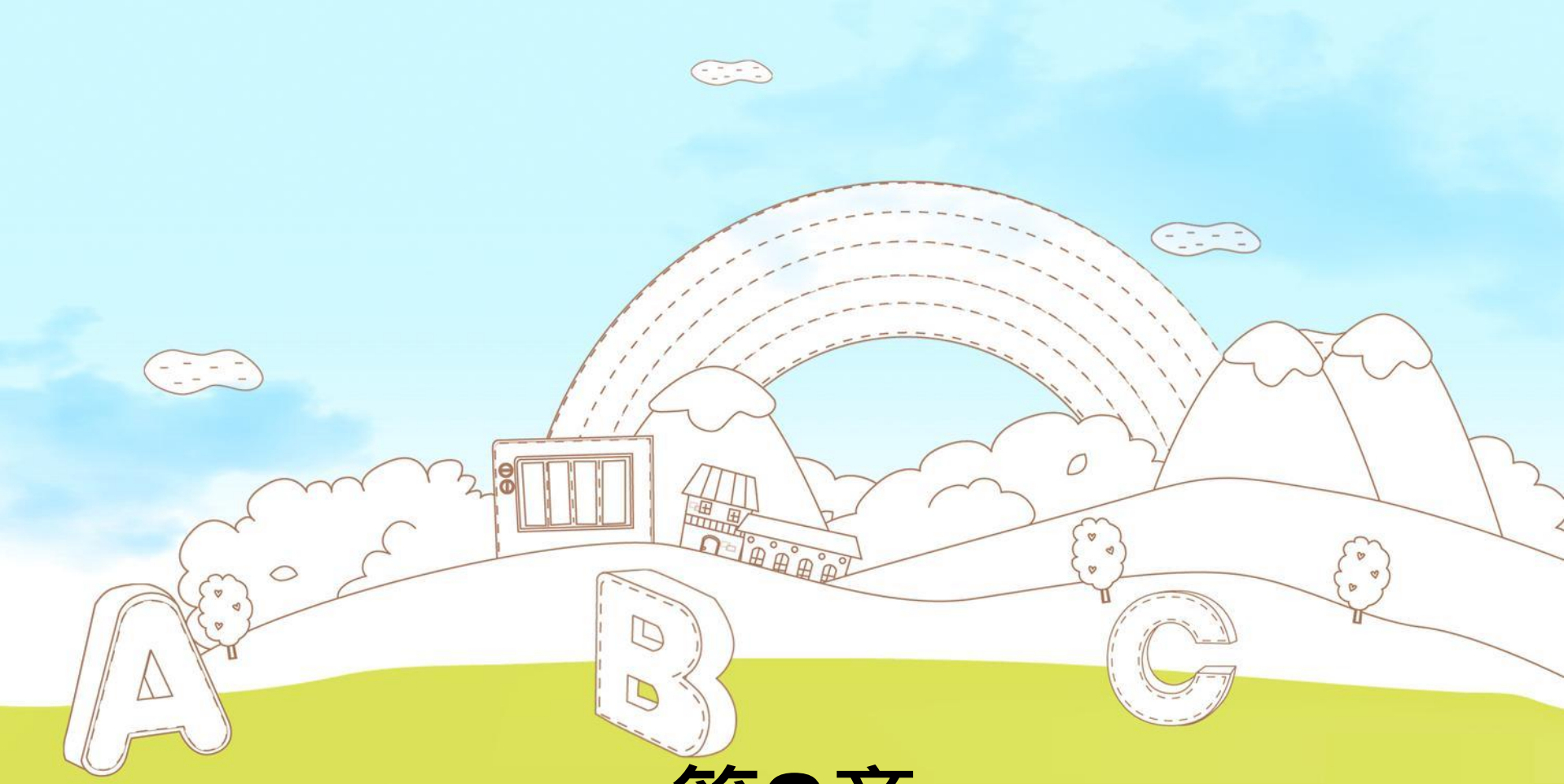
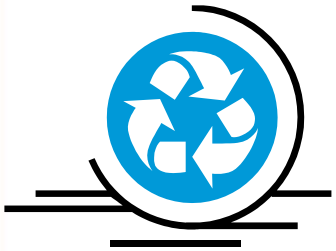


第3章
2层技术和设计



交换机概述

网桥与交换机

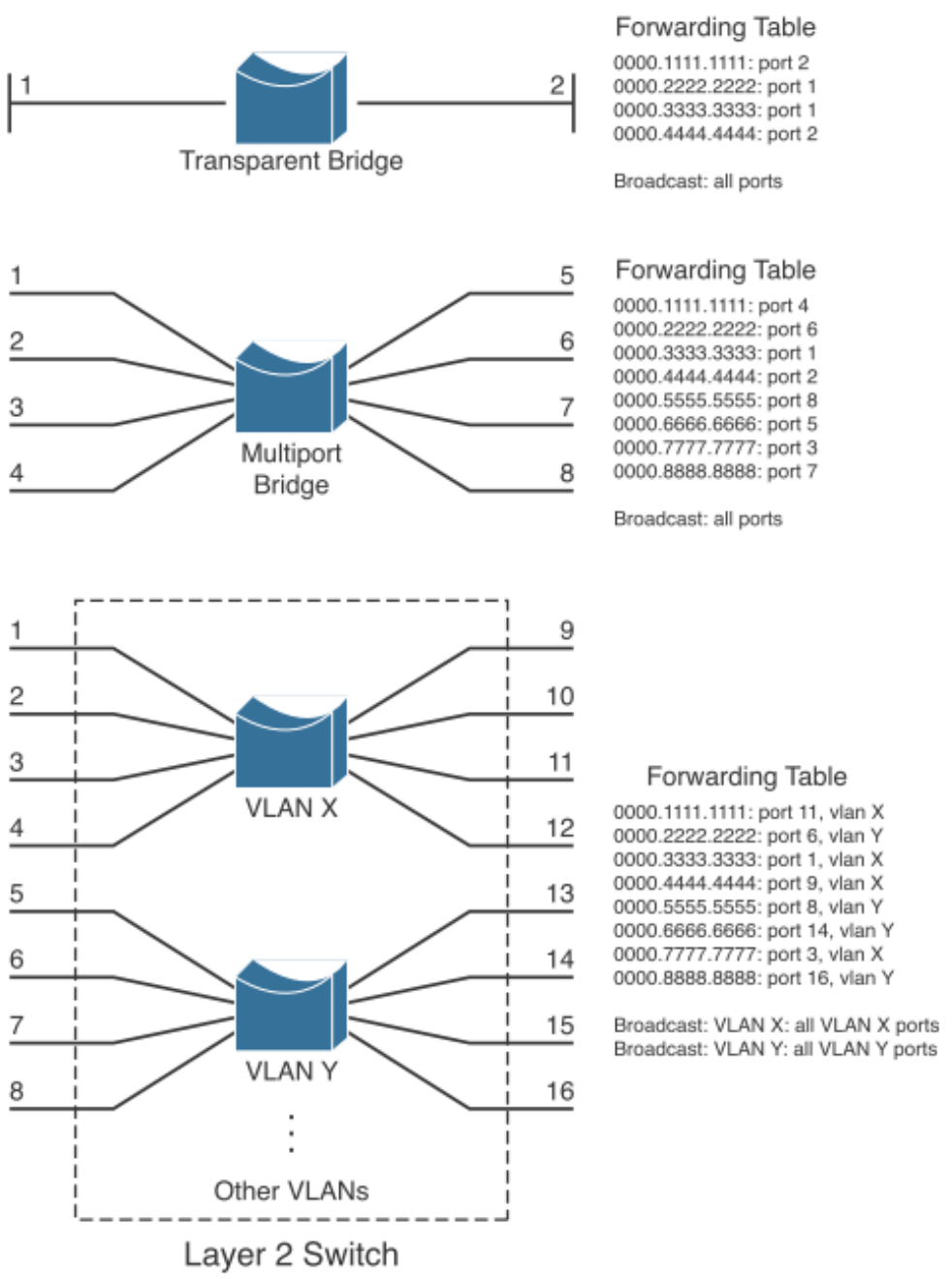
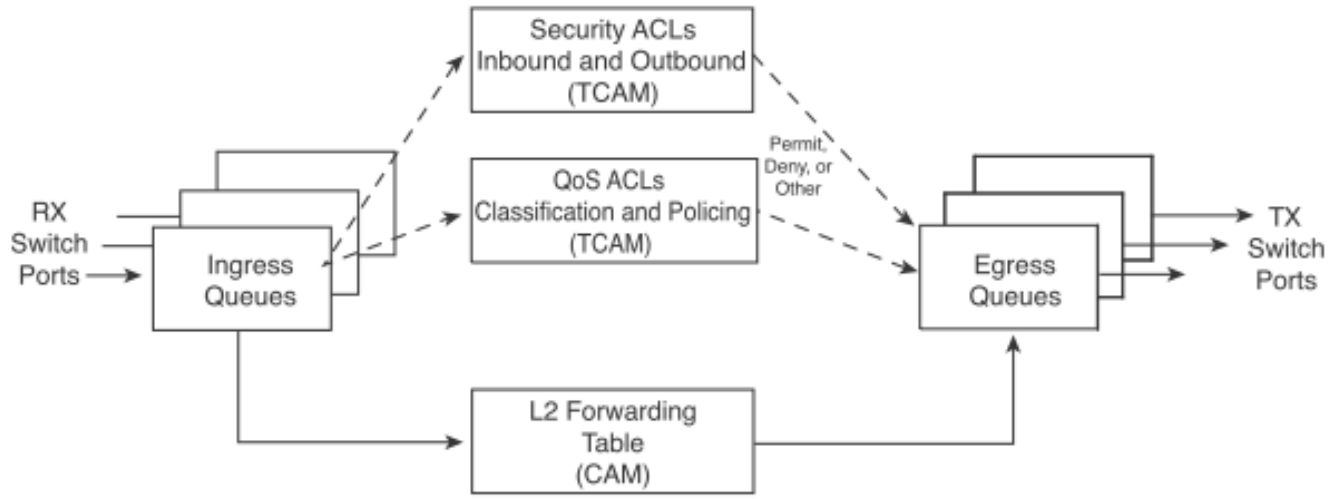


Figure 2-1 *A Comparison of Transparent Bridges and Switches*

二层交换内部构成



CAM Table

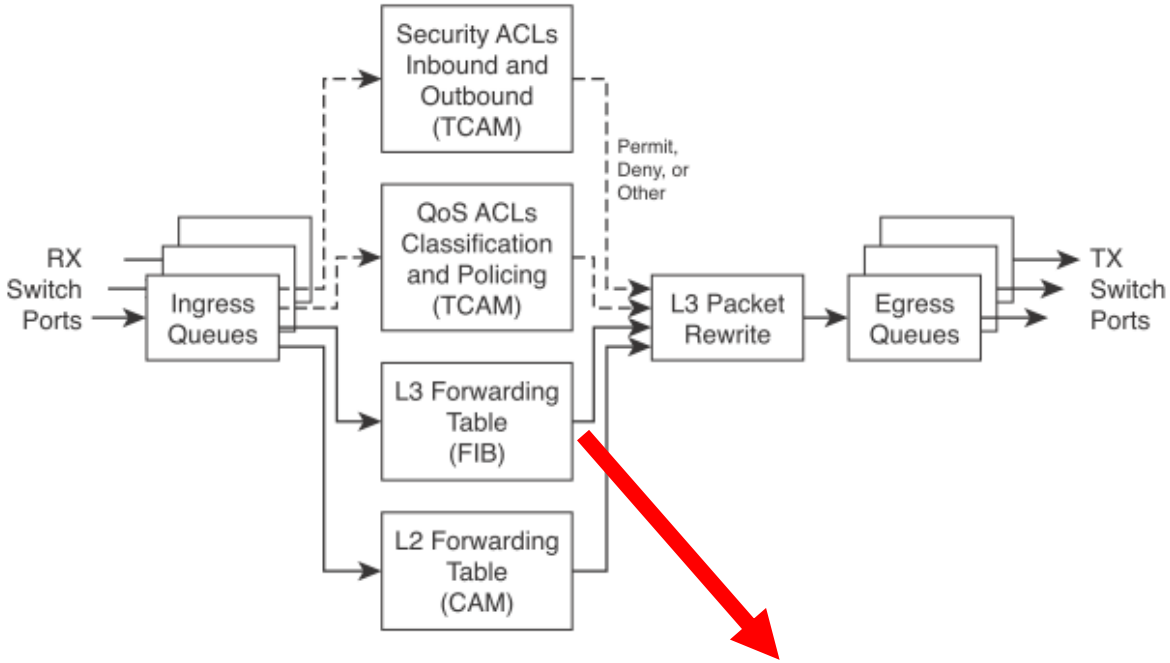
MAC Address	Egress Port	VLAN

Ternary Content Addressable Memory，“三态内容寻址存储器

可以使用掩码

Figure 2-3 *Operations Within a Layer 2 Catalyst Switch*

三层交换内部构成



CAM Table

MAC Address	Egress Port	VLAN

FIB Table

IP Address	Next-Hop IP Addr	Next-Hop MAC Addr	Egress Port

Figure 2-4 *Operations Within a Multilayer Catalyst Switch*

多层交换

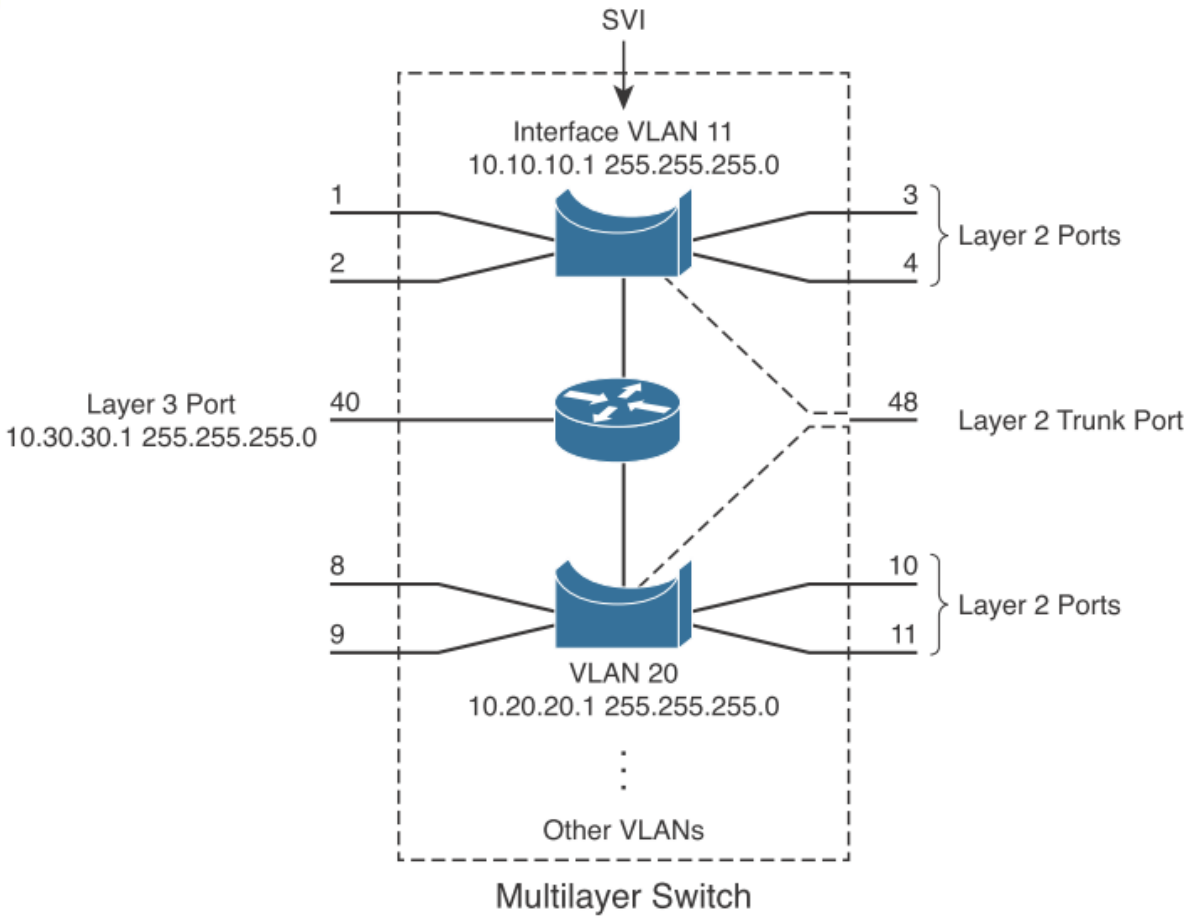
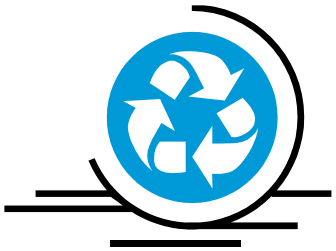


Figure 11-2 *Catalyst Switch with Various Types of Ports*

2层技术和设计



L2设计

二层网络设计★

- 二层网络设计主要使用的技术：
  - Layer 2 control protocols
    - such as Spanning Tree Protocol
  - VLANs and trunking
  - Link aggregation
  - Switch fabric

L2设计

STP

How STP working ★

- **STP完成以下操作:**
  - **Identify path costs on links.**
  - **Identify the root bridge.**
  - **Select root ports (1 per switch).**
  - **Select designated ports (1 per segment).**
  - **Identify the blocking ports.**

Superior BPDU ★

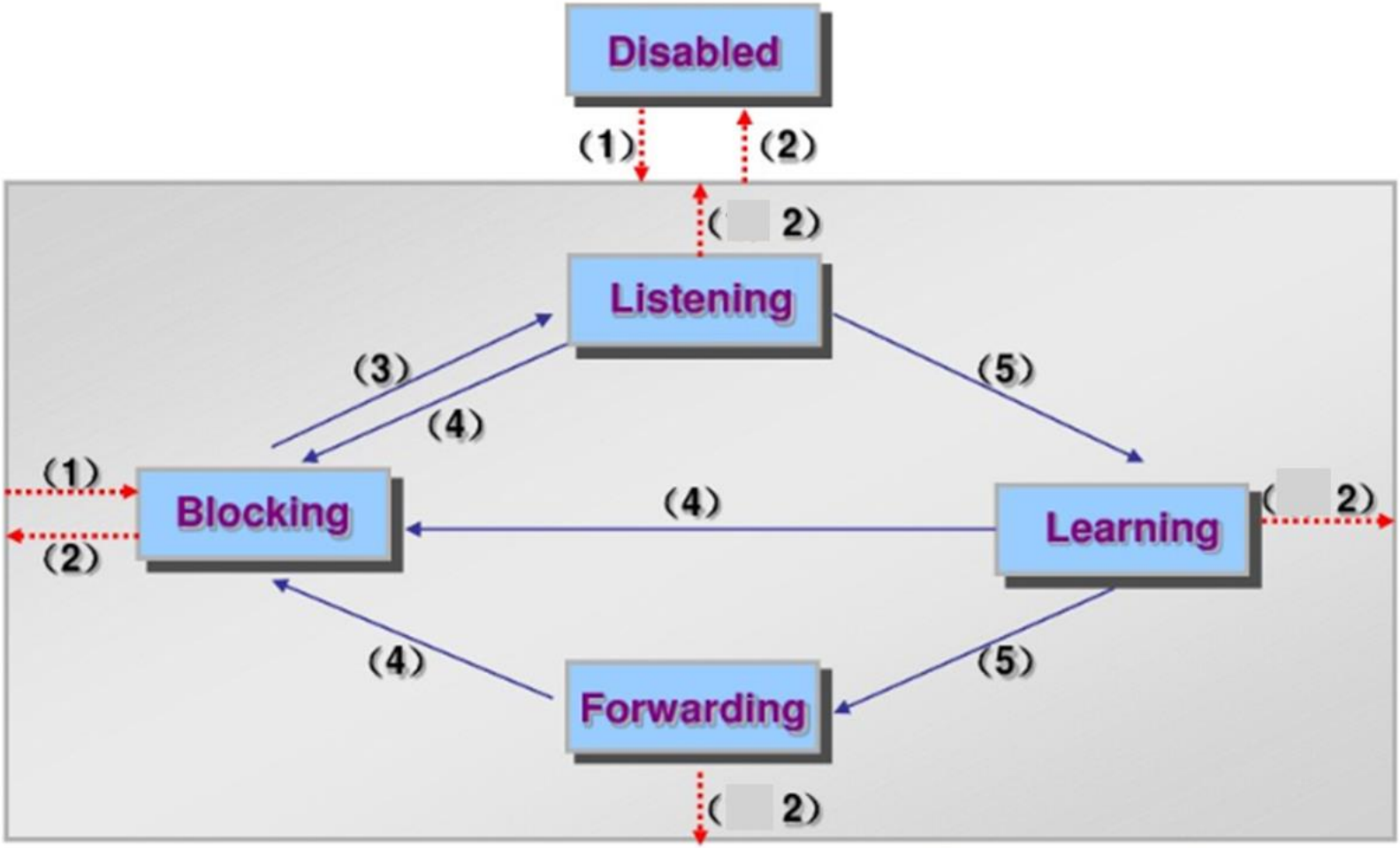
- **Superior BPDUs:**
  - **Root Bridge ID (RBID)**
  - **Root Path Cost (RPC)**
  - **Sender Bridge ID (SBID)**
  - **Sender Port ID (SPID)**
  - **Receiver Port ID (RPID)**
    - **not included in the BPDU, evaluated locally**

looking for the first occurrence of a lower value。

Port state ★

	BPDU packets		DATA packets		
	Receive BPDUS	**TRANSMIT BPDUS**	**LEARN ADDRESSES**	**FORWARD DATA FRAMES**	
Disabled	-	-	-	-	
Blocking	√	-	-	-	**Max Age (20 sec)**
Listening	√	√	-	-	**Forward Delay (15 sec)**
Learning	√	√	√	-	**Forward Delay (15 sec)**
Forwarding	√	√	√	√	

端口状态转换★



1）端口enabled

2）端口disabled

3）端口被选为根端口或指定端口

4）端口被选为备用端口（阻塞）

5）Forward Delay延时

L2设计

VLAN

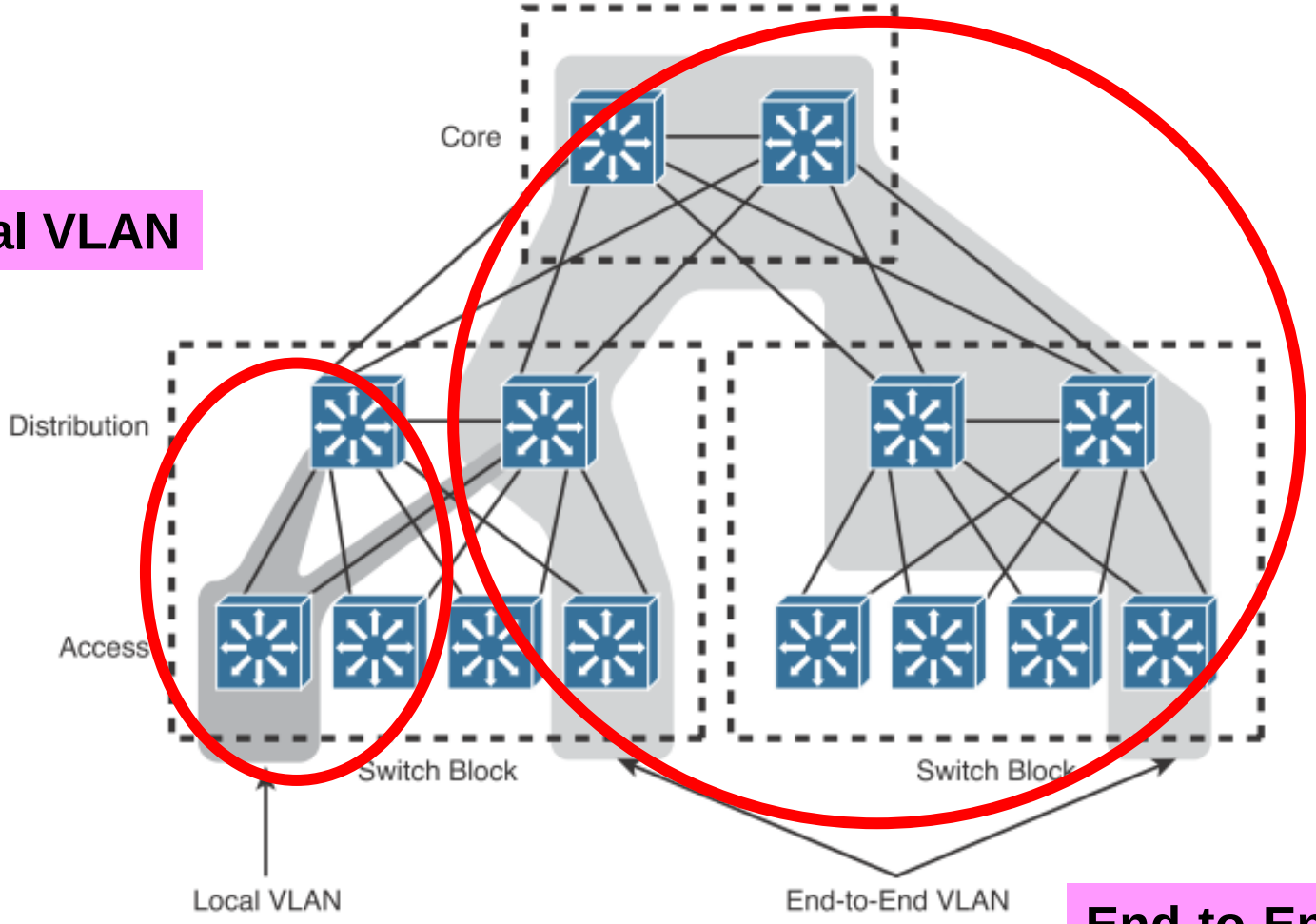
VLANs and Trunking

- **virtual local-area network (VLAN)**
  - 2层技术
  - 网络虚拟化技术
  - 提供广播域的逻辑分割
  - 提供策略控制
  - 另外：
    - 故障隔离
    - 优化性能、稳定性和可管理性。
- **Trunking**
  - 2层设备间单个物理链路传送多个VLAN的数据

设计策略

- 设计考虑：
  - （建议）VLAN 不跨越多个接入层交换机。
  - 部分应用可能要求必须跨多个接入层交换机
  - 需明确常用的2层拓扑和跨交换机VLAN产生的影响。

VLAN type



Local VLAN

End-to-End VLAN

Figure 4-2 *The Extent of an End-to-End VLAN*

Link Aggregation

Link Aggregation

- 汇聚链路使用多条物理链路来构成单条逻辑电路
  - 性价比高，
  - 可以在不更新硬件的前提下，增加累积链路可用带宽。
  - 消除单点故障，提高可靠性和可恢复性

IEEE 802.3ad

Link Aggregation Control Protocol (LACP)

分类

- 两种主要连接类型：
  - Single-chassis link aggregation
  - Multichassis link aggregation（mLAG）

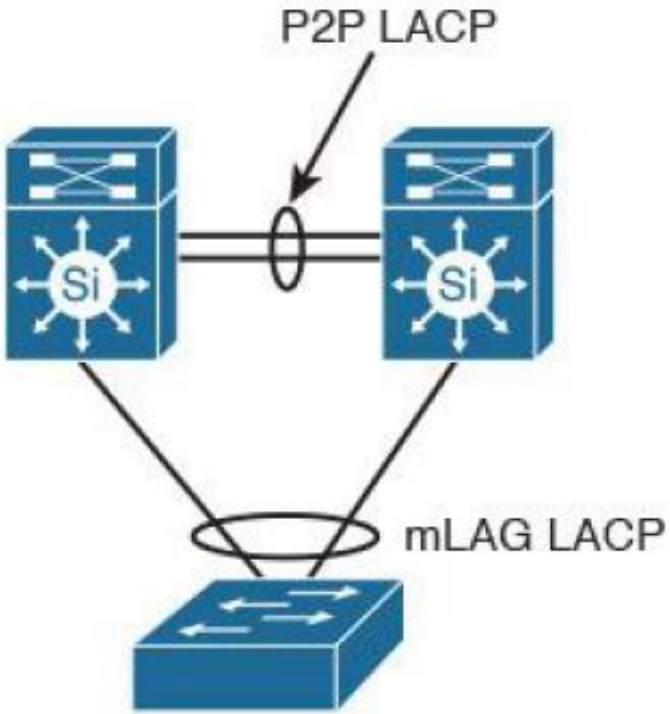
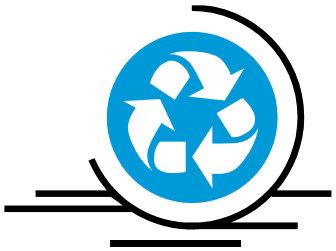


Figure 2-2 *Link Aggregation*



STP进价

STP进价

STP路径选择的影响

示例

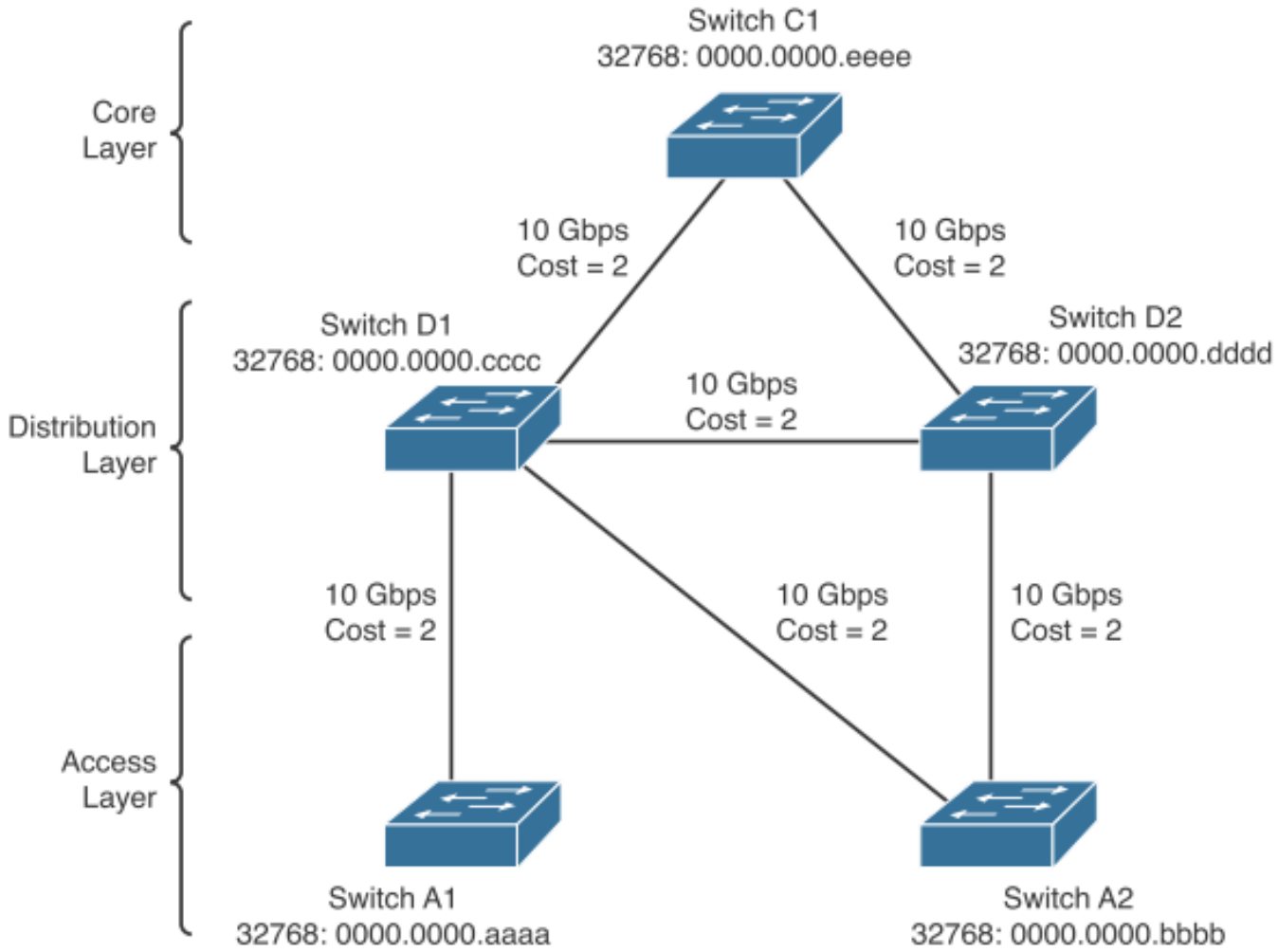


Figure 7-1 *Campus Network with an Inefficient Root Bridge Election*

使用默认配置

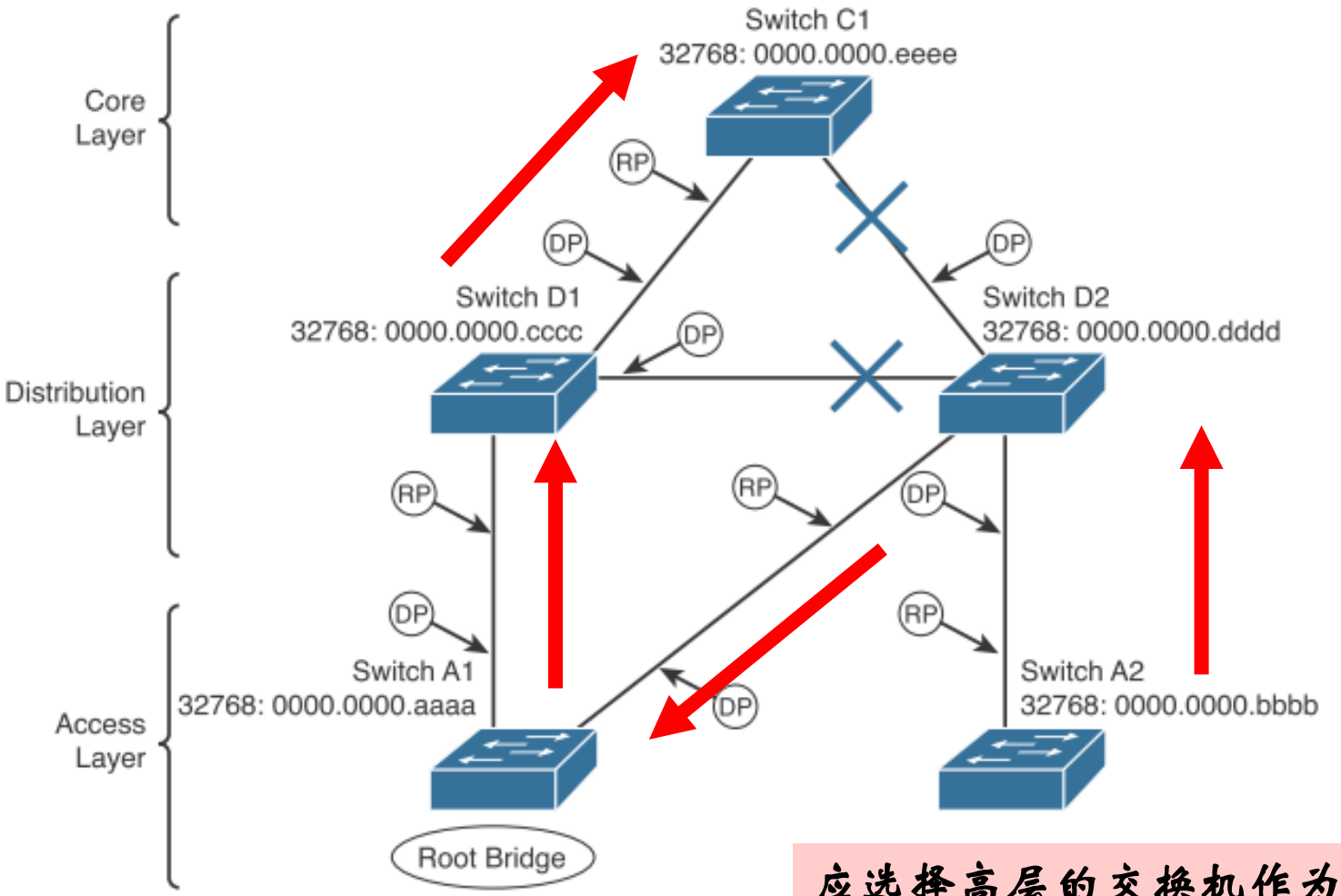


Figure 7-2 *Campus Network with STP Converged*

应选择高层的交换机作为根桥

更改网桥优先级

STP进价

STP拓扑变更

STP定时器★

Table 6-6 *STP Timers*

Timer	Function	Default Value
Hello	Interval between configuration BPDUs.	2 seconds
Forward delay	Time spent in Listening and Learning states before transitioning toward Forwarding state.	15 seconds
Max age	Maximum length of time a BPDU can be stored without receiving an update. Timer expiration signals an indirect failure with designated or root bridge.	20 seconds

TCN消息

- 拓扑改变
  - 从指定端口收到TCN BPDU
  - 当端口由其它状态进入转发状态
    - （有指定端口的交换机）
  - 当端口由转发/监听状态进入阻塞状态
  - 交换机变为根桥
- 拓扑改变时：
  - 端口状态改变的交换机通过根端口发送TCN BPDU

TCN 不包含具体拓扑改变信息

BPDU Packet

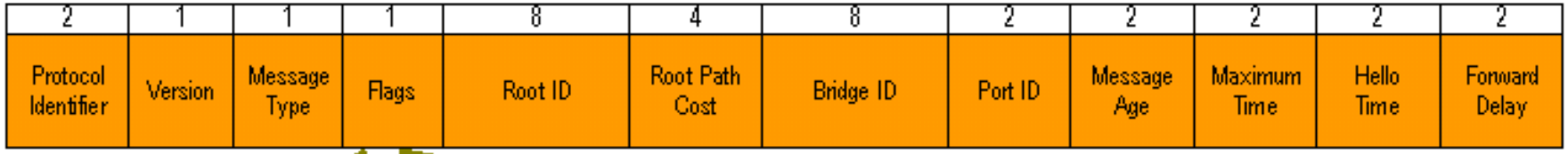
Configuration BPDU

BPDU Field	Length in Octets
Protocol Identifier	2
Protocol Version	1
BPDU Type	1
Flags	1
Root Bridge ID	8
Root Path Cost	4
Sending Bridge ID	8
Sending Port ID	2
Message Age	2
Max Age	2
Hello Time	2
Forward Delay	2

Topology Change Notification BPDU

BPDU Field	Length in Octets
Protocol Identifier	2
Protocol Version	1
BPDU Type	1

BPDU Flags



IEEE 802.1D

7	6	5	4	3	2	1	0

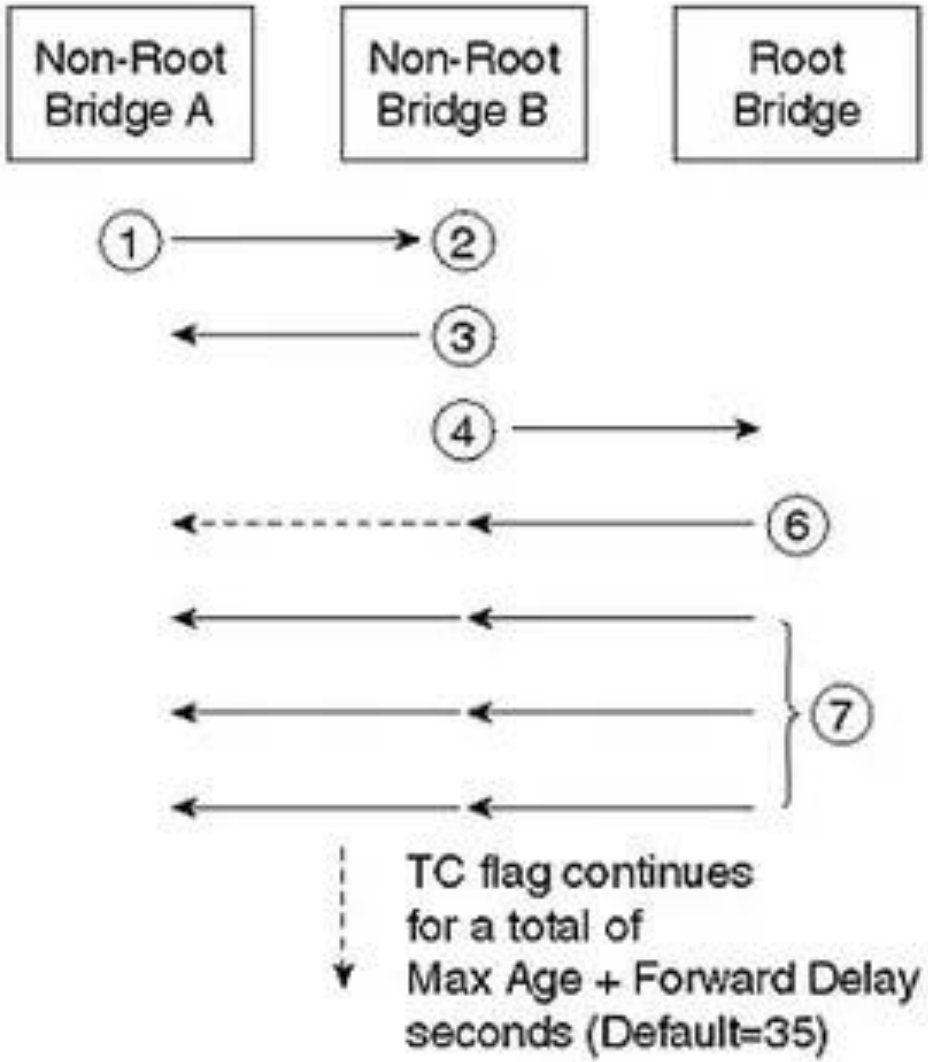
Bit	Function
7	Topology Change (TC)
6	Unused
5	Unused
4	Unused
3	Unused
2	Unused
1	Unused
0	Topology Change Ack (TCA)

IEEE 802.1w

7	6	5	4	3	2	1	0

Bit	Function
7	Toplogy Change (TC)
6	Proposal
5	**Port Role**: 00 – Uknown 01 – Alternate Port 10 – Root Port 11 – Designated Port
4	
3	Learning
2	Forwarding
1	Agreement
0	Topology Change Ack (TCA)

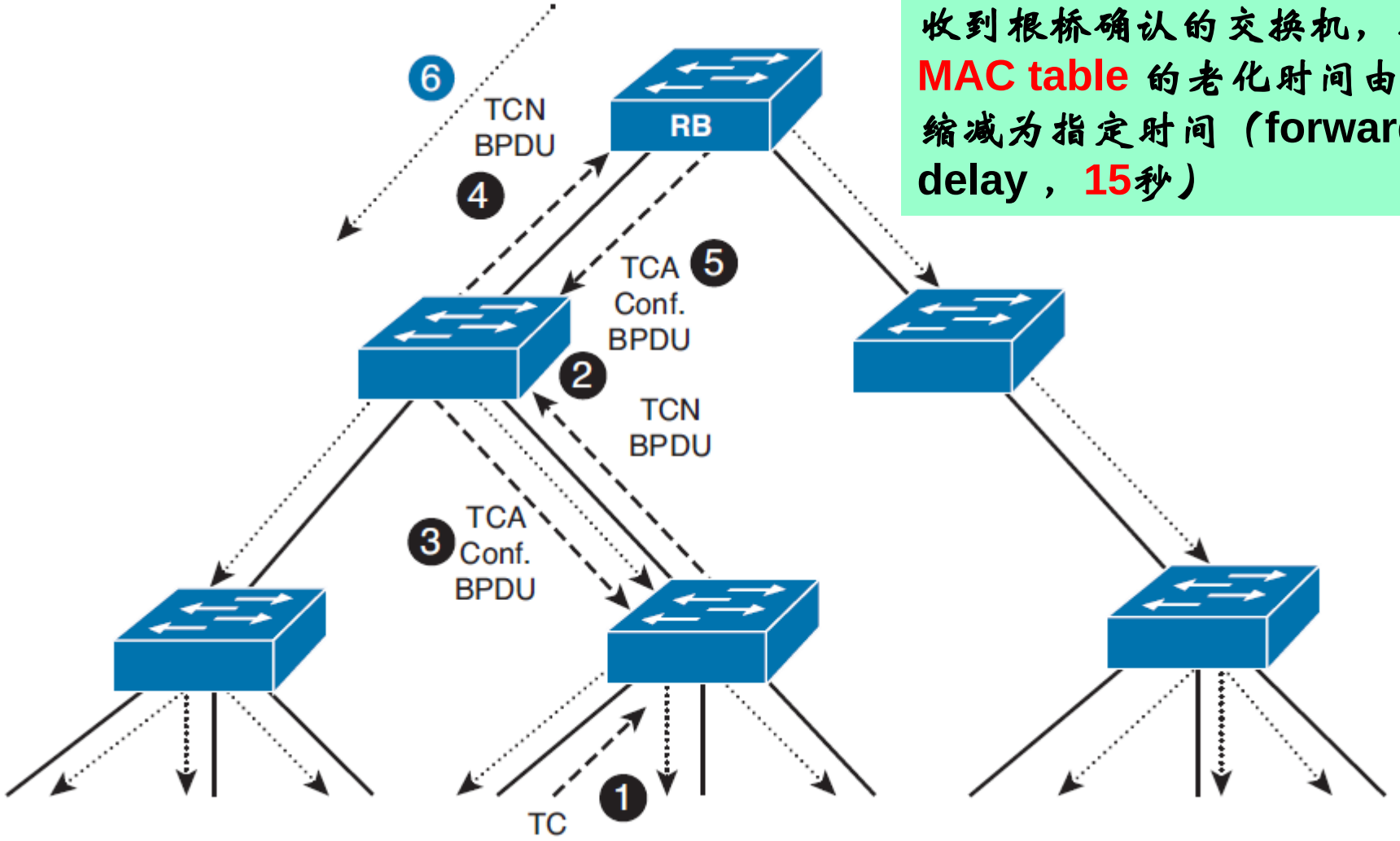
TCN 消息



收到根桥确认的交换机，将MAC table 的老化时间由300秒缩减为指定时间（forward delay，15秒）

拓扑更改消息

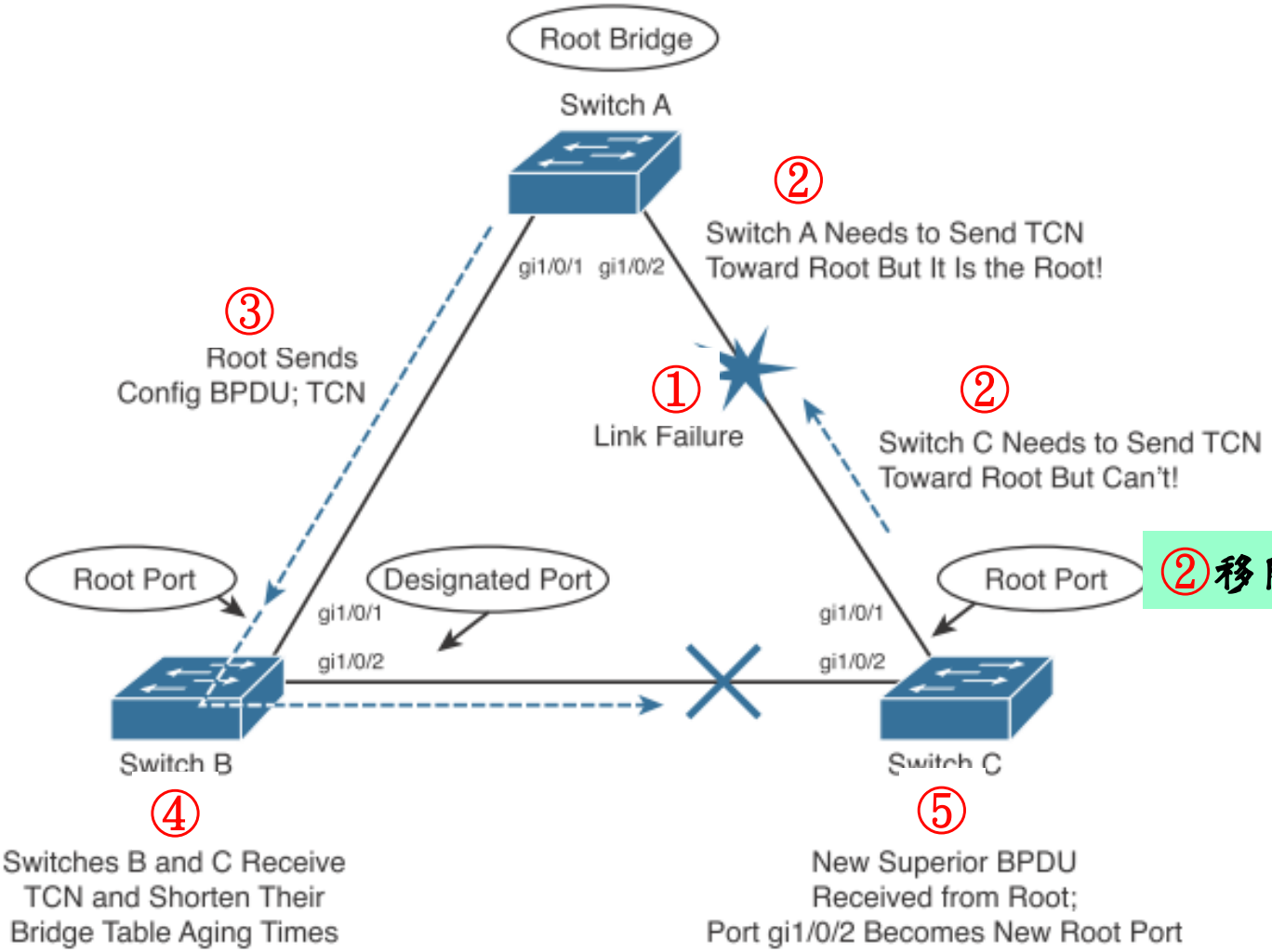
收到根桥确认的交换机，将MAC table 的老化时间由300秒缩减为指定时间（forward delay，15秒）



STP进价

STP收敛时间分析

直连拓扑故障★



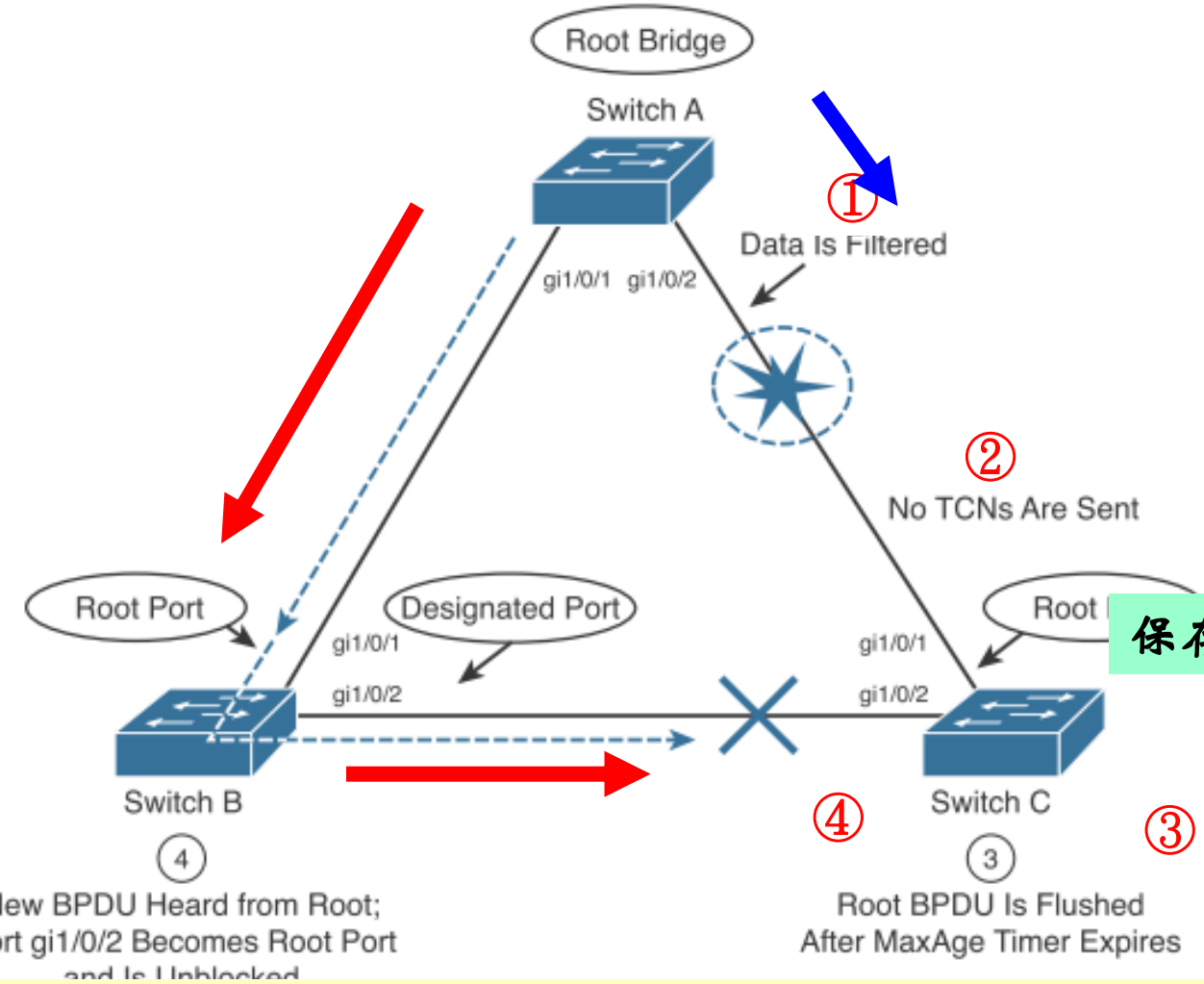
②移除保存的BPDU

5. Blocking -> Listening-> Learning-> Forwarding

30s

0 15 15

间接拓扑故障★



保存的BPDU

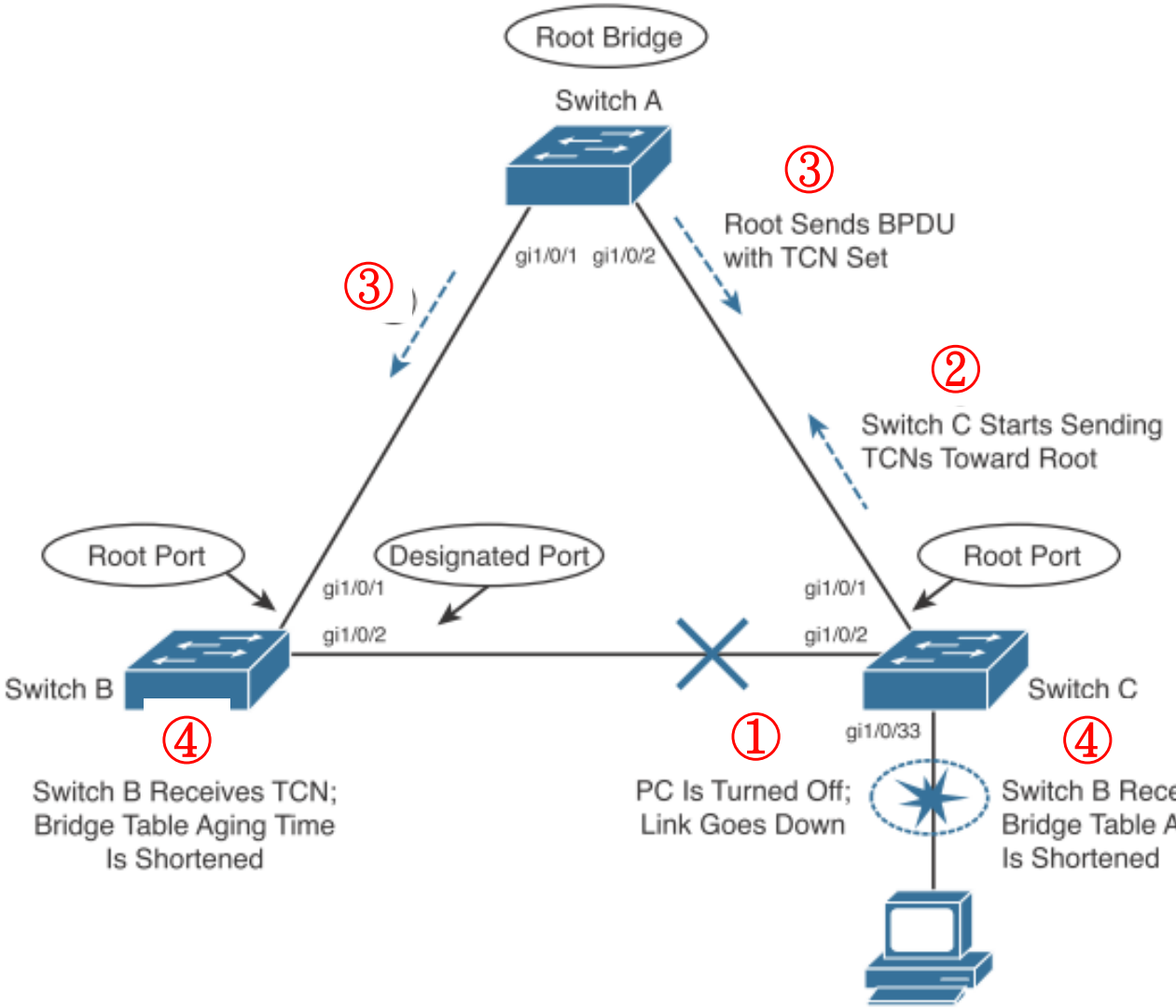
20

5. Blocking -> Listening-> Learning-> Forwarding

50s

07 Figure 6-8 20 ts of an Indir 15 ology Cha 15

非关键故障的影响★



STP 不变

MAC 老化时间缩短

单播导致的洪泛增加

Figure 6-9 *Effects of an Insignificant Topology Change*

STP进价

STP收敛加速技术

STP 收敛加速★

- **STP 收敛加速:**
  - **PortFast :**
    - **access layer switch ports to workstations**
  - **UplinkFast:**
    - **access layer switch when dual uplinks are connected into the distribution layer**
  - **BackboneFast**
    - **backbone or core layer switches**

STP

PortFast:

To Workstation

PortFast

- 用于接入层交换机和工作站（不是交换机）相连
  - 当工作站启用时，配置PortFast的接口直接进入Forwarding，无需经过Listening 和Learning。
  - 并且PortFast端口状态改变时不发送TCN BPDU。

UplinkFast: Access Layer Uplinks

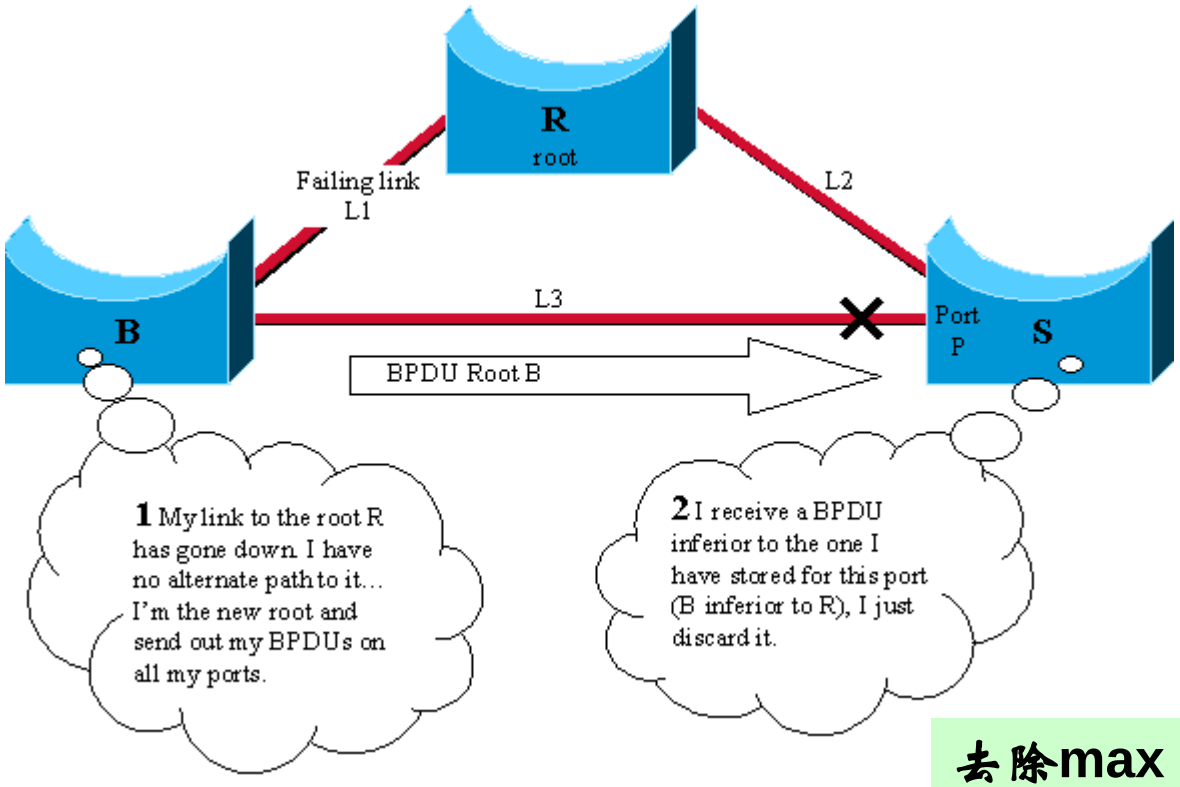
UplinkFast

- 用于STP中的叶子交换机根端口：
  - 在选择根端口时，保存备用根端口信息
  - 当前根端口出错，直接使用备用根端口作为新的根端口。

Backbone fast

Backbone fast

- 用于核心层和汇聚层之间



50s 收敛

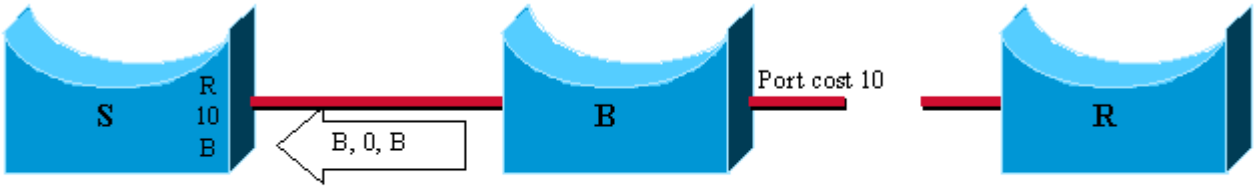
去除max age(20s)延时

方法：使用 Root Link Query，RLQ PDU。

方法：只适用于中继交换机的直连故障

直连和非直连故障

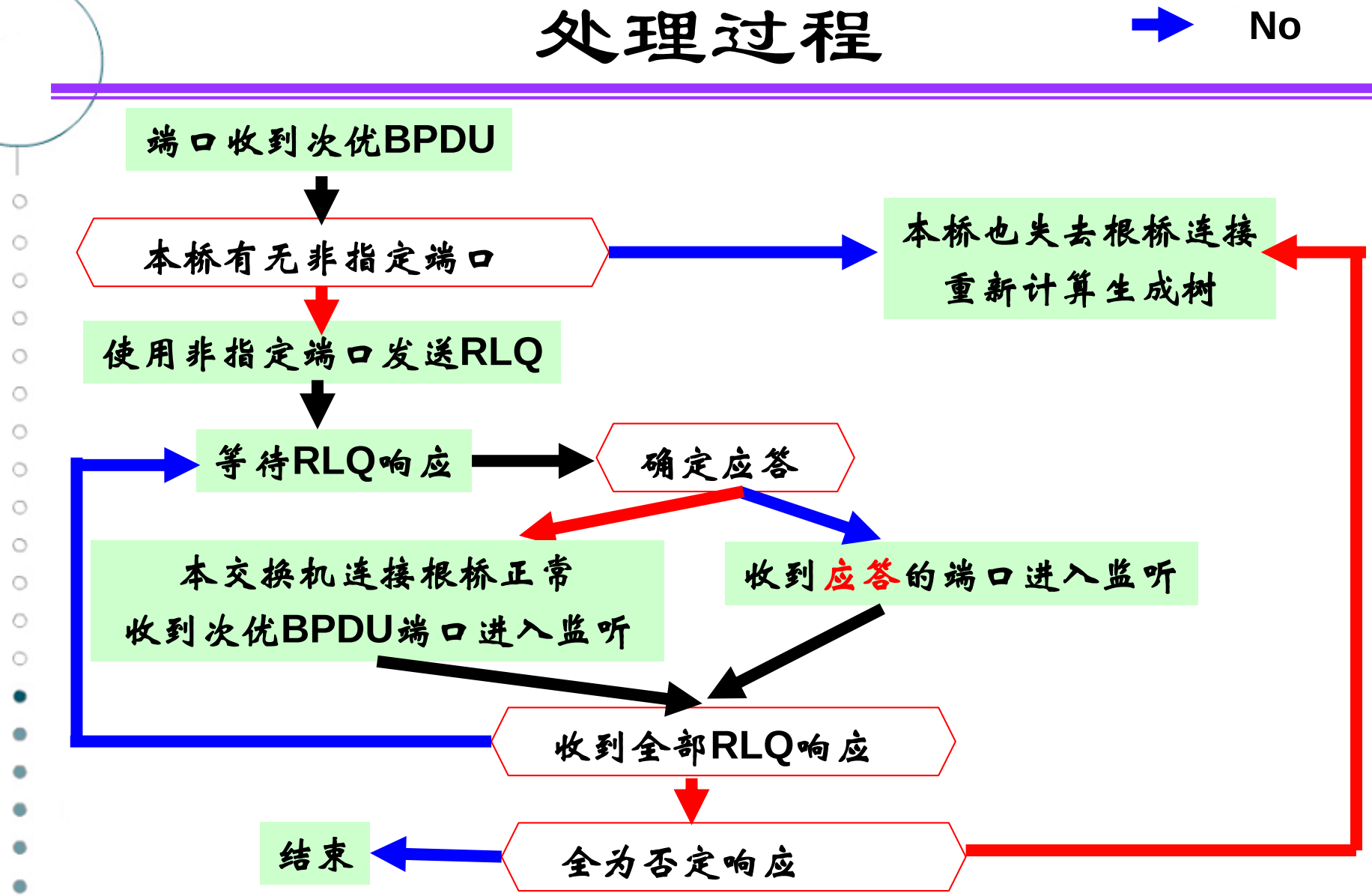
S 可以通过查询检测B的直连故障



1 in this case, B lost the root and sends a BPDU with root id B, path cost 0 and bridge id B. It is inferior to the one that S had stored, because R is a better root than B.

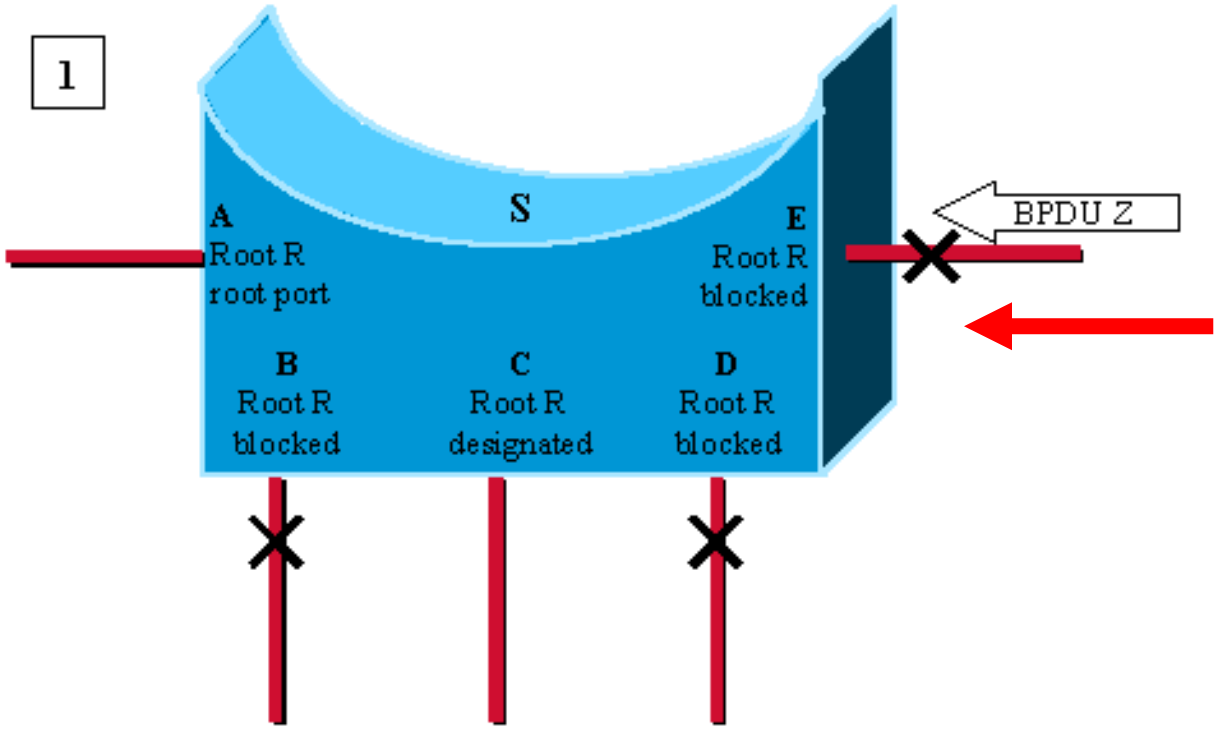
S 无法检测B的非直连故障

处理过程



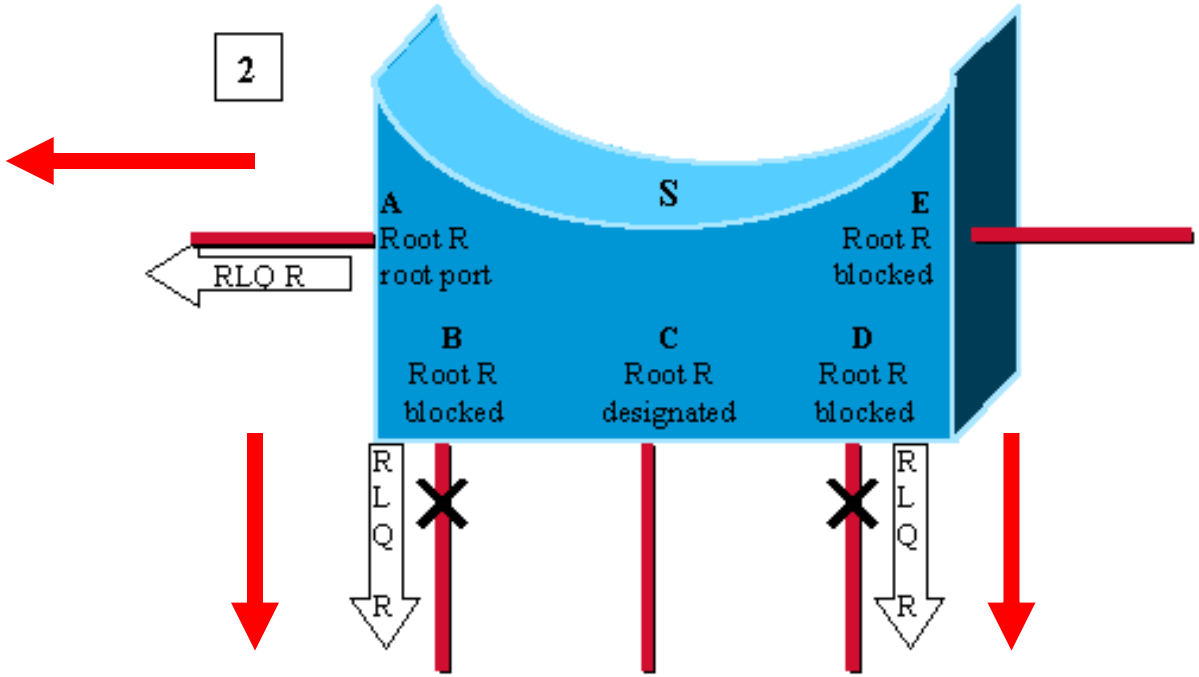
此处的非指定端口含根端口

1. 收到次优BPDU



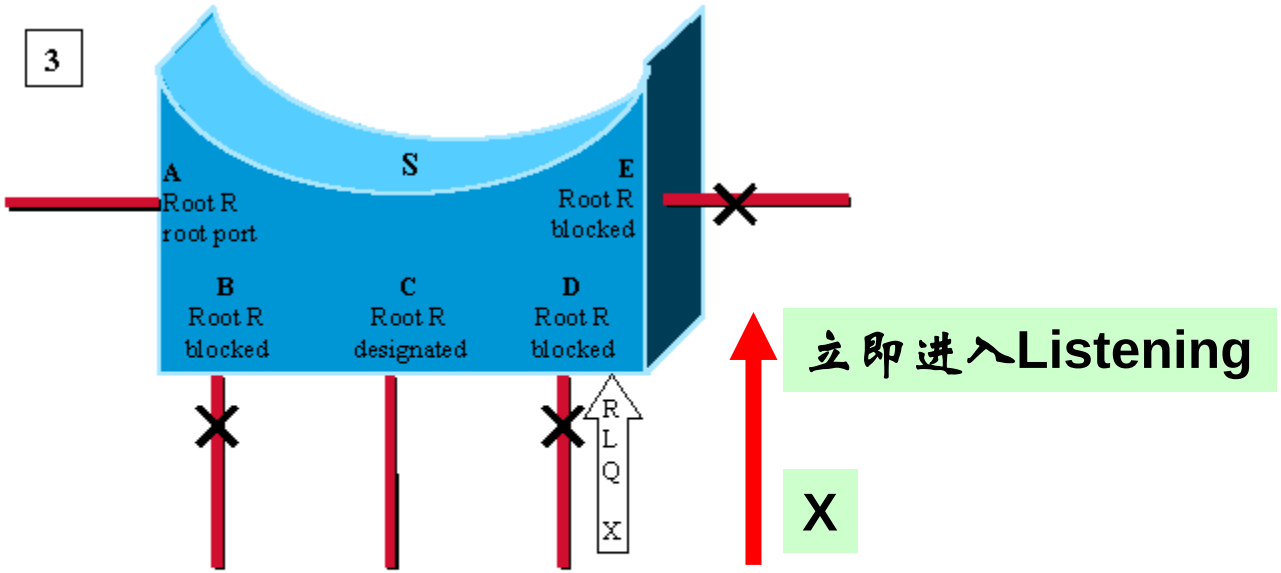
Port E receives an inferior BPDU, advertising root Z instead of root R stored on the different ports.

2. 向所有非指定端口发送RLQ



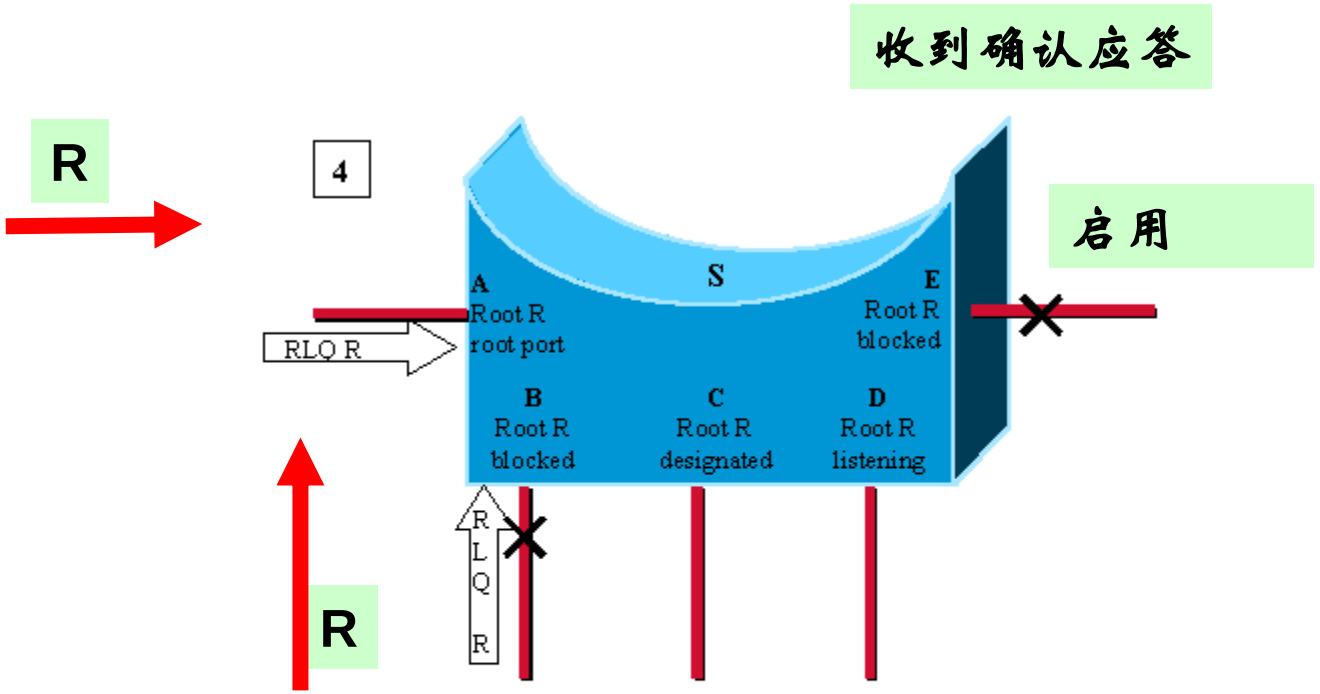
Switch S needs to recheck all its other non-designated ports. It sends out a RLQ request for root R on ports A,B and D.

3. 收到单个应答



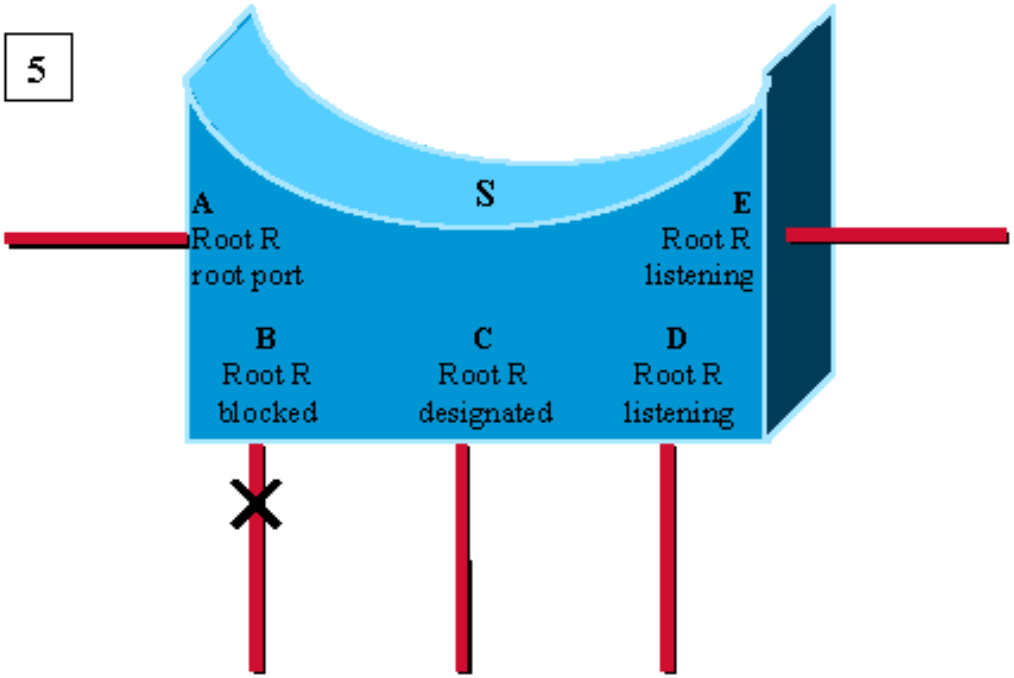
Port D is the first to receive and RLQ response from bridge X claiming to be the root. It is a negative response: D has lost connectivity to the root R. We age out immediately the BPDU on port D and go to listening. As we don't know if we still have connectivity to the root R, we don't age out port E yet.

4. 收到全部应答



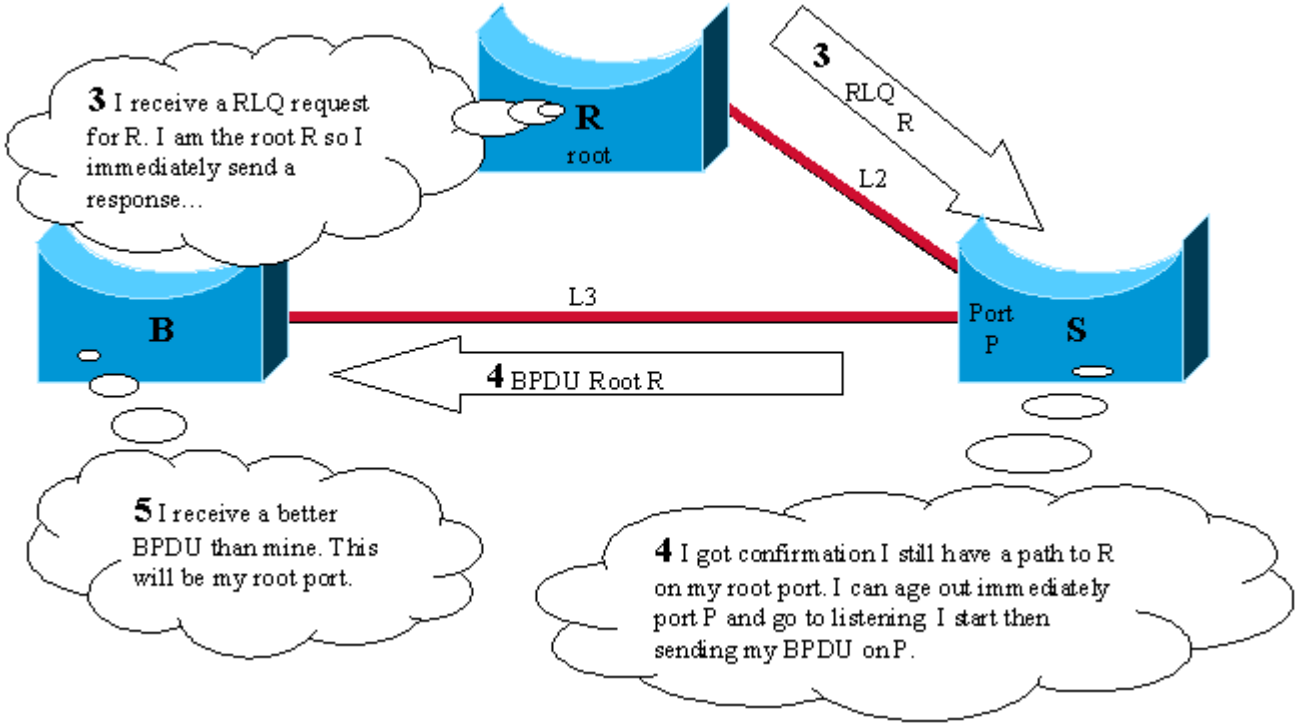
Here, A and B receive a RLQ response confirming R as being the root. As switch S still has connectivity to the root, we can age out immediately the BPDU stored on port E.

5. 使用STP 确定端口角色



Port E transitions to listening without waiting for max_age. Usual spanning-tree rules then apply to determine whether E and D will eventually go to blocking or forwarding.

使用RLQ加速


3 I receive a RLQ request for R. I am the root R so I immediately send a response…
R
root
3
RLQ
R
L2
B
L3
Port
P
S
4 BPDU Root R
5 I receive a better BPDU than mine. This will be my root port.
4 I got confirmation I still have a path to R on my root port. I can age out immediately port P and go to listening. I start then sending my BPDU on P.

STP进价

STP保护技术

STP保护技术★

- **STP保护技术**
  - **Root Guard**
  - **Loop Guard**
  - **BPDU Guard**
  - **BPDU Filter**

Root Guard

- 功能：配置该功能端口相连的交换机不会成为根桥
  - 当被保护的端口收到更优根桥BID
    - 表示相连的交换机要成为根桥
  - 该端口进入 root-inconsistent 状态。
    - 此时只接收BPDU，不会转发。
    - 根桥不会变更
  - 维持此状态，直至不再收到更优的BPDU
    - 然后使用STP确定其端口类型

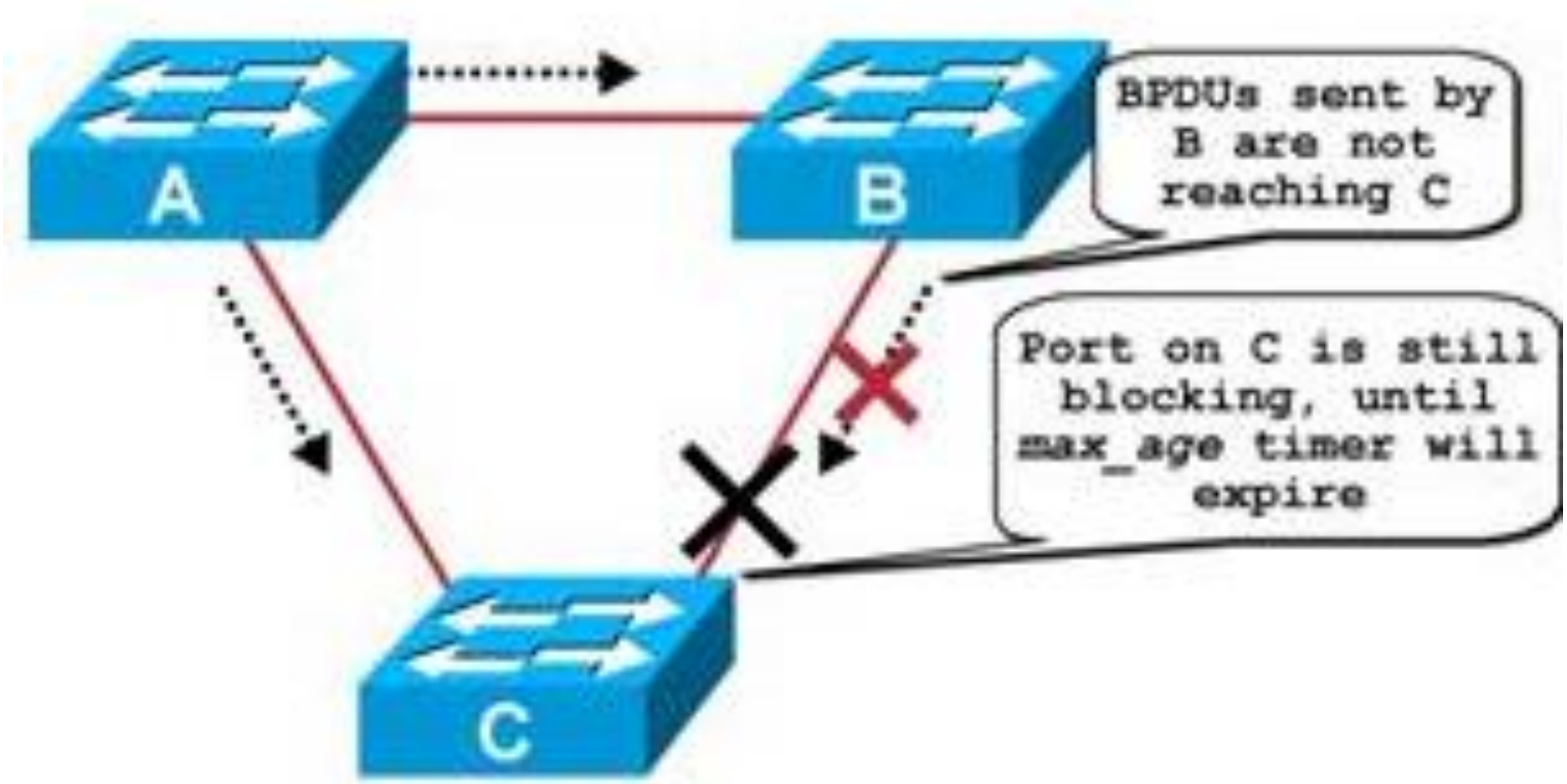
BPDU Guard

- 功能：和 PortFast 功能同时使用，配置端口收到BPDU时，端口停用。
  - 配置端口收到BPDU（无论是否更优）
  - 端口进入 errdisable 状态。
    - 需要手动 shut down 后重新启动。
    - 或配置 errdisable timeout 功能.

单向链路处理

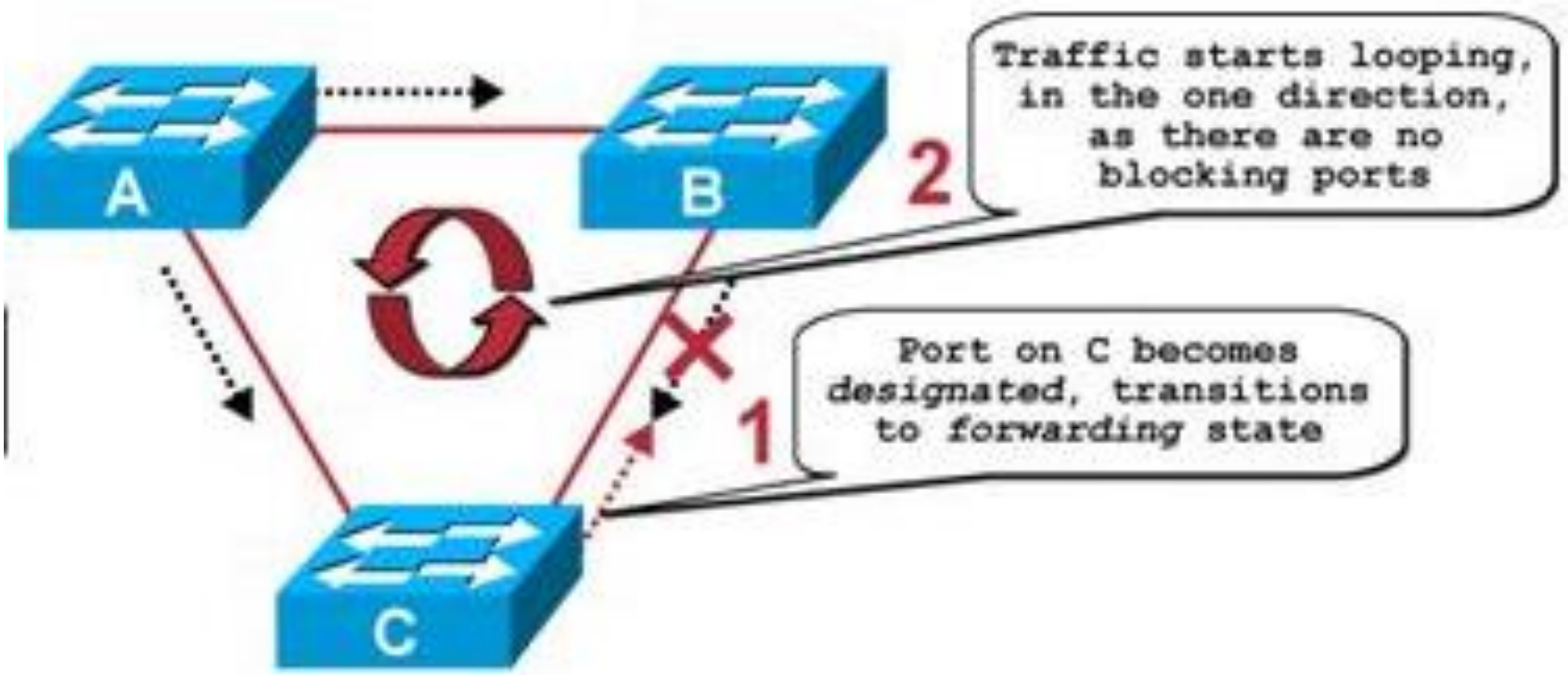
- 单向链路
  - 光纤使用两个介质实现双向传输，可能出现单向链路
  - UTP通常不会出现单向链路
- Cisco 使用以下两种技术避免单向链路导致的问题:
  - Loop Guard
  - Unidirectional Link Detection (UDLD)

单向连接导致的环路



当一个链路出现单向故障的时候

单向连接导致的环路


A
B
C
2
1
Traffic starts looping, in the one direction, as there are no blocking ports
Port on C becomes designated, transitions to forwarding state

Loop Guard

- 功能：配置端口收不到BPDU时停用。
  - 当端口处于非指定端口状态时跟踪BPDU
  - 收到BPDU，正常工作
  - BPDU丢失，进入 loop-inconsistent 状态。
    - 保持端口处于阻塞状体。
    - 避免出现环路

UDLD

- 功能：单向链路检测使用查询确认是否为双向连接。
  - 交换机通过定时发送2层的UDLD帧确定是否任然保持双向连接。
  - 链路另一端的交换机也需启用该功能，并在回应UDLD帧中加入其端口标识。

UDLD 工作模式

- UDLD 有两个工作模式：
  - Normal mode： 检测到单向链路时，
    - 仅将端口标识为undetermined（端口仍保持原有工作），
    - 并在syslog 中纪录
  - Aggressive mode： 检测到单向链路时，
    - 交换机将重新尝试建立连接：每秒发送一个UDLD，持续8秒。
    - 若仍无法收到应答，端口进入 errdisable 状态（停止使用）

BPDU filtering

- 功能：禁止端口发送和处理BPDU

常用STP保护和加速配置★

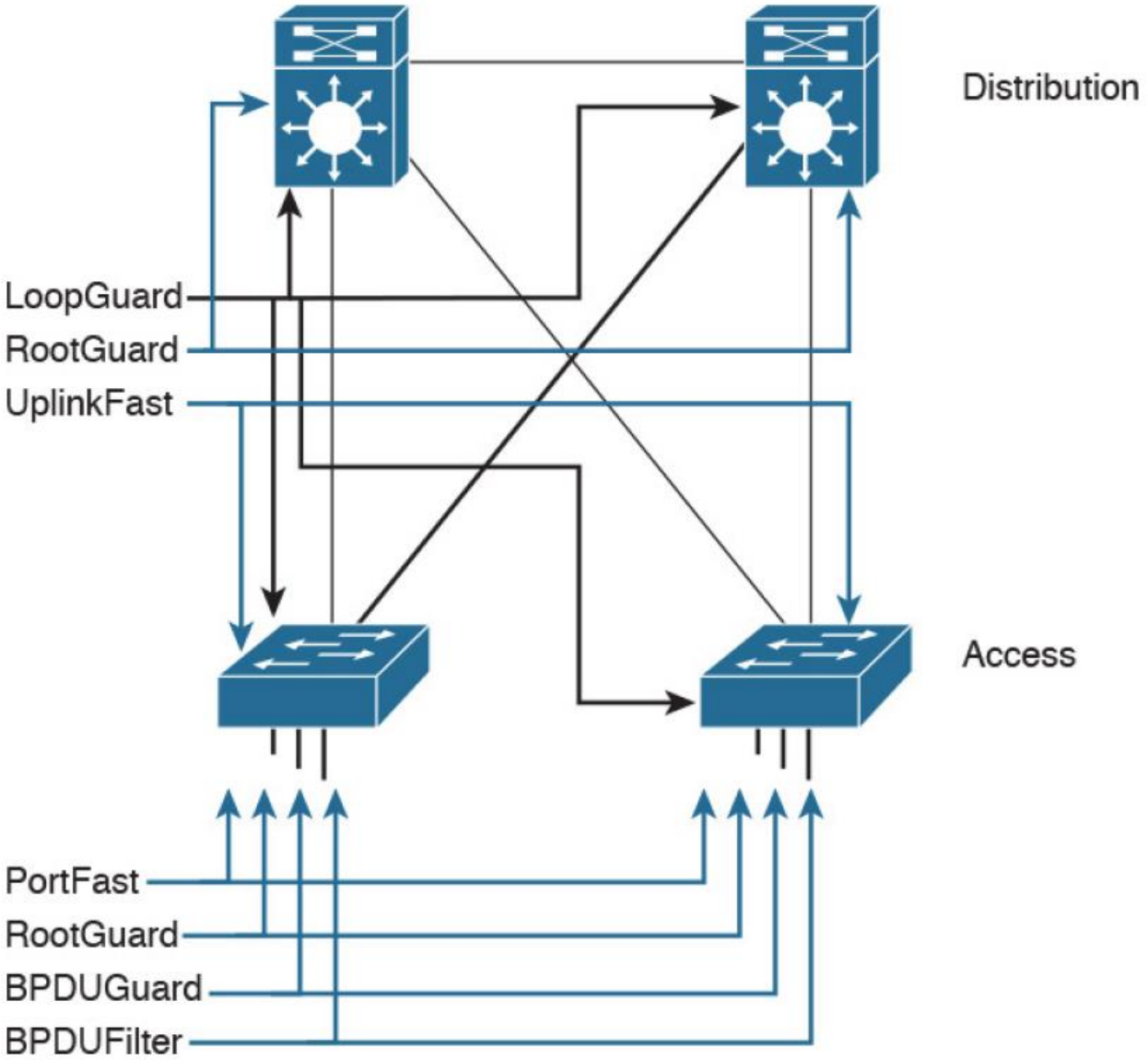
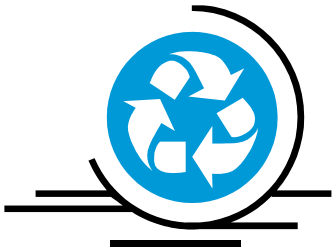


Figure 3-12 *Cisco STP Toolkit mechanisms*

总结

Mechanism	Improves STP Performance or Stability	Description
PortFast	STP performance	Bypasses listening-learning phases to transition directly to the forwarding state
UplinkFast	STP performance	Enables fast uplink failover on an access switch
BackboneFast	STP performance	Enables fast convergence in distribution and core layers when STP changes occur
Loop Guard	STP stability	Prevents an alternate or root port from being the designated port in the absence of bridge protocol data units (BPDUs)
Root Guard	STP stability	Prevents external switches from becoming the root of the STP tree
BPDU Guard	STP stability	Disables a PortFast-enable port if a BPDU is received
BPDU Filter	STP stability	Suppresses BPDU on ports

Table 3-10 Mechanisms Within the Cisco STP Toolkit

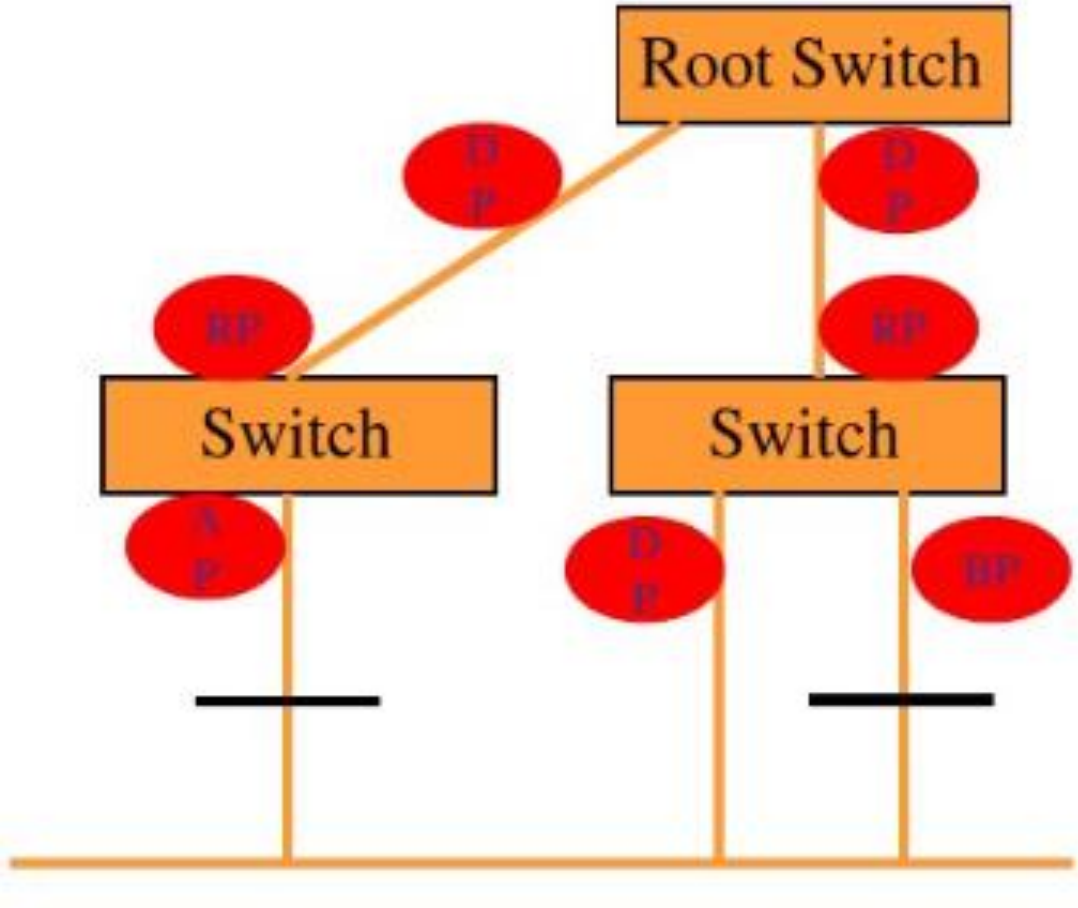


RSTP技术

RSTP 端口角色★

- **RSTP 端口角色（port roles）：**
  - **根端口，Root port : The one switch port on each switch that has the best root path cost to the root.**
  - **指定端口，Designated port : The switch port on a network segment that has the best root path cost to the root.**
  - **替代端口，Alternate port : A port that has an alternative path to the root, different from the path the root port takes. This path is less desirable than that of the root port**
  - **备用端口，Backup port : A port that provides a redundant (but less desirable) connection to a segment where another switch port already connects**

RSTP端口角色



RSTP 端口状态★

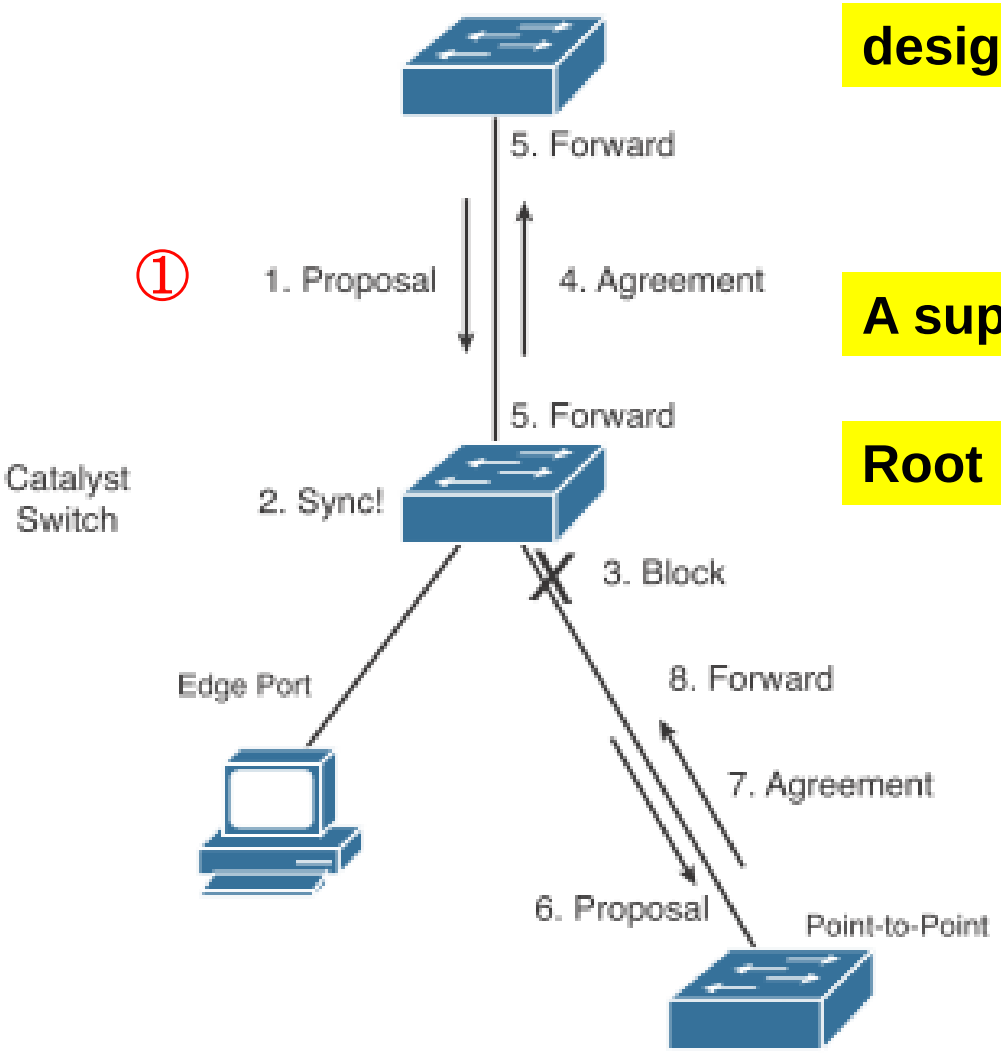
- **RSTP 端口状态（port states）：**
  - **丢弃，Discarding： Incoming frames simply are dropped; no MAC addresses are learned.**
  - **学习，Learning： Incoming frames are dropped, but MAC addresses are learned.**
  - **转发，Forwarding： Incoming frames are forwarded according to MAC addresses that have been (and are being) learned.**

RSTP 原理★

- RSTP 原理:
  - 可以手工或自动判断边缘端口edge ports，边缘端口直接进入转发
  - 可以手工或自动判断point-to-point link和shared links
    - point-to-point link，采用应答机制
      - 可以在3个 hello times检测到根桥故障。
      - 拓扑变更时，通过握手，快速进入转发
      - 纪录替代端口信息，根端口出错直接切换
    - shared links，使用STP机制

⑩端口收到STP BPDU（非RSTP），停止使用RSTP

握手



designated switch

A superior BPDU

Root port

Figure 9-1 *Sequence of Events During RSTP Convergence*

握手

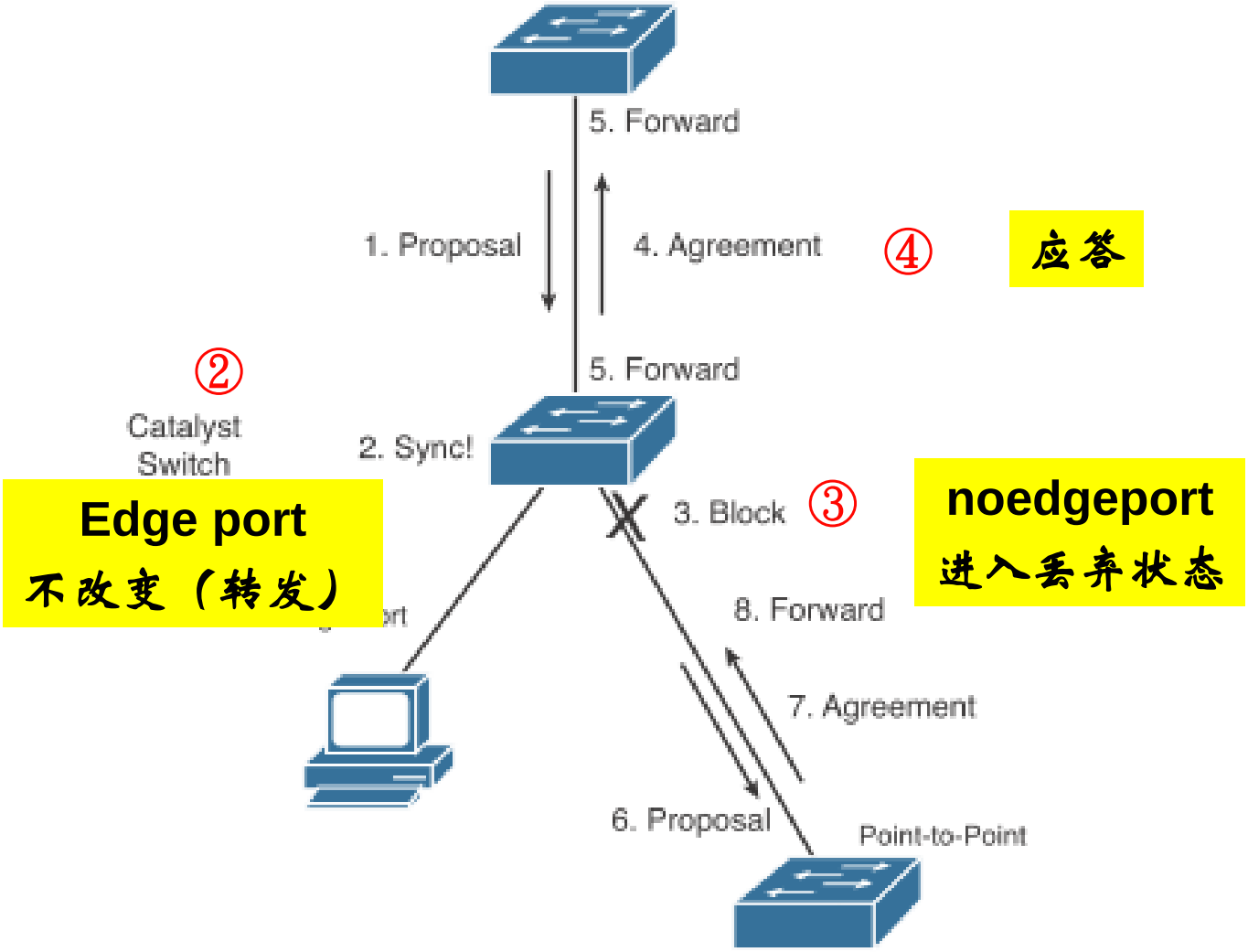


Figure 9-1 *Sequence of Events During RSTP Convergence*

握手

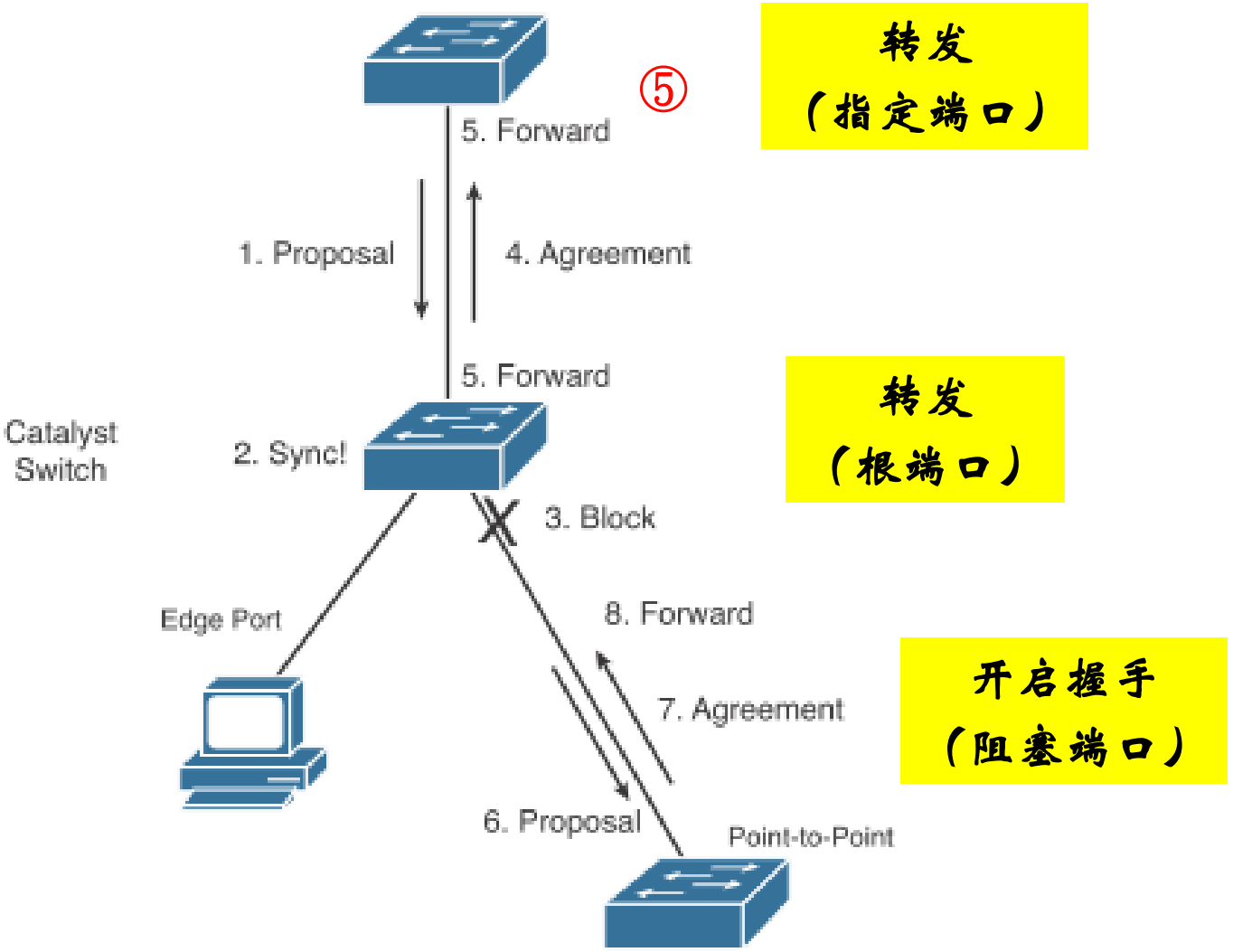


Figure 9-1 *Sequence of Events During RSTP Convergence*

握手级联

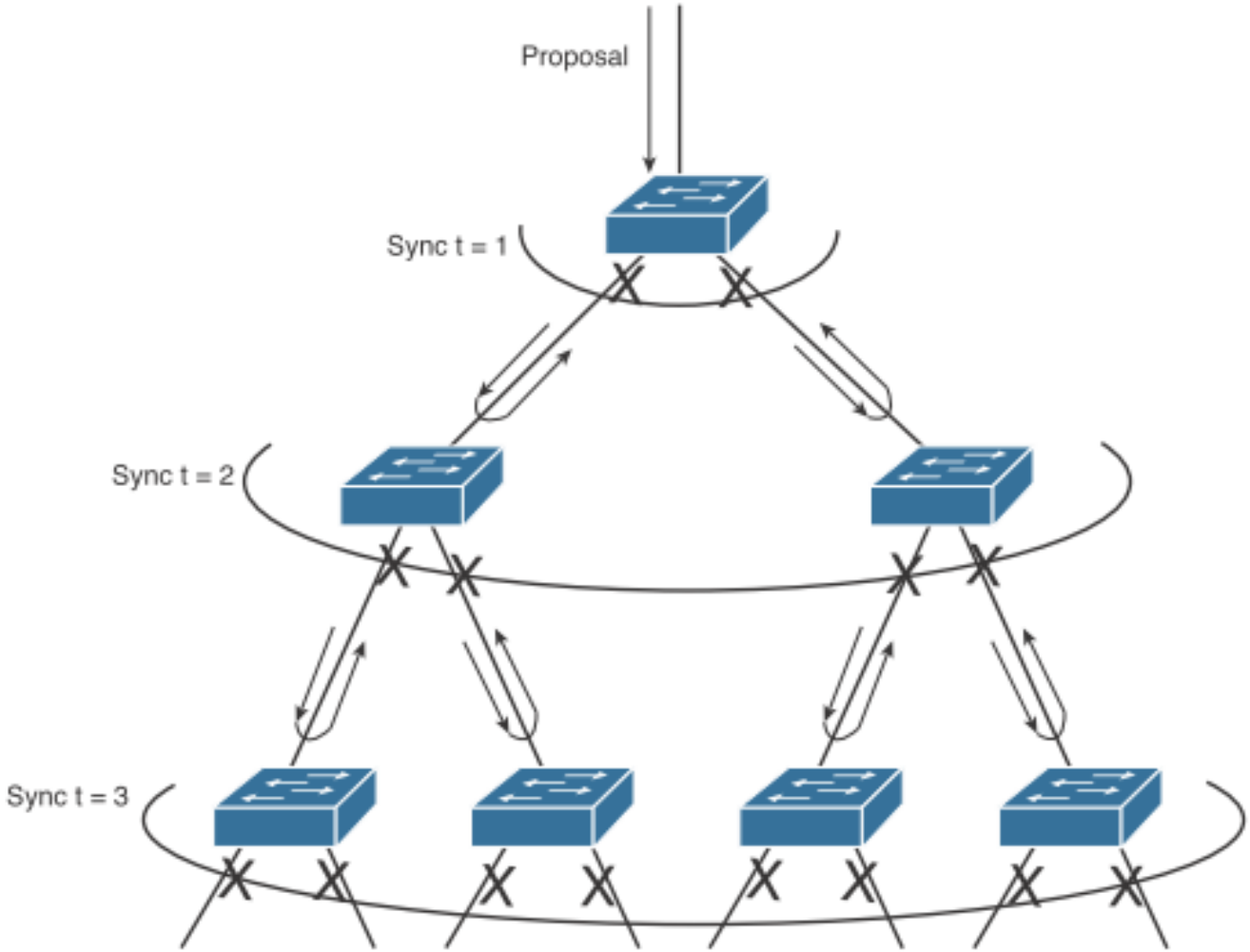
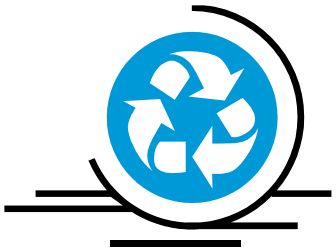


Figure 9-2 *RSTP Synchronization Traveling Through a Network*



私有VLAN

Private VLAN

单用户VLAN

- 通过给每个客户分配一个VLAN和相关的IP子网，可以使每个客户被从第2层隔离开，防止任何恶意的行为和Ethernet的信息探听。
- 这种方案的可扩展方面的局限：★
  - VLAN的限制：交换机固有的VLAN数目的限制；
  - 复杂的STP：对于每个VLAN，每个相关的Spanning Tree的拓扑都需要管理；
  - IP地址的紧缺：IP子网的划分势必造成一些IP地址的浪费；
  - 路由的限制：每个子网都需要相应的默认网关的配置。

PVLAN的基本概念

- 私有VLAN
  - 交换机上存在一个或多个primary vlan和多个secondary vlan。
  - 一个primary vlan包含几个secondary vlan，对于上层交换机只能见到primary vlan。
  - 一个primary vlan就是一个IP子网，即同一个primary vlan中包含的所有secondary vlan处在同一个子网中，节省了vlan资源。

PVLAN ★

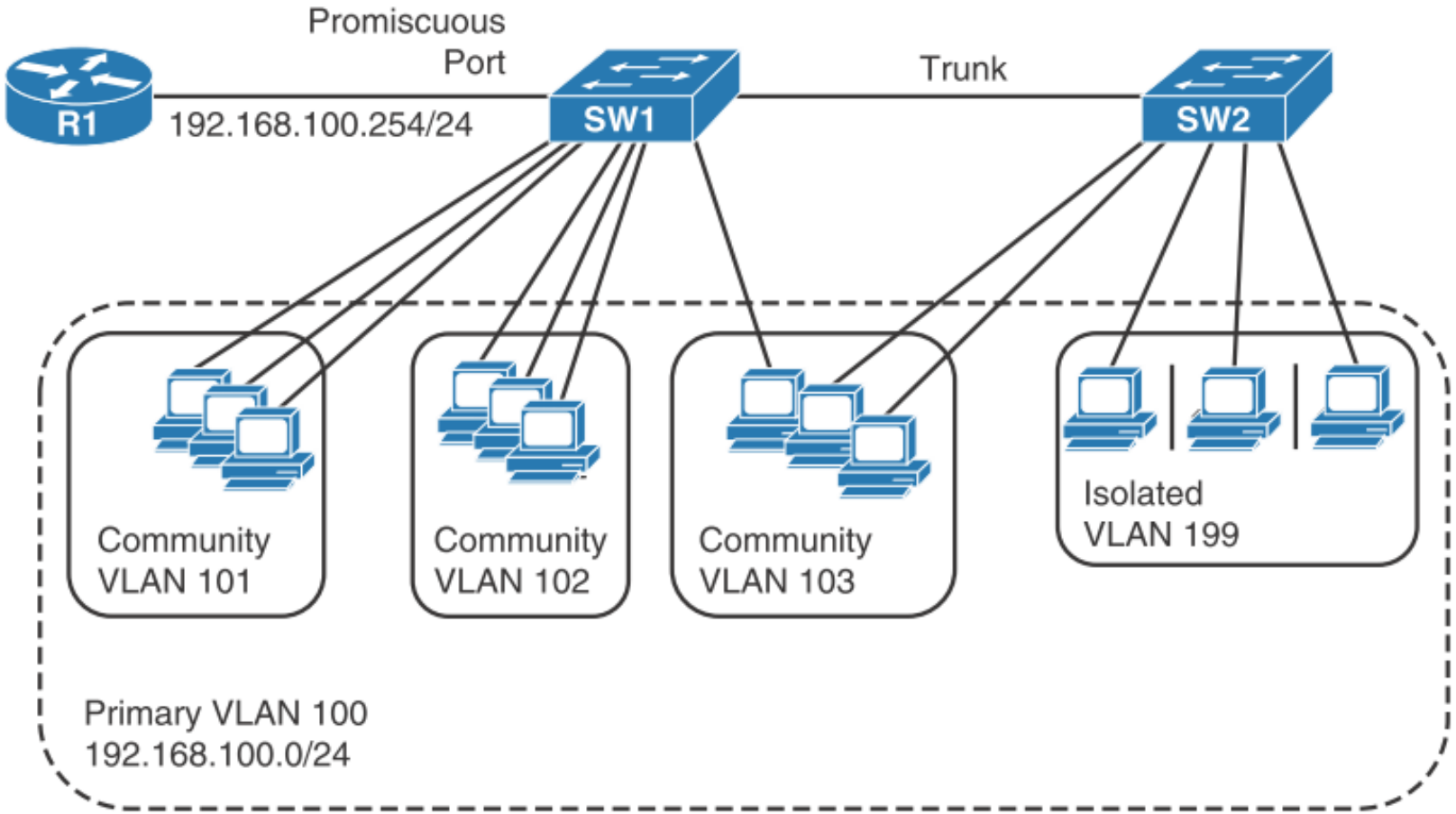


Figure 2-2 *Switched Network Utilizing Private VLANs*

端口功能★

■ 端口功能

Table 2-2 *Private VLAN Communications Between Ports*

Description of Who Can Talk to Whom	Primary VLAN Ports	Community VLAN Ports[1]	Isolated VLAN Ports[1]
Talk to ports in primary VLAN (promiscuous ports)	Yes	Yes	Yes
Talk to ports in the same secondary VLAN (host ports)	N/A[2]	Yes	No
Talk to ports in another secondary VLAN	N/A[2]	No	No
Talk to trunks	Yes	Yes	Yes

- **Isolated Port(隔离端口)、**
- **Community Port(团体端口)**
- **Promiscuous Port(混杂端口);**