(Rapid) Spanning Tree Protocol

(R)STP

Karst Koymans

Informatics Institute
University of Amsterdam

(version 19.4, 2019/11/20 11:51:56 UTC)

Friday, November 15, 2019

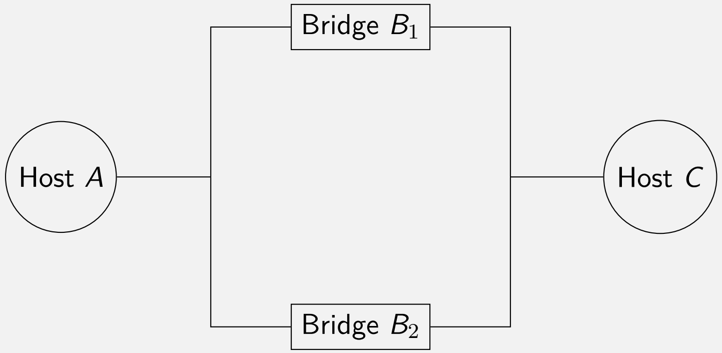
Table of Contents

Bridge loops
- Two bridges
- Three bridges

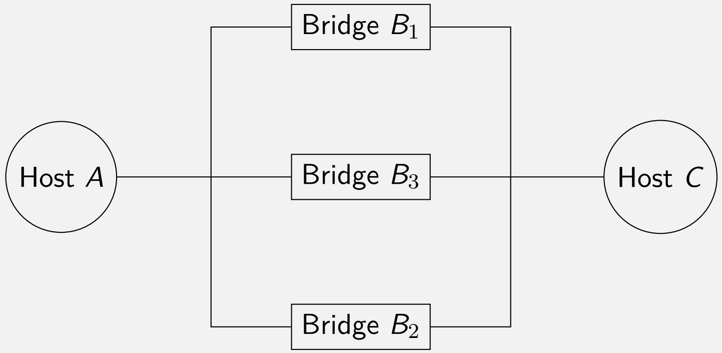
Network representation
- Incomplete representations
- Bridges as nodes
- LAN segments as nodes
- Complete representations

Spanning Tree
- The basic protocol
- Example (uniform cost)
- Parameters and timing
- Packet/Frame formats
- Enhancements

A simple bridge loop



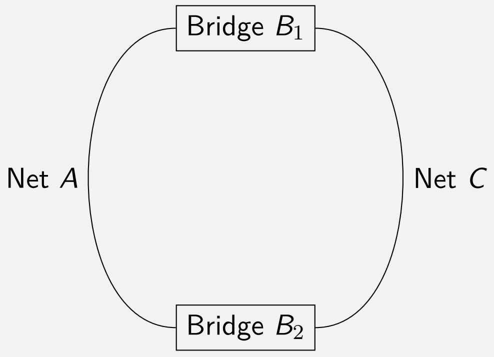
An even worse bridge loop



Naive graph representation

- Focus on the bridges
  - Nodes represent bridges
  - Edges represent network LAN segments
- Focus on the LAN segments
  - Nodes represent network LAN segments
    - In pictures we will use the phrase Nets for LAN segments
  - Edges represent bridges

Bridges as nodes (1)



Bridge B_1

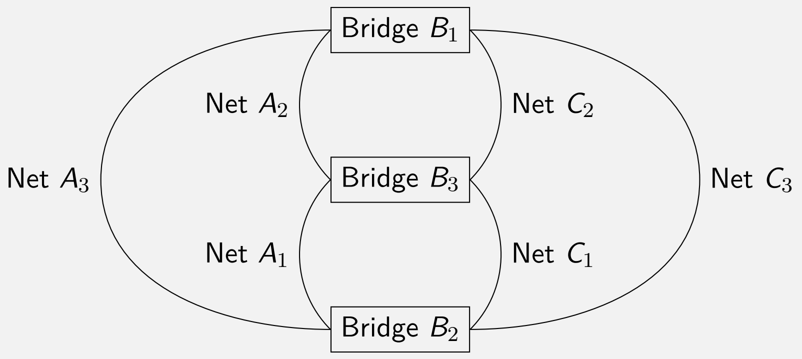
Net A

Net C

Bridge B_2

In this case this representation would be adequate

Bridges as nodes (2)



Bridge B_1

Net A_2

Net C_2

Net A_3

Bridge B_3

Net C_3

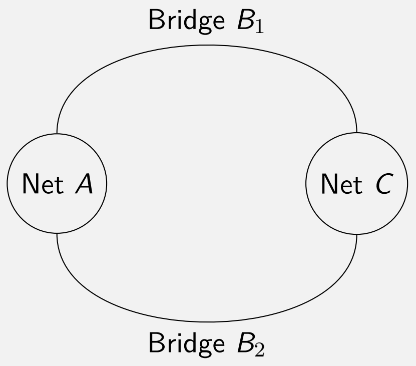
Net A_1

Net C_1

Bridge B_2

In this case the representation is not so adequate
One LAN segment is split into or represented by three edges

LAN segments as nodes (1)



Bridge B_1

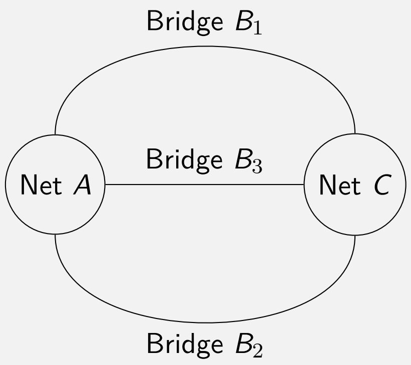
Net A

Net C

Bridge B_2

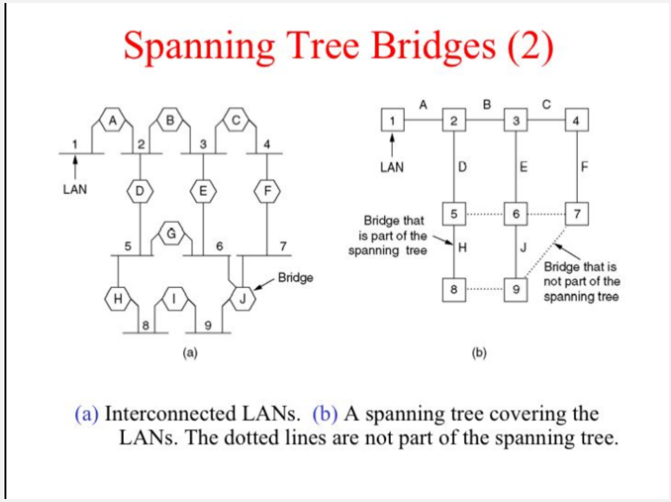
In this case again this representation would be adequate

LAN segments as nodes (2)



Also this case would be fine (for bridges with 2 ports)

Another incomplete representation



Spanning Tree Bridges (2)

A
B
C
D
E
F
G
H
I
J
1
2
3
4
5
6
7
8
9
LAN
Bridge
(a)
Bridge that is part of the spanning tree
Bridge that is not part of the spanning tree
(b)

(a) Interconnected LANs. (b) A spanning tree covering the LANs. The dotted lines are not part of the spanning tree.

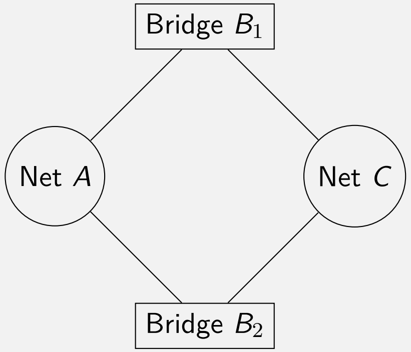
The problem is with bridge J having 3 ports

Source: “Computer Networks”, 4th edition, Tanenbaum (repaired in 5th edition)

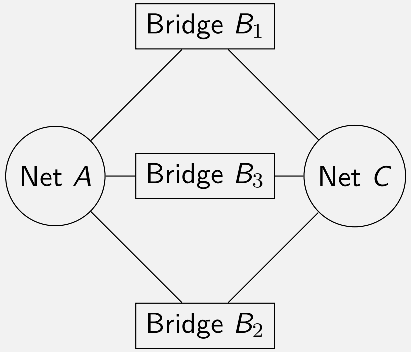
Complete graph representation

- Bridges and network LAN segments are both represented as nodes
- Interfaces of devices to network LAN segments are represented as edges

Two bridges, two LAN segments



Three bridges, two LAN segments



Bridge B_1

Net A

Bridge B_3

Net C

Bridge B_2

Another application for a bipartite graph

Spanning Tree Protocol Algorhyme

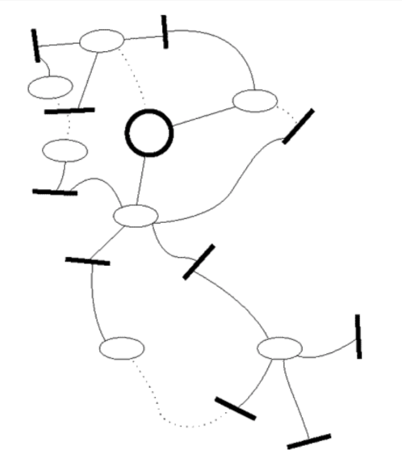
I think that I shall never see
A graph more lovely than a tree.
A tree whose crucial property
Is loop-free connectivity.
A tree that must be sure to span
So packets can reach every LAN.
First, the root must be selected.
By ID, it is elected.
Least-cost paths from root are traced.
In the tree, these paths are placed.
A mesh is made by folks like me,
Then bridges find a spanning tree.

—*Radia Perlman*

Spanning Tree Protocol

- Eliminate edges until the result is loop free
  - Actually add edges without creating loops
- This transforms the graph into a tree
- Changes in the topology cause the tree to change
- A root bridge is elected as the root of the tree

Example spanning tree



- Where are the networks?
  - Bold lines and circle
- Where are the bridges?
  - The thin oval shapes
- Where is the root?
  - Centre or bottom right
  - We assume a uniform cost of 1 here

Source: Interconnections, second edition, by Radia Perlman

Configuration messages

- Every bridge has an ID based on
  - A configurable priority (2 bytes)
  - One of its MAC addresses (6 bytes)
- A bridge transmits on all attached LAN segments
  - The ID of the currently perceived root (initially own ID)
  - Cost of the best path to the root (initially 0)
  - Its own ID (as first tie breaker)
  - The port ID of the transmission (as a second tie breaker)

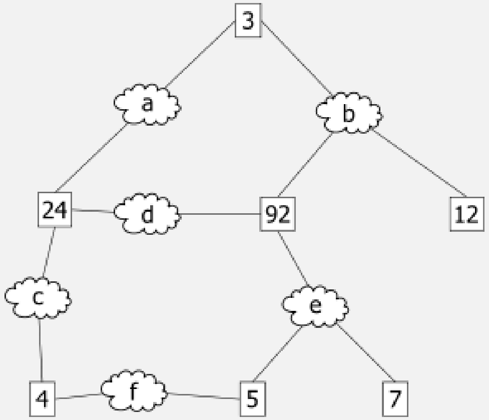
Designated bridge and port for a LAN segment

- Every LAN segment "chooses" the best route (path) towards the root using the following criteria in order
  - 1. Lower advertised root ID
  - 2. Lower advertised cost to root
  - 3. Lower transmitting bridge ID
  - 4. Lower port ID
- It would be reasonable to call this best port the LAN segment's root port but LAN segments are not able to "choose" anything ...
  - The bridge advertising the best route (path) becomes that LAN segment's designated bridge and the corresponding bridge port is called the bridge's designated port for the LAN segment

Root port for a bridge

- Every bridge except the root itself chooses the best route (path) towards the root advertised by attached networks (through their designated bridges) using the same criteria as before
  - 1. Lower advertised root ID
  - 2. Lower advertised cost to root
  - 3. Lower transmitting bridge ID
  - 4. Lower port ID
- The port corresponding to the best route (path) is called the bridge's root port

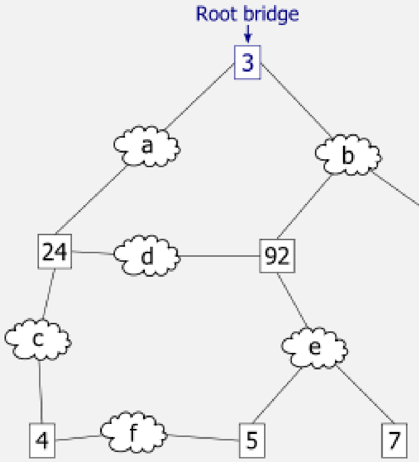
Example STP protocol execution



A. An example network. The numbered boxes represent bridges (the number represents the bridge ID). The lettered clouds represent network segments.

Source: Wikipedia (retrieved 20150218)
(`http://en.wikipedia.org/wiki/Spanning_tree_protocol`)

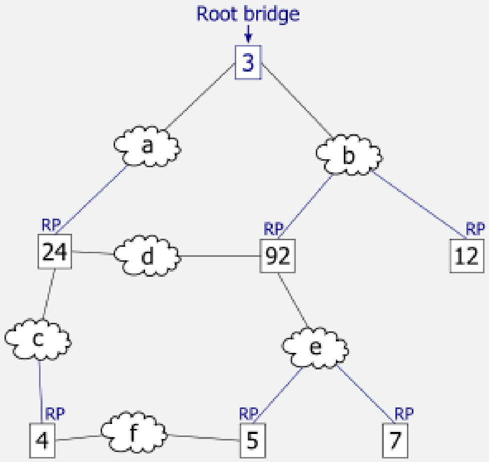
Example STP protocol execution



B. The smallest bridge ID is 3. Therefore, bridge 3 is the root bridge.

Source: Wikipedia (retrieved 20150218)
(`http://en.wikipedia.org/wiki/Spanning_tree_protocol`)

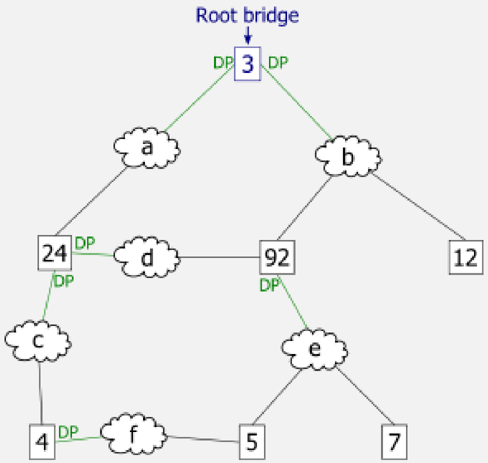
Example STP protocol execution



C. Assuming that the cost of traversing any network segment is 1, the least cost path from bridge 4 to the root bridge goes through network segment c. Therefore, the root port for bridge 4 is the one on network segment c.

Source: Wikipedia (retrieved 20150218)
(`http://en.wikipedia.org/wiki/Spanning_tree_protocol`)

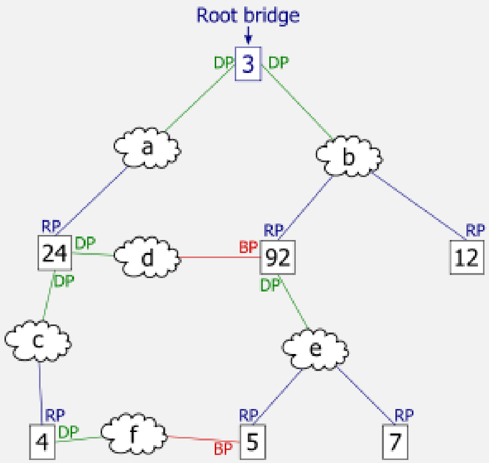
Example STP protocol execution



D. The least cost path to the root from network segment e goes through bridge 92. Therefore the designated port for network segment e is the port that connects bridge 92 to network segment e.

Source: Wikipedia (retrieved 20150218)
(`http://en.wikipedia.org/wiki/Spanning_tree_protocol`)

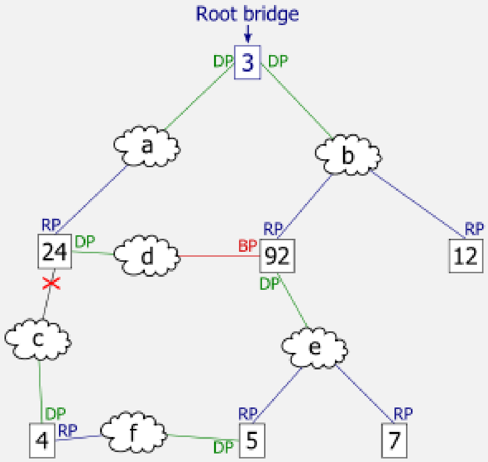
Example STP protocol execution



E. This diagram illustrates all port states as computed by the spanning tree algorithm. Any active port that is not a root port or a designated port is a blocked port.

Source: Wikipedia (retrieved 20150218)
(`http://en.wikipedia.org/wiki/Spanning_tree_protocol`)

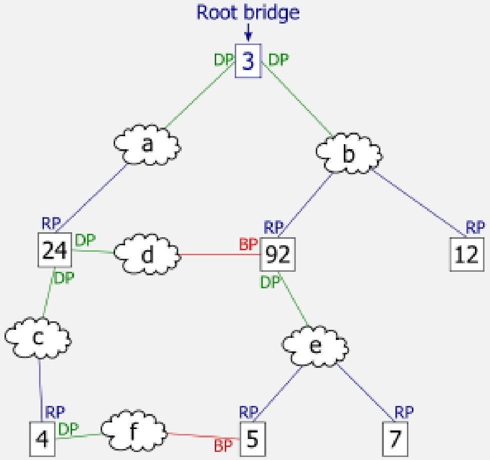
Example STP protocol execution



F. After link failure the spanning tree algorithm computes and spans new least-cost tree.

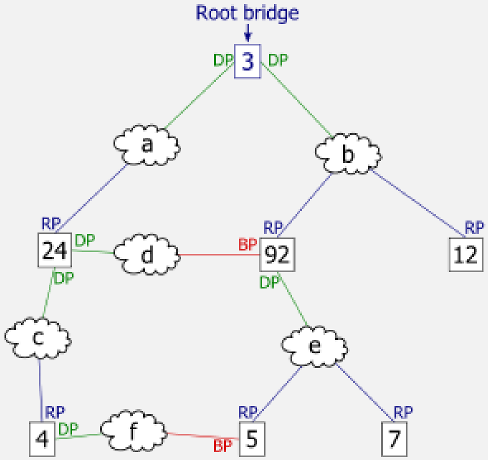
Source: Wikipedia (retrieved 20150218)
(`http://en.wikipedia.org/wiki/Spanning_tree_protocol`)

Example STP protocol execution



G. What happens when a new link comes up between bridge 92 and LAN segment f?

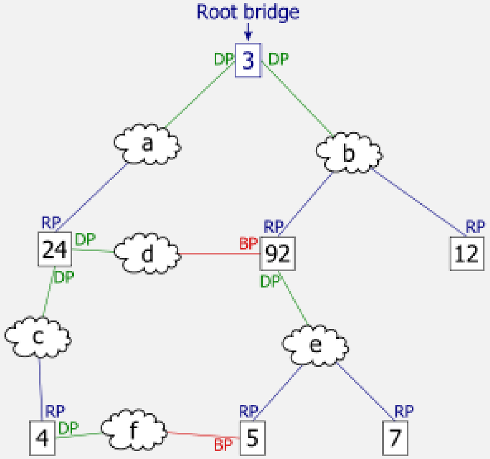
Example STP protocol execution



G. What happens when a new link comes up between bridge 92 and LAN segment f?

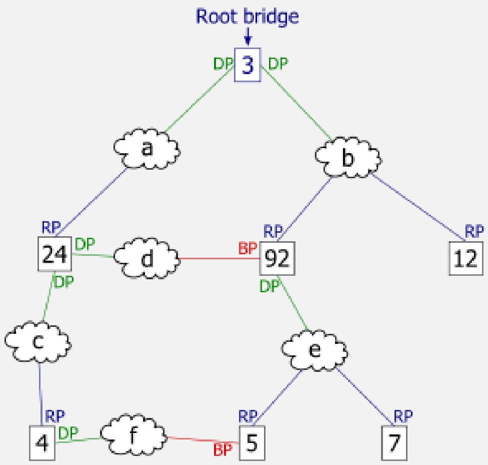
- DP between 92 and f
- DP changes into BP between 4 and f
- BP possibly exchanges with RP for 5, depending on port ids on 92

Example STP protocol execution



H. What happens when a new link comes up between bridge 24 and LAN segment f?

Example STP protocol execution



H. What happens when a new link comes up between bridge 24 and LAN segment f?

- DP between 24 and f
- RP exchanges with BP for 5
- DP becomes BP for 4 and possibly exchanges with RP, depending on port ids on 24

Timing parameters

- Message Age
  - Increased on each timer tick (1/256 second)
- Max Age
  - Discard configuration messages that are too old
- Hello Time
  - Time between two configuration messages
- Forward Delay
  - Half of the delay before transitioning from blocking to forwarding

Listening, learning and forwarding

- Every bridge waits for some period (twice the forward delay) to let the configuration messages spread and the topology converge,
  and in the mean time ...
  - ...it does not forward frames (very important)
  - ...it listens to neighbouring bridges in the first half
  - ...it learns the location of MAC addresses in the second half
- After this period it starts forwarding data frames
  - The root port and the designated ports are put into a forwarding state
  - All other ports are kept or put in a blocking state

Station learning and caching

- Bridges keep track of and cache where individual stations are located with respect to the current spanning tree
- Usually (when the topology is stable) there is a long caching time
- A short caching time is used when the topology of the spanning tree
  has changed anywhere in the graph
  - But how do bridges know this?
  - The change could have happened somewhere far away

Topology change mechanism

- Suppose any bridge notices a topology change (port up or down)
- This bridge (and, recursively, upstream bridges) send Topology Change Notification messages on their root ports
  - Upstream bridges set the TCA flag in their next configuration message downstream
- Finally such a TCN message reaches the root bridge
- The root bridge sets the TC flag in its configuration messages for a period of “forward delay + max age”
- If a bridge sees the TC flag it uses the short station cache timer
  - which is equal to the forward delay
  - until the topology has stabilized again

BPDU

- Bridge Protocol Data Unit, using 802.3 frame[1] format
- Configuration messages (hello)
  - DSAP = SSAP = 01000010 (= 0x42 :-))
  - Destination 01:80:C2:00:00:00
    - This is a group (multicast) address, but ...
    - ... it is not forwarded to other LAN segments
- Topology changed messages
  - To support “re-learning” after tree change

[1] A BPDU itself is called a packet; however some people call it a frame

Configuration (Hello) BPDU packet/frame format

<table>
<tr><th>0</th><th></th><th>7</th><th>8 … 15</th><th>16</th><th>31</th></tr>
<tr><td colspan="4">Protocol Identifier</td><td>Version</td><td>Message Type</td></tr>
<tr><td>T
C
A</td><td>RSTP flags</td><td>T
C</td><td colspan="3">These bits are absent</td></tr>
<tr><td colspan="6">Root ID (8 bytes)</td></tr>
<tr><td colspan="6">Cost of path to root</td></tr>
<tr><td colspan="6">Bridge ID (8 bytes)</td></tr>
<tr><td colspan="4">Port ID</td><td colspan="2">These bits are absent</td></tr>
<tr><td colspan="4">Message Age</td><td colspan="2">Max Age</td></tr>
<tr><td colspan="4">Hello Time</td><td colspan="2">Forward Delay</td></tr>
</table>

Topology Change Notification BPDU packet format

<table>
<tr><th>0 … 7</th><th>8 … 15</th><th>16</th><th>31</th></tr>
<tr><td colspan="2">Protocol Identifier</td><td>Version</td><td>Message Type</td></tr>
</table>

BPDU packet fields (type)

BPDU packet fields (1)

Protocol Identifier	0
Version	0(STP), 2(RSTP)
Message Type	0(Hello), 128(TCN), 2(RSTP)
TCA	Topology Change Ack(STP), 0(RSTP)
RSTP Flags	Proposal, Agreement, ...(RSTP)
TC	Topology Change

BPDU packet fields (metric data)

BPDU packet fields (2)

Root ID	Root bridge
Cost	Cost of path to root bridge
Bridge ID	Bridge transmitting BPDU
Port ID	Port on which BPDU is transmitted

BPDU packet fields (parameters)

BPDU packet fields (3)

Message Age	Age of BPDU information
Max Age	Typically 20 seconds (minimally 6 seconds)
Hello Time	Typically 2 seconds
Forward Delay	Typically 15 seconds (minimally 4 seconds)

Rapid spanning tree (802.1w, now part of 802.1D)

- Backward compatible with STP
- Has special RSTP BPDUs
- Uses incoming BPDUs as keep-alive
- Does not use the TCA flag at all
- Introduces a proposal and agreement flag in order to enable forwarding mode as early as possible
- Starts forwarding on root ports immediately and on point to point designated ports

VLANs and STP

- Global STP valid for all VLANs
- Running STP separately for every VLAN
  - PVST(+) (Per-VLAN Spanning Tree (Plus); Cisco proprietary)
  - VSTP (VLAN Spanning Tree Protocol; Juniper proprietary)
  - MSTP (Multiple Spanning Tree Protocol)
- Side note: security measures
  - Disable STP on “host” ports
  - Disable tagged traffic on “host” ports

Multiple Spanning Tree Protocol (802.1s, now integrated into 802.1Q)

- Divides a LAN into multiple regions
- Creates a MSTI (multiple spanning tree instance) inside each region separately for each VLAN
- Defines a global spanning tree between regions each treated as a single “pseudo” or “virtual” bridge by creating a CIST (Common Internal Spanning Tree) consisting of
  - a CST (Common Spanning Tree) between regions
  - an IST (Internal Spanning Tree) inside a region