UNIVERSITÄT LEIPZIG

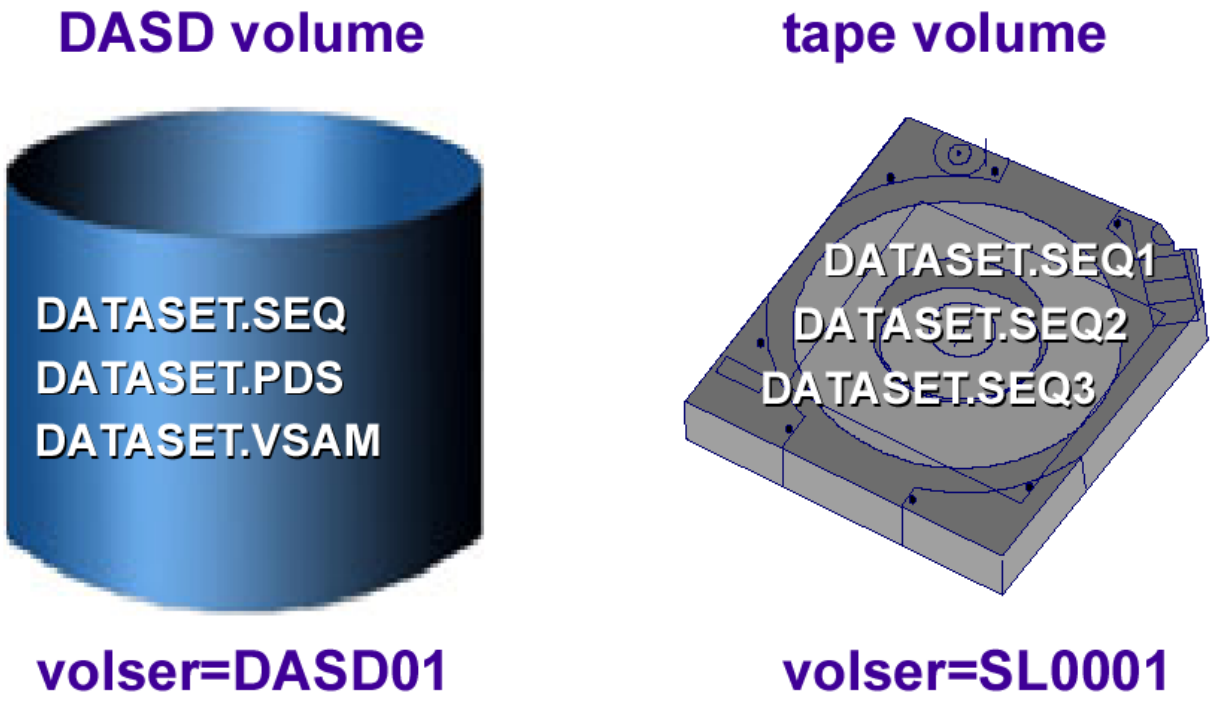
Enterprise Computing
Einführung in das Betriebssystem z/OS

Prof. Dr. Martin Bogdan
Prof. Dr.-Ing. Wilhelm G. Spruth

WS2012/2013

Input/Output Teil 3

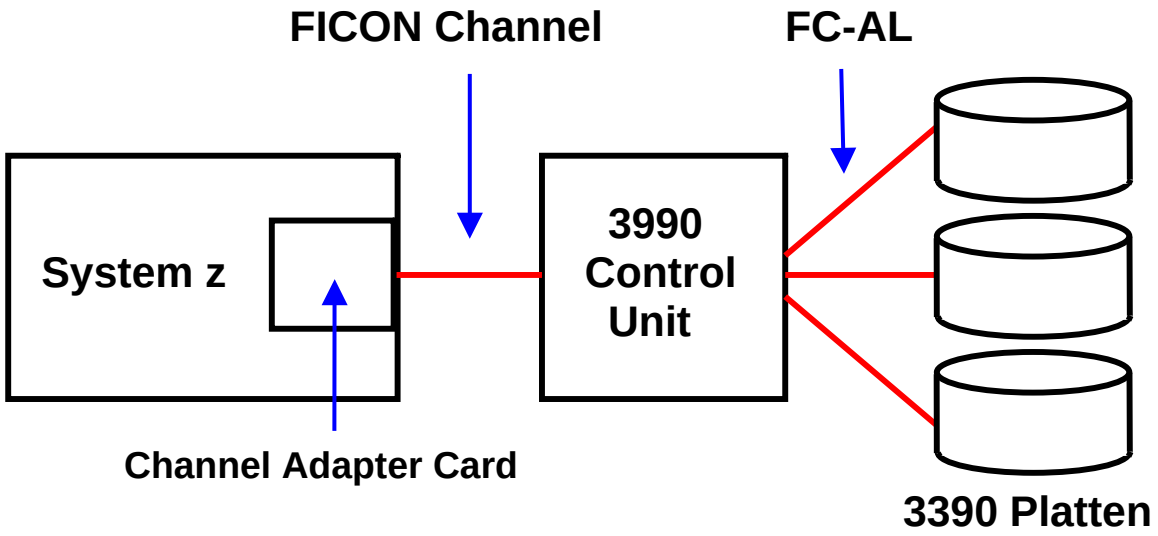
Mainframe I/O



Volume

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 Volumes 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.



Channel und Control Unit

deutsche Begriffe: Kanal und Steuereinheit bzw. Plattenspeicher Steuereinheit

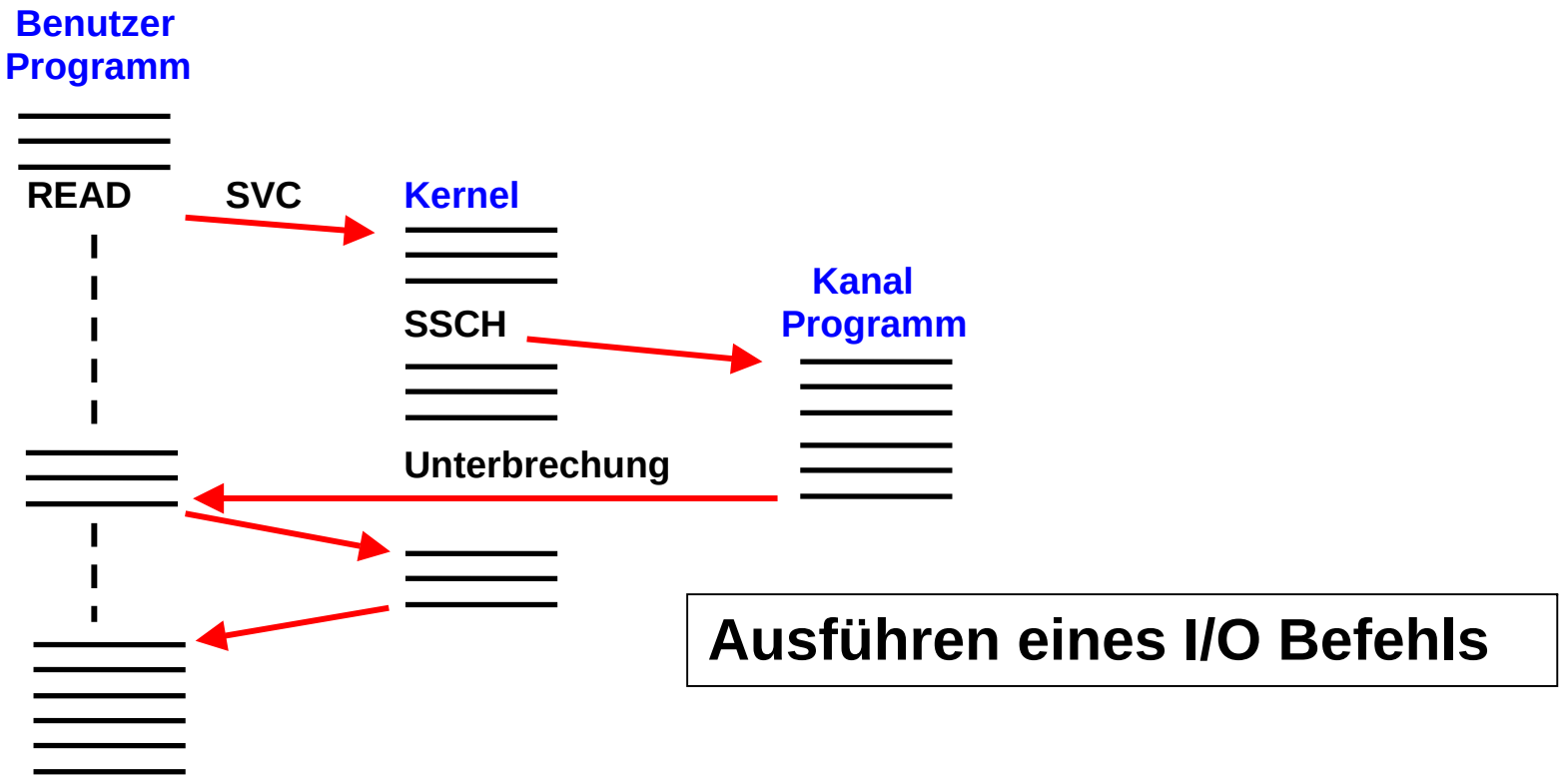
Ein Kanal (Channel) ist ein Verbindungskabel zwischen einem Rechner und einer Plattenspeicher-Steuereinheit. Ein Channel Adapter ist eine Steckkarte im I/O Cage eines Rechners, die Steckkontakte für Kanalkabel enthält. Kanäle werden heute als FICON Fibre Channel Verbindungen implementiert.

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 im I/O Drawer eines Rechners, die zum Anschluss von Ethernet Verbindungen dient.

Ausführen einer I/O Operation

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.

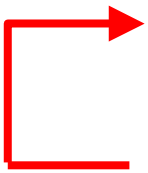


Ausführen eines I/O Befehls

Dargestellt ist 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.



SEEK	**Richtige Spur (Zylinder) finden**
SEARCH	**Richtigen Datensatz auf der Spur finden**
TIC	**(Transfer in Channel) Nochmals versuchen**
READ oder **WRITE**	**Daten übertragen**

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.

3390 Plattenspeicher

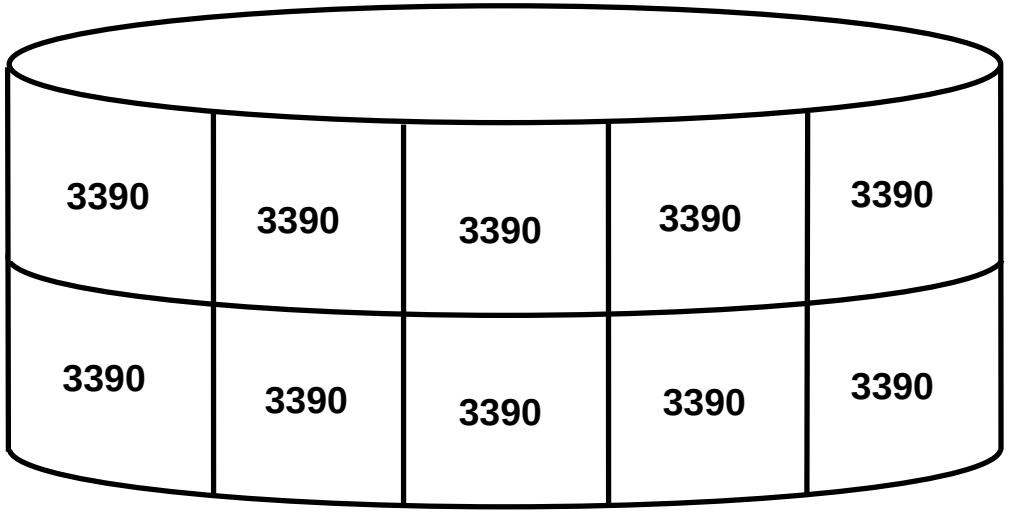
Control Units entlasten die CPUs, indem der I/O Driver Code außerhalb der CPUs in den Control Units ausgeführt wird. Zum Betriebssystem gehören nur Skelette für einige wenige I/O Driver Routinen, z.B. je eine generische Routine für Plattenspeicher, Magnetbänder, Drucker und Netzanschlüsse. Diese Routinen 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.

Beispiel Plattenspeicher

Um dies zu erreichen, arbeitet der OS Kernel und die Control Unit mit der Illusion eines generischen Plattenspeichers, als „3390 Plattenspeicher“ bezeichnet. Das Betriebssystem kennt nur 3390 Plattenspeicher mit deren Struktur bezüglich Anzahl Zylinder, Anzahl Spuren, Bytes pro Spur usw. 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.

Parameter wie Spuren pro Zylinder und Bytes pro Spur sind festgelegt, und sind bei allen 3390 Plattenspeichern identisch. Die Anzahl der Zylinder pro Plattenspeicher kann als Parameter definiert werden, woraus sich unterschiedliche Speicherkapazitäten ergeben.



Beispiel:

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

Physical and Logical Volume

Emulierte 3390 Plattenspeicher werden als „Logical Volumes“ (LV) bezeichnet. Sie existieren als zwei unterschiedliche standardirierte 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.

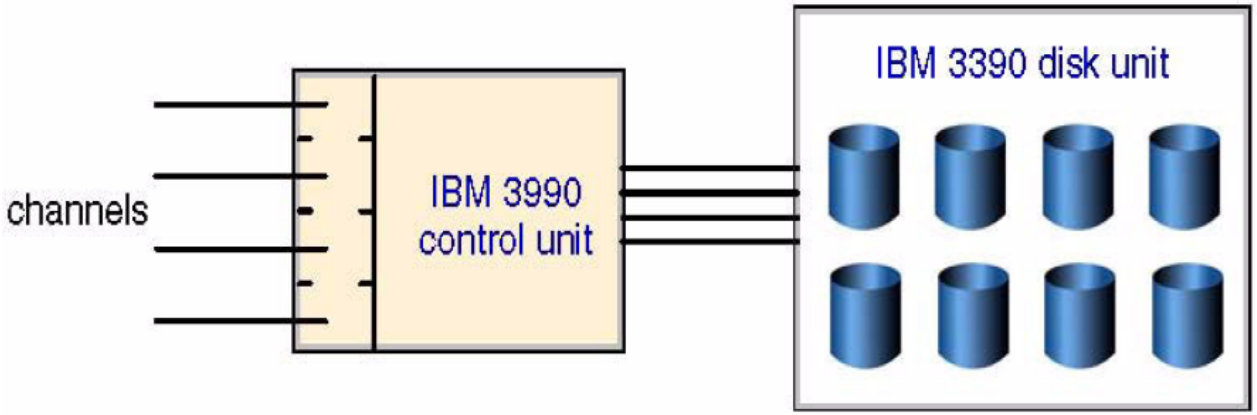
	3390-3	3390-9
Zylinder pro Plattenstapel	3 339	10 017
Spuren pro Zylinder	15	15
Bytes pro Spur	56 664	56 664
Bytes pro Zylinder	849 960	849 960
MByte pro Plattenstapel	2 838	8 514
4096 Byte Blocks pro Spur	12	12

DASD Speicherkapazität der Modelle 3390

Die hier wiedergegebenen Daten sind für das Allocate von Data Sets interessant. Beispielsweise sollte der Parameter BLKSIZE (siehe Tutorial 1a) so gewählt werden, dass es bei 56 664 Bytes/Spur möglichst wenig Verschnitt gibt. Bei der Optimierung hilft ein BLKSIZE Calculator, siehe http://webspace.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) ermöglichst 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 %.

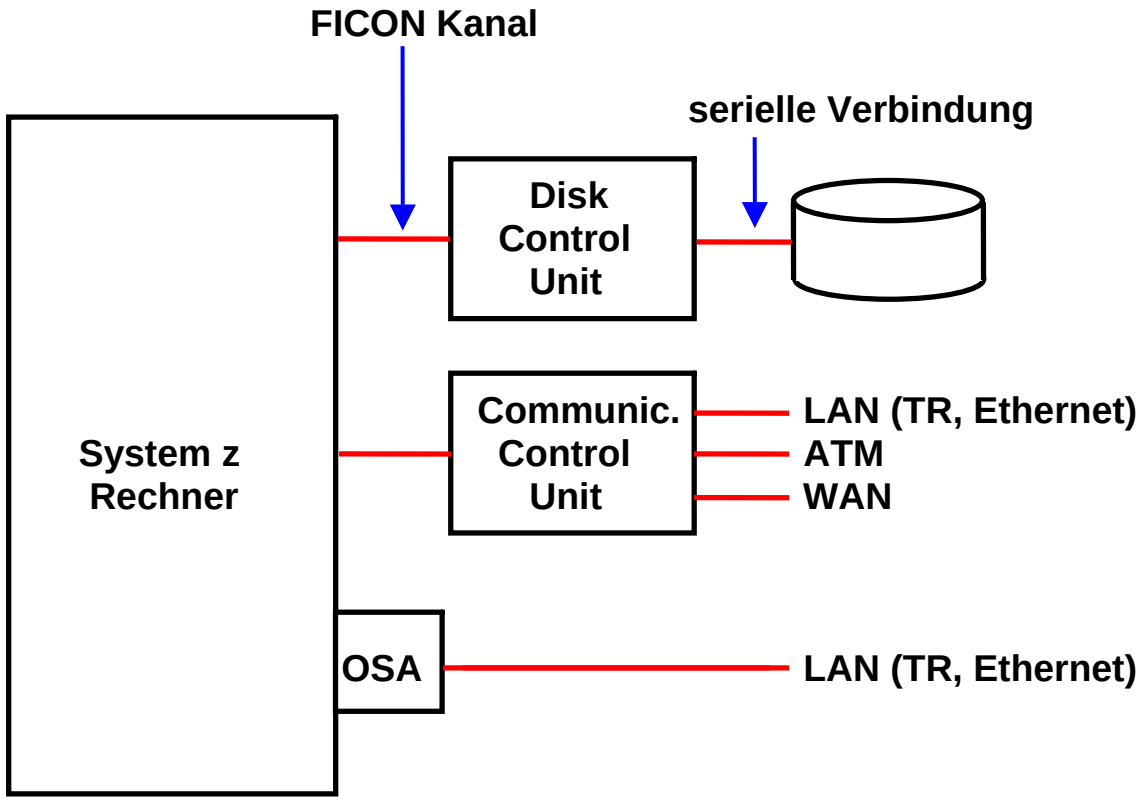
Eine populäre Wahl ist, 349 logische Record zu je 80 Bytes in einen Block von 27920 Bytes zu packen.



Ursprüngliche IBM 3390 Plattenspeicher Implementierung

Die in den 90er Jahren von IBM vertriebenen 3390 Plattenspeicher wurden über eine 3990 Control Unit an einen S/390 Rechner angeschlossen. Die Parameter dieser physischen Plattenspeicher wurden für die Definition der logischen 3390 Plattenspeicher übernommen. 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 Multiprozessoern 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.



Einige Steuereinheiten können in den System z Rechner integriert werden. Das wichtigste Beispiel ist der OSA Adapter für den Anschluß von Local Area Networks (LAN), z.B. Ethernet.

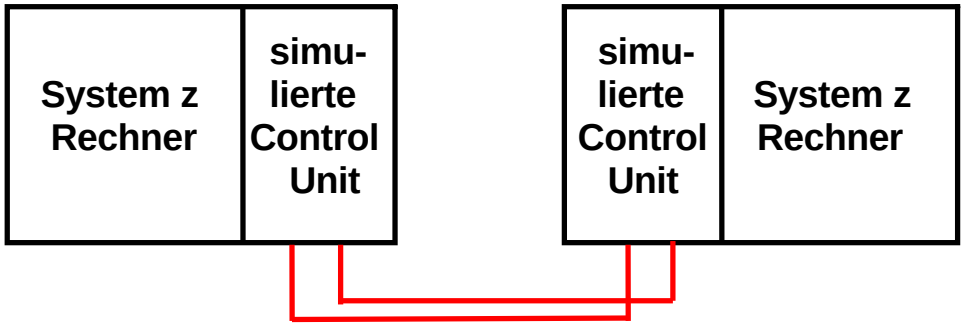
System z I/O-Konfiguration

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.

Channel- to Channel Verbindung (CTC)
Cross-System Coupling Facility (XCF)



Die Channel- to Channel 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.

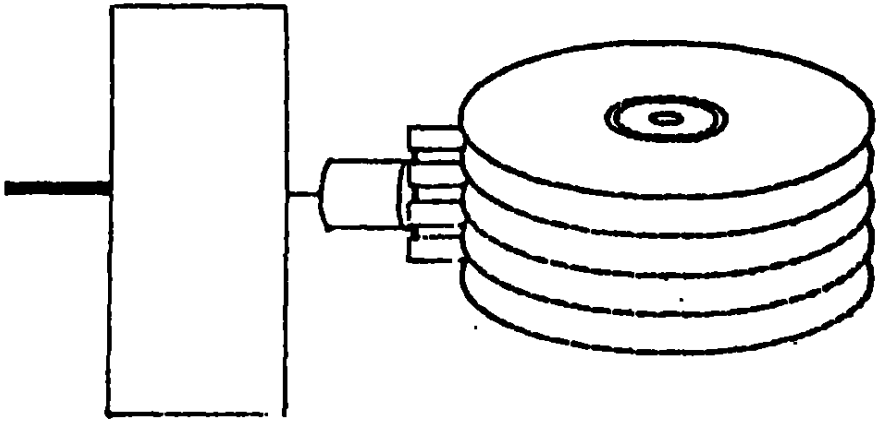
Die „Cross-System Coupling Facility“ (XCF) ist eine Komponente des z/OS Betriebssystems. Sie verwendet das CTC Protokoll und die CTC Hardware. Sie stellt die Coupling Services bereit, mit denen z/OS Systeme innerhalb eines Clusters (Sysplex) miteinander kommunizieren.

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 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.

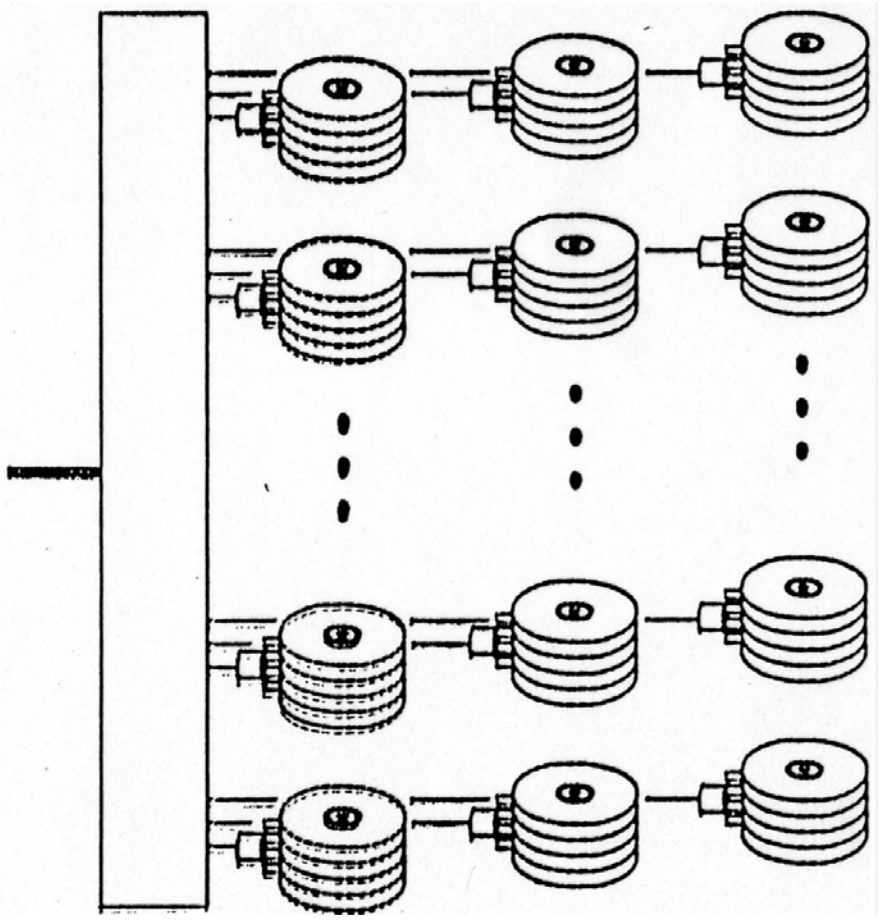
Die weitere Entwicklung des RAID-Konzepts führte zunehmend zum Einsatz in Serveranwendungen, die den erhöhten Datendurchsatz und die Ausfallsicherheit nutzen. Der Aspekt der Kostenersparnis wurde dabei aufgegeben. Heute existiert meistens die Möglichkeit, in einem solchen System einzelne Festplatten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



Single large Disk

Der ursprüngliche Vorschlag von Patterson, Gibson und Katz bestand darin, eine einzige 14 Zoll Plattedurch eine Gruppe (Array) kostengünstiger, aber weniger zuverlässiger Seagate 5,25-Zoll Platten zu ersetzen.

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



Array of small Disks

RAID-Level

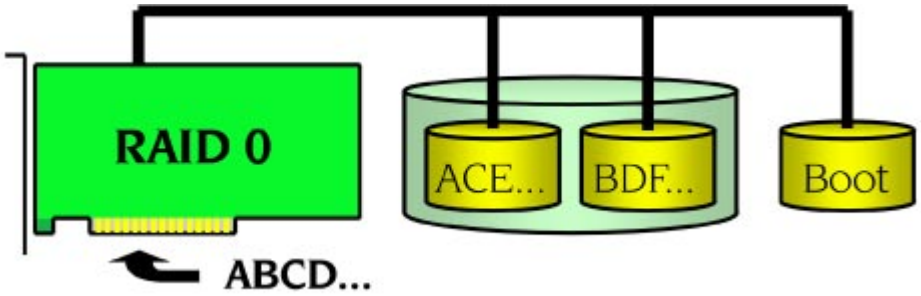
Der Betrieb eines RAID-Systems setzt mindestens zwei Festplatten voraus. Die Festplatten werden gemeinsam betrieben und bilden einen Verbund, der leistungsfähiger ist als die einzelnen Festplatten. Mit RAID-Systemen kann man folgende Vorteile erreichen:

- **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. Sie werden unten beschrieben.

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

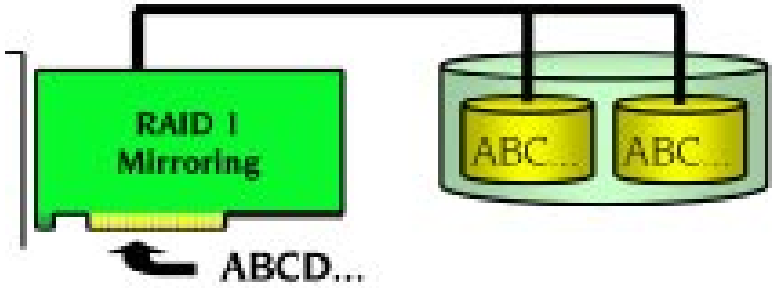
RAID 0



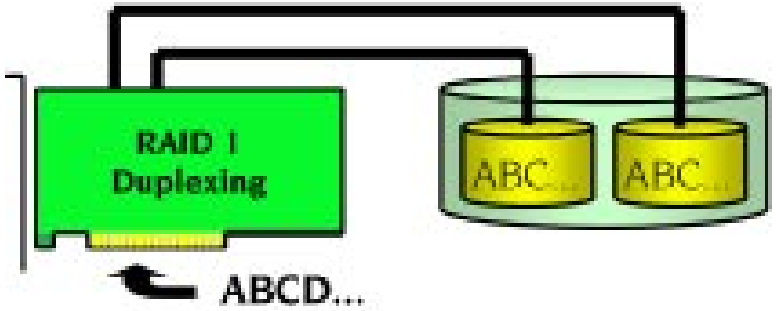
RAID 0 bietet gesteigerte Transferraten, indem die beteiligten Festplatten in zusammenhängende Blöcke gleicher Größe (z.B. 16KB) in Streifen (eng. stripes) aufgeteilt werden, wobei jeder Streifen eines Datenblocks auf einer separaten Festplatte gespeichert wird. Diese Blöcke werden quasi im Reißverschlussverfahren 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.

RAID 1



Disk Mirroring

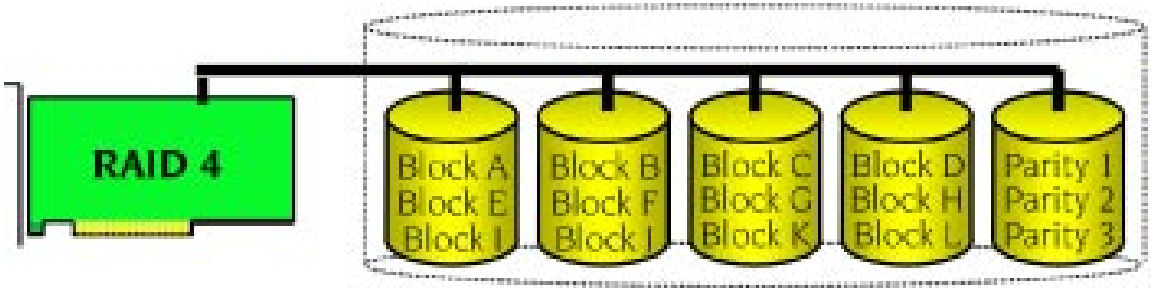


Disk Duplexing

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.

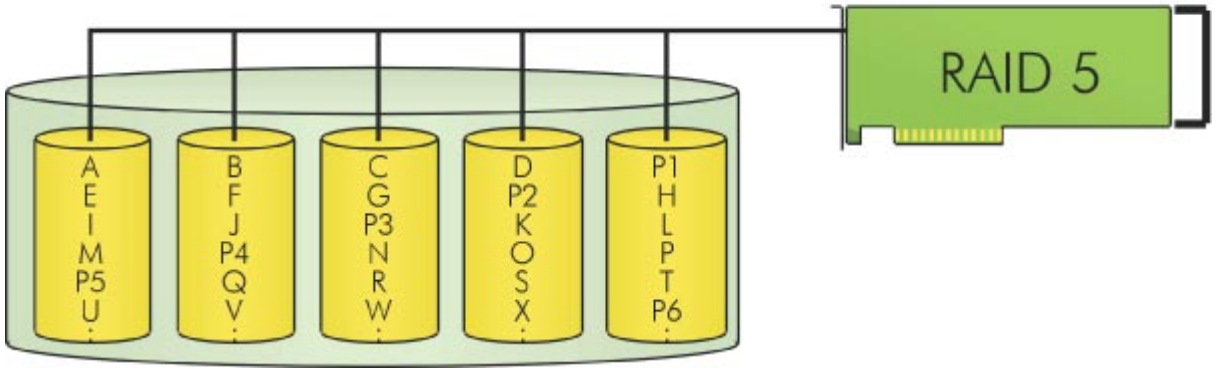
RAID 3 und 4



Wie bei RAID 0 werden die Daten auf den Festplatten verteilt. 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 RAID 3 oder 4 Verband mit 5 Festplatten stehen 80 Prozent der Gesamtkapazität als Nutzkapazität zur Verfügung.

Beim Schreiben kleiner Datenblöcke wird das Paritätsylaufwerk 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.

RAID 5



Anders als bei RAID 4 werden die Paritätsdaten P1 .. P6 auf allen Festplatten im Verband gleichmäßig verteilt. Dies garantiert bei allen Zugriffen eine optimale Auslastung der Laufwerke. Selbst bei zufallsbedingten (random) Zugriffen, wie sie für ein multitasking multiuser Bestriebssystem 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.

RAID 6

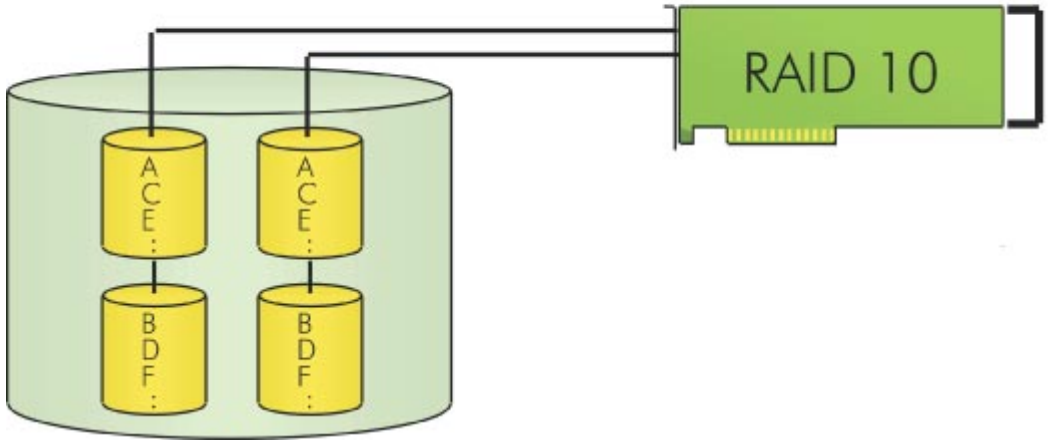
Wenn bei RAIR 5 eine Festplatte ausfällt, wird der laufende Betrieb dadurch zunächst nicht beeinträchtigt. Die fehlerhafte Festplatte muss baldigst durch eine neue Festplatte ersetzt werden. Nachdem dies geschehen ist, kann ein Unterprogramm des Enterprise Storage Servers den Inhalt der neue Platte aus 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, verkraftet 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 dier 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 erkauft.

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.

RAID 10



Aus einer Kombination von RAID 0 (Performance) und RAID 1 (Datensicherheit) ist der RAID Level 10 entstanden. Raid 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.

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 Lokationen 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.