The IETF TRILL Protocol

Transparent Interconnection of Lots of Links

Donald E. Eastlake, 3rd

Co-Chair, TRILL Working Group

Principal Engineer, Huawei

d3e3e3@gmail.com

HUAWEI

Donald E. Eastlake, 3rd

- Principal Engineer at Huawei Technologies
  - Formerly with Cisco and before that Motorola.
- Co-Chair of the IETF TRILL Working Group
  - Chair of the IETF PPPEXT Working Group
  - Chair of the IEEE 802.11ak Task Group
- Author of 59 IETF RFCs.

Note:

This tutorial represents my personal views, not those of the TRILL WG or Huawei. It is a high level technical overview. It is not practical to include all the details in the specification documents in a presentation of this length.

CONTENTS

What is TRILL?

- A Compatible Protocol
  - Attached end nodes just think it is Ethernet.

- The more bridges you convert to TRILL switches, the better your network's stability and bandwidth utilization.

- Terminates Spanning Tree Protocols

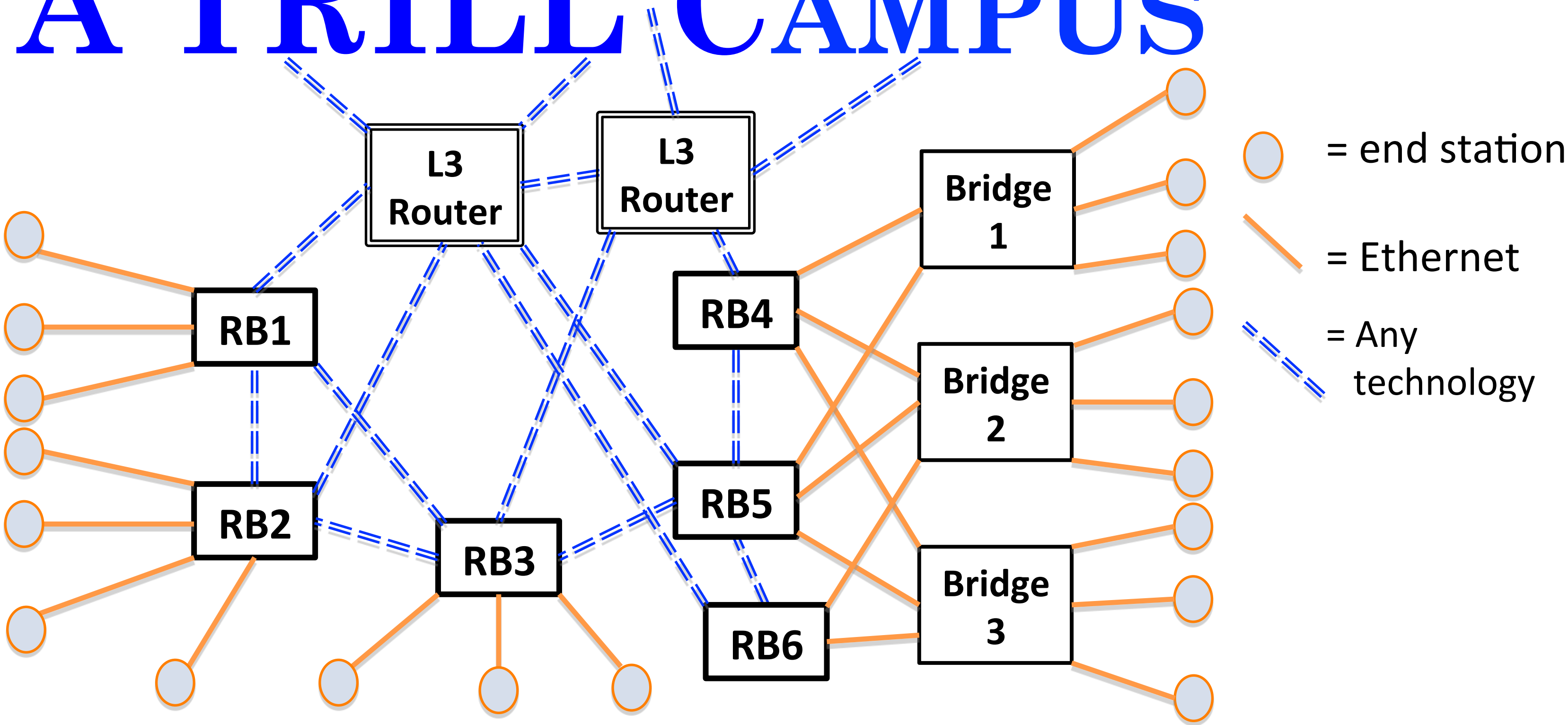
Inspired by a Real Life Incident

- 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.

What is TRILL?

- Basically a simple idea:
  - Encapsulate native Ethernet 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 traffic
  - Multi-pathing of multi-destination traffic

A TRILL CAMPUS



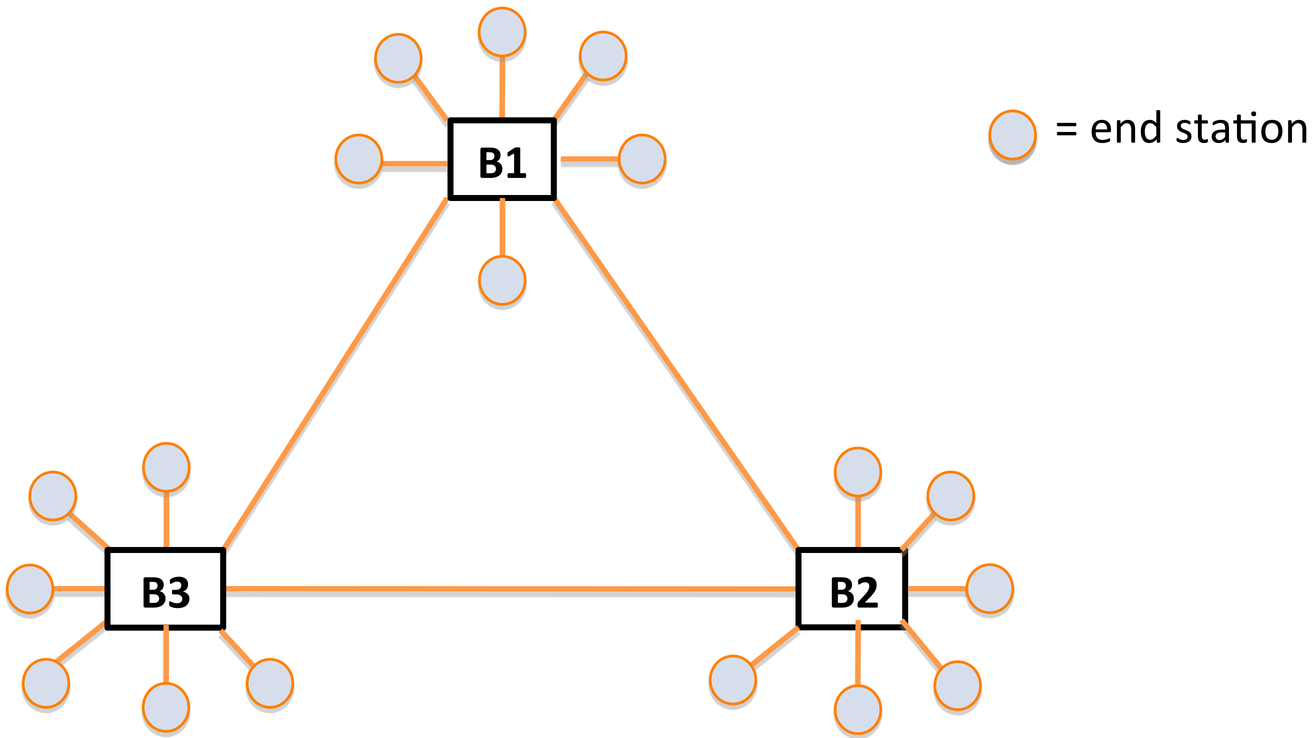
- End stations and Layer 3 routers are connected to TRILL switches by Ethernet.
- TRILL switches can be connected to each other with arbitrary technology.
- In both cases, the connection can be a bridged LAN.

What is TRILL?

- TRansparent Interconnection of Lots of Links
  - TRILL WG Charter
    - http://www.ietf.org/dyn/wg/charter/trill-charter.html
  - Standardized by IETF TRILL Working Group:
    - Donald E. Eastlake 3rd (Huawei), Co-Chair
    - Erik Nordmark (Cisco), Co-Chair
    - Jon Hudson (Brocade), Secretary
- TRILL Switch/ RBridge (Routing Bridge)
  - Device that implements TRILL
- TRILL/RBridge Campus
  - A network of RBridges, links, and any intervening bridges, bounded by end stations / layer 3 routers.

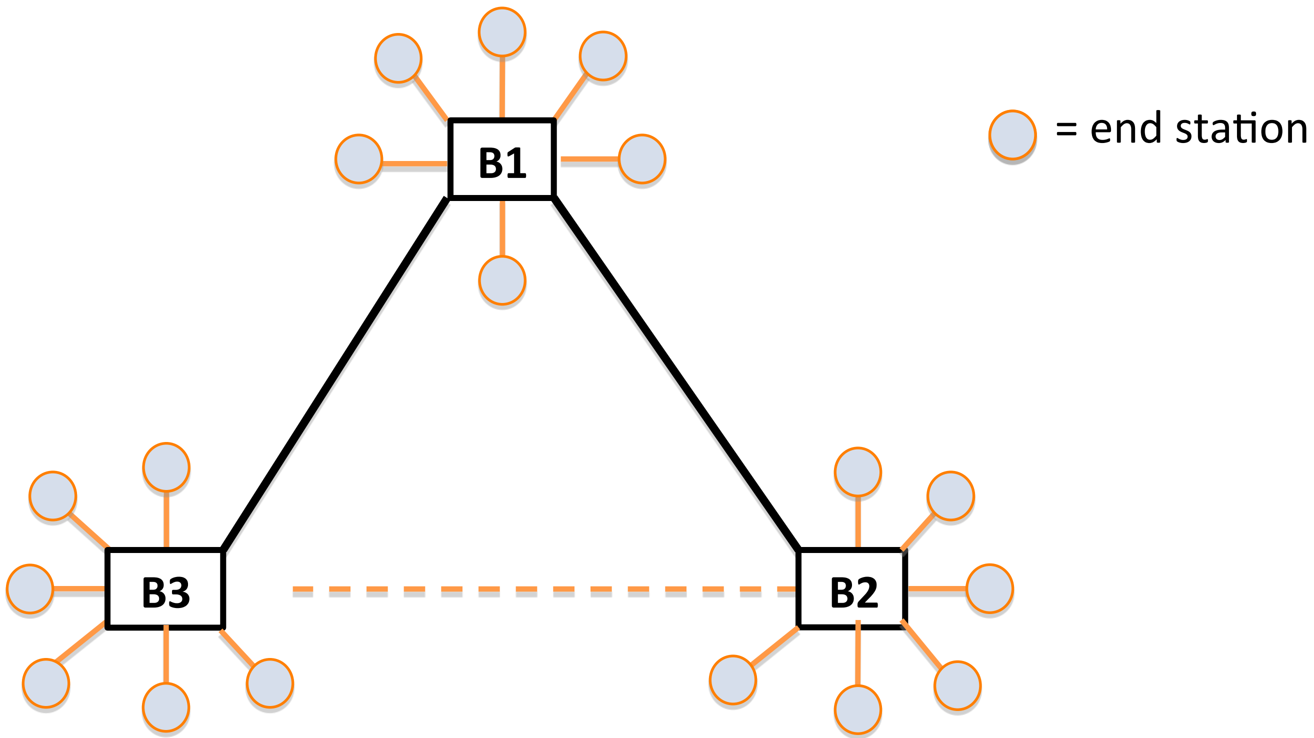
CONTENTS

Unicast Least Cost Paths



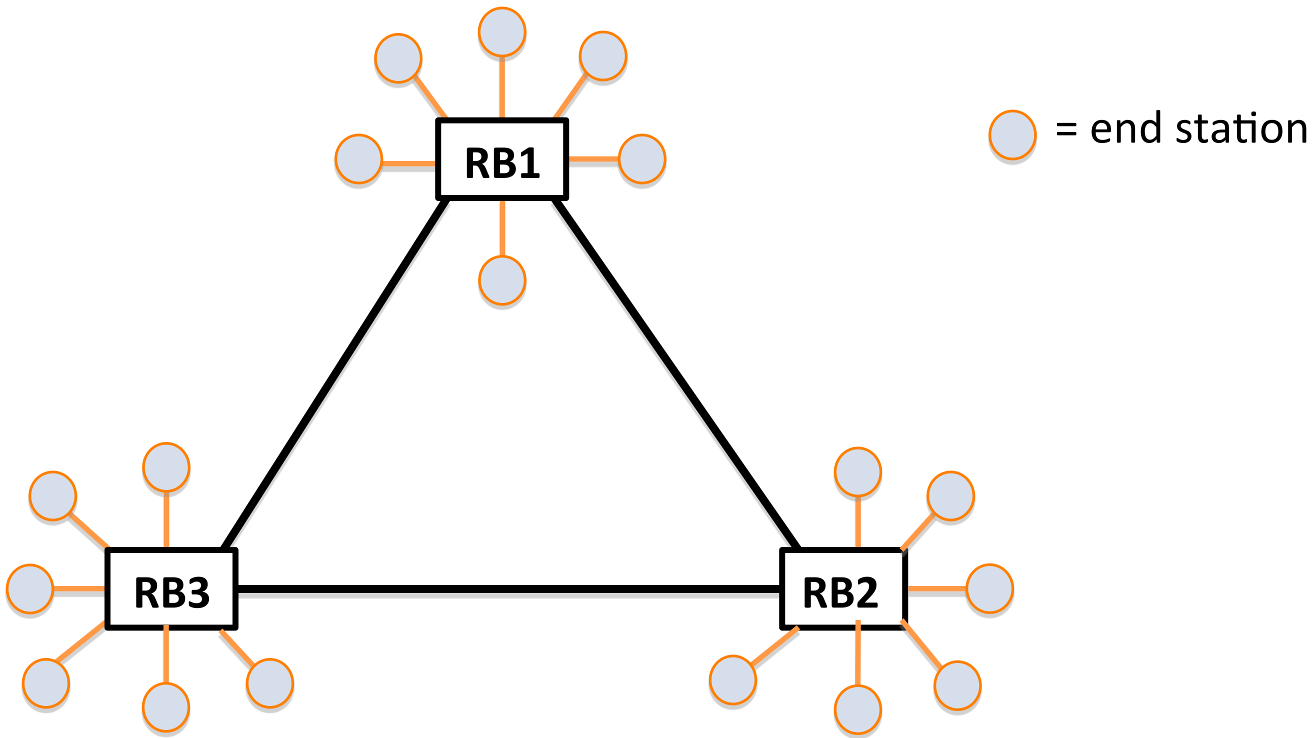
A three bridge network

Unicast Least Cost Paths



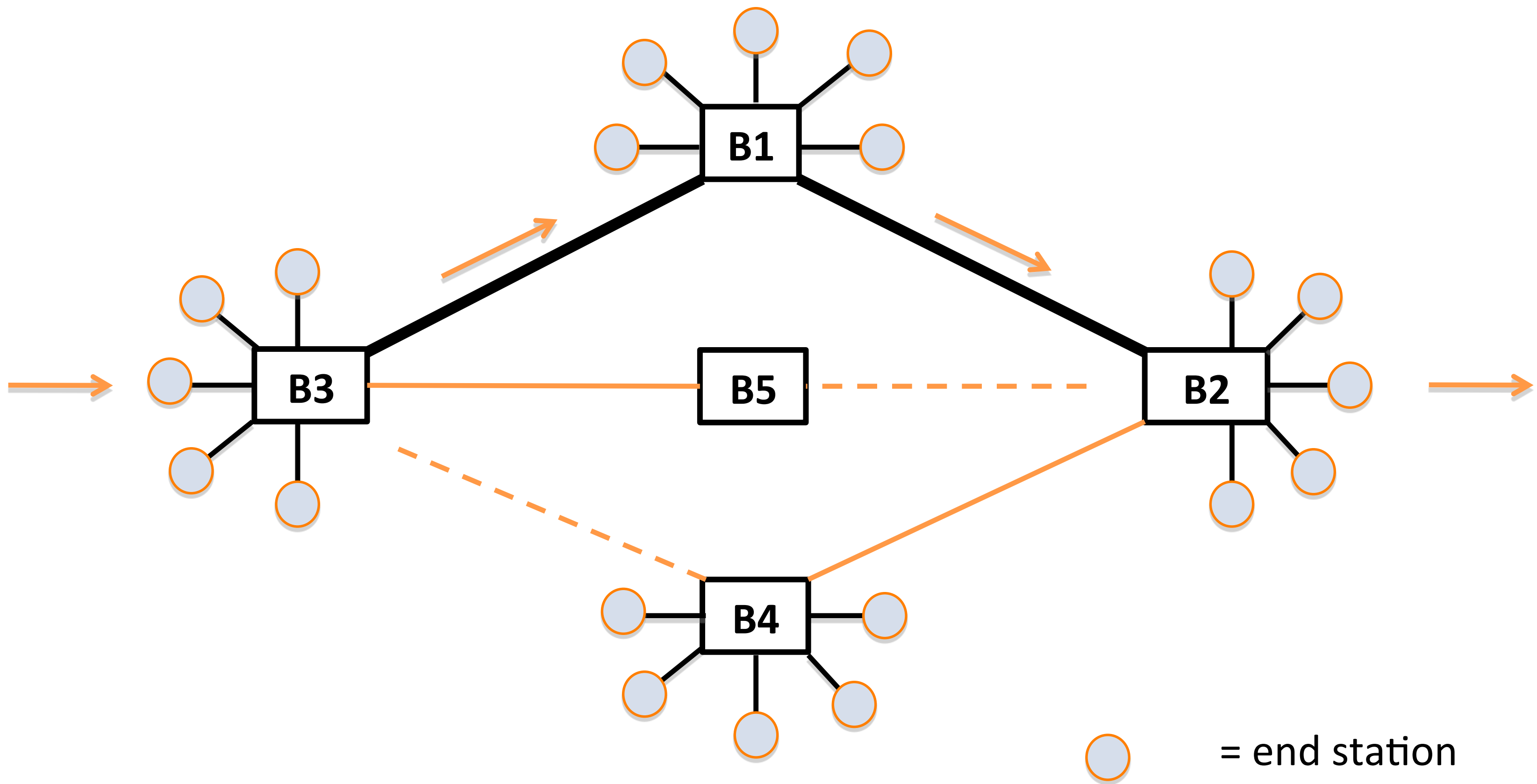
Spanning tree eliminates loops by disabling ports

Unicast Least Cost Paths



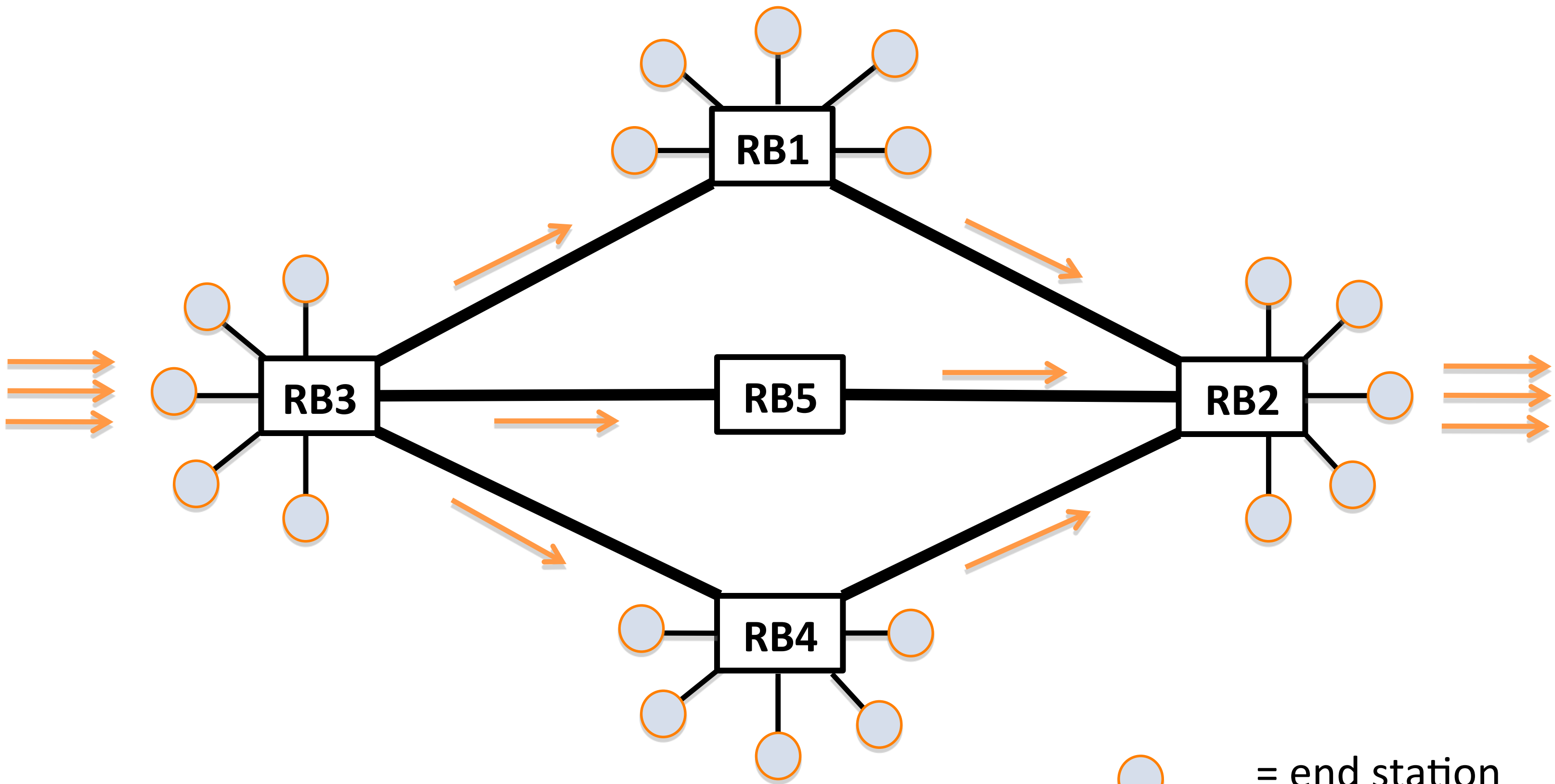
A three RBridge network: better performance using all facilities

Unicast Multi-Pathing



Bridges limit traffic to one path

UNICAST MULTI-PATHING



RBridges support
multi-path for higher throughput

TRILL FEATURES

- 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
- 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.
- Has a poem.
  - The only other bridging or routing protocol with a poem is Spanning Tree (see Algorhyme).

Algorhyme V2 (TRILL)

- 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)

CONTENTS

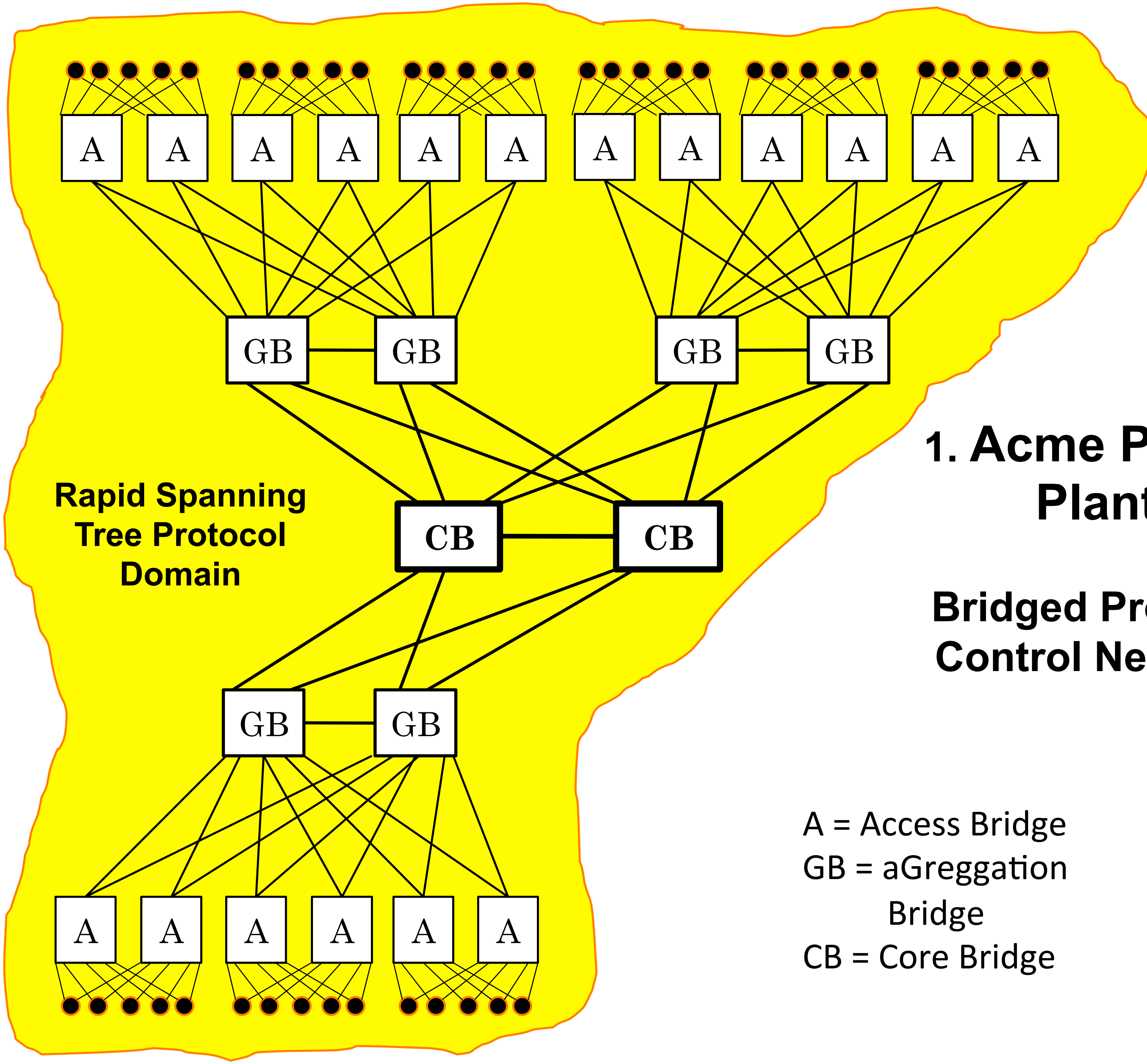
TRILL History up to 2009

- 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 invented by Radia Perlman, presented at Infocom.
- 2005: TRILL presented to IEEE 802 by Radia Perlman, rejected.
- 2005: IETF Charters the TRILL Working Group.
- 2008: MTU problem delays protocol while fix is incorporated.
- 2009: RFC 5556 "TRILL: Problem and Applicability Statement"
- 2009: TRILL Protocol passed up to IESG for Standards Approval.

TRILL IN 2010 TO 2013

- 2010: TRILL approved as IETF Standard (15 March 2010)
  - Ethertypes, Multicast addresses & NLPID assigned
- 2010: Successful TRILL control plane plugfest at UNH IOL
- 2011: TRILL Protocol base document set published:
  - RFC 6325: "RBridges: TRILL Base Protocol Specification" (Includes TRILL over Ethernet)
  - RFC 6326: "TRILL Use of IS-IS"
  - RFC 6327: "RBridges: Adjacency"
  - RFC 6361: "TRILL over PPP"
  - RFC 6439: "RBridges: Appointed Forwarders"
- 2012: 2nd TRILL plugfest at UNH IOL
- 2013: RFC 6847: "FCoE over TRILL"
- 2013: RFC 6850: "Definition of Managed Objects for RBridges"

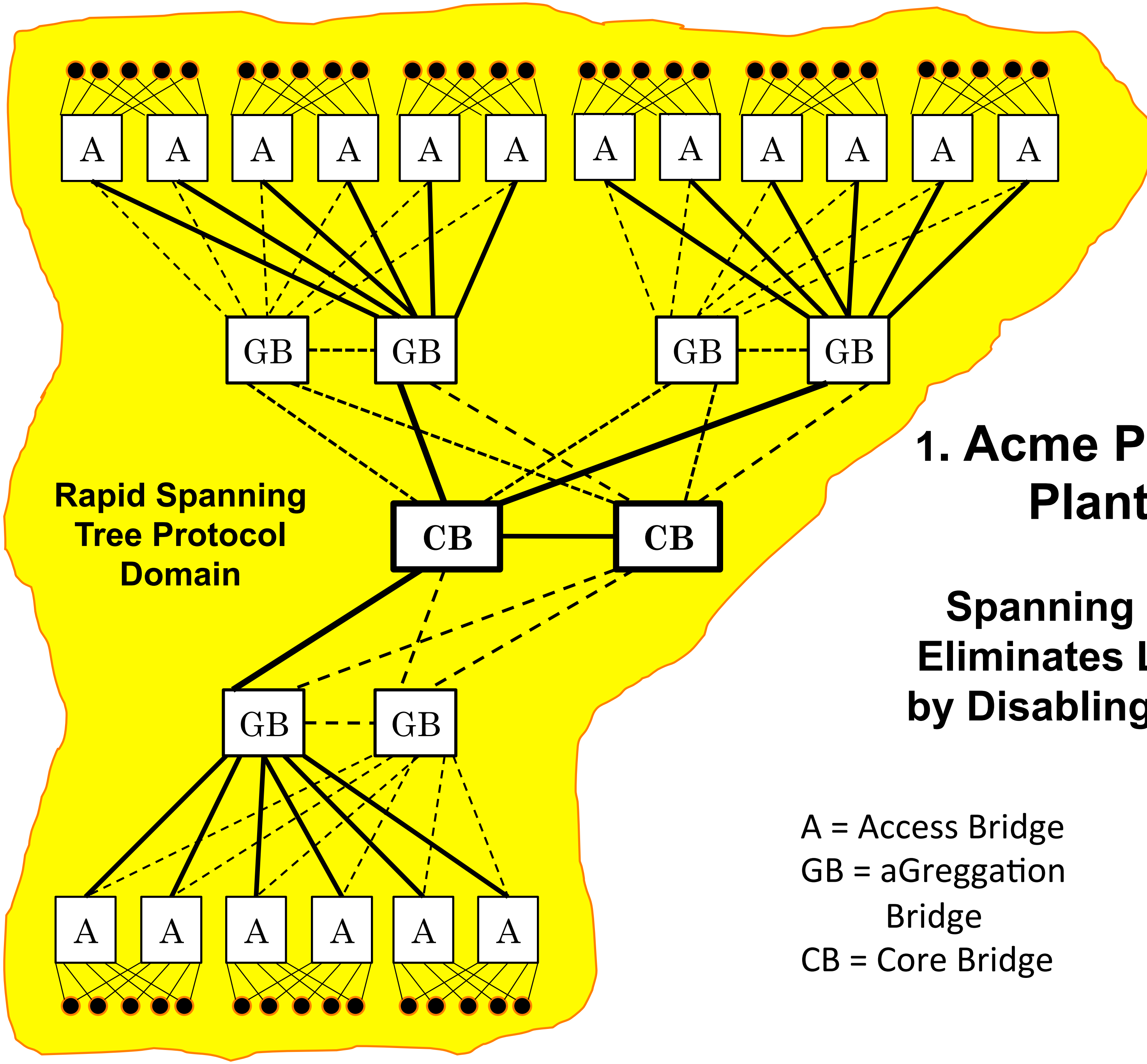
CONTENTS



1. Acme Power Plant

Bridged Process Control Network

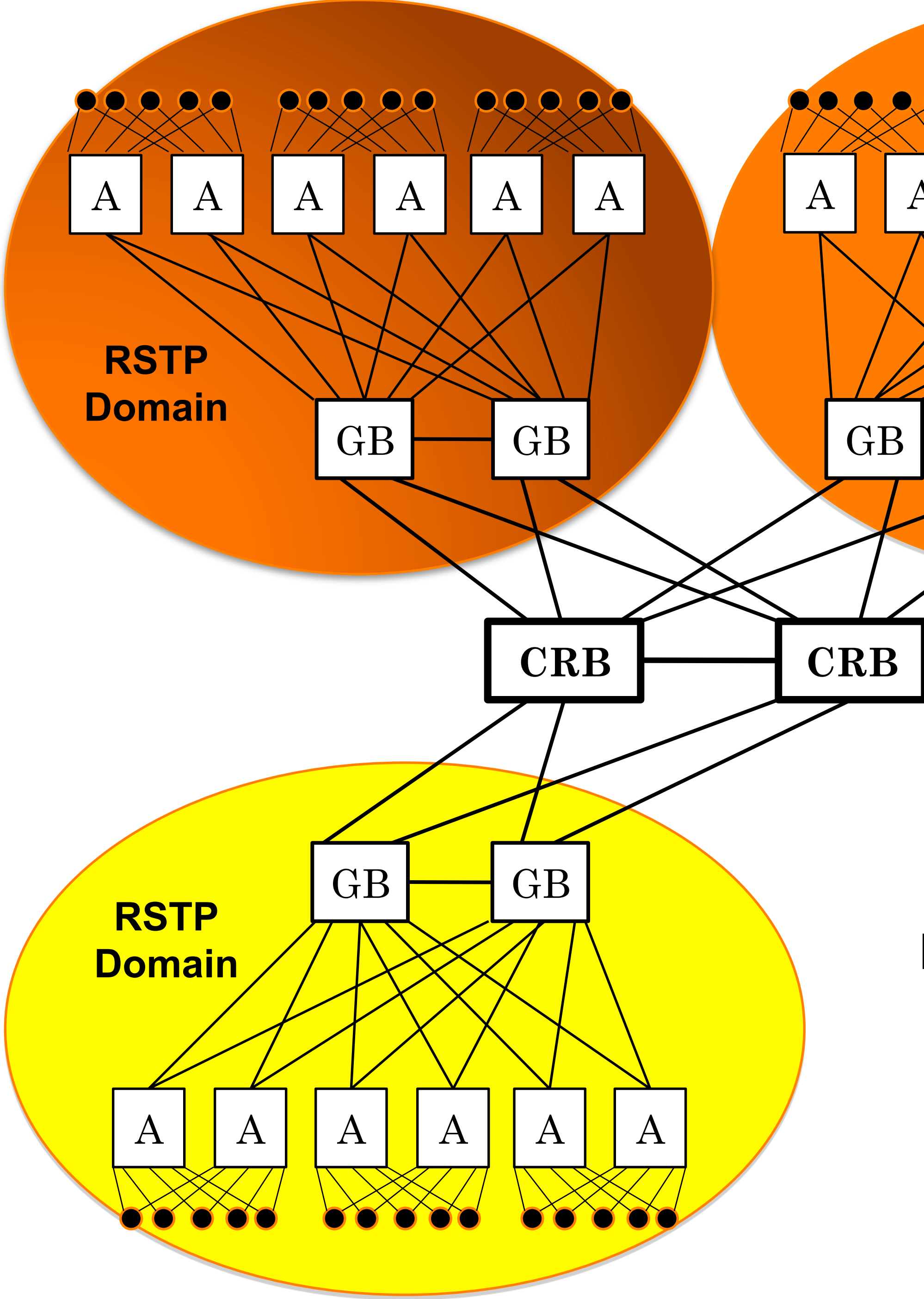
A = Access Bridge
GB = aGreggation Bridge
CB = Core Bridge



1. Acme Power Plant

Spanning Tree Eliminates Loops by Disabling Ports

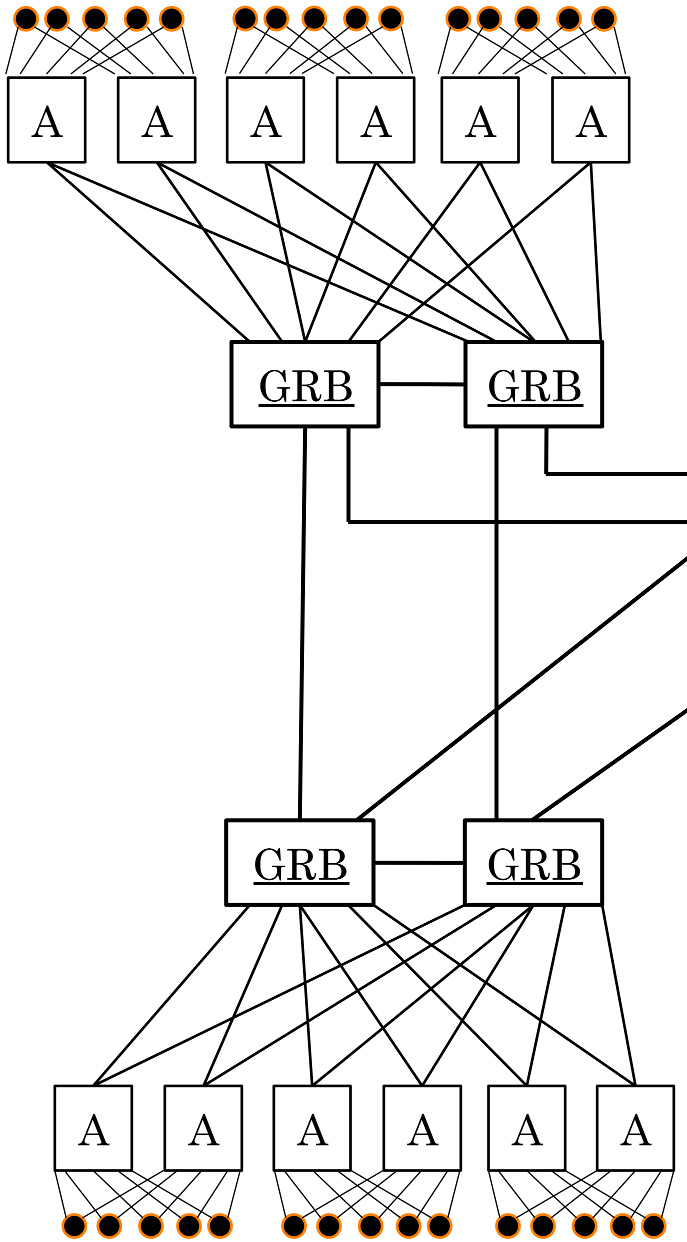
A = Access Bridge
GB = aGreggation Bridge
CB = Core Bridge



1. Acme Power Plant

Process Control Network with RBridge Core breaking up spanning tree

A = Access Bridge
GB = aGreggation Bridge
CRB = Core RBridge



1. Acme Power Plant

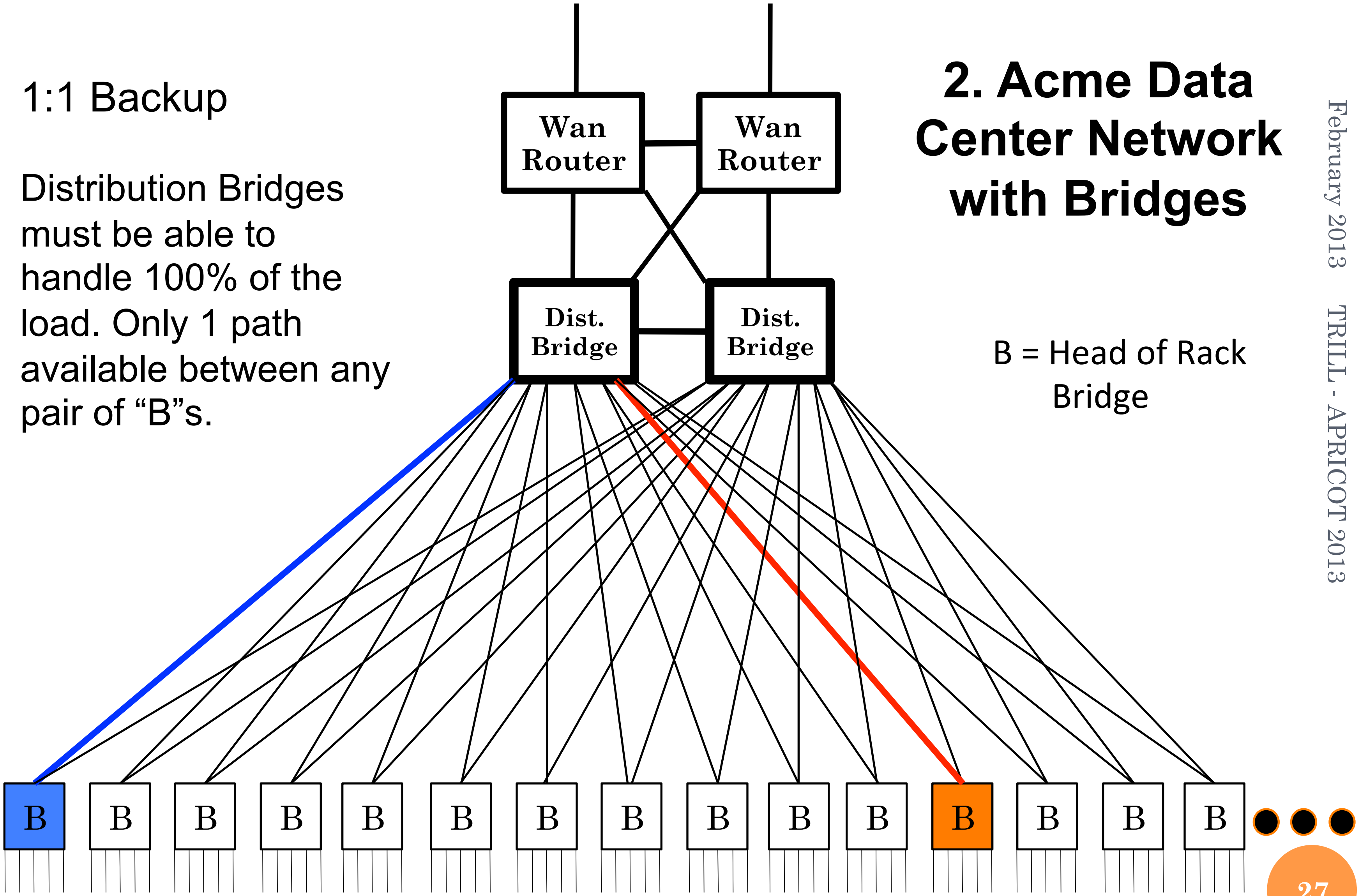
Process Control Network with RBridge Mesh eliminating spanning tree

A = Access Bridge
GRB = aGreggation RBridge

2. Acme Data Center Network with Bridges

1:1 Backup

Distribution Bridges must be able to handle 100% of the load. Only 1 path available between any pair of "B"s.



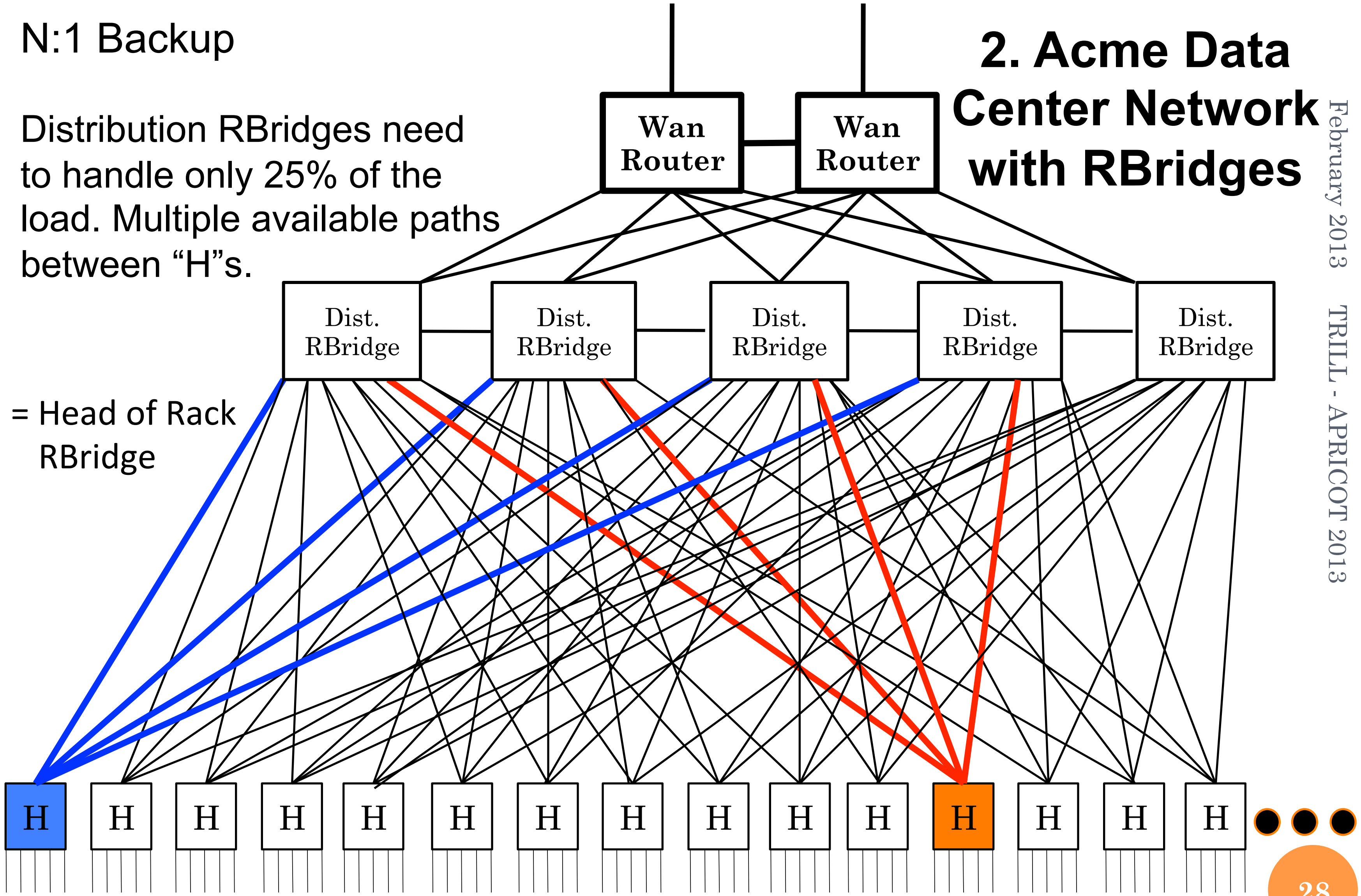
B = Head of Rack Bridge

N:1 Backup

Distribution RBridges need to handle only 25% of the load. Multiple available paths between “H”s.

H = Head of Rack RBridge

2. Acme Data Center Network with RBridges



Two TRILL Examples

- "Acme Power Plant" Process Control
  - Large process control commonly uses Ethernet
  - Some process control protocols interpret network interruption >1 second as equipment failure
  - Even Rapid Spanning Tree Protocol can take >3 second to recover from root bridge failure
  - Core RBridges reduce/eliminate spanning tree
- "Acme Data Center"
  - 1:1 to N:1 Backup Improvement

CONTENTS

TRILL IS BASED ON IS-IS

- TRILL switches (RBridges) use IS-IS (Intermediate System to Intermediate System) link state routing
  - Neighbor RBridges find each other by exchanging Hellos
  - This information is flooded so all RBridges in the campus know about all adjacencies. Then all RBridges can
    - calculate the topology for least cost unicast forwarding, including Equal Cost Multi-Pathing
    - Calculate the same distribution trees for multi-destination frames
  - Other flooded information supports nicknames (see later slide), optimization of multicast distribution based on VLAN attachment and multicast listeners, etc.

TRILL IS BASED ON IS-IS

- The IS-IS (Intermediate System to Intermediate System) link state routing protocol was chosen for TRILL over IETF OSPF (Open Shortest Path First), the only plausible alternative, 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.

TRILL NICKNAMES

- TRILL switches are identified by 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.

TRILL Encapsulation and Header

- TRILL Data frames between RBridges are encapsulated in a TRILL Header and then in a local link header.
  - On Ethernet links, the link header is addressed from the local source RBridge to the next hop RBridge for known unicast frames or to the All-RBridges multicast address for multi-destination frames.
  - The TRILL header specifies the first/ingress RBridge and either the last/egress RBridge for known unicast frames or the distribution tree for multi-destination frames.

TRILL Encapsulation and Header

- Reasons for encapsulation:
  - Provides a hop count to mitigate loop issues
  - To hide the original source address to avoid confusing any bridges present as might happen if multi-pathing were in use
  - To direct unicast frames toward the egress RBridge so that forwarding tables in transit RBridges need only be sized with the number of RBridges in the campus, not the number of end stations
  - To provide a separate outer VLAN tag, when necessary, for forwarding traffic between RBridges, independent of the original VLAN of the frame

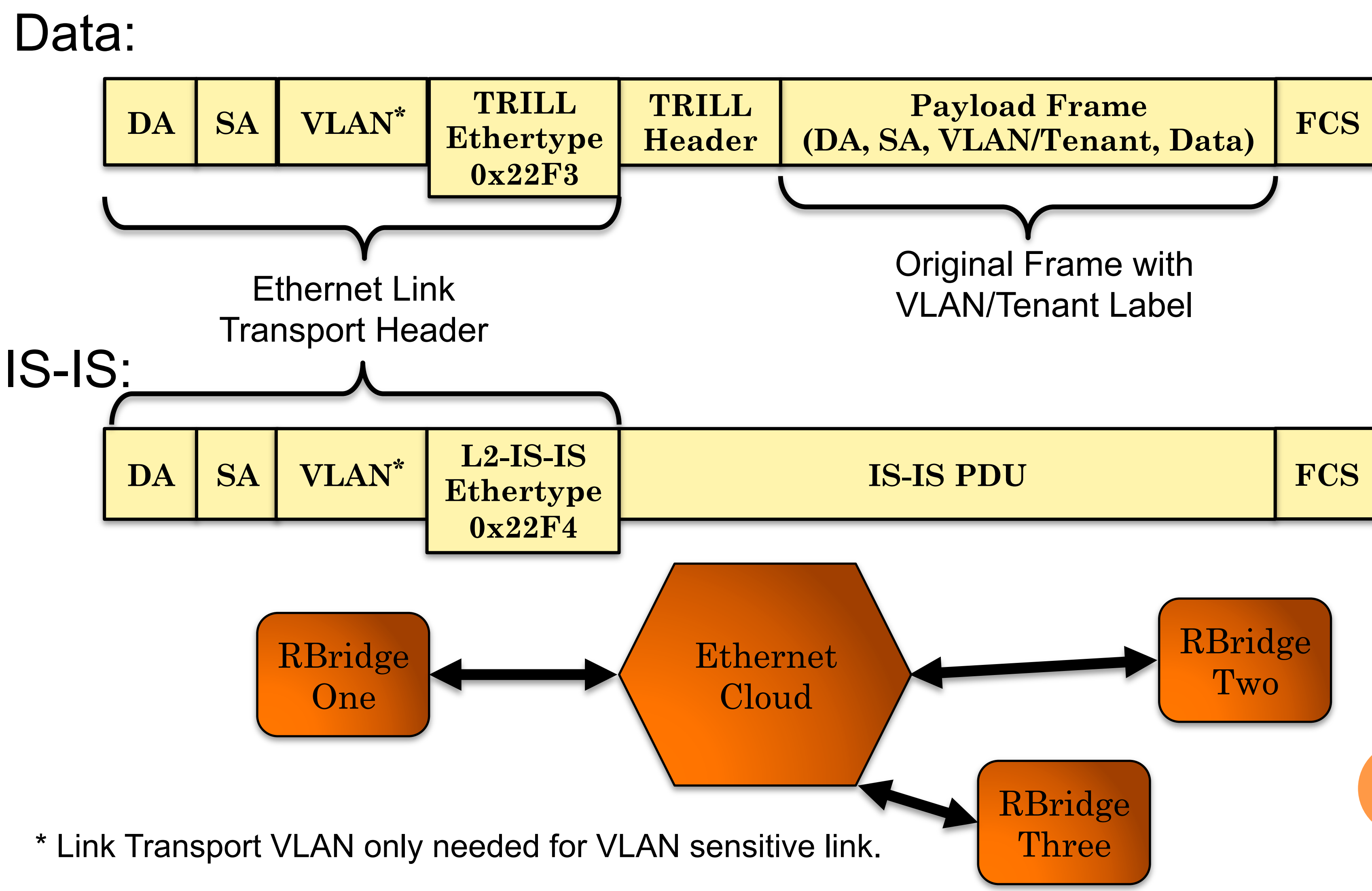
TRILL HEADER DETAILS

- TRILL Header

TRILL Ethertype					
	V	R	M	ExtLng	Hop
Egress RBridge Nickname	Ingress RBridge Nickname				

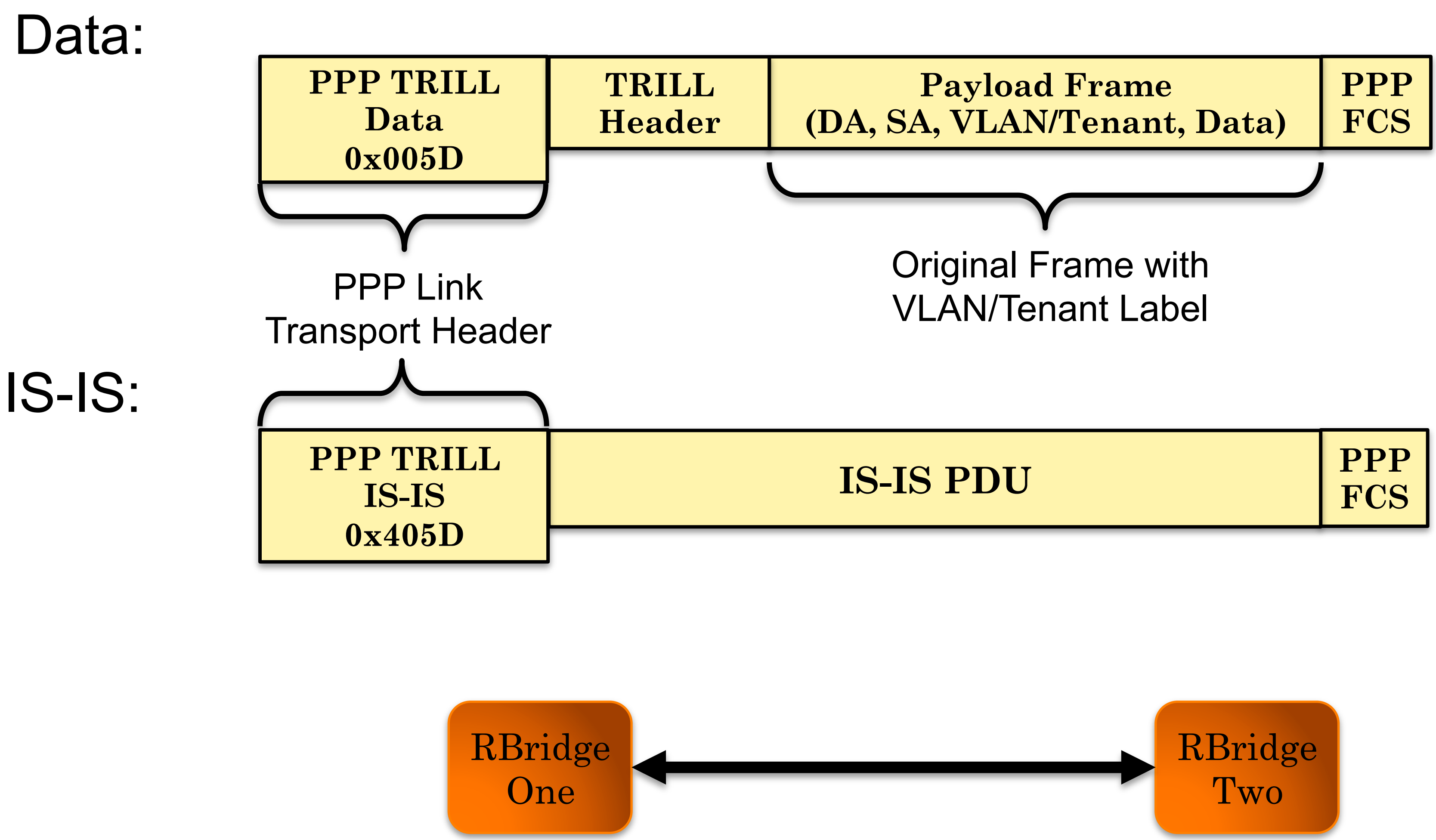
  - Nicknames – auto-configured 16-bit campus local names for RBridges
  - V = Version (2 bits)
  - R = Reserved (2 bits)
  - M = Multi-Destination (1 bit)
  - ExtLng = Length of TRILL Header Extensions
  - Hop = Hop Limit (6 bits)

TRILL OVER ETHERNET



* Link Transport VLAN only needed for VLAN sensitive link.

TRILL OVER PPP

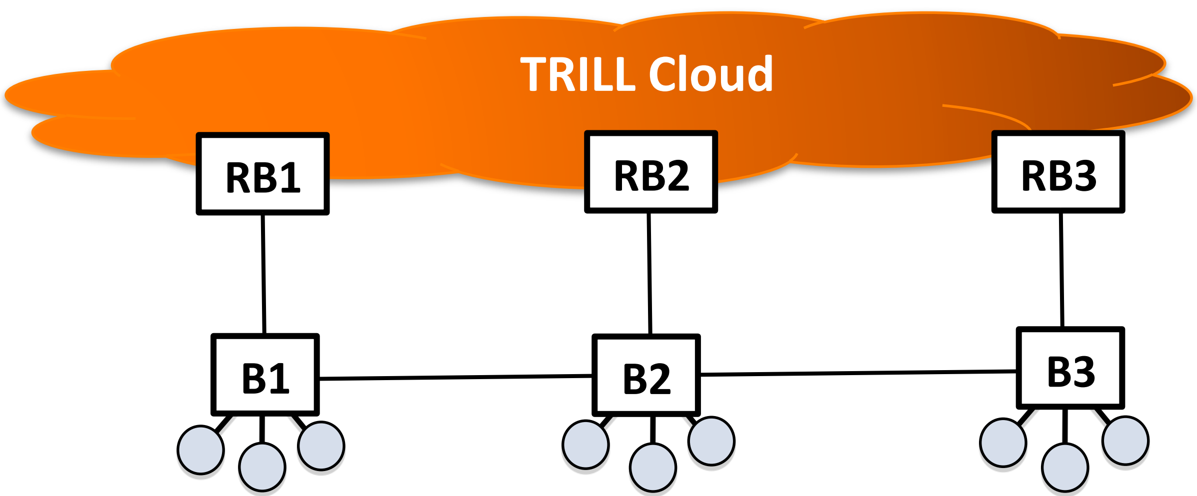


MAC ADDRESS LEARNING

- 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.
  - Optionally, MAC addresses can be passed through the control plane.
  - MAC addresses can be statically configured.
  - Transit TRILL switches do not learn end station addresses.

RBridges & Access Links

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



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.

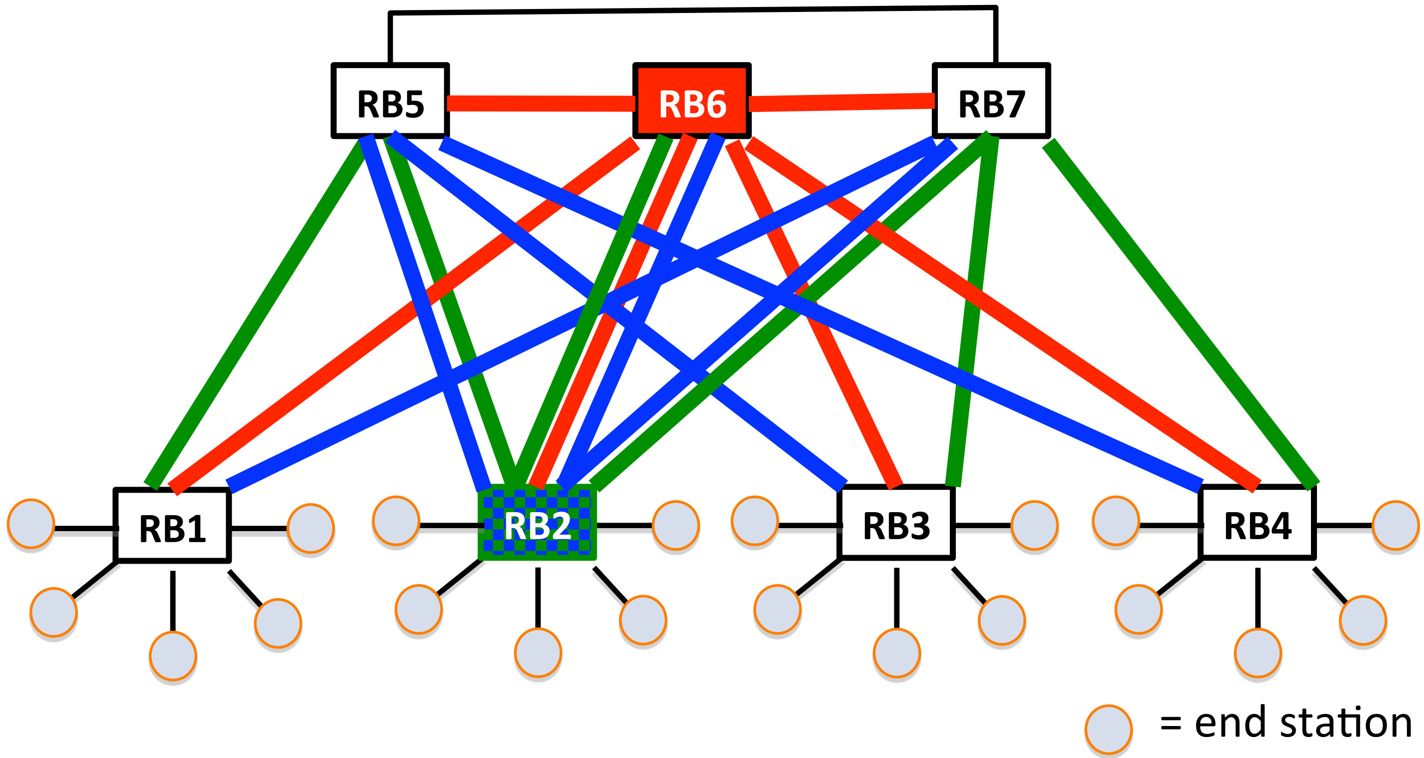
MULTI-DESTINATION TRAFFIC

- Multi-destination 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 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.

Multi-Destination Traffic



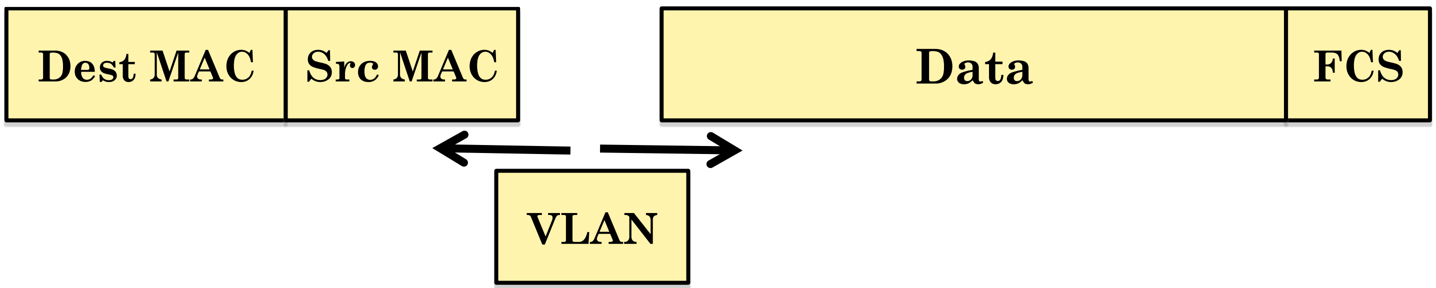
RBridges support multiple distribution trees.
The encapsulating RBridge chooses which to use. Each RBridge can split multi-destination traffic over three trees.

CONTENTS

INPUT PORT PROCESSING

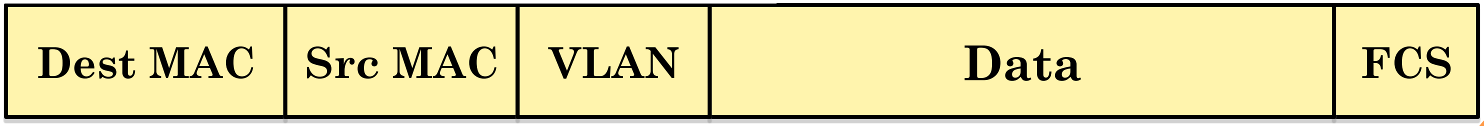
- Detailed example of unicast frame TRILL routing on an Ethernet link

Input Native Frame on link:

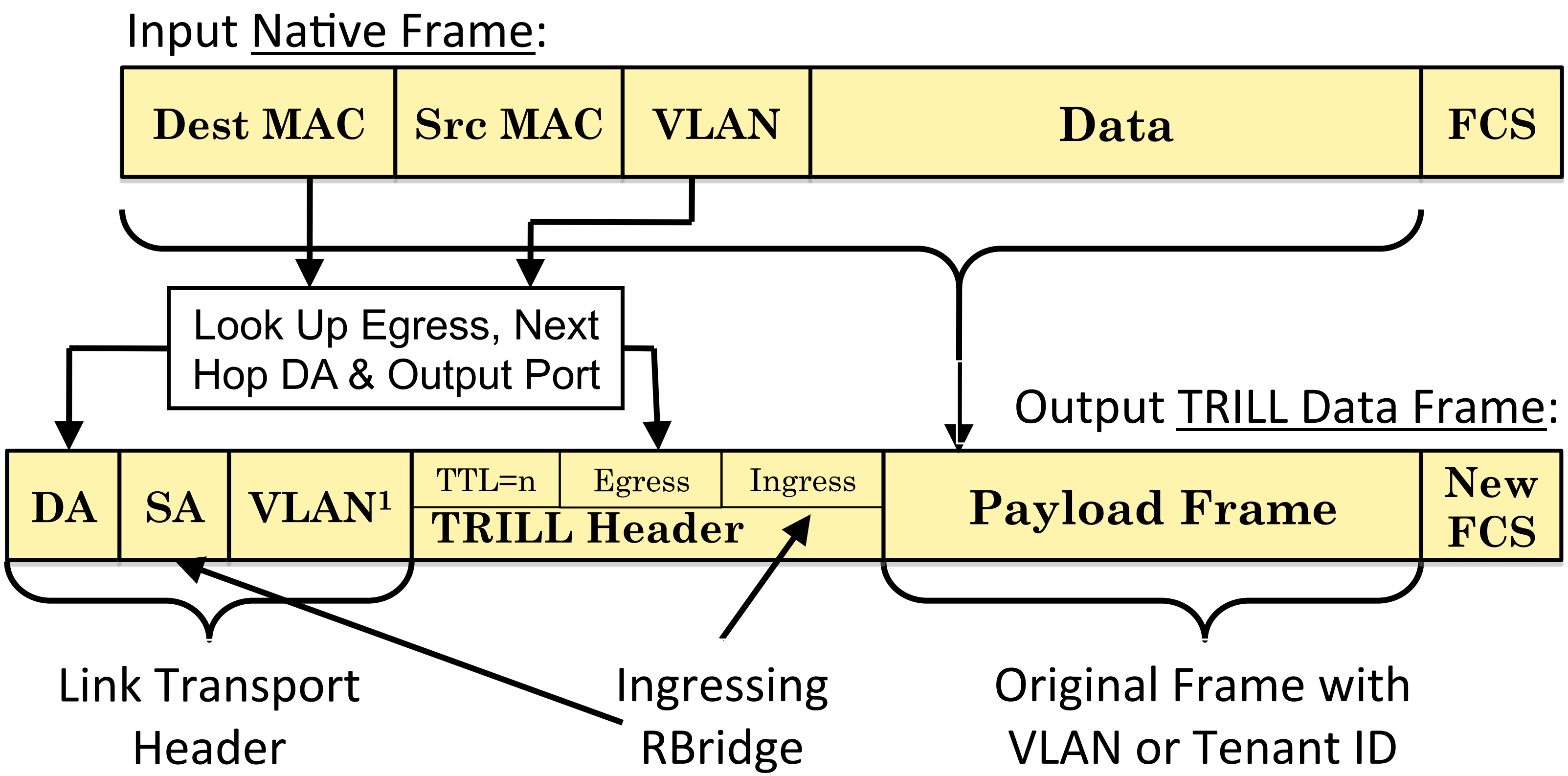


- Input port adds VLAN-ID and priority if frame untagged

Input Native Frame after input port:

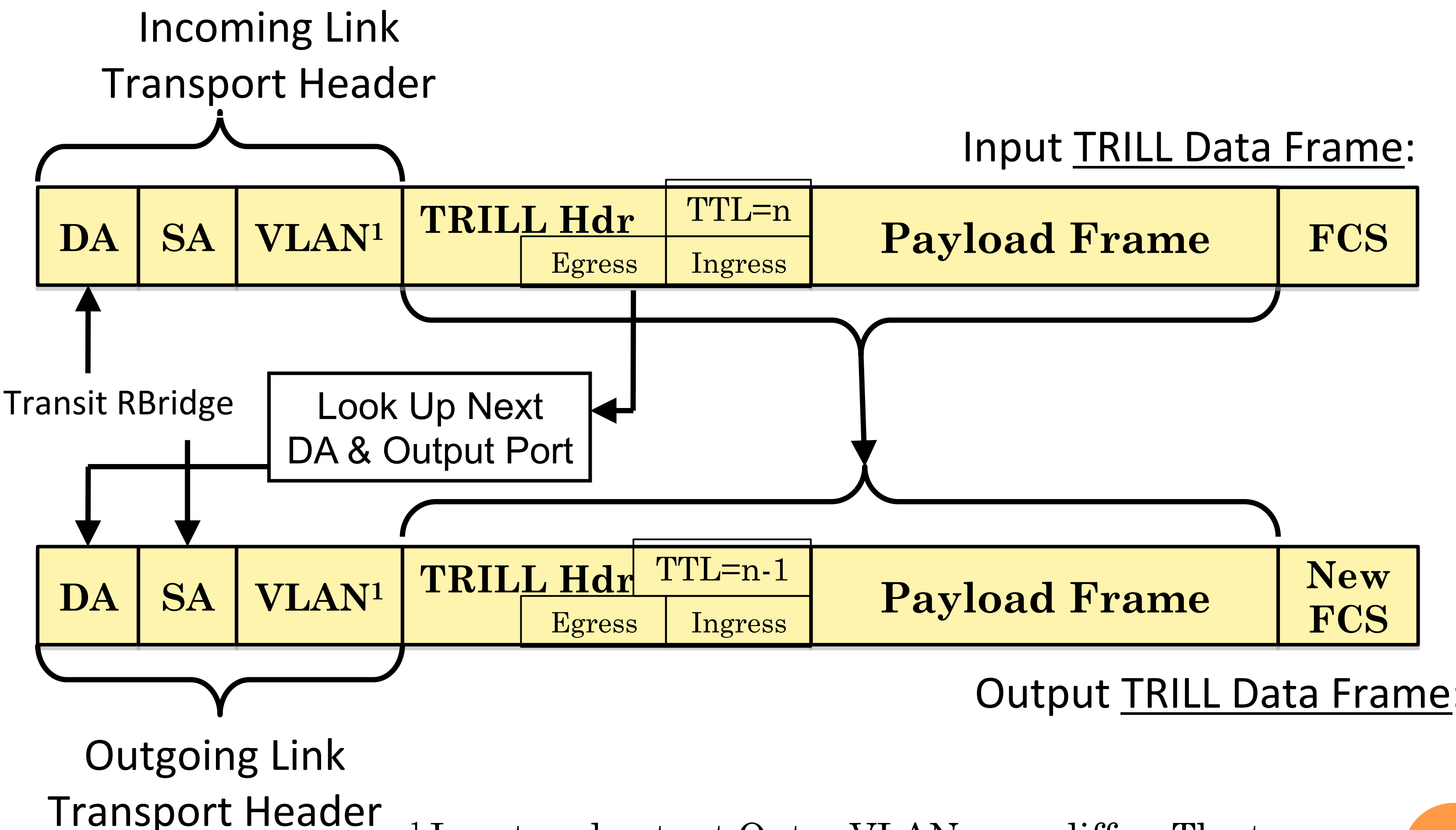


TRILL UNICAST INGRESS



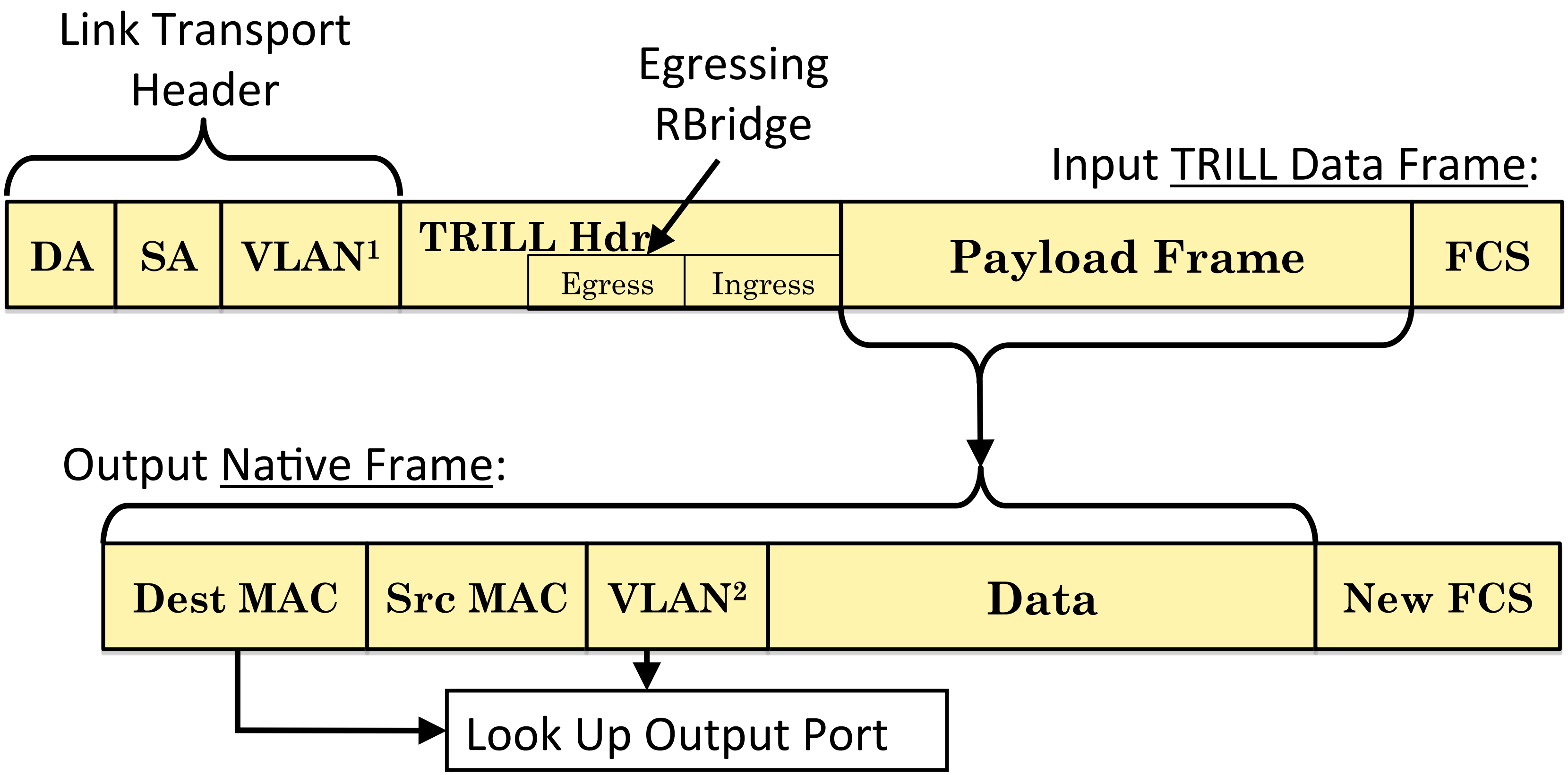
[1] Outer VLAN tag is a transport artifact and only needed if RBridges are connected by a bridged LAN or carrier Ethernet requiring a VLAN tag or the like.

TRILL Unicast Transit



[1] Input and output Outer VLANs can differ. The true VLAN or Tenant ID of the data is inside the payload frame. Outer VLAN only needed if link is VLAN sensitive.

TRILL Unicast Egress



[1] Outer VLAN only needed if RBridges are connected by a bridged LAN or carrier Ethernet requiring a VLAN tag or the like
[2] Final native frame VLAN tag may be omitted depending on RBridge output port configuration.

Output Port Processing

Output Native Frame before output port:

- Output port may be configured to output untagged and will do so by default for the port VLAN ID

CONTENTS

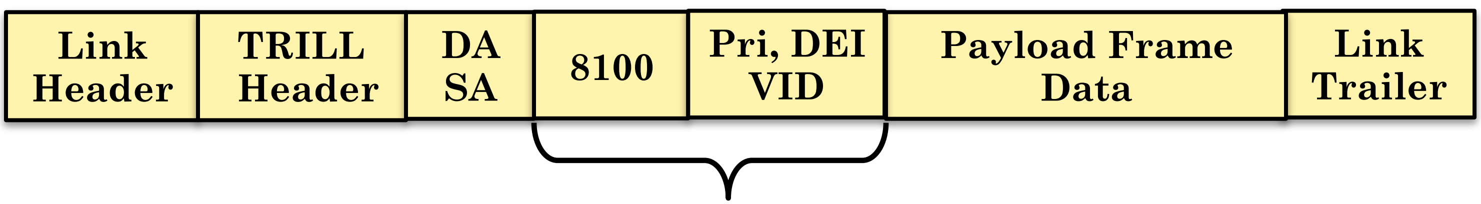
Fine Grained Labeling

- With Fine Grained Labeling (FGL), the inner data label is an extended 24-bit label.
- Ingress/egress TRILL switches that support FGL:
  - Map native frame VLAN and input port into a fine grained label on ingress and
    - do the reverse mapping on egress.
  - Remember the priority and DEI of native frames on ingress and restores them on egerss.
- Fine Grained Label TRILL switches are a superset of a base protocol TRILL switch. They support VLANs as in the base standard on a port if not configured to do Fine Grained Labeling

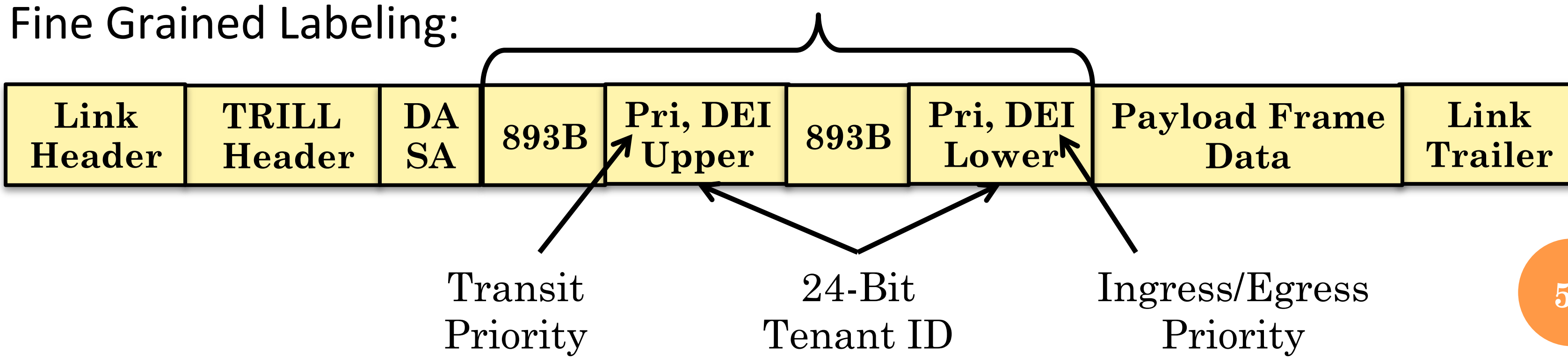
Fine Grained Labeling

- From the current draft:
  - https://datatracker.ietf.org/doc/draft-ietf-trill-fine-labeling/

Base protocol VLAN Labeling:

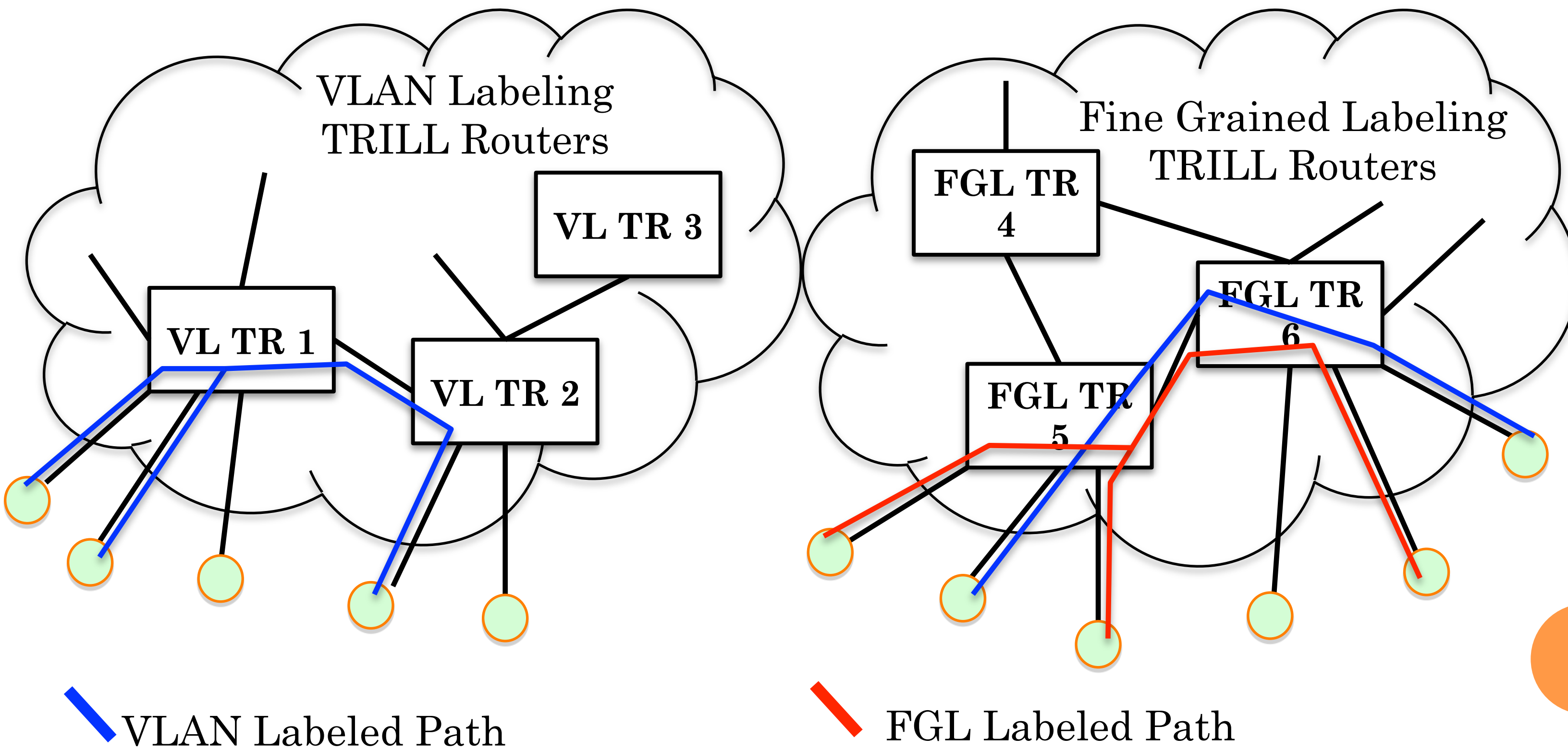


Fine Grained Labeling:



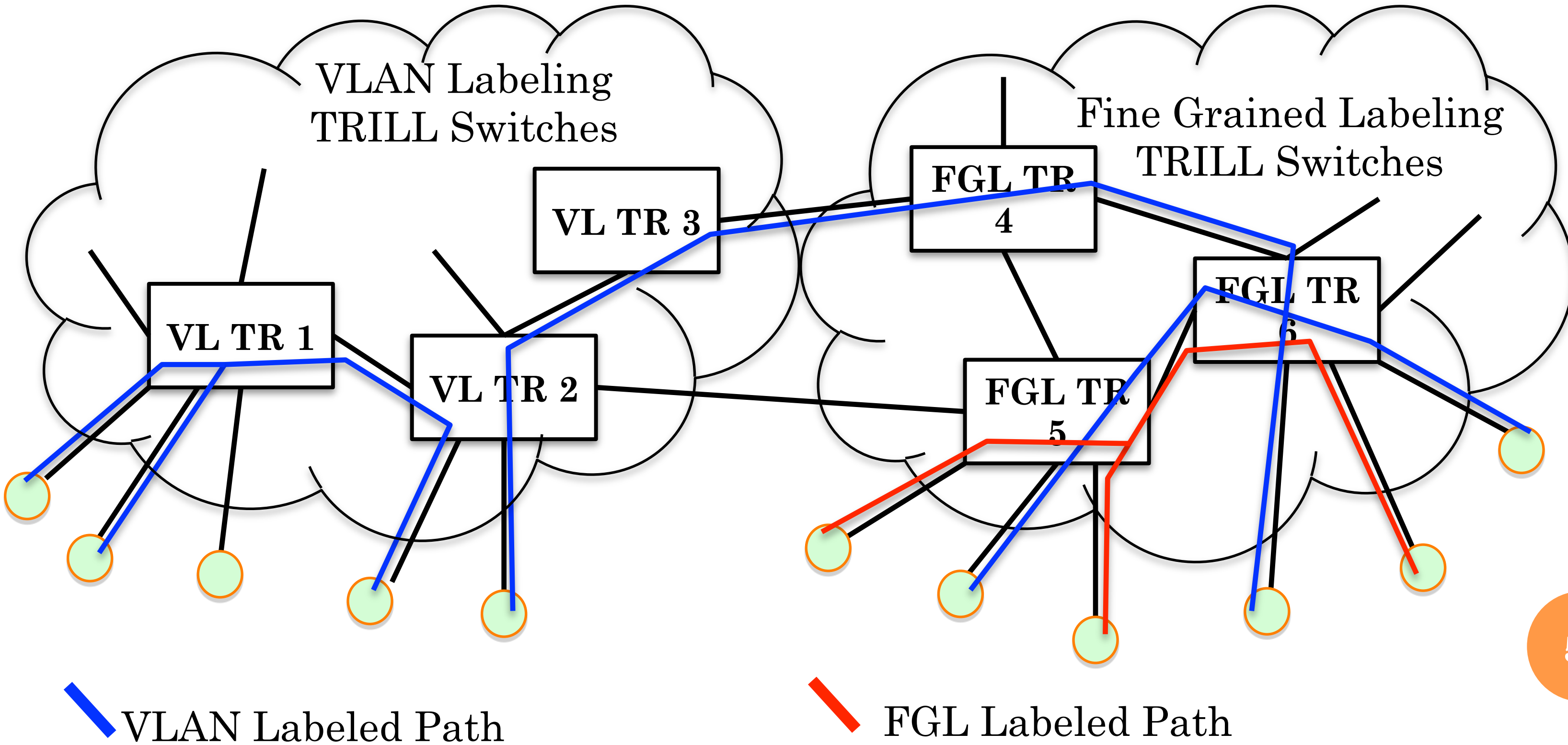
Fine Grained Labeling Migration

- An initial deployment of VLAN labeling TRILL routers can be smoothly extended to Fine Grained Labeling



Fine Grained Labeling Migration

- Some VL TRILL switches are convertible to FGL-safe RBridges (FGL transit only) with a software upgrade.
- Even if not upgradable, they can generally be connected.



CONTENTS

TRILL OAM PROTOCOLS

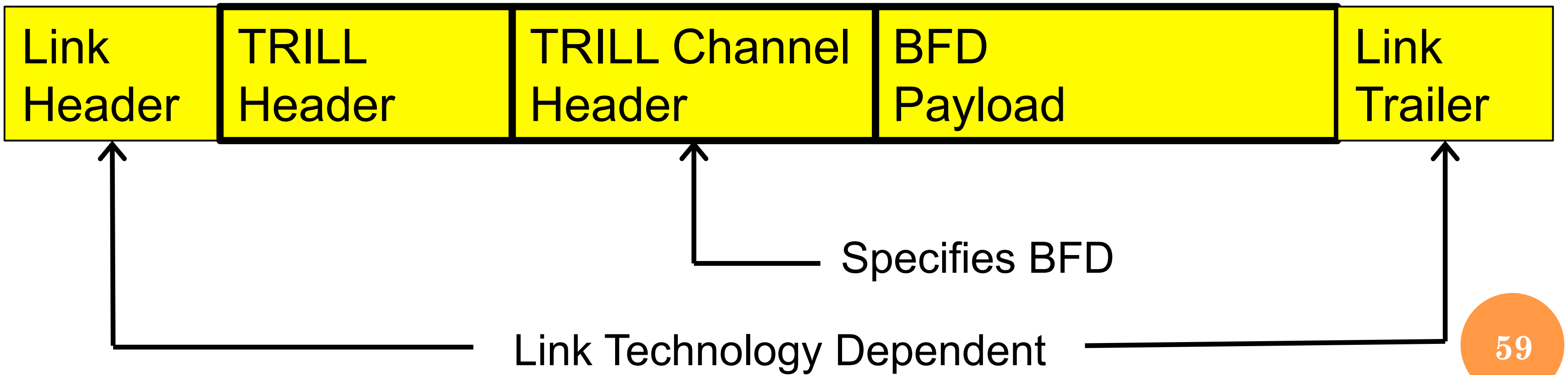
- **SNMP**
  - Used primarily to read the status and configuration of the TRILL switches but can be used to set configuration parameters.
- **BFD over TRILL**
  - Bound to TRILL port at the transmitting TRILL router. Primarily used for rapid one-hop failure detection but multi-hop supported.
- **TRILL OAM**
  - Operates between TRILL switches and is focused on
    1. testing TRILL Data paths (both fault and performance management) and
    2. reporting errors in TRILL Data frames.

OAM Documents Status

- **SNMP**
  - RFC 6850, "Definitions of Managed Objects for RBridges" (MIB)
- **BFD over TRILL**
  - In RFC Editor's queue:
    - draft-ietf-trill-rbridge-bfd-07.txt
    - draft-ietf-trill-rbridge-channel-08.txt
- **TRILL OAM**
  - draft-ietf-trill-oam-req-05.txt (Requirements)
  - draft-ietf-trill-oam-framework-00.txt (Framework)
  - draft-tissa-trill-oam-fm-00.txt (Fault Management)
  - TBD (Performance Management)

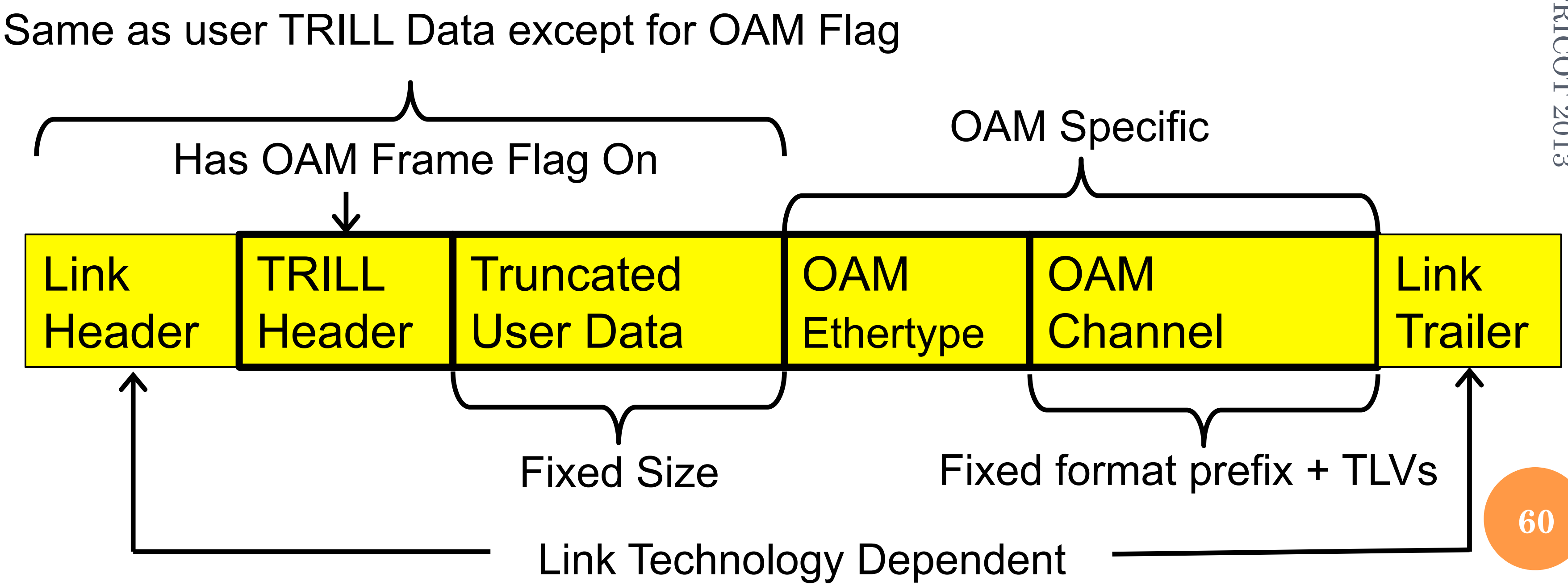
TRILL BFD Frame Format

- The BFD standard does not specify an envelope. One must be specified for each technology using BFD.
- The TRILL BFD envelope uses the RBridge Channel facility, a general method for sending typed messages between TRILL routers.



TRILL OAM Frame Format

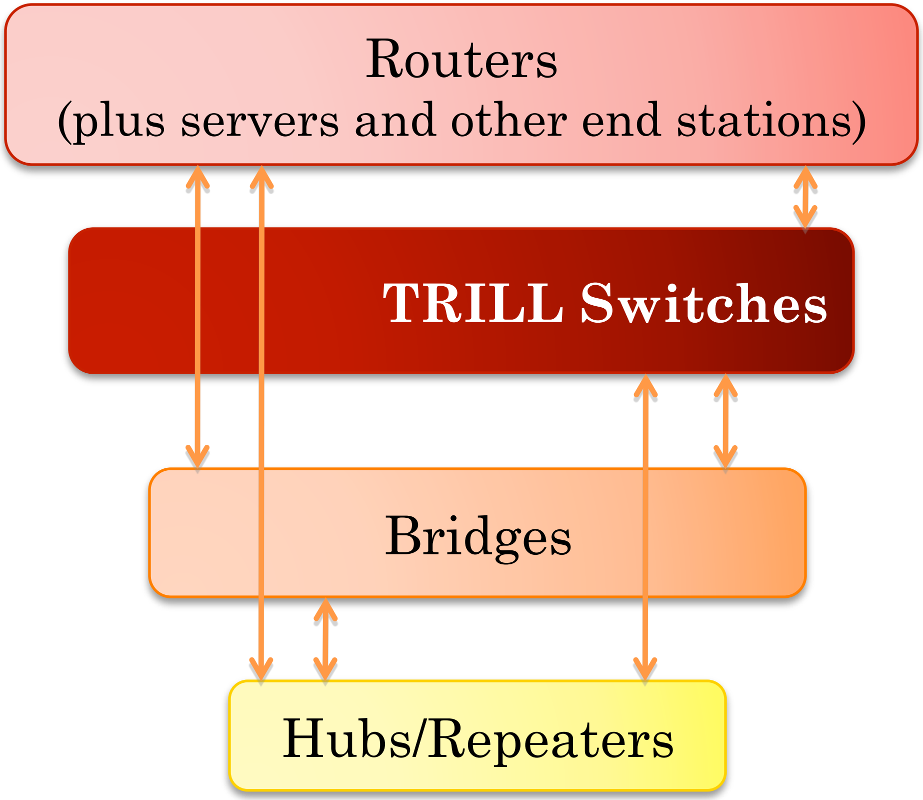
- Because TRILL OAM frames must be able to follow the same paths and get the same processing as TRILL Data frames, their format is very similar.



CONTENTS

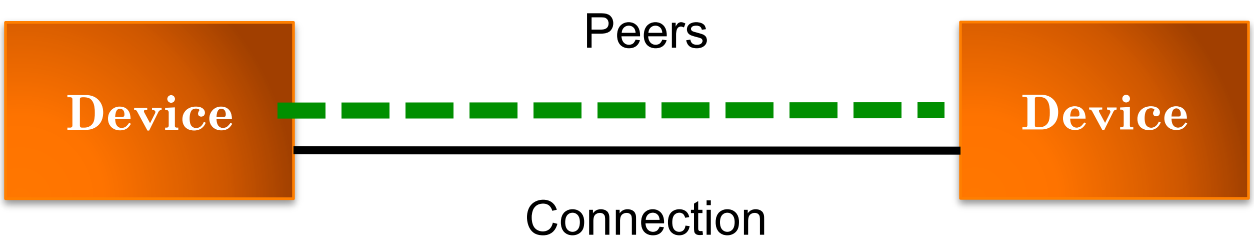
Peering: Are RBridges Bridges or Routers?

- Really, they are a new species, between IEEE 802.1 bridges and routers:



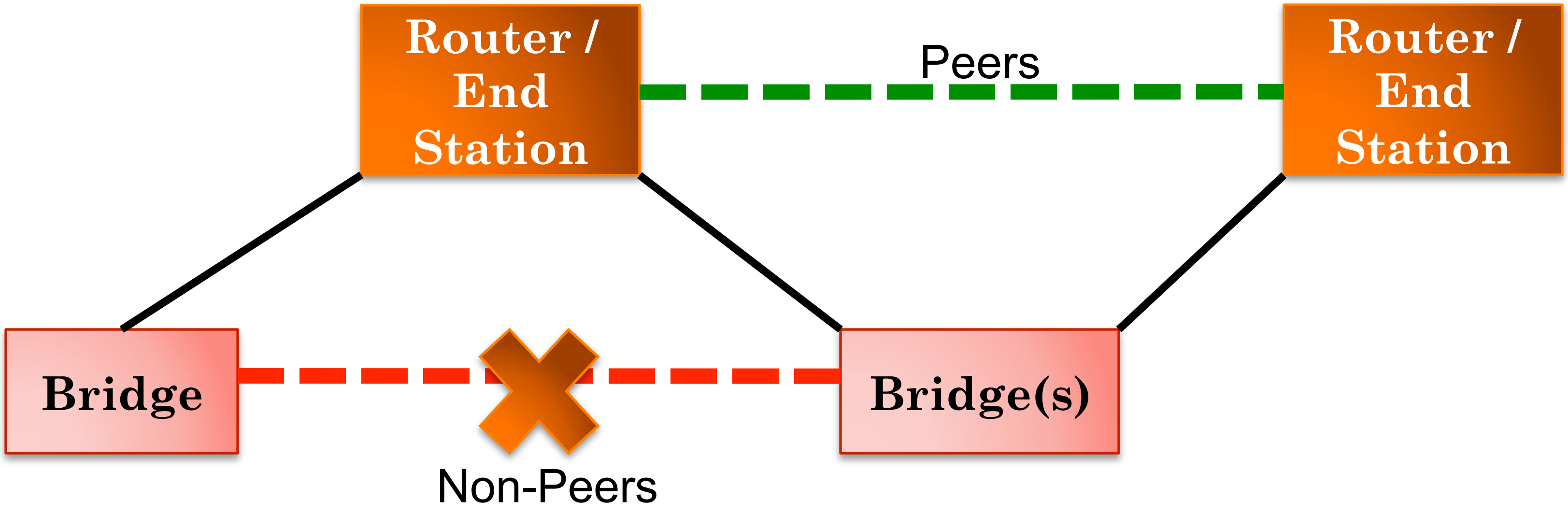
Peering

- Direct Connection



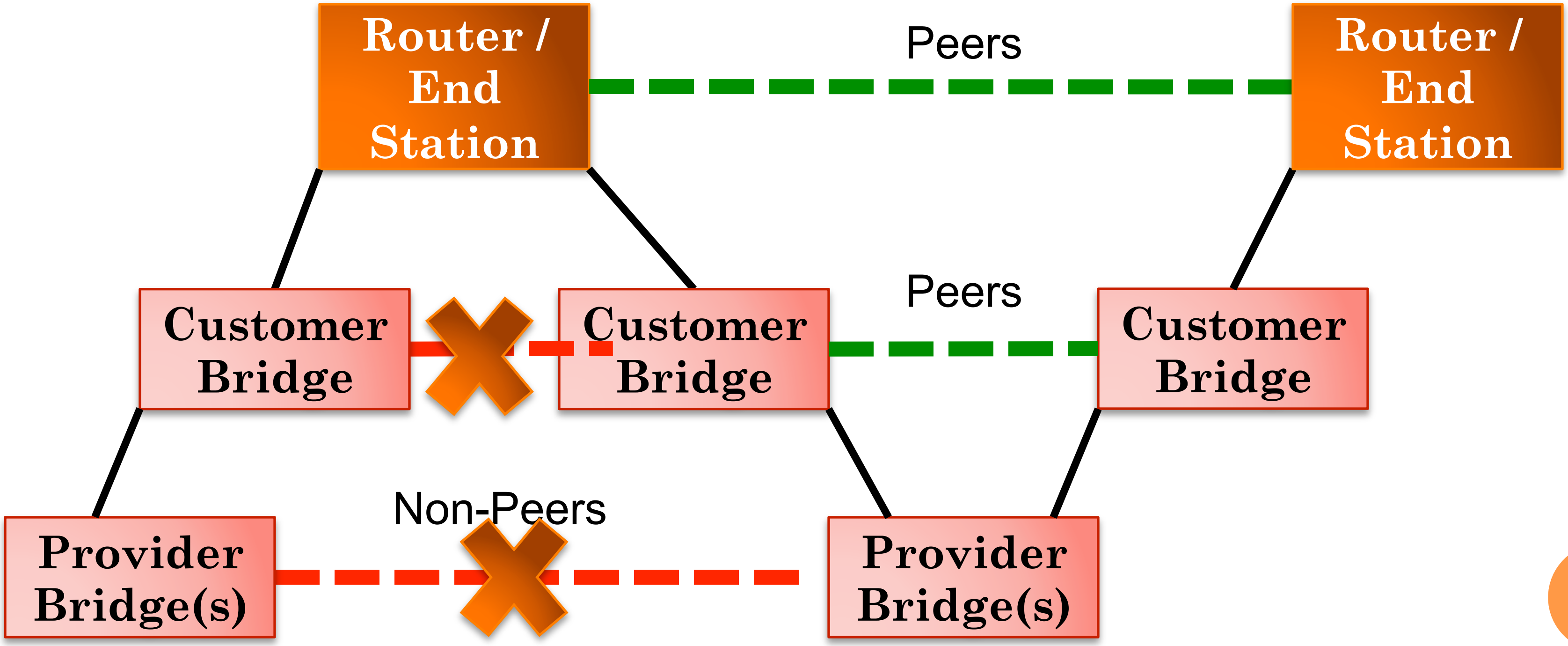
Peering

- Former Situation



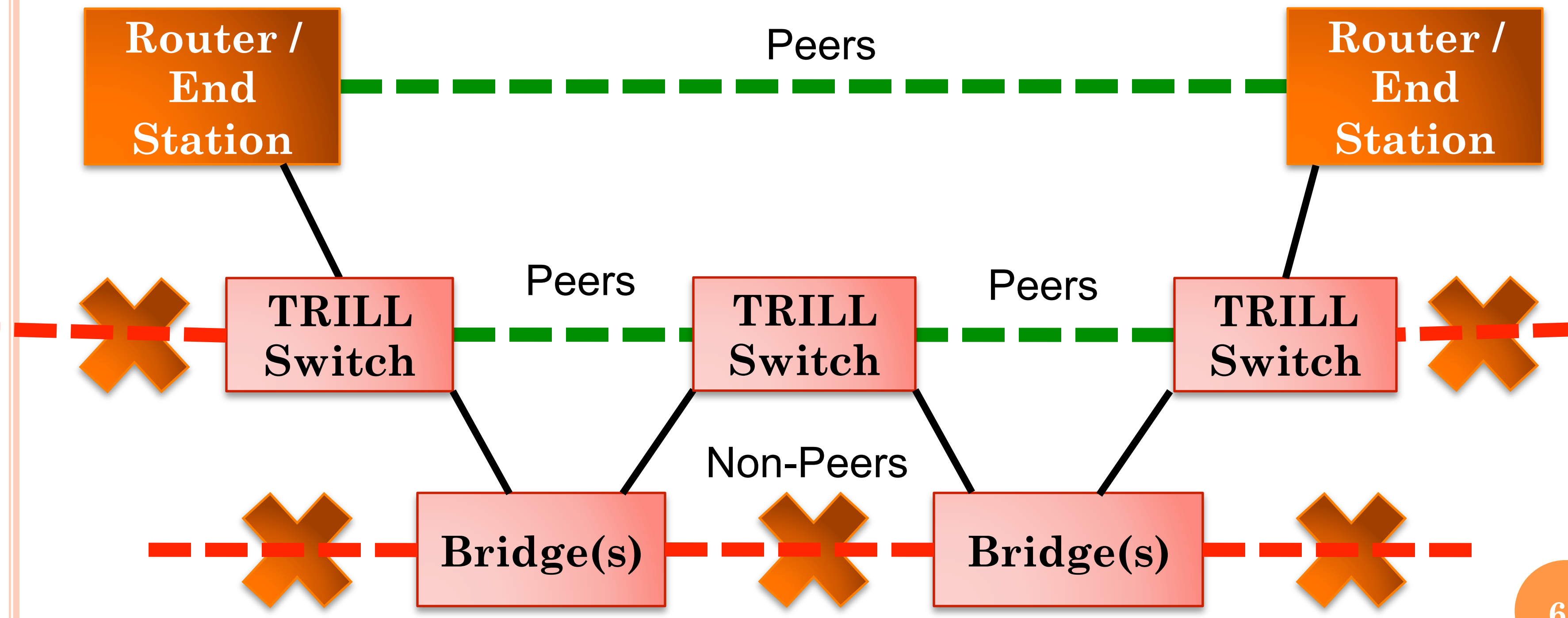
Peering

- Former Situation
  - Or perhaps



Peering

- With RBridges



CONTENTS

Comparison with 802.1aq

- TRILL is an IETF project with the base protocol specification (RFC 6325) approved 15 March 2010.
- Shortest Path Bridging (SPB) is IEEE 802.1aq, a project of the IEEE 802.1 Working Group, approved as an IEEE standard in March 2012.

Comparison with 801.1aq

- OAM
  - SPB: Supports SNMP and Continuity Fault Management (CFM).
  - TRILL: Currently supports SNMP and BFD (Bidirectional Forwarding Detection) with additional OAM under development.
- Data Label Granularity
  - SPB: Supports 4K VLANs or 2**24 Service Identifiers.
  - TRILL: Supports 4K VLANs with support for 2**24 Fine Grained Labels progressing well toward standardization.

Frame Overhead Details

- For point-to-point links with multi-pathing:
  - TRILL:
    - 20 bytes for Ethernet (+ 8 TRILL Header (including Ethertype) + 12 outer MAC addresses)
    - 8 bytes for PPP
  - SPBM:
    - 22 bytes for Ethernet (+ 18 802.1ah tag – 12 for MAC addresses inside 802.1ah + 4 B-VLAN + 12 outer MAC addresses)
    - 24 bytes for Ethernet over PPP, native PPP not supported
- For complex multi-access links with multi-pathing:
  - TRILL: 24 bytes (20 + 4 for outer VLAN tag)
  - SPBM: multi-access links not supported

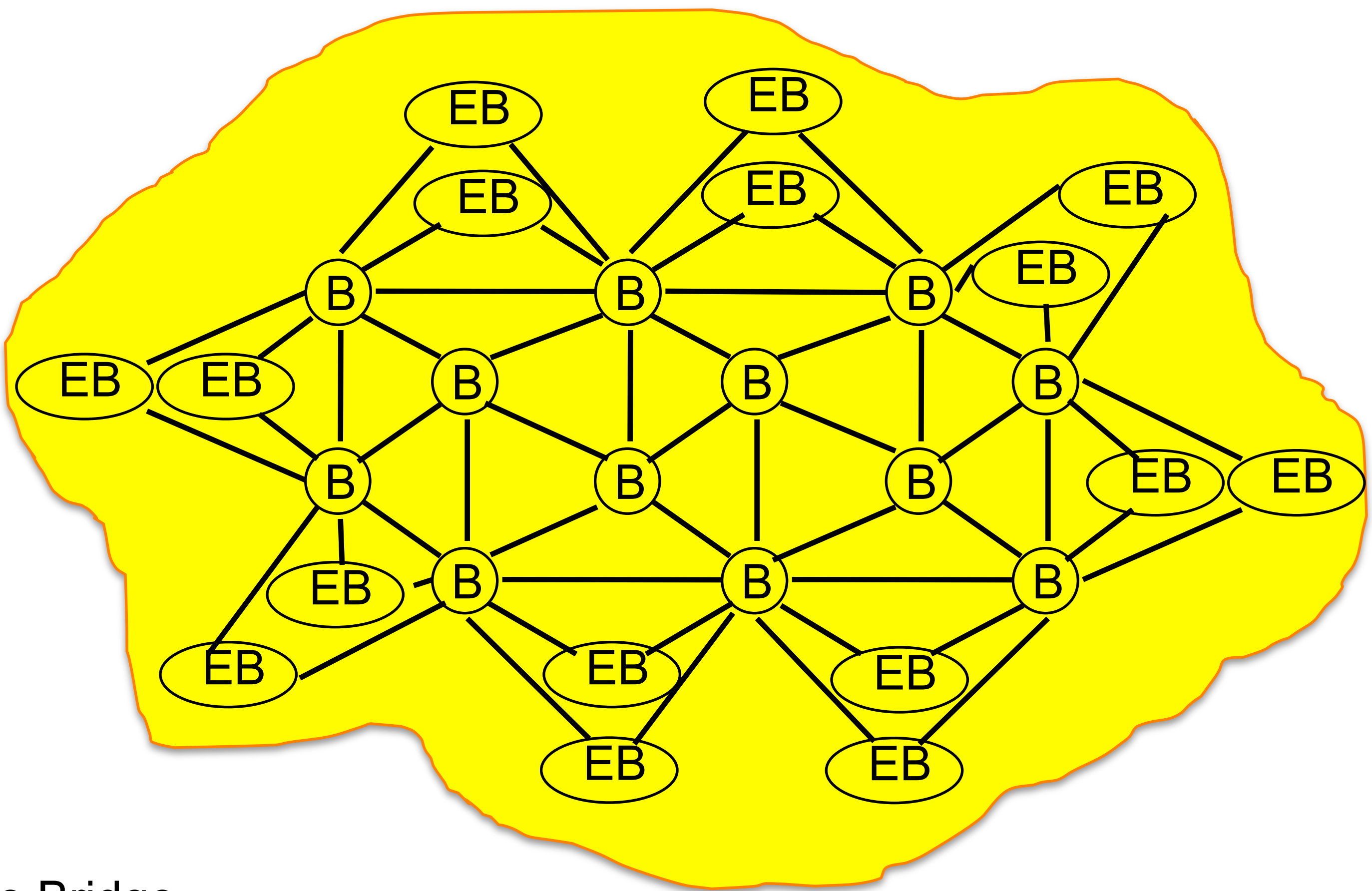
Routing Computation

- N = number of switches
  k = number of multi-paths
- IETF TRILL
  - For unicast frames, $O(N \times \log(N))$.
    - Arbitrary multi-pathing available by just keeping track of equal cost paths.
  - For multi-destination frames, $O(k \times N \times \log(N))$ to have k distribution trees available.
- IEEE 802.1aq
  - Unicast and multi-destination unified: $O(k \times N^2 \times \log(N))$ for k-way multi-pathing.

Comparison with 801.1aq

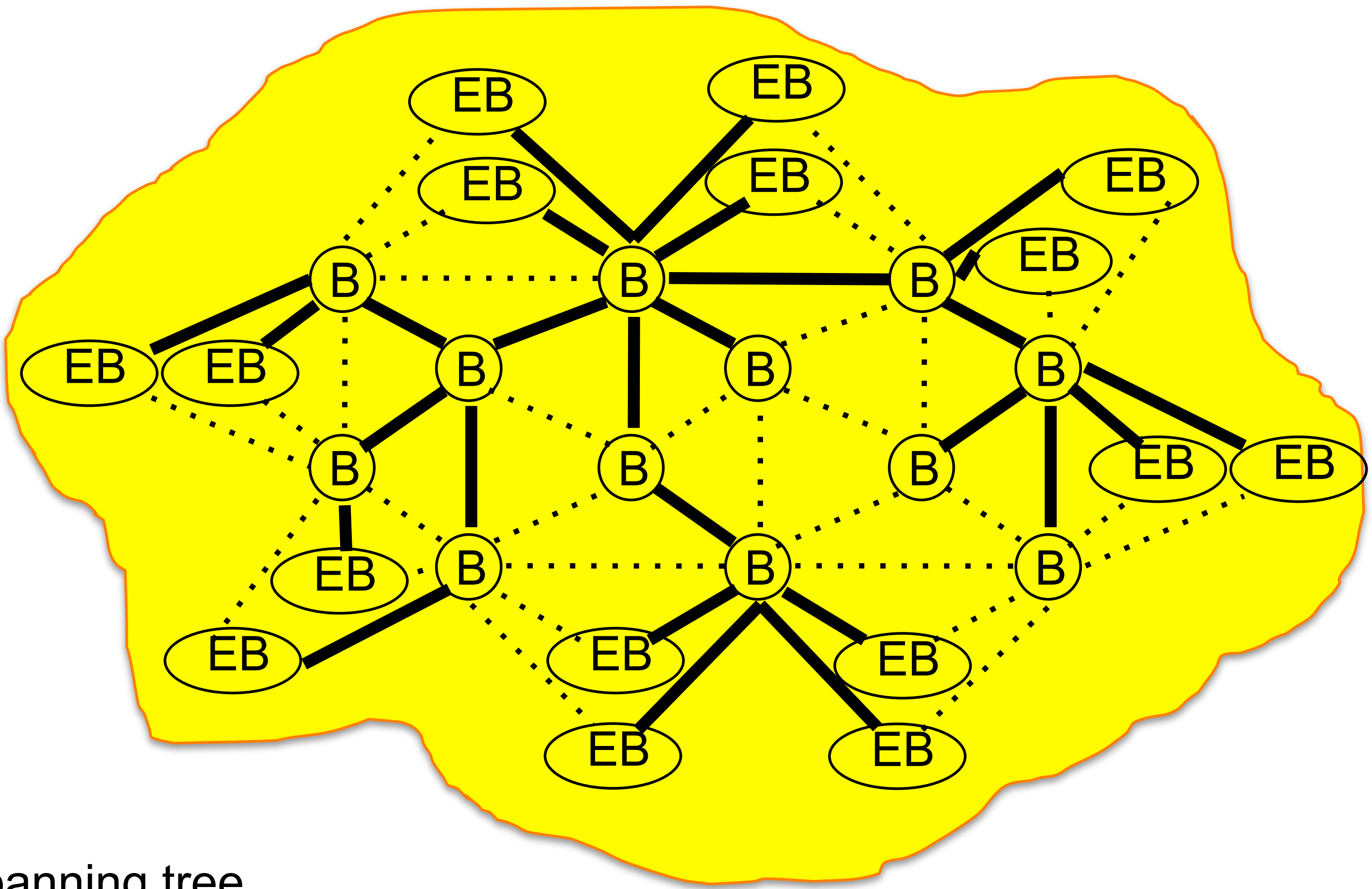
- Peering:
  - TRILL peers through any intervening bridges.
  - SPB bridges must be directly connected and only peer within a contiguous SPB region.
- Spanning Tree:
  - TRILL blocks spanning tree and provides a new level above all bridging but below Layer 3 routing.
  - SPB bridges run at the bridging level. It continues to maintain a spanning tree (or multiple spanning trees) hooking together any attached bridging to produce a bigger spanning tree. Frames are forwarded by spanning tree or by shortest path depending on VLAN.

Comparison with 801.1aq



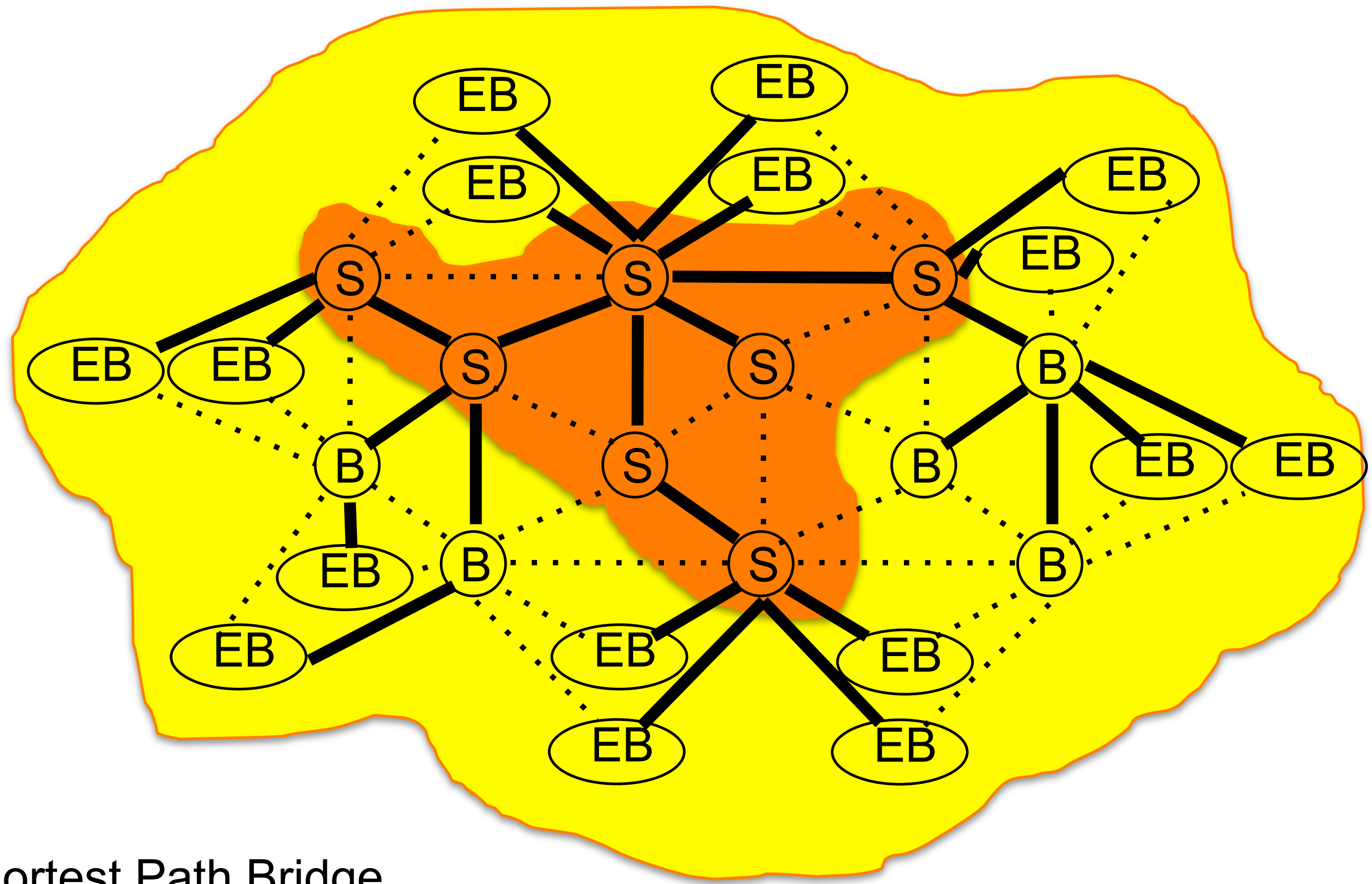
B = core Bridge
EB = edge Bridge – where many end stations are connected
Yellow = Ordinary Bridging

Comparison with 801.1aq



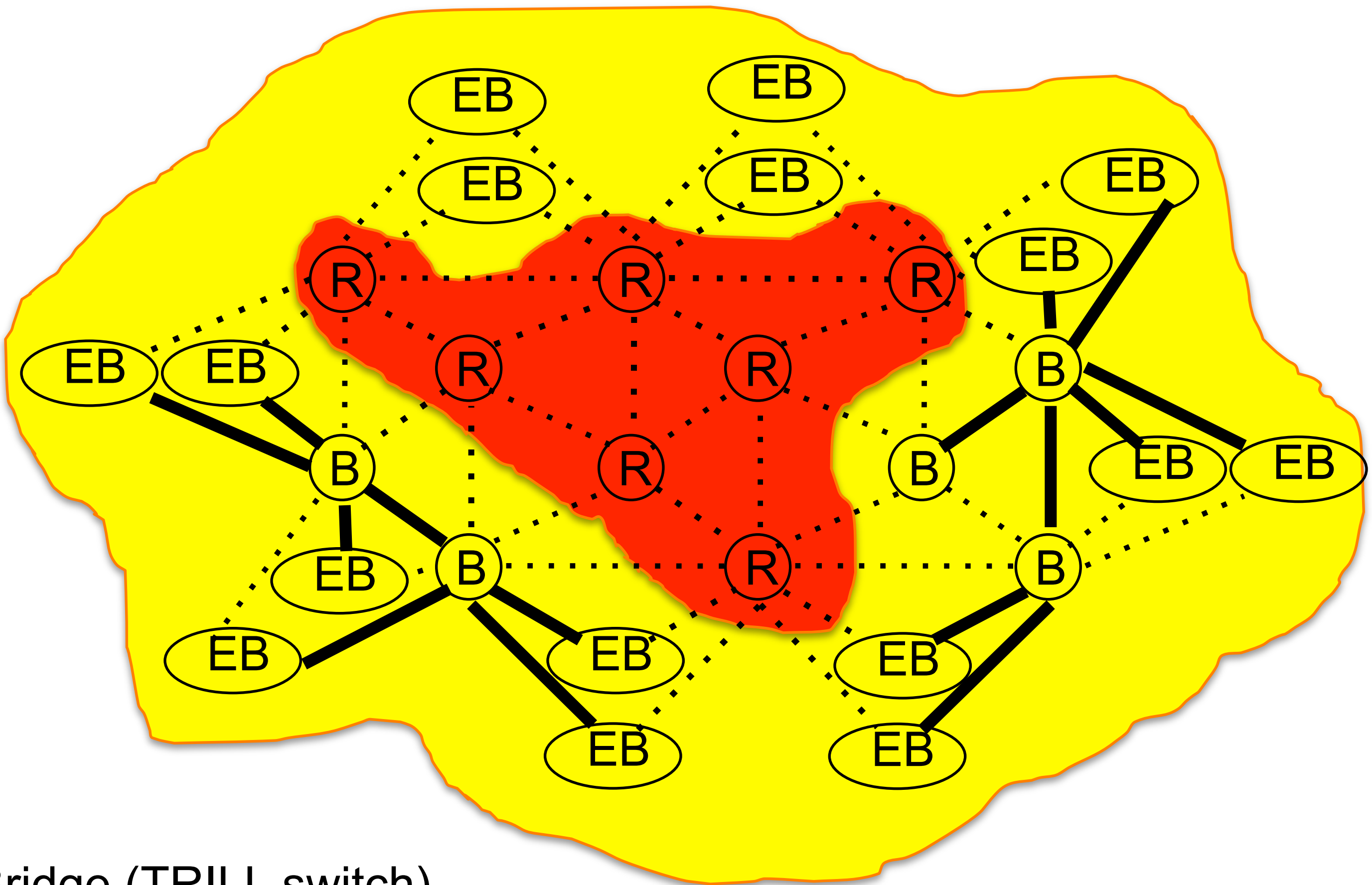
One spanning tree
(there could be multiple)

Comparison with 801.1aq



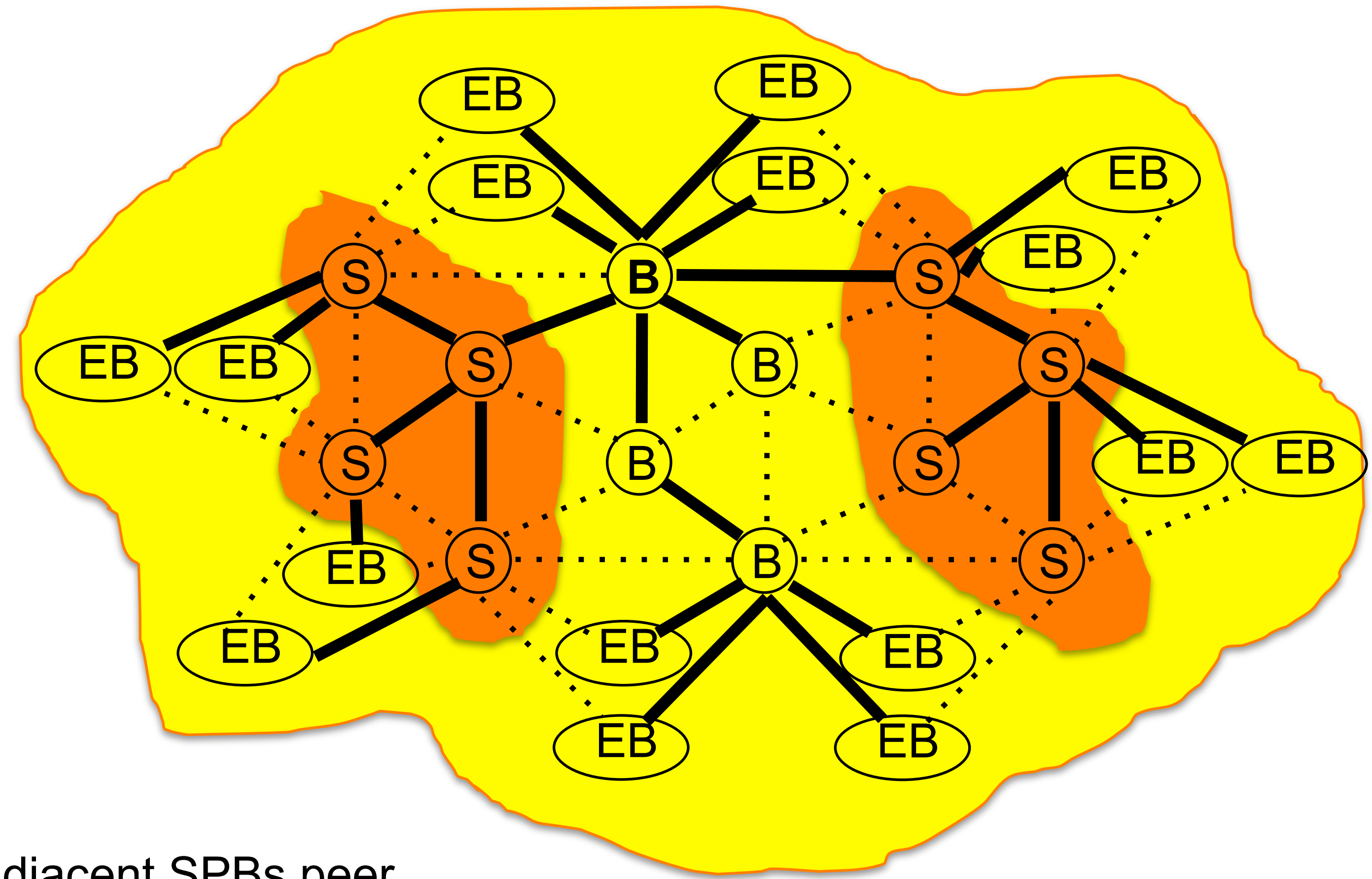
S = Shortest Path Bridge
Orange = SPB Region
Spanning Tree Penetrates SPB Regions

Comparison with 801.1aq



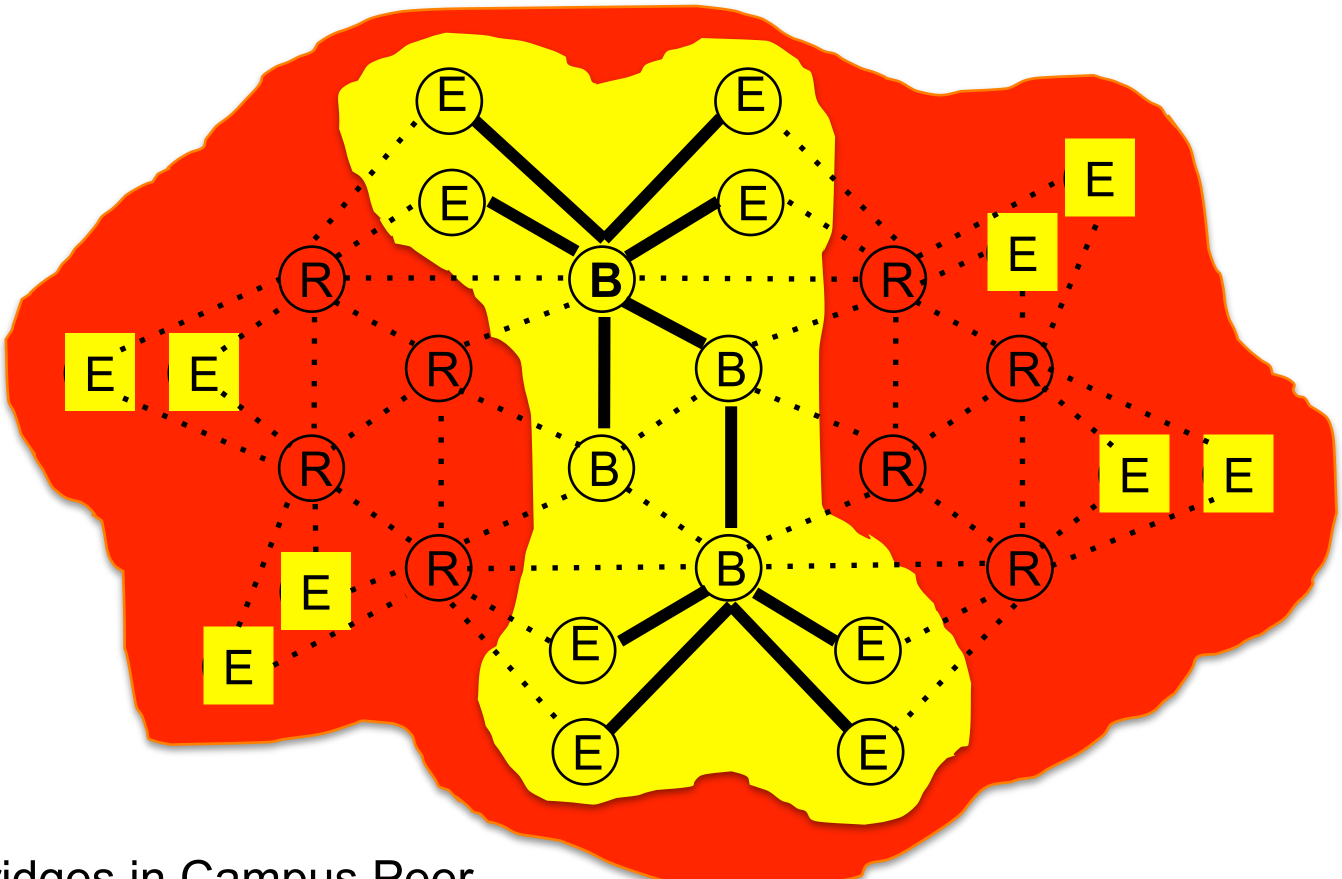
R = RBridge (TRILL switch)
Spanning Tree terminated by RBridges

Comparison with 801.1aq



Only adjacent SPBs peer
Spanning Tree Penetrates SPB Regions

Comparison with 801.1aq



All RBridges in Campus Peer
Spanning Tree terminated by RBridges
(Not recommended to have multiple islands of RBridges. Best to deploy from the core outwards.)

CONTENTS

TRILL PRODUCTS

- "Pre-standard" products
  - Cisco FabricPath
  - Brocade VCS
- Some TRILL Standard Products Announced on the Web
  - Broadcom – StrataXGS Trident (BMC5680)
  - Cisco – Nexus, Catalyst 6500
  - IBM / Blade Networks – RackSwitch G8264
  - HP – 5900 ToR Switches
  - Huawei – Cloud Engine 5800, 12800
  - Mellanox – SwitchX
  - ZTE – ZXR10 5800 Series

TRILL SILICON

- Here are six publicly known independent silicon implementations of the TRILL Fast Path. In some cases there are multiple different chips.
  - Broadcom – merchant silicon
  - Brocade – products
  - Cisco – products
  - Fulcrum – merchant silicon
  - Marvell – merchant silicon
  - Mellanox – merchant silicon

Open Source TRILL

- Oracle: TRILL for Solaris
  - http://hub.opensolaris.org/bin/view/Project+rbridges/WebHome
- TRILL Port to Linux (in process):
  National University of Sciences and Technology (NUST),
- Dr. Ali Khayam
- Islamabad, Pakistan

CONTENTS

Standardization Status

- The TRILL protocol RFCs
  - RFC 5556, "TRILL Problem and Applicability"
  - **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"
  - RFC 6847, "FCoE over TRILL"
  - RFC 6850, "Definitions of Managed Objects for RBridges" (MIB)

STANDARDIZATION STATUS

- Ethertypes assigned by IEEE Registration Authority:
  - TRILL Data: 0x22F3
  - L2-IS-IS: 0x22F4
  - Fine Grained Labeling: 0x893B
  - RBridge Channel: 0x8946
- Block of TRILL multicast addresses assigned by IEEE Registration Authority:
  - 01-80-C2-00-00-40 to 01-80-C2-00-00-4F
- TRILL NLPID (Network Layer Protocol ID) assigned from ISO/IEC: 0xC0

STANDARDIZATION STATUS

- Document that are fully approved and in the RFC Editor's Queue. These are expected to issue as RFCs:
  - "TRILL: Clarifications, Corrections, and Updates"
    - https://datatracker.ietf.org/doc/draft-ietf-trill-clear-correct/
  - "TRILL: Bidirectional Forwarding Detection (BFD) Support"
    - https://datatracker.ietf.org/doc/draft-ietf-trill-rbridge-bfd/
  - "TRILL: RBridge Channel Support"
    - https://datatracker.ietf.org/doc/draft-ietf-trill-rbridge-channel/
  - "TRILL: Header Extension"
    - https://datatracker.ietf.org/doc/draft-ietf-trill-rbridge-extension/

CONTENTS

TRILL References

- List of TRILL and closely related IETF RFCs: 5556, 6165, **6325**, 6326, 6327, 6329, 6361, 6439, 6847, 6850
  http://www.rfc-editor.org
- Fully approved Internet Drafts not yet published as IETF RFCs, see slide #86 with URLs
- All TRILL WG drafts:
  - https://datatracker.ietf.org/doc/search/?name=&activeDrafts=on&by=group&group=trill&search_submit=

More TRILL References

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

- The initial 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

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.
- Spanning Tree Protocol (Bridging) forwards data out all ports (except the one where the data was received) unless it receives control messages on that 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 SUPPORT OF DCB

- The goal is "loss-less" Ethernet. That is, no loss due to queue overflow

- Basic Ethernet PAUSE "works", but is a very blunt instrument
  - Interference with loss dependent flow control such as TCP
  - Blocking of high priority control frames
  - Congestion spreading

TRILL SUPPORT OF DCB

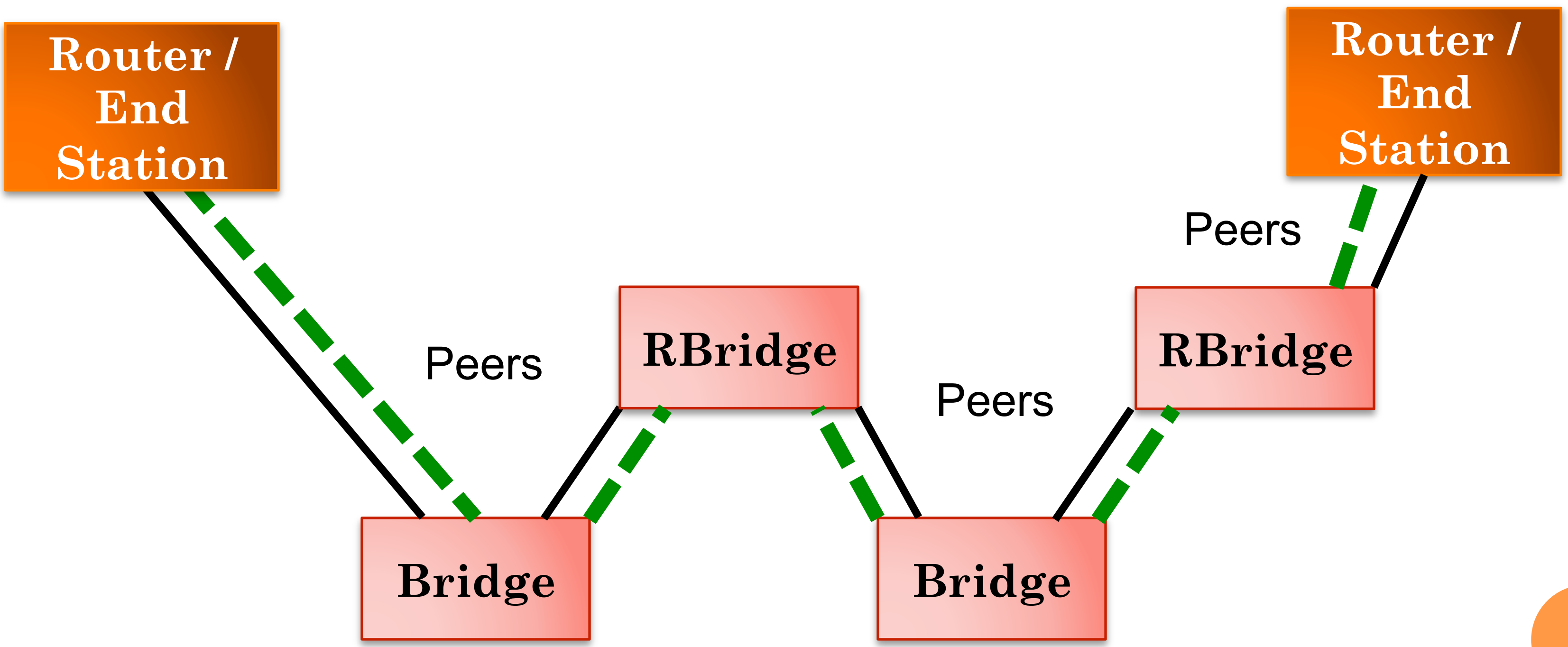
- Answer 1:
  - Consider different frame priorities as different pipes
    - Priority Based Flow Control (802.1Qbb, PFC), Separate PAUSE per priority
    - Enhanced Transmission Selection (802.1Qaz, ETS), Ability to allocate bandwidth between these pipes
- Answer 2:
  - Provide back pressure on the origin of congesting flows
    - 802.1Qau (CN): Congestion Notification
- Some people think CN is better but that you also need PFC as a backup.

TRILL Support of DCB

- Answer 1: Consider different frame priorities as different pipes
  - PFC: Separate PAUSE per priority
    - Don’t enable for priorities where urgent control frames are sent or where loss dependent flow control is in use
    - Enable for priorities where loss-less flow is more important.
  - ETS: Ability to allocate bandwidth between these pipes
    - Highest priority frames not restricted
    - Remainder of bandwidth can be carved up and frames can be selected in preference to “higher priority” frames if they have not used the allocation for their pipe.
  - The above are implemented in port queuing. Can be applied to bridges, RBridges, routers, end stations.

TRILL Support of DCB

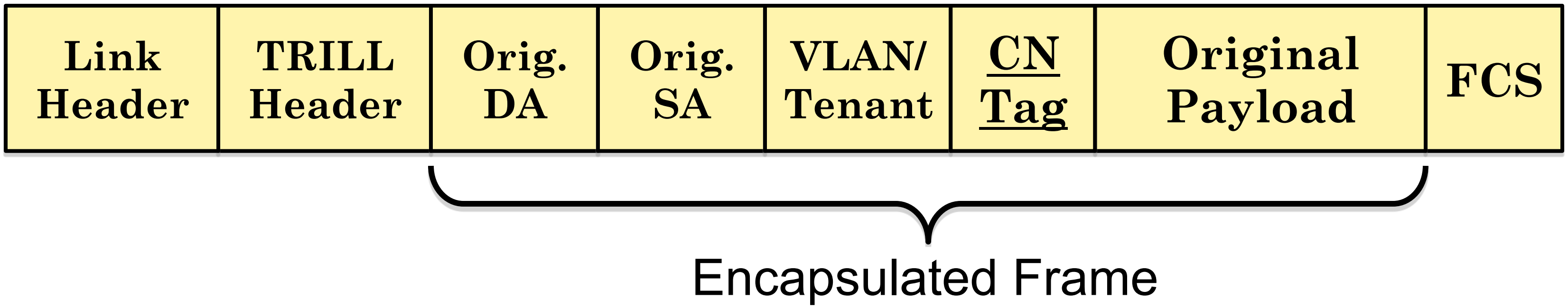
- For PFC and ETS, all successive stations are considered peers:



TRILL Support of DCB

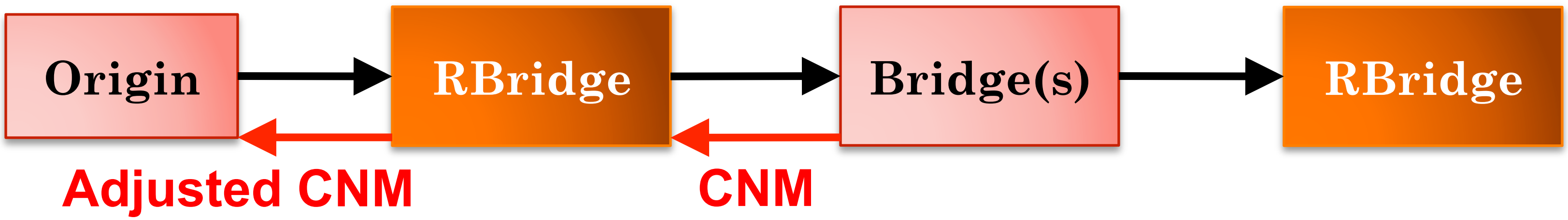
- Answer 2: CN: Provide back pressure on the origin of congesting flows
  - When queue depth exceeds a bound, send a Congestion Notification Message (CNM) back to source MAC address in the congesting frame's VLAN
  - Enabled per priority. (CNM itself usually priority 6.)
  - Frames can be labeled with a CN tag for more fine grained flows
  - Mostly implemented in port logic
- In TRILL a CN tag, if present, goes inside the encapsulated frame and a CNM is just a native frame, except for one corner case.

TRILL Support of DCB



- However, TRILL switches have to handle CNMs generated by TRILL ignorant bridges between TRILL switches. Such a CNM will be initially addressed to the previous hop TRILL switch, not the original end station.

TRILL Support of DCB



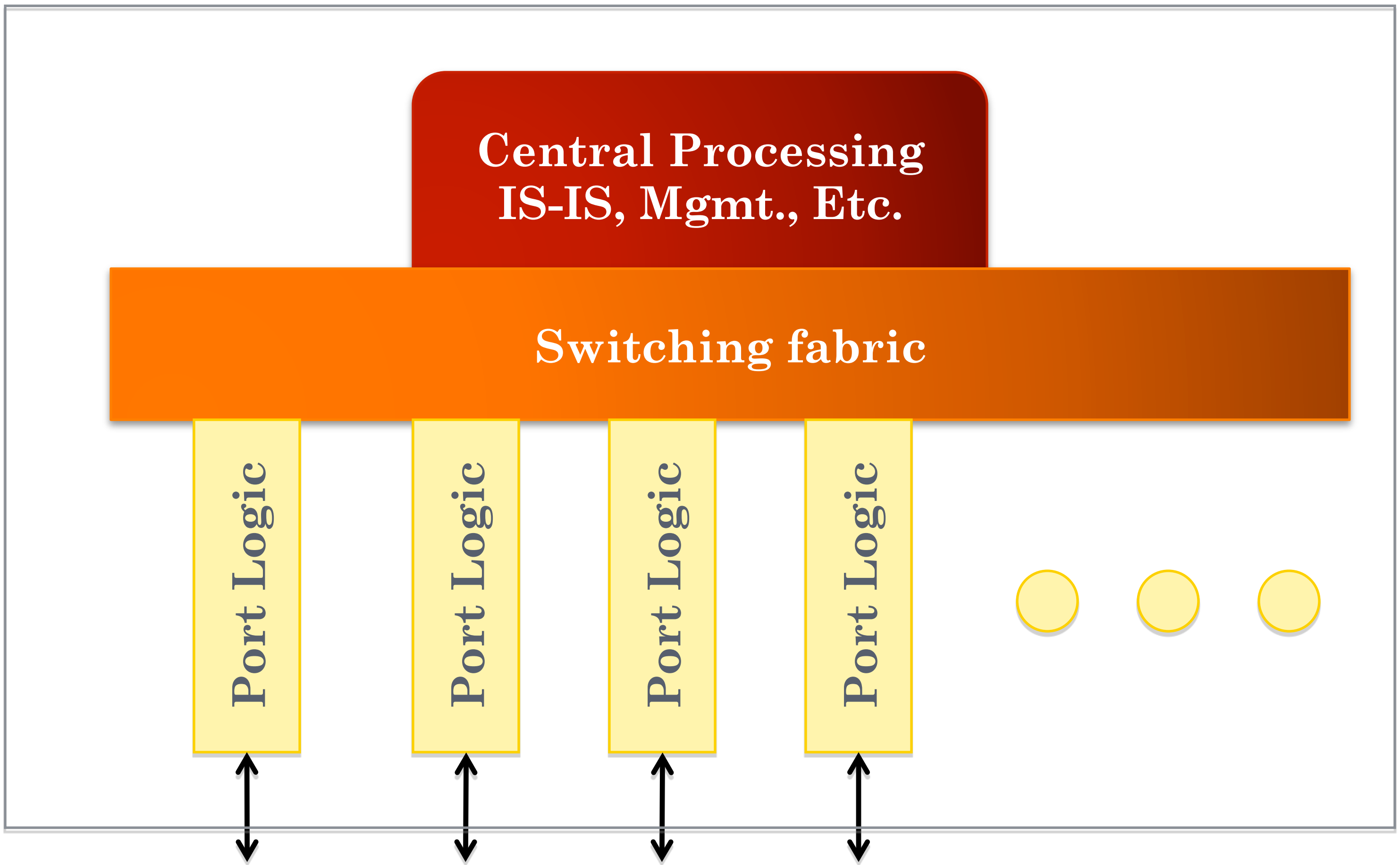
- Previous hop RBridge has to adjust the CNM so that it goes back to the origin end station.
- Note: All of the DCB facilities depend on appropriate engineering, limited delay bandwidth product, etc., to actually provide "loss-less" service.

TRILL Support of DCB

- The Internet draft discussing TRILL support of DCB has been split into two parts under new names, so the version numbers of each part got reset to zero. They are

- For PFC and ETS
  - https://datatracker.ietf.org/doc/draft-eastlake-trill-pfc-ets/
- For Congestion Notification
  - https://datatracker.ietf.org/doc/draft-eastlake-trill-cn/

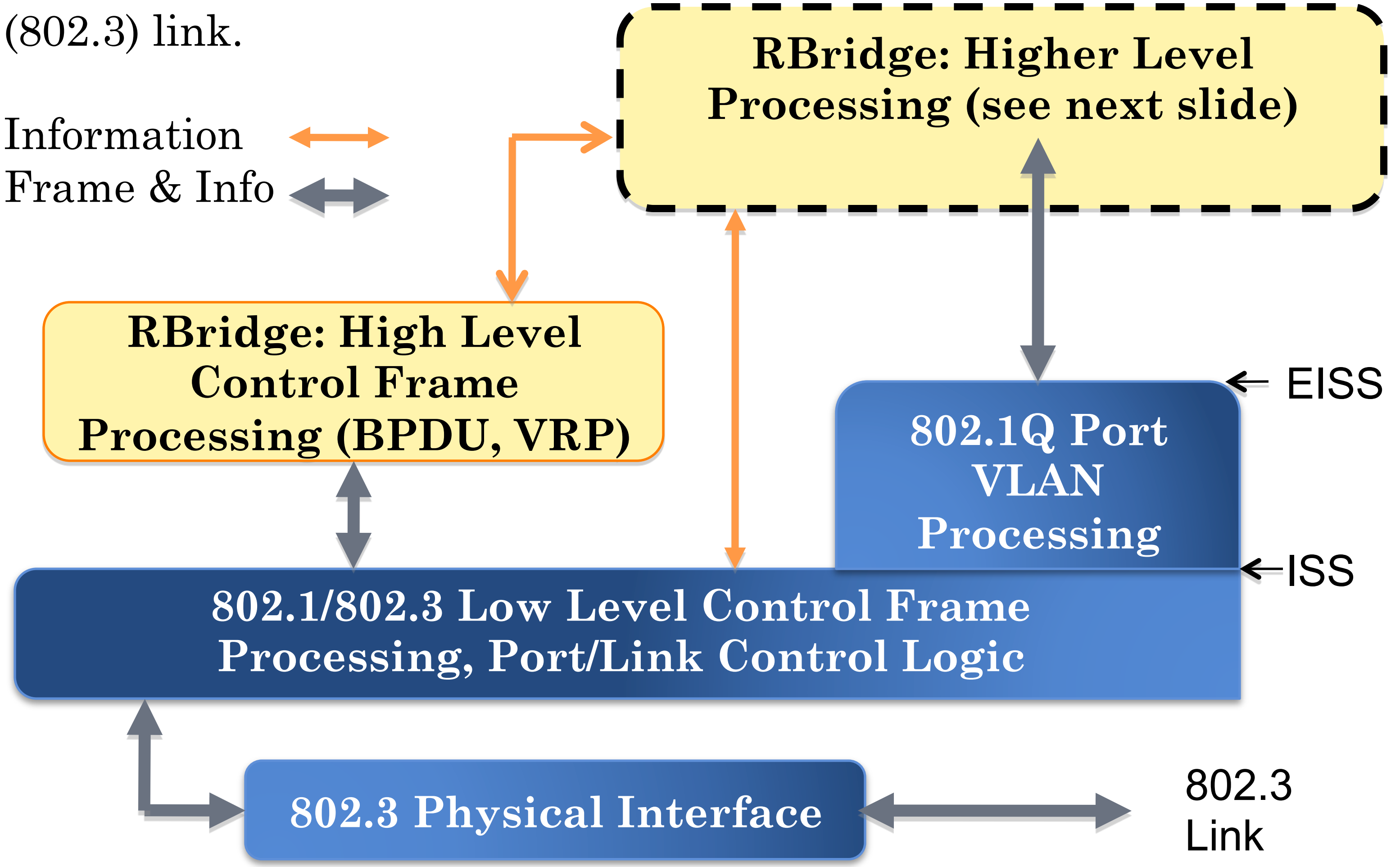
Structure of an RBridge



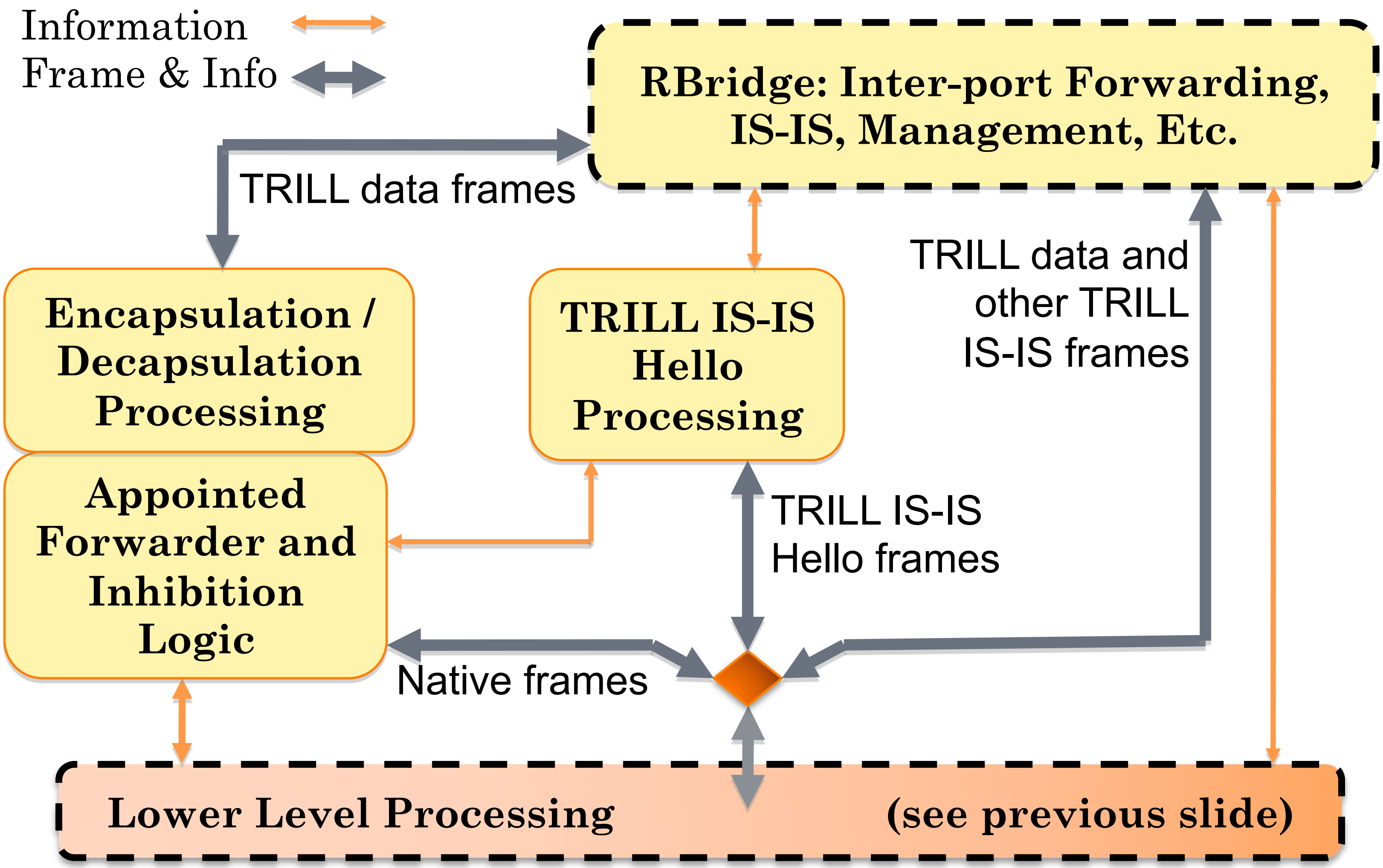
Links to other devices. Could be 802.3 (Ethernet), 802.11 (Wi-Fi), PPP, …

STRUCTURE OF AN RBRIDGE PORT

Assumes an Ethernet (802.3) link.



Structure of an RBridge Port



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

END

Backup Slides

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