UNIVERSITÄT LEIPZIG

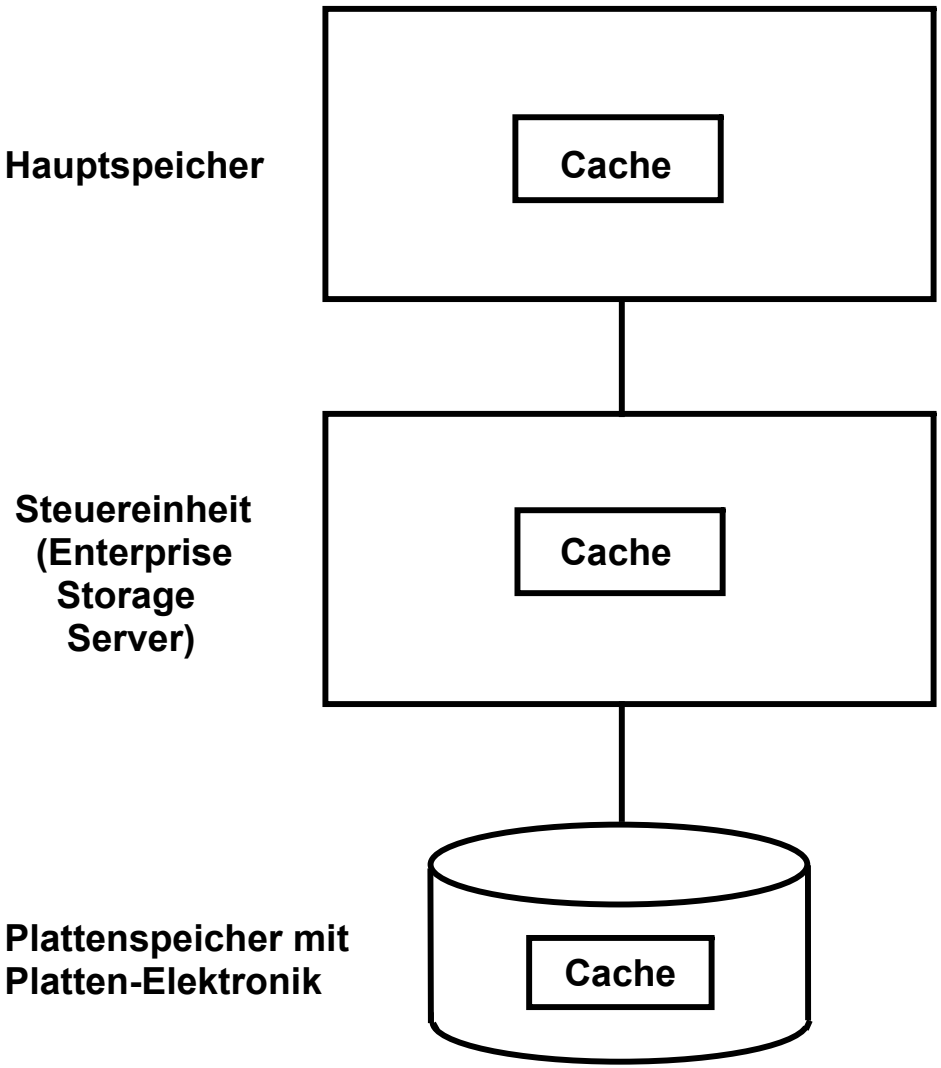
Enterprise Computing
Einführung in das Betriebssystem z/OS

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

WS2012/2013

Input/Output Teil 4

Enterprise Storage Server



Plattenspeicher-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.

Disk 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.

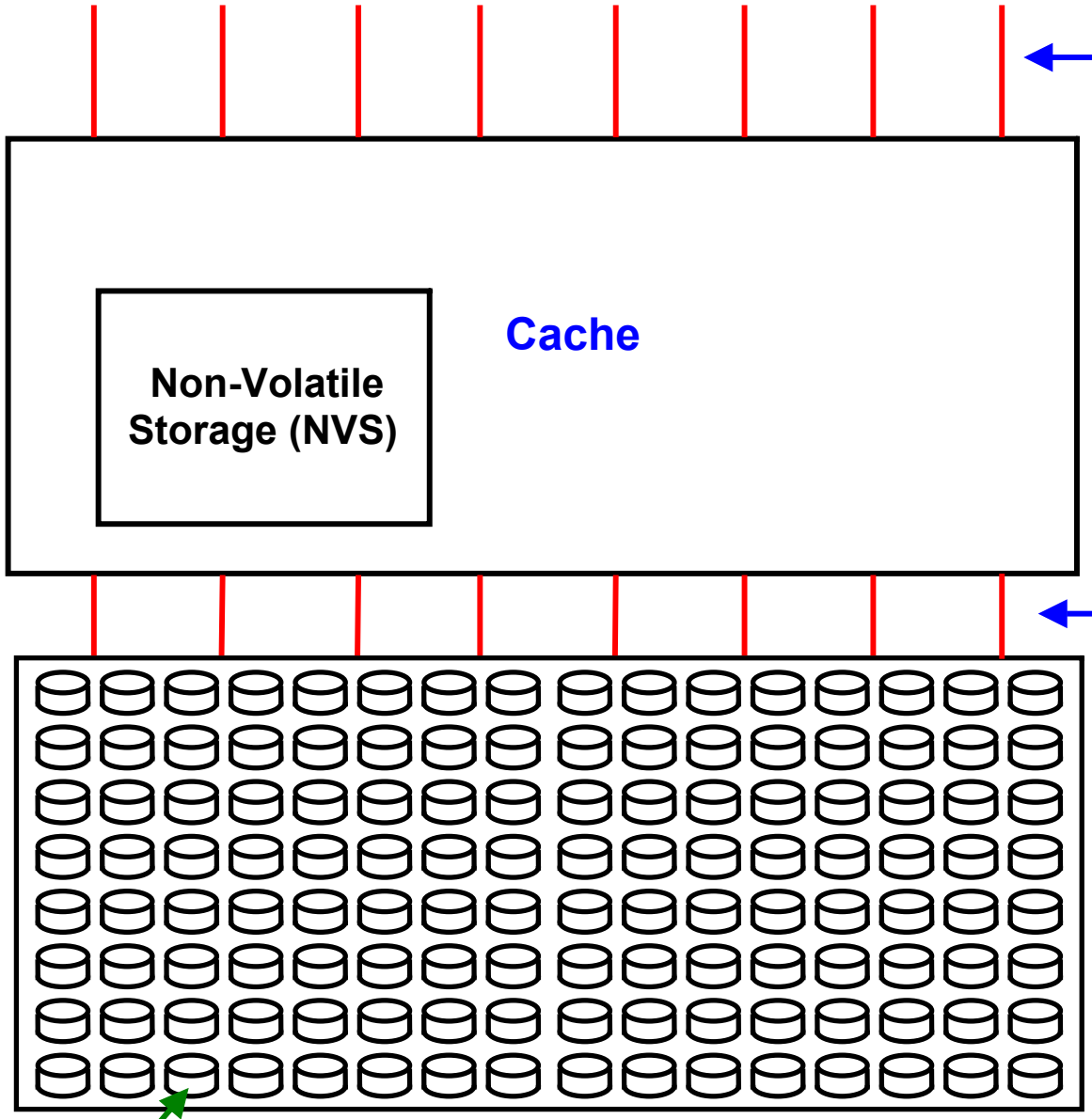
In den nächsten Schritten wurden die Caches grösser, und dann wurden mit dem Aufkommen von preisgünstig zu produzierenden 3 ½ Platten die großen 14 Zoll Platten ersetzt. Jetzt wurden auch die Datenstrukturen und Plattenformate wie Key-Count-Data als real existierende Formate aufgelöst und durch die Control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

Lese-Zugriffe können mit Hilfe eines Caches deutlich beschleunigt werden. Wünschenswert ist es, auch Schreibvorgänge 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 entgültige Datentransfer zum Plattenspeicher erfolgt dann asynchron und unbemerkt vom Betriebssystem.

Front end “Host Adapter” Anschlüsse für den Anschluss an FICON Kanäle



Aufbau eines Enterprise Storage Servers

Der Enterprise Storage Server übernimmt die Funktion meherer Control Units.

Normale 3 ½ Zoll (oder 2 ½ Zoll) Festplatten, jede mit einem eigenen Festplatten Cache

Enterprise Storage Server

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.**

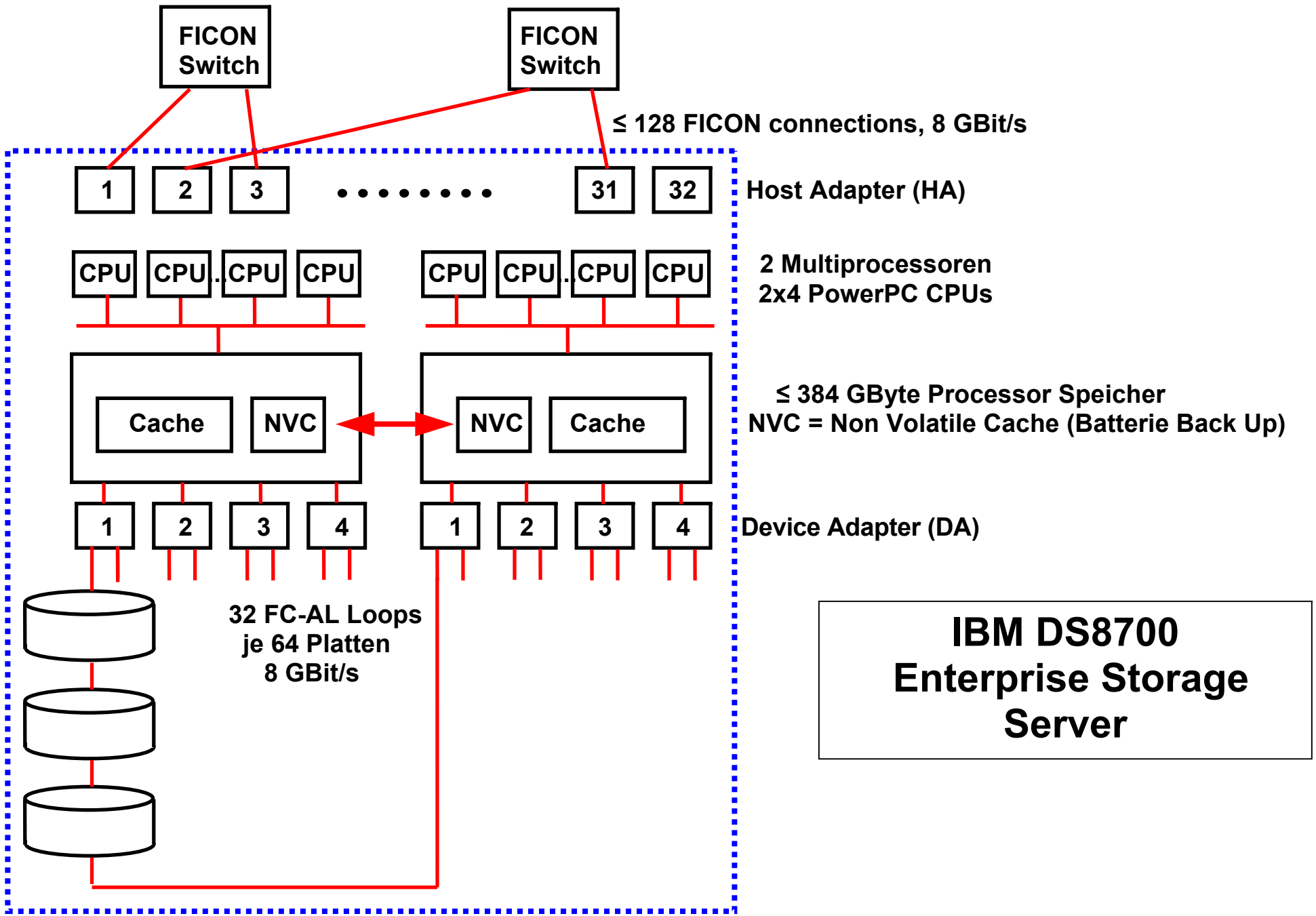
2. **Ein oder (aus Zuverlässigkeitsgründen) zwei Multiprozessoren plus Cache, welcher aus zwei Teilen besteht: Einem Cache für Daten, die gelesen werden können, und einem Cache für Daten, die geschrieben werden sollen. Letzterer heißt Non-Volatile Storage (NVS) und bezeichnet damit einen Cache, der extra 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.

Die bedeutensten 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.


FICON Switch
FICON Switch
≤ 128 FICON connections, 8 GBit/s
1
2
3
31
32
Host Adapter (HA)
CPU
CPU
CPU
CPU
CPU
CPU
CPU
CPU
2 Multiprocessoren
2x4 PowerPC CPUs
Cache
NVC
NVC
Cache
≤ 384 GByte Processor Speicher
NVC = Non Volatile Cache (Batterie Back Up)
1
2
3
4
1
2
3
4
Device Adapter (DA)
32 FC-AL Loops
je 64 Platten
8 GBit/s
IBM DS8700
Enterprise Storage
Server

IBM DS8700 Enterprise Storage Server

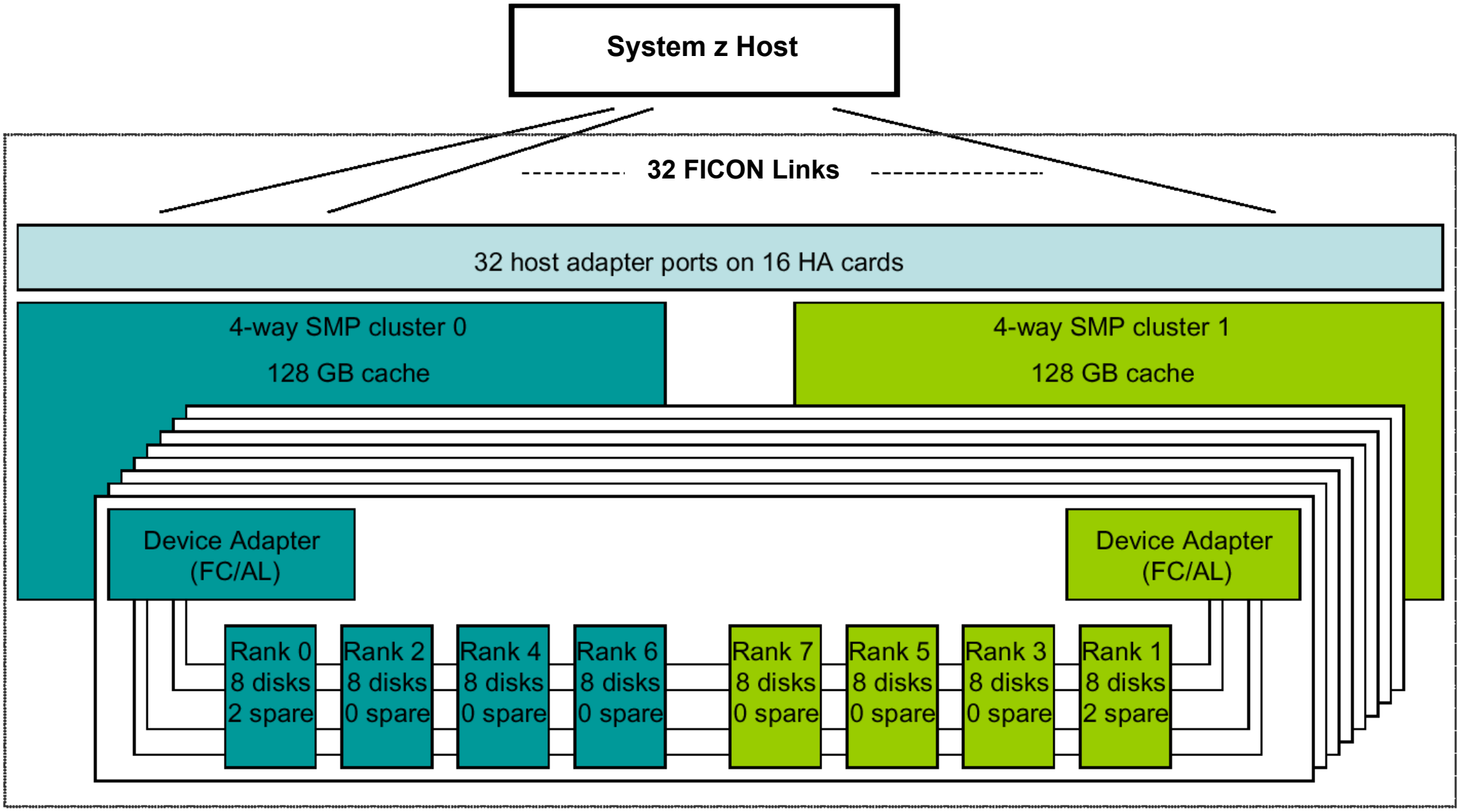
Zur Verbindung mit dem Host besitzt der DS8700 Enterprise Storage Server bis zu 32 Host Adapter (HA), die entweder je 4 FC-SCSI- Verbindungen oder je 4 FICON bzw. Fibre Channel Verbindungen nach aussen implementieren.

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 Enerigie als rotierende Festplatten.



IBM DS8700 Enterprise Storage Server Beispiel

IBM DS8700 Enterprise Storage Server

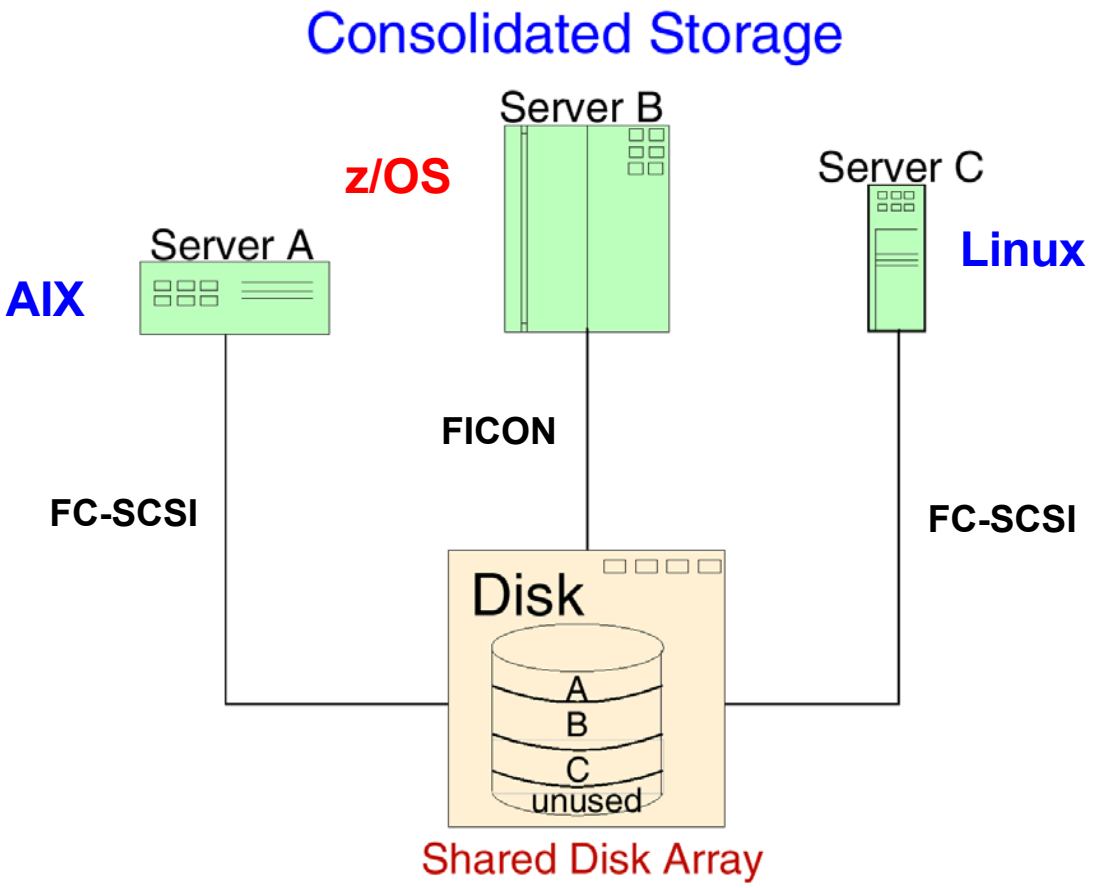
Die obige Abbildung demonstriert die Gliederung der Ranks in einer DS8700 Loop.

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 (Spares), die im Fehlerfall automatisch aktiviert (zugeschaltet) werden können. Die FC-AL Loop enthält somit 8 x 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.



Enterprise Storage Server

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 Storage Area Network (SAN). Mainframes werden über FICON Front End Host Adapter, Unix und Linux Rechner über FC-SCSI Front End Host Adapter angeschlossen.

Hot Plug

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**

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 Magnetbandkassetten (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) Kasseten können nur einmal beschrieben werden.**

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

Früher hat ein menschlicher Operator die Magnetbandkassetten in eine Magnetbandeinheit eingelegt und wieder entfernt. Heute verwendet man hier Magnetbandroboter, auch als Tape Libraries bezeichnet. Eine Tape Library in einer Mainframe Installation kann viele Tausend Kassetten verwalten.

IBM 3592 WORM and R/W Kasseten

Die IBM 3592 Magnetbandkassette (Cartridge) hat Abmessungen von 24.5 mm H x 109 mm W x 125 mm D und verwenden ein ½ 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 garaniert eine Lebensdauer der Kasseten 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).

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

IBM 3494 Tape Library

Die 3494 Enterprise Tape Library unterstützt ein Maximum von 6,240 Kassetten für eine Speicherkapazität von mehreren PetaBytes (PByte).

Die 3494 Tape Library unterstützt bis zu 132 Tape Drives in einer System z Umgebung.

Der 3494 Cartridge Accessor mit dem dualen Gripper holt Kassetten aus einem Regal und legt sie in eine Magnetband-einheit ein.

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.

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 Channel Adapter 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 Formattierungen 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.

Fiducia AG, Karlsruhe
Druck- und Kuvertierzentrum

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 Kontoauszugs-sendungen

1.500 Kunden für Druck