RBRIDGES
AND THE TRILL PROTOCOL

Donald E. Eastlake 3rd
Co-Chair, IETF TRILL Working Group
d3e3e3@gmail.com, +1-508-333-2270

NOTE:

- This presentation is a high level technical overview. It is not possible to include all the details in the specification documents.
- The -16 version of the TRILL base protocol specification was approved as an IETF standard on 15 March 2010. The protocol is being implemented by many companies.

CONTENTS

What/Why/Who TRILL?

- What is TRILL?
  - TRILL is a new standard protocol to perform Layer 2 bridging using IS-IS link state routing.

- Who invented TRILL?
  - Radia Perlman of Intel, the inventor of the Spanning Tree Protocol, a major contributor to link-state routing, and the inventor of DECnet Phase V from which IS-IS was copied.

What/Why/Who TRILL?

- TRILL –
  TRansparent Interconnection of Lots of Links
  - A standard specified by the IETF (Internet Engineering Task Force) TRILL Working Group co-chaired by
    - Donald E. Eastlake 3rd
    - Erik Nordmark, Oracle
- RBridge – Routing Bridge
  - A device which implements TRILL
- RBridge Campus –
  - A network of RBridges, links, and any intervening bridges, bounded by end stations / layer 3 routers.

What/Why/Who TRILL?

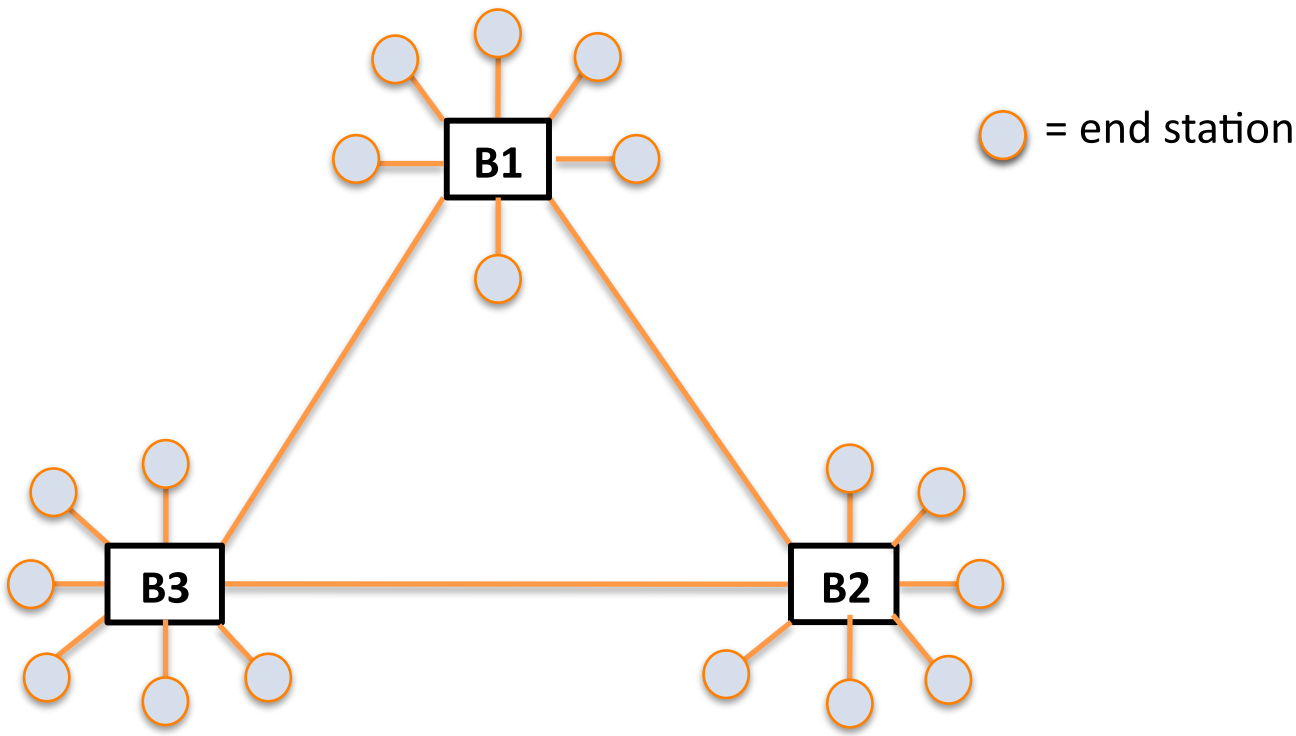
- Basically a simple idea:
  - Encapsulate native frames in a transport header providing a hop count.
  - Route the encapsulated frames using IS-IS.
  - Decapsulate native frames before delivery.

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.

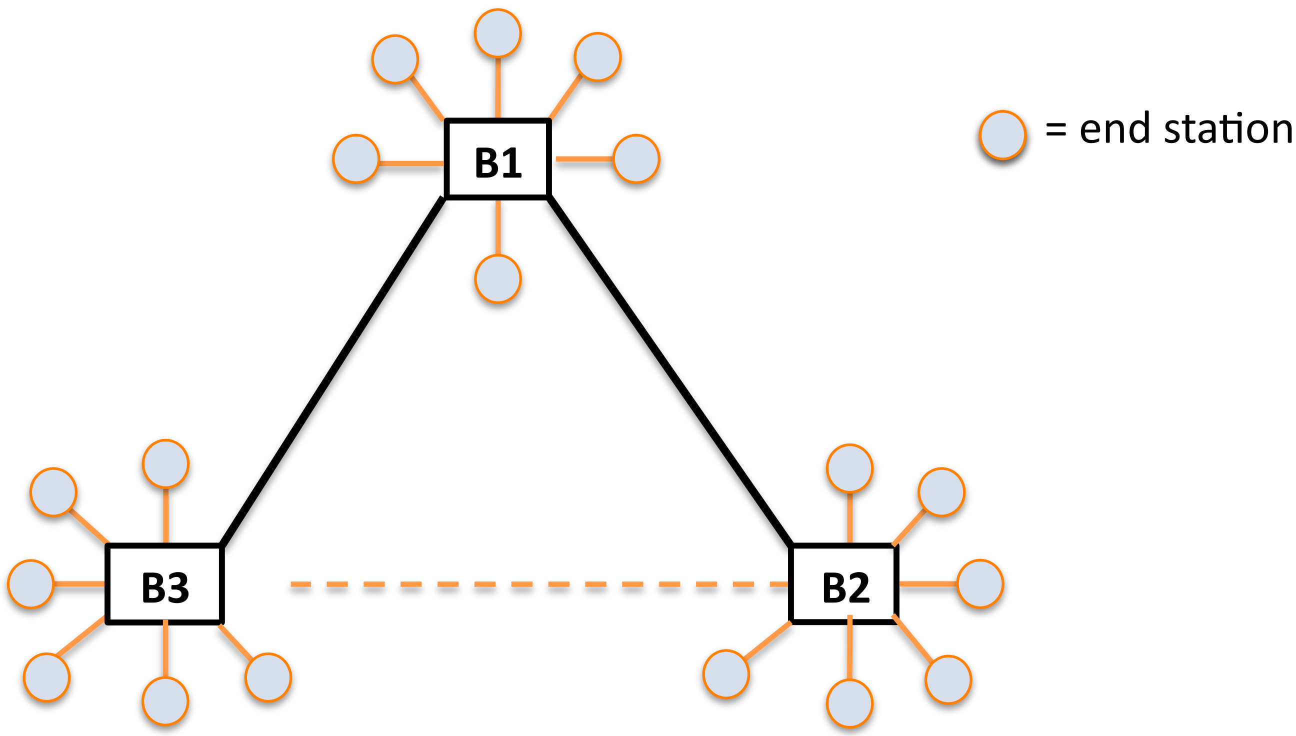
CONTENTS

Optimum Point-to-Point Forwarding



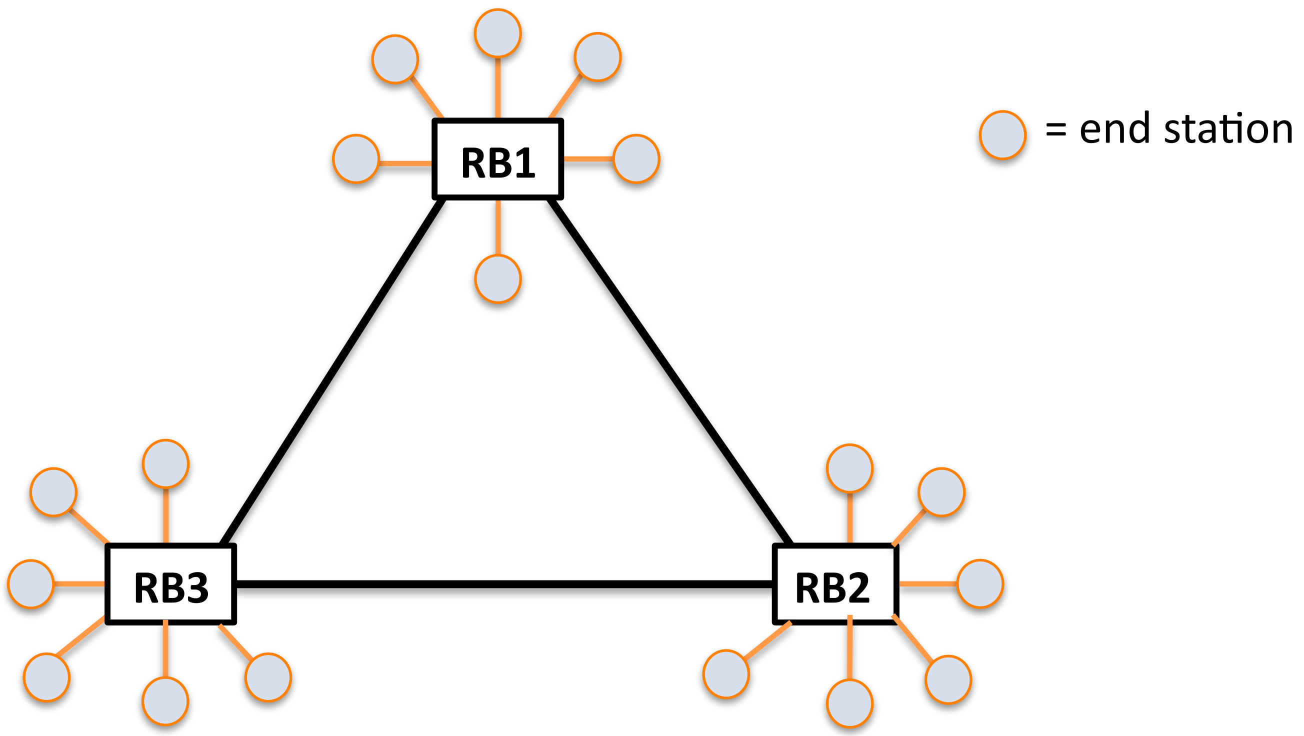
A three bridge network

Optimum Point-to-Point Forwarding



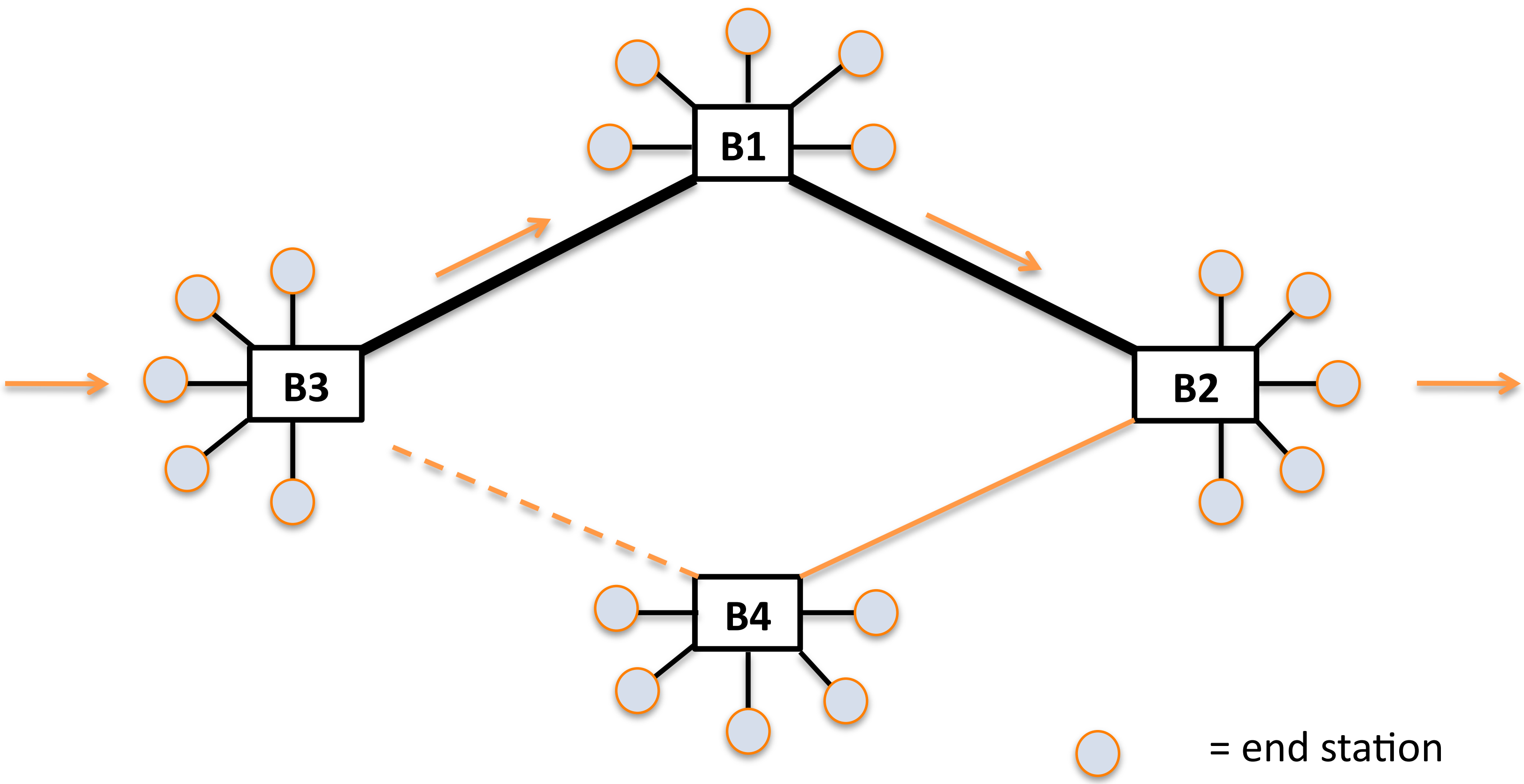
Spanning tree eliminates loops by disabling ports

Optimum Point-to-Point Forwarding



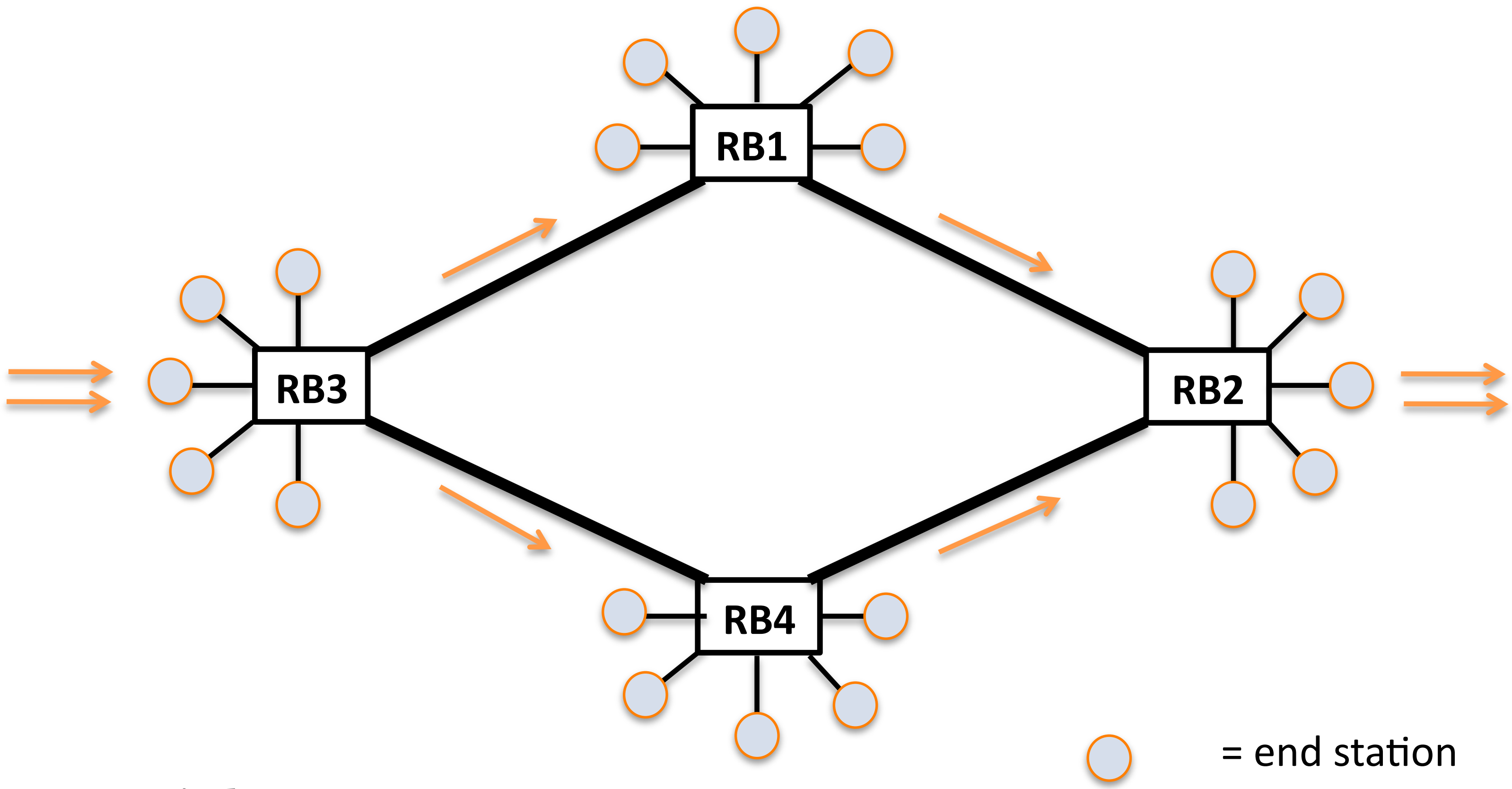
A three RBridge network: better performance using all facilities

MULTI-PATHING



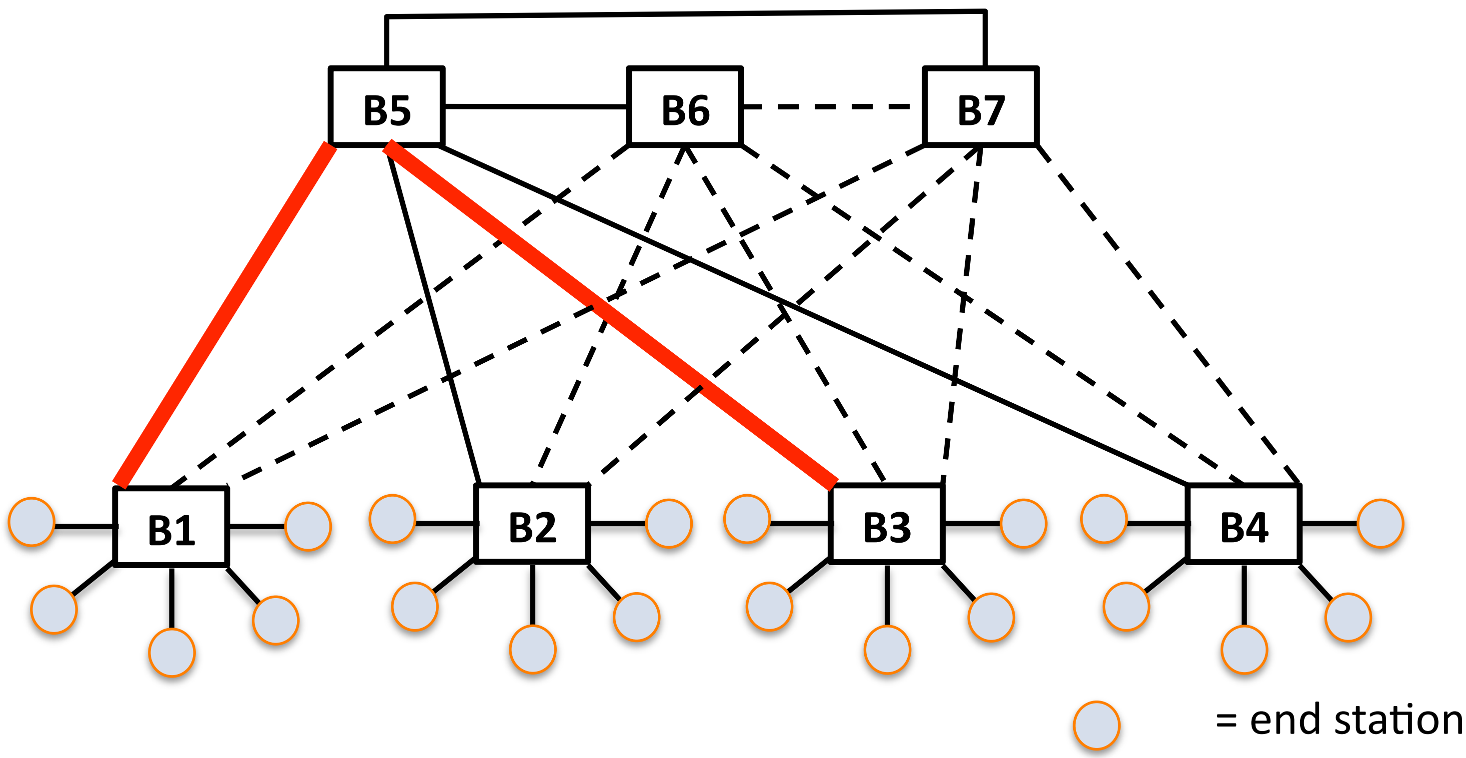
Bridges limit traffic to one path

MULTI-PATHING



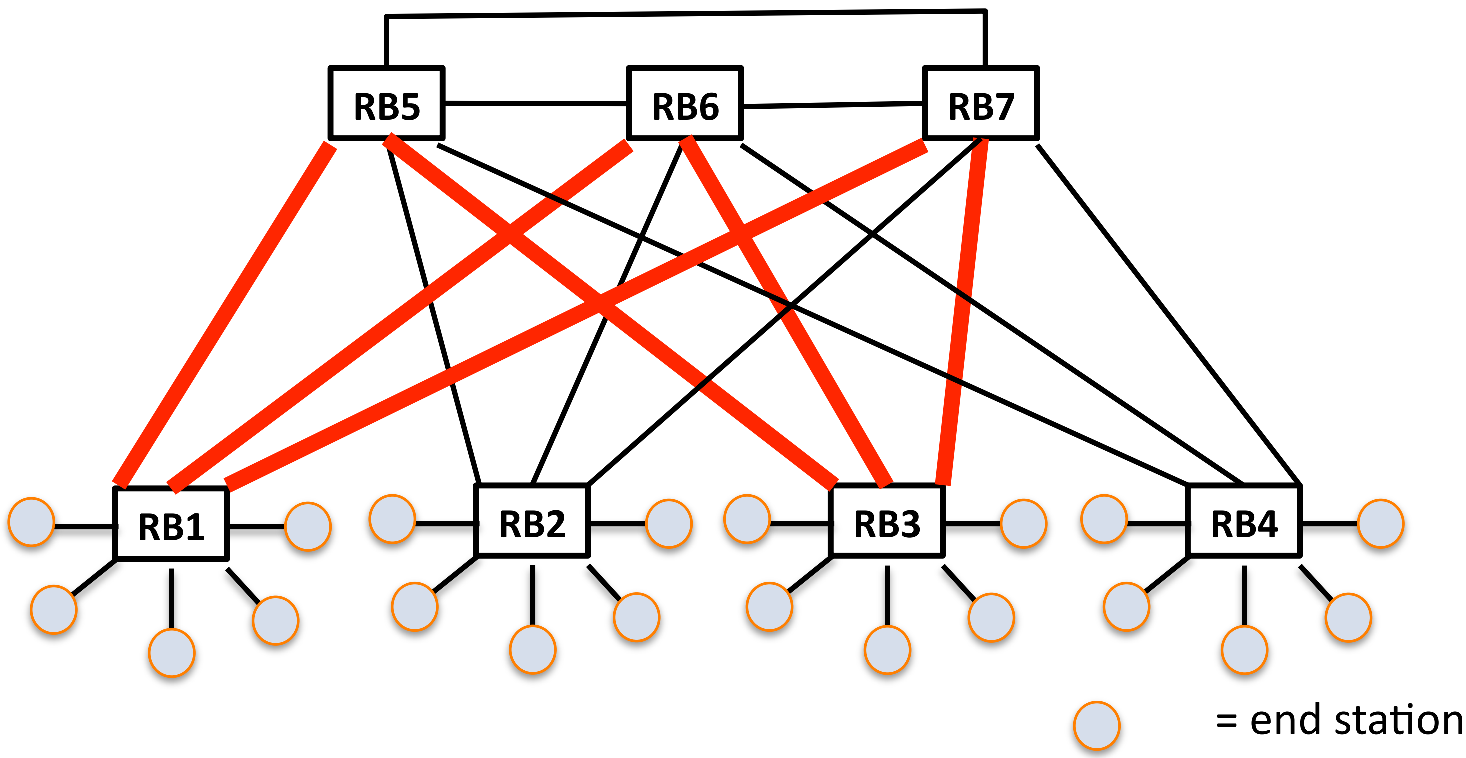
RBridges support
multi-path for higher throughput

Multi-Pathing



Bridges limit traffic to one path

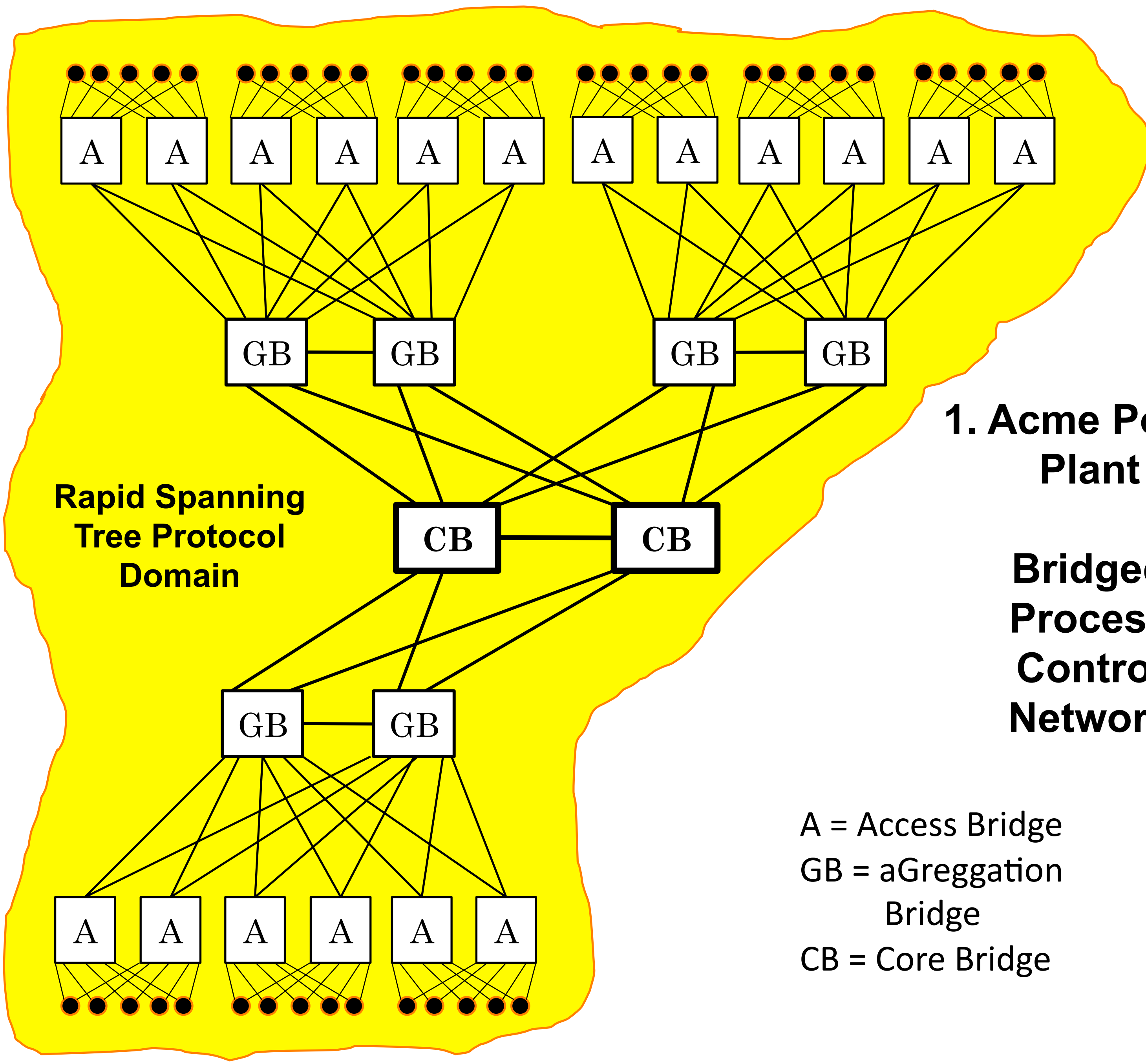
Multi-Pathing



RBridges support
multi-path for higher throughput

Some Other TRILL Features

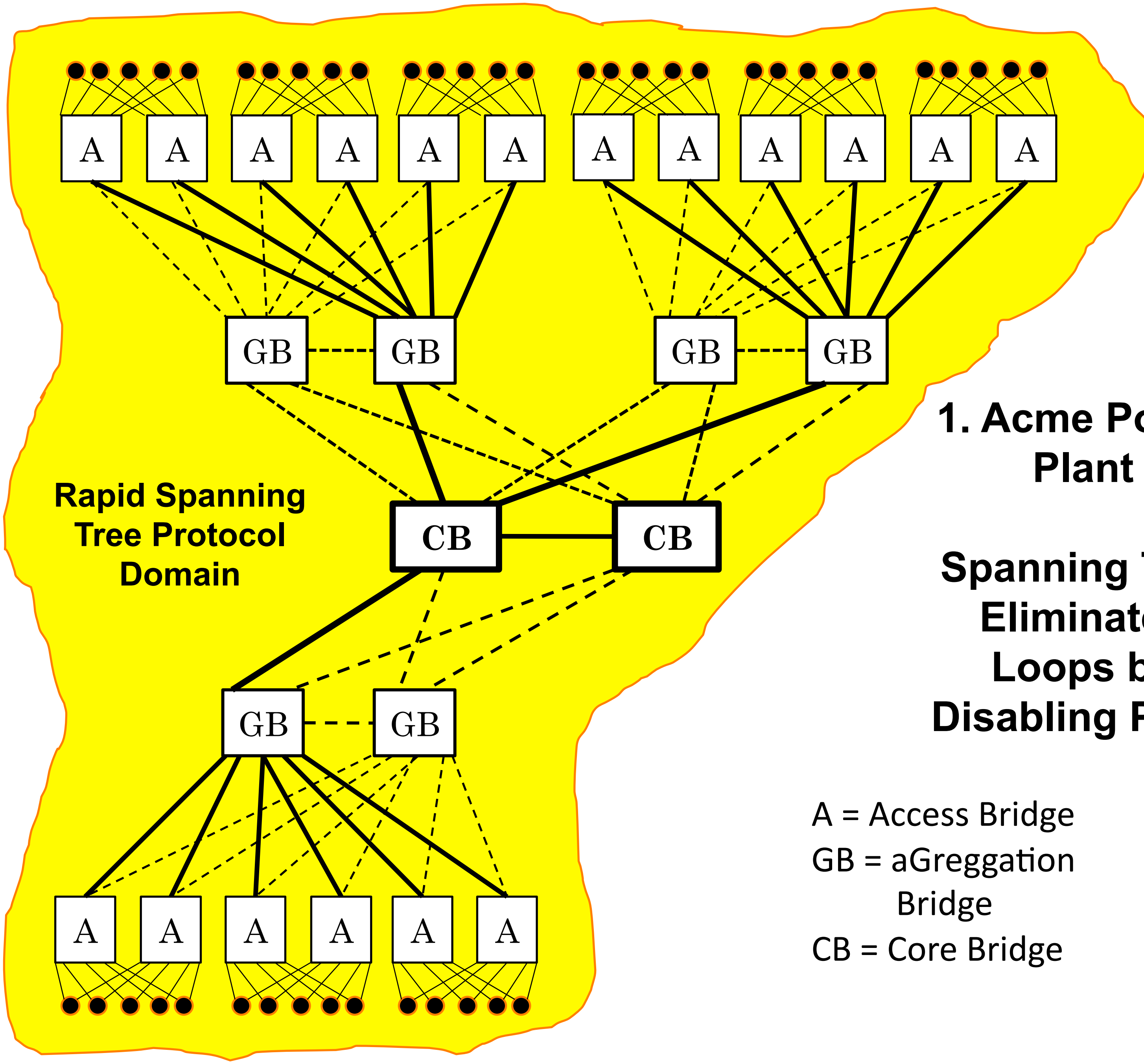
- Compatible with classic bridges. RBridges can be incrementally deployed into a bridged LAN.
- Unicast forwarding tables at transit RBridges scale with the number of RBridges, not the number of end stations. Transit RBridges do not learn end station addresses.
- A flexible options feature. RBridges know what options other RBridges support.
- Globally optimized distribution of IP derived multicast.



1. Acme Power Plant

Bridged Process Control Network

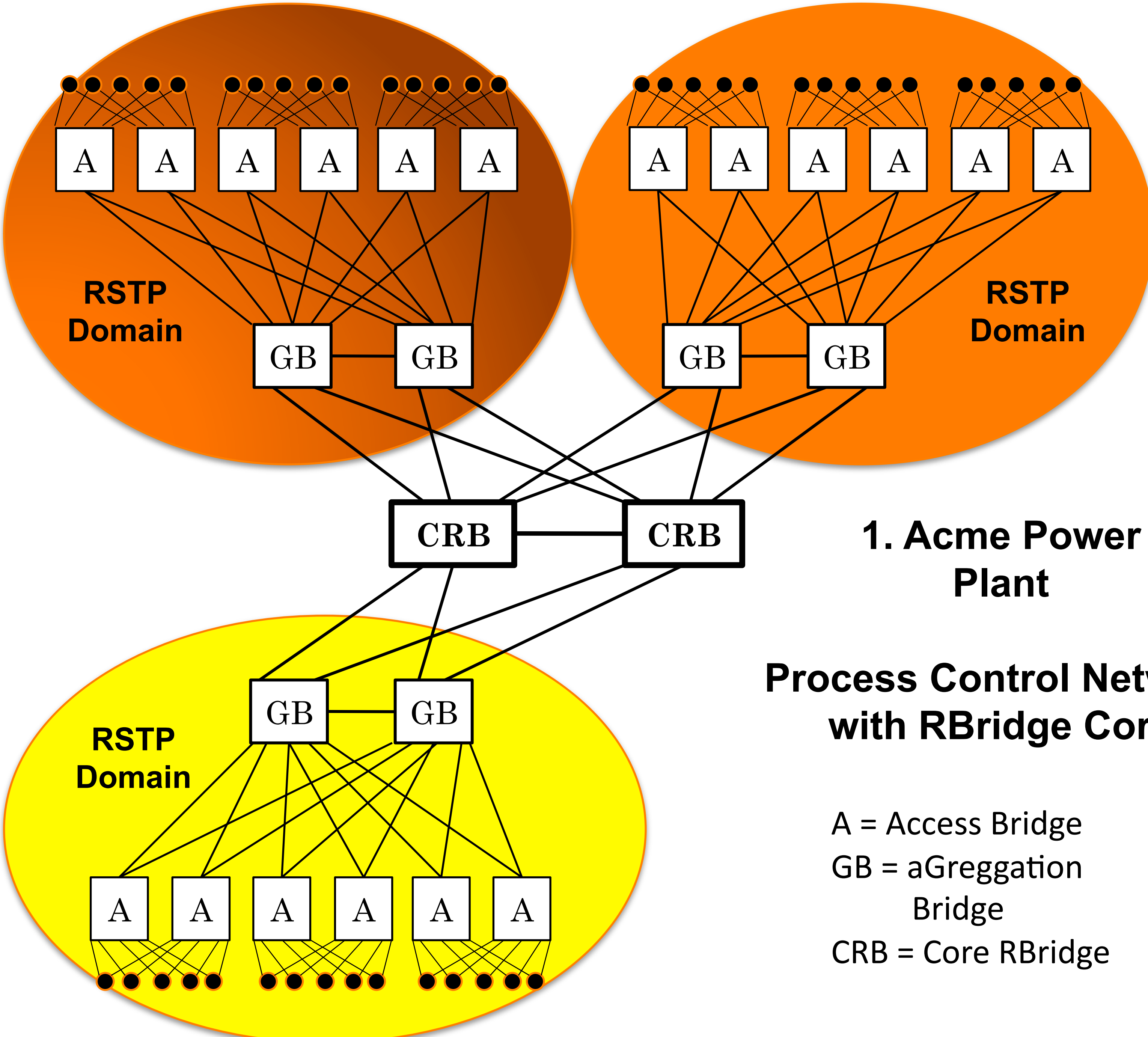
A = Access Bridge
GB = aGreggation Bridge
CB = Core Bridge



1. Acme Power Plant

Spanning Tree Eliminates Loops by Disabling Ports

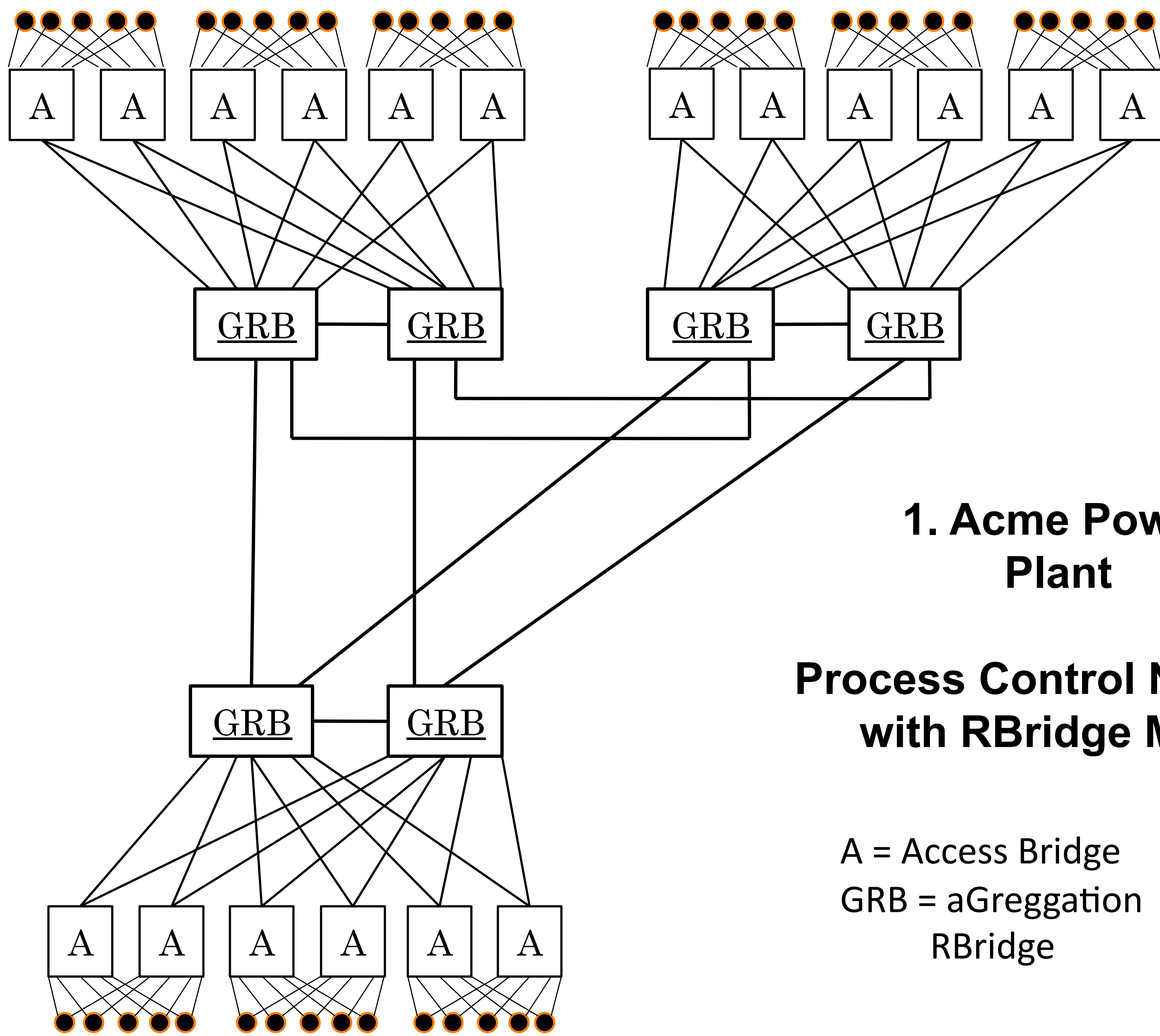
A = Access Bridge
GB = aGreggation Bridge
CB = Core Bridge



1. Acme Power Plant

Process Control Network with RBridge Core

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



1. Acme Power Plant

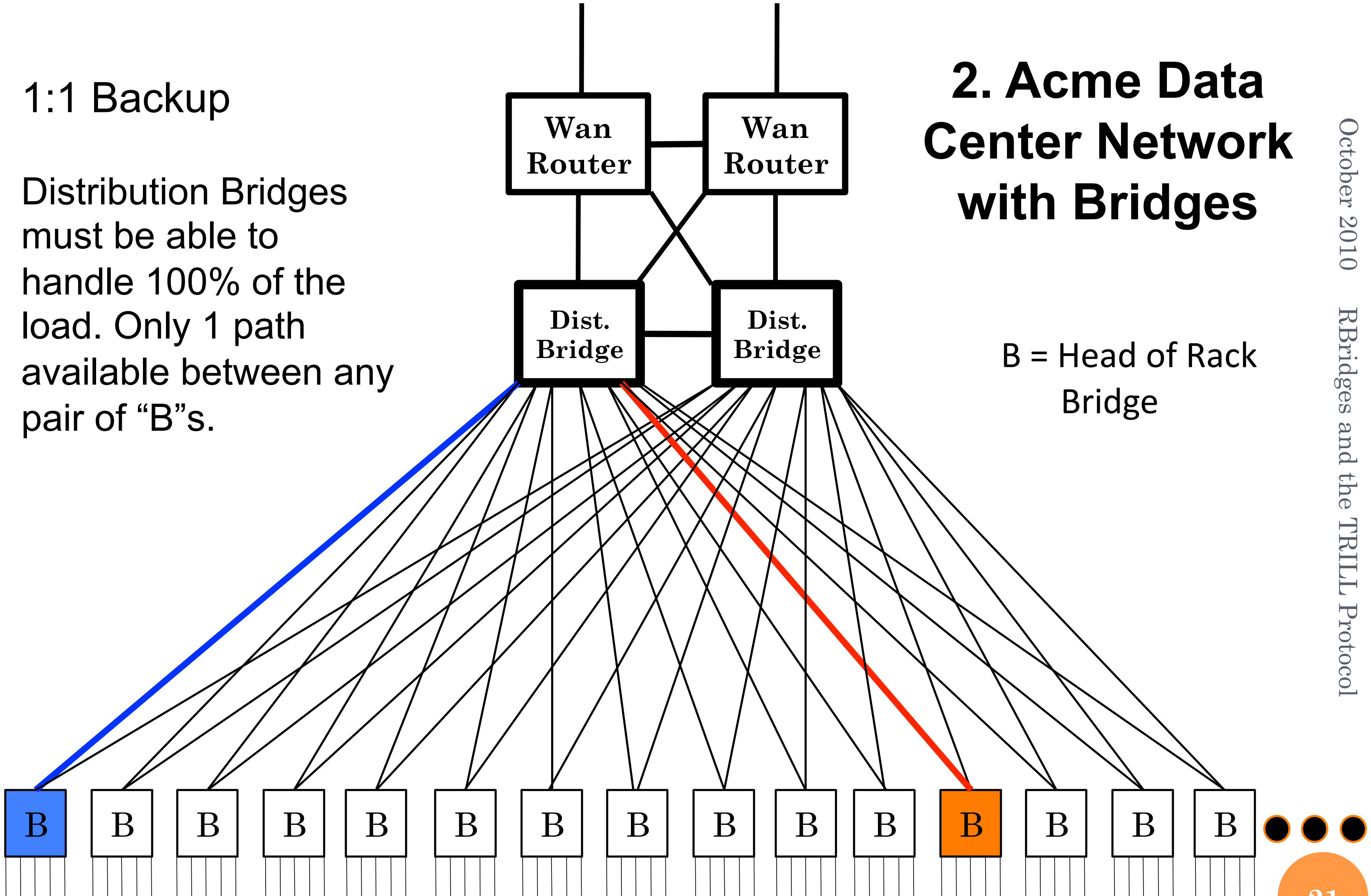
Process Control Network with RBridge Mesh

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.



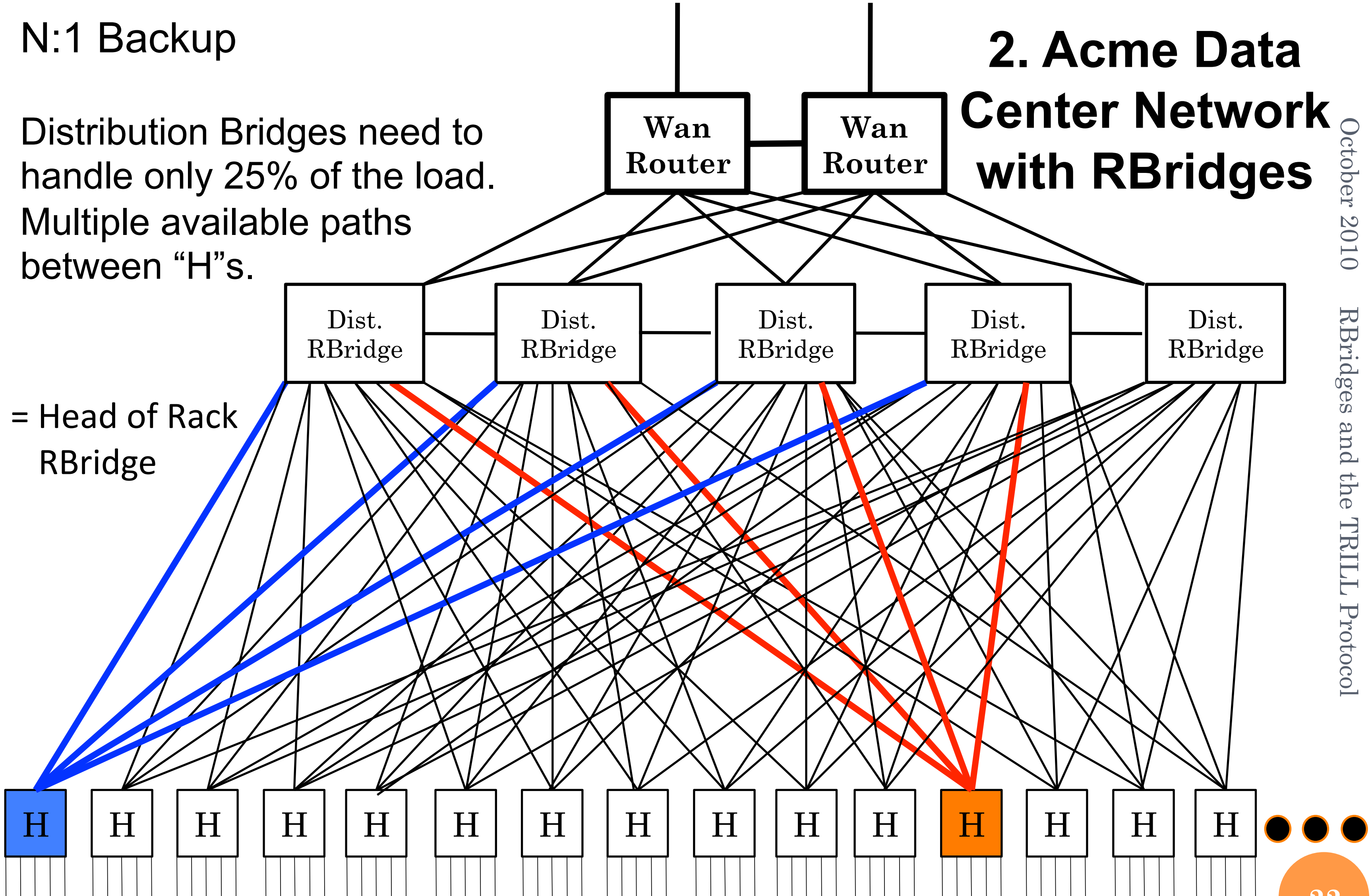
B = Head of Rack Bridge

2. Acme Data Center Network with RBridges

N:1 Backup

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

H = Head of Rack RBridge



TRILL Features

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

- Multi-pathing
- Optimal Paths
- Rapid Fail Over
- The safety of a TTL
- Options

CONTENTS

TRILL ENCAPSULATION AND HEADER

- TRILL Data frames between RBridges are encapsulated in a local link header and TRILL Header.
  - The local 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 multidestination frames.
  - The TRILL header specifies the first/ingress RBridge and either the last/egress RBridge for known unicast frames or the distribution tree for multidestination frames.

FRAME TYPES

- Frame Type Names Used in TRILL
  - TRILL IS-IS Frames – Used for control between RBridges.
  - TRILL Data Frames – Used for encapsulated native frames.
  - Layer 2 Control Frames – Bridging control, LLDP, MACSEC, etc. Never forwarded by RBridges.
  - Native Frames – All frames that are not TRILL or Layer 2 Control Frames.

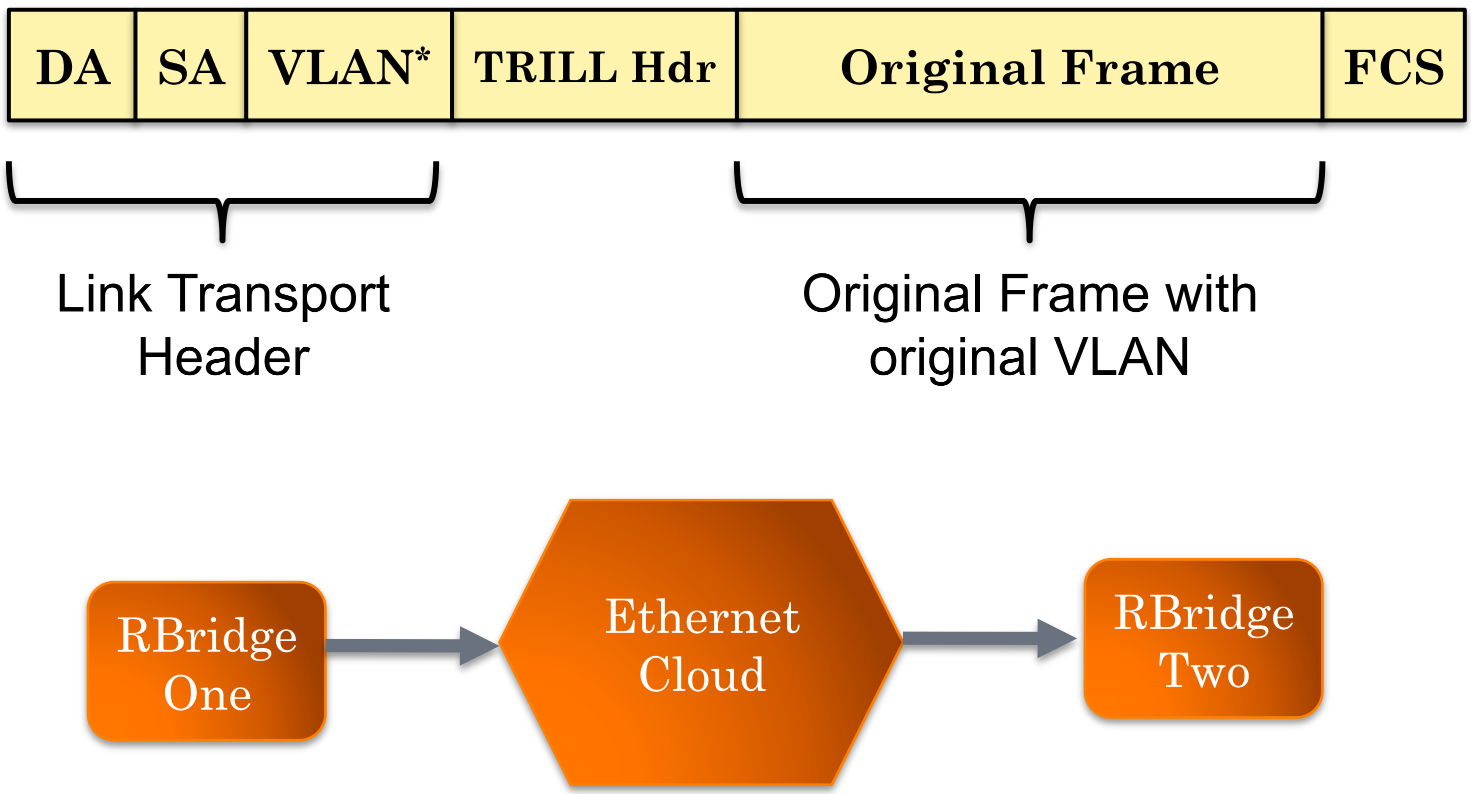
Frame Types, More Detail

- TRILL Frames – Ethernet local link encoding
  - TRILL Data Frames – Have the TRILL Ethertype and are sent to a unicast address or, if multi-destination, to the All-RBridges multicast address.
  - TRILL Control frames – Have the L2-IS-IS Ethertype and are sent to the All-IS-IS-RBridges multicast address.
  - (TRILL Other frames – sent to a TRILL multicast address but not a TRILL Data or TRILL IS-IS frame. Not currently used.)

Frame Types, More Detail

- Layer 2 Control Frames – Destination Address is 01-80-C2-00-00-00 to 01-80-C2-00-00-0F or 01-80-C2-00-00-21
  - High Level – BPDU (01-80-C2-00-00-00) & VLAN Registration (01-80-C2-00-00-21)
    - RBridges handle these differently from bridges. The spanning tree protocol never runs through an RBridge but an RBridge port is not prohibited from participating in spanning tree as a leaf node.
  - Low Level – all other control frames
    - An RBridge port may implement other Layer 2 protocols such as LLDP (Link Layer Discovery Protocol), 802.1X (Port Based Access Control), MACSEC (MAC Security), and the like.

TRILL ENCAPSULATION AND HEADER

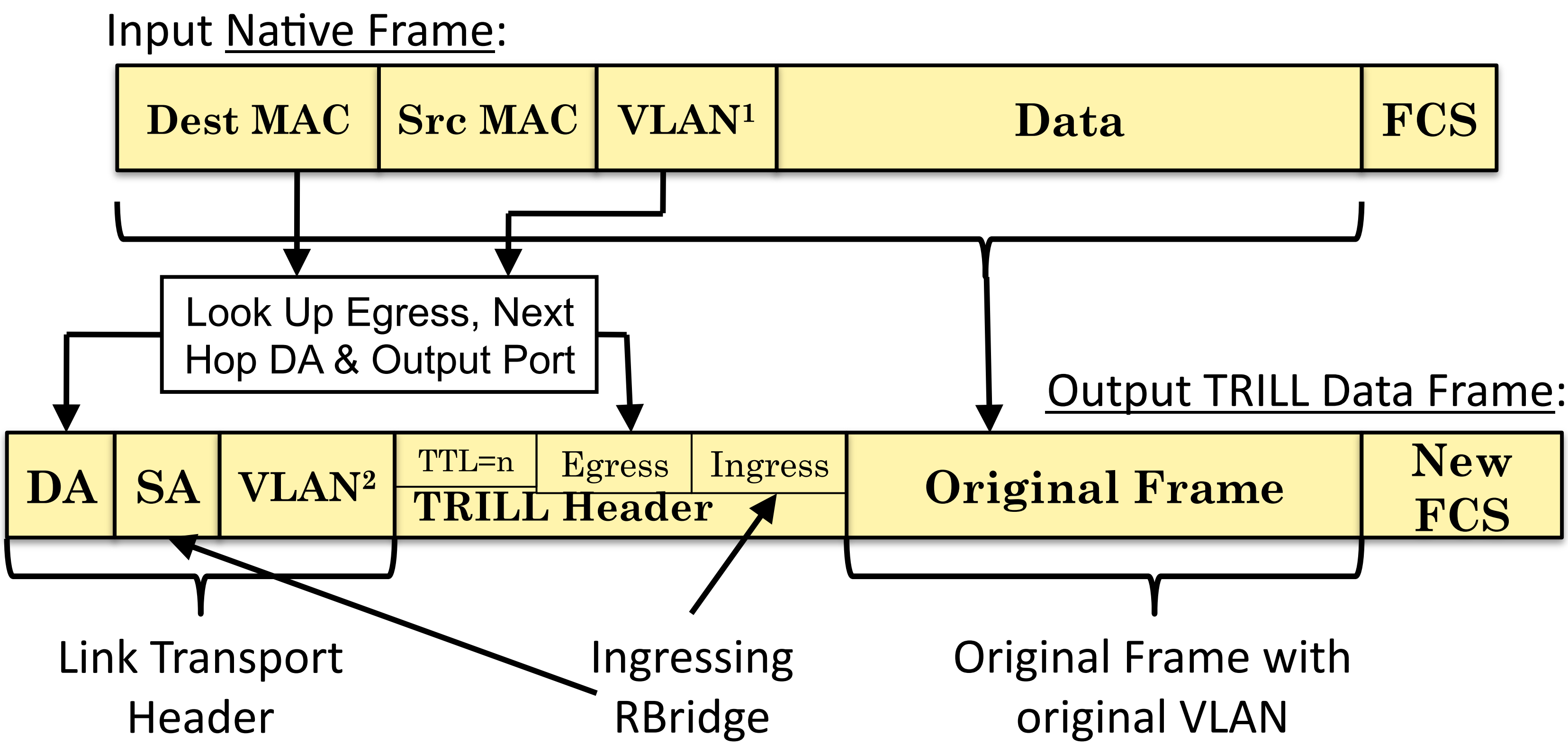


* Link Transport VLAN need only be present if RBridges are connected by a bridged LAN or carrier Ethernet requiring a VLAN tag or the like.

TRILL Encapsulation and Header

- Assuming the link is Ethernet (IEEE 802.3) the encapsulation looks like:
  1. Outer Ethernet Header
     - Source RBridge One, Destination RBridge Two
  2. (Outer VLAN Tag if Necessary)
  3. TRILL Header
  4. Inner Ethernet Header
     - Original Source and Destination Addresses
  5. Mandatory Inner VLAN Tag
  6. Original Payload
  7. Frame Check Sequence (FCS)

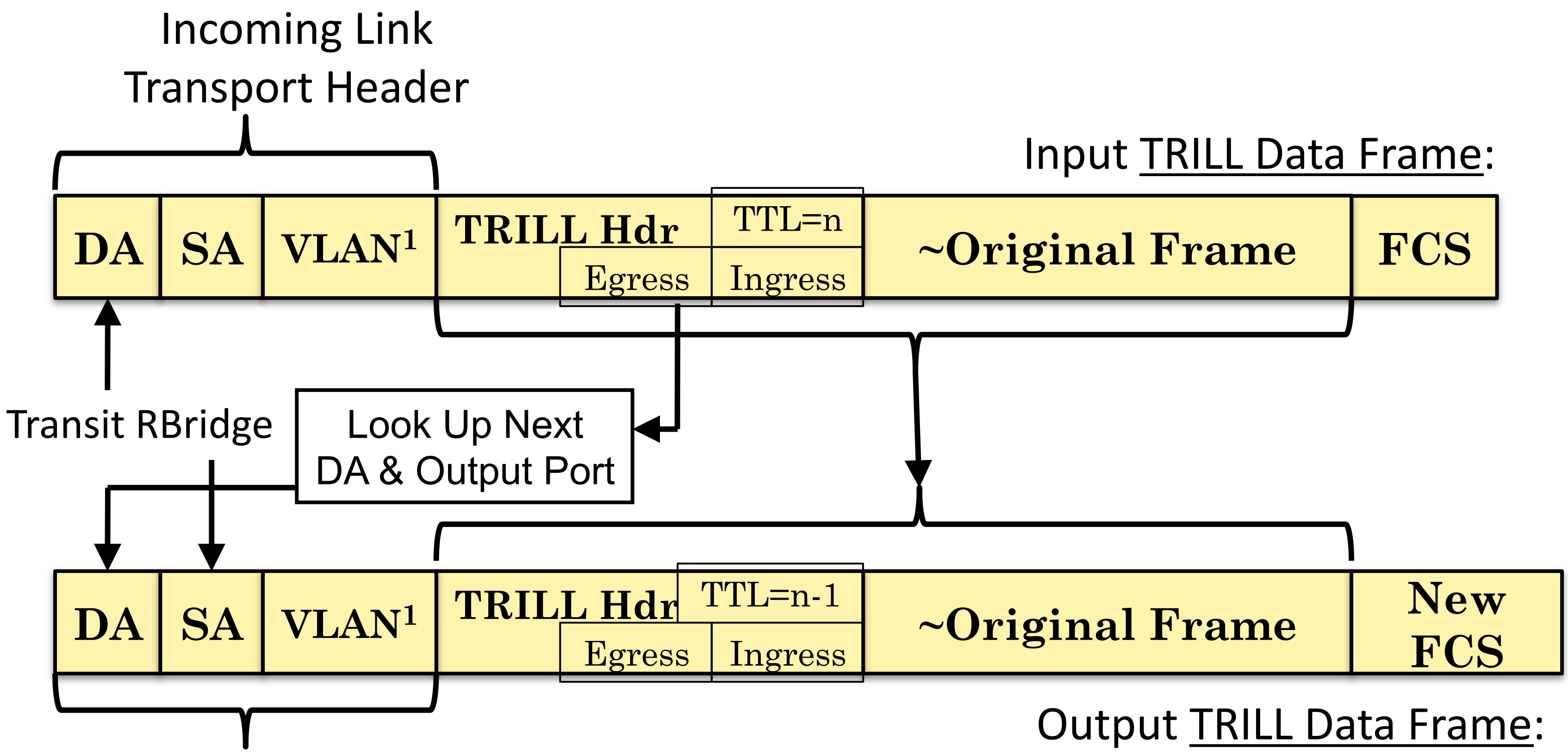
TRILL UNICAST INGRESS



[1] If initial VLAN tag not present, it will be added in encapsulation.

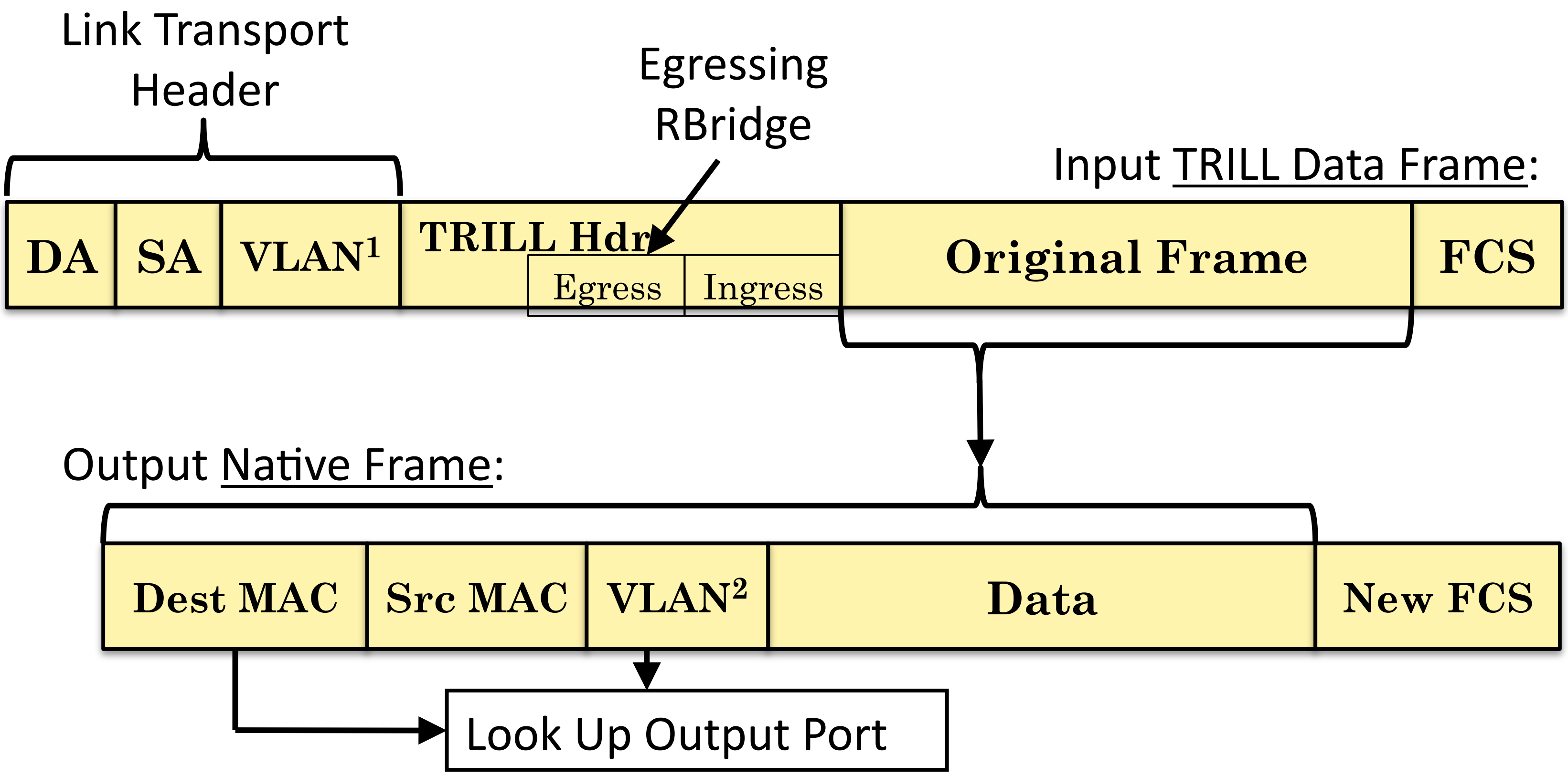
[2] Outer VLAN tag (Designated VLAN) is a transport artifact and is only needed if RBridges are connected by a bridged LAN or carrier Ethernet requiring a VLAN tag or the like.

TRILL Unicast Transit



[1] Input and output Outer VLANs can differ. The true VLAN of the data is inside the Original frame. Outer VLAN only needed if RBridges are connected by a bridged LAN or carrier Ethernet requiring a VLAN tag or the like.

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.

TRILL Encapsulation and Header

- TRILL Header – 8 bytes

TRILL Ethertype					
TRILL Ethertype	V	R	M	OpLng	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)
  - OpLng = Length of TRILL Options
  - Hop = Hop Limit (6 bits)

TRILL ENCAPSULATION AND HEADER

- Summary of 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

CONTENTS

Peering: Are RBridges Bridges or Routers?

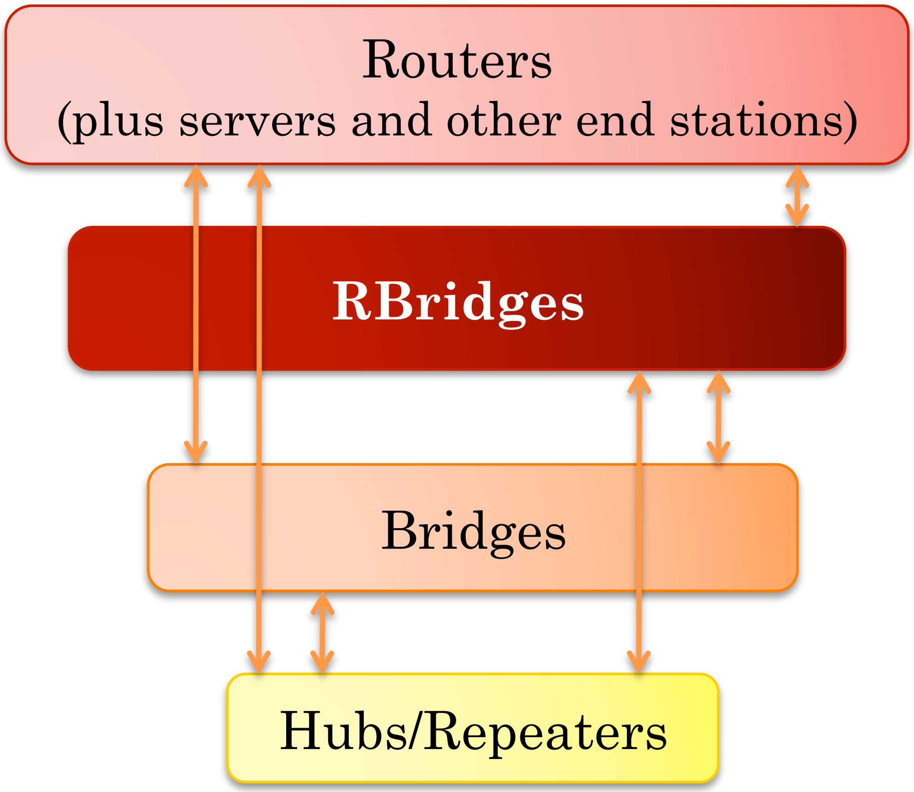
- They are obviously Bridges because
  - RBridges deliver unmodified frames from the source end station to the destination end station
  - RBridges can operate with zero configuration and auto-configure themselves
  - RBridges provide the restriction of frames to VLANs as IEEE 802.1Q-2005 bridges do
  - RBridges support frame priorities as IEEE 802.1Q-2005 bridges do
  - RBridges, by default, learn MAC addresses from the data plane

Peering: Are RBridges Bridges or Routers?

- They are obviously Routers because
  - RBridges swap the outer addresses on each RBridge hop from the ingress RBridge to the egress RBridge
  - RBridges decrement a hop count in TRILL frames on each hop
  - RBridges use a routing protocol rather than the spanning tree protocol
  - RBridges optionally learn MAC addresses by distribution through the control plane
  - RBridges can act based on IP multicast control messages (IGMP, MLD, and MRD) and restrict the distribution of IP derived multicast frames

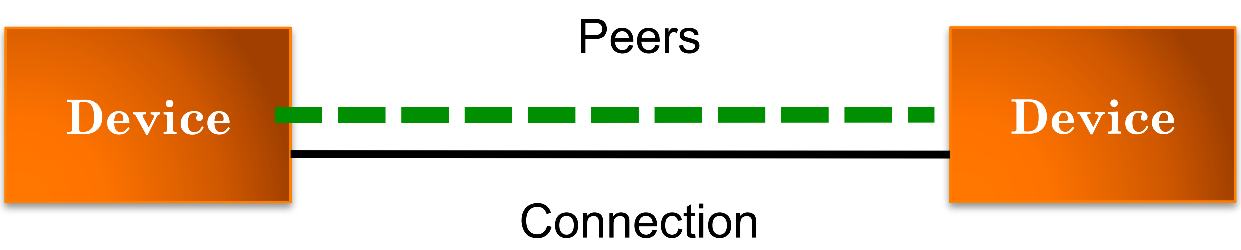
PEERING: ARE RBRIDGES BRIDGES OR ROUTERS?

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



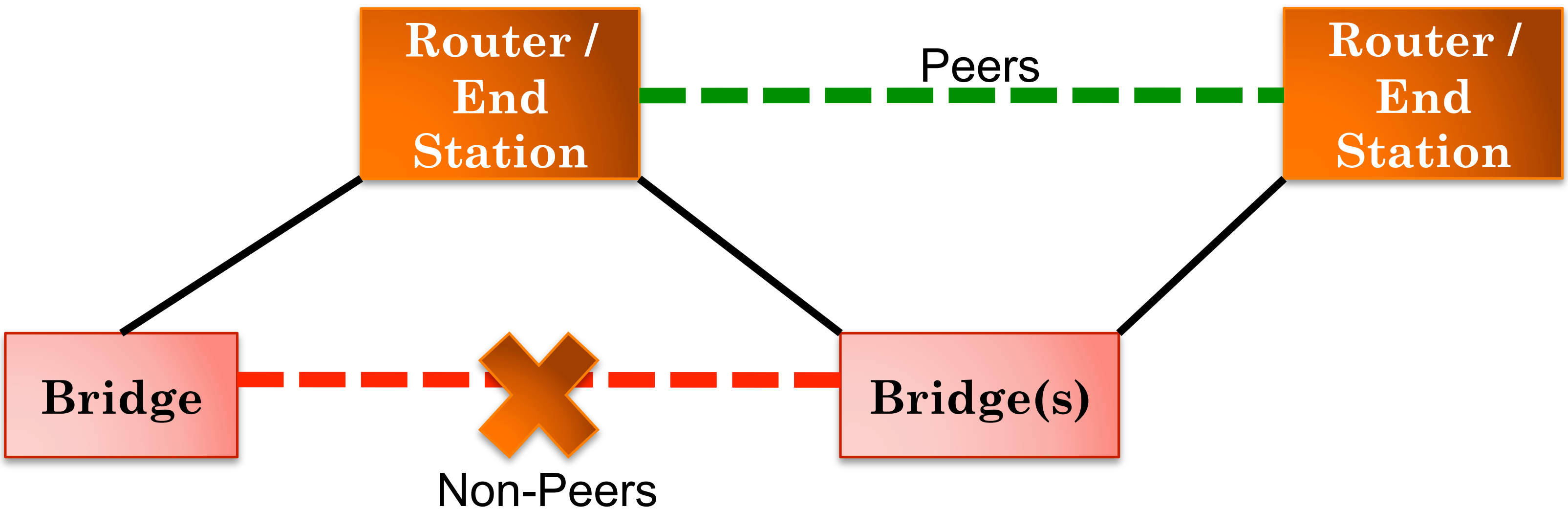
Peering

- Direct Connection



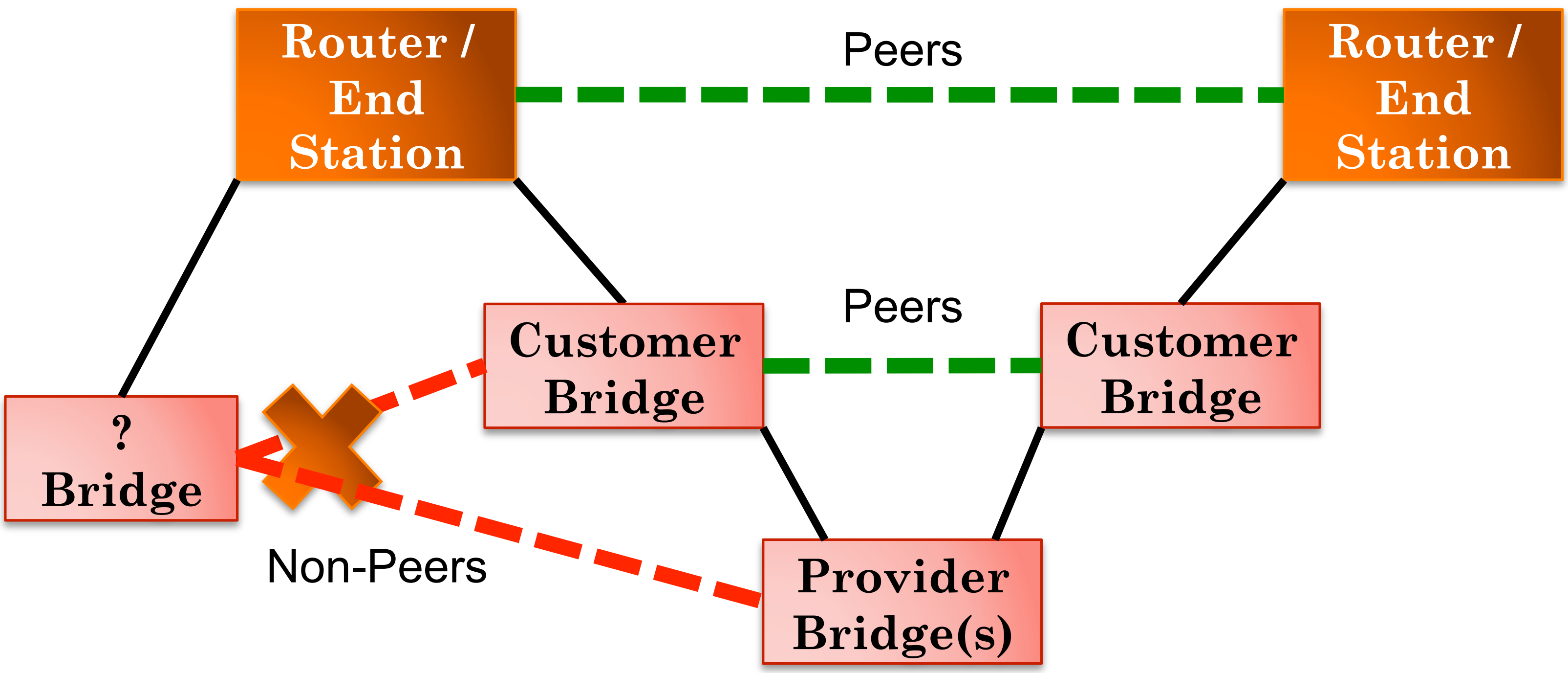
PEERING

- Former Situation



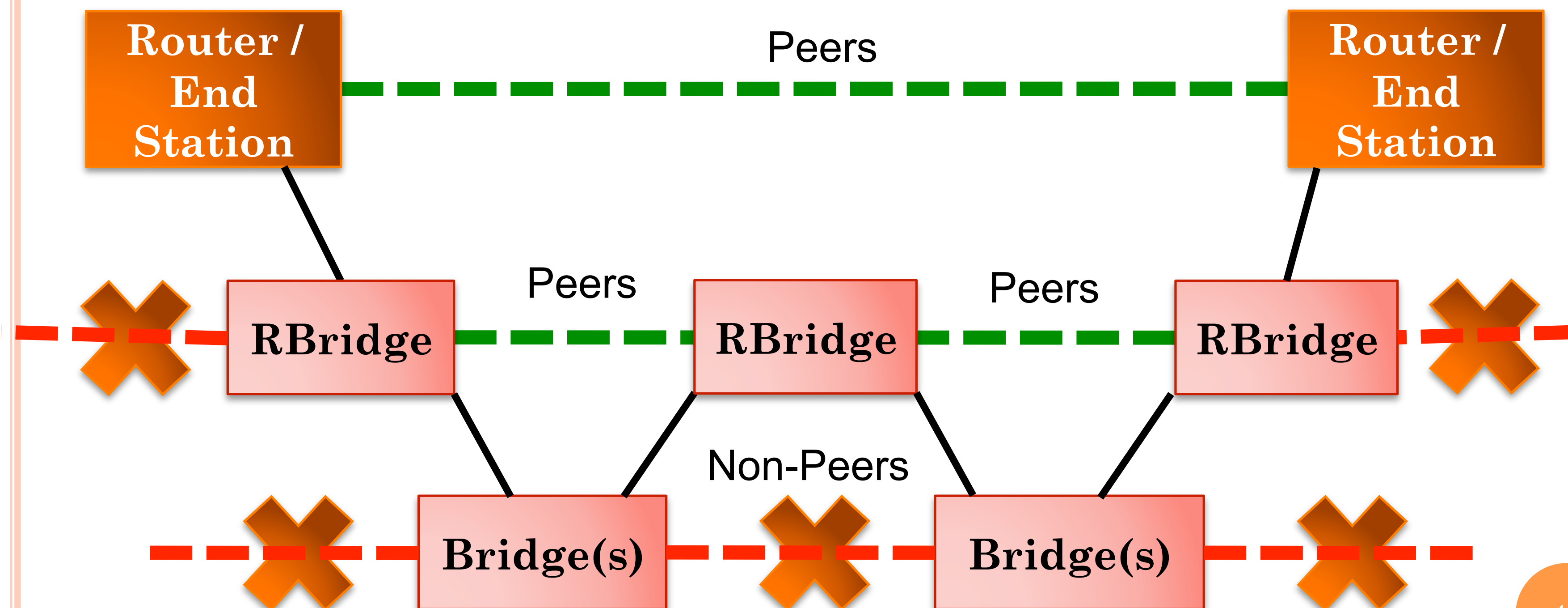
PEERING

- Former Situation
  - Or perhaps



PEERING

- With RBridges



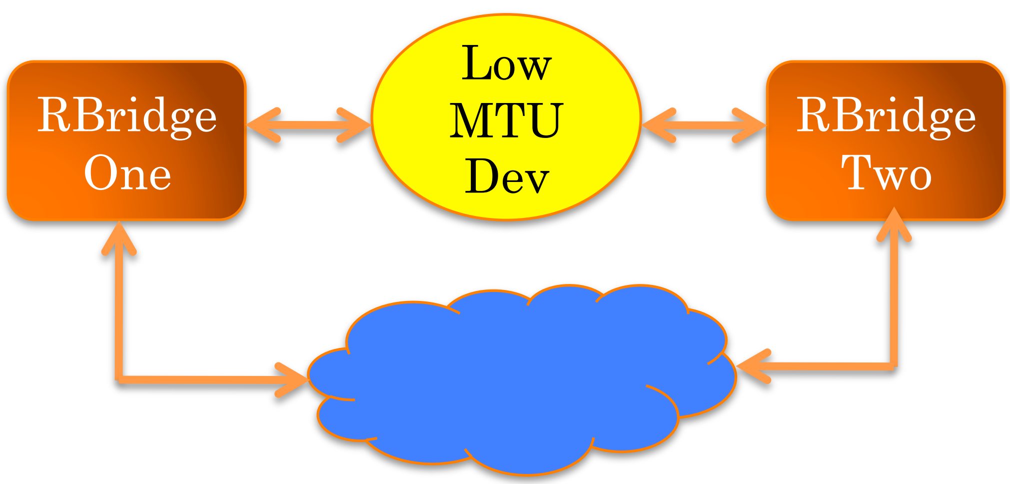
CONTENTS

How RBridges Work

- RBridges find each other by exchanging TRILL IS-IS Hello frames.
  - Like all TRILL control frames, TRILL Hellos are sent 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 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 RBridges Work

- 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 RBridges Work

- 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 information beyond connectivity and link cost. Information such as VLAN connectivity, multicast listeners and multicast router attachment, claimed nickname(s), ingress-to-egress options supported, and the like.
  - The 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 RBridges Work

- 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 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?

- General 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, 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 RBridges Work

- TRILL Data frames
  - That have known unicast ultimate destinations are forwarded RBridge hop by RBridge hop toward the egress RBridge.
  - That are multi-destination frames (broadcast, multicast, and unknown destination unicast) are forwarded on a distribution tree selected by the ingress RBridge.
    - For loop safety, a Reverse Path Forwarding Check is performed on multi-destination TRILL Data frames when received at each RBridge.
    - Distribution trees should be pruned based on VLAN and multicast group.

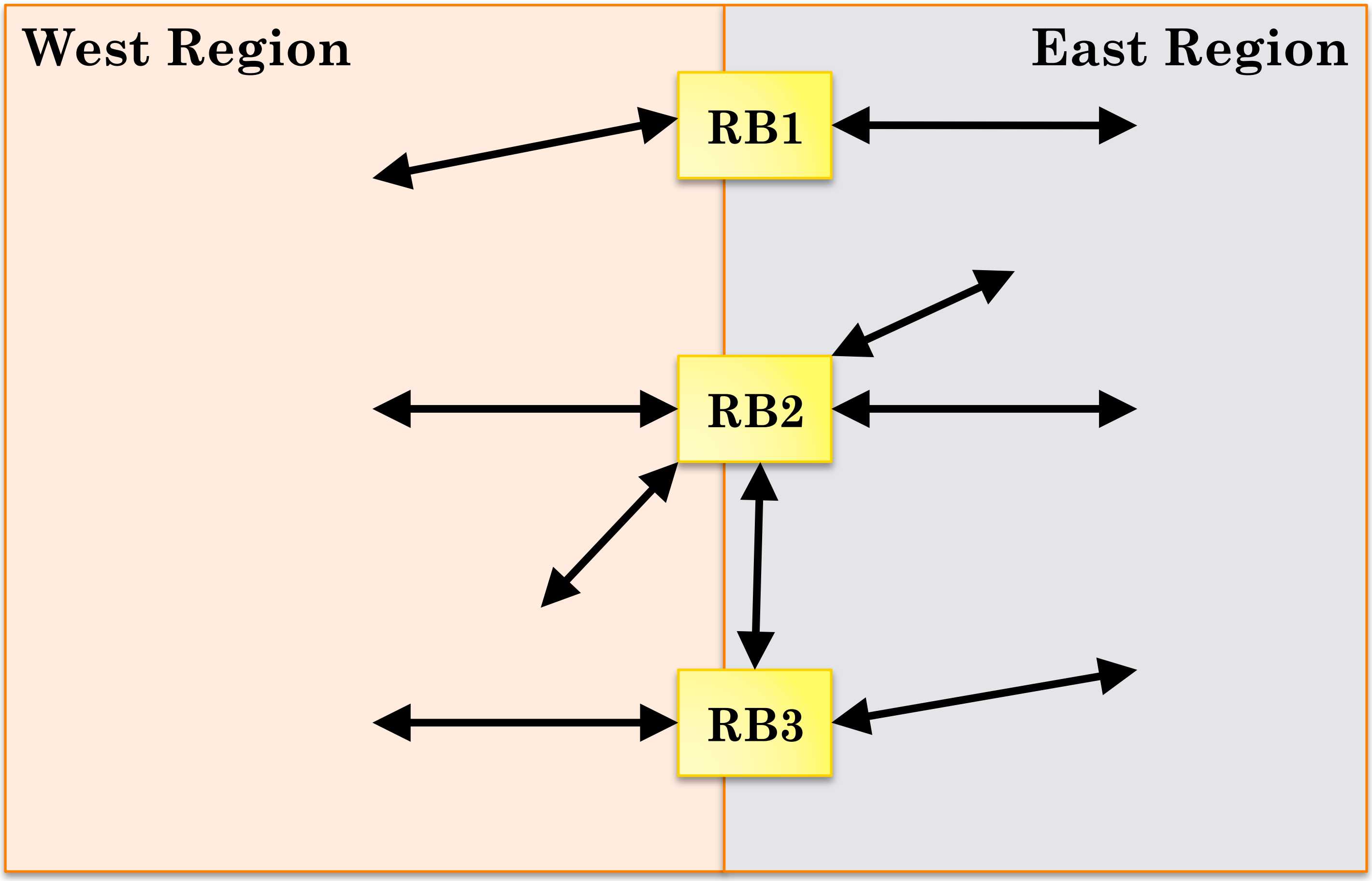
VLANs

- TRILL tries hard to glue together all end stations in a particular VLAN within the campus.
  - In an RBridge campus, any two end stations in the same VLAN that can each reach an RBridge will be able to communicate with each other.
    - I.E., TRILL glues together any VLAN islands
  - Surveys of customers have found this to be generally desirable but there are instances where you want the same VLAN ID in different parts of your RBridge campus to be different or different VLAN IDs in different parts of the campus to be connected.

VLAN Mapping Feature

- Optional feature: Divide the RBridge campus into regions such that VLANs (and priorities) can be mapped or blocked at the inter-regional border RBridges.
  - RBridges internal to a region don't need to do anything about it.
- Since RBridge ports are 802.1Q ports, why doesn't normal bridge port VLAN ID and priority mapping solve this problem?
  - Because TRILL encapsulation means that the <u>inner</u> VLAN tag VLAN ID and priority need to be changed if a TRILL Data frame is involved.

VLAN Mapping Feature



VLAN Mapping Feature

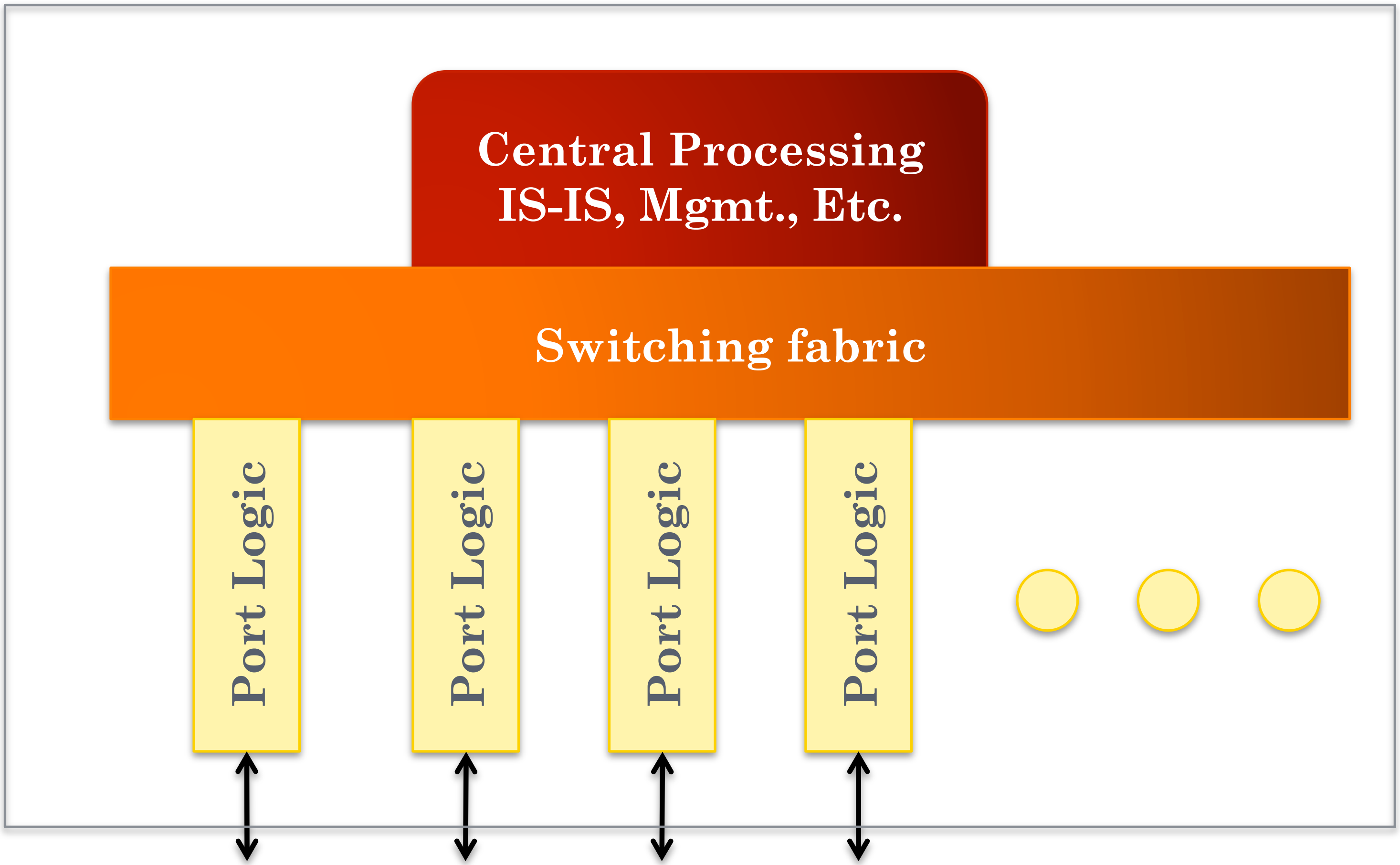
- Mapping at border RBridges is all that is needed for traffic to known unicast addresses.
  - Each port is assigned a region and the inter-regional mapping configured.
- For multi-destination frames, you need to do a little more to be sure all necessary frames get to border RBridges.
  - A border RBridge mapping VLAN X to VLAN Y must report connection to both VLANs so frames it needs to get won't be prematurely pruned.
  - A border RBridge must report connection to a multicast router in each VLAN it is mapping so that multicast frames won't be prematurely pruned.

VLAN Mapping Feature

- The campus still operates as a single routing area.
- The mapping at a border has no effect on the computation or use of shortest paths, distribution trees or the like.

CONTENTS

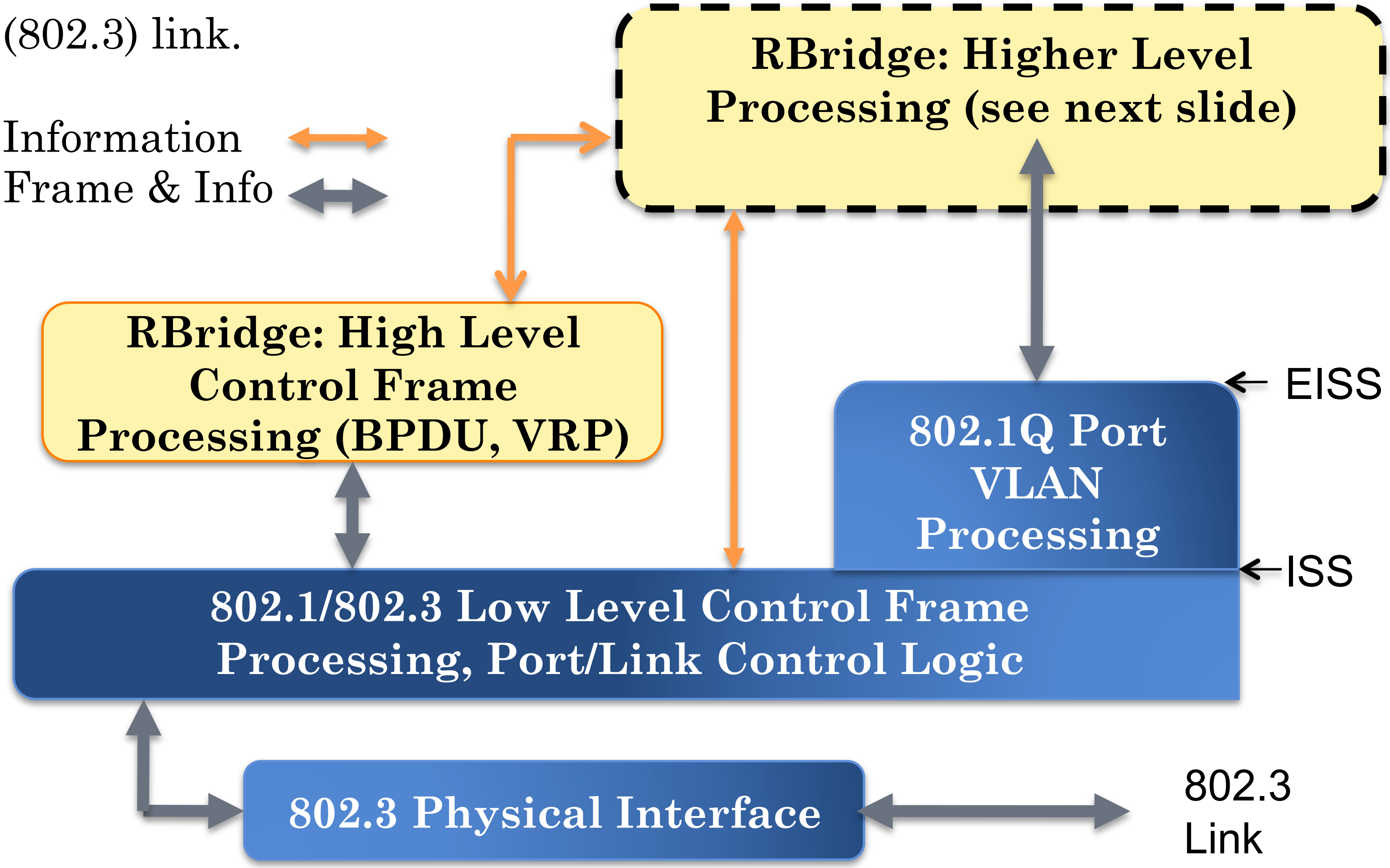
Structure of an RBridge



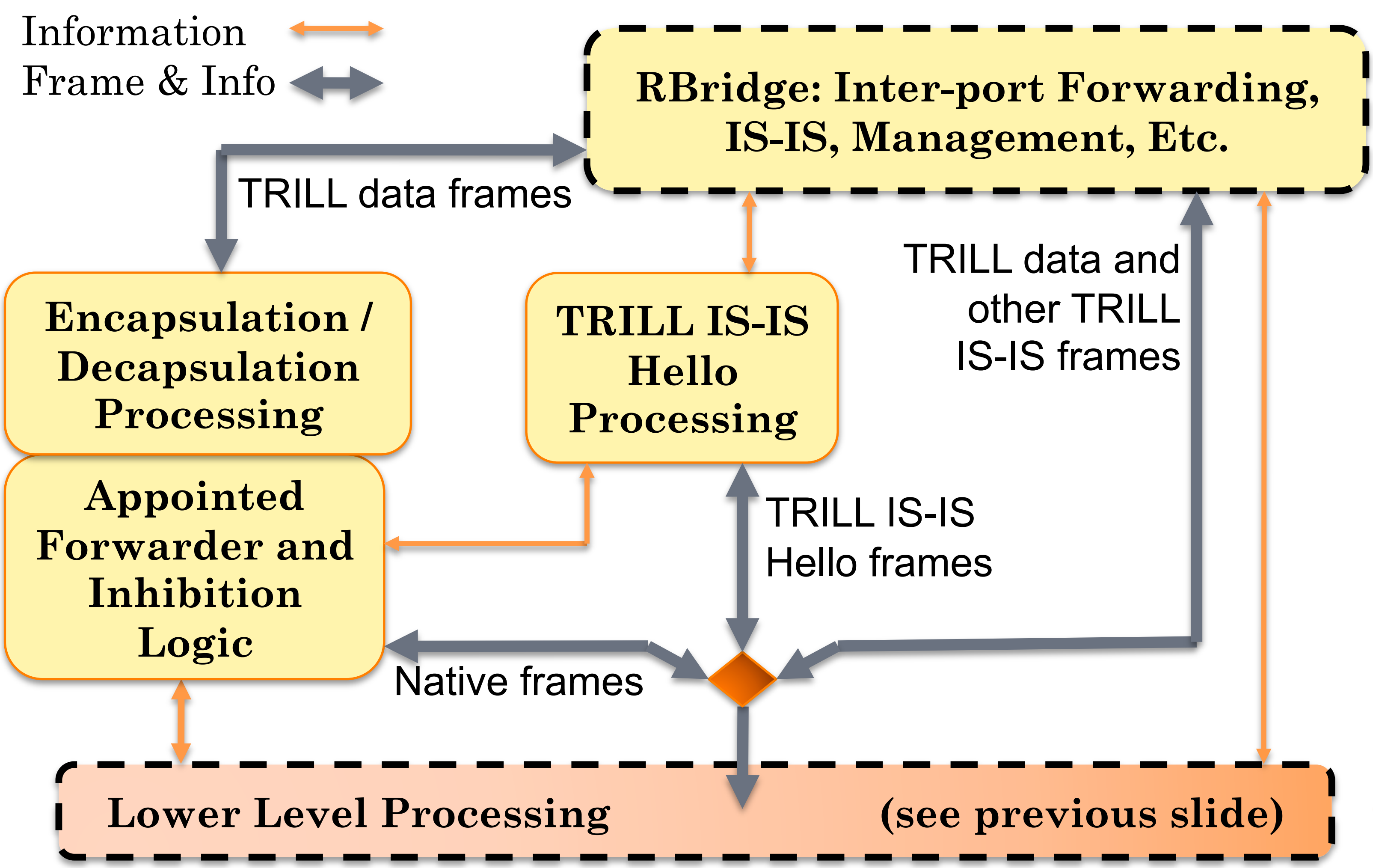
Links to other devices. Could be 802.3 (Ethernet), 802.11 (Wi-Fi), PPP, …

Structure of an RBridge Port

Assumes an Ethernet (802.3) link.



Structure of an RBridge Port



CONTENTS

RBridge Support of Data Center Bridging

- "Data Center Ethernet"

1. Priority Based Flow Control
   - Per Priority PAUSE
2. Enhanced Transmission Selection
3. Congestion Notification

4. TRILL

(Items 1–3:) Data Center Bridging

RBRIDGE SUPPORT OF DATA CENTER BRIDGING

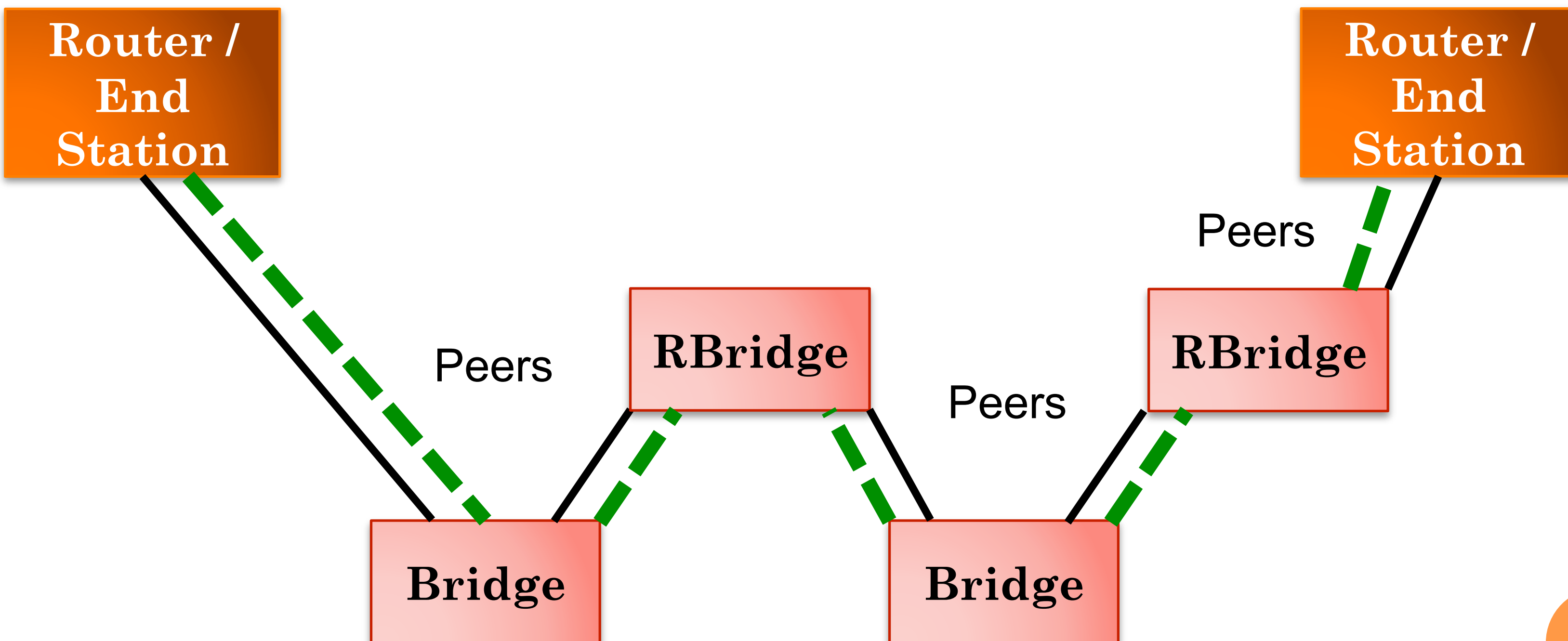
- The goal is "loss-less" Ethernet, that is, no loss due to queue overflow
- Basic Ethernet PAUSE is a very blunt instrument
  - Interference with loss dependent flow control such as TCP
  - Blocking of high priority control frames
  - Congestion spreading

RBRIDGE SUPPORT OF DATA CENTER BRIDGING

- Answer 1: Consider different frame priorities as different pipes
  - 802.1Qbb: Seperate 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.
  - 802.1az: 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.

RBridge Support of Data Center Bridging

- For purposes of Data Center Bridging, all nodes are considered peers:



RBRIDGE SUPPORT OF DATA CENTER BRIDGING

- Answer 2: Congestion Notification (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.

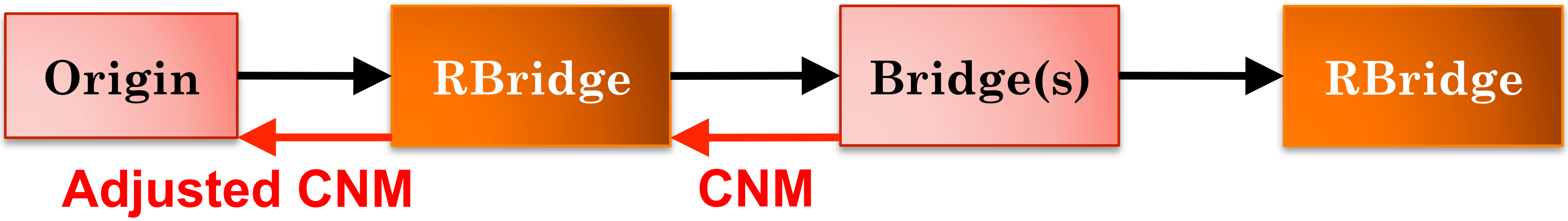
RBridge Support of Data Center Bridging

Link Header	TRILL Header	Orig. DA	Orig. SA	Orig. VLAN	CN Tag	Original Payload	FCS

Encapsulated Frame

- However, RBridges have to handle CNMs generated by TRILL ignorant bridges between RBridges. Such a CNM will be initially addressed to the previous hop RBridge, not the original end station.
- (Provider Backbone Bridges have a similar problem when generating a CNM.)

RBridge Support of Data Center Bridging



- Previous hop RBridge has to adjust the CNM so that it goes back to the origin end station.

- Note: All of the Data Center Bridging facilities depend on appropriate engineering, limited delay bandwidth product, etc., to actually provide "loss-less" service.

CONTENTS

Address Learning

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

ESADI

- The optional End Station Address Distribution Information (ESADI) protocol:
  - Provides a VLAN-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.

What About Re-Ordering?

- RBridges are required to maintain frame ordering internally, modulo flow categorization.
- When multi-pathing is used, all frames for an order-dependent flow must be sent on the same path if unicast or the same distribution tree if multi-destination.
- Unicast re-ordering can occur briefly when a destination address transitions between being known and unknown, or a topology change occurs.
  - This can be minimized with keep-alives, ESADI, distribution tree per RBridge, or configured addresses.

LOOP AVOIDANCE

- All possible forms of looping behavior within a TRILL campus can be grouped into one of three classes:
  - TRILL Data Frame Loops, in which the frame would remain TRILL encapsulated.
  - Hybrid TRILL Data / Native Frame Loops, in which the frame is repeatedly encapsulated and decapsulated.
  - Native Frame Loops, in which the frame is never encapsulated.
    - Since TRILL always encapsulates data, if you have this problem, no RBridges are involved so it is not TRILL's fault.

Loop Avoidance

- TRILL Data Frame Loops:
  - Known unicast frames have a hop count and are always unicast to the next hop RBridge towards their destination.
  - Multi-destination frames must be received on a port which is part of their distribution tree, the ingress RBridge nickname and input port must pass a Reverse Path Forwarding Check, and they have a hop count.

LOOP AVOIDANCE

- Hybrid TRILL Data / Native Frame Loops:
  - Such a loop would require, at some point around the loop, that a TRILL Data frame be decapsulated onto a link by one RBridge and then picked up and re-encapsulated by another RBridge.
  - TRILL takes great care to minimize the probability of there being two uninhibited appointed forwarders on the same link for the same VLAN.
    - Under certain conditions, an RBridge appointed forwarder is inhibited from accepting or sending native frames. This only affects native frames. An RBridge port is never inhibited or blocked from sending or receiving TRILL Data or TRILL IS-IS frames except by very low level link flow control mechanisms such as PAUSE or if the port has been manually configured as disabled.

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.
-   - By Ray Perlner

CONTENTS

Standardization Status

- Time span of effort: 5 ½ + years
- Earlier organizational meetings in late 2004
- First TRILL WG meeting: March 2005
- Base protocol draft pass up from TRILL Working Group December 2009
- Base protocol approved as a standard by the IETF March 15th 2010

Standardization Status

- Non-IETF Assignments:
  - TRILL Ethertype: 0x22F3
  - L2-IS-IS Ethertype: 0x22F4
  - Block of Multicast Addresses for TRILL: 01-80-C2-00-00-40 to 01-80-C2-00-00-4F
  - TRILL NLPID: 0xC0

- Final approval of IS-IS code points and data structures pending.

STANDARDIZATION STATUS

- First open interoperability testing (plug fest) was held at the University of New Hampshire Interoperability Laboratory (UNH IOL) 3-5 August 2010:
  - http://www.iol.unh.edu/services/testing/bfc/grouptest/TRILL_plugfest.php
- Second planned for Q1, 2011.

- Some ongoing standards work:
  - RBridge MIB
  - TRILL over PPP
  - RBridge VLAN Mapping
  - RBridge Support of DCB
  - OAM
  - TRILL Header Options

CONTENTS

- Introduction
- TRILL Features
- TRILL Encapsulation and Header
- Are RBridges Bridges or Routers?
- How RBridges Work
- Structure of RBridge Ports
- RBridge Support of Data Center Bridging
- Some Additional Details
- Standardization Status
- References

References (newer)

- Standard: “Rbridges: Base Protocol Specification”
  - http://tools.ietf.org/html/draft-ietf-trill-rbridge-protocol-16
- “RBridges: Campus VLAN and Priority Regions”
  - https://datatracker.ietf.org/doc/draft-ietf-trill-rbridge-vlan-mapping/
- “Definitions of Managed Objects for RBridge”
  - https://datatracker.ietf.org/doc/draft-ietf-trill-rbridge-mib/
- “PPP TRILL Protocol Control Protocol”
  - https://datatracker.ietf.org/doc/draft-ietf-pppext-trill-protocol/

References (older)

- "TRILL: Problem and Applicability Statement"
  - http://www.ietf.org/rfc/rfc5556.txt
- TRILL WG Charter:
  Current (out of date) and proposed
  - http://www.ietf.org/dyn/wg/charter/trill-charter.html
  - http://www.postel.org/pipermail/rbridge/2010-May/003986.html
- Original Paper by Radia Perlman:
  "Rbridges: Transparent Routing"
  - http://www.postel.org/rbridge/infocom04-paper.pdf

END

Donald E. Eastlake 3rd
Co-Chair, IETF TRILL Working Group
d3e3e3@gmail.com, +1-508-333-2270