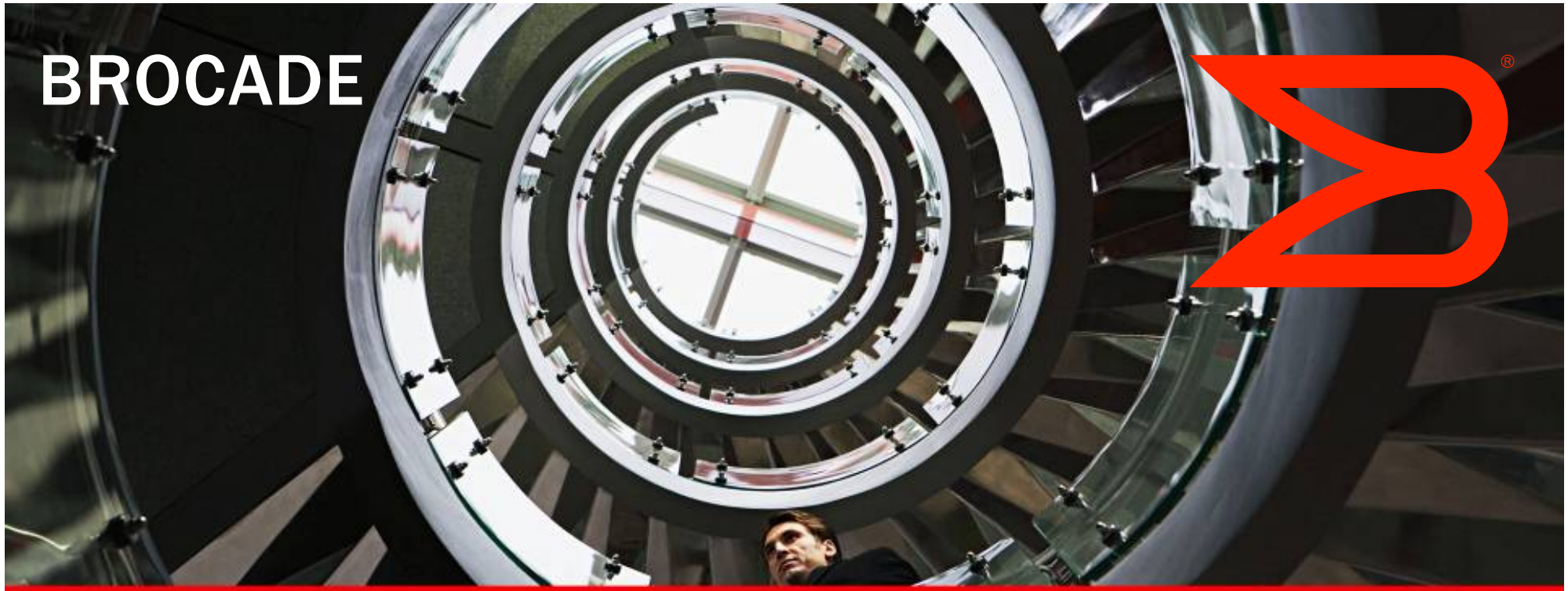


TRILL & Convergence in the Data Center

Marcelo M. Molinari

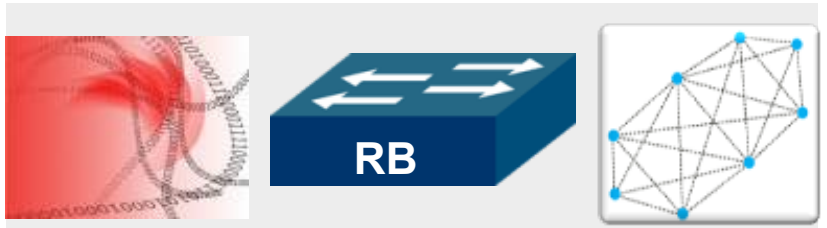
Maio de 2010

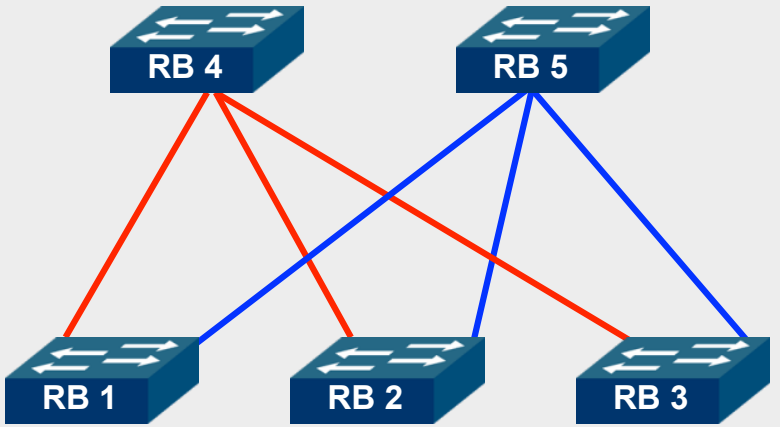
Data Center Networks

Algorhyme V2

by Ray Perlner

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.

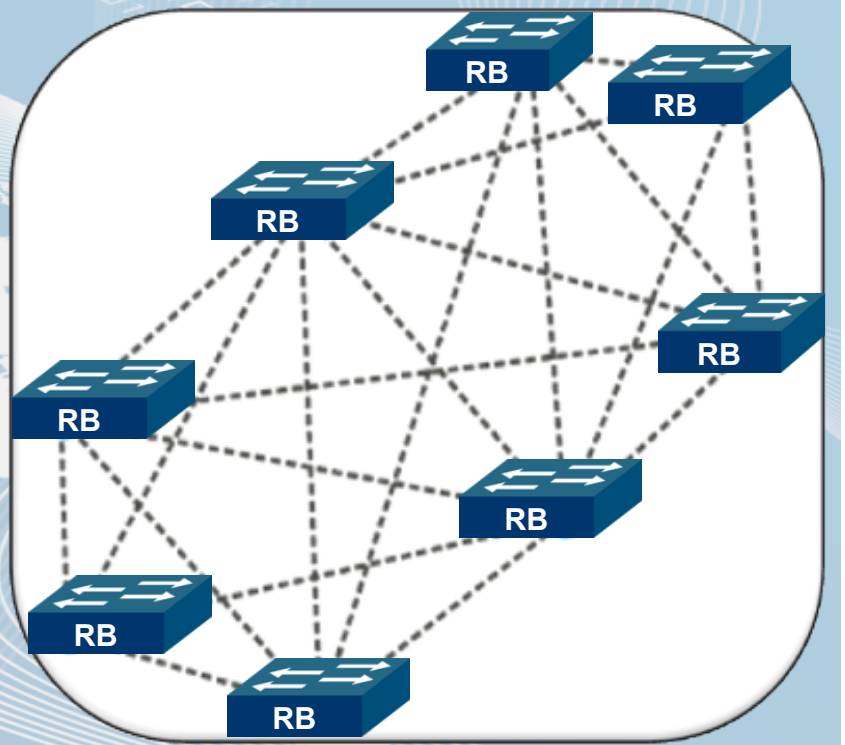




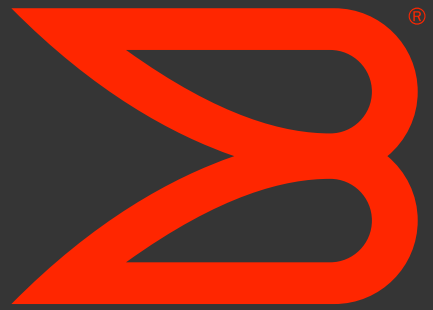
TRILL Ethertype	V	R	M	OpLng	Hop
Egress RBridge Nickname	Ingress RBridge Nickname				

Agenda

- Motivation
- Introduction
- TRILL discussion
  - STP limitations
  - TRILL L2 multipathing
- Brocade solution
  - Resilient L2 for LANs and SANs
- Summary



BROCADE



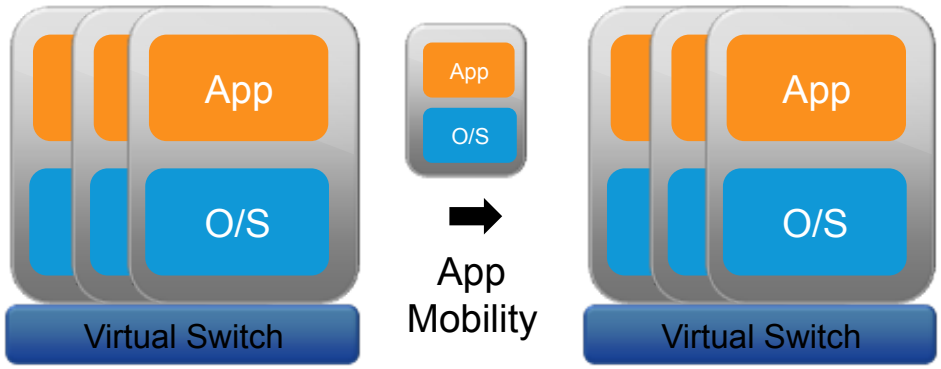
Motivation

Motivation

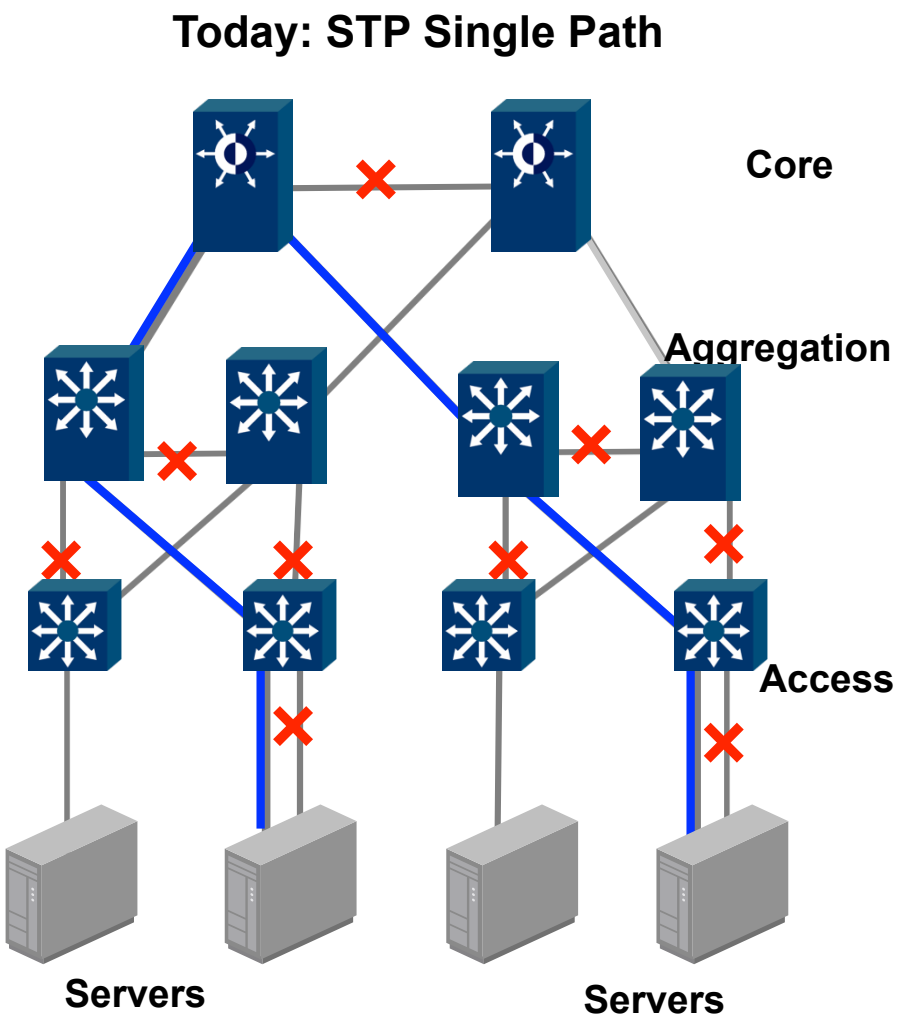
Why develop TRILL?

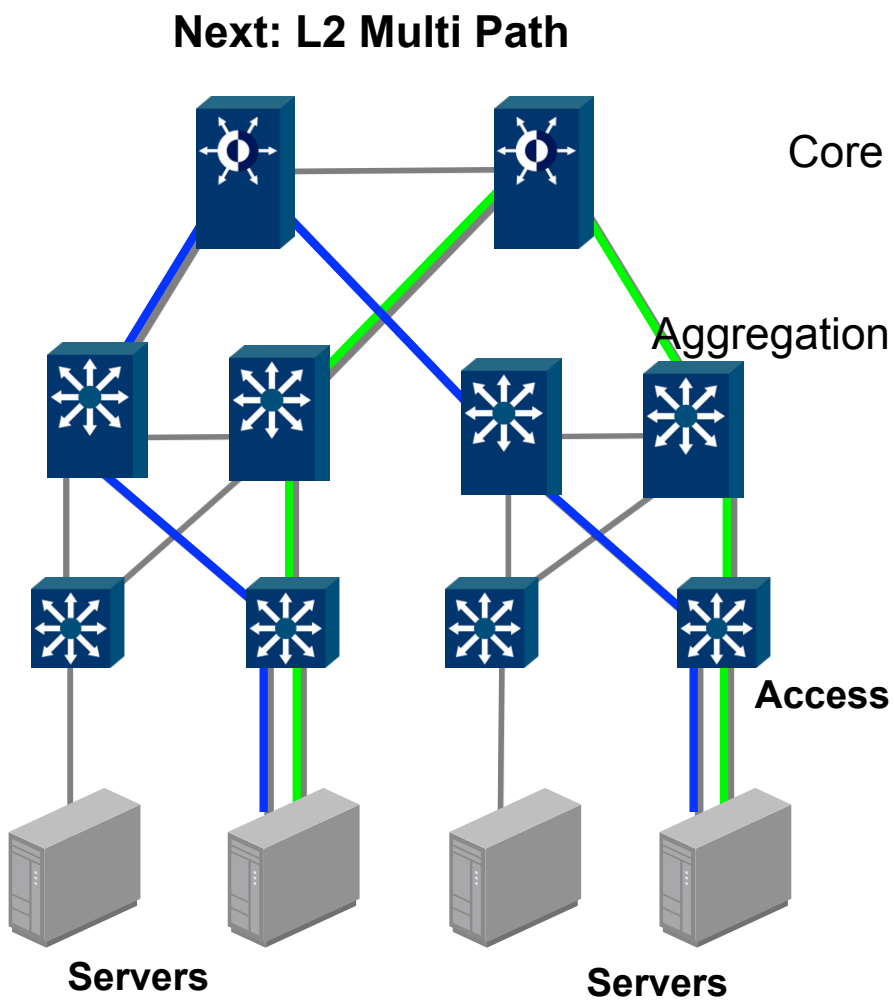
Convergence of LAN & SAN traffic over Ethernet transport

Server virtualization requires a more resilient L2 infrastructure

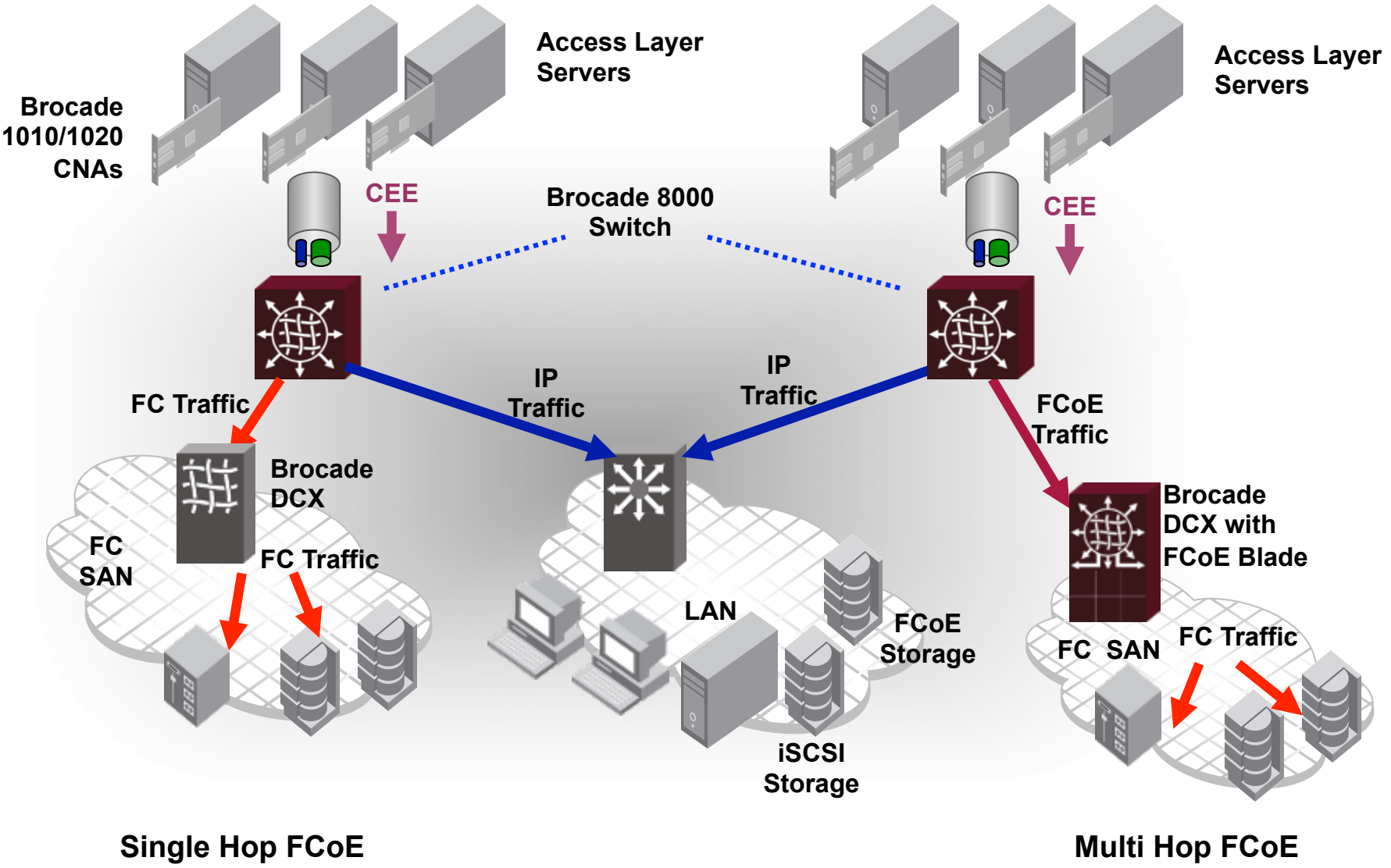


LANs Get L2 Multipathing

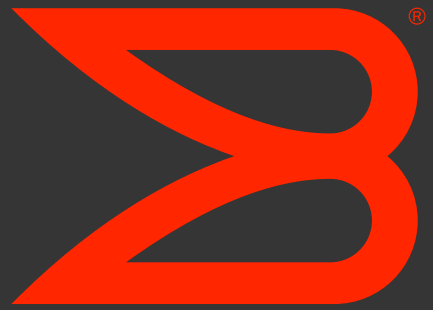




FCoE Gets Multi Hop Support



BROCADE



Introduction to TRILL

TRILL

Transparent Interconnection of Lots of Links

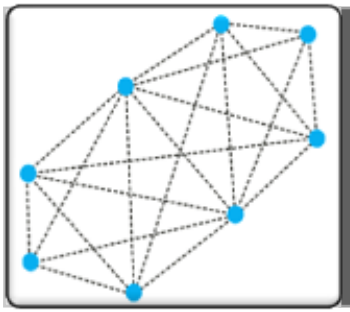
I E T F	a proposed data center L2 protocol being developed by an Internet Engineering Task Force (IETF) workgroup
Mission	"*The TRILL WG will design a solution for **shortest-path frame routing** in multi-hop IEEE 802.1-compliant Ethernet networks with arbitrary topologies, using an existing link-state routing protocol technology.*" - source IETF
Scope	"TRILL solutions are intended to address the problems of …, inability to multipath, … within a *single Ethernet link subnet*" - source IETF

TRILL Solution

Overview

Devices are Routing Bridges (RBridges or Rbridges)

Data Plane is TRILL protocol

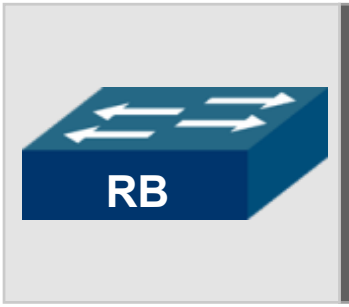


Control Plane is link state routing protocol

TRILL Solution

Functions

- Link state protocols
  - Flood configuration information to RBridges
  - Used for shortest path calculations
  - Used to distribute configuration database

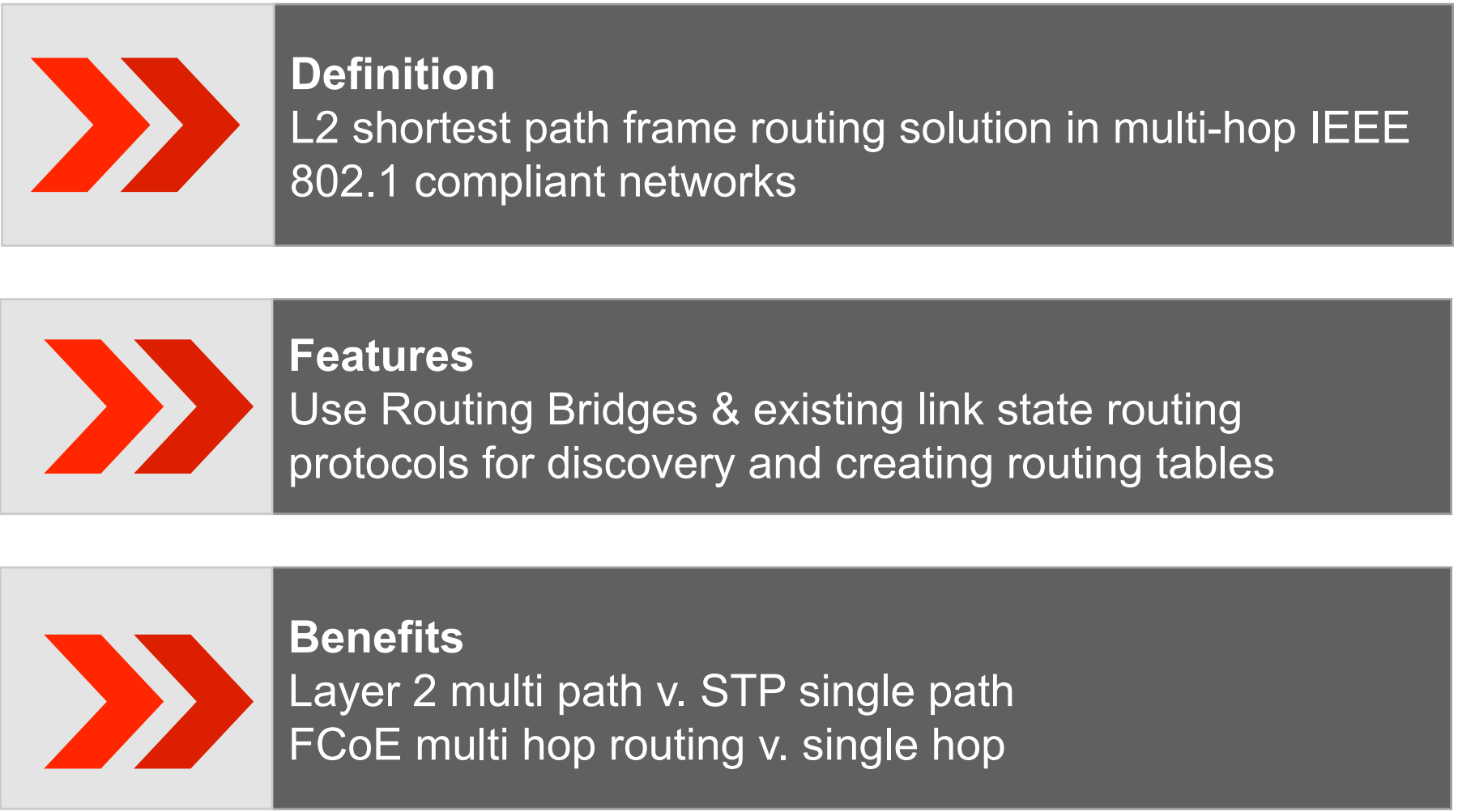


- RBridges
  - Use link state Hellos to find each other
  - Calculate shortest paths to all other RBridges
  - Build routing tables

- TRILL
  - Ingress RBridges encapsulate TRILL data
  - Egress RBridges decapsulate TRILL data

TRILL

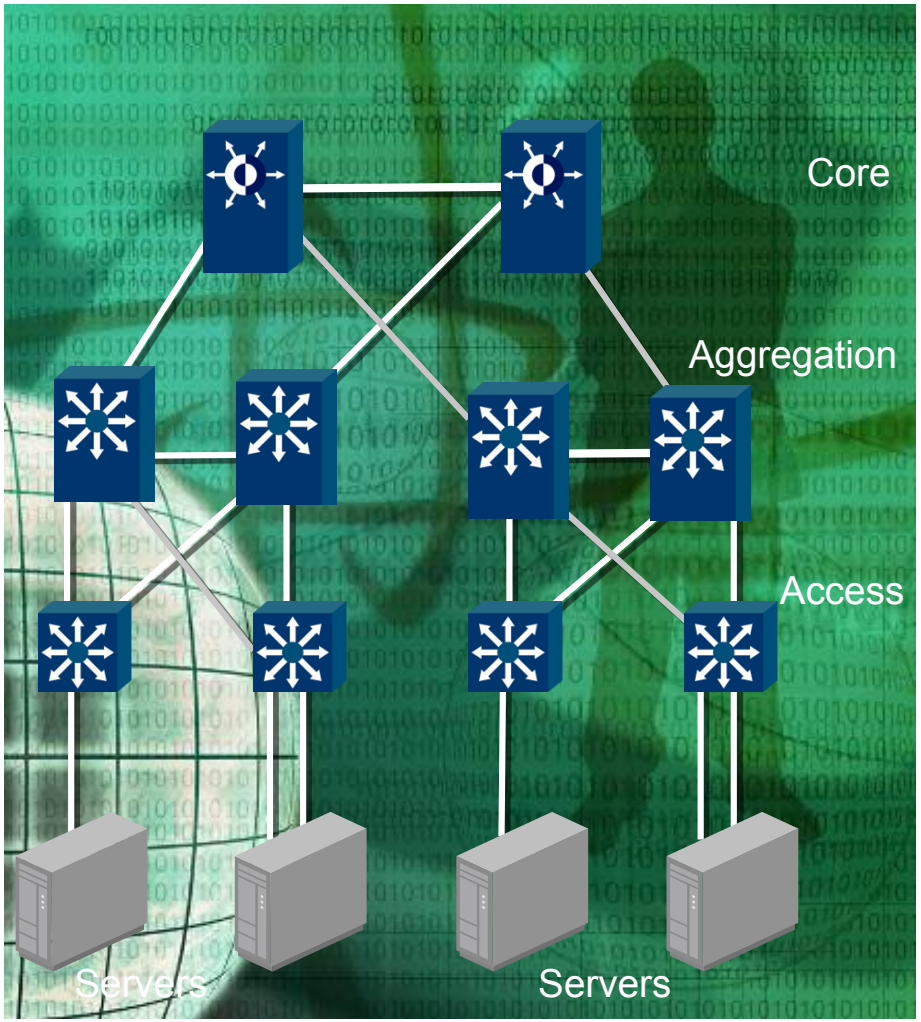
All You Need to Know. Almost ☺



Data Center Networks

Built with good intentions

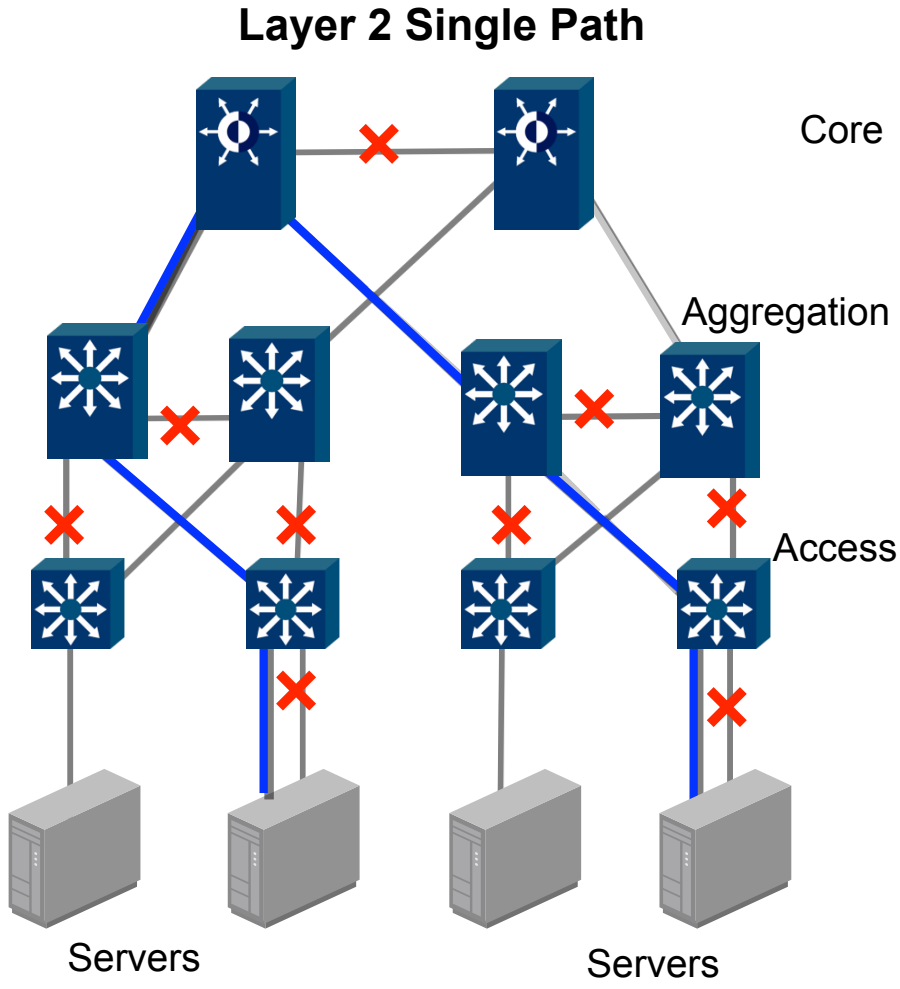
- Designed to meet business needs in a dynamic and cost effective fashion
- Data centers use richly connected topologies, such as fat trees
- Contain many Equal-Cost Multi-Paths (ECMP) between any given endpoints



The STP Effect

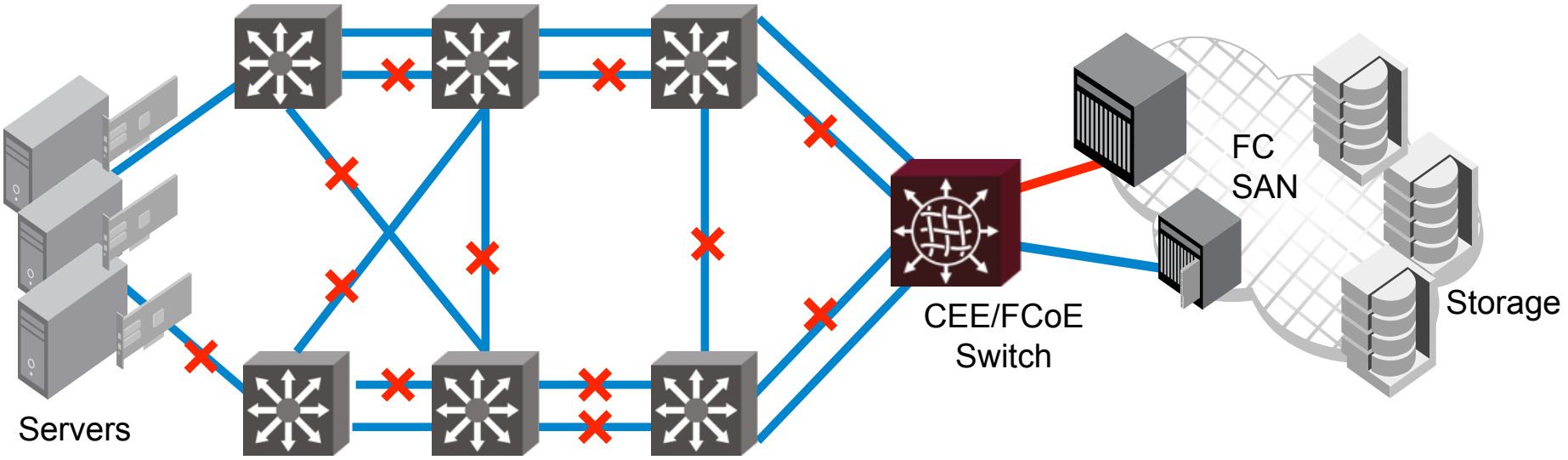
Data center with Spanning Tree Protocol (STP)

- STP is an Ethernet protocol that establishes and maintains a single loop free spanning tree among all the bridges on a VLAN
- All alternate paths are blocked
- Inefficient use of available links reduces aggregate bandwidth
- Reacts to small topology changes
- VLANs may partition due to connectivity changes
- The Ethernet header does not contain a hop count (or TTL) field



Converged Environments

*STP **Impacts Storage** Too*



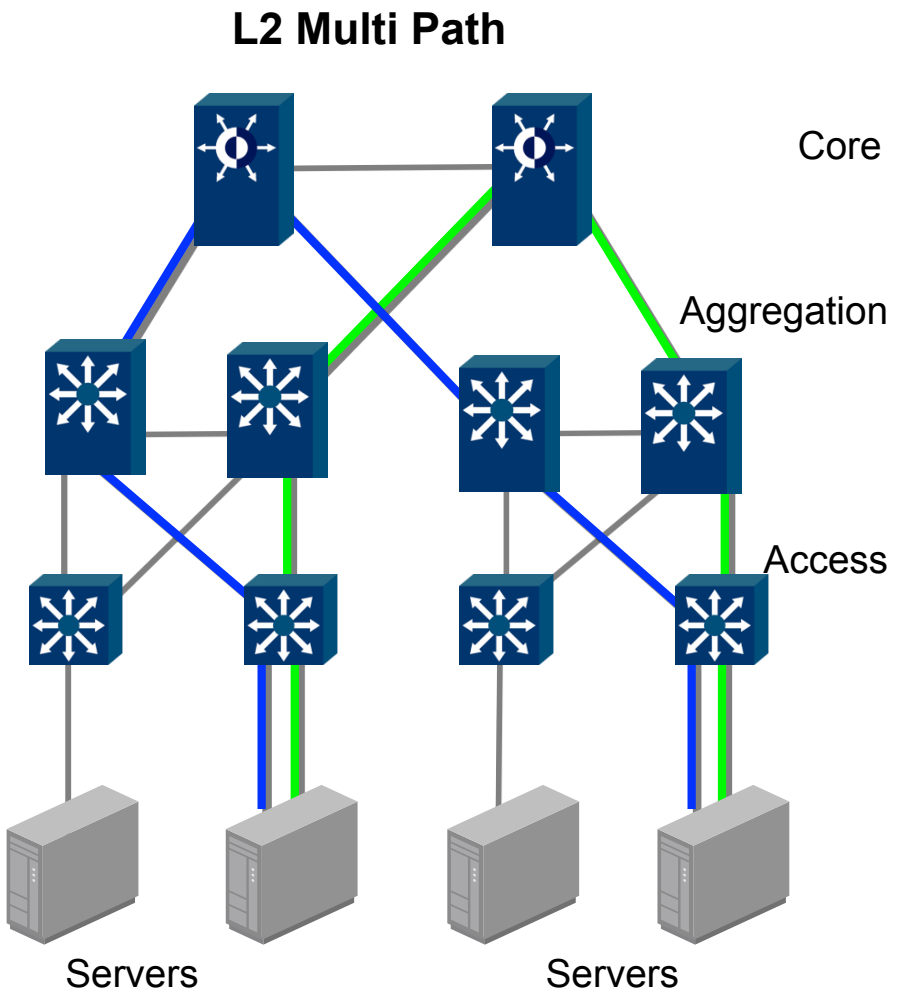
- STP shuts down redundant paths
  - Bandwidth limited to unique path
  - No Active-Active or fault tolerance ability
- Recovering from link or node failure may take 40 sec. or more
  - **Storage traffic may be broadcasted during relearning**
  - Network instability: lost transactions

FC
Ethernet
Disabled

The TRILL Solution

L2 multipathing and multi hop FCoE

- Enables L2 multiple paths via load splitting among paths
- Reclaims network bandwidth and improves utilization
- Improves efficiency with L2 shortest path and ECMP
- Faster response to failures
- TRILL is backward-compatible with existing infrastructures
- Delivers multiple hop FCoE with Routing Bridges and link state routing



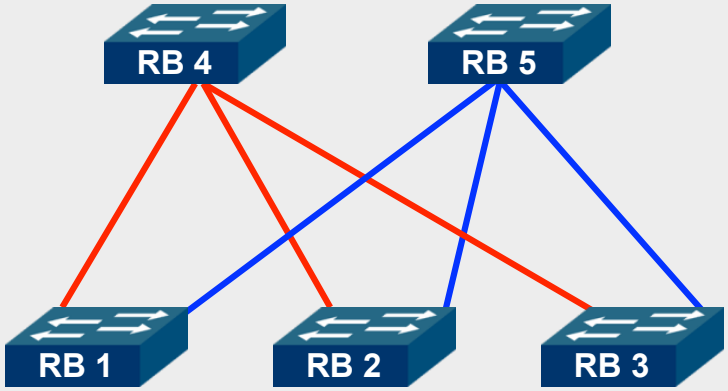
Link State Protocols

IS-IS and FSPF

IS-IS NOT approved yet

IS - IS

- Intermediate System to Intermediate System
- Used by SP and less by enterprises
-
-
-
-
-

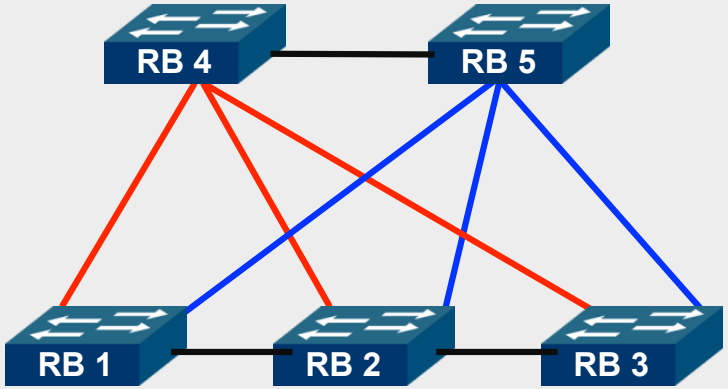


RBridges using IS-IS

  - A routing table update
- Stable, needs modification for TRILL

FSPF

- Fabric Shortest Path First
-
-
-
-
-
-

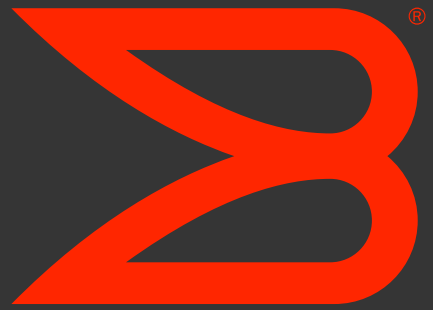


ol

RBridges using FSPF

  - A routing table update
- Stable and tested standard

BROCADE



TRILL Discussion

TRILL Solution

New Concepts

- TRILL Encapsulation
  - TRILL frame
  - TRILL EtherType
  - TRILL header

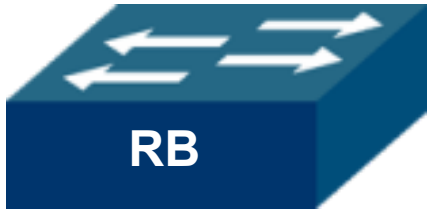
- TRILL Header
  - 64-bit field
  - Contains ingress and egress RBridge nicknames
  - Contains hop count

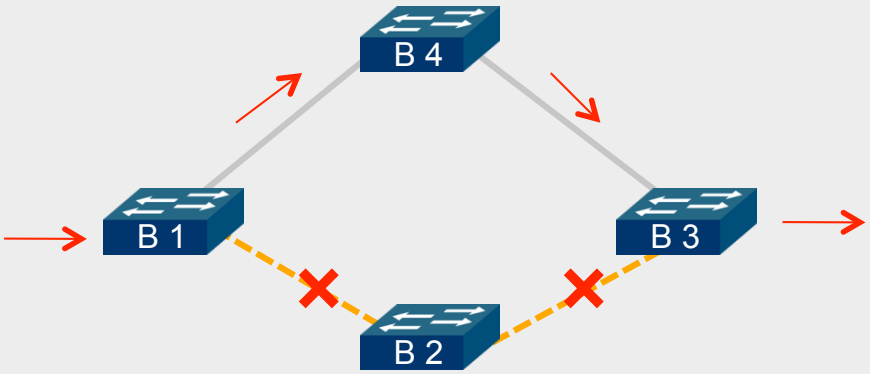
TRILL Ethertype	V	R	M	OpLng	Hop
Egress RBridge Nickname	Ingress RBridge Nickname				

- Routing Bridges
  - Identified by a 16-bit “nickname”
  - Nicknames are auto configured local names
  - Ingress RBridge encapsulates TRILL frames
  - Egress RBridge decapsulates the TRILL frames

- Link state protocols
  - Discover configurations
  - Calculate shortest paths

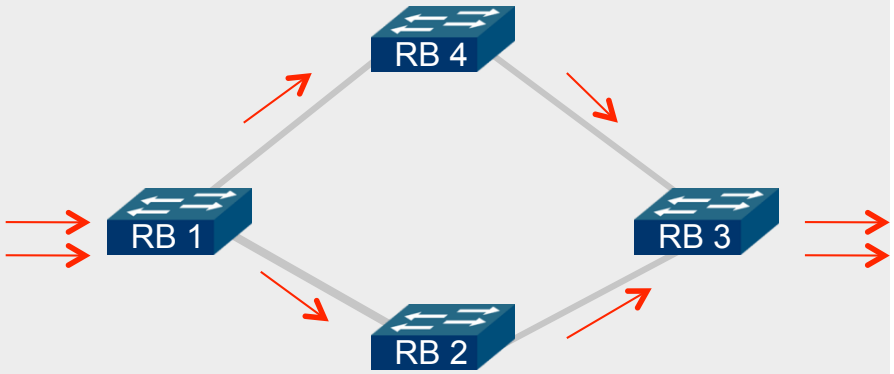
Classic Bridges and RBridges





Classic bridges use spanning tree

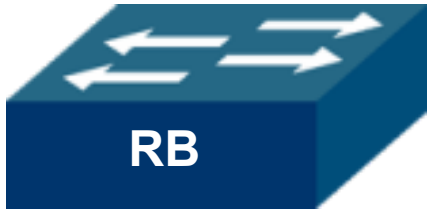
Eliminate loops by disabling ports



RBridges use routing protocols

RBridges support multipathing for better utilization of links

RBridges
Overview



- Implement TRILL protocol
- Perform L2 forwarding
- Use link state routing
- Provide point-to-point forwarding with zero configuration
- Can auto configure themselves
- RBridges forwarding tables scale with the number of RBridges
- RBridges know what options other RBridges support

- Support multi-pathing for unicast and multicast traffic
- Compatible with classic bridges and can be deployed in bridged LANs
- Ingress RBridge adds *TRILL* & *outer MAC headers* to frames
- Outer MAC header is modified hop-by-hop as with routing
- Egress RBridge decapsulates the frame and learns the association of the “Inner MAC SA” with the Source RBridge nickname

RBridges

Personality & Behavior!!

Routers?

- Decrement a hop count in TRILL frames on each hop
- Swap the outer addresses on each RBridge hop from ingress to egress
- Use routing protocols, not STP
- Optionally learn MAC addresses by distribution through the control messages
- Use IP multicast control messages such as IGMP and restrict the distribution of IP multicast frames

Bridges?

- Deliver frames from the source RBridge to the destination RBridge without modification
- Support restricting frames to VLANs like IEEE 802.1Q bridges
- Support frame priorities like IEEE 802.1Q bridges
- By default, learn MAC addresses from the data frames they receive
- Can operate with zero configuration and auto configure themselves

Role of Link State Routing

Discovery & Shortest Path

- Link-state routing protocols are used to
  - Discover RBridge peers
  - Determine RBridge VLAN topology
  - Establish L2 delivery using shortest path calculations
  - Routers tell every router on the network about their closest neighbor
  - The routers only distribute parts of the routing table containing its neighbors

- Link-state routing neighbor information
  - Gathered continuously
  - The list is flooded to all neighbors
  - Neighbors in turn send it to all of their neighbors and so on
  - Flooded whenever there is a (routing-significant) change
  - Allows routers to calculate the best path to any router on the network

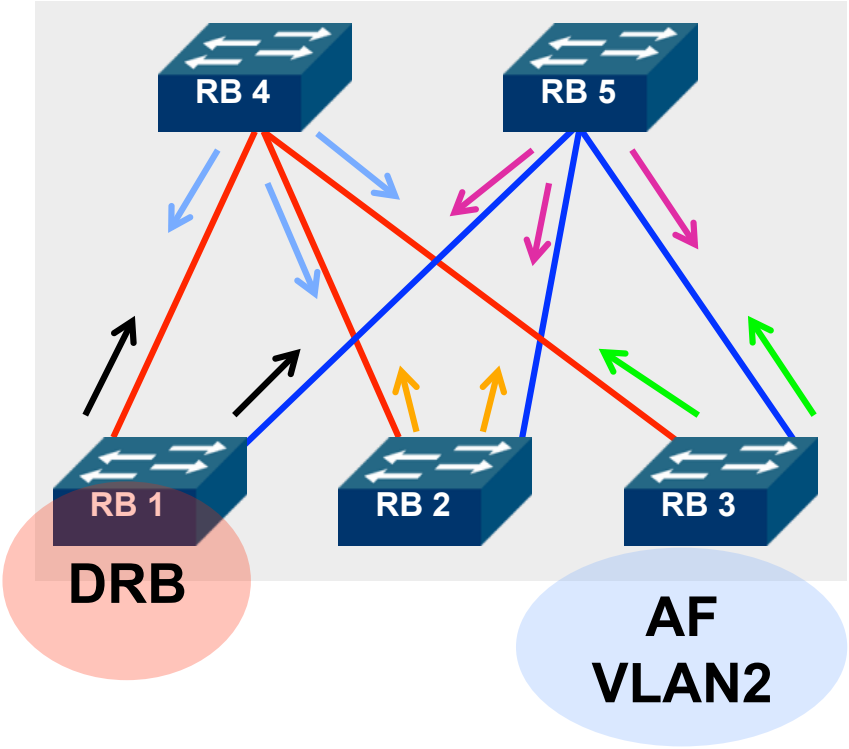
How RBridges Work?

Building the Routing Table

- RBridges need to maintain information about
  - Peer information
  - Topology information
  - Forwarding information : unicast, flooded, and multicast
- Link state protocols used to carry routing information about MAC addresses devices connected to VLANs
- Each RBridges uses the flooded information to
  - Construct a map of the VLAN
  - Calculates the shortest path from it to every RBridge on the VLAN
- The collection of the next best hop maps form the RBridge Routing Table
- Each RBridge has a copy of the global “link state” database

How RBridges Work?

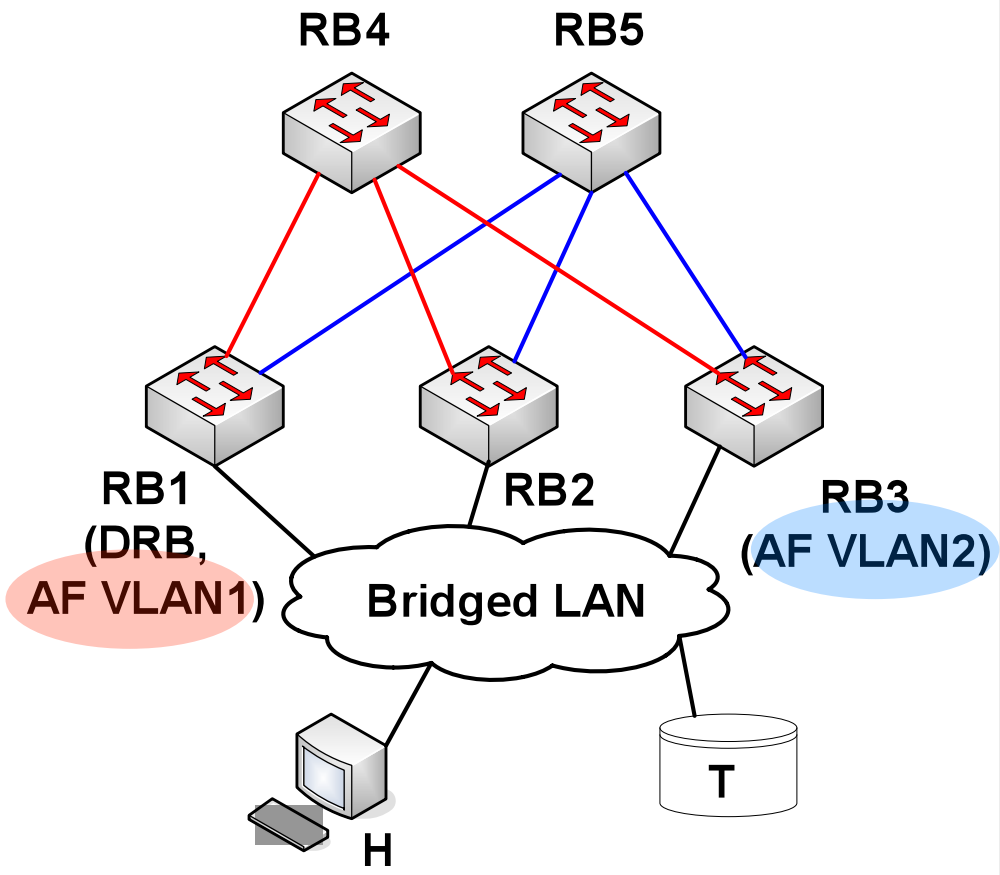
Designated RBridge - DRB



- RBridges discover each other by exchanging TRILL IS-IS (or FSPF) Hello frames
  - TRILL Hellos are sent to the All-IS-IS-RBridges multicast address
- Using link state protocol (IS-IS or FSPF), a single Designated RBridge (DRB) is elected from among all RBridges on the LAN
  - The DRB specifies the Appointed Forwarder (AF) for each VLAN
  - The DRB also specifies the Designated VLAN for inter-RBridge communication

How RBridges Work?

Appointed Forwarder - AF

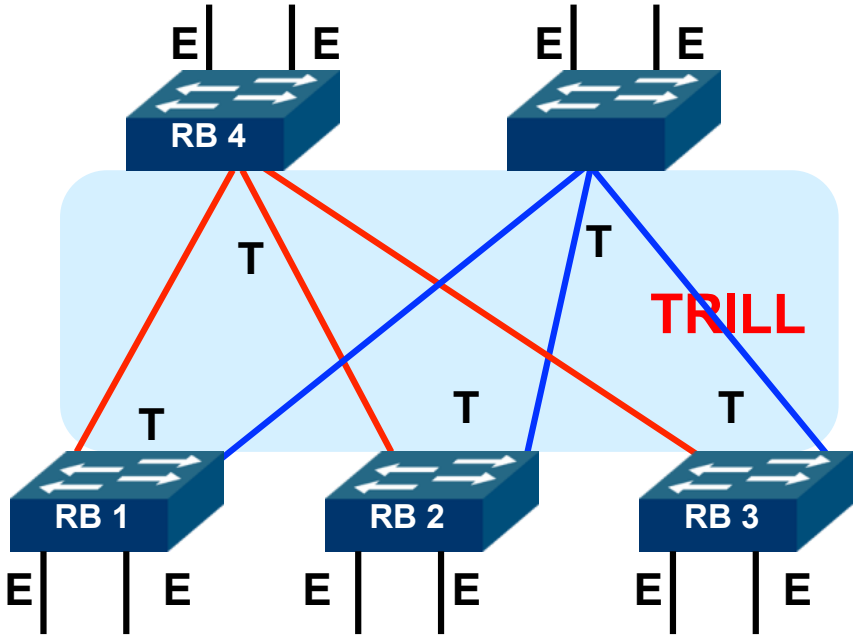


- The DRB specifies the Appointed Forwarder (AF) for each VLAN
  - DRB can also be the AF
- Only *ONE* AF can be appointed per VLAN; One VLAN - One AF
- The AF is in charge of handling all native frames in the VLAN
  - Ingress RBridge function: Encapsulates TRILL data frame
  - Egress RBridge function: Decapsulates TRILL data frames

TRILL Ports & Processing

Ethernet & TRILL Ports

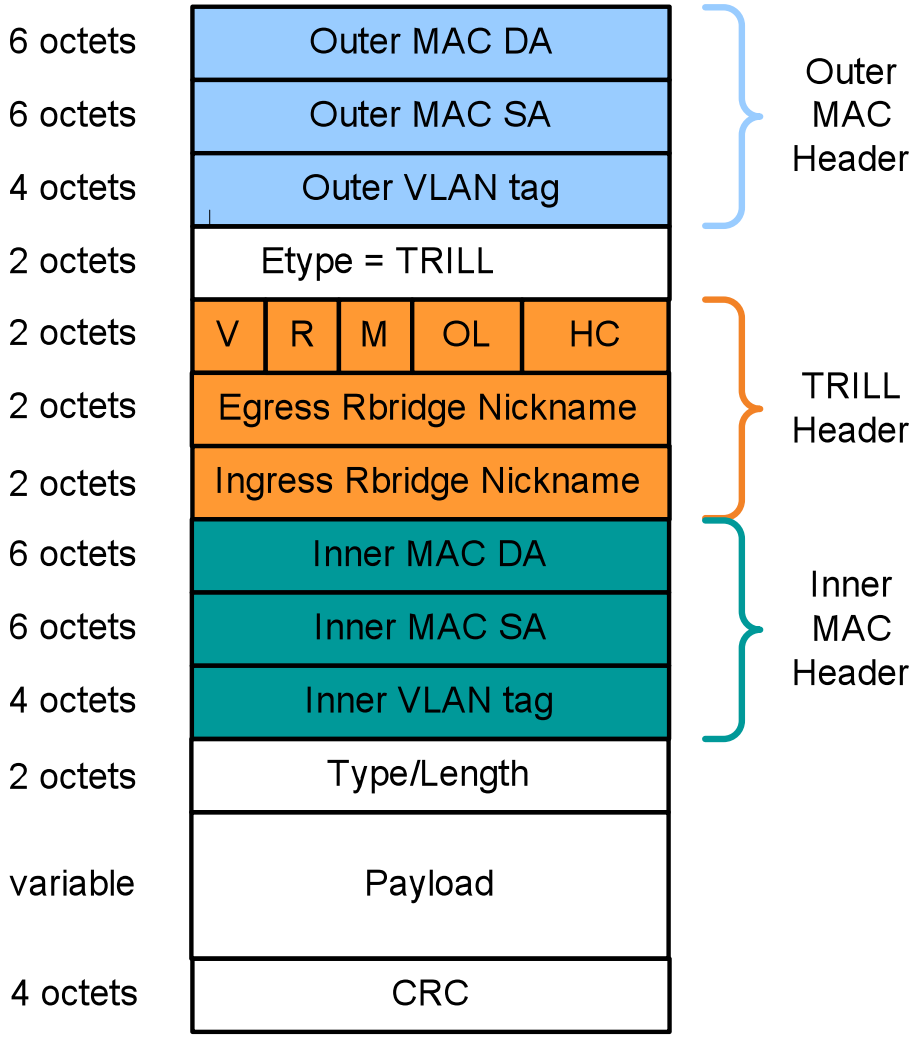
- Ports designations
  - Ethernet ports; E ports
  - TRILL ports; T ports
- Processing categories
  - Ingress: from E port to T port; E-T
  - Core: between T ports; T-T
  - Egress: from T port to E port; T-E



TRILL is layered above the ports of RBridges

TRILL Frame Format

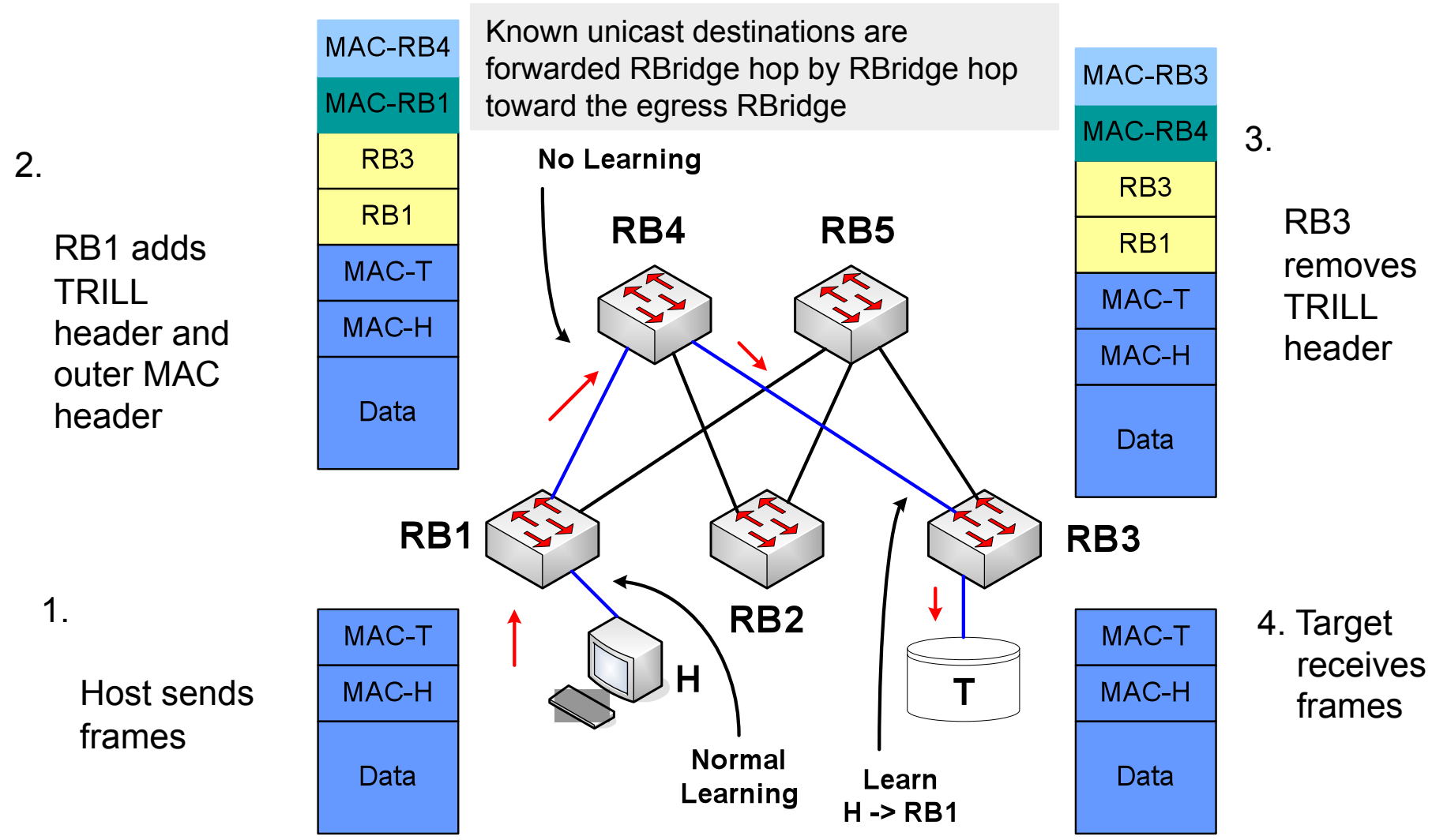
Header length: 64 bits



- Nickname: auto-configured 16-bit local names for RBridges
- V = Version (2 bits)
- R = Reserved (2 bits)
- M = Multi-destination (1 bit)
- OL = Options Length of TRILL options (5 bits)
- HC = Hop Count (6 bits)
- If M = 0, egress Nickname is the egress Rbridge
- If M = 1, egress Nickname is that of the RBridge that is the root of the tree

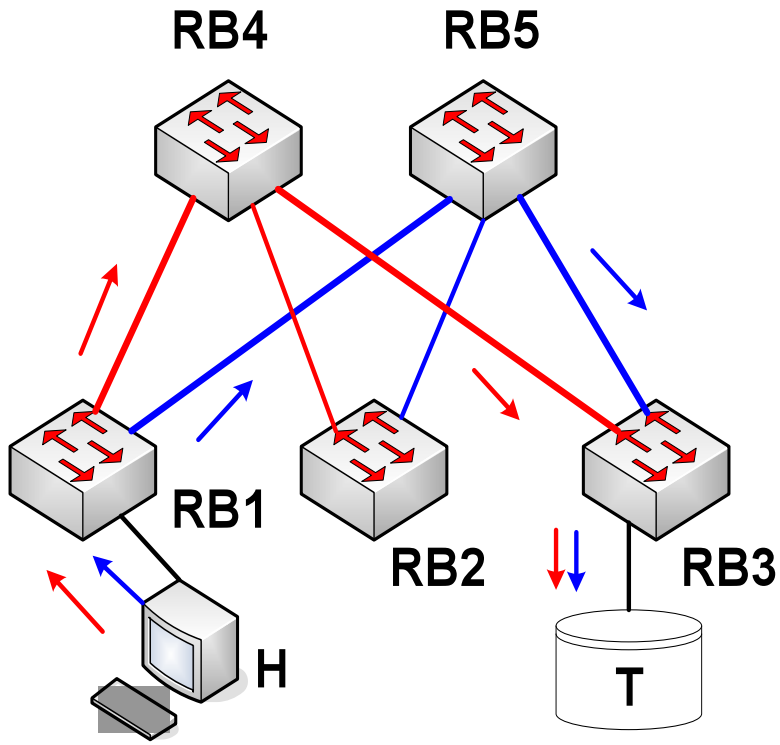
TRILL Encapsulation

Unicast data path

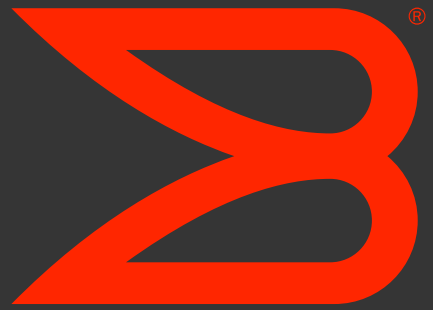


Utilizing ECMP Paths & Reordering

- TRILL supports up to 64 ECMP paths
  - Packet (frames) ordering maintained within flows
- RBridges are required to maintain frame ordering internally
- 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
- Re-ordering can occur when
  - A destination address transitions between being known and unknown
  - A topology change occurs

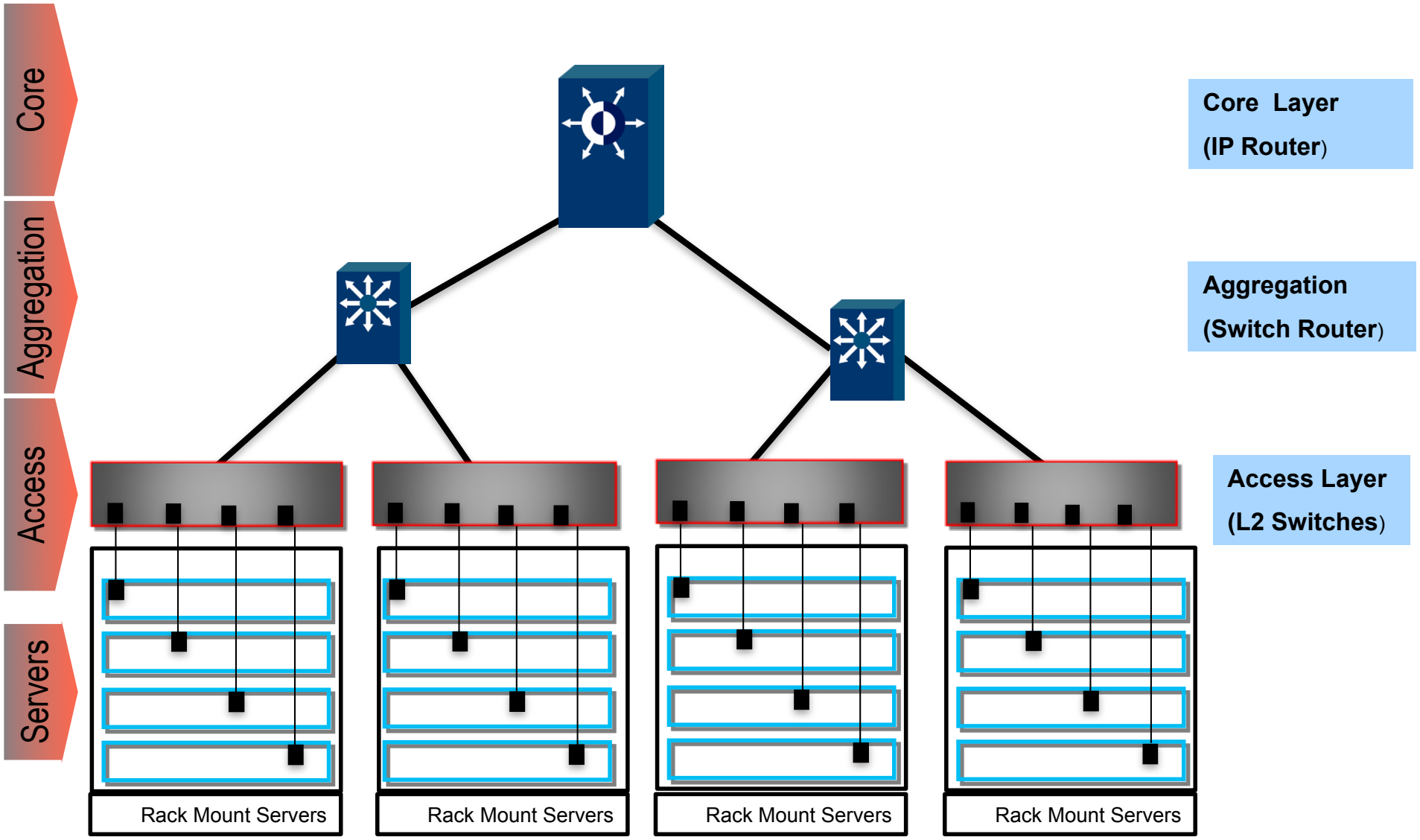


BROCADE

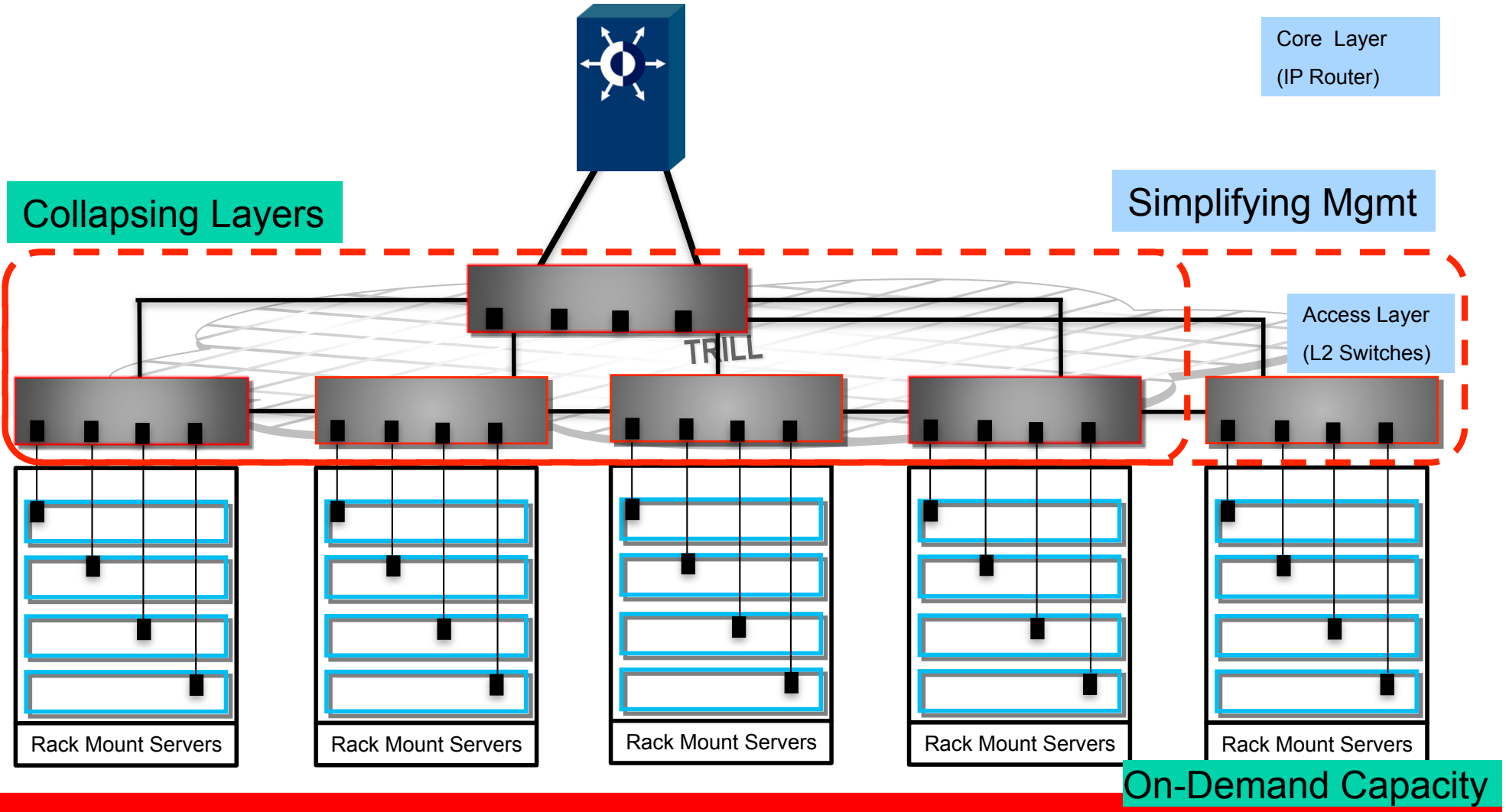


Brocade Solution

Multi-Tier Data Center Network



Next – Simplified & Scalable



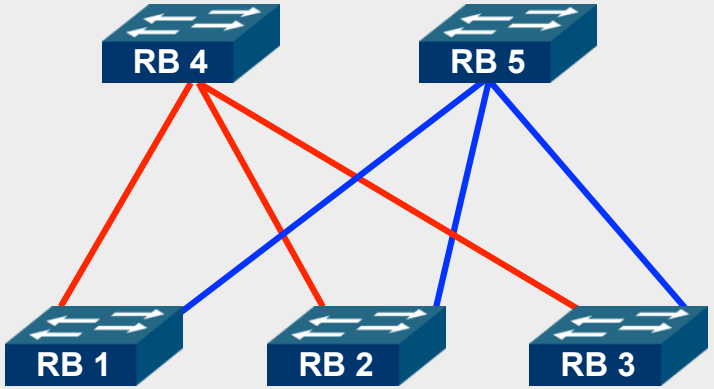
The Brocade Solution

TRILL + Brocade Enhancements

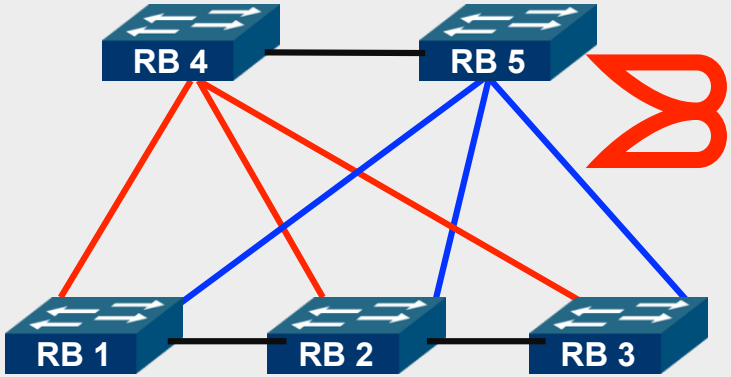
- TRILL-ready hardware
  - Layer 2 multipathing capabilities and multi hop FCoE support
- Brocade Fabric OS (FOS) storage services on each hop
  - Built on storage-aware ASICs and protocol stack (FOS)
  - FSPF running over CEE: multipathing using Fibre Channel standard routing protocol on CEE
  - Link protocol-agnostic fabric services: zoning and name server
- End-to-end troubleshooting and manageability
  - Across link layer protocols: CEE/FCoE and Fibre Channel
  - Across routing protocols: TRILL and FSPF
  - DCFM: Single management tool for all Brocade IP and SAN fabrics

TRILL Control Plane in VCS

- FC Fabric formation protocols are used to form the TRILL fabric
- FSPF is the link state routing protocol used to calculate ECMP capable shortest path routes among RBridges
- RBridge nickname assignment
  - A TRILL RBridge nickname = FC Domain
- Unicast path computation
  - Unicast forwarding is done by combination of domain routing generated by FSPF and MAC-to-RBridge learning generated by MAC learning and a distributed MAC database
  - Low overhead for computing ECMP paths
- Multicast path computation
  - Multicast forwarding uses one tree rooted at the principal switch



RBridges using IS-IS



RBridges using IS-IS

L2 Issues Addressed by TRILL & VCS

L2 Networks

1. Flooding and loop mitigation
2. Control planes are not designed for fast convergence times though RSTP improves it
3. Control planes and data planes are not designed for achieving shortest path and ECMP
4. Though delivers plug-n-play a small mis-configuration can bring down the network
5. Spanning tree is not suited for building large bridged networks

VCS & TRILL

1. TRILL can mitigate loops using the hop count field in the header. VCS will minimize and in some cases avoid flooding
2. Faster convergence using FSPF based control plane
3. ECMP capable control path and data-path
4. VCS will attempt to address some of these issues
5. TRILL & VCS will allow customers to build large flat L2 networks

L3 Issues Addressed by TRILL & VCS

L3 Networks

1. Very configuration centric
2. Have to divide up the address space per subnet/interface
3. Cannot carry other L3 traffic like FCoE
4. Issue for VM mobility as the current Hypervisors assume they have a direct L2 reach-ability in the mobility domain.
5. CN only works in a single L2 domain

VCS & TRILL

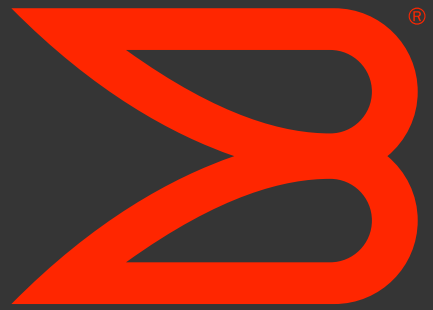
1. Minimal configuration
2. Flat address space
3. L3 agnostic so is capable of carrying FCoE over TRILL
4. Attractive for Virtualization centric data-center deployments
5. End-to-End congestion notification can be deployed

Summary of Key Messages

Superior Layer 2 Brocade solutions

- TRILL solution
  - Layer 2 multipathing and multi-hop FCoE support
- Brocade will support and exceed TRILL
  - Extending FOS storage services to every hop
    - FSPF over CEE brings Fibre Channel routing to Layer 2
- Brocade delivers unified & superior solution
  - All Ethernet and Fibre Channel resources unified under DCFM
  - New FOS services based on proven Fibre Channel technology

BROCADE



THANK YOU