

5. Input/Output

5.1 Festplattenspeicher Technologie

5.1.1 IBM 350 aus dem Jahre 1956

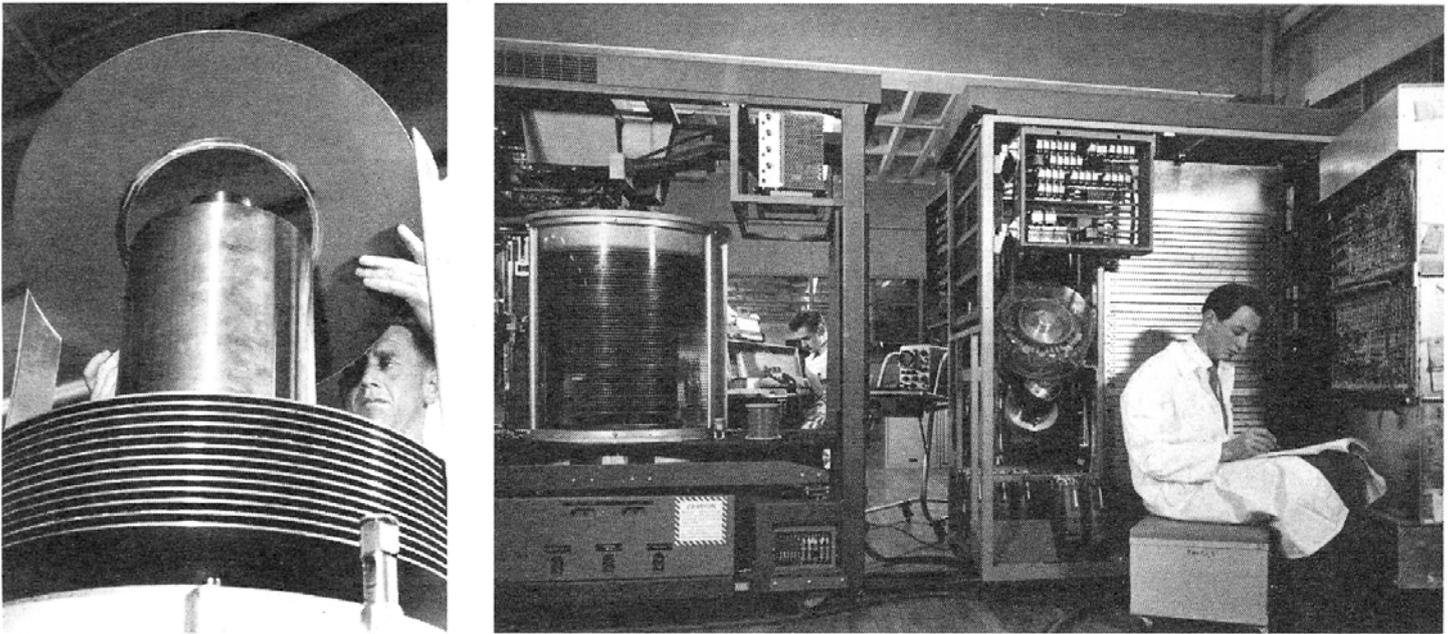


Abb. 5.1.1
IBM 350 Festplattenspeicher von 1956

Am 13. September 1956 stellte IBM das erste Festplattenspeichersystem der Welt vor, die "IBM 350 Plattenspeichereinheit", die zusammen mit dem "IBM 305 RAMAC" (Random Access Method of Accounting and Control)-Rechner ausgeliefert wurde. Der Festplattenspeicher im Kleiderschrankformat wog fast eine Tonne.

Auf einundfünfzig mit Eisenoxyd-beschichteten Platten mit einem Durchmesser von 61 cm (24-Zoll) speicherte die RAMAC sechs Millionen Zeichen. Hätte man damals schon Daten in Form von Bytes abgelegt, entspräche dies einer Kapazität von fünf Megabyte. Die IBM 350 Plattenspeichereinheit verbrauchte 2,5 Kilowatt an elektrischer Energie und wurde seinerzeit auf Mietbasis für 10 000 D-Mark pro Monat vertrieben.

Die modernen Festplattenscheiben verfügen nur über einen Durchmesser von 2,5 Zoll (6,35 cm) oder weniger. Das derzeit stärkste aktuelle Speichersystem der IBM, das System Storage DS8870, hat mit 2,3 PByte Kapazität eine um ca. 400 Millionen mal größere Speicherkapazität als die RAMAC.

5.1.2 Entwicklung der Festplattentechnologie

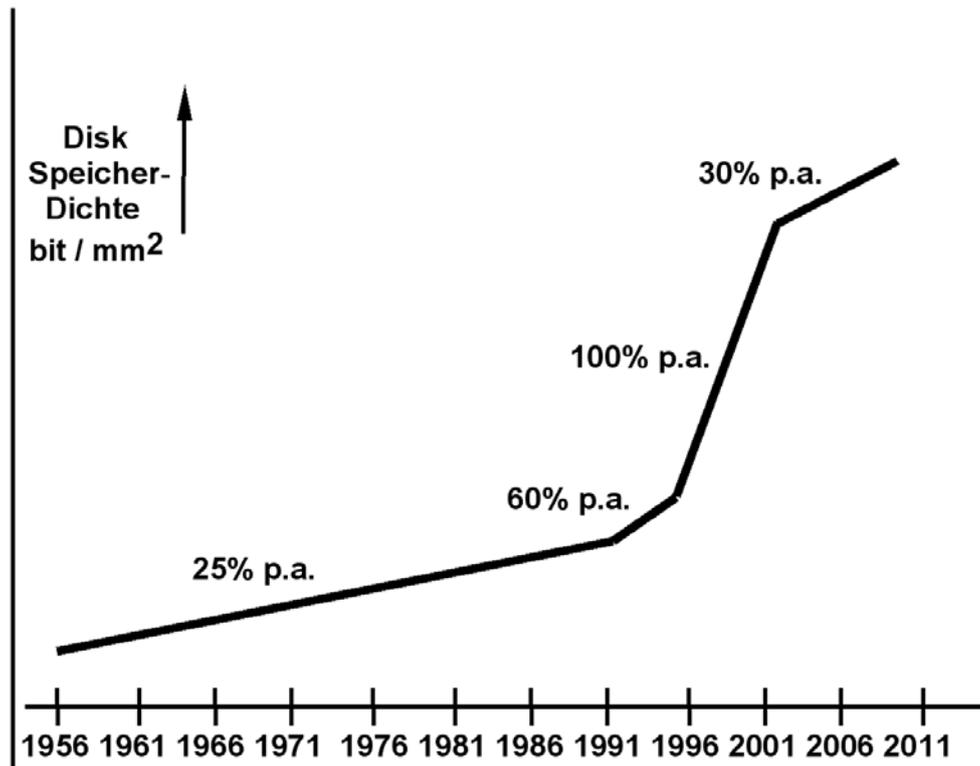


Abb. 5.1.2
Wachstumsrate der Festplatten Speicherdichte

Bei den magnetischen Festplattenspeichern wird der technologische Fortschritt daran gemessen, wie viele Bit auf einem Quadratmillimeter Oberfläche gespeichert werden können.

Dargestellt ist die 50 jährige Entwicklung mit dem jährlichen Wachstum der Plattenspeicherdichte. Auf der y-Achse ist die Speicherdichte mit einer logarithmischen Skala dargestellt. Das bedeutet, ein exponentielles Wachstum wird durch eine gerade Linie dargestellt.

Über mehrere Jahrzehnte betrug das Wachstum konstant etwa 25 % pro Jahr. In den 90er Jahren beschleunigte sich das Wachstum auf etwa 100 % pro Jahr. Ab Anfang der 2000er Jahre verlangsamt sich das Wachstum wieder auf etwa 30 % pro Jahr.

5.1.3 Baugruppen eines Festplattenspeichers

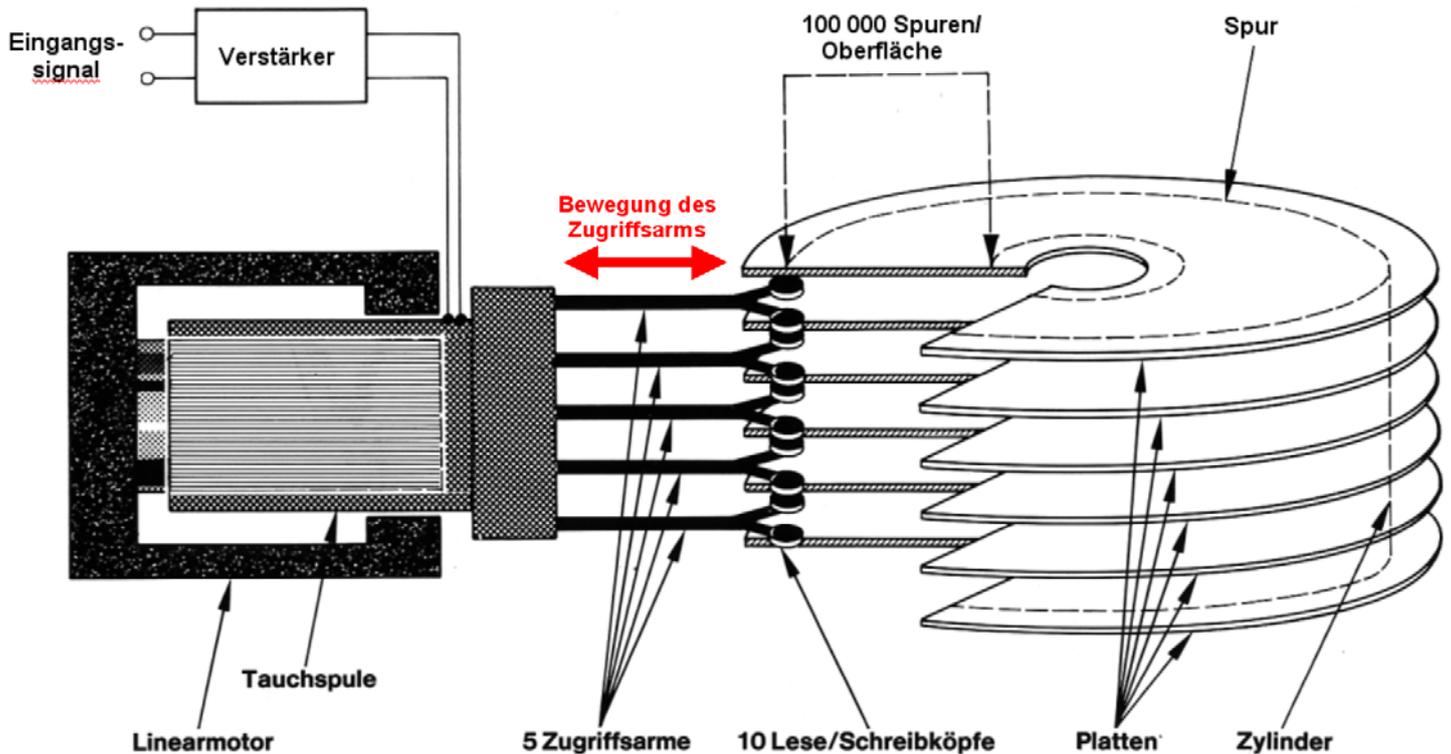


Abb. 5.1.3
Schnitt durch einen Festplattenspeicher

An der ursprünglichen Form eines Festplattenspeichers hat sich in den Jahrzehnten nicht viel geändert. Der Plattenspeicher besteht in der Regel aus mehreren Platten, von denen beide Oberflächen für die Aufzeichnung von Information genutzt wird. Die Information ist in kreisförmigen Spuren (Tracks) auf der Oberfläche angeordnet. Jeder Spur ist eine Spuradresse zugeordnet. Die Spuren selbst sind in Sektoren aufgeteilt. Die meisten Festplatten arbeiten heute mit einer Sektor Größe von 4096 Bytes. Früher waren 512 Byte Sektoren populär.

Pro Oberfläche gibt es mindestens einen Lese/Schreib-Kopf. Die Köpfe für alle Oberflächen sind auf einer gemeinsamen Zugriffsarmstruktur (slider) befestigt, welche alle Köpfe mittels eines Linearmotors (Actuator) auf eine der vielen Spuren bewegt.

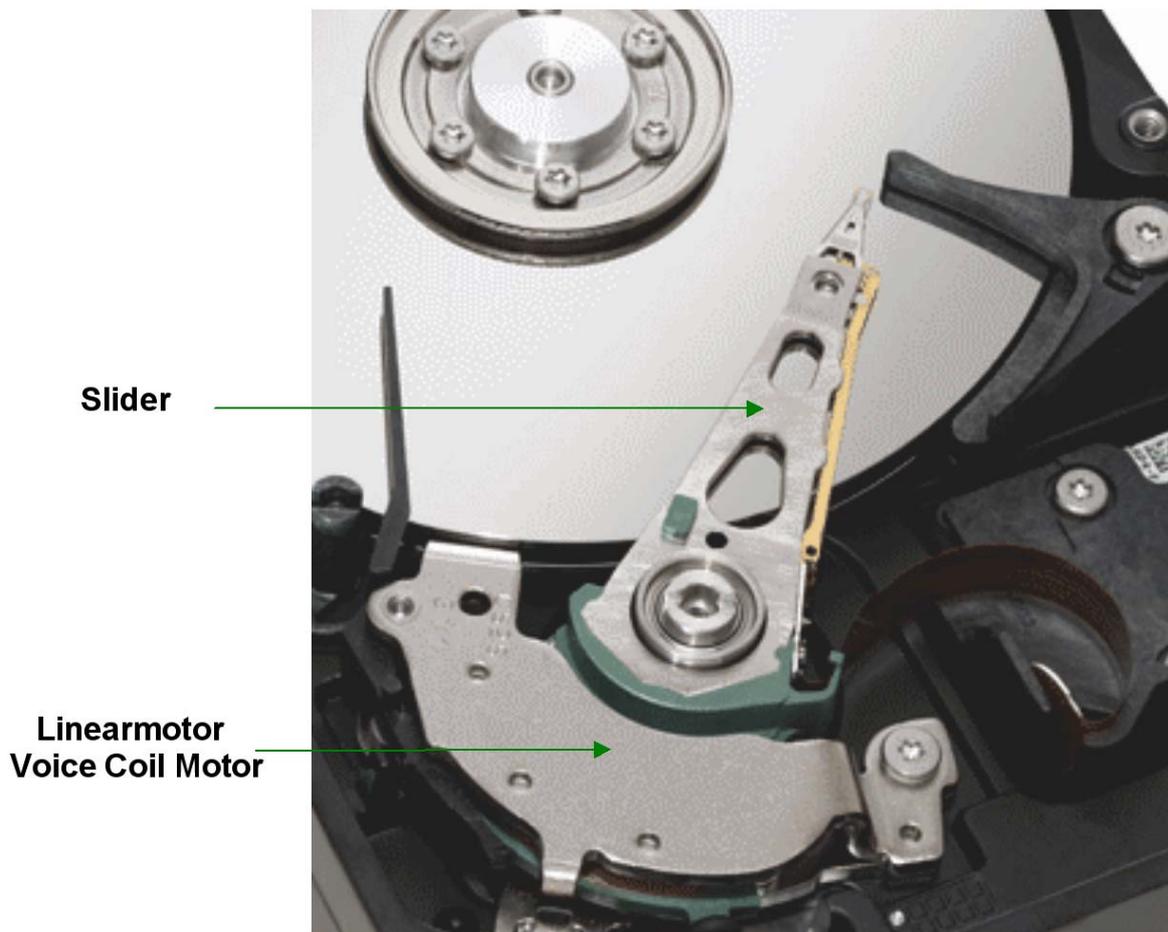


Abb. 5.1.4
Moderne Implementierung eines Festplattenspeichers

Abb. 5.1.4 zeigt eine Seagate Barracuda 7200.7 Festplatte der Firma Seagate (2006). Der Linearmotor bewegt sich senkrecht zur Achse des Sliders.

Eine Life Demo kann unter <http://www.youtube.com/watch?v=L0nbo1VOF4M> angeschaut werden.

Die Zugriffsarme (slider) werden durch einen Linearmotor angetrieben. Der Linearmotor wird oft auch als Voice Coil Motor bezeichnet, weil er das gleiche Prinzip benutzt wie der Linearmotor in einem Lautsprecher, der die Membrane antreibt. Eine Ansteuerungselektronik setzt eine Spuradresse in die entsprechenden Signale für den Linearmotor um.

5.1.4 Festplattenspeicher Zugriffszeiten

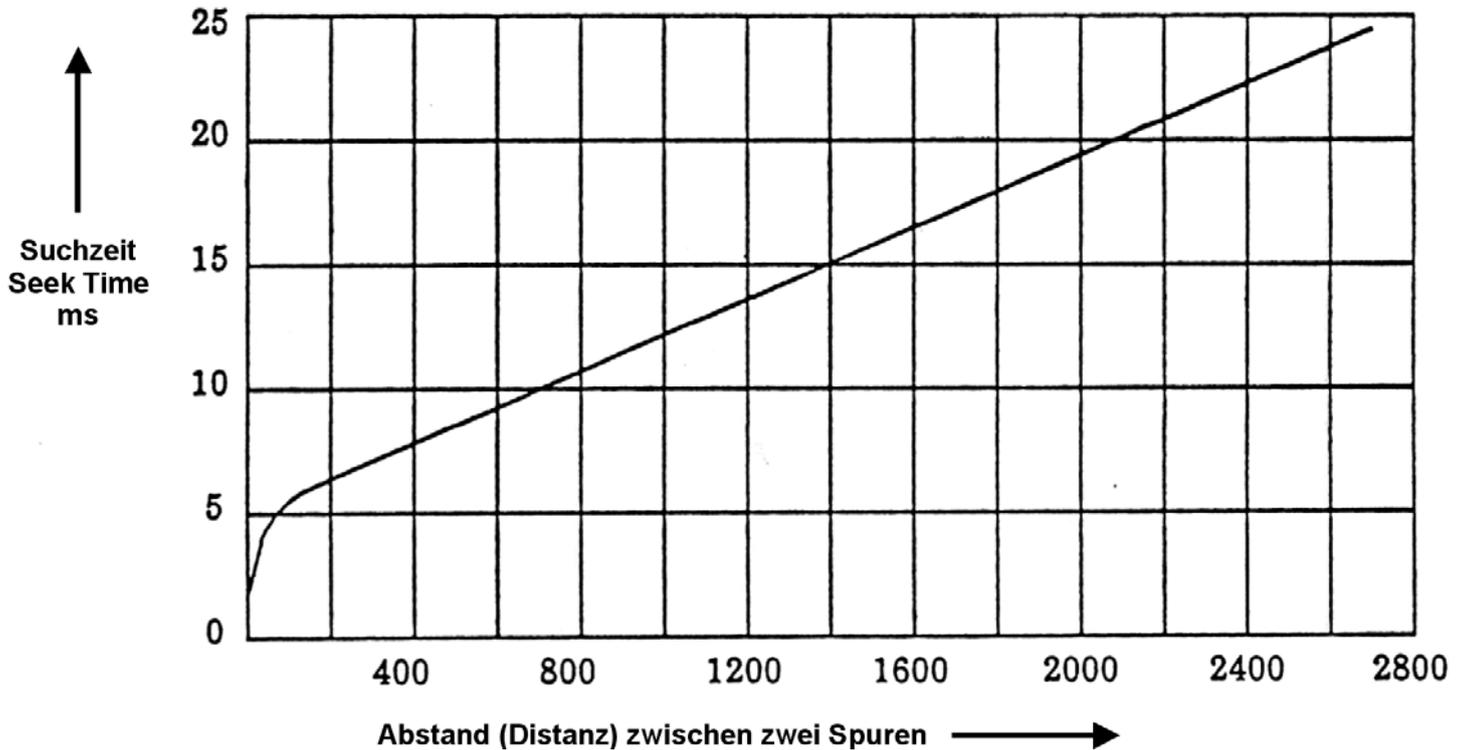


Abb. 5.1.5
Zugriffscharakteristik - Profil der Suchzeiten (Seek Time)

Die Bewegung durch den Linearmotor erfolgt relativ langsam. Der Wechsel von einer Spur auf eine unmittelbar **benachbarte** Spur kann mehrere Millisekunden betragen. Der Wechsel von einer Spur auf eine **entfernte** Spur kann bis zu 100 Millisekunden dauern. Diese Zeit ist kritisch für das Leistungsverhalten eines Plattenspeichers; sie wird als „Suchzeit“ (seek time) bezeichnet. Hersteller von Festplatten geben eine „mittlere Zugriffszeit“ an, die aus einer Mischung von vielen benachbarten und wenigen entfernten Zugriffen ermittelt wird. Zu fragen ist, ob diese Mischung realitätsnahe ist.

Die Speicherkapazität von Plattenspeichern ist in mehreren Jahrzehnten enorm gestiegen, von 5 MByte bei der ursprünglichen IBM RAMAC 305 bis zu mehreren TByte heute (2013). Spuren werden sehr viel enger gepackt und auch die Anzahl der Bits auf einer Spur ist sehr viel größer geworden.

An der Zugriffscharakteristik hat sich aber nur wenig geändert.

Der Lese/Schreibkopf wird mit etwa 100 g beschleunigt und abgebremst (1 g = Anziehungskraft auf der Erdoberfläche). Jenseits von 100 g beginnt Metall sich zu verformen. Es ist daher in den letzten Jahrzehnten nicht möglich gewesen, die Geschwindigkeit der mechanischen Bewegung des Linearmotors deutlich zu verbessern.

Die Platte dreht sich gegenüber dem Kopf mit etwa $\frac{2}{3}$ Schallgeschwindigkeit. Es entsteht ein Luftstrom zwischen Platte und Kopf. Der Kopf hat eine Fläche von etwa 1 mm^2 . Er ist aerodynamisch wie der Flügel eines Flugzeuges geformt und erhält einen Auftrieb gegenüber der Plattenoberfläche. Der Auftrieb ist umso größer, je geringer die Entfernung Kopf – Platte ist. Eine Feder drückt den Kopf gegen die Plattenoberfläche. Es stellt sich eine Flughöhe ein, bei der Auftrieb und Federdruck im Gleichgewicht sind. Die Flughöhe beträgt etwa 50 nm.

Ein kleines Gedankenexperiment soll dies verdeutlichen. Vergrößern wir diese Struktur um einen Faktor 100 000. Das Ergebnis ist ein Flugzeug-ähnliches Gebilde mit einer Flügelweite von 100 Meter, das mit nahezu Schallgeschwindigkeit mit einem Abstand von 5 mm über die Erdoberfläche fliegt. Jede Berührung des Kopfes mit der Plattenoberfläche bewirkt einen irreparablen Plattenspeichercrash.

5.1.5 Zylinder Organisation

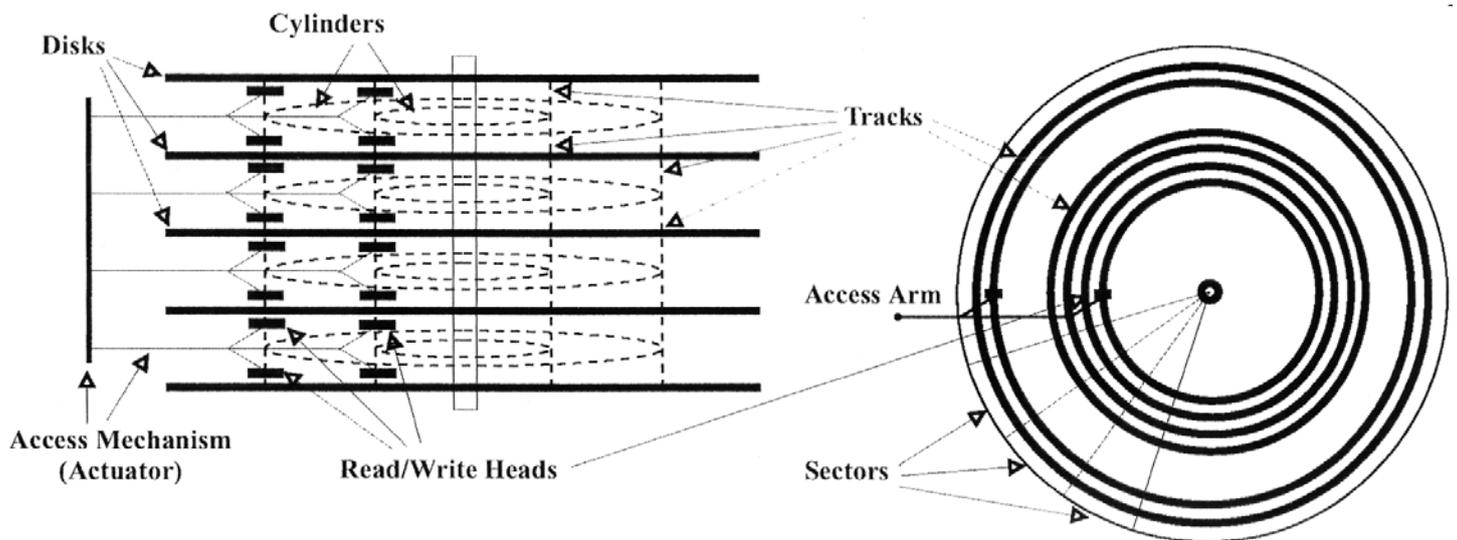


Abb. 5.1.6
a) Übereinander liegende Spuren

Um die Zugriffszeit zu verbessern werden häufig zwei (oder mehr) Lese/Schreibköpfe pro Plattenoberfläche eingesetzt. Dabei kann ein Kopf z.B. nur die äußeren Spuren und der andere Kopf nur die inneren Spuren abdecken. Hierdurch wird die maximale Distanz zwischen zwei Spuren halbiert.

Zusätzlich fasst man Spuren, die auf zwei oder mehr Oberflächen übereinander liegen, zu einem „Zylinder“ zusammen. Dadurch entstehen größere Speichereinheiten, die ohne eine Bewegung des Zugriffsarms adressiert werden können. Beim Wechsel von einer Spur zu einer anderen Spur des gleichen Zylinders muss nur die Elektronik der Lese/Schreibköpfe umgeschaltet werden, was im μs Bereich erfolgen kann.

Gebräuchliche Umdrehungsgeschwindigkeiten bei modernen Platten betragen:

3600 U/min	16,67 ms/Umdrehung
5400 U/min	11,11 ms/Umdrehung
7200 U/min	8,33 ms/Umdrehung
10 000 U/min	6 ms/Umdrehung
15 000 U/min	4 ms/Umdrehung

Bei diesen Umdrehungszeiten bewegt sich der Kopf gegenüber der Plattenoberfläche mit etwa $\frac{2}{3}$ Schallgeschwindigkeit. Bei einer größeren Geschwindigkeit würde der laminare Luftstrom zwischen Plattenoberfläche und Kopf abreißen und zu Turbulenzen führen. Dann kann der geringe Abstand zur Plattenoberfläche nicht mehr eingehalten werden.

Füllt man einen Plattenspeicher mit Helium statt mit Luft, sind höhere Umdrehungszahlen möglich. Die Schallgeschwindigkeit beträgt in Luft etwa 340 m/s, in Helium etwa 980 m/s .

In der Hardware und Betriebssystem-Software eines Rechners existieren sehr weitgehende Einrichtungen, die helfen, das an sich sehr schlechte Zugriffsverhalten eines Plattenspeichers zu verbessern. Es hat auch in den letzten Jahrzehnten viele Versuche gegeben, Plattenspeicher auf Grund Ihres schlechten Zugriffsverhaltens durch eine alternative Technologie zu ersetzen. Das ist über viele Jahre nicht erfolgreich gewesen. Mit der Einführung von Solid State Drives (SSD) wird sich das in der absehbaren Zukunft wahrscheinlich ändern.

5.1.6 Solid State Drive

Ein Solid State Drive (SSD,) ist ein Speichermedium, das wie eine herkömmliche magnetische Festplatte eingebaut und angesprochen werden kann, ohne eine rotierende Scheibe oder andere bewegliche Teile zu enthalten, da nur Halbleiterspeicherbausteine verwendet werden.

Der englische Begriff solid state in der Geschichte der Elektronik bedeutet, dass keinerlei bewegliche mechanische Teile wie Relais oder Röhren usw. verwendet werden, sondern Halbleiterbauteile, die durch die Entwicklungen in der Festkörperphysik möglich wurden. Insofern erscheint die Bezeichnung paradox, da ein eben gerade ohne bewegliche Teile auskommendes Medium als 'Drive' bzw. 'Disk' angesprochen wird. Dies geschieht in Analogie zu anderen Festplatten.

In SSD-Laufwerken werden in aller Regel Flash-Speicher Chips vom Typ "Nand" eingesetzt. Dabei wird die Information in Form von elektrischen Ladungszuständen der einzelnen Flash-Speicherezellen gespeichert. Das Speichern der Daten erfolgt nicht flüchtig, das heißt die Ladungen bleiben erhalten unabhängig davon, ob das Medium an die Stromzufuhr angeschlossen ist oder nicht. Das bedeutet, dass sich SSD-Speicher in der gleichen Art und Weise nutzen lassen wie Festplatten.

In den letzten Jahren werden zusätzlich zu (und bei mobilen Geräten an Stelle von) Plattenspeichern Solid State Drives eingesetzt.

Vorteile eines Solid State Drive sind Robustheit, schnelle Zugriffszeiten und Energieverbrauch. Nachteile sind Kapazität und Preis. In großen Mainframe Installationen findet man häufig zusätzlich zu den Plattenspeichern eine begrenzte Anzahl von SSDs für die Auslagerung von Daten für besonders performance-kritische Vorgänge.

Wenn sich SSDs als Alternative zu magnetischen rotierenden Plattenspeichern durchsetzen ist es denkbar, diese anders als herkömmliche Plattenspeicher zu nutzen.

5.1.7 Die Zukunft des Festplattenspeichers

Plattenspeicher behaupten ihre dominierende Rolle gegenüber SSDs, weil der Preisunterschied pro Byte Speicherkapazität etwa einen Faktor 10 beträgt.

Seit Jahrzehnten wird angenommen, dass SSDs in der Zukunft kostenmäßig gleichziehen werden. Das ist bis heute nicht geschehen.

Dennoch ist mit einem Zeithorizont von mehr als 10 Jahren anzunehmen, dass man dann den magnetischen Plattenspeicher als genauso archaisch ansehen wird, wie man heute die Lochkarte sieht. Dies wird in einem erheblichen Maße zu einer Änderung der Betriebssystem Struktur führen.

Panta rhei (griechisch ,Πάντα ῥεῖ), „Alles fließt“) sagte schon der griechische Philosoph Heraklit.

5.1.8 Direct Memory Access

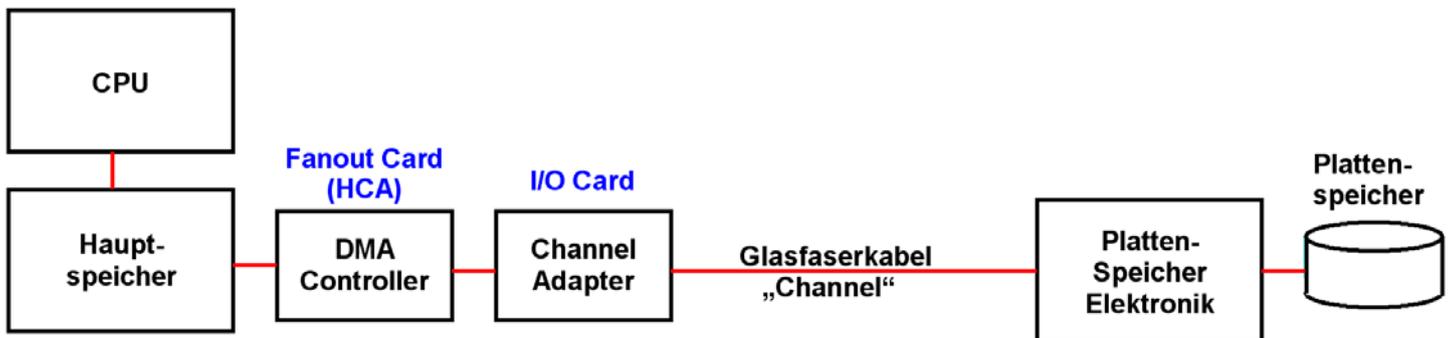


Abb. 5.1.7
Baugruppen einer Festplattenverbindung

Plattenspeicher werden mittels eines „Direct Memory Access“ (DMA) Controllers an den Hauptspeicher angeschlossen. In einem zEC12 Mainframe ist dies eine Fanout Card innerhalb eines Books. Der DMA Controller (die Fanout Card) ist wiederum mit einer Channel Adapter Card in einem I/O Drawer verbunden.

Die Channel Adapter Card stellt mittels eines Glasfaserkabels die Verbindung zur Elektronik eines Plattenspeichers her. Das Glasfaserkabel wird als Channel, Ficon Channel oder Channel Path bezeichnet (die Begriffe sind weitgehend austauschbar). Ein Mainframe kann mehrere 100 Glasfaseranschlüsse aufweisen. Die Distanz Rechner – Plattenspeicher kann viele Kilometer betragen.

Plattenspeicher werden in getrennten Gehäusen, sog. „Enterprise Storage Servern“ (ESS) untergebracht. Ein ESS kann Hunderte von Plattenspeichern enthalten. Für die Ansteuerung enthält ein ESS mehrere Spezialrechner, deren Firmware von außen nicht zugreifbar ist.

Der DMA Controller ist Bestandteil der Fanout Card (Host Channel Adapter). Der Channel Adapter ist als I/O Card implementiert, und befindet sich in einem PCIe I/O Drawer, siehe Abschnitt 4.3.6.

5.1.9 DMA Steuerung

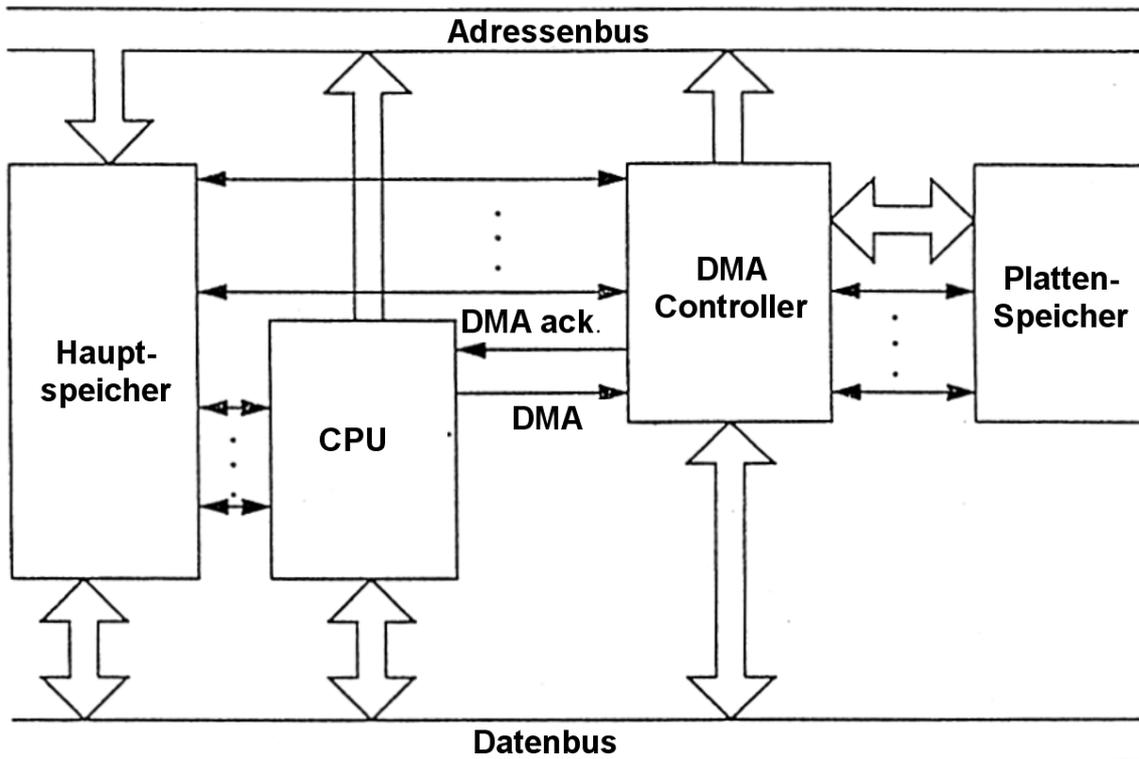


Abb. 5.1.8

Paralleler Zugriff auf den Hauptspeicher durch die CPU(s) und den DMA Controller

Der Processor (CPU) und der DMA (Direct Memory Access) Controller greifen mit Adress-, Daten- und Steuerleitungen gleichzeitig und parallel auf den Hauptspeicher zu. Jeder Hauptspeicherzyklus wird entweder von der CPU oder von dem DMA Controller genutzt.

Die Memory Bus Adapter (MBA) in dem z9 Rechner, bzw. Host Channel Adapter (HCA) in den z10/z196/zEC12 Rechnern werden gemeinsam auch als Fanout Card bezeichnet, und enthalten einen DMA Controller.

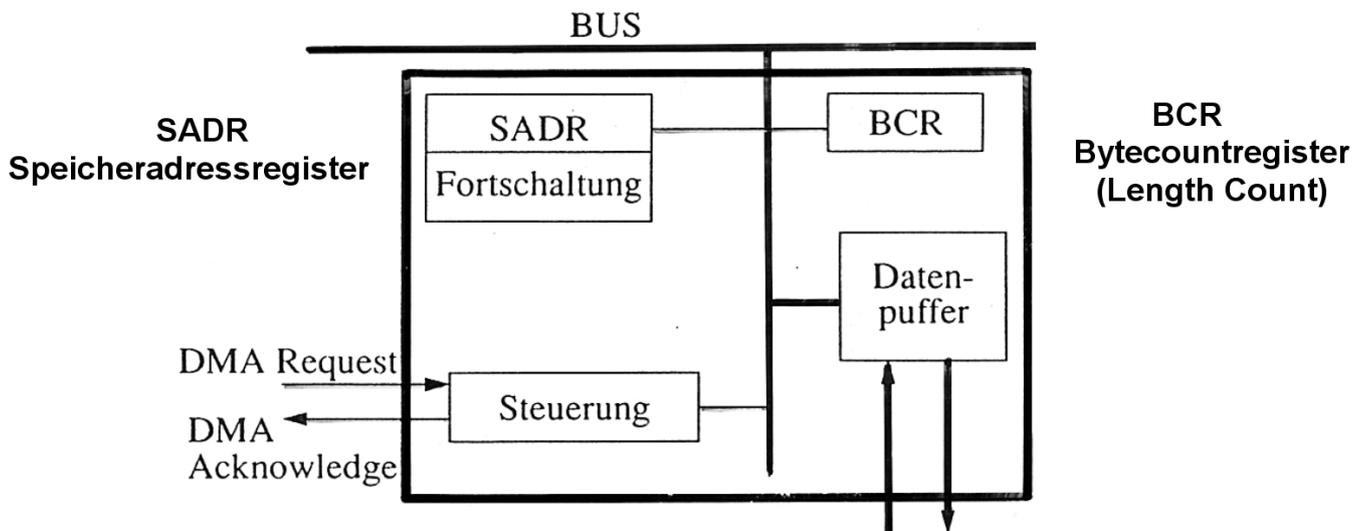


Abb. 5.1.9

Storage Address Register und Byte Count Register

Aufgaben der DMA-Steuerung:

- Adressieren des Hauptspeichers durch Adressfortschaltung
- Adressieren der Geräteschnittstelle
- Steuerung der Buszugriffe für Lesen oder Schreiben
- Zählen der übertragenen Bytes
- Rückmelden an CPU

Das Storage Address Register (SADR) enthält die Hauptspeicheradresse des zu übertragenden Bytes. Für die Übertragung eines größeren Datenblocks wird das SADR automatisch hochgezählt. Das Byte Count Register (BCR) enthält die Länge des zu übertragenden Datenblocks.

Für einen Übertragungsvorgang werden DMA-Steuerung und I/O Controller (I/O card) von der CPU initialisiert, z.B. Laden der Steuer(Control)- und Adressregister.

5.1.10 Unterbrechungsgesteuerte Ein/Ausgabe

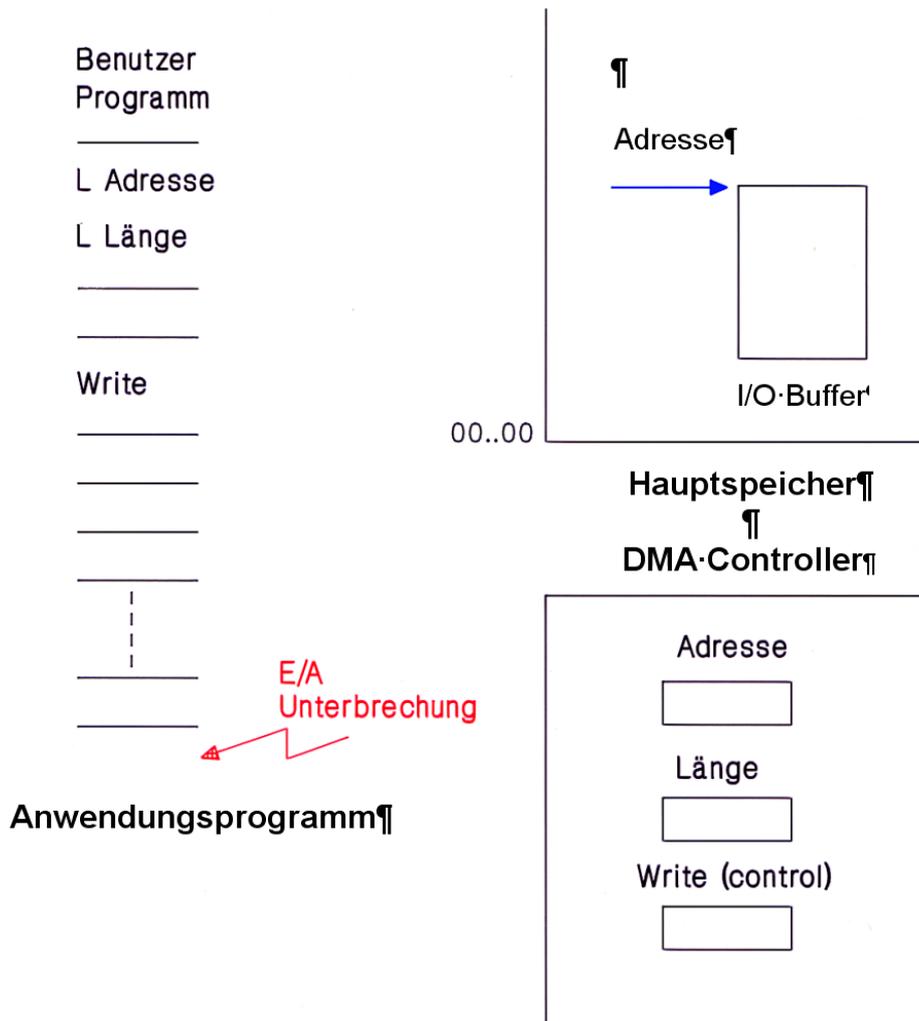


Abb. 5.1.10

Das Adress Register des DMA Controllers enthält die Hauptspeicheradresse des I/O Puffers

Die zu übertragenden Daten befinden sich in einem als I/O Buffer bezeichnetem Bereich des Hauptspeichers. Das Anwendungsprogramm definiert die Adresse und Länge des I/O Buffers, und führt z.B. einen WRITE Befehl aus.

Der DMA Controller übernimmt Adresse und Länge in seine SADR und BCR Register und setzt sein Control Register auf Write.

Danach liest er automatisch und unabhängig von der CPU die Daten aus dem I/O Puffer und überträgt sie zum Plattenspeicher. Nach der Übertragung eines jeden Bytes wird der Length Count in dem BCR Register (Byte Count Register) heruntergezählt und das SADR Register wird inkrementiert. Die CPU arbeitet während dieser Zeit unabhängig weiter.

Wenn der Wert in dem Längenregister den Wert 0 erreicht, ist die I/O Operation beendet. Dies wird der CPU über eine I/O Unterbrechung mitgeteilt.

5.2 SCSI und FICON

5.2.1 Moderne Plattenspeicher-Anschlussarten

Moderne Plattenspeicher werden heute fast ausschließlich über ein serielles Protokoll angeschlossen. Die wichtigsten Protokolle sind:

- SATA (serial ATA) Nachfolger für parallel ATA (anderer Name ist IDE)
- SAS (serial attached SCSI) Nachfolger für parallel SCSI
- iSCSI Internet SCSI (benutzt Ethernet, von Mainframes nicht unterstützt)
- FC-SCSI Fibre Channel SCSI

SATA dominiert beim PC und anderen Arbeitsplatzrechnern. Die verschiedenen SCSI Arten dominieren bei Servern. Fibre Channel SCSI kann sowohl für Punkt-zu-Punkt Verbindungen als auch als FC-AL Version (Fibre Channel Arbitrated Loop) eingesetzt werden.

Seagate z.B. bietet z.B. (2010) die Barracuda Familie von Plattenspeichern mit der SATA und der SAS Schnittstelle an; Speicherkapazität bis zu 4,0 TByte. Für „Mission Critical Applications“ ist die Cheetah Familie von Plattenspeichern mit einer 4-Gb/s Fibre Channel interface verfügbar; Speicherkapazität bis zu 600 GByte. Cheetah Plattenspeicher sind laut Seagate für Anwendungen vorgesehen, „**where system availability and reliability is of utmost importance**“.

Letzteres ist vor allem bei Mainframes gegeben. Dafür wird die deutlich geringere Speicherkapazität in Kauf genommen. In anderen Worten: Plattenspeicher in Ihrem PC haben pro Einheit in der Regel eine deutlich höhere Speicherkapazität als die Plattenspeicher eines Mainframes.

5.2.2 S/360 Channel und SCSI



Abb. 5.2.1

b) Historische Entwicklung OEMI Channel und SCSI

Für den Anschluss von I/O Geräten System führte das System/360 im Jahre 1964 den „Selektor“ und 1970 den „Block Multiplex“ Kanal ein. IBM veröffentlichte diese I/O Schnittstelle unter dem Namen OEMI (Original Equipment Manufacturer Interface), was dazu führte, dass viele unabhängige Hersteller I/O Geräte für den Anschluss an die damaligen Mainframes entwickelten und installierten. Das ist auch heute noch der Fall.

Ab 1982 wurde auf Initiative der Firmen Shugart und NCR eine Modifikation des OEMI Standards durch das ANSI (American National Standards Institute) unter dem Namen SCSI (Small Computer System Interface) für den Einsatz in kleineren Systemen veröffentlicht. Hierbei wurde die auf extreme Zuverlässigkeit ausgelegte OEMI Verkabelung (elektrische Interface) vereinfacht. Die logische Interface wurde weitestgehend beibehalten.

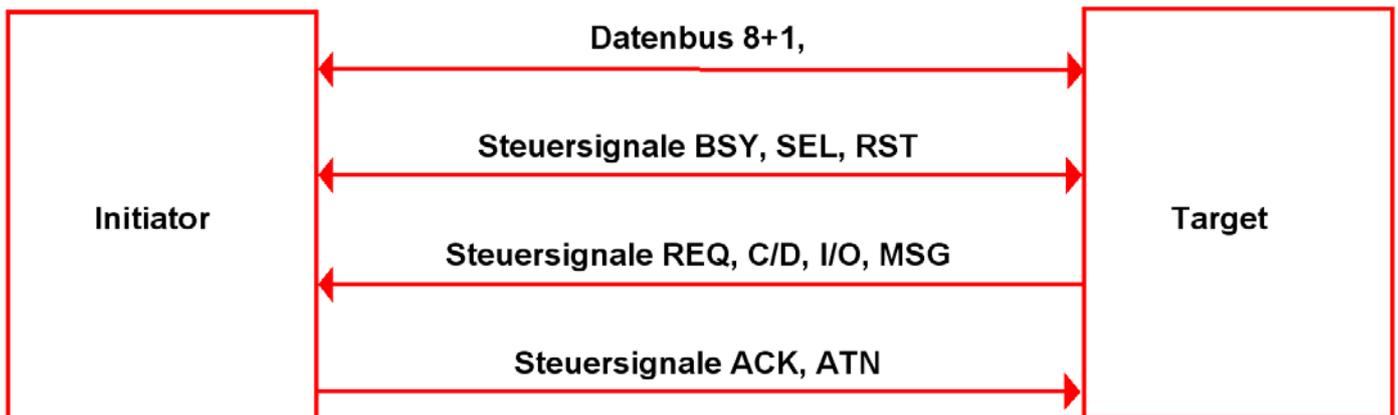


Abb. 5.2.2

OEMI Schnittstelle

Die logische OEMI und SCSI Interface bestand aus einem 8 Bit (plus Parity) Datenbus sowie einer Reihe von teils unidirektionalen, teils bidirektionalen Steuersignalen. Die OEMI Schnittstelle wird heute von IBM als „Parallel Channel“ bezeichnet. Ein Parallel Channel hat eine Datenrate von max. 4.5 MByte/s. und überbrückt Distanzen bis zu 130 m. Der Parallel Channel benutzt zwei Kupfer Kabel: *Bus* und *Tag*. Ein Bus Kabel überträgt Information (ein

Byte in jeder Richtung). Daten auf dem Tag Kabel definieren die Bedeutung der Information auf dem Bus Kabel.

Später entwickelten sich die OEMI und SCSI Schnittstellen unabhängig voneinander weiter. Der ursprünglich 8 Bit breite Datenbus der SCSI-1 Schnittstelle wurde auf 16 und später 32 Bit verbreitet. Anschließend entstanden serielle Versionen, die bei IBM zu den Glasfaser-gestützten ESCON und dann den FICON Kanälen führten. Bei SCSI entstanden die Serial SCSI und die Glasfaser FC-SCSI Versionen.

Wegen der höheren Anforderungen im Großrechnerbereich hatten die IBM Kanäle immer einen deutlich höheren Funktionsumfang als die SCSI Schnittstellen. Auch heute kann der mit einem FICON Netzwerk erzielbare Durchsatz durch ein SCSI Netzwerk nicht erreicht werden.

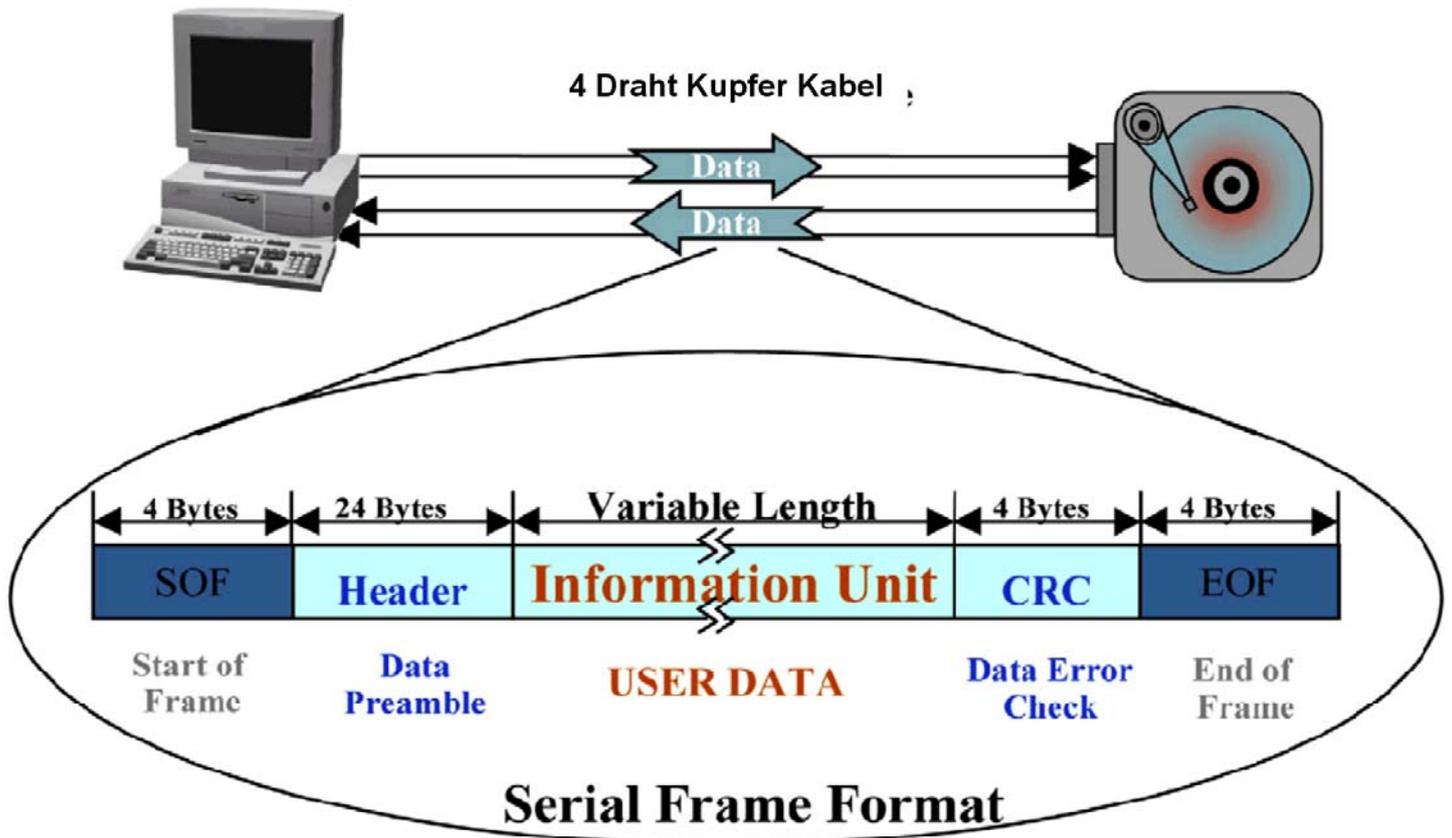


Abb. 5.2.3
Serial SCSI

Die ursprünglich parallel übertragenen OEMI oder SCSI Daten werden bei SAS (serial attached SCSI) in einen Rahmen (Frame) gepackt und über ein Kupferkabel seriell übertragen. Die Fibre Channel SCSI (FC-SCSI) Version verwendet statt dessen zwei Glasfaserkabel (für Hin- und Rückleitung).

Das aus dem Parallel Channel (OEMI) entwickelte serielle Fibre Channel FICON Protokoll benutzt ebenfalls Glasfasern. Es hat im Vergleich zu FC-SCSI einen wesentlich höheren Funktionsumfang und damit eine bessere I/O Leistung.

Der Parallel Channel dominierte die Mainframe I/O Konfigurationen bis zum Anfang dieses Jahrhunderts. Er hat viel Ähnlichkeit mit der parallelen SCSI Interface. Heutige Mainframes unterstützen den Parallel Channel nicht mehr.

Serielle Channels haben den älteren parallel Channels abgelöst. Es existieren zwei Serial Channel Typen:

- **Der (ältere) „ESCON Channel erlaubt Datenraten von 17 MByte/s. Er wurde 1990 eingeführt, und wird in zukünftigen Mainframe Modellen nicht mehr verfügbar sein.**
- **Ein FICON Channel erlaubt Datenraten bis zu 800 MByte/s. Das FICON Protokoll wurde 1997 eingeführt.**

Die über 10 Jahre alte „Shark“ Plattenspeichereinheit unseres eigenen Mainframe Rechners jedi.informatik.uni-leipzig.de wurde über ESCON Verbindungen angeschlossen. Unsere neuere DS6800 Plattenspeichereinheit verwendet FICON Verbindungen.

Serielle Channel verwenden Glasfaser Kabel an Stelle von Kupferverbindungen. Es können Entfernungen bis zu 100 km überbrückt werden. Außerdem verfügen sie über eine erweiterte I/O Adressierung.

Formal wird ein Channel als „Channel Path“ bezeichnet, und durch einen 8 Bit CHPID (Channel Path Identifier) gekennzeichnet. In der Umgangssprache werden Channel Path nach wie vor als Kanäle (Channels) bezeichnet, und auch Experten kennen den Unterschied nicht genau.

5.2.3 Fibre Channel

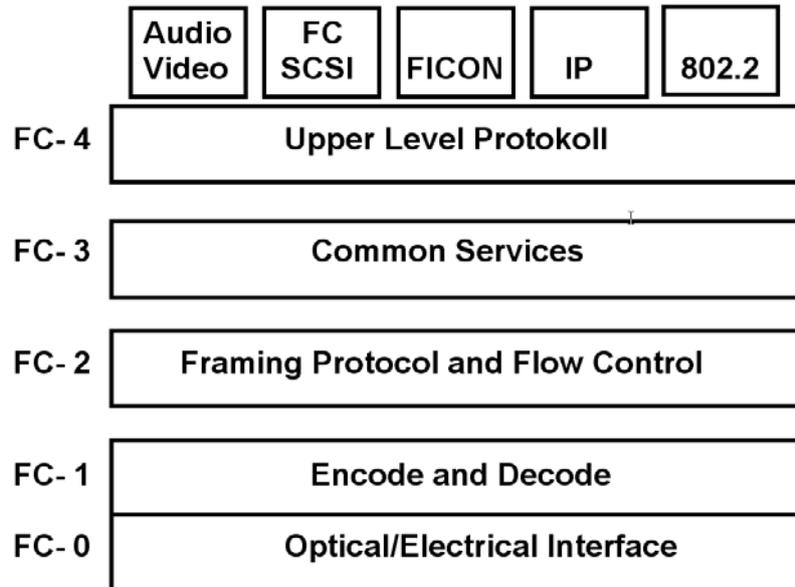


Abb. 5.2.4
c) Fibre Channel Architektur

Fibre Channel (FC) ist für serielle, kontinuierliche Hochgeschwindigkeitsübertragung großer Datenmengen konzipiert worden. Die erreichten Datenübertragungsraten liegen heute bei 8 Gbit/s, was im Vollduplex-Betrieb für Datentransferraten von 800 MB/s ausreicht. Als Übertragungsmedium findet man gelegentlich Kupferkabel (hauptsächlich innerhalb von Storage-Systemen; überbrückt bis zu 30 m), meistens aber Glasfaserkabel. Letzteres wird vor allem zur Verbindung von Rechnern mit Storage-Systemen oder aber von Storage-Systemen untereinander eingesetzt. Hierbei werden Entfernungen bis zu 10 km überbrückt. Der Zugriff auf die Festplatten erfolgt blockbasiert.

Es können generell zwei Arten von Fibre-Channel-Implementierungen unterschieden werden, die Switched Fabric, die meist als Fibre Channel Switched Fabric (FC-SW) bezeichnet wird und die Arbitrated Loop, kurz als FC-AL bekannt.

Bei der Fibre Channel-Switched Fabric werden Punkt-zu-Punkt-Verbindungen (Point To Point) zwischen den Endgeräten geschaltet. Beim Fibre Channel-Arbitrated Loop handelt es sich um einen logischen Bus, bei dem sich alle Endgeräte die gemeinsame Datenübertragungsraten teilen.

Das Fibre Channel-Switched Fabric ist die leistungsfähigste und ausfallsicherste Implementierung von Fibre Channel. In den meisten Fällen ist Switched Fabric gemeint, wenn nur von Fibre Channel gesprochen wird. Im Zentrum der Switched Fabric steht der Fibre Channel Switch (von IBM als „Director“ bezeichnet). Über dieses Gerät werden alle anderen Geräte miteinander verbunden, so dass es über den Fibre Channel Switch möglich wird, direkte logische Punkt-zu-Punkt-Verbindungen zwischen je zwei beliebigen angeschlossenen Geräten zu schalten.

FC-AL erlaubt es, bis zu 127 Geräte an einem logischen Bus zu betreiben. Dabei teilen sich alle Geräte die verfügbare Datenübertragungsrate (bis 8 GBit/s). Die Verkabelung kann sternförmig über einen Fibre Channel Hub erfolgen. Es ist auch möglich, die Geräte in einer Schleife (Loop) hintereinander zu schalten (Daisy Chain), da viele Fibre-Channel-Geräte über zwei Ein- bzw. Ausgänge verfügen. Dies ist z.B. beim IBM DSS 8700 Enterprise Storage Server der Fall.

Die Fibre Channel Architektur verwendet ein Schichtenmodell, vergleichbar mit (aber unabhängig von) den TCP/IP oder OSI Schichtenmodellen. Die unterste Schicht verwendet in den meisten Fällen optische Kabel. Wichtig ist besonders die oberste Schicht FC4. Hierüber ist es möglich, unterschiedliche Protokolle zu betreiben (siehe Abb. 5.2.4).

FC-SCSI ist eine serielle Form des SCSI Protokolls, die über Fibre Channel Verbindungen erfolgt. (Das als „Serial SCSI“ bezeichnete Protokoll benutzt keinen Fibre Channel).

FICON ist das universell von Mainframes eingesetzte Protokoll, um Rechner miteinander und mit I/O Geräten zu verbinden.

IP over Fibre Channel und **Ethernet over Fibre Channel** (IEEE 802.2) wurde standardisiert, wird aber nur wenig benutzt.

Ein spezieller Fibre Channel Adapter wird für die Echtzeitübertragung von Fernsehprogrammen benutzt.

Literatur:

<http://www.answers.com/topic/fibre-channel?cat=technology>

In der Praxis verwenden SCSI Platten bessere mechanische und elektronische Komponenten als SATA Platten. Dies bewirkt:

- schnellere Zugriffszeiten
- höhere Zuverlässigkeit
- deutlich höhere Kosten

Aus Zuverlässigkeitsgründen verwenden SCSI Platten selten die neueste Plattenspeichertechnologie. Dies ist einer der Gründe, warum für einen Personal Computer Plattenspeicher mit einer höheren Speicherkapazität erhältlich sind als dies im Mainframe Bereich der Fall ist.

FATA-Laufwerke sind SATA-Plattenlaufwerke mit einem Fibre Channel (FC) Anschluss. Sie arbeiten mit den mechanischen Komponenten von ATA-Festplatten, jedoch mit vorgeschalteter FC-Schnittstelle. In anderen Worten, es sind SATA Platten, bei denen die elektrische serielle ATA Schnittstelle durch eine Fibre Channel Schnittstelle ersetzt wurde. Die FATA-Technologie hat den Vorteil, dass sie in einer Mischung mit anderen FC-Laufwerken betrieben werden können.

FATA Platten werden auch als „Nearline-Platten“ bezeichnet. Sie werden im Großrechnerbereich dann eingesetzt, wenn Zuverlässigkeit und Zugriffszeit weniger wichtig sind, z.B. um Bilddateien (Images) zu archivieren.

5.2.4 Plattenspeicher Anschluss Alternativen

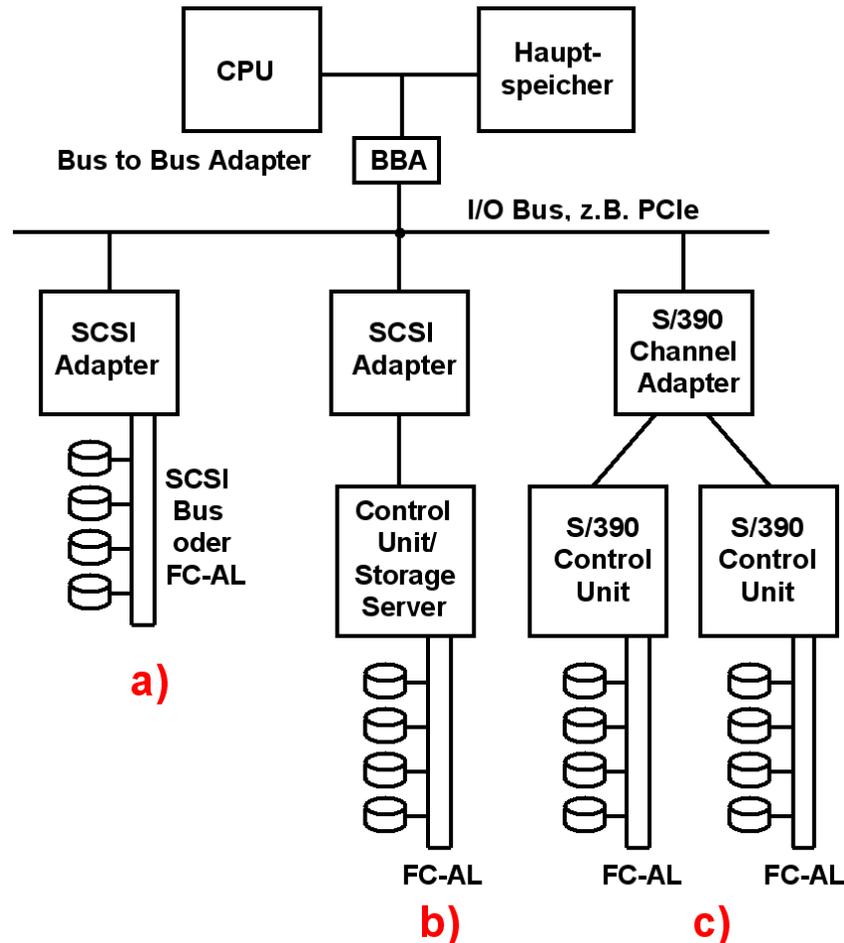


Abb. 5.2.5
Mögliche Plattenspeicher Anschlüsse an einen Server

- a)** Ein SCSI Adapter kann eine oder mehrere SCSI Platten mit dem I/O Bus eines Rechners verbinden.
- b)** Bei größeren Mengen an Plattenspeichern sind diese über eine Control Unit oder einen Storage Server mit dem SCSI Adapter verbunden.
- c)** Mainframes ersetzen den SCSI Adapter durch einen Channel Adapter. Plattenspeicher sind grundsätzlich über Control Units mit dem Channel Adapter verbunden .

5.2.5 S/360 I/O Konfiguration

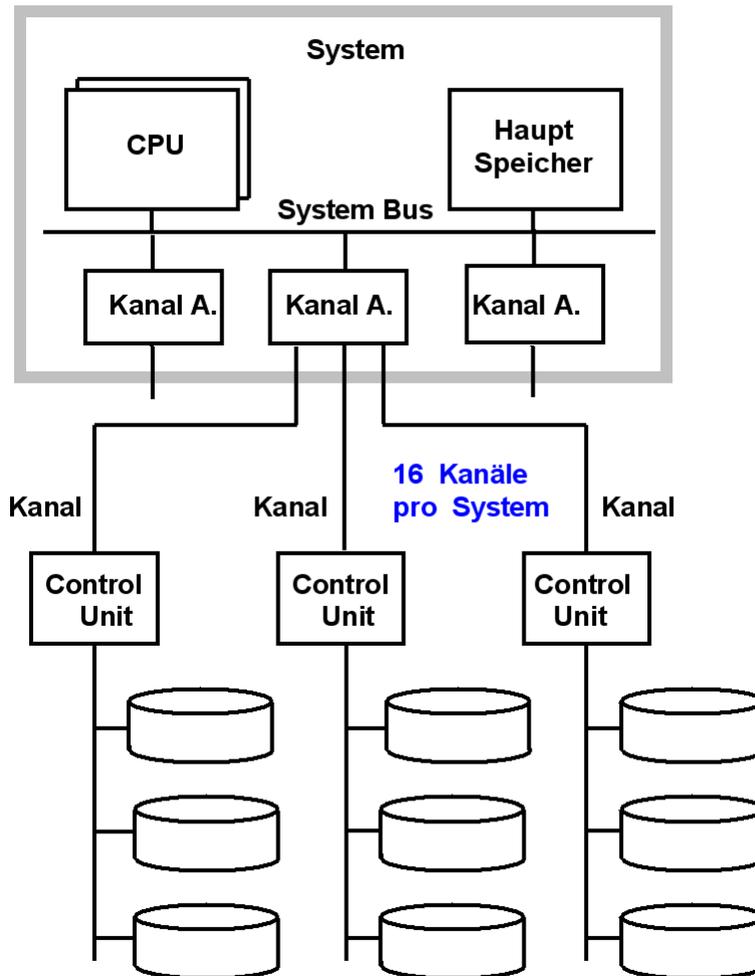


Abb. 5.2.6
S/360 Plattenspeicher Anschluss

Dargestellt ist die ursprüngliche S/360 I/O Konfiguration.

Plattenspeicher sind über Control Units, Kanal-Verbindungskabel (Channel Cables) und Kanal-Adapter mit dem Hauptspeicher des Systems verbunden. Die Verbindungskabel des „Parallel Channels“ waren bis zu 400 Fuß (130 m) lang. Die Kanal-Adapter konnten mittels DMA direkt auf den Hauptspeicher des Systems zugreifen.

Die Control Unit führte Befehle aus, die vom Kanal-Adapter aus dem Hauptspeicher ausgelesen und zwecks Ausführung an die Control Unit übergeben wurden.

Magnetbandgeräte und Drucker werden ähnlich wie Plattenspeicher über Control Units an einen Mainframe Rechner angeschlossen.

5.2.6 Aufgaben der Plattenspeicher-Steuereinheit

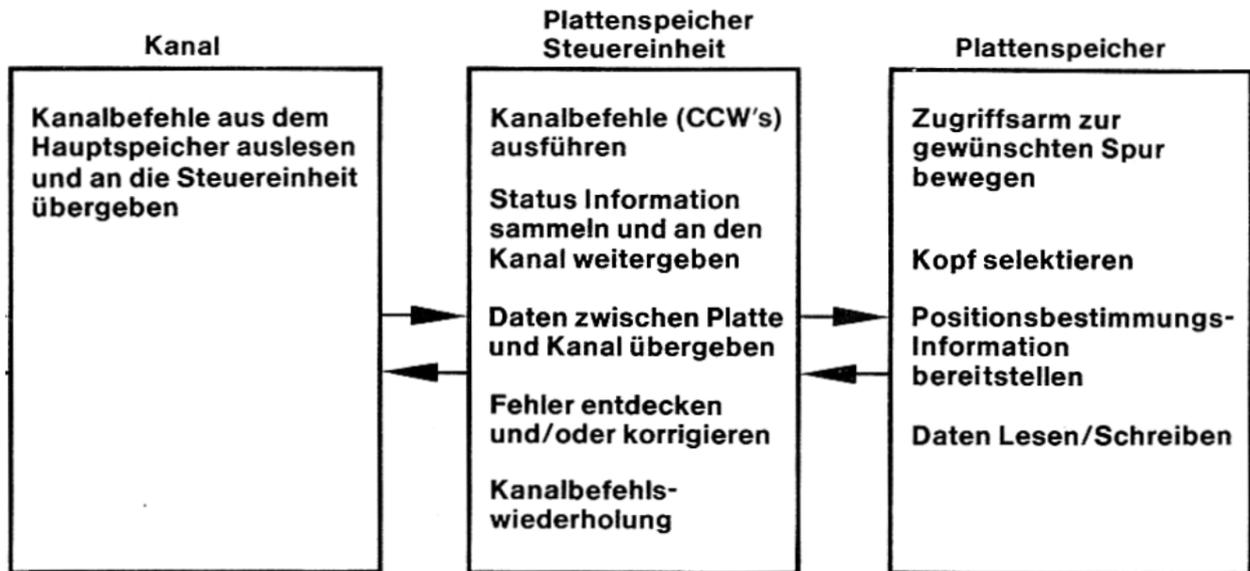


Abb. 5.2.7
Funktionen von Kanal Adapter, Steuereinheit und Festplatte

Dargestellt ist die Aufgabenteilung zwischen Kanal-Adapter, Steuereinheit (Control Unit) und der Plattenspeicher-Elektronik.

Der Kanal (die Kanal-Adapter Karte) ist nur dazu da, um stellvertretend für die Steuereinheit per DMA Daten und Kanalbefehle (CCWs) aus dem Hauptspeicher auszulesen und an die entfernte Steuereinheit weiter zu geben.

Der Plattenspeicher selbst enthält umfangreiche elektronische und logische Komponenten. Sie führen Steuerfunktionen aus wie:

- Zugriffsarm auf die richtige Spur bewegen
- Verifizieren, dass dies geschehen ist
- Einen von mehreren Lese/Schreib-Köpfen (und damit eine Plattenoberfläche) selektieren

5.2.7 Was ist Firmware ?

Komponenten, die früher mittels hart verdrahteter Transistorlogik erstellt wurden, verwenden heute häufig statt dessen einen dedizierten Mikroprozessor mit speziellem Code. Dieser Code hat die Eigenschaft, dass ein normaler Benutzer nicht darauf zugreifen, ihn ändern oder erweitern kann. Derartiger Code wird wahlweise als Microcode oder als Firmware bezeichnet ist.

Firmware Code eines Mainframes wird von Prozessoren mit der System z Architektur ausgeführt. Microcode eines Mainframes wird von nicht-System z Prozessoren ausgeführt, z.B. von PowerPC Prozessoren. Auf den in Abschnitt 12.3 erwähnten SAPs (System Assist Prozessor) läuft Firmware; auf der I/O Card befindet sich ein PowerPC Prozessor, der Microcode ausführt.

Außerhalb der Mainframes Welt ist der Unterschied zwischen Firmware und Microcode weniger sauber definiert. Firmware/Microcode läuft z.B. auf dem Prozessor, der einen WLAN Access Point, der Mikroprozessor Steuerung einer Geschirrspülmaschine oder der Mikroprozessor Steuerung des elektrischen Fensterhebers Ihres Mercedes S-Klasse Autos steuert.

Das in Abb. 5.2.8 erwähnte „Channel Subsystem“ besteht aus „System Assist Processoren (SAP)“ plus Firmware. Control Units und Enterprise Storage Server verwenden sehr leistungsfähige Prozessoren aber keine Firmware, da ihr Code nicht auf System z Prozessoren läuft.

5.2.8 System Assist Processor

Ein z196 Rechner hat bis zu $4 \times 24 = 96$ Prozessoren (Cores). Von diesen wird nur der größere Teil als CPUs eingesetzt. Mehrere Prozessoren, als „System Assist Processoren“ (SAP) bezeichnet und vorkonfiguriert, und führen ausschließlich Firmware Code aus. Bei einem maximal hochgerüsteten z196 Rechner sind dies mindestens 14 SAPs. Die SAPs weisen die gleiche Hardware Architektur auf wie die CPUs; auf ihnen läuft aber Firmware und kein z/OS. Bei der Installation eines neuen System z Rechners wird über eine Konfigurationsdatei eingestellt, wie viele der vorhandenen Prozessoren als CPUs bzw. SAPs eingesetzt werden.

SAPs und ihr Firmware Code wird vor allem für drei Funktionen benötigt:

1. Fehlerbehandlungs- und Recovery Funktionen
2. Channel Subsystem
3. PR/SM Hypervisor Software für LPAR virtuelle Maschinen (siehe Abschnitt 12.3).

Für Firmware (und dazugehörige Daten) sind in einem z196 Rechner 16 oder 32 GByte Speicherplatz vorgesehen. Von dem installierten physischen Speicher werden 16 oder 32 GByte abgeteilt und steht als „Hardware System Area“ (HSA) für Firmware Zwecke zur Verfügung. Der Rest kann als Hauptspeicher genutzt werden.

Bitte beachten: Den Begriff SAP (System Assist Processor) nicht verwechseln mit dem Namen der Firma SAP AG in Walldorf (Baden).

5.2.9 System z Festplattenspeicher Anschluss

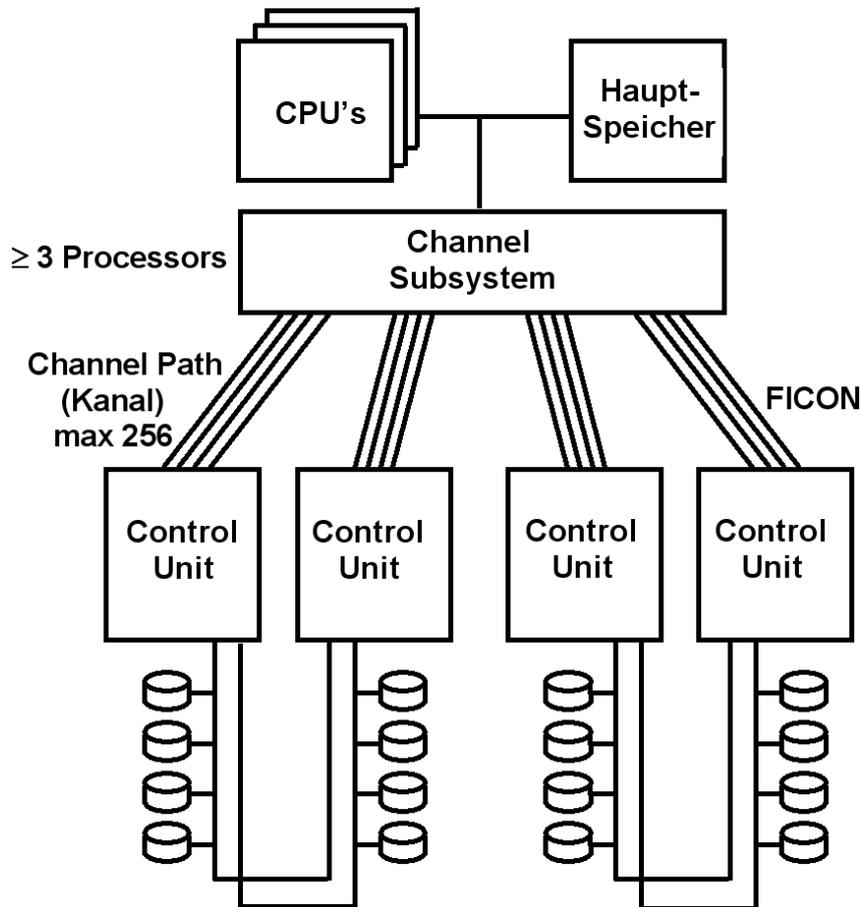


Abb. 5.2.8
Rolle des Channel Subsystems

Ein Mainframe kann über mehrere (bis zu 8) Kanäle mit einer bestimmten Control Unit verbunden werden, und ein I/O Gerät kann an mehr als eine Control Unit angeschlossen werden.

Das Channel Subsystem bildet die logische I/O Konfiguration, wie sie das Betriebssystem sieht, auf die physische Konfiguration ab.

Heute werden fast immer mehrere Control Units und angeschlossene Plattenspeicher zu einem physischen „Enterprise Storage Server“ zusammengefasst. Control Units (und das Channel Subsystem) sind jedoch Teil der System z Architektur-Spezifikation; das Enterprise Storage System ist lediglich eine Implementierung, und ist nicht Teil der Architektur.

Ein Plattenspeicher ist typischerweise an zwei Control Units angeschlossen. Die Kommunikation mit der CPU erfolgt wahlweise über eine der beiden Control Units.

5.2.10 System z Disk Storage Attachment

System z I/O Geräte werden über Steuereinheiten (Control Units) an das Kanal Subsystem angeschlossen. Viele Funktionen, die auf Ihrem PC von der CPU als I/O Driver Code ausgeführt werden, sind bei einem Mainframe in den Control Units implementiert. Sie belasten die CPUs nur wenig, und ermöglichen die Ansteuerung einer sehr großen Anzahl von Festplatten.

Steuereinheiten können über mehr als einen Kanalpfad an das Kanalsubsystem angeschlossen werden und I/O Geräte (z.B. Plattenspeicher) können an mehr als eine Steuereinheit angeschlossen werden.

z/OS kann auf ein beliebiges I/O Gerät über bis zu 8 unterschiedliche Kanalpfade zugreifen (und umgekehrt).

Der Zugriffsweg kann dynamisch geändert werden (DPS - Dynamic Path Selection). Eine I/O Operation muss nicht auf dem gleichen Weg abgeschlossen werden, auf dem sie gestartet wurde. Hiermit kann erreicht werden, dass sequentielle und zufallsbedingte Zugriffe sich nicht gegensätzlich beeinträchtigen.

Z.B. angenommen zwei Festplattenspeicher, die an die gleiche Steuereinheit angeschlossen sind. Ein Plattenspeicher überträgt einen großen Block sequentieller Daten, während der zweite Plattenspeicher gleichzeitig viele kurze Datenpakete mit einem zufallsbedingten Zugriffsmuster überträgt. Kein Plattenspeicher soll die Nutzung einer Verbindung für einen längeren Zeitraum usurpieren.

In anderen Worten, es gibt mehrere Wege, auf denen Daten (und Steuerinformation) zwischen Platte und CPU übertragen werden können.

Diese dynamische Weg-Steuerung ist rechenaufwendig. Deswegen belastet man mit dieser Aufgabe nicht die CPUs und das Betriebssystem, sondern überträgt sie einer getrennten Verarbeitungseinheit, dem **Channel Subsystem**. Eine Beispiel-Konfiguration ist in Abb. 5.2.9 wiedergegeben.

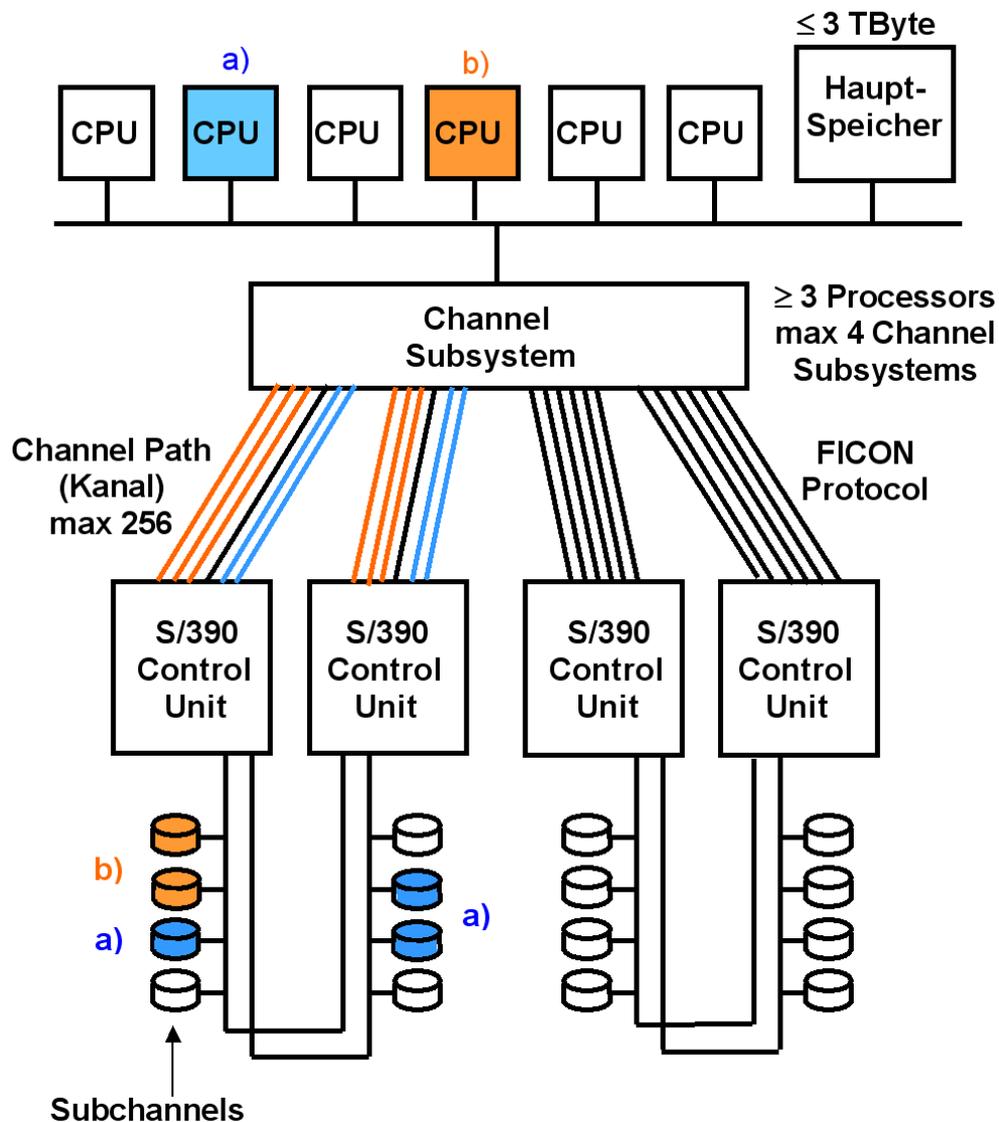


Abb. 5.2.9
Mehrfache Pfade zu einem I/O Gerät

Der auf der blauen CPU **a)** laufende Prozess greift auf die drei blauen Plattenspeicher **a)** zu. Diese können über die ersten beiden Control Units erreicht werden. Der blaue Prozess verwendet hierzu vier blau gezeichnete Kanäle, von denen jeweils zwei mit den beiden Control Units verbunden sind.

Ähnliches gilt für den Prozess auf der orange markierten CPU **b)**.

An eine Control Unit können bis zu 8 Kanäle angeschlossen sein.

Das Channel Subsystem kann die Zuordnung von Kanälen zu Prozessen dynamisch ändern.

5.2.11 FICON Director

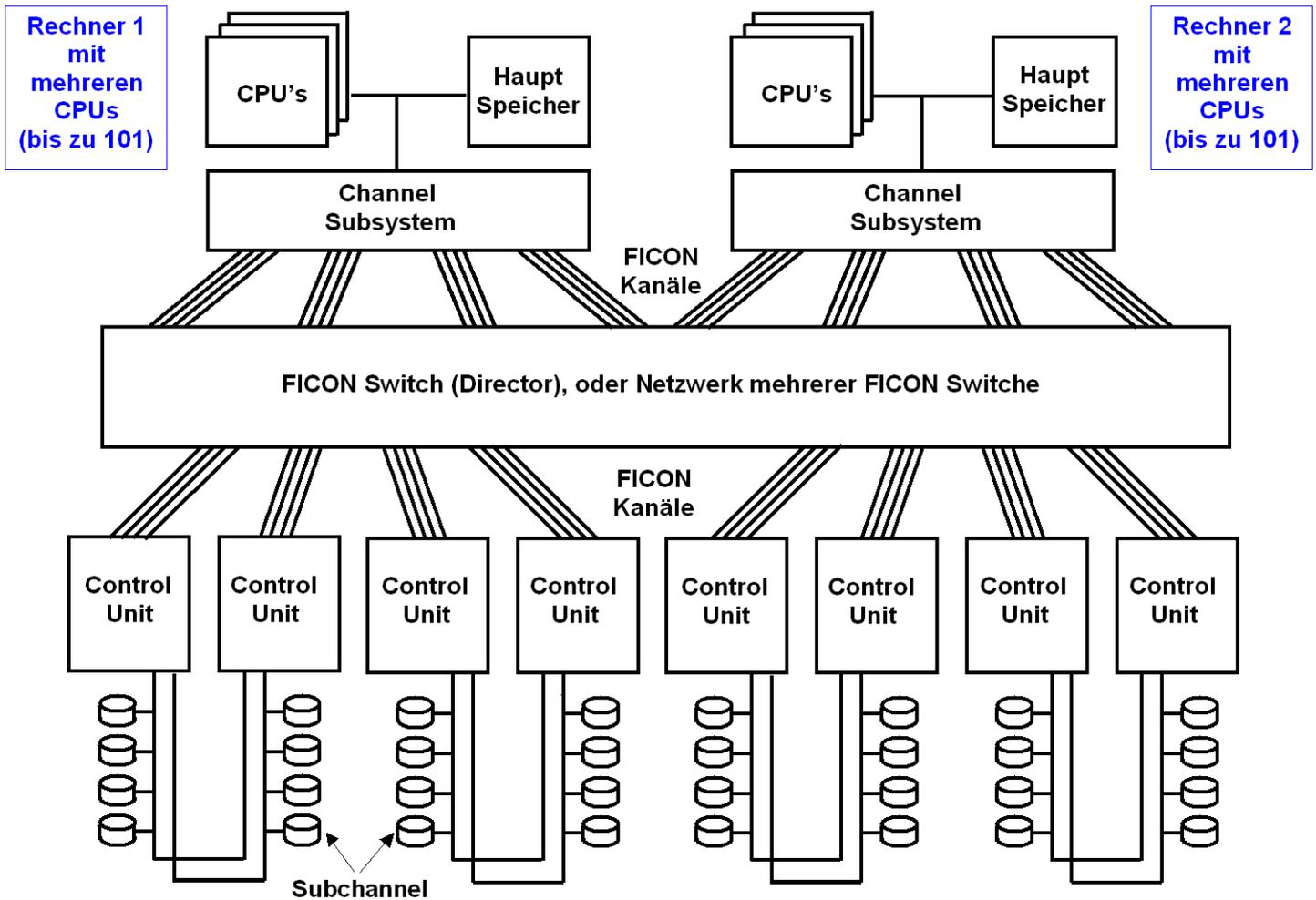


Abb. 5.2.10
FICON Netzwerk

Die obige Abbildung zeigt zwei Mainframe Rechner und zahlreiche Control Units mit ihren Plattenspeichern. Jeder zEC12 Mainframe Rechner kann bis zu 101 CPUs enthalten und verfügt über ein eigenes Channel Subsystem.

Es ist in der Regel notwendig, beide (bis zu 32) Mainframe Rechner mit allen Control Units und allen Plattenspeichern zu verbinden.

Da die Verkabelung sehr unübersichtlich ist, schaltet man einen FICON Switch (offizielle Bezeichnung „FICON Director“) zwischen die Channel Subsysteme und die Control Units. Vielfach wird an Stelle eines einzelnen FICON Switches ein ganzes Netzwerk von FICON Switches eingesetzt (nicht gezeigt).

Große Mainframe Installationen können über 10 000 Glasfaserkabel aufweisen.

Die I/O Anforderungen der Prozesse auf den einzelnen CPUs müssen ihren Weg durch das Netzwerk zu dem gewünschten Plattenspeicher finden. Die Channel Subsysteme der einzelnen Rechner haben die Aufgabe, das Routing der I/O Anforderungen über das FICON Netzwerk zu übernehmen. Ähnlich wie im Internet kann sich der Weg zwischen CPU und Plattenspeicher während der Ausführung einer I/O Operation dynamisch ändern.

Ein derartiges Netzwerk wird als „Storage Area Netzwerk“ (SAN) bezeichnet. Ein SAN benutzt das Fibre Channel Protokoll an Stelle der in Kommunikationsnetzen üblichen TCP/IP oder SNA Protokolle.

Die in Abb. 5.2.10 gezeigten beiden Channel Subsysteme haben eine weitere Aufgabe: Sie verbergen die komplexe I/O Konfiguration vor den z/OS Betriebssystemen, die auf den beiden Rechnern laufen. Ein z/OS Betriebssystem arbeitet mit der Illusion, dass alle Plattenspeicher über individuelle Punkt-zu-Punkt Verbindungen direkt an seine CPUs angeschlossen sind. Diese Verbindungen werden als „Subchannels“ bezeichnet; je ein Subchannel pro Plattenspeicher. Es ist die Aufgabe des Channel Subsystems, logische Subchannels auf physische Kanalpfade und die Topologie des FICON Netzwerkes dynamisch abzubilden.

5.2.12 Enterprise Storage Server

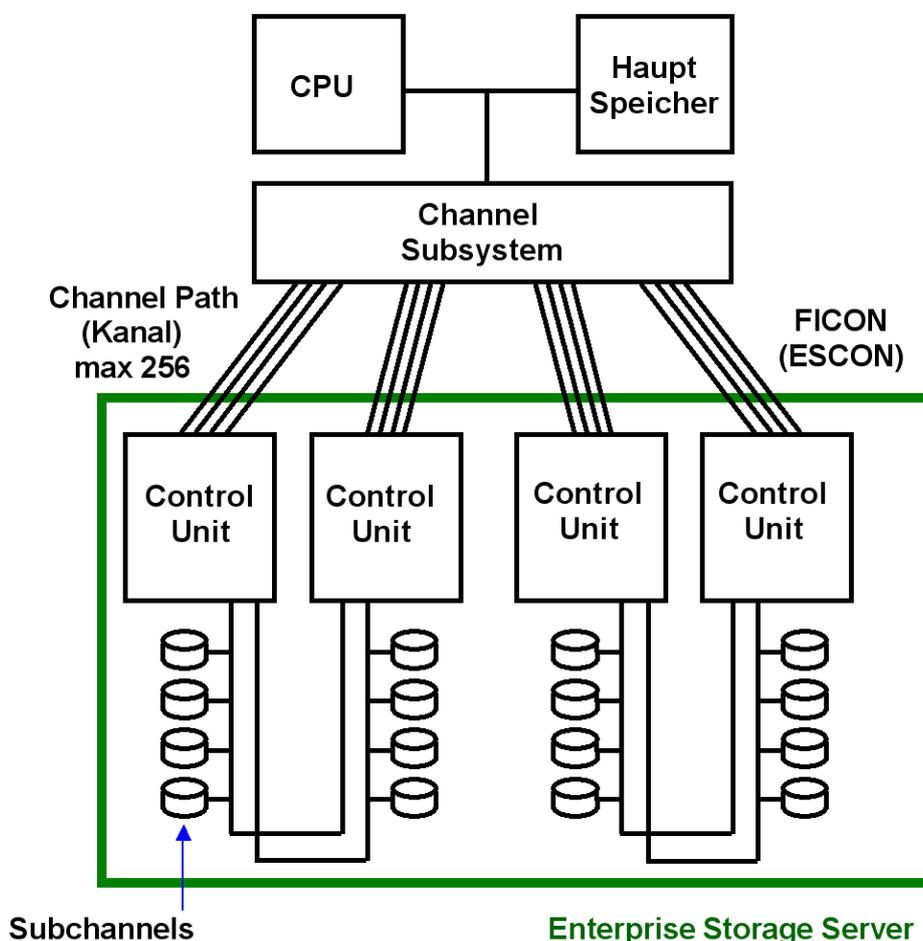


Abb. 5.2.11

Ein Enterprise Storage Server emuliert mehrere Control Units

Früher waren Control Units getrennte physische Einheiten in ihren eigenen Gehäusen.

Heute werden mehrere Control Units und angeschlossene Plattenspeicher zu einem physischen „Enterprise Storage Server“ (ESS) zusammengefasst, der auch die angeschlossenen Plattenspeicher enthält.

Der ESS emuliert mehrere logische Control Units, hat Anschlüsse für zahlreiche FICON Kanäle und bringt zahlreiche Plattenspeicher im gleichen Gehäuse unter.

Es können beliebig viele ESS angeschossen werden.

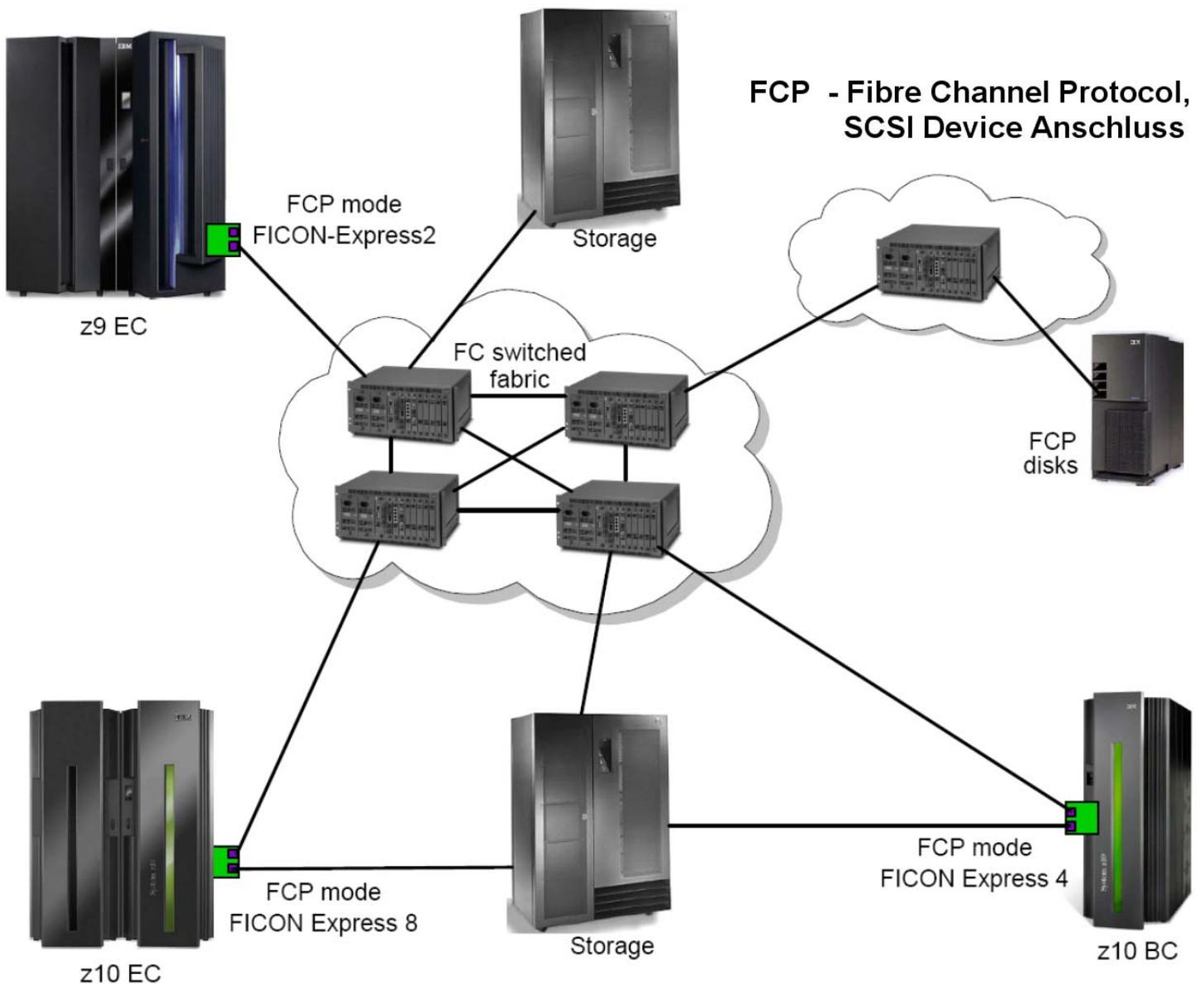


Abb. 5.2.12
Storage Area Netzwerk

Heutige Mainframes können alternativ neben FICON Plattenspeichern auch Fibre Channel SCSI Plattenspeicher anschließen. Dies ist z.B. beim Einsatz von zLinux üblich. Der Großteil der Daten einer Mainframe Installation wird aber auf FICON Platten gespeichert.

Über das Fibre Channel Storage Area Netzwerk können parallel FICON und FC SCSI Verbindungen geschaltet werden.

5.3 Mainframe I/O

5.3.1 Volume

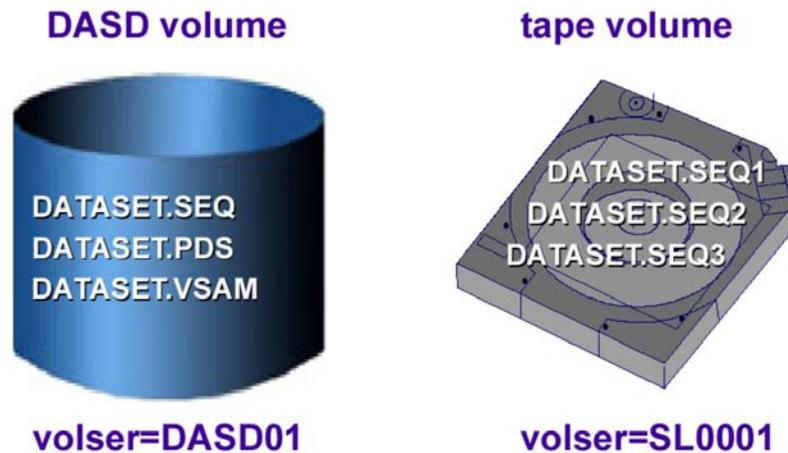


Abb. 5.3.1
Festplatten- und Magnetband - Volumen

Ein **“Volume”** (Datenträger) ist eine logische externe Speicher-Einheit, beispielsweise ein Festplattenspeicher oder eine Magnetbandkassette. Ein Volume speichert zahlreiche Dateien.

In einer Mainframe Installation ist ein jedes Volume durch eine eindeutige „Volume Serial Number“ (**volser**) gekennzeichnet, die auf dem Datenträger an einer bestimmten Stelle aufgezeichnet ist. Nicht alle vom Betriebssystem erfassten Volumen müssen in jedem Augenblick für das Betriebssystem zugreifbar sein. Beispielsweise kann eine Nachricht auf der Konsole den System-Administrator auffordern, eine Magnetbandkassette mit einer bestimmten Volume Serial Number manuell aus einem Regal zu entnehmen, und in eine angeschlossene Magnetbandeinheit einzulegen.

5.3.2 Channel und Control Unit

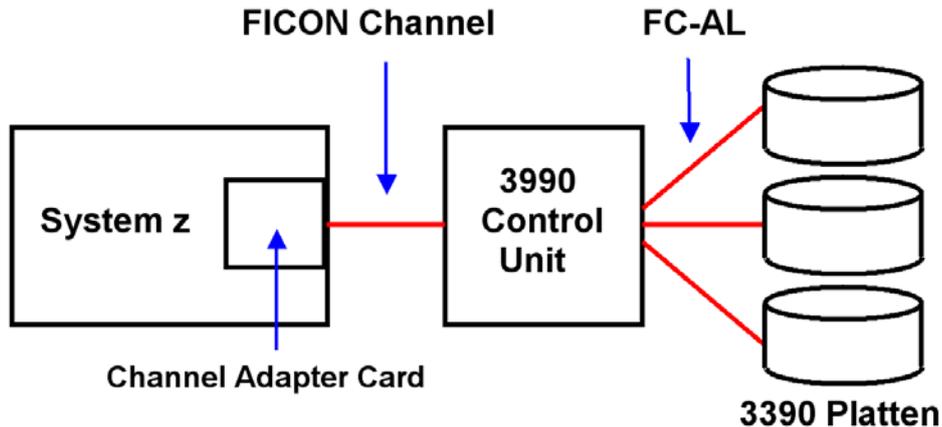


Abb. 5.3.2
Anschluss der Festplatten über eine Control Unit

Ein Kanal (Channel) ist ein Verbindungskabel zwischen einem Rechner und einer Plattenspeicher-Steuereinheit. Ein Channel Adapter ist eine Steckkarte im I/O Cage (I/O Drawer) eines Rechners, die Steckkontakte für das Kanalkabel enthält. Kanäle werden heute als FICON Fibre Channel Verbindungen implementiert.

Channel Adapter und Control Unit sind zwei physische Einheiten, die über ein Kanal-Kabel miteinander verbunden sind. Die Kombination stellt jedoch eine einzige logische Einheit dar. Die Aufteilung ist erforderlich, weil aus Platzgründen die Control Units in einer gewissen Entfernung voneinander und vom Rechner aufgestellt werden müssen.

In manchen Fällen ist es möglich, die Control Unit im gleichen Gehäuse wie die CPUs unterzubringen. In diesem Fall werden Control Unit und Channel Adapter als eine einzige Baugruppe implementiert. Ein Beispiel ist der OSA Adapter. Dies ist eine Steckkarte (I/O Card) im I/O Drawer eines Rechners, die zum Anschluss von Ethernet Verbindungen dient.

Note: Innerhalb IBM wird eine Festplatte regelmäßig als „DASD“ (Direct Access Storage Device, ausgesprochen „dasdi“ oder „dädsdi“) bezeichnet.

5.3.3 Ausführen einer I/O Operation

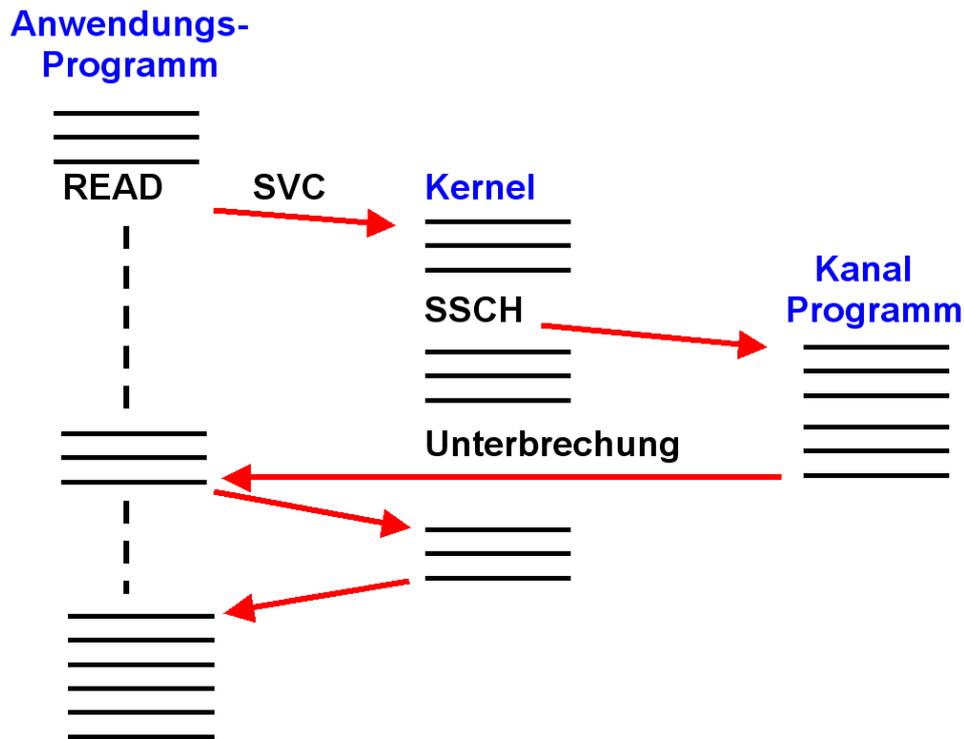


Abb. 5.3.3
d) Ausführen eines I/O Befehls

Unix Systeme steuern ihre Plattenspeicher über ein als I/O Driver bezeichnetes Programm, welches von der CPU ausgeführt wird. Die Funktion eines Unix I/O Drivers wird bei einem Mainframe zum allergrößten Teil in die Control Unit ausgelagert. Das dort ablaufende Programm wird als **Channel Program** bezeichnet. Ein Channel Program besteht aus einzelnen Befehlen, die als **Channel Command Words** (CCW) bezeichnet werden. Beim Starten einer I/O Operation werden die Befehle des Channel Programms aus dem Hauptspeicher ausgelesen, an die Control Unit übergeben und dort ausgeführt.

Abb. 5.3.3 zeigt das Zusammenspiel des Anwendungsprogramms im User State, des Kernels im Kernel State (Supervisor State) und des Kanalprogramms.

Das Anwendungsprogramm führt einen READ Befehl aus. Dies führt zu einem Aufruf des Kernels (Supervisor) über einen SVC Maschinenbefehl. Der Kernel ruft seine I/O Routinen auf und veranlasst zunächst die Erstellung, und danach die Ausführung des Kanalprogramms durch Kanal und Control Unit mittels eines Start Subchannel (SSCH) Maschinenbefehls. Die Control Unit führt die I/O Befehle (CCWs) der Reihe nach aus.

Der Abschluss der Kanalprogrammverarbeitung wird der CPU von der Control Unit mittels einer I/O Unterbrechung (Channel End Device End, CEDE) mitgeteilt.



Abb. 5.3.4
Einfaches Plattenspeicher - Steuerprogramm

Dargestellt ist ein einfaches Kanalprogramm und seine CCWs für die Datenübertragung vom/zum Plattenspeicher, wie es in den S/370 Rechnern gebräuchlich war. Heutige Kanalprogramme sind deutlich komplexer.

5.3.4 3390 Plattenspeicher

Control Units entlasten die CPUs, indem der I/O Driver Code außerhalb der CPUs in den Control Units ausgeführt wird. Zum Betriebssystem gehören nur Skelette für einige wenige I/O Driver Routinen, z.B. je eine generische Routine für Plattenspeicher, Magnetbänder, Drucker und Netzanschlüsse. Die I/O Driver Code Routinen werden als „Channel Programs“ bezeichnet, die jeweils aus einer Gruppe von Anweisungen, den „Channel Command Words“ (CCW) bestehen.

Das Anwendungsprogramm ergänzt das Channel Programm um Daten wie Adresse und Länge des I/O Buffers im Hauptspeicher oder die logische Adresse des I/O Gerätes (das Channel Subsystem bildet die logische auf die physische I/O Adresse ab). Der Betriebssystem Kernel überträgt daraufhin den Channel Program Code an die Control Unit, wo er ausgeführt wird.

Über viele Jahre war IBM – als Erfinder der Festplattenspeicher Technologie - der führende Hersteller von magnetischen Festplattenspeichern. Während dieser Zeit hatten IBM Festplattenspeicher einen Durchmesser von 14 Zoll.

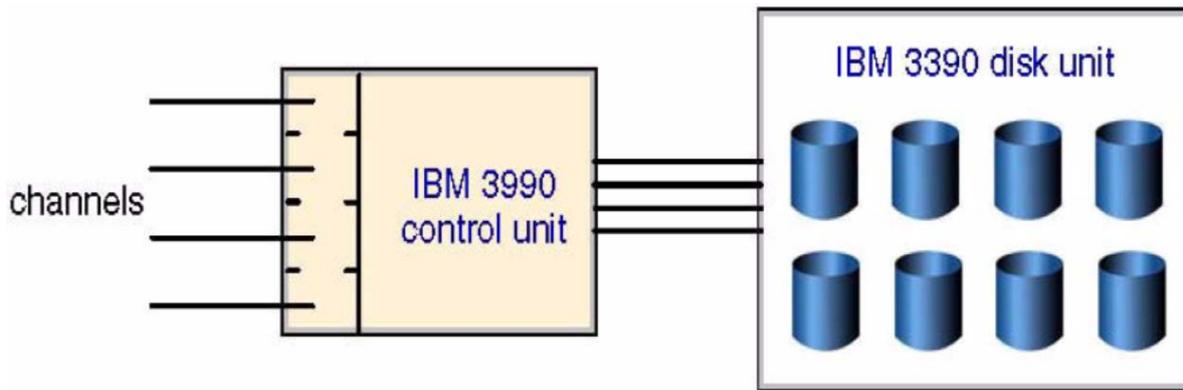


Abb. 5.3.5

Klassischer Anschluss eines 3390 Speichers an eine 3990 Control Unit

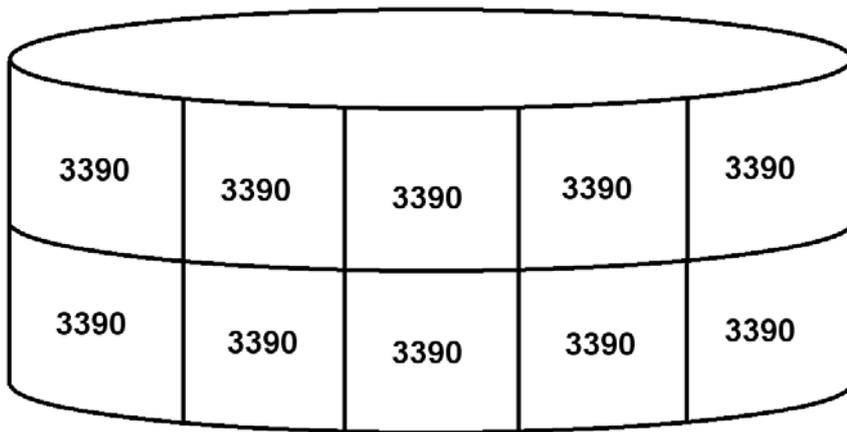
1993 wurde das letzte Modell dieser Art als IBM 3390 Plattenspeicher und der IBM 3990 Control Unit herausgebracht. Es hatte eine Speicherkapazität von 8,5 GByte. In der Folge wurde 5 ½ und 3 ¼ Zoll Festplatten eingesetzt.

Bis 1993 wurden für alle neuen Festplattenmodelle eigene Control Units und eigene Channel Programme entwickelt. Davon nahm man ab 1993 Abstand und standardisierte die Software Unterstützung auf das 3390 Modell.

Um dies zu erreichen, arbeitet der OS Kernel und die Control Unit mit der Illusion eines generischen 3390 Festplattenspeichers. Die Eigenschaften dieses physischen Plattenspeichers wurden für die Definition der logischen 3390 Plattenspeicher übernommen. Parameter wie Spuren pro Zylinder und Bytes pro Spur sind festgelegt, und sind wie bei allen 3390 Plattenspeichern identisch. Die Anzahl der Zylinder pro Plattenspeicher kann als Parameter definiert werden, woraus sich unterschiedliche Speicherkapazitäten ergeben.

Das Betriebssystem kennt nur 3390 Festplatten mit deren Struktur bezüglich Anzahl Zylinder, Anzahl Spuren, Bytes pro Spur usw. Das Nachfolgemodell des 3390 Plattenspeichers war der IBM „Enterprise Storage Server“ (ESS) . Das ursprüngliche „Shark“ Modell wurde später durch die „DS6000“ und „DS8000“ Enterprise Storage Server abgelöst.

Diese ESS benutzen bis zu 1024 Standard 3 ½ Zoll oder 2 ½ Zoll SCSI Festplatten mit bis zu je 600 GByte Speicherkapazität sowie zwei 4-way PowerPC Multiprozessoren für die Emulation mehrerer 3990 Control Units, für die Emulation der angeschlossenen 3390 Plattenspeicher und für weitere fortschrittliche Funktionen, u.A. die Verwaltung sehr großer Plattenspeicher-Caches. Ein Enterprise Storage Server kann unterschiedliche Arten von Festplatten enthalten, mit unterschiedlicher Speicherkapazität und Struktur. Es ist die Aufgabe des Enterprise Storage Servers, die virtuellen 3390 Plattenspeicher auf die tatsächlich eingesetzten physischen Festplatten und deren Struktur abzubilden. Spezifisch haben heutige Festplatten eine größere Speicherkapazität als die virtuellen 3390 Plattenspeicher. Es werden deshalb immer mehrere virtuelle 3390 Platten auf einer physischen Festplatte emuliert.



Beispiel:
 Ein etwa 100 GByte
 großer 3 ½ Zoll
 Festplatte, die zehn
 Plattenspeicher vom
 Typ 3390 - 9 emuliert

Abb. 5.3.6
 Physisches und Logisches Volume

Emulierte 3390 Plattenspeicher werden als „**Logical Volumes**“ (LV) bezeichnet. Sie existieren als zwei unterschiedliche standardisierte Größen, die als „Modell 3“ und „Modell 9“ bezeichnet werden. Die Spur- (track) Geometrie innerhalb einer Modellserie ist immer identisch.

Eine 3390-Modell 3 Plattenspeicher ist in 3339 Cylinder aufgeteilt, mit 15 Tracks pro Cylinder. Ein Track hat hierbei eine Kapazität von 56,664 Bytes, woraus sich eine Gesamtkapazität von 2.84 GByte für die Festplatte ergibt. Das Modell 9 hat die 3-fache Anzahl von Spuren (10 017) und damit die 3-fache Kapazität des Modell 3 (8,51 GByte). Auch hier sind 56,664 Bytes pro Spur vorhanden. Das z/OS Betriebssystem ist häufig auf Model 3 Platten installiert.

Tatsächlich hängt die effektive Kapazität von der Größe der Blocksize ab, mit der Data Sets angelegt (allocated) werden, und der Struktur der Data Sets. So können z.B. VSAM Data Sets maximal 2,3 GByte an Daten auf einer Modell 3 Festplatte speichern; der Rest wird für Verwaltungsinformation benötigt.

	3390-3	3390-9
Zylinder pro Plattenstapel	3 339	10 017
Spuren pro Zylinder	15	15
Bytes pro Spur	56 664	56 664
Bytes pro Zylinder	849 960	849 960
MByte pro Plattenstapel	2 838	8 514
4096 Byte Blocks pro Spur	12	12

Abb. 5.3.7
 DASD Speicherkapazität der Modelle 3390

Die in Abb. 5.3.6 wiedergegebenen Daten sind für das Allocate von Data Sets interessant. Beispielsweise sollte der Parameter BLKSIZE so gewählt werden, dass es bei 56 664 Bytes/Spur möglichst wenig Verschnitt gibt. Bei der Optimierung hilft ein BLKSIZE Calculator, siehe <http://webspaces.webring.com/people/lp/programmingstuff/blksize.htm> .

Es bietet sich das "half-track blocking" an: 2 Blöcke pro Spur. Eine Blockgröße von 27 998 Bytes ($55\,996 / 2 = 27\,998$) ermöglicht 2 Blöcke/Spur, oder eine Nutzung von 99 %. Eine Blockgröße von 28 000 Bytes gestattet nur einen Block pro Spur. Dies bedeutet einen Verschnitt von etwa 50 %.

Eine populäre Wahl ist, 349 logische Record zu je 80 Bytes in einen Block von 27920 Bytes zu packen.

5.3.5 System z I/O-Konfiguration

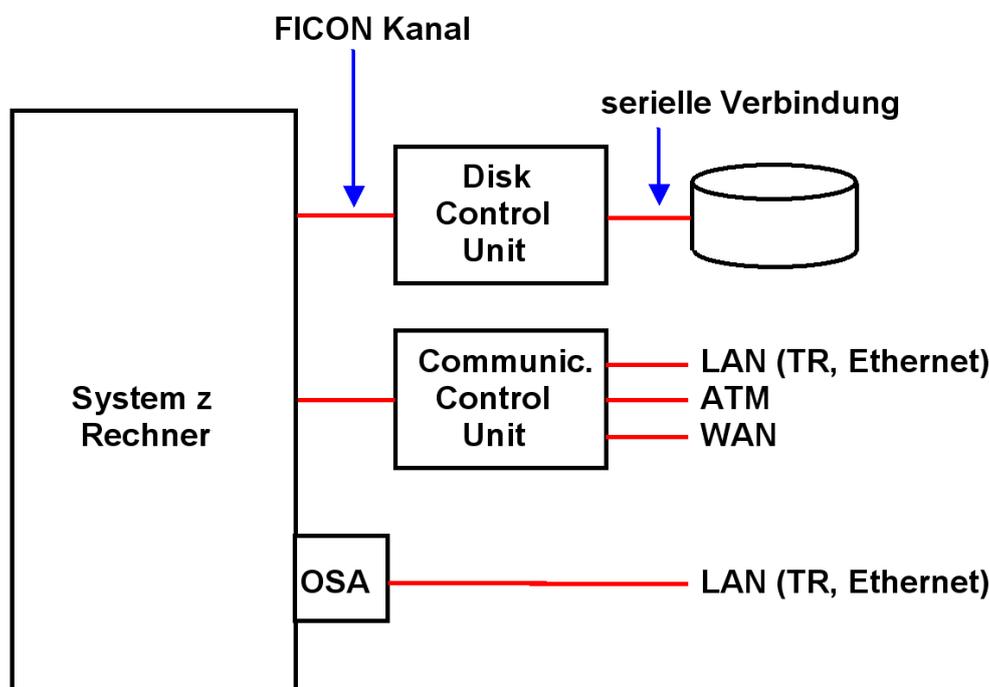


Abb. 5.3.8
Der OSA Adapter integriert die Control Unit Funktion

I/O Geräte werden grundsätzlich über Steuereinheiten (Control Units) angeschlossen. Steuereinheiten sind meistens in getrennten Boxen untergebracht, und über Glasfaser (z.B. FICON) an den System z Rechner angeschlossen.

Es existieren viele unterschiedliche Typen von Steuereinheiten. Die wichtigsten schließen externe Speicher (Platten, Magnetband-Archivspeicher) und Kommunikationsleitungen an.

Es existieren Steuereinheiten für viele weiteren Gerätetypen. Beispiele sind Belegleser für Schecks oder Drucker für die Erstellung von Rentenbescheiden.

5.3.6 Channel- to Channel Verbindung

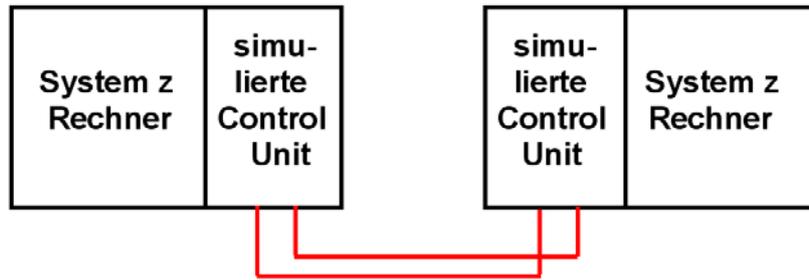


Abb. 5.3.9
Full Duplex Channel to Channel Verbindung

Die **Channel- to Channel** (CTC) Verbindung wird durch eine Hardware Einrichtung eines zSeries Systems verwirklicht, die dieses System gegenüber einem anderen zSeries System wie eine I/O Einheit erscheinen lässt.

Für eine Full Duplex Verbindung werden normalerweise zwei CTC Verbindungen eingesetzt.

Die „Cross-System Coupling Facility“ (**XCF**) ist eine Komponente des z/OS Betriebssystems. Sie verwendet das CTC Protokoll und die CTC Hardware. Sie stellt die Coupling Services bereit, mit denen z/OS Systeme innerhalb eines Clusters (Sysplex) miteinander kommunizieren.

5.3.7 RAID-System

Ein RAID-System (ursprünglich redundant array of inexpensive disks, heute redundant array of independent disks) dient zur Organisation mehrerer physischer Festplatten eines Computers zu einem logischen Laufwerk, das eine höhere Datensicherheit bei Ausfall einzelner Festplatten und/oder einen größeren Datendurchsatz erlaubt als eine physische Festplatte. Während die meisten in Computern verwendeten Techniken und Anwendungen darauf abzielen, Redundanzen (das Vorkommen doppelter Daten) zu vermeiden, werden bei RAID-Systemen redundante Informationen gezielt erzeugt, damit beim Ausfall einzelner Komponenten das RAID System als Ganzes seine Funktionalität behält.

Der Begriff wurde von Patterson, Gibson und Katz 1987 an der University of California, Berkeley in ihrer Arbeit „A Case for Redundant Array of Inexpensive Disks (RAID)“ zum ersten Mal verwendet. Darin wurde die Möglichkeit untersucht, kostengünstige Seagate Festplatten im Verbund als logisches Laufwerk zu betreiben, um die Kosten für einen großen (zum damaligen Zeitpunkt sehr teuren) 14 Zoll IBM Festplattenspeicher einzusparen. Dem gestiegenen Ausfallrisiko im Verbund sollte durch die Speicherung redundanter Daten begegnet werden, die einzelnen Anordnungen wurden als RAID-Level diskutiert.

Die weitere Entwicklung des RAID-Konzepts führte zunehmend zum Einsatz in Serveranwendungen, die den erhöhten Datendurchsatz und die Ausfallsicherheit nutzen. Der Aspekt der Kostenersparnis wurde dabei aufgegeben. Heute existiert meistens die Möglichkeit, in einem solchen System einzelne Festplatten im laufenden Betrieb zu wechseln.

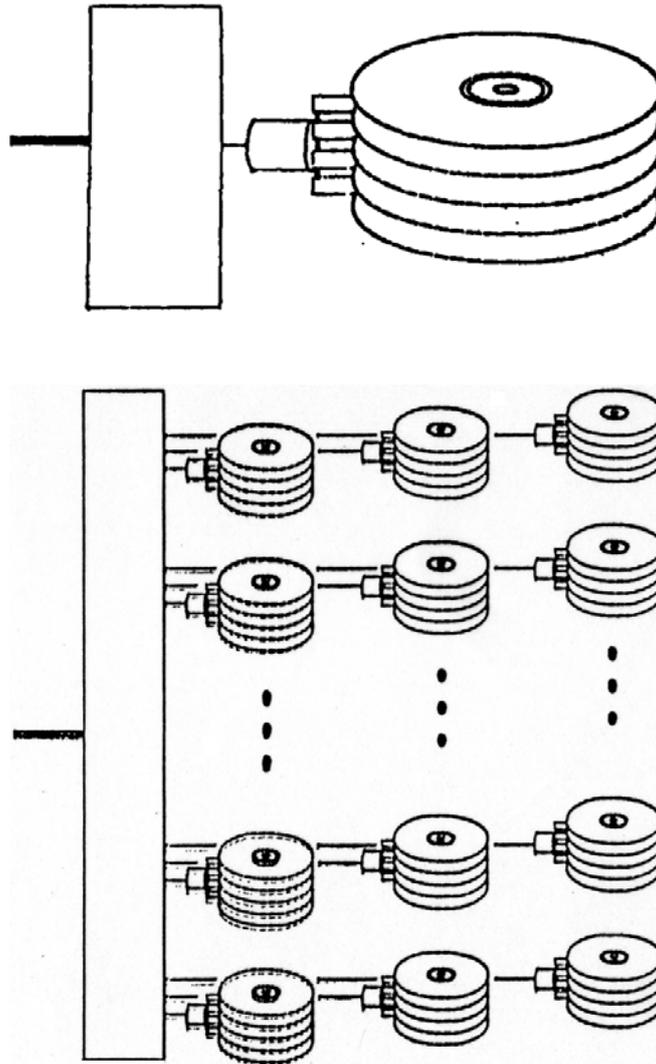


Abb. 5.3.10
Ersatz eines einzigen 14 Zoll Festplattenspeichers

Die Idee ist, dass das Disk Array für das Betriebssystem wie eine einzige Platte aussieht.

5.3.8 RAID-Level

Der Betrieb eines RAID-Systems setzt mindestens zwei Festplatten voraus. Die Festplatten werden gemeinsam betrieben und bilden einen Verbund, der leistungsfähiger ist als die einzelnen Festplatten. Mit RAID-Systemen kann man folgende Vorteile erreichen:

- Erhöhung der Ausfallsicherheit (Redundanz)
- Steigerung der Daten-Transferraten (Leistung)
- Aufbau großer logischer Laufwerke
- Austausch von Festplatten und Erhöhung der Speicherkapazität während des Systembetriebes
- Kostenreduktion durch Einsatz mehrerer preiswerter Festplatten

Die genaue Art des Zusammenwirkens der Festplatten wird durch den **RAID-Level** spezifiziert. Die gebräuchlichsten RAID-Levels sind RAID 0, RAID 1, RAID 5, RAID 6 und RAID 10.

Aus Sicht des Benutzers, eines Anwendungsprogramms oder des Betriebssystems unterscheidet sich ein logisches RAID-Laufwerk nicht von einer einzelnen physischen Festplatte.

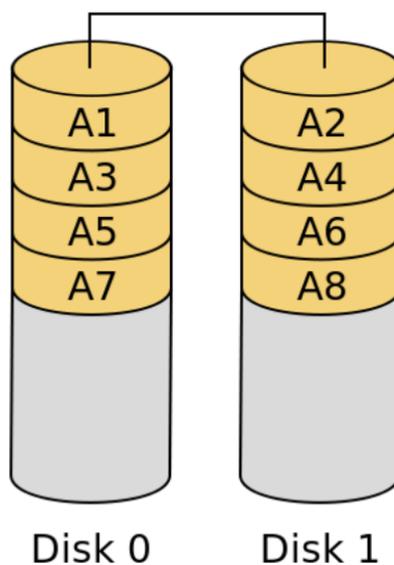


Abb. 5.3.11
RAID 0

RAID 0 bietet gesteigerte Transferraten, indem die beteiligten Festplatten in zusammenhängende Blöcke gleicher Größe (z.B. 16KB) in Streifen (eng. stripes) aufgeteilt werden, wobei jeder Streifen eines Datenblocks auf einer separaten Festplatte gespeichert wird. Diese Blöcke werden quasi im Reißverschlussverfahren zu einer großen logischen Festplatte angeordnet. Somit können Zugriffe auf allen Platten parallel durchgeführt werden (engl. striping, was „in Streifen zerlegen“ bedeutet).

Die Datendurchsatz-Steigerung (bei sequentiellen Zugriffen, aber besonders auch bei hinreichend hoher Nebenläufigkeit) beruht darauf, dass die notwendigen Festplatten-Zugriffe in höherem Maße parallel abgewickelt werden können.

Streng genommen handelt es sich bei RAID 0 nicht um ein wirkliches RAID, da es keine Redundanz gibt. Beim Ausfall einer Festplatte sind die Daten des gesamten RAID 0 Verbandes verloren.

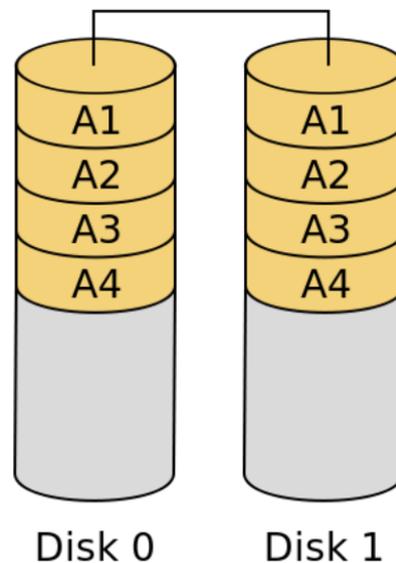


Abb. 5.3.12
RAID 1

RAID 1 ist der Verbund von mindestens 2 Festplatten. Ein RAID 1 speichert auf beiden Festplatten die gleichen Daten auch als Spiegelung bezeichnet. Beim Ausfall einer Platte sind die Daten identisch auf der zweiten Festplatte vorhanden. Beim Spiegeln von Festplatten an einem Kanal spricht man von Disk Mirroring, beim Spiegeln an unabhängigen Kanälen von Disk Duplexing (zusätzliche Sicherheit).

Fällt eine der gespiegelten Platten aus, kann die andere weiterhin alle Daten liefern. Besonders für sicherheitskritische Echtzeitanwendungen ist das unverzichtbar. RAID 1 ist eine einfache und schnelle Lösung zur Datensicherheit und Datenverfügbarkeit, besonders geeignet für kleinere Nutzkapazitäten. Lediglich die Hälfte der Gesamtkapazität steht als nutzbarer Bereich zur Verfügung.

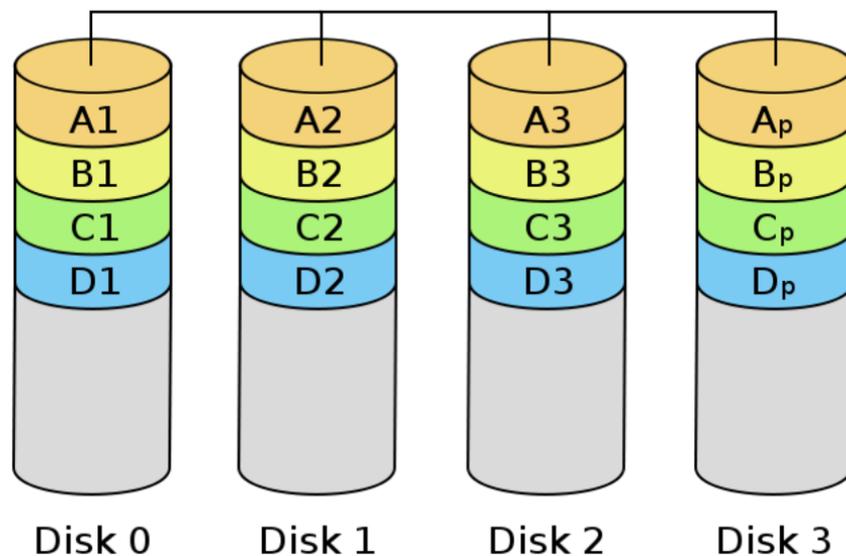


Abb. 5.3.13
RAID 3 und 4

RAID 4 verteilt wie bei RAID 0 die Daten auf mehrere Festplatten. Auf einem Sicherheitslaufwerk werden Paritätsdaten abgelegt. Durch diese Parität stehen selbst bei einem Ausfall einer Festplatte alle Daten weiterhin zur Verfügung. Lediglich die Kapazität einer Festplatte geht für die Redundanz verloren. Bei einem 4 Verbund mit 5 Festplatten stehen 80 Prozent der Gesamtkapazität als Nutzkapazität zur Verfügung.

Beim Schreiben kleiner Datenblöcke wird das Paritätslaufwerk sehr stark belastet was die Performance deutlich negativ beeinflusst. RAID 4 arbeitet im Gegensatz zu RAID 3 mit unabhängigen Datenpfaden für jedes Laufwerk. Dies bringt vor allem beim Schreiben und Lesen großer Dateien eine bessere Performance.

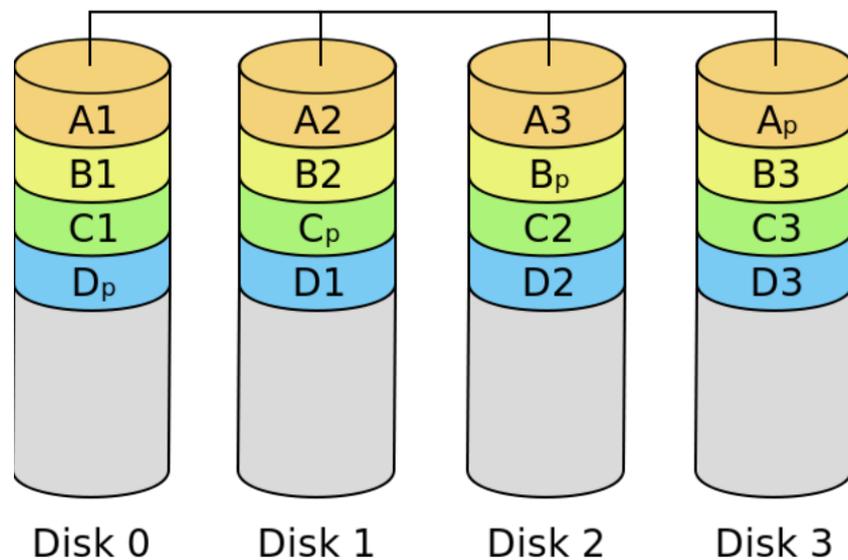


Abb. 5.3.14
RAID 5

Anders als bei RAID 4 verteilt **RAID 5** die Paritätsdaten A_p, B_p, C_p und D_p gleichmäßig auf allen Festplatten im Verbund. Dies garantiert bei allen Zugriffen eine optimale Auslastung der Laufwerke. Selbst bei zufallsbedingten (random) Zugriffen, wie sie für ein multitasking multiuser Betriebssystem typisch sind, kann somit eine optimale Performance erreicht werden. RAID 5 bietet beim Ausfall einer Festplatte die gleiche Sicherheit und Datenverfügbarkeit wie RAID 4.

In schreibintensiven Umgebungen mit kleinen, nicht zusammenhängenden Änderungen ist RAID 5 benachteiligt, da bei zufälligen Schreibzugriffen der Durchsatz aufgrund des zweiphasigen Schreibverfahrens deutlich abnimmt (an dieser Stelle wäre eine RAID-0+1-Konfiguration vorzuziehen).

RAID 5 ist eine der kostengünstigsten Möglichkeiten, Daten auf mehreren Festplatten redundant zu speichern und dabei das Speichervolumen effizient zu nutzen.

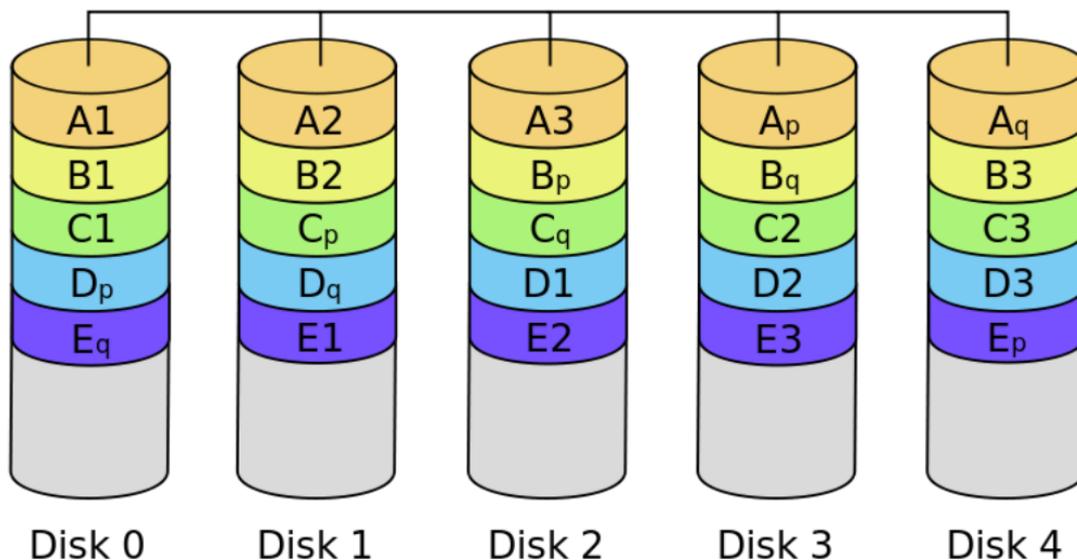


Abb. 5.3.15
RAID 6

Wenn bei RAID 5 eine Festplatte ausfällt, wird der laufende Betrieb dadurch zunächst nicht beeinträchtigt. Die fehlerhafte Festplatte muss baldigst durch eine neue Festplatte ersetzt werden. Nachdem dies geschehen ist, kann ein Unterprogramm des Enterprise Storage Servers den Inhalt der neuen Platte aus den Inhalten der restlichen RAID 5 Platten automatisch generieren. Bis dies geschehen ist, ist die volle Ausfallsicherheit nicht mehr gewährleistet.

RAID 6 funktioniert ähnlich wie RAID 5, verkräftet aber den gleichzeitigen Ausfall von bis zu zwei Festplatten. Insbesondere beim intensiven Einsatz hochkapazitiver Festplatten kann die Wiederherstellung der Redundanz nach einem Plattenausfall viele Stunden dauern. Während dieser Zeit besteht kein Schutz gegen einen weiteren Ausfall.

Im Gegensatz zu RAID 5 gibt es bei RAID 6 mehrere mögliche Implementierungsformen, die sich insbesondere in der Schreibleistung und dem Rechenaufwand unterscheiden, und von unterschiedlichen Herstellern unter dem Namen RAID 6 vertrieben werden. Im allgemeinen gilt: Bessere Schreibleistung wird durch erhöhten Rechenaufwand erkaufte.

Im einfachsten Fall wird eine zusätzliche XOR-Operation über eine orthogonale Datenzeile berechnet. Auch die zweite Parität wird rotierend auf alle Platten verteilt. Eine andere RAID 6 Implementierung rechnet mit nur einer Datenzeile, produziert allerdings keine Paritätsbits, sondern einen Zusatzcode, der 2 Einzelbit-Fehler beheben kann. Das Verfahren ist rechnerisch aufwändiger.

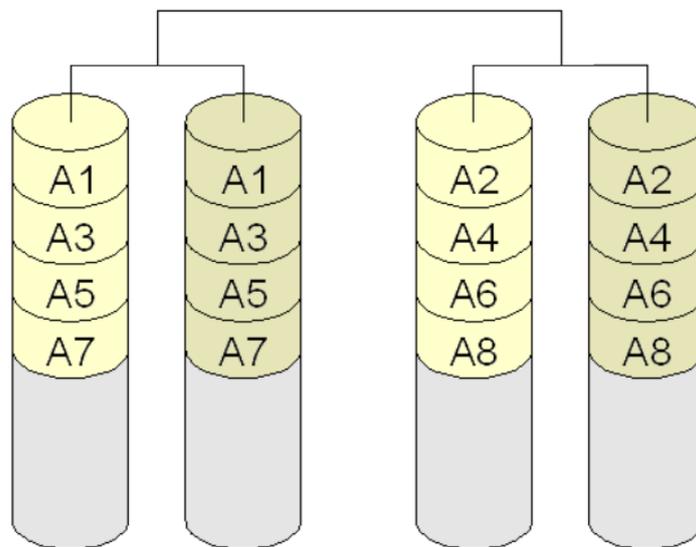


Abb. 5.3.16
RAID 10

Aus einer Kombination von RAID 0 (Performance) und RAID 1 (Datensicherheit) ist der RAID Level 10 entstanden. Rapid 10 ist eine Kombination aus RAID 0 + 1. Dabei werden immer $2 * n$ Platten zu einem RAID 0 zusammen gefasst - und dann per RAID 1 miteinander verbunden. Dabei ist $n \geq 2$.

RAID 10 Verbände bieten optimale Performance bei optimaler Ausfallsicherheit. Wie bei RAID 0 wird die optimale Geschwindigkeit allerdings nur bei sequentiellen Zugriffen erreicht und wie bei RAID 1 gehen 50 Prozent der Gesamtkapazität für die Redundanz verloren.

http://en.wikipedia.org/wiki/Standard_RAID_levels

5.3.9 Persistenz

Auf Festplatten abgelegte Daten haben den Vorteil, dass bei einem Stromausfall (oder beim Abschalten eines Rechners) Daten nicht verloren gehen.

Mit einem ausreichend hohem RAID Aufwand kann erreicht werden, dass Daten beliebige und auch sehr seltene Fehlerfälle intakt überstehen. Dieses Ziel wird in Mainframe Installationen häufig mit RAID 6 Systemen erreicht, die in zwei unterschiedlichen geografischen Standorten gespiegelt werden. Eine moderne Mainframe Installation geht heute davon aus, dass RAID Daten beliebig hohe Sicherheitsanforderungen erfüllen und nie verloren gehen.

Derartig gespeicherte Daten werden als persistent bezeichnet. Daten im Hauptspeicher eines Rechners sind nicht persistent, da sie bei einem Stromausfall verloren gehen, oder bei einem Hardware oder Software Fehler beschädigt werden können.

Die persistente Speicherung von Daten ist besonders bei der Transaktionsverarbeitung ein wichtiges Kriterium.

5.4 Enterprise Storage Server

5.4.1 Festplattenspeicher-Cache

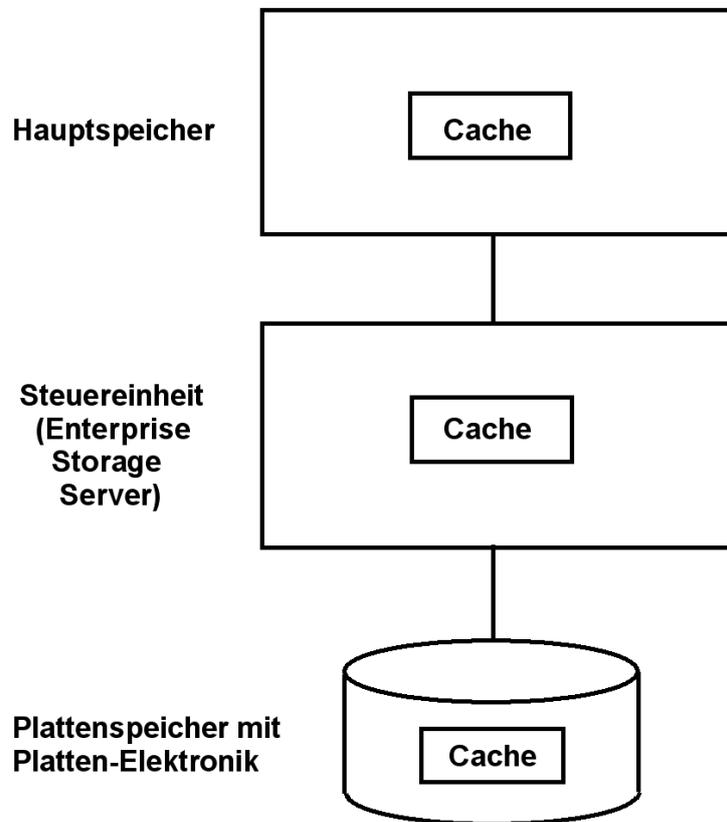


Abb. 5.4.1
Drei unterschiedliche Arten des Festplatten Cache

Zugriffe zu einem Festplattenspeicher benötigen Millisekunden. Um die Zugriffseigenschaften zu verbessern, wird bei jedem Lesezugriff ein größerer Block an Daten (evtl. eine ganze Spur oder ein ganzer Zylinder) in einen Cache Speicher gelesen. Vor allem bei einer Folge von sequentiellen Zugriffen können diese dann aus dem Cache befriedigt werden

Ein Cache für Plattenspeicherdaten kann sich im Hauptspeicher (Buffer Pool bei Datenbanken), und/oder in der Steuereinheit (Enterprise Storage Server) und/oder auf dem Plattenspeicher selbst befinden.

Unter z/OS wird der Plattenspeicher Cache im Hauptspeicher meistens als „Buffer Pool“ realisiert. Unabhängig davon unterhält der Enterprise Storage Server einen umfangreichen Plattenspeicher Cache. Auch die Elektronik eines Plattenspeichers unterhält heute in der Regel einen weiteren Cache.

In der historischen Entwicklung wurden zuerst Control Units um einen Cache-Speicher erweitert. Dieser ermöglichte es ihnen, einen Teil der Lese Zugriffe auf die physischen Platten zu vermeiden. Die ersten Control Unit Caches hatten Größen von nur 16 bis 32 MByte. Mit den Caches wurde es notwendig, Mikrocode zu entwickeln, der in der Control Unit ablief und eine effiziente Datenpufferung ermöglichte.

In den nächsten Schritten wurden die Caches größer, und dann wurden mit dem Aufkommen von preisgünstig zu produzierenden 3 ½ Platten die großen 14 Zoll Platten ersetzt. Jetzt wurden auch die Datenstrukturen und Plattenformate wie Count-Key-Data als real existierende Formate aufgelöst und durch die Control Units emuliert. Diese Control Units hatten damit bereits eine Komplexität erreicht, die sie zu eigenständigen Systemen machte.

Lese-Zugriffe können mit Hilfe eines Caches deutlich beschleunigt werden. Wünschenswert ist es, auch **Schreib**vorgänge zu beschleunigen, in dem Daten zunächst in den Cache, und dann asynchron auf die Platte geschrieben werden.

Dies ist grundsätzlich problematisch. Wenn ein Problem auftritt, ehe das Schreiben der Daten aus dem Cache auf den Plattenspeicher abgeschlossen ist, gehen die Daten verloren. Ein typisches Beispiel ist ein Stromausfall. Wenn zum Zeitpunkt eines Stromausfalls der Cache nicht vollständig geleert wurde, sind Daten verloren gegangen.

Als Lösung bildet man einen Teil des Enterprise Storage Server (ESS) Caches als Non-Volatile Storage (NVS) aus. Dies ist ein Halbleiterspeicher mit einer eigenen Batterie zur Stromversorgung. Letztere stellt sicher, dass bei einem Stromausfall (oder in anderen Fehlerfällen) die NVS Daten nicht verloren gehen. Weitere Einrichtungen stellen sicher, dass eine I/O Operation als abgeschlossen gelten kann, wenn die Daten im NVS gelandet sind. Der entgeltliche Datentransfer zum Plattenspeicher erfolgt dann asynchron und unbemerkt vom Betriebssystem.

5.4.2 Enterprise Storage Server

Front end "Host Adapter" Anschlüsse für den Anschluss an FICON Kanäle

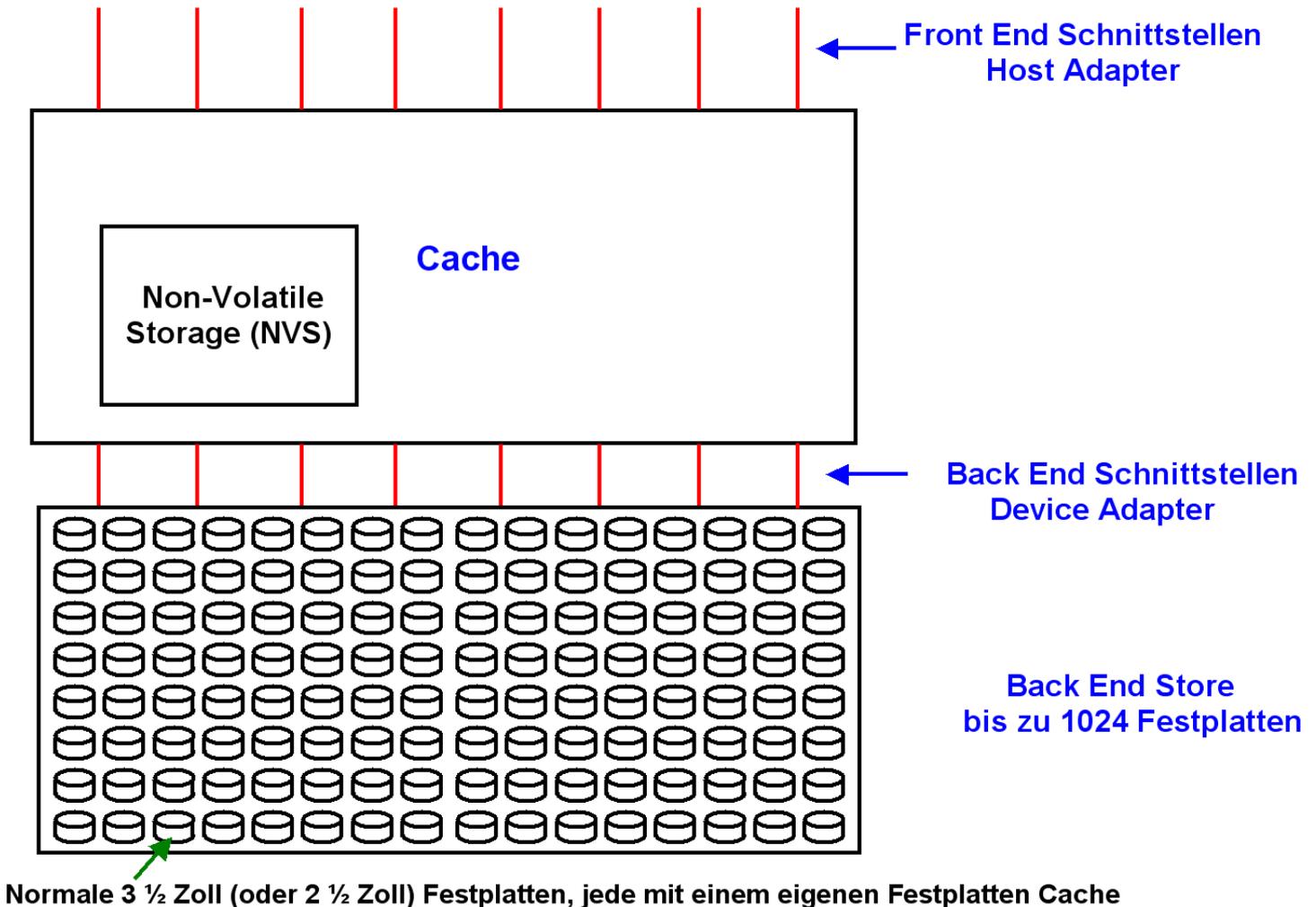


Abb. 5.4.2
Aufbau eines Enterprise Storage Servers

Im Wesentlichen besteht jeder Enterprise Storage Server aus vier Teilen:

1. Front End, welches die Schnittstelle zu den Rechnern darstellt (Host Adapter). Beim Anschluss an System z Rechner sind dies Host Adapter für FICON-Kanäle. UNIX-Systeme verwenden Host Adapter für das Fibre Channel SCSI Protokoll.
2. Ein oder (aus Zuverlässigkeitsgründen) zwei Multiprozessoren plus Cache, welcher aus zwei Teilen besteht: Einem Cache für Daten, die gelesen werden können, und einem Cache für Daten, die geschrieben werden sollen. Letzterer heißt Non-Volatile Storage (NVS) und bezeichnet damit einen Cache, der extra gegen Stromausfälle und andere Störfälle gesichert ist, z.B. durch eine Pufferung mit Batterien .
3. Back-End-Schnittstellen (Device Adapter), welche bei den meisten heutigen Enterprise Storage Servern FC-AL (Fibre Channel Arbitrated Loop) Anschlüsse sind.

4. Back End Store. Dieser besteht aus zahlreichen Festplatten und kann unterschiedlich sein. So bauen einige Hersteller SCSI-Platten ein, während andere Hersteller FATA oder SATA Disk Arrays bevorzugen. Jede der Platten verfügt noch einmal, ähnlich wie PC-Platten, über einen eigenen kleinen Cache.

Heutige Enterprise Storage Server besitzen sehr große Caches von z.B. 256 GByte. Der Non-Volatile Storage kann deutlich kleiner sein, da er nur zum vorübergehenden Zwischenspeichern der Schreibzugriffe benötigt wird. Zu schreibende Daten werden dann asynchron auf den Back End Store geschrieben, ohne dass die Anwendung davon etwas bemerkt.

Der Enterprise Storage Server übernimmt die Funktion mehrerer Control Units.

Die bedeutendsten Hersteller von Enterprise Storage Servern sind die Firmen EMC, IBM, Hitachi, MaxData und StorageTek, die im internen Aufbau alle große Ähnlichkeiten haben. Als Beispiel wird im folgenden der IBM DS8700 Enterprise Storage Server beschrieben. In vielen Fällen setzen Mainframe Installationen Enterprise Storage Server anderer Hersteller ein; besonders Enterprise Storage Server der Firma EMC sind häufig anzutreffen.

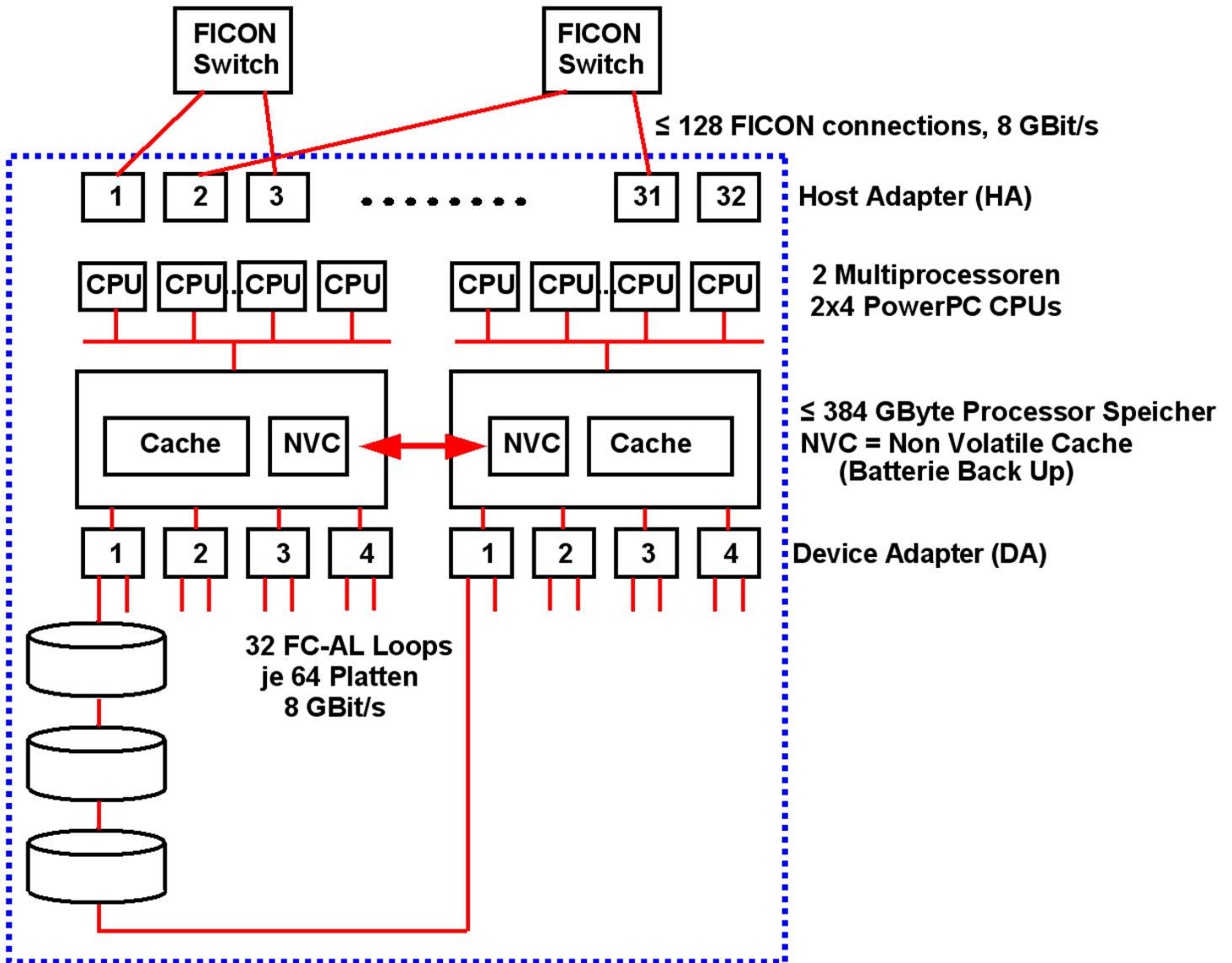


Abb. 5.4.3
IBM DS8700 Enterprise Storage Server

Zur Verbindung mit dem Host besitzt der DS8700 Enterprise Storage Server bis zu 32 Host Adapter (HA), die entweder je 4 FC-SCSI-Verbindungen oder je 4 FICON bzw. Fibre Channel Verbindungen nach außen implementieren.

Intern besteht der Storage Processor aus 2 unabhängigen Rechnern (Cluster), die jeweils aus (bis zu 4) PowerPC Prozessoren, dem Cache Storage und dem Non-Volatile Storage bestehen. Der Non-Volatile-Cache wird für die Zwischenspeicherung von Schreiboperationen benutzt. Die Idee ist: Wenn Daten einmal im ESS angekommen sind, gelten sie als sicher (persistent). Die Anwendung muss nicht das Schreiben auf den Festplatte abwarten.

Die beiden Cluster emulieren mehrere 3390 Control Units. Sie verfügen über getrennte Stromversorgungen und verhalten sich wie zwei unabhängige Rechner in der gleichen Box, außer dass sie über ein internes Netzwerk miteinander in Verbindung stehen. Zum Back Store besitzt jeder Cluster 8 Device Adapter mit je 4 FC-AL Ports. Die Adapter arbeiten immer paarweise, und die Plattenstränge (Disk Arrays) oder **Ranks** sind über eine FC-AL Loop mit den Device Adaptern verbunden.

FC-AL stellt eine serielle Kreisverbindung (Loop) für SCSI-Platten dar. Es existieren 2 Lese- und 2 Schreibverbindungen, von denen jede mit 40 MByte/s arbeitet, was eine Gesamtkapazität von 160 MByte/s ergibt.

Es werden 300, 450 oder 600 GByte Festplatten eingesetzt. Alternativ kann ein Teil auch aus Solid State Drives (SSD) bestehen. SSDs sind sehr teuer, bewähren sich aber für I/O-intensive Workloads. Sie ermöglichen eine bis zu 100fache Verbesserung des Throughput und bis zu 10fache bessere Antwortzeit als mit 15K U/min rotierende Festplatten. Sie verbrauchen auch weniger Energie als rotierende Festplatten.

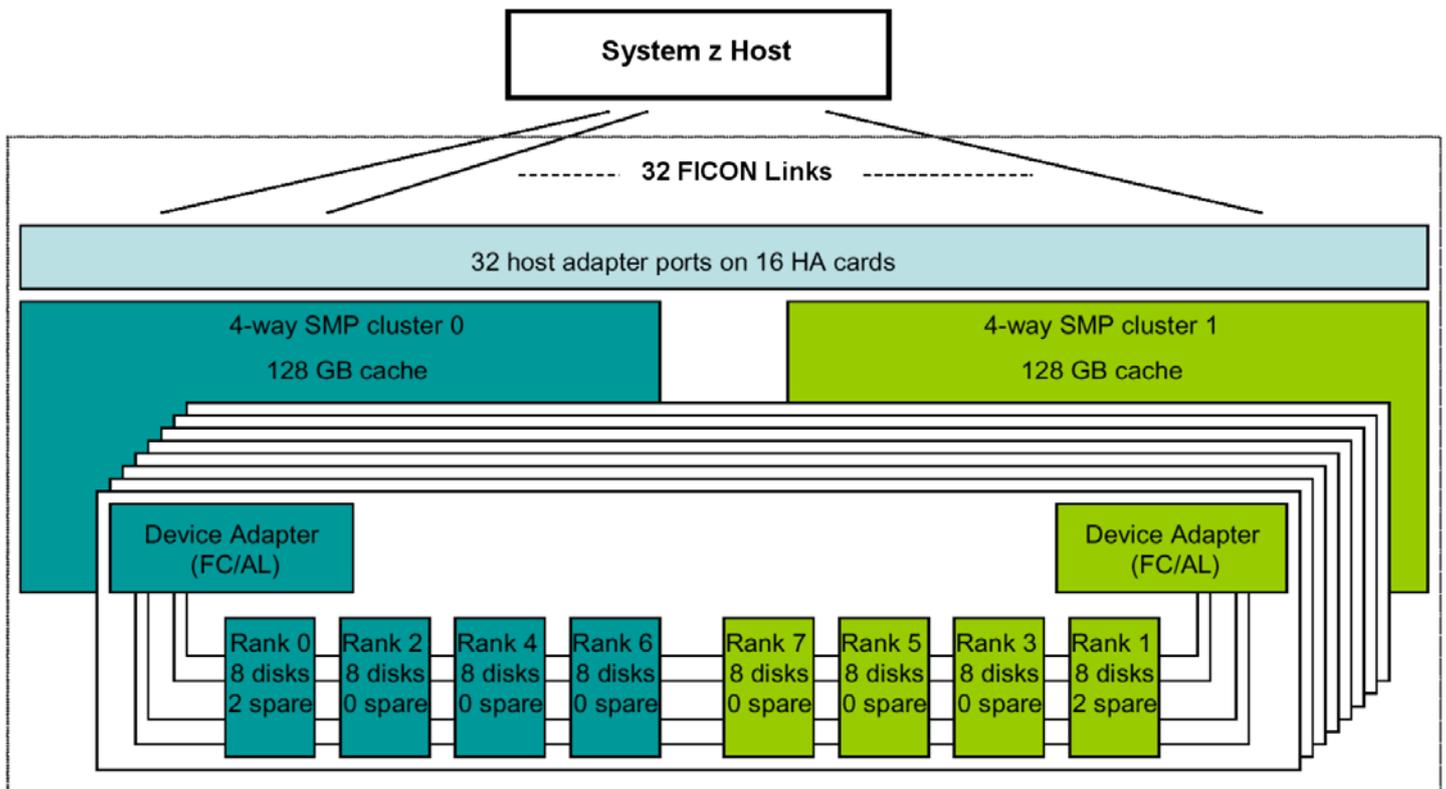


Abb. 5.4.4
Ranks eines DS8700 Enterprise Storage Servers

Abb. 5.4.4 demonstriert die Gliederung der Ranks in einer DS8700 Loop.

Dargestellt ist ein DS8700 Enterprise Storage Server mit 2 Cluster Prozessoren, je 4 x SMP, PowerPC, je 128 GByte Hauptspeicher/Cache sowie 2 x 8 Device Adaptoren . Jeweils eine FC-AL Loop mit 64 Festplatten ist an 2 Device Adaptoren angeschlossen. Jede FC-AL Loop besteht aus 2 Lese- und zwei Schreibverbindungen.

Jede FC-AL Loop besteht aus 64 aktiven Festplatten, aufgeteilt in 8 Ranks zu je 8 Platten. Die Ranks können als RAID 5, RAID 6 oder RAID 10 konfiguriert werden. Ein RAID 5 Rank würde aus 7 Platten für die Daten und einer Platte für Parity bestehen. Zusätzlich zu den 64 aktiven Platten einer FC_AL Loop existieren 2 Reserveplatten (Spare), die im Fehlerfall automatisch aktiviert (zugeschaltet) werden können. Die FC-AL Loop enthält somit $8 \times 8 = 64 + 2 = 66$ Festplatten.

Alle Festplatten sind als Hot Plug Steckplatten ausgeführt. Während des laufenden Betriebs können Platten entfernt und neu zugesteckt werden.

In dem hier gezeigten Beispiel enthält der Enterprise Storage Server 8 x 64 = 512 aktive Festplatten. Manche Modelle verfügen über 1024 aktive Platten. Wenn alle Plattenplätze mit 600 GByte Festplatten bestückt sind, ergibt dies eine gesamte Speicherkapazität von 633 TByte.

5.4.3 Nutzung durch unterschiedliche Server

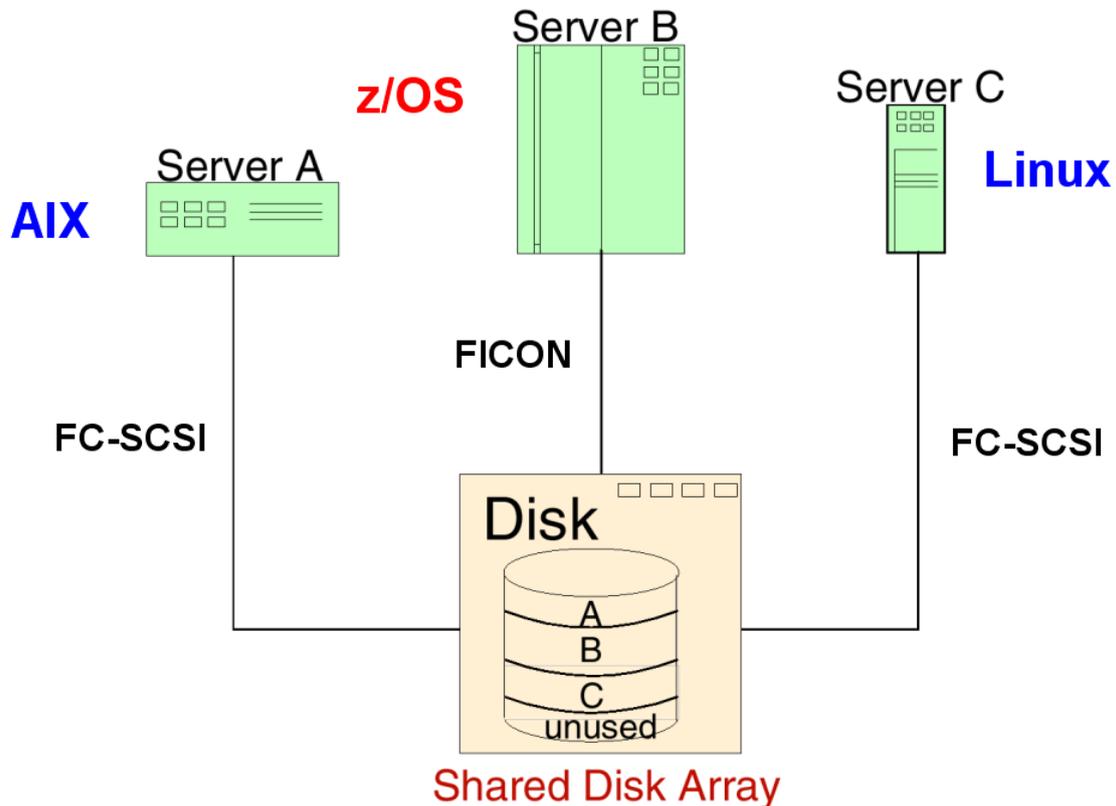


Abb. 5.4.5
Consolidated Storage

In einigen Situationen ist es attraktiv, mehrere physische Rechner mit unterschiedlichen Architekturen (z.B. z/OS, AIX, Solaris, Linux) an einen gemeinsamen Enterprise Storage Server, z. B. eine IBM DS8700 anzuschließen. Die Verbindungen zwischen den Servern und dem (den) Enterprise Storage Server(n) erfolgen über ein Fibre Channel Storage Area Network (SAN). Mainframes werden über FICON Front End Host Adapter, Unix und Linux Rechner über FC-SCSI Front End Host Adapter angeschlossen.

Unter **Hot Plug** versteht man den Austausch einer defekten Festplatte eines RAID Verbundes im laufenden Betrieb (oftmals auch als Hot Swap bezeichnet).

- Beim Hot Plug wird die Ersatzfestplatte im Betrieb manuell getauscht
- Während Hot Plug besteht weiterhin volle Datenverfügbarkeit
- Die Einbindung der Ersatzfestplatte geschieht automatisch durch das Hot Plug Programm

5.4.4 Datenarchivierung

In der Wirtschaft und öffentlichen Verwaltung hat die Datenarchivierung eine große Bedeutung. Der Gesetzgeber verlangt für manche Daten eine Archivierungsdauer von 30 Jahren. Bei Versicherungen kann die Archivierungsdauer auch noch länger sein.

Gelegentlich werden CDs und DVDs für die Archivierung eingesetzt. Das Standard Archivierungsmedium sind jedoch Magnetband Kassetten (Cartridges).

Bei manchen archivierten Daten fordert der Gesetzgeber eine Garantie, dass Daten nicht nachträglich modifiziert worden sind. Magnetbandkassetten sind deshalb in zwei Ausführungen verfügbar:

- Read/Write Kassetten können mehrfach beschrieben werden. Daten können überschrieben werden.
- WORM (Write Once, Read Many) Kassetten können nur einmal beschrieben werden.

Ein in jede Kassette eingebauter Mikroprozessor stellt die Eigenschaften einer WORM Kassette sicher. Weitere technologische Eigenschaften garantieren, dass eine betrügerische Änderung von Daten in einer WORM Kassette unmöglich ist.



Abb. 5.4.6
IBM 3592 WORM and R/W Kassetten

Die IBM 3592 Magnetbandkassette (Cartridge) hat Abmessungen von 24.5 mm H x 109 mm W x 125 mm D und verwenden ein 1/2 Zoll breites Magnetband mit einer Länge von 825 Meter. Die Speicherkapazität beträgt bis zu 4 TByte. Mainframes benutzen Datenkomprimierung für die Magnetbandspeicherung, was die Speicherkapazität zusätzlich um einen Faktor 2 – 3 erhöht.

IBM garantiert eine Lebensdauer der Kassetten von 10 Jahren. 30 Jahre sind wahrscheinlich. Die meisten Unternehmen kopieren archivierte Magnetband-Daten viel häufiger um, um auf der sicheren Seite zu sein (z.B. alle 5 Jahre).



Abb. 5.4.7
Magnetbandeinheit für 3592 WORM and R/W Kassetten

Der IBM System TS1130 Tape Drive (Magnetbandeinheit) liest und schreibt 3592 Kassetten mit einer Datenrate von 160 MByte/s.

5.4.5 Magnetband Library



Abb. 5.4.8
Ansicht der IBM 3494 Tape Library

Früher hat ein menschlicher Operator die Magnetbandkassetten in eine Magnetbandeinheit eingelegt und wieder entfernt. Heute verwendet man hier Magnetbandroboter, auch als Tape Libraries bezeichnet.

Die 3494 Enterprise Tape Library unterstützt ein Maximum von 6,240 Kassetten für eine Speicherkapazität von mehreren PetaBytes (PByte). Sie kann bis zu 132 Tape Drives in einer System z Umgebung unterstützen.

Der 3494 Cartridge Accessor mit dem dualen Gripper holt Kassetten aus einem Regal und legt sie in einen der Tape Drives ein.



Abb. 5.4.9
Dual Gripper 3494 Cartridge Accessor

Es können bis zu 265 Cartridge Exchanges/Stunde mit einem einzige Greifarm (Gripper), und bis zu 610 Exchanges/Stunde mit einem dualen Gripper und dualen aktiven Accessors durchgeführt werden.

5.4.6 Virtual-Tape-Server

Ein Virtual-Tape-Server (VTS, anderer Namen Virtual Tape Library oder Virtual I/O, VIO) ist ein Speicher auf Basis eines Festplatten Arrays.

Ein VTS stellt sich für angeschlossene Computer wie eine Tape-Library dar. Die Anzahl der Libraries und darin enthaltener Slots und Bandlaufwerke sind dabei frei konfigurierbar. Virtuelle Bänder können zumeist von dem VTS direkt auf „echte“ Bänder geschrieben werden, ohne dass die Backup-Software oder ein Server an diesem Vorgang beteiligt ist.

Magnetbandkassetten tolerieren nur eine begrenzte Anzahl von Zugriffen. Wenn überhaupt auf die Daten einer Magnetbandkassette zugegriffen wird, sind zahlreiche Zugriffe auf einzelne physische Datensätze oder Slots häufig die Folge. Eine VTL arbeitet so etwa wie ein mittels einer Festplatte implementierter Cache für eine Magnetbandkassette.

Weiterhin ist es möglich, Backup-To-Disk-Konzepte in bestehenden Datensicherungsumgebungen, die in der Regel auf Bandlaufwerken basieren, einzubinden. Ein Beispiel sind temporär genutzte Data Sets in der Stapelverarbeitung. JES nimmt hierfür in der Regel Magnetbandlaufwerke an.

5.4.7 Drucker Ausgabe

Mainframe Installationen haben sehr unterschiedliche Anforderungen bezüglich Printer I/O.

Die Annahme, dass mit wachsender Bildschirmausgabe und wachsender digitaler Speicherung von Dokumenten der Papierverbrauch rückläufig sein würde, hat sich fast nirgendwo bewahrheitet. Im Allgemeinen kann man davon ausgehen, dass der Papierverbrauch in Unternehmen und staatlichen Organisation Jahr für Jahr nach wie vor steigt.

Beispiele sind Konto Auszüge und Überweisungsbelege im Bankenbereich, Mitteilungen der staatlichen Rentenversicherung oder Abrechnungen von Krankenkassen. Derartige Dokumente werden teilweise auf dem Postweg in Briefumschlägen (Kuvert) versandt.

Drucker können wie Plattenspeicher über eine Printer-Control Unit und FICON Channel Kabel mit einer I/O Card eines Mainframe Systems verbunden werden. Mehrere Hersteller liefern derartige Produkte. Unter z/OS existiert ein generischer I/O Driver für die Drucker-Ansteuerung. Dieser überträgt ein entsprechendes Kanalprogramm an die Printer Control Unit.

Alternativ kann Print Output als Datei über einen Netzanschluss an einen getrennten Druck Server übergeben werden. Weitergehende Formatierungen erfolgen dezentral und unabhängig von z/OS.

Es existieren weitere exotische Ein/Ausgabegeräte. Ein Beispiel sind zentrale Check Lese Geräte für die automatische Verarbeitung von großen Mengen von Überweisungen und Bank-Schecks.



Abb. 5.4.10
Fiducia AG, Karlsruhe, Druck- und Kuvertierzentrum

Die Fiducia AG in Karlsruhe ist einer der großen deutschen IT Dienstleister. Ihr Rechenzentrum mit mehreren Mainframes bedient vor allen genossenschaftliche Einrichtungen wie Volksbanken und Raiffeisenbanken.

Das Druck- und Kuvertierzentrum der Fiducia AG in Karlsruhe ist über Control Units an die dortigen Mainframes angeschlossen. Dies sind einige der Leistungsmerkmale:

- **19 Mainframe Drucksysteme mit einer Gesamtleistung von 9 200 Seiten pro Minute.**
- **14 Kuvertierstraßen mit einer Gesamtleistung von 40.500 Kuvertierungen pro Stunde**
- **378 Millionen DIN-A4-Seiten pro Jahr**
- **140 Millionen kuvertierte Sendungen pro Jahr, davon 90 Millionen kuvertierte Kontoauszugssendungen**
- **1.500 Kunden für Druckdienstleistungen**

5.5 Weiterführende Information

Life Demo eines Festplattenspeicher Zugriffsarms

<http://www.youtube.com/watch?v=L0nbo1VOF4M>

Ein sehenswertes Video demonstriert ein altes Rechenzentrum in den 90er Jahren

<http://www.youtube.com/watch?v=Cwj6pfhWBps&feature=related>

In 1992 IBM released the 3495 Tape Library using a bright yellow robot which this video shows in great detail. This is a great look back to technology in the early 1990s.

<http://www.youtube.com/watch?v=GwMn7YpF8r8&feature=fvwrel>

Eine ähnliche Demo mit der Ultrium Tape Unit wird gezeigt in

http://www.youtube.com/watch?v=INa_D1HIjww&feature=related

Youtube Video zu den Vorteilen eines Storage Area Network:

<http://www.youtube.com/watch?v=KjggHyIfGL4>

Youtube Video zum Thema IBM Impact Printer circa 1990

<http://www.youtube.com/watch?v=kEWcvmUluyE>

IBM war Marktführer mit mechanischen High Speed Druckern der Serien 1403 und 3203 von 1959 bis in die 1990's, siehe auch

<http://www.youtube.com/watch?v=Cwj6pfhWBps>