



Traffic Engineering in Ethernet-based Access Networks



István Moldován



MUSE Autumn School 2006
(October 19-20, Bilbao)

- > Good old, proven technology
 - Standardized long time ago, continuously developed
 - Mass production made it cheap
 - Available at high speeds
- > Widely used in Enterprise/Campus
 - Local area network solution
 - With the apparition of optical transmission standards grove out of campus
- > In the provider network, used as
 - Point-to-point between router ports
 - In the aggregation network

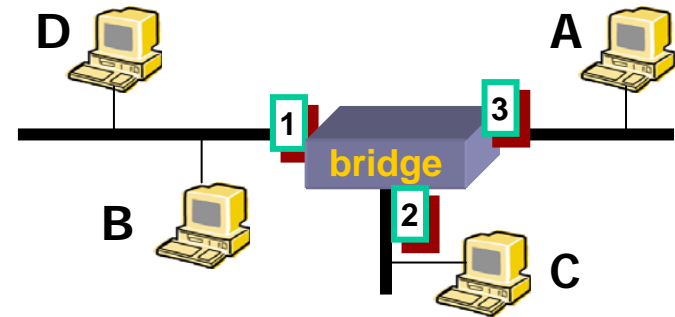
- > More than 90% of Internet traffic starts from Ethernet
 - Ethernet services appear
- > Takes the advantage that Ethernet is cheap
 - High bandwidth for low cost
- > However requires enhancements for:
 - QoS
 - Scalability
 - Reliability, better resilience
 - Traffic separation
 - Operation, Administration & Management

- > Bridge operation
 - MAC learning
 - Forwarding
- > Virtual LANs
 - VLANs
- > Spanning Tree Protocols
 - STP
 - RSTP
 - MSTP
- > Traffic Engineering and protection based on MSTP

Ethernet Bridge Operation

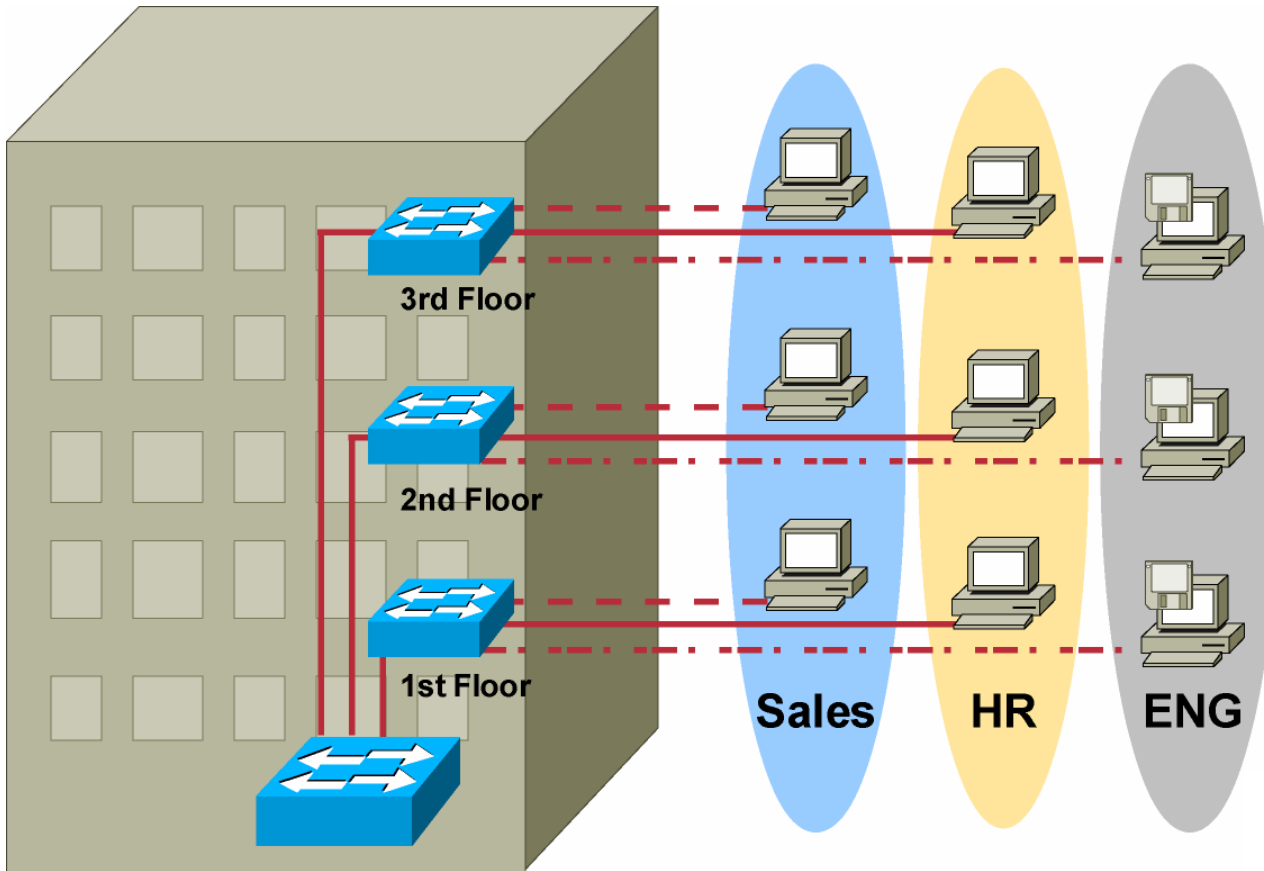
- > Frame forwarding based on destination MAC address
 - MAC addresses supposed to be unique
- > If destination not known: flooding
 - and learn the source MAC
- > If destination MAC is already learned, forward only to that port

- > Example:
 - A->D: broadcast
 - D->A: port 3
 - learn D's MAC
 - C->D: port 1



| MAC addr. | Port |
|-----------|------|
| A | 3 |
| B | 1 |
| C | 2 |
| | |

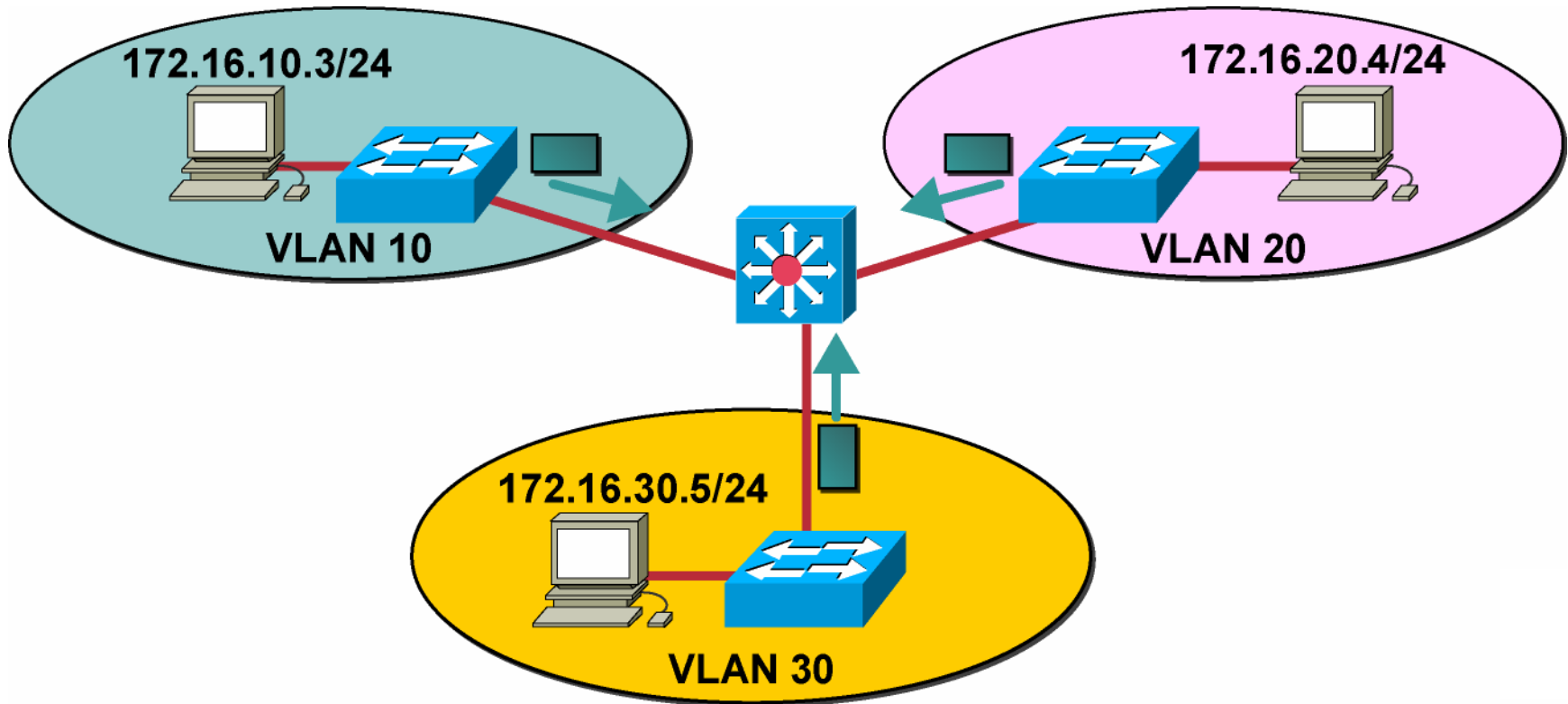
VLAN Overview



- Layer 2 connectivity
- Logical organizational flexibility
- Single broadcast domain
- Management
- Basic security

A VLAN = A Broadcast Domain = Logical Network (Subnet)

Routing Between VLANs



Problem: Isolated Broadcast Domains

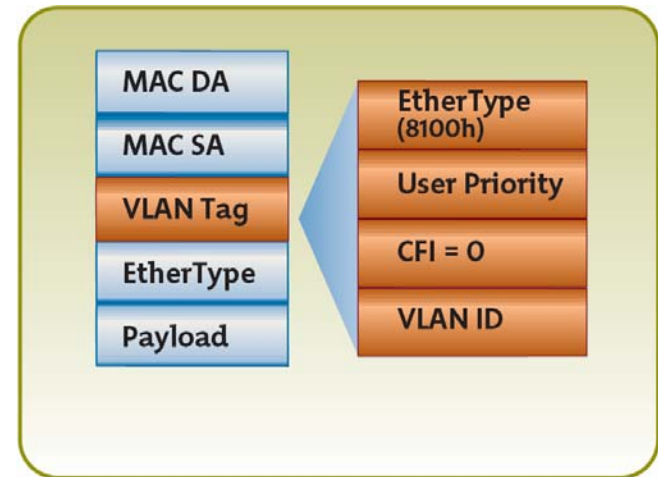
- Because of their nature, VLANs inhibit communication between VLANs.
- Communications between VLANs requires a Layer 3 services module.

VLAN standard: IEEE 802.1Q

- > IEEE 802.1Q adds a VLAN tag to frames
 - allows to extend VLANs over multiple bridges
 - VID = 12 bit = 4096 VLANs possible
 - adds also 3 priority bits (p-bits)

- > Tagging usually is based on port
 - the port “tags” the packet to a predefined value
 - the uplinks carrying multiple VLANs are *trunk ports*

- > VLAN bridge operation
 - ingress filtering
 - switching
 - egress filtering



Spanning Tree (802.1D, 802.1w, 802.1s)



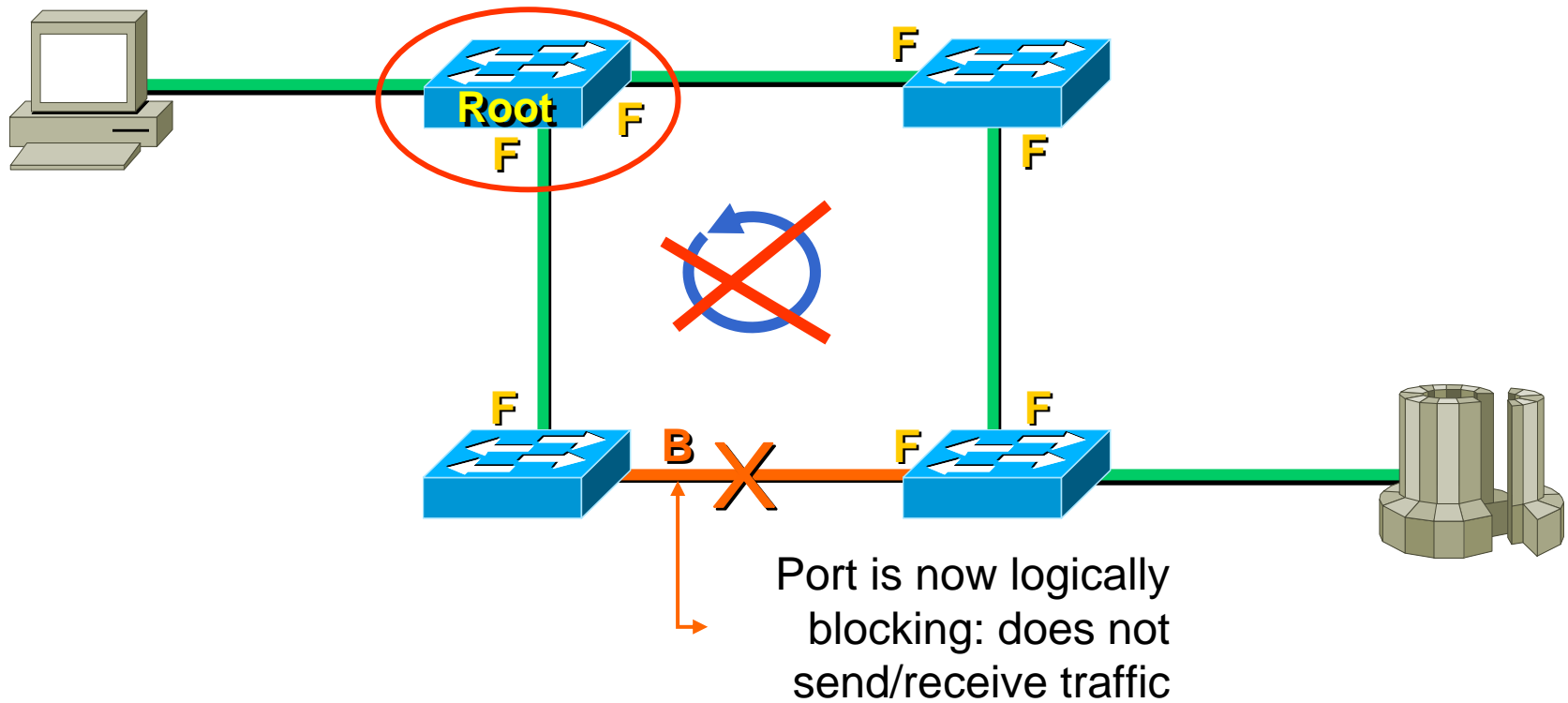
- > Purpose:
 - to maintain loop-free topologies in redundant/resilient Layer 2 infrastructures
- > Provides automatic path recovery services upon link or device failure
- > By default, 802.1D (STP) implements
 - pessimistic & safe behavior => slow convergence, typically 30 to 50 seconds, timer-based operations
- > Many vendor specific enhancements
 - available for scalability, safety and convergence speed
- > IEEE Standards 802.1w (RSTP), 802.1s (MSTP) include a superset of these enhancements



Spanning Tree Basics

Loop-free
Connectivity

Root is elected based on lowest bridge ID
(priority and MAC address concatenated)



IEEE 802.1D STP operation



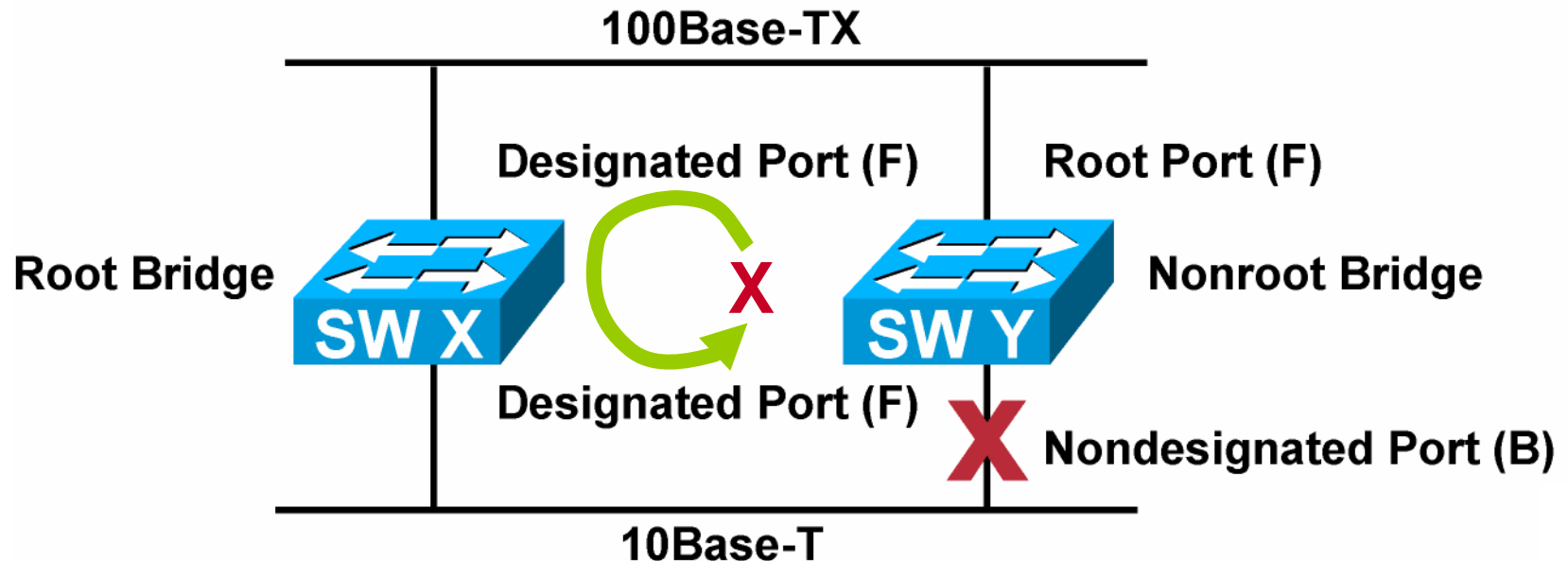
- > Bridges first elect a *root bridge*
 - based on lowest bridge ID (priority and MAC address concatenated)
- > Starting from the root they build the tree
 - each port has a *cost*
 - the tree is spanned using the minimum cost paths
 - each port not used by the tree will be *blocked*
 - 15 seconds to build the tree
- > After spanning the tree starts the MAC address *learning*
 - 15 seconds again
- > The bridges start *forwarding*



Spanning Tree Basics

- One root bridge per network
- One root port per non-root bridge
- One designated port per segment
- Non-designated ports are blocked

Loop-free
Connectivity

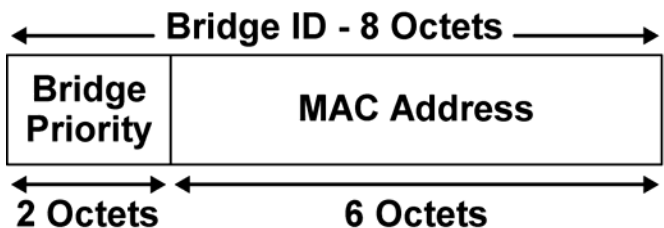


Bridge Protocol Data Unit (BPDU)

| Bytes | Field |
|-------|---------------|
| 2 | Protocol ID |
| 1 | Version |
| 1 | Message Type |
| 1 | Flags |
| 8 | Root ID |
| 4 | Cost of Path |
| 8 | Bridge ID |
| 2 | Port ID |
| 2 | Message Age |
| 2 | Maximum Time |
| 2 | Hello Time |
| 2 | Forward Delay |

The BPDU used by STP is responsible for:

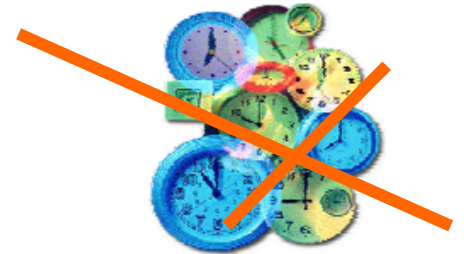
- Electing a root bridge
- Determining the location of loops
- Blocking to prevent loops
- Notifying the network of changes
- Monitoring the state of the spanning tree



| Link Speed | Cost (Revised IEEE Spec) | Cost (Previous IEEE Spec) |
|------------|--------------------------|---------------------------|
| 10 Gbps | 2 | 1 |
| 1 Gbps | 4 | 1 |
| 100 Mbps | 19 | 10 |
| 10 Mbps | 100 | 100 |

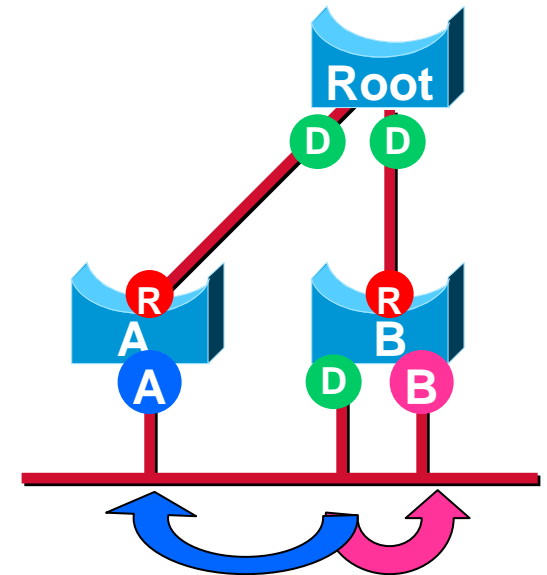
802.1w - Rapid STP Basics

- > New IEEE protocol (802.1w)
Compatible with 802.1D
- > Faster Convergence & Timer independent
 - IF only point-to-point FDX Links are used
 - IF all edge ports are correctly identified
 - IF no 802.1D interaction required
- > Includes functionality equivalent to vendor's proprietary solutions
- > What is new
 - New Port Role
 - Modified BPDU
 - BPDU handling
 - Rapid port state transition
 - New topology change mechanism
 - 802.1D Compatibility

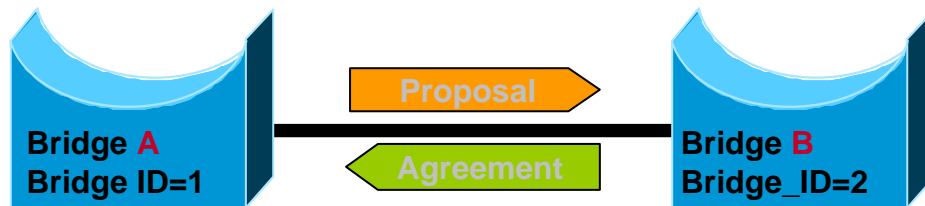


RSTP Port Roles

- R** Root Port (Fwd):
Port receiving the best BPDUs for the bridge
– shortest path to the Root in terms of path cost
- D** Designated Port (Fwd):
Port sending the best BPDUs on a segment
- A** Alternate Port (Disc):
Port blocked by BPDUs from a different bridge
– redundant path to the Root
- B** Backup Port (Disc):
Port blocked by BPDUs sent from the same bridge
– redundant path to a segment



Rapid Transition to Forwarding

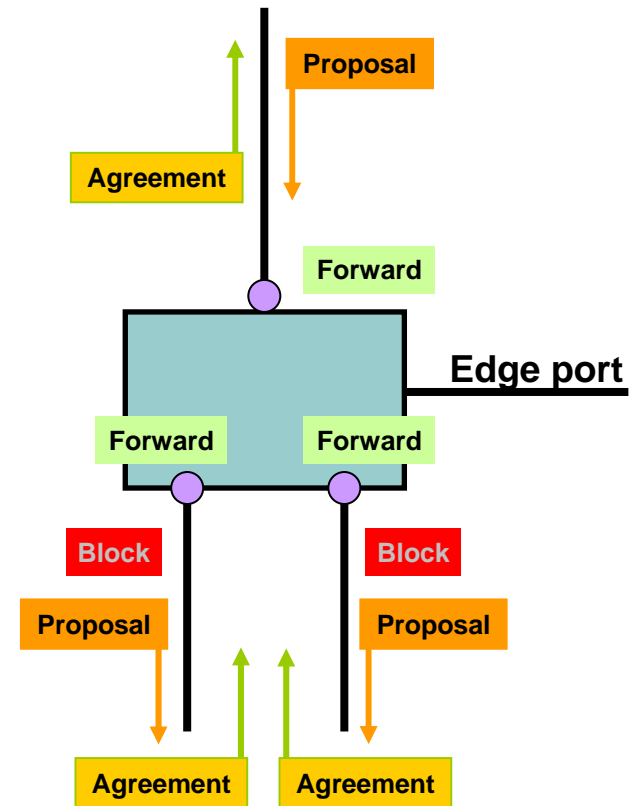


- > A has a better id than B
- > A sends a proposal to B to become designated
- > B compares the received priority and replies with an agreement
- > B's port becomes Root port -> forwarding
- > A's port becomes Designated -> forwarding

- > When port comes up,
 - bridge sends BPDU with **proposal** flag set to become designated for that segment
- > Response is
 - BPDU with **agreement** flag set if remote bridge selects the port on which it received the proposal as its root port
- > As soon as agreement is received, port moves to forwarding

IEEE 802.1w sequence of events

- > Receive a proposal
 - Block all other non-edge ports
- > Send an agreement back
 - Put the new root port to forwarding
- > Send out proposals on other ports
- > Receive agreement from others
 - Put ports into forwarding



Evolution to multiple trees & regions - MSTP



> Why regions?

- Different administrative control over different parts of the L2 network
- Not all switches in the network might run/support MST - different kinds of STP divide network into STP regions
- All benefits of MST are available INSIDE the region, outside it is single instance (topology) for all vlans

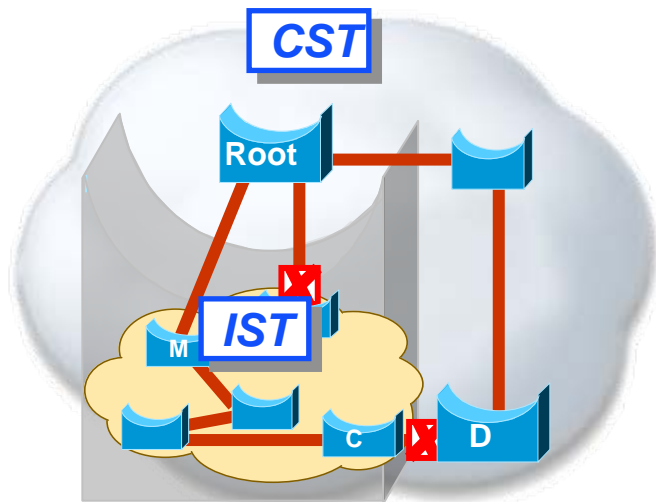
> MST region is a linked group of MST switches with same MST configuration

- Inside region: many instances
 - IST – Internal Spanning Tree (instance 0), always exists on ALL ports
 - MSTI - Multiple Spanning Tree Instance
- Outside of region: one instance

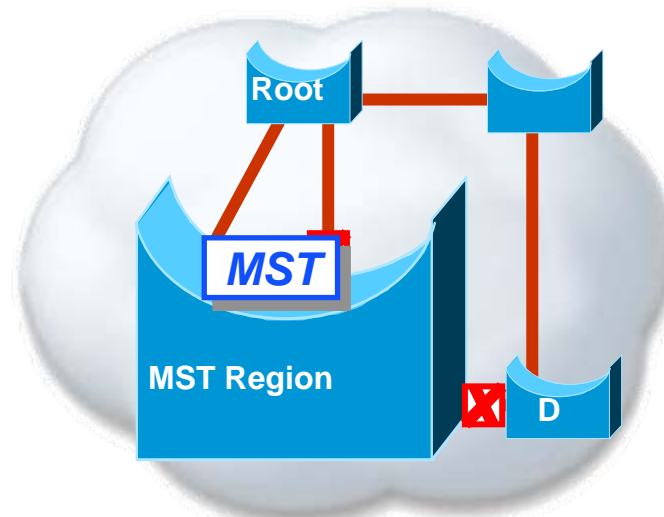


802.1s: CST, IST, MST - Lots of Trees ...

Inside View



World View

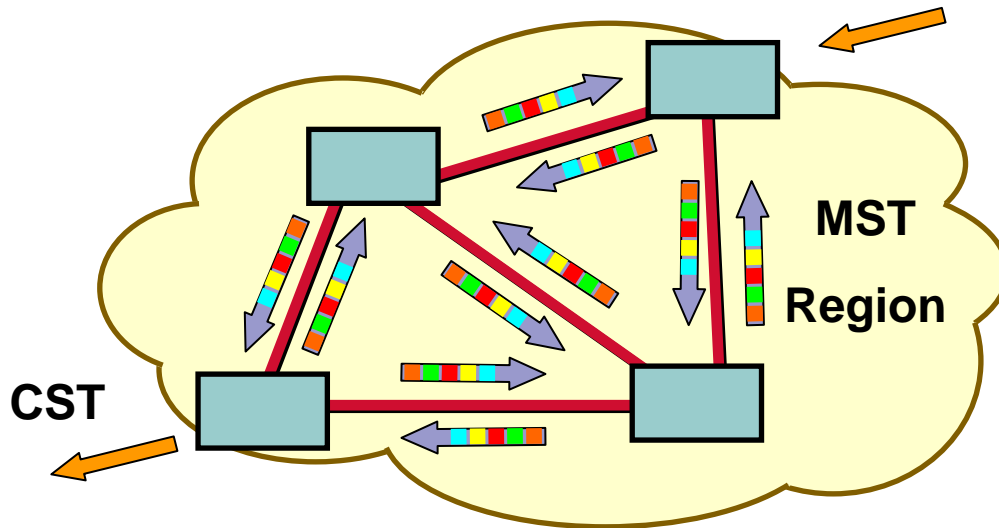


- **CST 802.1Q Common SPT => Single Instance only**
- **IST 802.1s Internal SPT => receives and sends BPDUs to the CST represents the MST to the Outside World as CST Bridge**
- **MST 802.1s Multiple SPT => represent several VLANs mapped to a single MST Instance**



MST instances

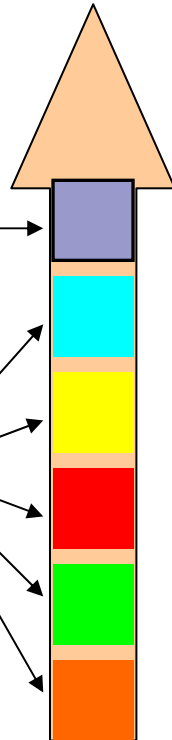
- MSTIs are STP instances existing **only inside** the region
- MSTIs do not interact directly with the outside of the region
- MST only send **1 BPDU** for all the instances with one M-record per instance
- Only one instance has timer related parameters (the IST instance)
- MST BPDUs are **sent on every port**
- **BPDUs are sent both ways** on a link as opposed to 802.1D where BPDUs are only sent by designated bridge



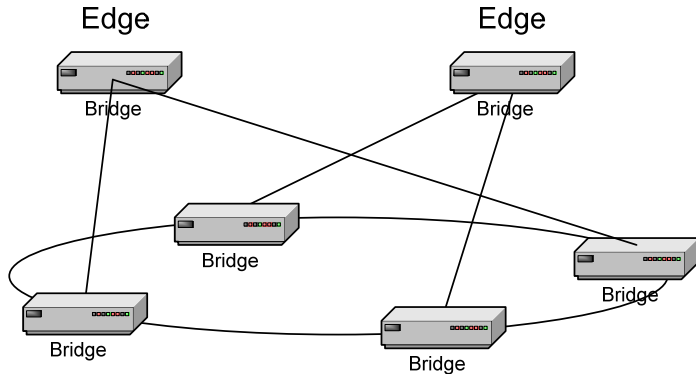
Protocol information for the IST

Protocol information for the MSTIs

MST BPDU

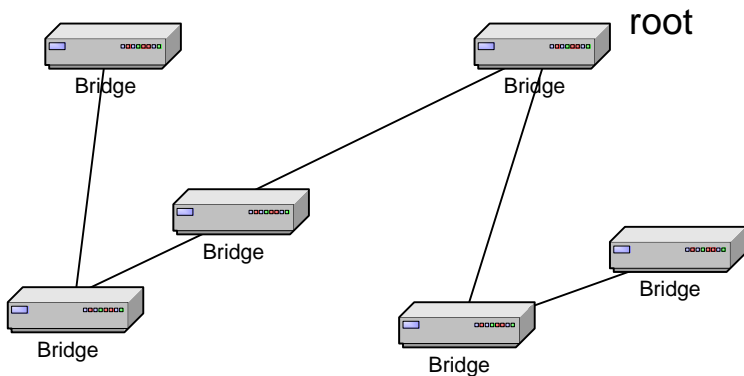


MSTP advantages



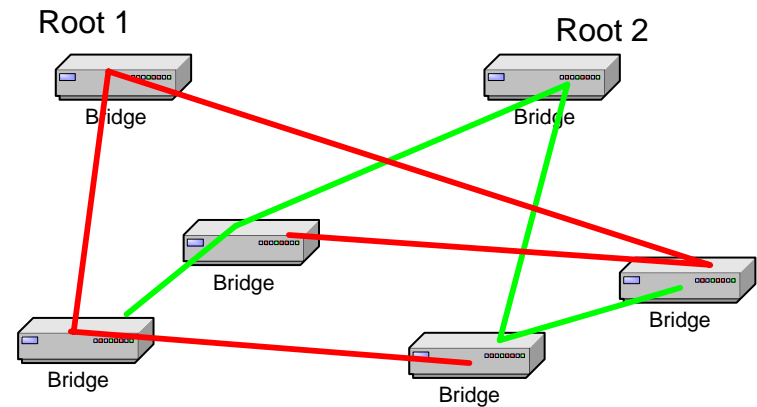
- > Network topology with 2 Edges
- > The ring provides redundancy

> Spanning Tree Protocol



> MSTP Protocol

- 2 trees



With MSTP we can do



> Traffic Engineering

- Load balancing
- paths can be “engineered”
- traffic mapping to different engineered paths

> Protection

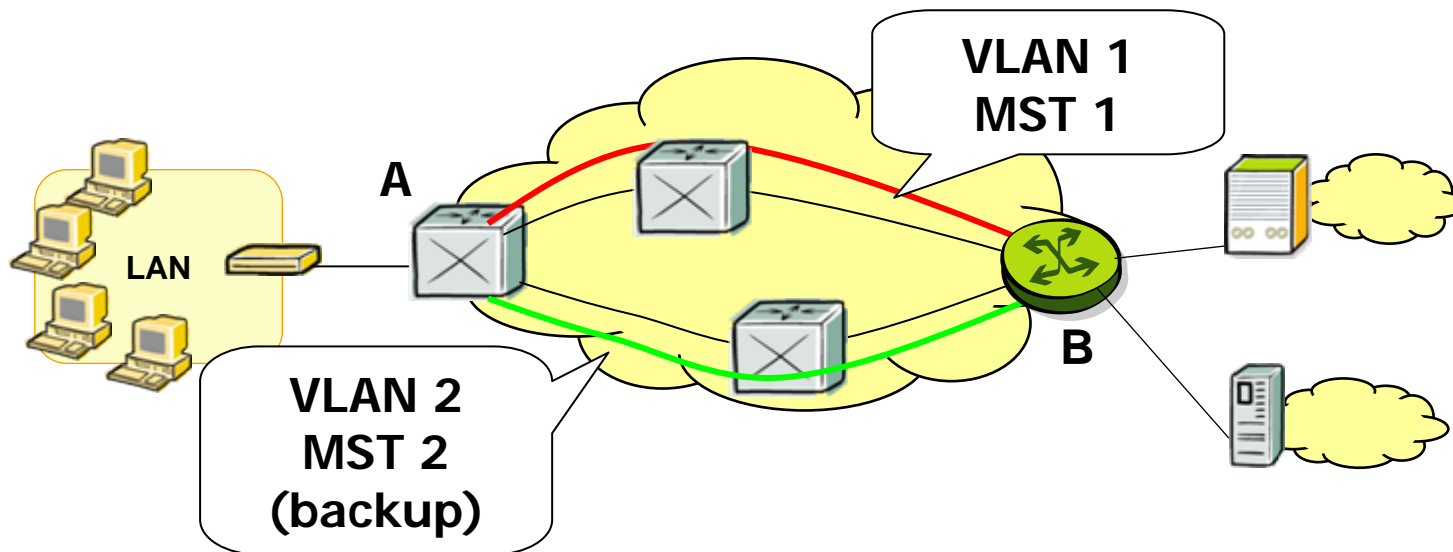
- Multiple disjoint trees
- VLAN 1 assigned to primary tree, VLAN 2 to backup tree
- On failure, traffic is switched to VLAN 2, using the backup tree
- (requires IP level switching/failover logic)

> Of course, for a simple tree physical topology it is useless



Protection switching

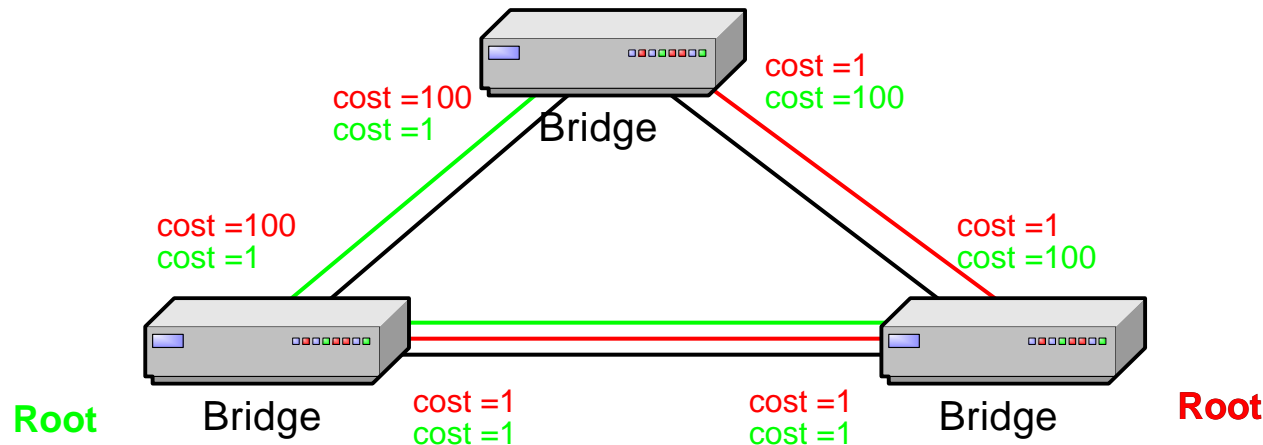
- > 802.3ad Link Aggregation
 - uses redundant links for load balancing and protection
- > Using MSTP
 - 2 MSTI trees, two paths: red and green
 - VLAN 1 -> MST 1, VLAN 2 -> MST 2
 - **A** and **B** uses VLAN 1, in case of failure switch to VLAN 2



- > Usually done at Layer 3
 - We need to do at Layer 2!
- > Traffic Engineering using MSTP:
 - Set up multiple trees
 - Multiple trees use different cost set
 - results different trees
 - Different VLANs assigned to different trees -> different paths for VLANs
 - load sharing by VLAN assignment
- > Enables:
 - multiple paths between two nodes
 - shorter paths between nodes

MSTP optimization

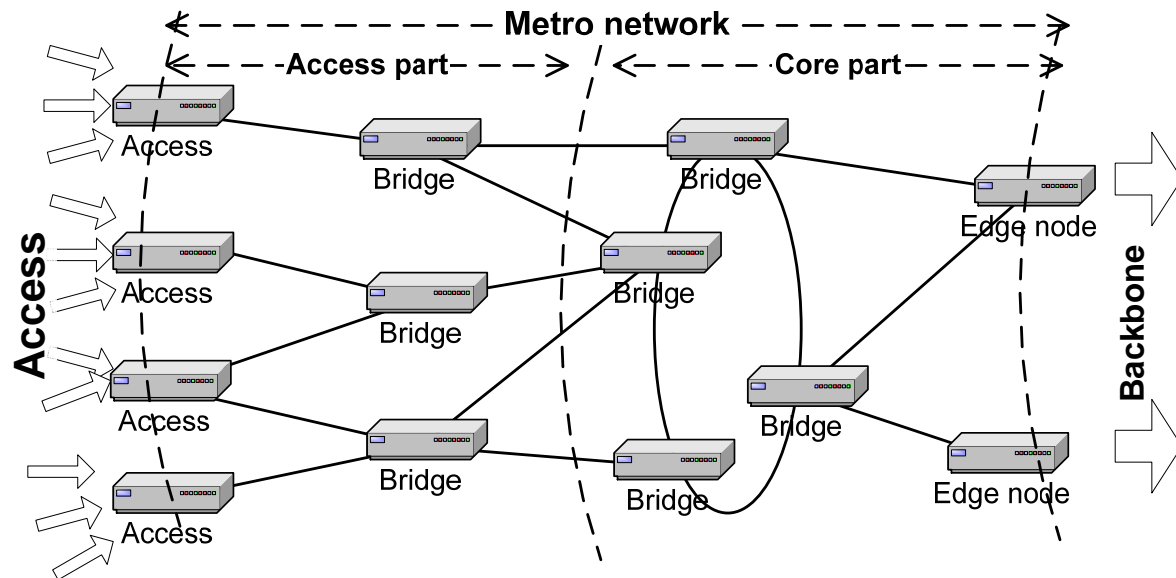
- > MSTP requires configuration
- > Trees are set up by setting different port costs



- > Port cost assignment:
 - 1 for forwarding, #of bridges+1 for blocking

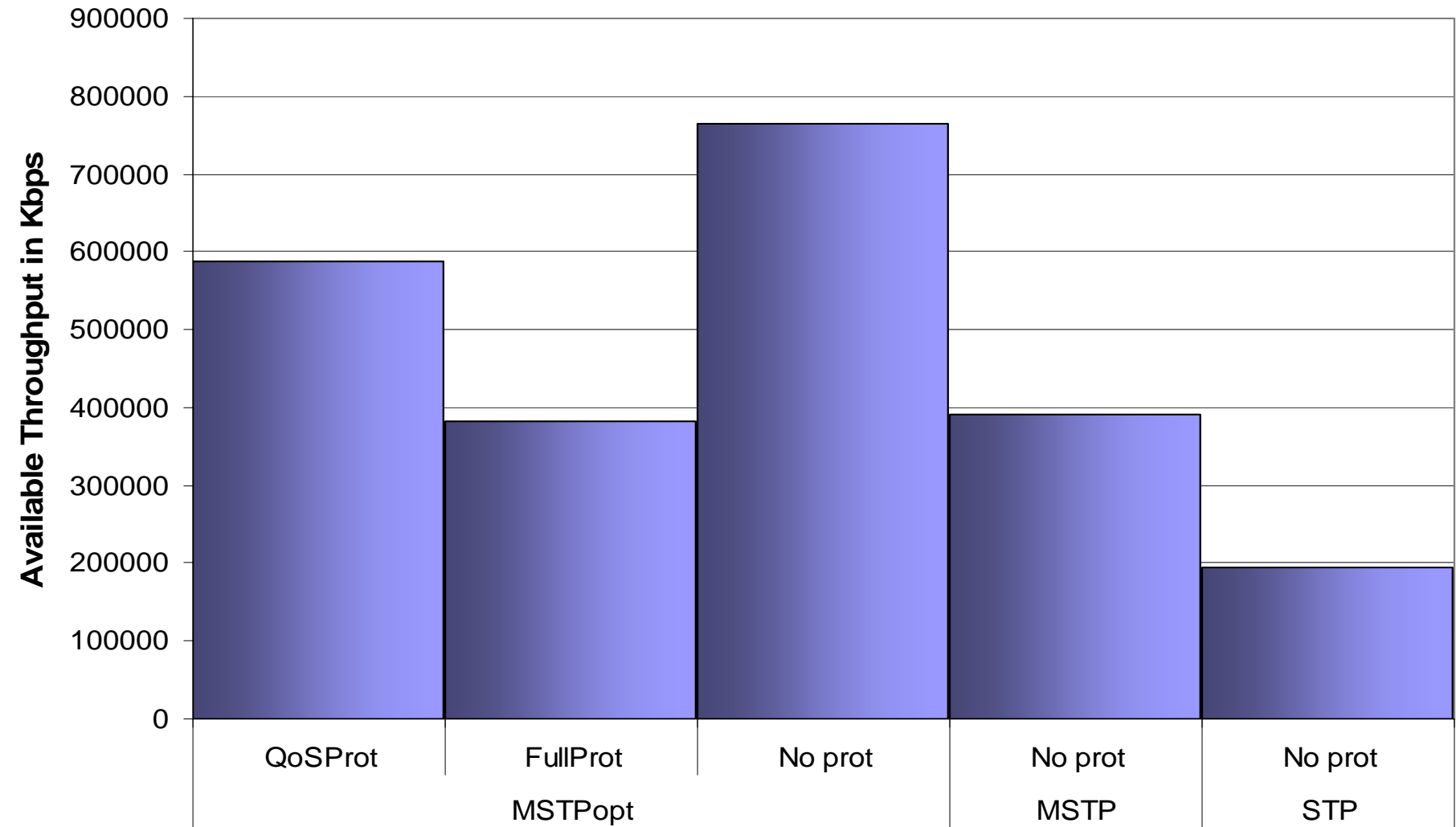
An example for Traffic Engineering

- > We target **OPTIMAL UTILIZATION** of the network
 - utilize alternate paths
 - take into consideration traffic parameters too
 - keep QoS guarantees
- > Multiple trees rooted in Edge Nodes
- > Optimization
 - offline
- > Set up port costs for all bridges to get the desired trees

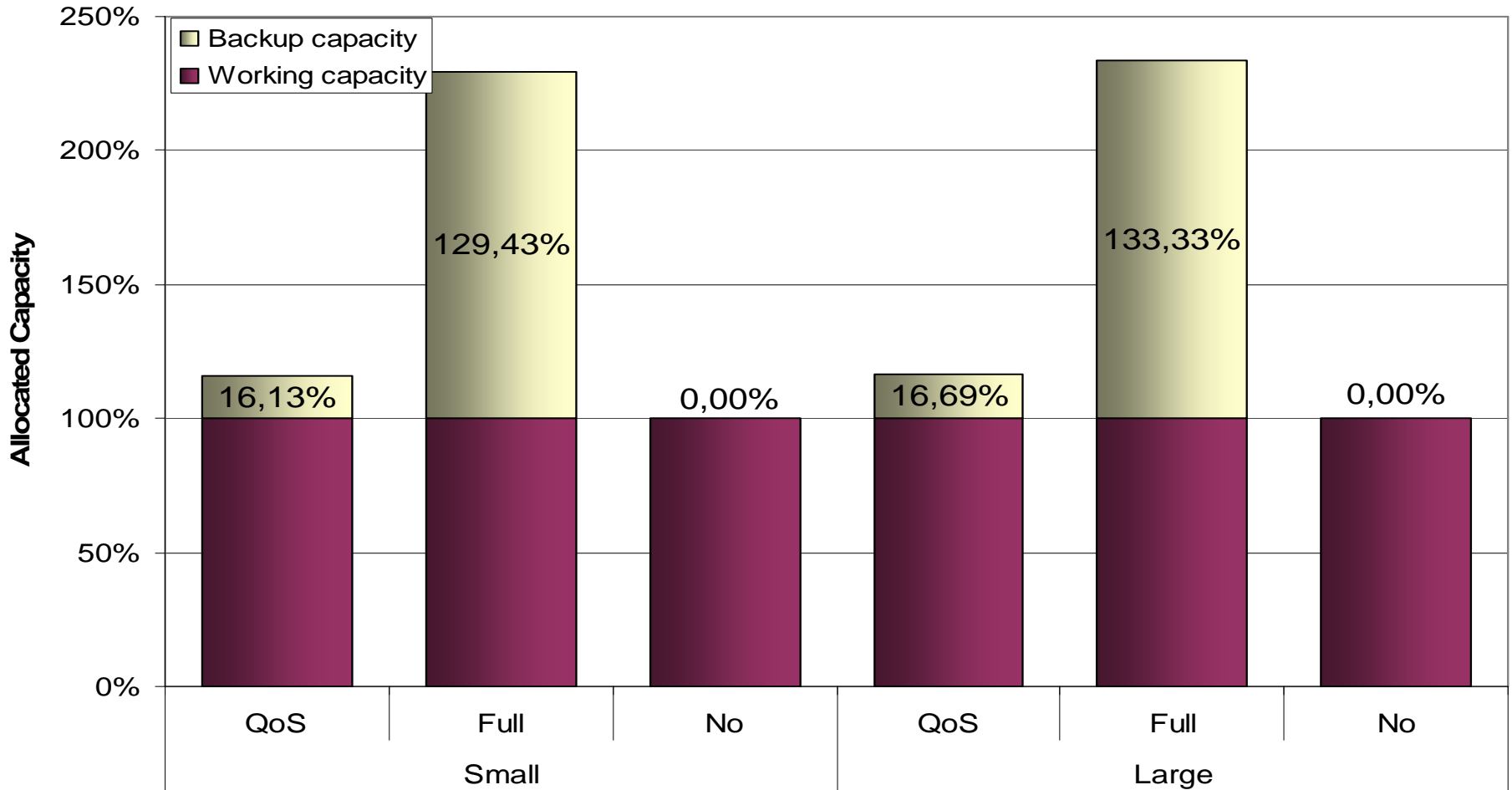


- > Optimize the spanning trees: the goal is to minimize the used network resources to maximize network throughput
- > Provide 1:1 protection
 - To protect all traffic is expensive - protect only a part of the traffic: the prioritized traffic
 - Best Effort can use the protection paths
- > Optimization algorithms developed within MUSE
 - Integer Linear Programming formulation is given
 - the solution is optimal spanning trees
 - Slow for practical sized networks
 - Heuristic solution
 - near optimal, but much faster

Achievable throughput



Capacities allocated



- > Ethernet in the MAN—„carrier grade” requirements
- > STP protocol has disadvantages: single tree inhibits load sharing
 - MSTP provides means for TE
 - Enhances scalability
- > We present an optimization TE Framework for QoS and protection
- > We show that:
 - With optimization we can use redundant links
 - Throughput doubles compared to STP
 - Optimized 1:1 protection
 - The same throughput as STP but all protected
 - Protecting QoS traffic only is reasonable tradeoff

Thank you for your attention!



Questions?



Publications on Ethernet TE Optimization



11th European Conference on Networks & Optical Communications (MSAN2005)
" Optimizing QoS Aware Ethernet Spanning Trees" ,

First International Conference on Multimedia Services Access Networks (DRCN2005)
" Optimized QoS Protection of Ethernet Trees" ,

Conference on Next Generation Internet Design and Engineering (EuroNGI2006)
" On the Optimal Configuration of Metro Ethernet for Triple Play"

IEEE Symp. on Computers and Communications (ISCC2006)
" Scalable Tree Optimization for QoS Ethernet"

World Telecommunications Congress (WTC2006)
" Traffic-driven Optimization of Routing for Metropolitan Ethernet Networks"

