Cloud Networking (VITMMA02)
Data Center Bridging, Network virtualization technologies

Markosz Maliosz PhD

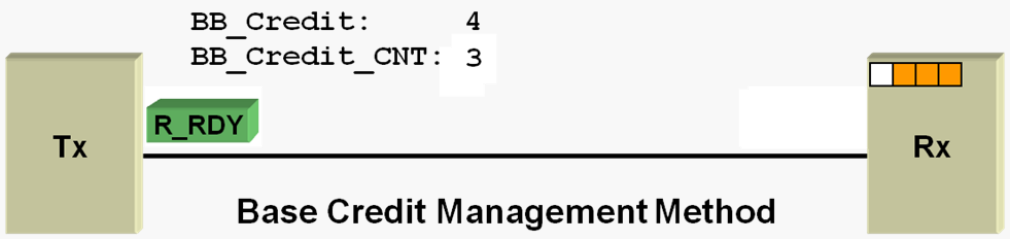
Department of Telecommunications and Media Informatics
Faculty of Electrical Engineering and Informatics
Budapest University of Technology and Economics

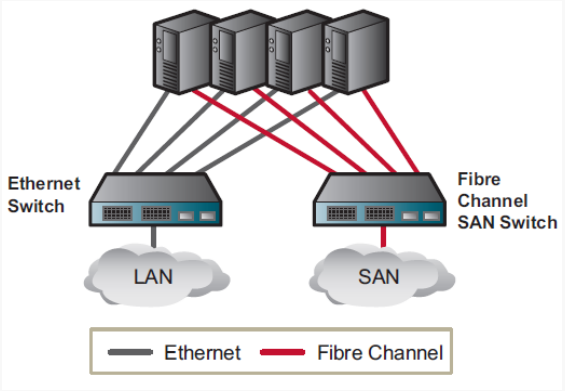
Spring 2018

DATA CENTER BRIDGING

Storage traffic in the data center

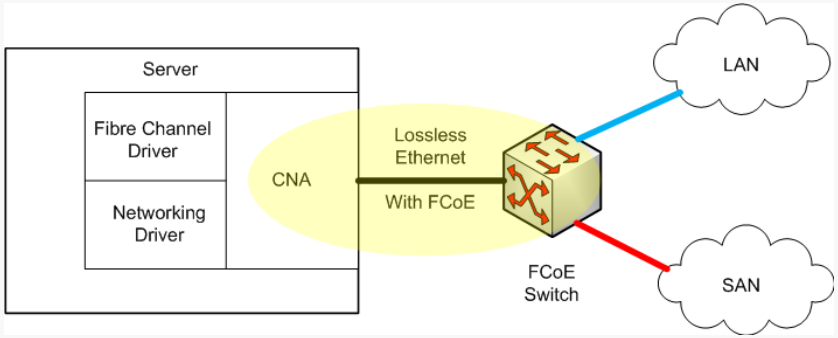
- Earlier data centers
  - Ethernet for data traffic
  - Fibre Channel for storage traffic (SAN – Storage Area Network)
    - different dedicated networks
    - optical or electronic interface
    - 2, 4, 8, 16 Gbps
    - in case of congestion no packet drops
      - buffer credit based flow control
      - buffer to buffer credit

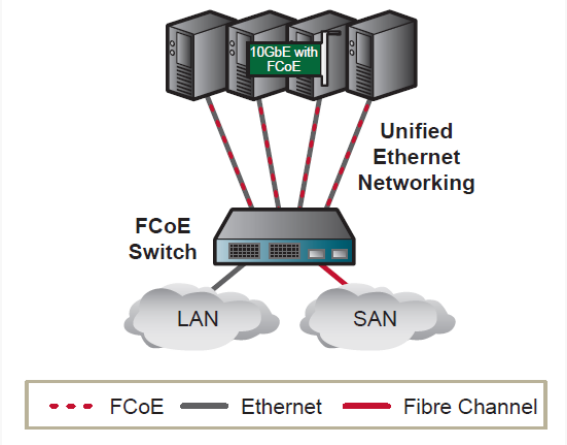




Fibre Channel over Ethernet (FCoE)

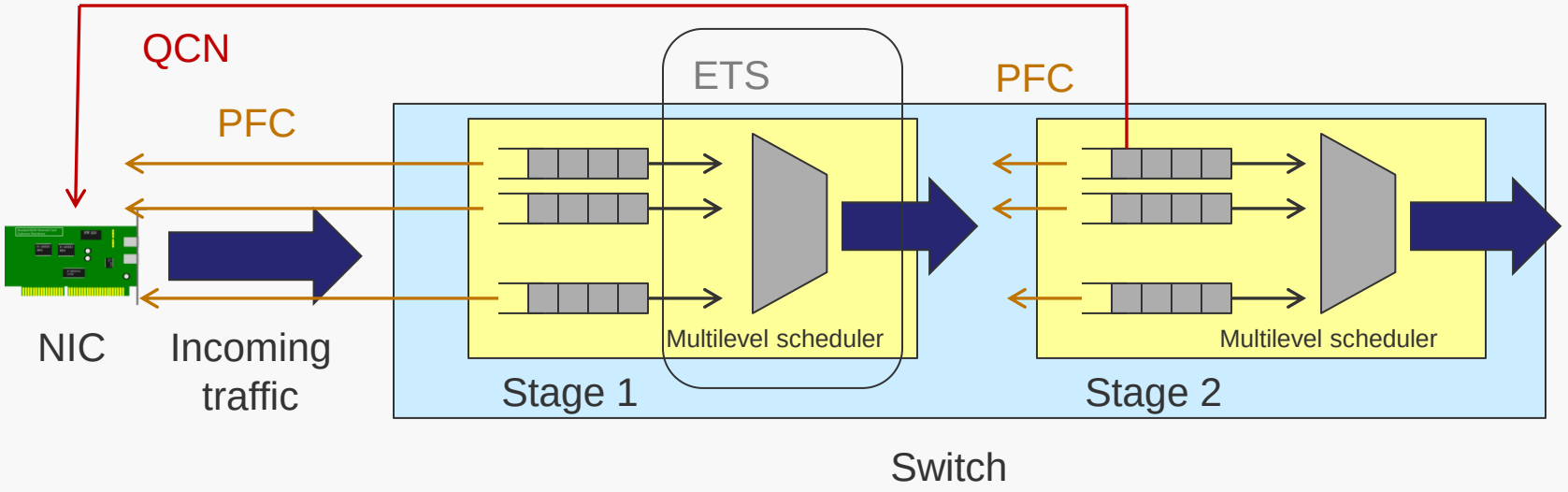
- Ethernet
  - in case of congestion packets might be dropped
  - TCP: reliable delivery (retransmission)
    - delay jitter
    - not ideal for video and storage traffic
  - required extensions: DCB





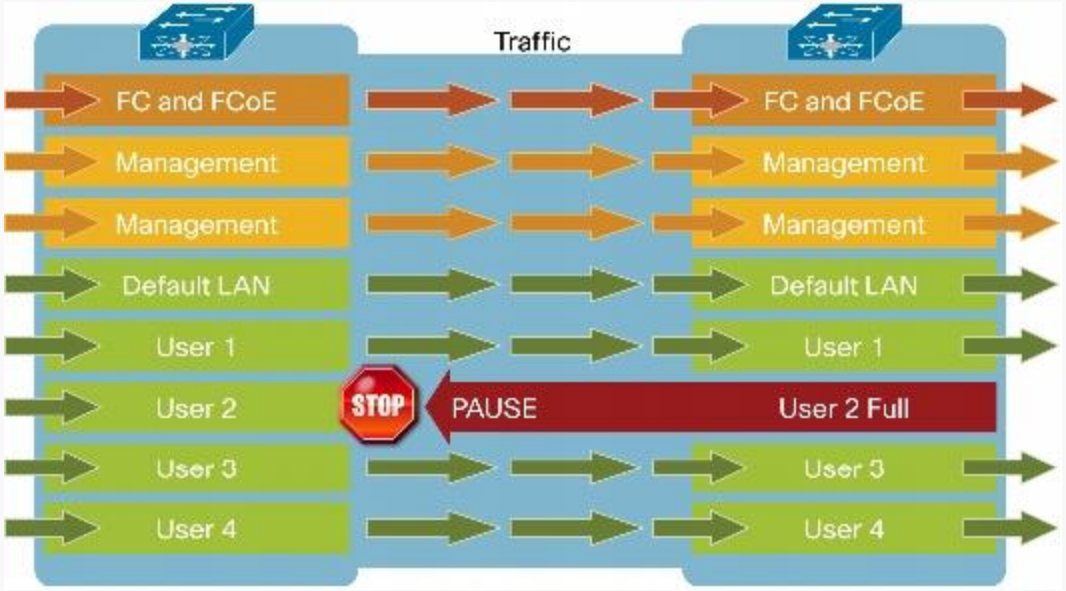
Data Center Bridging

- Ethernet extensions: (more) reliable delivery without the complexity of TCP
  - Priority based Flow Control (PFC)
  - Enhanced Transmission Selection (ETS)
  - Quantized Congestion Notification (QCN)
  - Data Center Bridging exchange (DCBx) protocol



Priority based Flow Control

- To provide lossless operation
- IEEE 802.1Qbb
  - link level
  - between switches or switch stages
- 8 priority class (802.1p): virtual lanes
- inside switch: allocated memory partitions
  - check if watermark is crossed
- pause message includes a duration



Source: Cisco

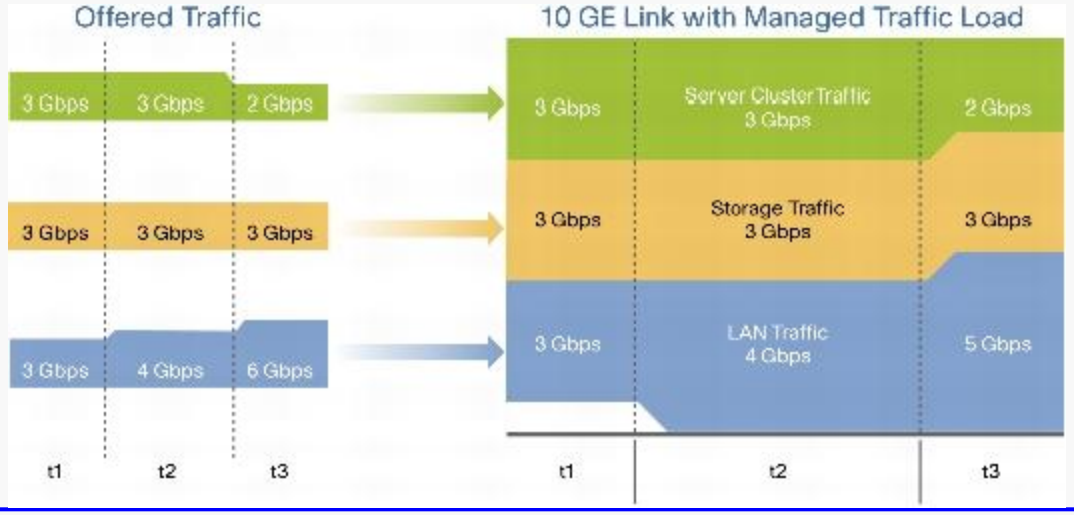
Enhanced Transmission Selection

- IEEE 802.1Qaz
- Traffic classes
  - classification
    - rule based header matching: Access Control List (ACL)
    - 3-bit priority filed in VLAN tag
  - scheduling may by applied to Traffic Class Groups (TCG)
    - an ETS capable switch is required to support for at least three traffic classes

Prio1: 2 Gbps

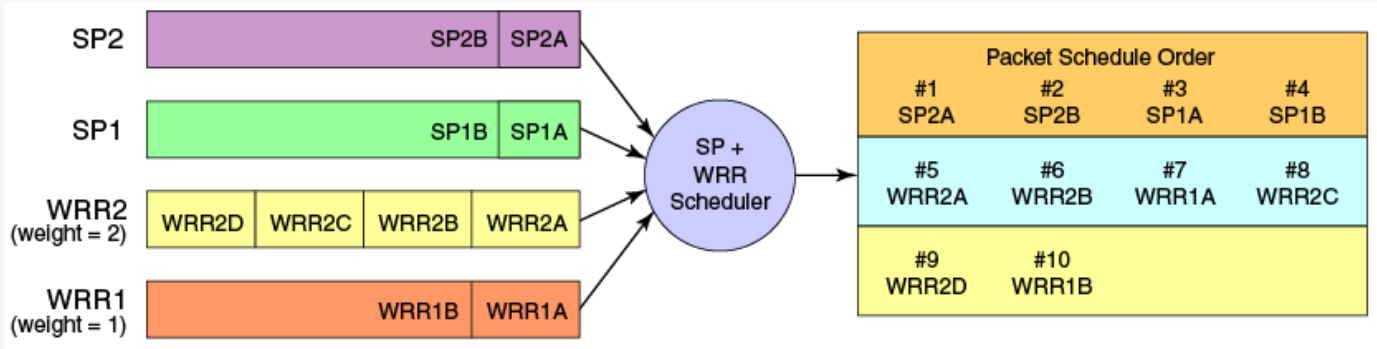
Prio2: 3 Gbps

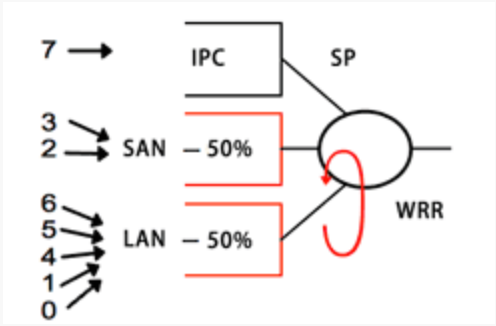
Prio3: 5 Gbs



Enhanced Transmission Selection

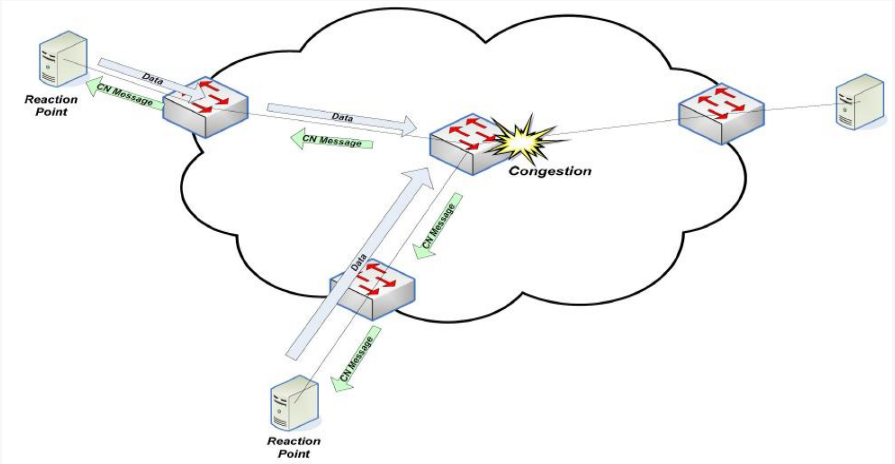
- Bandwidth allocation
  - bandwidth to be configured for each traffic class (max. 8)
    - with a granularity of 1% with allowed deviation of +/-10%
  - any unused bandwidth is available to other traffic classes
- Implementation: scheduling and rate limiting, shaping





Quantized Congestion Notification

- PFC + ETS
  - lossless transmission and bandwidth guarantees
  - quick reaction time
  - however: many hops through switches and multi-stage switches inside the data center
- QCN (802.1Qau): for minimizing transient congestions
  - feedback to the source (ent-to-end)
  - larger time-scale
  - congestion point
    - reads the queue length from the switch, random sampling (depending on queue fill level)
    - calculates a feedback value based on the queue fill level info (quantized to 6 bits)
    - sends back to source MAC address (reaction point)
      - with probability of 1-10%
    - updates the queue sampling rate
  - reaction point
    - rate limiting traffic based on the feedback value
    - then slowly increased again



Quantized Congestion Notification

- Rarely implemented in data centers
  - the control loop is highly dependent on factors such as
    - congestion point reaction time, time to send the QCN frame back through the network, and the reaction point queue throttling time
  - requires a lot of fine tuning
  - ideal for long lived flows
  - uncertainty: frames are randomly sampled
  - at the source one queue should be allocated for each potential congestion point
  - operates inside L2 subnets
    - traffic crossing a router lands in another QCN domain
  - for high traffic rates the proper implementation is by hardware
    - replacement of all NICs and switches

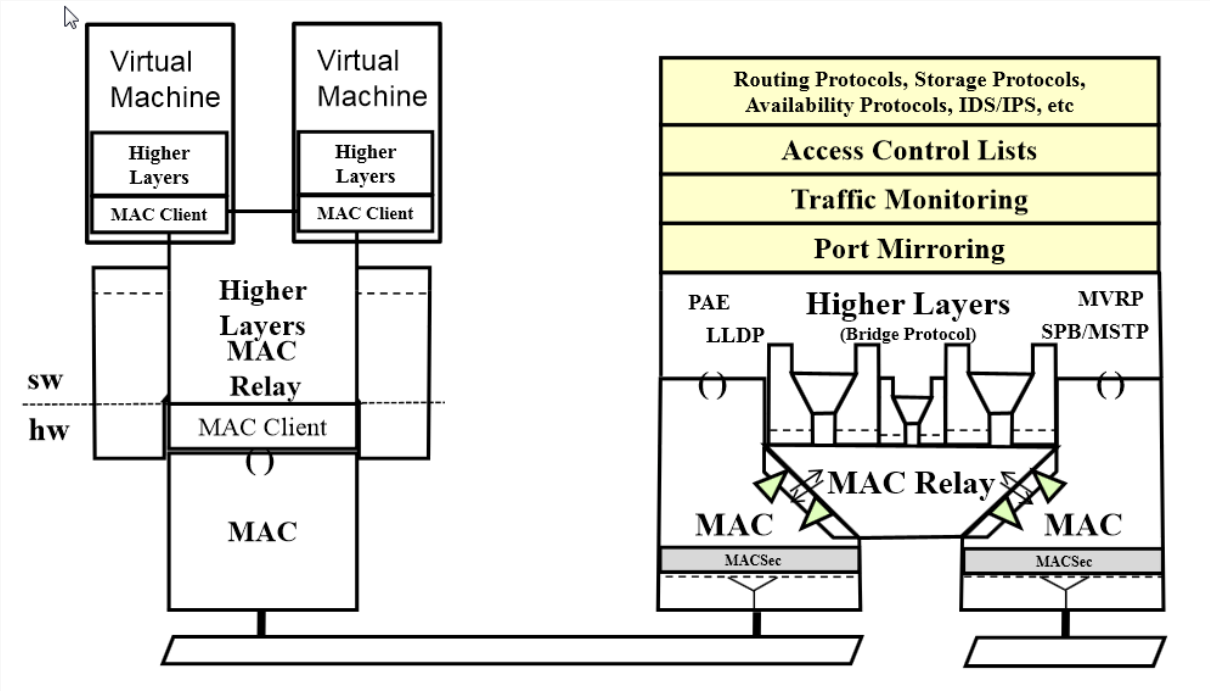
Data Center Bridging exchange (DCBx)

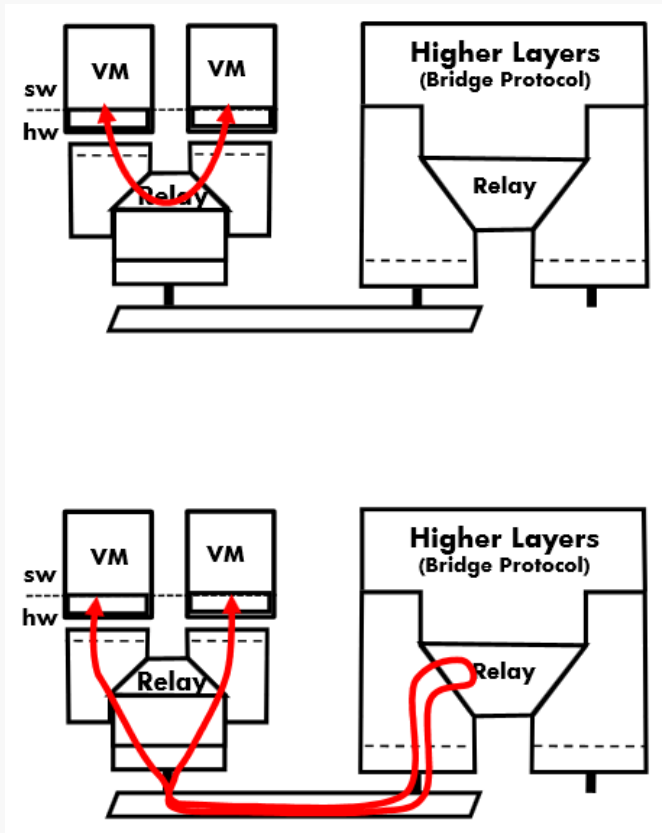
- Coordination between neighboring devices
  - PFC
    - number of priorities or traffic classes
  - ETS
    - allocated bandwidth units
- Link Level Discovery Protocol (LLDP) messages
  - Type-Length-Value structure
- Operation
  - sending side
    - *suggests* parameter settings to the remote end
    - sent at a periodic rate
  - receiving side
    - setting up parameters taking into consideration of the configuration received from the other side
    - database update based on received data
  - does not expect, process, or generate acknowledgements
    - does not care what the remote side does

NETWORK VIRTUALIZATION TECHNOLOGIES

Edge Virtual Bridging

- ToR physical switch ⇔ virtual switch (Virtual Ethernet Bridge – VEB) capabilities
  - filtering, security, monitoring, etc.

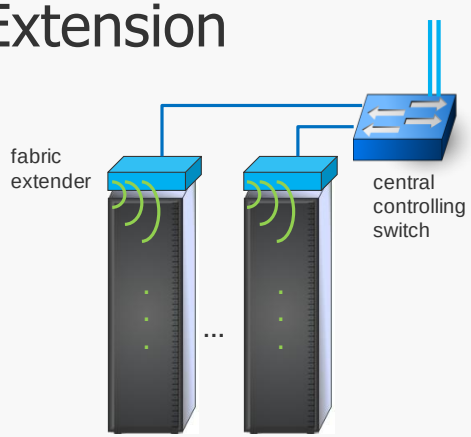




Forrás: Pat Thaler et al., IEEE 802 Tutorial: Edge Virtual Bridging, 2009.

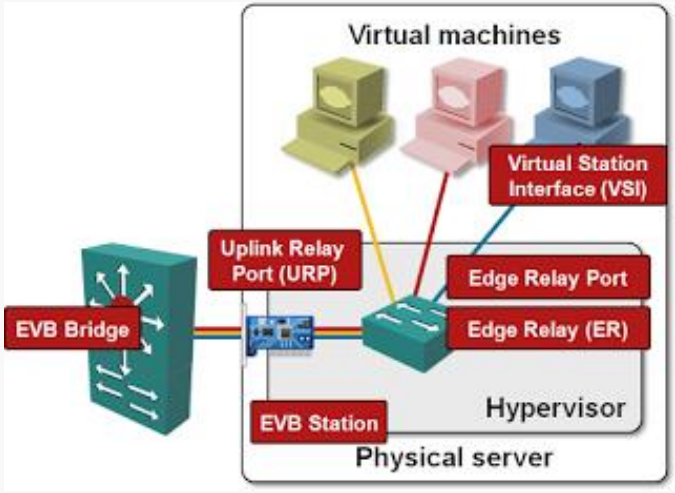
Edge Virtual Bridging

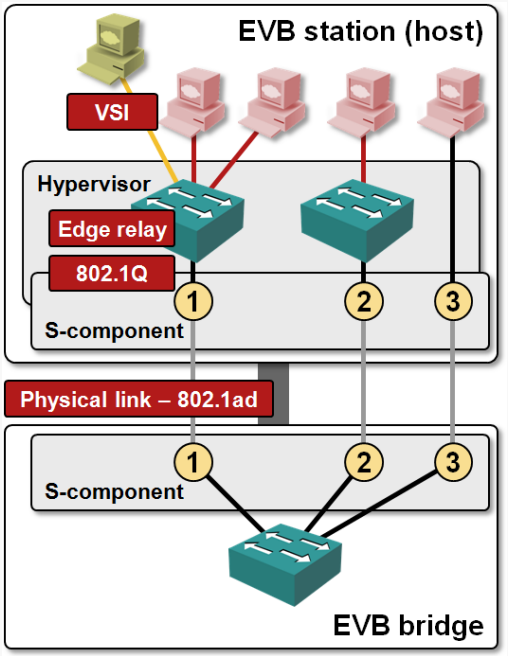
- EVB: IEEE standard
  - interaction between the physical and virtual switches
    - capability of the physical switch
    - goal: handle all traffic uniformly
  - Virtual Ethernet Port Aggregation (VEPA) 802.1Qbg
    - server side capability
    - all traffic is forwarded to the neighboring physical switch
    - multi-channel: S-Tag (Q-in-Q)
- Identifying virtual interfaces on a physical port
  - Virtual Network Tag (VN-Tag), Bridge Port Extension 802.1Qbh, 802.1BR (E-Tag)
    - ports configured by central controlling switch
      - on fabric extender (S-Tag)
      - on NIC of server (VN-Tag)
        - for each vNIC a separate VN-Tag
        - extra header containing Virtual Interface (VIF)



L2 configuration automation

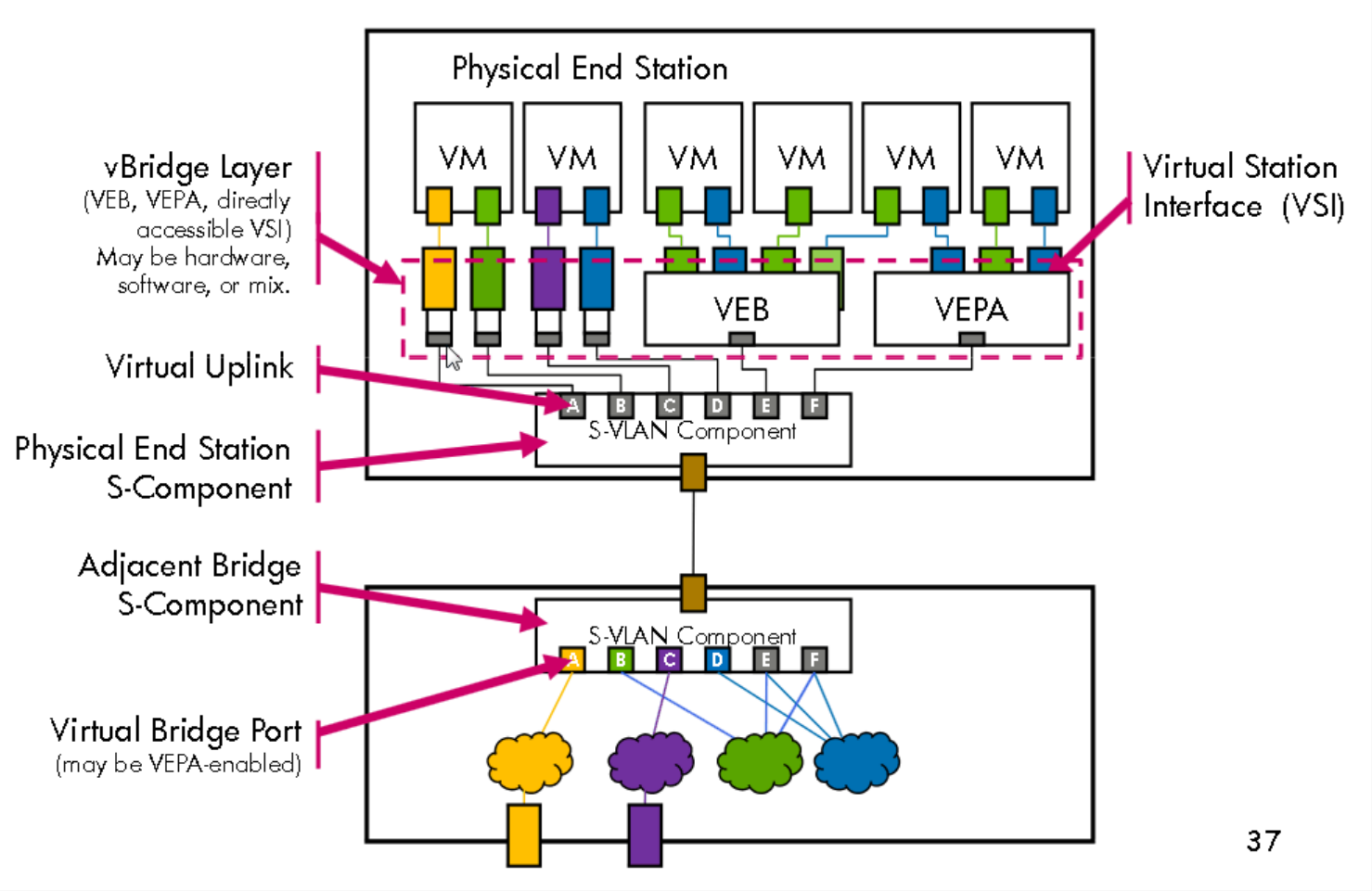
- Edge Virtual Bridging
  - Virtual Station Interface (VSI): VM NIC
  - VSI Discovery and Configuration Protocol (VDP)
    - EVB bridge receives info from the hypervisor before starting the VM
  - VN-Tag: extra header for identifying vNIC (Cisco)
    - local tag between the controlling switch and the fabric extender
  - S-component
    - multiplication of logical 802.1Q links over a physical link (Q-in-Q)





Edge Virtual Bridging

» Combining different technologies



Forrás: Pat Thaler et al., IEEE 802 Tutorial: Edge Virtual Bridging, 2009.

Evaluation

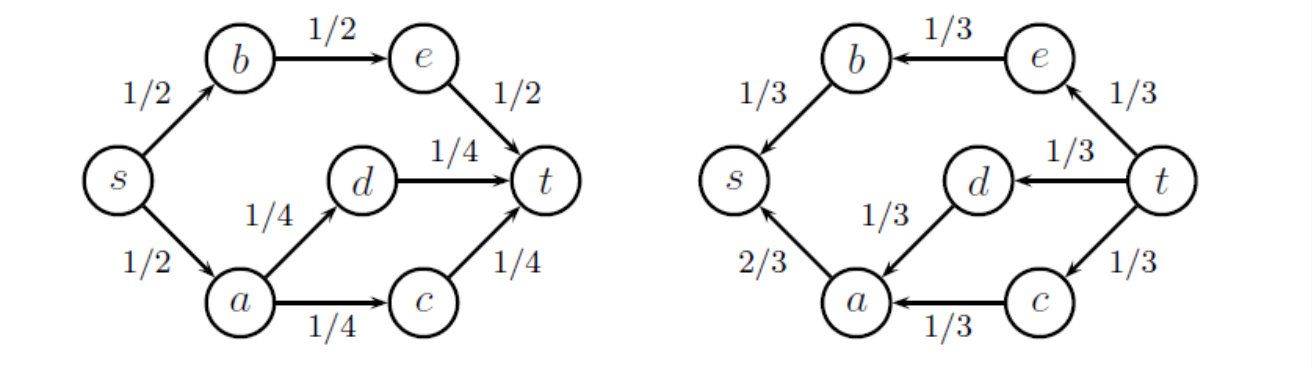
- Virtual switch (VEB)
  - forwarding by MAC + VID
  - not needed
    - MAC address learning, because VM addresses can be preconfigured
    - STP, because located at the edge of the network
  - traffic kept inside the server
    - not visible, analyzable, filterable from outside
    - better performance for VMs residing on the same server
  - no common management with the physical switched
  - CPU and RAM usage on the server
- EVB
  - all traffic crossing the physical switches (more advanced features)
  - less network configuration task
  - more traffic and delay in the network
  - VEPA
    - forwarding by MAC + VID
    - function of virtual switch is kept
    - Ethernet frames
    - capability for sending the traffic back on the input port (hairpin)
  - VN-Tag
    - forwarding by tags
    - new frame format
- Applicability of technologies
  - VEPA: hypervisor support required
  - VN-Tag: special NIC required
  - other directions
    - physical switch features integrated into virtual switches
    - other network virtualization and tunneling technologies (VXLAN, NVGRE, etc.)

Improving network utilization

- Ethernet Spanning Tree Protocol
  - spanning tree: unused links
  - Rapid STP (RSTP)
  - Multiple STP (MSTP)
  - ideal for arbitrary and changing topologies
- But not ideal for data centers
  - structured and not frequently changing
  - new standards
    - Equal Cost MultiPath (ECMP) routing
    - Shortest Path Bridging (SPB)
    - Transparent Interconnection of Lots of Links (TRILL)

ECMP

- Equal Cost MultiPath

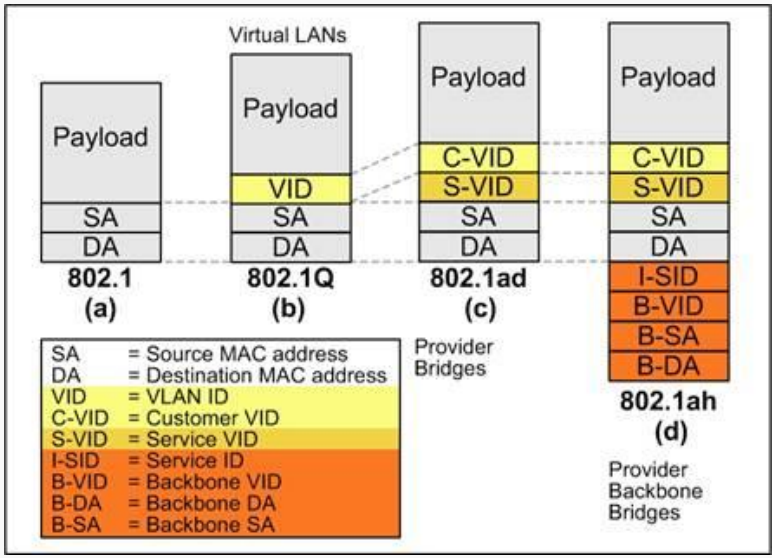


- Layer3 routing or tunneling between Layer2 domains
  - L2 over L3
- generally not used in networks
  - if routes join before the destination, only the complexity is enlarged, but not the bandwidth utilization
  - virtual network ⇔ physical network

Shortest Path Bridging

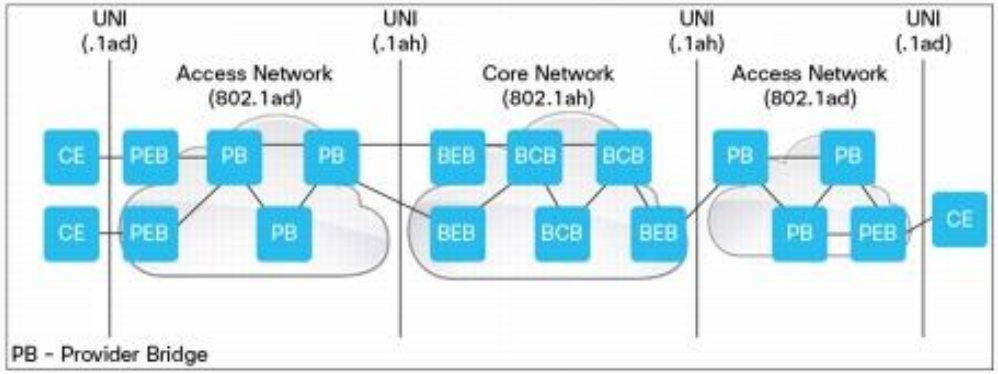
- Origins: Carrier Ethernet
  - Provider Bridging (PB) 802.1ad
  - Provider Backbobe Bridging (PBB) 802.1ah
- Shortest Path Bridging (SPB) 802.1aq

Carrier Ethernet



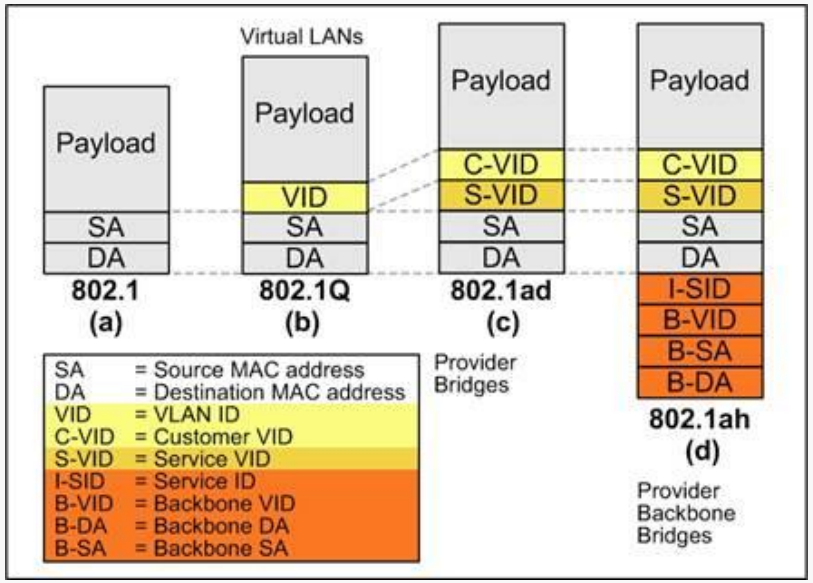
Ethernet in carrier networks (MAN, WAN)

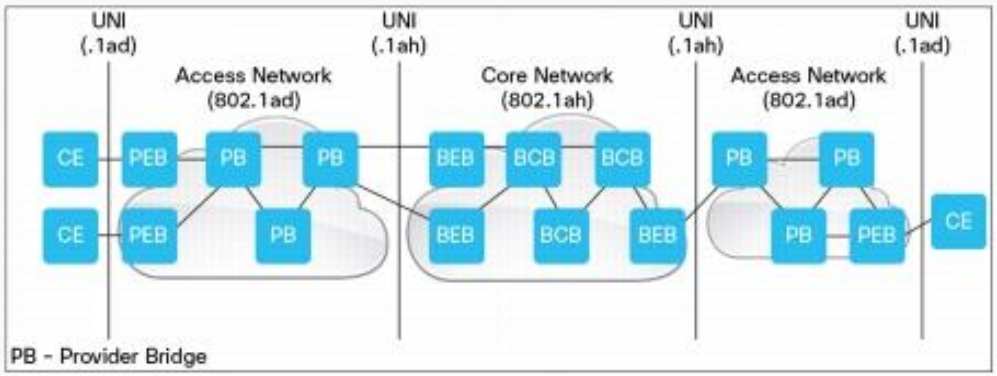
- Ethernet service for many customers
  - separating customers
- tunneling by additional tags
  - keeping customer VLAN information
  - separating service instances (customers) (PB)
    - Q-in-Q: Customer tag, Service tag
    - two VLAN IDs (VID)
    - 4096 service instance (upper bound)
  - complete separation of customer and provider domains (PBB)
    - MAC-in-MAC: separated address space
      - customer addresses are not seen by switches in the carrier network
    - service tag: 24 bit I-SID (service identifier) 16M service instances
    - separating service and transport layers: I-SID and B-VID



Carrier Ethernet

- Mapping virtual networks at the edge
  - C-VID ⇨ S-VID ⇨ I-SID ⇨ B-VID
  - Edge Brigdes
- In the core network: forwarding based on VLAN ID and destination MAC address
  - Core Bridges

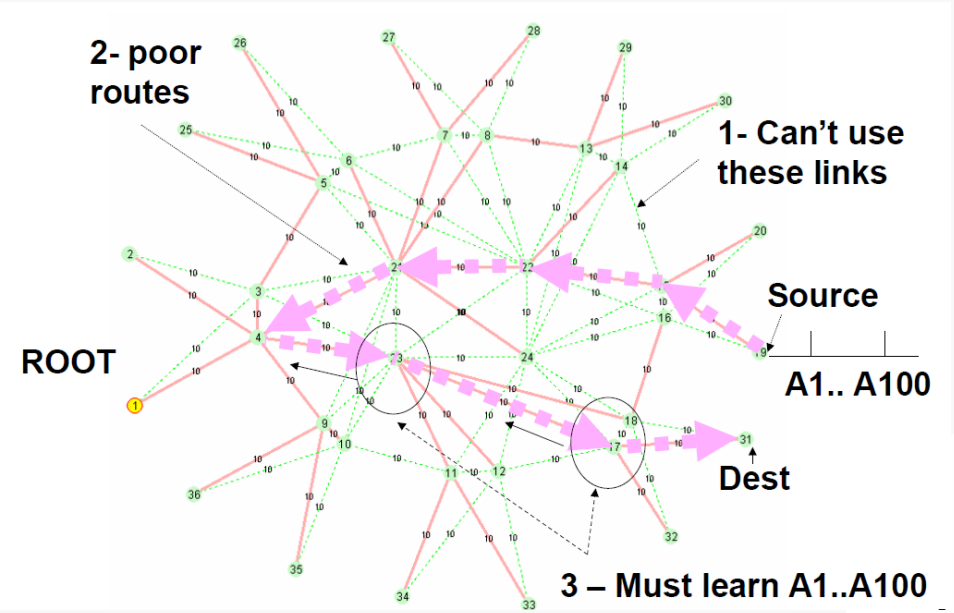


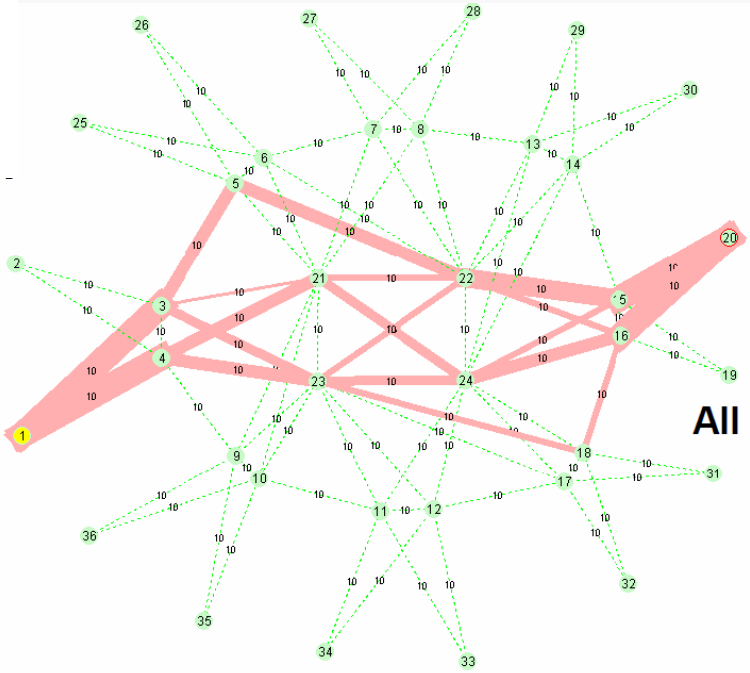


Shortest Path Bridging

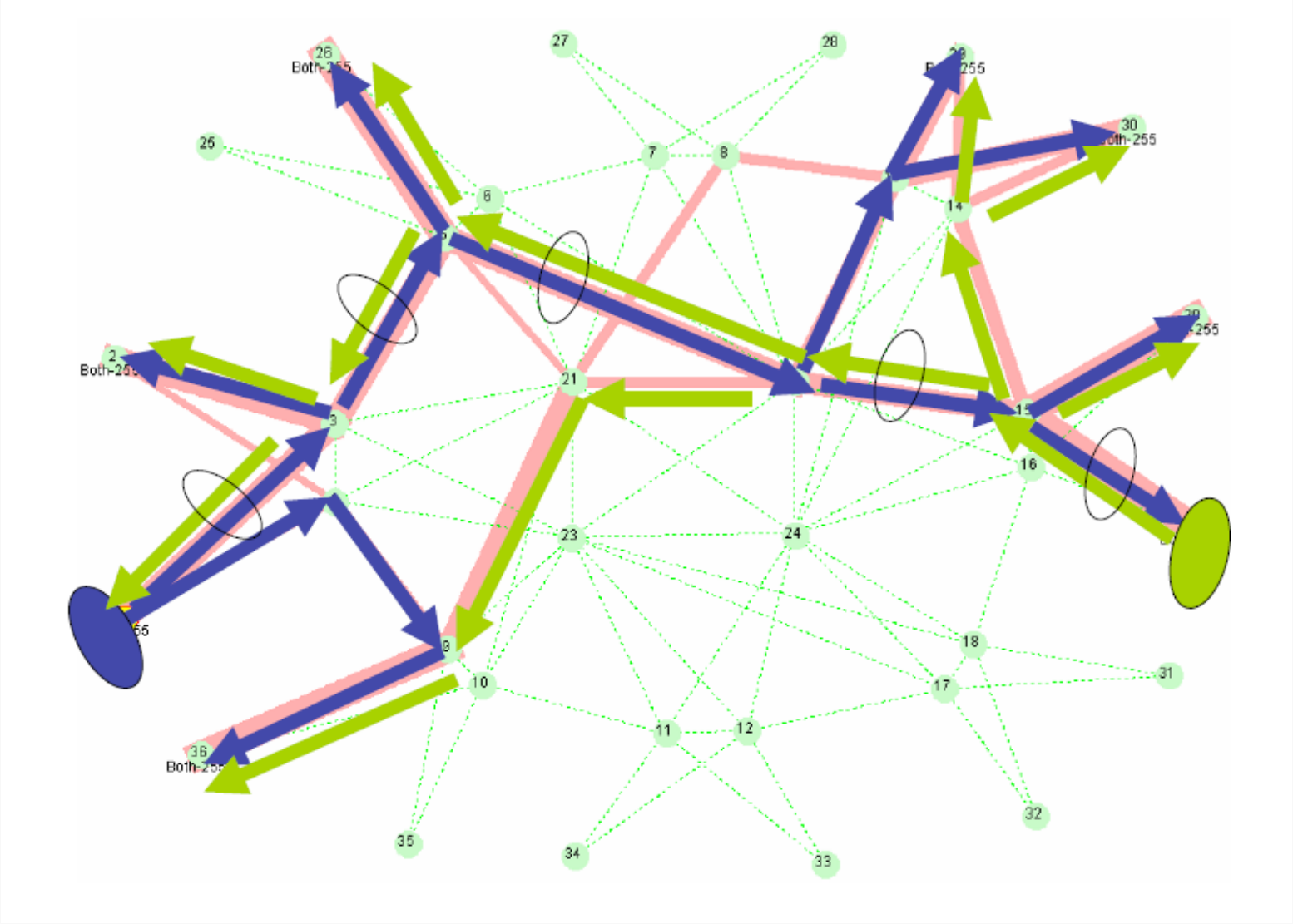
- Replacing STP with a new control plane
  - providing logical networks over native Ethernet
  - link state protocol advertising the topology and the logical network membership
    - Intermediate System to Intermediate System (IS-IS) with extensions with link state protocol: IS-IS
      - runs directly at Layer 2
      - no IP addresses are needed, as they are for OSPF
      - IS-IS can run with zero configuration
      - with TLV (type, length, value) encoding new types of data
    - automatic link state discovery
    - no blocked ports, links
  - using equal cost *multiple* shortest paths
  - sources calculate a shortest path tree
    - symmetric forward-backward paths
- Encapsulation
  - PB ⇨ SPB Vlan ID (SPBV)
  - PBB ⇨ SPB MAC (SPBM)
- vendors: Avaya, Huawei, Alcatel-Lucent

STP vs. SPB


2- poor routes
1- Can't use these links
Source
ROOT
A1.. A100
Dest
3 – Must learn A1..A100


All links usable

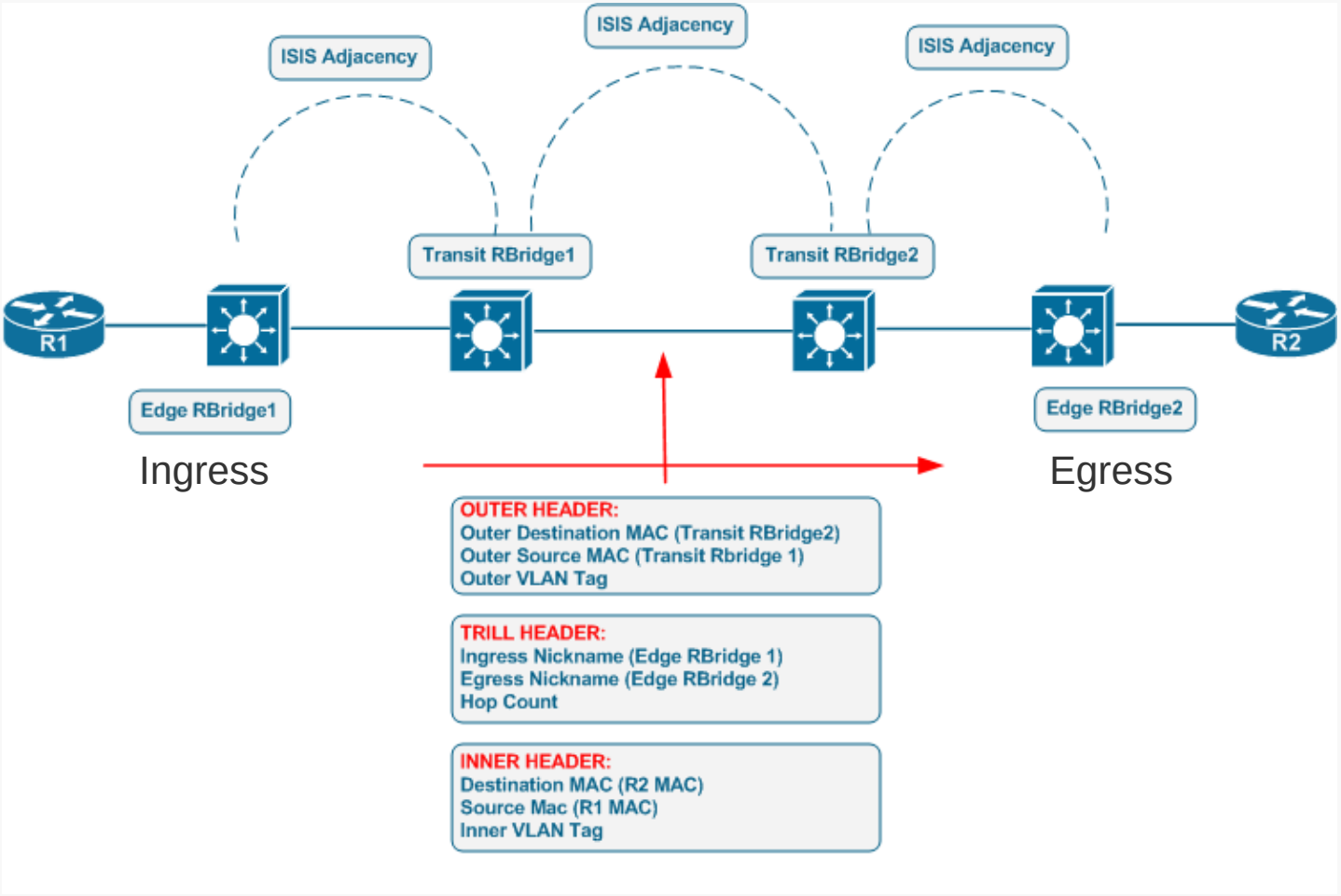
SPB



TRILL

- Transparent interconnection of lots of links
  - RBridge: routing bridge
    - *multipath* (ECMP) tunnels over L2 domain
  - with link state protocol: IS-IS
    - for same reasons as in SPB
  - extra headers
    - TRILL header
      - hop count
      - RBridge ingress, egress nickname
    - outer Ethernet header
      - RBridge source, destination MAC
      - VLAN tag
  - transit/relay RBridges swap the outer Ethernet header to the next hop RBridge MAC address
    - standard Ethernet switches forward traffic by outer MAC address
- vendors: Cisco, Brocade

TRILL


ISIS Adjacency
ISIS Adjacency
ISIS Adjacency
Transit RBridge1
Transit RBridge2
R1
R2
Edge RBridge1
Edge RBridge2
Ingress
Egress
OUTER HEADER:
Outer Destination MAC (Transit RBridge2)
Outer Source MAC (Transit Rbridge 1)
Outer VLAN Tag
TRILL HEADER:
Ingress Nickname (Edge RBridge 1)
Egress Nickname (Edge RBridge 2)
Hop Count
INNER HEADER:
Destination MAC (R2 MAC)
Source Mac (R1 MAC)
Inner VLAN Tag

SPB vs. TRILL

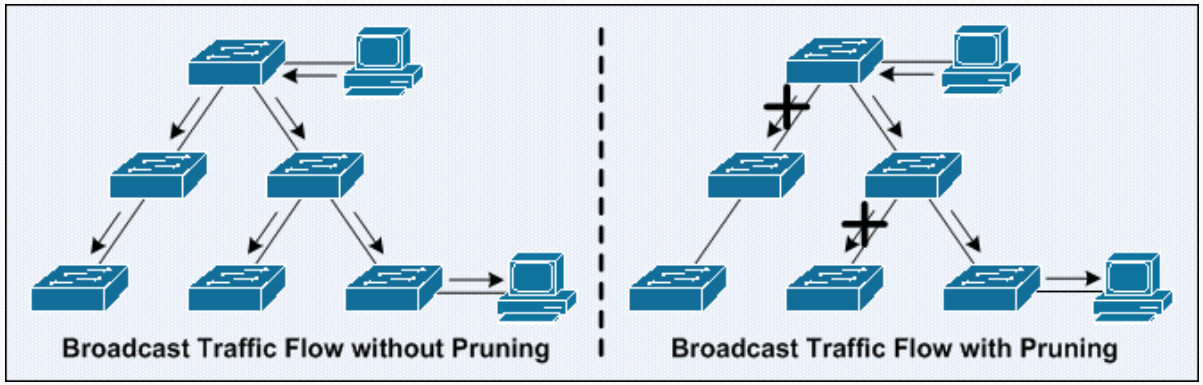
	SPB	TRILL
Standardization	IEEE	IETF
Data forwarding	Ethernet switching Without MAC address swapping	Forwarding by RBridge nicknames MAC address swapping hop-by-hop
Virtual networks	SPBM: 16 million	4096
Hardware	Existing, low cost Ethernet ASIC	New hardware
Loop protection	Reverse Path Forward Checking (RPFC)	RPFC + hop count
ECMP	Yes, 16 way	Yes, 16 way

Reverse Path Forward Checking

- checking whether a source addresses can be reached via the input interface (is there an entry in the forwarding table in the opposite direction), i.e. it arrived on the shortest path
  - if yes: forward
  - if not: drop
- conditions
  - correct forwarding information in a converged state
  - symmetric forward-backward paths
- unicast and multicast

Network virtualization technologies

- STP problems: routing (e.g. IS-IS) with MAC addresses
  - Shortest Path Bridging MAC (SPBM)
- limited number of VLANs: add another VLAN tag
  - Q-in-Q, provider bridging, (IEEE 802.1ad)
- MAC address limit: add another MAC address header
  - Provider Backbone Bridges (PBB), 802.1ah
  - Transparent Interconnection of Lots of Links (TRILL)
    - bridging + routing
- to avoid hypervisor flooding: consider VMs
  - VLAN pruning: elimination of unnecessary traffic
- to avoid flooding the core network: VLAN pruning in the core network



Network virtualization

- VN-Tag identifies the VM, but not the tenant
- Support for tenant separation
  - Virtual Extensible LAN (VXLAN) – RFC 7348
    - Cisco, VMware
    - transport of virtual L2 traffic over physical L3 network
  - Network Virtualization using Generic Routing Encapsulation (NVGRE)
    - Microsoft, Intel, HP, Dell
  - Generic Network Virtualization Encapsulation (GENEVE)
    - superset of VXLAN and NVGRE
  - Stateless Transport Tunneling (STT)
    - Nicira ⇨ VMware

Sources

- Pat Thaler et al., IEEE 802 Tutorial: Edge Virtual Bridging, 2009.
- Overlay Virtual Networking Explained, Ivan Pepelnjak, NIL Data Communications, 2011.
- Shortest Path Bridging, IEEE 802.1aq, Tutorial and Demo, NANOG 50 Oct 2010, Peter Ashwood-Smith, Huawei
- Radhika Niranjan Mysore, Andreas Pamboris, Nathan Farrington, Nelson Huang, Pardis Miri, Sivasankar Radhakrishnan, Vikram Subramanya, and Amin Vahdat. PortLand: a scalable fault-tolerant layer 2 data center network fabric. *SIGCOMM Comput. Commun. Rev.* 39, 4 (August 2009)
- Ankit Singla, Chi-Yao Hong, Lucian Popa, and P. Brighten Godfrey. 2012. Jellyfish: networking data centers randomly. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation* (NSDI'12). USENIX Association, Berkeley, CA, USA.