

## 7 Layer-2-Verfahren: Load Balancing

Dieses Kapitel behandelt Lastverteilungsverfahren auf Ebene 2. Dies sind im Wesentlichen Link Aggregation als symmetrische Lastverteilung und Adapter Load Balancing als asymmetrische Lastverteilung.

### 7.1 Link Aggregation

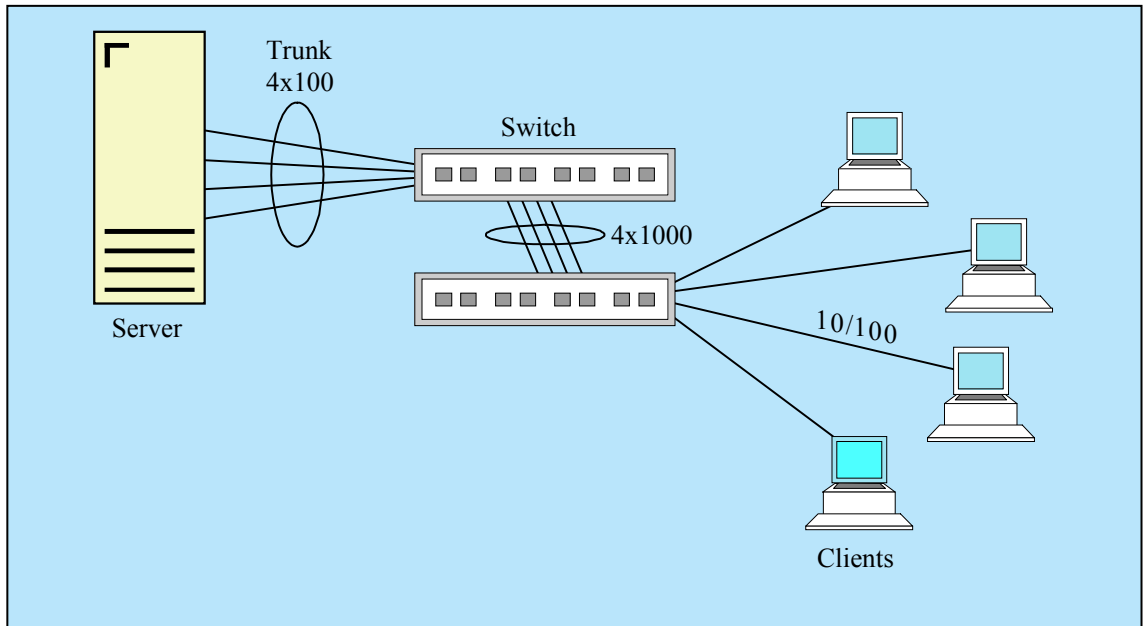
Zur Beseitigung von Spanning-Tree-Nachteilen entwickelten Internetworking-Hersteller so genannte Trunking-Konzepte. Dabei werden mehrere physikalische Interfaces und die angeschlossenen Leitungen zu einer logischen Verbindung, einem "Trunk" gebündelt. Da der Begriff "Trunk" nicht eindeutig ist, sondern auch in der VLAN-Technik und bei MPLS Verwendung findet, wurde von der Standardisierung als formaler Begriff "Link Aggregation" gewählt. Im Weiteren wird dieser Begriff verwendet, um Missverständnisse zu vermeiden.

Die Link Aggregation arbeitet im Normalbetrieb lastverteilt über alle Verbindungen und schaltet im Fehlerfall den Datentransport im Millisekunden- bis Sekundenbereich auf die verbleibenden Leitungen um. Allerdings sind nur parallele Punkt-zu-Punkt-Verbindungen, Interfaces mit gleicher Übertragungsrates und full-duplex Betrieb erlaubt, um komplexe Rechenverfahren für die Lastverteilung zu vermeiden und den Overhead möglichst gering zu halten.

Link Aggregation hat sich als Ethernet-Technik etabliert, für FDDI und Token Ring gibt es keinen Standard und kaum Herstellerlösungen.

Herstellereigene Verfahren für Link Aggregation wurden zuerst von Cisco vermarktet (EtherChannel), später von Avaya (damals noch unter dem Namen Lucent), Enterasys (Cabletrons SmartTrunk) und Nortel (Multilink Trunking), 3Coms Link Aggregation kam mit Verspätung hinterher. Es folgten jüngere Hersteller wie Extreme und Foundry.

Link Aggregation ist für verschiedenste Switch-Switch-, Switch-Server- oder Server-Server-Kopplungen einsetzbar (siehe Abbildung 7.1). Die maximale Anzahl paralleler Verbindungen liegt je nach Lösung zwischen 4 bis 16 Leitungen bei Switches und zwei bis 32 Leitungen bei Server-Interfaces. Damit erhöht sich die Kapazität einer Backbone-Kopplung z.B. von 1-Gigabit-Ethernet auf 4- oder 16-Gigabit-Ethernet. Beim Einsatz von 100-Mbit-Verbindungen ergibt sich immerhin die Möglichkeit, den Backbone auf bis zu 1,6 Gbit/s aufzurüsten und zwar bei einer maximalen Entfernung von bis zu 2 km und Multimodefasern als Kabelmedium. Dies ist in Geländebereichen sehr hilfreich, wo keine Singlemode-Glasfaser verfügbar ist und die Entfernungen die Längenrestriktion von 550 m für Gbit-Ethernet überschreiten.



**Abbildung 7.1: Einsatzszenario für Link Aggregation in Ethernet-Umgebungen**

## 7.2 Der Standard IEEE 802.3ad - IEEE 802.3/2002

### 7.2.1 Überblick

Der Standard für Link Aggregation (Link Aggregation) ist seit Mitte 2000 unter IEEE 802.3ad verabschiedet und inzwischen in den übergeordneten Ethernet-Standard IEEE 802.3 von 2002 als "Clause 43" eingearbeitet. Er ist sowohl für NICs als auch für Switches implementierbar und bietet die Möglichkeit ohne Lizenzprobleme und Herstellerallianzen (siehe hierzu das Kapitel "Adapter Load Balancing") kompatible Leitungsbündelung zwischen Server-Adapttern, RZ-Switches oder Backbone-Switches zu schalten. Im Wesentlichen adaptiert der Standard existierende Herstellerlösungen und erweitert sie um eine automatische Konfiguration und Kontrolle von Link Aggregationen.

Die aggregierten Verbindungen (Links) werden im Standard *Link Aggregation Group* (LAG) genannt. Neu im Vergleich zu Herstellerlösungen ist, dass die Link Aggregation nicht manuell konfiguriert werden muss, sondern via Link Aggregation Control Protocol (LACP) von den beiden beteiligten Geräten automatisch ausgehandelt werden kann. Hierfür wird eine Link Aggregation Kontrolle (LAC) als Protokollsoftware zwischen MAC-Kontrollebene und MAC-Clientinterface eingefügt - wie in Abbildung 7.2 gezeigt - und steuert über das LACP, welche und wie viele Links zu einer LAG zusammengeschaltet werden. Unterstützen zwei benachbarte Switches Link Aggregation, wird diese automatisch aktiv und nutzt in einem maximalen Umfang vorhandene Leitungen parallel und lastverteilt.

Allerdings ist diese automatische Konfiguration bei den Herstellern aktuell noch kaum implementiert, d.h. Link Aggregierungsgruppen müssen nach wie vor manuell konfiguriert werden und die Anwender dürfen auf neue Software-Releases warten.

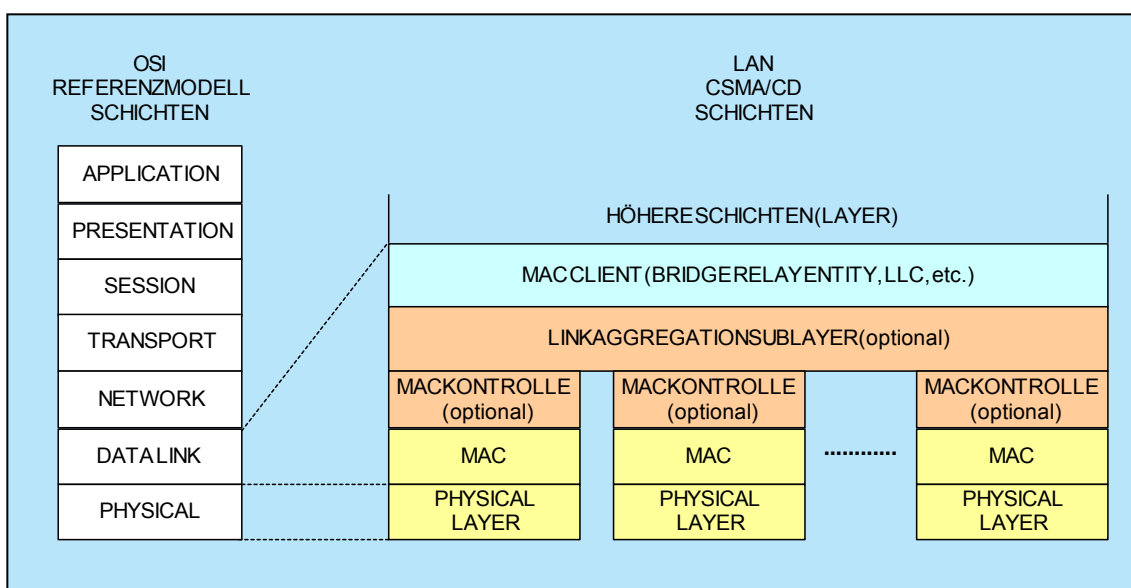


Abbildung 7.2: Link Aggregation Protokollschicht für IEEE 802.3

### 7.2.2 Ziele, Anforderungen, Einschränkungen

In diesem Kapitel werden Ziele, Anforderungen und Einschränkungen für Link Aggregationen beschrieben, wie sie der Standard sieht.

#### Ziele

Ziele des Link-Aggregierungs-Standards sind:

**Bandbreiten-Erhöhung:** Die Kapazitäten mehrerer Verbindungen (Links) werden zu einer gemeinsamen logischen Verbindung gebündelt.

**Lineare Steigerung:** Die Bandbreite kann in Vielfachen etablierter Standards erhöht werden (2mal oder 4mal 100 Mbit/s), nicht allein in Sprüngen einer Größenordnung (10, 100, 1000 Mbit/s).

**Erhöhung der Verfügbarkeit:** Der Ausfall einer Verbindung innerhalb der Link Aggregation führt zu keinem Ausfall des MAC-Dienstes.

**Lastverteilung:** Der MAC-Verkehr kann über mehrere Verbindungen verteilt werden.

**Automatische Konfiguration:** Fehlt eine manuelle Konfiguration, so konfiguriert sich automatisch eine geeignete Menge von Link Aggregierungsgruppen gemäß folgender Maximalbedingung: Falls mehrere Verbindungen aggregiert werden können, werden sie auch alle aggregiert.

**Schnelle Konfiguration und Rekonfiguration:** Ändert sich die physikalische Topologie, wird die Link Aggregation schnell zu einer neuen Konfiguration konvergieren, typischerweise innerhalb 1 Sekunde oder schneller.

**Deterministisches Verhalten:** Die aktive Konfiguration ergibt sich unabhängig von der Reihenfolge, in der die Ereignisse auftreten, die zu einer Änderung führen. Dieselbe Ausgangssituation führt stets zu derselben Endkonfiguration.

**Geringes Risiko für Frame-Duplizierung oder Änderung der Reihenfolge:** Sowohl im konvergierten Zustand als auch bei einer Rekonfiguration ist der Anteil duplizierter oder reihenfolge-vertauschter Frames sehr niedrig.

**Unterstützung der existierenden MAC-Dienste:** Für höhere Protokolle sind keinerlei Änderungen erforderlich.

**Rückwärtskompatibilität:** Verbindungen mit Komponenten, die Link Aggregationen nicht unterstützen, arbeiten als normale Ethernet-Verbindungen.

**Anpassung unterschiedlicher Funktionen und Einschränkungen:** Komponenten mit unterschiedlichen Hardware- und Software-Möglichkeiten und Einschränkungen auf beiden Seiten werden maximal möglich aneinander angepasst ("größter gemeinsamer Nenner").

**Keine Änderung des IEEE 802.3 Frameformats:** Link Aggregation nimmt keine Einträge in ein vorhandenes Frame vor oder fügt Protokollfelder hinzu.

**Unterstützung von Netzwerkmanagement:** Der Standard definiert geeignete Parameter (Management Objekte) für Konfiguration, Monitoring und Kontrolle einer Link Aggregation.

### **Anforderungen**

Als charakteristische Anforderungen eines MAC-Dienstes müssen erfüllt werden:

- Einhaltung der Paketreihenfolge für einzelne Sessions,
- deterministische Konfiguration und Arbeitsweise einer Link Aggregation,
- möglichst gleichmäßige Lastverteilung,
- schnelle Umschaltung im Fehlerfall.

Einhaltung der Paketreihenfolge bedeutet: alle Frames einer Session ("Conversation") werden über dieselbe physikalische Verbindung transportiert, um eine Vertauschung der Reihenfolge zu vermeiden. Ein Beispiel hierfür, das der Standard informell beschreibt, ist die Verteilung von Frames auf verschiedene aggregierte Links anhand der MAC-Adressen. Weitere mögliche Verteilalgorithmen sind im Detail der jeweiligen Implementierung freigestellt, sofern die zuvor genannten Anforderungen eingehalten werden.

Zur gleichmäßigen Lastverteilung kann ein Switch verschiedene Konfigurationsmöglichkeiten unterstützen, zur Auswahl stehen in jeder Richtung unabhängig die MAC-Zieladresse, MAC-Quelladresse, IP-Zieladresse, IP-Quelladresse, TCP/UDP-Portnummern etc.. Für verschiedene Konfigurationen sind unterschiedliche Verfahren optimal: Bedient ein einzelner Server viele Clients, macht die Aufteilung nach MAC-Zieladressen von den Clients zum Server keinen Sinn, die Aufteilung nach Quelladressen jedoch schon. Vom Server zu den Clients macht umgekehrt nur die Aufteilung nach MAC-Zieladressen Sinn (siehe Abbildung 7.3).

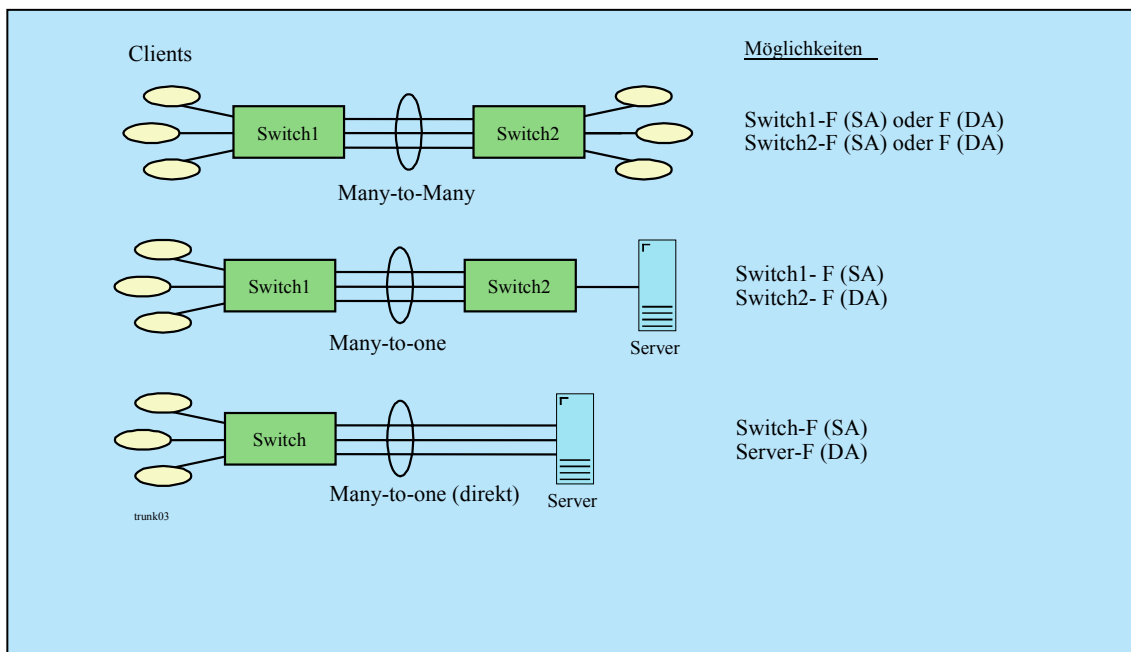


Abbildung 7.3: Aufteilung von Frames anhand von MAC-Adressen

### Einschränkungen

Für Link Aggregationen gelten folgende Einschränkungen:

- parallele Leitungsführung (alle Verbindungen verlaufen zwischen genau zwei Komponenten),
- nur full duplex, kein half duplex,
- gleiche Kapazität für alle Verbindungen einer Link Aggregation,
- ausschließlich Ethernet, der Standard unterstützt keine anderen MAC-Dienste wie Token Ring, FDDI, ATM,
- keine Multipunkt-Aggregation, der Standard unterstützt keine Aggregation über mehr als zwei Systeme.

### 7.2.3 Funktionsweise der IEEE 802.3 Link Aggregation

Welche Funktionen sind erforderlich, um eine Link Aggregation zu konfigurieren und zu betreiben? Eine Gruppe von Ports muss, manuell oder automatisch, zu einer Link Aggregierungsgruppe (LAG) konfiguriert werden. Die Empfängerlogik muss die ankommenden Frames in einer geordneten Reihenfolge an den darüber liegenden MAC-Dienst weitergeben, die Sendelogik muss die vom MAC-Dienst zum Senden übergebenen Frames möglichst effizient und unter Einhaltung der Reihenfolge auf mehrere Verbindungen (Links) verteilen. Die Verbindungen einer LAG müssen sich gegenseitig überwachen, um einen Ausfall zu erkennen und eine Fehlerumschaltung auf die verbleibenden Links durchzuführen. Umgekehrt muss die zuvor ausgefallene Verbindung wieder mitbenutzt werden, sobald der Fehler behoben ist.

Zur Implementierung der erforderlichen Funktionen wurde eine Architektur mit mehreren Hardware- und Softwaremodulen definiert (siehe auch Abbildung 7.4), die im Weiteren näher beschrieben werden:

**Frame Collection** (Frame Collector, Marker Responder): Übergibt die Frames, die auf den Ports einer LAG empfangen wurden, an den MAC-Dienst.

**Frame Distribution** (Frame Distributor, Marker Generator): Nimmt Frames, die gesendet werden müssen, vom MAC-Dienst entgegen und verteilt sie mit Hilfe eines Lastverteilungsalgorithmus auf einen entsprechenden Port.

**Aggregator Multiplexer** (Aggregator Parser): Der Multiplexer leitet Frame Requests vom Distributor oder Marker Generator einfach an den entsprechenden Ausgangsport weiter. Beim Empfangen von Frames unterscheidet der Aggregation Parser/Multiplexer nach Datenframes und Marker Request/Marker Response und leitet die Frames an das erforderliche Funktionsmodul (Collector, Marker Responder, Marker Generator) weiter.

**Aggregator**: Die Kombination von Distributor, Collector und Multiplexer/Parser wird als Aggregator bezeichnet.

**Aggregation Kontrolle** (inkl. Link Aggregation Control Protocol): Das ist die Implementierung des LACP und handhabt sowohl die Konfiguration als auch die Kontrolle der Link Aggregation.

**Control Multiplexer** (Control Parser): Leitet beim Senden Frames, die vom Aggregator und der Kontrollfunktion kommen, an den gewünschten Ausgangsport weiter. Beim Empfangen von Frames unterscheidet der Control Parser/Multiplexer, ob es sich um LACP-Frames (LACPDUs) oder andere Frames handelt. LACPDUs werden an die Kontrollinstanz weitergegeben, andere Frames an den Aggregator (und dort an die Frame Collection).

Durch die Zwischenschaltung des Link Aggregierungs-Sublayer zwischen MAC-Dienst und physikalischer Schicht wird Folgendes erreicht: Ein MAC-Client kommuniziert mit einer Gruppe von Ports über einen Aggregator. Der

Aggregator seinerseits bietet dem MAC-Client eine Standard-MAC-Dienstschnittstelle. An den Aggregator sind ein oder mehrere Ports gebunden und er ist dafür verantwortlich, einerseits zu sendende Frames vom MAC-Client auf diese Ports zu verteilen und andererseits Frames, die an diesen Ports ankommen, transparent an den MAC-Client weiterzuleiten. Ein Switch kann mehrere Aggregatoren beinhalten, die ihrerseits mehrere MAC-Clients bedienen. Jeder Einzelport ist jederzeit an einen einzelnen Aggregator gebunden. Ein MAC-Client zu einem bestimmten Zeitpunkt wird nur von einem einzelnen Aggregator bedient.

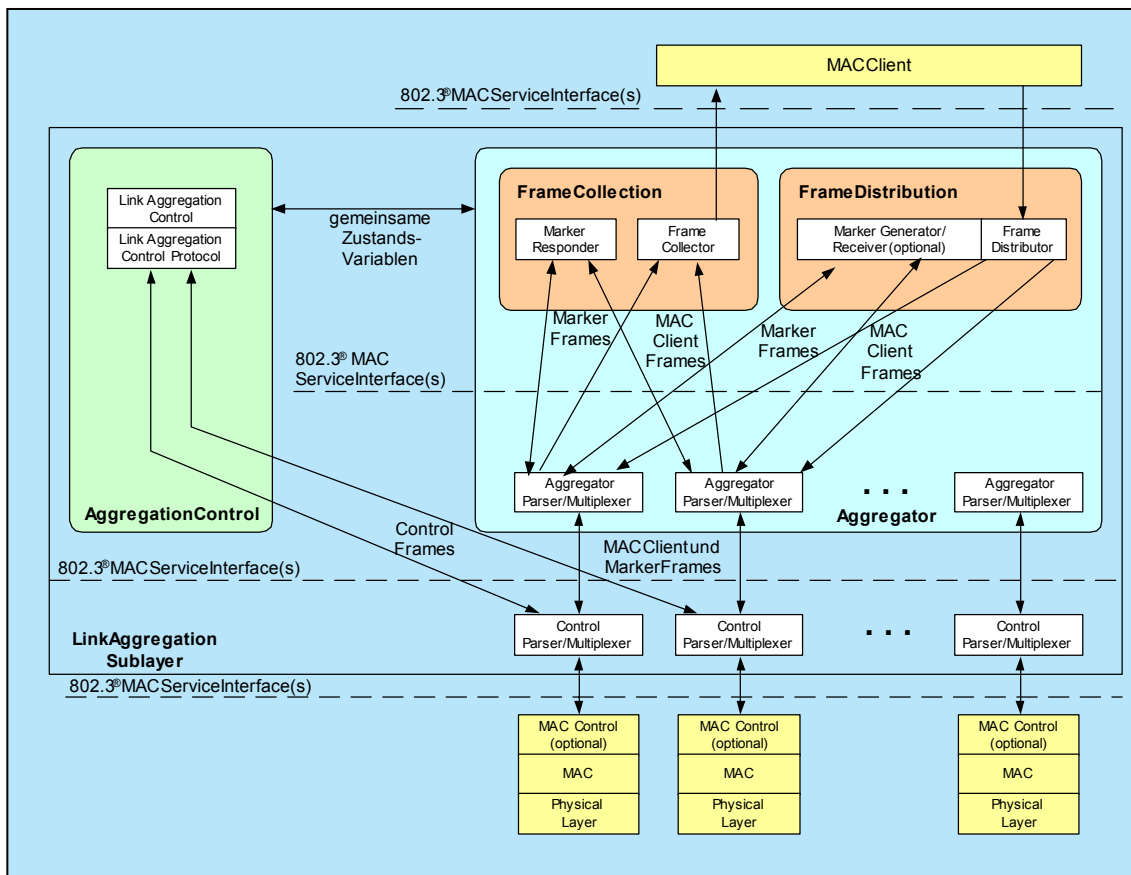


Abbildung 7.4: Blockdiagramm des Link Aggregierungs-Sublayer

Wie die Ports eines Switches oder Servers an Aggregatoren gebunden werden, das regelt die Link Aggregation Kontrolle. Sie entscheidet, welche Links aggregiert werden können, führt die Aggregation durch, bindet die Ports des Switches oder Servers an einen geeigneten Aggregator und überwacht den Betrieb, um zu entscheiden, wann Bedingungen eingetreten sind, die eine Änderung innerhalb einer Aggregation erfordern. Die Entscheidung sowie das Binden von Ports an einen Aggregator können mittels manueller Konfiguration durch einen Netzwerkmanager oder automatisch mittels des Kontrollprotokolls LACP erfolgen. LACP nutzt einen regelmäßigen Informationsaustausch zwischen den Peers einer Link Aggregation (die beiden beteiligten Netzwerkkomponenten), um den Betriebszustand der beteiligten Links zu

überwachen und stets die maximal mögliche Aggregation zwischen einem Komponentenpaar zu erreichen.

Der Distributor stellt sicher, dass alle Frames einer bestehenden Conversation (Session, Flow) an genau einen (Ausgangs-)Port weitergeleitet werden. Der Collector stellt sicher, dass Frames einer Conversation genau in der Reihenfolge, in der sie an einem Port empfangen wurden, an den MAC-Dienst weitergeleitet werden. Diese Einschränkung ist ausreichend, um die Beibehaltung der Frame-Reihenfolge für jede beliebige Anwendung zu garantieren: Für Frames, die nicht zur selben Conversation gehören, ist der Collector nicht an die Reihenfolge gebunden. Er kann z.B. erst zehn Frames von Conversation 1 an den MAC-Dienst weiterleiten und danach zwei Frames von Conversation 2, obwohl die Frames von Conversation 2 zuerst empfangen wurden.

Conversations können innerhalb einer Link Aggregation im laufenden Betrieb auf einen anderen Link umgeschaltet werden, sowohl zur Fehlerumschaltung als auch zur besseren Lastverteilung. Hieraus ergibt sich kein Reihenfolgeproblem, wenn der Distributor bei der Umschaltung eine kurze Wartezeit einhält, so dass alle Frames der Conversation, die bis zur Umschaltung empfangen wurden, inzwischen von der Gegenseite verarbeitet sind.

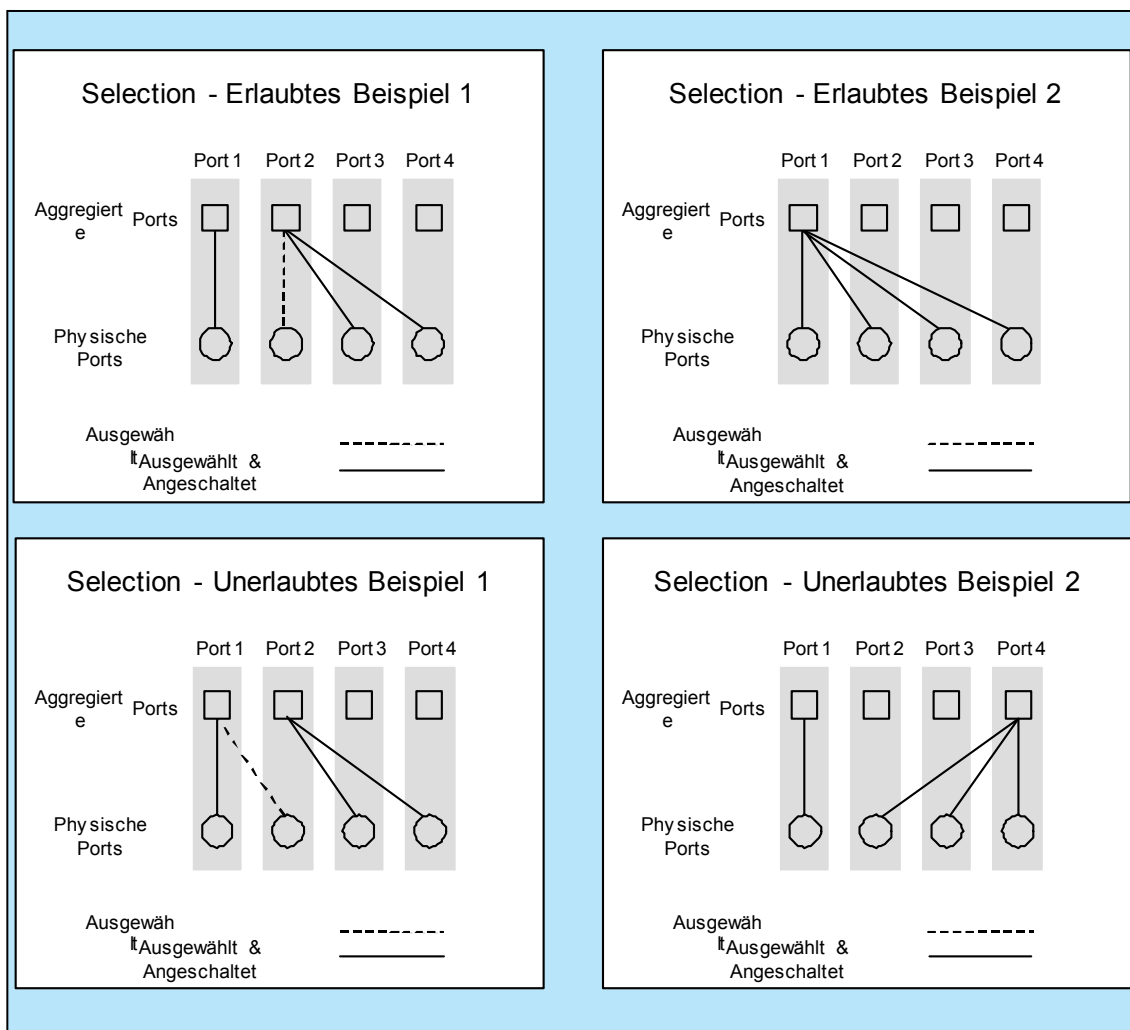
Der Standard schreibt keine festen Verteilalgorithmen für den Distributor vor, sondern überlässt die Details den Herstellern, solange das Verteilverfahren den Standard-MAC-Dienst d.h. die zuvor genannten Anforderungen unterstützt. Allerdings wird informativ im Anhang eine Verteilung auf Basis der MAC-Adressen oder der TCP/UDP-Portnummern vorgeschlagen, die alle wesentlichen Fälle abdeckt. Hierauf wird im Kapitel "Link Kontrolle und Link Aggregation Control Protocol" näher eingegangen.

### **MAC-Adressierung**

Jeder Port erhält eine im gesamten Netzwerk eindeutige MAC-Adresse. Diese MAC-Adresse wird als Source in Frames eingetragen, die vom LACP oder Marker Protokoll generiert und über diesen Port gesendet werden (Achtung: Sowohl LACP als auch Marker Protokoll nutzen als Destination-Adresse den MAC-Multicast 01-80-C2-00-00-02!). Ebenso erhält jeder Aggregator eine netzwerkweit eindeutige MAC-Adresse. Diese Adresse wird von der gesamten Aggregation, also aus der Perspektive des (übergeordneten) MAC-Client als Source-Adresse zum Senden und als Destination-Adresse zum Empfangen genutzt: Erhält der Aggregator ein MAC-Frame vom MAC-Dienst zum Senden, für das der MAC-Dienst keine Source-Adresse spezifiziert hat, so fügt er die Aggregator-MAC-Adresse als Source ein.

Die Aggregator-MAC-Adresse kann die Port-MAC-Adresse eines Ports aus der zugehörigen Link Aggregation sein. Typischerweise wird dies auch so gehandhabt, obwohl die Vorschrift optional ist. Ein Hersteller könnte für eine LAG auch eine logische, "virtuelle" MAC-Adresse nutzen. Daraus ergibt sich: Alle Ports in einem Switch oder Server, die zu einer Link Aggregation

konfiguriert sind, benutzen aus Sicht der Endgeräte oder der Layer-3-Komponenten dieselbe MAC-Adresse, die wir auch Stellvertreteradresse nennen können. Welche das ist, berechnet sich im Regelfall automatisch: Alle Ports eines Switches oder Servers haben eine eindeutige Kennung (Priorität plus MAC-Adresse). Eine als Link Aggregation gebündelte Portgruppe erhält die MAC-Adresse des Ports mit der "höchsten Priorität" d.h. dem numerisch niedrigsten Wert. Dadurch wird ein kompliziertes Verfahren zur Auswahl einer MAC-Adresse für einen LAG vermieden. Das Kapitel "Link Aggregation Kontrolle" geht hierauf näher ein.



**Abbildung 7.5: Erlaubte und nicht erlaubte Aggregation**

Erlaubte und nicht erlaubte Zuweisung einer Stellvertreter-MAC-Adresse für eine Link Aggregation auf Basis des zuvor beschriebenen Auswahlverfahrens ist in Abbildung 7.5 exemplarisch dargestellt. Da eine Link Aggregation mit einer einzigen MAC-Adresse arbeitet, ergibt sich natürlich ein Analyseproblem: Im Fehlerfall ist nicht mehr anhand der MAC-Adresse erkennbar, welche physikalische Leitung Probleme macht. Hierfür muss die Abfrage zusätzlicher Managementparameter in entsprechende Analysetools implementiert werden.