



Spanning tree protocol

Mario Baldi

Politecnico di Torino

<http://staff.polito.it/mario.baldi>

Pietro Nicoletti

Studio Reti

<http://www.studioreti.it>

Based on chapter 4 of:

M. Baldi, P. Nicoletti, "Switched LAN", McGraw-Hill, 2002, ISBN 88-386-3426-2

Copyright note

These slides are protected by copyright and international treaties. The title and the copyrights concerning the slides (inclusive, but not only, every image, photograph, animation, video, audio, music and text) are the author's (see Page 1) property.

The slides can be copied and used by research institutes, schools and universities affiliated to the Ministry of Public Instruction and the Ministry of University and Scientific Research and Technology, for institutional purpose, not for profit. In this case there is not requested any authorization.

Any other complete or partial use or reproduction (inclusive, but not only, reproduction on discs, networks and printers) is forbidden without written authorization of the author in advance.

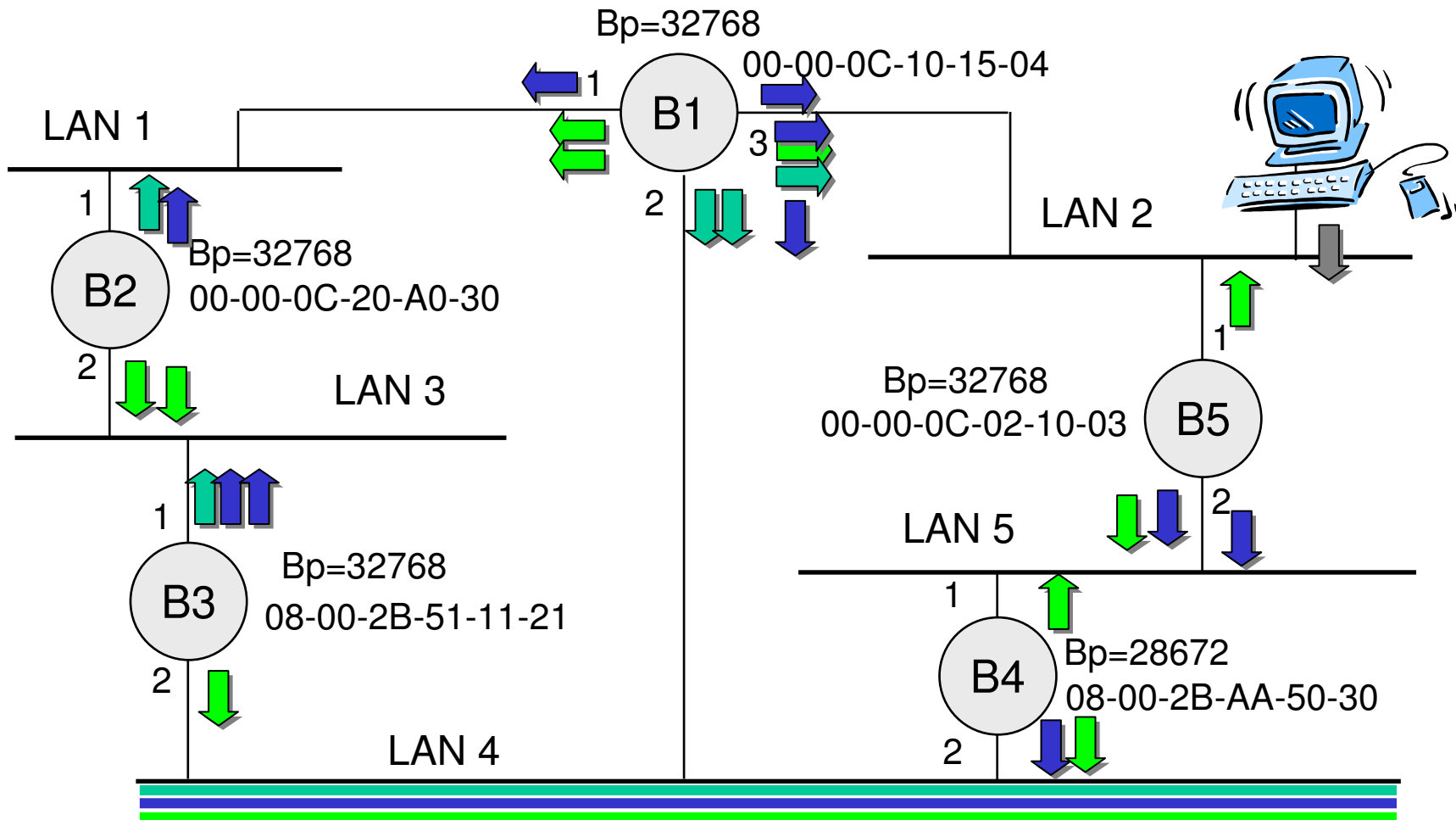
The information contained in these slides are believed correct at the moment of publication. They are supplied only for didactic purpose and not to be used for installation-projects, products, networks etc. However, there might be changes without notice. The authors are not responsible for the content of the slides.

In any case there can not be declared conformity with the information contained in these slides.

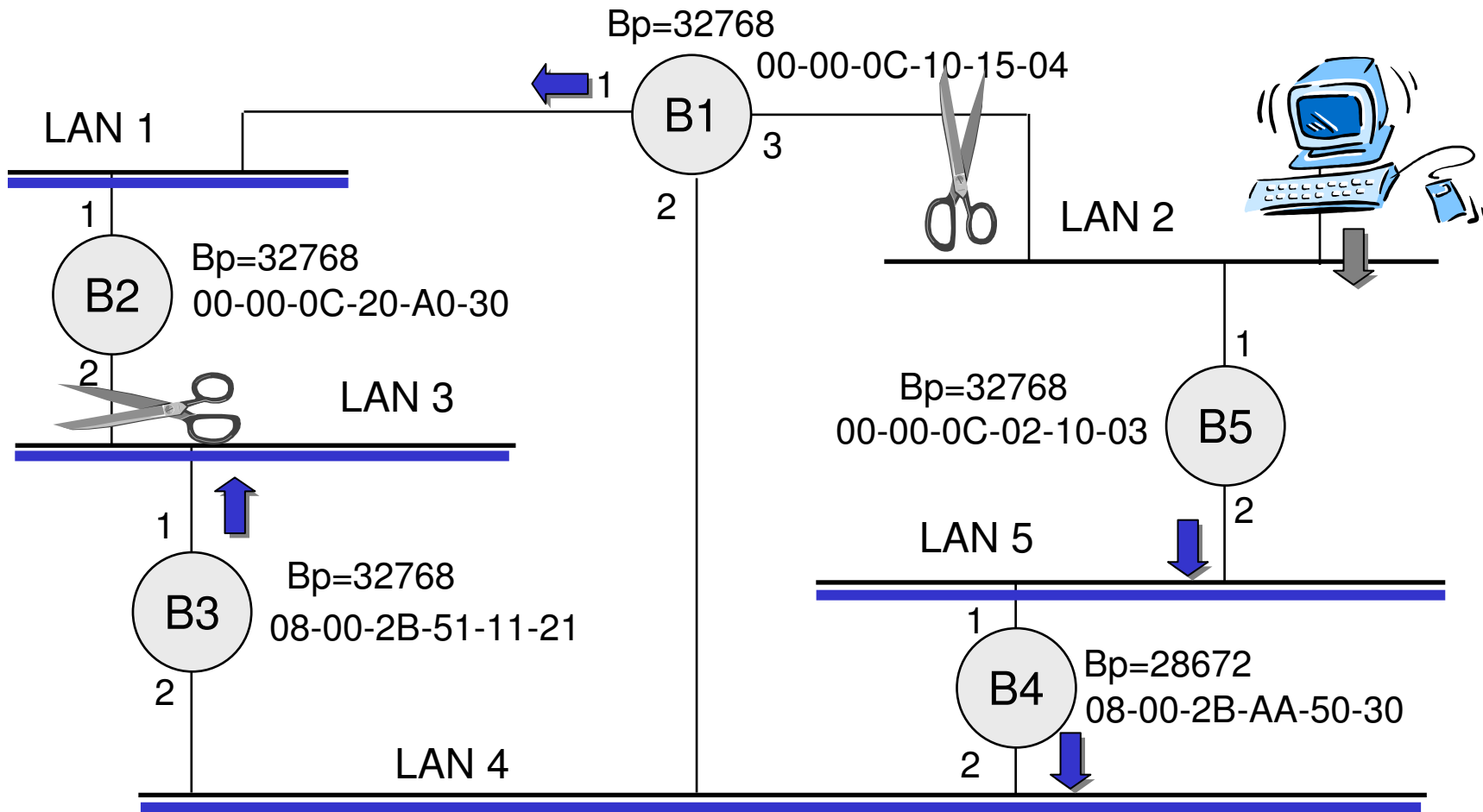
In any case this note of copyright may never be removed and must be written also in case of partial use.

Loops on the network

Broadcast are not filtered (quick network saturation) : **broadcast storm**



Solution: avoid loops



Spanning Tree Protocol (STP)

STP is executed by the Spanning Tree procedure

- STP is defined by IEEE 802.1D

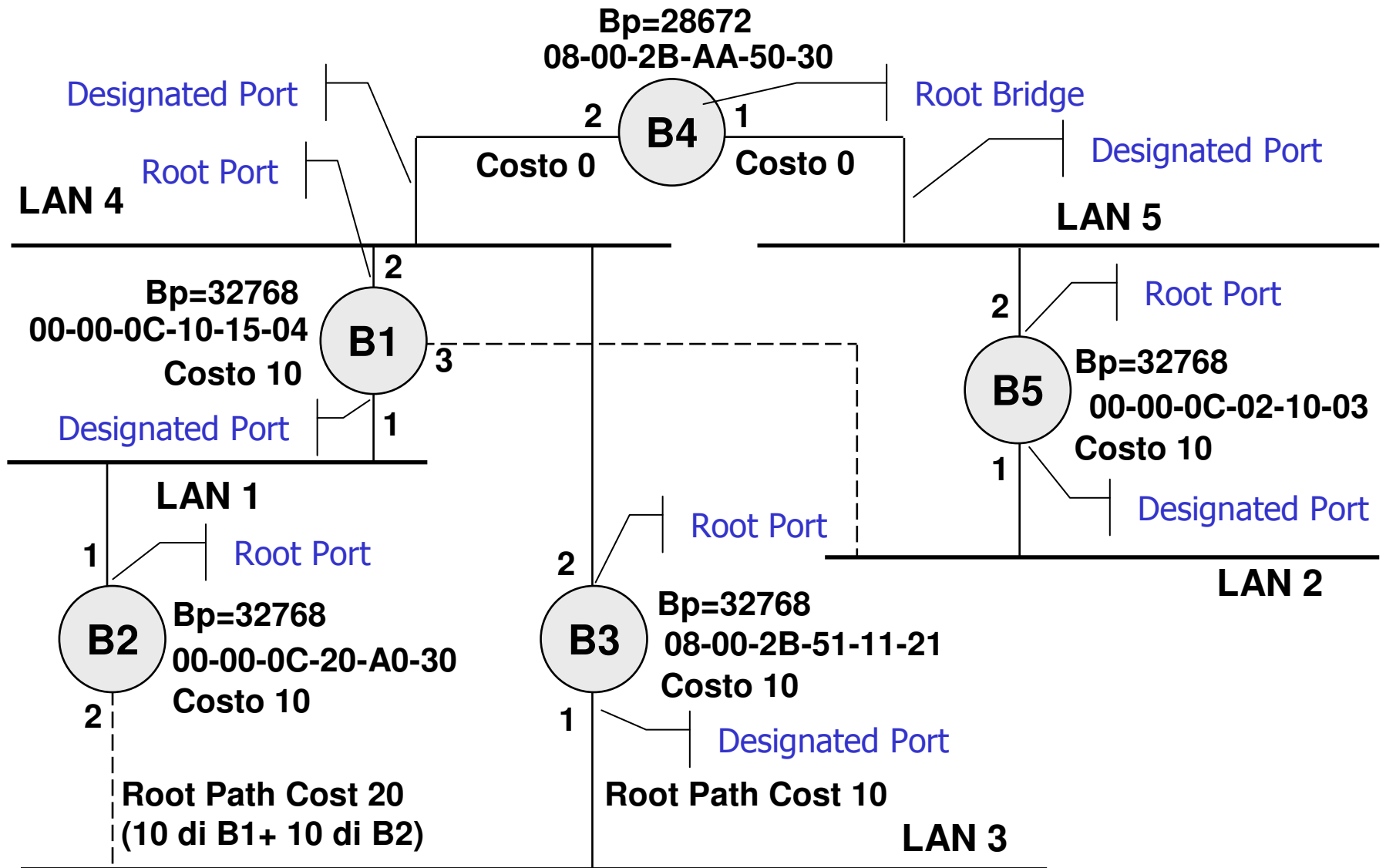
STP turns a meshed network into a tree

- STP is used to avoid loops

The spanning tree algorithm operates is made up of the following steps:

- **Root Bridge** election
 - Root of the tree (spanning tree) under construction
- **Root port** selection
 - A port of each bridge is used to reach the root Bridge
- **Designated port** selection
 - One of the ports connected to each LAN is used to receive and forward packets

Spanning tree protocol result





For a better comprehension... a little poetry

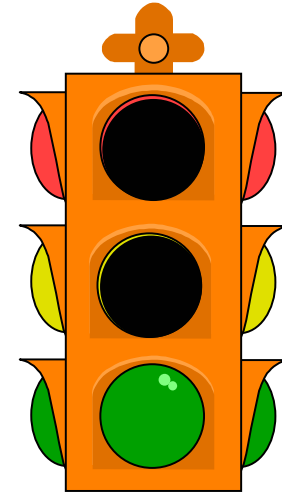
I think that I shall never see
A graph more lovely than a tree.
A tree whose crucial property
Is loop-free connectivity.
A tree which must be sure to span
So packets can reach every LAN.
First the Root must be selected
By ID it is elected.
Least cost paths from Root are traced
In the tree these paths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

[Radia Perlman]

States of the ports

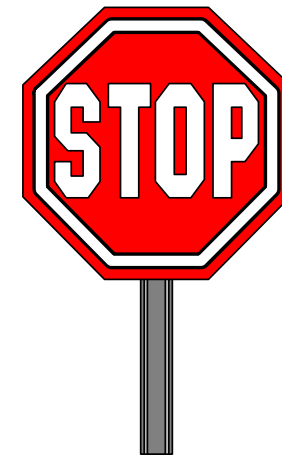
■ *Forwarding*

- Root port and designated port
- The port is used to forward packets
- Packets received by the port are processed and, if it is the case, they are forwarded by the Bridge



■ *Blocking*

- All the remaining ports
- The packets received by the port are not forwarded
- The packets received are discharged



Bridge Protocol Data Unit (BPDU)

- BPDUs are sent periodically by each Bridge to a predefined multicast address
- **Configuration** BPDU
- **Topology Change Notification** BPDU

Dest. Addr.	Source Addr.	Length	DSAP	SSAP	Control	BPDU	
Multicast 01-80-C2 00-00-00	Singlecast Indirizzo Bridge	XY	042H	042H	XID	Configuration BPDU or Topology Change Notification BPDU	FCS

BPDU: Bridge Protocol Data Unit
DSAP: Destination Service Access Point
SSAP: Source Service Access Point

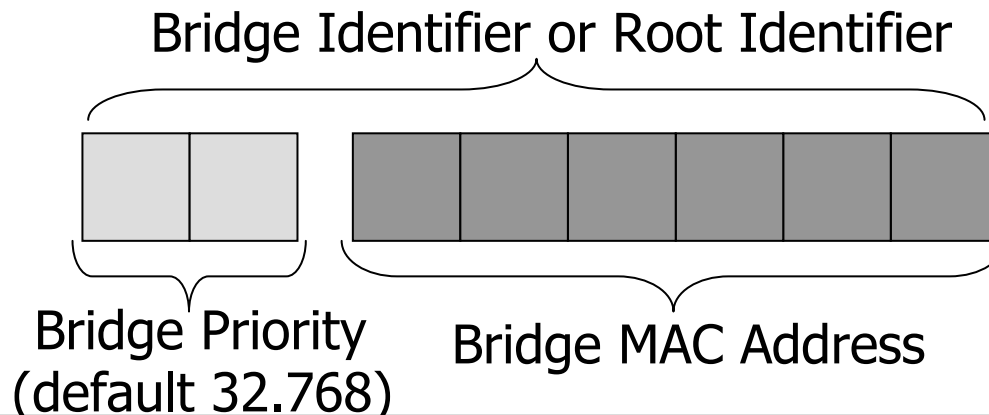
Byte 1÷2	Protocol Identifier:	00-00
3	Protocol Version Identifier:	00
4	BPDU Type:	00
5	TC	Flags
		TCA
6÷13	Root Identifier	
	first 2 bytes	= Bridge Priority
	last 6 bytes	= MAC address of the Root Bridge
14÷17	Root Path Cost	
	Bridge Identifier	
18÷25	first 2 bytes	= Bridge Priority
	last 6 bytes	= MAC address of the Bridge that sends BPDU
26÷27	Port Identifier	
	first byte	= Port Priority
	second byte	= port number
28÷29	Message Age	
30÷31	Max Age	
32÷33	Hello Time	
34÷35	Forward Delay	

Configuration BPDU

- **Root Identifier**
 - Bridge identifier of root Bridge
- **Root Path Cost**
 - Cost to reach the Bridge that created the Configuration BPDU on the path used by the message
- **Bridge Identifier**
 - Identifier of the Bridge that sent the Configuration BPDU
- **Port Identifier**
 - Identifier of the port belonging to the Bridge that created the Configuration BPDU

Bridge Identifier

- Each Bridge has a MAC address for each interface
 - One of them is chosen to create the bridge identifier
- A priority is assigned to each Bridge: **Bridge priority**
 - Default value to guarantee "plug&play" operation
- The Bridge with the lowest identifier is chosen as root
 - The lowest bridge priority determines root bridge election
 - If Bridge priority is the same in every bridge the lowest MAC address determines the root bridge election



Root Bridge Election

- At first each Bridge assumes it is the root Bridge
 - Each Bridge places its Bridge identifier in the root identifier field
 - Each Bridge sends Configuration BPDUs with a `hello time` periodicity
 - Default: 2 seconds
- Each Bridge compares its bridge identifier with the root identifier field of the Configuration BPDUs received
 - If it is lower, the Bridge keeps sending Configuration BPDUs
 - Otherwise, the Bridge places its Bridge identifier in the root identifier field of the candidate to become the root Bridge
- In the end all the Configuration BPDUs contain the Bridge Identifier of the root Bridge

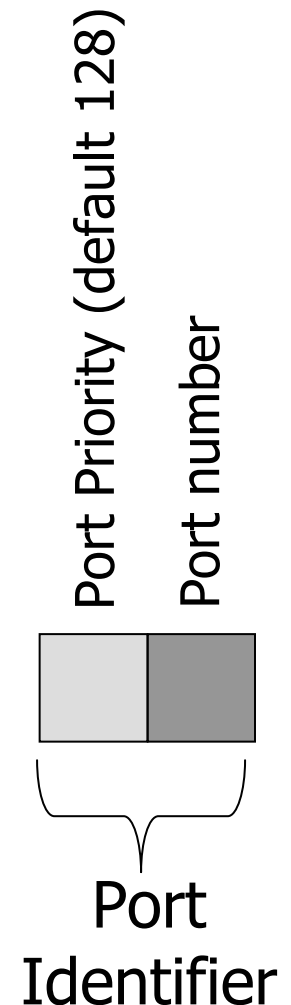
Root Port Selection

- Port on the path with the lowest cost up to the root Bridge
 - This port receives Configuration BPDUs sent by the root Bridge
 - This port forwards frames up to the root Bridge
- Every port has a related cost
- Configuration BPDUs store the cost of the traversed path
 - `Root path cost` field
 - The cost related to the receiving port is added to the `root path cost` by each traversed Bridge
- The port chosen as root port is the one that receives Configuration BPDUs with the highest priority
- The Configuration BPDUs received by the root port are forwarded on the other ports

Priority of Configuration BPDUs

A Configuration BPDU has higher priority than another one if it meets the following requirements in this order:

- The value in the `root path cost` field is lower
 - The value is updated by adding the cost (path cost) related to the port where the BPDU has been received
- The value in the `bridge identifier` field is lower
- The value in the `port identifier` field is lower
- The ***port identifier*** value related to the receiving port is lower



Designated port selection

- Many copies of each Configuration BPDUs arrive on a LAN connected to more than one non root port
 - The BPDUs took different paths from the root Bridge to the LAN
- A bridge connected to a LAN by a non root port receives Configuration BPDUs forwarded by other bridges
- A port is chosen as designated if the Configuration BPDUs forwarded have a higher priority than the BPDUs received
 - All the remaining ports are set in the blocking state
 - Only the designated port sends BPDUs to the LAN

STP Configuration parameters

- The bridges are plug&play
 - They may work with default configuration parameters
- Bridge priority
 - Range: 0 - 61440
 - Default/recommended: 32768
 - Suggested increment (IEEE 802.1t): 4096
- Port priority
 - Range: 0 - 240
 - Default/recommended: 128
 - Suggested increment (IEEE 802.1t): 16
- Port Path cost
 - Range: 0 - 65535
 - Recommended (IEEE 802.1D): $1000/(\text{Speed in Mb/s})$

Path cost recommended by IEEE 802.1D rev1998

Port speed	Recommended value	Recommended range values	Accepted range values
4Mb/s	250	100 - 1000	1 - 65535
10 Mb/s	100	50 - 600	1 - 65535
16 Mb/s	62	40-400	1 - 65535
100 Mb/s	19	10 - 60	1 - 65535
1 Gb/s	4	3 - 10	1 - 65535
10 Gb/s	2	1 - 5	1 - 65535

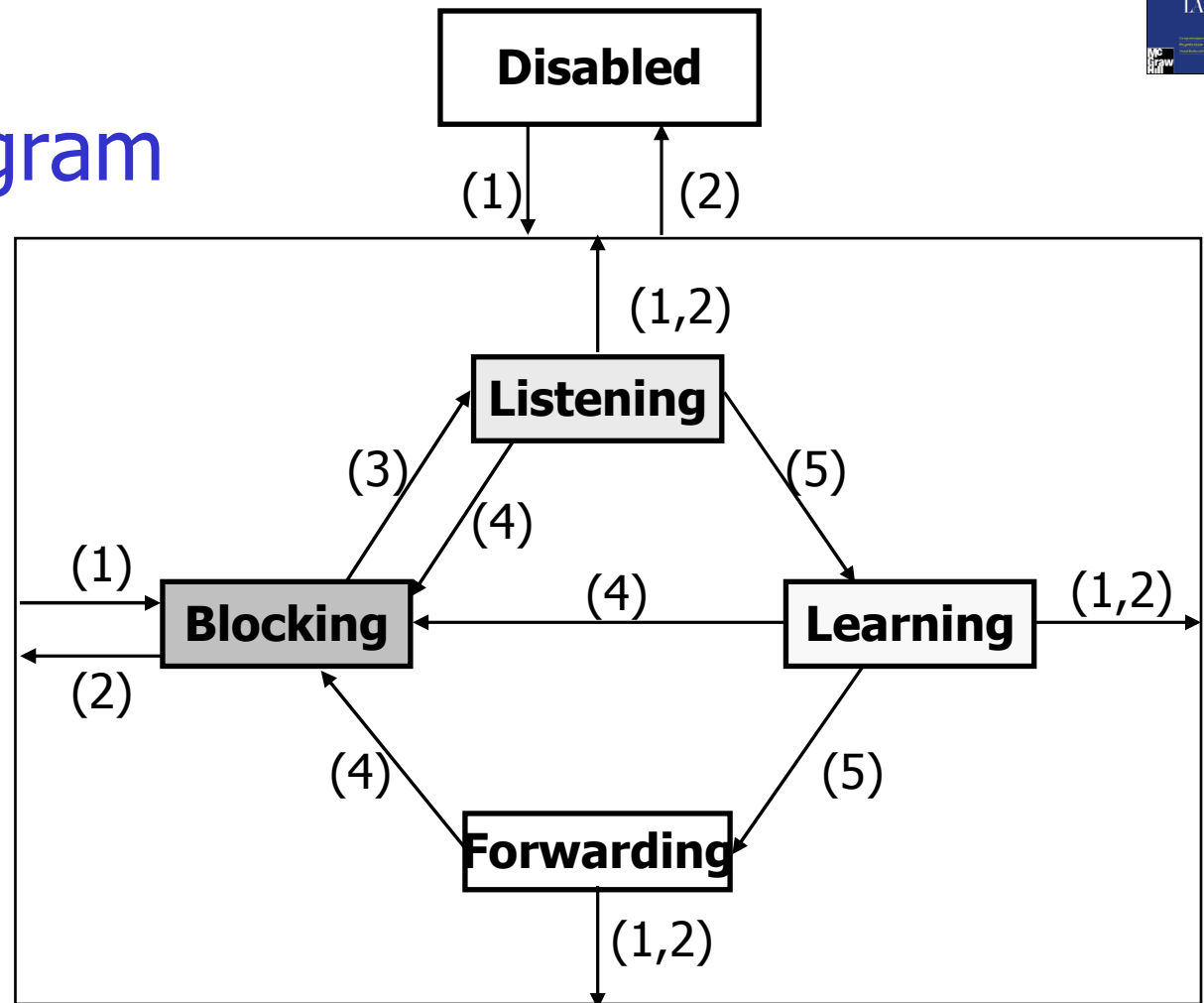
Configuration BPDU processing

- Each configuration BPDU received by the root port is forwarded on the other ports (flooding)
- The Bridge updates each copy:
 - The Bridge adds to the `root path cost` the root path cost value related to the receiving port
 - The Bridge places its Bridge identifier in the proper field
 - The Bridge places the port identifier of the port that will be used to send the BPDU in the proper field

Timers

- Hello time
 - Configuration BPDU generation interval
 - Range: 1 - 10 seconds – Recommended: 2 seconds
- Forward delay timer
 - Delay port state change: from listening to learning, from learning to forwarding state
 - short time to remove the entries from filtering database
 - Range: 4 - 30 seconds – Recommended: 15 seconds
- Max age
 - The maximum age of received protocol information before it is discarded (unblocking timer port).
 - Range: 6 - 40 seconds – Recommended: 20 seconds
- The timers are announced by the root bridge on the configuration BPDU

Port state diagram



- (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)

Port state

■ Listening

- Receive frames
- Do not forward frames
- Do not update the forwarding data base
- Compute received BPDU
- Transmit BPDU

■ Forwarding

- Receive frames
- Forward frames
- Update the forwarding data base
- Compute received BPDU
- Transmit BPDU

■ Learning

- Receive frames
- Do not forward frames
- Update the forwarding data base
- Compute received BPDU
- Transmit BPDU

■ Blocking

- Receive frames
- Do not forward frames
- Do not update the forwarding data base
- Compute received BPDU
- DO not transmit BPDU

Topology Change

Byte	
1÷2	Protocol Identifier: 00-00
3	Protocol Version Identifier: 00
4	BPDU Type: 80

- The Filtering database can be out of date
 - Entries are dropped to guarantee reachability through fllooding
 - New entries are learned
- If a Bridge detects a topology change send a Topology Change Notification BPDU through the root port
 - A bridge that receives a TCN BPDU, forwards that frame through the root port
- Root Bridge answers with a Configuration BPDU with **topology change** bit set
 - The Bridge receiving that BPDU from root port forwards a BPDU with **topology change acknowledgment**
- A Bridge detecting or informed about a topology change sets a short timer to remove the entries equal to forward delay timer

Topology change discovery

- A malfunction concerning the physical layer is found
 - Link Integrity Test failure
- No Configuration BPDU received in the expected time
 - A port in the blocking state sets a timer to the max age parameter value
 - If the timer fires, the ports turns to the listening state
 - The path between the root Bridge and the LAN is down
 - When the forward delay fires, it turns to the learning state
 - When the forward delay fires, it turns to the forwarding state
 - Delayed transitions to avoid oscillations
 - The failure recovery takes 50 seconds

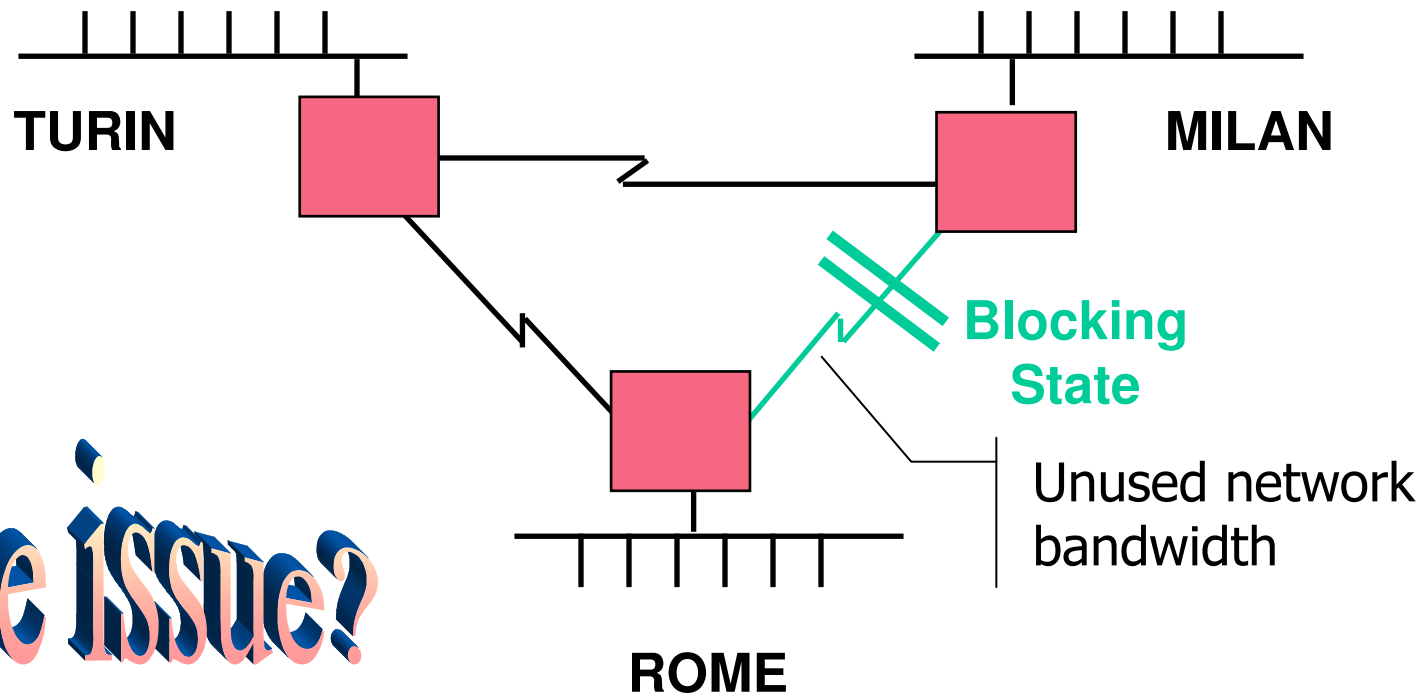
Timers limit

- The recommended timers assure a STP convergence of 50 seconds
- Changing the timers may:
 - Increase or decrease the STP convergence
 - Increase or decrease the maximum bridge diameter
- Not easy to do
 - Suboptimal values can:
 - Reduce network reactivity to topology changes
 - Impair network functionality (loops!!!)
- Hard to manage
 - If the root bridge hardware changes, it is useful to update the timers

Timers limit

- IEEE 802.1D shows ho to get optimal parameter values starting from:
 - `max bridge diameter` → maximum number of Bridges between any two points of attachment of end stations.
 - `maximum bridge transit delay` → maximum time needed by a BPDU to cross a Bridge
 - It starts from the arrival to the departure, including processing
- The recommended timers assure a correct STP operation with a maximum bridge diameter up 7 to bridge
- The results for all the timers must be computed
 - The hello time is usually twice the maximum bridge transit delay
 - The hello time is often set to 1 s for optimization purposes

Spanning Tree for extended LANs



The issue?

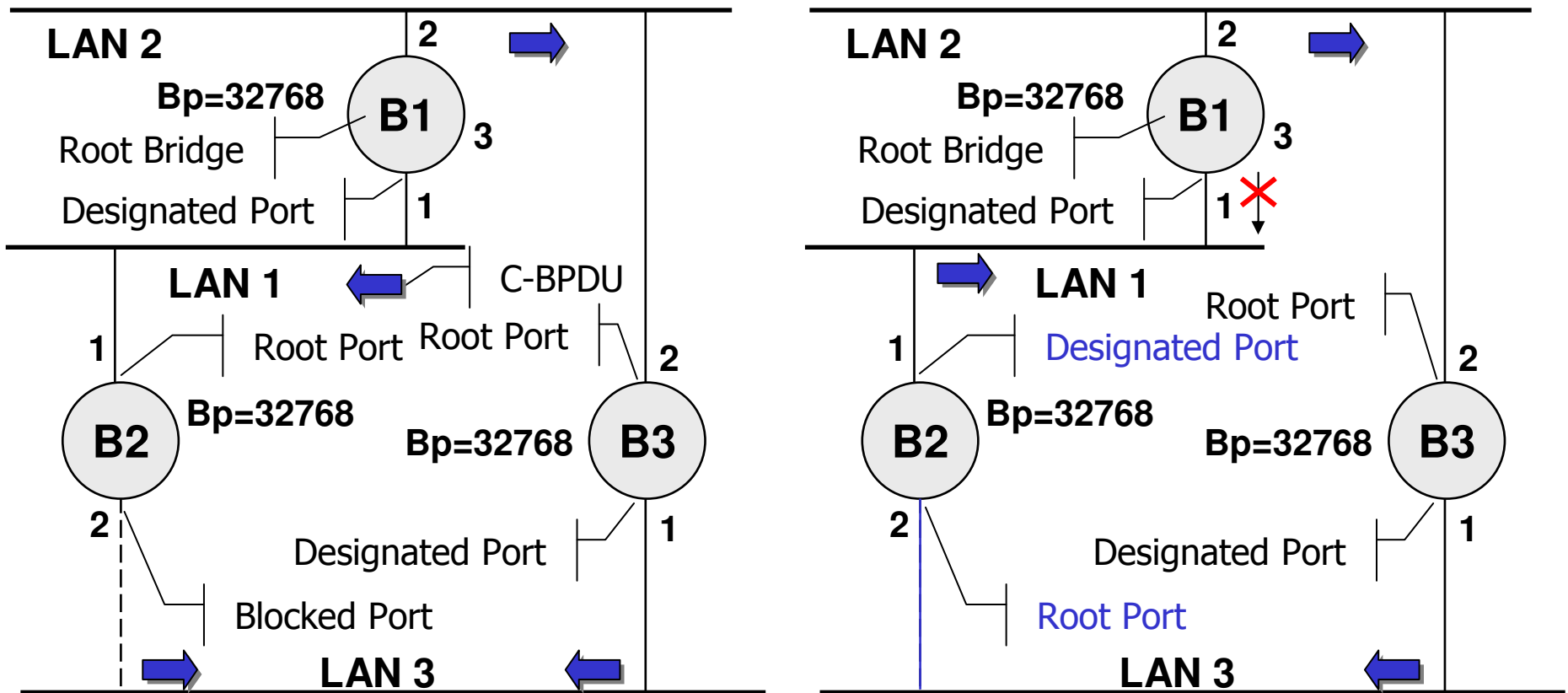
Just a tree for the whole traffic

The solution? A tree for each sender

A big issue: unidirectional paths

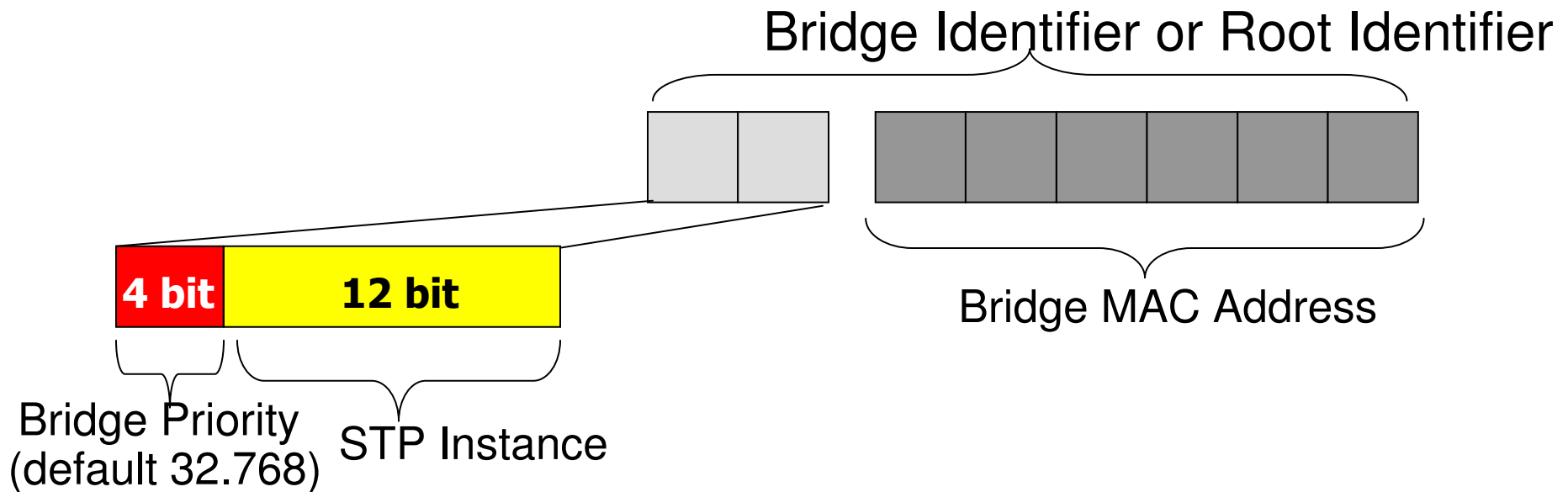
- The BPDUs are forwarded from the root to the leaves
 - The propagation is unidirectional
- If the “directed” propagation breaks, the other endpoint assumes that the port is a designated one
 - A broken bridge does not detect the failure
 - A loop is inserted
- Maybe it is for a limited period, by the way a broadcast storm is generated

Broadcast storm after a failure



IEEE 802.1t: new STP parameters adopted by 802.1w and 802.1s standards

- Extended Port Path Cost in a range from 1 to 200.000.000
- Extended System Identifier



Path Cost IEEE 802.1t & 802.1w



Port speed	Recommended value	Recommended range	Accepted range values
<= 100 Kb/s	200.000.000	20.000.000 - 200.000.000	1 - 200.000.000
1 Mb/s	20.000.000	2.000.000 - 200.000.000	1 - 200.000.000
10 Mb/s	2.000.000	200.000 - 20.000.000	1 - 200.000.000
100 Mb/s	200.000	20.000 - 2.000.000	1 - 200.000.000
1 Gb/s	20.000	2.000 - 200.000	1 - 200.000.000
10 Gb/s	2000	200 - 20.000	1 - 200.000.000
100 Gb/s	200	20 - 2000	1 - 200.000.000
1 Tb/s	20	2 - 200	1 - 200.000.000
10 Tb/s	2	1 - 20	1 - 200.000.000