Switching - Spanning Tree Protocol (STP)

Netzwerkgrundlagen (NWG2)

Markus Zeilinger[1]

[1]FH Oberösterreich
Department Sichere Informationssysteme

Sommersemester 2023

FH OBERÖSTERREICH
UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Wichtiger Hinweis

Alle Materialien, die im Rahmen dieser LVA durch den LVA-Leiter zur Verfügung gestellt werden, wie zum Beispiel Foliensätze, Audio-Aufnahmen, Übungszettel, Musterlösungen, ... dürfen ohne explizite Genehmigung durch den LVA-Leiter NICHT weitergegeben werden!

Spanning Tree Algorhyme [1]

I think that I shall never see
A graph more lovely than a tree.

A tree whose curcial property
Is loop-free connectivity.

A tree that must be sure to span
So packets can reach every LAN.

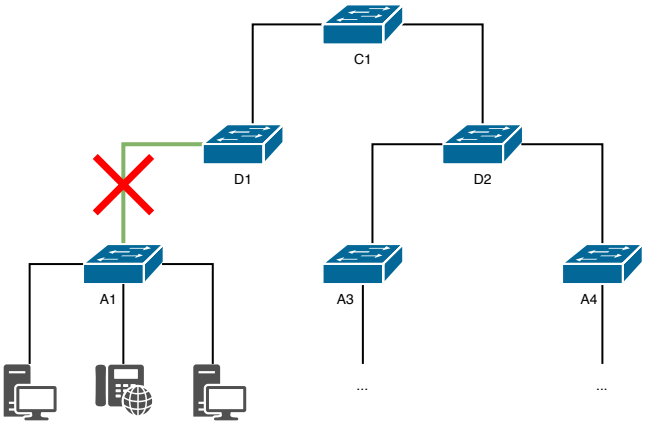
First, the root must be selected
By ID, it is elected.

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.

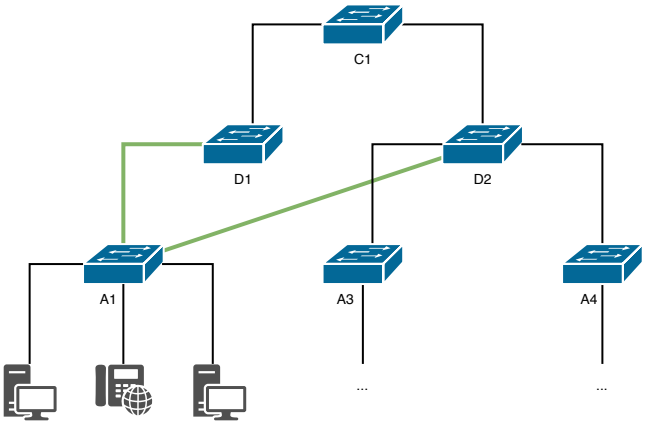
Redundanz in geswitchten Netzwerken I

- Problem: In geswitchten Netzen stellen insb. Inter-Switch Links problematische Single Point of Failures dar (Access $\leftrightarrow$ Distribution, Distribution $\leftrightarrow$ Core).



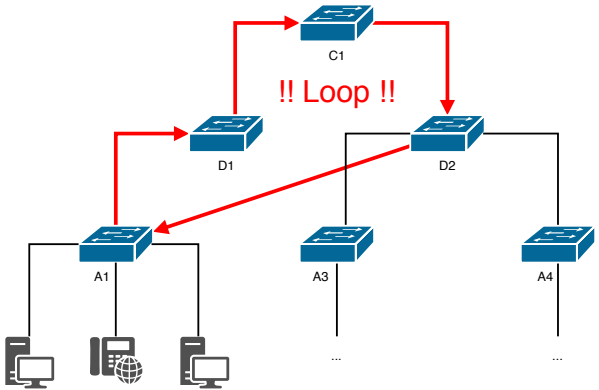
Redundanz in geswitchten Netzwerken II

- Lösung: Switches werden physisch redundant miteinander verbunden (z. B. Access Switch mit zwei physischen Verbindungen zu zwei Distribution Switches).



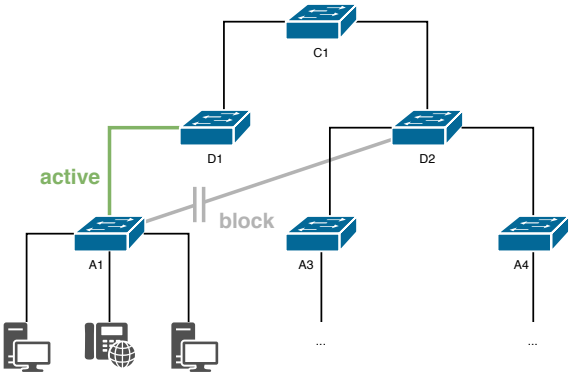
Redundanz in geswitchten Netzwerken III

- Neues Problem: Ethernet Frames besitzen kein "Lebenszeit" (vgl. TTL- bzw. Hop-Limit-Feld in IPv4- bzw. IPv6-Paketen) + durch physische Redundanz entstehen Schleifen (Loops) im geswitchten Netz → Broadcast Storm



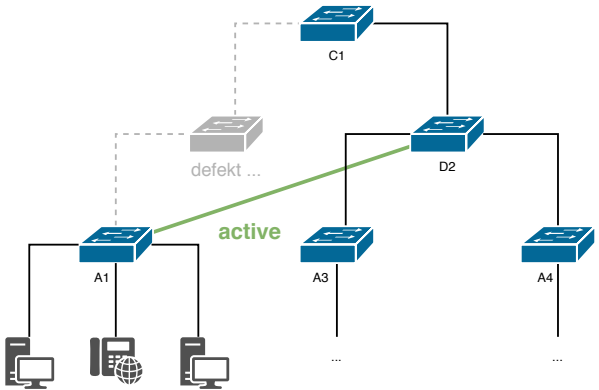
Redundanz in geswitchten Netzwerken IV

- Neue Lösung: Die gewollten, physischen Redundanzen müssen logisch unterbrochen werden.
  - Kein Lastausgleich, vorhandene physische Wege werden nicht genutzt ($\rightarrow$ RBridges/TRILL, IEEE 802.1aq Shortest Path Bridging, Cisco FabricPath)!



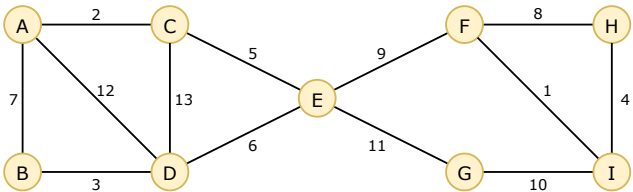
Redundanz in geswitchten Netzwerken V

- Neue Lösung: Im Fehlerfall werden die logisch unterbrochenen redundanten Verbindungen reaktiviert.



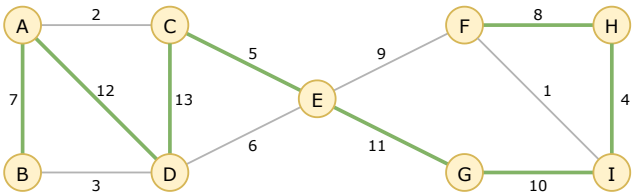
Graphentheorie I

- Ein geswitchtes Netzwerk kann als ungerichteter, gewichteten und zusammenhängender Graph G aufgefasst werden.
  - Graph $G = (V, E)$ wobei die Knoten (Vertices) die Switches und die Kanten (Edges) die Inter-Switch Links sind.
  - Ungerichtet bedeutet, das die Kanten keine Richtung aufweisen und in beide Richtungen genutzt werden können.
  - Gewichtet bedeutet, dass den Kanten Werte im Kontext des Graphen zugeordnet sind (z. B. Distanz zwischen zwei Orten).



Graphentheorie II

- Zum Zweck der Schleifenfreiheit (loop-free) kann das Konzept des minimalen Spannbaums (Minimum Spanning Tree, MST) angewendet werden.
  - Ein Spannbaum (auch Gerüst genannt) ist ein Teilgraph T eines ungerichteten, gewichtete und zusammenhängende Graphen, der ein Baum ist und alle Knoten des Graphen G enthält.
  - Minimal ist ein Spannbaum, wenn die Summe der Kantengewichte für den Graphen G minimal ist.
  - Berechnung des MST durch den Algorithmus von Kruskal.



Basic Facts

- Das Spanning Tree Protocol (STP) wurde von Radia Perlman[1] entwickelt, 1990 in IEEE 802.1D standardisiert und 2014 in IEEE 802.1Q (aktuell IEEE 802.1Q-2022) integriert.
  - Verfahren läuft dezentral und kooperativ im geswitchten Netz ab, kein Lastausgleich!
- Varianten von STP
  - Per-VLAN Spanning Tree Protocol (PVSTP): Ein eigener Spanning Tree pro VLAN.
  - Rapid Spanning Tree Protocol (RSTP): Schneller Konvergenz bei Änderungen in der Topologie als STP, ursp. IEEE 802.1w, jetzt Teil von IEEE 802.1Q-2022.
  - Multiple Spanning Tree Protocol (MSTP): Mehrere Spanning Trees, auch mehrere VLANs in einem Spanning Tree möglich.

[1] Radia Perlman: Internet Hall of Fame Pioneer https://internethalloffame.org/inductees/radia-perlman

Grundsätzlicher Algorithmus

1. Unter allen Switches in einer Broadcast Domain wird eine Root Bridge als Wurzel des Spanning Trees gewählt (Root Election).
   - Auswahlkriterium ist primär die s. g. Priority (+ ggf. eine MAC-Adresse des Switches).
   - Die Priority ist default 32768 [2, S. 535] und sollte durch den Administrator festgelegt werden; niedrigerer Wert höhere Priorität (4096er Schritte [2, S. 535]).
2. Jeder Switch bestimmt seinen Port mit dem geringsten Abstand zur Root Bridge als Root Port (Root Port Election).
3. Für alle anderen Ports eines Switches gilt (Designated Port Election):
   - Kann der Switch an einem Port den kürzesten Abstand zur Root Bridge anbieten, dann bleibt der Port aktiv und wird zum Designated Port.
   - Kann der Switch das an einem Port nicht, wird der Port zu einem Blocking Ports (hier würden sonst Schleifen entstehen).

Bridge Protocol Data Units (BPDUs) I

- Der Austausch von Informationen im Spanning Tree Protocol passiert über Bridge Protocol Data Units (BPDUs).
- BPDUs werden sofort ab Abschluss des Starts eines Switches gesendet und dienen folgenden konkreten Zwecken:
  - Bestimmung der Root Bridge in der Broadcast Domain (Root Election),
  - Bestimmung der Root Ports der Switches (Root Port Election),
  - Bestimmung der anderen Ports der Switches (Designated Port Election) und
  - zur Signalisierung und Bestätigung von Topologieänderungen (Topology Change Notification [TCN], Topology Change Acknowledged [TCA]).

Bridge Protocol Data Units (BPDUs) II

- Ein Switch teilt mit dem Senden einer BPDU im Wesentlichen semantisch folgendes mit:
  - *Ich Switch `Transmitter ID` teile auf meinem Port `Port ID` mit, dass ich den Switch `Root ID` für die Root Bridge in der Broadcast Domain halte und dass mein Abstand zum Switch `Root ID` den Wert `Root Path Costs` hat.*

- BPDU Type: 0x00 für Configuration BPDUs, 0xA0 für TCN
- Flags: 0xA0 für TCA, 0x01 für TC, ...
- Root Identifier (Root ID): Bridge ID (Priorty + MAC-Adresse) der vermuteten Root Bridge

Bridge Protocol Data Units (BPDUs) III

- Root Path Costs: "Pfad-Kosten" (Abstand) zu denen der sendende Switch die Root Bridge erreichen kann.
- Transmitter Identifier (Transmitter ID): Bridge ID des sendenen Switches (Priority + MAC-Adresse).
- Port Identifier (Port ID): Port über den der sendende Switch die BPDU geschickt hat.
- Hello Time: default 2s [2, S. 544], BPDU-Intervall der (vermeintlichen) Root Bridge.
- Forward Delay: default 15s [2, S. 544], Zeit im Übergang vom State `Listening` zu `Learning` und von `Learning` zu `Forwarding` ($\rightarrow$ Konvergenzzeit für STP bei 30s!).

Path Costs

- "Pfad-Kosten" sind abhängig von der Geschwindigkeit eines Ports und summieren sich auf von Switch zu Switch.

Port Speed Mbps	Port Speed	STP	RSTP [2, S. 536]
1	1 Mbps	1000	20.000.000
10	10 Mbps	100	2.000.000
100	100 Mbps	19	200.000
1000	1 Gbps	4	20.000
10.000	10 Gbps	2	2.000
100.000	100 Gbps		200
1.000.000	1 Tbps		20
10.000.000	10 Tbps		2

Priority

- Priority ist default 32768 [2, S. 535] und sollte durch den Administrator festgelegt werden, insb. zur Bestimmung der Root Bridge.
- Niedriger Priority-Wert bedeutet Bevorzugung bei der Auswahl der Root-Bridge.
- Switch mit der niedrigsten Priority wird Root Bridge in der Broadcast Domain.
- Bei gleichen Priority-Werten dient die MAC-Adresse des Switches als zusätzlicher Tie Breaker (wertmäßig kleinere MAC-Adresse gewinnt).
- Zusammen bilden Priority und MAC-Adresse die Bridge ID (BID).

Root Election

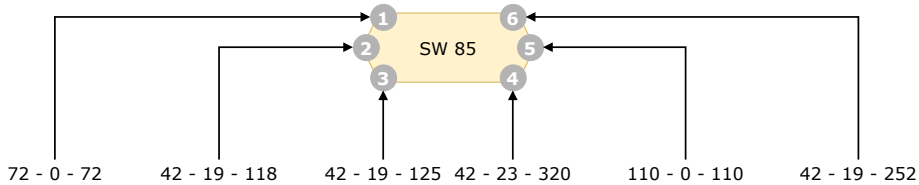
- Zu Beginn nach dem Starten glaubt jeder Switch, selbst die Root Bridge zu sein.
- Jeder Switch sendet im Hello-Time-Intervall [2, S. 544] BPDUs an alle anderen und behauptet selbst die Root Bridge zu sein (Merkmal: Root ID == Transmitter ID).
- Bei Empfang einer BPDU von einem Nachbarn vergleicht ein Switch seine gespeicherte Root ID mit der in der BPDU behaupteten. Ist diese kleiner als die gespeicherte Root ID, aktualisiert der Switch seine gespeicherte Root ID und sendet in zukünftigen BPDUs auch diese neue Root ID.
- Der Prozess der Root Election endet, wenn alle Switches in der Broadcast Domain die tatsächliche Root Bridge identifiziert und deren Root ID gespeichert haben.

Root Port Election

- Im Vorgang der Root Election identifizieren Switches auch ihren Root Port, d. h. den Port, über den sie die Root Bridge mit den geringsten Kosten (dem geringsten Aufwand) erreichen können.
- Zentrales Entscheidungskriterium sind die Bridge IDs und die Root Path Costs, weitere Tie Breaker sind Transmitter ID und Port Identifier.

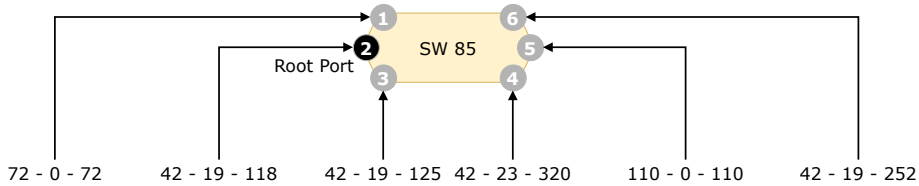
Root und Root Port Election - Beispiele I

- Switch mit Bridge ID 85 glaubt zu Beginn, selbst die Root Bridge zu sein ...



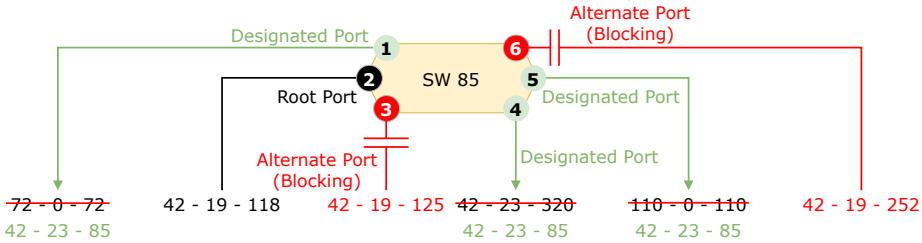
- BPDUs in der (verkürzten) Form `<RootID> - <RootPathCosts> - <TransmitterID>` treffen ein.
- An jedem Port wird nur die "beste" BPDU gezeigt.
- Alle Anschlüsse sind Gigabit-Ethernet-Anschlüsse.

Root und Root Port Election - Beispiele II



- Root Election: Zwar glaubt Switch 85 selbst die Root Bridge zu sein, am Port 2 bekommt er aber ein besseres Angebot (niedrigere Bridge ID), nämlich Switch 42 mit den Pfadkosten von 19.
- Zwar gibt es an anderen Ports auch Angebote für Bridge ID 42, allerdings sind dort die Pfadkosten schlechter weil wertmäßig höher oder die Transmitter ID ist wertmäßig höher.
- Switch 85 aktualisiert seine Root ID, setzt seine Port Roles und sendet nun seinerseits BPDUs an seine Nachbar aus ...

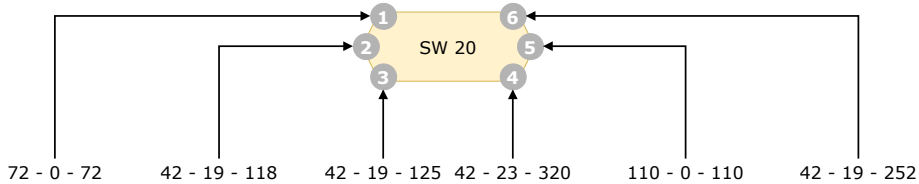
Root und Root Port Election - Beispiele III



- Port 2: Ist nun der Root Port von Switch 85.
- Ports 1, 4 und 5: Hier kann Switch 85 nun ein besseres als das erhaltene Angebot (42 - 23 - 85) machen und schickt daher entsprechende BPDUs. Die Port Role wird Designated Port.
- Ports 3 und 6: Hier kann Switch 85 kein beseres als das erhaltene Angebot machen. Er schickt daher keine BPDUs und die Port Role wird Blocking Port (hier würden sonst Schleifen entstehen).

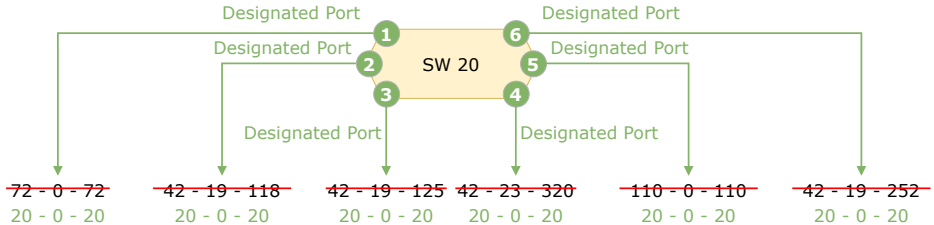
Root und Root Port Election - Beispiele IV

- Switch mit Bridge ID 20 glaubt zu Beginn, selbst die Root Bridge zu sein ...



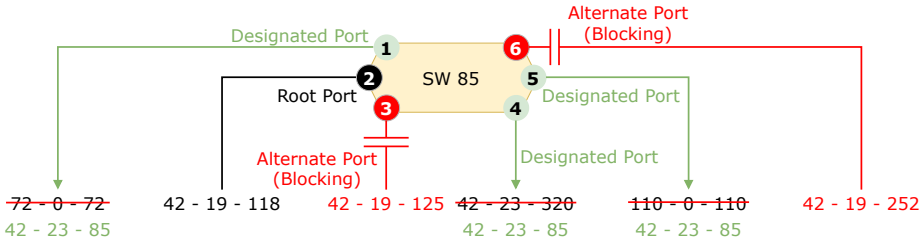
- BPDUs ein in der (verkürzten) Form `<RootID> - <RootPathCosts> - <TransmitterID>` treffen ein.
- An jedem Port wird nur die "beste" BPDU gezeigt.
- Alle Anschlüsse sind Gigabit-Ethernet-Anschlüsse.

Root und Root Port Election - Beispiele V



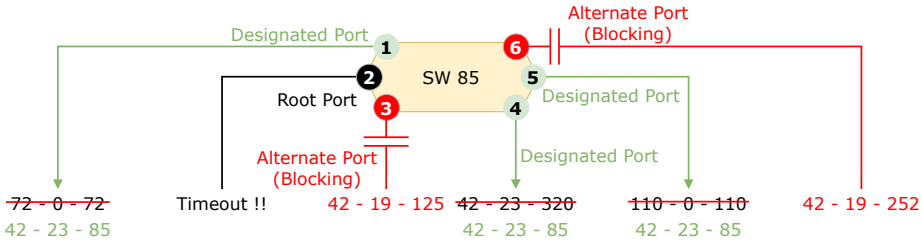
- Root Election: Nachdem kein anderer Switch ein besseres Angebot machen kann, ist Switch 20 offensichtlich die Root Bridge.
- Alle Ports werden Designated Ports und Switch 20 schickt entsprechende BPDUs (20 – 0 – 20) über alle Ports.

BPDU Timeouts - Beispiel I



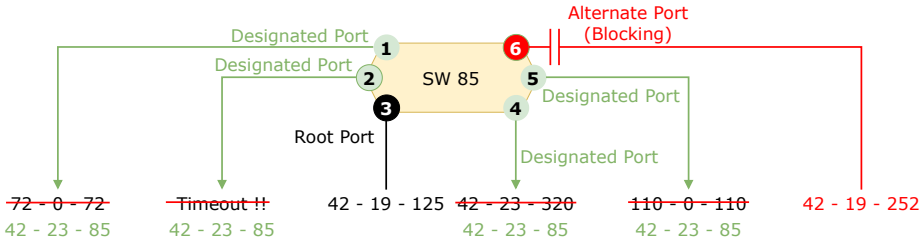
- Die Root Bridge schickt in periodischen Abständen (Hello Time) BPDUs ("Hello"-Nachricht).
- Jeder Switch schickt bei Empfang ebenfalls eine "Hello"-Nachricht auf alle Designated Ports.
- Empfängt eine Bridge auf ihrem Root Port eine gewisse Zeit (Max Age) keine "Hello"-Nachricht, geht sie davon aus, das ihr Pfad zu Root nicht mehr gegeben ist und konfiguriert sich um.

BPDU Timeouts - Beispiel II



- Switch 85 läuft auf seinem Root Port in einen Timeout, er betrachtet seinen Pfad zu Root als nicht mehr gegeben.
- Er wird einen neuen Root Port wählen (der Pfad zu Root wird dadurch schlechter, weil er ja nun das zweitbeste Angebot wählen muss) und auch die weiteren Port Rollen überdenken.

BPDU Timeouts - Beispiel III



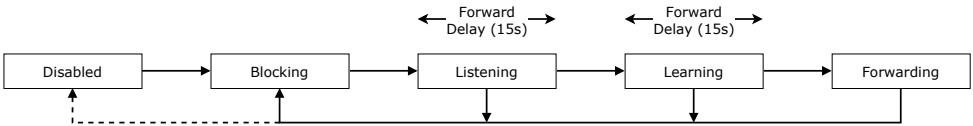
- Root Election: Bridge ID 42 bleibt für Switch 85 die Root Bridge, der neue Root Port wird Port 3.
- Ports 1, 2, 4 und 5: Switch 85 kann hier ein besseres als das erhaltene Angebot (43 - 23 - 85) machen und schickt daher entsprechende BPDUs. Die Port Role wird/bleibt Designated Port.
- Port 6: Hier kann Switch 85 kein beseres als das erhaltene Angebot machen. Er schickt daher keine BPDUs und die Port Role wird Blocking Port.

Port Roles

- Root Port:
  - Switch Port mit dem besten/kürzesten Pfad zur Root Bridge.
  - Die Root Bridge hat keinen Root Port.
- Designated Ports:
  - Auf der Root Bridge sind alle Ports Designated Ports.
  - Auf allen anderen Switches sind die Ports Designated Ports, über die der beste/kürzeste Pfad zur Root Bridge realisiert werden kann. Über Designated Ports werden Frames empfangen und weitergeleitet.
- Nondesignated/Alternate Ports:
  - Alle Blocking Ports sind Alternate Ports.
  - Über diese Ports würden Schleifen entstehen und sie werden daher logisch blockiert.
- Disabled Port:
  - Administrativ deaktivierter Port (`SW01(config-if)# shutdown`).

Port Status I

- Jeder Switch Port durchläuft im STP-Vorgang folgende Stati (soll verhindern, dass Ports vor Feststellung des Spanning Trees aktiv werden und dadurch - versehentlich - eine Schleife bilden):



- Disabled: Administrativ deaktiviert (`SW01(config-if)# shutdown`), kein STP, kein Forwarding.
- Blocking: BPDUs empfangen + verarbeiten, kein Forwarding.
- Listening: + BPDUs senden, kein Forwarding.
- Learning: + MAC Learning für SAT, kein Forwarding.
- Forwarding: + Forwarding → volle Funktionalität.

Port Status II

- Cisco bietet für Access Ports das Feature PortFast, welches den Anschluss direkt vom Blocking-Status in den Forwarding-Status setzt.
- Dieses Feature soll einem Access Port schnellere Konnektivität zum Netzwerk, z. B. für einen DHCP-Vorgang, ermöglichen.

```
SW01(config)# int fa0/1
SW01(config-if)# description ACCESS_PORT
SW01(config-if)# switchport mode access
SW01(config-if)# spanning-tree portfast
%Warning: portfast should only be enabled on ports connected to a single
host. Connecting hubs, concentrators, switches, bridges, etc... to this
interface  when portfast is enabled, can cause temporary bridging loops.
Use with CAUTION

%Portfast has been configured on FastEthernet0/1 but will only
have effect when the interface is in a non-trunking mode.
```

Topologieänderung

- Normaler Zustand: Switches erhalten von der Root Bridge an ihrem Root Port BPDUs senden dort aber selbst nie BPDUs hin.
- Erkennt ein Switch eine Topologieänderung, dann signalisiert er diese aber über eine Topology Change Notification (TCN) BPDU an seinem Root Port.
- Der empfangende Switch (s. g. Designated Switch) antwortet mit einer BPDU mit gesetztem Topology Change Acknowledgement Bit und schickt die TCN weiter an seine Designated Bridge (Fortsetzung zur Root Bridge).
- Sobald die Root Bridge von der Topologieänderung erfährt, schickt sie BPDUs mit gesetztem Topology Change (TC) Bit. Diese werden von allen Switches weitergegeben und damit werden alle Switches über Topologieänderungen informiert.

Zusatzmaterial

Graphentheorie III

- Algorithmus von Kruskal zur Berechnung des minimalen Spannbaums T eines Graphen G:

1. Alle Kanten des Graphen sind "ungewählt".
2. Die "kürzeste" Kante (= kleinstes Kantengewicht) wird gewählt.
3. Von den noch nicht gewählten Kanten wird die gewählt, die (a) die kürzeste Kante ist und (b) mit den bereits gewählten Kanten keine Schleife bildet.
4. Schritt 3 wird solange wiederholt, solange Kanten verfügbar sind.

Graphentheorie IV