Evolution of the Spanning Tree Protocol

Abstract

The Spanning Tree Protocol that is the basis for the IEEE standard 802.1D was designed to provide "plug-and-play" operation for large Layer 2 networks based on half duplex shared Ethernet, which was the prevalent LAN technology throughout the late 1980s and early 1990s. As Ethernet evolved to become a switched full duplex technology, it soon became evident that 802.1D needed to be upgraded in order to keep pace with the new design models and switch features (e.g., VLANs) that emerged to allow optimization of switched networks.

The needed enhancements have now been standardized as IEEE 802.1w Rapid Spanning Tree Protocol (RSTP) and IEEE 802.1s Multiple Spanning Tree Protocol (MSTP). RSTP provides the rapid convergence needed to optimize high availability and network resiliency, while MSTP provides the VLAN-awareness and VLAN-scalability required for standards-based traffic isolation and load-sharing over redundant links. Both protocols build on the foundation of 802.1D, and retain sufficient backward compatibility to allow interoperability with switches and bridges that support the older protocol.

The enhancements described in the body of the document are significant because they allow both enterprises and service providers to design switched Ethernet access networks capable of supporting current and future generations of mission critical applications and network services.

Introduction

In the early days of Ethernet, the extent of the Ethernet Layer 2 network was determined by the maximum end-to-end delay that would allow collision detection in the shared segment. Maximum delay was specified in terms of the maximum number of repeaters and other delay-producing elements in the segment.

The size limitations of Ethernet were relaxed significantly when the Digital Equipment Corporation (DEC) introduced the first 2-port Ethernet bridge in the mid-1980s. With the Ethernet bridge, arbitrarily large Layer 2 networks could be constructed by concatenating multiple collision domains into a large broadcast domain. At the time, the architectural limit on the size of the broadcast domain was chosen (rather arbitrarily) as seven bridge hops.

The DEC LANbridge 100 was a store-and-forward, "transparent bridging" device that used the DEC Spanning Tree Protocol (STP) to prevent traffic from being forwarded in loops that would be created by redundant paths in the bridged network. STP is a distributed algorithm that requires only minimal bridge configuration in order to allow bridges to automatically adjust the active topology of the Layer 2 network to accommodate bridge and link failures or the installation of new bridges in the network. The ability of STP to support auto-configuration and re-configuration after changes in physical topology allowed "plug and play" operation of Layer 2 networks that was reminiscent of a home stereo system.

The DEC proprietary version of STP became the basis for the IEEE standard 802.1D, but the two versions of the protocol were different enough to prevent interoperability.

As Ethernet has evolved into a full duplex switched network interconnected with Layer 3 switches, networks are generally designed to conform to structured models based on replicated modules optimized for the wiring closet, backbone, and data center. As these changes have occurred, 802.1D has been enhanced to keep pace with the requirements of the modern Ethernet LAN. The primary enhancements are 802.1w Rapid Spanning Tree Protocol (RSTP) and 802.1s Multiple Spanning Tree Protocol (MSTP). These protocols provide the rapid convergence and VLAN-awareness required for today's switched Ethernet networks, while building on the foundation of, and retaining backward compatibility with, 802.1D.

Overview/Review of STP

The IEEE 802.1D Spanning Tree Protocol (STP) eliminates loops in the logical topology by allowing only a single active path between any two stations on the Layer 2 network. This is accomplished by placing all redundant paths in a standby, non-forwarding state (i.e., blocked). In the event that a bridge or link in the active path fails, STP provides the mechanism for auto fail-over to one of the redundant paths, enhancing the fault tolerance of the network. The combined benefits of "plug and play" operation and fault tolerance have made STP a basic requirement for any network device that supports Layer 2 switching/bridging.

STP eliminates loops by controlling the state of network links in accordance with the spanning tree algorithm (STA). STA is a distributed algorithm based on exchange of control messages among bridge interfaces or switch ports using special Ethernet packets called Bridge Protocol Data Units (BPDUs). Based on the information in the BPDUs, the bridges elect a Root Bridge. The bridge with the highest priority Bridge Identifier is selected as the Root and then each bridge in the contiguous Layer 2 network independently computes the spanning tree stemming from the Root. The spanning tree defines a logical topology of the network that is loop-free and provides full inter-bridge connectivity as shown in Figure 1.

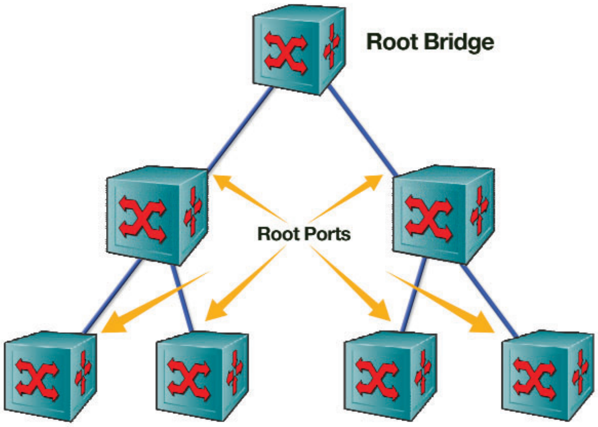


Figure 1. Spanning Tree

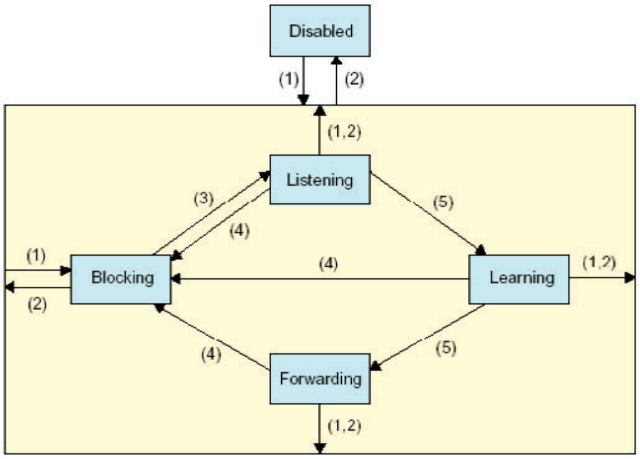
To calculate the spanning tree, each bridge/switch in the network determines which of its interfaces has the lowest path cost to the Root Bridge, and designates that interface as its Root Port. A bridge/switch can only have one Root Port active at any one time. The Root Path Cost to the Root Bridge is calculated summing the Path Costs assigned (possibly by default) to every port on the least cost path. Assignments are generally made as a function of the port bandwidth (e.g., a 10 GbE port might be assigned a cost of 2, while a 10 Mbps Ethernet port might be assigned a cost of 100).

Each individual LAN has a Bridge Port connected to it that forwards frames from that LAN towards the Root, and forwards frames from the direction of the Root onto that LAN. This port is known as the Designated Port for that LAN, and the bridge which it is part of is the Designated Bridge for that LAN. The Designated Port is chosen as the one that has the lowest cost path to the Root for that segment.

If an active port is neither a Root Port nor a Designated Port, then it is placed in a blocked state that does not forward traffic and the spanning tree is complete. From this description it is obvious that the Root Bridge is the Designated Bridge for all the LAN segments to which it is connected and that the only ports on each bridge that are in a forwarding state are the Root Port and the Designated Ports.

If a link or interface fails, the process of computing the spanning tree begins anew. Convergence occurs when enough time has elapsed so that it may be safely assumed that all switches and bridges have independently computed the same stable spanning tree and traffic forwarding can begin. The 802.1D version of STP has a convergence time = ((2 x Forward Delay) + Max_Age) or ((2 * 15) + 20) = 50 seconds based on default timer values. Until the network has converged, all interfaces remain in a blocking state and the contiguous Layer 2 network as a whole is temporarily inoperable.

Figure 2 shows the possible states in which a bridge/switch port may be placed. The numbered arrows indicate the state transitions during normal operation of STA.



1. Port enabled, by management or initialization
2. Port disabled, by management or failure
3. Algorithm selects as Designated or Root Port
4. Algorithm selects as Alternate Port
5. Protocol timer expiry (Forwarding Timer)

Figure 2. State diagram for STP

The purpose of each of these states is summarized as follows:

- ***Blocking:*** A blocking port does not forward traffic, discards any received frames, and does not add new station addresses to the filtering database. It receives and processes BPDUs
- ***Listening:*** A listening port is preparing to participate in frame forwarding. This state is entered from the Blocking State when the STA determines that the port should participate in forwarding
- ***Learning:*** A learning port is still blocked, but it is in the process of updating its filtering database by learning new station addresses. This state is entered from the Listening State after expiration of a forward delay timer
- ***Forwarding:*** This is the normal state of an active port, forwarding traffic and learning new station addresses. This state is entered from the Learning State after expiration of a forward delay timer
- ***Disabled:*** A port in this state does not participate in frame relay or the operation of the Spanning Tree Algorithm and Protocol. This state is entered from any other state by the operation of management

The 802.1D Configuration BPDU includes the following information fields:

Root Identifier: The unique Bridge Identifier of the Bridge assumed to be the Root by the transmitting Bridge

Root Path Cost: The cost of the path to the Root Bridge denoted by the Root Identifier from the transmitting Bridge

Bridge Identifier: The unique Bridge Identifier of the Bridge transmitting the Configuration BPDU

Port Identifier: The Port Identifier of the Port on the transmitting Bridge through which the Configuration BPDU was transmitted

Message Age: The age of the Configuration Message, being the time since the generation of the Configuration BPDU by the Root that instigated the generation of this Configuration BPDU

Max Age: A timeout value set by the Root used to limit the time period for which the last configuration BPDU message is considered valid. The default value of Max Age is 20 seconds

Hello Time: The time interval between generation of Configuration BPDUs by the Root. The default Value is 2 seconds

Forward Delay: The delay set by the Root between transitioning a port to the forwarding state and the beginning of the forwarding process. This parameter is also used as waiting time for aging out dynamic entries in the Filtering Database following changes in active topology. The default value of Forward Delay is 15 seconds

Topology Change: A flag set by the Root in all Configuration BPDUs transmitted for a period of time following the notification or detection of a topology change

Topology Change Acknowledgment: A flag set in a Configuration Message transmitted in response to Topology Change Notification received on a Designated Port

The Root Bridge sends out Configuration BPDUs every "Hello" time, downstream switches/bridges forward this BPDU over all their designated ports. In 802.1D a topology change occurs whenever a port transitions into or out of a blocking state. When this occurs, the bridge in question sends Topology Change Notification BPDUs toward the Root Bridge until an acknowledgement is received from the Root. After the Root receives the TCN, it forwards the notification to the rest of the bridges.

Shortcomings of IEEE 802.1D STP in Today's Networks

STP as standardized in 802.1D has provided critical functionality for Layer 2 bridged networks and continues to be useful to the present time. However, over the nearly two decades since the introduction of STP, Ethernet and the network design models have changed significantly:

- Switched full duplex Ethernet has gradually displaced shared Ethernet at both the core and the edge of the network. As this has occurred, structured network topologies based on replicated Layer 2 and Layer 3 modules have displaced more ad hoc approaches to network design. Enterprise and service provider Ethernet access networks are typically designed to conform to 3-tiered (access/distribution/core) or 2-tiered (access/core) design models that simplify management and limit the size of failure domains. Structured design principles typically employ a maximum of 2-3 bridge hops for traffic traversing Layer 2 modules.
- Virtual LANs (VLANs) in Layer 2 switched networks are frequently employed to further sub-divide broadcast/failure domains, to isolate traffic based on user group or application type, and to support load balancing across redundant connections.

As noted in the previous section of this document, the original versions of STP used very conservative timer values for Max Age and Forward Delay in order to ensure that a network of arbitrary physical topology has converged before ports are placed in the forwarding state.

The long convergence time represents an unnecessarily long waiting time in most structured networks and an excessive interruption of traffic for today's mission critical applications. Adjusting the protocol timers to shorter delays provides a partial remedy to the problem, but requires careful tuning to avoid introducing instability into the network. Tuning protocol timers is also rather inconsistent with the prevailing "plug-and-play" philosophy underlying Layer 2 networking. The long convergence times of 802.1D has led to a number of vendors to develop proprietary STP enhancements designed to reduce long convergence times in specific network design scenarios. The IEEE standard solution to the problem is 802.1w, also known as Rapid Spanning Tree (RSTP), which circumvents the use of protocol timers in normal operation and is applicable in arbitrary network topologies.

802.1D (in conjunction with 802.1Q) also places severe limitations on the diversity of VLANs that may be configured. In particular, 802.1D assumes that there should be only a single logical topology for the bridged network. This means that using STP in a network where VLANs span multiple switches using a trunking protocol such IEEE 802.1Q requires that all VLANs share the same logical topology. This reduces the degree of traffic isolation that VLANs can provide, and wastes bandwidth during broadcasts and flooding of packets. In addition, the single topology forces any redundant paths in the Layer 2 network to be in a blocked state for all traffic, a waste of bandwidth capacity that can be avoided if multiple logical topologies can co-exist on the bridged network, as will be shown toward the end of the document.

This lack of VLAN-awareness has led to the development of another set of proprietary enhancements to make STP VLAN-aware, such as Per-VLAN Spanning Tree (PVST), which creates a separate STP instance for each VLAN. With PVST, 1,000 VLANs means 1,000 instance of STP will be run, which places a significantly load on the switch CPU and wastes bandwidth with BDPUs sent for each VLAN.

The IEEE standard solution to the problem is 802.1s Multiple Spanning Tree Protocol (MST or MSTP) which builds upon 802.1w (RSTP) and 802.1Q to support multiple instances of RSTP, each of which may be shared by multiple VLANs. Shared instances of RSTP can greatly reduce the processing burden placed on the CPU of Layer 2 switches, while supporting large numbers of VLANs with disparate logical topologies.

Overview of RSTP

IEEE 802.1w Rapid Spanning Tree is an amendment to 802.1D that greatly improves convergence times and provides rapid recovery from switch, port, and link failures in switched Ethernet networks. With RSTP, convergence (or transition of the network to the forwarding state) normally occurs on a port-by-port basis, progressing down the spanning tree.

RSTP makes a number of changes to STP that minimize convergence times after changes in the physical topology or failures of network elements. This allows RSTP to provide very short fail-over times in high availability networks, especially in structured designs where the Layer 2 network module has a limited number of bridge hops.

BDPU Hellos as Keepalives: With RSTP, each bridge independently generates RSTP Configuration BPDUs every Hello Time. This is in contrast to STP where each bridge only relays a Hello generated by the Root. The locally generated BDPU serves as a keepalive that verifies the connectivity among neighboring switches/bridges. For example, when a bridge ceases to receive Hellos from a directly-attached neighbor, it can safely assume that connectivity has been lost on that port without waiting for protocol timers to expire. Connectivity loss is assumed when three consecutive Hellos have been missed. A switch can further accelerate the fail-over process by monitoring its interfaces to detect port and link failures without the need to wait for missing RSTP Hellos.

Port States/Roles: RSTP provides a clear differentiation between the state of a port (e.g., forwarding or blocking) and the role it plays in the operation of STP (e.g., Root Port or Designated Port). With STP, there are only three states a port can be in: discarding, learning, and forwarding. In addition to Root Ports and Designated Ports, RSTP defines two new port roles: the Alternate Port and the Backup Port. An Alternate Port is a normally discarding port that provides redundancy for a Root Port. The Backup Port is a normally discarding port that provides redundancy for a Designated Port connected to a shared LAN or where two ports are connected in loopback by a point-to-point link.

RSTP also embraces the concept of an Edge Port as a port that is directly connected to an end station or is connected to a shared LAN segment that does not include other bridges. If a port receives BDPUs, it cannot be an Edge Port. Edge Ports can be transitioned to the forwarding state immediately without causing a topology change notification because the change does not affect other switches/bridges.

With RSTP, rapid transition to the forwarding state is possible only under the following conditions:

- The port is an Alternate Port (the LAN segment to which it is attached can be either point-to-point or shared)
- The port is a Designated Port attached to a point-to-point segment (at most one other bridge can be attached to the segment)
- The port is an Edge Port

The LAN segment type is automatically derived from the duplex mode of a port. A port operating in full-duplex mode is assumed to be connected point-to-point, while a half-duplex port will be considered as a shared LAN port by default. In today's switched networks, most links are operating in full-duplex mode and are therefore treated as point-to-point links by RSTP.

Modified Topology Change Process: With RSTP, a Topology Change (TC) occurs only if a non-edge port is moved to a forwarding state. When an RSTP bridge detects a Topology Change, it flushes the MAC addresses associated with all its non-edge designated ports and its root port and sets the TC bit in its Hello BDPUs sent to neighbors. When a neighboring bridge receives a BPDU with the TC bit set, it clears the MAC addresses on all its ports except the one that received the TC and sends BPDUs with TC set on its Root Port and all its designated ports. As a result, the TC Notification is propagated very quickly across the whole network without reliance on the Root Bridge. In just a few Hello Times the stale entries in the CAM tables of the entire Layer 2 network are flushed.

Convergence without Timers: Port transitions to the discarding state can be made without the risk of creating a data loop. On the other hand, port transitions to the forwarding state are riskier and need to be consistent with the port states of neighboring bridges.

RSTP specifies that an Alternate Port may be placed in the forwarding state immediately on detection of the failure of the Root Port because the only other change in the network that is needed is to flush the address tables of the upstream switch in the tree.

For transitioning a Designated Port to forwarding state, consistency of Port Roles is achieved through an explicit handshake exchange between adjacent switches. The handshake process involves the following steps: The upstream switch (e.g., Switch A in Figure 3) sends a Proposal BPDU to its adjacent downstream switch (or switches). The downstream switch then makes sure that all its other ports are in sync with the proposal. Ports are in sync if they are in a discarding state or they are edge ports. This means that the downstream switch will have to place any designated forwarding ports into a Discarding or Blocked state. When this is done, an Agreement BPDU can be sent back to the upstream switch and the corresponding port placed in the forwarding state. Now the downstream switch initiates the handshake process with its downstream neighboring switches in order to attempt to transition its blocked designated ports into a forwarding state. The handshake process progresses tier-by-tier down the spanning tree as shown in the figure. At each step, the process ensures that no temporary loops have been introduced as ports are progressively transitioned to the forwarding state. When the edge of the tree is reached, full convergence has been achieved, without any need to wait for the expiration of protocol timers.

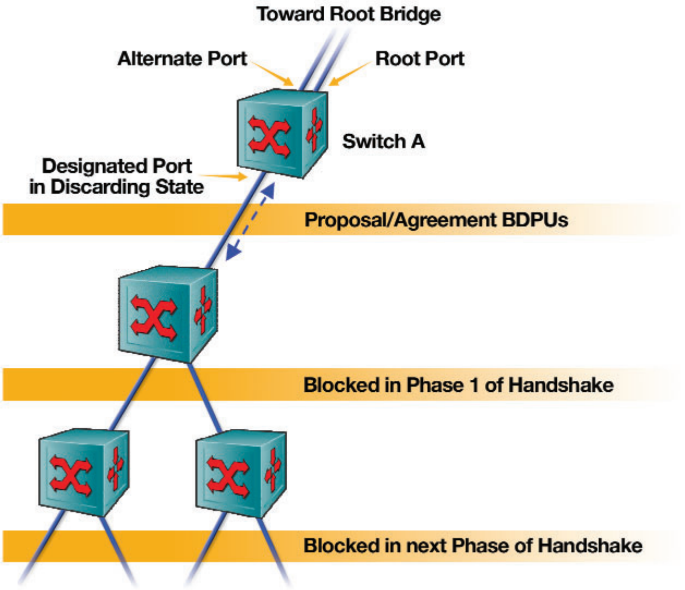

Toward Root Bridge
Alternate Port
Root Port
Switch A
Designated Port in Discarding State
Proposal/Agreement BDPUs
Blocked in Phase 1 of Handshake
Blocked in next Phase of Handshake

Figure 3. **RSTP handshake process**

Overview of Multiple Spanning Tree Protocol

The IEEE 802.1s Multiple Spanning Tree Protocol (MSTP) was developed to overcome the lack of VLAN awareness of 802.1D and the inefficiencies of PVST. MSTP provides for multiple instances of RSTP to coexist within a Layer 2 network with a group of VLANs sharing each instance of RSTP. Since the number of different logical topologies is generally much smaller than the number of VLANs, comparatively few RSTP instances are needed. For example, a typical enterprise access network might need only two RSTP instances, with each instance supporting up to 2,048 VLANs. MSTP therefore represents a significant improvement over proprietary STP extensions that support a separate instance of STP for each VLAN, especially where the network involves a large number of VLANs.

In order to allow the Layer 2 network to support a wide diversity of VLANs, 802.1s introduces the concept of MST regions. An MST region is defined to be a group of switches that share the same set of VLAN configuration attributes consisting of a configuration name, configuration revision number, and a configuration table that lists up to 4,096 VLANs and their association with MST instances of RSTP. Within a region, there may be multiple MST instances which run RSTP for the associated set of VLANs. In addition, 802.1s specifies an Internal Spanning Tree (IST), which is an additional instance of RSTP that spans all of the bridges in the region.

The boundaries of the regions are determined by boundary ports that connect one region to another. The boundary ports are discovered by BDPU exchange. The MSTP BDPU includes the configuration name, configuration revision number, and a digest (hash) of the VLAN association table. When two switches disagree on any of the configuration parameters, the ports between them are identified as boundary ports.

The MST regions of the Layer 2 network are connected by an 802.1Q Common Spanning Tree (CST), as shown in Figure 4. The IST instance is simply an RSTP instance that extends the CST within each MST region. Accordingly, the IST instance in each region receives and sends BPDUs to the CST. The IST represents the entire MST region as a single virtual bridge that participates in the CST. Therefore, the three regions shown in Figure 5 could be depicted as three virtual bridges. MST instance BDPUs are not forwarded over boundary ports, only CST BDPUs are forwarded by boundary ports.

In order to establish backward compatibility with 802.1D/802.1Q bridges, 802.1s bridges listen for 802.1D format BDPUs on their ports. When 802.1D format BDPUs are detected, the port uses standard 802.1D behavior to ensure compatibility.

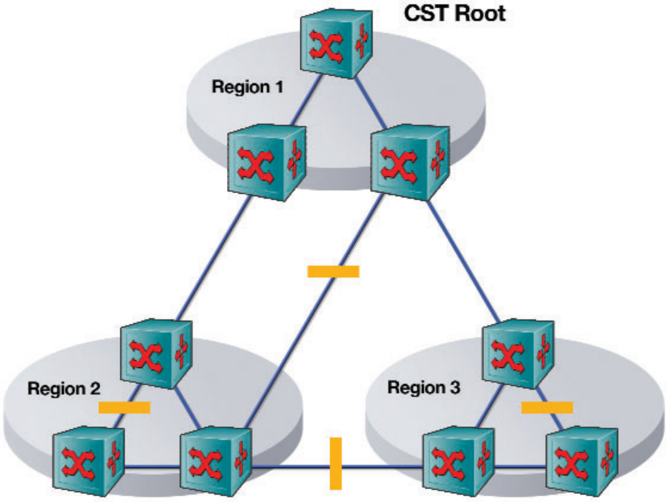


Figure 4. **CST/IST spanning three MSTP regions**

RSTP and MSTP in High Availability Networks

The combination of RSTP and MSTP allow switched network designs that feature fast convergence and active load sharing across redundant Layer 2 links and switches, as shown in Figure 5. In the figure there are two RSTP instances, one for the Red VLANs and one for the Blue VLANs. Each instance has its own Root Bridge. In normal operation, all four of the uplinks are carrying traffic and sharing the load based on VLAN color. Since approximately half of the end systems attached to each access switch belong to each color of LAN, the load is shared efficiently. The use of two colors of VLAN with separate RSTP instances avoids the overhead costs of

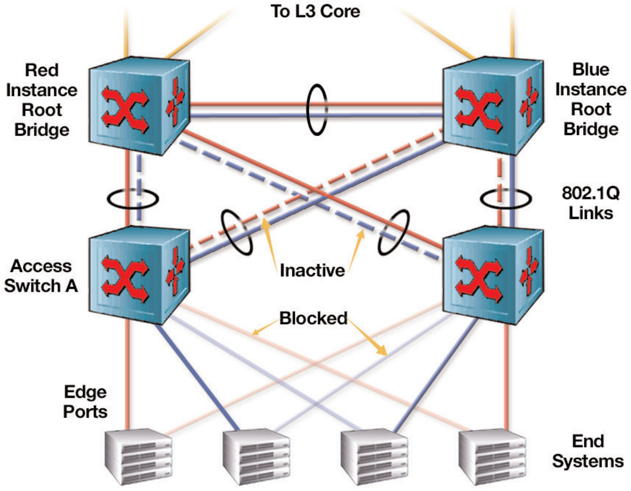


Figure 5. **RSTP and MSTP in a high availability network**

inactive redundant networking elements that standby in a passive mode until a failure occurs.

In addition, each end system is redundantly attached to the access switches using a dual port NIC that automatically fails over from the active port to the back up port (inactive) in the event of link or switch failure. Therefore, the access network is fully redundant with no single points of failure between the end system and the Core of the network.

When a failure does occur, IEEE Rapid Spanning Tree Protocol (RSTP) allows traffic to fail-over rapidly from primary to secondary paths:

- If the uplink on Switch A carrying Red VLAN traffic fails, the Red VLAN traffic will fail over to the uplink on its RSTP Alternate Port, which is currently carrying only Blue VLAN traffic, and the 802.1w topology change mechanism clears the appropriate entries in the upstream bridge's address tables. In this failure scenario, the Red VLAN traffic would then traverse the 802.1Q trunk to the Red Root Bridge and from there be forwarded to its destination. Because the failure of directly-attached links can be sensed from the interface, Switch A does not necessarily need to wait three Hello Intervals before failing over to the Alternate Port.
- If the Red VLAN Root Bridge fails, the Red VLAN traffic will again be immediately diverted to its Alternate Port while the Blue Root Bridge also assumes the role of the Root Bridge for the Red as well as the Blue VLANs. Again, fail-over can occur in as little as tens or hundreds of milliseconds. In this scenario, the remaining Root switch will be forwarding traffic for both Red and Blue VLANs, until the failed switch is restored.
- When an end system's active link fails or its primary access switch fails, it detects the failure and begins to send traffic over its backup NIC port. As soon as the secondary access switch receives traffic on the previously blocked Edge Port it can immediately place this port into the Forwarding state. Fail-over time in this scenario is almost entirely due to the inherent delay in the end system's dual port NIC fail-over protocol.

Conclusion

The combination of 802.1w (RSTP) and 802.1s (MSTP) provides the needed enhancements to 802.1D and 802.1Q that allow both enterprises and service providers to design switched Ethernet access networks that can deliver the high availability, resiliency, and VLAN scalability needed to support current and future generations of mission critical applications and services.

Force10 Networks, Inc.
350 Holger Way
San Jose, CA 95134 USA
www.force10networks.com

408-571-3500 PHONE
408-571-3550 FACSIMILE

WP19 107 v1.1