Spanning Tree Protocol (IEEE 802.1D 1998), Rapid STP (IEEE 802.1D 2004), Cisco PVST+, MSTP
STP Tuning – LAN Network Design
Agenda

• **Spanning Tree Protocol (STP)**
  – Introduction
  – Details
  – Convergence
  – Some more details

• **Rapid Spanning Tree Protocol (RSTP)**

• **Cisco PVST, PVST+**

• **Multiple Spanning Tree Protocol (MSTP)**

• **Spanning Tree Tuning – LAN Design**
Problem Description

• We want redundant links in bridged networks
• But transparent bridging cannot deal with redundancy
  – Broadcast storms and other problems
• Solution: STP (Spanning Tree Protocol)
  – Allows for redundant paths
  – Ensures non-redundant active paths
• Invented by Radia Perlman as general "mesh-to-tree" algorithm
• Only one purpose:
  cut off redundant paths with highest costs
I think that I shall never see
a graph more lovely than a tree
a graph whose crucial property
is loop-free connectivity.
A tree which must be sure to span
so packets can reach every lan.
first the root must be selected
by ID it is elected.
least cost paths to root are traced,
and in the tree these paths are place.
mesh is made by folks like me;
bridges find a spanning tree.

Radia Perlman
STP in Action (1)
No Broadcast Storm

DA = Broadcast address or nonexistent host address
STP in Action (2)
Bridge Failure – New STP Topology

DA = Broadcast address or non-existent host address

1  
2  
3  
4
Agenda

• **Spanning Tree Protocol (STP)**
  – Introduction
  – **Details**
  – Convergence
  – Some more details

• **Rapid Spanning Tree Protocol (RSTP)**

• **Cisco PVST, PVST+**

• **Multiple Spanning Tree Protocol (MSTP)**

• **Spanning Tree Tuning – LAN Design**
• Takes care that there is always exact only one active path between any 2 stations
• Implemented by a special communication protocol between the bridges
  – Using BPDU (Bridge Protocol Data Unit) frames with MAC-multicast address as destination address
• Three important STP parameters determine the resulting tree topology in a meshed network:
  – Bridge-ID
  – Interface-Cost
  – Port-ID
• **Bridge Identifier (Bridge ID)**
  - Consists of a priority number and the MAC-address of a bridge
    - Bridge-ID = Priority# (2 Byte) + MAC# (6 Byte)
  - Priority number may be configured by the network administrator
    - Default value is 32768
  - Lowest Bridge ID has highest priority
  - If you keep default values
    - The bridge with the lowest MAC address will have the highest priority
Parameters for STP

- **Port Cost (C)**
  - Costs in order to access local interface
  - Inverse proportional to the transmission rate
  - Default cost = \( 1000 \) / transmission rate in Mbit/s
    - With occurrence of 1Gbit/s Ethernet the rule was slightly adapted
    - May be configured to a different value by the network administrator

- **Port Identifier (Port ID)**
  - Consists of a priority number and the port number
    - Port-ID = port priority#.port#
    - Default value for port priority is 128
    - Port priority may be configured to a different value by the network administrator
Comparison Table For Port Costs:

<table>
<thead>
<tr>
<th>Speed [Mbit/s]</th>
<th>OriginalCost (1000/Speed)</th>
<th>802.1D-1998</th>
<th>802.1D-2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>100</td>
<td>100</td>
<td>2000000</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>19</td>
<td>200000</td>
</tr>
<tr>
<td>155</td>
<td>6</td>
<td>14</td>
<td>(129032 ?)</td>
</tr>
<tr>
<td>622</td>
<td>1</td>
<td>6</td>
<td>(32154 ?)</td>
</tr>
<tr>
<td>1000</td>
<td>1</td>
<td>4</td>
<td>20000</td>
</tr>
<tr>
<td>10000</td>
<td>1</td>
<td>2</td>
<td>2000</td>
</tr>
</tbody>
</table>

- Also different cost values might be used
  - See recommendations in the IEEE 802.1D-2004 standard to comply with RSTP and MSTP
  - 802.1D-2004 operates with 32-bit cost values instead of 16-bit
STP Parameter Example (1)
Spanning Tree Algorithm Summary

• Select the root bridge
  – Bridge with the lowest Bridge Identifier

• Select the root ports
  – By computation of the shortest path from any non-root bridge to the root bridge
  – Root port points to the shortest path towards the root

• Select one designated bridge for every LAN segment which can be reached by more than one bridge
  – Bridge with lowest root path costs on the root port side
  – Corresponding port on other side is called designated port

• Set the designated and root ports in forwarding state

• Set all other ports in blocking state
STP Parameter Example (2)

Root Bridge

LAN 1

C=10

B-ID 42

LAN 2

C=10

Designated Port

Root Port

C=10

B-ID 45

LAN 5

C=05

B-ID 57

LAN 4

C=05

B-ID 83

LAN 3

C=05

B-ID 97

LAN 3

Designated Bridge for LAN 3

Designated Port

C=05

Designated Bridge for LAN 4

Designated Port

C=05

Designated Port

C=10

Root Port

B-ID 42

LAN 1

C=10

Designated Port

C=10

Designated Port

© 2016, D.I. Lindner / D.I. Haas

Spanning-Tree Details, v6.0
BPDU Format

- Each bridge sends periodically BPDUs carried in Ethernet multicast frames
  - Hello time default: 2 seconds
- Contains all information necessary for building Spanning Tree

<table>
<thead>
<tr>
<th>Prot. ID</th>
<th>Prot. Vers.</th>
<th>BPDU Type</th>
<th>Flags</th>
<th>Root ID (R-ID)</th>
<th>Root Path Costs (RPC)</th>
<th>Bridge ID (O-ID)</th>
<th>Port ID (P-ID)</th>
<th>Msg Age</th>
<th>Max Age</th>
<th>Hello Time</th>
<th>Fwd. Delay</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Byte</td>
<td>1 Byte</td>
<td>1 Byte</td>
<td>1 Byte</td>
<td>8 Byte</td>
<td>4 Byte</td>
<td>8 Byte</td>
<td>2 Byte</td>
<td>2 Byte</td>
<td>2 Byte</td>
<td>2 Byte</td>
<td>2 Byte</td>
</tr>
</tbody>
</table>

The Bridge I regard as root

The total cost I see toward the root

My own ID
BPDU Fields in Detail (1)

- Protocol Identifier:
  - 0000 (hex) for STP 802.1D
- Protocol Version:
  - 00 (hex) for version 802.1D (1998)
  - 02 (hex) for version 802.1D (2004) - RSTP
- BPDU Type:
  - 00 (hex) for Configuration BPDU
  - 80 (hex) for Topology Change Notification (TCN) BPDU
- Root Identifier:
  - 2 bytes for priority (default 32768)
  - 6 bytes for MAC-address
- Root Path Costs in binary representation:
  - range 1-65535
- Bridge Identifier:
  - Structure like Root Identifier
BPDU Fields in Detail (2)

- Port Identifier:
  - 1 byte priority (default 128)
  - 1 byte port number
- Message Age (range 1-10s):
  - Age of Configuration BPDU
  - Transmitted by root-bridge initially using zero value, each passing-on (by designated bridge) increases this number
- Max Age (range 6-40s):
  - Aging limit for information obtained from Configuration BPDU
  - Basic parameter for detecting idle failures (e.g. root bridge = dead)
  - Default 20 seconds
- Hello Time (range 1-10s):
  - Time interval for generation of periodic Configuration BPDU by root bridge
  - Default 2 seconds
BPDU Fields in Detail (3)

- **Forward Delay (range 4-30s):**
  - Time delay for putting a port in the forwarding state
  - Default 15 seconds
  - That actually means:
  - 15 seconds LISTENING for allowing STP topology to converge after a topology change
  - plus
  - 15 seconds LEARNING to fill the empty MAC address table with locally seen MAC addresses in order to avoid flooding for any local MAC addresses
  - After that the ports are set to forwarding
- **Hello Time, Max Age, Forward Delay** are specified by Root-Bridge
- **Maximum Bridge Diameter**
  - Maximum number of bridges between any two end systems is 7 using default values for hello time, forward delay and max age
BPDU Fields in Detail (4)

- Flags (a "1" indicates the function):
  - Bit 8 ... Topology Change Acknowledgement (TCA)
  - Bit 1 ... Topology Change (TC)
  - Used in TCN BPDUs for signaling topology changes
    - TCN ... Topology Change Notification
    - The bridge recognizing the topology change sends a TCN BPDU on the root port until a CONF BPDU with TCA is received on its root port
    - Bridge one hop closer to the root passes TCN BPDU on towards the root bridge and acknowledges locally to the initiating bridge by usage of CONF BPDU with TCA
    - When the root bridge is reached a flushing of all bridging table is triggered by the root bridge by usage of CONF BPDUs with TC and TCA set
    - Now the new location (port) can be dynamically relearned by the actual user traffic
  - Note: In case of a topology change the MAC addresses should change quickly to another port of the corresponding bridging table (convergence) in order to avoid forwarding of frames to the wrong port/direction and not waiting for the natural timeout of the dynamic entry
BPDU MAC Addresses / LLC DSAP-SSAP

• Bridges use for STP-communication:
  – Multicast address:
    0180 C200 0000 hex
    0180 C200 0001 to 0180 C200 000F are reserved
    0180 C200 0010 hex All LAN Bridges Management Group Address
  – Note:
    • All addresses in Ethernet canonical format
  – The DSAP/SSAP of LLC header
    42 hex … Bridge Spanning Tree Protocol
Selection of Root Bridge

- R-ID=42
  - RPC=0
  - O-ID=42
  - P-ID=1

- R-ID=45
  - RPC=0
  - O-ID=45
  - P-ID=1

- R-ID=45
  - RPC=0
  - O-ID=45
  - P-ID=2

- R-ID=83
  - RPC=0
  - O-ID=83
  - P-ID=2

- R-ID=83
  - RPC=0
  - O-ID=83
  - P-ID=1

- R-ID=97
  - RPC=0
  - O-ID=97
  - P-ID=1

- R-ID=97
  - RPC=0
  - O-ID=97
  - P-ID=2

Port 1

Port 2
Root Bridge Selected, Triggers RPC Calculation

Root Bridge

R-ID=42
RPC=0
O-ID=42
P-ID=1

B-ID 42
C=10
Port 1

R-ID=42
RPC=0
O-ID=42
P-ID=2

B-ID 42
C=10
Port 2

B-ID 45
C=10

B-ID 57
C=05

B-ID 83
C=05

B-ID 97
C=10

C=05

C=05

C=05
Root Port Selection based on RPC (1)

- **Root Port Selection based on RPC (1)**
  - **Port 1**
    - R-ID=42
    - RPC=10
    - O-ID=45
    - P-ID=2
  - **Port 2**
    - R-ID=42
    - RPC=0
    - O-ID=42
    - P-ID=2
  - **R-ID=42**
    - RPC=0
    - O-ID=42
    - P-ID=2
  - **RPC=0**
    - B-ID 42
  - **RPC=5**
    - B-ID 45
  - **RPC=10**
    - B-ID 97
  - **RPC=5**
    - B-ID 83
  - **RPC=5**
    - B-ID 57
  - **RPC=10**
    - B-ID 45
  - **RPC=10**
    - B-ID 97
  - **RPC=10**
    - B-ID 57
  - **RPC=5**
    - B-ID 83
  - **RPC=5**
    - B-ID 45
  - **RPC=10**
    - B-ID 97
  - **RPC=5**
    - B-ID 57
  - **RPC=10**
    - B-ID 97
  - **RPC=5**
    - B-ID 57

© 2016, D.I. Lindner / D.I. Haas
Spanning-Tree Details, v6.0
Root Port Selection based on RPC (2)

- R-ID=42, RPC=0, O-ID=42, P-ID=1
- R-ID=42, RPC=10, O-ID=45, P-ID=2
- R-ID=42, RPC=0, O-ID=42, P-ID=2
- R-ID=42, RPC=10, O-ID=97, P-ID=3
- R-ID=42, RPC=10, O-ID=97, P-ID=3
- R-ID=42, RPC=5, O-ID=83, P-ID=1
- R-ID=42, RPC=0, O-ID=57, P-ID=2
- R-ID=42, RPC=10, O-ID=97, P-ID=2
- R-ID=42, RPC=0, O-ID=57, P-ID=2
- R-ID=42, RPC=5, O-ID=57, P-ID=2
- R-ID=42, RPC=5, O-ID=83, P-ID=1
- R-ID=42, RPC=5, O-ID=57, P-ID=2
- R-ID=42, RPC=5, O-ID=83, P-ID=1
Designated Bridge Selection (based on O-ID)

- **B-ID 42**: C=10, Port 1
- **B-ID 45**: C=10
- **B-ID 57**: C=05
- **B-ID 83**: C=05
- **B-ID 97**: C=10, Port 2

R-ID=42, RPC=0, O-ID=42, P-ID=1
R-ID=42, RPC=0, O-ID=42, P-ID=2
R-ID=42, RPC=5, O-ID=83, P-ID=1
R-ID=42, RPC=5, O-ID=57, P-ID=2
R-ID=42, RPC=10, O-ID=97, P-ID=3
R-ID=42, RPC=10, O-ID=97, P-ID=2

Final Topology

```
C=10  C=05  C=10  C=05
R-ID=42  RPC=0  O-ID=42  P-ID=1
B-ID 42

C=10  Port 1  Designated Port
RPC=0
B-ID 42

C=10  Port 2  Designated Port
RPC=0
B-ID 42

RPC=10  RPC=5
Root Port  Root Port
C=10  C=05
B-ID 45  B-ID 57

RPC=5
B-ID 83
Designated Bridge

RPC=5
C=05  Root Port
B-ID 97

RPC=10
C=05  C=05
R-ID=42  R-ID=42
RPC=10  RPC=5
O-ID=97  O-ID=57
P-ID=3  P-ID=2
```

© 2016, D.I. Lindner / D.I. Haas
Port States

At each time, a port is in one of the following states:
- Blocking, Listening, Learning, Forwarding, or Disabled

Only Blocking or Forwarding are final states (for enabled ports)

Transition states
- 15 s Listening state is used to converge STP
- 15 s Learning state is used to learn MAC addresses for the new topology

Therefore it lasts 30 seconds until a port is placed in forwarding state
Agenda

• **Spanning Tree Protocol (STP)**
  – Introduction
  – Details
  – **Convergence**
  – Some more details

• **Rapid Spanning Tree Protocol (RSTP)**

• **Cisco PVST, PVST+**

• **Multiple Spanning Tree Protocol (MSTP)**

• **Spanning Tree Tuning – LAN Design**
STP Error Detection

• The root bridge generates (triggers)
  – Every 1-10 seconds (hello time interval) a Configuration BPDU to be received on the root port of every other bridge and carried on through the designated ports
  – Bridges which are not designated are still listening to such messages on blocked ports

• If triggering ages out two scenarios are possible
  – Root bridge failure
    • A new root bridge will be selected based on the lowest Bridge-ID and the whole spanning tree may be modified
  – Designated bridge failure
    • If there is an other bridge which can support a LAN segment this bridge will become the new designated bridge
STP Convergence Time – Failure at Designated Bridge

- Time = max age (20 sec) to be waited until new STP is triggered
STP Convergence Time – Failure at Designated Bridge – New Topology

Convergence time = max age (20 sec) + 2 * forward delay (15 sec Listening + 15 sec Learning) = 50 sec
STP Convergence Time – Failure of Root Bridge

- Time = max age (20 sec) + 2*forward delay (15 sec Listening + 15 sec Learning) = 50 sec

© 2016, D.I. Lindner / D.I. Haas

Spanning-Tree Details, v6.0
STP Convergence Time – Failure of Root Bridge – New Topology

- Time = max age (20 sec) + 2*forward delay (15 sec Listening + 15 sec Learning) = 50 sec
**STP Convergence Time – Failure of Root Port**

- **Time = max age (20 sec) has not to be waited until new STP is triggered**

![Diagram of STP Convergence Time – Failure of Root Port](image-url)
STP Convergence Time – Failure of Root Port - Interruption of Connectivity D->A

- Convergence Time = 2*forward delay (15 sec Listening + 15 sec Learning) = 30 sec
STP Convergence Time – Failure of Root Port – Topology Change Notification (TCN)

LAN 2

C=10

Route Bridge
B-ID 42

C=10

MAC A

TCN to flush MAC entries in Bridging Table

LAN 1

MAC D

LAN 5

RP

MAC A

B-ID 57

B-ID 83

LAN 3

B-ID 97

LAN 4
STP Disadvantages

• Active paths are always calculated from the root, but the actual information flow of the network may use other paths
  – Note: network-manager can control this via Bridge Priority, Path Costs und Port Priority to achieve a certain topology under normal operation
  – Hence STP should be designed to overcome plug and play behavior resulted by default values

• Redundant paths cannot be used for load balancing
  – Redundant bridges may be never used if there is no failure of the currently active components
  – For remote bridging via WAN the same is true for redundant WAN links

• Convergence time between 30 and 50 seconds
  – Note: in order to improve convergence time Rapid Spanning Tree Protocol has been developed (802.1D version 2004)
Agenda

• **Spanning Tree Protocol (STP)**
  – Introduction
  – Details
  – Convergence
  – Some more details

• **Rapid Spanning Tree Protocol (RSTP)**

• **Cisco PVST, PVST+**

• **Multiple Spanning Tree Protocol (MSTP)**

• **Spanning Tree Tuning – LAN Design**
Usage for a Port-ID

• The Port-ID is only used as last tie-breaker
• Typical situation in highly redundant topologies: Multiple links between each two switches
  – Same BID and Costs announced on each link
  – Only local Port-ID can choose a single link
Importance of details…

• Many people think STP is a simple thing – until they encounter practical problems in real networks

• Important Details
  – STP State Machine
  – BPDU format details
  – TCN mechanism
  – RSTP
  – MSTP
Note: STP is a port-based algorithm

- Only the root-bridge election is done on the bridge-level
- All other processing is port-based
  - To establish the spanning tree, each enabled port is either forwarding or blocking
  - Additionally two transition states have been defined
STP State Machine: Port Transition Rules

- **STP is completely performed in the Listening state**
  - Blocking ports still receive BPDUs (but don’t send)

- **Default convergence time is 30-50 s**
  - 20s aging, (15+15)s transition time

- **Timer tuning: Better don't do it !**
  - Only modify timers of the root bridge
  - Don’t forget values on supposed backup root bridge

802.1d defines port roles and states:

<table>
<thead>
<tr>
<th>Port Roles</th>
<th>Port States</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root</td>
<td>Disabled</td>
</tr>
<tr>
<td>Designated</td>
<td>Blocking</td>
</tr>
<tr>
<td>Nondesignated</td>
<td>Listening</td>
</tr>
<tr>
<td></td>
<td>Learning</td>
</tr>
<tr>
<td></td>
<td>Forwarding</td>
</tr>
</tbody>
</table>
Example with L2 Switches

Three steps to create spanning tree:
1. Elect Root Bridge (Each L2-network has exactly one Root Bridge)
2. Elect Root Ports (Each non-root bridge has exactly one Root Port)
3. Elect Designated Ports (Each segment has exactly one Designated Port)

To determine root port and designated port:
1. Determine lowest (cumulative) Path Cost to Root Bridge
2. Determine lowest Bridge ID
3. Determine lowest Port ID
Components of the Bridge-ID

- The recent 802.1D-2004 standard requires only 4-bits for priority and 12 bits to distinguish multiple STP instances
  - Typically used for MSTP, where each set of VLANs has its own STP topology
- Therefore, ascending priority values are 0, 4096, 8192, ...
  - Typically still configured as 0, 1, 2, 3 ...

Old:
- 2 Bytes: Priority
- 6 Bytes: Lowest MAC Address

New:
- 4 Bits: Priority
- 12 Bits: Extended System ID
- 6 Bytes: Lowest MAC Address

To allow distinct BIDs per VLAN as used by MSTP or Cisco per VLAN-STP

Typically derived from Backplane or Supervisor module
### Detailed BPDU Format

#### When first booted, Root-ID == BID
- A TCN-BPDU only consists of these 3 fields !!!

#### If value increases, then the originating bridge lost connectivity to Root Bridge
- Predetermined by root bridge
- Affect convergence time
- Misconfigurations cause loops

<table>
<thead>
<tr>
<th>Bytes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protocol ID</td>
<td>2</td>
</tr>
<tr>
<td>Version</td>
<td>1</td>
</tr>
<tr>
<td>Message Type</td>
<td>1</td>
</tr>
<tr>
<td>Flags</td>
<td>1</td>
</tr>
<tr>
<td>Root ID</td>
<td>8</td>
</tr>
<tr>
<td>Root Path Cost</td>
<td>4</td>
</tr>
<tr>
<td>Bridge ID</td>
<td>8</td>
</tr>
<tr>
<td>Port ID</td>
<td>2</td>
</tr>
<tr>
<td>Message Age</td>
<td>2</td>
</tr>
<tr>
<td>Maximum Age = 20</td>
<td>2</td>
</tr>
<tr>
<td>Hello Time = 2</td>
<td>2</td>
</tr>
<tr>
<td>Forward Delay = 15</td>
<td>2</td>
</tr>
</tbody>
</table>

- Always zero
- Always zero
- Configuration (0x00) or TCN BPDU (0x80)
- LSB = Topology change flag (TC), MSB = TC Ack flag (TCA)
- Who is Root Bridge?
- How far away is Root Bridge?
- ID of bridge that sent this BPDU
- Port-ID of sending bridge (unique: Port1/1=0x8001, 1/2=0x8002, ...)
- Time since Root generated this BPDU
- BPDU is discarded if older than this value (default: 20 seconds)
- Broadcast interval of BPDU (default: 2 seconds)
- Time spent in learning and listening states (default: 15 seconds)

**BPDUs are sent in 802.3 frames**
- DA = 01-80-C2-00-00-00
- LLC has DSAP=SSAP = 0x42 ("the answer")

**Configuration BPDU**
- Originated by Root Bridge periodically (2 sec Hello Time), flow downstream
Topography Change Notification (TCN)

- **Special BPDUs, used as alert by any bridge**
  - Flow upstream (through Root Port)
  - Only consists of the first three standard header fields!
  - It is transported as TCN BPDU

- **Sent upon**
  - Transition of a port into Forwarding state and at least one Designated Port exists
  - Transition of a port into Blocking state (from either Forwarding or Learning state)

- **Sent until acknowledged by TC Acknowledge (TCA)**
  - Which is actually a Conf BPDU from the upstream bridge
**Topology Change Notification (TCN)**

- Only the Designated Ports of upstream bridges processes TCN-BPDUs and send TC-Ack (TCA) downstream
- Finally the Root Bridge receives the TC and sends Configuration BPDUs with the TC and TCA flag set to 1 (=TCA) downstream for (Forward Delay + Max Age = 35) seconds
  - This instructs all bridges to reduce the default bridging table aging (300 s) to the current Forward Delay value (15 s)
  - Thus bridging tables can adapt to the new topology
Cisco Port Fast

- **Optimizes switch ports connected to end-station devices**
  - Usually, if PC boots, NIC establishes L2-link, and switch port goes from Disabled=>Blocking=>Listening=>Learning=>Forwarding state ... 30 seconds!!!

- **Port Fast allows a port to immediately enter the Forwarding state**
  - STP is NOT disabled on that port!

- **Port Fast only works once after link comes up!**
  - If port is then forced into Blocking state and later returns into Forwarding state, then the normal transition takes place!
  - Ignored on trunk ports

- **Alternatives:**
  - Disable STP (often a bad idea)
  - Use a hub in between => switch port is always active
Cisco Uplink Fast (1)

• **Accelerates STP to converge within 1-3 seconds**
  – Cisco patent
  – Marks some blocking ports as backup uplink

• **Typically used on access layer switches**
  – Only works on non-root bridges
  – **Requires some blocked ports**
  – Enabled for entire switch (and not for individual VLANs)
Cisco Uplink Fast (2): The Problem

• When link to root bridge fails, STP requires (at least) 30 seconds until alternate root port becomes active
Cisco Uplink Fast (3): Idea

• When a port receives a BPDU, we know that it has a path to the root bridge
  – Put all root port candidates to a so-called "Uplink Group"
• Upon uplink failure, immediately put best port of Uplink group into forwarding state
  – There cannot be a loop because previous uplink is still down
Cisco Uplink Fast (4): Incorrect Bridging Tables

- But upstream bridges still require 30 s to learn new topology
- Bridging table entries in upstream bridges may be incorrect
Cisco Uplink Fast (5): Actively Correct Tables

- Uplink Fast corrects the bridging tables of upstream bridges
- Sends 15 multicast frames (one every 100 ms) for each MAC address in its bridging table (i.e., for each downstream host)
  - Using SA=MAC: All other bridges quickly reconfigure their tables; dead links are no longer used
  - DA=01-00-0C-CD-CD-CD, flooded throughout the network

MAC B is at g3/17

Packet for MAC B
DA=01-00-0C-CD-CD-CD
SA=MAC B

MAC A

MAC B

Packet for MAC B
DA=01-00-0C-CD-CD-CD
SA=MAC B

g1/3

g3/17
Cisco Uplink Fast (6): Additional Details

• When broken link becomes up again, Uplink Fast waits until traffic is seen
  – That is, 30 seconds plus 5 seconds to support other protocols to converge (e. g. Etherchannel, DTP, …)

• Flapping links would trigger uplink fast too often which causes too much additional traffic
  – Therefore the port is "hold down" for another 35 seconds before Uplink Fast mechanism is available for that port again

• Several STP parameters are modified automatically
  – Bridge Priority = 49152 (don't want to be root)
  – All Port Costs += 3000 (don't want to be designated port)
Cisco Backbone Fast (1)

- Complementary to Uplink Fast
- Safes 20 seconds when recovering from indirect link failures in core area
  - Issues Max Age timer expiration
  - Reduce failover performance from 50 to 30 seconds
  - Cannot eliminate Forwarding Delay
- Should be enabled on every switch!
Cisco Backbone Fast (2): The Problem

• Consider initial situation
• Note that blocked port (g0/1) always remembers "best seen" BPDU – which has best (=lowest) Root-BID
Now backup-root bridge loses connectivity to root bridge and assumes root role
Port g0/1 does not see the BPDUs from the original root bridge any more
But for MaxAge=20 seconds, any inferior BPDU is ignored
Cisco Backbone Fast (4): The Problem (cont.)

• Only after 20 seconds port g0/1 enters listening state again
• Finally, bridge A unblocks g0/1 and forwards the better BPDUs to bridge B
• Total process lasts 20+15+15 seconds
Cisco Backbone Fast (5): The Solution

• If an inferior BPDU is originated from the local segment's Designated Bridge, then this probably indicates an indirect failure
  – (Bridge B was Designated Bridge in our example)

• To be sure, we ask other Designated Bridges (over our other blocked ports and the root port) what they think which bridge the root is
  – Using Root Link Query (RLQ) BPDU

• If at least one reply contains the "old" root bridge, we know that an indirect link failure occurred
  – Immediately expire Max Age timer and enter Listening state
Other CISCO STP Tuning Options

- **BPDU Guard**
  - Shuts down PortFast-configured interfaces that receive BPDUs, preventing a potential bridging loop

- **Root Guard**
  - Forces an interface to become a designated port to prevent surrounding switches from becoming the root switch

- **BPDU Filter**

- **BPDU Skew Detection**
  - Report late BPDUs via Syslog
  - Indicate STP stability issues, usually due to CPU problems

- **Unidirectional Link Detection (UDLD)**
  - Detects and shuts down unidirectional links

- **Loop Guard**
Agenda

• **Spanning Tree Protocol (STP)**
  – Introduction
  – Details
  – Convergence
  – Some more details

• **Rapid Spanning Tree Protocol (RSTP)**

• **Cisco PVST, PVST+**

• **Multiple Spanning Tree Protocol (MSTP)**

• **Spanning Tree Tuning – LAN Design**
Introduction

• **RSTP is part of the IEEE 802.1D-2004 standard**
  – Originally defined in IEEE 802.1w
  – Old STP IEEE 802.1D-1998 is now superseded by RSTP

• **Computation of the Spanning Tree is identical between STP and RSTP**
  – Conf-BPDU and TCN-BPDU still remain
  – New BPDU type "RSTP" has been added
    • Version=2, type=2

• **RSTP BPDUs can be used to negotiate port roles on a particular link**
  – Only done if neighbor bridge supports RSTP (otherwise only Conf-BPDUs are sent
  – Using a **Proposal/Agreement** handshake

• **Designed to be compatible and interoperable with the traditional STP – without additional management requirements**
Major Features

• BPDUs are no longer triggered by root bridge
  – Instead, each bridge can generate BPDUs independently and immediately (on-demand)

• Much faster convergence
  – Few seconds (typically within 1 – 5 seconds)

• Better scalability
  – No network diameter limit

• New port roles and port states
  – Non-Designated Port role split in Alternate and Backup
  – Root Port and Designated Port role still remain the same
  – Port state discarding instead of disabled, learning and blocking
<table>
<thead>
<tr>
<th>STP (802.1d) Port State</th>
<th>RSTP (802.1w) Port State</th>
<th>Is Port included in active Topology?</th>
<th>Is Port learning MAC addresses?</th>
</tr>
</thead>
<tbody>
<tr>
<td>disabled</td>
<td>discarding</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>blocking</td>
<td>discarding</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>listening</td>
<td>discarding</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>learning</td>
<td>learning</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>forwarding</td>
<td>forwarding</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Backup and Alternate Ports

• If a port is neither Root Port nor Designated Port
  – It is a **Backup Port** – if this bridge is a Designated Bridge for that LAN
  – Or an **Alternate Port** otherwise
BPDU Types (Old and New)

Configuration BPDU

<table>
<thead>
<tr>
<th>Protocol ID</th>
<th>Protocol Version</th>
<th>BPDU Type</th>
<th>Root Bridge ID (BID of bridge believed to be the root by the transmitter)</th>
<th>Root Path Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Topology Change BPDU

<table>
<thead>
<tr>
<th>Protocol ID</th>
<th>Protocol Version</th>
<th>BPDU Type</th>
<th>Root Bridge ID (BID of bridge believed to be the root by the transmitter)</th>
<th>Root Path Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

RSTP BPDU

<table>
<thead>
<tr>
<th>Protocol ID</th>
<th>Protocol Version</th>
<th>BPDU Type</th>
<th>Root Bridge ID (BID of bridge believed to be the root by the transmitter)</th>
<th>Root Path Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: The RSTP BPDU replaces the Configuration BPDU and the Topology Change BPDU

Flags

<table>
<thead>
<tr>
<th>TCAck</th>
<th>agree</th>
<th>fwd</th>
<th>learn</th>
<th>prop</th>
<th>TCN</th>
</tr>
</thead>
</table>

Port Role:

- 0 0 = Unknown
- 0 1 = Alternate or Backup
- 1 0 = Root
- 1 1 = Designated

Version 1 Length

<table>
<thead>
<tr>
<th>Version 1 Length</th>
<th>0000 0000 indicates that there is no Version 1 protocol information present</th>
</tr>
</thead>
</table>

all set to zero means RSTP but also STP!

λSTP BPDU: 0000 0000 λRSTP BPDU: 0000 0010

λ must be less than Max Age

- 20 seconds
- 2 seconds
- 15 seconds

© 2016, D.I. Lindner / D.I. Haas

Spanning-Tree Details, v6.0
BPDU Flag Field – New Values

- TC and TCA were already introduced by old STP
- Other bits were unused by old STP
- RSTP also uses the 6 remaining bits
Proposal/Agreement Sequence

- Suppose a new link is created between the root and switch A and a new switch B is inserted.

**Diagram:**

1. **Proposal**
   - Root to A
   - P1 Designated Port
   - Forwarding State

2. **Agreement**
   - P1 Root Port

1. **Proposal**
   - Root to A
   - P2 Root Port

2. **Agreement**
   - P3 Designated Port
   - Forwarding State
   - P4 Root Port
NEW BPDU Handling

• Faster Failure Detection
  – BPDUs acting now as keepalives messages
    • Different to the 802.1D STP a bridge now sends a BPDU with its current information every `<hello-time>` seconds (2 by default), even if it does not receive any BPDU from the root bridge
  – If hellos are not received for 3 consecutive times, port information is invalidated
    • Because BPDU's are now used as keepalive mechanism between bridges
    • If a bridge fails to receive BPDUs from a neighbor, the connection has been lost
  – Max age not used anymore
    • For listening and waiting for STP to converge
Algorithm Overview

• Designated Ports transmit Configuration BPDUs periodically to detect and repair failures
  – Blocking (aka Discarding) ports send Conf-BPDUs only upon topology change

• Every Bridge accepts "better" BPDUs
  – From any Bridge on a LAN or revised information from the prior Designated Bridge for that LAN

• To ensure that old information does not endlessly circulate through redundant paths in the network and prevent propagation of new information
  – Each Configuration Message includes a message age and a maximum age

• Transitions to Forwarding is now confirmed by downstream bridge
  – Therefore no Forward-Delay is necessary!
Link Types and Edge Port

• **Shared Link (Half Duplex !!!)**
  – Are not supported by RSTP (ambiguous negotiations could result)
  – Instead slow standard STP is used here

• **Point-to-point Link (Full Duplex !!!)**
  – Supports proposal-agreement process

• **Edge Port**
  – Hosts reside here
  – Transitions directly to the Forwarding Port State, since there is no possibility of it participating in a loop
  – May change their role as soon as a BPDU is seen

• **RSTP fast transition**
  – Only possible on edge ports or point-to-point links
Main Differences to STP

• BPDUs are sent every hello-time, and not simply relayed anymore
  – Immediate aging if three consecutive BPDUs are missing

• When a bridge receives better information ("I am root") from its DB, it immediately accepts it and replaces the one previously stored
  – But if the RB is still alive, this bridge will notify the other via BPDUs
A new link between A and Root is being added to the bridged network.

New port activated, BPDUs start traveling.

New port coming up on the root will immediately cause switch A to enter the listening state hence blocking all traffic.

BPDUs from the root start propagating towards the leaves through A hence blocking also downstream links.

Current Spanning Tree.
Very quickly, the BPDUs from the root bridge reach D that immediately blocks its port P1.

The topology has now converged, but the network is disrupted for twice forward delay because all switches needs time for listening (STP convergence time) and learning.

30 seconds no network connectivity !!!
Slow Convergence with Legacy STP

New Spanning Tree
A new link between A and Root is being added to the bridged network.

Both ports on link between A and the root are put in so-called designated blocking as soon as they come up.

As soon as A receives the root's BPDU, it blocks its non-edge designated ports until synchronization is achieved. Through the agreement A explicitly authorizes the root bridge to put its port in forwarding.
Now the link between switch A and the root is put in forwarding state.

The network below switch A is still blocking until port roles are negotiated at the next stage between switch A and switch B or A and C.

Switch B and C will enter the new spanning tree and A will put its ports in the forwarding state and the negotiations will proceed between C and D.
Switch C blocks its port to D because it root path costs of D are better than the root path costs of C.

We have reached the final topology, which means that port P1 on D ends up blocking. It's the same final topology as for the STP example.

But we got this topology just time necessary for the new BPDU's to travel down the tree. No timer has been involved in this quick convergence.

Convergence Time < 1 second
Rapid Transition in Detail

**Basic Principle**

- The new rapid STP is able to **actively confirm** that a port can safely transition to forwarding without relying on any timer configuration
  - Feedback mechanism
- **Edge Ports connect hosts**
  - Cannot create bridging loops
  - Immediate transition to forwarding possible
  - No more Edge Port upon receiving BPDU
- **Rapid transition only possible if Link Type is point-to-point**
  - No half-duplex (=shared media)

**Legacy STP Details**

- Upon receiving a (better) BPDU on a blocked/previously-disabled port, 15+15 seconds transition time needed until forwarding state reached
- But received BPDUs are propagated immediately downstream; some bridges below may detect a new Root Port candidate and also require 15+15 seconds transition time
- Network in between is unreachable for 30 seconds!!!

**NEW: Sync Operation**

- Not the Root Port candidates are blocked, but the designated ports downstream—this avoids potential loops, too!
- Bridge explicitly authorizes upstream bridge to put Designated Port in forwarding state (sync)
- Then the sync-procedure propagates downstream

**More Details**

1) A new link is created between the root and Switch A.
2) Both ports on this link are put in a designated blocking state until they receive a BPDU from their counterpart.
3) Port p0 of the root bridge sets "proposal bit" in the BPDU (step 1)
4) Switch A then starts a sync to ensure that all of its ports are in-sync with this new information (only blocking and edge-ports are currently in-sync). Switch A just needs to block port p3, assigning it the discarding state (step 2).
5) Switch A can now unblock its newly selected root port p1 and reply to the root by sending an agreement message (Step 3, same BPDU with agreement bit set)
6) Once p0 receives that agreement, it can immediately transition to forwarding.
7) Now port 3 will send a proposal downwards, and the same procedure repeats.
**Topology Change**

- **802.1d: When a bridge detects a topology change**
  - A TCN is sent towards the root.
  - Root sends Conf-BPDU with TC-bit downstream (for 10 BPDUs).
  - All other bridges can receive it and will reduce their bridging-table aging time to forward-delay seconds, ensuring a relatively quick flushing of stale information.

- **RSTP: Only non-edge ports moving to the forwarding state cause a TCN**
  - Loss of connectivity NOT regarded as topology change any more.
  - TCN is immediately flooded throughout whole domain.
  - Every bridge flushes MAC addresses and sends TCN upstream (RP) and downstream (DPs).
  - Other bridges do the same: Now, the TCN-process is a one-step procedure, as the TCNs do not need to reach the root first and require the root for re-origination downstream.
IEEE 802.1w is an improvement of 802.1d
- Vendor-independent (Cisco's Uplink Fast, Backbone Fast, and Port Fast are proprietary)

The three 802.1d states disabled, blocking, and listening have been merged into a unique 802.1w discarding state

Nondesignated ports on a LAN segment are split into alternate ports and backup ports
- A backup port receives better BPDUs from the same switch
- An alternate port receives better BPDUs from another switch

Other changes:
- BPDU are sent every hello-time, and not simply relayed anymore.
- Immediate aging if three consecutive BPDUs are missing
- When a bridge receives inferior information ("I am root") from its DB, it immediately accepts it and replaces the one previously stored. If the RB is still alive, this bridge will notify the other via BPDUs.
Agenda

• Spanning Tree Protocol (STP)
  – Introduction
  – Details
  – Convergence
  – Some more details

• Rapid Spanning Tree Protocol (RSTP)

• Cisco PVST, PVST+

• Multiple Spanning Tree Protocol (MSTP)

• Spanning Tree Tuning – LAN Design
• In over 70% of all enterprise networks you will encounter Cisco switches

• Cisco extended STP and RSTP with a per-VLAN approach: "Per-VLAN Spanning Tree"

• Advantages:
  – Better (per-VLAN) topologies possible
  – STP-Attacks only affect current VLAN

• Disadvantages:
  – Interoperability problems might occur
  – Resource consumption (800 VLANs means 800 STP instances)
Example

- Remember that root bridge should realize the center of the LAN
  - Attracts all traffic
  - Typically servers or Internet-connectivity resides there
- Different VLANs might have different cores
- PVST+ allows for different topologies
  - Admin should at least configure ideal root bridge BID manually
Scalability Problem

- Typically the number of VLANs is much larger than the number of switches
- Results in many identical topologies
- In the above example we have 400 VLANs but only three different logical topologies
  - 400 Spanning Tree instances
  - 400 times more BPDUs running over the network
PVST (Classical, OLD!)

- Cisco proprietary (of course)
- Interoperability problems when also standard CST is used in the network (different trunking requirements)
- Provides dedicated STP for every VLAN
- Requires ISL
  - Inter Switch Link (Cisco's alternative to 802.1Q)
PVST+

- Today standard in Cisco switches
  - Default mode
  - Interoperable with CST

- The PVST BPDUs are also called SSTP BPDUs

- The messages are identical to the 802.1d BPDU but uses SNAP instead of LLC plus a special TLV at the end
PVST+ Protocol Details

• For native VLAN on trunk, normal (untagged) 802.1d BPDUs are sent
  – Also to the IEEE destination address 0180.c200.0000

• For tagged VLANs, PVST+ BPDUs use
  – SNAP, OID=00:00:0C, and EtherType 0x010B
  – Destination address 01-00-0c-cc-cc-cd
  – Plus 802.1Q tag

• Additionally a "PVID" TLV field is added at the end of the frame
  – This PVID TLV identifies the VLAN ID of the source port
  – The TLV has the format:
    • type (2 bytes) = 0x00 0x34
    • length (2 bytes) = 0x00 0x02
    • VLAN ID (2 bytes)
    • Also usually some padding is appended
PVST+ Compatibility Issues

• PVST+ switches can act as translators between groups of Cisco PVST switches (using ISL) and groups of CST switches
  – Sent untagged over the native 802.1Q VLAN
  – BPDUs of PVST-based VLANs are practically 'tunneled' over the CST-based switches using a special multicast address (the CST based switches will forward but not interpret these frames)

• Not important anymore…
Agenda

• **Spanning Tree Protocol (STP)**
  – Introduction
  – Details
  – Convergence
  – Some more details

• **Rapid Spanning Tree Protocol (RSTP)**

• **Cisco PVST, PVST+**

• **Multiple Spanning Tree Protocol (MSTP)**

• **Spanning Tree Tuning – LAN Design**
Overview

• Also the MSTP standard contains contributions from Cisco
  – IEEE 802.1Q-2003 (former 802.1s)

• Solves the cardinality mismatch between the number of VLANs and the number of useful topologies

• Switches are organized in Regions

• In each Region sets of VLANs can be independently assigned to one out of 16 Spanning Tree Instances

• Each Instance has its own Spanning Tree topology
Example

- Compared to PVST+ only three Spanning Tree Topologies (=Instances) required
- Each STP instance has assigned 200 VLANs
  - Each VLAN can only be member of one instance of course
MSTP Details

• Each switch maintains its own MSTP configuration which contains the following mandatory attributes:
  – The configuration name (32 chars),
  – The revision number (0..65535),
  – The element table which specifies the VLAN to Instance mapping

• All switches in a region must have the same attributes
Regions

- The bridges checks attribute equivalence via a digest contained in the BPDUs
  - Note that the attributes must be configured manually and are NOT communicated via the BPDUs
- If digest does not match then we have a region boundary port
- Regions are only interconnected by the Common Spanning Tree (CST)
  - Instance 0
  - Uses traditional 802.1d STP
Region Example

- Only the logical STP topologies are shown (not the physical links)
- Each region has internal STP instances (red and blue)
- One CST instance interconnects all regions (black)
Note

• When enabling MSTP, per default the CST (instance zero) has all VLANs assigned
• Each region must be MSTP-aware
  – Since only a subset of VLANs is assigned to the CST
  – Old-STP switched always create a general (all-VLAN) topology
  – Don't let MSTP-unaware switch become root bridge
Agenda

• Spanning Tree Protocol (STP)
• Rapid Spanning Tree Protocol (RSTP)
• Cisco PVST, PVST+
• Multiple Spanning Tree Protocol (MSTP)
• Spanning Tree Tuning – LAN Design
  – Design Considerations
  – Design Solution – Best Practices
  – Failover Handling
  – Advanced Techniques - Teaming
  – LAN – WAN Interconnection
Design Principles

• Avoiding of “Single Point of Failure”
  – Physical link failure
    • Access link
    • Trunk link (LAN or WAN)
  – Network component failure
    • L2 Switch
    • Router, DHCP Server, DNS Server, Production Server

• Load balancing in normal situations

• Server with two or more NIC’s
  – OS must support parallel operation and/or switch over between cards

• Clients with two network outlets
  – Two NIC’s and special OS aspects may not economically be justified
Physical Layer

2 workgroup switches on every floor

Network outlets – at least 2 in every room

© 2016, D.I. Lindner / D.I. Haas
Spanning-Tree Details, v6.0 99
Spanning Tree Problem Plug & Play

One Single VLAN

S1

1GE FO

1GE FO

1GE FO

10GE FO

S2

10GE FO

Root Bridge

FE CU

GE CU
Slower trunks are taken and 5 hops must be passed in order to reach the server.
Spanning Tree Problem Plug & Play

Possible packet path for packet coming from outside

Slower trunks are taken and 4 hops must be passed in order to reach the server.
Figure Out STP in Complex Scenarios

Therefore to limit the complexity of STP design and the amount of hops for reaching an other device on LAN or WAN, a layered design is necessary.
Figure Out STP in Complex Scenarios
Spanning Tree – Good Luck
Figure Out STP in Complex Scenarios
Spanning Tree – Bad Luck with Bridge IDs

Root Bridge
- ID = 12

S1
- ID = 16

S2
- ID = 29
- ID = 19
- ID = 15
- ID = 22
- ID = 13

ID = 36
Spanning Tree Problem Unequal Load Balancing with Single VLAN

S1 = Root Bridge

Most trunks of S2 are unused

Faster trunks are taken and only 3 hops must be passed in the worst case
One Single VLAN
Improvement on Trunk S1 – S2 by Using GEC / LCAP

S1 = Root Bridge

Most trunks of S2 are unused

1GE FO

1GE FO

1GE FO

1GE FO

10GE FO with GEC

Faster trunks are taken and only 3 hops must be passed in the worst case

One Single VLAN
Agenda

• Spanning Tree Protocol (STP)
• Rapid Spanning Tree Protocol (RSTP)
• Cisco PVST, PVST+
• Multiple Spanning Tree Protocol (MSTP)
• **Spanning Tree Tuning – LAN Design**
  – Design Considerations
  – **Design Solution – Best Practices**
  – Failover Handling
  – Advanced Techniques - Teaming
  – LAN – WAN Interconnection
Best Practices

• **Build at least two separated VLAN's**
  – In case of IP that means two IP subnets

• **How to achieve?**
  – Per VLAN STP (Cisco)
  – MIST (Multiple Instances Spanning Tree)
    • IEEE 802.1d

• **Tune STP parameters**
  – In order to use all trunks and all switches in a similar way
Build STP for VLAN 1

S1 = Root Bridge
Red VLAN

W1
W2
WX
WY

S2

Mac A

Mac A

F1
F2

© 2016, D.I. Lindner / D.I. Haas
Spanning-Tree Details, v6.0
Build STP for VLAN 2

S1

Mac B

S2 = Root Bridge
Blue VLAN

W1
W2
WX
WY

F1
F2

Mac B

Mac B

Mac B

Mac B

© 2016, D.I. Lindner / D.I. Haas
Spanning-Tree Details, v6.0

111
Solution – Load Balancing using 2 VLAN's

Load balancing on every trunk if clients are equally distributed to workgroup switches and perform with similar statistic to server farm.
Agenda

• Spanning Tree Protocol (STP)
• Rapid Spanning Tree Protocol (RSTP)
• Cisco PVST, PVST+
• Multiple Spanning Tree Protocol (MSTP)
• **Spanning Tree Tuning – LAN Design**
  – Design Considerations
  – Design Solution – Best Practices
  – Failover Handling
  – Advanced Techniques - Teaming
  – LAN – WAN Interconnection
Link Failure (Trunk)

Trunk W1-S2 is down
Red VLAN is not affected
Link Failure (Trunk) - Solution

STP for Blue VLAN will converge to trunk via W1-S1

Trunk of W1-S1 will additionally take the load of all affected PC’s

S1, S2 switches will learn MAC addresses of PC’s with first frames sent by the PC’s (TCN may help)
Switch Failure (Access)

Switch W1 is down.
Red VLAN is not affected if PCs are connected to W2.
Switch Failure (Access) - Solution

Switch PC to the other network outlet of the given VLAN (Blue)

W2, S2 will learn MAC address of PC with first frame sent by this PC and trunk of W2-S2 will take the additional load of this PC
Link Failure (Access)

Link to W1 is down
Link Failure (Access) - Solution

Switch PC to the other network outlet of the given VLAN (Blue)

W2, S2 will learn MAC address of PC with first frame sent by this PC and trunk of W2-S2 will take the additional load of this PC
Switch S2 is down

Red VLAN is not affected if PCs are connected to W2
STP for Blue VLAN will converge to new root bridge S1; All trunks will additionally take the load of all blue PC’s and all switches will learn MAC address of blue PC with first frames sent; special care for WAN traffic !!!
Trunk Failure (F-Switches)

Trunk F1-S2 is down
Red VLAN is not affected
Trunk Failure (F-Switches) - Solution

STP for Blue VLAN will converge to trunk via F2-S1

Trunk of F2-S1 will additionally take the load of all affected PC’s

S1, S2 switches will learn MAC address of server with first frame sent by the server (TCN may help)
Switch Failure (F-Switches)

Switch F2 is down
Red VLAN is not affected
Switch server to the other network outlet of the given VLAN (Blue); In the meantime the server is still reachable via second port!!! S2 will learn MAC address of server with first frame sent by this server and trunk of F1-S2 will finally take the load to the server.
Link Failure (F-Switches)

Switch F2 is down
Red VLAN is not affected
Link Failure (F-Switches) - Solution 2A

Switch server to the other network outlet of the given VLAN (Blue); In the meantime the server is still reachable via second port!!! S2 will learn MAC address of server with first frame sent by this server and trunk of F1-S2 will finally take the load to the server.
Switch server to the other network outlet of the given VLAN (Blue); In the meantime the server is still reachable via second port !!!
Blue clients must go via router !!!
Agenda

- Spanning Tree Protocol (STP)
- Rapid Spanning Tree Protocol (RSTP)
- Cisco PVST, PVST+
- Multiple Spanning Tree Protocol (MSTP)
- **Spanning Tree Tuning – LAN Design**
  - Design Considerations
  - Design Solution – Best Practices
  - Failover Handling
  - Advanced Techniques - Teaming
  - LAN – WAN Interconnection
Server Connections to F-Switches – Advanced Techniques

Outlet for normal operation

Outlet for link failure
Outlet for F-switch failure

Standby NIC’s: can take over MAC address of normal NIC (OS support!!!)
Switch Failure (F-Switches)

Outlet for normal operation

Red VLAN not affected

Server switches to blue backup NIC – special NIC needed
New Server Connection for Blue VLAN to F-Switches

F1 learns MAC of server’s blue backup NIC

Outlet for normal operation

Server switches to blue backup NIC – special NIC needed
Link Failure (F-Switches)

Outlet for normal operation

Blue VLAN not affected
New Server Connection for Blue VLAN to F-Switches

F2 learns MAC of server’s red backup NIC

Outlet for normal operation
Configuration Options for Redundant NIC

VLAN Blue / IP NET Blue

**Type-A**

- Only one NIC active, the other NIC in standby (will take over in case of failure)
- 1 server IP address and 1 server MAC address, 1 Default Gateway

VLAN Red / IP NET Red

Type-C

- Both NIC’s active, load balancing possible
- 2 server IP addresses and 2 server MAC addresses, 2 Default Gateways, strategy which NIC to use is necessary

VLAN Blue / IP NET Blue

Type-B

- Both NIC’s active, load balancing possible
- 2 server IP addresses and 2 server MAC addresses, 1 Default Gateway, strategy which NIC to use is necessary

VLAN Blue / IP NET Blue

Type-D

- Both NIC’s active and connected to internal router, load balancing via routing protocol is possible
- 1 server IP addresses on the loopback interface, which connects the server to the internal router
Redundant NIC Type-A, Intel Teaming

Intel Teaming Basic AFT (Adapter Fault Tolerance) Feature;
one primary adapter, up to 7 secondary adapters
primary for transmitting and receiving all traffic from and to the unique server MAC address;
multicast/broadcast probes to ensure a secondary adapter is available, therefore all links are physically active and all secondary adapters have their own MAC address;
secondary will take over in case primary fails that means it starts sending traffic with the server MAC address and receiving traffic for the server MAC address;
STP must be switched off

Intel Teaming SFT (Switch Fault Tolerance) Feature;
one primary adapter, one secondary adapter
primary for transmitting and receiving all traffic from and for the unique server MAC address;
multicast/broadcast probes to ensure a secondary adapter is available, therefore all links are active and all secondary adapters have their own MAC address;
secondary will take over in case primary fails that means it starts sending traffic with the server MAC address and receiving traffic for the server MAC address;
STP must be switched on
Redundant NIC Type-A, Intel Teaming

Intel Teaming ALB (Adaptive Load balancing) Feature; one primary adapter, up to 7 secondary adapters primary for receiving all traffic to the unique server MAC address and unique server IP address; secondary are used for balancing the load for transmit traffic; all links are active and all secondary adapters have their own MAC address; secondary send with their own MAC address and will not answer ARP Requests to the server IP address; thus the server MAC address will not be seen on switch ports leading to secondary adapters (? Doing so will not solve the ARP cache problem of the client-PCs because every received Ethernet frame at the client-PC will refresh/change the ARP cache ?)

Intel Teaming RLB (Receiver Tolerance) Feature; same as ALB, but now secondary answer ARP requests based on an internal scheduling decision hence populating the ARP cache of different client-PCs with different MAC addresses for the same unique server IP address; Tricky procedure in case the server itself sends an ARP request for a client with its unique server MAC address -> client-PC ARP caches would be refreshed and traffic would be directed to the primary -> hence appropriate ARP replies must be sent out to correct ARP cache again
Redundant NIC Critical Aspect

In case of active / standby it is important that both sides (PC and the switch(es)) have the same sight who is active and who is standby (symmetric view)

switch sees this

server sees this

switch sees this

server sees this
Intel Advanced Network Services Software (ANS)

- **What is Intel ANS?**
  - Implemented as an intermediate driver in the servers driver stack
  - Windows and Linux supported
Teaming Features

• **Fault Tolerance**
  – 1 or more secondary adapter take over if primary fails

• **Link Aggregation**
  – Combine multiple adapters into a single channel
  – Bandwidth increase only available to multiple destination addresses
  – Must be supported by connected switch!

• **Load balancing**
  – Distribution of transmission and reception load among aggregates network adapters
  – Agent in ANS analyzes traffic and distributes the packets based on destination addresses
Teaming Modes

- **Adapter Fault Tolerance (AFT)**
  - 2-8 adapter supported
  - If primary fails -> secondary takes over
  - All adapters must be connected to same network

- **Switch Fault Tolerance (SFT)**
  - Failover relationship between 2 Adapters connected to different switches
  - STP must be enabled on the switches
  - STP must be disabled on connected Ports
Teaming Modes

• **Adaptive Load Balancing (ALB)**
  – Load balancing of transmit traffic
  – Receive Load Balancing (RLB) is advanced feature – enabled by default

• **Static Link Aggregation (SLA)**
  – IEEE 802.3ad static and dynamic mode
    • Needs compatible switch!
  – Intel Link Aggregation (LA), Cisco Fast EtherChannel (FEC), Gigabit EtherChannel (GEC) replaced by static link aggregation mode
  – 2-8 Adapters – all ports same speed
  – Incorporates AFT and ALB modes
## Teaming Features and Modes

<table>
<thead>
<tr>
<th>Features</th>
<th>Modes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AFT</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Fault Tolerance</td>
<td>X</td>
</tr>
<tr>
<td>Link Aggregation</td>
<td></td>
</tr>
<tr>
<td>Load Balancing</td>
<td></td>
</tr>
<tr>
<td>Layer 3 Address Aggregation</td>
<td></td>
</tr>
<tr>
<td>Layer 2 Address Aggregation</td>
<td></td>
</tr>
<tr>
<td>Mixed Speed Adapters</td>
<td>X</td>
</tr>
</tbody>
</table>

Details – How does it work

• How To Detect State And Health Of Adapters
  – Probe Packets
    • Adapters send and receive them to determine presence and state of other adapters
    • Either broadcast or multicast – configurable in software
  – Activity Based Tolerance
    • If probe packets are not used or do not reach their destination -> sensing activity on the line
  – Link Based Tolerance
    • Used if neither probe packets nor activity based tolerance are available or successful
Probe Packets Details

• 2 different types of user configurable probes
• Each member uses 2 flags – Send and Receive – to track status
• When adapter sends probe sets both flags to Pending state
• When packet is received by a member of same team – it sets its receive flag to ReceiveComplete and sets sending Flag to SendComplete
• If Primary Adapter is set to disabled -> Secondary Adapter takes this role – new Secondary will be elected
Server Load Balancing Methods

- **Adaptive Load Balancing (ALB)**
  - Receive Load Balancing (RLB) is a subset of ALB
  - Transmit Traffic balanced by Hash Table of last Octet of receivers IP address
  - New Dataflows are assigned to least loaded team member
  - After timeout of load bal. timer Dataflows are rebalanced
  - ALB without RLB uses Primary Team Members MAC in ARP Reply Packets
  - Send Packets include Team Members MAC as source
  - Failover: Secondary Adapter gets MAC of Primary
  - Do not Hotplug Primary and reuse somewhere else until Server Reboot
Receive Load Balancing (RLB)

• When receiving ARP Request -> Intel ANS answers with MAC Address of the port which is chosen to service this client
• Clients are allocated in a “Round-Robin” manor
• RLB client table is refreshed after ReceiveBalancing Interval
• OS ARP requests are send through primary port
  – Receive load collapses to primary
  – ANS sends gratuitous ARP to all clients in hash to restart RLB
Static Link Aggregation (SLA)

- All Ports share same MAC Address
- For the switch this is a single link
- No designated primary port in the team
- Links must be same speed
- Switch handles receive load balancing
802.3ad Dynamic Mode

- All members share same MAC
- Switch ports must use LACP protocol
- Switch communicates with Intel ANS to add or remove members of team
- No designated primary – but first teamed port is initiator to switch
- Removal of Initiator could lead to packet loss
- To avoid this -> preconfigure the switch ports for added or removed members
## Teaming Modes Comparison

<table>
<thead>
<tr>
<th>Function</th>
<th>Intel (AFT, SFT, ALD)</th>
<th>Intel (SLA)</th>
<th>Intel (802.3ad)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of NICs per team</td>
<td>8 (2 for SFT)</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>NIC Fault Tolerance</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Switch Fault Tolerance</td>
<td>2</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Tx Load Balance</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Rx Load Balance</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Requires compatible switch</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Heartbeats to check connectivity</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>NICs with different media</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>NICs with different speeds</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Load balances TCP/IP</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Load balances other protocols</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Same MAC address for all team members</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Same IP address for all team members</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Load balancing by IP address</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Load balancing by MAC address</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>802.1Q tagged VLANs</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Untagged VLANs</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
 Agenda

• Spanning Tree Protocol (STP)
• Rapid Spanning Tree Protocol (RSTP)
• Cisco PVST, PVST+
• Multiple Spanning Tree Protocol (MSTP)
• **Spanning Tree Tuning – LAN Design**
  – Design Considerations
  – Design Solution – Best Practices
  – Failover Handling
  – Advanced Techniques - Teaming
  – **LAN – WAN Interconnection**
LAN – WAN Interconnection
VLAN Interconnection

• Now let us look to Layer 3 (IP)

• We need routers
  – For connecting the two VLANs
  – For connecting the LAN infrastructure of a site to the WAN infrastructure

• Be very careful to differentiate between
  – L1 look of your network
  – L2 look of your network (VLAN, STP)
  – L3 look of your network (IP, ARP)
Router Connections

Two physical interfaces:
- one for VLAN Red
- one for VLAN Blue
One physical interface:
Router “on a stick”
with 802.1Q Tagging
Layer 3 View Dual Homed Server

R1
VLAN Red
IP Subnet Red
VLAN Blue
IP Subnet Blue

R2
Layer 3 View Dual Homed Server
NET-ID, IP Addressing

WAN-Domain

R1
IP: 192.168.1.251/24
Mac: R1_Red

IP: 192.168.2.253/24
Mac: R1_Blue

IP NET-ID Red: 192.168.1.0/24

VLAN Red
IP: 192.168.2.253/24
Mac: R2_Blue

R2
IP: 192.168.2.254/24
Mac: R2_Blue

IP: 192.168.1.252/24
Mac: R2_Red

IP NET-ID Blue: 192.168.2.0/24

VLAN Blue
IP: 192.168.2.253/24
Mac: R1_Blue

IP: 192.168.2.1/24
Mac: S1_Blue

IP: 192.168.1.1/24
Mac: S1_Red

IP: 192.168.2.253/24
Mac: R1_Blue
Layer 3 View Dual Homed Server
Routing Table, ARP Cache

Routing Table R1
- 192.168.1.0/24 Int Vlan Red
- 192.168.2.0/24 Int Vlan Blue
- ...... IP Next Hop
- ...... IP Next Hop

ARP-Cache R1 Red
- 192.168.1.1 MAC S1_Red
- ...... ......

IP NET Red: 192.168.1.0/24
- IP: 192.168.1.251/24 Mac: R1_Red

Routing Table Client PC Red
- 192.168.1.1 MAC S1_Red
- 192.168.1.0/24 Int Vlan Red
- 0.0.0.0/0 192.168.1.251

ARP-Cache Client PC Red
- 192.168.1.1 MAC S1_Red
- 192.168.1.251 MAC R1_Red

IP NET Blue: 192.168.2.0/24
- IP: 192.168.1.1/24 Mac: S1_red
Layer 3 Load Balancing to WAN Using HSRP

WAN-Domain

Left

R1

Red VLAN: VR_GR_1
SR_GR_2

Blue VLAN: VR_GR_3
SR_GR_4

Default-GW: VR_GR_1

Left Domain

Default-GW: VR_GR_3

Right Domain

Default-GW: VR_GR_2

Right

R2

Red VLAN: VR_GR_2
SR_GR_1

Blue VLAN: VR_GR_4
SR_GR_3

Default-GW: VR_GR_2

Default-GW: VR_GR_4

Default-GW: VR_GR_3

Default-GW: VR_GR_4

© 2016, D.I. Lindner / D.I. Haas

Spanning-Tree Details, v6.0

158
Layer 3 Simple Backup to WAN Using HSRP

WAN-Domain

R1
Red VLAN: VR_GR_1
Blue VLAN: VR_GR_4
Default-GW: VR_GR_1

R2
Red VLAN: SR_GR_1
Blue VLAN: SR_GR_4
Default-GW: VR_GR_1

Default-GW: VR_GR_1

Default-GW: VR_GR_1

Default-GW: VR_GR_4

Default-GW: VR_GR_4

Default-GW: VR_GR_4

Default-GW: VR_GR_4
Layer 3 Simple Backup to WAN Using HSRP

R1

WAN-Domain

R2

Red VLAN: VR_GR_1

Blue VLAN: VR_GR_4

Default-GW: VR_GR_1

Default-GW: VR_GR_4

Default-GW: VR_GR_4

Default-GW: VR_GR_4

Default-GW: VR_GR_1

Default-GW: VR_GR_4

Default-GW: VR_GR_4
Layer 3 Simple Backup & Load Balancing WAN Using HSRP

R1
- Red VLAN: VR_GR_1
- Default-GW: VR_GR_1

R2
- Blue VLAN: SR_GR_4
- Default-GW: VR_GR_4

WAN-Domain

© 2016, D.I. Lindner / D.I. Haas
Spanning-Tree Details, v6.0