Ethernet Switching

Easterhegg 2007

Falk Stern
falk@fourecks.de

Inhalt

- Was ist Ethernet?
- Was sind das für Bitmuster auf dem Draht?
- Wofür brauche ich aktive Komponenten?
- Was ist dieses “Spanning Tree Protocol”?
- Was sind Virtual Local Area Networks?

Inhalt

- **Was ist Ethernet?**
- Was sind das für Bitmuster auf dem Draht?
- Wofür brauche ich aktive Komponenten?
- Was ist dieses “Spanning Tree Protocol”?
- Was sind Virtual Local Area Networks?

Ethernet Standards

- 10Base5
  - "Yellow Cable"
- 10Base2
  - "Cheapernet", RG-58 Kabel mit 50Ω Terminatoren
- 10BaseT
  - Benötigt eine aktive Komponente oder Crossoverkabel

FastEthernet Standards

- 100BaseT
  - 2 verdrillte Adernpaare wie 10BaseT
- 100Base4
  - 4 verdrillte Adernpaare
- 100BaseVG
- 100BaseFX

Gigabit Standards

- 1000BaseT
- 1000BaseSX
- 1000BaseLX
- 1000BaseZX
- Über Glasfaser sind noch mehr Wellenlängen möglich

1000base-T

- Gigabit über Kupfer
  - weit verbreitet
  - erfordert Autonegotiation und MDIX
  - PHYs sind abwärtskompatibel

1000base-SX

- Multimode Glasfaser
- 850nm Wellenlänge
- Reichweite: bis zu 220m

1000base-LX

- Singlemode Glasfaser
- 1310nm Wellenlänge
- Reichweite: bis zu 2km

1000base-ZX

- Singlemode Glasfaser
- 1550nm Wellenlänge
- Reichweite: bis zu 70 Kilometer

Mehr Infos:

- http://en.wikipedia.org/wiki/Gigabit_Ethernet

TenGigabit Ethernet

- Glasfaser
  - 10Gbase-SR, -LR, -ER, -ZR, -LX4
- Kupfer
  - 10Gbase-T, 802.3an

10Gbase-SR

- “Short Range”
- Wellenlänge: 850nm
- Läuft über Multimode Glasfaser
- Reichweite zwischen 26 und 82 Metern

10Gbase-LR

- “Long Range”
- Wellenlänge: 1300nm
- Reichweiten zwischen 10 und 25 Kilometern

10Gbase-ER

- “Extended Range”
- Wellenlänge: 1550nm
- Reichweite bis zu 40 Kilometer

10Gbase-ZR

- Noch kein Standard
- Erweiteter -ER Transceiver
- Reichweiten bis zu 80 Kilometer möglich

10Gbase-LX4

- Integriertes Wavelength Division Multiplexing
- 10 Gigabit über existierende Multimodeverkablung
- 4 separate Laser mit unterschiedlichen Wellenlängen
- Reichweite zwischen 240 und 300 Metern

10Gbase-T

- 10 Gigabit über Kupfer
- Lange Zeit für “undenkbar” gehalten
- Mindestens Cat6a Kabel, besser Cat7
- DSQ128 Kodierung
  - 2 Dimensionales PAM-16
  - Pulse-Amplitude Modulation mit 16 Symbolen

Mehr Infos:

- http://en.wikipedia.org/wiki/10_gigabit_Ethernet

CSMACD (IEEE 802.3)

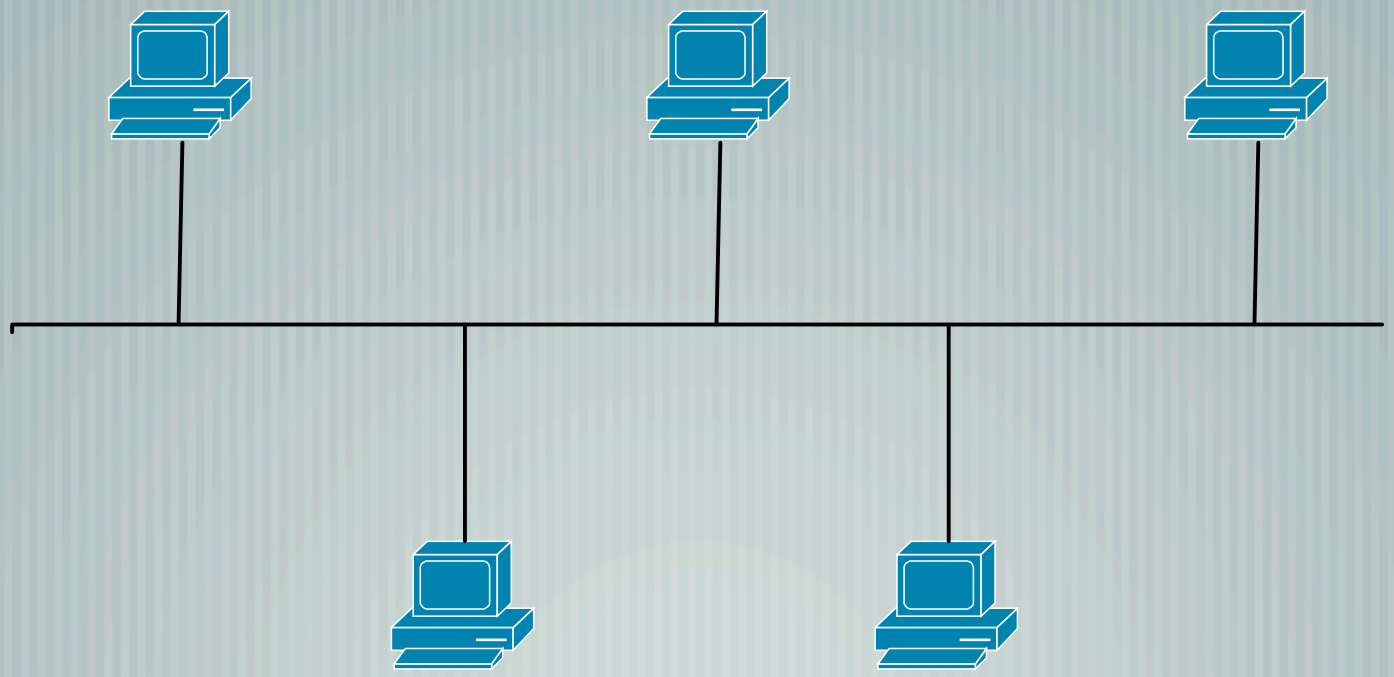
- Carrier Sense Multiple Access Collision Detection
- Jede Station erkennt, ob sie ans Netz angeschlossen ist
- Ethernet ist ein Broadcast Medium
- Kollisionen werden erkannt, Pakete nach einem zufälligen Intervall neu gesendet

Duplex

- Bei Half Duplex wartet die Station auf ein freies Zeitfenster auf dem Segment, bevor sie Daten sendet
- Full Duplex (gleichzeitiges Senden und Empfangen) ist nur mit aktiven Komponenten möglich

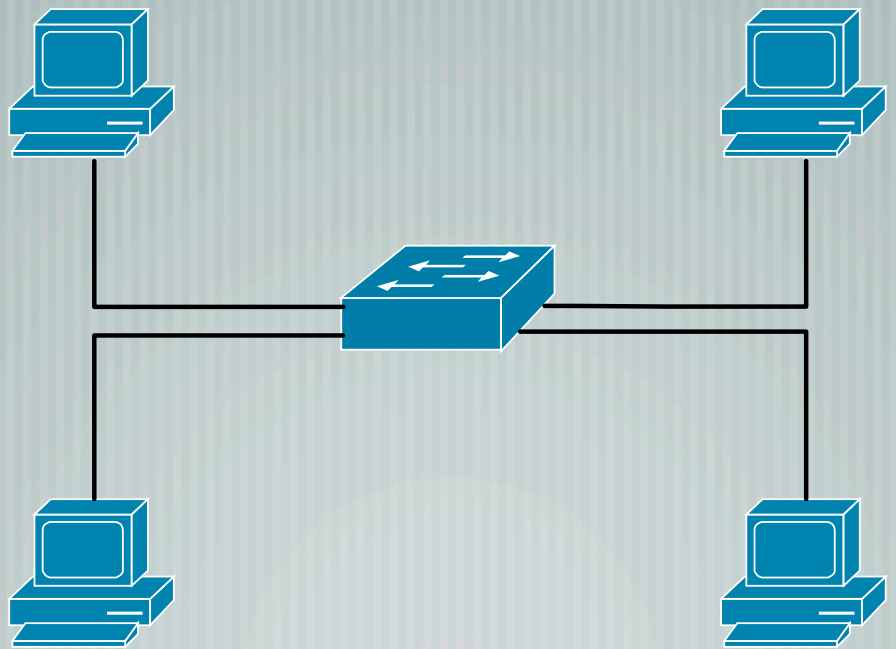
Topologien

Bus



Topologien

Stern



Aktive Komponenten

- Repeater
- Bridges
- Hubs
- Switches

Repeater

- Passive Netzwerkkomponente
- Verlängern die maximale Stranglänge
- Vergrößern Kollisionsdomänen
- Vergrößern Broadcastdomänen
- Es dürfen sich nicht mehr als 4 Repeater in einem Segment befinden

Bridges

- Aktive Netzwerkkomponenten
- Reduzieren Kollisionen in Netzwerksegmenten
- Trennen Kollisionsdomänen
- Vergrößern Broadcastdomänen

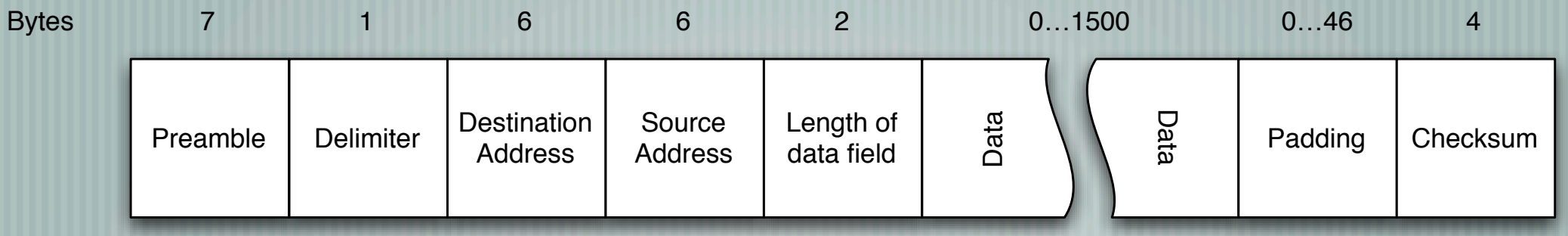
Hubs & Switches

- Hubs sind Multiport Repeater
- Switches sind Multiport Bridges

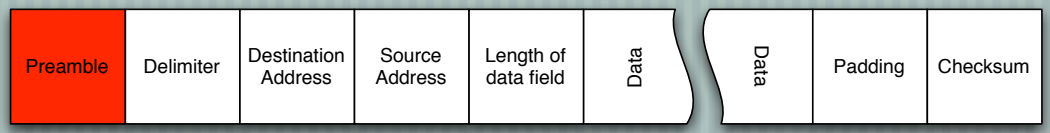
Inhalt

- Was ist Ethernet?
- **Was sind das für Bitmuster auf dem Draht?**
- Wofür brauche ich aktive Komponenten?
- Was ist dieses “Spanning Tree Protocol”?
- Was sind Virtual Local Area Networks?

Ethernet Frames

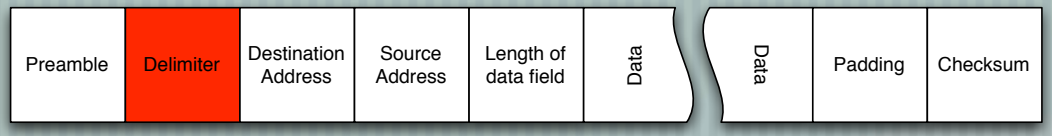


Frame: Präambel



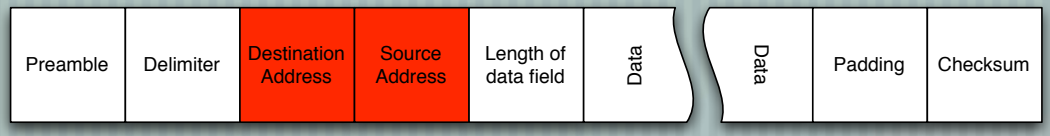
- Jeder Frame startet mit der 7 Byte langen Präambel
- Bitmuster: 10101010
- Dient zur Synchronisation von Sender und Empfänger

Frame: Delimiter



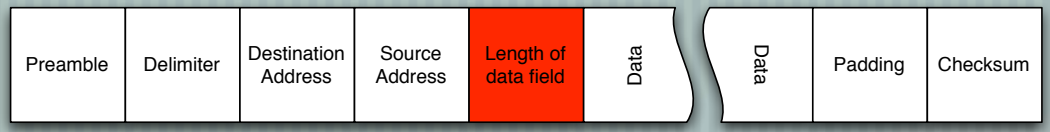
- Leitet den Beginn eines Frames ein
- Ab hier folgen Daten
- Bitmuster: 10101011

Frame: Adressen



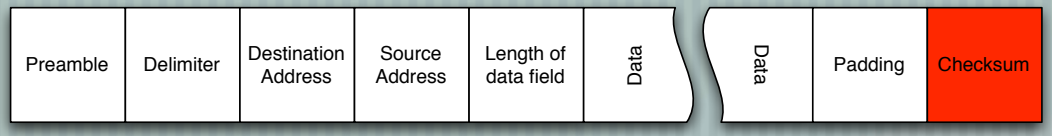
- Adressen sind 6 Byte lang
- Standard erlaubt 2 Byte lange Adressen, werden nicht benutzt
- Wenn das höchste Bit 1 ist folgt ein Multicastframe
- Wenn alle Bits 1 sind folgt ein Broadcastframe

Frame: Length of Datafield



- Maximale Datenlänge bei IE^3 802.3 sind 1500 Bytes
- Minimale Framelänge sind 64 Bytes
  - Dient zur Unterscheidung bei Kollisionen
- Sollten die Daten weniger als 46 Bytes sein, wird im Padding aufgefüllt.
- Ethernet benutzt das Feld als “Type” Feld

Frame: Checksumme



- Die Checksumme wird als 32 Bit Hashwert aus dem Datensegment berechnet
- Bildet einen Cyclic Redundancy Check
- Bei Fehlern wird das Paket verworfen

Inhalt

- Was ist Ethernet?
- Was sind das für Bitmuster auf dem Draht?
- **Wofür brauche ich aktive Komponenten?**
- Was ist dieses “Spanning Tree Protocol”?
- Was sind Virtual Local Area Networks?

Begriffe

- Runts (Zwerg, Ferkel)
  - Frames < 64 Byte
- Baby Giants
  - Frames zwischen 1518 und 1522 Bytes
- Giants
  - Frames > 1522 Bytes

Begriffe

- ASIC
  - Application Specific Integrated Circuit
- Jumboframes
  - Die Gigabit Ethernet Spezifikation 802.3z erlaubt Framegrößen bis zu 9000 Bytes

Warum Switches?

- Switches verringern die Latenz, die auftritt, wenn eine Station warten muß, bis sie senden kann.
- Zwischen Switch und Station entsteht eine eigene Kollisionsdomäne
- Fehlerhafte Stationen können nicht das ganze Netz blockieren

Switching Strategien

- Store and Forward
- Fragment Free
- Cut Through

Store and Forward

- Der gesamte Frame wird empfangen, zwischengespeichert und auf Gültigkeit geprüft
- Langsamste Switching Strategie, da der komplette Frame im Buffer des ASICs zwischengespeichert werden muß

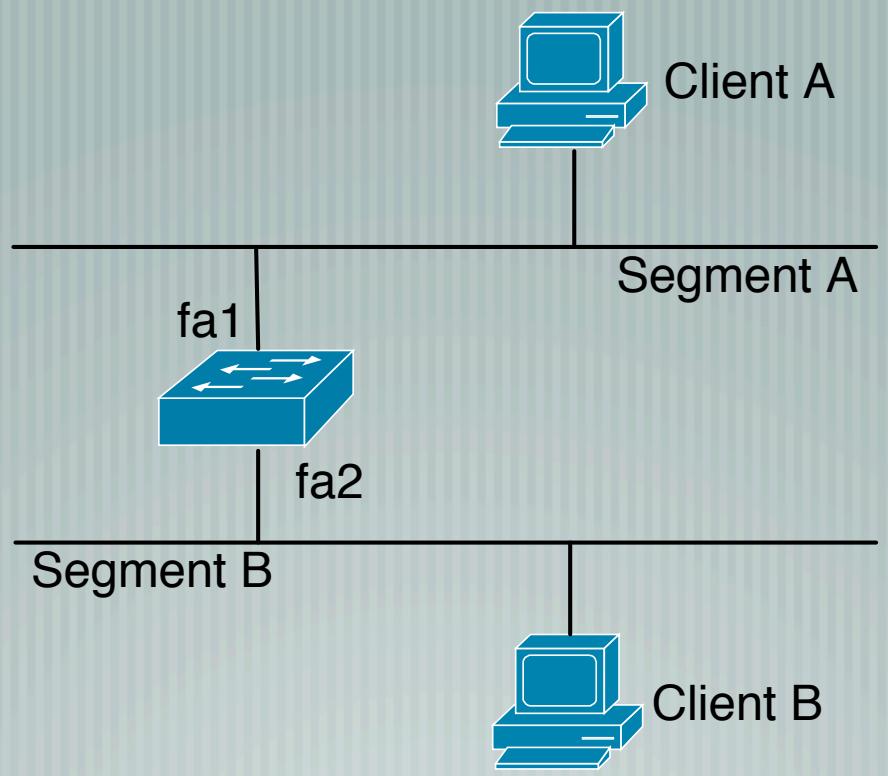
Fragment Free

- Hier werden nur die ersten 64 Byte des Frames auf Gültigkeit geprüft, Runts werden sofort verworfen
- Die Checksumme wird geprüft, nachdem der Frame übertragen wurde
- Bei fehlerhafter Checksumme wird ein Fehlerzähler inkrementiert, der nach Überschreiten einen Schwellenwertes auf Store and Forward zurückschaltet

Cut Through

- Der Frame wird, sobald die Zieladresse empfangen wurde, in den Ausgangspuffer des ASICs geschrieben
- Checksumme wird wie bei Fragment Free geprüft

Transparentes Bridging



Ein Switch verhält sich wie eine transparente Bridge

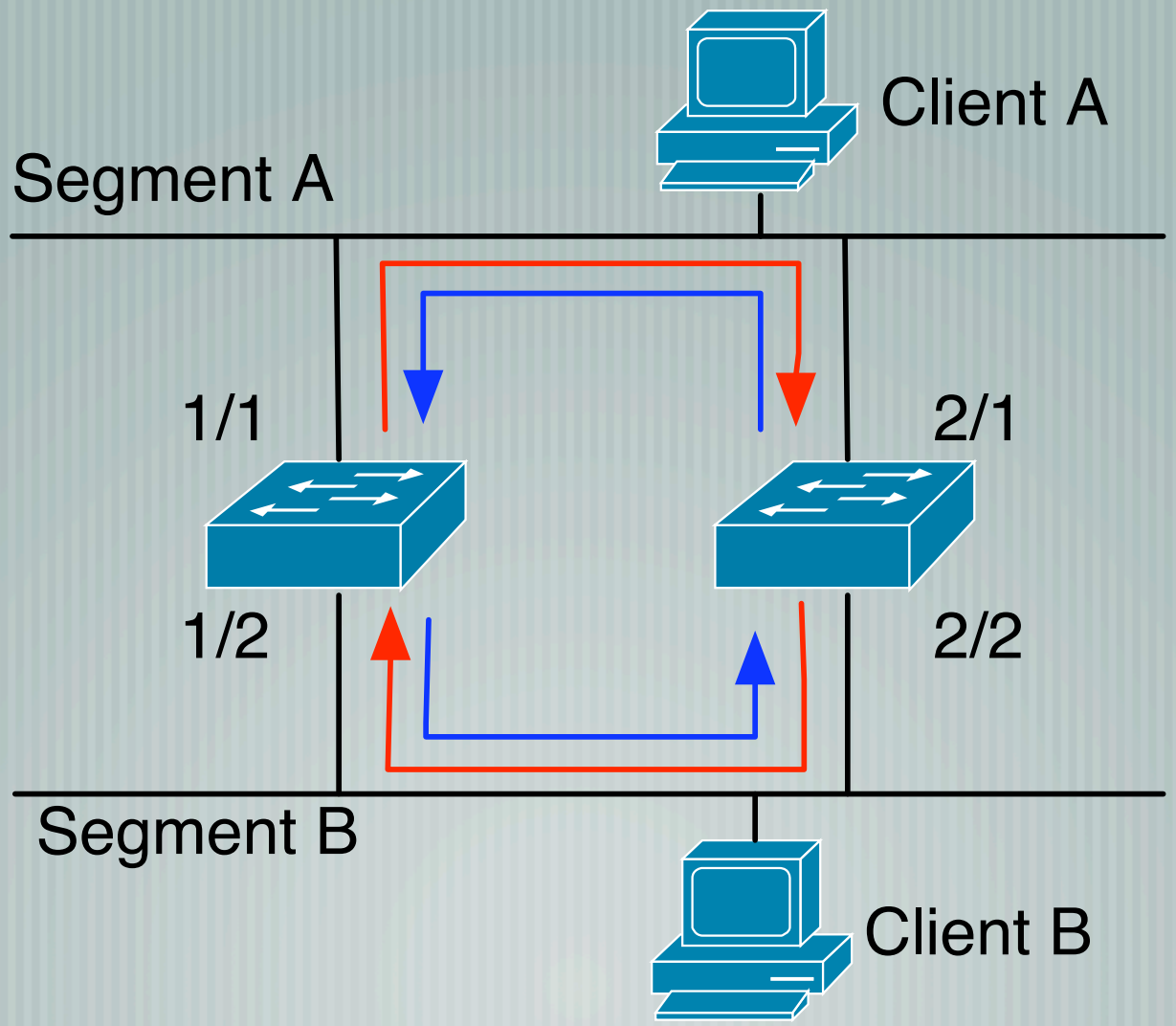
Transparentes Bridging

- Dürfen weitergeleitete Frames nicht modifizieren
- Lernen MAC Adressen durch “hören” auf den Ports
  - Aufbau einer Bridge Table
  - Auf einem Port gelernte Adressen können auch durch diesen Port wieder erreicht werden
- Eine Bridge “lernt” und “hört” immer

Transparentes Bridging

- Broadcasts müssen an alle Ports außer dem eingehenden weitergeleitet werden
- Bei unbekannter Zieladresse wird der Frame auf allen Ports außer dem eingehenden ausgegeben ("Flooding")
- Sobald ein redundanter Pfad dem Netzwerk hinzugefügt wird, kommt es zu Problemen

Bridgeschleifen



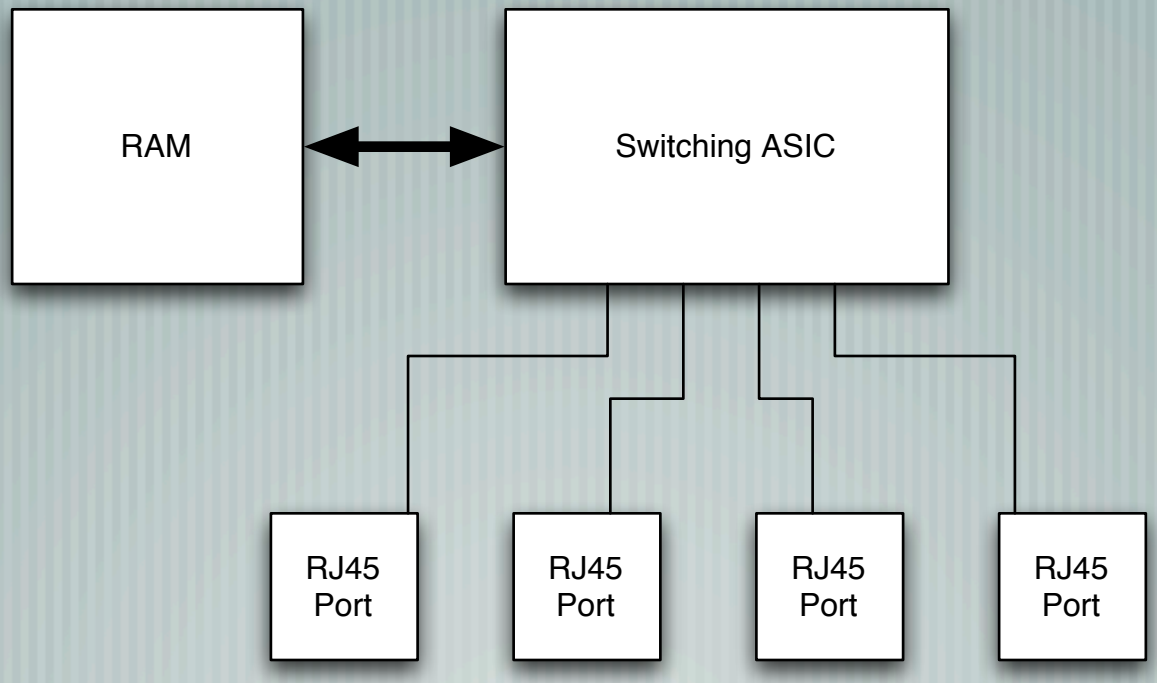
Bridgeschleifen

- Die Bridges sehen einen Frame von Station A auf 1/1 und 2/1
- Dieser Frame wird auf 1/2 und 2/2 weitergeleitet
- Da die Bridges nichts voneinander wissen, sehen beide die MAC-Adresse von Station A auf Segment B, da sie den Frame vom jeweils anderen Switch bekommen
- Der Frame wird jetzt wieder auf Segment A gekippt

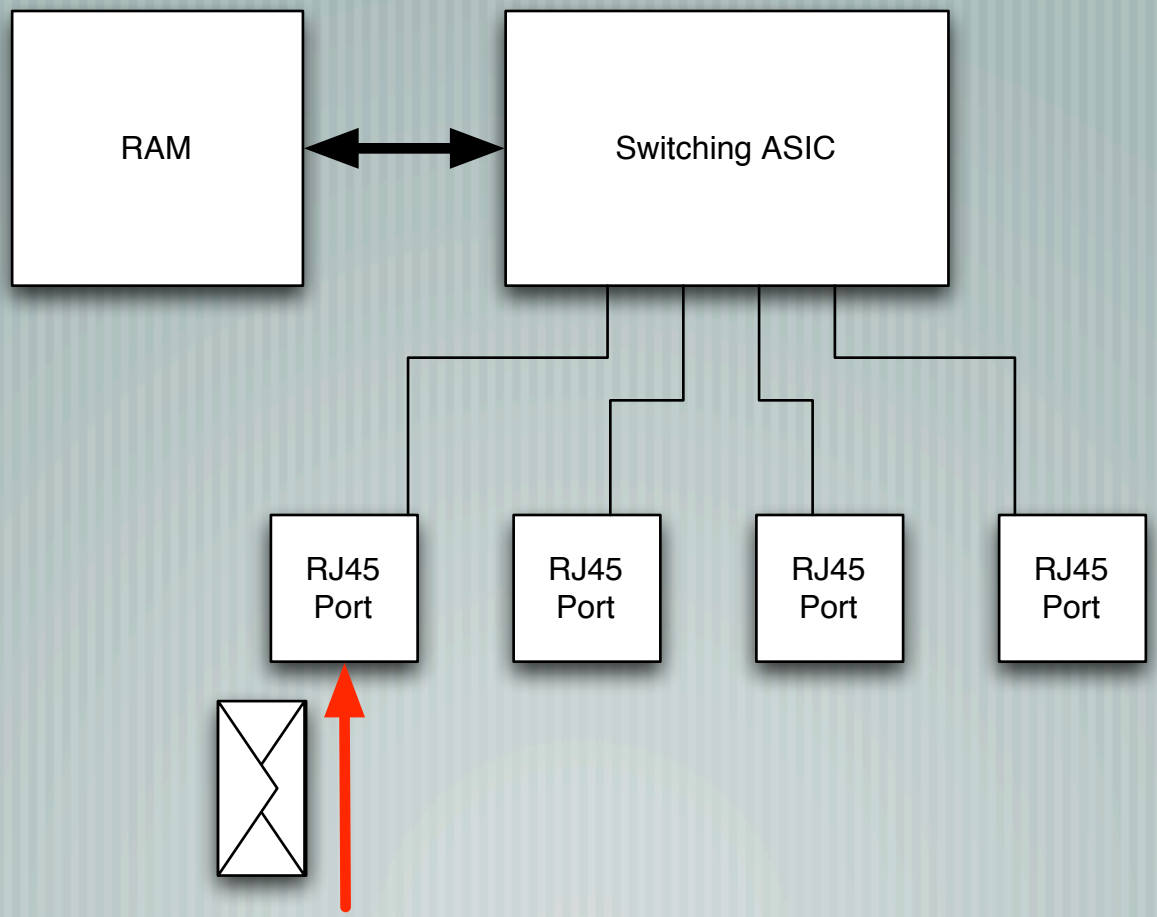
Switcharchitektur

- Beispiele
  - “dummer” Switch
  - Cisco 3500XL

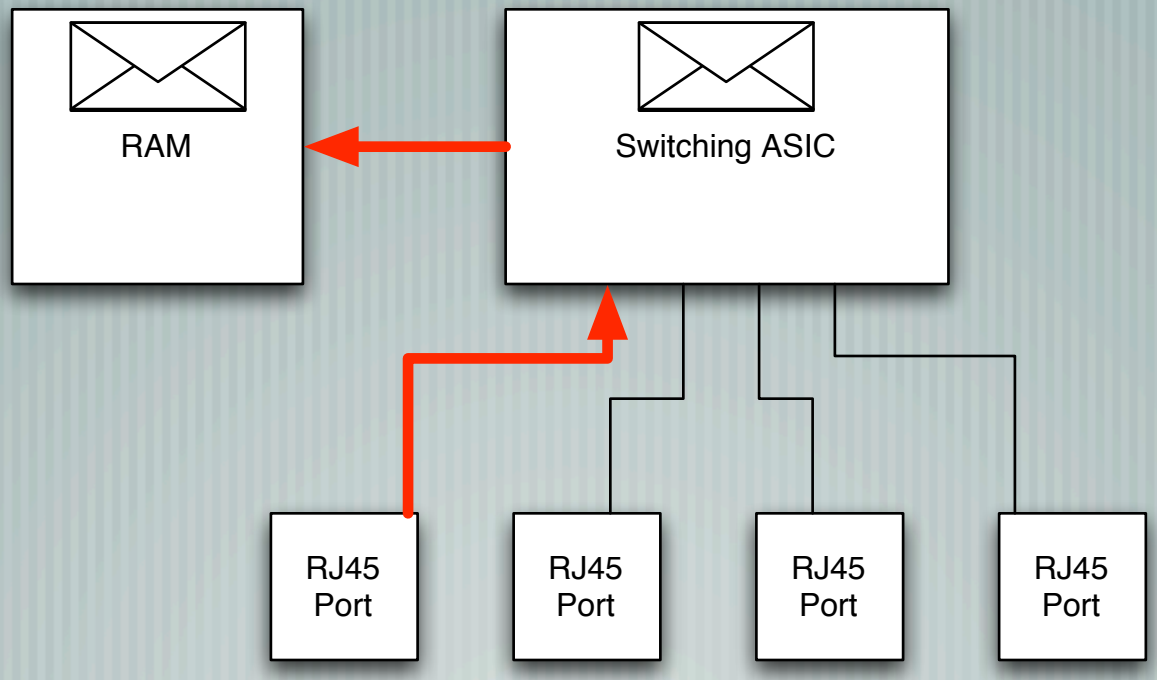
"Dummer" Switch



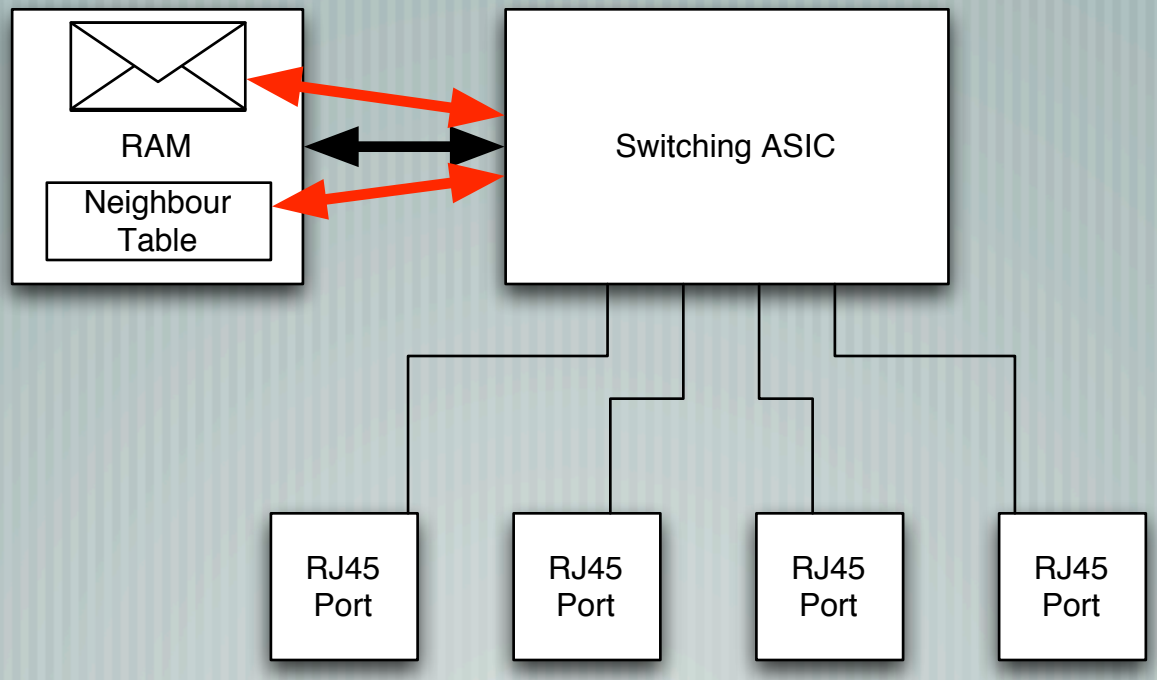
Receive



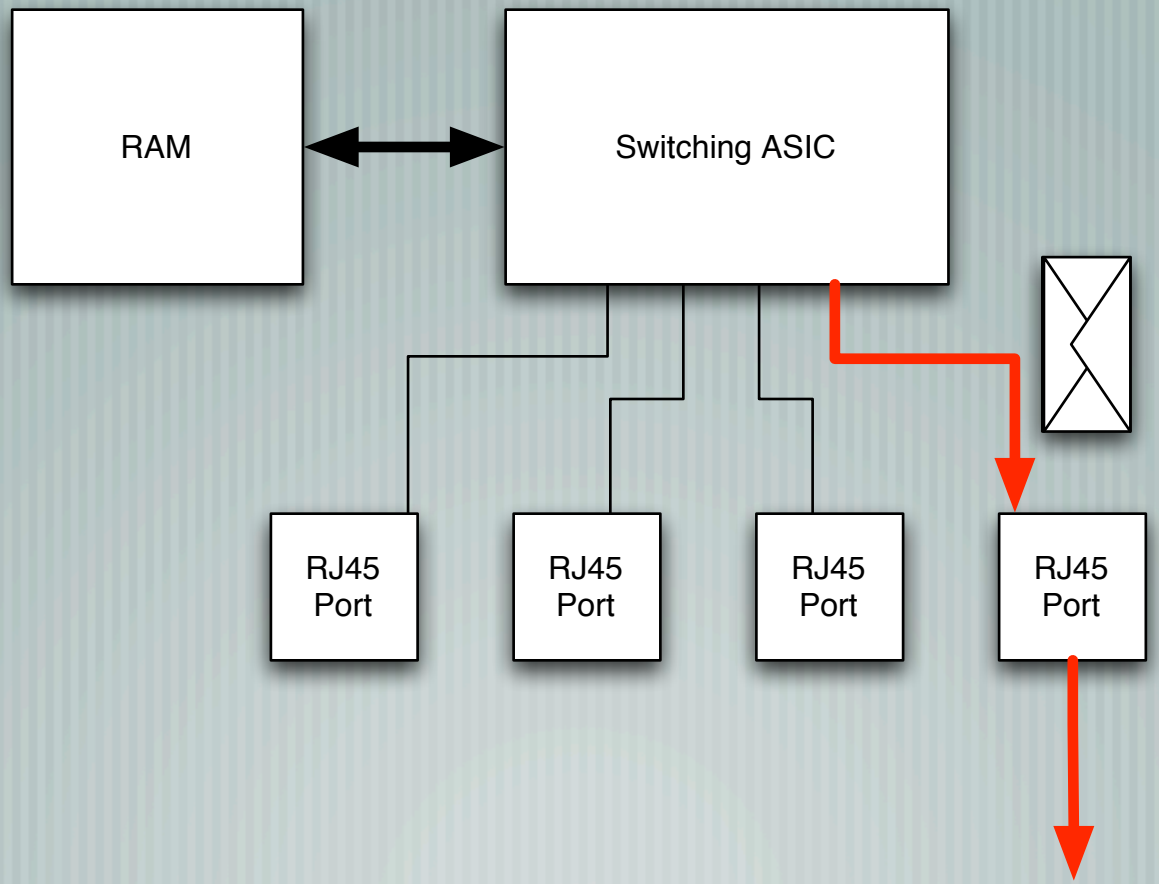
Store



Lookup


RAM
Neighbour Table
Switching ASIC
RJ45 Port
RJ45 Port
RJ45 Port
RJ45 Port

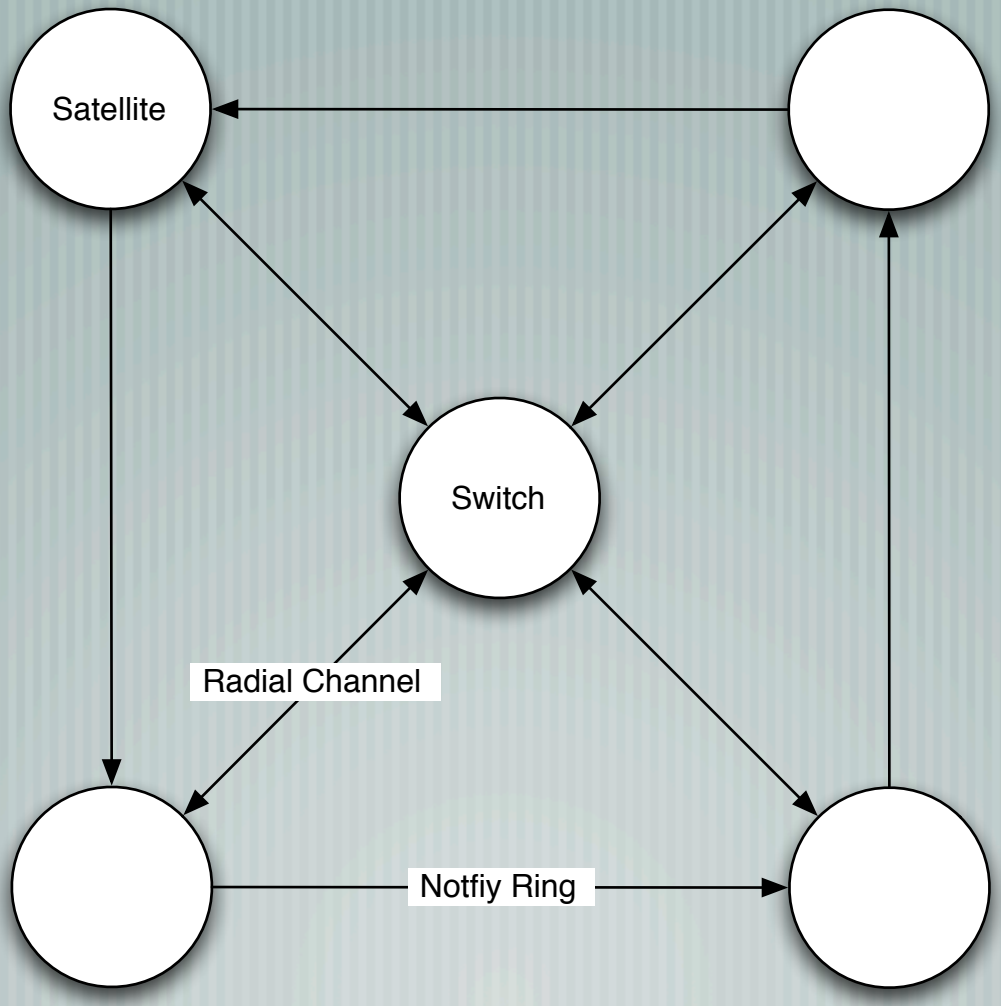
Forward



Cisco Catalyst XL

- Hier als Beispiel, weil gut dokumentiert
- Besteht aus mehreren ASICs
- Deutlich komplexere Architektur

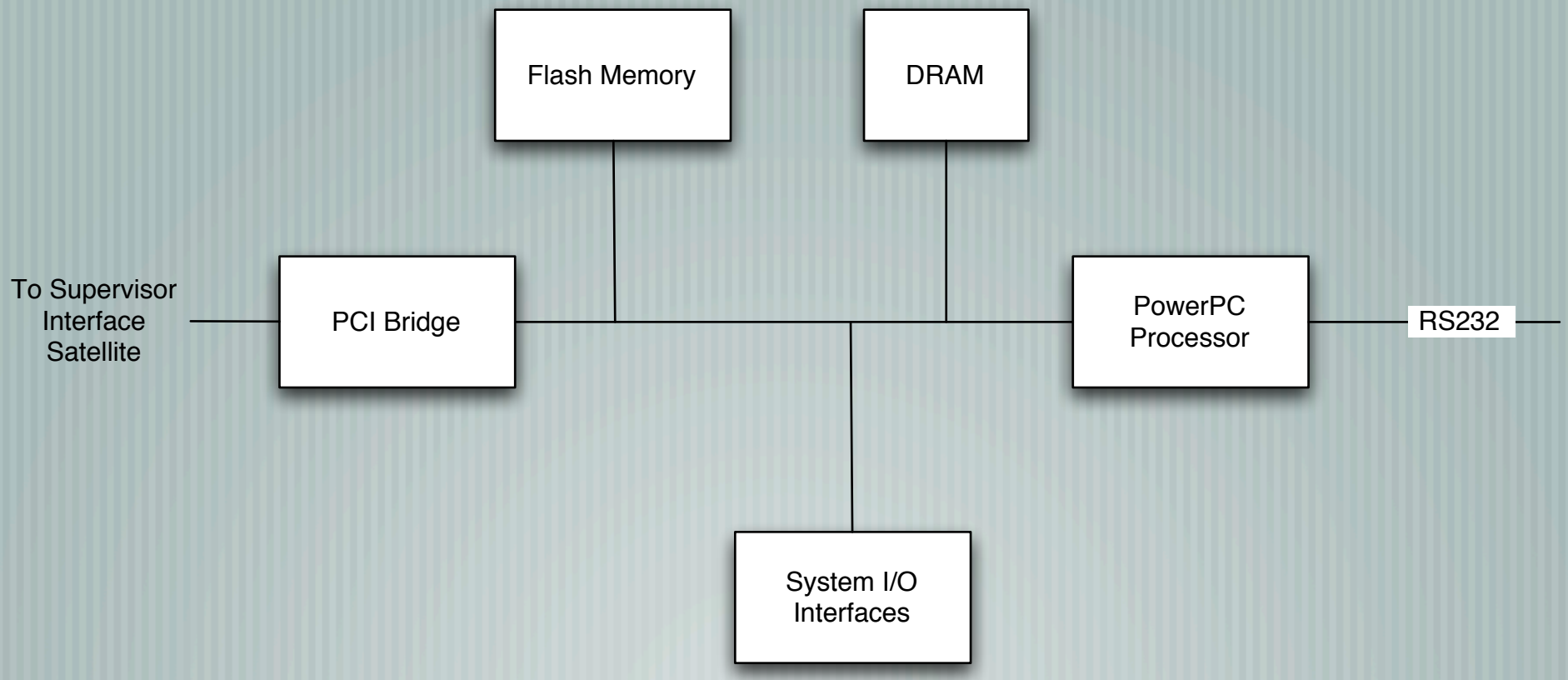
Cisco XL Architektur



Cisco XL Architektur

- Satelliten kommunizieren untereinander mit einem Bus
- Die Switching Fabric greift auf einen Shared Memory Buffer zu, der allen Satelliten zur Verfügung steht
- Satelliten sind mit bis zu 8 Kanälen mit der Fabric verbunden
- 1 “Radial Channel” hat 200Mbps pro Richtung, 160 Mbps netto

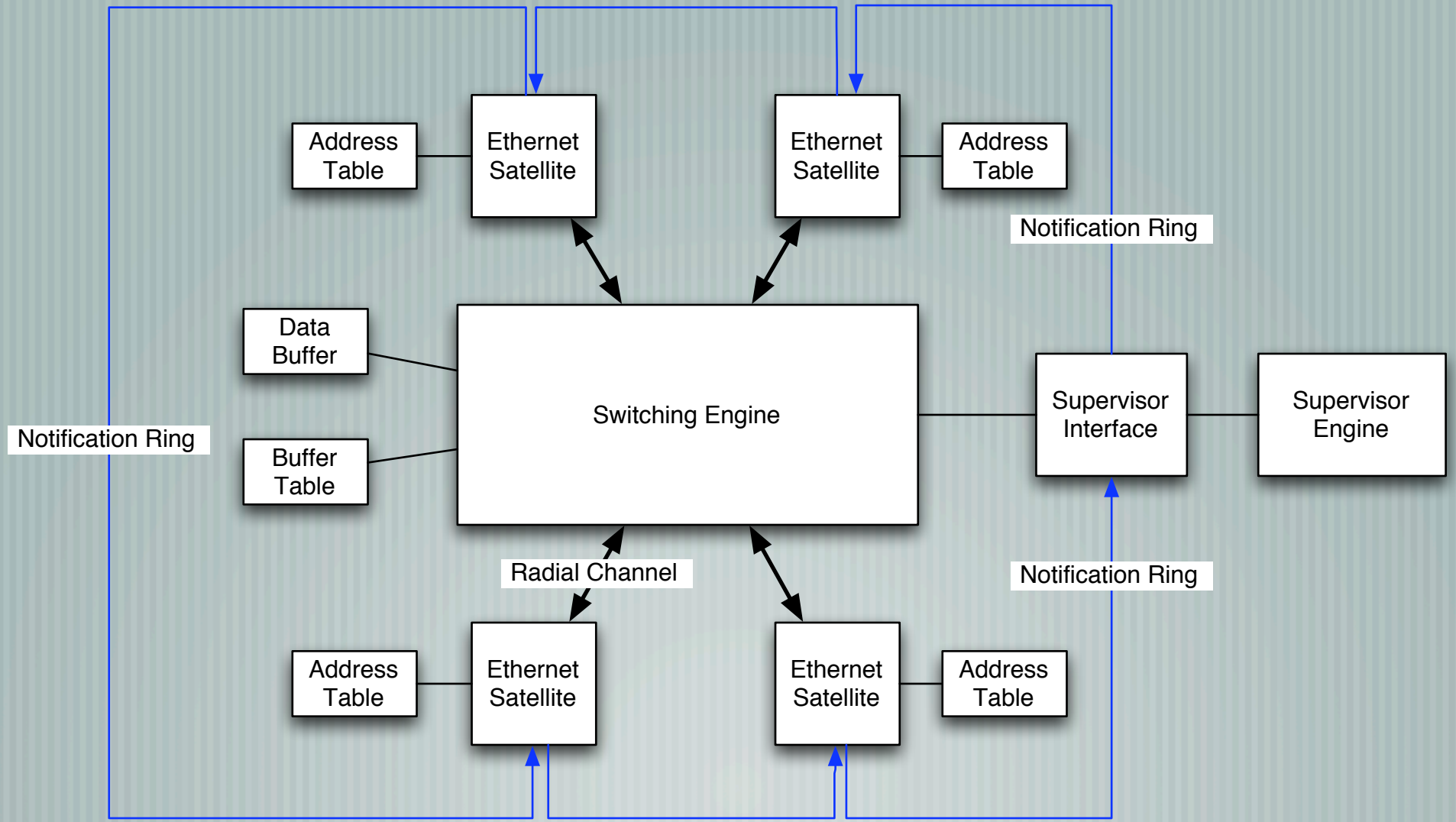
Cisco XL: Supervisor Engine



Cisco XL: Supervisor Engine

- Programmiert Adress- und VLANtabellen in die Satelliten
- Übernimmt keine Switching-Funktionalität
- Steuert Lüfter, RPS, Diagnosefunktionen
- Übernimmt Managementaufgaben (VLANs, STP)

Cisco XL: Switching Engine



Cisco XL Switching Engine

- Verwaltet den zentralen Datenpuffer
- Kontrolliert Pakete sobald sie in den zentralen Puffer geschrieben werden
- Beim Einlesen des Pakets wird ein temporärer Eintrag in der Buffer Table erstellt
- Daten werden immer in derselben Zellengröße gelesen und geschrieben

Cisco XL Switching Engine

- Bandbreite zwischen Engine und Puffer: 10 Gbps Vollduplex
- Shared Buffer erlaubt hohe Bandbreite, niedrige Latenz
- Höhere Geschwindigkeit gegenüber Per-Port-Buffers
- Quellsatellit informiert Zielsatellit über Notification Ring
- Kann der Zielsatellit das Paket nicht zustellen, verwirft die Switching Engine das Paket

Inhalt

- Was ist Ethernet?
- Was sind das für Bitmuster auf dem Draht?
- Wofür brauche ich aktive Komponenten?
- **Was ist dieses “Spanning Tree Protocol”?**
- Was sind Virtual Local Area Networks?

Schleifenvermeidung

- Der einzige Weg zur Schleifenvermeidung sind “intelligente” Geräte
- Für redundante Anbindungen sind “Schleifen” notwendig
- Dafür wurde das “Spanning Tree Protocol” entwickelt
  - Standardisiert nach IEEE 802.1d

Spanning Tree Protocol

- Wurde entwickelt um Bridging-Schleifen in einem Netzwerk mit redundanten Pfaden zu vermeiden
- Um Spanning Tree und Switches zu verstehen, sollte man transparentes Bridging kennen

Spanning Tree Protocol

- Switches tauschen untereinander Topologiedaten aus
- Sie wählen eine “Rootbridge” um eine Baumstruktur aufzubauen
- Eine Grundlage zur Berechnung des Baums sind die Geschwindigkeiten der Links untereinander
  - Je geringer die Kosten, desto besser der Link

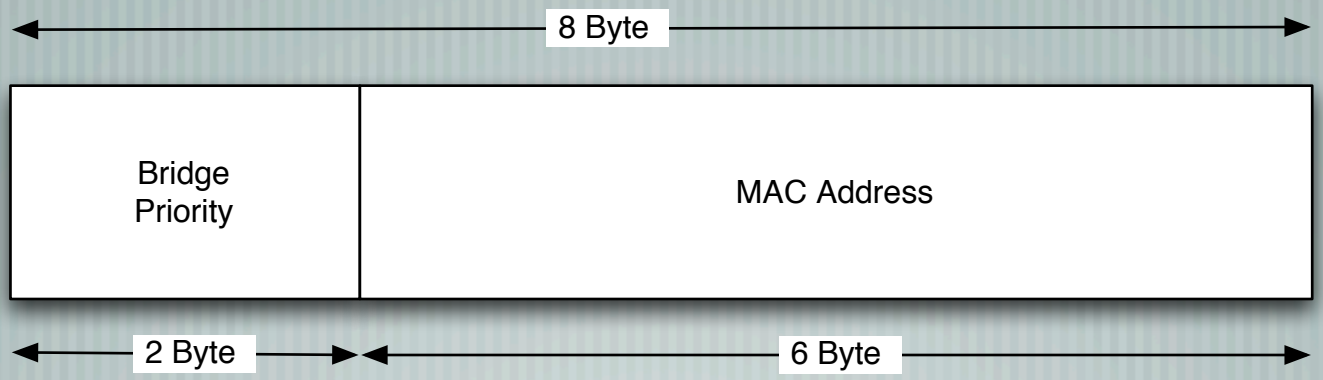
Spanning Tree Protocol

- Ports durchlaufen mehrere Stadien bis sie auf “Forwarding” geschaltet werden
- Die Rolle eines Switchports wird anhand empfangener BPDUs entschieden
  - Bridge Protocol Data Unit
- Switches senden BPDUs aus, in denen die MAC-Adresse der Rootbridge, die Pfadkosten und die Priorität enthalten sind

STP: Wer ist Rootbridge?

- Rootbridge wird der Switch mit
  - der niedrigsten Priorität (Standardwert: 32768)
  - der niedrigsten MAC-Adresse
- also der niedrigsten Bridge-ID

STP: Bridge ID



STP: Pfadkosten

Bandbreite	Kosten (neu)	Kosten (alt)
10 Gbps	2	1
1 Gbps	4	1
100 Mbps	19	10
10 Mbps	100	100

Mit dem weiten Gebrauch von 1 Gbps Links wurde die 802.1d Spezifikation erweitert um auch höhere Bandbreiten zu berücksichtigen

STP: Begriffe

- Root Port
  - Port, der die niedrigsten Pfadkosten zur Rootbridge hat
- Designated Port
  - Port mit den niedrigsten Kosten zur Rootbridge im Segment
- Nondesignated Port
  - Geblockter Port

STP: Definitionen

- Ein Rootport pro Bridge
- Ein Designated Port pro Netzsegment
- Eine Rootbridge im Netzwerk

STP: Designated Port Wahl

- Designated Port wird
  - der Port mit den niedrigsten Pfadkosten zur Rootbridge
  - bei gleichen Pfadkosten gewinnt die niedrigste Bridge ID

STP: Portzustände

- Blocking
- Listening
- Learning
- Forwarding
- Disabled

STP: Blocking

- Blocking
  - In diesem Zustand nimmt der Port nicht am Layer 2 Forwarding teil

STP: Listening

- Listening
  - Der erste Übergangszustand nachdem durch den Spanning Tree Algorithmus entschieden wurde, daß der Port am Layer 2 Forwarding teilnehmen soll

STP: Learning

- Learning
  - In diesem Zustand bereitet sich der Port darauf vor, am Layer 2 Forwarding teilzunehmen
  - In diesem Zustand werden bereits auf dem Netzsegment vorhandene MAC-Adressen gelernt

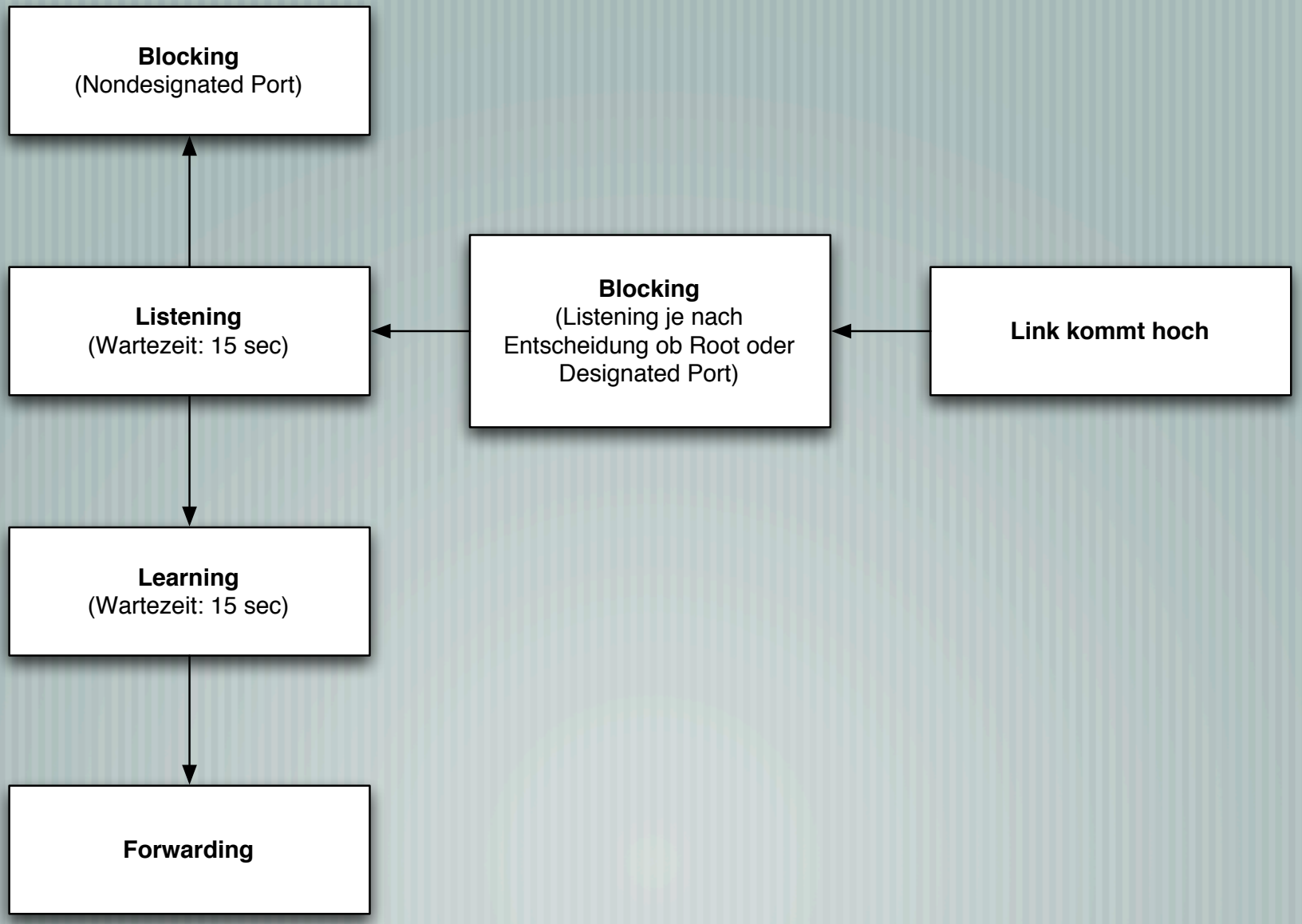
STP: Forwarding

- Forwarding
  - In diesem Zustand arbeitet der Port “normal”
  - Er nimmt ganz normal am Bridging-Prozess teil

STP: Disabled

- Disabled
  - In diesem Zustand nimmt der Port nicht am Layer 2 Forwarding und am Spanning Tree teil
  - Der Port ist “tot”

STP: Zustandsdiagramm



STP: BPDU Format

Bytes	Feld
2	Protocol ID
1	Version
1	Message Type
1	Flags
8	Root ID
4	Cost of Path
8	Bridge ID
2	Port ID
2	Message Age
2	Maximum Time
2	Hello Time
2	Forward Delay

BPDUs werden alle 2 Sekunden versandt

Werden nicht an die Rootbridge geschickt

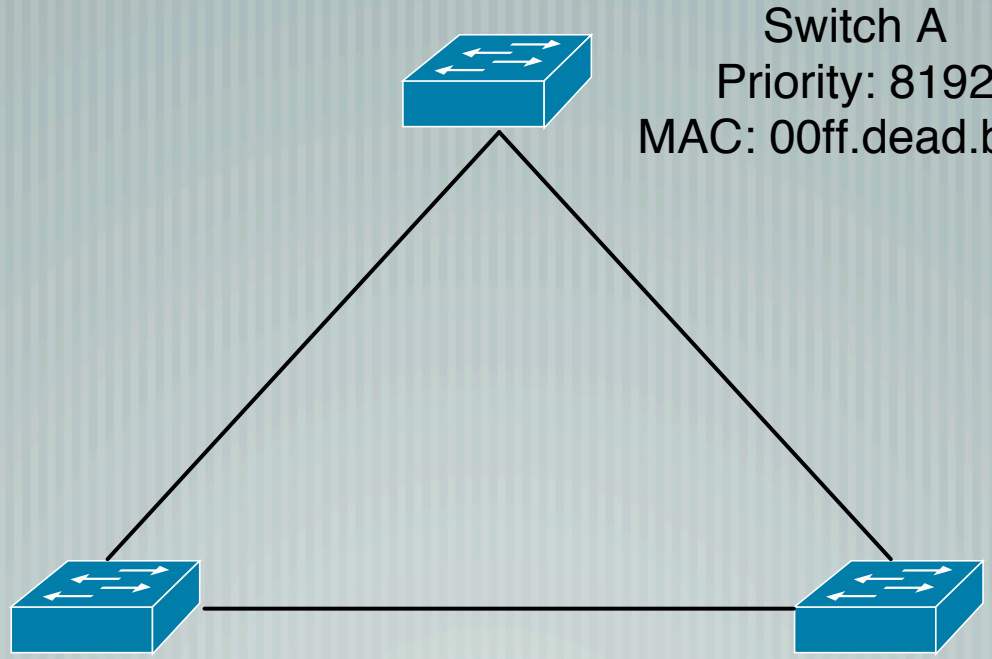
Bei Topologieänderungen werden TCN BPDUs geschickt

TCN: Topology Change Notification

STP: Topology Change

- TCNs werden verschickt, wenn
  - Ein Link ausfällt (von Forwarding/Learning zu Blocking)
  - Ein Port “Forwarding” wird und ein “Designated Port” existiert
  - Ein TCN auf einem “Designated Port” empfangen wird
- TCNs werden mit “Topology Change Acknowledge” beantwortet

STP: Wer wird Rootbridge?

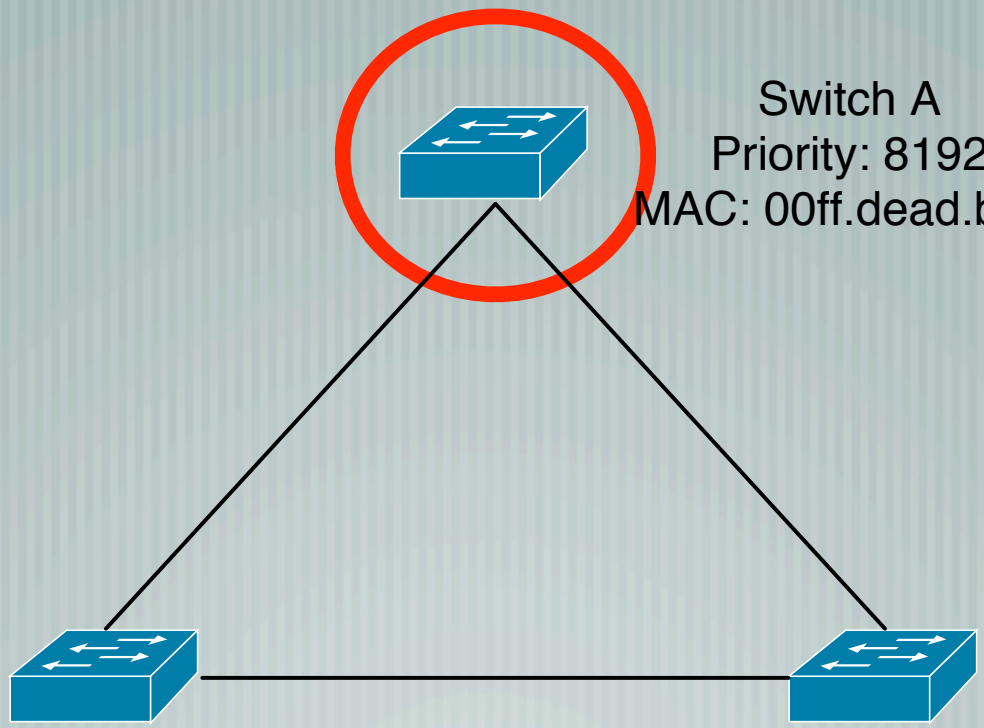


Switch A
Priority: 8192
MAC: 00ff.dead.beef

Switch B
Priority: 32768
MAC: 00cc.dead.beef

Switch B
Priority: 32768
MAC: 00aa.dead.beef

STP: Wer wird Rootbridge?



Switch A
Priority: 8192
MAC: 00ff.dead.beef

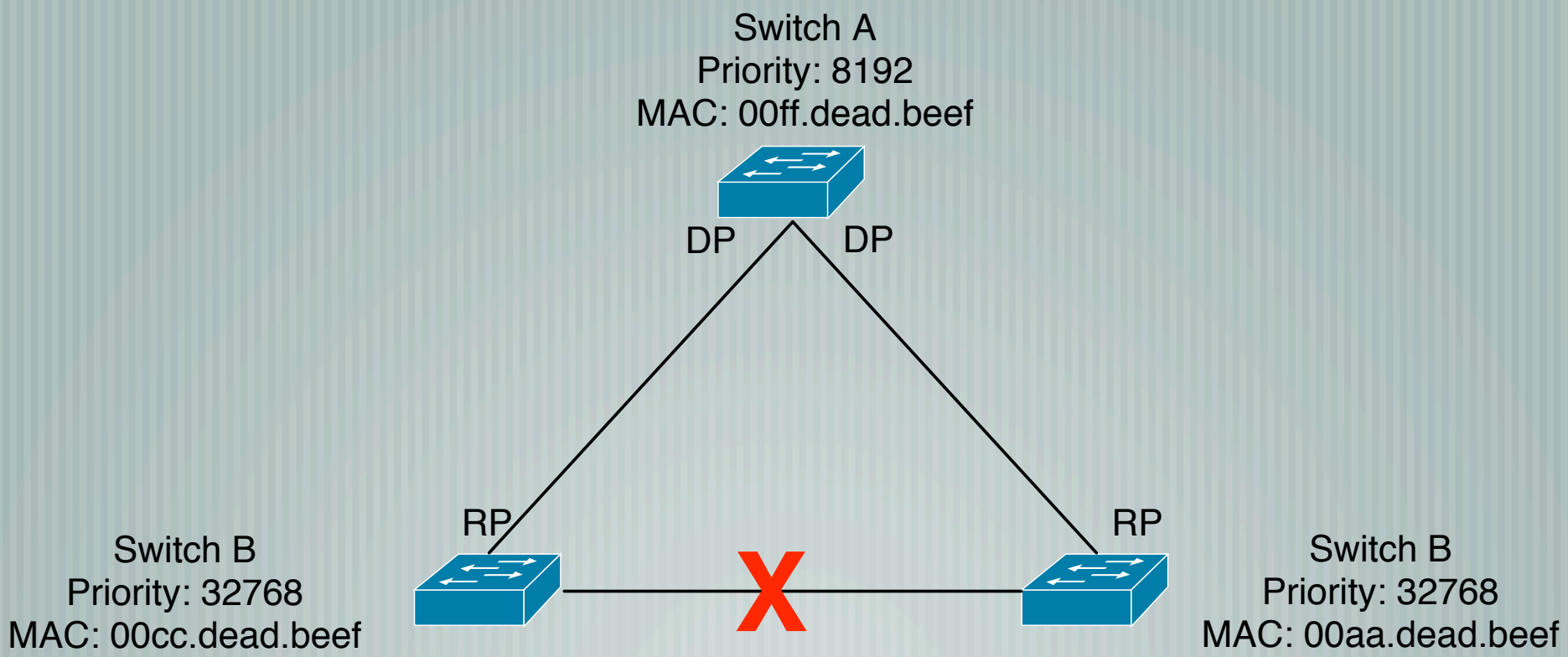
Switch B
Priority: 32768
MAC: 00cc.dead.beef

Switch B
Priority: 32768
MAC: 00aa.dead.beef

STP: Zustandsentscheidung

- Wenn STP mehr als zwei Wege zur Rootbridge kennt wird anhand dieser Kriterien entschieden, welcher Port Rootport wird
- Niedrigste Rootbridge ID
- Niedrigste Pfadkosten zur Rootbridge
- Niedrigste Sender BID
- Niedrigste Port ID

STP: Zustand


Switch A
Priority: 8192
MAC: 00ff.dead.beef
DP
DP
RP
RP
X
Switch B
Priority: 32768
MAC: 00cc.dead.beef
Switch B
Priority: 32768
MAC: 00aa.dead.beef

STP: Rapid Spanning Tree

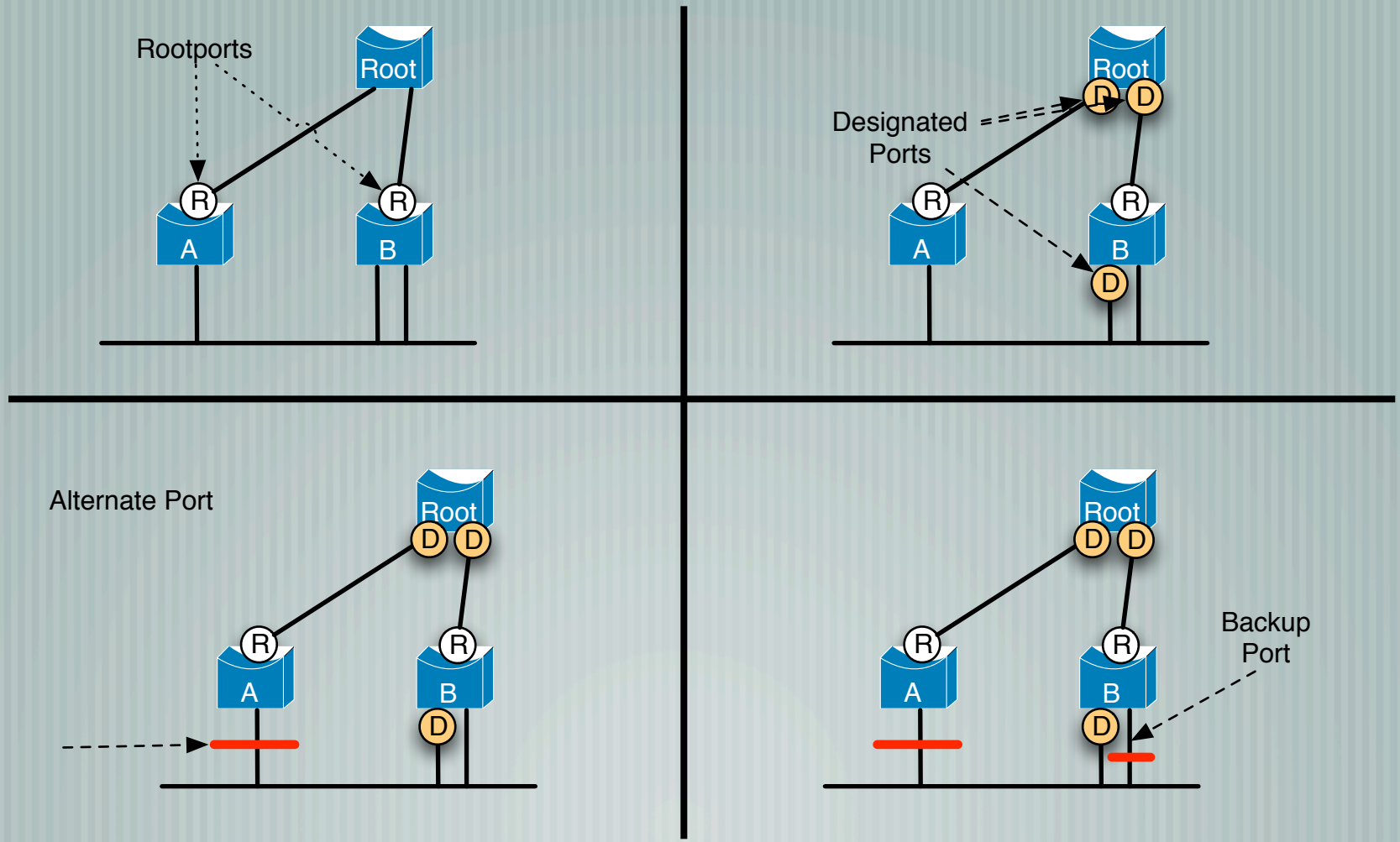
- Schnellere Variante des Spanning Tree Protocols
- IEEE Standard 802.1w
- Vorteile
  - Schnelle Zustandswechel
  - Ein Ausfall legt das Netz nicht 30 Sekunden lahm
  - Nur 3 Zustände

STP: Rapid Spanning Tree

- Änderungen gegenüber Spanning Tree
  - Wenn auf einem Port nach 3 “Hello”-Zeiten keine BPDU empfangen wurde wird der Spanning Tree neu berechnet
  - BPDUs werden als “Keep-alives” zwischen den Bridges verwendet

Rapid Spanning Tree

Port Zustände



Rapid Spanning Tree

- “Backup Ports” werden auf “Forwarding” geschaltet, sobald der “Designated Port” den Zustand auf “down” wechselt
- “Alternate Ports” werden auf “Forwarding” geschaltet, sobald der “Designated Port” keine BPDUs mehr schickt
  - Umschaltzeit sind 6 Sekunden, nicht mehr 30

Rapid Transition to forwarding

- Alle Ports an die direkt Endgeräte angeschlossen sind, überspringen die “Listening” und “Learning” Zustände
  - Sogenannte “Edge Ports”
  - Zustandswechsel von “Edge Ports” generieren keine TCNs
  - Sobald ein “Edge Port” eine BPDU empfängt, durchläuft er die ganz normalen Spanning Tree Mechanismen

Rapid Spanning Tree

- Vollduplex Ports sind “point-to-point”-Links
- Halbduplex Ports sind “shared medium”-Links

Rapid Spanning Tree

STP Zustand	RSTP Zustand	Port aktiv?	Lernt der Port?
Disabled	Discarding	nein	nein
Blocking	Discarding	nein	nein
Listening	Discarding	ja	nein
Learning	Learning	ja	ja
Forwarding	Forwarding	ja	ja

STP: PVST & PVST+

- “Per VLAN Spanning Tree” ist eine Cisco eigene Implementation
- PVST fährt für jedes VLAN einen eigenen Spanning Tree Prozess
- PVST+ ist PVST für RSTP
- Jedes VLAN kann seine eigene Rootbridge haben

Multiple Instance STP

- Ähnliches Prinzip wie PVST+
- Mehrere VLANs können in einer STP Instanz zusammengefaßt werden
- PVST skaliert nur bis 64 VLANs
- Wird von allen größeren Herstellern unterstützt

Vendor C und andere...

- Cisco Switches fahren pro VLAN eine eigene STP Instanz
- HP Switches fahren eine STP Instanz auf dem “native” VLAN
  - außer, man konfiguriert MST

Vendor C und andere...

- Wenn ein nicht-Cisco Gerät an eine PVST Wolke angeschlossen wird, benutzen Cisco Switches den Spanning Tree auf VLAN 1 um den Rest des Netzes in die Berechnungen einzubeziehen
- Die auf nicht-Cisco Geräten konfigurierte Bridge Priorität gilt nur für VLAN 1
- Wenn man mischt, sollte man auf MST setzen

Inhalt

- Was ist Ethernet?
- Was sind das für Bitmuster auf dem Draht?
- Wofür brauche ich aktive Komponenten?
- Was ist dieses “Spanning Tree Protocol”?
- **Was sind Virtual Local Area Networks?**

Virtual LANs

- VLANs erlauben eine Segmentierung des Netzes in mehrere Broadcast Domänen
- VLANs können ohne einen Router keine Daten untereinander austauschen
- VLANs werden zwischen Switchen über “Trunks” transportiert
- Auf einem “Trunk” Port sind alle Broadcasts aller VLANs sichtbar

Von A nach B

- Um VLANs von einem Switch in den nächsten zu bekommen gibt es 2 standardisierte Protokolle
  - ISL
    - InterSwitchLink - Cisco proprietär
  - IEEE 802.1q (dot1q)
    - Internationaler Standard den alle Hersteller implementieren

VLANs: 802.1q

- Nach dem Längen/Typfeld wird ein 4 Byte langes Feld eingefügt, das die VLAN-Id enthält
- Die Framechecksumme muß neu berechnet werden, da sich der Inhalt des Frames ändert
- Getaggte Frames müssen auch von nicht VLAN-fähigen Geräten transportiert werden

VLANs: ISL

- Es wird ein neuer Header vor dem Frame eingefügt und eine neue FCS angefügt
- ISL Frames sind 2048(?) Bytes groß
- Cisco proprietäres Protokoll

VLANs: Trunk Ports

- Das “native VLAN” wird ohne Tag übertragen
- Alle anderen VLANs werden mit Tag übertragen
- Trunks müssen mindestens “FastEthernet” sein

VLAN Trunking Protocol

- VTP ist ein Cisco spezifisches Protokoll um VLANs in einem Netz synchron zu halten
- In einer “VTP Domain” gibt es einen Server und mehrere Clients
- VLANs werden auf dem “Server” eingerichtet und können dann auf den “Clients” auf Ports konfiguriert werden

VLAN Trunking Protocol

- Es gibt 3 Rollen
  - Server
  - Client
  - Transparent

VTP Server

- VTP Server broadcasten ihre VLAN Konfiguration alle XX Sekunden an Clients
- Cisco Switches sind ab Werk als “VTP Server” konfiguriert
  - Erst den VTP Mode umstellen, dann die VTP Domäne setzen

VTP Client

- VTP Clients bekommen in regelmäßigen Intervallen konfigurierte VLANs vom Server
- VTP Clients leiten empfangene VTP Pakete weiter
- VTP Clients müssen eine Domäne und ein gesetztes Passwort haben

VTP Transparent

- Auf “VTP transparent” konfigurierte Switche ignorieren VTP Pakete, leiten sie aber weiter
- Sollte der Auslieferungszustand von Cisco Switchen sein
  - “ip http server” ist allerdings auch immer noch Standard

GVRP

- GARP VLAN Registration Protocol
  - GARP: Generic Attribute Registration Protocol
- Ermöglicht die herstellerübergreifende Weitergabe von 802.1q VLANs
- Voraussetzung für VLAN Pruning

GVRP: Mehr Infos

- http://www.javvin.com/protocolGVRP.html
- Ist im IEEE 802.1q und 802.1p Standard definiert
- Wird unterstützt von HP, Extreme Networks, Foundry und Cisco
  - bei Cisco nur in CatOS, nicht IOS!

VLAN Pruning

- Wird durch VTP und GVRP ermöglicht
- Switche “kennen” die auf anderen Switchen konfigurierten VLANs
- Broadcasts auf einem bestimmten VLAN werden nur an Switche weitergeleitet, die dieses VLAN auch auf einen Port konfiguriert haben

Cisco Discovery Protocol

- Wird von Cisco und HP Switchen gesprochen
- CDP benutzt Ethernet Multicasts
- Es gibt eine Implementation für Unix
  - ftp://ftp.lexa.ru/pub/domestic/snar/cdpd-1.0.2.1.tgz
- Hilfreich zur Netzwerkdiagnose

Cisco Discovery Protocol

- HP wechselt mittlerweile zu LLDP
  - in neuerer ProCurve Software ist es nicht mehr vorhanden
- Cisco betreibt es weiterhin

Cisco Discovery Protocol

- In einem CDP Paket sind folgende Informationen enthalten
  - Bridgename
  - Capabilities
  - Port ID

CDP: Capabilities

- Router
- Transparent Bridge
- Switch
- Host
- IGMP
- Repeater

Danke an:

- #t42 fürs Querlesen
- Cisco für die Dokumentation im Web