IBM Academic Initiative

Sistemi Storage ed Architettura di IO in ambiente zSeries

Stefano Ricci
IBM Italia
Systems and Technology Group

Novembre 2007

© 2007 IBM Corporation

IBM Systems
Simplify your IT.

© 2007 IBM Corporation

Agenda

Prima lezione

- L'architettura di I/O in ambiente zSeries
- Sistemi Storage a disco

Seconda lezione

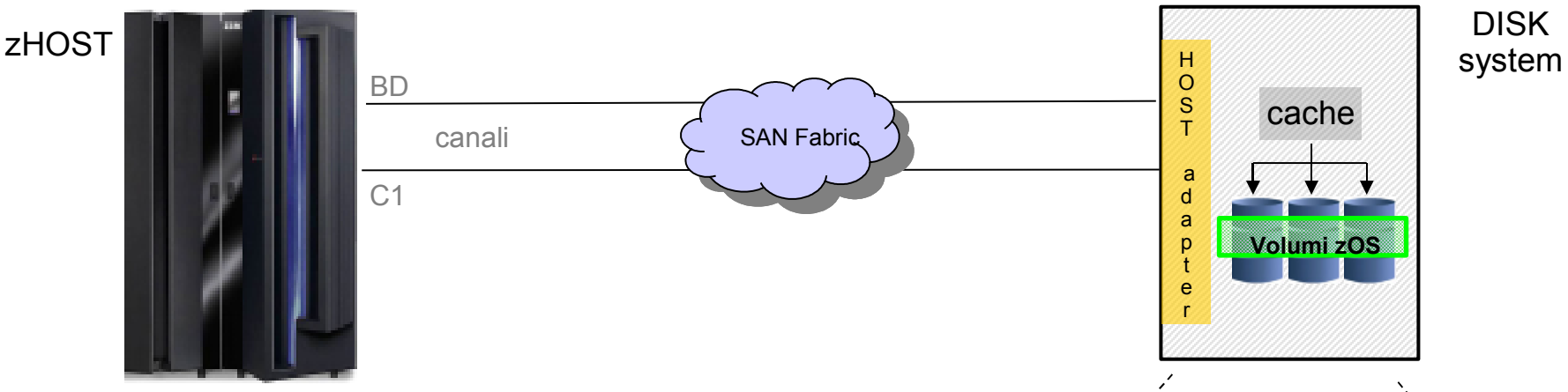
- Sistemi Storage a nastro
- Le funzioni di copia dei dati
- Connettivita' e trasporto dei dati
- Cenni sulle prestazioni dei sistemi a disco

© 2007 IBM Corporation

- L’architettura di I/O in ambiente zSeries
- *Sistemi Storage a disco*
- *Sistemi Storage a nastro*
- *Le funzioni di copia dei dati*
- *Connettivita’ e trasporto dei dati*
- *Cenni sulle prestazioni dei sistemi a disco*

© 2007 IBM Corporation

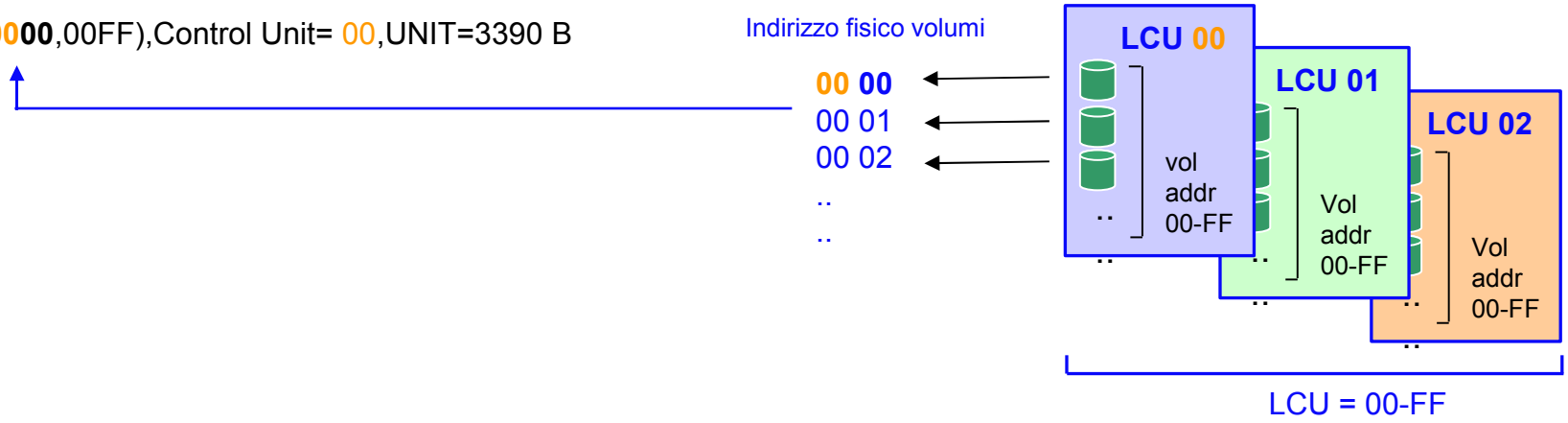
Indirizzamento di un volume logico in ambiente zSeries



Generazione dello IO allo zOS (IOCP)

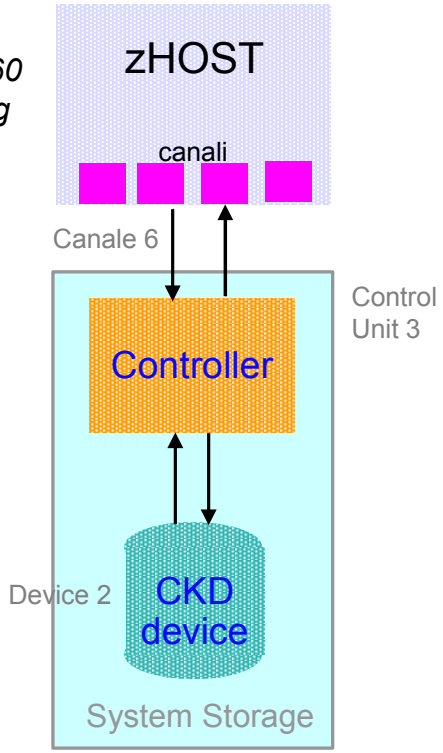
Control Unit =00, Path =(BD,C1), UNIT=DISK

IO device=(**0000**,00FF),Control Unit= 00,UNIT=3390 B

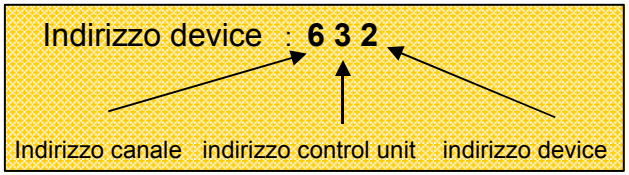


Fondamenti dell'architettura di IO in ambiente zSeries

System/360 addressing



Architettura	Central storage addressing	IO device addressing (per LPAR)	Canali (per LPAR)	Funzioni introdotte
System 360 (1960)	**24 bit**	**16 bit (4k)**	**256**	▪**Concetto di LCU (piu' paths al device)**
System 370	**24 bit**	**16 bit (65k)**	**256**	▪**IO multiple sullo stesso canale** ▪**Selezione del path e riconnessione dinamica (DPS e DPR)**
System 370 XA ed ESA	**31 bit**	**16 bit (65k)**	**256**	▪**Canali condivisi tra i processori** ▪**System Managed Storage**
System 390	**31 bit**	**16 bit (65k)**	**256**	▪**Nuova tecnologia dii canale (ESCON e FICON)**
z/Architecture (2000)	**64 bit**	**2*16 bit (65k+65k)**	**256**	▪**FICON piu' veloci (Express 2 e 4)** ▪**Uso del secondo set di indirizzi per gli alias**



© 2007 IBM Corporation

Architettura CKD - DASD

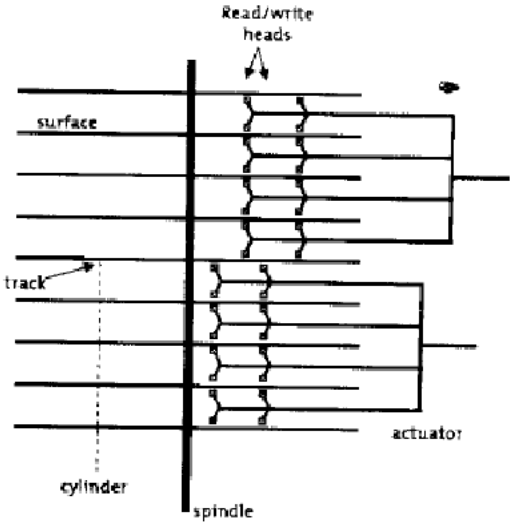
- I dischi CKD sono nati per il mercato mainframe (FBA e' l'architettura nel mondo Open)

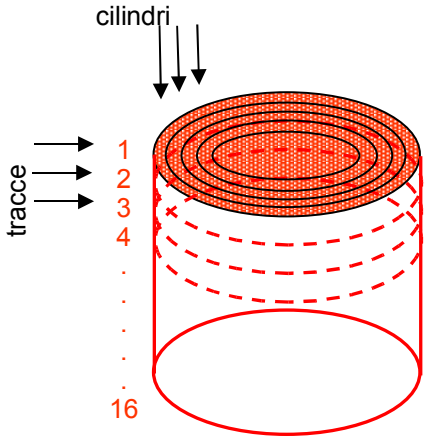
- Anche indicati con il nome DASD (Direct Access) consistono in una serie di superfici, disposte in parallelo a formare un **cilindro**, per la memorizzazione dei dati. Dei bracci (actuator) muniti di testine di lettura/scrittura si muovono simultaneamente sulle superfici dove i dati vengono memorizzati in **tracce** concentriche.

- Ogni movimento del braccio munito di testine permette di posizionarsi contemporaneamente su 16 tracce e di trasferire senza altri movimenti meccanici un intero cilindro (costituito dalle tracce che sulle 16 superfici risultano nella stessa posizione).

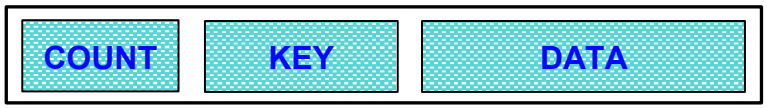
- 16 tracce formano 1 cilindro; una traccia e' di servizio e le altre 15 sono usate per memorizzare i dati; ogni traccia contiene un numero fisso di Byte (56.664) organizzati nel **tracciato record CKD**. Un record utente e' indirizzato indicando cilindro, traccia e numero record.

- Ogni modello di disco (3390-3, 3390-9, ...) ha un numero diverso di cilindri, quindi fornisce una diversa capacita' di memorizzazione





Il **tracciato record CKD** venne introdotto dal System 360 per avere una corrispondenza univoca tra i comandi eseguiti dai programmi ed il formato fisico dei dati sul disco



COUNT :	Identifica il record number (CCHHR)
KEY :	Campo opzionale per la ricerca
DATA :	Area dati; lunghezzza variabile per ciascun record

© 2007 IBM Corporation

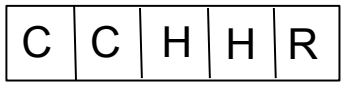
Architettura CKD – formato traccia

Ogni traccia disco e' formattata con una disposizione precisa di record:

- 'index' e' l'inizio della traccia
- 'home address' e' l'indirizzo fisico della traccia
- 'R0' e' il record zero che descrive la traccia
- Seguono i record di dati in formato *Count, Key, Data* oppure *Count, Data*

Es: RECFM=FB, LRECL=80, BLKSIZE= 4560
ogni Data Area conterra' 4560 bytes, divisi in record logici da 80 bytes
ogni traccia (device type 3390) conterra' 11 record fisici per 50.160 bytes

indirizzamento del disco fisico



CC – con 16 bit unsigned indirizza max un device type 3390 Mod. 54

HH – 16 testine lettura/scrittura

R – numero del record nella traccia

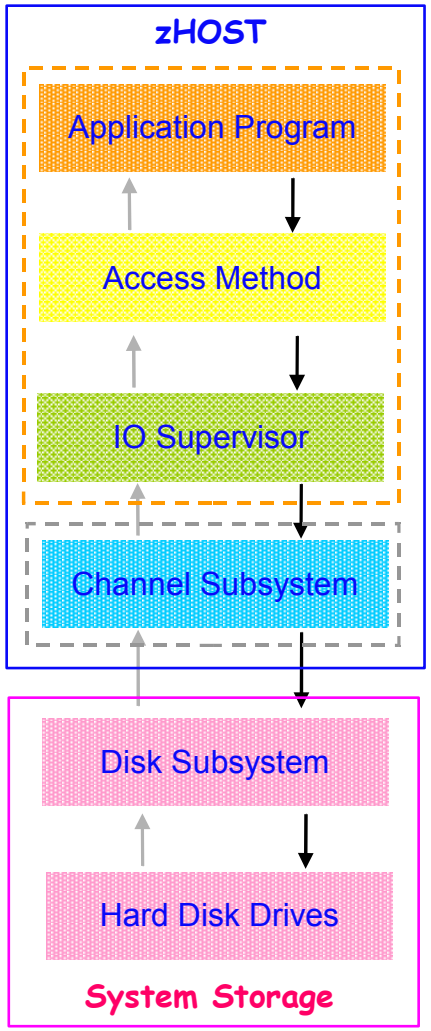
Modello	Cylinders	Tracks	Bytes/volume	Bytes/track
3390-1	1113	16695	946 MB	56664
3390-2	2226	33390	1.89 GB	56664
3390-3	3339	50085	2.83 GB	56664
3390-9	10017	150255	8.51 GB	56664
3390-27	32760	491400	27.84 GB	56664
3390-54	65520	982800	55.68 GB	56664

© 2007 IBM Corporation

Flusso logico della richiesta di un dato

componenti zOS

componente zSeries



Il programma richiede la lettura o scrittura di un record logico da un dataset allocato nelle JCL (DDname)
L'applicazione conosce il formato del record grazie ai parametri indicati in JCL

I logical records vengono accorpati in blocchi, come sono in realta' sul disco
Costruisce il programma di canale che descrive l'operazione di IO (channel command words **CCW)**
Viene passata la richiesta di IO al Supervisor tramite EXCP driver

Gli indirizzi virtuali dei puntatori ai dati sono tradotti in reali
E' gestito l'accodamento (se device non disponibie) in IOS Queue
E' dato il via all'operazione di IO con l'istruzione Start Subchannel (**SSCH**)
Gestisce lo IO Interrupt e notifica allo A.M. il completamento

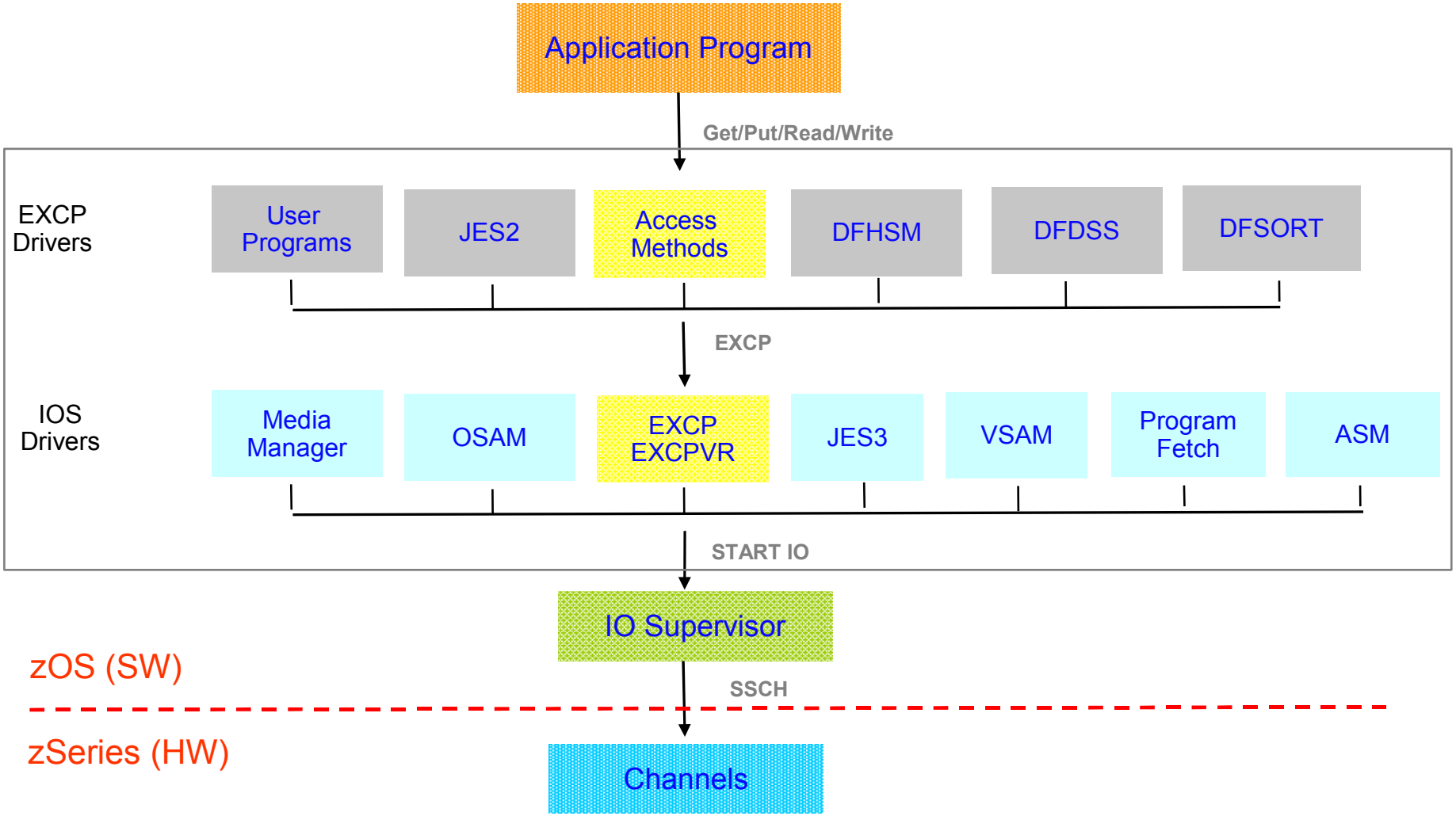
Seleziona il channel path al device usando le informazioni dello IOCP
Attende che il path risulti disponibile (Pending Time)
Passa le CCW's al Disk Subsystem

Decodifica il Channel Program in operazioni riconoscibili dal device
Verifica se l'operazione di IO puo' essere risolta in cache
Traduce i comandi CKD in FBA SCSI per richiedere operaziono ai Disk Drives

Legge o scrive sui dischi fisici usando comandi SCSI
Realizza smart caching ed error checking per segnalare errori HW e SW

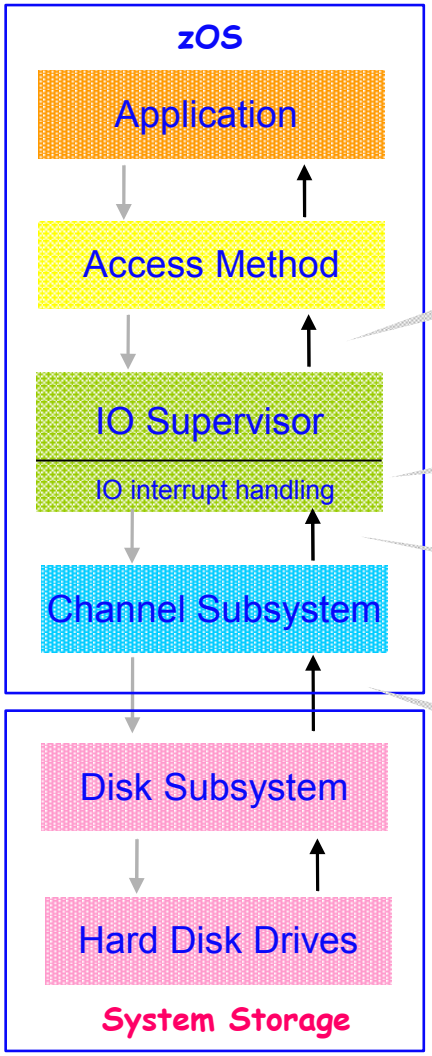
© 2007 IBM Corporation

Funzioni zOS per l'accesso ai dati



© 2007 IBM Corporation

IO interruption – l'applicazione ottiene il dato richiesto



POST

SRB

IO interrupt

Channel End

4. IO Supervisor informa l' Access Method dell'esito della IO

3. zOS verifica che l'IO e' completata correttamente e notifica lo IO Supervisor

2. zOS viene notificato quando l'IO e' completata

1. IO e' completata in modo asincrono dal Disk Subsystem

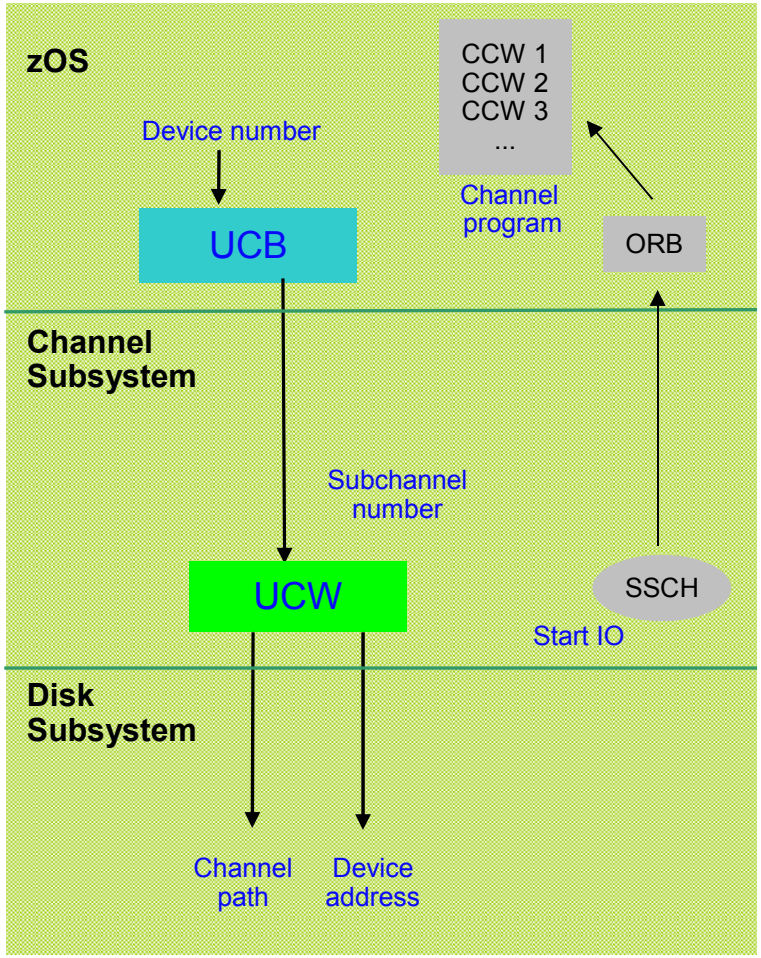
© 2007 IBM Corporation

Start dell'operazione di IO

Lo zOS usa il device number per identificare il volume logico definito nella generazione di IO e memorizza le informazioni del device nello Unit Control Block (UCB)

Il Channel Subsystem usa i subchannel number per identificare i device e le Unit Control Word UCW's contengono le informazioni sui device, compresi i path disponibili (equivalenti alle LCU)

Attraverso il canale prescelto viene acceduto il volume logico richiesto identificato dal device address



Start IO

Quando avviene lo Start di una operazione di IO (IO Supervisor fa SSCH) il subchannel number e' passato allo zOS.

Altre informazioni, come il puntamento al channel program e allo UBC, sono contenute nello ORB.

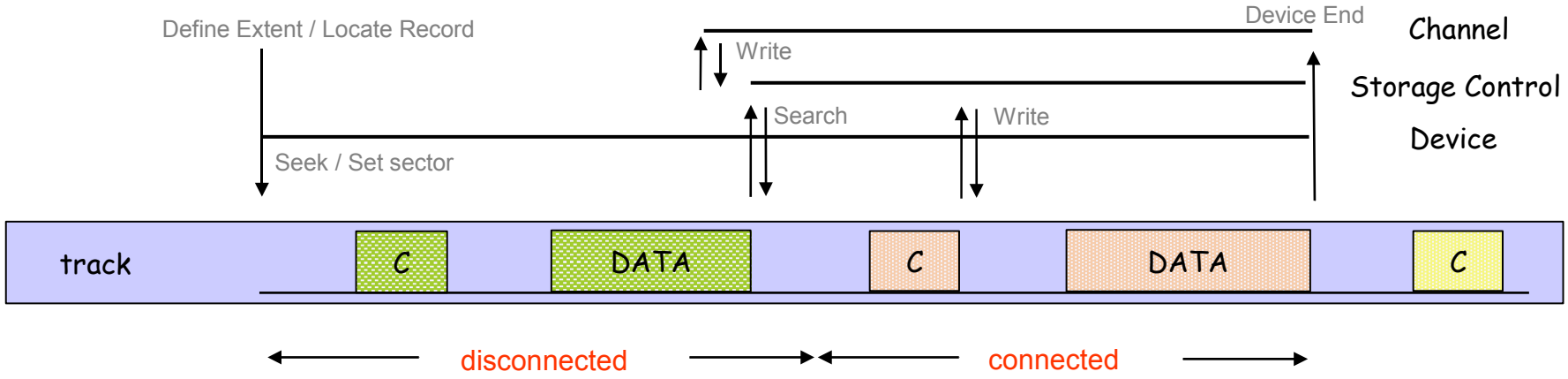
Le CCW contengono le istruzioni per gestire l'operazione sul canale e sul device, oltre a quelle per manipolare i dati.

© 2007 IBM Corporation

Esecuzione di un programma di canale non sincrono

Il tempo necessario al device per posizionarsi sulla traccia richiesta permette la disconnessione e l'acquisizione dei dati dal canale

- I comandi di Seek e Set Sector CCW (ricerca settore) vengono gestiti dal device in modo autonomo
- Prima di ricevere la Search CCW (posizionamento Data Area per scrittura) il device si deve riconnettere alla Storage Control
- Terminata l'operazione la Storage Control richiede la CCW successiva al canale

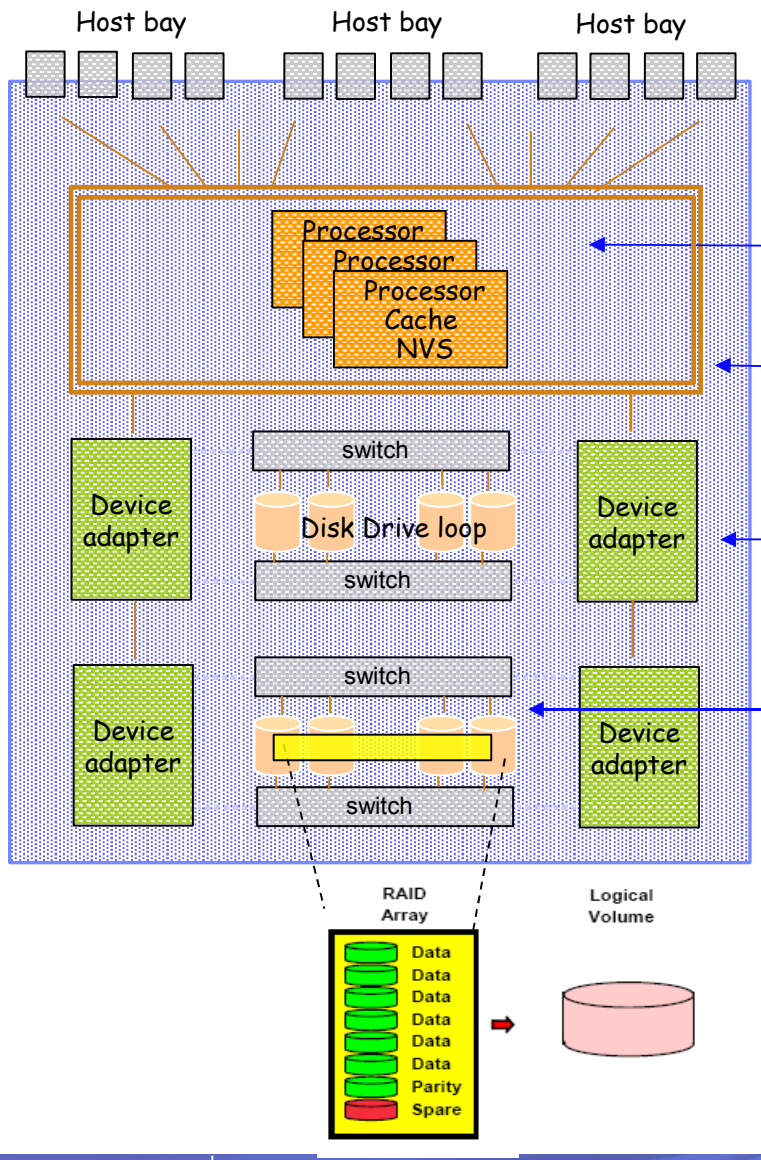


© 2007 IBM Corporation

- *L'architettura di I/O in ambiente zSeries*
- Sistemi Storage a disco
- *Sistemi Storage a nastro*
- *Le funzioni di copia dei dati*
- *Connettivita' e trasporto dei dati*
- *Cenni sulle prestazioni dei sistemi a disco*

© 2007 IBM Corporation

Disk Subsystem architecture



Host Adapter : processori specializzati (Frontend) per connessione ad host o per funzioni di replica

Processori e Cache : gestione delle operazioni. Per performance e data integrity

Network interna ad alta velocita' per il colloquio tra Frontend adapters, Processori e Backend drives

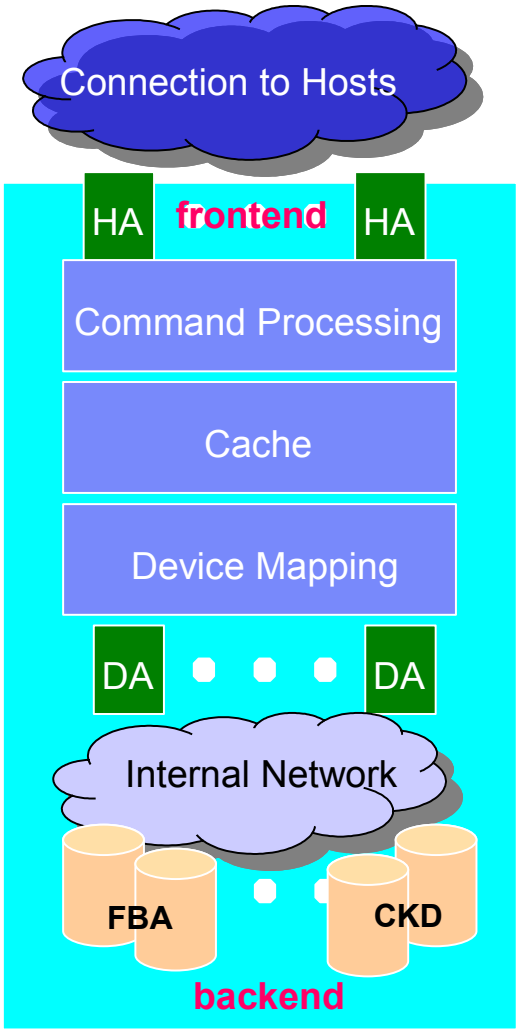
Device Adapter : processori specializzati per la connessione ai loop di dischi in backend (DDM) attraverso SSA o switched FC AL.

Disk Drives (DDM) : dischi in architettura SCSI di capacita' e velocita' diversa, organizzati in RAID e connessi ai DA attraverso rete interna ad alta velocita'

© 2007 IBM Corporation

Disk Subsystem - caratteristiche

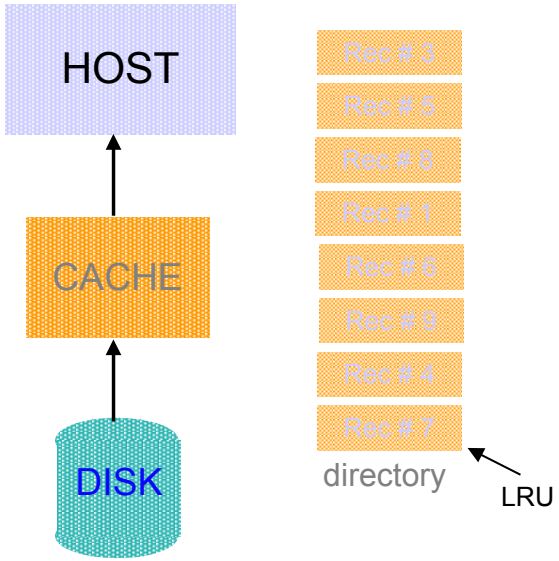
- Supporto per ambienti zSeries (CKD) ed Open (FBA)
- Connettivita' Fiber Channel/Ficon ed ESCON
- Il formato CKD e' usato per interfacciare i comandi Host ma non e' usato come formato dei dati a disco.
- Virtualizzazione degli SCSI device in volumi logici raggruppati in Logical Subsystem (LCU). Simulazione della geometria 3390 su device SCSI in architettura RAID
- Funzioni microcodificate di copia dei dati (locale e remota) e di ottimizzazione delle prestazioni per l'accesso ai dati (es. PAV)
- Gran parte delle operazioni risolte a velocita' di memoria/network senza coinvolgere i device in backend : tutte le scritture avvengono in cache : 100% write hit
- Operazioni effettuate dai dischi in backend :
  - Stage : lettura dal device per cache miss pre-stage
  - Destage: scrittura a disco di dati modificati dati non rimossi dalla cache
  - Demote : rimozione di dati dalla cache (non modificati)



© 2007 IBM Corporation

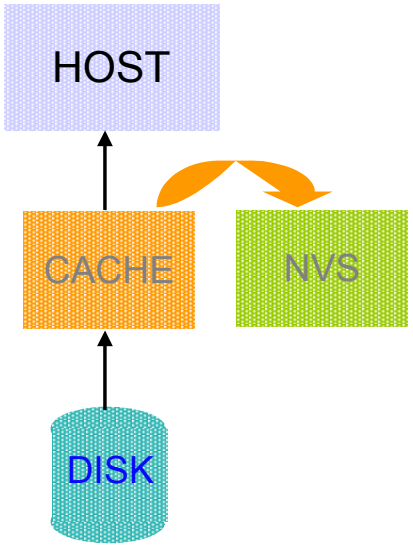
Caching concepts

Caching read operation



- La cache e' una memoria elettronica che permette di memorizzare i dati piu' utilizzati ed accederli piu' velocemente
- Mantiene una lista con i blocchi usati piu' di recente cosi' da rimuovere i meno utilizzati
- La porzione di cache volatile contiene solo dati scritti sui dischi in backend

Caching write : a safe copy



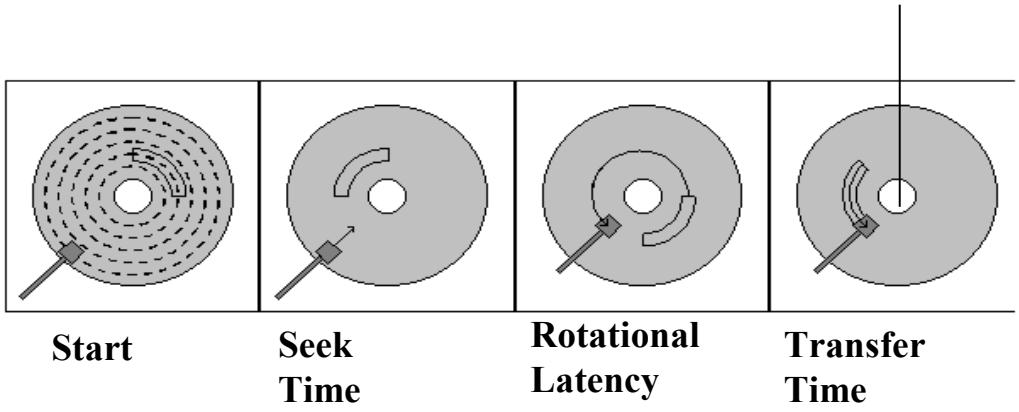
L'integrita' dei dati e' preservata garantendone la loro scrittura permanente:

3. il dato e' scritto in cache
4. Una seconda copia e' scritta nella parte non volatile (NVS) della cache
5. E' segnalata al host l'avvenuta scrittura
6. In maniera asincrona il dato verra' scritto sui dischi in backend

© 2007 IBM Corporation

Disk Drive - fondamenti

- read / write cache hits sono nell'ordine di ~ 1ms
- physical disk I/O operations sono nell'ordine di > 4ms poiche' richiedono l'intervento di componenti meccaniche come il movimento delle testine e la rotazione del disco
- ogni disk drive puo processare un numero ben preciso di I/O al secondo, determinato da:
  - *average seek time* = posizionamento della testina sulla traccia richiesta
  - *rotational latency* = rotazione della superficie del disco finche il primo settore indirizzato passa sotto le testine di read/write heads (avg. time = meta' rotazione)
  - *transfer time* = read/write dei settori di dati (1 sector = 512 Byte)



© 2007 IBM Corporation

Architettura RAID - Concetti

- Utilizzo di dischi economici pur mantenendo una buona affidabilita' del sistema complessivo
  - *mettere insieme piu' dischi SCSI di size ridotto risulta piu' conveniente che assemblare un Single Large Expensive mainframe Disk (SLED)*

- Ridondanza delle componenti per permette di ricoverare i dischi in errore anche a fronte di piu' eventi contemporanei
  - *vengono implementate diverse tecniche di RAID per rispondere a requisiti di performance / availability / capacity*

- Tecnica di ridondanza delle informazioni piu' efficente del semplice mirroring
  - la *tecnica del bit di parity risulta piu' conveniente della duplicazione del dato*

- Fornire un 'data rate' elevato grazie alla distribuzione dei dati su piu' dischi
  - *avere piu' meccaniche che lavorano in parallelo su dati 'striped' garantisce un throughput elevato*

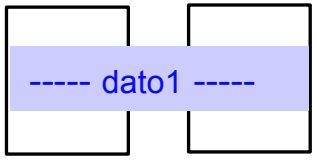
- RAID10, RAID5, RAID6 sono le tecniche maggiormente diffuse

© 2007 IBM Corporation

RAID types

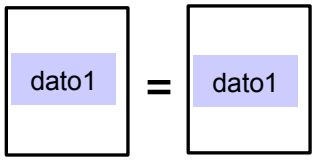
RAID 0 (striped, no RAID)
Dato distribuito su un set di dischi

Data rate=elevata
Availability=bassa



RAID1 (Mirroring)
Dato scritto su un disco in replica

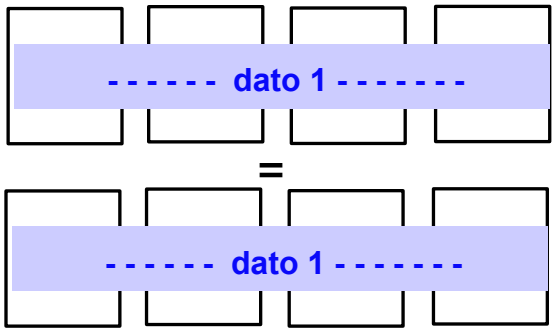
Write overhead=100%
Read overhead / benefit=0



RAID10 (striped + mirroring)
Il dato e' distribuito su un set di dischi duplicato

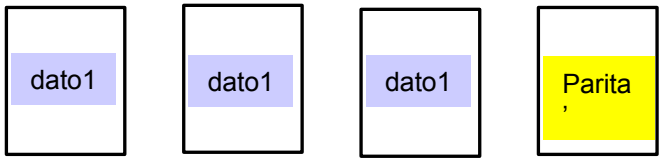
Capacita' utile dimezzata

Overhead scrittuta = 100%
Availability=eccellente



RAID 5 (striped + parita')
Data distribuito su un set di dischi

Data rate=elevata in lettura
Overhead scrittuta = da 114% a 400%
Availability= uso di un bit di parity (alta)
Capacita' persa = da 14% a 33%



RAID 6 (striped + doppia parita')
Data distribuito su un set di dischi

Data rate=elevata in lettura
Overhead scrittura = da 133% a 600%
Availability= uso di doppio bit di parity (eccellente)
Capacita' persa = 33%