Spanning Tree

Von Wolfgang Schulte

Den optimalen Weg in einer komplexen Topologie zu bestimmen und dies möglichst schnell, ist das Ziel des Spanning-Tree-Protokolls (STP) und des Rapid-Spanning-Tree-Protokolls (RSTP). Nicht nur einen Pfad, sondern gleich mehrere Wege durch das Netz werden per Multiple-Spanning-Tree-Protokoll (MSTP) ermöglicht.

Das Institute of Electrical and Electronic Engineers (IEEE) beschreibt in ihren 802-Standards drei Protokolle der Wegewahl durch größere lokale Ethernet/CSMA/CD-Netze (LANs) mit Brücken oder Switches.

Auf der einen Seite ist Redundanz in einem Netz besonders wichtig, weil damit Netzwerke fehlertolerant werden. Redundante Topologien schützen vor unerwünschten Ausfallzeiten im Netz auf Grund von Fehlern einer einzigen Verbindung, eines Anschlusses oder einer Netzeinheit. Andererseits wird durch diese Redundanz in der Topologie die Möglichkeit für die fehlerhafte Doppelübertragungen von Informationen eröffnet.

Anfang der 80er Jahre hat die Firma Digital Equipment Corporation das Konzept der transparenten Brücken entwickelt, das in die IEEE-Standards eingeflossen ist. Das heißt, nicht die Rechner kennen den Weg vom Sender zum Empfänger, sondern die Brücken oder Switches.

Ob über Brücken oder Switches, beide Schicht-2-Einheiten „lernen“ die Topologie des Netzes, das heißt die Medium-Access-Control-Adressen (MAC) von sendenden Stationen durch die hindurchgehenden Pakete, selbstständig kennen. Das gesamte Netz wird als Baum (Tree) betrachtet, in dem die angeschlossenen Stationen die Blätter des Baumes darstellen und die Brücken oder Switches die Äste und den Stamm bilden.

Das IEEE hat mit dem ersten Standard 802.1d für STP 1998 die Grundlage dieser Protokollfamilie geschaffen. Im Jahr 2002 stellte IEEE 802.1w das RSTP und 802.1s das MSTP vor. Beide neuen zusätzlichen Protokolle unterstützen folgende Ziele:

- Verhinderung von kreisenden Paketen in Topologien die eine Schleife, und damit mehrere mögliche Pfade bilden. Eine Kompatibilität mit dem vorhandenen STP ist einzuhalten.
- Bestimmung eines oder mehrerer optimalen Pfade zum Beispiel für Lastausgleich bei mehreren möglichen Wegen.
- Im Falle von Störung oder Ausfall eines festgelegten Pfades, die schnelle Bereitstellung eines alternativen Weges.

In Token-Ring-LANs ist das so genannte Source-Routing bekannt, in dem die Stationen den besten Weg vom Sender zum Empfänger kennen. Deshalb ist in diesen LANs die Baumstruktur für das Versenden von Paketen nicht notwendig. Parallele Brücken sind in der Topologie erlaubt.

Spanning-Tree-Protokoll

In einem zuverlässigen und stabilen Netz wird der Datenverkehr vom Sender zum Empfänger effizient geleitet, wenn im Falle von Störungen redundante Pfade bereit stehen. In den vermaschten Weitverkehrsnetzen (WAN) werden an Hand eines Teils der IP-Adresse auf Schicht 3 Routing-Protokolle wie Routing-Information-Protokoll (RIP) oder Open Shortest Path First (OSPF) diese Aufgaben der Bestimmung des Pfades übernehmen. Bei Ausfall eines Weges, oder zum Lastausgleich, wird ein alternativer Weg eröffnet.

In einem Schicht-2-Netz, einem LAN, stehen diese Protokolle nicht zur Verfügung. Diese Aufgabe übernimmt also STP in Ethernet/CSMA/CD-LANs. Brücken und Switches nutzen die Schicht-2-Adresse (MAC-Adresse), um ihre Entscheidung zur Weiterleitung zu treffen.

Beispiel: Wenn Rechner A einen Rahmen an Rechner B sendet, empfangen beide Brücken den Rahmen und erkennen zunächst, das Rechner A sich im Netz 1 befindet. Beide Brücken leiten den Rahmen an Rechner B im Netz 2 weiter. Rechner B erhält also zwei Rahmen. Der Rahmen von Rechner A über Brücke A erreicht auch Brücke B per Netz 2, und umgekehrt erreicht der Rahmen von Rechner A über Brücke B per Netz 2 auch Brücke A. Brücke A und B nehmen nun an, dass Rechner A sich jetzt im Netz 2 befindet. Antwortet jetzt Rechner B und will die Nachricht an Rechner A senden, so leiten die Brücken B den Rahmen nicht in Netz 1 weiter, weil sie annehmen das Ziel ist im gleichen Netz wie die Quelle, das heißt, der Rechner A ist aus Netz 2 für Rechner B nicht mehr erreichbar. Da Brücken Broadcastsendungen zum Bei-

Schleifenbildung

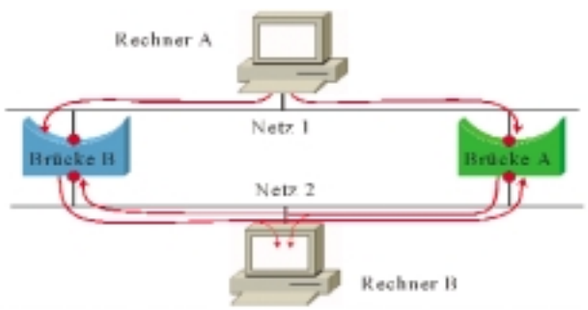

Bild: W. Schulte

Die Abbildung zeigt die Verbindung zweier Rechner in verschiedenen Netze, die über redundante Wege erreicht werden können

Wolfgang Schulte war 28 Jahre bei IBM im Labor als Manager verschiedener Abteilungen (Hard- und Software) tätig. Zuletzt vertrat er die IBM Europa in mehreren externen Ausschüssen zum Beispiel bei ETSI, ITU-T, ISO, ECMA und DVB. Zurzeit ist er Dozent an der BA-Stuttgart und arbeitet freiberuflich für Siemens und Cisco.

spiel ARP Requests über alle Ports weiterleiten, wird mit einer Broadcastsendung das gesamte Netz total ausgelastet.

Spanning Tree Algorithmus

Der beim STP implementierte Spanning-Tree-Algorithmus (STA) legt, zur Beseitigung des Problems, eine schleifenfreie logische Topologie fest, das heißt nur ein aktiver Pfad besteht zwischen zwei Rechnern im Netz. Bestimmte Anschlüsse der Brücken werden in einen speziellen Blocking-Zustand gesetzt, um somit eine Schleife zu verhindern. Fällt die gewählte primäre Verbindung aus, so kann der Anschluss, der geblockt wurde, wieder aktiviert werden, um so eine neue Baumstruktur zu bilden.

Nach der Initialisierung gehen alle Ports zunächst in den Status Blocking, das heißt, nur Konfigurationsrahmen von Brücke zu Brücke, Bridged Protocol Data Unit (BPDU) genannt, werden akzeptiert. Brücken generieren von sich aus keine Konfigurationsrahmen, sondern senden nur BPDUs aus, wenn sie von einer so genannten Root-Brücke eine BPDU erhalten haben. Dabei wird auch von der BPDU-Weiterleitung (Relaying) gesprochen.

Teilnehmerdaten werden jetzt nicht weitergeleitet. Die Brücken senden danach Konfigurationsrahmen aus. Mit Hilfe der Information der BPDU wechseln die Brückenanschlüsse ihren Status.

Mit den Port-Zuständen Listening und Learning wird eine temporäre Schleife während einer Rekonfiguration vermieden. Während des Status Listening wird die „aktive“ Topologie gebildet, keine Nutzerdaten werden weitergeleitet. Im Status Learning wird die Bridging-Tabelle aus den gelesenen Adressen zusammengestellt, Nutzerdaten werden nicht weitergeleitet.

Ports, die die Nutzerdaten weiterleiten sollen, sind im Status Forwarding. Im Status Disabled sind die Anschlüsse heruntergefahren, es werden weder Nutzerdaten noch BPDUs empfangen oder weitergeleitet.

Die Felder Protocol ID, Version und Message Type sind auf 0 gesetzt. Protokoll ID auf 0 bedeutet hier IEEE 802.1d STP. Message Type 0 signalisiert eine Konfigurations-BPDU.

Im Flag-Feld wird Bit 0 verwendet, um Topologieänderungen anzuzeigen, und Bit 7, um die Topologieänderung zu bestätigen (ACK).

Die Root-ID identifiziert die Root-Brücke mit Angabe von Priorität (2 Byte) und ID (6 Byte). Das nächste Feld trägt als Kennung die Pfadkosten der Brücke, die die BPDU an die Root Bridge gesendet hat.

Der STA ermöglicht allen Brücken die eindeutige Vergabe eines Identifikators: Bridge ID (BID). Dieser Identifikator besteht aus der 6 Byte großen Schicht-2-Adresse, der MAC-Adresse der Brücke, und ein 2-Byte-Prioritäts-Feld.

Allen Anschlüssen einer Brücke, Ports genannt, werden ebenfalls eindeutige Identifikatoren innerhalb der Brücke zugewiesen. Jedem Port wird, entweder vom Netzadministrator oder selbstbestimmend aus der Geschwindigkeit des Netzes, ein Wert, die Kosten eines Pfades im LAN, zugewiesen.

Das Feld Message Age (Meldungsalter) enthält den Wert der Zeit seit dem Absenden der Konfigurationsmeldung. Mit Max. Age wird der Zeitpunkt angegeben, an dem die Meldung gelöscht werden soll. Das Feld Hello-Zeit gibt die Zeitspanne zwischen den Konfigurationsmeldungen der Root-Brücke an, der Standardwert ist zwei Sekunden.

Im Feld Vorwärtsverzögerung wird die Wartezeit nach einer Änderung der Topologie angegeben. Durch diese Wartezeit ist die Konvergenz des Netzes auf die neue Topologie sichergestellt.

Wenn eine schleifenfreie logische Topologie zu bilden ist, nutzt STP vier Kriterien zur Bestimmung der höchsten Prioritäten der Brücken beziehungsweise der Ports:

- die kleinste Root-Brücken-ID (Brücken-Priorität und MAC-Adresse),
- die geringsten Wegekosten zur Root-Brücke,
- die kleinste Sender-Brücken-ID,
- die kleinste Port-ID.

Drei Schritte sind notwendig, um zu einer redundanten, aber logisch schleifenfreie Topologie zu kommen.

Wahl der Root-Brücke

Um einen effektiven Weg im LAN zu bestimmen, wird von allen Brücken im LAN die so genannte Root-Brücke oder Switch bestimmt. Bei der Initialisierung, zum Beispiel nach Netz-Ein, geht jede Brücke davon aus, dass sie Root-Brücke ist. Die Bestimmung der Root-Brücke erfolgt mit der BPDU an Hand der kleinsten Priorität beziehungsweise bei gleicher Priorität entscheidet die kleinste MAC-Adresse. Die Ports einer Brücke oder Switches sind im Blocking-Status, das heißt, sie leiten keine Rahmen weiter, registrieren aber die BPDUs. Nur eine Brücke im Netz kann Root-Brücke werden. Von dort aus werden alle anderen Wege bestimmt. Die Brücken, die nicht Root-Brücke geworden sind, müssen einen Anschluss, den Root-Port, in Richtung der Root-Brücke bestimmen.

Cisco setzt die Priorität standardmäßig auf 32768, dies ist bei zwei Byte (2^{16} = 65.536) exakt die Hälfte des möglichen Adressbereichs.

Die Festlegung von Root-Ports

Die Bestimmung der Root-Ports in den Brücken, die nicht Root-Brücke

STP-Port-Status

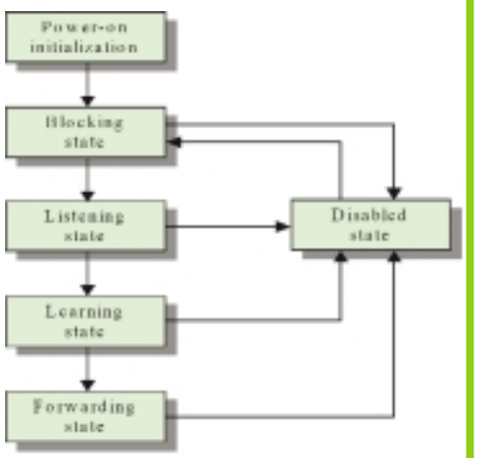

Bild: W. Schulte

Die Ports werden nach Bedarf geschaltet

Bridged Protocol Data Unit (BPDU)

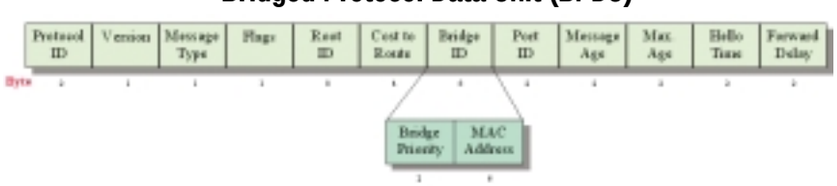

Bild: W. Schulte

Die Grafik zeigt die Felder der BPDU und die Größe der Felder in Bytes

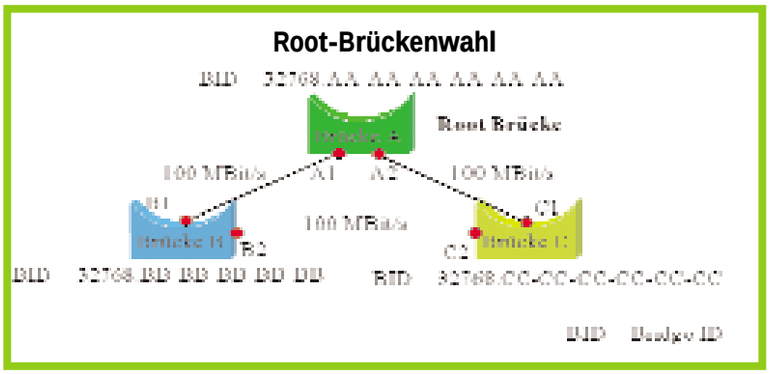

Bild: W. Schulte

Das Bild zeigt ein Beispiel zur Bestimmung der Root-Brücke. Mit der MAC-Adresse AA-AA-AA-AA-AA-AA ist die Brücke A die Root-Brücke bei gleicher Priorität aller Brücken

sind, wird durch den „billigsten“, das heißt „schnellsten“ Weg ermittelt. Jede Nicht-Root-Brücke muss einen Root-Port ausweisen.

Die Tabelle rechts zeigt, dass Standard-Ethernet mit 10 MBit/s langsamer ist, das heißt teuerer als Fast-Ethernet mit 100 MBit/s. Die Gesamt-Wegekosten ist die Summe der einzelnen Wegekosten. Aus der Übertragung der BPDU wird dies ermittelt.

Festlegung der Root-Ports: Die Root-Brücke sendet über A1 BPDUs aus mit den Wegekosten = 0. Wenn Brücke B diese BPDU auf B1 erhält, addiert sie zu den Wegekosten ihre Wegekosten dazu: 0 + 19 = 19. Diese neuen Wegekosten sendet Brücke B über B2 and Brücke C Port C2. Brücke C addiert zu diesen 19 noch einmal 19 dazu: 19+19=38. Gleichzeitig sendet Brücke A über A2 die gleiche BPDU an Brücke C mit dem Port C1. C1 hat also die Wegekosten 19, wie B1. Brücke C sendet ebenfalls diese BPDU per C2 an Brücke B auf Port B2.

Die Wegekosten für B2 zur Root-Brücke sind also ebenso 38 wie C2. Ports B1 und C1 sind die bestimmten Root-Ports.

IEEE spezifizierte ursprünglich die Wegekosten als 1.000 MBit/s dividiert durch die Bandbreite in MBit/s. Beispielsweise liegen für 10-Base-T die Kosten bei 100 (1000/10). Durch die Einführung von GBit/s LAN wurde der Algorithmus durch die Tabelle ersetzt.

Root-Ports empfangen also die „beste“ BPDU an einer Brücke und sind der Root-Brücke, bezogen auf die Wegekosten, am nächsten.

Bestimmung der Designate-Ports

Jedes Segment im Netz, zum Beispiel die Verbindung A1 - B1 oder B2 - C2, hat einen Designate-Port. Die Bestimmung eines solchen Ports in den Brücken geschieht ebenso durch die geringsten Wegekosten, das heißt, A1 und A2 werden Designate-Ports. Was aber ist mit dem Segment B2 - C2? Die Wegekosten sind für beide Ports gleich, die Priorität ist gleich, es entscheidet jetzt die kleinste MAC-Adresse. Dies ist im Beispiel die Brücke B, dies bedeutet Port B2 wird Designate-Port. Nach der Zuordnung der Root-Brücke und der Funktion der Ports wird Port C2 in den Blocking-Status gesetzt, das heißt, an diesem Port werden BPDUs nur empfangen aber keine gesendet. Es werden auch keine Nutzerdaten weitergeleitet. Die Designate-Ports senden also die „beste“ BPDU in einem angeschlossenen Segment.

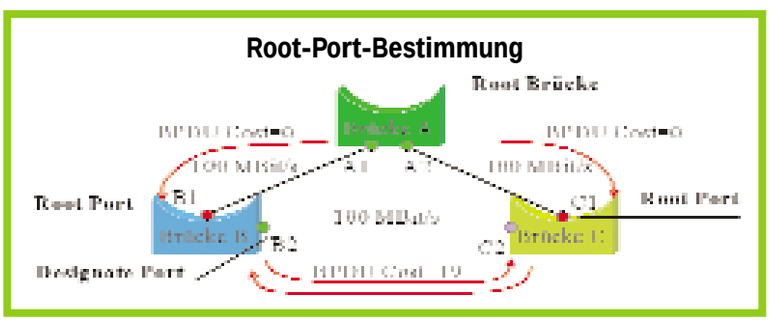

Bild: W. Schulte

Der schnellste Weg entscheidet

Root-Ports und Designate-Ports gehen nach dem Learning-Status in den Forwarding-Status. Damit die Adresstabellen stets aktuell sind, erhalten sie einen Zeitstempel. Empfängt die Brücke innerhalb einer einstellbaren Zeit kein Paket von der eingetragenen MAC-Adresse, wird der Eintrag gelöscht.

Der IEEE-802.1d-Standard unterscheidet, bezogen auf die Anschlüsse, den Port-Status wie Blocking, Listening, Learning und Forwarding, das heißt, ob ein Port blockiert oder weiterleitet und der Aufgabe eines Ports in der aktiven Topologie wie Root-Port, Designated-Port oder Blocking-Port. Zunächst ist es kein Unterschied, ob ein Port im Listening- oder im Blocking-Status ist, es werden in beiden Fällen keine neuen Adressen gelernt. Weiterhin kann einigermaßen sicher angenommen werden,

Wegekosten in Abhängigkeit von der Geschwindigkeit

Bandbreite	STP-Kosten
4 MBit/s	250
10 MBit/s	100
16 MBit/s	62
45 MBit/s	39
100 MBit/s	19
155 MBit/s	14
622 MBit/s	6
1 GBit/s	4
10 GBit/s	2

dass ein Port in Listening-Status entweder Root-Port oder Designated-Port wird und auf dem Weg ist, in den Forwarding-Status zu wechseln.

Das neue Protokoll Rapid Spanning Tree (RSTP) trennt den Status der Ports von seiner eigentlichen Aufgabe. Die Aufgaben der Ports ist bei RSTP in den Flag-Bits codiert.

Rapid-Spanning-Tree-Protokoll

Als das STP eingeführt wurde, waren die zeitlichen Anforderungen an eine Wiederinbetriebnahme eines Netzes nach einem Fehler oder bei eine Neukonfiguration im Minutenbereich. Cisco führte schon bald proprietäre Lösungen wie „Port Fast“ und „Uplink Fast“ in seinen Produkten ein, um schneller eine Konvergenz, das heißt, alle Brücken kennen eine einheitliche Topologie, zu erreichen. Die IEEE hat unter anderem durch die Mitarbeit

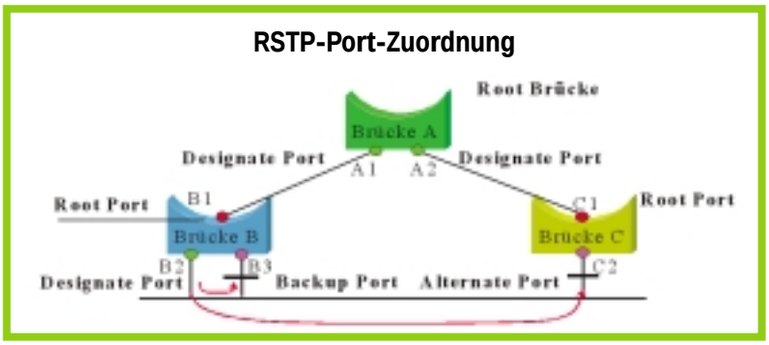

Bild: W. Schulte

RSTP hat eine neue Port-Anordnung

von Cisco im Jahr 2002 den neuen IEEE-802.1w-Standard veröffentlicht. RSTP ist eine natürliche Weiterentwicklung von STP mit einem erweiterten BPDU-Format. Werden in einem Netz RSTP und STP zusammen auf verschiedenen Brücken betrieben, so wird STP verwendet, allerdings unter Verlust des Vorteils von RSTP, der schnellen Konvergenz.

Der Port-Status bei 802.1d und 802.1w ist in der Tabelle unten angegeben. Der neue Standard führt den Status Discarding ein und verzichtet auf Disabled, Blocking und Listening.

Die Aufgaben des Root- und des Designated-Ports sind unverändert. Der Blocking-Port wird nun aufgeteilt in Backup- und Alternate-Port. Der Alternate-Port blockiert die BPDUs von anderen Brücken und bietet einen alternativen Weg zur Root-Brücke, falls der Root-Port ausfällt. Der Backup-Port blockiert den Empfang von BPDUs von der eigenen Brücke.

Status der Ports 802.1d versus 802.1w

STP (802.1d) Port-Status	RSTP (802.1w) Port-Status
Disabled	Discarding
Blocking	
Listening	
Learning	Learning
Forwarding	Forwarding

Im BPDU-Format wurden die Bits im Flag-Byte wie folgt erweitert:

- Die Aufgabe und Status des BPDU sendenden Ports wurde codiert,
- Es gibt einen Proposal/Agreement-Handshake-Mechanismus.

Der RSTP Message Type ist mit 2 angegeben, die Version ebenfalls mit 2. Dies bedeutet, dass Brücken mit STP die neuen RSTP BPDU verwerfen.

Neue BPDU-Behandlung

Bei 802.1d werden BPDUs von „Nicht-Root"-Brücken nur dann weitergeleitet, wenn sie selbst vorher eine BPDU von der Root-Brücke erhalten haben. Bei RSTP senden die „Nicht-Root"-Brücken, entsprechend der vom Netzadministrator gewählten Zeit, BPDUs selbstständig. BPDUs werden damit auch als ein „Keep-Alive"-Mechanismus zwischen den Brücken angewendet. Eine Brücke verliert ihre Verbindung zur Nachbarbrücke, wenn sie drei BPDUs hintereinander nicht sieht (Hello-Zeit). Dieser neue Mechanismus erlaubt somit viel schneller und direkter, den Verlust einer Verbindung festzustellen als bei STP.

Wenn ein Anschluss beim STP ein Designate-Port wurde, wartet er das 2-fache Forward Delay (2 x 15 Sekunden), bevor er in den Forward-Status wechselt. Bei RSTP wird ein Port im Discarding- oder Learning-Status das Proposal-Bit in der BPDU einschalten. Damit wird angezeigt, dass eine Neukonfiguration eingeleitet wurde.

Multiple-Spanning-Tree-Protokoll

Das neue MSTP von IEEE wurde unter anderem durch Beiträge von Cisco entwickelt und stützt sich weitgehend auf RSTP. Bisher wurde von IEEE mit STP eine gemeinsame, schleifenfreie Topologie für alle virtuellen LANs (VLANs) spezifiziert. Cisco kannte als proprietäre Lösung ein Per-VLAN-Spanning-Tree-Protokoll (PVST).

Mit MSTP kann ein effektiver Lastausgleich zwischen den VLANs erreicht werden, da eine Zuordnung von mehreren VLANs zu einer Instanz (Zusammenfassung mehrerer VLANs) 1 oder 2 erfolgt. Es wird jetzt nur eine Baumstruktur pro Instanz zugelassen.

Die Zuordnung der VLANs auf die Instanzen wird per MSTP-Regionen durchgeführt. Eine Region ist die Zusammenfassung von Switches unter einer gemeinsamen Administration. Jeder Switch benutzt eine eigene MSTP-Konfiguration, welche durch einen Namen, eine Revisionsnummer und einer Elemententabelle identifiziert ist. In der Elemententabelle befindet sich die Zuordnung von VLAN zur entsprechenden Instanz.

Um zu einer Gruppe zu gehören, muss jeder zugehörige Switch gemeinsame Attribute unterstützen. Die Parameter der Region sind in der BPDU enthalten. Erhält ein Switch eine BPDU, so vergleicht er die erhaltene BPDU mit den eigenen Eigenschaften. Sind diese Charakteristika ungleich, so wurde die BPDU aus einer anderen Region empfangen, das heißt, dieser Switch befindet sich an dem Übergang zweier Regionen.

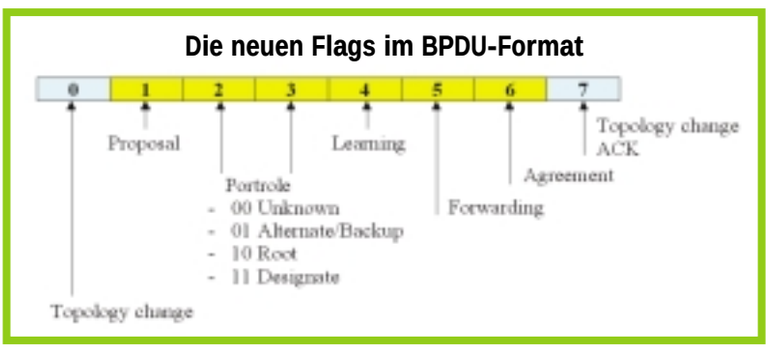

Bild: W. Schulte

Jedes Flag hat seine eigene Bedeutung

Zusammenfassung

IEEE hat mit den beiden neuen Standards 802.1w (RSTP) und 802.1s (MST) die immer größer und komplexer werdenden LANs noch mehr unter Kontrolle gebracht. Parallele Pfade, zum Beispiel zum Lastausgleich, und eine schnellere Konvergenz sind mit diesen beiden neuen Standards erfolgreich eingeführt worden. *(WM)*