

# DAS ETHERBOOK

Eine umfassende Einführung in die Netzwerktechnik

*Steve Graegert*





*für Harri & Rita*

Oktober 2001, Revision 1.1

**Copyright** Dieses Werk ist unter einem *Creative Commons Namensnennung-Keine kommerzielle Nutzung-Keine Bearbeitung 3.0 Deutschland* Lizenzvertrag lizenziert. Um die Lizenz anzusehen, gehen Sie bitte zu <http://creativecommons.org/licenses/by-nc-nd/3.0/de/> oder schicken Sie einen Brief an Creative Commons, 171 Second Street, Suite 300, San Francisco, California 94105, USA.

Jedoch dürfen Sie das Werk bzw. den Inhalt vervielfältigen, verbreiten und öffentlich zugänglich machen. Dabei sind folgende Bedingungen zu beachten:

#### **Namensnennung**

Sie müssen den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen.

#### **Keine kommerzielle Nutzung**

Dieses Werk bzw. dieser Inhalt darf nicht für kommerzielle Zwecke verwendet werden.

#### **Keine Bearbeitung**

Dieses Werk bzw. dieser Inhalt darf nicht bearbeitet, abgewandelt oder in anderer Weise verändert werden.

Wobei gilt:

#### **Verzichtserklärung**

Jede der vorgenannten Bedingungen kann aufgehoben werden, sofern Sie die ausdrückliche Einwilligung des Rechteinhabers dazu erhalten.

#### **Sonstige Rechte**

Die Lizenz hat keinerlei Einfluss auf die folgenden Rechte:

1. Die gesetzlichen Schranken des Urheberrechts und sonstigen Befugnisse zur privaten Nutzung;
2. Das Urheberpersönlichkeitsrecht des Rechteinhabers;
3. Rechte anderer Personen, entweder am Lizenzgegenstand selber oder bezüglich seiner Verwendung, zum Beispiel Persönlichkeitsrechte abgebildeter Personen.

#### **Hinweis**

Im Falle einer Verbreitung müssen Sie anderen alle Lizenzbedingungen mitteilen, die für dieses Werk gelten. Am einfachsten ist es, an entsprechender Stelle einen Link auf diese Seite einzubinden.

# Inhaltsverzeichnis

<b>I Ethernet</b>	<b>1</b>
<b>1 Local Area Networks - Eine Einführung</b>	<b>3</b>
1.1 Token Ring . . . . .	3
1.2 Fibre Distributed Data Interface . . . . .	4
1.3 Fibre Channel & ATM . . . . .	4
1.4 Schichtenmodell . . . . .	4
1.4.1 LAN-Schichten . . . . .	4
1.5 Basiskomponenten für LANs . . . . .	5
1.5.1 DTEs . . . . .	6
1.5.2 Transportmedien . . . . .	6
1.5.3 Network Interface Card . . . . .	6
1.5.4 Gerätetreiber . . . . .	6
1.6 SNMP, Monitoring und RMON . . . . .	6
1.6.1 SNMP-Architektur . . . . .	7
1.6.2 Management Information Base ( <i>MIB</i> ) . . . . .	7
1.6.3 SNMP Transport . . . . .	8
<b>2 LAN MAC Adressen</b>	<b>9</b>
2.1 Universal-MAC-Adressen . . . . .	9
2.2 Lokale MAC-Adressen . . . . .	10
2.3 Broadcast- und Gruppenadressen . . . . .	10
2.4 Individual/Group und Universal/Local Flag Bits . . . . .	10
2.5 Ethernet Adresskonventionen . . . . .	11
2.5.1 Ethernet Multicast Adressen . . . . .	11
2.6 Token Ring Adresskonventionen . . . . .	11
2.7 FDDI-Adresskonventionen . . . . .	12
<b>3 LAN-Strukturen</b>	<b>13</b>
3.1 Ethernet-LAN-Architektur . . . . .	13
3.1.1 1-Segment-Ethernet mit CSMA/CD . . . . .	13
3.1.2 Multisegment-LANs mit Repeatern . . . . .	14
3.1.3 Twisted-Pair-Verkabelung und Hubs . . . . .	15

3.1.4	Höhere Leistung durch den Einsatz von Bridges . . . . .	15
3.1.5	Collision Domains . . . . .	15
3.1.6	Mit Switches höhere Leistung erreichen . . . . .	16
3.1.7	Full-Duplex-Kommunikation mit Switches . . . . .	17
3.1.8	Router . . . . .	17
<b>4</b>	<b>Das CSMA/CD MAC Protokoll</b>	<b>19</b>
4.1	Klassische Ethernet-LANs . . . . .	19
4.1.1	Preamble und Interframe Gap . . . . .	19
4.1.2	Das CSMA/CD Protocoll . . . . .	20
4.1.3	Random Backoff . . . . .	20
4.2	Ethernet MAC-Frames . . . . .	21
4.2.1	Preamble und Startframe-Muster . . . . .	21
4.2.2	Ethernet MAC-Framegrößen . . . . .	21
4.2.3	MAC-Frameformate . . . . .	22
4.2.3.1	Quell- und Zieladressen . . . . .	22
4.2.3.2	Typ oder Länge . . . . .	22
4.2.3.3	Ethernet-Typen . . . . .	22
4.2.4	802.3 LLC/SNAP-Frames . . . . .	23
4.2.5	Herkunft der LLC- und SNAP-Header . . . . .	23
<b>5</b>	<b>Ethernet 10 Mbps PHY-Layer</b>	<b>25</b>
5.1	Basisband-Ethernet über 10Base5 . . . . .	25
5.1.1	10Base5-Coax-Verbindungen . . . . .	26
5.1.2	Transceiver (MAU)-Funktionen . . . . .	26
5.1.3	10Base5 Collision Domain Parameter . . . . .	26
5.1.3.1	Propagation Delay . . . . .	28
5.1.4	Probleme mit 10Base5 . . . . .	28
5.2	Basisband-Ethernet über Thin-Coaxkabel (10Base2) . . . . .	28
5.2.1	10Base2 Collision Domain Parameter . . . . .	30
5.2.2	Probleme mit 10Base2 . . . . .	30
5.3	10Mbps Twisted Pair (10Base-T) . . . . .	30
5.3.1	10Base-T-Segmente . . . . .	30
5.3.2	Halb-Duplex-Betrieb . . . . .	30
5.3.2.1	10Base-T Collision Domain Parameter . . . . .	31
5.3.2.2	10Base-T Collision Domain Topologie . . . . .	31
5.3.3	Voll-Duplex-Betrieb . . . . .	33
5.3.4	10Mbps Twisted-Pair-Kommunikation . . . . .	33
5.3.5	Hub- und Switch-Verbindungen . . . . .	34
5.3.6	Twisted-Pair Link Integrity Test (LIT) . . . . .	35

5.4	10 Mbps über Fibre Optic . . . . .	35
5.4.1	Merkmale der Fibre Optic . . . . .	36
5.4.2	FOIRL in der Collision Domain . . . . .	36
5.4.3	10Base-FL . . . . .	37
5.4.4	10Base-FB . . . . .	37
5.4.5	Struktur von Fibre Optic . . . . .	37
5.4.6	Multimode-Transmission . . . . .	38
<b>6</b>	<b>Ethernet 100 Mbps PHY-Layer</b>	<b>41</b>
6.1	Einführung . . . . .	41
6.2	100 Mbps Ethernet über Twisted-Pair . . . . .	41
6.2.1	100Base-TX . . . . .	42
6.2.2	Verwendung von CDDI . . . . .	42
6.2.3	100Base-T4 . . . . .	43
6.2.4	100Base-T2 . . . . .	44
6.3	100Base-FX und FDDI . . . . .	44
6.4	100Mbps Collision Domain Diameter . . . . .	45
6.4.1	Repeater-Klassen . . . . .	45
6.4.2	Collision-Domain-Konfiguration . . . . .	45
6.4.3	Vermeiden des Diameter-Problems mit Switches . . . . .	46
<b>7</b>	<b>Gigabit Ethernet-Architektur</b>	<b>47</b>
7.1	Einführung . . . . .	47
7.2	Gigabit-Konfigurationen . . . . .	47
7.3	Vollduplex Gigabit Ethernet . . . . .	47
7.3.1	Vollduplex-Repeater . . . . .	48
7.3.2	Jumbo Frames . . . . .	49
7.3.2.1	Auswirkungen auf den Durchsatz . . . . .	50
7.3.2.2	Vor- und Nachteile der Jumbo Frames . . . . .	50
<b>8</b>	<b>Gigabit PHY-Layer</b>	<b>51</b>
8.1	Merkmale des Gigabit Ethernet . . . . .	51
8.1.1	Auto-Negotiation . . . . .	52
8.1.2	Bidirektionale Gigabit-Transmissionen . . . . .	52
8.2	Physikalische Eigenschaften des Gigabit Ethernet . . . . .	52
8.3	1000Base-X-Technologie . . . . .	53
8.3.1	Die 8B/10B-Kodierung . . . . .	53
8.3.2	1000Base-SX- und -LX-Transmission: Laser und VCSELs . . . . .	54
8.4	1000Base-T-Technologie . . . . .	54
8.4.1	1000Base-T-Encoding . . . . .	54
8.4.2	Verkabelung . . . . .	54
8.4.3	Encoder, Decoder und Hybride . . . . .	55
8.4.4	Master/Slave-Timing . . . . .	55
8.4.5	Auto-Negotiation und Crossover . . . . .	55

<b>9</b>	<b>Standards</b>	<b>57</b>
9.1	Standardisierungsgremien ( <i>Standard Bodies</i> ) . . . . .	57
9.2	TIA/EIA-Kategorien . . . . .	57
9.2.1	Kabellayout . . . . .	58
9.3	UTP-Leistungsparameter . . . . .	59
9.3.1	Parameter für alle Ethernet-LANs . . . . .	59
9.3.2	Parameter für Hochgeschwindigkeits-LANs . . . . .	59
9.3.3	Parameterbeschreibung . . . . .	60
9.3.3.1	Jitter . . . . .	60
9.3.3.2	Dämpfung in Dezibel (dB) . . . . .	60
9.3.3.3	Near End Crosstalk (NEXT) . . . . .	60
9.3.3.4	Far End Crosstalk (FEXT) . . . . .	60
9.3.3.5	PSNEXT, PSELFEXT und worst Pair-to-Pair ELFEXT . . . . .	61
9.3.3.6	Dämpfung/Crosstalk-Verhältnis (ACR) . . . . .	61
9.3.3.7	Structural Return Loss und Return Loss . . . . .	61
9.3.3.8	Propagation Delay . . . . .	62
9.3.3.9	Delay Skew . . . . .	62
<b>10</b>	<b>Auto-Negotiation</b>	<b>63</b>
10.1	Auto-Negotiation für TP-Schnittstellen . . . . .	63
10.1.1	AN-Funktionalität von TP-Schnittstellen . . . . .	64
10.1.2	AN-Unterstützung für Flußkontrolle . . . . .	65
10.1.3	Ermitteln der Master/Slave-Timing-Regeln . . . . .	65
10.1.4	Parallel Detection für TP-Kabel . . . . .	65
10.1.5	Datenaustausch während der Auto-Negotiation über TP . . . . .	65
10.1.6	Base Page, Message Page und Unformatted Pages . . . . .	66
10.1.7	Message Pages für 1000Base-T . . . . .	68
10.2	Auto-Negotiation für 1000Base-X-Interfaces . . . . .	68
10.2.1	Implementation der 1000Base-X Auto-Negotiation . . . . .	69
<b>II</b>	<b>Routing und Switching</b>	<b>71</b>
<b>11</b>	<b>Bridges und 2nd Layer Switching</b>	<b>73</b>
11.1	Hauptfunktionen . . . . .	73
11.2	Zusatzfunktionen . . . . .	73
11.2.1	Collision Domains . . . . .	74
11.2.2	Transparente Bridges . . . . .	74
11.2.2.1	Bridges in Twisted-Pair-Umgebungen . . . . .	75
11.3	Interna . . . . .	76
11.3.1	Lernen, Lernen und nochmals Lernen . . . . .	76

11.3.2	Statische Filterinformationen . . . . .	77
11.3.2.1	Effizientes Forwarding mit statischen Filtereinträgen . . . . .	77
11.3.2.2	Beispiel einer statischen Filtertabelle . . . . .	77
11.3.2.3	Sicherheitsaspekte . . . . .	78
11.3.2.4	Leistungsverbesserung durch Protokollfilter . . . . .	78
11.4	Architektur von Layer-2-Switches . . . . .	79
11.4.1	Store-and-Forward und Cut-Through . . . . .	79
11.4.2	Pallele Verarbeitung durch Asics . . . . .	80
11.5	Redundanz im LAN . . . . .	80
11.5.1	Spanning Tree Protocol . . . . .	80
11.5.2	Link Aggregation . . . . .	81
11.6	Multicast-Verkehr in LANs . . . . .	81
11.6.1	IGMP Snooping und GARP . . . . .	81
11.6.2	Funktionale Struktur eines Layer-2-Switches . . . . .	82
<b>12</b>	<b>Das Spanning Tree Protocol</b>	<b>83</b>
12.1	Initialisierung der Topologie . . . . .	84
12.2	Änderungen der Topologie . . . . .	84
12.3	Portstatus . . . . .	85
12.4	Zustandsänderungen . . . . .	85
12.4.1	Beispiel . . . . .	86
<b>13</b>	<b>Switches und Multicast-Traffic</b>	<b>93</b>
13.1	Multicasting . . . . .	93
13.1.1	Warum müssen wir Multicasting steuern? . . . . .	94
13.1.2	IP-Multicasting . . . . .	94
13.2	IGMP Snooping . . . . .	95
13.2.1	Nachteile des IGMP-Snooping . . . . .	95
13.3	GARP Multicast Registration Protocol . . . . .	96
13.3.1	Die Registrierung . . . . .	96
13.4	Generic Attribute Registration Protocol . . . . .	96
<b>14</b>	<b>Link Aggregation</b>	<b>99</b>
14.1	Link Aggregation im Einsatz . . . . .	99
14.2	Konzepte und Vorgehensweisen . . . . .	100
14.3	Parameter von Link Aggregation . . . . .	100
14.4	Das Link Aggregation Control Protocol . . . . .	101
14.4.1	LACP-Nachrichten . . . . .	101
14.4.2	Rahmenübermittlung . . . . .	102

<b>15 Routing und Switching</b>	<b>103</b>
15.1 Routing – Ein Überblick . . . . .	103
15.1.1 IP-Adressen und MAC-Adressen . . . . .	104
15.1.2 Routing außerhalb des LANs . . . . .	104
15.2 Wie funktioniert Routing? . . . . .	105
15.2.1 Aufbau einer Routing-Tabelle . . . . .	105
15.3 Router-Architektur und MPLS . . . . .	107

# Abbildungsverzeichnis

1.1	Token Ring Topologie . . . . .	3
1.2	TCP/IP und OSI-Kommunikationsmodelle . . . . .	5
1.3	LAN Layers . . . . .	5
1.4	Einfaches Ethernet Frame-Format . . . . .	5
1.5	Die Rolle des Gerätetreibers. . . . .	6
1.6	Das SNMP-Modell . . . . .	7
2.1	Bedeutung der Address Flag Bits . . . . .	10
2.2	Initiale Bits in Token Ring Frames. . . . .	12
3.1	Übertragung eines Frames über ein Ethernet-Segment. . . . .	14
3.2	Drei Segmente, die über einen Repeater verbunden sind. . . . .	14
3.3	Sterntopologie. . . . .	15
3.4	3-Segment-LAN mit einer Bridge. . . . .	16
3.5	Layer-2-Switch. . . . .	16
3.6	Verbindung zweier LANs und Bereitstellung eines WAN-Links . . . . .	17
4.1	Frame Preambles und Interframe Gaps. . . . .	19
4.2	Allgemeines Ethernet Frameformat. . . . .	21
4.3	Preamble, SFD und MAC-Frame. . . . .	21
4.4	DIX- und 802.3-Frames. . . . .	22
4.5	Format eines 802.3-Frame mit Protokollfeld und gesetztem Ethertype-Feld. . . . .	23
4.6	Sublayer des Data Link Layer. . . . .	23
4.7	Format des LLC-Header. . . . .	24
4.8	Format des SNAP-Header. . . . .	24
5.1	Verzweigtes Bus-Ethernet. . . . .	25
5.2	Eine Station ist mit dem Coax-Ethernet verbunden. . . . .	26
5.3	Thick-Ethernet-Tranceiver. . . . .	27
5.4	Segmente mit Repeatern verbinden. . . . .	27
5.5	Coax-Ethernet-LAN erweitert durch Repeater. . . . .	28
5.6	Ethernet NIC mit integriertem Transceiver und BNC-Verbinder. . . . .	29
5.7	10Base2-Segmente als Daisy-Chains. . . . .	29

5.8	10Base-T-Segmente. . . . .	31
5.9	Halb-Duplex-Kommunikation via Hub. . . . .	31
5.10	Ein einfaches 10Base-T-LAN. . . . .	32
5.11	Komponenten eines 10Base-T-LANs. . . . .	32
5.12	Interconnect und Cross-Connect. . . . .	32
5.13	Verkabelung eines Gebäudes. . . . .	33
5.14	Senden und Empfangen über Twisted-Pair. . . . .	34
5.15	Pins, Verdrahtung und RJ45-Stecker. . . . .	34
5.16	Port- und Kabeltypen. . . . .	35
5.17	Senden und Empfangen. . . . .	36
5.18	FOIRL-Längenrestriktionen für einen Pfad mit 5 Segmenten. . . . .	37
5.19	Struktur eines optischen Datenträgers. . . . .	38
5.20	Übertragungsmodi SIF und GIF. . . . .	38
6.1	100Base-T4-Transmission. . . . .	43
6.2	100Base-T4 Cross-Connection. . . . .	44
6.3	DTE-zu-DTE-Verbindungen. . . . .	45
6.4	Entfernungen für einen Klasse-1-Hub. . . . .	46
6.5	Entfernungen für einen Klasse-2-Hub . . . . .	46
6.6	Entfernungen für zwei Klasse-2-Hubs . . . . .	46
7.1	100Mbit mit Gigabit verbinden. . . . .	48
7.2	Switch und <i>Buffered/Full-Duplex</i> Repeater. . . . .	48
7.3	Gigabit Ethernet Vollduplex-Repeater. . . . .	49
8.1	Gigabit-Übertragungsmechanismen. . . . .	52
8.2	1000Base-T-Übertragung mit 4 Twisted-Pair-Kabeln. . . . .	55
8.3	Crossover MDI/MDI-X-Pinlayout für 1000Base-T. . . . .	56
9.1	Near End Crosstalk (NEXT). . . . .	60
9.2	Far End Crosstalk (NEXT). . . . .	61
10.1	NLPs und FLPs . . . . .	66
10.2	Format einer TP Base Page. . . . .	66
10.3	Format der Message Pages und Unformatted Page. . . . .	67
10.4	Eine Auto-Negotiation Page. . . . .	69
10.5	AN Base Message für 1000Base-X. . . . .	70
11.1	Ethernet LAN mit 4 Segmenten bestehend aus drei Collision Domains. . . . .	75
11.2	Ethernet LAN mit drei Collision Domains auf Twisted-Pair-Basis. . . . .	75
11.3	Twisted-Pair-Ethernet mit Switches. . . . .	76
11.4	Beispielnetzwerk für statisches Forwarding. . . . .	78

11.5	Redundantes Switching in LANs . . . . .	80
11.6	Schichten eines Layer-2-Switches . . . . .	82
12.1	Zustände der Ports und die möglichen Zustandsänderungen . . . . .	86
12.2	Herleitung der aktiven Topologie anhand eines Beispielnetzwerks. . . . .	87
12.3	Ermittlung der Root Ports. . . . .	88
12.4	Ermittlung der Designated Ports. . . . .	89
12.5	Aktive Topologie des Beispielnetzwerks. . . . .	90
12.6	Aktive Topologie des Beispielnetzwerks. . . . .	90
12.7	Aktive Topologie des Beispielnetzwerks. . . . .	91
12.8	Aktive Topologie des Beispielnetzwerks. . . . .	91
14.1	Link Aggregation in der Praxis. . . . .	99
15.1	Routing zwischen zwei Netzwerken mit unterschiedlichen Architekturen. . . . .	103



# Tabellenverzeichnis

4.1	Ethernet-Typen. . . . .	23
5.1	10Base5 Collision Domain Parameter. . . . .	27
5.2	10Base2 Collision Domain Parameter. . . . .	30
5.3	10Base-T Collision Domain Parameter. . . . .	31
6.1	Fast Ethernet Technologien. . . . .	41
7.1	Vollduplex-Ethernet-Parameter. . . . .	48
8.1	Vollduplex-Ethernet-Parameter (MMF = <i>multimode fibre</i> , SMF = <i>singlemode fibre</i> ). . . . .	53
9.1	Vollduplex-Ethernet-Parameterd . . . . .	58
10.1	Vollduplex-Ethernet-Parameter (* wurde nicht implementiert). . . . .	64
10.2	Base Page Technology Ability Fields. . . . .	67
10.3	Felder der Technology Ability Fields . . . . .	68
10.4	Master/Slave-Zuordnung . . . . .	68
10.5	Message Codes. . . . .	69
11.1	Beispiel einer typischen Filtertabelle. . . . .	77
11.2	Beispielhafte statische Filtertabelle. . . . .	78



Teil I

Ethernet



# Kapitel 1

## Local Area Networks - Eine Einführung

Local Area Networks (*LANs*) verbinden eine Gruppe Computer untereinander, so dass sie miteinander kommunizieren können. Der Begriff *local* stammt aus früheren Zeiten, in denen die nur Überbrückung kurzer Strecken möglich war. Heutzutage sind LANs in ganze Gebäudekomplexe integriert und dienen der unternehmensinternen Kommunikation. Mögliche Anwendungen sind Intranet-Lösungen für UM (*Unified Messaging*) oder Groupware, verteilte Storage-Netzwerke und verteiltes Rechnen (*Number Crunching*).

### 1.1 Token Ring

Entstanden ist TR (*Token Ring*) im Jahre 1981 in den Züricher IBM Research Labs und wurde 1985 dem Markt zugänglich gemacht. TR wurde als Herausforderung gegenüber Ethernet entwickelt und floss nach Vorlage der Spezifikation bei der IEEE als ANSI-, ISO- und IEEE-Standard in die Ethernet-Gruppe ein.

TR wird im Ring betrieben, was der Name bereits verdeutlicht. Die Topologie ist sehr komfortabel, denn die Arbeitsplätze sind an sog. *Concentrators* angebunden, die ihrerseits in einem Verteiler untergebracht sein können, um entfernte Standorte über festverdrahtete Netzwerke erreichen zu können. Abbildung 1.1 illustriert den einfachen Aufbau eines TR-Netzwerkes.

Eine spezielle Nachricht, *Token* genannt, wird innerhalb des Rings von Station zu Station gereicht. Jede Station kann Transaktionsanfragen stellen, so dass beim Eintreffen des Tokens die jeweilige Station mit dem Datentransfer beginnen kann, allerdings nur eine begrenzte Zeit, denn das Token wird der Maschine bei Ablauf eines Timers entzogen und von der nächsten Maschine mit einer Transaktionsanfrage aufgenommen, die ihrerseits mit dem Transfer beginnt.

Das ursprüngliche TR operierte zunächst mit 4 Mbps, und später mit 16 Mbps. Eine 100 Mbps-Version (HSTR, *High Speed Token Ring*) wurde 1998 eingeführt, obwohl es bereits Anfang der 90er entwickelt wurde. Die Kunden begrüßten zwar den Leistungszuwachs, jedoch verhinderten die teuren Schnittstellen eine größere Verbreitung.

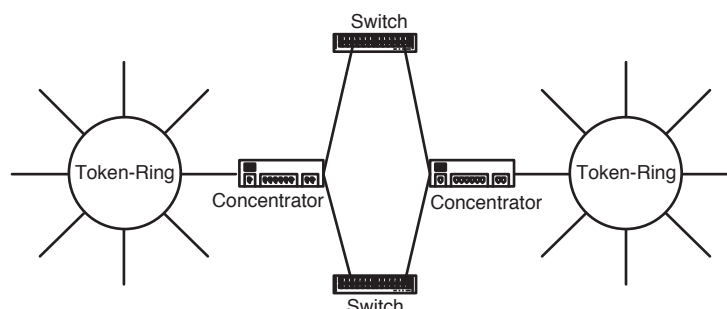


Abbildung 1.1: Token Ring Topologie

Es existiert eine Token Ring Protokoll Version, die speziell für die Bedürfnisse an Switches angeschlossener Arbeitsstationen angepasst wurde. In einem geschwichten Netzwerk kann jeder gleichzeitig kommunizieren, was Tokens zur Steuerung der Transfers überflüssig macht. Allerdings sind TR-Switches deutlich teurer als gewöhnliche Ethernet-Switches und mit der Einführung von Gigabit- und 10Gigabit-Ethernet sahen die meisten Nutzer diese Technologie als überholt an. Ethernet ist preiswert, schnell und skalierbar.

## 1.2 Fibre Distributed Data Interface

Im Jahre 1983 erblickte eine revolutionäre Technologie das Licht der Welt. Es wurde als *Supernetwork* bezeichnet, weil es bereits in frühen Jahren mit einer Geschwindigkeit von 100 Mbps operierte. FDDI wurde durch die *ANSI X3T9.5 Task Group* spezifiziert und als unmittelbarer TR-Konkurrent gehandelt. FDDI setzt genau wie Token Ring auf eine Ring-Topologie auf und wird durch Concentrators verteilt. Das Übertragungsprotokoll ist mit dem von Token Ring vergleichbar.

FDDI ist extrem komplex und erfordert einen hohen Kostenaufwand für die Installation. Daher beschränkt sich der Einsatz dieser Technologie auf Hochverfügbarkeitslösungen in Form von Doppelringkonstruktionen und Backbone-Funktionen zur Verbindung unterschiedlicher Netzwerke wie etwa Token Ring und Ethernet.

## 1.3 Fibre Channel & ATM

Fibre Channel Netzwerke operieren mit enormen Geschwindigkeiten und basieren entweder auf Ring-, Stern- und Mischtopologien zwischen beiden. Die hohen Transferraten sind besonders für den Einsatz sog. *Storage Attached Networks* (SANs) interessant, um beispielsweise schnellen Zugriff auf Speicher- und Tape-Farmen zu gewährleisten. Desweiteren wird es zur Verbindung lokaler Netzwerke und als Backbone-Verbindungen für andere periphere Netzwerke und Geräte verwendet.

Der *Asynchrone Transfer Modus* (ATM) wurde als WAN-Technologie eingeführt und später für LANs aufgebohrt. Somit kann ein ATM-Netzwerk über viele WAN-Verbindungen bis in das LAN hinein aufgebaut werden und ermöglicht Kommunikation mit Token Ring und Ethernet Netzwerken.

Die Schlüssel-Features von Fibre Channel und ATM sind:

- beide benötigen einen expliziten Verbindungsaufbau bevor die Kommunikation beginnen kann
- beide bieten QoS-Funktionalität, so dass Fehler- und Delay-Levels in bestimmten Grenzen gehalten werden können.

## 1.4 Schichtenmodell

In den frühen 80er Jahren begann die ISO (*International Standards Organization*) eine Serie von OSI-Dokumenten (*Open Systems Interconnect*) herauszugeben, die neue Standards für die Kommunikation zwischen Systemen unterschiedlicher Architektur beschreiben. Eines dieser Dokumente beschreibt das OSI-Schichtenmodell. Die IETF (*Internet Engineering Task Force*) entwickelte ein einfacheres Modell, das als TCP/IP-Modell in die Geschichte der Netzwerktechnik eingegangen ist. Abbildung 1.2 stellt beide Modelle gegenüber.

### 1.4.1 LAN-Schichten

Die LAN-Technologie operiert auf Level 1 und 2. Die Aufgabe von Layer 1 (PHY) ist die Übertragung von 0s und 1s über das Transport-Medium. Der LAN-PHY-Layer-Standard beschreibt folgende Spezifikationen:

- die Eigenschaften der Übertragungsmedien

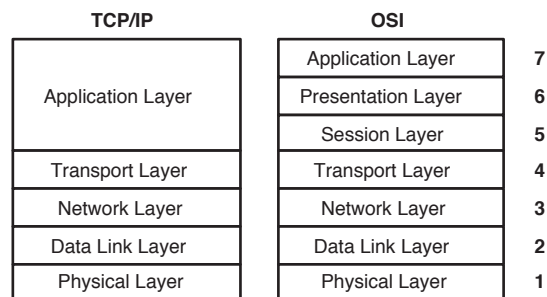


Abbildung 1.2: TCP/IP und OSI-Kommunikationsmodelle

- die physikalischen Signale, die verwendet werden, um die digitale Information zu repräsentieren
- andere physikalische Spezifikationen, wie z.B. Stecker und Steckverbindungen, maximale Länge des Mediums und optimale Umweltbedingungen (Temperatur, EM-Strahlung) für einen reibungslosen Betrieb.

Layer 2, der *Data Link Layer*, teilt die 0s und 1s in logische Dateneinheiten, die als Rahmen (engl. *Frames*) bezeichnet werden. Abbildung 1.3 zeigt die Aufgaben der LAN-Schichten.

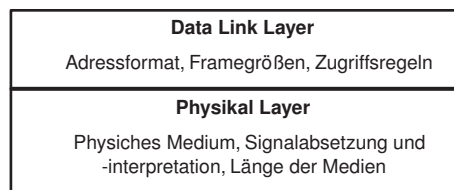


Abbildung 1.3: LAN Layers

Der Data-Link-Layer-Standard für LANs beschreibt folgende Spezifikationen:

- das allgemeine Format der Datenframes und die Bedeutung eines jeden Feldes, das in dem Frame enthalten ist
- die Regeln, die eingehalten werden müssen, um Zugriff auf das Medium zu erhalten, damit der Frame übertragen werden kann.
- das Format und die Interpretation der hardwareseitigen LAN-Adressen, die als MAC-Adressen bezeichnet werden.

MAC-Adressen (physische Adressen) sind lokal, d.h. Sie haben nur innerhalb des gleichen LANs Bedeutung. Sie dienen den Kommunikationspartnern zur Übermittlung von Nachrichten innerhalb eines LANs. Das allgemeine Frame-Format wird in Abbildung 1.4 illustriert.



Abbildung 1.4: Einfaches Ethernet Frame-Format

## 1.5 Basiskomponenten für LANs

Bevor wir tiefer in die technischen Details eintauchen und die Transportprotokolle analysieren, klären wir zunächst die Mindestanforderungen, die für den Betrieb eines lokalen Netzwerks erforderlich sind. Es sei hier angemerkt, daß die Netzwerkkomponenten *Router*, *Switches*, *Hubs* und *Repeater* später besprochen werden.

### 1.5.1 DTEs

Systeme, die Frames absetzen und empfangen können, werden *Stationen* oder auch *Nodes* genannt. Stationen können Client- und Serversysteme gleichermaßen sein, wie auch Router, Switches oder andere aktive Komponenten. Ein System, das Frames empfangen und senden kann, wird auch als DTE (*Data Termination Equipment*, dt. Datenendeinrichtung) bezeichnet.

### 1.5.2 Transportmedien

Es gibt eine Vielzahl Transportmedien, die LANs verwendet werden können. Einige davon sind: UTP/STP (*Unshielded/Shielded Twisted Pair*), Coaxialkabel, Lichtwellenleiter (LWL) und inzwischen auch diverse Wireless-Lösungen.

### 1.5.3 Network Interface Card

Der Verbindungspunkt der eigentlichen Station und dem LAN wird als NI (*Network Interface*, dt. Netzwerkschnittstelle) bezeichnet. Die Netzwerkoperationen werden dann auf der NIC (*Network Interface Card*, Netzwerkschnittstellenkarte) ausgeführt, die über eine Hardwareschnittstelle mit dem DTE verbunden ist. Allgemein kann die NIC auch als Adapter, Adapterkarte oder einfach Netzwerkkarte bezeichnet werden. Jeder Netzwerkadapter hat eine einzigartige MAC-Adresse. Alle Frames, die gesendet werden, beinhalten die Quell- und Zieladressen der jeweiligen NICs. Daher wird oft auch von der NIC-Adresse gesprochen.

### 1.5.4 Gerätetreiber

Jedes Gerät in einem Computer benötigt spezielle Software, die zwischen der Systemsoftware (Betriebssystem) und der Software, die auf der Hardware integriert ist, vermittelt. Diese Softwareteile bezeichnet man allgemein als Gerätetreiber. Die Protokollschichten interagieren mit den Gerätetreibern, um die Daten übertragen zu können. Dabei wissen die Schichten nicht, wie der Treiber die angeforderten Dienste bereitstellt, sondern greifen über ein API (*Application Programming Interface*, dt. Programmierschnittstelle für Anwendungen) auf die Funktionen der Treiber zu. Über diese Schnittstellen werden Daten empfangen und gesendet. Der Kommunikationsweg wird in Abbildung 1.5 dargestellt.

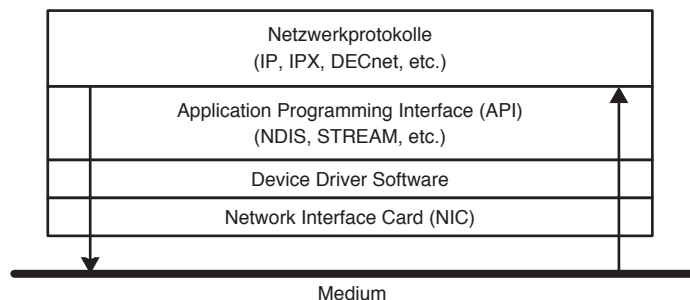


Abbildung 1.5: Die Rolle des Gerätetreibers.

Jeder Netzadapter benötigt den jeweiligen Treiber des Herstellers, da die Hardware-Implementationen oft sehr unterschiedlich sind, bzw. die Einsatzgebiete der NICs andere Software erfordern. Die NICs kooperieren hingegen mit allen Protokollen, wenn die vereinbarten Standards eingehalten werden, d.h. Standard-APIs müssen eingesetzt und ein spezifischer Befehlssatz bereitgestellt werden. Damit ist es möglich, NICs in einem System zu ersetzen, ohne Veränderungen am Netzwerk vornehmen zu müssen.

## 1.6 SNMP, Monitoring und RMON

Moderne LANs bestehen oft aus hunderten und tausenden von Workstations, die alle verwaltet, überwacht und gewartet werden müssen. Das SNMP (*Simple Network Management Protocol*) ist die am

weitesten verbreitete Netzwerkmanagement-Technologie. Theoretisch ist jedem Netzwerkgerät möglich über SNMP administriert zu werden, bzw. im SNMP-Netzwerk teilzunehmen. Eine Networkmanagement-Station übernimmt gewissermaßen die Masterstellung im Netzwerk, so daß der Administrator alle sich im SNMP-Netzwerk befindlichen Workstations erreichen und verwalten kann.

Monitors, auch *Probes* genannt, haben die Aufgabe im Netz mitzulauschen, um Fehler möglichst früh zu erkennen und den zuständigen Administrator umgehend zu benachrichtigen. Desweiteren sind Aufzeichnung und Analyse des Netzverkehrs auf den unterschiedlichsten Ebenen realisierbar, so daß tiefe Einblicke in den Netzwerkbetrieb gewährt werden. Aus dem Grund ist in diesem Zusammenhang besonderes Augenmerk auf die Sicherheitsaspekte zu richten.

*Remote Monitoring* (RMON) beschreibt die Kooperation eines Monitors (*Probes*) mit einem *SNMP-Manager*. Hier werden die Funktionalitäten vereint, um die Überwachung und Steuerung des Netzwerks zu optimieren. Die einzelnen Features solcher Technologien wird in den folgenden Abschnitten beschreiben.

### 1.6.1 SNMP-Architektur

SNMP folgt dem Datenbankmodell, d.h. alle Geräte und Arbeitsstationen stellen eine Informationsdatenbank zur Verfügung, mit deren Hilfe der SNMP-Manager folgende Standardinformationen über die SNMP-Agents abfragt:

- Konfigurationseinstellungen
- Statusinformationen (Betriebszustand, Auslastung, etc.)
- Leistungsstatistiken

Abbildung 1.6 zeigt die Elemente des SNMP-Modells. *Requests* von Anwendungen des Managers an den *Agent* können die Informationen abfragen und updaten, indem an den Agent via SNMP entsprechende Nachrichten gesendet werden. Falls ein signifikantes Ereignis, wie etwa ein Reboot oder ein ernster Fehler auftritt, kann der Agent dieses Problem durch eine sog. *Trap* dem Manager mitteilen. Es gibt eine Vielzahl Hersteller, die die SNMP-Funktionalität um graphische Darstellungen und ähnliche Hilfsmittel mit Hilfe eigener Software erweitern.

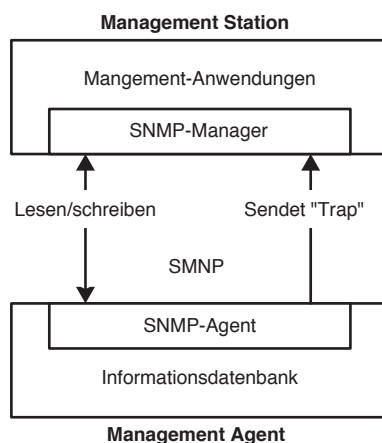


Abbildung 1.6: Das SNMP-Modell

### 1.6.2 Management Information Base (MIB)

Wie bereits oben beschreiben, verfügen alle *Management Agents* über eine Informationsdatenbank die alle notwendigen Variablen und Attribute enthält, die der Administrator benötigt, um die Arbeitsstationen zu verwalten. Eine solche Datenbank wird als *Management Information Base* (MIB) bezeichnet. Einige Variablen werden im Folgenden genannt:

- Beschreibung des Gerätes oder Arbeitsplatzes
- Anzahl der Netzwerkadapter, Schnittstellen und deren Typen
- Zahl der ein- und ausgehenden Frames.

Die MIBs sind standardisiert, so daß jeder Hersteller auf diese Technologie mit seiner eigenen Software aufsetzen kann. Es existieren eine Vielzahl an Dokumenten, die z.T. von IETF als RFCs (*Request For Comment*) veröffentlicht wurden und maßgeblich für die Verbreitung des Standards verantwortlich sind.

Die MIBs werden von einer Expertengruppe innerhalb der IETF verwaltet und weiterentwickelt. Die entstehenden Elemente der Datenbank sind in Einheiten aufgeteilt, die als *MIB Modules* bezeichnet werden, da für jede neu eingeführte Technologie ein neues MIB Modul von der Expertengruppe entwickelt wird. Der Administrator kann die MIB Modules auf die zu managenden Geräte aufspielen und den Agent sofort über SNMP verwalten.

### 1.6.3 SNMP Transport

SNMP-Anfragen, Updates, Antworten und Traps werden zwischen den Systemen durch irgendein geeignetes Transportprotokoll übertragen. SNMP wurde für die Bedürfnisse des Internets angepaßt, was den Einsatz einfacher TCP/IP-Protokolle wie z.B. UDP ermöglicht. Andere Protokolle können SNMP beispielsweise auch über ATM-Netzwerke senden und empfangen.

# Kapitel 2

## LAN MAC Adressen

Wenn ein System mit einem anderen innerhalb des eigenen LAN kommunizieren möchte, muß das Zielsystem eindeutig identifiziert werden können. *Media Access Control*-Adressen (auch physikalische Adressen genannt) ermöglichen eine Adressierung im einfachsten Level des Protokollstapels. Jede NIC, die mit einem LAN verbunden ist, wird von allen anderen Systemen mit Hilfe der MAC-Adressen eindeutig identifiziert.

Der Ansatz der IEEE sah vor, daß jede vergebene MAC-Adresse weltweit einzigartig sei und somit das häufige Wechseln der NIC von einem Netz in das andere problemlos realisiert werden kann. Dadurch können MAC-Inkompatibilitäten nicht auftreten und eine zusätzliche Fehlerquelle ist eliminiert worden. Wir wollen uns in diesem Abschnitt mit der Arbeit der IEEE befassen, die die einheitliche Vergabe der Adressen sicherstellt.

### 2.1 Universal-MAC-Adressen

Während des Herstellungsprozesses jeder NIC für Ethernet, Token Ring oder FDDI wird die MAC-Adresse festgelegt. Diese Adresse besteht aus einer 48 Bit (6 Bytes) Zeichenkette. Es ist allgemein üblich eine MAC-Adresse mit einem X' gefolgt von 6 Paaren Hexadezimalzahlen getrennt durch Dashes („-“). Eine Beispiel-Adresse könnte demnach etwa so aussehen: X'03-7F-BE-V9-7C-1E.

Die IEEE hat den Herstellern von NICs gegenüber administrative Funktion bei der Vergabe der MAC-Adressen eingeräumt. Jeder Hersteller muß sich an die Vorgaben der IEEE halten, damit jede NIC mit einer einzigartigen Adresse ausgestattet wird. Der Vorgang ist folgendermaßen zu verstehen:

- Der Hersteller muß einen schriftlichen Antrag auf die Vergabe von MAC-Adressen stellen und kann dann gegen eine Registrierungsgebühr die MAC-Adressen erhalten.
- Die IEEE im Gegenzug wird dem Hersteller ein 3-Byte-Prefix (*Organizational Unique Identifier*, OUI) zuweisen, so daß der Hersteller jeder NIC eindeutig identifiziert werden kann.

Beispiele sind:

X'00-60-08 (3Com) X'00-00-00 (Xerox)

Für genauere Angaben über die OUIs besuchen Sie bitte die offizielle Homepage der IEEE [?].

Nachdem der Hersteller seine OUI erhalten hat, können die übrigen 24 Adress-Bits für die Auflösung des zweiten Teils der MAC-Adresse verwendet werden. In Zahlen bedeutet dies, daß der Hersteller nun  $2^{24}$  (16.777.216) Adapter mit eindeutigen MAC-Adressen herstellen kann. Diese global einzigartigen Adressen werden aus diesem Grund als Universaladressen bezeichnet. Der Hersteller ist außerdem berechtigt eine weitere OUI zu beantragen, falls die erste bereits erschöpft sein sollte.

## 2.2 Lokale MAC-Adressen

MAC-Adressen im lokalen Netzwerk sind in den Augen der Administratoren oft nur eine zufällige Aneinanderreihung von Zeichen. Viele Admins schreiben die MAC-Adressen um, damit z.B. die Standorte bestimmter Geräte besser bestimmen kann. Die gesammte Adresse könnte in mehrere Teile aufgeteilt werden, die eine Einteilung in Gebäude, Etage, Raum und Patchfeld ermöglichen.

Diese Form der eigenständigen Vergabe der Adressen erfordert viel Sorgfalt und ausreichende Planung, damit sich der erhebliche Mehraufwand auszahlt. Allerdings sind die Vorteile, die sich ergeben besonders für das Troubleshooting von Netzwerken und Arbeitsstationen überragend.

Die IEEE hat für die Selbstvergabe lokaler MACs eine Regelung getroffen, die auch in diesem Bereich Sicherheit bringen soll, damit konsistente MAC-Umgebungen nicht durch Universal-MACs kompromittiert werden. Ein Bit im ersten Adressfeld muß 1 für local MACs und 0 für alle Universal-MACs sein.

## 2.3 Broadcast- und Gruppenadressen

Eine Adresse, die nur ein System adressiert, wird auch als Unicast-Adresse bezeichnet. Ein Frame, der an eine Unicast-Adresse gesendet wird, hat nur genau ein Sytem als definiertes Ziel. Eine Workstation, die an eine Multiaccess-LAN angebunden ist, kann einen oder mehrere Frames an eine oder mehrere Maschinen senden. Dazu wird eine Broadcast-Adresse verwendet. Alle NICs absorbieren Frames mit der Broadcast-Adresse X'FF-FF-FF-FF-FF-FF.

Diese Adresse besteht aus 48 Bits, die alle auf eins (1) gesetzt sind. Manchmal ist es interessant einen Frame nicht an alle Systeme eines LANs zu senden, sondern eine ausgewählte Gruppe mit Informationen zu versorgen. Gruppenadressen (Multicast-Adressen) bieten hier Abhilfe. Ein Bit in dem ersten Adress-Byte ist 0 für individuelle Adressen und 1 für Gruppenadressen gesetzt.

Um einen Adapter mit der Fähigkeit, Multicast-Frames zu absorbieren, auszustatten, ist es erforderlich, das eine höhere Schicht dem Gerätetreiber eine Nachricht sendet, die diesen veranlaßt, die Multicast-Fähigkeit der NIC zu aktivieren und die neue Multicast-Adresse hinzuzufügen.

## 2.4 Individual/Group und Universal/Local Flag Bits

Die untersten beiden Bits in dem ersten Byte einer MAC-Adresse sind speziell für die Identifizierung der Adresse vorgesehen:

- Individualadressen (0) oder Gruppenadresse (1)
- Universal (0) oder lokale Adresse (1)

Abbildung 2.1 zeigt die Positionen dieser Bits, wenn jedes Byte der MAC-Adresse in eine binäre Form übertragen wird, wobei die untersten Bits links dargestellt werden.

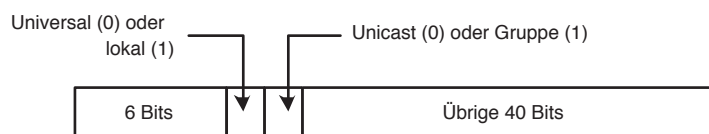


Abbildung 2.1: Bedeutung der Address Flag Bits

## 2.5 Ethernet Adresskonventionen

Das IEEE 802.3 Komitee spezifizierte, daß jedes Byte eines Ethernet-Frames mit dem untersten Bit zuerst gesendet werden muß. Diese Konvention wird oft auch als *little endian order* (i386, AMD) bezeichnet. In der Prozessortechnik existiert daneben auch die *big endian order* (z.B. SPARC, Alpha), was vielen Programmierern Kopfzerbrechen bei der Fehlersuche bereitet. Die Standards spezifizieren nicht in welcher Reihenfolge die Bytes in einem Computer gespeichert werden müssen oder übertragen werden sollen. Jedes System, daß mit einem anderen kommuniziert weiß genau, welche „Order“ erforderlich ist.

Nehmen Sie zur Kenntnis, daß die Übertragung des untersten Bits zunächst sicher stellt, daß das erste Bit, das abgesetzt wird, die Unicast/Multicast-Zugehörigkeit ausdrückt und das zweite somit die Unicast-Adresse bzw. lokale Adresse repräsentiert. Ein Beispiel: wenn X'02 das erste gesendete Byte darstellt, ist das erste Bit, das gesendet wird eine 0 (Unicast-Adresse) und das zweite Bit eine 1 (lokale Adresse). Die Reihenfolge der Transmission ist durch die Buchstaben a bis h dargestellt. Das mit a markierte Bit ist das zuerst gesendete:

```

h g f e d c b a
1 1 0 0 0 0 1 0

```

### 2.5.1 Ethernet Multicast Adressen

Eine Große Anzahl an Ethernet-MAC-Adressen – genauer gesagt: die Hälfte – sind Multicast-Adressen. Wenn eine Organisation ein OUI für Unicast-Adressen registriert hat, kann diese OUI auch für Multicast-Adressen verwendet werden. Die Organisation kann Multicast-Adressen für jede beliebige Anwendung definieren, die sie für geeignet hält.

Das IEEE 802 Komitee besitzt die Registratur der Unicast-OUI X'00-80-C2. Schauen Sie sich die folgenden beiden Beispiele an:

- X'01-80-C2-00-00-00 Genutzt als Gruppenadresse für Bridges mit Unterstützung für das *Spanning Tree Protocol*.
- X'01-80-C2-00-01-10 Genutzt als Gruppenadresse für FDDI-Stationen mit *Status Report Frames*.

Die *Digital Equipment Corporation* (DEC) besitzt die Unicast-OUI X'08-00-2B, so daß die Multicast-Adresse mit X'09-00-2B starten kann. DEC hat eine Menge Spezial-Multicast-Adressen definiert, darunter auch X'09-00-2B-00-00-0F als *Local Area Transport* (LAT) Adresse. LAT ist ein Terminal-Zugriffsprotokoll, das in DEC-Netzwerken eingesetzt wurde und unter Tru64 Unix (ursprünglich DEC, dann Compaq, heute HP) weit verbreitet ist.

## 2.6 Token Ring Adresskonventionen

IBM entwickelte das Token Ring Protokoll als evolutionäre Erweiterung der bisherigen Protokolle, das im Gegensatz zu Ethernet das oberste Bit des ersten Adress-Bytes zuerst überträgt (*big endian order*). IBM wollte diese Praxis nicht ablegen, andererseits wollte IBM seine Token-Ring-Spezifikation bei der IEEE einreichen, die jedoch die *little endian order* spezifiziert hat.

Der Kompromis sah vor:

- Transmission der Bytes im Adressfeld des MAC-Header in der *little endian order* (Ethernet). Transmission der Bytes im Informationsfeld des Frames in der *big endian order* (Token Ring).
- IBM ging noch einen Schritt weiter, indem die Adressen reinterpretiert werden, so daß Bits der Adress-Bytes immer in umgekehrter Reihenfolge gesendet werden. Damit wird das unterste Bit automatisch das oberste.

Abbildung 2.2 illustriert die „Order“ der Token Ring Bits. Da eine Quelladresse immer eine Individual-Adresse sein muß, entschied sich IBM das Unicast-/Multicast-Bit für andere Zwecke zu verwenden. Sie benutzten das Bit als ein Flag, daß anzeigt, ob weitere Frame-Routing-Informationen (*Routing Information Field*, RIF) in einem weiteren Feld folgen. Der untere Teil der Abbildung 2.2 zeigt die Verwendung dieses Bits.

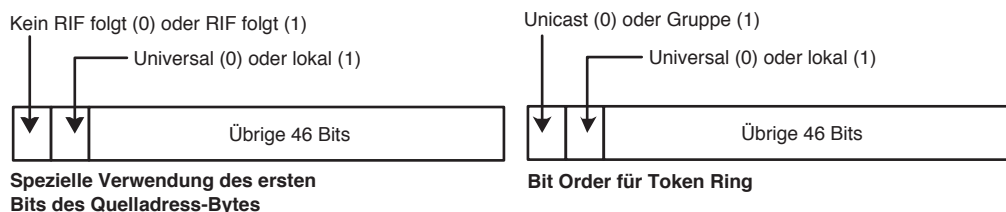


Abbildung 2.2: Initiale Bits in Token Ring Frames.

## 2.7 FDDI-Adresskonventionen

Die FDDI-Frames sind den Token Ring Frames sehr ähnlich. Auch FDDI folgt den IBM-Regeln, so daß auch hier die *big endian order* verwendet wird und die MAC-Adresse in umgekehrter Reihenfolge codiert wird. Ein fundamentaler Unterschied zu Token Ring besteht in der Verbindung von FDDI mit Token Ring und FDDI mit Ethernet, da die Adress-Bytes der MAC-Adresse umgekehrt und „richtig herum“ transportiert werden:

- Ist FDDI mit Token Ring verbunden, wird das nicht-kanonische Format verwendet.
- Ist FDDI jedoch mit Ethernet verbunden, muß unbedingt das kanonische Format verwendet werden.

Befindet sich FDDI in einer Mischung aus Ethernet und Token Ring, so können beide Formate gleichzeitig auftreten.

# Kapitel 3

## LAN-Strukturen

1980 entwickelten die Firmen DEC, Intel und Xerox die Ethernet Version 1, dessen Spezifikation im *Ethernet Blue Book* veröffentlicht wurde. Diese Spezifikation sah eine Geschwindigkeit von 10 Mbps vor, die über Coaxialkabel via Basisbandtransmission übertragen werden sollte. Eine verbesserte Version, bekannt unter dem Namen *DIX Ethernet* oder Ethernet Version 2 kam kurz danach im Jahre 1982 auf den Markt. *DIX* steht übrigens für DEC, Intel, Xerox.

Während dieser Phase etablierte die IEEE eine Expertengruppe mit dem Namen 802, die sich mit der Standardisierung und Erweiterung der Spezifikationen für sog. MANs, *Metropolitan Area Networks* beschäftigte. Der Name 802 entstand aus dem Gründungsdatum: Februar 1980 (80-2). Eine Untergruppe mit der Bezeichnung 802.3 ging aus dem Komitee hervor und befaßte sich mit der Entwicklung und Standardisierung eines DIX-Ethernet-Nachfolgers.

### 3.1 Ethernet-LAN-Architektur

Bevor wir uns einige Architekturen genauer ansehen, klären wir erst einige Begrifflichkeiten, die für das Verständnis unabdingbar sind, zumal die Ethernet-Architektur stetig gewachsen ist und ein 20-jähriges Erbe mitbringt.

#### 3.1.1 1-Segment-Ethernet mit CSMA/CD

Das einfachste lokale Netzwerk besteht aus einem einzigen Segment, das einige Desktop-Computer enthält und evtl. einen Server, der mit dem Client über Coaxialkabel verbunden ist.

Möchte ein System Daten auf dem Medium übertragen, muß es warten, bis das Medium „frei“ (*quiet*) ist und kein Frame eines anderen Systems das Medium gerade passiert. Damit sind alle anderen Stationen solange besetzt, bis der letzte Frame der sendenden Station das Medium verlassen hat. Das Medium erlaubt nur die Übertragung eines einzelnen Frames zu jedem Zeitpunkt. Möchten zwei Systeme gleichzeitig senden, kann es zu Kollisionen der Frames kommen, die dann ungültig sind und verworfen werden. Beide Systeme müssen nach einer Kollision einen zufälligen Zeitabschnitt zwischen 1 und 25 ms warten, bis sie die Transmission wiederholen dürfen. Dieses Regelwerk wird als CSMA/CD (*Carrier Sense Multiple Access/Collision Detection*) bezeichnet.

Alle Systeme lauschen ununterbrochen dem Medium und „sehen/hören“ jeden Frame der über das Medium übertragen wird. Jeder Frame hat einen Header, der Quell- und Zieladresse enthält und einen *Trailer*, der die CRC-Informationen trägt.

Abbildung 3.1 zeigt ein Desktop-System, daß versucht einen Frame über das Coaxialkabel an Server C zu übertragen. Server C liest die Header-Informationen aus und erkennt sich selbst als Empfängerstation, so daß der Frame angenommen wird. Andernfalls wird er verworfen. Alle anderen Stationen (B und D) sehen diesen Frame auch, lehnen ihn jedoch ab, weil die Zieladresse nicht mit der eigenen Adresse übereinstimmt. Die in Abbildung 3.1 dargestellte Variante wird als *Bus-Topologie* bezeichnet.

Die Abbildung macht drei Dinge deutlich:

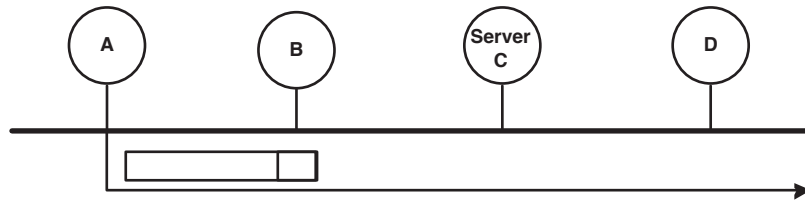


Abbildung 3.1: Übertragung eines Frames über ein Ethernet-Segment.

- Das System ist mit allen anderen Stationen über ein Coaxialkabel zu einem *Segment* verbunden.
- Es ist eine Mehrfachzugriffsumgebung (*Multiaccess Environment*), d.h. mehrere Systeme verwenden das gleiche Medium (*shared medium*) zur Übertragung aller Daten.
- Jede Zieladresse jedes Frames wird von allen Stationen analysiert. Damit ist es möglich, mit Broadcast-Adressen alle Maschinen oder mit Multicast-Adressen eine ausgewählte Gruppe von Stationen mit Informationen zu versorgen.

Daher spricht man in diesem Fall oft auch von einer „Broadcast Multiaccess“-Umgebung.

### 3.1.2 Multisegment-LANs mit Repeatern

Frühe LANs waren klein und kompakt, sie funktionierten sehr gut, solange die Zahl der Benutzer ein bestimmtes Maß nicht überschritt. Das Signal, das die Daten in Form von 0s und 1s transportiert, neigt dazu, kontinuierlich mit der Entfernung an Intensität abzubauen, mit dem Resultat, daß gegen Ende des Mediums das Signal der 1 nicht von dem der 0 unterschieden werden kann. Folglich sind fehlerhafte Frames Grund für starke Leistungseinbrüche und unterschiedlichste Fehlererscheinungen.

*Repeater* können dieses Problem beseitigen indem Sie eingehende Signale „auffrischen“, bevor Sie das Gerät wieder verlassen. Mit Repeatern kann man den Umfang des Netzwerkes leicht erweitern und bisher unerreichbare Teile in das bestehende Netzwerk integrieren. Es sei noch an dieser Stelle genannt, daß die Begriffe *Hub*, *Concentrator* und *Repeater* im Grunde dasselbe Gerät beschreiben, Hub und Concentrator werden jedoch meist dann verwendet, wenn sie mehr als zwei Segmente miteinander verbinden.

Abbildung 3.2 zeigt ein Netzwerk, bestehend aus drei Segmenten, die über einem Repeater miteinander verbunden sind. Jedes Segment wird durch einen Port an den Repeater angeschlossen. Das bedeutet, daß ein von Arbeitsplatz A abgesetzter Rahmen, der an Server X adressiert ist, von allen Arbeitsplätzen in allen Segmenten gesehen werden kann.

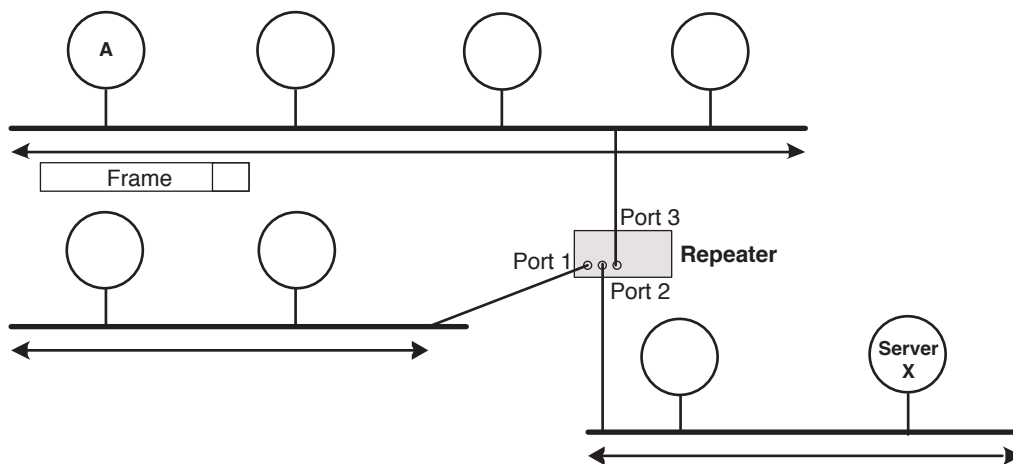


Abbildung 3.2: Drei Segmente, die über einen Repeater verbunden sind.

### 3.1.3 Twisted-Pair-Verkabelung und Hubs

Ein Coaxialkabel ist steif und schwer. Außerdem sind Coax-Netzwerke schwer zu warten, denn eine Unachtsamkeit kann unterschiedlichste Nebeneffekte, wie *Near-End-Crosstalk* (NEXT) oder Dipolverhalten, zur Folge haben.

Seit Jahren schon setzen Telefongesellschaften weltweit Twisted-Pair-Kabel ein, um die Teledienste (Telefon, Facsimile, Voice Mail, etc.) abzuwickeln. In den frühen 80ern haben sich einige Hersteller, darunter AT&T, durchgerungen, eine Twisted-Pair-LAN-Technologie zu entwickeln, die das Verlegen und Warten der Datenleitungen deutlich vereinfacht, zumal eine signifikante Materialersparnis erreicht wurde. Wer beim Anschluß einer Telfondose etwas genauer hinsieht, kann erkennen, daß das Haus voll mit Twisted-Pair-Kabeln ist. Es ist ein bereits sehr ausgereiftes Medium, dargestellt in Abbildung 3.3.

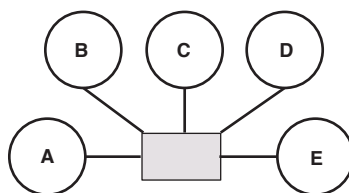


Abbildung 3.3: Sterntopologie.

Das Resultat der Bemühungen war *StarLAN*, eine 1 Mbps Twisted-Pair-LAN-Lösung. Diese Variante wurde schnell aufgegriffen, da die Verteilung der Datenkabel von einem zentralen Punkt aus Sinn macht. Die meisten Gebäude sind mit Twisted-Pair-Kabeln für die Teledienste ausgestattet, die für die LAN-Infrastruktur verwendet werden können.

Später fanden die Ingenieure heraus, wie man bis zu 1000 Mbps über ungeschirmtes Twisted-Pair -Kabel transportieren kann. Neue Verfahren verlangen nach neuen Geräten, z.B. nach dem 10 Mbps-Hub.

Der Wechsel in der Topologie schlägt sich besonders in dem Begriff *Segment* nieder. Jede Arbeitsstation ist mit dem Hub verbunden, so daß jede Strecke zwischen Hub und Arbeitsstation ein Segment bildet. Entscheidend ist, daß zwischen Einspeisung des Signals und Zielstation nur zwei Knoten existieren: der Hub und die Station selbst. Ein großer Vorteil für die Wartung und Verwaltung kleiner und großer Netzwerke.

### 3.1.4 Höhere Leistung durch den Einsatz von Bridges

Die Ethernet-Spezifikation erlaubt maximal 1.024 Stationen in einem LAN. Alle Benutzer eines LANs teilen sich die gesamte Bandbreite, was bei zunehmender Teilnehmerzahl unweigerlich zu unerträglichen Latenzzeiten führt.

Eine *Bridge* (dt. *Brücke*), auch *Switch* genannt, ist ein Layer-2-Gerät, das die Erweiterung des lokalen Netzwerk stark vereinfacht, ohne die Bandbreite weiter einzuschränken. Eine Brücke blockt alle Frame ab, die in einem Segment nicht benötigt werden, wenn keine Adressaten in diesem Segment vorhanden sind.

Wenn beispielsweise Station A in Abbildung 3.4 einen Frame an Station C senden möchte, besteht keine Notwendigkeit, diesen Frame in die Segmente 1 und 2 zu transportieren. So wird ein Frame der von Station E an Station H gesendet wird, nicht auf Segment gelangen.

Das Resultat ist, daß die Bandbreite optimaler ausgenutzt wird, da die Frames nicht unnötiger Weise das ganze Netzwerk belasten, obwohl die Zielstation auf dem gleichen Segment angesiedelt ist. Somit kann die Bandbreite in diesem 3-Segment-LAN um den Faktor 3 gesteigert werden.

### 3.1.5 Collision Domains

Die Segmente in Abbildung 3.2 und 3.3 sind durch Repeater miteinander verbunden. Jeder gesendete Frame einer Station kann von allen anderen gesehen werden (das ist der große Unterschied zu Bridges, die

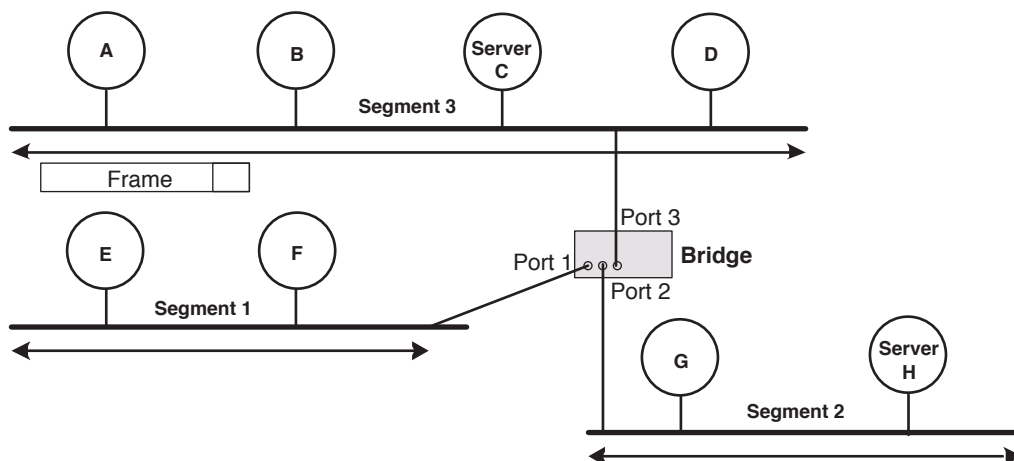


Abbildung 3.4: 3-Segment-LAN mit einer Bridge.

nur adressierte Stationen mit Rahmen beliefern), folglich führt ein gleichzeitiger Zugriff auf das Medium durch zwei Stationen zu einer Kollision, auch, wenn die Stationen unterschiedlichen Segmenten zugeordnet werden können. Daher ist ein Verbund aus Segmenten, die durch einen oder mehrere Repeater verbunden sind, eine *Collision Domain*.

**Anmerkung:** Zu jedem Zeitpunkt, kann nur eine Station in einer Collision Domain einen Frame erfolgreich absetzen. Die anderen Stationen sollten während dieser Zeit „zuhören“ (*listen*). Daher spricht man bei CSMA/CD auch von *Half-Duplex-Ethernet*.

Die Brücke in Abbildung 3.4 teilt die ursprüngliche Collision Domain in drei unabhängige Teile. Sollte ein Frame in ein anderes Segment transportiert werden, kann die Bridge den Frame puffern (*buffering*), bis das Medium im Zielsegment „still“ ist und den Transport durchführen kann.

Bridges lesen die MAC-Adressen der gesendeten Frames und merken sich mit internen Tabellen, welche Station auf welchem Segment lokalisiert sind, so daß die Bridge Frames auf MAC-Basis weiterleiten kann.

### 3.1.6 Mit Switches höhere Leistung erreichen

1993 erfand eine Firma namens Kalpana den ersten LAN-Switch. Ein Switch ist eine Multiport-Brücke, der mehrere Frames gleichzeitig weiterleiten kann, und nicht nur einen zu einem bestimmten Zeitpunkt.

Die ursprünglichen Switches entwickelten sich zu den heutige Layer-2-Switches. Jede Verbindung zwischen zwei Workstation wird gewissermaßen gebridged. Damit besteht zu jedem Zeitpunkt eine direkte *Peer-to-Peer*-Verbindung zwischen den Stationen. Die interne Architektur eines Switches wird in Abbildung 3.5 dargestellt.

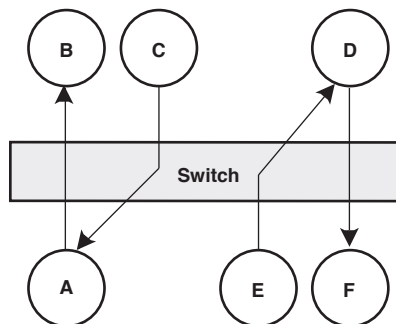


Abbildung 3.5: Layer-2-Switch.

**Anmerkung:** Der Name *Twisted-Pair-Hub* ist nur ein Marketing-Name für Repeater. Genauso ist auch die Bezeichnung *Layer-2-Switch* nur ein Marketing-Name für eine moderne Brücke.

### 3.1.7 Full-Duplex-Kommunikation mit Switches

Der 802.3-Standard erweiterte die Ethernet-Funktionalität um den Voll-Duplex-Betrieb. Die beiden Kommunikationspartner sind durch den Switch direkt miteinander verbunden, und können gleichzeitig senden und empfangen. Dadurch entfallen unnötige Wartezeiten, bis das Medium wieder frei ist.

**Anmerkung:** Voll-Duplex-Kommunikation kann zwischen allen Systemen stattfinden, die keine Repeater sind. D.h. eine Voll-Duplex-Verbindung kann zwischen zwei Hosts, zwischen zwei Switches (Bridges), einem Switch und einem Router und zwischen zwei Routern bestehen.

### 3.1.8 Router

Die Benutzer von LANs verlangten bald nach einer einfachen Möglichkeit, entfernte LANs miteinander zu verbinden oder in Zeiten des Internets über andere Netzwerke das Zielnetzwerk zu erreichen. Diesen Zweck erfüllen heute die sog. *Router*.

In Abb. 3.6 sind zwei entfernte LANs durch einen Router miteinander verbunden und tauschen Daten über eine Weitverkehrsverbindung aus. Diese könnte zu einem anderen Firmenstandort führen oder zwei universitäre Netzwerke verbinden.

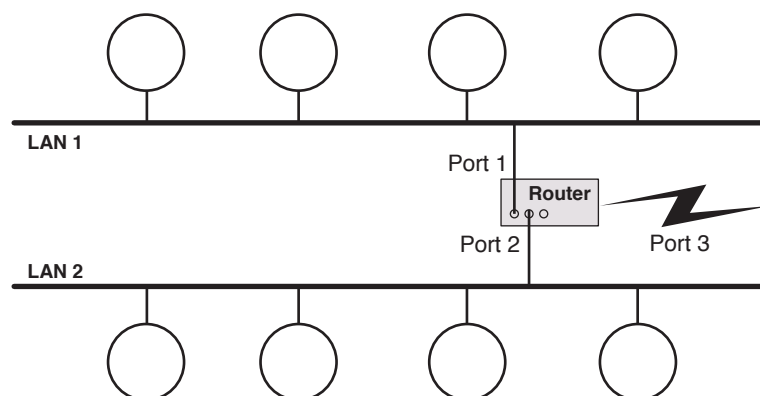


Abbildung 3.6: Verbindung zweier LANs und Bereitstellung eines WAN-Links

Ein Router ist ein Layer-3-Gerät mit vielen Features. Einige sind:

- Sie verbinden unterschiedliche Netzwerke.
- Ein Router leitet keinen Broadcast-Verkehr der LANs weiter wie eine Bridge.
- Sie können verschiedene LANs über Weitverkehrsverbindungen unterschiedlichen Typs konnektieren. Beispielsweise Dial-up, ATM, Frame Relay und Standleitungen.
- Sie können Sicherheitsaufgaben übernehmen und gezielt Teile des Verkehrs filtern.

Es ist wichtig zur Kenntnis zu nehmen, daß Frames, die das LAN verlassen sollen und den Router erreichen in ein anderes Frame-Format umgewandelt werden, indem die MAC-Header und -Trailers entfernt und diese Teile neu hinzugefügt, bevor sie weitergeleitet werden.



# Kapitel 4

## Das CSMA/CD MAC Protokoll

Das Ethernet MAC Protokoll ist sehr robust. Es wurde für nahezu alle Medien, Topologien und Geschwindigkeiten, von 1 Mbps bis 1000 Mbps, seit dem Jahre 1998 von der Arbeitsgruppe 802.3 spezifiziert.

Damit die Kommunikation über Netzwerke mit unterschiedlichen Architekturen einwandfrei arbeitet, muß ein allgemeines Regelwerk existieren, das die Rahmenbedingungen für jede Architektur definiert. Dazu zählt auch eine ausgeprägte Abwärtskompatibilität, sodaß unternehmenskritische Netzwerke mit der Zeit bestehen bleiben können, während andere Teile des Unternehmens mit neuer Technologie und anderer Topologie ausgestattet werden.

### 4.1 Klassische Ethernet-LANs

Zielstationen erkennen Frames, die für sie bestimmt sind, durch die in dem MAC-Header hinterlegte Zieladresse. Der Frame enthält neben Header und Trailer auch ein Datenfeld in dem die zu übermittelnden Daten eingeschlossen sind.

#### 4.1.1 Preamble und Interframe Gap

Jeder Frame wird durch eine klar definierte Folge von 0s und 1s, *Preamble* genannt, eingeleitet. Abbildung 4.1 illustriert die Anordnung der Preambles innerhalb der Frames. Frames müssen durch einen gewissen zeitlichen Abstand von einander getrennt übertragen werden, damit nicht zufällige Folgen von 0s und 1s eine falsche Preamble bilden kann, die zu Fehlinterpretationen führen könnten. Diese zeitliche Lücke wird als *Interframe Gap* bezeichnet (dt. *Rahmenvonraum*). Es ist selbstverständlich möglich, daß der zeitliche Abstand zwischen zwei Frames größer als das Interframe Gap ist, kann jedoch nicht kleiner sein.

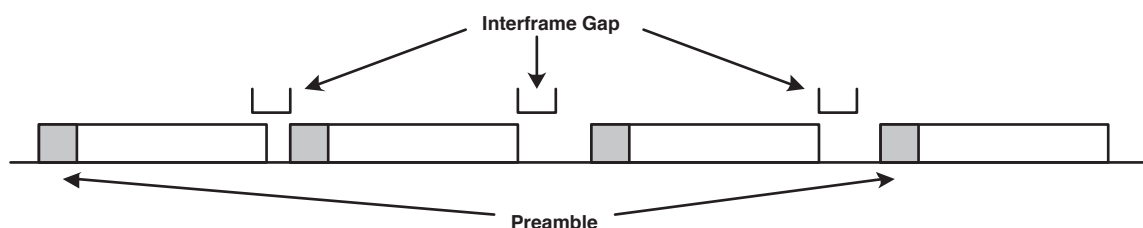


Abbildung 4.1: Frame Preambles und Interframe Gaps.

Die Lücke entspricht der Zeit, die 96 Bits benötigen, um auf dem Medium mit einer bestimmt Geschwindigkeit übertragen zu werden. Von daher unterscheiden sich die Lücken in den unterschiedlichen Geschwindigkeitsraten wie folgt:

- 1 Mbps:  $96.000 \mu s$
- 10 Mbps:  $9.600 \mu s$
- 100 Mbps:  $0.960 \mu s$
- 1000 Mbps:  $0.096 \mu s$

Die Lücke kann variieren, wenn Frames Repeater passieren, um anderen Segmenten zugeführt zu werden. Dieses Verhalten ist von zwei Faktoren abhängig:

1. Verzögerungen (*delays*) können eine Varianz während des Empfangs und Weiterleitens der Frames bewirken.
2. Ein Frame muß erst eine vollständige Preamble erkennen, bevor der Frame verarbeitet werden kann. Die Erkennungszeit kann von Frame zu Frame unterschiedlich sein.

### 4.1.2 Das CSMA/CD Protocoll

Die Arbeitsstationen teilen sich ein Medium, dessen Zugriff durch das CSMA-Protokoll (*Carrier Sense Multiple Access*) gesteuert wird. Dieses Protokoll definiert folgende Regeln:

#### Carrier Sense

Eine Station „lauscht“ dem Medium, ob eine andere Station Daten transportiert. Ist das Medium für eine vordefinierte Zeit „still“, initiiert sie den Transfer.

#### Collision Detection

Wenn zwei Station versuchen, gleichzeitig zu senden, kommt es zur Kollision. Beide Stationen warten eine zufällige Zeit, bis sie es erneut versuchen. Kollodierte Rahmen werden sofort verworfen, da sie unbrauchbar sind.

#### Jamming

Nachdem eine Kollision entdeckt wurde, muß der Sender weiterhin Bits übertragen, damit alle Teilnehmer die Kollision entdecken können. Die Anzahl an Bits, die für diesen Zweck abgesetzt werden sollen, wird als *Jam-Size* bezeichnet und beträgt 32 Bits.

#### Waiting

Eine Station, die an einer Kollision beteiligt ist, muß eine zufällige Zeit warten, bis sie die Transmission erneut aufnehmen darf.

### 4.1.3 Random Backoff

Eine Station, die eine Kollision verursacht hat, versucht den Frame erneut zu tranprotieren, der ebenfalls zu einer Kollision führen kann. Es bedarf also eines definierten Verhaltens, sodaß keine erneuten Kollision auftreten.

Die Grundidee ist, eine zufällige Zeit zu warten, bis die Retransmission wieder aufgenommen werden kann. Genauer gesagt ist diese Zeit ein vielfaches des sog. *Slot-Time*-Parameters.

Für 10 und 100 Mbps Netzwerke ist die Slot-Time 512 Bit. Das ist die Zeit, die ein Frame mit der kleinsten erlaubten Länge für den Transport über das Medium bei einer definierten Geschwindigkeit benötigt.

Nachdem die Kollision statt fand, wählt eine Zufallsroutine aus, ob sofort ein neuer Versuch unternommen, oder eine Zufallszeit gewartet werden sollte. Im Falle einer zweiten Kollision, wird zufällig ein Slot-Time-Integer zwischen 0 und 3 gewählt, und die Transmission nach dieser Slot-Time erneut aufgenommen. Sollte eine dritte Kollision auftreten, erhöht sich der Slot-Time-Integer auf 7.

Die Verdopplung des Slot-Time-Integer erfolgt bei weiteren 10 Kollisionen und bis der Slot-Time-Integer-Wert 1023 erreicht. Danach gibt die Station auf.

Mathematisch bedeutet dies folgendes:

- für den  $n$ -ten Versuch, mit  $n \leq 10$ , wähle eine Zahl  $r$  mit:  $0 \leq r \leq 2n$
- übertrage den Frame erneut nachdem eine Zeit von  $rx$  (*slot time*) verstrichen ist.

Der formale Name dieser Prozedur ist *Truncated Binary Exponential Backoff* und wird meistens als *Random Backoff* bezeichnet. Die Anzahl der Verdopplungen wird *Backoff Limit* und die der maximalen Versuche *Attempt Limit* genannt.

## 4.2 Ethernet MAC-Frames

Wie in Abbildung 4.2 dargestellt, besteht ein MAC-Frame aus einem Header, einem Informationsfeld und einem Trailer. Im Header sind Quell- und Zieladresse des Frames enthalten, während der *Trailer* (Anhang) die *Frame Check Sequence* trägt. Dieses wird benötigt, um Übertragungsfehler aufzudecken.

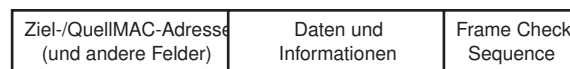


Abbildung 4.2: Allgemeines Ethernet Frameformat.

### 4.2.1 Preamble und Startframe-Muster

Eine Station beginnt mit der Transmission, indem zuerst zwei spezielle Bitmuster vorrausgeschickt werden: die Preamble und der Start-Frame. Abb. 4.3 zeigt einen MAC-Frame der mit einer Preamble und einem Start-Frame ausgestattet ist. Beachten Sie jedoch, daß die Bitmuster zwischen den unterschiedlichen Geschwindigkeiten variieren können.

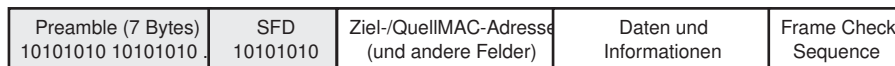


Abbildung 4.3: Preamble, SFD und MAC-Frame.

Die Preamble teilt anderen Stationen mit, daß eine Station einen Frame senden möchte und veranlaßt den *PHY-Layer* der anderen Stationen, ihre Bit-Timings mit dem der Sendestation zu synchronisieren, indem Sie das Signal aufnehmen (*Lock On Signal*).

Die Preamble besteht aus 7 Bytes:

```
10101010 10101010 10101010 10101010 10101010 10101010 101010
```

Der spezielle *Start Frame Delimiter* (SFD), der der 802.3-Preamble folgt gibt an, daß die MAC-Adresse als nächstes gesendet wird. Das SFD-Bit-Muster sieht aus wie folgt:

```
10101011
```

### 4.2.2 Ethernet MAC-Framegrößen

Bereits bei der Entwicklung der ersten Ethernet-Generation wurde die minimale und die maximale Framegröße festgelegt.

Die Minimalgröße eines MAC-Frames muß mindestens 64 Bytes betragen. Manchmal muß das Informationsfeld aufgefüllt werden, wenn die Anzahl der Informationsbits plus die der Steuerbits geringer als 512 Bits ist. Frames, die durch Fehlverhalten (Kollisionen, Error-Bursts) gekürzt worden sind, werden als *runts* bezeichnet. Die Maximalgröße eines MAC-Frames beträgt 1518 Bytes.

### 4.2.3 MAC-Frameformate

Die linke Seite der Abbildung 4.4 zeigt den ursprünglichen DIX Ethernet-Frame. Die rechte Seite zeigt das Frameformat des aktuellen 802.3-Ethernet-Frame. Die Bytes eines Ethernetframes werden von oben nach unten, und von links nach rechts übertragen. Das unterste Bit eines jeden Bytes wird zuerst übertragen (siehe *little endian order*).

Ethernet 2 (DIX)					
Ziel-MAC-Adresse (6 Bytes)	Quell-MAC-Adresse (6 Bytes)	Type (2 Bytes)	Informationen (maximal 1500 Bytes)	Padding (falls notwendig)	FCS (4 Bytes)

Ethernet 802.3					
Ziel-MAC-Adresse (6 Bytes)	Quell-MAC-Adresse (6 Bytes)	Typ / Länge (2 Bytes)	Informationen (maximal 1500 Bytes)	Padding (falls notwendig)	FCS (4 Bytes)

Abbildung 4.4: DIX- und 802.3-Frames.

Der Frame-Header besetzt 14 Bytes und der FCS-Trailer 4 Bytes. Folglich ist das Informationsfeld auf 1500 Bytes beschränkt.

#### 4.2.3.1 Quell- und Zieladressen

Die ersten beiden Felder eines Frames halten 6-Byte für die Quell- und Ziel-MAC-Adressen besetzt. Die Quelladresse ist immer eine Unicast-Adresse und die Zieladresse kann eine Multicast-, Broadcast- oder Unicast-Adresse sein.

#### 4.2.3.2 Typ oder Länge

Der DIX-Frame-Header und der ursprüngliche 802.3-Ethernet-Frame sind sehr ähnlich. Nur in der Verwendung des 2-Byte-Feldes, die der Quell-MAC-Adresse folgen, unterscheiden sie sich.

Bei DIX-Frames enthält dieses Feld die Protokolleinheit, die in dem Informationsfeld verwendet wird. Bei Frames der ursprünglichen 802.3-Ethernet-Spezifikation wird dieses Feld für die Anzeige der Größe des Informationsfeldes verwendet.

**Anmerkung:** Es ist wichtig für einen Frame, die Protokollinformationen identifizieren zu können. Wenn das dritte Feld eines Ethernet-Frames die Länge des Informationsfeldes enthält, muß der Protokolltyp-Identifizierer in einen anderen Header am Anfang des Informationsfeldes versetzt werden.

Da das 802-Komitee nicht auf die Verwendung des Protokolltyp-Identifizierers bestand, verweigerten die Benutzer die Aufgabe des DIX-Protokolls. DIX funktionierte hervorragend, jedoch war es nicht in der Lage anzuzeigen, welchem Protokoll die Daten im Informationsfeld unterliegen. DIX hatte einen kleineren Overhead und wurde letztendlich Teil des 802.3-Standards.

#### 4.2.3.3 Ethernet-Typen

Ein Ethernet-Frame darf jeden beliebigen Protokolltyp transportieren. Wenn ein System einen Frame empfängt, muß es herausfinden von welchem Protokolltyp das Informationsfeld ist, damit es zu der entsprechenden Verarbeitungseinheit weitergeleitet werden kann. Die Ethernet-Typen - genannt EtherTypes - sind in Tabelle 4.1 aufgelistet.

Eine vollständige Liste der Ethertypes können Sie unter [2] beziehen:

0x0800h (2048)	IP Version 4 Datagramm
0x0806h (2054)	ARP (Address Resolution Protocol ) Nachricht
0x0BADh (2989)	Banyan VINES
0x809Bh (32923)	Apple Talk Dateneinheit
0x80D5h (32981)	IBM SNA-Dienste über Ethernet
0x8137h (33079), 0x8138h (33080)	NetWare DataUnits
0x86DDh (34525)	IP Version 6 Datagramm
0x6003h (24579)	DECnet Phase IV Routing Information
0x6004h (24580)	DEC LAT (Local Area Transport, beispielsweise Tru64 Unix)

Tabelle 4.1: Ethernet-Typen.

Ethernet 802.3

Ziel-MAC (6 Bytes)	Quell-MAC (6 Bytes)	Typ / Länge (2 Bytes)	LLC-Header X'AA-AA-03	SNAP-Header X'00-00-00 (Ethertype)	Daten (max. 1492 Bytes)	Padding	FCS (4 Bytes)
-----------------------	------------------------	--------------------------	--------------------------	---------------------------------------	----------------------------	---------	------------------

Abbildung 4.5: Format eines 802.3-Frame mit Protokollfeld und gesetztem Ethertype-Feld.

#### 4.2.4 802.3 LLC/SNAP-Frames

Für die Identifikation des Protokolltyps sind weitere Header notwendig, die im Frame eingeschlossen sind. Abbildung 4.5 zeigt das Format eines 802.3-Ethernet-Frames mit Extra-Headern.

Nach dem Längenfeld folgt der 3-Byte-LLC-Header (*Logical Link Control*) und der 5-Byte-SNAP-Header (*Subnetwork Access Protocol*). Der Wert des LLC-Feldes ist X'AA-AA-03. Der SNAP-Header trägt den Wert X'00-00-00, gefolgt von dem Ethernet-Type-Code der eingeschlossenen Protokollinformationen. Damit wird das Informationsfeld auf 1492 Bytes begrenzt.

#### 4.2.5 Herkunft der LLC- und SNAP-Header

Die Extra-Header sind aus der Arbeit der 802.2- Arbeitsgruppe entstanden. Sie entschied sich, den Data Link Layer in 2 Sublayer aufzuteilen: Logical Link Control (LLC) und einem MAC-Sublayer, verdeutlicht in Abbildung 4.6.

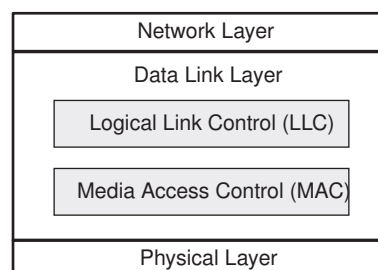


Abbildung 4.6: Sublayer des Data Link Layer.

Der LLC-Sublayer wurde für zwei Anwendungen entwickelt:

- Bereitstellung einer konsistenten Schnittstelle zwischen Network und Data Link Layer, der unabhängig von den Kommunikationsmodi (Ethernet, FDDI, Token Ring, WAN) ist.

- Bereitstellung dreier verschiedener Service-Typen:

**Type 1** Behandlung individueller PDUs (*Protocol Data Units*). Keine Extrafunktionen werden dem Data Link Layer hinzugefügt.

**Type 2** Ermöglicht die Etablierung eines zuverlässigen Links mit einem Kommunikationspartner. Die übertragenen Daten werden protokolliert und jede PDU muß bestätigt werden. Type 2 ist eng an das *X.25 LAPB* angelehnt.

**Type 3** Unterstützt einfache Senden/Empfangen-Interaktionen. Die Ankunft einer PDU erfordert die Verifizierung durch den Empfänger und ermöglicht neues Versenden fehlerhafter Frames nach einem definierten Timeout.

Type 1 ist der dominanteste Data Link-Dienst. IBM verwendet Type 2 für SNA und für den Type 3 sind mir keine Verwendungen bekannt.

*Link Service Access Points* (LSAPs): Ein Computer kann mehrere Data Link-Verbindungen aufsetzen. Die 802.2-Arbeitsgruppe führte dazu Data-Link-Adressen - Link Service Access Points genannt - ein, die dem Computer helfen sollte, die Verbindungen besser überwachen zu können. LSAPs sitzen im LLC-Header. Abbildung 4.7 zeigt das LLC-Headerformat.



Abbildung 4.7: Format des LLC-Header.

Das erste Feld des Headers enthält die Nummer des *Destination Service Access Point* (DSAP) der Zielstation. Das zweite Feld enthält die Nummer des *Source Service Access Point* der Quellstation. Das dritte Feld führt Kontrollinformationen mit. Für die Type-1-Kommunikation wird ein 1-Byte-Wert (*X'03*) eingetragen, was *unnumbered information* (unbestätigte Information) bedeutet. Für die Type-2-Kommunikation identifiziert das Kontrollfeld den Nachrichtentyp (z.B. Informationen oder Flußkontrolle). Es trägt auch die Sequenznummern, falls zuverlässige Kommunikation stattfindet (*Acknowledged Data*).



Abbildung 4.8: Format des SNAP-Header.

Werden die Felder DSAP und SSAP auf *X'AA* gesetzt, bedeutet dies, daß ein SNAP-Header folgt.

*SNAP-Header*: Die Abbildung 4.8 zeigt auf der linken Seite das Format der SNAP-Header. Die ersten 3 Bytes enthalten einen „organisatorisch einzigartigen Bezeichner“, den sog. OUI (*Organizational Unit Identifier*), der durch die IEEE vergeben wird. Der Rest des SNAP-Headers kann durch die jeweilige Organisation verwendet werden.

Die IEEE hat beispielsweise den OUI *X'00-00-00* der Firma Xerox zugewiesen. Die rechte Seite der Abbildung 4.8 zeigt einen entsprechenden SNAP-Header dessen 2-Byte-Ethertype-Code gesetzt ist.

## Kapitel 5

# Ethernet 10 Mbps PHY-Layer

Es werden nun die physikalischen Eigenschaften von Ethernet-LANs unter der Verwendung von Coaxialkabel, Twisted-Pair und Fibre Optic besprochen. Einige Netzwerke setzen nur ein Medium ein, andere sind Mischumgebungen.

Obwohl Coaxialkabel nicht mehr als Medium für neue Netzwerke eingesetzt werden, ist der Markt für diese Produkte immernoch vorhanden und die Nutzer solcher Netze sind oft zufrieden, so daß nicht immer Bedarf nach Twisted-Pair-Verkabelung besteht. Dennoch werden wir mit der Besprechung der Standards für *Thick Ethernet* (10Base5) und *Thin Ethernet* (10Base2). Die derzeit am weitesten verbreitete Verkabelungstechnik ist die Twisted-Pair-Verkabelung, die im weiteren Verlauf erläutert wird.

### 5.1 Basisband-Ethernet über 10Base5

Das ursprüngliche Ethernet-Design aus den 80ern, sah vor, daß die Segmente über ein schweres 50 Ohm Kabel aneinandergereiht werden. Ein LAN-Segment besteht entweder aus einem oder mehreren Kabeln, die durch Verbinder, genannt *Barrel Connectors*, zu einem Segment zusammengeführt werden (die Verbinder sahen früher aus wie kleine Fässer, engl. *barrel*). Abbildung 5.1 zeigt den typischen Aufbau eines 10Base5-LANs.

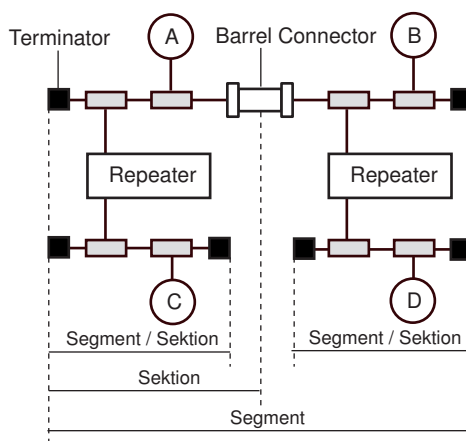


Abbildung 5.1: Verzweigtes Bus-Ethernet.

Am Ende eines jeden Segmentes muß ein spezieller Baustein, *Terminator* genannt, angebracht werden. Jeder 50 Ohm-Terminator absorbiert die Signale bevor sie am Ende des Mediums reflektiert werden. Das Coaxialkabel muß an genau einer Stelle geerdet sein. Auch dafür eignet sich ein *Terminator*.

Für Thick-Ethernet 10Base5-LANs gilt:

- 10 steht für die Geschwindigkeit von 10 Mbps.

- *Base* drückt aus, daß die Daten via Basisbandtransmission übertragen werden.
- 5 entspricht der maximalen Segment-Länge, hier 500 Meter.

Bei Basisbandtransmissionen kann immer nur ein Signal zu einem bestimmten Zeitpunkt auf dem Medium transportiert werden. Die 0s und 1s werden nach *Manchester*-Verfahren codiert.

### 5.1.1 10Base5-Coax-Verbindungen

Ethernet-Stationen sind nicht direkt mit dem Medium verbunden, sondern werden mit Hilfe zweier Mechanismen angebunden: ein Ende eines Verbinders, das *Attachment Unit Interface* (AUI) wird mit der NIC verbunden, das andere Ende, wird an einen Transceiver, *Medium Attachment Unit* (MAU) genannt, angeschlossen. MAU ist der offizielle IEEE-Begriff, der auch in vielen Texten und RFCs verwendet wird. Ein hängendes Kabel für die Anbindung der Workstations ist praktisch, da das Ethernet-Backbone-Kabel über die Fußböden oder Decken verlegt werden kann. Die Stationen werden dann über das hängende Kabel mit dem Backbone verbunden. Die Verbindung zwischen Transceiver und der AUI wird durch 15-Pin DB15-Stecker realisiert, wie Abbildung 5.2 zeigt.

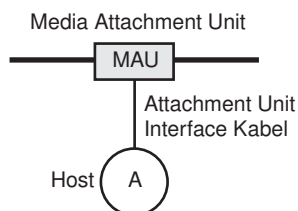


Abbildung 5.2: Eine Station ist mit dem Coax-Ethernet verbunden.

### 5.1.2 Transceiver (MAU)-Funktionen

Die Transceiver-Funktionen gehen weit über die Realisierung der Verbindung hinaus; sie senden und empfangen Bits, entdecken Kollisionen und senden Jam-Signale. Ein Transceiver übernimmt außerdem noch zwei weitere Aufgaben:

#### Jabber Control

Wenn eine Station zu lange „redet“, nennt man das *Jabber*. Ein Station redet zu lange, wenn Sie für die Übertragung eines Frames länger als die dafür vorgesehene Zeit benötigt. Der Transceiver bricht die Kommunikation ab, so daß das Medium wieder frei ist.

#### Signal Quality Error Messages

Nachdem das letzte Bit eines Frames übertragen wurde, wird eine spezielle Nachricht an die eigene Station abgesetzt, die die korrekte Funktion des Transceivers bestätigt und sicherstellt, daß die Daten sauber übermittelt wurden. Diese Funktion wird *SQE*-Test genannt.

Eine detaillierte Ansicht des AUI und des Transceivers illustriert Abbildung 5.3.

Die Transceiver-Funktionen werden über das Ethernet-Medium propagiert. Jabber Control, SQE Test und Jam Control sind für Coax, Twisted-Pair und Fibre Optic spezifiziert. Die Verbindung zwischen verschiedenen Segmenten durch AUI-Kabel wird in Abbildung 5.4 dargestellt.

### 5.1.3 10Base5 Collision Domain Parameter

Eine 10Base5 Collision Domain besteht aus verschiedenen Segmenten, die durch Repeater miteinander verbunden sind. Einige Parameter bestimmen die maximale Ausdehnung und den Aufbau des Netzwerkes. Diese sollten unbedingt in der Planung des Netzes beachtet werden.

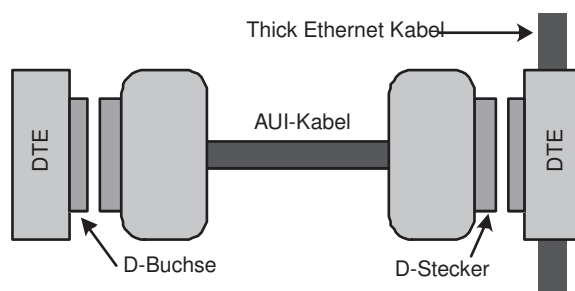


Abbildung 5.3: Thick-Ethernet-Tranceiver.

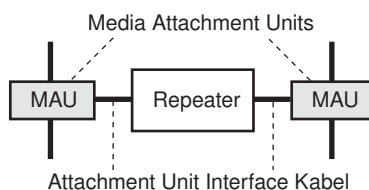


Abbildung 5.4: Segmente mit Repeatern verbinden.

Die Platzierung der Workstations innerhalb eines Segmentes ist nicht gleichgültig. Das 10Base5-Kabel, das oft einen roten oder gelben Mantel besitzt, hat alle 2,5m eine Markierung, die die optimale Anschlußstelle des Transceivers kennzeichnen.

Die letzten drei Zeilen der Tabelle 5.1 repräsentieren die berühmte 5-4-3-Regel, die besagt, daß ein Pfad durch eine Collision-Domain folgende Eigenschaften haben sollte:

- es dürfen maximal 5 Segmente durchlaufen werden.
- maximal 4 Repeater sind in einem Pfad erlaubt.
- nur 3 Segmente, die Stationen enthalten, dürfen passiert werden.

In Abbildung 5.5 wird die Regel illustriert. Das abgebildete Netzwerk enthält 7 Segmente; das ist kein Verstoß gegen die 5-4-3-Regel, denn sie gilt nur für Pfade. Aufgrund der Baumstruktur, wird ein Pfad zwischen zwei Stationen nicht mehr als 5 Segmente passieren. Z.B. ein Pfad von Station A zu Station B kreuzt 5 Segmente und passiert 4 Repeater. Nur drei dieser Segmente sind mit Stationen belegt. Ähnlich ist es mit dem Pfad von Station B zu Station C enthält 5 Segmente, nur drei davon sind mit Stationen belegt.

Parameter oder Charakteristik	Wert
Topologie	verteilte Hubs
Segmentkabel	Thick, 50-Ohm Coax
Verbinder	AUI-Kabel & Transceiver
Länge des AUI-Verbindungskabels	5-50m
Max. Segmentlänge	500m
Max. <i>Propagation Delay</i> eines Segmentes	2165ns
Max. Anzahl an Knoten pro Segment	100
Min. Distanz zwischen zwei Transceivern	2,5m
Max. Collision-Domain-Durchmesser	2500m
Max. Anzahl der Segmente im Pfad zwischen zwei Stationen	5
Max. Anzahl der Repeater im Pfad zwischen zwei Stationen	4
Max. Anzahl der Segmente im Pfad, die mit Stationen besetzt sein dürfen	3

Tabelle 5.1: 10Base5 Collision Domain Parameter.

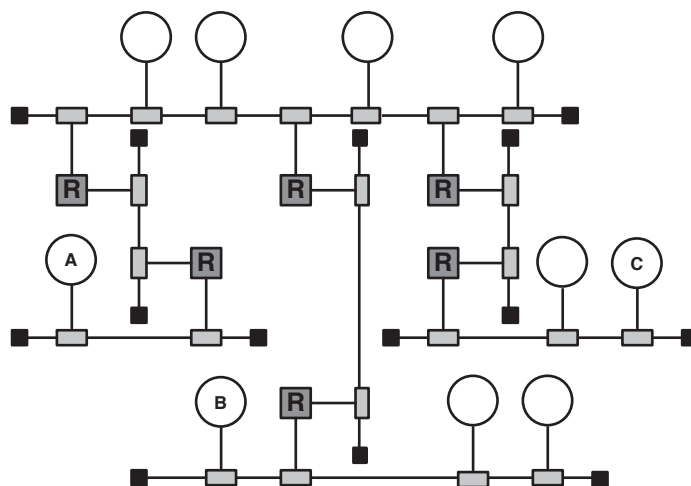


Abbildung 5.5: Coax-Ethernet-LAN erweitert durch Repeater.

### 5.1.3.1 Propagation Delay

*Propagation Delay* ist die Verzögerung (*delay*), die ein Signal zwischen Sender und Empfänger benötigt, während es über das Medium transportiert wird. Dieser Parameter ist letztendlich der begrenzende Faktor einer Collision Domain. Wenn zwei Stationen nahezu gleichzeitig senden wollen, werden die Signale kollidieren. Das Netzwerk muß klein genug sein, damit alle Stationen, die an der Kollision beteiligt sind, von dieser Notiz nehmen können. Alle empfangenden Stationen müssen ebenfalls in der Lage sein, die Kollision zu entdecken.

Der Netzwerkdurchmesser ist der längste Pfad zwischen zwei Punkten innerhalb einer Collision Domain. Das 5-Segment-Maximum limitiert den größten Durchmesser einer 10Base5-Collision-Domain auf 2.500 Meter. Bedenken Sie, daß die Zählung spätestens an einer Bridge endet, das Netzwerk mit ihnen jedoch einfach und effektiv erweitert werden kann.

### 5.1.4 Probleme mit 10Base5

Zunächst muß man bedenken, daß das Thick-Coax sehr dick, starr, schwer und damit unhandlich bei der Arbeit war. Hinzu kam der relativ hohe Preis.

Viele Firmen und Organisationen begannen mit kleinen LANs. Nachdem die Personal Computer erschwinglich wurden, wuchs die Zahl der Arbeitsplätze kontinuierlich. Meist sind die Arbeitsplätze einzeln über das Netzwerk verteilt und spätestens nach der dritten Erweiterung wußte keiner der Admins mehr, wie viele Kabel an welcher Stelle verlegt wurden, sodaß eine Erweiterung fast unmöglich wird.

Die Erweiterung des LANs war ebenfalls alles andere als trivial. Neue Transceiver mußten an einer markierten Stelle auf dem Kabel platziert werden. Während der Installation eines neuen Transceivers oder eines Barrel-Connectors ist das gesamte Segment vom Rest des Netzwerkes abgetrennt.

Eine schadhafte Stelle auf dem Medium, ein defekter Terminator, ein lockerer Barrel-Connector oder ein schlechter Transceiver kann die Kommunikation für alle Teilnehmer eines Segmentes unterbrechen. Das Problem kann auf Grund der hohen Zahl der Fehlerquellen nicht leicht aufgespürt werden.

## 5.2 Basisband-Ethernet über Thin-Coaxkabel (10Base2)

Eine neue Version des Coax-Ethernetstandards wurde entwickelt, um die o.g. Probleme von 10Base5 zu beseitigen. Diese neue Version basiert auf dem leichteren und preiswerteren Thin-Coax (auch als *thin net* oder *Cheapernet* bezeichnet). Die Bezeichnung gibt uns erneut einige Informationen über die Spezifikation des Standards. *10* steht für 10 Mbps, *Base* bedeutet Basisbandtransmission und *2* repräsentiert 200

Meter maximale Segmentlänge. 200 Meter sind der aufgerundete Wert von eigentlich 185 Meter, laut Spezifikation.

Obwohl AUI-Kabel und externe Transceiver mit 10Base2 genutzt werden können, ersetzte eine neue Art der Verkabelung diese Form. Die Transceiver-Funktionen wurden in die NICs integriert, die wiederum direkt mit dem Bussystem des Computers und dem Coaxialkabel via sog. *Bayonet Neil-Concelman* (BNC) T-Konnektoren verbunden. Diese Anschlußvariante wird in Abbildung 5.6 illustriert.

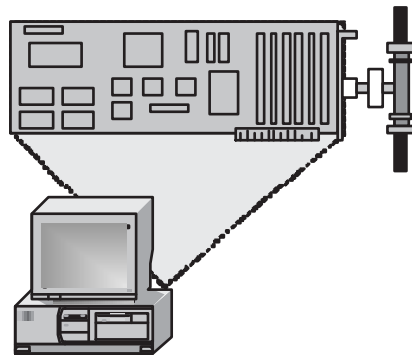


Abbildung 5.6: Ethernet NIC mit integriertem Transceiver und BNC-Verbinder.

Bitte beachten Sie, daß ein formales Interface, AUI genannt, weiterhin zwischen Transceiverchip und den anderen Komponenten einer NICs existiert, es jedoch auf die NIC montiert und wird nicht durch ein Kabel repräsentiert.

Darstellung 5.7 zeigt den Aufbau eines einfachen 10Base2-Netzwerkes. Ein 10Base2-Segment besteht aus einer Reihe kurzer, durch BNC-T-Konnektoren miteinander verbundener Kabel. Man spricht auch von einer *Daisy Chain* aus kurzen Kabeln und T-Konnektoren.

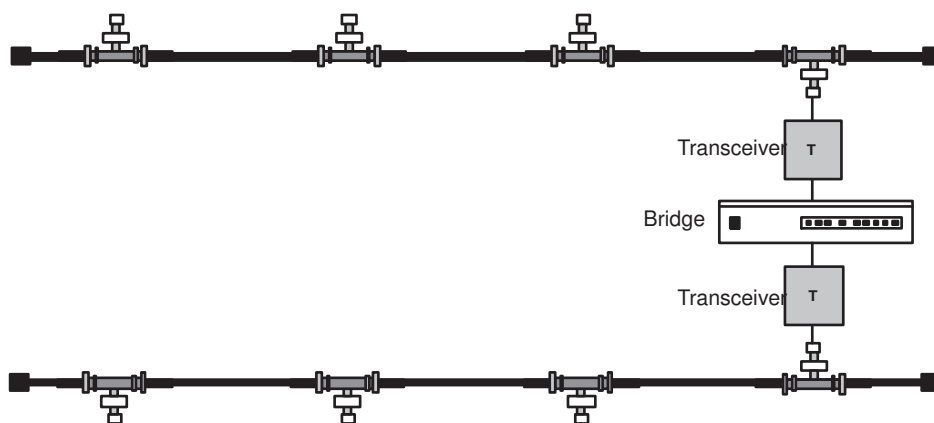


Abbildung 5.7: 10Base2-Segmente als Daisy-Chains.

Die T-Konnektoren werden direkt mit einer Steckvorrichtung an der NIC verbunden. Manche NICs sind mit D-Buchsen ausgestattet, die dann an ein AUI-Kabel angebunden werden müssen. Das in Abbildung 5.7 abgebildete Netzwerk besteht aus Coaxialkabeln, die durch T-Konnektoren miteinander verbunden sind. Ein Transceiver hat Kontakt zu dem Repeater, der wiederum Kontakt zu dem anderen Segment hat.

Der Preis, den wir für das neue, dünnere Kabel zu zahlen hatten ist die Tatsache, daß sich die Anzahl der Nodes, die ein Thin-Coax-Segment aufnehmen kann auf 30 begrenzt ist, incl. Repeater und Stationen.

### 5.2.1 10Base2 Collision Domain Parameter

Tabelle 5.2 hat die wichtigsten Parameter aufgelistet. Die 5-4-3-Regel bleibt weiterhin bestehen. Deshalb hat eine 10Base2-Collision-Domain einen maximalen Durchmesser von 925m. Die Collision Domain kann in der Größe erweitert werden, indem innerhalb einiger Segmente Thick-Ethernet-Kabel oder Fibre Optic verwendet, bzw. Repeater platziert werden.

### 5.2.2 Probleme mit 10Base2

Eine weitere Station in ein Segment hinzuzufügen ist einfach: es wird ein Kabel über ein T-Konnektor in das Segment eingesetzt und die Station angeschlossen. Nachteil ist, daß das gesamte LAN, während dieser Zeit stillgelegt ist.

Die Einfachheit kann aber auch ein Fallstrick sein. Einige User beginnen, den Arbeitsplatz umzuräumen und entfernen dafür das Netzkabel, ohne zu wissen, daß die Kommunikation unterbrochen wird.

Die Verbindungsstücke und Kabel sind relativ anfällig für kleinere und größere Schäden. Eine fehlerhafte Station kann den Netzwerkverkehr empfindlich stören. Es ist also wichtig, vor der Errichtung des Netzwerkes eine entsprechende Service-Infrastruktur bereitzustellen.

## 5.3 10Mbps Twisted Pair (10Base-T)

10Base-T wurde 1990 von der IEEE standardisiert, nachdem Wissenschaftler die 10Mbps-Kommunikation über unshielded Twisted-Pair (UTP-Kabel) entwickelt haben. Nun ist das Durcheinander von in der Decke, Wand und im Boden verlegten Kabeln vorbei. Gebäude können jetzt sternförmig verkabelt werden, ähnlich dem Verkabelungssystem für die Teledienste (Fax, Telefon, etc.). Oft können bereits existierende Telefonkabel verwendet und somit Kosten eingespart werden. Das Label für 10Mbps-TP-Ethernet ist 10Base-T. Verdrehte Adernpaare kommen zum Einsatz.

### 5.3.1 10Base-T-Segmente

Ein TP-Segment besteht naturgemäß immer aus 2 Nodes (Knoten) an jedem Ende des TP-Kabels. Die max. Segmentlänge beträgt 100m. Abbildung 5.8 zeigt diverse Segmente innerhalb eines sternförmigen Netzwerkes. Es ist auch möglich Stationen direkt zu verkabeln, dazu benötigt man ein TP-Crossover-Kabel (Kreuzkabel), da zwischen Hub/Switch immer gepacht wird.

### 5.3.2 Halb-Duplex-Betrieb

Hubs wurden entworfen, um verschiedene Collision Domains miteinander zu verbinden. Beispielsweise leitet der Hub in Abbildung 5.8 die Bits weiter und frischt damit die Signale auf. Nur eine Maschine

Parameter oder Charakteristik	Wert
Topologie	verteilter Bus
Segmentkabel	Thin, 50-Ohm Coax
Verbinder	BNC
Max. Segmentlänge	185m
Max. <i>Propagation Delay</i> eines Segmentes	950ns
Max. Anzahl an Knoten pro Segment	30
Mindestabstand zwischen zwei Nodes	0,5m
Max. Collision-Domain-Durchmesser	925m
Max. Anzahl der Segmente im Pfad zwischen zwei Stationen	5
Max. Anzahl der Repeater im Pfad zwischen zwei Stationen	4
Max. Anzahl der Segmente im Pfad, die mit Stationen besetzt sein dürfen	3

Tabelle 5.2: 10Base2 Collision Domain Parameter.

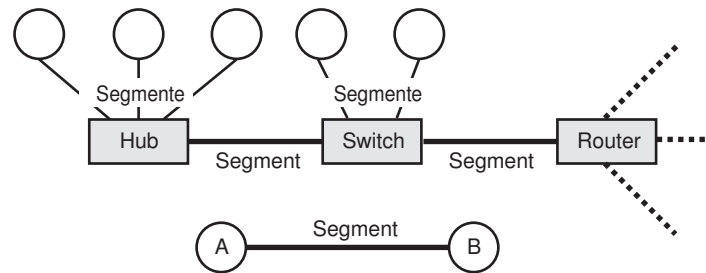


Abbildung 5.8: 10Base-T-Segmente.

kann zu einem bestimmten Zeitpunkt senden. Daher wird vom Halb-Duplex-Betrieb gesprochen. Die Kommunikation wird vom CSMA/CD-Protokoll überwacht, welches bekannterweise im Halb-Duplex-Modus arbeitet (Abbildung 5.9).

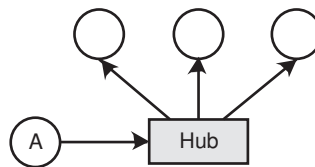


Abbildung 5.9: Halb-Duplex-Kommunikation via Hub.

### 5.3.2.1 10Base-T Collision Domain Parameter

Tabelle 5.3 listet die Parameter für 10Base-T Collision Domains auf. Eine Collision Domain wird durch die Verbindung jeder Station an einen Hub aufgebaut und durch diese ggf. erweitert. Es ist wichtig, darauf zu achten, daß das Kabel gewissen Qualitätsstufen entspricht um eine reibungslose Kommunikation zu garantieren und unsichtbare Fehlerquellen auszuschließen.

Auch die 5-4-3-Regel ist gültig, denn die maximale Segmentlänge beträgt 100m, der Durchmesser der Collision Domain darf somit max. 500m betragen (5 Segmente zu je 100m).

### 5.3.2.2 10Base-T Collision Domain Topologie

Allgemein kann die Topologie eines 10Base-T-Netzwerkes als Baum aus Sternen beschrieben werden. Jeder Stern nimmt sowohl verschiedene Stationen als auch andere Hubs auf, die allerdings einem einzigen Hub entspringen. Abbildung 5.10 illustriert diese Variante.

Parameter oder Charakteristik	Wert
Topologie	Baum, Stern
Segmentkabel	Twisted Pair, Cat. 3+
Verbinder	RJ45 (Western)
Max. Segmentlänge	180m
Max. Anzahl an Knoten pro Segment	2
Max. Collision-Domain-Durchmesser	500m
Max. Anzahl der Segmente im Pfad zwischen zwei Stationen	5
Max. Anzahl der Repeater im Pfad zwischen zwei Stationen	4
Max. Anzahl der Segmente im Pfad, die mit Stationen besetzt sein dürfen	3

Tabelle 5.3: 10Base-T Collision Domain Parameter.

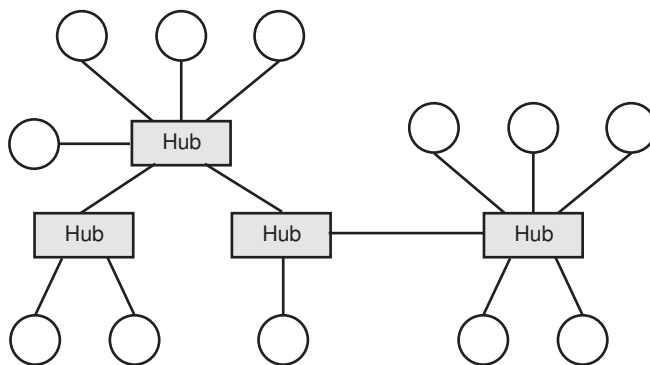


Abbildung 5.10: Ein einfaches 10Base-T-LAN.

Beachten Sie, daß die Regel, daß max 3 Segmente eines Pfades Stationen enthalten dürfen, immernoch angewendet werden kann, denn jeder Pfad besteht aus 2 Stationen. Alle anderen *Nodes* (Knoten) sind Repeater (Hubs).

Ein ununterbrochenes 100m-Kabel wird meist verwendet, um Hubs oder Bridges zu verbinden. Die Kabel zu den Stationen sind jedoch in einzelne Teile aufgeteilt, wie Abbildung 5.11 zeigt. Die Workstation wird über einen Kabelkanal oder eine Bodendose via *Patch Cord* (Patchkabel) an die interne Verkabelung und dann erneut über ein Patch Cord an den Hub im Kabelschrank (*Wiring Closet*) angeschlossen. Die Gesamtlänge dieser Konstruktion beträgt typischerweise maximal 90m.

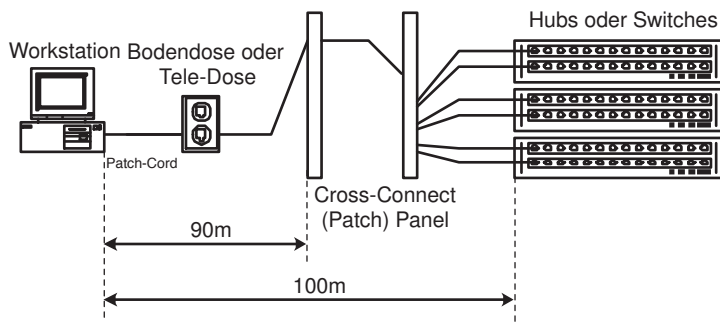


Abbildung 5.11: Komponenten eines 10Base-T-LANs.

Ein Patch-Panel im Kabelschrank (*Wiring Closet*) sorgt für die Anbindung der Arbeitsstationen an den Hub. Diese Vorrichtung ermöglicht es, Arbeitsplätze leichter zu reorganisieren und neu einzubinden, denn jede Bodendose oder Netzwerkdose ist mit einer Nummer versehen, die mit der Nummerierung am Patch-Panel korrespondiert. So genügt es, ein Kabel am Patch-Panel in den Steckplatz des neuen Arbeitsplatzes zu stecken damit dieser sofort am Netzwerk partizipieren kann. Ein weiterer Vorteil der Stern-Topologie ist offensichtlich: der Netzwerkverkehr wird nicht unterbrochen, für alle anderen Stationen ist der Vorgang völlig transparent, es sei denn natürlich, jemand kommunizierte mit dem alten Arbeitsplatz.

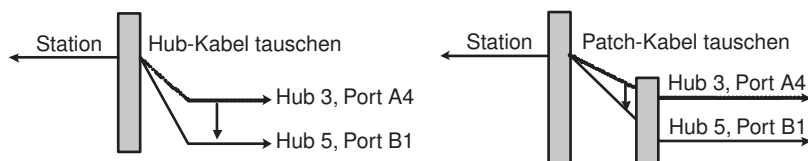


Abbildung 5.12: Interconnect und Cross-Connect.

Die Patches in Abbildung 5.12 werden als *interconnect* bezeichnet. Eine Station wird an einen neuen Hub angeschlossen, indem die Verbindung vom alten Hub in den neuen übernommen wird. Der untere Teil der Abb. 5.12 zeigt sogenannte *cross-connect* Patches. Die am Hub befindlichen Kabel werden nicht

verändert, sondern es wird über ein Patch-Feld einfach ein anderer Steckplatz gewählt. Cross-connect Patches sind daher deutlich komfortabler, können jedoch die Signalqualität negativ beeinflussen, denn es entstehen viele Verlustelemente (Konnektoren, Adapter, etc.) durch diese Bauweise.

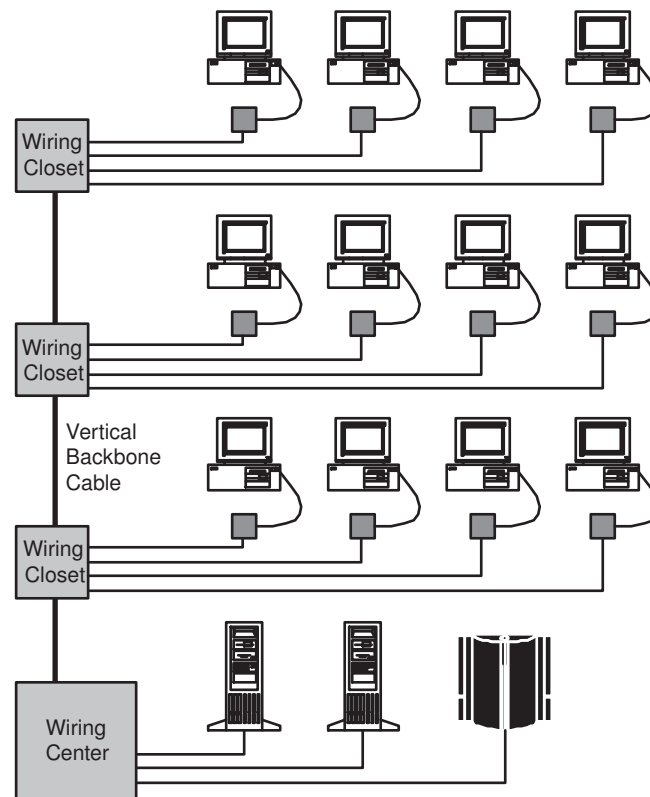


Abbildung 5.13: Verkabelung eines Gebäudes.

### 5.3.3 Voll-Duplex-Betrieb

Viele LANs im heutigen Betrieb wurden für das CSMA/CD-Protokoll und dessen 5-4-3-Regel entwickelt. Im Laufe der Zeit hat sich der Preis für Switches erheblich verringert, während Hubs langsam an Bedeutung verloren. Weil die Daten über unterschiedliche Adernpaare gesendet und empfangen werden, ist Voll-Duplex-Betrieb inzwischen ein wichtiger Faktor für die Effizienz eines Netzwerkes.

Heutige Switches funktionieren mit 10, 100 oder 1000 Mbps. Auto-Negotiation (Kapitel 10) ermöglicht den NICs die erforderlichen Parameter für die Kommunikation auszuhandeln.

Die allgemeinen Vorteile von Switches lassen sich wie folgt nennen:

- Voll-Duplex-Kommunikation für jeden Port
- Bandbreite vergrößert sich proportional zur Anzahl der Ports
- es können keine Kollisionen auftreten
- die 5-4-3-Regel ist nicht mehr gültig.

### 5.3.4 10Mbps Twisted-Pair-Kommunikation

10Base-T transportiert Daten über zwei Adernpaare: eines zum Senden, das andere zum Empfangen von Signalen. Diese Adernpaare sind miteinander verdreht (*twisted*) und minimieren den sog. *Crosstalk-Effekt*, der durch Interferenzen benachbarter Adern ausgelöst wird. Diesen Effekt konnte man in frühen

und leider noch heutigen Zeiten beim Telefonieren beobachten, wenn plötzlich ein zweites Gespräch auf einer anderen Leitung gehört werden kann.

Wenn zwei Drähte, die dieses umgedrehte Signal tragen, miteinander verdrillt werden, verringern sich die elektromagnetischen Strahlungen auf ein Minimum, da sich die Interferenzen gegenseitig fast auslöschen. Dieses Arrangement wird auch als *balanced cable* bezeichnet. Je mehr *Twists* (Verdrillungen) pro Meter auftreten, desto besser ist die Signalqualität. Die Anzahl der Verdrillungen pro Meter liegen zwischen 6 und 40. Abbildung 5.14 illustriert den Zusammenhang zwischen den Adern. Der Draht mit der Beschriftung Tx+ trägt das übertragende Signal, der andere Draht mit der Beschriftung Tx- trägt das invertierte Transmissionssignal. Der Draht Rx+ trägt das Signal, das vom entfernten Ende gesendet wurde und Rx- trägt wiederum das invertierte Signal.

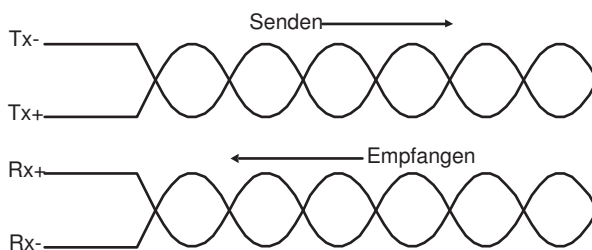


Abbildung 5.14: Senden und Empfangen über Twisted-Pair.

Ein weiterer guter Grund für die TP-Variante ist, daß Rauschen weitestgehend unterbunden werden kann. Denn wenn ein Rauschen über TP wandert, sind beide Signale eines Paares dem gleichen Rauschen ausgesetzt. Beim Eintreffen des Rauschens wird das invertierte Signal von Rx- kommend erneut invertiert. Damit wird das Rauschen ebenfalls invertiert, mit dem Ergebnis, daß das invertierte Rx- nun Rx+ hinzugefügt wird und damit das Rauschen sich selbst auslöscht.

Die zwei verdrillten Adernpaare werden durch die RJ45-Stecker an jedem Ende terminiert. Der RJ45-Stecker hat genau wie die Buchse 8 Kontakte nummeriert von 1 bis 8. Allerdings werden nur vier verwendet:

1 → Tx+ 2 → Tx- 3 → Rx+ 6 → Rx-

Jeder Draht ist mit einem Transmit-Pin und einem Receive-Pin auf der anderen Seite verbunden. Die Figur in Abbildung 5.15 zeigt dieses Arrangement. Beispielsweise ist das Transmit-Pin 1 auf der Senderseite mit dem Receive-Pin 3 auf Empfängerseite verbunden. Rechts ist eine zugehörige RJ45-Buchse abgebildet.

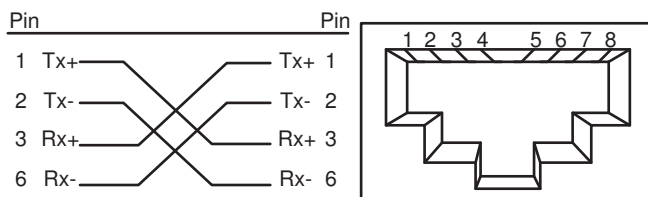


Abbildung 5.15: Pins, Verdrahtung und RJ45-Stecker.

### 5.3.5 Hub- und Switch-Verbindungen

Ein Port, der über Pin 1 und 2 überträgt und über 3 und 6 empfängt, wird als *medium dependent interface port* (MDI) bezeichnet. Alle Stationen haben MDI-Ports. Sie erwarten vielleicht, daß Sie ein Crossover-Kabel benötigen, um einen Hub mit einer Station zu verbinden. Das ist nicht der Fall. Hub-Ports sind „strukturiert“, das heißt, daß das *crossover* innerhalb des Hubs vorgenommen wird. Diese Ports sind vom Typ MDI-X (*medium dependent interface crossover*). Ein *straight-through*-Kabel wird verwendet, um einen MDI-Port eines Endsystems mit dem MDI-X-Ports eines Hubs zu verbinden.

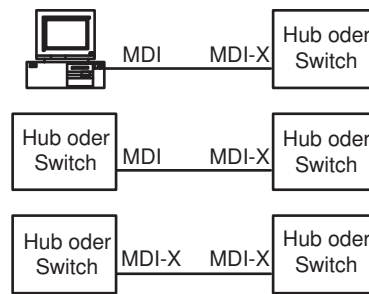


Abbildung 5.16: Port- und Kabeltypen.

Doch wie wird ein Hub mit einem anderen Hub verbunden? Ein Straight-Through-Kabel kann nicht zwischen zwei MDI-X-Ports verwendet werden, da Tx mit Tx und Rx mit Rx verbunden wäre. Die meisten Hubs verfügen daher über einen speziellen MDI-Port (*Up-Link*), der die Verwendung eines straight-through-Kabels ermöglicht, weil im anderen Hub ein MDI-X-Port das Up-Link aufnimmt. Manche Hubs können einen Port sowohl für MDI als auch für MDI-X durch setzen eines DIP-Schalters bereitstellen. So wird kein Port verschwendet. Dies gilt auch für Switches.

### 5.3.6 Twisted-Pair Link Integrity Test (LIT)

Da unterschiedliche Kabel zum senden und empfangen verwendet werden, wurde mit 10Base-T ein sinnvoller Selbsttest eingeführt. TP-Schnittstellen erkennen Fehler auf dem Medium selbständig, indem ständig eine Integritätstest durchgeführt wird. Jede Schnittstelle implementiert diese Funktion mit Hilfe sog. *integrity signals*, die alle 16 ms durch jede Ader gesendet werden, sobald das Medium „stumm“ ist.

Das Signal besteht aus einem besonderen datenlosen Puls (*normal link pulse*). Solange das Signal auf der anderen Seite empfangen werden kann, wird das Medium als betriebsbereit eingestuft. Anderenfalls wird das Medium deaktiviert, Datentransfer ist nun nicht mehr möglich.

## 5.4 10 Mbps über Fibre Optic

Fibre Optic setzt neue Maßstäbe in den Übertragungsraten und möglichen Distanzen. Desweiteren wird es nicht durch elektromagnetische Felder beeinflusst und es entstehen keine *NEXTs* (*near end cross-talks*). Die Sicherheit wird auf ein Maximum erhöht, da „Lauschangriffe“ sofort bemerkt werden und keine Signale emittiert werden, die abhörbar sind.

Die Verwendung von Fibre Optic geht auf die späten 70er Jahre zurück, als das Medium für Weitverkehrsverbindungen einsetzte. Erst in den 80er Jahren hielt es auch in Ethernet Einzug.

Die ersten Anwendungen waren Verbindungen zwischen Gebäuden, die evtl. weit voneinander entfernt sind, oder eine sichere Datentransmission erfordern. Hersteller begannen proprietäre Lösungen für diese Anwendungen zu entwickeln und bündelten ihre Bemühungen in der 802.3-Arbeitsgruppe, die den *Fibre Optic Inter-Repeater Link* (FOIRL) etablierten. FOIRL konnte für die Verbindung zweier Repeater oder eines Repeaters mit einer Station verwendet werden.

Zu dieser Zeit war die Herstellung von Glasfaser relativ aufwendig, teuer und schwer zu verarbeiten. Neue, ausgefeilte Komponenten vereinfachten die Integration von Fibre Optic, so daß nach kurzer Zeit Glasfaser das Maß aller Dinge in Verbindung mit leistungsfähigen Netzwerken wurde.

Das Set der 802.3-Spezifikation wurde um einige Standards erweitert, die sich im Wesentlichen wie folgt darstellen:

#### 10Base-FL

Eine erweiterte Version des FOIRL, die mit der vorangegangenen vollständig kompatibel ist. 10Base-FL unterstützt die Station-zu-Station-, Repeater-zu-Station- und die Repeater-zu-Repeater-Verbindungen.

**10Base-FB**

Eine Spezifikation für Repeater-zu-Repeater-Backbone-Verbindungen.

**10Base-FP**

Spezifikation eines passiven optischen Gerätes, daß mehrere Stationen miteinander verbinden kann. Diese Spez. wurde von den Herstellern nicht aufgegriffen und fand daher keine Verbreitung; wir werden nicht weiter darauf eingehen.

**5.4.1 Merkmale der Fibre Optic**

Ein Fibre Optic Kabel besteht aus zwei Fasern, eine für jede Richtung der Transmission (Abbildung 5.17). Das ausgesprochen sensible Design der 802.3-Layer erlaubte es, problemlos die Kupfertransceiver durch optische Transceiver auszutauschen, ohne ein Redesign des Protokollstapels nach sich zu ziehen.

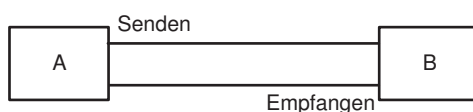


Abbildung 5.17: Senden und Empfangen.

FOIRL-, 10Base-FL- und 10Base-FB-Schnittstellen haben diverse gemeinsame Merkmale mit 10 Mbps über Coax oder Kupfer:

- es wird das *Manchester*-Encoding für die Transmission der 0s und 1s verwendet
- *Jabber control* - JC beschreibt die Fähigkeit, Transmissionen nach dem überschreiten eines Zeitlimits abzubrechen. So wird ein längerer Stau vermieden.
- die Verwendung von SQE-Signalen, um Collisionen, Staus oder andere Ereignisse bekannt zu machen.

Einige Faktoren sind jedoch einzigartig für Fibre Optic:

- Es werden spezielle Idle-Signale verwendet, die gesendet werden, wenn das Medium „stumm“ ist. Damit wird die Integrität der Datenleitung gemessen und garantiert.
- Eine sog. *low-light detection* sorgt dafür, daß Idle-Signale gesendet werden, sobald die Intensität des emittierten Lichts einen Grenzwert unterschreitet
- Für 10Base-FB besteht die Möglichkeit, ein Signal über das Medium zu senden, daß eine Fehlfunktion anzeigt.

**5.4.2 FOIRL in der Collision Domain**

Die maximale Länge eines FOIR-Links innerhalb einer Collision Domain hängt von der Topologie der Collision Domain ab:

- ein FOIR-Link in einem Pfad mit 5 Segmenten und 4 Repeatern kann bis zu 500 Meter lang sein.
- für einen Pfad mit 4 Segmenten und 3 Repeatern kann die maximale Länge 1000 Meter betragen.

Die Abbildung 5.18 zeigt eine gültige Konfiguration eines Pfades (von Station A zu Server X), der insg. 5 Segmente und 4 Repeater passiert. Er enthält außerdem 3 500 Meter FOIR-Links.

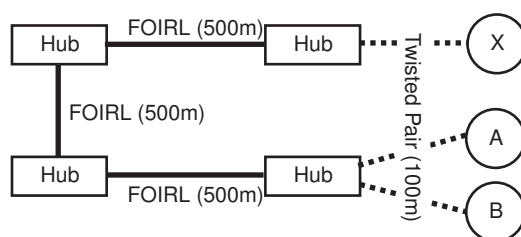


Abbildung 5.18: FOIRL-Längenrestriktionen für einen Pfad mit 5 Segmenten.

### 5.4.3 10Base-FL

10Base-FL ist im Grunde eine Erweiterung des FOIRL mit folgenden Eigenschaften:

- es basiert auf weitaus aktuelleren Komponenten und Verbindungen, die sowohl für Repeater-zu-Repeater als auch für Station-zu-Repeater geeignet sind
- es werden maximale Längen bis zu 2000 Meter unterstützt
- 10Base-FL ist abwärtskompatibel und kann somit problemlos mit FOIRL-Schnittstellen über eine Länge von 1000 Meter zwischen zwei Repeatern kommunizieren.

Ein 10Base-FL-Link zwischen zwei DTEs kann im Voll-Duplex-Modus operieren. Dennoch wird Auto-Negotiation nicht unterstützt, und beide Stationen müssen manuell für die Voll-Duplex-Kommunikation konfiguriert werden.

- Ein 10Base-FL-Link kann bis zu 2000 Meter lang sein, aber die Distanz muß kleiner in einer Collision Domain sein, wenn der längste Pfad 3 oder 4 Repeater enthält.
- Ein 10Base-FL-Repeater-zu-Repeater-Link über einen Pfad mit 5 Segmenten und 4 Repeatern kann maximal 500 Meter betragen.
- Ein 10Base-FL-Repeater-zu-Repeater-Link über einen Pfad mit 4 Segmenten und 3 Repeatern kann bis zu 1000 Meter betragen. Ein 10Base-FL-Repeater-zu-Station-Link über einen Pfad mit 4 Segmenten und 3 Repeatern kann maximal 400 Meter betragen.

### 5.4.4 10Base-FB

Das *B* im Namen der Spezifikation läßt den Schluß zu, es handle sich um die Definition eines Backbone-Interfaces. Das ist grundsätzlich richtig, wird allerdings nur zur Verbindung zweier Repeater verwendet und eignet sich daher für diesen Anwendungsbereich. Ein 10Base-FB-Link kann 2000 Meter betragen, aber innerhalb einer Collision Domain verringert sich die Maximallänge eines Pfades, abhängig von der Zahl der Repeater:

- Ein 10Base-FB-Link über einen Pfad mit 5 Segmenten und 4 Repeatern kann maximal 500 Meter in der Länge betragen.
- Ein 10Base-FB-Link über einen Pfad mit 4 Segmenten und 3 Repeatern kann maximal 1000 Meter lang sein.

### 5.4.5 Struktur von Fibre Optic

Abb. 5.19 zeigt eine Faser im Querschnitt. Erkennbar ist der Kern, bestehend aus Glas oder Kunststoff, die Reflektionsschicht (*cladding*), die einen anderen Brechungsindex besitzt und somit das Licht innerhalb des Faserkerns hält. Eine zusätzliche Schutzschicht (*buffer*) verhindert die Einwirkung von Umwelteinflüssen.

Zur Zeit existieren zwei geläufige optische Transmissionsmechanismen:

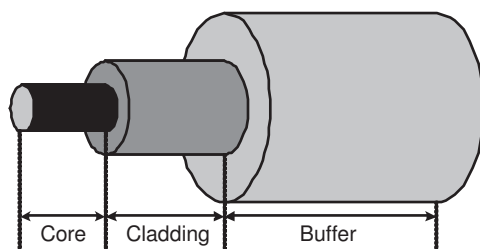


Abbildung 5.19: Struktur eines optischen Datenträgers.

### Single Mode

Ein Laser emittiert einen fokussierten Lichtstrahl durch den etwa  $8\text{-}10\ \mu\text{m}$  schmalen Lichtkern. Eine Faser ist in der Lage, enorme Entfernungen von bis zu mehreren hundert km zu überbrücken. Single-Mode wird auch in Gigabit-LANs eingesetzt.

### Multimode

Eine lichtemittierende Diode (LED) emittiert mehrere Lichtstrahlen mit unterschiedlichen Reflektionswinkeln in das Medium mit einem etwas größeren Kern, ca.  $62\ \mu\text{m}$ . Diese Lichtstrahlen werden Modes genannt. Ein Multimode-System kann Entfernungen von bis zu 2 km überbrücken. Laser sind relativ teuer, LEDs hingegen recht preiswert, was sie für die Verwendung von 10 Mbps attraktiv macht.

Es wurde noch eine weitere Übertragungsform entwickelt, die als VCSEL (*Vertical Cavity Surface Emitting Laser*) bezeichnet wird. Es kann ebenfalls für die Übertragung über Multimode verwendet werden. VCSEL ist relativ preiswert und wird in Gigabit-Ethernet eingesetzt.

Optische Übertragungsmedien werden mit zwei durch einen Slash getrennten Parametern charakterisiert. Die erste Zahl steht für den Durchmesser der Faserkerns, die zweite Zahl beschreibt den Durchmesser der Reflektionsschicht. Folgende Typen sind für die 10Mbps-Transmission vorgesehen: Multimode  $62.5/125\ \mu\text{m}$  wird für 10Mbps empfohlen und ist sehr populär. Teilweise wird auch auf  $50/125\ \mu\text{m}$  und  $100/140\ \mu\text{m}$  sowie  $85/125\ \mu\text{m}$  zurückgegriffen.

#### 5.4.6 Multimode-Transmission

Eine LED emittiert Licht in den Faserkern. Diese Lichtstrahlen werden während der Wanderung durch das Medium von der Reflektionsschicht reflektiert und erreichen durch unterschiedlich Reflektionswinkel das Ziel zu unterschiedlichen Zeitpunkten. Sollte die Dispersion einen Grenzwert überschreiten, ist das Signal beschädigt und kann auf Empfängerseite nicht mehr korrekt interpretiert werden.

Abbildung 5.20 illustriert die Dispersion des Lichts innerhalb einer Multimode-Faser. Der obere Teil zeigt ein älteres Verfahren namens *Multimode Step Index Fibre* (SIF), während der untere Teil das neue Verfahren mit der Bezeichnung *Multimode Graded Index Fibre* (GIF) zeigt. Bei SIF ist der Brechungsindex auf der gesamten Strecke der Transmission konstant. Der Index bei GIF schwankt hingegen während der Transmission. Das Resultat dieses Verfahrens ist, daß die Daten schneller transportiert werden, weil sich die Photonen an den Kerngrenzen schneller bewegen. Somit können die Photonen, die einen längeren Weg zurücklegen, den Abstand „aufholen“.

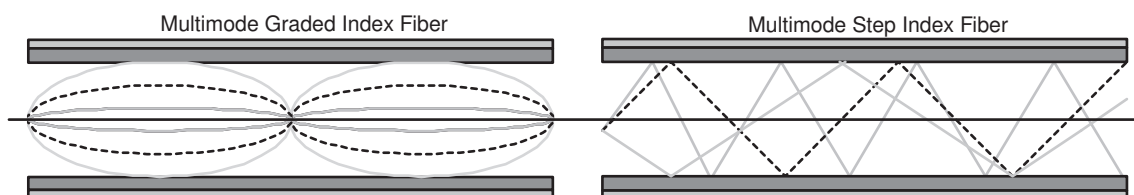


Abbildung 5.20: Übertragungsmodi SIF und GIF.

Die Wellenlänge des Lichts variiert in Abhängigkeit vom Abstand des Kernrandes. Die folgenden Wellenlängen sind für den Einsatz empfohlen:

- 800-950 *nm* für 10Base-FB und
- 10Base-FL 790-860 *nm* für FOIRL.



# Kapitel 6

## Ethernet 100 Mbps PHY-Layer

Die 802.3-Arbeitsgruppe befaßte sich erstmals 1990 mit 100 Mbps Ethernet. Verschiedener Hersteller bündelten ihre Energien in der *Fast Ethernet Alliance* und warteten im Jahre 1993 mit der ersten 100 Mbps-Lösung auf. Die Spezifikation ist vollständig abwärtskompatibel zu 10Base-T und 10Base-F. Die IEEE 802.3-Arbeitsgruppe veröffentlichte diese Spezifikation zunächst als 802.3u und wurde 1993 in den 802.3-Standard übernommen.

Das Überraschende ist, daß die wesentlichen Eigenschaften erhalten blieben. So ist die Frame-Größe konstant geblieben und die Regeln zur Kollisionsbehandlung sind unverändert. Die Unterschiede sind im PHY-Layer besonders deutlich: der Durchmesser der Collision Domain ist aufgrund der hohen Übertragungsgeschwindigkeit signifikant gesunken.

### 6.1 Einführung

Fast Ethernet funktioniert auf der Basis von Twisted-Pair Kupferkabel und Fibre Optic. Tabelle 6.1 zeigt eine Übersicht der verfügbaren Twisted-Pair und Fibre-Optic-Technologien.

Die ersten beiden Kupfertechnologien, 100Base-TX und 100Base-T4 und die optische Variante 100Base-FX wurden in verschiedenen Produkten implementiert. 10Base-T2 hingegen hatte keinen Markt und fand somit keine Verbreitung. Der PHY-Layer der 100Base-FX-Spezifikation basiert auf dem der 100Base-TX-Variante. Es wurde unter dem Namen FDDI (*Fibre Distributed Data Interface*) in LANs integriert. 100Base-TX ist die Kupferversion, genannt CDDI (*Copper Distributed Data Interface*). Das X3T9.5-Komitee der ANSI definierte CDDI für die Integration von Twisted-Pair in FDDI.

### 6.2 100 Mbps Ethernet über Twisted-Pair

Während der Ausarbeitung des Fast Ethernet-Standards arbeiteten 3 Teams an unterschiedlichen Spezifikationen:

#### 100Base-TX

Diese Ethernet-Variante ist am populärsten. Es funktioniert mit 2 TP-Kabeln, entweder Cat. 5 oder IBMs Type 1 TP.

Technologie	Kabeltyp	Länge in m	Operationsmodus
100Base-TX	2-Paar-UTP, Cat. 5 oder 2x IBM Typ 1 STP	100	Halb-/Vollduplex
100Base-T4	4-Paar-UTP/STP, Cat 3+	100	Halbduplex
100Base-T2	2-Paar-UTP/STP, Cat 3+	100	Halb-/Vollduplex
100Base-FX	2 Multimode Fibre Optic 412/2000	412/2000	Halbduplex/Vollduplex

Tabelle 6.1: Fast Ethernet Technologien.

**100Base-T4**

Da hier Cat.-3-Kabel eingesetzt werden können, ist auch dieser Standard sehr weit verbreitet. Er eignet sich für relativ einfache Installationen oder dort, wo keine Cat.-5-Verkabelung vorgenommen werden kann. Negativ jedoch ist, daß es keinen Vollduplex-Betrieb unterstützt.

**100Base-T2**

Das Ziel dieser Spezifikation war die Wahrung der Abwärtskompatibilität zu 10Base-T, also 100 Mbps über 2 TP-Kabel Cat. 3+. Aufgrund des harten Marktkampfes zwischen den verschiedenen 100Base-TX und 100Base-T4-Herstellern sanken die Preise schnell. Durch den hohen Entwicklungsaufwand konnte 100Base-2 nicht mit aktuellen Produkten mithalten, sodaß letztendlich zwar die Marktreife erreicht wurde, jedoch keine Abnehmer für diese Technologie existierten. Der PHY-Layer des Gigabit Ethernet wiederum basiert auf dieser Spezifikation.

Alle o.g. Versionen operieren mit maximalen Kabellängen von 100m.

**6.2.1 100Base-TX**

Der 100Base-TX-PHY-Layer hat viel mit dem von 10Base-T gemein:

- Er arbeitet mit 2 TP-Kabeln
- Segmentlänge von max. 100m unterstützt
- TP-Kabel werden durch RJ-45-Stecker terminiert
- Vollduplexkommunikation kann bei Bedarf aktiviert werden.

Die Unterschiede zwischen 100Base-TX und 100Base-T4 sind folgende:

- 100Base-TX erfordert Cat. 5 UTP oder IBMs Type 1 STP
- Ein anderer Kodierungsmechanismus wird verwendet
- Link-Pulse werden während der Initialisierung übertragen. Sie enthalten Autokonfigurationsinformationen. Im normalen Betrieb wird ein konstanter Strom spezieller Idle-Signale (*Idle Symbols*) zwischen den einzelnen Frames gesendet.

**6.2.2 Verwendung von CDDI**

Hersteller waren schnell in der Lage 100Base-TX-Produkte zu entwickeln da sie auf dem PHY-Layer-Standard CDDI basierten. Somit konnten bereits existierende Komponenten eingesetzt werden, was sich äußerst vorteilhaft auf den Wettbewerb auswirkte.

Für CDDI wird das 4B/5B-Schema zur Kodierung der Signale verwendet. Das gleiche Kodierungsschema wird auch für 100Base-TX verwendet. Es wandelt 4 Bits (Nibbles) in 5-Bit-Muster aus 0s und 1s um, bevor sie über das Medium transportiert werden. Der Vorteil ist, daß nun weitere 5 Bit-Muster zur Verfügung stehen, die als spezielle Kontrollcodes verwendet. Drei wichtige Bit-Muster sind:

**Idle (11111)**

Das Idle-Symbol wird kontinuierlich zwischen den Rahmen gesendet.

**Start-of-Stream-Delimiter (11000 10001)**

Ein neuer Frame wird mit diesem Muster eingeleitet.

**End-of-Stream-Delimiter (01101 00111)**

Das Ende eines Frames wird mit diesem Muster gekennzeichnet.

Die Verwendung dieser Kontrollmuster macht einige signifikante Unterschiede zwischen 100Base-TX und 10Base-T deutlich:

100Base benötigt keinen LIT. Die Idle-Symbole reichen aus. Das erste Byte einer Preamble ist der Start-of-Frame-Delimiter (SFD). Ein End-of-Stream-Delimiter (EOS) wird am Ende des Frames transportiert.

### 6.2.3 100Base-T4

Dies ist die einzige Spezifikation, die mit 4 Adernpaaren über Cat. 3+ arbeiten kann. Es existieren noch einige Verkabelungen mit Cat. 3, die meisten Netzwerke wurden inzwischen jedoch auf Cat. 5 umgerüstet. Folgende Regeln gelten für 100Base-T4:

- Ein Segment darf eine Maximallänge von 100m nicht überschreiten.
- Ein RJ-45-Stecker terminiert das Twisted-Pair-Kabel.

Die Unterschiede zwischen 100Base-T4 und 100Base-TX sind:

- Es werden vier Twisted-Pair-Adern verwendet.
- Vollduplexkommunikation wird nicht unterstützt.
- Die Bits werden vor der Transmission nach dem 8B/6T-Schema kodiert.
- Der *line integrity test* wird während „stummer“ Phasen übertragen.

Die Verwendung von vier Adernpaaren hat keinen Einfluss auf die Steckerform, auch hier wird mit RJ-45 terminiert, es sind allerdings alle acht Kontakte belegt. Die gesamte Bandbreite von 100 Mbps wird gleichmäßig auf drei Adern je Transmissionsrichtung verteilt. Drei Adern senden (Tx) und drei Adern empfangen (Rx) Signale. Abbildung 6.1 zeigt das Pinlayout des 100Base-T4-Standards.

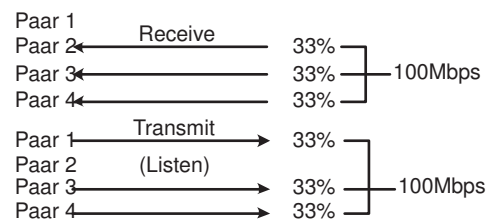


Abbildung 6.1: 100Base-T4-Transmission.

Paar zwei wird immer als Empfangspaar arbeiten und hat Kontrollfunktion. Ein System „lauscht“ auf Paar zwei, um herauszufinden, ob eine Station senden möchte. Wenn ein lokales System nicht sendet und die Preamble-Bits auf Paar zwei eintreffen, werden Paar drei und vier als zusätzliche Empfangspaare für eintreffende Daten verwendet. Wenn das lokale System Daten übertragen möchte, lauscht es auf Paar zwei, um sicher zu stellen, daß das Medium frei ist. Die Station überträgt nun auf Sendepaare eins, drei und vier. Es wird weiterhin auf Paar zwei gelauscht, falls eine Kollision eintritt. Sobald die Preamble-Bits auf Paar zwei eintreffen, während das System sendet, wird eine Kollision eintreten, so daß der Jabber-Control-Mechanismus Jam-Bits über die anderen Paare sendet.

Die Paare eins und zwei übernehmen eine spezielle Aufgabe. Paar eins kann nur senden und Paar zwei nur empfangen. Durch das „Lauschen“ auf Paar zwei, kann das System feststellen, ob das Medium frei ist, so daß sie Übertragung sofort beginnen kann. Sobald die Transmission begonnen hat, vereinen sich die Paare eins, drei und vier zu einem sendenden Medium. In Empfangsrichtung übernehmen dann die Paare zwei, drei und vier diese Aufgabe.

Nun ist auch klar, warum keine Vollduplexkommunikation möglich ist. Die Paare drei und vier können zwar auf Senden und Empfangen eingestellt werden, jedoch besteht keine Möglichkeit beide gleichzeitig senden und empfangen zu lassen.

Genau wie bei 10Base-T müssen die Kabel gekreuzt werden damit Tx auf Rx und Rx auf Tx enden. Das Crossover wird innerhalb des Switches/Hubs vorgenommen, manchmal auch durch ein Crossover-Kabel. Abb. 6.2 zeigt die Kreuzung der TP-Adern. Beachten sie, daß die Pinnummern nicht in sequenziell aufgezeigt sind, diese Variante der Darstellung ermöglicht eine übersichtliche und logische Darstellung.

Spezielle Start-of-Stream- und End-of-Stream-Delimiter werden über die Paare eins, drei und vier als Kennzeichen für den Beginn oder das Ende eines Rahmens (*Frame*). Idle-Signale (manchmal auch *Symbol*

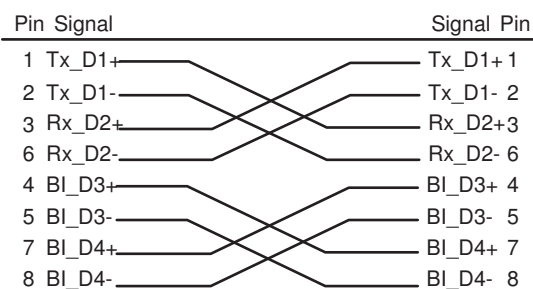


Abbildung 6.2: 100Base-T4 Cross-Connection.

genannt) werden nicht eingesetzt. Während einer Ruhephase senden die Schnittstellen LIT-Pulse über Tx\_D1 aus, die auf Rx\_D2 empfangen werden. Wenn die Schnittstellen das erste Mal initialisiert werden, senden die gleichen Paare (Tx\_D1

→

Rx\_D2) spezielle *Auto-Negotiation-Pulse* aus.

#### 6.2.4 100Base-T2

Durch die geringe Verbreitung, de facto wurde es nie implementiert, erwähne ich den Standard hier nur der Vollständigkeit halber.

100Base-T2 unterscheidet sich von 100Base-TX durch die Fähigkeit Cat. 3 UTP verwenden zu können und das durch komplexe Kodierungsschema. Die Implementierung war sehr aufwendig, teuer und erforderte spezielle elektronische Bauteile.

Dennoch hat 100Base-T einige Gemeinsamkeiten mit 100Base-TX:

- Ein Segment hat eine maximale Länge von 100 m.
- Cat 3+ UTP/STP werden durch RJ-45 terminiert.
- Daten werden über eine Ader empfangen und eine andere gesendet. 1 und 3 sind Tx, 3 und 6 sind Rx.
- Vollduplexkommunikation kann zwischen zwei Stationen etabliert werden.

### 6.3 100Base-FX und FDDI

Kurz nach der Einführung der 100 Mbps über Kupfer erblickte auch eine Fibre-Optic-Variante das Licht der Welt, die sich den PHY-Layer der Kupferversion zu Nutze machte. Das ANSI X3T9.5-Komitee entwickelte folgende Charakteristika für 100Base-FX:

- Es arbeitet mit 2 Adern zum Senden und Empfangen. Ein Durchmesser von  $62.5/125 \mu m$  ist innerhalb der 802.3-Spezifikation erlaubt.
- Eine Segmentlänge von 412 m bei Halbduplexkommunikation ist erlaubt. 2000 m sind bei Vollduplexkommunikation möglich.
- Das Fibre-Optic-Fasern werden durch SC- oder ST-Verbinder terminiert.
- Vollduplexkommunikation kann zwischen zwei DTEs etabliert werden.

Für 100Base-FX gilt das gleiche Kodierungsschema, das 100Base-TX einsetzt: 4B/5B

**Idle (11111)**

Das Idle-Symbol wird kontinuierlich zwischen den Rahmen gesendet.

**Start-of-Stream-Delimiter (11000 10001)**

Ein neuer Frame wird mit diesem Muster eingeleitet.

**End-of-Stream-Delimiter (01101 00111)**

Das Ende eines Frames wird mit diesem Muster gekennzeichnet.

## 6.4 100Mbps Collision Domain Diameter

Der kleinste Ethernet-Rahmen ist 64 Bytes (512 Bits) groß. Um sicher zu stellen, daß jede Station innerhalb einer Collision Domain eine Kollision erkennen kann, muß die *Round-Trip-Time* (Umlaufzeit eines Rahmens, RTT) zwischen den am weitesten entfernten Stationen kleiner als 512 Bit-Zeiten sein. Bei 100Mbps liegt die RTT von 512 Bit-Zeiten bei 0.00000512 Sekunden.

Die Lichtgeschwindigkeit ( $c$ ) liegt bei 299.792.458  $m/s$ . Die Geschwindigkeit der Elektronen im TP-Kabel liegt bei etwa 0,6  $c$ -0,9  $c$ . Durch diverse Einflüsse wie z.B. Umwandlungen, Zwischenspeichern und Auswerten, kann die Zeit von 512 Bit-Zeiten leicht überschritten werden. Daher muß der Durchmesser einer Collision Domain relativ klein sein.

### 6.4.1 Repeater-Klassen

Die Zeit, die ein Router benötigt, um eine bestimmte Anzahl von Rahmen zu verarbeiten, ist bei 100Mbps vergleichenermaßen groß. Der Einfluß dieser Geräte ist also nicht ohne Bedeutung. Für die verschiedenen Ansprüche an Verarbeitungszeiten wurden zwei Klassifikationen für Hubs entwickelt:

**Klasse 1** Unterstützt einen Mixbetrieb aus 100Base-T4 und 100Base-X (TX/FX). Da allerdings die Kodierungsschemata für 100Base-T4 und 100Base-X sehr verschieden sind, muß jedes eintreffende Signal auf den Typ geprüft werden und danach auf das Ausgangsmuster umkodiert werden. Das hat zur Folge, daß nur ein Klasse-1-Router innerhalb einer 100Mbps Collision Domain verwendet werden kann.

**Klasse 2** Alle Hub-Schnittstellen sind vom gleichen Typ (entweder 100Base-X oder 100Base-T4). Daher können alle Signale sofort weitergeleitet werden, ohne zusätzlichen Zeitverlust. Die Verarbeitungszeit der Klasse-2-Hubs ist schnell genug, um zwei Hubs diesen Typs in einer 100Mbps Collision Domain einzusetzen.

### 6.4.2 Collision-Domain-Konfiguration

Die Abb. 6.3 zeigt eine einfache Konfiguration einer 100Mbps Collision Domain. Wir beschränken uns hier auf die Abbildung von DTEs. Für Cat. 5 kann die Länge eines Segments max. 100  $m$  betragen. Für Fibre Optic sind die Längen auf 412  $m$  (Halbduplex) und 2000  $m$  (Vollduplex) begrenzt.

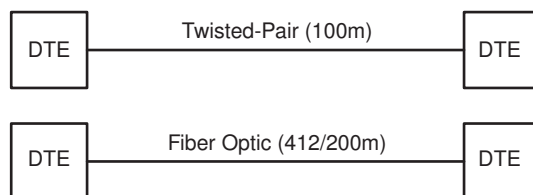


Abbildung 6.3: DTE-zu-DTE-Verbindungen.

Abbildung 6.4 zeigt die möglichen Längen beim Einsatz unterschiedlicher Medienkombinationen. Beachten Sie, daß die Distanzen vom verwendeten Medium abhängen.

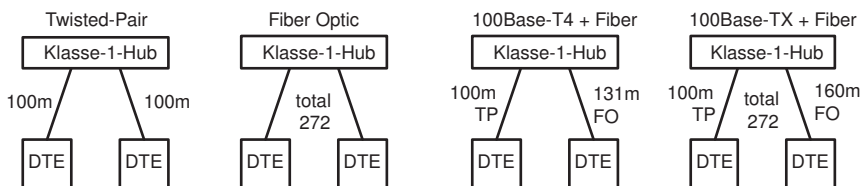


Abbildung 6.4: Entfernungen für einen Klasse-1-Hub.

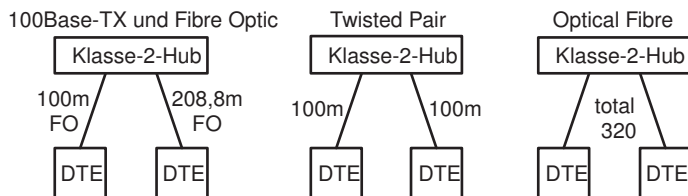


Abbildung 6.5: Entfernungen für einen Klasse-2-Hub

Abbildung 6.5 illustriert die maximalen Distanzen zwischen DTEs für Pfade, die einen Klasse-2-Hub durchlaufen. Die Kabellängen sind größer, denn es wird weniger Verarbeitungszeit benötigt.

In Abbildung 6.6 sind die maximalen Distanzen zwischen DTEs mit Pfaden, die zwei Klasse-2-Hubs durchlaufen. Der Collision Domain Diameter ist durch die Präsenz des zweiten Hubs kleiner geworden. Es macht Sinn einen Hub-Stapel zu verwenden, der sich wie einen einzigen Hub verhält.

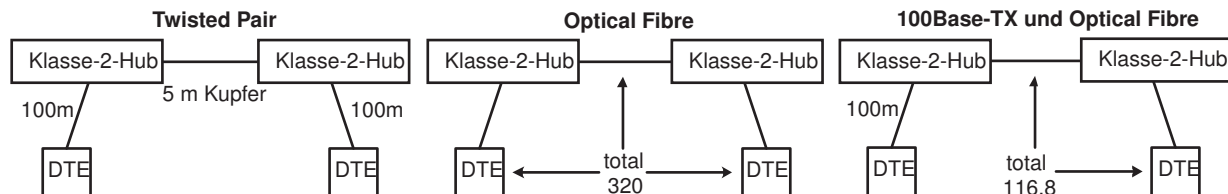


Abbildung 6.6: Entfernungen für zwei Klasse-2-Hubs

### 6.4.3 Vermeiden des Diameter-Problems mit Switches

Switches erlauben es, viele Collision Domains zusammenzufassen. Das in Abbildung 6.6 dargestellte LAN verbindet alle Arbeitstationen über einen Hub mit 60m TP. Ein 250 m Fibre-Optic-Link verbindet einen Server, der weit entfernt positioniert ist. Die RTTs für jede Komponente sind in der Figur aufgezeigt. Die zwei DTEs verwenden 100 Bit-Zeiten, der Klasse-2-Hub benötigt 92 Bit-Times, die Twisted-Pair-Verkabelung beansprucht 66.72 Bit-Zeiten und die Fibre-Optic-Verkabelung benötigt 205 Bit-Zeiten. Es ergibt sich eine Gesamtdauer von 508.72 Bit-Zeiten, was innerhalb des zulässigen Maximums von 512 liegt.

## Kapitel 7

# Gigabit Ethernet-Architektur

Im Jahre 1995 formten einige Hersteller das *Gigabit Ethernet Consortium* um den 1000Mbps-Standard auszuarbeiten. Innerhalb von 13 Monaten wurden die Spezifikationen für Optical Fibre und speziell geschirmte Kupferkabel entwickelt. Kurz darauf wurden die Entwürfe von der IEEE 802.3-Arbeitsgruppe anerkannt. 1999 wurde die noch fehlende Spezifikation für *Twisted Pair* veröffentlicht.

Die Vollduplexvariante ist vollständig abwärtskompatibel zu den 10Mbps- und 100Mbps-Versionen. Allerdings ist es unmöglich eine kompatible Halbduplex-Version des CSMA/CD-Protokolls für Gigabit Ethernet zu entwickeln. Der Aufwand einer nutzbaren CSMA/CD-Implementierung war enorm, leider fand sich kein Hersteller, der sie implementierte.

### 7.1 Einführung

Normalerweise werden die Stationen innerhalb eines Vollduplex-LANs über Switches an das Netz angebunden. Da Gigabit-Switches relativ teuer sind, haben die Hersteller eine kostengünstige Lösung entwickelt. Das alternative Netzwerkgerät wird als *full-duplex repeater* oder *buffered distributor* bezeichnet.

Kunden, die Gigabit-Ethernet-Komponenten erworben haben, verlangen hohe Leistung von ihren Komponenten, so daß die Hersteller ihre Mechanismen durch eine weitere Funktion, *Jumbo Frame* genannt, erweiterten und Durchsatz weiter erhöhten. Die Größe eines MAC Jumbo Frames ist 9018 bytes groß. Es sei angemerkt, daß Jumbo Frames nicht standardisiert sind und es besteht desweiteren keine Aussicht, daß die IEEE 802.3 dies jemals in die Definition aufnimmt.

### 7.2 Gigabit-Konfigurationen

Abbildung 7.1 zeigt einen Ausschnitt einer Netzwerkumgebung, wie sie sich in einem Gebäude befinden könnte. Die Server werden über 100Mbit-Switches angebunden. Diese integrieren auch gleichzeitig die anderen Workstations über 10Mbps-Switches/-Hubs. Die 100Mbps-Switches verfügen über jeweils einen Gigabit Ethernet Port, sodaß die beiden Switches mit 1000Mbps verbunden sind. Somit wird die Kommunikation zwischen Servern und Clients enorm beschleunigt, ohne den Einsatz eines „echten“ Gigabit-Switches.

Abbildung 7.2 zeigt eine andere Anordnung. Die Server verfügen über Hochleistungs-NICs, die wiederum an einen 100/1000Mbps-Switch angeschlossen sind. Gigabit NICs wurden auch in die Clients integriert, die beispielsweise netzwerkintensive Berechnungen durchführen. Diese Workstations teilen sich die Bandbreite des *full-duplex (duplex) repeater*. Ein Satz 10Mbps-Switches verbindet die „einfachen“ Workstations mit dem zentralen 100Mbps-Switch via 100Mbps-Uplinks.

### 7.3 Vollduplex Gigabit Ethernet

Vollduplex Gigabit Ethernet ist der heutige Stand der Dinge. Außerdem ist es völlig natürlich, daß auf den Vollduplex-Betrieb zurückgegriffen wird, da die PHY-Layer der Kupfer- und Fibre-Optic-Spezifikation

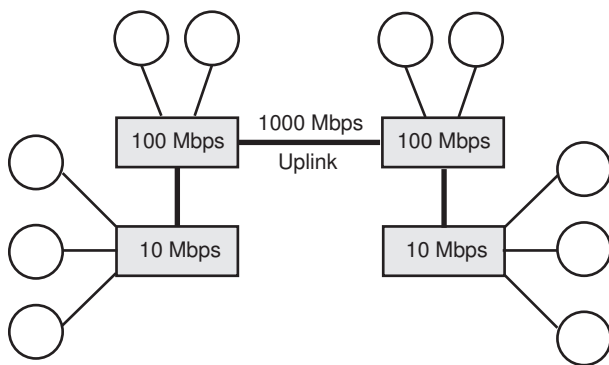


Abbildung 7.1: 100Mbit mit Gigabit verbinden.

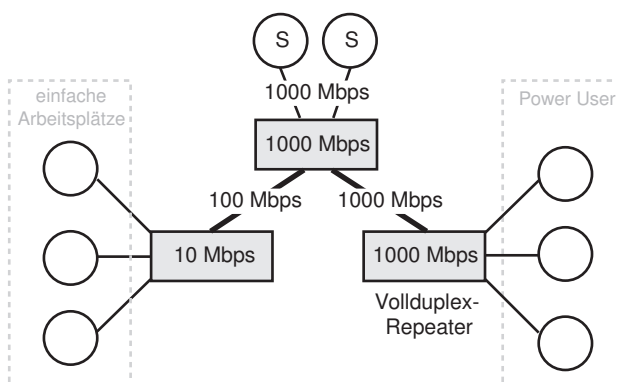


Abbildung 7.2: Switch und *Buffered/Full-Duplex* Repeater.

diesen Modus unterstützen. Gigabit Ethernet implementiert Vollduplex nicht anders als es 10 oder 100Mbps tun. Der Jumbo Frame ist allerdings eine gravierender Unterschied, der weiter unten erläutert wird. Tabelle 7.1 listet die Parameter des Vollduplex-Modus der verschiedenen Geschwindigkeiten auf.

### 7.3.1 Vollduplex-Repeater

Dieses Gerät wurde entwickelt, um die Engpässe der normalen Hub zu beseitigen. Anders als Switches, haben die „buffered repeater“ keine Filtermechanismen oder den hohen Durchsatz durch die parallele Verarbeitung.

Vollduplex-Repeater haben folgende Aufgaben:

- jeder Frame wird an alle Ports weitergeleitet
- alle Systeme werden durch Vollduplex-Links angebunden
- puffert eintreffende Rahmen, um sie für die Weiterleitung nach dem FIFO-Prinzip (*first-in first-out*) zu verarbeiten

Parameter	10Mbps	100Mbps	1000Mbps
InterFrameGap	9,6µm (96 Bit-Zeiten)	0,96 (96 Bit-Zeiten)	0,096 (96 Bit-Zeiten)
maxFrameSize	1518 Bytes	1518 Bytes	1518/9018 Bytes
minFrameSize	64 Bytes	64 Bytes	64 Bytes

Tabelle 7.1: Vollduplex-Ethernet-Parameter.

- teilt eine Bandbreite eines Vollduplex-Links (insg. 2000Mbps) mit allen Geräten.
- Es wird ein PAUSE-Rahmen als Flußkontrollmechanismus an die Endsysteme gesendet, sobald sich der Puffer (*Queue*) derselben dem Ende neigt. So wird eine Overflow-Situation vermieden, die das Verwerfen der Rahmen zur Folge hätte.

Abbildung 7.3 liefert eine Darstellung, wie ein Vollduplex-Repeater arbeitet. Die Details der I/O-Queues und deren Management liegen beim Hersteller. In der Figur repräsentiert jeder Großbuchstabe einen Rahmen. Die alphabetische Order liegt dem Eintreffen der Frames zugrunde. Der obere Teil der Abb. stellt die I/O-Queue, die schon fast voll ist, der Station 1 dar. Der Repeater sendet einen PAUSE-Frame an Station 1, so daß die Übermittlung der Frames vorübergehend ausgesetzt wird.

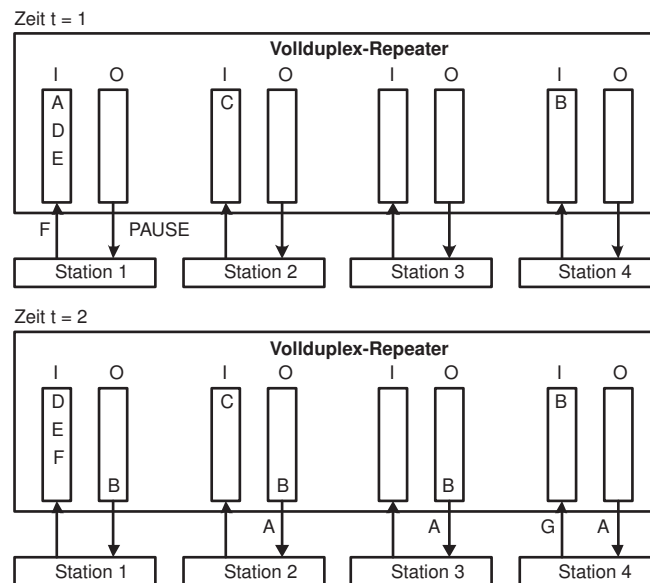


Abbildung 7.3: Gigabit Ethernet Vollduplex-Repeater.

Im unteren Teil der Abbildung befindet sich der gleiche Repeater kurzeZeit später. Jeder eintreffende Rahmen wurde gespeichert und wird an alle Ports weitergeleitet. Frame A ist auf dem Weg zu den Stationen 2, 3 und 4, während sich Frame B in der Queue auf dem Weg zu Station 1, 2 und 3 befindet. Beachten Sie, daß Station G gleichzeitig einen Frame an den Repeater sendet.

### 7.3.2 Jumbo Frames

Die Standardframegröße ist 1518 Bytes, zusätzlich muß die Preamble, der Start-of-Stream-Delimiter, der End-of-Stream-Delimiter und das Interframe Gap hinzugerechnet werden. Das bedeutet, daß 12.304 Bit-Zeiten durch die maximalen Framegrößen aufgewendet werden. Über Gigabit-Ethernet können maximal 81.274 Rahmen maximaler Größe übertragen werden. Die Zahl erhöht sich weiter bei unterschiedlichen Framegrößen. Jeder Frame muß von der Endstation bearbeitet werden, mit dem Resultat, daß sie von den Gigabit-Verbindungen völlig überfordert sind.

Viele Hersteller versuchten, das Problem mit nicht-standardisierten Framegrößen zu beseitigen. Jumbo Frames haben eine Gesamtgröße von 9018 Byte, 9000 Byte umfaßt das Informationsfeld, das ist die gleiche Menge, die 6 „normale“ Standardframes tragen würden.

Die Jumbo Frames bewirken eine Senkung der Anzahl der gesendeten Frames, als auch die Zeit, die die Endgeräte benötigen um jeden Frame im Verhältnis zu seiner Größe zu verarbeiten. Es müssen nun nicht viele kleine Frames bestätigt werden, dadurch sinkt der Netzwerkverkehr auf ein Minimum. Diese Verfahren machen sich auch andere Netzwerktechnologien zu Nutze. So ist die Framegröße bei 16Mbps-Token-Ring und ATM relativ ähnlich.

### 7.3.2.1 Auswirkungen auf den Durchsatz

Der größte Gewinn wird durch die geringe Verarbeitungszeit der Rahmen an den Endstationen gemacht. Ein Jumbo Frame enthält die Daten von 6 Standardframes, sodaß eine Bestätigung für 6 MAC-Frames gesendet werden muß. Dadurch sind die Endgeräte in der Lage effektiv mehr Daten zu verarbeiten. Das bedeutet:

Wenn man die 18 Frame-Header- und -Trailer-Bytes mit den 20 Overhead-Bytes addiert, werden 38 Bytes (304) als Overhead für jeden Frame aufgewendet. Für einen konstanten Datenstrom von 1518-Bytes Frames, wird in dieser Overhead auf 3.100.000 Bytes pro Sekunde anwachsen. Mit Hilfe der Jumbo Frames kann der Overhead auf weniger als 526.000 Bytes pro Sekunde reduziert werden. Diese Differenz kann für zusätzliche Daten verwendet werden.

### 7.3.2.2 Vor- und Nachteile der Jumbo Frames

Das IEEE 802.3-Komitee hat sich mit der Empfehlung der Jumbo Frames sehr zurückgehalten und wird in absehbarer Zeit diese Strategie nicht ändern. Der Grund sind Kompatibilitätsprobleme. Beispielsweise können diese Framegrößen nicht eingesetzt werden, wenn nicht jeder Port des Buffered Repeaters und nicht jeder NIC diese Größe unterstützen.

Eine Applikation hingegen profitiert sehr von den Jumbo Frames. Das NFS (*Network File System*) Protokoll verwendet Framegrößen von 8.192 Bytes. Das UDP-Protokoll wird normalerweise für die Übertragung dieser Daten verwendet. Oft müssen bei konventionellen Framegrößen die UDP-Pakete in sechs Teile fragmentiert werden. Ein Jumbo Frame macht diesen Prozess überflüssig, eine Reassemblierung auf Empfängerseite wird ebenfalls entfallen.

# Kapitel 8

## Gigabit PHY-Layer

Es existieren vier standardisierte Gigabit-Ethernet-Implementierungen

### 1000Base-SX

Eine Fibre-Optic-Implementation über ein optisches Multimode-Adernpaar. SX entspricht der Bezeichnung *short wavelength* (kurze Wellenlänge). Laser emittieren Licht mit der Wellenlänge von 850 nm.

### 1000Base-LX

Diese Variante ist sowohl für den Betrieb über ein optisches Multimode- oder Singlemode-Adernpaar ausgelegt. LX korrespondiert mit *long wavelength* (lange Wellenlänge). Laser emittieren das Licht bei einer Wellenlänge von 1300 nm.

### 1000Base-CX

Hier wird ein kurzes geschirmtes Kupferkabel aus der zentralen Verkabelung in die entsprechenden Verteiler geführt. CX bedeutet in diesem Zusammenhang *copper* (Kupfer).

### 1000Base-T

Ein ungeschirmtes Twisted-Pair-Kabel (UTP) mit vier TP-Adern.

Die ersten drei Spezifikationen werden gewöhnlich einer Gruppe 1000Base-X zugeordnet. Sie basieren grundsätzlich auf der Fibre Channel Technologie und haben viele Gemeinsamkeiten wie etwa die Kodierungsschemata.

1000Base-T, die Twisted-Pair-Version, verwendet eine hochkomplexe Kodierungstechnik und erfordert die Entwicklung eines neuen Transmissionsmechanismus. Desweiteren benötigt 1000Base-T eine ausgefeilte Hardware.

## 8.1 Merkmale des Gigabit Ethernet

Folgende Eigenschaften sind für das Gigabit Ethernet und deren Implementation gültig:

- Zu Beginn der Initialisierung oder einem Reset, wird das Gigabit-Interface eine Auto-Negotiation durchführen, um mit den Link-Partnern die Grundregeln der Kommunikation auszuhandeln.
- Nach dieser Phase, werden auf dem PHY-Level kontinuierlich Signale in beide Richtungen gesendet.

Diese Merkmale werden in den folgenden Abschnitten erläutert.

### 8.1.1 Auto-Negotiation

Alle Gigabit-Technologien beherrschen den Austausch sog. *Link Parameter* über das Auto-Negotiation-Protokoll. Folgende Aspekte sind für Auto-Negotiation wichtig:

- Auf beiden Seiten wird der Betrieb bei 1000Mbps bestätigt.
- Jede Schnittstelle gibt bekannt, ob sie im Halb- oder Vollduplexmodus (CSMA/CD) arbeiten wollen. Es wird immer der Vollduplexmodus angewendet, wenn es die Schnittstellen erlauben.
- Für Vollduplex-Verbindungen, wird zusätzlich ermittelt, ob die PAUSE-Flußkontrolle verwendet wird und ob sie symmetrisch oder asymmetrisch ist.

### 8.1.2 Bidirektionale Gigabit-Transmissionen

Alle aktuellen Gigabit-Ethernet-Technologien sind Vollduplex-fähig. Auf dem physikalischen Level werden Signale in beiden Richtungen kontinuierlich übertragen. Das Interface sendet Idle-Symbole aus, wenn das Medium „stumm“ ist. Wenn ein Gigabit-Link auf Halbduplex-Betrieb konfiguriert wurde, muß verhindert werden, daß Frames während des Sendes empfangen werden. Der obere Teil der Abb. 8.1 zeigt, wie 1000Base-SX- und -LX-Transmissionen funktionieren. Der mittlere Teil illustriert die 1000Base-CX-Implementation, bei der ein separates Paar für jede Richtung verwendet wird. Auch hier werden die Idle-Symbole eingesetzt, wenn das Medium „stumm“ ist. Unten wird zusätzlich die 1000Base-T-Funktionalität aufgezeigt.

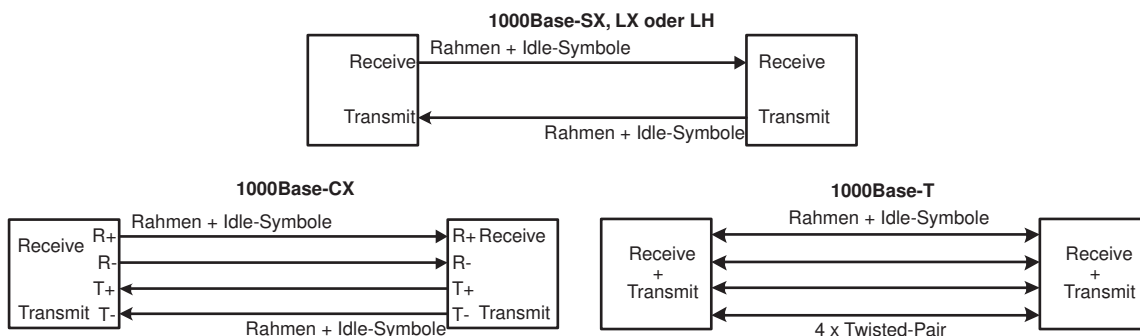


Abbildung 8.1: Gigabit-Übertragungsmechanismen.

## 8.2 Physikalische Eigenschaften des Gigabit Ethernet

Es existieren Implementierungen für Single-/Multimode-Fibre Optic, kurze geschirmte und ungeschirmte Twisted-Pair-Kabel. Tabelle 8.1 listet die entsprechenden Längen der verschiedenen Gigabit-Ethernet-Varianten auf.

Beachten Sie, daß alle SX-Implementationen und alle LX-Implementierung bis auf eine, über Multimode-Fibre-Optic übertragen. Eine LX- und eine alle LH-Varianten erfordern Singelmode-Fibre-Optic.

Der Grund für die verschiedenen LX und SX-Einträge sind die Abhängigkeiten der Kabellänge von der *modalen Bandbreite*. Sie beeinflusst die Fähigkeit bestimmter Multimode-Fibre-Optic-Kabel, Signale über verschiedene Distanzen zu übertragen. Formal wird die modale Bandbreite als *worst case* Bandbreite bei  $-3\text{ dB}$ , die ein Medium erreichen kann, definiert. Diese modale Bandbreite wird in  $\text{MHz} \times \text{km}$  angegeben. Um die Bandbreite eines Mediums in MHz zu berechnen, muß die modale Bandbreite durch die Länge des Kabels in km dividiert werden. Beispiel: die Bandbreite eines 500 m langen Kabels (0,5 km) mit einer modalen Bandbreite von 500 entspricht:  $500/0,5 = 1000\text{ MHz}$ .

Die *modale Dispersion* ist der Hauptfaktor, der die modale Bandbreite bestimmt. Diese Dispersion entsteht durch die Tatsache, daß einige Lichtstrahlen, die in einen relativ großen Kern emittiert werden

Transceiver	Kabeltyp	Kerndurchmesser	Modale Bandbreite	Kabellängen in <i>m</i>
1000Base-CX	STP	N/A	N/A	0.1-25
1000Base-SX	MMF	62.5	160	2-220
1000Base-SX	MMF	62.5	200	2-275
1000Base-SX	MMF	50	400	2-500
1000Base-SX	MMF	50	500	2-550
1000Base-LX	MMF	62,5	500	2-550
1000Base-LX	MMF	50	400	2-550
1000Base-LX	MMF	50	500	2-550
1000Base-LX	SMF	10	N/A	2-5000
1000Base-LH	SMF	9	N/A	1000-100.000
1000Base-T	4 UTP	N/A	N/A	100

Tabelle 8.1: Vollduplex-Ethernet-Parameter (MMF = *multimode fibre*, SMF = *singlemode fibre*).

einen längeren Weg nehmen und später als andere auf Empfängerseite eintreffen. Ein hoher Dispersionsfaktor verringert die Fähigkeit des Empfängers die Signale sauber zu erkennen proportional. Desweiteren wird der Dispersionsfaktor von der Qualität, der Länge und Art, wie das Licht in das Kabel gelangt, bestimmt.

Für den 1000Base-LH-Eintrag in der Tab. 8.1 sind viele Kabellängen möglich. Einige 1000Base-LH-Transceiver sind einfach nur hochwertige 1000Base-LX-Transceiver, deren Hersteller eine zuverlässige Maximallänge von 10km garantieren. Einige LX-Variationen unterstützen ebenfalls sehr große Kabellängen. Zur Zeit existieren Implementationen für Singlemode-Kabellängen bis zu 100 *km*. Teilweise sind die Wellenlängen der hier verwendeten Laserstrahlen größer als 1300 *nm*, die für LX definiert sind.

## 8.3 1000Base-X-Technologie

Wie bereits bei 100Mbps hat sich auch diese Gruppe einige existierende Standards geerbt. Die Kodierungsmechanismen sind bereits vorhanden und in einigen 100Mbps-Standards (100Base-LX/SX/CX) definiert. Sie ließen sich sehr einfach anpassen und implementieren. Somit konnten die Produkte schnell zur Marktreife gelangen.

### 8.3.1 Die 8B/10B-Kodierung

Bei 1000Base-SX, 1000Base-LX und 1000Base-CX wird jedes Byte in einem Frame in einen 10-Bit-Code umgewandelt (*mapping*), bevor es auf die Reise geht. Diese Umwandlung wird als 8B/10B-Encoding bezeichnet.

Die jeweiligen 10-Bit-Codes werden als *Codegruppen* bezeichnet. es sind nur 256 8-Bit-Muster und 1024 10-Bit-Muster vorhanden. Nur ein Subset der gesamten Palette wird effektiv für die Repräsentierung der Daten verwendet. Dieses Subset wurde sorgfältig ausgewählt:

- Die ausgewählten Codegruppenmuster enthalten ein gute Durchmischung von 0s und 1s. Das hilft dem Empfänger die Synchronisation zu wahren.
- Desweiteren verbessert die Durchmischung mit 0s und 1s das elektrische Verhalten der Signale über das Kupferkabel und verhindert die Überhitzung der Laser optischer Kabel.
- Treten ein oder mehrere Bitfehler auf, wird das 10-Bit-Muster in ein ungültiges Muster verwandelt, daß nicht die eigentlichen Daten repräsentiert. Somit wird die Fehlererkennung erweitert.
- Es existieren relativ viele 10-Bit-Muster, so daß es leicht ist, neben den Idle-Symbolen und Rahmenerweiterung neue Codegruppenmuster zu definieren.

### 8.3.2 1000Base-SX- und -LX-Transmission: Laser und VCSELs

Für die Transmission über Langstrecken mit 100Mbps und 10Mbps werden Laser eingesetzt, kurze Strecken werden mit der weniger teuren LEDs betrieben. LEDs können jedoch nicht für 1000Mbps eingesetzt werden, weil sie einfach zu träge sind.

Es wurde eine neue Lasertechnologie entwickelt, die als *Vertical Cavity Surface Emitting Laser* (VCSEL) bekannt wurde. Diese Laser emittieren Licht mit einer Wellenlänge von 850 nm und sind in der Lage Entfernungen von mehreren hundert Meter zu überbrücken. Die hochwertigeren Laser, die für die Transmission über Singlemode-Phasern mit einer Wellenlänge von 1300 nm verwendet werden, sind in 1000Base-LX implementiert. LX-Laser arbeiten sowohl mit Multimode und Singlemode. Die maximale Länge ist auf 550 m beschränkt. Im Kontrast dazu stehen die LX-Singlemode-Implementationen, die mehrere Kilometer überbrücken.

## 8.4 1000Base-T-Technologie

Die Wahl über ein geeignetes Cat. 5 kann angesichts der Vielfalt nicht schwer fallen. Es werden mindestens 4 ungeschirmte Twisted-Pairs benötigt, um die Geschwindigkeit zu liefern.

### 8.4.1 1000Base-T-Encoding

Der bidirektionale Charakter von 1000Base-T macht die Kodierung nicht einfach. Es muß nicht nur eine gewisse Balance zwischen 0s und 1s garantieren, sondern auch das elektromagnetische Verhalten der Twisted-Pairs ausgleichen.

Die Kodierung erfordert mehrere Schritte:

- Jedes Bit wird „gescrambled“.
- Das gescramblete Bit wird in ein Quadrupel von speziellen Symbolen via *8B1Q4*-Mapping.
- Jedes Symbol wird auf dem Medium durch eine Spannung repräsentiert. Fünf unterschiedliche Spannungsmuster werden unterstützt. Der physikalische Transmissionsmodus heißt *4-Dimensional 5-Level Pulse Amplitude modulation* (4D-PAM5).

### 8.4.2 Verkabelung

Eine Twisted-Pair-Gigabit-Verkabelung muß strikten Qualitätsansprüchen genügen. Sie muß eine Vielzahl an Tests über sich ergehen lassen, bis der eigentliche Betrieb aufgenommen werden kann. Cat. 5E-Kabel (*Enhanced Category 5*) sind gewöhnliche Twisted-Pair-Kabel, die diese Test bestanden haben. Neue Kabel müssen mit Vorsicht installiert werden, wenn wieder Cat. 5 eingesetzt wird, muß auch dieses erneut vermessen werden.

Ein Beispiel für den Einfluß der Kabelqualität ist das Crosstalk von anderen Kabeln (*alien crosstalk*). Schlechte Abschirmung verhindert den reibungslosen Betrieb von Gigabit Ethernet. Zusätzlich müssen folgende Probleme berücksichtigt werden:

- Zu viele Verbinder im Netzwerk verringern die Signalqualität.
- Nichtstandardisierte Konnektoren sollten grundsätzlich vermieden werden.
- Keine losen Installationen durchführen, da sich die Zahl der möglichen Fehlerquellen exponentiell erhöht.
- Die Kabel müssen unbedingt den in Cat. 5E aufgeführten Qualitätsansprüchen genügen.

### 8.4.3 Encoder, Decoder und Hybride

Abbildung 8.2 zeigt ein Signal das über eine 1000Mbps-Schnittstelle gesendet wird. Das Dreieck mit dem T repräsentiert den Transmit-Encoder. Das andere Dreieck enthält ein R und stellt den empfangenden Receive-Decoder dar. Während des Encoding-Prozesses wird jedes Byte in ein 4-Bit-Muster umgewandelt. Jedes der 4 Paare enthält ein Code-Symbol und werden gleichzeitig übertragen. Dann wird das Quadrupple in ein Byte decodiert. Die Hybridkomponenten in der Figur vermeiden Interferenzen der lokal transportierten mit eintreffenden Signalen.

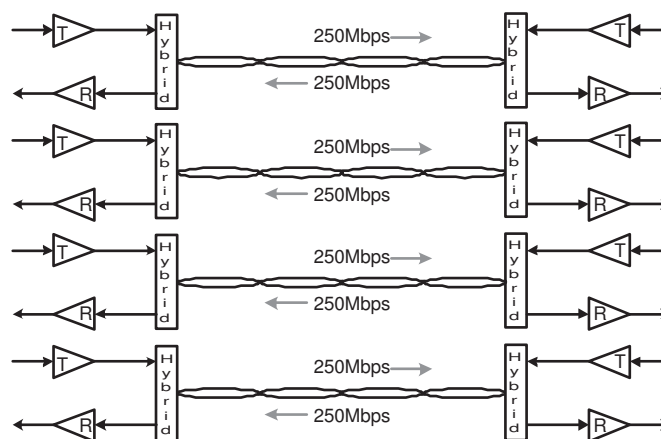


Abbildung 8.2: 1000Base-T-Übertragung mit 4 Twisted-Pair-Kabeln.

### 8.4.4 Master/Slave-Timing

Die Synchronisation der Timings ist ganz besonders wichtig, weil in beiden Richtungen transportiert wird. Die Bits werden zwischen den Partnern synchronisiert, in dem ein Teilnehmer die „Masterstellung“ (*master-timing*) einnimmt. Der interne Taktgeber wird für die Erzeugung des Timings verwendet. Das andere Ende übernimmt die Funktion des „Slaves“ (*slave-timing*) und taktet die ankommenden Bits und benutzt diesen Takt für eigene Transmissionen. Die Aufgabenverteilung kann manuell im Vorfeld der Kommunikation durchgeführt werden oder über *Auto-Negotiation* zwischen den Partnern.

Nachdem die Auto-Negotiation durchgeführt wurden, initiiert der Master eine Trainingsperiode. Während dieser Phase sendet der Master eine Sequenz von Idle-Symbolen, um das Timing zu synchronisieren. Nach dem Training beginnt die Transmission.

### 8.4.5 Auto-Negotiation und Crossover

Die Abb. 8.3 kann etwas irreführend sein, da man meinen könnte, es wird ein Straight-through-Kabel verwendet. Tatsächlich müssen wie für 10Base-T, 100Base-TX und 100Base-T4 auch für 1000Base-T Crossover-Kabel verwendet werden. Der Grund dafür ist, daß die Schnittstellen in der Initialisierungsphase in der Lage sein müssen die Auto-Negotiation durchzuführen. So wird sichergestellt, daß zum einen die andere Seite als TP-fähig erkannt wird und die Kompatibilität zu den anderen TP-Ethernet-Standards gewahrt bleibt. Abb. 8.3 zeigt das Diagramm der Pinbelegung nach bekanntem Schema.

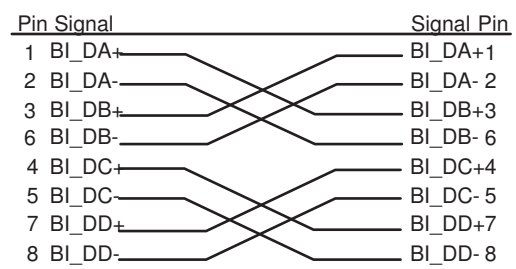


Abbildung 8.3: Crossover MDI/MDI-X-Pinlayout für 1000Base-T.

# Kapitel 9

## Standards

Kabelstandards sind wichtig, um eine konsistente und wettbewerbsfähige Basis für die Entwicklung von hochwertigen Produkten zu gewährleisten. Kabelstandards geben auch Messgeräteherstellern die Möglichkeit, sich auf die geltenden Regeln einzustellen und Prüfgeräte auf dieser Grundlage zu entwickeln.

Twisted-Pair-Verkabelung ist das mit Sicherheit am weitesten verbreitete Medium für LANs. Aus diesem Grund fokussiert dieser Abschnitt TP-Kabel und -Standards. Auf Optic Fibre wird am Ende eingegangen.

### 9.1 Standardisierungsgremien (*Standard Bodies*)

In den Vereinigten Staaten von Amerika hat die TIA (*Telecommunications Industry Alliance*) die Hoheit über die Kabelstandards. Die Muttergesellschaft der TIA ist die EIA (*Electronic Industry Alliance*). Die Anforderungen an die Verkabelung werden in dem Standarddokument 568A *Commercial Building Telecommunications Cabling Standards* aus dem Jahre 1991 veröffentlicht. Diese Dokumente, respektive Standards, werden regelmäßig angepaßt.

Die kanadische Einrichtung ist die CSA deren Standard als CSA T529 bezeichnet wird. In Europa sind die ISO/IEC-Gremien für die Standardisierung der Kabel zuständig und fassen diese in ISO-IEC 11801 zusammen. Sie basieren auf den Empfehlungen der TIA/EIA, sind allerdings an die europäischen Bedürfnisse angepaßt.

Eine große Zahl anderer Gremien wurden gegründet, die unter anderem Vorschläge und Entwürfe den Standardisierungsgremien zukommen lassen. Darunter sind die 802.3-Arbeitsgruppe der IEEE, das ATM Forum, das *Comite Europeen de Normalisation Electrotechnique* (CENELEC) und viele andere.

### 9.2 TIA/EIA-Kategorien

Die Hauptaktivität liegt bei der Erstellung verschiedener Kategorien (*Categories*, Abk. Cat.), die für verschiedene Anwendungen geeignet sind. Es existieren Nummern zur Bestimmung der Kategorie und der beteiligten Komponenten, die verwendet werden. Hier sind Kabel, Kabelsysteme, Kabelkanäle, Buchsen, Stecker, Verbinder und Patch-Felder zu nennen.

Die Ingenieure sind bemüht die höheren Kapazitäten in die Standards einfließen zu lassen. Es existieren derzeit 7 Hauptkategorien für Twisted-Pair, eine 8. wird spezifiziert. Je höher die Kategorie, desto besser ist die Qualität. Jede Kategorie über 1 garantiert eine Bandbreite, die in MHz angegeben wird. Fast alle aktuellen Installationen bestehen aus Cat. 5, 5E und 3.

Beachten Sie, daß diese Kategorien Ihre Kaufentscheidungen beeinflussen werden, denn bestimmte Geräte erfordern bestimmte Kategorien. Die besten Komponenten und schnellsten Systeme können das gesamte Netzwerk fast unbrauchbar machen, da an den verschiedensten Stellen Flaschenhalse durch eine falsche Kabelwahl entstehen können.

Folgende Kategorien haben diese Eigenschaften:

**Category 1**

Manchmal auch unter *barbed wire* (Stacheldraht) bekannt. Existiert nicht mehr und hat keinerlei Besonderheiten.

**Category 2**

Eine verbesserte Version des Cat. 1. Wird in PBXs (Telefonvermittlungsanlagen) und für digitale Voice-Dienste eingesetzt.

**Category 3**

Weit verbreitet in 10Base-T-LANs. Enthält ein oder mehrere zusätzliche Kupferadern in einer einzelnen Isolierung.

**Category 4**

Nahezu indentisch mit Cat. 3. Die Kabelqualität ist dennoch eindeutig besser.

**Category 5**

Sehr populär, da es für 100Mbps-Kommunikation geeignet ist, Vollduplexfähigkeit und besondere Abschirmung besitzt. Gigabit Ethernet ist auch für Cat. 5 vorgesehen. Die Netzwerkgeschwindigkeit ist dann allerdings sehr von einer durchgehend hohen Qualität der Kabel abhängig.

**Category 5E**

wurde für den wachsenden Bedarf nach Gigabit-Ethernet entwickelt und stellt eine höherwertige Verarbeitung, sowohl im Prozess und der Wahl der Komponenten dar.

**Category 6**

Müssen eine Bandbreite von 250 MHz über 100 m bereitstellen und die RJ-45-Steckerform unterstützen.

**Category 7**

Die bisherigen Informationen geben darüber Auskunft, daß es immernoch auf Kupfer basiert, allerdings doppelt geschirmt ist, sodaß es möglicherweise schwerer als andere TP-Kabel ist und teurer als Fibre Optic sein könnte.

### 9.2.1 Kabellayout

Ein typisches Beispiel für eine Kabelvariante ist in *Ethernet 10Mbit PHY-Layer* beschrieben. Dieses Layout ist konform zu den Standards der ANSI/TIA/EIA 568-A, unterstützt das 10 m Patch-Kabel und 90 m Crossover. Desweiteren existieren eine Menge anderer Kabelvarianten, die in ANSI/TIA/EIA TSB 75 definiert sind.

Diese sieht die Verwendung sog. *Multiuser telecommunications outlet* (MUTO) vor. Sie können bis zu 24 Arbeitsgruppenkabel aufnehmen, solange die 20 m nicht überschritten werden. Sie laufen direkt auf die MUTOs, die maximal 70 m von dem Verteiler (*wiring closet*) entfernt sein dürfen. Da sie eine zusätzliche Schnittstelle ersetzen, wird keine weitere Verkabelung mehr benötigt. Es ist auch üblich bis zu 25 Kabel zu einem Strang zusammenzufassen, was man mit 1000Mbps tunlichst vermeiden sollte.

Cat.	Bandbreite	Beschreibung
1	N/A	N/A
2	N/A	160
3	16MHz	Geeignet für 10Base-T Ethernet, 4Mbps TR, 100Base-T4, digitale Voice-Dienste.
4	20MHz	Ausgelegt für 10Base-T Ethernet und 16Mbps Token Ring.
5	100MHz	Zertifiziert für 100Mbps Ethernet, 1000Mbps Ethernet, auch für CDDI
5E	100MHz	Geeignet für 100/1000Mbps Ethernet, zertifiziert für Hochgeschwindigkeitsnetze
6	250 MHz	ausgelegt für 1000Mbps und höher, höherwertige Fertigung, RJ-45-kompatibel
7	600-1200	Zertifiziert für 1000Mbps und höher, schwer und teuer.

Tabelle 9.1: Vollduplex-Ethernet-Parameterd

## 9.3 UTP-Leistungsparameter

Leistungsparameter sind Kenngrößen, die Bedingungen, Zustände und Verhalten des Übertragungsmediums beschreiben. Sie sind maßgeblich für die Leistungsfähigkeit verantwortlich.

Das Signal kann abgeschwächt werden, während es das Medium durchwandert. Wenn die verdrehten Kupferpaare nicht eng genug verdreht sind, können die verschiedenen Level für positive und negative Signale nicht mehr korrekt erkannt werden, da die elektromagnetischen Interferenzen sich z.T. auslösen. Schlechte Kabelstrukturen können für Reflektionen am Kabelende sorgen.

### 9.3.1 Parameter für alle Ethernet-LANs

Es müssen einige Basisanforderungen der TIA/EIA erfüllt werden. Sie ermöglichen eine Einstufung in diverse Leistungskategorien:

#### Dämpfung

Das Verhältnis zwischen der Ausgangsleistung und der Leistung, die während der Signalwanderung über das Medium verloren geht.

#### Near-end-Crosstalk (NEXT)

Eines starkes ausgehendes Signal erzeugt ein Rauschen, daß das schwache Signal auf der benachbarten Ader beeinflusst.

#### Widerstand

Beschreibt „Gegenkraft“, die ein Medium dem Elektronenfluß entgegenbringt. Unreinheiten im Kabel beeinflussen den Widerstand und damit die Übertragungsgeschwindigkeit.

Die Prüfung des Kabels beinhaltet auch das Testen der Adern. Folgende Faktoren werden berücksichtigt:

- Korrekte Pin-Terminierung an jedem Ende Kontinuität am entfernten Ende
- Kreuzpaare und umgedrehte Paare Splitpaare (TP-Adern werden getrennt und kreuzverbunden)

Die 802.3-Spezifikationen sehen noch andere Faktoren vor:

- Kabellänge Jitter (Abweichung zwischen aufeinanderfolgender Signalflanken)
- Transportverzögerung (propagation delay)

### 9.3.2 Parameter für Hochgeschwindigkeits-LANs

Die Anforderungen für High-Speed-LANs sind weitaus restriktiver. Für 1000Base-T Ethernet-Transmissionen, die gleichzeitig stattfinden ist dies besonders wichtig. Einige Parameter sind:

- *Far end crosstalk* (FEXT)
- *Equal level far end crosstalk* (ELFEXT)
- *Power sum near end crosstalk* (PSNEXT)
- *Worst pair-to-pair* (ELFTEXT)
- *Power sum* (ELFEXT)
- *Attenuation to crosstalk ration* (ACR), Verhältnis Dämpfung/Crosstalk
- *Delay skew Return loss*

### 9.3.3 Parameterbeschreibung

Es folgt eine detailliertere Beschreibung der Parameter für Low- und High-Speed-LANs

#### 9.3.3.1 Jitter

Jede Transmission basiert auf einer zuvor ausgehandelten Geschwindigkeit bei der ein Bit übermittelt wird. Zwischen jedem übermittelten Signal wird eine Pause eingelegt, so daß sich bei Abweichungen im Intervall Probleme ergeben können, da das Signal nicht eindeutig zugeordnet werden kann. Diese Abweichung im Intervall wird als *Jitter* bezeichnet und sollte nicht größer als 5 *nm* sein.

#### 9.3.3.2 Dämpfung in Dezibel (dB)

Während der Transmission verliert das Signal an Schärfe, es wird gedämpft. Je größer die Dämpfung ist, desto schwerer bis unmöglich ist die Erkennung des Signals auf der Gegenseite. Die Dämpfung wird mit Hilfe der folgenden Logarithmen beschrieben:

$$\text{Dämpfung} = 10 \log_{10}(\text{Eingangssignal} / \text{Ausgangssignal}) \text{dB}$$

Wenn das Ausgangssignal 1/10 des Eingangssignal ist, beträgt die Dämpfung:

$$10 \log_{10}(\frac{1}{10}) = 10 \log_{10}(10^{-1}) = 10(-1) = -10 \text{dB}$$

Sollte das Ausgangssignal 1/100 des Originals betragen, ist die Dämpfung folglich:

$$10 \log_{10}(\frac{1}{100}) = 10 \log_{10}(10^{-2}) = 10(-2) = -20 \text{dB}$$

Je weiter das Signal wandert, desto mehr wird es gedämpft. Es sei angemerkt, daß immer eine Dämpfung existiert und daß sie immer negativ ist. Oftmals wird das Minuszeichen in der Fachliteratur weggelassen. Die Dämpfung ist außerdem auf jeder Frequenzstufe anders. Höhere Frequenzen erfahren höhere Dämpfungen.

#### 9.3.3.3 Near End Crosstalk (NEXT)

Ein starkes Signal von einer Station kann ein schwaches, eintreffendes Signal beeinflussen, wenn die übertragenden Adern benachbart sind. Das starke Signal erzeugt ein elektromagnetisches Feld, das die Signale der anderen Ader teilweise auslöscht.

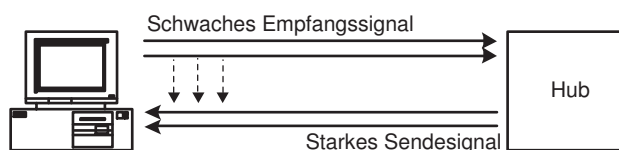


Abbildung 9.1: Near End Crosstalk (NEXT).

Die Stärke des NEXT wird in dB gemessen, genauer gesagt ist das NEXT die Menge Energie, die auf das schwache Signal übertragen wird. Die Menge steigt bei höheren Transferraten, da die Frequenzen steigen.

#### 9.3.3.4 Far End Crosstalk (FEXT)

FEXT ist die Menge Energie, die von einem starken ausgehendem Signal auf ein eintreffendes schwaches Signal gegen Ende des Segments überträgt. Enge Verdrillung der TPs in Verbindung mit guten Konnektoren kann der FEXT in Zaum gehalten werden. In Abbildung 9.2 ist erkennbar, daß das Signal auf dem Weg an Leistung verloren hat, jedoch immernoch stark genug ist, um auf das andere TP einzuwirken.

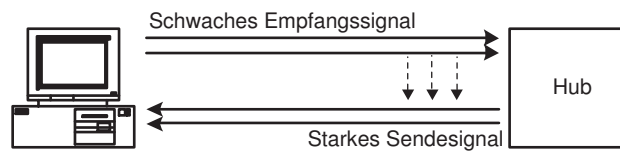


Abbildung 9.2: Far End Crosstalk (NEXT).

### 9.3.3.5 PSNEXT, PSELFEXT und worst Pair-to-Pair ELFEXT

100Base-T4 und 1000Base-T verwenden 4 TP-Kabel für die Kommunikation. Jedes Paar wird für die Transmissionen verwendet und erzeugt deshalb Interferenzen, die die anderen drei Kabel beeinflussen. Daher gilt für jedes Paar A, B, C und D das folgende:

- B, C und D verursachen FEXT auf A
- A, C und D verursachen FEXT auf B
- A, B und D verursachen FEXT auf C
- A, B und C verursachen FEXT auf D

Das ergibt zusammen 24 FEXTs. Man hat die Liste anhand diverser Faktoren verringert, um ein Kabel zu bewerten:

- Power sum near end crosstalk (PSNEXT)
- Power sum equal level far end crosstalk (PSELFEXT)
- Worst pair-to-pair equal level far end crosstalk

Der PSNEXT ist die Summe der NEXT-Effekte auf ein Paar durch die anderen 3 Paare. Beispielsweise ist der PSNEXT von A die Summe der NEXTs von B, C und D. Es existiert zusätzlich noch ein PSEFEXT.

Ähnlich verhält sich das mit PSEFEXT. Er ist die Summe der ELFEXT-Effekte auf ein Paar durch die anderen 3 Paare. Beispielsweise ist der PSEFEXT von A die Summe der ELFEXTs von B, C und D.

Der *worst pair-to-pair ELFEXT* ist der größte ELFEXT-Effekt durch ein Kabel auf ein anderes.

### 9.3.3.6 Dämpfung/Crosstalk-Verhältnis (ACR)

Dieses Verhältnis basiert auf der Messung der Dämpfung und der Crosstalks auf Empfängerseite eines Paares. Dieser Wert wird in dB ausgedrückt. Je stärker ein Signal ist, desto kräftiger ist das Rauschen und der ACR-Wert positiv. Hohe ACR-Werte repräsentieren bessere Signalstärken.

Die ACR-Messung ist eigentlich eine ELFEXT-Messung und kann bei Bedarf leicht in eine solche umgewandelt werden.

### 9.3.3.7 Structural Return Loss und Return Loss

Ein Übertragungsmedium ist nicht uniform, so daß die gewünschte Impedanz abweichen kann. Eine ungleichmäßige Impedanz veranlaßt das Signal, an Schärfe zu verlieren, denn ein Teil wird zur Quelle reflektiert. *Structural Return Loss* (SRL) ist ein Maß für diese Reflektionen und wird in dB ausgedrückt.

Return Loss ist ein Maß für den relativen Wert die Reflektion an die Signalquelle. Es ist relativ zu einer dem Kabel typischen Ausgangsimpedanz von beispielsweise 100 Ohm im Fall von Twisted-Pair.

#### 9.3.3.8 Propagation Delay

Propagation Delay ist die Zeit, die ein Signal benötigt, um von einem Ende des Kabels an das andere zu gelangen. Dieser Wert hängt auch von der Länge des Kabels und der Geschwindigkeit der Elektronen. Die Verzögerung kann mit Hilfe der sog. *nominal velocity of propagation* (NVP) berechnet werden. Die NVP ist die Übertragungsgeschwindigkeit relativ zur Lichtgeschwindigkeit im Vacuum. Es wird in % der Lichtgeschwindigkeit angegeben.

#### 9.3.3.9 Delay Skew

1000Base-TX überträgt Signale über 4 TP gleichzeitig. Multiple eintreffende Signale müssen synchronisiert werden, damit die Rekombinierung erfolgreich durchgeführt werden kann. Der Empfänger kann kleinere Verzögerungen ausgleichen, aber wenn die Abstände zu groß werden, bricht die Kommunikation zusammen.

Der Delay Skew ist der Unterschied zwischen der Propagation Delay der langsamsten und der schnellsten Paare.

# Kapitel 10

## Auto-Negotiation

Die Vielseitigkeit des Ethernets erlaubt Benutzern unterschiedlichster Geschwindigkeiten und Technologien miteinander zu kommunizieren. Dies könnte leicht in einem Administrationsalptraum enden. Die Fähigkeit, bestimmte Features, Geschwindigkeiten und Mechanismen selbständig auszuhandeln ist die große Errungenschaft von Ethernet. *Auto-Negotiation* ermöglicht genau diese Vorgänge, ohne das Eingreifen des Benutzers oder eines Admins möglich.

Die Funktionen wurden erstmals Ende der 90er entwickelt. Sie erlauben den Kommunikationspartnern, selbständig die Verbindung zu initialisieren, zurückzusetzen, zu rekonfigurieren oder zu beenden. Die *Negotiation* (Verhandlung) erlaubt die Auswahl optimaler Parameter für den Betrieb.

Zur Zeit existieren 2 unterschiedliche Auto-Negotiation-Familien:

### **10Mbps, 100Mbps und 1000Mbps TP-Schnittstellen**

Hersteller entwickelten Schnittstellen mit unterschiedlichen Geschwindigkeiten, einige unterstützen nur eine, andere mehrere Geschwindigkeiten. Auto-Negotiations-Partner können untereinander die besten Parameter aushandeln.

### **1000Base-SX, 1000Base-LX und 1000Base-CX Schnittstellen**

Alle Mitglieder dieser Familie unterstützen ein einheitliches Auto-Negotiationsprotokoll, das nur verwendet werden kann, wenn die angeschlossenen Geräte die gleiche Schnittstelle besitzen. Beispielsweise muß sich 1000Base-SX mit 1000Base-SX verbinden damit das Protokoll verwendet werden kann.

## 10.1 Auto-Negotiation für TP-Schnittstellen

In der Tat haben einige Hersteller vor der Einführung eines Standardprotokolls einige proprietäre Protokolle implementiert, so daß eine Verbreitung derselben nicht möglich war. Erst mit den Bemühungen der 802.3-Arbeitsgruppe wurde ein Dokument (IEEE 802.3u) für die Spezifizierung des Auto-Negotiations-Protokolls, das auf dem Entwurf von *National Semiconductor's* namens *NWay Negotiation* basiert, entwickelt.

Der IEEE-Auto-Negotiation-Entwurf sah diese Funktion als optional für die Standards 10Base-T, 100Base-TX und 100Base-T4 vor. 1000Base-T und 1000Base-X erfordern diese Funktion zwingend.

Mit Hilfe der Auto-Negotiation (im weiteren Verlauf als *AN* bezeichnet) können folgende Parameter ermittelt werden:

- Die Übertragungsgeschwindigkeit
- Sind beide Partner Vollduplex-fähig?
- Soll Flußkontrolle in einer Richtung, beide Richtungen oder gar nicht für die Kommunikation verwendet werden?

Weiterhin können mit dem AN-Protokoll zusätzliche Informationen übermittelt werden:

- Berichten über fehlerhaftes Verhalten.
- Übertragen Hersteller-/Produkt-spezifische Informationen.
- Sind beide Partner Vollduplex-fähig?
- Soll Flußkontrolle in einer Richtung, beide Richtungen oder gar nicht für die Kommunikation verwendet werden?

Im Detail bedeutet dies, daß der Austausch und das Aufrüsten von Komponenten problemlos möglich ist und ein sofortiger Leistungszuwachs verbucht werden kann. Es ist immer wichtig, sich an den o.g. Regeln zu halten. Die Schnittstellen müssen das verwendete Protokoll unterstützen. Einen 10Mbps-Hub durch einen 10/100Mbps-Hub zu ersetzen, aber einen Client mit 10Mbps anzuschließen ist nicht sehr effektiv, da sich alle Mitglieder einer Collision-Domain auf die langsamste Verbindung einstellen. Anders sieht es mit einem 10/100Mbps-Switch aus. Er bringt einen vollen Leistungszuwachs, auch wenn weiterhin 10Mbps-Clients betrieben werden. Switches errichten zu Beginn der Kommunikation zwischen jedem Kommunikationspartner einen Pfad, der spezifische Parameter besitzt, unabhängig von den anderen Systemen.

### 10.1.1 AN-Funktionalität von TP-Schnittstellen

Die Prozedur der AN ist sehr einfach:

- Jede Partei sendet ihren Partnern eine „Checkliste“ mit allen unterstützten halb- und vollduplex-fähigen Betriebsmodi.
- Ein Satz Regeln wird auf der Basis der Checkliste erstellt.
- Das Ergebnis der Regelauswertung legt für beide Partner die Bedingungen der Kommunikation fest.

Tabelle 10.1 listet die Werte der aushandelbaren Betriebsmodi auf. Die Einträge sind nach geordnet nach dem besten (bevorzugten) und dem schlechtesten (einfachsten) Modus. So wird sichergestellt, daß der erstbeste Modus auch der optimale für beide Partner ist.

Es mag merkwürdig erscheinen, daß 100Base-T4 halbduplex vor 100Base-TX halbduplex rangiert. Der Grund dafür ist, daß 100Base-T4 auch mit Cat. 3 arbeiten kann, während 100Base-TX Cat. 5 erfordert. Wenn beide Modi ausgehandelt werden können, wird immer 100Base-T4 vorgezogen. Es wird die weniger anspruchsvolle Kategorie gewählt, da die Umgebung nicht immer auf dem neuesten Stand sein kann.

Technologie	unterstützte Kabel
1000Base-T vollduplex	4 Paar Cat. 5 UTP
1000Base-T halbduplex	4 Paar Cat. 5 UTP
100Base-T2 vollduplex*	Cat. 3 UTP
100Base-TX vollduplex	2 Paar Cat. 5 UTP
100Base-T2 halbduplex*	Cat. 3 UTP
100Base-T4 halbduplex	4 Paar Cat. 5 UTP
100Base-TX halbduplex	2 Paar Cat. 5 UTP
10Base-T vollduplex	2 Paar Cat. 3 UTP
10Base-T halbduplex	2 Paar Cat. 3 UTP

Tabelle 10.1: Vollduplex-Ethernet-Parameter (\* wurde nicht implementiert).

### 10.1.2 AN-Unterstützung für Flußkontrolle

Partner, die Vollduplexbetrieb ausgehandelt haben, überprüfen zusätzlich, ob Flußkontrolle aktiviert werden kann. Sie kann symmetrisch (beide Partner senden PAUSE) oder asymmetrisch (einer von beiden sendet PAUSE) sein. Die PAUSE-Funktion wird in *Full-Duplex-Kommunikation* beschrieben.

Zwei Bits werden für die Beschreibung von Flußkontrollnachrichten verwendet:

- Unterstützt keine Flußkontrolle: [00]
- Möchte PAUSE senden, aber nicht empfangen: [01]
- Möchte PAUSE senden und empfangen: [10]
- Möchte PAUSE senden und empfangen, oder nur empfangen: [11]

### 10.1.3 Ermitteln der Master/Slave-Timing-Regeln

1000Base-T-Kommunikation erfordert mind. einen Partner, der das Timing als Link Master (*timing master*) übernimmt. Der andere Partner arbeitet fortan als Slave (*timing slave*). Der Link Master verwendet die interne Systemuhr (*system clock*) für die Dauer der Transmission. Mit diesem Timing-Signal synchronisiert der Slave seinen Takt.

Die Wahl der Rollen basiert auf drei Regeln:

- Wenn der Admin die Partner manuell auf eine Stellung (Master oder Slave) eingestellt hat, wird ihr der Vorzug gegeben.
- Wenn beide Seiten nicht voreingestellt sind, wird ein sog. Multiport Device (Hub oder Switch) als Master arbeiten, wenn es mit einem Singleport Device (z.B. NIC) kommuniziert.
- Sollten beide Partner Multiport/Singleport oder nicht vorkonfiguriert sein, sendet jeder Endpunkt eine Zufallszahl (*random seed*) an den Partner. Die Partei mit dem höheren Wert, übernimmt den Master. Bei gleichen Werten, wird der Vorgang wiederholt.

### 10.1.4 Parallel Detection für TP-Kabel

Was passiert, wenn eine Seite AN unterstützt, aber die andere nicht. Eine Funktion namens *Parallel Detection* kann aushelfen. Sie ist Bestandteil der Auto-Negotiation-Spezifikation. Mit Hilfe einer Abtastung der eintreffenden Signale, kann ein Partner ermitteln, ob das Gegenüber mit 10Base-T, 100Base-T4 oder 100Base-TX arbeitet, wenn keine AN unterstützt wird.

Mit dieser Variante kann allerdings nur im Halbduplex-Modus gearbeitet werden, da es keine Möglichkeit gibt, zu ermitteln, ob Vollduplex-Kommunikation unterstützt wird. Der einzige Ausweg ist die manuelle Konfiguration beider Seiten auf Vollduplexbetrieb und das Deaktivieren der AN-Funktion.

### 10.1.5 Datenaustausch während der Auto-Negotiation über TP

Wie können Informationen über Twisted-Pair-Verbindungen ausgetauscht werden, wenn die Geschwindigkeit noch nicht bekannt ist? Und wie werden die Signale codiert?

Die Lösung bestand darin, eine *Low-Level-Implementierung* seit 10Base-T vorzunehmen, da 10Base-T kontinuierlich einen Link-Check (*link integrity pulse*) durchführt. Dieses Signal wird gesendet, sobald das Medium stumm ist. Der Link-Check wird oft auch als NLP (*normal link pulse*) bezeichnet. Er wird alle 16 ms durchgeführt.

Alte 10Base-T-Schnittstellen operieren korrekt, wenn ein Burst von Signalen gesendet wird und kein einzelner Pulse. Der Burst wird als *Fast Link Pulse* beschrieben. Alle TP-Geräte, die über AN verfügen senden FLPs während der Initialisierung aus. Die Konfigurationsparameter werden als FLP-codiert

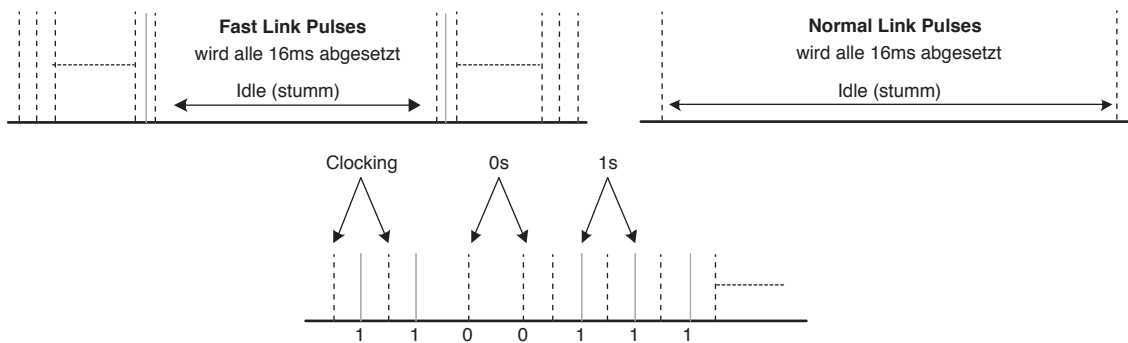


Abbildung 10.1: NLPs und FLPs

übertragen. Die Abbildung 10.1 illustriert diese Signale, die während der Auto-Negotiation übermittelt werden.

Jeder FLP enthält 33 Pulse und trägt eine 16-Bit-Nachricht. 17 ungerade Pulse übernehmen die Synchronisierung (*clocking*) und die übrigen 16 repräsentieren ein Datenbit:

Ein Pulse mit einer geraden Nummer ist 1. ein Pulse mit einer ungeraden Nummer ist 0.

Der untere Teil der Abb. zeigt das Bitmuster, daß mit 1100111 beginnt. Die Linien, die das Clocking repräsentieren sind gestrichelt, die Datensignale sind gurchgehend gezeichnet.

### 10.1.6 Base Page, Message Page und Unformatted Pages

Die erste 16-Bit-Nachricht wird als Base Page bezeichnet, manchmal auch als Base Link Code Word. Die weiteren Nachrichten bestehen aus folgenden Komponenten:

Eine einführende *Message Page* mit dem Code, der den Typ der Nachricht angibt. Abhängig vom Nachrichtentyp, enthält die Nachricht auch zusätzliche *Unformatted Pages* (unformattierte Seiten) mit bestimmten Formatierungen, die dem Nachrichtentyp entsprechen. Abbildung 10.2 zeigt das Layout der Message Base. Sie besteht aus diesen Komponenten:

- *5-Bit Selector Field* - Es enthält den Typ der Base Page. Ein Ethernet-Rahmen (802.3) hat den Code 00001.
- *Technology Ability Bits* - Dieses Feld enthält 8 Bits, die für die unterstützten Technologien verwendet werden.
- *ACK Bit* - Das Bit (ACK = Acknowledgment, Bestätigung) wird auf 1 gesetzt, wenn mind. 3 Kopien der Base Pages des Partners eingetroffen sind.
- *Remote Fault (RF) Bit* - Wird auf 1 gesetzt, wenn ein Fehler während der Verbindung auftritt.
- *Next Page (NP) Bit* - Wenn es auf 1 gesetzt wird, teilt der Sender mit, daß ein oder mehrere Seiten folgen werden.

Selector Field (5 Bits)	Technology Ability Bits A0 A1 A2 A3 A4 A5 A6 A7	RF Bit	ACK Bit	NP Bit
----------------------------	--	-----------	------------	-----------

Abbildung 10.2: Format einer TP Base Page.

Die übrigen Bits des *Base Page Technology Ability Fields* wird in Tabelle 10.2 beschrieben. Wird das Bit auf 1 gesetzt, ist der Adapter in der Lage, diese Technologie bereitzustellen. Ein Adapter teilt seinem Partner mit, welche Alternativen zur Verfügung stehen. Beispielsweise kann ein 10/100Mbps-Adapter

Bit	Technologie
A0	10Base-T
A1	10Base-T voll duplex
A2	100Base-TX
A3	100Base-TX voll duplex
A4	100Base-T4
A5	erstes PAUSE-Bit
A6	zweites PAUSE-Bit
A7	reserviert

Tabelle 10.2: Base Page Technology Ability Fields.

das Bit auf A0, A1, A2 und A3 auf 1 setzen und gibt zu verstehen, daß er vier verschiedene Modi von 10Base-T Halbduplex bis 100Base-TX Voll duplex unterstützt.

**Format der Message und Unformatted Pages:** Beachten Sie, daß 1000Base-T nicht in der Tabelle 10.2 nicht aufgelistet ist. 1000Base-T verhandelt über die zu verwendende Technologie über zusätzliche Messages, die aus eine Nachricht und zwei darauf folgenden Unformatted Pages.

Der obere Teil der Abb. 10.3 zeigt das allgemeine Layout einer Message. Die ersten 11 Bits eine Nachricht enthalten einen Message Code, die den Nachrichtentyp identifiziert. Die übrigen fünf Bits (von rechts nach links) haben folgende Bedeutung:

- *Next Page Bit* - Der Wert ist 1 wenn eine weitere Page (Seite) folgen wird.
- *Message Page Bit* - Wird auf 1 gesetzt, falls es sich um eine Message Page handelt.
- *ACK Bit* - Wird auf 1 gesetzt, wenn mind. 3 Kopien der Base Pages des Partners eingetroffen sind.
- *ACK2 Bit* - Wenn es auf 1 gesetzt ist, kann der Empfänger der Nachricht mit dem Nachrichtentyp umgehen und die Daten ordnungsgemäß verarbeiten, die sie erhalten hat.
- *Toggle (T) Bit* - Wechselt zwischen 0 und 1 bei übertragenen Seiten und stellt einen einfachen Zählmechanismus dar. Der Startwert ist das Gegenteil des A6-Bits der Message des Senders.

Der rechte Teil der Abb. 10.3 listet eine Unformatted Page auf. Die Inhalte der ersten 11 Bit sind abhängig vom vorangegangenen Message Code. Die letzten 5 Bits der Nachricht sind identisch mit denen der letzten Page Message. Das Message Page Flag ist für eine Unformatted Page immer 0.

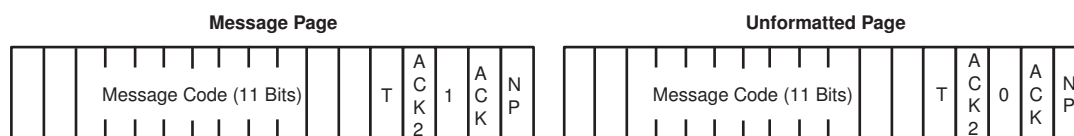


Abbildung 10.3: Format der Message Pages und Unformatted Page.

**Page Exchange Protocol** Das Protokoll erfordert, daß die Pages mehrmals gesendet werden, um sicher zustellen, daß keine Übertragungsfehler oder andere Einflüsse die Aushandlung behindern.

Die Schritte sind werden nun aufgelistet:

1. Beide Systeme senden ihre Base Pages immer wieder mit dem auf 0 gesetzten ACK Bit.
2. Nachdem mind. drei aufeinander folgende Base Page Messages empfangen wurden, wird das ACK Bit auf 1 gesetzt.
3. Wenn beide Partner setzen die NP Bits auf 1 in ihren Base Pages, werden weitere zusätzliche Nachrichten folgen.
4. Wenn die Kommunikation vollständig ist und beide Partner eine kompatible Konfiguration ausgehandelt haben, wird die Verbindung etabliert. Andernfalls bricht das Setup ab.

### 10.1.7 Message Pages für 1000Base-T

Tabelle 10.3 listet das Layout der 3 Pages, die über die Funktionalität der 1000Base-T-Technologie Auskunft geben. Ein Message Code von 1 gibt an, das 1000Base-T-Informationen folgen werden. Die U3 und U4 in der ersten Unformatted Page machen deutlich, daß das 1000Base-T-Interface im Halb- oder Vollduplexbetrieb arbeiten kann. Die Bits U0, U1 und U2 stellen fest, welche Rollen, Master oder Slave, eingenommen werden sollen.

Bit oder Feld	Beschreibung
<i>Message Page</i>	
11 Bits	Message Code = 8; zeigt 1000Base-Fähigkeiten an
<i>Unformatted Page 1</i>	
U0, U1, U2	Master/Slave Bits
U0	1 = Master/Slave manuell konfiguriert; 0 = Master/Slave nicht manuell konfiguriert
U1	1 = Master wenn U0=1; 0 = Slave wenn U0=0
U2	1 = Multiport wenn U=0; 0 = Singleport wenn U=1
U3	1000Base-T vollduplex
U4	1000Base-T halbduplex
U5-U10	Reserviert, auf 0 gesetzt
<i>Unformatted Page 2</i>	
U0-U10	Seed für Master/Slave-Negotiation

Tabelle 10.3: Felder der Technology Ability Fields

Tabelle 10.4 enthält eine Übersicht der Master/Slave-Zuordnungen.

U0 U1 U2 (A)	U0 U1 U2 (B)	Beschreibung
1 1 X	1 0 X	Partner A ist Master, B ist Slave, manuell konfiguriert
1 1 X	1 1 X	Nicht erlaubt, Konfiguration wird beendet, beide sind Master
1 0 X	1 0 X	Nicht erlaubt, Konfiguration wird beendet, beide sind Slave
0 X 1	0 X 0/1 0 X	A = Multiport, B = Singleport/Slave. A = Master, B = Slave
0 X 1	1 1 X	A und B sind Multiport, B manuell Master, A ist Slave
0 X 1	0 X 1	Partner A und B sind Multiport, Highest Seed = Master
0 X 0	0 X 0	Partner A und B sind Singleport, Highest Seed = Master

Tabelle 10.4: Master/Slave-Zuordnung

Die oberen Abschnitte beschreiben die Message Pages. Nun folgt ein Abhandlung der Message Codes für 1000Base-T Message Code 5 enthält einen OUI. Message Code 6 enthält den PHY-Layer Identifier. Dieser trägt Bits 3-24 der OUI (denken Sie daran, daß das 3-Byte-Prefix von der IEEE verwaltet wird).

## 10.2 Auto-Negotiation für 1000Base-X-Interfaces

Für 1000Base-T ist AN notwendig. Das schließt 1000Base-LX-, 1000Base-SX-Fibre und 1000Base-CX-Kupfer ein. Grundsätzlich ist es für physikalisch ungleiche Geräte miteinander zu kommunizieren, daher ist der Technologie-Typ nicht aushandelbar. Die aushandelbaren Fähigkeiten sind:

- Voll-/Halbduplex-Kommunikation
- Flußkontrolle für Vollduplexverbindungen. 2 Bits zeigen an, ob PAUSE verwendet werden soll.

Diese Nachrichten können auch Remote Fault Conditions und hersteller- oder produktspezifische Daten transportieren.

Code	Beschreibung
0	reserviert
1	Code für Null-Messages, die gesendet werden, wenn die Transmission beendet ist, aber der andere Partner noch Daten überträgt
2	reserviert
3	reserviert
4	Fehlerbeschreibungen ( <i>Remote Faults</i> ), eine Unformatted Page folgt, mit dem Fehlercode: <ol style="list-style-type: none"> <li>1. Testen der Remote Fault Funktion</li> <li>2. Link Loss</li> <li>3. Jabber (siehe <i>10Mbit PHY-Layer</i>)</li> <li>4. Parallel Detection Fault (siehe <i>Parallel Detection über TP</i>)</li> </ol>
5	OUI, 4 Unformatted Pages folgen
6	PHY-Layer Identifier, 4 Unformatted Pages folgen
7	100Base-T2 Technology Message, Fähigkeiten in 2 Unformatted Pages folgen
8	1000Base-T2 Technology Message, Fähigkeiten in 2 Unformatted Pages folgen

Tabelle 10.5: Message Codes.

### 10.2.1 Implementation der 1000Base-X Auto-Negotiation

Die 1000Base-X AN wird nicht in Form spezieller Pulse durchgeführt, so daß alle Daten in codierten Datenbits übertragen werden weil noch keine Pakete erzeugt werden können. Bei 1000Base-T werden die Daten zuerst in 10-Bit Codegruppen umgewandelt, die dann über das Medium transportiert werden.

Zusätzlich zu den Coderuppen existieren eine Menge anderer Muster, die für verschiedene Zwecke eingesetzt werden können, z.B. Idle-Symbole, End-of-Frame- und Start-of-Frame-Muster. 2 Spezielle Codegruppen markieren den Anfang einer Auto-Negotiation Page. Beim Eintreffen der Page werden die beiden Codegruppen erkannt und entfernt, so daß die folgenden beiden Codegruppen in 2 Datenbytes konvertiert werden können. Diese 2 Bytes bilden die Negotiation Page. Abb. 10.4 zeigt das Framelayout.

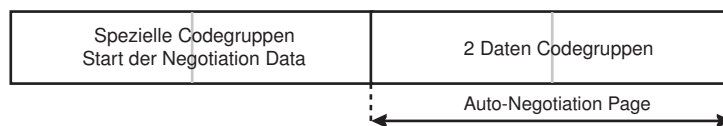


Abbildung 10.4: Eine Auto-Negotiation Page.

Abb. 10.5 zeigt das Format einer 1000Base-X Base Page. Wir können ihr folgende Informationen entnehmen:

- Die Geschwindigkeit muß 1000Mbps betragen, deshalb sind keine Bits zur Angabe der Technologie oder Geschwindigkeit vorhanden.
- Ein Paar von 1-Bit-Flags werden verwendet, um Halb- oder Vollduplexbetrieb anzuzeigen.
- Zwei PAUSE-Bits zeigen an, daß die PAUSE-Funktion unterstützt wird und ob sie symmetrisch oder asymmetrisch verwendet werden soll.
- 2 RF-Bits existieren, statt nur eines:
  - 00 = Link OK oder keine Fehlererkennung (RF)
  - 01 = Gerät geht offline 10 = Verbindungsfehler (*Link Failure*)
  - 11 = AN-Fehler, Kommunikation nicht möglich Flußkontrolle für Vollduplexverbindungen.

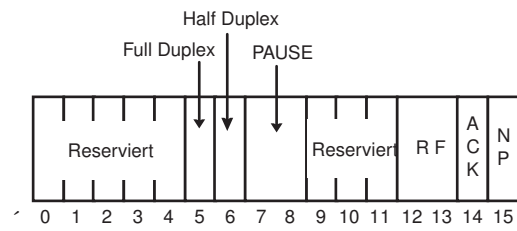


Abbildung 10.5: AN Base Message für 1000Base-X.

- 2 Bits zeigen an, ob PAUSE verwendet werden soll.

Die nachfolgenden Message Pages und Unformatted Pages werden nach dem gleichen Format, wie in Abbildung 10.5 gezeigt, verwendet. Diese Implementierung ist weitgehend abwärtskompatibel mit den Twisted-Pair-Varianten so weit möglich.

Teil II

# Routing und Switching



# Kapitel 11

## Bridges und 2nd Layer Switching

Im ersten Teil beschäftigten wir uns mit den technischen Grundlagen der Datenübertragung von Ethernet. Die Themenbereiche Switching und Routing sind bisher nur rudimentär besprochen worden. Nun wollen wir alle Aspekte dieser wichtigen Eigenschaften beleuchten.

Bridges sind seit der Einführung von Ethernet Bestandteil solcher LANs. Sie entwickelten sich bis heute zu so genannten *Layer 2 Switches*, obwohl sie technisch gesehen Bridges sind. Kunden sind eher von der Notwendigkeit eines neuen Produktes zu überzeugen, wenn sie durch einen modischen Begriff belegt sind. Außerdem vermittelt es den Eindruck, daß dieses Produkt eine Neuerung erfahren hat. Tatsächlich verfügen L2-Switches über mehr Ports als Bridges, sind schneller, bieten besseren Zugriff auf interne Informationen und Statistiken und unterstützen teilweise auch virtuelle Netzwerke. Dennoch handelt es sich noch immer um Bridges.

Im weiteren Verlauf sprechen wir von *Switches*, obwohl *Bridges* oder *Layer 2 Switches* gemeint sind.

### 11.1 Hauptfunktionen

Grundsätzlich lassen sich die Aufgaben eines Switches in drei Bereiche aufgliedern:

- Minimierung des Netzwerkverkehrs durch Begrenzung der Daten auf bestimmte Teile des LANs.
- Es ist nun möglich, LANs mit unterschiedlichen Geschwindigkeiten zu etablieren. Jedes Segment kann mit verschiedenen Geschwindigkeiten operieren.
- Switches erweitern die Collision Domains, da jeder eingehende Netzwerkrahmen erneuert wird, sobald er den Port eines Switches verläßt.

In frühen Jahren waren Bridges einfache Geräte, die problemlos in das bestehende Netzwerk integriert werden konnten. Schon bald forderten Administratoren und Integratoren andere Funktionalitäten, die besser auf die Bedürfnisse der Unternehmen abgestimmt sind. In diesem Kapitel werden wir auf die Kernfunktionen eingehen und Administratoren mit den Features eines vollständigen Layer 2 Switches vertraut machen.

### 11.2 Zusatzfunktionen

Diverse Zusatzfunktionen erhöhen die Geschwindigkeit, Sicherheit, Verfügbarkeit und Wartbarkeit der Switches. Besondere Fähigkeiten sind:

- Unterbindung der Vermittlung bestimmter Frames in andere Teile des Netzwerkes als Verbesserung der Sicherheit. Der Administrator kann Filterregeln zur Steuerung der Rahmen erstellen.

- Kritische Fehlerpunkte (*Single Points of Failure*) können durch Installation mehrerer redundanter Bridges ausgeschaltet werden, da im Fall der Fälle andere Routen zur Verfügung stehen. Mit Hilfe des *Spanning Tree Protocols* werden die verfügbaren Routen ermittelt und deren Status verwaltet.
- Verbindungen können zu Gruppen kombiniert werden und als Backup eingesetzt werden. Dieser Vorgang wird als *Link Aggregation* bezeichnet.
- Virtuelle LANs (VLANs) erlauben die Zusammenfassung von Benutzern des Netzwerkes nach dem Kommunikationsbedarf. Dabei wird die Art der auszutauschenden Informationen unterschieden und ist nicht auf den Standort derselben begrenzt.
- Prioritäten helfen Administratoren, Benutzern zeitkritische Daten zu liefern und hohen Anforderungen an das Netzwerk gerecht zu werden. Die *Priority Tags* sind ein VLAN-Feature und werden in Abschnitt (X – Oops! Noch nicht fertig.) besprochen.
- Statistiken und Einstellungen sind nun per SNMP abrufbar. Zwar ist das Protokoll Hardware-unabhängig, jedoch bieten längst nicht alle Geräte Zugriff auf die Datenbanken des Geräts.

Alle Hersteller versuchen stets Neuentwicklungen in ihre Produkte einfließen zu lassen. Oftmals zu Lasten existierender Standards mit der Folge, daß sich nicht alle Funktionen in jeder Umgebung einsetzen lassen. Um den Wildwuchs solch proprietärer Lösungen den Riegel vor zu schieben, hat das IEEE 802.1 Committee eine Reihe von Spezifikationen unter der Bezeichnung 802.3 veröffentlicht, die beispielsweise festlegen, wie die Aggregation von Links zu erfolgen hat, damit eine reibungslose Kommunikation zwischen verschiedenen Partnern möglich ist und die Investitionssicherheit der IT-Manager erhöht wird. Spezifikationen, die sich nur auf Bridges beziehen, wurden unter 802.1 veröffentlicht und in diversen Komitees entwickelt.

### 11.2.1 Collision Domains

Eine Station innerhalb eines Koax-Netzwerkes (BNC oder Ether-Tap) kann alle Rahmen, die das Medium passieren „sehen“. Befindet sich ein weiteres Segment durch einen Repeater in Reichweite, so ist auch dieses betroffen. Wandern zwei Rahmen gleichzeitig über das Medium, so kollidieren sie. Aus diesem Grund sprechen wir bei Mehrsegmentnetzwerken von *Collision Domains*.

Bridges teilen das Netzwerk im Gegensatz zu Repeatern in unabhängige Bereiche auf, damit lokaler Netzwerkverkehr die Collision Domain nicht verlassen kann. Wenn beispielsweise Arbeitsgruppen jeweils in ihrer eigenen Collision Domain untergebracht sind, wird das Datenaufkommen in allen anderen Domains minimiert, da kein Rahmen, der nicht für Stationen angrenzender Domains bestimmt ist, durch Bridges vermittelt wird. Die Vorteile liegen auf der Hand und machen sich durch höhere Bandbreite und schnelle Antwortzeiten bemerkbar.

### 11.2.2 Transparente Bridges

Alle Bridges eines Ethernet-LANs sind *transparent*. Rahmen, die für ein benachbartes Segment bestimmt sind, werden durch die Bridge automatisch basieren auf den Kenntnissen über die Netzwerktopologie (Bridges „lernen“ den Aufbau des Netzes im Laufe der Zeit kennen) weitergeleitet. Das jeweilige Ziel des Rahmens wird durch den Port bestimmt, an dem das Zielsegment angeschlossen ist. Auf diese Weise ist die Bridge in der Lage, zu entscheiden, an welchen Ports der Rahmen an das gewünschte Netzwerk ausgegeben werden muß. Betrachten wir dazu das Beispielnetzwerk in Abbildung 11.1.

Die Bridge lernt die Topologie des Netzwerkes auf folgende Art und Weise kennen:

- Alle Frames, die für Domain B bestimmt sind, werden über Port 1 vermittelt. Dabei wird jede MAC-Adresse der Quellrahmen aufgezeichnet.
- Ebenso kann die Bridge alle Rahmen sehen, die über Port 2 vermittelt werden und für Collision Domain C bestimmt sind. Jede Quelladresse der Rahmen wird in einer internen Tabelle gespeichert.
- Port 3 leitet alle Rahmen in die Collision Domain A und zeichnet auch hier die jeweiligen MAC-Adressen auf.

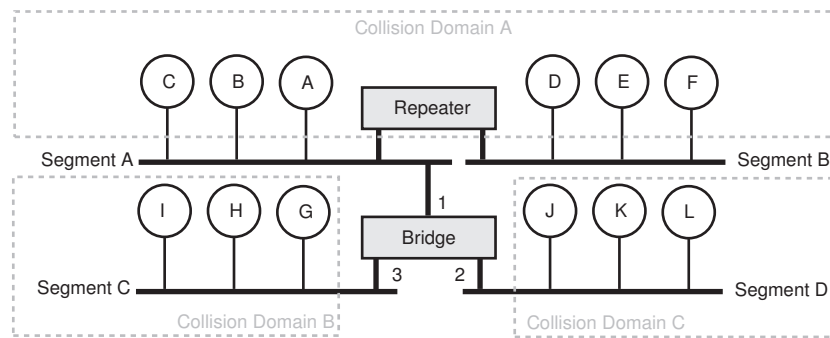


Abbildung 11.1: Ethernet LAN mit 4 Segmenten bestehend aus drei Collision Domains.

Angenommen, Station H (Collision Domain B) möchte Rahmen an Station D senden, erkennt die Bridge anhand der eintreffenden Pakete, welcher Port für die Weiterleitung verwendet werden muß. Das funktioniert nur, wenn die angeschlossenen Stationen zuvor ihre Anwesenheit in einer bestimmten Domain anzeigen.

Durch die Isolation des lokalen Netzwerkverkehrs ist es möglich, drei Frames gleichzeitig im Netzwerk zu übertragen: jeweils einer pro Domain.

### 11.2.2.1 Bridges in Twisted-Pair-Umgebungen

In Twisted-Pair-Umgebungen ist die Funktion der Bridges dieselbe. Abbildung 11.2 ist das gleiche Netzwerk wie zuvor abgebildet, jedoch unter Verwendung der Twisted-Pair-Technologie. Die beiden Hubs im oberen Teil der Zeichnung sind miteinander verbunden und bilden zusammen eine Collision Domain. Die Brücke teilt das Netzwerk in insgesamt drei Domains.

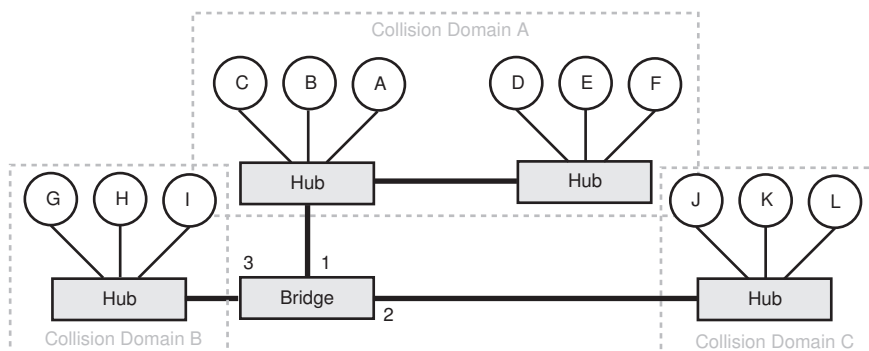


Abbildung 11.2: Ethernet LAN mit drei Collision Domains auf Twisted-Pair-Basis.

Mit Hilfe von Bridges kann ein Netzwerk in Domains mit unterschiedlichen Geschwindigkeiten aufgeteilt werden. Durch Layer-2-Switches erhält jede Workstation ihr eigenes Segment, welches mit jeweils einem Port des Switches verbunden ist. Arbeitet ein NIC einer Workstation im Halbduplexmodus so fungiert dieses Segment als eigenständige Collision Domain, da das CSMA/CD-Protokoll zu Anwendung kommt. Im Vollduplexmodus ist das nicht der Fall. Das Netzwerk aus Abbildung 11.3 basiert auf Layer-2-Switches. Jede Station kann, solange im Vollduplexmodus gearbeitet wird, senden und empfangen wann sie wollen.

Switches nutzen die verfügbare Bandbreite eines Netzwerkes besser aus, als Bridges, denn mehrere Rahmen können alle Segmente gleichzeitig durchwandern und von den Switches vermittelt werden. Besser noch im Vollduplexbetrieb: hier sind zwei Rahmen in jede Richtung erlaubt.

Aus Sicht eines Switches ist das LAN in zwei Komponenten aufgeteilt: die Station, welche durch ein TP-Kabel mit einem Switch-Port verbunden ist und der restliche Teil des LAN selbst. Der obere Teil des Netzwerkes aus Abbildung 11.3 ist durch einen Switch über Port 1 und Port 2 verbunden. In beiden Fällen sind die beiden sichtbaren Bestandteile des Netzwerkes die jeweils angeschlossene Workgroup. Aus

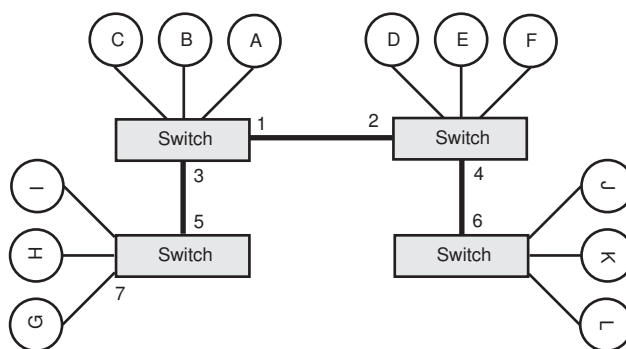


Abbildung 11.3: Twisted-Pair-Ethernet mit Switches.

Sicht des Switches ist Host G über Port 7 angeschlossen und stellt eine Komponente, der gesamte Rest des Netzwerks die zweite Komponente dar.

## 11.3 Interna

Die Hauptaufgabe einer Bridge liegt in der Vermittlung von Frames deren Quell- und Zieladressen auf unterschiedlichen Seiten der Bridges liegen. Aber, wenn das Ziel eines Frames über den Eingangsport erreicht werden kann, wird der Rahmen verworfen, andernfalls wird er vermittelt.

### 11.3.1 Lernen, Lernen und nochmals Lernen

Der beschriebene Mechanismus funktioniert nur, wenn die Bridge in der Lage ist, zu lernen, welche Station (mit einer eindeutigen MAC-Adresse) über welchen Port erreicht werden kann. Die Methode wird oft als *Lernen* bezeichnet und ist relativ einfach:

- Die Brücke beobachtet den gesamten Verkehr an allen Ports.
- Dabei wird die Quell-MAC-Adresse des Rahmens aufgezeichnet und welcher Port ihn empfangen hat.
- Alle Vorgänge werden aufgezeichnet und in eine Filterdatenbank eingetragen. Solche Einträge werden als dynamische Einträge bezeichnet.

Viele Hersteller haben andere Namen für die Filterdatenbank, wie z.B. Filtertabelle oder Weiterleitungstabelle (*forwarding table*). Die IEEE nennt sie Filterdatenbank, wobei manuell konfigurierte Einträge als statische und selbst erlernte als dynamische Einträge bezeichnet werden.

Das Vergleichen der Filterdatenbank mit eintreffenden Rahmen selbst ist ebenfalls schnell erklärt:

1. Wenn die Zieladresse eines Rahmens bereits in der Datenbank enthalten ist, das Ziel aber durch den eingehenden Port erreichbar ist, wird der Rahmen verworfen.
2. Ist das Ziel in der Filterdatenbank und der ausgehende Port unterschiedlich vom eingehenden, so wird der Rahmen über den in der Datenbank eingetragenen Port vermittelt.
3. Befindet sich die Zieladresse nicht in der Datenbank, so wird der Rahmen über alle Ports außer dem eintreffenden vermittelt.

Die genannten Punkte machen deutlich, wieso Bridges als *transparent* bezeichnet werden. Sie lassen sich einfach ins Netzwerk integrieren und lernen völlig eigenständig die Topologie kennen.

Alle erlernten Einträge werden in einer Filtertabelle abgelegt. Ein Beispiel ist in Tabelle 11.1 zu finden.

Ziel-MAC-Adresse	Ausgangsport	Status
00-17-06-2B-AE-32	2	gelernt
00-17-06-BD-7D-1A	3	gelernt
00-60-08-AE-27-1D	1	abgelaufen

Tabelle 11.1: Beispiel einer typischen Filtertabelle.

Die Tabelle läßt sich sehr einfach lesen. In der ersten Spalte stehen die MAC-Adressen, für die das analysierte Paket bestimmt war. In der zweiten Spalte ist der Port eingetragen, über den das betreffende Paket ausgeliefert wurde. Die letzte Spalte zeigt an, welcher Status mit dem jeweiligen Eintrag assoziiert ist. Solange kein Timeout abgelaufen ist, hat der Eintrag den Status gelernt.

Nach einer bestimmten Zeit werden Einträge als *abgelaufen* markiert. Das passiert immer, wenn die NIC mit der Zieladresse schon lange nicht mehr, in der Regel etwa 5 Minuten, adressiert worden ist (der Wert kann durch den Netzwerkadministrator verändert werden). Jedesmal wenn die Zieladresse von der Bridge erkannt wird, setzt sie sie den Timer für diesen Eintrag zurück. Das macht insgesamt auch Sinn: zum einen wird die Filtertabelle nicht unnötig groß, denn der Algorithmus durchsucht die Filtertabelle immer von oben nach unten. Zum anderen kann die NIC einer Station ausgetauscht werden und die Bridge ist in der Lage, den alten Eintrag selbständig zu entfernen und den neuen einzufügen.

### 11.3.2 Statische Filterinformationen

Obwohl Adminsitrator das Plug-and-Play transparenter Bridges lieben, erkannten Hersteller schnell die Forderungen nach mehr Kontrolle der Filtertabellen und ergänzten ihre Produkte um die Fähigkeit, den Netzwerkverkehr gezielt steuern zu können. Dazu setzten Administratoren *statische* Filtereinträge, die nicht, wie ihre dynamischen Varianten, ablaufen können.

Statische Filtereinträge sind für folgende Anwendungen vorgesehen:

- Um das Forwarding zum Server zu granularisieren.
- Um die Sicherheit zu erhöhen.
- Erhöhung des Datendurchsatzes, indem bestimmte Rahmen ausgegrenzt werden.

#### 11.3.2.1 Effizientes Forwarding mit statischen Filtereinträgen

Manchmal kann es von Vorteil sein, Filtereinträge wichtiger Server vor dem Ablaufen durch den internen Timeout zu bewahren. Mit statischen Einträgen ist das einfach: Sie erstellen einen Filtereintrag, der die MAC-Adresse des Servers auf den gewünschten Ausgabeport legt.

Beispielsweise könnten Sie einen Webserver mit der MAC-Adresse 00-60-08-AE-27-1D, der über Port 2 erreichbar ist als statischen Eintrag konfigurieren, so daß andere Segmente nicht mit Frames überflutet werden. So reduzieren Sie unnötigen Verkehr auf dem Medium.

#### 11.3.2.2 Beispiel einer statischen Filtertabelle

In Tabelle 11.2 finden Sie eine Beispieltabelle statischer Filtereinträge. Die Einträge sind nur illustrativ und können von Bridge zu Bridge (oder Switch zu Switch) voneinander abweichen. Eintrag eins ist dynamisch durch den Lernprozess entstanden. Das Gerät hat gelernt, daß die MAC-Adresse 00-60-08-AE-27-1D durch den Port 2 erreicht werden kann.

Als Grundlage der Tabelle dient das beispielhafte Netzwerk aus Abbildung 11.4. Zu Gunsten der Übersichtlichkeit ist Segment 1 über Port 1, Segment 2 über Port 2 und Segment 3 über Port 3 erreichbar.

Der zweite Tabelleneintrag ist statisch und regelt die Weiterleitung von Rahmen an den Server B. Alle Rahmen, die von einem anderen Port außer Port 1 stammen, aber an Server B adressiert sind, werden über Port 1 weitergeleitet (*forwarding*).

Nr	Zieladresse	Protokoll	Quellport	Zielport	Status
1	00-60-08-AE-27-1D	Alle	Jeder	2	gelernt
2	08-00-20-5A-01-6C	Alle	Jeder	1	statisch
3	01-00-83-23-BC-11	Alle	2	Keiner	statisch
4	01-00-83-23-BC-11	Alle	1	3	statisch
5	Jede	IPX	Jeder	Keiner	statisch

Tabelle 11.2: Beispielhafte statische Filtertabelle.

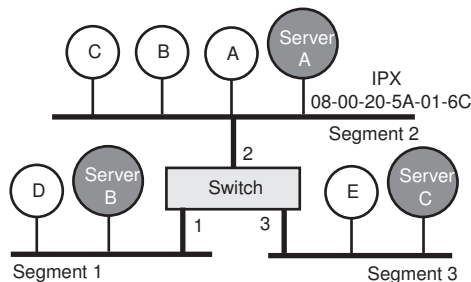


Abbildung 11.4: Beispielnetzwerk für statisches Forwarding.

Eintrag drei weist keinen Zielport auf, so daß Pakete, die Segment 2 verlassen wollen, die Bridge nicht passieren können. Der vierte Eintrag regelt beschränkt den Zugriff auf die Segmente 2 und drei. Der letzte Eintrag ist sorgt dafür, daß IPX nicht auf andere Segmente weitergeleitet wird.

Jeder Eintrag in der Tabelle identifiziert nur einen ausgehenden Port. Es ist jedoch möglich, und oft erwünscht, mehrere Ports zu spezifizieren. Beispielsweise müßten Multicast-Rahmen über mehrere Ausgangsport vermittelt werden.

Der Nachteil in der Verwendung von statischen Einträgen ist offensichtlich. Administratoren müssen die Einträge sorgfältig verwalten, denn jedesmal, wenn ein Host hinzugefügt oder entfernt wird, muß die Tabelle aktualisiert werden. Das gleiche gilt natürlich auch, wenn eine NIC ausgetauscht werden muß.

Wenn es einen statischen Eintrag für eine MAC-Adresse gibt, wird anschließend keine dynamische Information für die betreffende Adresse in die Tabelle aufgenommen. Wird also ein System an einen anderen Standort verbracht, so wird keine dynamische Information für das System am neuen Standort ermittelt.

### 11.3.2.3 Sicherheitsaspekte

Administratoren verspüren oft das Verlangen, den Rahmenfluß zwischen verschiedenen Bridges zu kontrollieren, um Sicherheitsmechanismen bereits auf Layer 2 zu implementieren. So könnten sie den Zugriff auf eine bestimmte Ziel-MAC-Adresse über den Eingangsport an dem die Pakete eintreffen beschränken.

Wie der Verkehr gezielt auf Segmente eingeschränkt oder zugelassen werden kann, zeigten die Einträge aus Tabelle 11.2. Durch den geschickten Aufbau einer effizienten Filtertabelle, kann der Administrator sehr einfach, ausgefeilte Sicherheitsrichtlinien implementieren.

### 11.3.2.4 Leistungsverbesserung durch Protokollfilter

Ein äußerst nützliches Feature hat bereits kurze Zeit nach Einführung der Layer-2-Bridges Einzug gehalten. Sie erlauben neben der Analyse der Eingangs- und Ausgangsport auf Basis der MAC-Adressen, die Analyse des Protokollfeldes auf Layer 3 des Rahmens. Beispielsweise haben wir in der Filtertabelle 11.2 das IPX-Protokoll aktiv von der Weiterleitung auf andere Segmente ausgeschlossen, da kein anderer Knoten im Netzwerk Zugriff auf dieses Protokoll benötigt. Der übrige Netzwerkverkehr findet über TCP/IP statt, sodaß es keine Veranlassung der Weiterleitung von anderen Layer-3-Protokollen in andere Segmente besteht.

Sollte ein IPX-Rahmen auf Port 2 eintreffen und dessen MAC-Adresse nicht auf dem lokalen Segment bekannt sein, so wird er nicht über Ports 1 und 2 ausgeliefert und stillschweigend verworfen. IP-Rahmen jedoch, werden ohne weiteres weitergeleitet.

Ein weiterer Nebeneffekt ist, daß IPX-Verkehr aus anderen Segmenten niemals die Bridge passieren werden, so daß die Beschränkung des Protokolls auch in die Gegenrichtung aufrecht erhalten wird.

## 11.4 Architektur von Layer-2-Switches

Bevor wir mit der Besprechung Architektur von Layer-2-Switches beginnen, wollen wir uns die Frage stellen, was Hersteller von Bridges dazu veranlaßte, eine neue Gerätegeneration zu entwickeln und einen eigenen Namen für sie zu entwickeln:

- Lokale Netzwerke wurden immer größer
- Die Bandbreite in den LANs nahm drastisch zu.
- Hersteller (oder eher ihre Marketingabteilungen) sind mit der Zeit schlauer geworden.
- Die Hardware ist leistungsfähiger und gleichzeitig preiswerter geworden.

Die wohl wichtigste Innovation ist die Einführung von ASICs (*Application Specific Integrated Circuits*). Sie tragen gewissermaßen die Softwarelogik wie sie von Anwendungsprogrammen her bekannt ist in sich. Somit wird die Ausführung der Programme beschleunigt, da die Hardwareabstraktionsschicht entfällt.

Das ist eigentlich keine neue Idee. Bereits seit einigen Jahren sind spezialisierte Chips für die Ausführung rechenintensiver Aufgaben im Einsatz, insbesondere für kryptographische Anwendungen. Neu ist nun, daß die Herstellung und Programmierung solcher Chips deutlich vereinfacht wurde und preiswerter ist. Aus diesem Grund drängt sich die Integration in Switches geradezu auf.

### 11.4.1 Store-and-Forward und Cut-Through

Der Standard IEEE 802.1D schreibt vor, daß die Verarbeitung von Rahmen nach dem *Store-and-Forward*-Prinzip geschehen sollte. Das Verfahren ist folgendermaßen spezifiziert:

1. Warte bis ein vollständiger Rahmen empfangen wurde.
2. Prüfe das FCS-Feld und verwerfe den Rahmen, wenn Sie nicht intakt ist.
3. Verwerfe Rahmenfragmente, die durch fehlerhafte Übertragungen und Kollisionen aufgetreten sind.
4. Ist der Rahmen gültig, prüfe die Filtertabelle und, falls notwendig, leite den Rahmen über einen oder mehrere Ports weiter.

Dieser Operationsmodi ermöglicht die Auslöschung von Netzwerkfehlern, anstatt sie weiterzuleiten. Allerdings wird die Auslieferungsgeschwindigkeit von Rahmen verringert. Viele Hersteller bieten oftmals beide der folgenden Optionen der Fast-Forwarding- und *Cut-Through*-Technik:

- Die Rahmenübertragung wird vorgenommen, sobald die Ziel-MAC-Adresse ermittelt und mit der Filtertabelle verglichen wurde. Dadurch werden die Bits am Anfang des Rahmens bereits auf dem Medium sein, bevor der Rest vollständig empfangen wurde. Das bedeutet natürlich auch, das fehlerhafte Rahmen weitergeleitet werden bevor durch Analyse der FCS die Integrität des Rahmens bestätigt werden könnte.
- Alternativ kann die Rahmenübertragung beginnen, sobald die ersten 64 Bit empfangen wurden. Somit wird das Risiko, fehlerhafte Rahmen weiterzuleiten, minimiert, da bereits eine geringe Zahl von Bits genügen kann, um Fehler zu entdecken.

Einige Produkte vereinen Verfahren aus den besten Eigenschaften der beiden Techniken. Wird eine zuvor definierte Fehlerrate erreicht, so wechselt der Switch von Cut-Through auf Store-and-Forward.

### 11.4.2 Parellele Verarbeitung durch Asics

Unabhängig davon, ob Cut-Through oder Store-and-Forward eingesetzt wird, muß zuerst immer die Filtertabelle befragt werden, bevor entschieden werden kann, was mit dem Rahmen geschehen soll. Mit wachsender Größe von LANs sind die Switches schnell überfordert. Da kommen nun die ASICs ins Spiel.

Normalerweise enthält jede Switch-Karte (genauer gesagt jede *Switch Line Card*) eines Layer-2-Switches einen Satz von ASICs, die eintreffende Rahmen verarbeiten. Obwohl der Zugriff auf die Filtertabelle durch einen zentralen Prozessor geregelt wird, befinden sich stets aktuelle Kopien der Tabelle in den Speicherbereichen der ASICs. Dadurch sind parallele Abfragen der Filtertabellen durch die ASICs möglich. Der Leistungszuwachs dieser Arbeitsweise ist ganz erheblich.

## 11.5 Redundanz im LAN

Der Ausfall einer Verbindung oder eines Switches kann von einer leichten Unbequemlichkeit bis zum totalen Ausfall des Netzwerkes reichen und die Produktivität einer Abteilung oder eine Unternehmens empfindlich beeinflussen. Viele Administratoren richten wichtige Elemente meist redundant aus, um einen zentralen Fehlerpunkt (*Single Point of Failure*) auszuschalten.

Das *Spanning Tree Protocol* ein allgemein anerkannter Standard zur Etablierung von Redundanz in LANs. Es erlaubt Switches, alternative Verbindungen für die Weiteleitung oder den Empfang von Rahmen bei Ausfall einer anderen zu nutzen.

Ein zweiter Mechanismus wird als *Link-Aggregation* bezeichnet. Dabei werden zwei oder mehr physikalische Verbindungen zu einem logischen Link zusammengefaßt. Fällt eine Verbindung aus, wird die Übertragung nicht unterbrochen, sondern die andere weiter genutzt. Das geschieht für den Benutzer völlig transparent.

### 11.5.1 Spanning Tree Protocol

Transparente Brücken und redundante Pfade sind nicht miteinander vereinbar. Warum das so ist, und warum das Spanning Tree Protocol so wichtig ist, möchte ich an der Abbildung 11.5 illustrieren. Darin sind zwei LANs, das linke als Twisted Pair und das rechte als Bus-Ausführung, abgebildet. Beide verfügen über je zwei Bridges zur redundanten Anbindung der Collision Domains A und B.

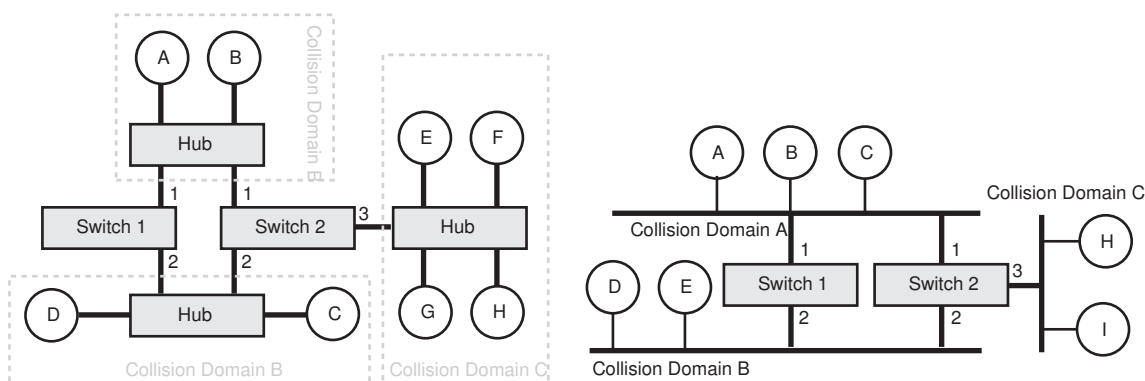


Abbildung 11.5: Redundantes Switching in LANs

Wenn Station A in Collision Domain A einen Rahmen, der an Station C in Collision Domain B adressiert ist, absetzt, leiten beide Bridges diesen Rahmen weiter. Wenn Switch 2 seine Kopie vermittelt, beobachtet Port 2 von Switch 1 (denken Sie daran, das Hubs die Rahmen über alle Ports weiterleiten) den Rahmen, liest seine Zieladresse und schließt daraus, das Station A sowohl in Collision Domain A als auch in Collision Domain B vorhanden ist. Das gleiche passiert auch mit Bridge 2, wenn Port 2 des Switches die Kopie des Rahmens erhält. Es ist nur eine Frage der Zeit, bis das Netzwerk durch diese Verwirrung zusammenbricht.

Tatsächlich darf zwischen zwei Stationen innerhalb eines Ethernet-LANs immer nur ein einziger Pfad existieren. Zwei Pfade erzeugen eine Schleife. Ein LAN, welches mit transparenten Brücken aufgebaut wurde, darf keine Schleifen aufweisen und wird daher als Baumstruktur organisiert. Glücklicherweise liefert uns das Spanning Tree Protocol einen Baum und bietet gleichzeitig Redundanz.

Mit Hilfe des Spanning Tree Protocols wird ein LAN, das physikalisch aus mehreren Schleifen bestehen kann, in ein logisch strukturiertes LAN auf Basis eines Baumes umgewandelt. Dabei werden zwischen den Bridges *Bridge Protocol Data Units* ausgetauscht, die ihnen erlauben, sich über einen bestimmten Pfad zu einigen und bei Bedarf auf einen anderen Pfad umzuschalten. Die genaue Funktionsweise des Spanning Tree Protocols wird in Kapitel 12 *Das Spanning Tree Protocol* erläutert.

### 11.5.2 Link Aggregation

Ein interessantes Feature, daß von vielen, aber nicht allen Herstellern, unterstützt wird, ist die Fähigkeit der Switches, mehrere physikalisch Verbindungen zu einem logischen Link zusammenzufassen. Dieser Vorgang wird als *Link Aggregation* bezeichnet. Viele Hersteller haben dafür andere Begrifflichkeiten eingeführt, darunter *Inverse Multiplexing*, *Port Aggregation*, oder *Trunking*. Sollten Ihnen diese Begriffe in der Literatur begegnen, so ist in jedem Fall *Link Aggregation* gemeint.

Das Prinzip ist einfach. Es werden mehrere Ports eines Switches zu einer Verbindung zusammengefaßt, wodurch sich der Durchsatz jeweils um die Anzahl der eingesetzten Ports vervielfacht. Würden beispielsweise drei 100 Mbps Ports zu einem Link zusammengefaßt, erhöht sich die Bandbreite auf 300 Mbps. Logischerweise muß auch die Gegenseite über die Aggregation der Links informiert werden. Zwischen zwei Switches funktioniert das ohne weitere Eingriffe des Administrators, sind jedoch Hosts im Spiel so muß das Betriebssystem diesen Modus unterstützen. Dabei werden zwei oder mehr NICs durch einen Mechanismus, der gemeinhin als *Channel Bonding* bekannt ist, zusammengefaßt.

## 11.6 Multicast-Verkehr in LANs

Der Lernprozess von Switches erhöht indirekt auch die Leistungsfähigkeit des Netzwerkes: ein Unicast-Rahmen, dessen Ziel durch die Weiterleitung dynamisch erlernt wurde, wird nur durch den Port, der ihn „gesehen“ hat weitergeleitet; er wird also nicht in andere Teile des LANs weitergeleitet. Der Administrator kann statische Informationen zur Reduzierung unnötigen Unicast-Verkehrs einsetzen.

Den Fluß von Multicast-Rahmen zu kontrollieren, ist eine weit größere Herausforderung und erfordert den Einsatz ergänzender Protokolle und Mechanismen. Der ursprünglichen Regel von Multicast folgend, werden alle Stationen die Multicast-Rahmen sehen, wenn mindestens eine Station einer Multicast-Gruppe beitrifft. Mit dem zunehmenden Aufkommen von Multicast-Anwendungen werden die Fortschritte neuerer Switches leider wieder zu nichte gemacht.

### 11.6.1 IGMP Snooping und GARP

Um der soeben genannten Problematik entgegenzutreten, haben die Hersteller einen Mechanismus namens *IGMP Snooping* entwickelt, der die Weiterleitung von Multicast-Rahmen an die Multicast-Gruppen kanalisiert. Das IGMP-Protokol (*Internet Group Message Protocol*) ist Teil des TCP/IP-Protokollstapels und wird für die Organisation von Multicast-Gruppen verwendet. Es ermöglicht Systemen, einer Gruppe beizutreten oder eine zu verlassen. Bridges hören den IGMP-Verkehr ab, um Tabellen aufzubauen, die alle Systeme einer Multicast-Gruppe identifizieren.

IGMP-Snooping ist alles andere als Perfekt. Aus diesem Grund wurde von der IEEE-Arbeitsgruppe 802.1 eine längerfristige Lösung namens *GARP Multicast Registration Protocol* entwickelt. Mit Hilfe von GMRP registrieren sich Systeme, die einer Multicast-Gruppe beitreten oder diese verlassen möchten bei einer benachbarten Bridge. Sie tauschen anschließend die Informationen mit anderen Bridges aus, so daß Multicast-Rahmen nur an Mitglieder einer Gruppe ausgeliefert werden.

Eine genaue Abhandlung über die Funktionsweise von IGMP Snooping und GMRP finden Sie in Kapitel 13 „Switches und Multicast-Traffic“.

### 11.6.2 Funktionale Struktur eines Layer-2-Switches

Ursprünglich wurden Switches mit der Aufgabe, Pakete entweder weiterzuleiten oder zu verwerfen, entworfen. Sie generieren keine Pakete oder sind das Ziel von ihnen. Mit der Einführung des Spanning Tree Protocols, Link Aggregation, GMRP und SNMP hat sich das grundlegend geändert. Switches tauschen nun Informationen mit anderen Bridges aus und nehmen grundlegende Einstellungen automatisch vor.

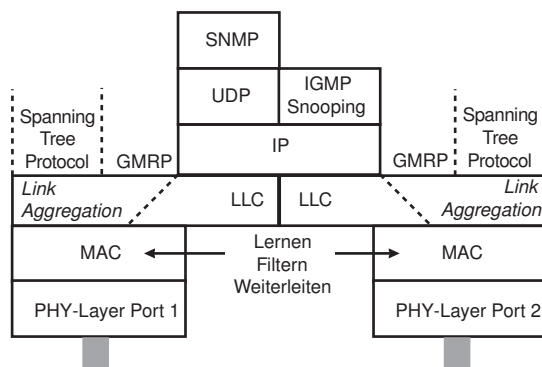


Abbildung 11.6: Schichten eines Layer-2-Switches

Um die Struktur eines Layer-2-Switches zu verdeutlichen, betrachten wir Abbildung 11.6. Jeder Port hat seine eigene MAC-Adresse, so daß der Port als Quelle und Ziel eines Rahmens identifiziert werden kann.

Die Link Aggregation läuft direkt oberhalb des MAC-Layers ab. Die Rahmen, welche für das Spanning Tree Protocol und GMRP benötigt werden, enthalten auch LLC-Header, da Sie auf Layer 3 operieren.

Im Gegensatz dazu findet das IGMP Snooping oberhalb von IP statt und SNMP selbst ist UDP-basiert. Die jeweiligen Protokolldetails werden im nächsten Kapitel erläutert.

## Kapitel 12

# Das Spanning Tree Protocol

Mit Hilfe des Spanning Tree Protocols (STP) tauschen benachbarte Switches spezielle Nachrichten (Rahmen), sogenannte *Bridge Protocol Data Units* (BPDUs), aus. Jede Bridge verfügt über einen eindeutigen Bezeichner und einer Priorität. In einem vollständig mit Bridges ausgestatteten LAN wird zufällig eine Bridge als Ursprung (*Root*, Wurzel) für den Aufbau des Baums ausgewählt. Dieser Bridge wird der niedrigste Bezeichner und die höchste Priorität zugeordnet. Diese Daten werden in regelmäßigen Abständen verifiziert.

Nachdem die Root-Bridge ausgewählt wurde, wird der Preis<sup>1</sup>(engl. *cost*) für den Pfad von der Root-Bridge zu allen andere Bridge-Ports ermittelt. Dieser Wert wird als *Root Path Cost* bezeichnet und ist der geringste *Cost Path* dem ein Rahmen folgt, der von der Root-Bridge zum Bestimmungsport übermittelt wird. Jede Bridge ermittelt anschließend welcher ihrer Ports den geringsten Path Cost zur Wurzel ist. Dieser Port wird als *Root Port* bezeichnet, denn ab diesem Zeitpunkt werden alle BPDUs über diesen Port empfangen.

Der Path Cost, der einem Port zugeordnet ist, wird durch die Bitrate des Segments bestimmt (engl. *Designated Cost*). Je höher die Bitrate ist, desto geringer der Designated Cost. Gibt es beispielsweise zwei alternative Pfade von der Wurzel, einer mit 100 Mbps, der andere mit 10 Mbps, so weist der Link mit 100 Mbps einen geringeren Path Cost als der mit 10 Mbps auf. Sind die Bitraten der beiden Links gleich, so wird der Bezeichner der Bridge als Entscheidungshilfe herangezogen.

Nachdem der Root Port für jede Bridge ermittelt wurde, wird ein Bridge-Port für die Weiterleitung von Rahmen auf jedes Segment ausgewählt. Diese Bridge wird als *Designated Bridge* bezeichnet. Die Auswahl basiert auf der geringsten Path Cost zur Root Bridge über das Segment, welches gerade untersucht wird. Haben die beiden Bridge Ports des Segments den gleichen Path Cost, wird die Bridge mit dem kleineren Bezeichner ausgewählt. Der Bridge Port, der das Segment mit seiner Designated Bridge verbindet, wird als *Designated Port* bezeichnet. Die Root-Bridge ist immer die Designated Bridge für alle Segmente, die mit ihr verbunden sind. Aus diesem Grund sind alle Ports der Root-Bridge auch gleichzeitig Designated Ports.

Wenn der Designated Port für ein bestimmtes Segment ermittelt wurde und dieser als Root Port fungiert, so wird er nicht mehr in die Auswertungsprozedur zur Ermittlung eines Designated Ports einbezogen. Somit findet die Auswahl des Designated Ports eines Segments nur unter den Ports statt, die nicht Root Ports sind.

Nachdem nun die Root-Bridge, die Wurzel und die Designated Ports ermittelt wurden, kann der Zustand aller Bridge-Ports entweder auf *Forwarding* oder *Blocking* gesetzt werden. Die Ports der Root-Bridge befinden sich zu Beginn stets im Zustand Forwarding, da sie alle Designated Ports sind. Bei allen anderen Bridges werden nur die Root Ports und Designated Ports auf Forwarding gesetzt; alle anderen sind im Zustand Blocking. Als Resultat sind alle Bridges in einem LAN baumartig strukturiert.

---

<sup>1</sup>In Fachkreisen wird stets der englische Begriff „Cost“ verwendet. Der Einfachheit halber werde ich dieser Gewohnheit folgen und ebenfalls von Cost sprechen.

## 12.1 Initialisierung der Topologie

Alle Bridges eines LANs verfügen über eine eindeutige MAC-Gruppenadresse, die für den Versand der BPDUs zwischen Bridges von Bedeutung ist. Die empfangenen BPDUs werden nicht unmittelbar weitergeleitet. Stattdessen können die enthaltenen Informationen von der *Bridge Protocol Entity* dazu verwendet werden, eigene BPDUs zu erzeugen, die über die Ports der Bridge weitergeleitet werden.

Wenn eine Bridge das erste mal in Betrieb genommen wird, geht sie davon aus, die Root Bridge zu sein. Sie initiiert dann die Übertragung einer *Configuration BPDU* über alle ihre Ports (und damit auch über die angebotenen Segmente) in regelmäßigen Intervallen, die als *Hello Time* bezeichnet werden.

Jede Configuration BPDU enthält einige Felder, mit folgenden Inhalten:

- Der Bezeichner (Root-Bezeichner) der Bridge, die glaubt die Wurzel zu sein und die BPDU abgesetzt hat (zu Beginn die Bridge selbst).
- Der Path Cost zur Wurzel von dem Bridge Port, über das die BPDU empfangen wurde (zu Beginn immer null).
- Der Bezeichner der Bridge, die die BPDU abgesetzt hat.
- Der Bezeichner des Bridge Ports über den die BPDU abgesetzt wurde.

Beim Empfang einer Configuration BPDU, kann jede Bridge, die an die Segmente angeschlossen sind, die eine Configuration BPDU empfangen haben, feststellen, ob die Bridge eine höhere, gleiche oder niedrigere Priorität aufweist, in dem der empfangene Bezeichner mit dem eigenen Bezeichner verglichen wird. Ist der eigene Bezeichner kleiner, so wird dieser BPDU-Rahmen verworfen.

Wenn der Root-Bezeichner der empfangenen Configuration BPDU anzeigt, daß die Bridge nicht die Wurzel ist, macht die Bridge einfach weiter, indem der Cost des Ports, über den die BPDU empfangen wurde der bereits übermittelten hinzugefügt wird. Die Bridge kennt die Designated Costs der angeschlossenen Segmente aus den anderen Verwaltungsnachrichten, die sie empfangen hat. Sie erzeugt anschließend eine eigene Configuration BPDU mit diesen Informationen und ihren eigenen Bezeichnern und leitet eine Kopie über alle Ports and alle benachbarten Switches weiter. Diese Prozedur wird von allen Bridges im LAN durchgeführt. Auf diese Weise bewegen sich die BPDUs weg von der Wurzel in die weiter entfernten Teile des LANs.

Während die Configuration BPDUs von der Wurzel in die Äste des Netzwerks transportiert werden, wird der Path Cost für jeden Port alle Switches im Netzwerk berechnet. Deshalb können alle Bridges ihre Root Ports identifizieren und, da an jedem Punkt der Hierarchie zwei oder mehr Bridges mit dem gleichen Segment verbunden sein können und somit Configuration BPDUs austauschen, sind sie in der Lage mit Hilfe der Root Path Cost aus der BPDU-Nachricht, ihre Designated Bridge für das betreffende Segment zu ermitteln. Bei gleichem Path Cost wird erneut der Bezeichner der Bridge zur Entscheidung herangezogen.

Aus diesen Ausführungen können wir folgende Annahmen herleiten:

- Eine Bridge empfängt BPDUs über ihren Root Port und übermittelt sie über ihre Designated Ports.
- Alle Root Ports und Designated Ports befinden sich im Zustand Forwarding.
- Eine Bridge, die ein Segment aufweist, daß mit einem Root Port verbunden ist, kann nicht gleichzeitig der Designated Port für dieses Segment sein.
- Für jedes Segment steht nur ein Designated Port zur Verfügung.

## 12.2 Änderungen der Topologie

Wie bereits erwähnt werden redundante Bridges zur Erhöhung der allgemeinen Betriebssicherheit eines LANs integriert. Durch die soeben beschriebene Prozedur befinden sich einige Bridges im Zustand

Blocking, so daß es einen Mechanismus geben muß um den Zustand zu ändern, falls eine Bridge ausfällt und die andere in Betrieb genommen werden muß. Diese Prozedur wird als *Topology Change Procedure* bezeichnet.

Nachdem eine Root Bridge und die mit ihr verbundene Topologie ermittelt wurde, gehen Configuration BPDUs nur von der Wurzel aus. Sie werden in regelmäßigen Abständen (also nach Ablauf des Hello Timers) über alle Ports ausgesandt. Mit Hilfe dieser Informationen aktualisiert jede empfangende Bridge ihre Einstellungen auf Basis der Configuration BPDUs und bestätigt im Gegenzug die Zustände eines jeden Ports.

Um nun den Ausfall einer Bridge zu erkennen wird ein sogenannter *Message Age Timer* von jeder Bridge für alle ihre Ports verwaltet. Funktioniert das Gerät fehlerfrei so wird der Timer stets beim Empfang einer Configuration BPDUs zurückgesetzt. Wenn eine Designated Bridge oder ein aktiver Bridge Port ausfällt, kann die Configuration BPDUs nicht zugestellt werden, so daß der Timer an jener Bridges abläuft, die sich im Downstream der ausgefallenen Bridge oder eines Ports befinden.

Die Bridge Protocol Entity reagiert auf das Ablaufenden des Message Age Timers mit der *Become Designated Port Procedure*. Sie wird von allen betroffenen Bridges durchgeführt. Nach Abschluß der Prozedur wurden mindestens ein oder mehr neue Designated Ports etabliert. Sollte die Root Bridge ausfallen, so wird eine neue Root Bridge ausgewählt.

Immer wenn ein Port vom Zustand Blocking in den Zustand Forwarding wechselt, wird eine *Topology Change Notification BPDUs* durch den neuen Port in Richtung Root Bridge abgesetzt. Alle Designated Bridges zwischen dieser Bridge und der Wurzel verarbeiten die Nachricht und leiten sie über ihre Root Ports weiter an alle benachbarten Bridges. Dadurch werden alle Bridges eines Netzwerks über die Topologieänderung informiert. Um sicher zu stellen, daß die Wurzel in jedem Fall über diese Änderung informiert wird, ist mit der Topology Change Notification BPDUs ein Timer und eine Bestätigung assoziiert.

Nach der Topologieänderung können die Stationen an allen Segmenten über einen anderen Port erreicht werden als der akutelle als Forwarding markierte Port der internen Bridge-Tabelle. Da der Timer für den Ablauf der Einträge in der Switch-Datenbank relativ lang ist, wird bereits eine neue Configuration BPDUs als Reaktion auf die Topologieänderung empfangen. Dieses Paket enthält ein Bit, das die Bridges anweist, die Einträge in der Datenbank zu aktualisieren. Damit werden neue Einträge erstellt, sobald erneut Rahmen übertragen werden.

## 12.3 Portstatus

Damit keine Schleifen bei der Ermittlung der LAN-Topologie entstehen, ist den Bridges nicht erlaubt, direkt von dem Zustand Blocking in den Zustand Forwarding zu wechseln. Stattdessen sind zwei zusätzliche Zustände definiert: *Listening* und *Learning*. Ein letzter Zustand *Disabled* ermöglicht Netzwerkadministratoren, eine spezielle BPDUs abzusetzen, um einzelne Ports dauerhaft zu sperren.

Abhängig von dem jeweiligen Zustand können BPDUs weitergeleitet werden oder nicht. Folgende Regeln sind für die Weiterleitung in bestimmten Zuständen definiert:

- **Disabled:** Nur Management BPDUs können empfangen und verarbeitet werden.
- **Blocking:** Nur Configuration BPDUs und Management BPDUs werden empfangen und verarbeitet.
- **Learning:** Alle BPDUs werden empfangen und verarbeitet. Informationsrahmen werden empfangen aber nicht weitergeleitet.
- **Forwarding:** Alle BPDUs werden empfangen und verarbeitet. Alle Informationsrahmen werden empfangen, verarbeitet und weitergeleitet.

## 12.4 Zustandsänderungen

Zustandsänderungen werden von der Bridge Protocol Entity initiiert und sind das Resultat der Verarbeitung einer Port-bezogenen BPDUs oder eines abgelaufenen Timers. Mögliche Zustandsänderungen werden in Abbildung 12.1 dargestellt.

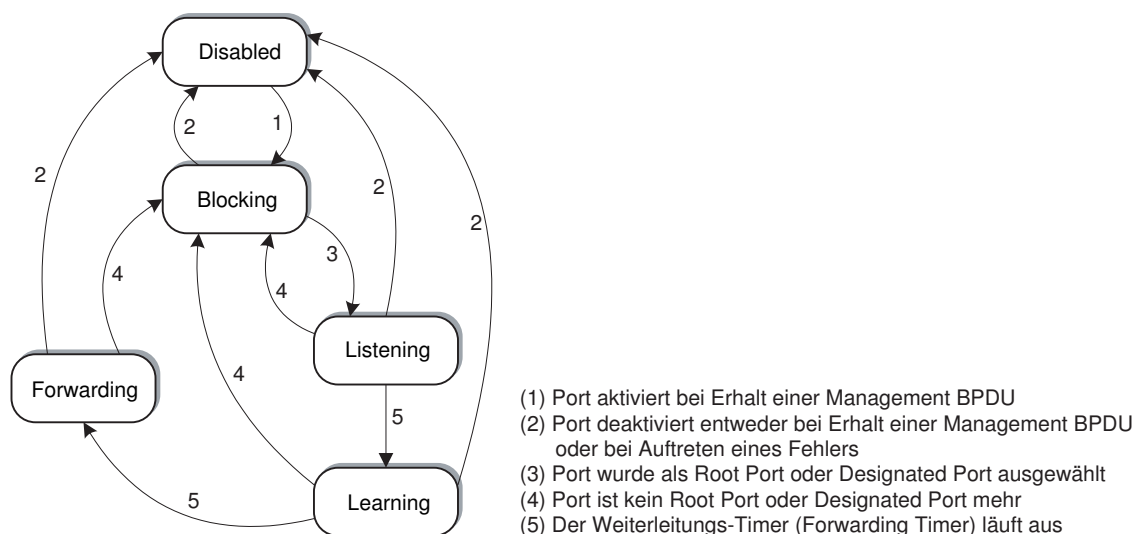


Abbildung 12.1: Zustände der Ports und die möglichen Zustandsänderungen

Der Zustand Disabled trifft für alle Bridges zu, die das erste mal in Betrieb genommen werden. Sie gehen in den Zustand Blocking über, wenn eine bestimmte Management BPDU von einer Management Station abgesetzt wird (1). Auf die gleiche Weise kann ein Netzwerk-Manager bei Bedarf einen bestimmten Port deaktivieren (2).

Nachdem eine Bridge das Initialisierungs-Kommando von dem Netzwerk-Manager erhalten hat, werden alle ihre Ports in den Zustand Blocking versetzt und sie beginnt, Configuration BPDUs auszusenden. Anschließend beteiligt sie sich an der Initialisierung der Topologie, wie zuvor beschrieben. Während dieser Prozedur werden entweder Root Ports oder Designated Ports etabliert. Dabei können sich die Zustände der Ports natürlich ändern. Beispielsweise könnten Bridges weiter hinten im Baum aufgrund erhaltener BPDUs annehmen, daß einige Ports Designated Ports oder Root Ports sind. Doch beim Durchlaufen der Baumstruktur könnten sich die Zustände erneut ändern. Statt also nun die Ports direkt vom Zustand Blocking in den Zustand Forwarding zu versetzen, gehen sie zuerst in die Zustände Listening und Learning über.

Wenn eine Bridge festgestellt hat, daß einer ihrer Ports ein Root Port oder Designated Port ist, wird es vom Zustand Blocking in den Zustand Listening versetzt (3) und es wird ein *Forwarding Timer* gesetzt. Ist ein Port noch immer ein Root oder Designated Port, wenn der Timer abläuft, wird sie in den Zustand Learning (5) versetzt. Daraufhin wird der Forwarding Timer zurückgesetzt und die gesamte Prozedur wiederholt, bis er erneut abläuft. Diesemal werden alle Ports, die noch immer Root oder Designated Ports sind in den Zustand Forwarding (5) versetzt. Ist einer der Ports in der Zwischenzeit weder Designated noch ein Root Port, so werden sie in den Zustand Blocking zurückgesetzt (4).

### 12.4.1 Beispiel

Um zu zeigen, wie der Spanning Tree Algorithmus arbeitet, betrachten wir Abbildung 12.2. Auf der linken Seite sehen Sie den Aufbau des Netzwerkes und die Beschriftungen B1 bis B6, welche die Bridges darstellen. Die Ziffern in den Kästen stehen für die Portnummern des Switches. Wie deutlich zu erkennen ist, sind die Bridges redundant ausgelegt, um die Ausfallsicherheit im Falle eines Fehlers zu verbessern.

Versuchen wir nun die Funktionsweise des Spanning Tree Protocols schrittweise nachzuvollziehen. Ermitteln Sie die aktive Topologie für die folgenden Fälle:

1. Alle Bridges weisen die gleiche Priorität auf und allen Segmenten ist der gleiche Designated Cost (also die gleiche Bitrate) zugeordnet.
2. Arbeiten Sie die gleiche Topologie wie in (1) aus, nur für den Fall, daß Bridge B1 ausgefallen ist.
3. Alle Bridges sind in Betrieb, aber die Segmente S1, S2, S4 und S5 haben den dreifachen Cost der Segmente S3 und S6 (sie verfügen über höhere Bitraten).

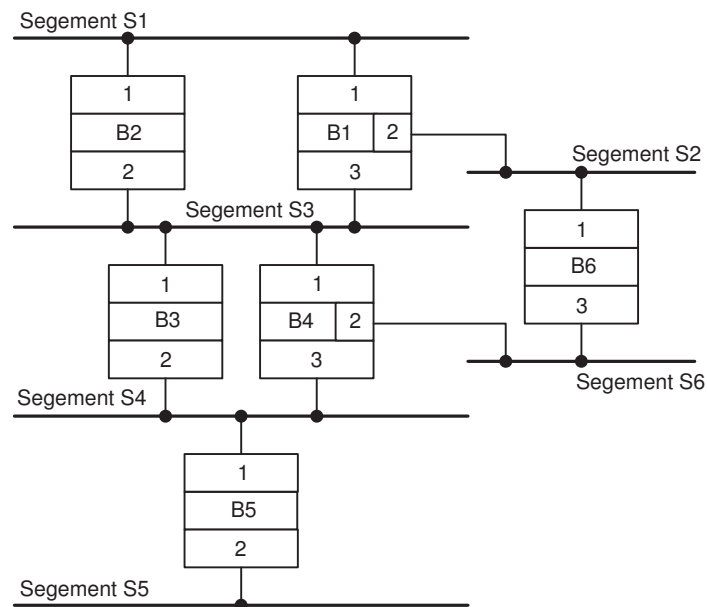


Abbildung 12.2: Herleitung der aktiven Topologie anhand eines Beispielnetzwerks.

4. Die Segmente verfügen über den gleichen Designated Cost wie in (3), aber die Priorität der Bridge B5 ist auf einen höheren Wert gesetzt worden als die anderen Bridges.

Die Lösung der Einzelfragen wird nun detailliert aufgeführt, um die Funktionsweise des Spanning Tree Protocols möglichst einleuchtend zu vermitteln.

1. (a) Der erste Austausch von Configuration BPDUs etabliert B1 als Root Bridge, da sie den kleinsten Bezeichner aufweist.
- (b) Nach dem Austausch von Configuration BPDUs ist der Root Path Cost (RPC) für jeden Port ermittelt worden. Sie sind der Abbildung 12.3 zu entnehmen.
- (c) Anschließend wird der Root Port (RP) für jede Brücke ausgewählt und ist immer der mit dem geringsten RPC. Für B3 hat Port 1 den geringsten RPC von 1 und Port 2 einen RPC von 2. Also ist Port 1 der Root Port. Für B2 ist der Fall nur unwesentlich anders, denn beide Ports haben den gleichen Cost, also wird Port 1 ausgewählt, weil er einen kleineren Bezeichner aufweist.
- (d) B1 ist die Root Bridge. Aus diesem Grund haben alle Ports eine DPC (Designated Port Cost) von 0. Also sind sie auch die Designated Ports für die Segmente S1, S2 und S3.
- (e) Auf Segment S4 ist Port 1 von B5 ein Root Port und wird daher nicht in der Auswahl des Designated Ports berücksichtigt. Die anderen beiden Ports für das Segment S4 haben beiden einen DPC von 1. Deshalb wird Port 2 von B3 als Designated Port ausgewählt, da er einen kleineren Bezeichner aufweist.
- (f) Der einzige mit S5 verbundene Port ist Port 2 von B6 und wird daher ausgewählt.
- (g) Schließlich haben beide Ports von S6 einen DPC von 1, so daß Port 2 von B4 statt Port 2 von B6 ausgewählt wird.

Die ermittelten DPCs werden in Abbildung 12.4 dargestellt und die entgültige aktive Topologie für Aufgabe (1) können Sie 12.5 entnehmen.

2. (a) Zuerst wird B2 als neue Root Bridge durch die BPDUs als Teil der Topologieerkennung ermittelt, da sie den kleinsten Bezeichner aufweist.
- (b) Anschließend wird der neue RPC für jeden Port und die Root Ports für alle Bridges ermittelt.
- (c) Da B2 nun die Root Bridge ist, sind ihre Ports Designated für die S1 und S3.
- (d) Mit S2 ist nur Port 1 von B6 verbunden und wird somit ausgewählt.

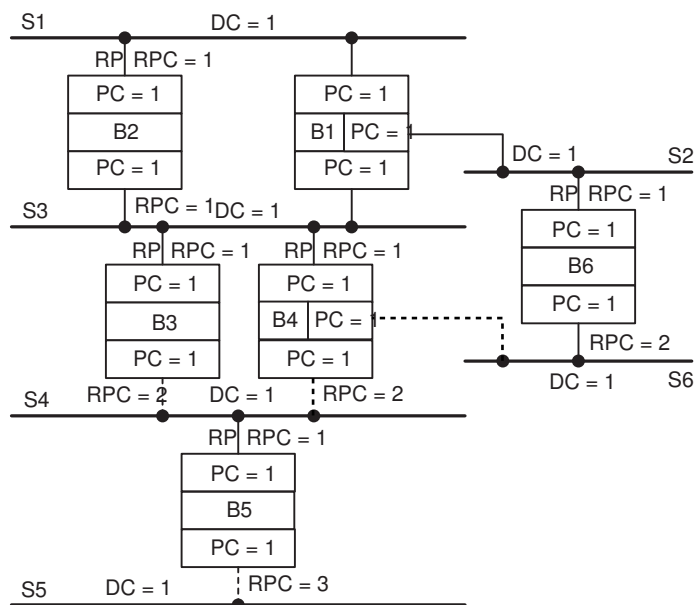


Abbildung 12.3: Ermittlung der Root Ports.

- (e) Beide Ports, die mit S4 verbunden sind weisen den gleichen PDC auf, so daß Port 2 von B3 ausgewählt wird.
- (f) Für S2 wird Port 2 von B5 ausgewählt.
- (g) Letztendlich wird Port 2 von B4 für S6 ausgewählt, da beide den gleichen DPC aufweisen.
- Die neue aktive Topologie wird in Abbildung 12.6 dargestellt.
3. (a) Die Root Bridge ist B1 mit dem kleinsten Bezeichner.
- (b) Die RPCs und RPs sind wie in Abbildung 12.7 dargestellt. Beachten Sie, daß Port 2 von B6 nun der Root Port ist, da er einen kleineren RPC als Port 1 aufweist.
- (c) B1 ist die Root Bridge und somit sind ihre Ports die Designated Ports für S1, S2 und S3.
- (d) Für S4 wird erneut Port 2 von B3 ausgewählt.
- (e) Für S5 wird erneut Port 2 von B5 ausgewählt.
- (f) Für S6 wird erneut Port 2 von B4 ausgewählt.
- Die geänderte aktive Topologie wird in Abbildung 12.7 dargestellt.
4. (a) Nun hat B5 die höchste Priorität und wird daher als neue Root-Bridge ausgewählt.
- (b) Da B5 die Root Bridge ist, sind ihre Ports die Designated Ports für S4 und S5.
- (c) Der neue RPC wird für jeden Port berechnet. Das Ergebnis ist in Abbildung 12.8 zu entnehmen.
- (d) Für S1 wird Port 1 von B1 ausgewählt.
- (e) Für S2 wird Port 2 von B1 ausgewählt.
- (f) Für S3 wird Port 1 von B3 ausgewählt.
- (g) Für S6 wird Port 2 von B1 ausgewählt.
- Die neue geänderte Topologie wird in Abbildung 12.8 dargestellt.

□

Aus Beispiel 12.4.1 können wir ableiten, daß wenn alle Bridges eines LANs die gleiche Priorität aufweisen, das Spanning Tree Protocol wahrscheinlich keine optimalen Aktivtopologien ermitteln wird, zumindest was die Bandbreite betrifft. Gerade in großen Netzwerken kann das von entscheidender Bedeutung sein,

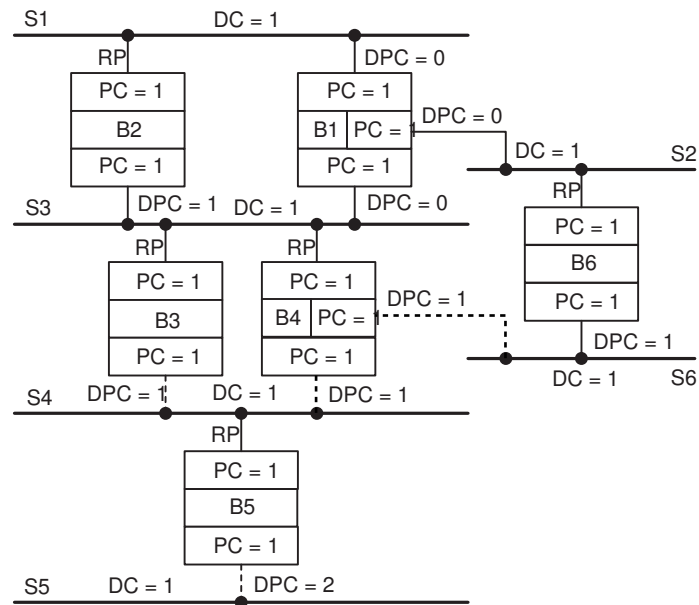


Abbildung 12.4: Ermittlung der Designated Ports.

denn hier ist die Ausnutzung der zur Verfügung stehenden Bandbreite wichtiges Kriterium eines Netzwerks.

Zur Unterstützung des Protokolls wird dem Bridge-Bezeichner auch eine Priorität zugeordnet, um dieses Ziel zu erreichen. Der Netzwerkmanager kann das Feld je nach Anwendung dynamisch setzen, was mit dem eindeutigen Bezeichner nicht geht, der bei der Herstellung der Bridge fest eingeschrieben wird.

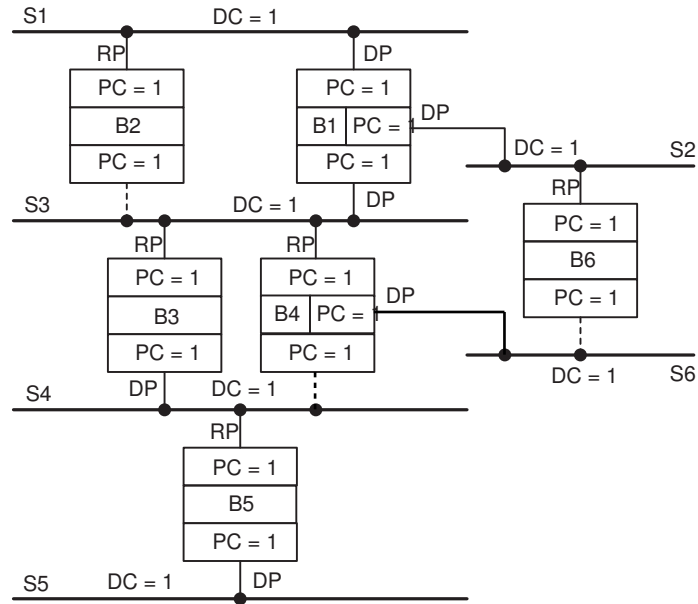


Abbildung 12.5: Aktive Topologie des Beispielnetzwerks.

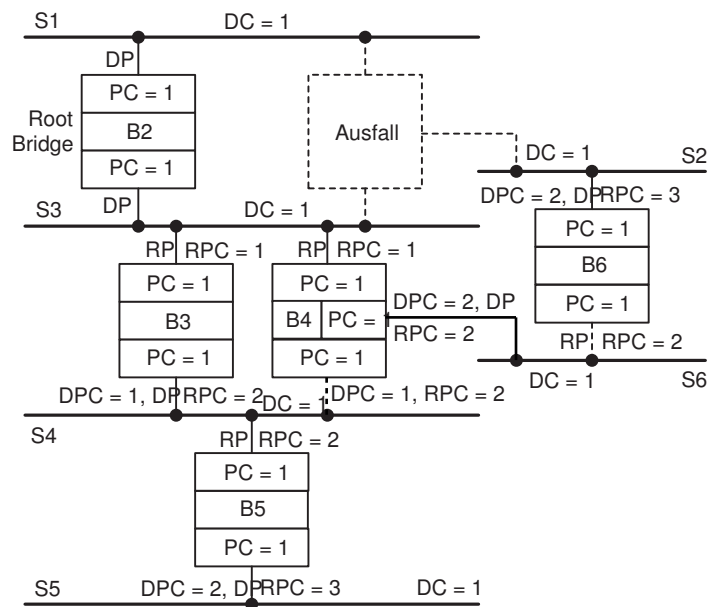


Abbildung 12.6: Aktive Topologie des Beispielnetzwerks.

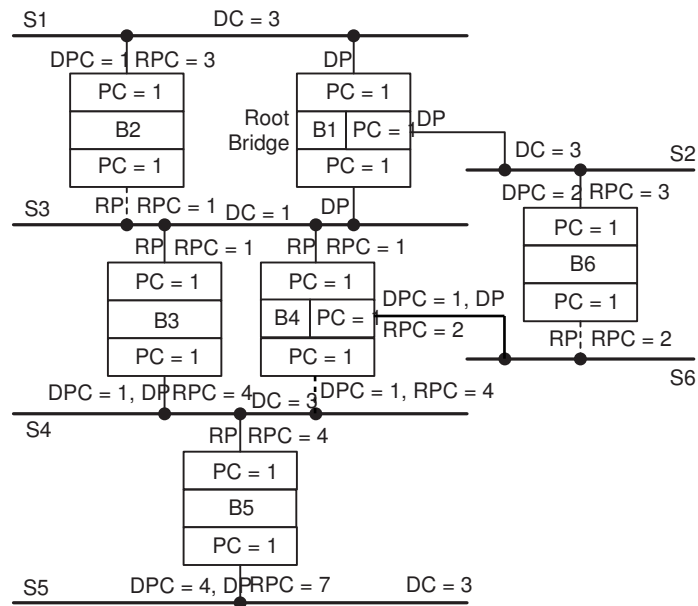


Abbildung 12.7: Aktive Topologie des Beispielnetzwerks.

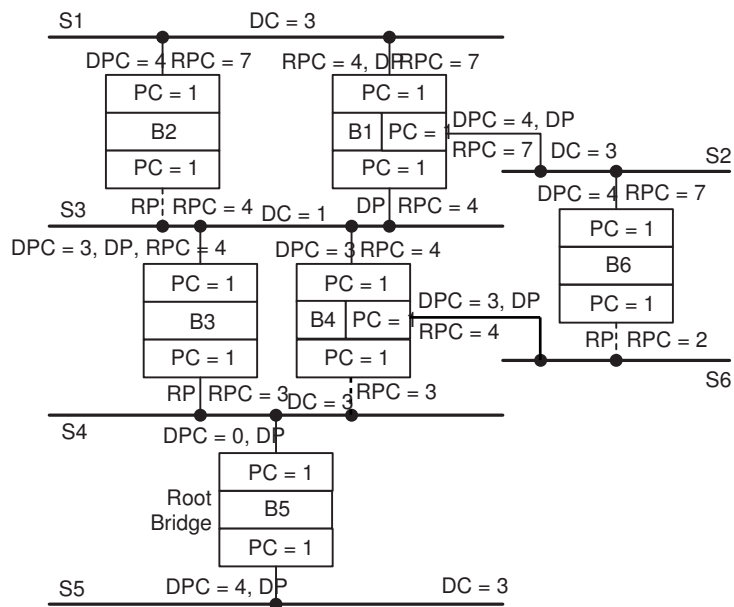


Abbildung 12.8: Aktive Topologie des Beispielnetzwerks.



# Kapitel 13

## Switches und Multicast-Traffic

Die Hauptaufgabe eines Switches ist die Bandbreite so gering wie möglich zu halten, indem unerwünschter Verkehr innerhalb einer Collision Domain gezielt herausgefiltert wird. In Kapitel 11 wurde gezeigt, das Switches über recht effektive Mechanismen zur Kontrolle des Netzverkehrs verfügen.

Multicast-Verkehr zu koordinieren ist eine weit aufwendigere Aufgabe. Ursprünglich wurde Multicasting für LANs nach dem Broadcasting-Prinzip entwickelt. Alle Stationen eines LANs sollten, genau wie beim Broadcasting alle Pakete sehen können. Die Idee wurde schnell verworfen, denn schließlich macht es für geographisch verteilte LANs wenig Sinn, das ganze Netz mit Paketen zu belasten, die evtl. nur für wenige Stationen bestimmt sind.

Immer wenn ein System einer Multicast-Gruppe beitreten möchte, sendet es IGMP-Nachrichten (*Internet Group Management Protocols*) aus, das den Routern mitteilt, daß alle Pakete, die an eine bestimmte Multicast-Adresse gerichtet sind, auch an dieses System weitergeleitet werden. In Bridges hat sich ein Mechanismus, der als *IGMP Snooping* bezeichnet wird, durchgesetzt. Bridges lauschen auf IGMP-Nachrichten, um herauszufinden, welche Teile des LANs Mitglied einer Multicast-Gruppe sind. Der Verkehr wird dann auf Basis der MAC-Adresse weitergeleitet.

IGMP Snooping ist nicht optimal, denn es belastet Bridges sehr stark und funktioniert nur mit IP. Die 802.1 Arbeitsgruppe der IEEE hat eine Alternative aufgezeigt. Sie entwickelte ein Layer-2-Protokoll, das Bridges ermöglicht, sich bei benachbarten Bridges zu registrieren, so daß sie wissen, welche Bridges Multicast-Verkehr mit einer bestimmten Multicast-MAC-Adresse erhalten.

Die Registrierungsinformationen werden von Bridge zu Bridge über das gesamte LAN verteilt und jede Bridge weiß dadurch über welche Ports der Multicast-Verkehr weitergeleitet werden muß. Die Arbeit der 802.1 Arbeitsgruppe wird als *GARP Multicast Registration Protocol* (GMRP) bezeichnet.

GARP steht übrigens für *Generic Attribute Registration Protocol* und ist das Fundament von GMRP. GARP bietet grundlegende Registrierungsmechanismen für Bridges in LANs. GMRP ist das erste Protokoll, daß auf GARP basiert.

In diesem Kapitel werden wir uns mit IGMP und GMRP auseinandersetzen.

### 13.1 Multicasting

Multicast MAC-Adressen erlauben es einem einzelnen Rahmen an mehrere Teilnehmer gesendet zu werden. Aus diesen Grund wird eine Multicast-Adresse auch als *Gruppenadresse* und teilnehmende Stationen als Mitglieder einer Multicast-Gruppe bezeichnet.

Multicast MAC-Adressen werden auf zwei unterschiedliche Weisen verwendet:

- Standardisierte Multicast-Adressen wurden speziellen Geräten und Netzwerkrollen zugeordnet. Beispielsweise wird ein Multicast-Paket mit der Ziel-MAC X'01-00-5E-00-00-02 an alle Router im LAN gesendet, die Multicast verarbeiten können.

- Es wird ein bestimmter Adressbereich für den flexiblen Einsatz von Ad-Hoc Multicast-Gruppen reserviert. Diese Adressen werden beispielsweise für Netzwerkanwendungen wie Konferenzen, Computer-basiertes Lernen oder Unterhaltungsprogramme verwendet. Eine solche Adresse kann aus dem Pool entnommen und für eine zuvor festgelegte Zeit verwendet werden.

Eine Bridge wird von Haus aus mit wichtigen standardisierten Multicast-Adressen ausgeliefert und kann sofort Multicast-Verkehr verarbeiten.

Selbstverständlich kann eine Station, die selbst nicht Mitglied einer Multicast-Gruppe ist, Daten an diese Gruppe senden. Allerdings verwirft sie alle Rahmen, die an die Gruppenadresse gerichtet sind. Das entspricht dem Redner-Zuhörer-Modell eines Parlaments. Der Redner richtet Worte an eine Gruppe von Menschen, akzeptiert aber keine Zwischenfragen. Die Mitglieder einer Gruppe dürfen sich aber sehr wohl untereinander unterhalten.

### 13.1.1 Warum müssen wir Multicasting steuern?

Ganz offensichtlich brauchen wir in einem kleinen LAN kein Multicasting. Eine Bridge mit vier Teilnehmern sendet alle eingehenden Rahmen an jedes angeschlossene Segment, egal ob Unicast oder Multicast. Damit sieht auch jede Station alle Rahmen, die eine Bridge zu sehen bekommt. Stationen, die nicht Mitglieder einer Multicast-Gruppe sind, werfen die Rahmen einfach.

Größere LANs laufen beim Betrieb mehrerer Multicast-Anwendungen Gefahr, vom Multicast-Verkehr aufgefressen zu werden. Es ist also notwendig, einen Mechanismus zu etablieren, der Teile eines LANs von Multicast-Verkehr freihält, wenn es keine Stationen gibt, die Mitglied einer Multicast-Gruppe sind.

Die Sichtbarkeit der Rahmen für Teile eines LANs einzuschränken wurde noch wichtiger, als Switches ins Spiel kamen, denn ihre Eigenschaft war es, nicht alle Rahmen an alle Ports weiterzuleiten. Aber genau das ist die Funktionsweise von Multicast. Es sendet alle Multicast-Rahmen an alle angeschlossenen Stationen, auch wenn nur eine tatsächlich Mitglied einer Gruppe ist.

IGMP Snooping kann zur Verwaltung von Multicast-Verkehr eingesetzt werden und wird im nächsten Abschnitt besprochen. GMRP ist ausgefeilter und bietet effektivere Mechanismen, die wir uns weiter hinten in den Abschnitten *Das GARP Multicast Registration Protocol* und *Das Generic Attribute Registration Protocol* anschauen wollen.

### 13.1.2 IP-Multicasting

Das TCP/IP-Protokoll ermöglicht das Aufsetzen von Multicast-Konferenzen. Es ist ein Protokoll mit sehr hoher Reichweite. So können beispielsweise Teilnehmer aus entfernten Teilen der Erde an Konferenzen teilnehmen, obwohl sehr viele Router zwischen Sender und Empfänger liegen.

IP-Multicasting kann in folgenden Umgebungen eingesetzt werden:

- In einem einzelnen lokalen Netzwerk.
- In einem privaten Netzwerk, das mehrere LANs umfaßt.
- Im Internet, das die Teilnehmer miteinander verbindet.

Ähnlich wie MAC-Adressen gibt es im IP-Stapel eine ganze Reihe von IP-Adressen, die für die Anwendung von Multicasting reserviert sind. Systeme, die an einer Sitzung teilnehmen, deren Pakete an eine bestimmte Multicast-Adresse gerichtet sind, bilden eine *IP Multicast-Gruppe*.

IP-Datenpakete (sogenannte *Datagramme*) müssen an alle Stationen eines LANs weitergeleitet werden, wenn sie an einer Multicast-Sitzung teilnehmen. Trifft ein IP-Datagramm mit einer 32-Bit IP-Multicast-Adresse an einem Router ein, muß der Router einen 48-Bit MAC-Header an das IP-Datagramm anfügen. Doch welche MAC-Adresse soll der Router wählen? In Ethernet-LANs ist das einfach: alle IP-Multicast-Adressen werden an die Multicast-MAC-Adresse gerichtet. Die ersten drei Bytes dieser Adresse sind immer X'01-00-5E. Die letzten drei Bytes sind immer die letzten drei Oktette der IP-Multicast-Adresse.

Wir brauchen nur die letzten drei Oktette, da das erste immer eine Zahl zwischen 224 und 239 repräsentiert. Dadurch entsteht bei einer Ethernet-Adresse mit den Bytes X'01-00-5E-00-00-02 und einer Multicast-IP von 224.2.3.4 einer Ziel-MAC-Adresse mit den Bytes X'01-00-5E-02-03-04.

Anders stellt sich die Situation mit FDDI dar. Grundsätzlich ist das Verfahren das gleiche, wie wir es gerade für Ethernet kennengelernt haben. Allerdings wird die Multicast-Adresse in nicht-kanonischer Schreibweise auftauchen, so daß sie verkehrt herum verarbeitet wird. Somit wird aus 224.2.5.9 die MAC-Adresse X'80-00-7A-40-A0-90.

Token Ring bietet nur sehr geringe Multicast-Fähigkeiten und ein einfaches Mapping, wie für Ethernet und FDDI gibt es auch nicht. Alle Multicast-Datagramme werden entweder in Rahmen mit verdrehten Bytes verpackt oder an eine MAC-Broadcast-Adresse gesendet. Alle Stationen eines Token Rings müssen dann die Pakete, die für die Gruppe an der sie teilnehmen bestimmt sind, absorbieren oder andernfalls verwerfen.

## 13.2 IGMP Snooping

IGMP definiert Verfahrensweisen, die IP-Endsysteme erlauben, Router darüber zu benachrichtigen, daß sie an einer Multicast-Sitzung teilnehmen wollen. Das funktioniert etwa so:

1. Ein Router sendet periodisch IGMP-Anfragen aus, ob ein Teilnehmer einer bestimmten Gruppe beitreten möchte.
2. Ein Host, der einer Gruppe beitreten oder seine Mitgliedschaft erneuern möchte, antwortet mit einem IGMP-Mitgliedschaftsbericht (*membership report*). Sie sind an die Multicastadresse der jeweiligen Multicast-Gruppe gerichtet.
3. Der Router hört die Berichte ab und verwendet sie als Entscheidungsgrundlage, ob mindestens eine Station im LAN Multicast-Daten erhalten möchte.

Die meisten Multicast-fähigen Switches erlauben das Zuschalten von IGMP-Snooping, so daß der Switch solche Nachrichten erhält und Einträge in seiner Filtertabelle machen kann. Auf diese Weise kannalisiert er die Nachrichten an jene Ports, an denen Mitglieder einer Multicast-Gruppe zu finden sind.

IGMP Snooping ist UDP-basiert und arbeitet auf Layer 4 des TCP/IP-Protokollstapels.

### 13.2.1 Nachteile des IGMP-Snooping

Switches, die Multicast-Teilnehmer über IGMP-Snooping aufspüren, unterliegen sehr hoher Prozessorlast. Membership-Reports werden an die Multicast-Adresse gesendet, der ein Host beitreten möchte. Somit muß der Switch jedes Multicast-Datagramm untersuchen, um herauszufinden, ob es ein IP-Datagramm enthält, um anschließend im Header nachzuschauen, ob es einen Membership Report enthält oder nicht. Ist das der Fall, muß der Switch evtl. seine Filtertabelle aktualisieren und sicherstellen, daß der Eingangsport des Reports auch in der Liste der Ausgabeports enthalten ist.

Außerdem muß ein Switch alle Multicast-Datagramme, die von einem Router stammen an anderer Router weiterleiten, wenn sie Zugang zu anderen Multicast-Teilnehmern bieten. Um herauszufinden, welche Router weitere Empfänger enthalten, hört der Switch zusätzlich auch die Router-zu-Router-Kommunikation ab. Unglücklicherweise existieren hier eine Menge unterschiedlicher Implementierungen, wie etwa DVMRP (*Distance Vector Multicast Routing Protocol*), Dense Mode PIM (*Protocol Independent Multicast*), Sparse Mode PIM oder MOSPF (*Multicast Extensions to Open Shortest Path First*).

Neben den soeben genannten rechenintensiven Problemen gesellt sich noch ein weiterer Nachteil dazu. IGMP bietet keine einfache Möglichkeit, Router zu identifizieren, die Rahmen einer bestimmten Multicast-Adresse empfangen möchten. Das gleiche gilt für Netzwerkmonitore, die Multicast-Verkehr für die Fehlersuche empfangen müssen.

### 13.3 GARP Multicast Registration Protocol

Da IGMP-Snooping nur mit IP zusammenarbeitet, hat die IEEE eine Layer-2-Variante eines Multicast-Management Protokolls entwickelt. GMRP wird im Standard 802.1D (Bridge Protocol Standard) definiert.

Möchte ein System Daten einer bestimmten Multicast-Gruppe empfangen, so verwendet es GMRP, um sich bei einem benachbarten Switch zu registrieren. GMRP erlaubt Hosts, Switches und Routern den Verkehr abzugreifen, den sie empfangen möchten. GMRP leitet den Verkehr nur über die Ports weiter, an denen entweder Empfänger von Multicast-Datagrammen anliegen, oder an denen Zugang zu einem Teil eines LANs besteht, der Multicast-Teilnehmer enthält.

#### 13.3.1 Die Registrierung

Ein System verwendet GMRP um den Wunsch anzuzeigen, daß es Multicast-Verkehr empfangen möchte. Es kann sich für folgende Daten registrieren:

- Multicast-Rahmen, die an eine Liste bestimmter Multicast-MAC-Adressen gerichtet sind.
- Alle Multicast-Rahmen. Wird meistens von Routern und Netzwerkmonitoren registriert.
- Alle Multicast-Rahmen, die nicht durch eine bestimmte Regel an der Weiterleitung gehindert werden.

Die Prozedur selbst läuft dann etwa so ab:

1. Ein Host registriert sich bei seinem nächstgelegenen Switch, und zeigt an, daß er Rahmen der Multicast-Adresse  $M$  empfangen möchte.
2. Dieser Switch leitet die Anfrage samt MAC-Adresse über seine aktiven Ports an seine Nachbarn weiter, die eine Filtertabelle mit Ein- und Ausgangsport der Anfrage aufbauen.
3. Trifft nun an einem der Ports im Baum des Netzwerkes ein Multicast-Rahmen ein, so wissen die Switches, über welchen Port die Daten weitergeleitet werden müssen.

Auf diese Weise wird das gesamte LAN mit den notwendigen Informationen versorgt, aber gleichzeitig sichergestellt, daß die Switches nicht unnötig belastet werden.

Als aktive Ports eines Switches bezeichnen wir solche, die nicht durch das Spanning Tree Protocol (Kapitel 12) blockiert werden. Da die Topologie als Baum aufgebaut wird, ist sichergestellt, daß Registrierungen und Deregistrierungen nicht aus Versehen wieder zum ursprünglichen Switch zurückkehren.

### 13.4 Generic Attribute Registration Protocol

Das GMRP-Protokoll ist eine Anwendung des allgemeinen *Generic Attribute Registration Protocols*. Es stellt Mechanismen bereit, die es einer Station erlauben, einen bestimmten Teil einer Information bei anderen Systemen eines LANs zu registrieren. Diese Registrierung wird in alle Teile eines LANs verteilt. Wie Sie dieser Erläuterung entnehmen können, ist GARP in der Tat sehr allgemein formuliert worden.

GARP ist ein Layer-2-Protokoll und kann daher als Teil des Treibers für NICs implementiert und durch Software in der Hardware der Komponenten unterstützt werden. Das könnte dann etwa so funktionieren, daß der Treiber automatisch eine GRMP-Registrierung an benachbarte Switches sendet, wenn ein System einer Multicast-Gruppe via IGMP beitrifft.

Prinzipiell verfolgten die Entwickler von GARP folgende Ideen:

- Ein System, daß direkt mit einem Switch verbunden ist, tritt einer Gruppe bei, indem es zwei JOIN-Nachrichten absetzt.
- Ein System muß seine Mitgliedschaft erneuern und erledigt das ebenfalls mit JOINS.
- Ein System verläßt eine Gruppe, indem es eine LEAVE-Nachricht sendet.
- Sollte eine LEAVE-Nachricht mal verloren gehen, geht der Switch weiterhin davon aus, daß das System noch immer Teilnehmer ist. Um das Problem zu umgehen, sendet der Switch nach einer definierten Zeitspanne eine LEAVEALL-Nachricht aus, die allen Teilnehmern sagt, daß wenn sie nicht bald eine neue JOIN-Nachricht schicken, sie automatisch aus der Gruppe ausscheiden.

Verbinden sich mehrere Systeme über einen Hub mit einem Switch, so genügt es, wenn nur ein System, eine JOIN-Nachricht absetzt, denn der HUB liefert die Multicast-Rahmen an alle Stationen aus.



# Kapitel 14

## Link Aggregation

Mit Hilfe von *Link Aggregation* sind wir in der Lage, mehrere Verbindungen zusammenzufassen, so daß sie sich wie eine einzige verhalten. Eine solche kombinierte Verbindung wird als *Trunk* oder *Aggregated Link*, und ein teilnehmender Port als *Aggregation Port* bezeichnet.

Link Aggregation hat viele Vorteile. Zwei davon wären:

- Erzielung höherer Bandbreite durch Kombination mehrerer Links mit geringerer Bandbreite.
- Höhere Ausfallsicherheit, da bei Ausfall eines Aggregated Ports die Bandbreite auf die übrigen verteilt wird und die Kommunikation nicht unterbrochen wird.

In der Vergangenheit wurden zur Steigerung der verfügbaren Bandbreite weitere Links in das LAN integriert, was mit hohem Materialeinsatz verbunden war. Oftmals waren diese Verbindungen ungenutzt, bis ein Ausfall auftrat und auf die Ersatzverbindungen umgeschaltet wurde.

Natürlich hat die IEEE eine Arbeitsgruppe (802.3ad) zur Ausarbeitung eines Standards für die Link Aggregation gebildet. Inzwischen haben die meisten Hersteller diesen Standard angenommen und in ihre Produkte integriert. Da sich die Link Aggregation auf der Ebene der Gerätetreiber abspielt, konnte sie leicht auf bestehende Produkte durch neue Treiber übertragen werden.

### 14.1 Link Aggregation im Einsatz

Abbildung 14.1 zeigt ein Beispielnetzwerk, daß illustriert, wie Link Aggregation eingesetzt werden kann. zwischen zwei Gigabit-Switches wurden zwei Links zusammengefaßt, um den Datendurchsatz zwischen den Switches unter Last besser zu verteilen. Zwei Server sind jeweils durch vier mal 100 Mbps angebunden und die Workstations erhalten eine Doppelverbindung.

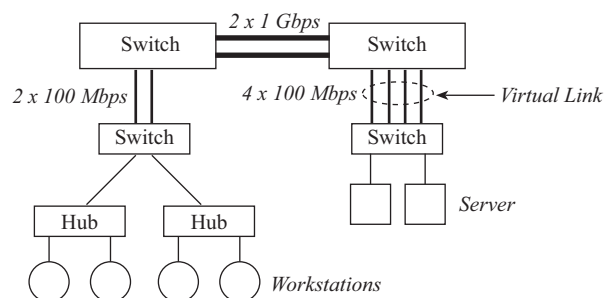


Abbildung 14.1: Link Aggregation in der Praxis.

Link Aggregation läßt sich vielseitig einsetzen. So können wir mehrere Links zwischen folgenden Netzwerkkomponenten kombinieren:

- ein Paar Layer-2-Switches.
- Layer-2-Switch und Server/Workstation
- zwei Server/Workstations
- zwei Router und Router/Layer-2-Switch

Wichtig ist, daß nur Links mit gleicher Geschwindigkeit kombiniert werden können.

Wenn der Administrator Link Aggregation aktiviert hat, sind beide Systeme in der Lage, die Modalitäten selbständig auszuhandeln, ähnlich wie das bei der Auto-Negotiation der Fall ist. Fällt eine der Links aus, so ignoriert das Protokoll diesen Link und führt die Kommunikation bei reduzierter Bandbreite über die anderen Links weiter.

Link Aggregation ist rückwärtskompatibel. Beim Versuch ein altes System ohne Unterstützung von Link Aggregation mit einem modernen System zu verbinden und einen aggregierten Link zu etablieren, schlägt die Aushandlung durch das Protokoll fehl und es findet gar keine Link Aggregation statt.

## 14.2 Konzepte und Vorgehensweisen

Aggregierte Links werden oftmals als *virtuelle Links* bezeichnet, da sie sich wie ein Link verhalten, aber in Wirklichkeit aus mehreren Einzelverbindungen bestehen. Das bedeutet aber, daß die Endsysteme eines virtuellen Links auch eine gültige MAC-Adresse benötigen. Nachwievor werden die Daten auf Basis der Quell- und Ziel-MAC-Adresse über den virtuellen Link transportiert. Einige Implementierungen wählen die kleinste MAC-Adresse unter den verfügbaren aus, andere nehmen einfach die des ersten Ports. In der Regel kann aber der Administrator über die jeweilige MAC-Adresse entscheiden. Egal welches Verfahren gewählt wird, am Ende verfügt der virtuelle Link auch über eine virtuelle *MAC-Adresse*.

Gewöhnlicher Datenverkehr wird über die virtuelle MAC-Adresse geleitet, während Nachrichten des *Link Aggregation Control Protocol* über die jeweilige MAC-Adresse des Segments transportiert werden. Das gleiche gilt auch für die Seite des Switches. Daten werden über die virtuelle MAC-Adresse, aber spezifische Informationen wie BPDUs oder SNMP-Nachrichten sind stets an eine bestimmte reelle MAC-Adresse gerichtet.

Für die Link Aggregation wurde ein weiterer Sublayer zwischen LLC- und der MAC-Schicht eingeführt. Er teilt die eingehenden Rahmen des Logical Link Control so ein, daß sie optimal auf die echten Links verteilt werden. Dabei werden die Daten nicht etwa in gleichgroße Einheiten aufgeteilt, sondern je nach Auslastung an einen bestimmten Link ausgegeben. Das alles erledigt das *Link Aggregation Control Protocol* (LACP).

## 14.3 Parameter von Link Aggregation

Die meisten Parameter, die wir in Link Aggregation Nachrichten antreffen sind dem Spanning Tree Protocol entliehen. Das STP setzt den Bezeichner für eine Brigade aus einem 2-Byte-Prioritätsfeld und einer MAC-Adresse (normalerweise die kleinste). Andere Systeme, wie Router und Server, erhalten einen Bezeichner auf die gleiche Weise.

Ports, die Teil einer Link Aggregation sind, erhalten ebenfalls einen Identifier, allerdings setzt er sich aus einer Priorität, Portnummer und einem Schlüssel zusammen. Der Schlüssel ist allen aggregierten Ports gemein und bilden zusammen eine Schlüsselgruppe. Jede Aggregation weist damit eine eigene Schlüsselgruppe auf. Wenn Ports einer Schlüsselgruppe zugeordnet werden können, heißt das noch nicht, daß auch die Aggregation erfolgreich durchgeführt werden kann. Das Protokoll des Gerätes muß in folgenden Punkten mit denen des LACP übereinstimmen:

- Gerätespezifische Regeln, die für die Aggregation notwendig sind, müssen anwendbar sein. Beispielsweise kann die Zahl der maximal aggregierbaren Ports eingeschränkt sein.

- Es muß feststellbar sein, ob alle Ports in einer Schlüsselgruppe auch tatsächlich mit dem entfernten System verbunden werden können.
- Es muß feststellbar sein, ob die entfernten Ports eines Systems einer gemeinsamen Schlüsselgruppe angehören.

Protokollnachrichten enthalten Flags, die anzeigen, ob ein Port zu einer bestimmten Gruppe zugeordnet werden kann und ob es evtl. schon einer Gruppe angehört.

## 14.4 Das Link Aggregation Control Protocol

Nun, da wir die grundlegende Funktionsweise der Link Aggregation kennengelernt haben, wollen wir nun ein Blick auf das LACP werfen.

Die LACP-Nachrichten werden über jeden Link der Aggregation übertragen. Sie sind also nicht den virtuellen Links sondern realen Links zugeordnet. Die Quelladresse in jedem LACP-Rahmen entspricht der des realen Ports über den es gesendet wurde.

Das LACP ist ein langsames Protokoll. Im Ethernet-Jargon ist ein langsames Protokoll durch folgende Eigenschaften definiert:

- Nicht mehr als fünf Rahmen können pro Sekunde übertragen werden.
- Protokollnachrichten werden in speziellen Multicast-Rahmen transportiert. Langsamen Protokollen ist die Multicast-Adresse X'01-80-C2-00-00-02 zugeordnet.
- Das Ethertype-Feld enthält den Wert des langsamen Protokolls und lautet für LACP X'88-09.

Das erste Byte des Datenfeldes eines langsamen Protokolls enthält den Untertyp der folgenden langsamen PDUs (*protocol data units*).

### 14.4.1 LACP-Nachrichten

Alle Informationen des LACP werden in *LACP Protocol Data Units* (LACPDU) transportiert. Der Wert des Subtype ist X'01. Die Parameter einer LACP-Nachricht beinhalten folgende Informationen:

- den lokalen Bezeichner des Systems (Priorität und MAC-Adresse)
- den lokalen Bezeichner des Ports (Priorität und Portnummer)
- Schlüssel, der dem Port zugeordnet ist.
- Lokale und Partnerstatusflags.
- Systembezeichner, Portbezeichner und Schlüssel des entfernten Systems.
- Zeitgeberinformationen.

Im Laufe des Nachrichtenaustauschs werden die Informationen ständig aktualisiert und zur Konfiguration genutzt, beispielsweise um einen Link aus der Gruppe zu entfernen.

Leider ist mir das genaue Rahmenformat nicht bekannt, da der Standard nicht frei verfügbar ist, sondern nur gegen ein Endgelt angefordert werden kann.

### 14.4.2 Rahmenübermittlung

Wenn die Datenrahmen über mehrere Linksegmente verteilt werden, könnten sie auf der Gegenseite in der falschen Reihenfolge oder mit Verzögerung eintreffen. Je nach Netzwerkarchitektur und verwendetem Protokoll kann es durchaus von Bedeutung sein. Während IBMs SNA empfindlich auf eine falsche Reihenfolge während einer Sitzung reagiert, ist TCP/IP weitgehend immun und kann flexibel darauf reagieren.

Wenn Rahmen in einer bestimmten Reihenfolge transportiert werden müssen, spricht man von einer Konversation. Rahmen einer Konversation werden immer über den gleichen Port gesendet und in dieser Reihenfolge den höheren Protokollschichten zugeführt. Es ist aber durchaus erlaubt, daß Rahmen einer Konversation durch Rahmen einer anderen Konversation unterbrochen werden, solange die Reihenfolge beibehalten wird.

Es stellt sich nun die Frage, wie eine Konversation im Falle eines Ausfalls eines der Links in der Aggregation weitergeführt werden kann. Die Prozedur wird folgendermaßen durchgeführt:

1. Wähle einen neuen Link für die Konversation aus.
2. Beende die Übermittlung von Rahmen der unterbrochenen Konversation. Alle eingetroffenen Rahmen der alten Konversation werden auf der Gegenseite verworfen.
3. Starte einen Zeitgeber, der es erlaubt, alle bereits abgeschickten Rahmen einer Konversation auf der Gegenseite einzutreffen.
4. Ist der Link der fehlerhaften Konversation noch in Takt, so wird ein Marker gesendet, der Anzeigt, daß keine weiteren Daten dieser Konversation auf dem Segment übermittelt werden.
5. Warte bis die Gegenstelle eine Antwort auf den Marker sendet. Sobald die Antwort eintrifft wird der Wechsel auf das andere Segment durchgeführt.
6. Fahre mit der Rahmenübermittlung fort. Die Gegenstelle weiß, daß die Konversation auf einem anderen Segment eintrifft, da die Konversation durch die Nachrichten vollständig beschrieben wird.

# Kapitel 15

## Routing und Switching

Aufgabe dieses Tutorials ist es, die Kommunikation zwischen Geräten eines LANs zu erläutern. Doch es ist auch wichtig zu verstehen, wie Router den Verkehr von und zu LANs transportieren. Das ist umso wichtiger, da immer mehr Layer-2/3-Switches den Weg in das LAN finden. Sie schalten Verkehr innerhalb des LANs auf Layer 2 um und leiten anderen Verkehr auf Layer 3 weiter.

In diesem Kapitel wollen wir uns auf die Funktionalität von Routern konzentrieren und beleuchten, welche Rolle sie bei der Datenübermittlung zwischen entfernten LANs spielen.

### 15.1 Routing – Ein Überblick

Jeder Netzwerkprotokollstapel enthält ein Adressierungsmechanismus und einen Satz von Funktionen, die für die Datenübermittlung verwendet werden. Router leiten Daten in das Zielnetzwerk weiter und durchqueren dabei oftmals mehrere andere Netzwerke und Router.

Viele Protokolle sind im Laufe der Zeit entstanden und alle haben eine Erfolgsgeschichte zu erzählen. Die wichtigsten sind: IBMs SNA, DEC's DECnet, Novell's IPX/SPX und Apple's AppleTalk. Doch als offenes Protokoll hat sich nun IP durchgesetzt, das Layer 3 des TCP/IP-Protokollstapels repräsentiert. Da es noch immer viele traditionelle Installationen gibt, reichen einige Router auch diese Protokolle weiter.

Aus der Tatsache, daß ein Router nicht nur ein bestimmtes Protokoll unterstützt, geht hervor, daß Router in der Lage sind unterschiedliche Architekturen und Medien miteinander zu verbinden. Abbildung 15.1 zeigt, wie zwei Netzwerke, eines mit Token Ring ein anderes mit Ethernet, über einen Router miteinander verbunden werden können. Aus Gründen der Redundanz sind hier zwei Router mit je drei Interfaces vorhanden. Fällt ein Link aus, greift eine andere Route und wird automatisch vermittelt.

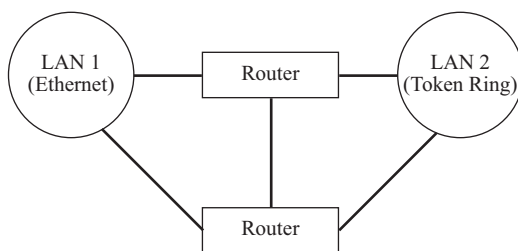


Abbildung 15.1: Routing zwischen zwei Netzwerken mit unterschiedlichen Architekturen.

Das IP Protokoll stellt ein globales Adressierungsschema zur Verfügung. Jedes Netzwerk verfügt über eine Adresse (die Network ID), so daß Daten zwischen Routern nicht auf Basis der IP-Adresse eines Hosts, sondern der Netzwerkadresse, der der Host angehört, weitergeleitet werden.

### 15.1.1 IP-Adressen und MAC-Adressen

Das Adressierungsschema des Network-Layers ist hierarchisch. Es kann wie ein Telefonnetzwerk verstanden werden. Zuerst kommt die Ländervorwahl, dann die Ortsvorwahl und anschließend die Nummer des eigentlichen Anschlusses. So ähnlich funktioniert auch IP. Eine IP-Adresse besteht immer aus einem Netzwerk- und einem Hostanteil. Adressen mit dem gleichen Netzwerkanteil gehören demnach auch zum gleichen logischen Netzwerk.

Das ist nur eine verallgemeinerte Sichtweise zur Vermittlung der Funktionalität von IP und ist keineswegs die Regel. Außerdem wird auch noch das Subnetz angegeben, das einen Teil des Netzwerkes repräsentiert. Routing ist ein sehr komplexes Thema und liegt nicht im Rahmen dieses Textes.

Wenn die Präfixe (der Netzwerkanteil) der Quell- und Netzwerkadressen übereinstimmen, können die Daten direkt zugestellt werden, denn beide Knoten befinden sich im gleichen Netzwerk. Der nächste Schritt in der Abwicklung der Kommunikation besteht darin, die MAC-Adresse des Zielsystems herauszufinden. Das Verfahren wird mit Hilfe des *Address Resolution Protocol* (ARP) abgewickelt.

Die MAC-Adresse finden wir heraus, indem wir einen Broadcast absetzen (*ARP-Request*), der alle Rechner eines LANs fragt, ob sie eine passende MAC-Adresse für die IP-Adresse des Zielsystems verfügen. Passen beide zusammen, gibt der jeweilige Host Antwort (*ARP-Reply*), so daß alle Daten, die für dieses Systems bestimmt sind, zugestellt werden können. Im weiteren Verlauf muß die MAC-Adresse nicht mehr abgefragt werden, denn der Router speichert die MAC-/IP-Adresspaare in einer Tabelle (*ARP-Cache*).

### 15.1.2 Routing außerhalb des LANs

Passen die Netzwerkadressen der Quelle und des Ziels nicht zusammen, so werden die Dateneinheiten an den Router des LANs gesendet, der es mit der Außenwelt verbindet. Der Router weiß, welche Schritte zu unternehmen sind. Die Prozedur läuft etwa so ab:

1. Die Station, die Daten senden möchte versucht, die IP-Adresse des Ziel herauszufinden und, wenn das fehlschlägt, die Daten an den Router sendet, dessen IP-Adresse entweder vom Administrator festgelegt wurde oder über einen Automatismus dynamisch festgelegt wird.
2. Falls notwendig, wird ein ARP-Request durchgeführt, um die MAC-Adresse des Routers abzufragen.
3. Die Station verpackt das IP-Datagramm in einem Ethernet-Rahmen mit der Ziel-MAC-Adresse des Routers und übermittelt es an den Router.
4. Der Router verarbeitet das Paket, wandelt es um und überträgt es an das Ziel (oder einen Router, der auf dem Weg zum Ziel liegt).

Der Router akzeptiert das eingehende Ethernet-Paket, denn die Ziel-MAC-Adresse entspricht dem Eingangsport des Routers. Nachdem die FCS (frame check sequence) geprüft wurde, liest der Router das Ethertype-Feld aus, um das Protokoll, das in dem Rahmen transportiert wird zu identifizieren. In unserem Fall handelt es sich natürlich um ein IP-Datagramm.

Der IP-Header des Datagramms enthält die Quell- und Ziel-IP-Adresse. Der Router versucht herauszufinden, welche Route (also über welchen Ausgangsport) das Datagramm gesendet werden muß, um zu seinem Ziel zu erreichen. Das erledigt er, in dem er seine Routing-Tabelle befragt. Anschließend wird das Datagramm neu verpackt, denn es benötigt eine neue Ziel-Adresse. Das Format des neuen Rahmens hängt von der Architektur des Netzwerks ab, das an dem Ausgangsport des Routers anliegt. Beispielsweise könnte es sich ja um ein Token Ring Netzwerk handeln, das ein anderes Rahmenformat als Ethernet verwendet. Die Ziel-MAC-Adresse ist aber immer die des nächsten Systems, das bis zum Ziel durchquert werden muß.

## 15.2 Wie funktioniert Routing?

Der Schlüsselschritt beim Layer-3-Routing (ja, es gibt noch eine andere Schicht auf der Routing-Entscheidungen getroffen werden können, wie wir weiter hinten sehen werden) ist der Blick in die Routing-Tabelle. Ein Router verwaltet eine Routing-Tabelle, die alle Netzwerkadressen der Ausgangsports enthält.

Das Routing wird durch den Router in folgenden Schritten durchgeführt:

1. Nachdem der Header und Trailer des Rahmens abgelegt wurden, wird das IP-Datagramm der IP-Schicht im TCP/IP-Protokollstapel zugeführt, wo er weiter verarbeitet wird.
2. Der IP-Header enthält neben der Ziel- und Quell-IP-Adresse auch eine Prüfsumme, die nachgerechnet und mit der des Datagramms verglichen wird. Stimmen beide nicht überein, wird das Datagramm verworfen.
3. Der IP-Header enthält auch eine maximal zulässige Anzahl von Zwischenstationen (*Hops*), die an jedem Hop reduziert wird.
4. Die Prüfsumme des Headers wird nach der Umformung des IP-Datagramms erneut berechnet und im Header plaziert.
5. Ein Blick in die Routing-Tabelle verrät dem Router, welcher Ausgangsport für die Weiterleitung gewählt werden muß. Dieser Port zeigt auf den nächsten Hop auf dem Weg zum Ziel.
6. Das Datagramm wird anschließend in einem neuen Rahmen verpackt und über den Ausgangsport auf den Weg zum Ziel geschickt.

Verglichen mit einem Layer-2-Switch ist das eine ganze Menge Arbeit.

### 15.2.1 Aufbau einer Routing-Tabelle

Eine Routing-Tabelle kann manuell aufgebaut und durch Router selbst erstellt werden. Ersteres Verfahren ist für sehr kleine Netzwerke durchaus praktikabel aber für komplexe Großnetze nicht durchführbar.

Router tauschen untereinander Nachrichten aus, die Informationen über die erreichbaren Adressen enthalten. Jeder Router nutzt diese Informationen, um seine Routing-Tabelle aufzubauen. Sie wird jedesmal aktualisiert, wenn benachbarte Router Änderungen mitteilen. Das passiert beispielsweise, wenn eine Verbindung nicht mehr verfügbar ist, so daß die Routeninformation auf eine alternative Route umgestellt werden kann.

Die Router kommunizieren untereinander auf Basis eines Routing-Protokolls. Es definiert den Inhalt und das Format der Nachrichten, sowie die Ereignisse, bei denen bestimmte Nachrichten abgesetzt werden sollen.

Im Laufe der Zeit haben sich mehrere Routing-Protokolle für lokale Netzwerke herausgebildet:

#### **Routing Information Protocol, Version 1 (RIP1)**

Zwar nicht mehr aktuell, aber noch immer im Einsatz, ist es sehr gut für kleinere Netzwerke geeignet. Leider kann die Router-zu-Router-Kommunikation zu einem großen Broadcast-Aufkommen führen und damit das Netzwerk unnötig belasten. Außerdem kann die Übernahme von Datenverkehr eines fehlerhaften Pfades auf einen anderen mehrere Minuten dauern.

#### **Routing Information Protocol, Version 2 (RIP2)**

Verbesserte Version von RIP1. Weist mehr Flexibilität bei der Adressierung und Authentifizierung der Router-zu-Router-Kommunikation auf. Die vielen Broadcasts der ersten Version können optional durch Multicasts ersetzt werden.

#### **Open Shortest Path First (OSPF)**

RIP1 und RIP2 überlegen. Skaliert auch in großen Netzwerken sehr gut und erlaubt Routern den Aufbau einer sehr detaillierten Netzwerktopologie. Aktuell wird Version 2 eingesetzt, aber Version 1 ist noch immer im Einsatz. Das Datenaufkommen der Router-zu-Router-Kommunikation ist sehr gering und die Datenübernahme auf einen anderen Pfad dauert nur Sekunden.

**Internet Gateway Routing Protocol (IGRP)**

Erste Version von Cisco's proprietären Routing-Protokoll. Das Datenaufkommen der Router-zu-Router-Information kann aufgrund der häufigen Broadcasts sehr hoch sein. Vorteilhaft ist, daß Router detaillierte Informationen über die Routen erhalten, wie etwa Geschwindigkeit, Durchsatz und Zuverlässigkeit.

**Enhanced Internet Gateway Routing Protocol (EIGRP)**

Aktuelle Version von Cisco's proprietären Protokoll. Datenaufkommen durch Routing-Informationen relativ klein. Die Datenübernahme auf einen anderen Pfad dauert nur wenige Sekunden.

Unterschiedliche Netzwerke eines Unternehmens können durchaus unterschiedliche Routing-Protokolle verwenden, denn die meisten Hersteller implementieren gleich mehrere Routing-Protokolle.

Das Routing über die Grenzen des lokalen Netzwerkes hinaus, wird durch zwei andere Protokolle realisiert:

**Exterior Gateway Protocol (EGP)**

Veraltet, aber noch immer teilweise im Einsatz. Es teilt uns mit, an welchem Punkt ein bestimmtes Netzwerk mit dem Internet verbunden ist.

**Border Gateway Protocol (BGP)**

Grundlegendes Protokoll zum Aufspüren von Pfaden über das Internet hinaus, die zu einem bestimmten Netzwerk führen. Aktuell ist Version 4 im Einsatz.

Zusätzlich gibt es noch vier weitere Routing-Protokolle für das Multicasting:

**Distance Vector Multicast Routing Protocol (DVMRP)**

Bietet einen effektiven Mechanismus für das Routing von Nachrichten zu Multicast-Gruppen im Internet. DVMRP ist ein *Interior Gateway Protocol* (IGP) einsetzbar in Routern, aber nicht für die Kommunikation zwischen Routern. Es basiert auf RIP und verwendet den *Truncated Reverse Path Broadcasting Algorithmus*. Die DVMRP-Pakete sind in IP-Datagrammen des IGMP-Protokoll gekapselt.

**Dense Mode Protocol Independent Multicast (DM-PIM)**

Entworfen für Multicast-LANs implementiert DM-PIM den gleichen Mechanismus wie das DVMRP. Der Unterschied zwischen DVMRP und PIM-DM ist, daß letzteres protokollunabhängig ist. PIM-DM kann die Routing-Tabelle eines anderen Unicast-Protokolls verwenden, um RPF-Checks (*Reverse Path Forwarding*) durchzuführen.

**Sparse Mode Protocol Independent Multicast (SM-PIM)**

Optimiert für das Routing von *Sparse Groups* (verteilte Gruppen). Kann Routeninformationen anderer Protokolle wie RIP oder OSPF verarbeiten, um Multicast-Traffic weiterzuleiten.

**Multicast OSPF (MOSPF)**

Erweiterung von OSPF2 zur Unterstützung von Multicast-Routing. Rückwärtskompatibel und in der Lage mit Routern ohne MOSPF-Unterstützung zu kommunizieren. Die Multicast-Informationen sind in *OSPF Link State Advertisements* (die Dateneinheiten von OSPF) enthalten.

In der heutigen Zeit unterstützt quasi jeder Router IP, einige warten zusätzlich noch mit traditionellen Protokollen wie IPX/SPX oder DECnet auf.

Die Eigenschaft der Router, den Verkehr auf Basis einer Tabelle umzuleiten, machen sich viele Netzwerk-Administratoren zu Nutze um minimale Sicherheitsmechanismen auf Ebene 3 des Protokollstapels zu etablieren, denn eine geschickt gewählte Routing-Tabelle kann ganze Netzwerk von der Außenwelt abgrenzen.

Desweiteren bieten viele Router Unterstützung von Prioritäten und *Quality of Service* (QoS), so daß einige Pakete anders behandelt werden als die übrigen.

## 15.3 Router-Architektur und MPLS

Router können auch als Layer-3-Switches bezeichnet werden, denn letztendlich tun sie nichts anderes als Layer-2-Switches, mit dem Unterschied, daß sie auf Layer 3 operieren. Router arbeiten sehr schnell, da sie hochspezialisierte ASICs enthalten, die oftmals parallel oder hintereinander ausgelegt sind.

Meist enthalten Router auch einen zentralen Prozessor, der folgende Aufgaben übernimmt:

- Erzeugung ausgehender Routing-Nachrichten für die Router-zu-Router-Kommunikation, Verarbeitung eingehender Nachrichten von Routing-Protokollen und Aktualisierung der Routing-Tabellen.
- Ausstattung der ASICs mit Kopien der Routing-Tabelle oder Aktualisierung, falls notwendig.
- Bereitstellung einer Benutzerschnittstelle zur Administration des Gerätes.
- Betrieb eines SNMP-Agents und Unterstützung von Remote Monitoring (RMON).

Viele Hersteller beschreiben die Leistungsfähigkeit ihrer Geräte auf Basis der Latenz und Anzahl der Paketweiterleitungen pro Sekunde. Zum einen ist das Wort *Paket* eine ungenaue Bezeichnung, denn es kann für einen Layer-2-Rahmen oder eine beliebige Dateneinheit der dritten Schicht beschreiben. Zum anderen hängt die Latenz von der Größe der zu verarbeitenden Pakete ab. Als Latenz wird die Zeit bezeichnet, die benötigt wird, um ein Paket vom Empfang bis zur Weiterleitung zu verarbeiten. Je kleiner die Pakete, desto geringer ist die Latenz.

Das Routing wird benötigt, um Datenverkehr nach einem globalen Schema weiterzuleiten. MAC-Adressen haben dabei nur eine lokale Bedeutung. Router benötigen für die Weiterleitung immer etwas Zeit, evtl. sogar einige Mikrosekunden. Im Zeitalter der Multigigabit-Netzwerke ist das eindeutig zu viel.

Hersteller haben sich darauf hin Gedanken über einen Mechanismus gemacht, der es erlaubt, Routing-Entscheidungen schneller zu treffen. Die IETF vermied durch das rechtzeitige Einsetzen einer Arbeitsgruppe, einen Wildwuchs proprietärer Mechanismen. So entstand der Standard *Multiprotocol Label Switching* (MPLS) und basiert auf folgenden Überlegungen:

- Klassifizierung aller Pakete, die in einem Router eintreffen und automatische Weiterleitung von Paketen gleicher Klassen.
- Weise eingehenden Paketen eine *Forwarding Equivalence Class* zu und stelle ihnen ein Label voran, das einen numerischen Bezeichner der Klasse enthält.
- Leite das Paket an einen entfernten Router auf Basis des Labels weiter. Es finden keine weitere Verarbeitung des Pakets statt.

Durch das Voranstellen eines Labels, genügt es, das Label zu untersuchen, um eine Routing-Entscheidung zu treffen. Aus diesem Grund bildete sich auch die Redewendung „route once, switch many“ heraus.

Vorteile dieser Technik sind:

- Einsetzbar mit allen Layer-3-Protokollen.
- Setzt auf alle physikalischen Architekturen auf.
- Datenverkehr kann einer *Forwarding Equivalence Class* zugeordnet werden. Die Klasse wird auf Basis unterschiedlicher Faktoren, wie Priorität oder ähnlichem, gebildet.
- *Label Switching* ist eine sehr einfache Funktionalität und kann ohne viel Leistungsverlust implementiert werden.



# Literaturverzeichnis

[1] IEEE, <http://www.ieee.org>

[2] Ethernet-Typen, <http://standards.ieee.org/regauth/ethertype/type-pub.html>