Co-Chair, TRILL Working Group
Principal Engineer, Huawei
d3e3e3@gmail.com

Algorhyme (Spanning Tree)

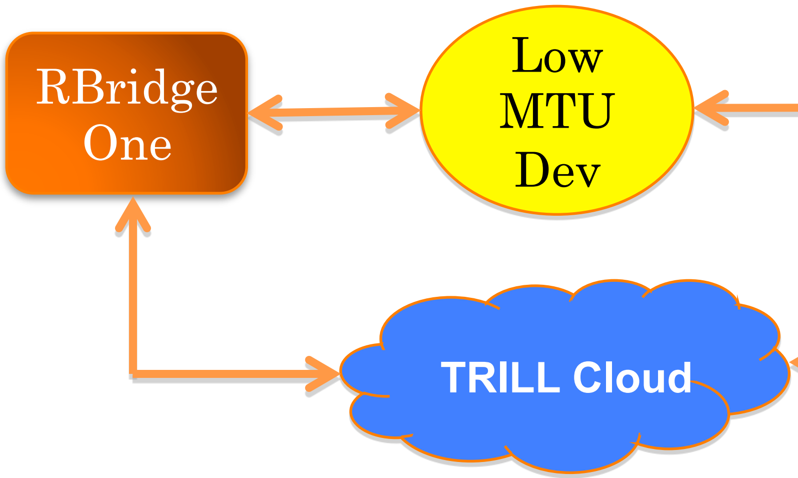
- I think that I shall never see
- A graph more lovely than a tree.
- A tree whose crucial property
- Is loop-free connectivity.
- A tree that must be sure to span
- So packets can reach every LAN.
- First, the root must be selected.
- By ID, it is elected.
- Least-cost paths from root are traced.
- In the tree, these paths are placed.
- A mesh is made by folks like me,
- Then bridges find a spanning tree.
- - By Radia Perlman

How TRILL Works

- RBridges find each other by exchanging TRILL IS-IS Hello frames.
  - Like all TRILL IS-IS frames, TRILL Hellos are sent on Ethernet to the multicast address All-IS-IS-RBridges. They are transparently forwarded by bridges, dropped by end stations including routers, and are processed (but not forwarded) by RBridges.
  - TRILL Hellos are different from Layer 3 IS-IS LAN Hellos because they are small, unpadded, and support fragmentation of some information.
    - Separate MTU-probe and MTU-ack IS-IS messages are available for MTU testing and determination.
  - Using the information exchanged in the Hellos, the RBridges on each link elect the Designated RBridge for that link (the link could be a bridged LAN).

How TRILL Works

- TRILL Hellos are unpadded and a maximum of 1470 bytes long to be sure RBridges can see each other so you don't get two Designated RBridges on the same link.



How TRILL Works

- RBridges use the IS-IS reliable flooding protocol so that each RBridge has a copy of the global "link state" database.
    - The RBridge link state includes the campus topology and link cost but also other information. Information such as VLAN/FGL connectivity, multicast listeners and multicast router attachment, claimed nickname(s), ingress-to-egress options supported, and the like.
    - The link state database is sufficient for each RBridge to independently and without further messages calculate optimal point-to-point paths for known unicast frames and the same distribution trees for multi-destination frames.

How TRILL Works

- The Designated RBridge specifies the Appointed Forwarder for each VLAN on the link (which may be itself) and the Designated VLAN for inter-RBridge communication.
- The Appointed Forwarder for VLAN-x on a link handles all native frames to/from that link in that VLAN. It is only significant if there are end station on the link.
  - It encapsulates frames from the link into a TRILL Data frame. This is the ingress RBridge function.
  - It decapsulates native frames destined for the link from TRILL Data frames. This is the egress RBridge function.

Why Designated VLAN?

- Ethernet links between RBridges have a Designated VLAN for inter-RBridge traffic. It is dictated by the Designated RBridge on the Link.
- For Point-to-Point links, usually no outer VLAN tag is needed on TRILL Data frames. For links configured as P2P, there is no Designated VLAN.
- However, there are cases where an outer VLAN tag with the designated VLAN ID is essential:
  - Carrier Ethernet facilities on the link restrict VLAN.
  - The link is actually a bridged LAN with VLAN restrictions.
  - The RBridge ports are configured to restrict VLANs.

How TRILL Works

- TRILL Data packets that have known unicast ultimate destinations are forwarded RBridge hop by RBridge hop toward the egress RBridge.

- TRILL Data packets that are multi-destination frames (broadcast, multicast, and unknown destination unicast) are forwarded on a distribution tree.

Multi-Destination Traffic

- Multi-destination data is sent on a bi-directional distribution tree:
    - The root of a tree is a TRILL switch or a link (pseudo-node) determined by a separate election and represented by nickname.
    - The ingress RBridge picks the tree, puts the tree root nickname in the "egress nickname" slot, and sets the M bit in the TRILL Header.
- All the TRILL switches in a campus calculate the same trees.
- All trees reach every TRILL switch in the campus.

Multi-Destination Traffic

- Multi-destination TRILL Data frames are more dangerous than unicast because they can multiply at fork points in the distribution tree.
  - So, in addition to the Hop Count, a Reverse Path Forwarding Check is performed. This discards the frame if, for the ingress and tree, it seems to be arriving on the wrong port.
  - To reduce the RPFC state, ingress RBridges can announce which tree or trees they will use.

MULTI-DESTINATION TRAFFIC

- As a TRILL Data frame is propagated on a distribution tree, its distribution can be pruned by VLAN and by multicast group since it is not useful to send a frame down a tree branch if
  - There are no end stations downstream in the VLAN of the frame, or
  - The frame is multicast and there is no multicast listener or multicast router downstream.

TRILL NICKNAMES

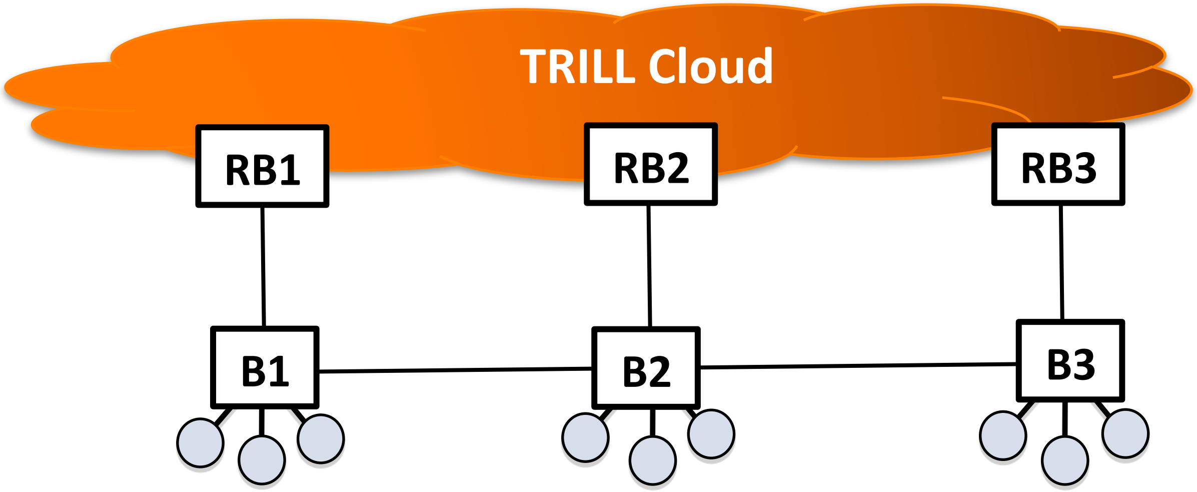
- TRILL switches are identified by 6-byte IS-IS System ID and by 2-bytes nicknames.
- Nicknames can be configured but by default are auto-allocated. In case of collisions, the lower priority RBridge must select a new nickname.
- Nicknames:
  - Saves space in headers.
  - An RBridge can hold more than one nickname so that
    - It can be the root of more than one different distribution tree.
    - May be used to distinguish frames following traffic engineered routes versus least cost routes.

Why IS-IS For TRILL?

- The IS-IS (Intermediate System to Intermediate System) link state routing protocol was chosen for TRILL over OSPF (Open Shortest Path First), the only other plausible candidate, for the following reasons:
  - IS-IS runs directly at Layer 2. Thus no IP addresses are needed, as they are for OSPF, and IS-IS can run with zero configuration.
  - IS-IS uses a TLV (type, length, value) encoding which makes it easy to define and carry new types of data.

RBridges & Access Links

- You can have multiple TRILL switches on a link with one or more end stations.
- The elected Designated RBridge is in charge of the link and by default handles end station traffic. But to load split, it can assign end station VLANs to other RBridges on the link.



MAC ADDRESS LEARNING

- By IS-IS all TRILL switches in the campus learn about and can reach each other but what about reaching end station MAC addresses?
  - By default, TRILL switches at the edge (directly connected to end stations) learn attached VLAN/MAC addresses from data as bridges do.
  - Optionally, MAC addresses can be passed through the control plane.
  - MAC addresses can be statically configured or learned from Layer 2 registration protocols such as Wi-Fi association or 802.1X.
  - Transit TRILL switches do not learn end station addresses.

MAC Address Learning

- Data Plane Learning
  - From Locally Received Native Frames
    - { VLAN, Source Address, Port }
  - From Encapsulated Native Frames
    - { Inner VLAN, Inner Source Address, Ingress RBridge }
    - The Ingress RBridge learned is used as egress on sending
- Via
  1. Optional End Station Address Distribution Information (ESADI) control plane protocol
  2. Via Layer-2 Registration protocol(s)
  3. By manual configuration
     - { VLAN, Address, RBridge nickname }

ESADI

- The optional End Station Address Distribution Information (ESADI) protocol:
  - Provides a VLAN/tenant scoped way for an RBridge to distribute control plane information about attached End Stations to other RBridges.
  - Highly efficient transmission because information is tunneled through transit RBridges encapsulated as if it was normal data.
  - Intended for use for attachment data that is either secure or that changes rapidly.
    - The source RBridge selects which addresses it wants to distribute through ESADI.
    - There is no particular advantage in using ESADI for large amounts of information learned from the data plane.