Linux Day Torino
28 ottobre 2023

Storage
anno 2023
dalle basi al futuro

Massimo Nuvoli

Mi presento

- Architetto di Sistemi
- Lavoro per me stesso, Progetto Archivio SRL, Dicobit
- Trainer certificato in ambiente tecnico

e …

- Co-fondatore di Adenda SRL (CTO)

  Provider di infrastrutture di rete, con il primo datacenter innovativo a Torino

ma anche...

Storage: cosa vediamo oggi?

- Un po' di teoria
- Dischi magnetici
- Dischi allo stato solido
- Cosa diavolo è un tebibyte
- Come si connette storage internamente ed esternamente
- Raid, ma non sono insetti
- La nuova rubrica “C'è ma non lo sai”
- Il futuro?

Storage, ma cos..

- Partiamo dalle basi, l'informatica può essere riassunta in un semplice

$$y=f(x)$$

quindi?

- abbiamo una funzione che prende un dato e lo elabora
- tutto dipende quindi da x e y
- così è un po' troppo semplice

$$y=f(x)$$

quindi?

- immaginiamo che il dato x cresca di dimensioni in modo evidente
- la trasformazione da x a y richiede di potere accedere a x sicuramente e potere scrivere in qualche modo y

$$y=f(x)$$

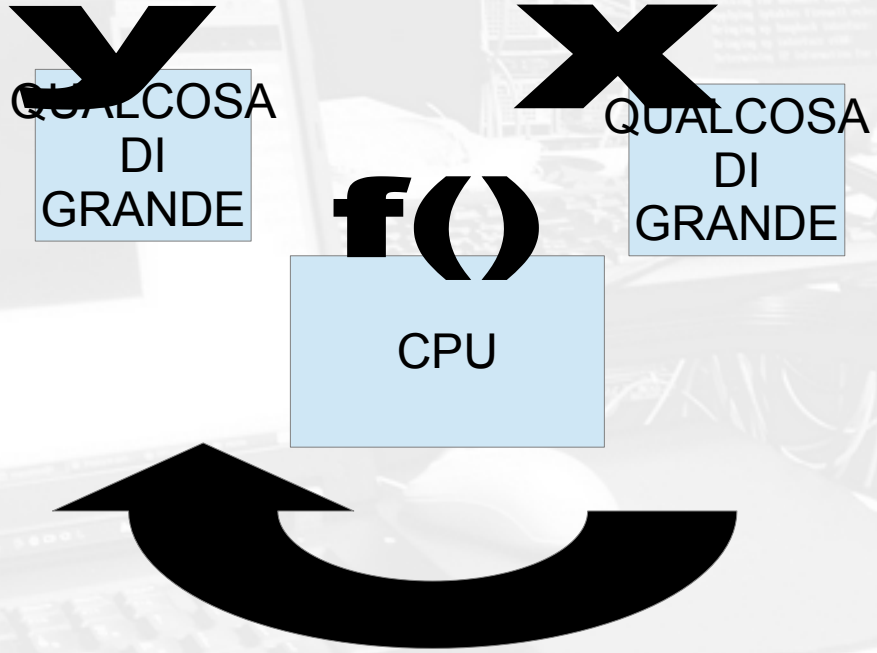
lo storage?

- quando x e y diventano enormi prima di tutto dovranno stare da qualche parte, dovranno essere letti, messi in memoria, quindi elaborati, quindi magari scritti da qualche parte

$$y=f(x)$$

lo storage?

- quando x e y diventano enormi prima di tutto dovranno stare da qualche parte, dovranno essere letti, messi in memoria, quindi elaborati, quindi magari scritti da qualche parte



Storage = Qualcosa di grande

- In realtà la percezione di grande è relativa
- Negli anni 80 era grande la cassetta su cui salvavamo i giochi dei primi PC
- Negli anni 90 erano grandi gli hard disk da pochi MB
- Negli anni 2000 erano grandi gli hard disk da pochi GB
- Negli anni 2010 erano grandi gli hard disk da pochi TB
- Oggi sono grandi gli hard disk da decine di TB

Un po' di matematica per capire

- Anni 80 → migliaia di caratteri → KB
- Anni 90 → alcuni milioni di caratteri → MB
- Anni 2000 → alcuni miliardi di caratteri → GB
- Anni 2010 → alcune migliaia di miliardi di caratteri → TB
- Anni 2020 → centinaia di migliaia di miliardi di caratteri
- futuro → PB milioni di miliardi di caratteri

Prima di tutto...

- Storicamente le dimensioni dei dati vengono espresse in potenze di 2 (così è fatta la memoria dei sistemi)
- 1KB = 1024
- 1MB = 1024*1024
- 1GB = 1024*1024*1024
- 1TB = 1024*1024*1024*1024
- 1PB = 1024*1024*1024*1024*1024

fino a quando quel 2,4% non ha dato fastidio

ed ecco IEEE 1541

- All'inizio la differenza tra 1000 e 1024 era qualcosa di trascurabile (2,4%)
- Nel 2002 ci si accorge che questa cosa non è sostenibile quindi si distingue tra le dimensioni “binarie” e le dimensioni “decimali”

  La storia è di lunghe discussioni ma nel 2005 IEEE 1541 diventa “ufficiale” standard IEEE internazionale
- Da qui in poi nascono infinite discussioni… soprattutto perché i dischi vengono venduti per MB e TB decimali mentre sul PC il sistema operativo continua a riportare spazi in base 2

Il nuovo sistema “i”

Decimale SI			Binario		
chilobyte	kB	10^3	kibibyte	Ki	2^10
megabyte	MB	10^6	mebibyte	Mi	2^20
gigabyte	GB	10^9	gibibyte	Gi	2^30
terabyte	TB	10^12	tebibyte	Ti	2^40
petabyte	PB	10^15	pebibyte	Pi	2^50
exabyte	EB	10^18	exbibyte	Ei	2^60
zettabyte	ZB	10^21	zebibyte	Zi	2^70
yottabyte	YB	10^24	yobibyte	Yi	2^80

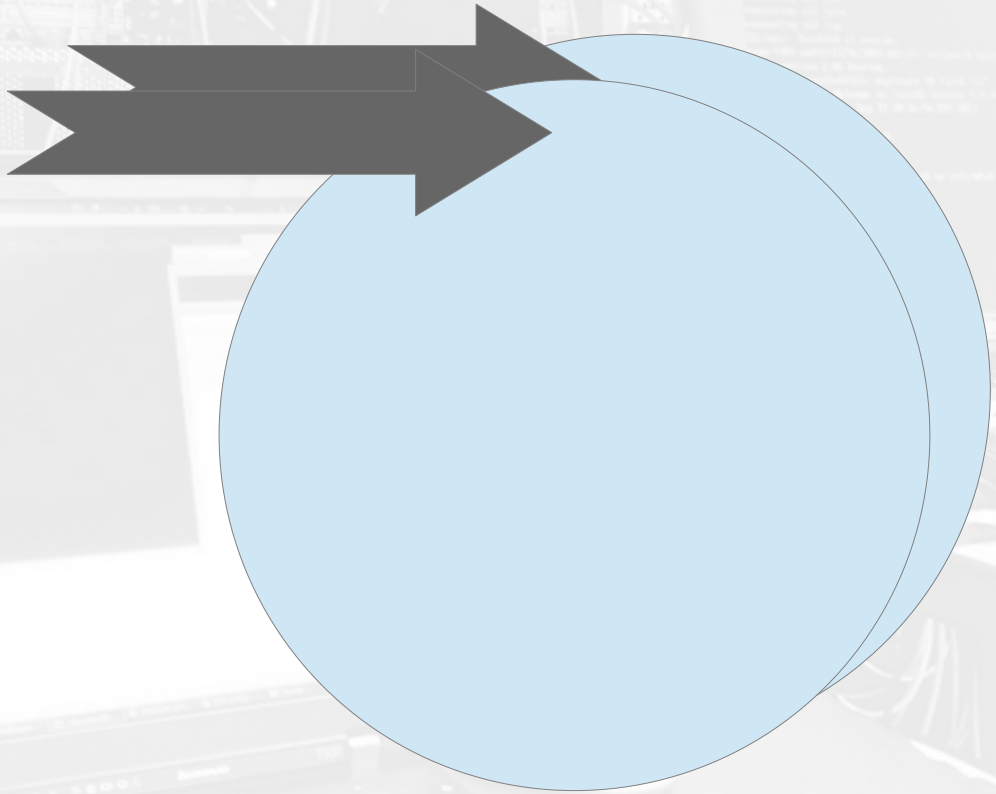
i dettagli e la guerra

- Dopo l'uscita dello standard IEEE 1541 in teoria tutto dovrebbe essere più semplice
- In alcuni casi è stato abbracciato il sistema SI, in altri si tiene in pista il binario, con un ambiguo uso di K M G T P che sono più probabilmente KiB o Ki non KB
- In realtà ovunque la differenza tra 1024 e 1000 (quel maledetto 2,4%) sia determinante è importante specificare quale unità si utilizza.
- Un esempio: gli hard disk, gli SSD etc. etc.

tipi di “storage”

- non permanente
  - ferriti
  - memoria
- permanente
  - nastro
  - disco
  - stato solido (memoria flash)
  - ottico

i dischi

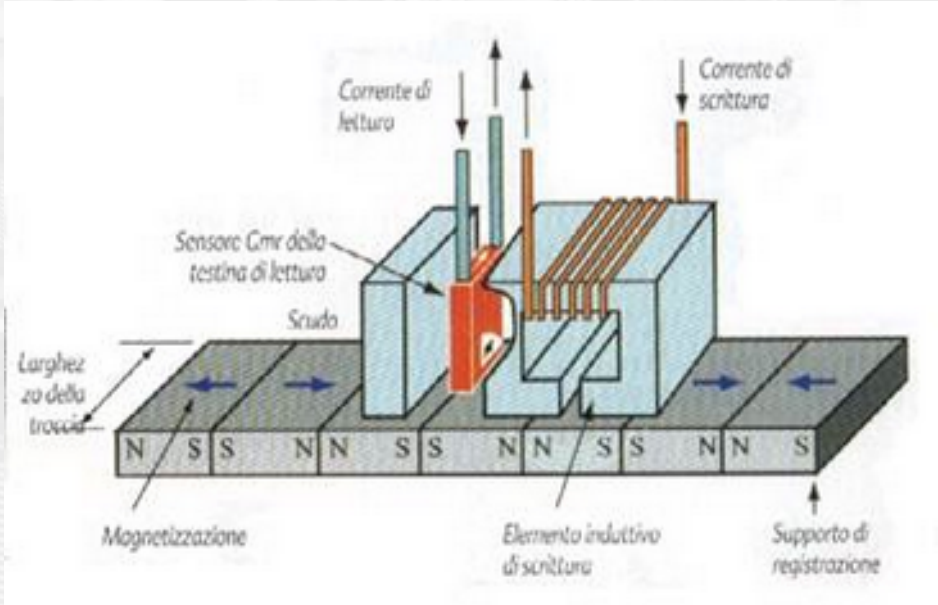


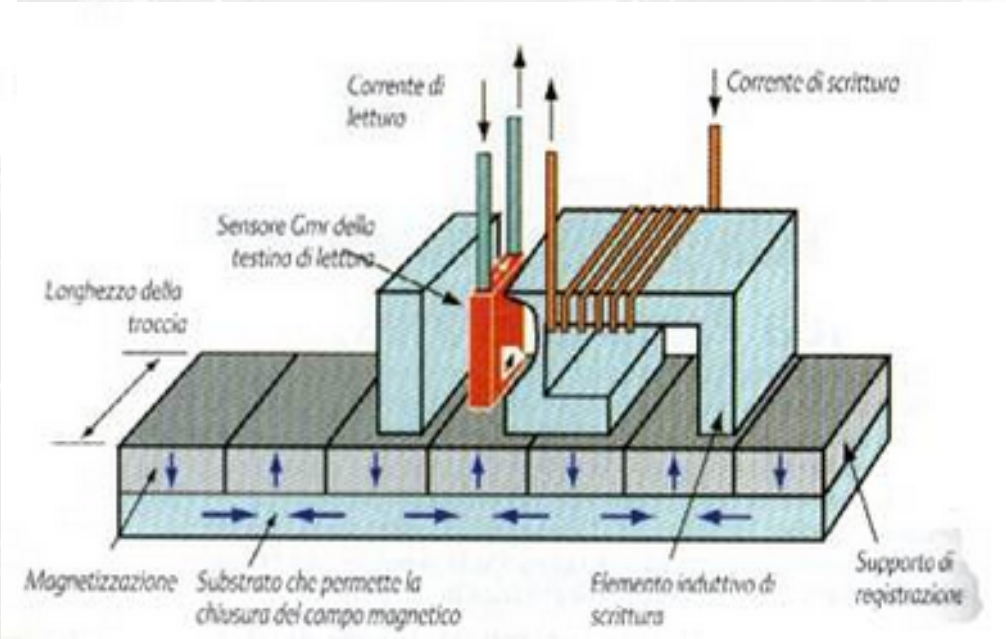
Limiti “fisici” dei dischi

- Latenza per accedere ai dati dovuta al posizionamento XY del dato
- Lo spostamento della testina non può avvenire in modo così rapido e bisogna anche aspettare che il dato passi sotto alle testine
- La velocità del disco angolare è costante, ma cambia quella lineare in base alla distanza dal centro
- Sia i dischi ottici che i dischi magnetici hanno una tecnologia estremamente complessa, affascinante, vediamo alcuni esempi

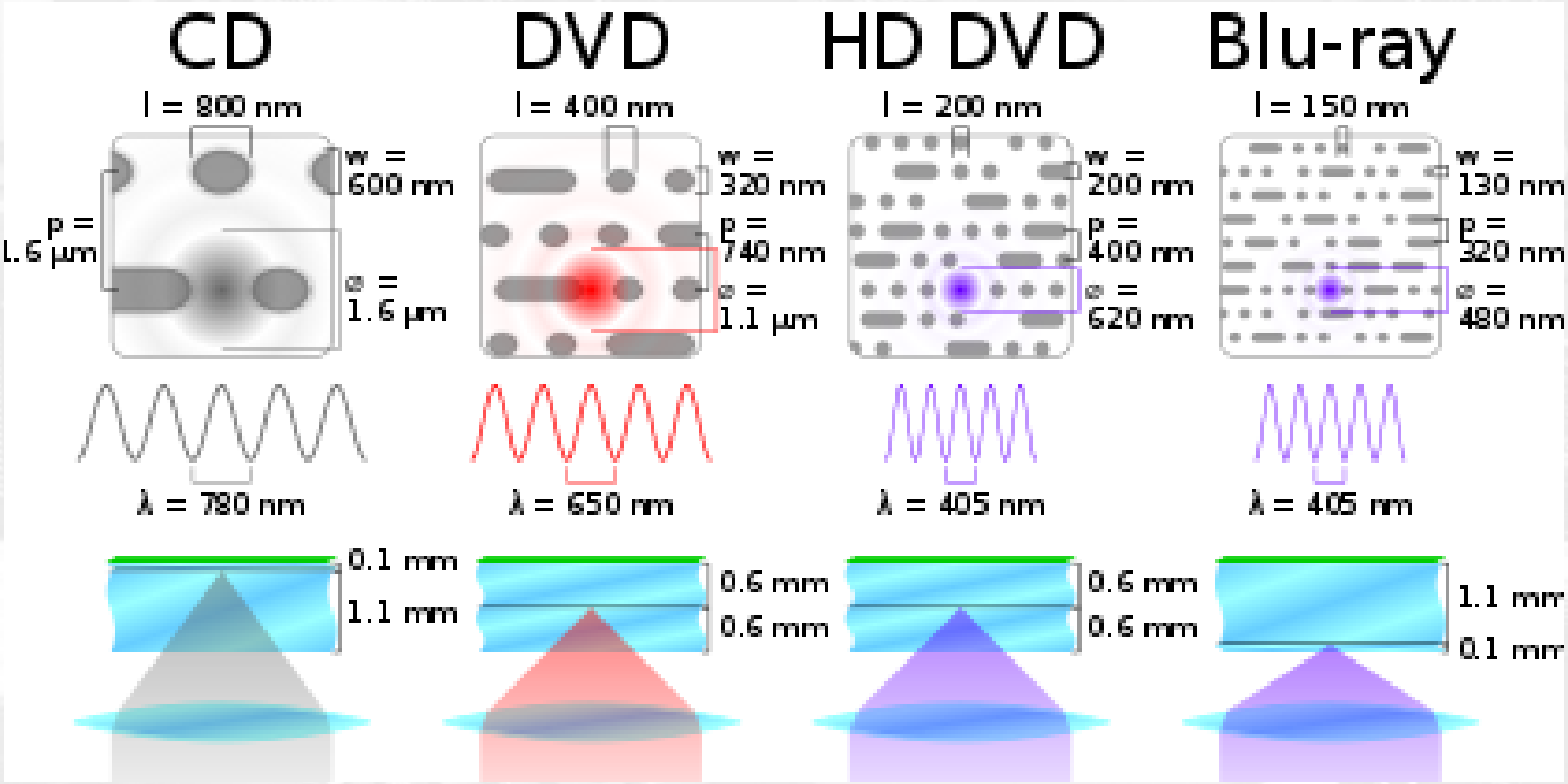
Le testine dei dischi magnetici

Dischi magnetici





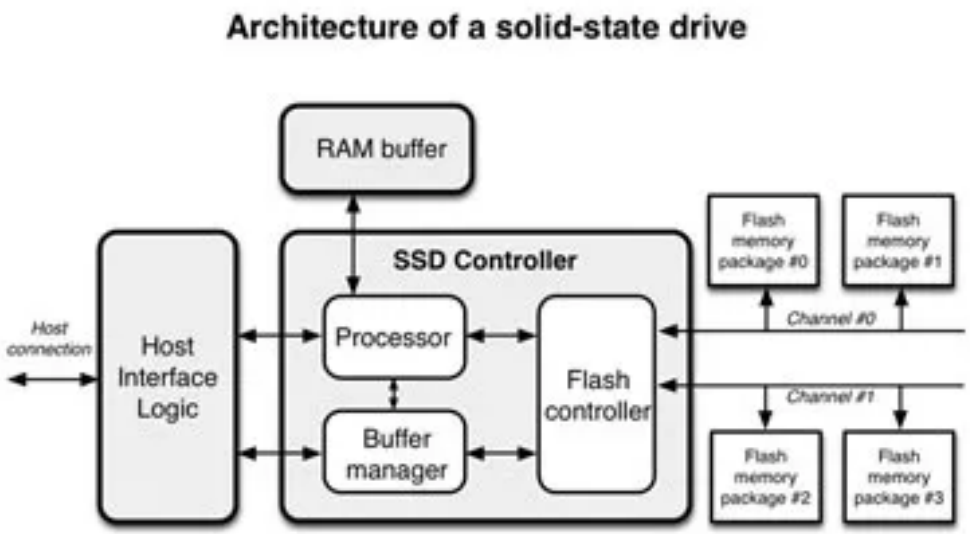
Dischi Ottici (laser)

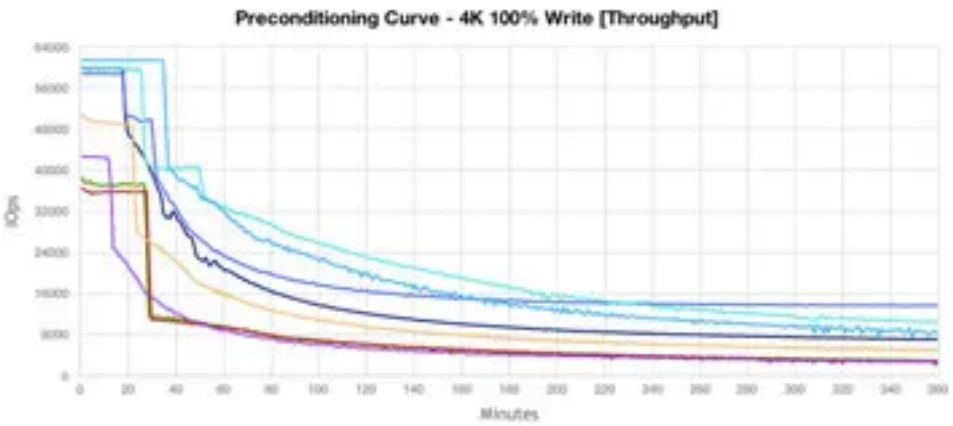


dischi “ibridi”

- Se non lo sapete esistono dischi ibridi, venuti fuori nel tempo
- per cercare di ovviare ai limiti dei dischi magnetici
- per cercare di ovviare ai limiti dei dischi ottici
- Magneto ottici da 300MB o 600MB
- Floppy da 120MB (stesso formato 3.5”)

Hard Disk allo stato solido





Hard Disk allo stato solido

- Hanno i loro limiti
- La qualità dei componenti determina la loro affidabilità che non è “infinita”
- A seconda dei tipi di componenti potrebbe sussistere il problema della “cancellazione” e della “riallocazione dei blocchi compromessi”
- Al crescere della velocità ritorniamo ad avere problemi di latenza, questa volta per via della logica di controllo

SSD e il problema riscrittura (a)

- Ogni tipo di disco SSD ha una caratteristica di costruzione che viene denominata in modo equivoco per non informare correttamente i consumatori
- Ad ogni scrittura l'affidabilità dei singoli blocchi di celle “viene consumata” ed in base alla tecnologia costruttiva dell'SSD questo significa che prima o poi l'SSD non riuscirà più ad essere affidabile
- Si chiama WEARING, ogni SSD ha un livello massimo di riscritture, raggiunto questo livello il disco diventa statisticamente inaffidabile

SSD e il problema riscrittura (b)

- Nel caos esistono più modi per capire quanto dura un SSD
  - DWPD per un certo periodo di garanzia, indica quante volte al giorno può essere riscritta la superfice del disco
  - TBW indica la quantità di TB che possono essere scritte sul disco prima di uscire dal suo effettivo target di vita o di garanzia
  - E altri

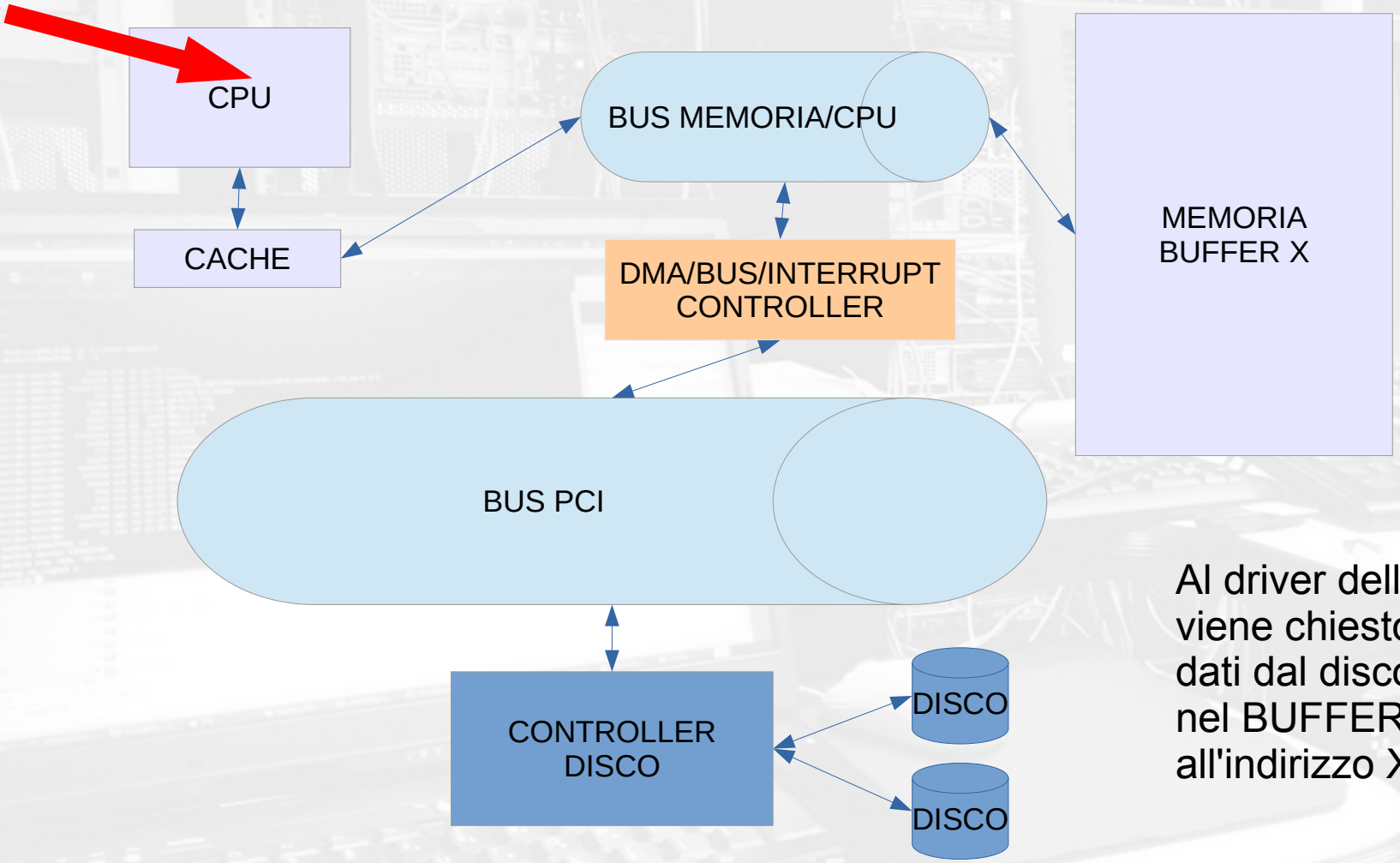
SSD e il problema riscrittura (c)

- Due SSD con DWPD identico non hanno la stessa durata in termini di TBW
  - Ad esempio il TBW con obiettivo 5 anni di un disco da 15TB con DWPD 1 è pari a 27375
  - Il TBW con obiettivo 5 anni di un disco da 1TB con DWPD 1 invece è solo 1825
- Sia nell'uso aziendale che in quello privato la sostituzione dell'SSD va pianificata in modo accurato, eventualmente monitorando il parametro WEARING intervenendo per tempo

DMA Direct Memory Access

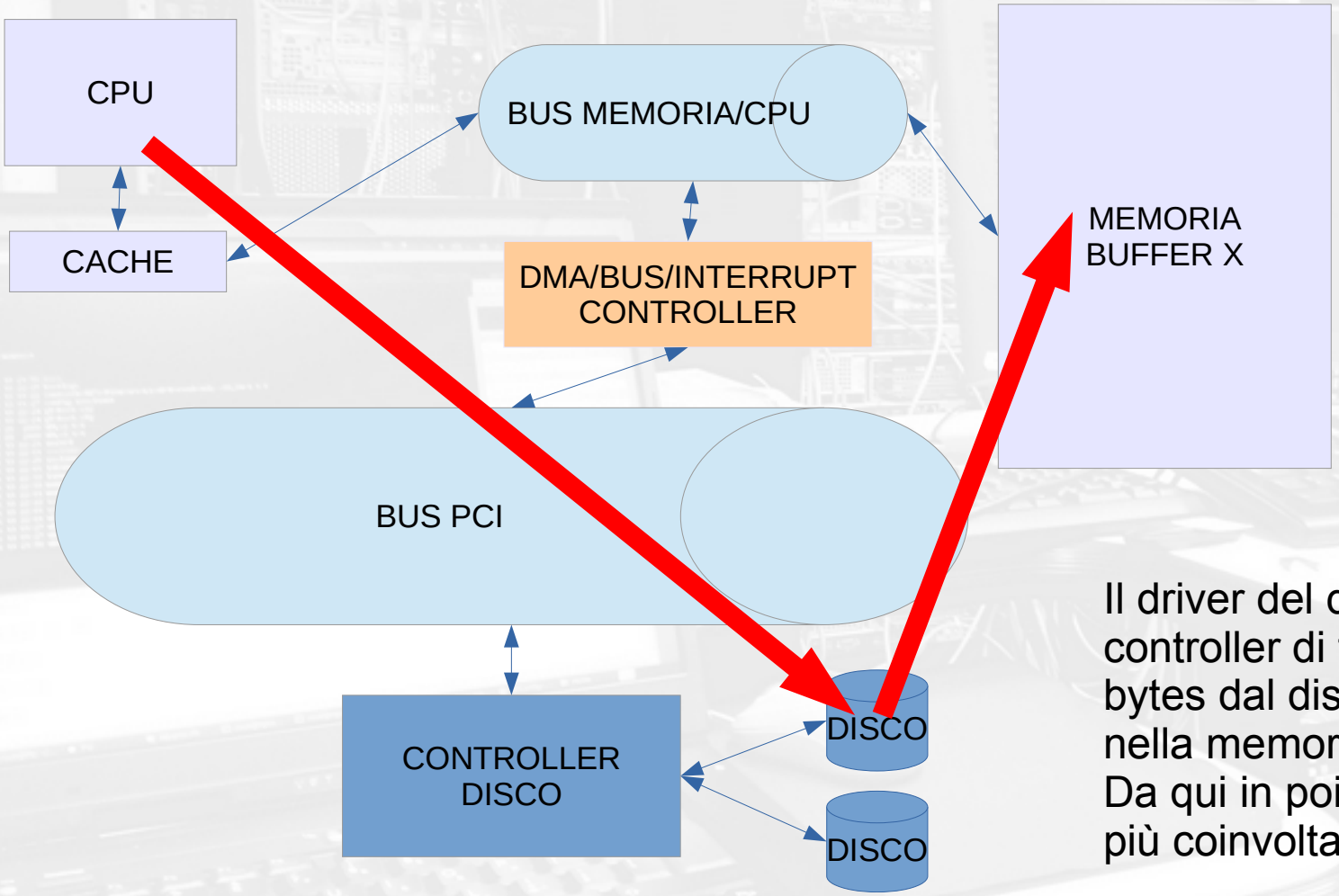
- Giusto per complicare un po' il talk.. :-)
- E' impensabile per mille motivi che la CPU si occupi di trasferire grandi quantità di dati direttamente, oltre al tempo di attesa si creerebbero latenze difficilmente gestibili
- Nasce nella notte dei tempi addirittura con IC dedicati sulle vecchie MB
- Ma come funziona?

DMA Direct Memory Access (1)



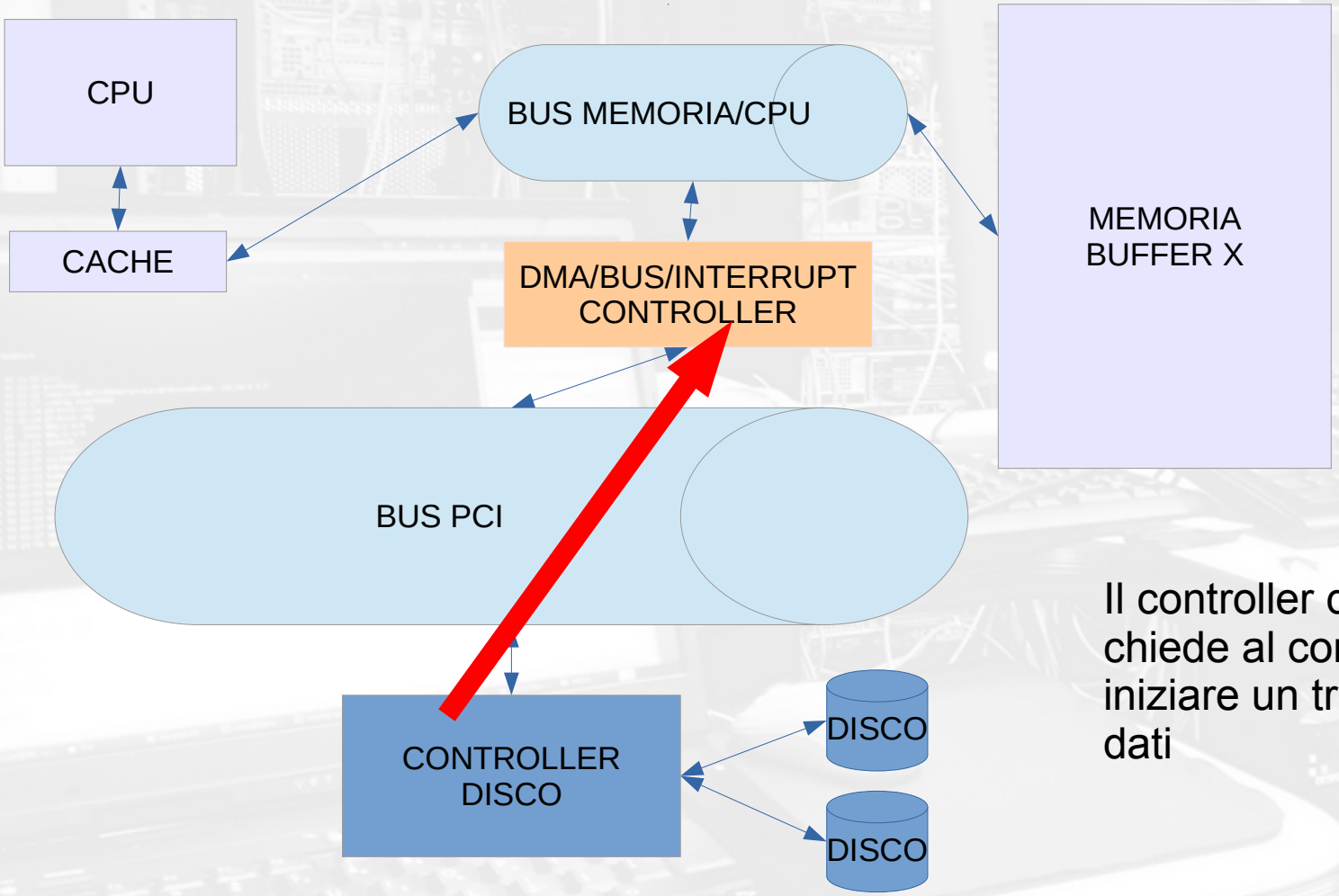
Al driver della periferica viene chiesto di leggere i dati dal disco e depositarli nel BUFFER in memoria all'indirizzo X

DMA Direct Memory Access (2)



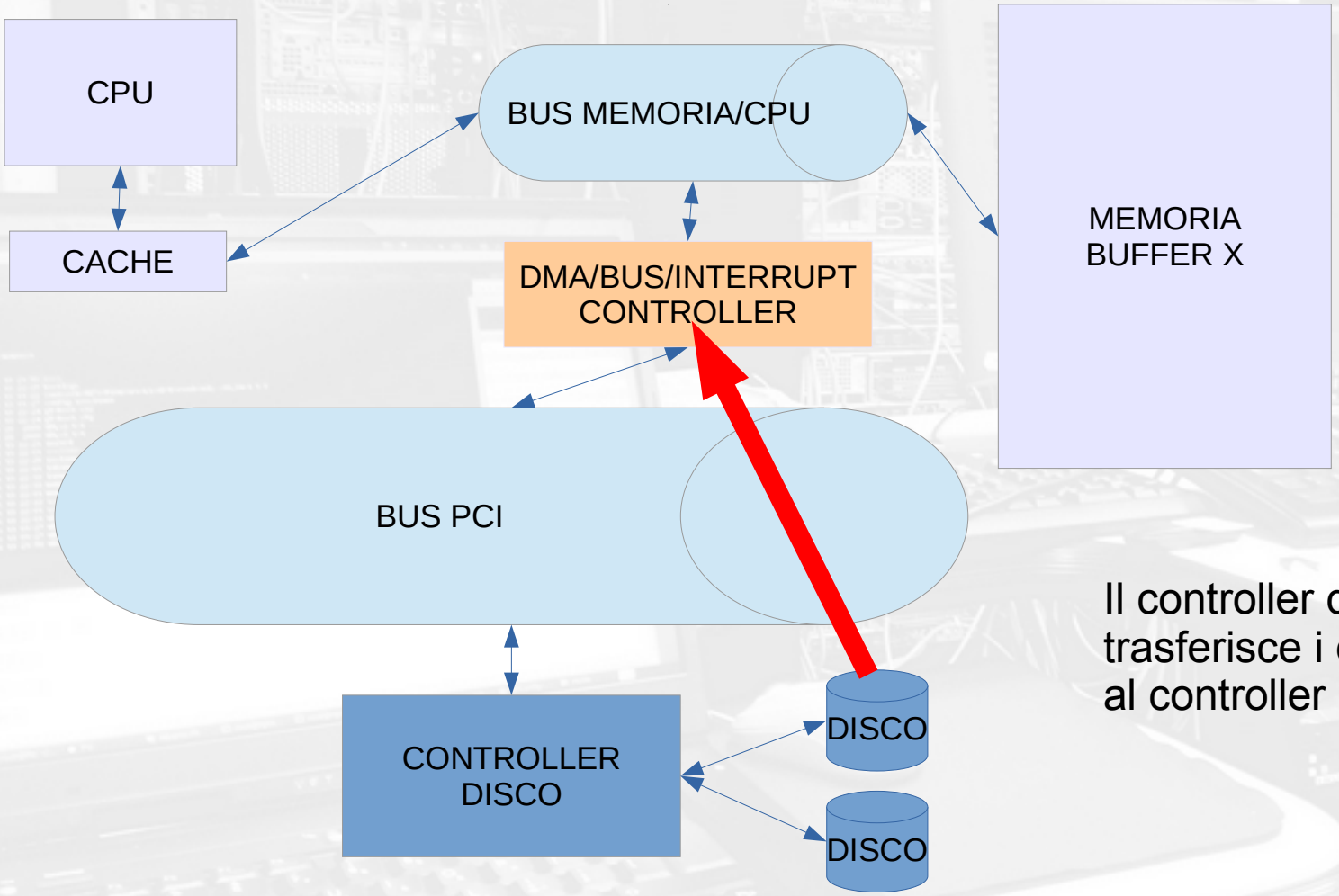
Il driver del disco dice al controller di trasferire C bytes dal disco al buffer X nella memoria
Da qui in poi la CPU non è più coinvolta

DMA Direct Memory Access (3)



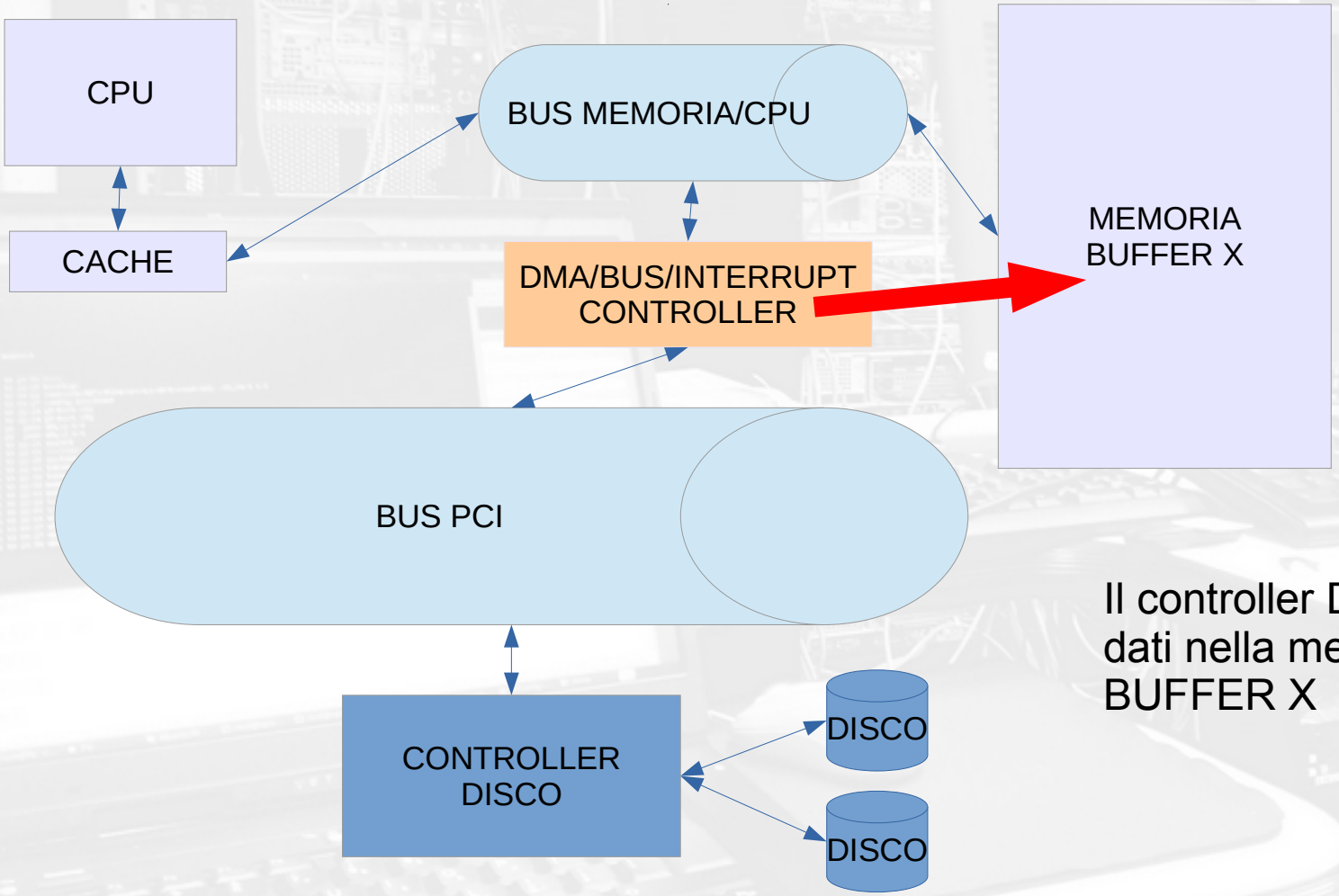
Il controller del disco chiede al controller DMA di iniziare un trasferimento di dati

DMA Direct Memory Access (4)



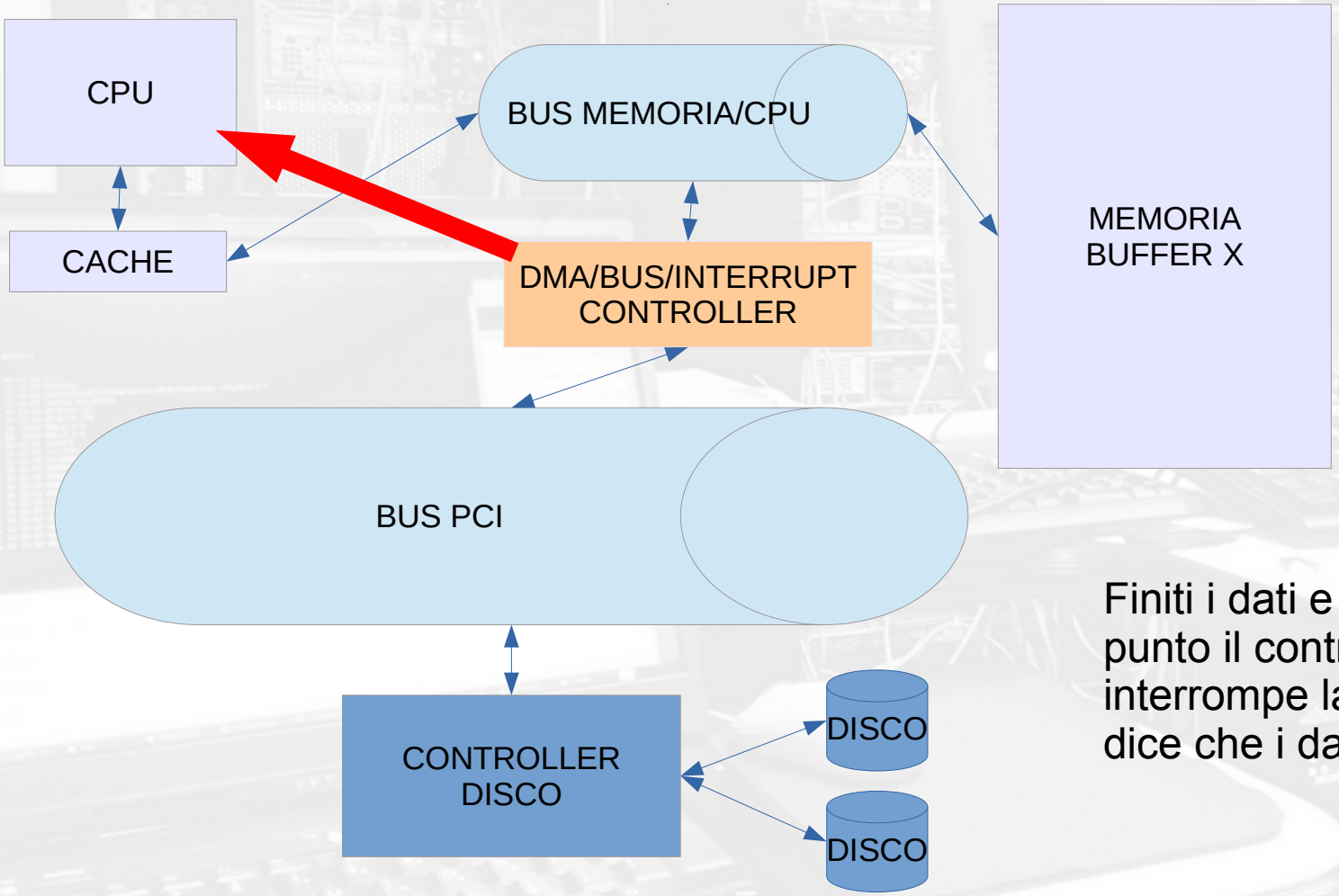
Il controller del disco trasferisce i dati dal disco al controller DMA

DMA Direct Memory Access (5)



Il controller DMA mette i dati nella memoria BUFFER X

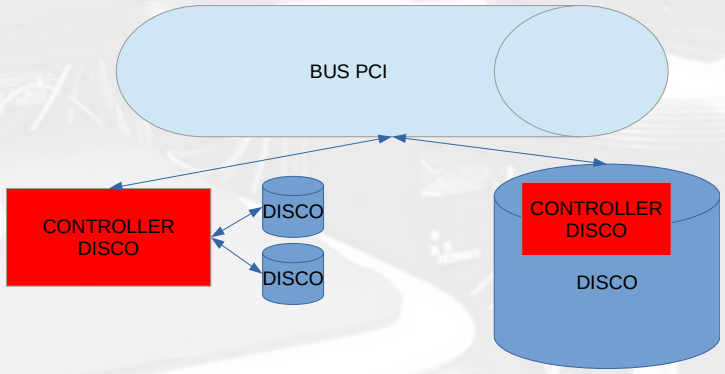
DMA Direct Memory Access (6)



Finiti i dati e solo a questo punto il controller del DMA interrompe la CPU e gli dice che i dati sono pronti

Perchè parlare di DMA?

- Ma perchè complicarvi la vita?
- Tutte le operazioni di IO anche quelle più banali sono estremamente complesse soprattutto per garantire la velocità e la latenza
- L'evoluzione dell'IO ha portato il controller PCI direttamente nel disco eliminando un elemento nella catena di controllo
- il DMA diventa fondamentale!!



Come si collega internamente?

- IDE
- SATA
- mSATA
- SAS
- PCIe
- NVME

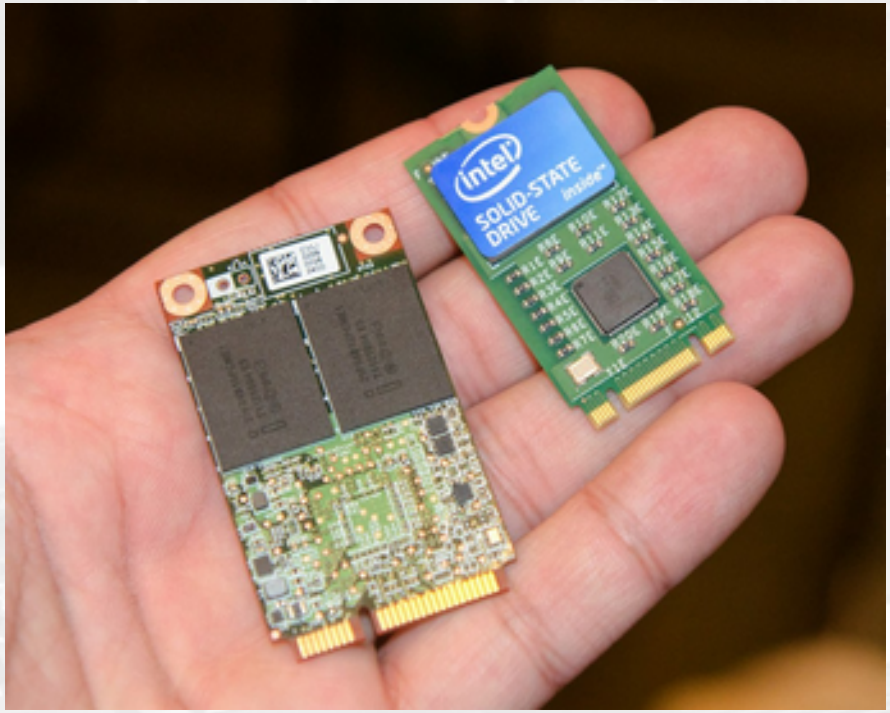
Lo standard SATA

- SATA significa Serial Advanced Tecnology Attachmet
- SATA → 1,5 Gbit al secondo (circa 125MB al secondo)
- SATA 2.0 → 3 Gbit al secondo (circa 250MB al secondo)
- SATA 3.0 → 6 Gbit al secondo (circa 500MB al secondo)
- Ora siamo a SATA 3.5
- Dispositivi indirizzati “per porta”

Solo dalla versione 2 inizia ad essere previsto NCQ (riordinamento dei comandi) per migliorare prestazioni con flussi video e accesso contemporaneo

Lo standard mSATA e M.2

- mSATA utilizza un connettore tipo Mini PCIe che al suo interno porta i segnali SATA
- M.2 implementa la nuova SATA Express (Pci Express) e anche un bus USB 3.0 su porta NGFF



Lo standard SAS

- SAS sta per Serial Attached SCSI
- versione 3.0 Gib/s → 300MiB/s
- versione 6.0 Gib/s → 600MiB/s
- versione 12.0 Gib/s → 1200MiB/s
- Dispositivi indirizzati per WWN (World Wide Name)
- Alcune caratteristiche:
  - Trasferimento Full-duplex con l'aggregazione di 8 link su wide ports 24/48/96 Gib/s.
  - 3.0 Gib/s per link nella versione introduttiva va fino a 12.0 Gib/s
  - Cavo esterno max 8 metri
  - 128 dispositivi per porta (16.384 in totale)
  - Compatibilità SAS-SATA

Dischi connessi in PCIe

- Con l'avvento dei dischi SSD diventano evidenti sia i limiti di velocità che di latenza del “sistema” SATA
- Si sceglie di interconnettere gli SSD direttamente al bus PCIe tramite controller proprietari
- Scelta parecchio performante ma che non è praticabile su qualsiasi PC
- Per “performare” la scheda deve utilizzare un numero elevato di canali PCIe (minimo 8x) e non tutti i PC sono dotati di abbastanza slot > a 1x
- Sono spariti dal mercato grazie a NVME

Lo standard NVME

- NVME sta per NVM express o meglio Non-Volatile Memory Host Controller Interface Specification NVMHCIS
- connessione AIC (scheda PCI aggiuntiva)
- connessione U.2 (connettore SAS) solo NVME
- connessione U.3 (connettore SAS) tricompatibile SATA/SAS/NVME
- M.2 NGFF (quello) PCIe 3.0

NVME over qualcosa

- Si parla di NVMEoFC (vedi sotto) ovvero della possibilità di adottare lo standard NVME utilizzando come trasporto FC (20Gbps)
- Ma si parla anche di NVMEoE (over Ethernet) o NVMEoIP utilizzando come trasporto L2 ethernet o addirittura L3

E la connessione esterna?

- USB
- eSATA
- SAS
- Thunderbolt
- FC
- iSCSI

USB

- USB sta per Universal Serial Bus, non nasce per lo storage, fa anche quello
- USB 1.0 → velocità 1Mbps 125KB/sec
- USB 1.1 → velocità 12Mbps 875KB/sec
- USB 2.0 → velocità 280Mbps 35MB/sec
- USB 3.0 → velocità 3,2Gbps 400MB/sec
- USB 3.1 → velocità 7,2Gbps 900MB/sec
- USB 3.2 → velocità teorica 20Gbps (2,5GB/sec)
- USB 4→ velocità teorica 40Gbps (5GB/sec)

eSATA

- standard poco adottato a livello mondiale
- arriva al massimo allo standard 2 (6Gib/s)
- esiste sia eSATA (solo dati) che eSATAp (con alimentazione) anche se questo non è uno standard ufficiale
- ampiamente surclassato da USB 3.0 e Thunderbolt

Thunderbolt

- E' una tecnologia standard sviluppata da Intel in collaborazione con Apple
- Combina i segnali PCIe e DP (Display Port) in due canali seriali, a cui si aggiunge alimentazione.
- Thunderbolt 1 2 utilizzano un connettore proprietario, dalla versione 3 il connettore è lo stesso di USB-C ed è compatibile con USB 3.1 gen 2
- Il problema è che “non è uno standard” in pratica è e resta una tecnologia di proprietà di Intel che in qualche modo da una parte cerca di diffonderlo ma dall'altra parte impedisce la certificazione ai competitor (vedi AMD)

FC o Fibre Channel

- E' uno standard storico, presente come standard dal 1988
- Partito da velocità sub/megabit/secondo oggi è approdato a 20GFC
- Connessioni di tipo ottico prevalentemente
- Esiste anche FCoE (over ethernet) e ovviamente tentativi di fare FCoIP
-

iSCSI

- Vi rimando alla mia analisi del 2022 sulle differenze tra FC e iSCSI
- Largamente adottato perchè più economico e convergente in termini di tecnologia con “rete”
- Non performa come FC e salvo dove non sia possibile fare altrimenti va evitato come la peste
- Chi fa sistemi di storage seri lo sta “mettendo da parte” per via delle difficoltà nella scalabilità

... ma il RAID?

- Molto più “aziendale” che non per uso comune, per qualche strana idea
- Redundant Array of Inexpensive Disk o Redundant Array of Indipendent Disk
- Definizione tecnica: tecnologia di virtualizzazione dello storage che combina differenti dispositivi fisici per formare una o più unità logiche implementando, tra le altre cose, la ridondanza, la velocità o entrambe
- I dati vengono “distribuiti” sui vari dischi fisici allo scopo di ottenere il risultato desiderato.

i livelli del RAID (1->4)

- RAID 0 denominato striping ovvero la distribuzione dei dati sui dischi in modo che il volume logico sia uguale alla somma dei volumi fisici, nessuna ridondanza, velocità elevata e rischio molto elevato di perdere i dati
- RAID 1 denominato mirroring, ovvero i dati vengono distribuiti in modo identico sui disichi, il volume logico sarà replicato quindi su ogni disco fisico, la capacità del volume è espressa come (dim min disco)*n/n (1)
- RAID 2 3 e 4 sono poco utilizzati per mille motivi (storici o pratici di utilizzo)

i livelli del RAID 5 6

- RAID 5 i dischi fisici vengono utilizzati distribuendo sia i dati che la parità. La capacità del volume logico diventa (dim min disco) * (n-1)
  In caso di fallimento di un disco non succede nulla, al secondo disco fallito si perdono i dati
  Minimo tre dischi, ideale da 5 in su
- RAID 6 i dischi fisici vengono utilizzati distribuendo sia i dati che la parità che raddoppia. La capacità del volume logico diventa (dim min disco) * (n-2)
  In caso di fallimento fino a due dischi non succede nulla, al terzo disco fallito si perdono i dati
  Minimo quattro dischi, ideale da 7 in su

i livelli del RAID 10 50 60

- I volumi logici in RAID 1 5 e 6 possono essere combinati utilizzando lo striping, per ottenere dischi di grandi dimensioni e ottimizzare le prestazioni
- Lo striping è e resta una pratica rischiosa
- Numero massimo di dischi? dipende dalla velocità del dispositivo che gestisce il RAID, potrebbe essere via software (occhio alla raggiungibilità di TANTI dischi) oppure potrebbe essere un controller RAID hardware, magari in SAS, o potrebbe essere FC, con numero massimo di dischi anche considerevole (>100) a seconda della tecnologia

RAID e SSD

- In generale utilizzando il RAID non si riesce facilmente a ricondurre il TRIM a livello di filesystem verso i dischi fisici
- si può fare con dispositivi aziendali, occhio alle caratteristiche degli SSD, non sono tutti uguali
  - read intensive
  - mixed use
  - write intensive
- In base al tipo di disco le prestazioni del RAID rischiano di essere da disastrose a assurde, oltre ad impattare sulla affidabilità

velocità e latenza

- breve riassunto, velocità e latenza impattano sulle performance di qualsiasi sistema di trasmissione
- i dischi introducono una loro latenza e anche il sistema operativo fa la sua parte
- quindi senza accorgimenti particolari anche al nascere di standard stratosferici (vedi USB 4 o Thunderbolt) il rischio è di non capitalizzare le velocità di punta dei dispositivi di storage utilizzati (SSD ovviamente)

nuova rubrica: C'è ma non lo sai

- Esistono delle aree di tecnologia largamente utilizzate che in realtà non sono per niente note ai più
- Potrebbe essere che sai che esiste, ma perchè esiste e quali sono le basi teoriche, quello magari no
- Lo sai, bene non spoilerare al vicino :-) grazie

C'è ma non lo sai: NCQ

- Native Command Queuing, che tradotto potrebbe essere una cosa tipo accodamento di comandi nativi
- E' un processo in cui il disco può riordinare comandi eccezionali per ridurre l'uso della meccanica ed migliorare la latenza delle operazioni di lettura/scrittura
- Sia il disco che la scheda madre o il controller PCI debbono supportare questa funzionalità
- Il software deve essere minimo multithread

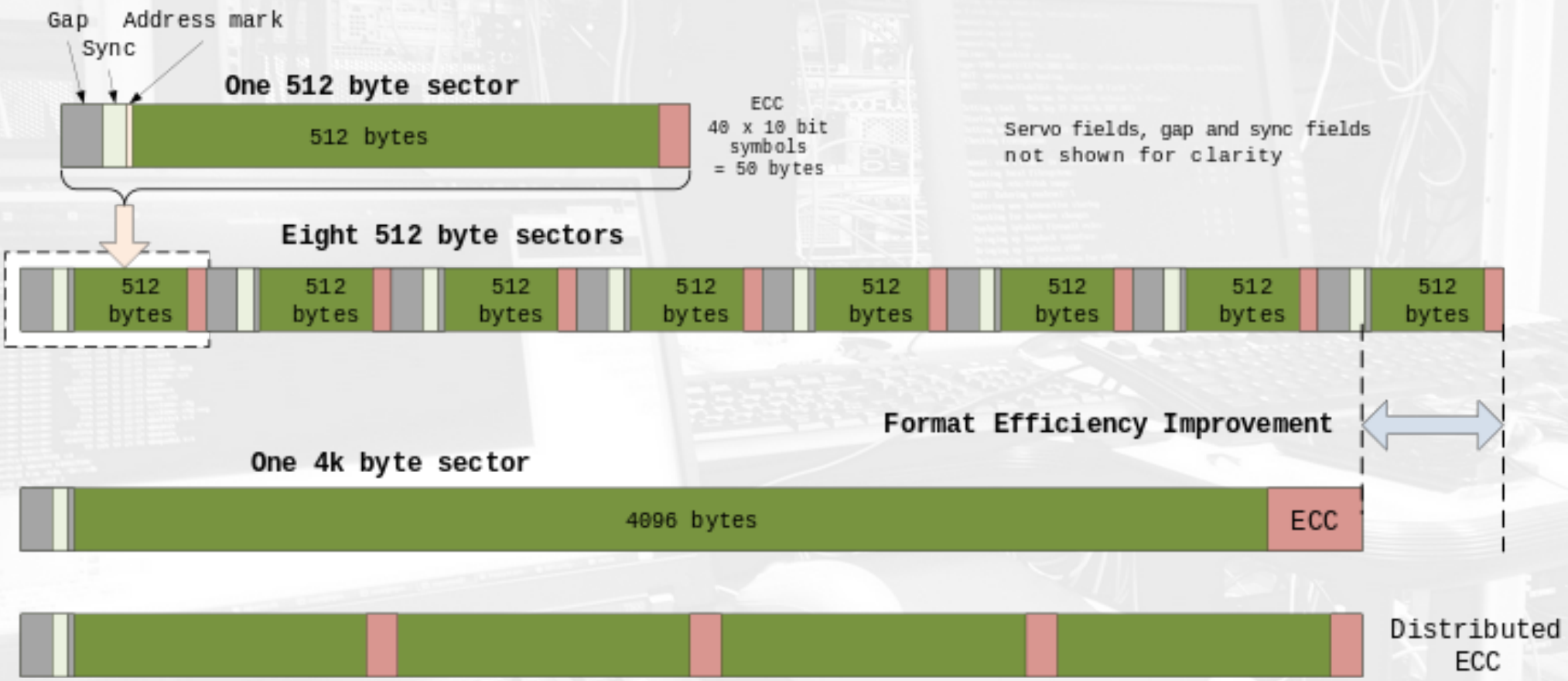
C'è ma non lo sai: NCQ

- Il sistema operativo presenta al controller i dati da leggere e scrivere, il controller del disco può dialogare con il disco dicendo “guarda queste operazioni falle tutte insieme” ma queste le puoi fare dopo “tutte insieme” e via così.
- In questo modo ogni processo può scrivere i propri dati in modo coerente senza forzare gli spostamenti meccanici del disco
- Sarà la logica di controllo del disco a decidere come ottimizzare gli spostamenti
- **Il sistema operativo è “molto probabile” che non conosca la collocazione fisica reale dei dati**

C'è ma non lo sai: Advanced Format (AF)

- Con il crescere delle dimensioni dei dischi è nata la necessità di mettere in ogni settore più di 512 byte, nasce Advanced Format, i normali 512 e 520 o 528 diventano 4096 4112 4160 e 4224
- I valori superiori a 512 e 4096 servono per memorizzare su dischi Advanced Format Drive (AFD) anche dati relativi alla correzione degli errori per mantenere coerenza dei dati al crescere della densità di scrittura

C'è ma non lo sai: Advanced Format (AF)



C'è ma non lo sai: Advanced Format (AF)

- Supporto da parte dei sistemi MA (importante) solo per dischi 512 e 4096, non i valori con AFD
- GNU/Linux dal 2.6.31 (circa 2010)
- Windows 8 e Windows Server 2012
- OS/X 10.8.2

Ah... nulla a che vedere con l'allocazione minima da parte del sistema operativo

C'è ma non lo sai: il TRIM

- Con gli SSD i termini “Garbage Collection” e “TRIM” diventano di uso comune
- Il Garbage Collect è di derivazione “linguaggi di programmazione”, è un operazione di ottimizzazione in cui spazi di “memoria” non più in uso vengono liberati e lo spazio in generale riorganizzato per fare spazio a nuovi dati
- Il TRIM invece è un comando introdotto nello standard dei dischi per dire “ok quel blocco ora non lo uso più”, in pratica si cerca di comunicare ai dispositivi a blocchi sottostanti la disponibilità di un blocco per l'ottimizzazione ovvero il Garbage Collect.

C'è ma non lo sai: il TRIM

- Il TRIM deve essere supportato dal sistema operativo e dai dispositivi fisici, anche dai controller RAID... ad esempio
- GNU/Linux dal (2.6.28 su flash) 2.6.33 su ATA
- OS/X dal 10.6.8-23 (ma solo su SSD Apple) 10.10.4 su qualsiasi SSD
- Windows dalla versione 7 e Windows server 2008r2 in poi

C'è ma non lo sai: il TRIM

- Ma se ho un sistema operativo che non supporta TRIM?
- Allo stato attuale esistono SSD in cui il TRIM non è più indispensabile per fare operare in modo veloce il disco, era un esigenza specifica dei primi SSD in cui non era possibile riscrivere un singolo blocco ma andava riscritta una quantità di dati maggiore, provocando un rallentamento generale
- Nel caso peggiore il disco SSD degrada di prestazioni fino a valori “worst case” spesso dichiarati in specifica

C'è ma non lo sai: il TRIM

- Il commando TRIM dovrebbe non essere utilizzato su dischi crittografati perchè rende comprensibile la struttura del disco
- L'uso “non corretto” di TRIM provoca spesso problemi, i dischi SSD possono piantarsi e corrompere i dati
- Anche sotto GNU/Linux esistono whitelist e blacklist di funzionalità TRIM per evitare comportamenti irresponsabili dei device fisici
- Al solito, fatto lo standard ognuno lo ha implementato a modo suo creando “gli standard” ovvero il caos

il futuro?

- Abbiamo assistito ad un evoluzione notevole in campo CPU/RAM per qualche strano motivo mancano gli stessi risultati in ambito storage
- RAM DDR → 100MT → 800MB/sec sono 1Gbit
- RAM DDR5 → 6400MT → 51200MB/sec sono 50Gbit al secondo

- Hard Disk meccanici circa 80/100MB/sec
- SSD attuali siamo a 1600MB/sec

il futuro?

- Abbiamo assistito ad un evoluzione notevole in campo CPU/RAM per qualche strano motivo mancano gli stessi risultati in ambito storage
- RAM DDR → 100MT → 800MB/sec [illegible] ono 1Gb[illegible]
- RAM DDR5 → 6400MT → 51200M[illegible]ec [illegible] o [illegible]5[illegible]bit [illegible] secondo

64X

- Hard Disk meccanici circa 80/100MB[illegible]c
- SSD attuali siamo a 600MB/sec

6X

il futuro?

- La dimensione dei dati sta crescendo in modo evidente, video, audio, e soprattutto oggetti come l'AI
- Ma per AI serve recuperare questo GAP, anzi, serve riuscire ad avvicinare la capacità delle CPU attuali che sta arrivando a valori anche sopra i 40GB al secondo
- E' facile capire che 0,6GB al secondo sono pochi

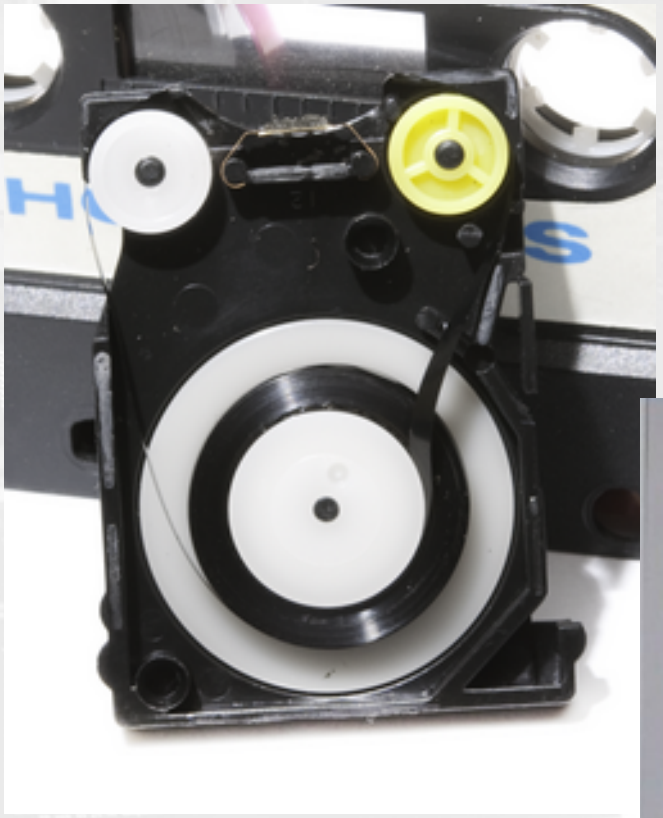
il futuro? il problema dei consumi!

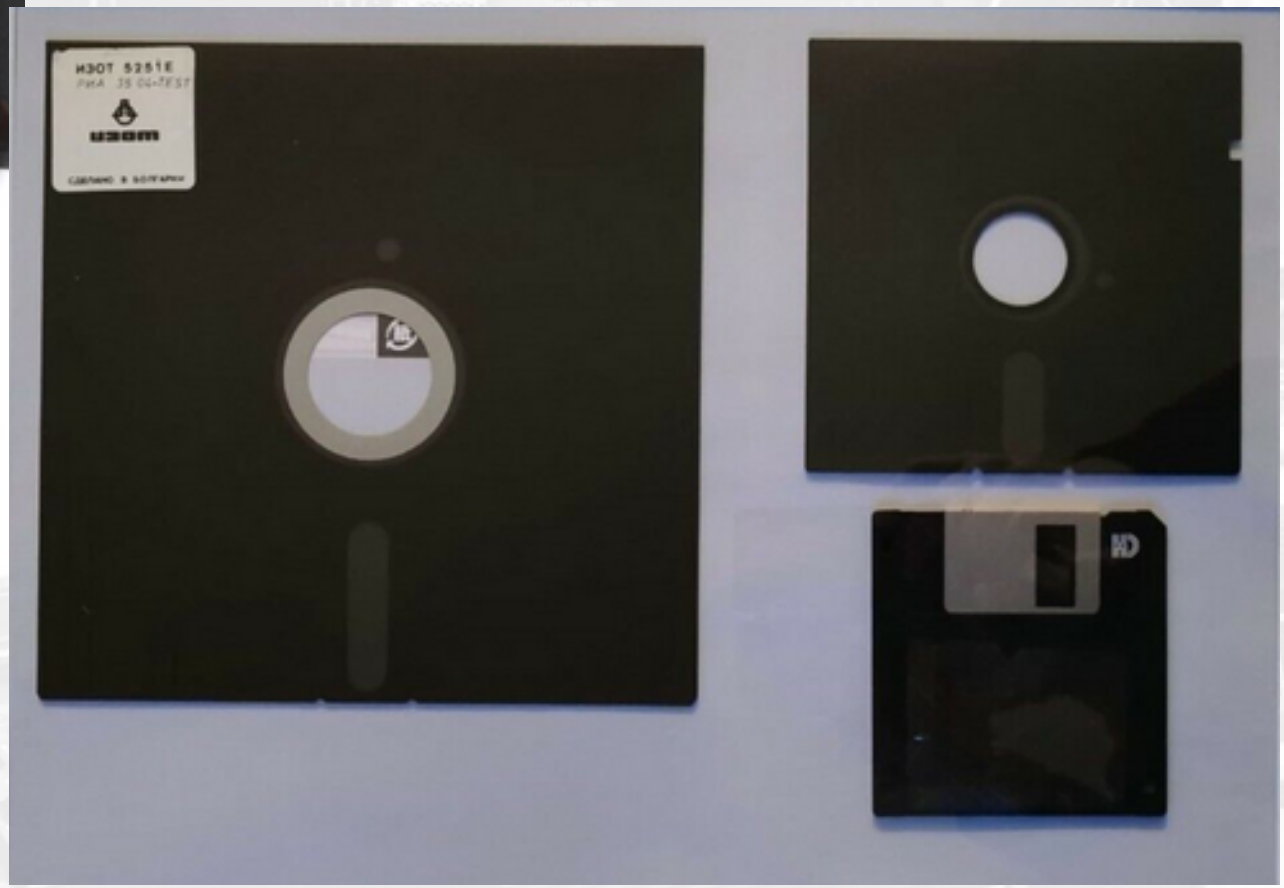
- Un hard disk meccanico da 22TB consuma circa 10W a pieno carico, 6 in stand/by con un indice di consumo medio di 0,6W per TB
- Un SSD da 30TB consuma circa 14W a pieno carico, 5 in stand/by con un indice di consumo pari a 0,53W per TB
- In pratica uno dei cavalli di battaglia degli SSD è ancora ben lontano dall'essere realizzato

  i dischi meccanici consumano meno e sono più longevi rispetto ai dischi SSD

il futuro? esiste anche il problema della durata nel tempo!

- Ogni supporto ha un suo dato specifico di “affidabilità” nel tempo
- Abbiamo esempi di scrittura che si sono conservati in condizioni proibitive per secoli o millenni
- Quanto durano sui nostri dispositivi attuali?
- Poco, gli SSD non alimentati 5 anni, forse 10, forse 20
- Hard disk e altri supporti arrivano a 10 anni forse 20
- Ed esiste sempre il problema: se prendo in mano un supporto così, come lo leggo?




ИЗОТ 5251Е
HD

DOMANDE?

per contattarmi:

maxnuv@linux.it