ISTITUTO NAZIONALE DI FISICA NUCLEARE

COMMISSIONE CALCOLO E RETI

INFN/20-04/CCR
28 aprile 2020

CCR-58/2020/P

ANALISI DELLA EVOLUZIONE DELLE TECNOLOGIE HARDWARE PER IL CALCOLO SCIENTIFICO

Alessandro Brunengo[1], Gianpaolo Carlino[2], Andrea Chierici[3], Luca Dell'Agnello[3], Alessandro De Salvo[4], Sergio Fantinel[5], Gaetano Maron[3], Enrico Mazzoni[6], Michele Michelotto[7], Vladimir Sapunenko[3], Stefano Zani[3]

[1)]INFN-Sezione di Genova, Via Dodecaneso, 33, I-16146 Genova
[2)]INFN-Sezione di Napoli, Via Cintia, I-80126 Napoli
[3)]INFN-CNAF, Viale Berti Pichat 6/2, I-40127 Bologna
[4)]INFN-Sezione di Roma, Piazzale Aldo Moro 2, I-00185 Roma
[5)]INFN-Lab. Naz. di Legnaro, Viale dell'Università 2, I-35020 Legnaro (PD)
[6)]INFN-Sezione di Pisa, Largo B. Pontecorvo 3, I-56127 Pisa
[7)]INFN-Sezione di Padova, Via Marzolo 8, I-35131 Padova

Abstract

Gli esperimenti della fisica HEP avranno bisogno, nel prossimo decennio, di strumenti di calcolo di enorme potenza, che le tecnologie attuali non potranno soddisfare.

Questo lavoro è un'analisi dell'evoluzione delle tecnologie di CPU, di storage e di rete, volta a valutare alcuni degli indicatori (prestazioni, consumi) su cui poter basare la progettazione di un centro di calcolo valida per i prossimi 10 anni.

Previsioni sull'evoluzione tecnologica di tale durata sono facilmente soggette ad errori di valutazione anche macroscopici: roadmap valide solo per tre-quattro anni, inaspettati balzi tecnologici, imprevedibili mutamenti di andamento del mercato, possono facilmente cambiare o anche ribaltare previsioni così a lungo protratte nel futuro, ed invalidare estrapolazioni per quanto accurate. Ciononostante, un'analisi della situazione e di quello che oggi può essere detto in relazione all'evoluzione di queste tecnologie è un punto di partenza senza il quale non sarebbe possibile fare alcun tipo di ipotesi.

Pubblicato da
Laboratori Nazionali di Frascati

1 INTRODUZIONE

Il calcolo è entrato in un'epoca in cui è largamente data-driven, e questo è vero non solo per LHC ma anche per esperimenti in altri ambiti di interesse per l'INFN, quali ad esempio la fisica astroparticellare e delle onde gravitazionali. È probabile che l'evoluzione tecnologica faticherà a soddisfare le future esigenze di calcolo per via del fatto che le leggi che hanno regolato l'evoluzione dell'elettronica per oltre tre decadi, la legge di Moore e la legge nota come Dennard Scaling, appaiono aver raggiunto limiti di validità (si veda [1], [2], [3]).

Queste problematiche sono state analizzate approfonditamente dalla comunità HEP, in un'analisi prodotta dalla HSF[1] che mostra quanto siano critiche e complesse le sfide necessarie per soddisfare le esigenze degli esperimenti HEP nel prossimo decennio ([4]).

In questo contesto il Comitato Coordinamento Calcolo Scientifico (C3S) del INFN, che ha il compito di formulare proposte di ricerca e sviluppo sul calcolo scientifico e le infrastrutture correlate, ha creato alla fine di aprile 2018 il gruppo di lavoro Technology Tracking (TT-C3S) con il mandato di studiare le soluzioni tecnologiche previste entro un arco temporale di dieci anni e dunque stimare la potenza di calcolo dei futuri processori, la densità di storage, sia disco che nastro, e la capacità della rete; il lavoro del gruppo, condotto in stretta collaborazione con realtà internazionali che già stanno svolgendo questa attività al CERN e in ambito HEPIX, è necessario per rispondere alle esigenze del corretto dimensionamento dei siti di calcolo dell'INFN e in particolare del nuovo sito del Tier1 previsto al Tecnopolo di Bologna.

Il lavoro è partito dall'analisi proposta nella presentazione fatta al Joint WLCG & HSF Workshop di marzo 2018 svoltosi a Napoli ([5]) dal gruppo coordinato dal CTO della divisione IT del CERN, ed è proseguito con la pianificazione di incontri e l'individuazione di *maker* e di *vendor* fra i più importanti nel panorama tecnologico mondiale per ottenere informazioni autorevoli attinenti al mandato del gruppo.

Tra maggio e giugno 2018 sono stati così ascoltati sul tema dell'evoluzione di CPU/GPU, Storage, Tape e Networking, i *chip-maker* AMD, ARM, Intel, Nvidia, Western Digital e i *vendor* DDN, Dell/EMC che hanno proposto, richiedendo specifiche sottoscrizioni NDA, la loro visione del futuro; importante per attrarre l'attenzione dei player menzionati e anche nell'organizzazione di alcuni incontri, la collaborazione tenuta con il CINECA di Bologna attraverso il responsabile della divisione Production Services. È parso subito chiaro che l'orizzonte temporale di previsione di dieci anni è estremamente complesso dato che anche i *chip-maker* fanno fatica a prevedere oltre i 5 anni e comunque, tra i più avventurosi, non si va oltre agli 8 anni (2026); tra i *vendor* ovviamente la situazione è ancor peggiore dato che il limite di previsione si attesta sui 3/5 anni al massimo.

Nei successivi paragrafi viene proposta l'analisi operata dal gruppo interpolando le informazioni raccolte dai vari canali.

[1] Hep Software Foundation: Fondazione volta a facilitare il coordinamento e gli sforzi comuni nelle attività di software e computing sviluppate in ambito High Energy Physics.

2 CPU E SERVER

2.1 Caratteristiche dei server di calcolo

I nodi di calcolo utilizzati in ambito HEP da anni sono nella forma di schede madri a doppio socket. Le schede a 4 o 8 socket sono troppo costose perché hanno bisogno di processori più evoluti e dotati di un maggior numero di interconnessioni, e sono usate solitamente in supercomputer o quando si desidera un nodo di calcolo con un numero elevatissimo di core, alla ricerca di latenze molto basse difficili da ottenere uscendo dal board.

I fattori di forma che all'inizio prevedevano un board in un enclosure 1U si sono evoluti verso sistemi twin (due board in un unico enclosure) o doppio twin (quattro board in un enclosure 2U) raddoppiando quindi la densità, oppure in soluzioni di tipo blade che permettono gestione semplificata, connessioni di rete su backplane e alimentazione ridondata N+1 invece che 2N.

Il set di istruzioni è quasi ovunque x86 a 64 bit (x86_64) dopo la migrazione della tecnologia da 32 a 64 bit avvenuta nel 2010. I produttori che sostengono questa tecnologia sul mercato sono Intel, che ha goduto di una posizione dominante, e AMD, meno competitiva, ma che si è da poco rilanciata grazie alla linea EPYC, molto promettente. Per un certo periodo c'è stato interesse per ARM ma gli sviluppi a 64 bit sono arrivati molto tardi e attualmente non sembrano esserci risorse sufficienti per effettuare una traduzione del codice degli esperimenti verso questa architettura. Rimane tuttavia un mercato interessante perché a livello globale ARM domina sia come numeri, grazie alla diffusione di smartphone e tablet, sia come risparmio energetico, a parità di prestazioni. Un'altra architettura ancora attiva è "POWER" spinta da IBM, con prestazioni molto buone ma con un rapporto prezzo prestazioni peggiore di x86_64.

2.2 Intel

2.2.1 Storia dei processi produttivi di Intel

Nel 2007 Intel ha adottato un modello chiamato "Tick-Tock" [6], secondo cui ogni cambiamento della microarchitettura viene seguito da un nuovo processo tecnologico di riduzione del die. Quindi dal processo a 65nm si è passati nella fase tick ai 45nm, poi 32nm, 22nm fino all'attuale 14nm (a volte eccezionalmente sono state inserite nuove istruzioni nella fase di Tick). Mentre nella fase di Tock si è passati da Core a Nehalem, quindi Sandy Bridge, Haswell ed ora Skylake (vedi Fig. 1).

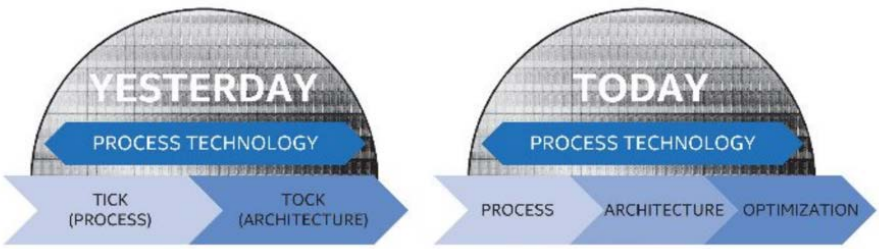


Fig. 1: Modelli evolutivi dei processori Intel

Tuttavia, difficoltà incontrate nel migliorare il processo produttivo hanno portato Intel a rivedere questo modello, divenuto insostenibile.

Questo cambiamento è stato formalizzato nel Marzo 2016, quando Intel ha chiarito di voler sostituire il modello Tick-Tock con un modello in tre passi "Process-Architecture-

Optimization" (PAO, vedi Fig. 1) secondo il quale tre generazioni di processori saranno prodotti in un singolo processo produttivo, dove la terza generazione viene dedicata all'ottimizzazione [7]

- La prima fase, chiamata *Process* (che precedentemente veniva chiamata "Tick"), consiste nell'introduzione di un nuovo processo produttivo abbinato ad un'architettura già matura, in modo da procedere all'aggiornamento dei prodotti correnti al nuovo processo produttivo, allo scopo di massimizzare la resa della futura generazione di processori.
- La fase "Architecture" è quella in cui, sfruttando il nuovo processo produttivo viene introdotta una nuova microarchitettura e tipicamente qui abbiamo nuove funzionalità, cache maggiori e nuove istruzioni.

Dopo aver introdotto una nuova architettura, l'anno successivo si procede a una sua revisione e ottimizzazione, allo scopo di sfruttarne al massimo tutte le potenzialità. In pratica, unendo affinamenti nel layout dei componenti, nelle logiche di funzionamento interno alla CPU e nella produzione dei transistor con uno specifico processo produttivo, si ottengono dei microprocessori ancora più ottimizzati che in genere dovrebbero essere in grado di migliorare l'efficienza di circa il 5-10%

2.2.2 Sunny Cove

A fine 2018, Intel ha presentato una nuova roadmap in fatto di architetture x86 e annunciato la microarchitettura "Sunny Cove" (vedi Fig. 2) nel corso dell'Architecture Day tenuto a Santa Clara. Come noto, il dominio dell'azienda sul mercato dei chip si è sempre basato sulla duplice leadership sul fronte delle architetture e dei processi produttivi.

In poche parole, con il progredire delle architetture si passava a processi produttivi più avanzati, capaci di migliorare ulteriormente il quadro generale, soprattutto i consumi. Quell'approccio ha mostrato i suoi limiti quando Intel ha incontrato evidenti problemi di messa a punto del suo processo produttivo a 10 nanometri.

Anziché portare sul mercato nuove architetture, l'azienda è stata costretta ad affidarsi ai 14 nanometri per ben quattro anni di fila, rifinendo costantemente il processo produttivo tramite versioni indicate da un "+". Ogni nuova versione dei 14 nanometri ha permesso a Intel di aumentare le frequenze e così le prestazioni, passando da 4,2 a 5,1 GHz. Questi miglioramenti hanno garantito fino al 70% di prestazioni in più dal debutto dei 14 nanometri nel 2014 a oggi, ma l'assenza di una nuova microarchitettura, che solitamente migliora il throughput IPC (istruzioni per ciclo di clock) del processore, ha rallentato i progressi.

Fig. 2: Architettura Sunny Cove e sua evoluzione

Intel ha iniziato a progettare nuove architetture che possano essere realizzate con più

processi produttivi. Questo le consentirà comunque di offrire soluzioni più avanzate anche se incontrerà ostacoli nel cammino volto a un costante miniaturizzazione dei transistor.

Sunny Cove è la prima microarchitettura che può essere usata su diversi processi produttivi, e anche se Intel ha dichiarato che il nuovo core debutterà con i 10 nanometri, non è certo che arriverà con i chip Ice Lake (la decima generazione di processori Intel, per rimpiazzare Sky Lake). In linea con le nuove specifiche di progettazione, Intel selezionerà differenti processi produttivi per prodotti diversi in base alle necessità del segmento. È un approccio simile a quello dei produttori di terze parti TSMC e GlobalFoundries e significa che Intel può scegliere di usare Sunny Cove anche per processori a 14 nanometri.

Sunny Cove debutterà nel 2019, offrendo maggiori prestazioni nei compiti single-thread, un nuovo "Instruction Set Architecture" (ISA) e un progetto pensato per la scalabilità. Willow-Cove seguirà con una migliorata gerarchia della cache, funzionalità di sicurezza e ottimizzazioni ai transistor. La microarchitettura Golden Cove debutterà nel 2021 garantendo ancora più prestazioni single-thread, prestazioni con carichi di intelligenza artificiale, miglioramenti legati al networking e nuove funzionalità di sicurezza. Intel ha pronti miglioramenti prestazionali generali, ma vuole compiere progressi anche in quei casi d'uso specializzati come AI, crittografia e carichi di compressione e decompressione.

La nuova microarchitettura Sunny Cove vede miglioramenti a ogni livello della pipeline. I passi avanti chiave al front end includono buffer più ampi per reorder, load e store, insieme a *reservation station*[2] più capienti. Ciò consente al processore di esaminare più in profondità l'insieme delle istruzioni in entrata per trovare operazioni indipendenti tra loro e che possono essere eseguite simultaneamente. Le operazioni sono poi eseguite in parallelo per migliorare l'IPC. Intel ha aumentato la cache dati L1 da 32 KiB a 48 KiB (+50%). Anche la cache L2 è più grande, ma la capacità è legata a ogni specifico tipo di prodotto – se il chip è destinato ai desktop oppure ai server, insomma. Intel ha inoltre espanso la micro-op cache (uop) e il *translation lookaside buffer*[3] (TLB) di secondo livello. Un aspetto chiave per migliorare le prestazioni è aumentare il parallelismo. Ciò inizia con il buffer più profondo e le reservation station già descritte, ma richiede anche più execution unit per elaborare le operazioni. Intel ha aumentato anche la quantità di memoria che il processore può indirizzare, un aspetto fondamentale dato il suo obiettivo di aumentare la capacità di memoria con le DIMM Optane DC Persistent Memory. I veloci moduli di memoria Optane forniscono fino a 512 GB di memoria indirizzabile per DIMM, il che significa che la capacità di memoria "esploderà" non appena un maggior numero di datacenter adotterà la tecnologia.

Sunny Cove passa a una struttura di paging a 5 livelli, dai quattro delle architetture precedenti. Questo aumenta lo spazio d'indirizzamento virtuale fino a 57 bit e lo spazio d'indirizzamento fisico fino a 52 bit, il che significa che supporta fino a 4 petabyte di memoria, rispetto ai 64 TB di Skylake.

2.2.3 Il nuovo corso di Intel

La nuova visione di Intel di disaccoppiare il progetto delle architetture dai processi produttivi è un passo avanti concreto che migliorerà la competitività dell'azienda in futuro. Di certo Intel non può permettersi un altro periodo di stagnazione come quello che abbiamo visto – e vediamo tuttora – nel passaggio dai 14 ai 10 nanometri. I produttori di terze parti si sono dimostrati concorrenti formidabili per Intel. TSMC sta producendo a 7 nanometri, con il

---

[2] Si veda "https://en.wikipedia.org/wiki/Reservation_station"

[3] Si veda "https://en.wikipedia.org/wiki/Translation_lookaside_buffer"

risultato che AMD, Apple, Qualcomm e Nvidia hanno o avranno presto chip a 7 nanometri.

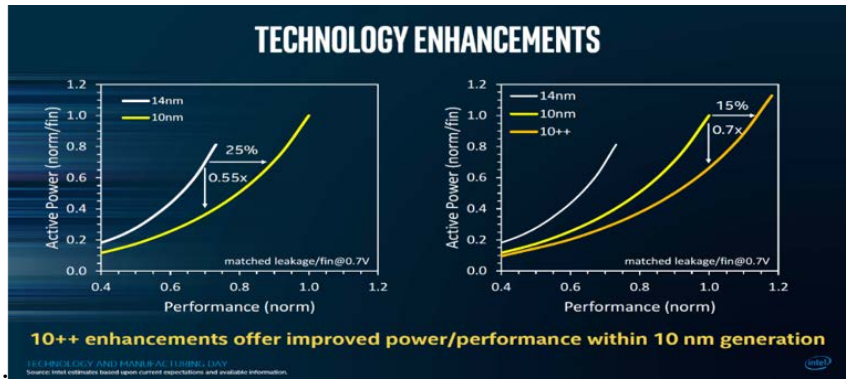


Fig. 3: Evoluzione del processo costruttivo e prestazioni

Queste aziende lavorano insieme a TSMC per portare i loro progetti sul mercato, il che significa che Intel non è in competizione con una sola azienda, ma con diversi colossi nella progettazione di chip. Intel ha intenzione di sconfiggere i suoi rivali sfruttando la sua ampia gamma di prodotti, ma per farlo ha bisogno anche di una parte software molto robusta. Per questo sta lavorando su "One API", progettata per semplificare la programmazione per acceleratori di AI, GPU, CPU e FPGA. Il software segue la filosofia del "nessun transistor deve essere lasciato indietro" e, dato l'obiettivo, è fondamentale.

Il nuovo software metterà a disposizione librerie unificate che consentiranno alle applicazioni di spostarsi senza problemi tra i diversi tipi di hardware di casa Intel. In caso di successo, potrebbe essere un fattore chiave di differenziazione che altre aziende non saranno in grado di eguagliare, almeno non in maniera così vasta.

2.2.4 Intel calata nella nostra realtà

Escludendo le ben note migliorie atte ad aumentare il numero di core per processore, ottimizzando contestualmente i consumi, le novità più significative di nostro interesse dovrebbero venire dall'avvicinamento alla CPU di sottosistemi attualmente periferici, al fine di ottimizzare le prestazioni. Andiamo ad analizzare alcune di queste soluzioni considerando come piattaforma di riferimento "Cascade Lake" (fino a 28 core fisici, seguita nel 2020 da "Cooper Lake"). Ai 6 canali di memoria per Cascade Lake, si dovrebbe aggiungere il supporto alla memoria persistente, denominata "Apache Pass". Con questa soluzione si vuole creare una via di mezzo tra memoria RAM e memoria di massa (con velocità paragonabile ad un buon SSD). I casi d'uso più immediati sono accessi a grandi data base, che possono essere mantenuti in memoria, velocizzando terribilmente l'accesso, senza rischiare alcuna perdita di dati in caso di interruzione di corrente. I tagli previsti inizialmente sono fino a 512GB, compatibili con piedinature DDR4 e avranno la cifratura hardware per prevenire furti di dati. Risulta molto difficile pensare ad un utilizzo di massa nel nostro ambiente, data la peculiarità della soluzione. Non si nasconde però che risulti una soluzione molto interessante.

Altra novità prevista riguarda l'inclusione di un modulo FPGA all'interno del socket del processore: l'obiettivo prestazionale di Intel è di arrivare ad un rapporto Pflop/watt superiore di 5 volte agli attuali Xeon. Tale modulo, generalmente molto esigente in termini di consumi, potrebbe anche venire spento in caso di inutilizzo.

Allo stesso modo del modulo FPGA, sarà integrato all'interno del processore, per una

specifica SKU, il controller OmniPath.

La soluzione più innovativa e interessante per la nostra realtà potrebbe realizzarsi nel server "Cascade Lake AP". Si tratta di una CPU in tecnologia BGA (quindi saldata direttamente sulla scheda madre) che dovrebbe contenere fino a 48 core per socket, consumando circa 350W. In questo caso non viene venduta la sola CPU, ma il sistema integrato. È in fase di sviluppo uno chassis proprietario, che viene proposto principalmente con tecnologia di raffreddamento a liquido. La dimensione prevista si attesta su 2U, con al massimo 4 unità indipendenti. Su ognuna di queste unità, che ospita 2 socket, si verrebbero ad avere 96 core, per un totale di 384 core in sole due rack unit!

Tutta la tecnologia realizzata da Intel richiede una specifica programmazione e per questo motivo da anni viene sviluppato Intel Parallel Studio, giunto alla versione xe2020. Intel riesce a indirizzare la definizione di nuovi standard così che i programmatori siano rapidamente in grado di sfruttare le istruzioni proprietarie. L'ultima versione implementa quasi tutto lo standard c++17 e completamente la versione c++14, openMP5 e OpenMP4.5. Per quanto riguarda le librerie standard si fa affidamento a gcc, che è un requisito all'installazione di Parallel Studio. Le specificità per tipologia di CPU possono creare diversi problemi alle nostre computing farm, poiché generalmente i job vengono mandati su un nodo, ignorandone il tipo di architettura. Questo non permette di sfruttare le ottimizzazioni introdotte da Intel, specificatamente per i suoi processori, in numerose librerie matematiche, tra le quali ad esempio Math Kernel Library o numpy, che sono dichiarate essere molto più veloci rispetto alle versioni non ottimizzate. È auspicabile la creazione di un livello intermedio in grado di astrarre le specificità dell'hardware su cui girano i job, senza tuttavia perdere, anzi abilitando, le ottimizzazioni realizzate dai produttori di chip.

2.3 AMD

AMD è tornata con forza nel mercato server abbandonato da qualche anno con una soluzione tecnologicamente avanzata, in grado di competere finalmente alla pari con le soluzioni Intel. Attualmente la soluzione per data center, denominata EPYC (core Naples), è presente anche in tutte le sedi INFN che hanno partecipato all'ultima gara CPU coordinata dal Tier1. Le caratteristiche peculiari sono l'elevato numero di core (fino a 32 core "Zen"), 8 canali di memoria e fino a 128 linee I/O. Grazie all'elevato numero di linee I/O, AMD vanta la peculiarità di richiedere una sola CPU per comandare numerose schede GPU (vedi Fig. 4), quando al contrario, Intel richiede sistemi bi-processori per poter garantire banda passante sufficiente: purtroppo ad oggi, come citato in altra parte di questo report (si veda § 3.3), la soluzione GPU di AMD appare molto meno competitiva.

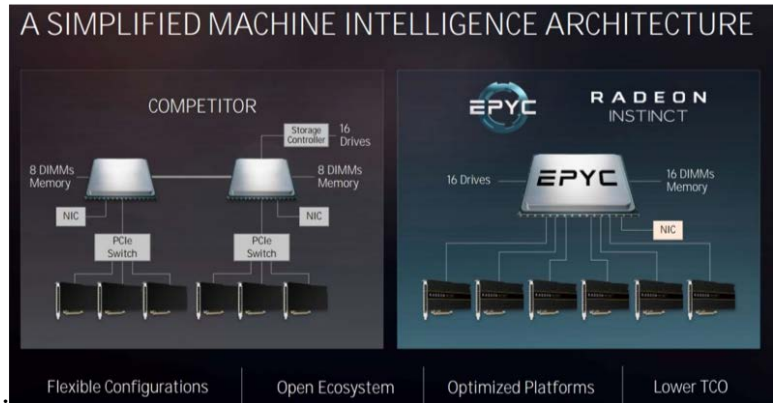


Fig. 4: Architettura EPYC

Come sempre le evoluzioni (vedi Fig. 5) prevedono, per le prossime iterazioni una riduzione del processo produttivo (si passa da 14nm a 7nm con il processore Rome), con la possibilità di vedere raddoppiati i core presenti su singolo processore, arrivando a fornire fino al 25% di prestazioni in più se paragonato al processore attuale a parità di core. Potrebbe anche venire introdotta la memoria DDR5 per la prima volta in una soluzione desktop, poiché questa tecnologia attualmente è appannaggio unicamente delle GPU. Questa CPU dovrebbe anche essere in grado di elaborare una istruzione AVX[4] per ciclo di clock, contro gli attuali due richiesti dalla CPU modello Naples. La successiva evoluzione, denominata Milan, prevede un doppio numero di thread per core rispetto alle soluzioni attuali: si passerà quindi da 2 a 4 thread per singolo core.

Fig. 5: Roadmap per il processore AMD

Sfruttare l'elevato numero di core e le nuove istruzioni ottimizzate per il calcolo vettoriale, così come per il caso di Intel, richiede un grosso sforzo di programmazione, oltre ad un insieme di compilatori in grado di abilitare tali caratteristiche. AMD ha un approccio molto aperto, contrapposto per certi aspetti a quello di Intel, poiché non fornisce strumenti specifici (spesso a pagamento) bensì collabora con il mondo open source al fine di includere le ottimizzazioni in compilatori ben conosciuti nel nostro ambiente (quali ad esempio *gcc*). Parimenti, anche le librerie matematiche seguono la stessa strategia: tra le più comuni possiamo citare *lapack, blas, fft*.

2.4 ARM

Diverse soluzioni stanno nascendo basate su tecnologia ARM con l'intento di entrare nei data center di tutto il mondo. Al momento le soluzioni sono molto poco diffuse ma vale sicuramente la pena riassumere le soluzioni più promettenti da tenere sotto osservazione nei prossimi anni.

[4] Advanced Vector Extension: instruction set sviluppato da Intel che permette l'esecuzione di operazioni intere su più variabili in un singolo ciclo di clock (vettorializzazione). Introdotto nel 2011 da Intel e poi da AMD per operare su vettori a 128 bit (es. 8 Floating Point a singola precisione o 4 FP a DP). Estesa a 256 bit con AVX2 (Haswell) ed ora a AVX512 dopo SkyLake.

2.4.1 Neoverse N1

Il Neoverse N1 di ARM Holding è il primo processore sotto proprietà intellettuale di ARM che si rivolge specificamente al data center. L'N1 è la prima generazione di una microarchitettura che ha due generazioni successive in sviluppo: è specificamente progettato per processi a 7 nm, con le generazioni future destinate a processi più avanzati. Il blocco Neoverse N1 IP è composto da due core ARM N1 e cache L1 e L2 come mostrato in Fig. 6.

Come nella normale pratica di ARM, il core N1 è un "IP block", un mattoncino concesso in licenza, che i licenziatari combinano tra loro con blocchi di altra tipologia, quali memoria, di interconnessione e di I/O al fine di creare sistemi più complessi.

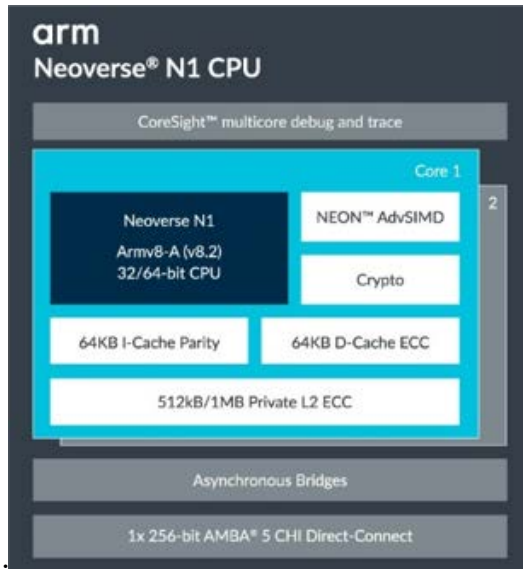


Fig. 6: ARM Neoverse N1

Il progetto di CPU di riferimento, suggerito dalla stessa ARM, per sfruttare i core N1, prevede quanto segue:

- 64 CPU Neoverse N1, ciascuna con 1 MB L2 riservata, che fornisce 64 thread di esecuzione parallela
- Configurazione di interconnessione mesh 8x8 coerente (CMN-600) con 64 MB di cache condivisa a livello di sistema
- Quattro collegamenti CCIX[5] per configurazioni multi-socket, chiplet e configurazioni con acceleratore grafico
- Otto canali di memoria DDR4

2.4.2 ThunderX2

Dei processori basati su ARM specificamente progettati per server di uso generale, Marvell's ThunderX2 attualmente sembra essere il processore che ha ottenuto la maggiore visibilità e diffusione sul mercato. La CPU è disponibile tra gli altri nel Cray XC-50 installato presso il Los Alamos National Laboratory e nel sistema HPE Apollo 70 basato su Astra HPC presso il Sandia National Laboratory. ThunderX2 è un'architettura Arm V8.1 a 64 bit, CPU a

[5] Cache Coherent Interconnect for Accelerators, si veda "https://en.wikichip.org/wiki/ccix"

doppio socket di Marvell. Le specifiche del processore sono le seguenti:

- Arm v8.1
- Fino a 32 core singolo socket
- Core SMT4: fino a 4 per core fisico
- Fino a 8 canali di memoria DDR4 per socket
- Fino a 56 linee PCIe Gen3 con 14 controller
- Processo FinFET a 16 nm.

2.4.3 ARM nel data center

Vari test sono stati eseguiti in ambito HPC/HTC sia al CERN che al Tier1 con CPU ARM (si veda il progetto COSA [8]) ma pur essendo soluzioni interessanti e promettenti, non si è mai raggiunto il livello di prestazioni tale da suggerire la migrazione in massa verso questa tecnologia. Considerando i nodi di calcolo in genere, non appare al momento quindi percorribile alcuna strada che porti verso la tecnologia ARM, sia a causa delle prestazioni, che delle necessarie modifiche al software per poterlo rendere compatibile. Diverso è invece il panorama considerando le soluzioni storage, dove appliance integrate o soluzioni che prevedono l'utilizzo di disk server, potrebbero beneficiare di soluzioni a basso consumo e altamente ottimizzate dal vendor, evitando gli ingenti costi legati solitamente alle CPU Intel.

Si deve comunque, a nostro avviso, continuare a monitorare la situazione, poiché tra le tante iniziative legate al mondo ARM, Apple sembra intenzionata a migrare le proprie soluzioni hardware verso questa tecnologia a breve termine, sfruttando le conoscenze acquisite in ambito mobile con Ipad e Iphone[6]. Essendo da anni un "*trend-setter*", se davvero venisse intrapresa questa strada, la situazione potrebbe cambiare considerevolmente, con una possibile ricaduta anche sul mondo server.

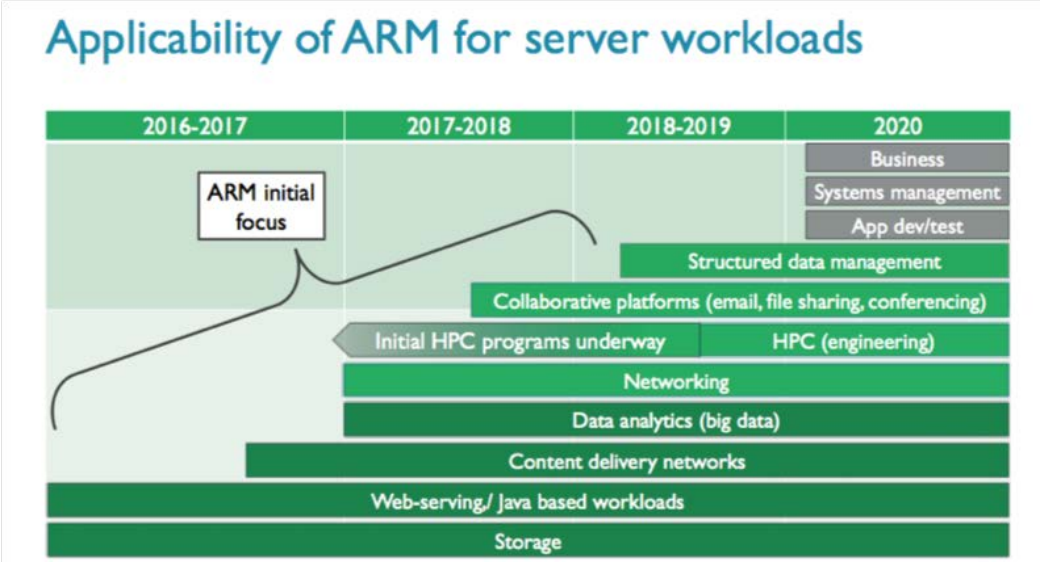


Fig. 7: Roadmap di ARM per il settore Server

L'evoluzione delle CPU ARM si sta orientando ad un migliore supporto delle

[6] Si veda a tal proposito https://www.bloomberg.com/news/articles/2018-04-02/apple-is-said-to-plan-move-from-intel-to-own-mac-chips-from-2020

applicazioni nei data center e nel mondo server in generale (vedi Fig. 7) grazie ad un aumento del numero di core e della profondità delle unità vettoriali, oltre che all'evoluzione dell'insieme di istruzioni a supporto del *Machine Learning*.

2.5 Prestazioni

Un parametro importante ai fini del dimensionamento di un centro è il consumo elettrico dei server, che costituisce la parte principale dei consumi elettrici di un data center. È stata fatta una analisi sui dati raccolti negli ultimi 15 anni in relazione al consumo per HepSpec06, considerando sia i dati del Tier1 del CNAF, sia i dati integrati con quelli di altri centri di calcolo HEP che hanno gentilmente messo a disposizione i dati.

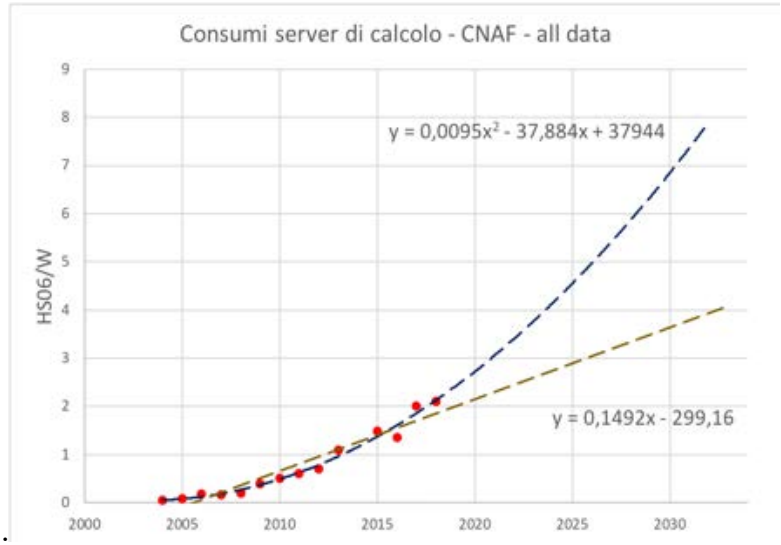


Fig. 8: HepSpec2006 per Watt (dati del CNAF)

In Fig. 8 è mostrato il grafico con tutti i dati disponibili a partire dal 2004. L'immagine mostra due fit, uno lineare ed uno quadratico. Benché il fit quadratico si adatti meglio ai dati, l'estrapolazione a lungo termine rischia di essere un risultato troppo ottimistico, e sembra più ragionevole basarsi su una estrapolazione lineare, o eventualmente considerare il fit lineare come conservativo, il fit quadratico come ottimistico.

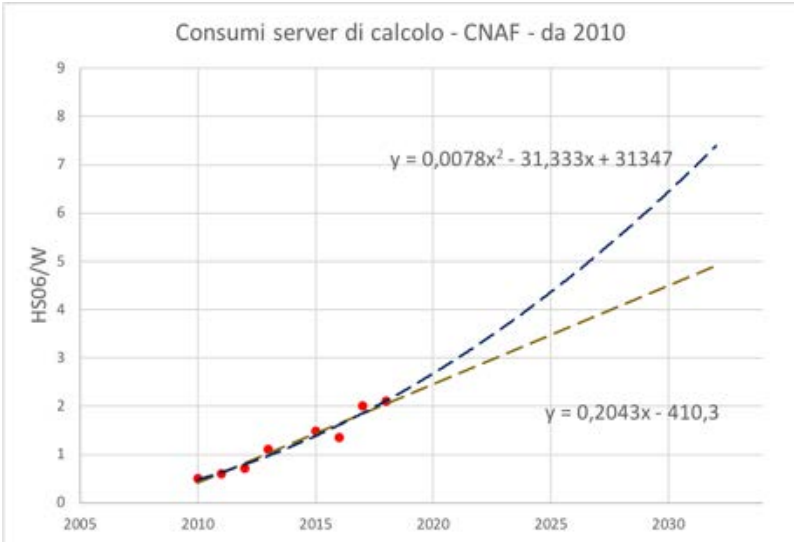


Fig. 9: HepSpec2006 per Watt (CNAF, dati dal 2010)

Il grafico in Fig. 9 mostra la stessa analisi operata sui soli dati presi a partire dal 2010, che comprendono misure su CPU più omogenee dal punto di vista tecnologico (multicore). L'andamento è simile al grafico con i dati completi, con un migliore adattamento del fit lineare.

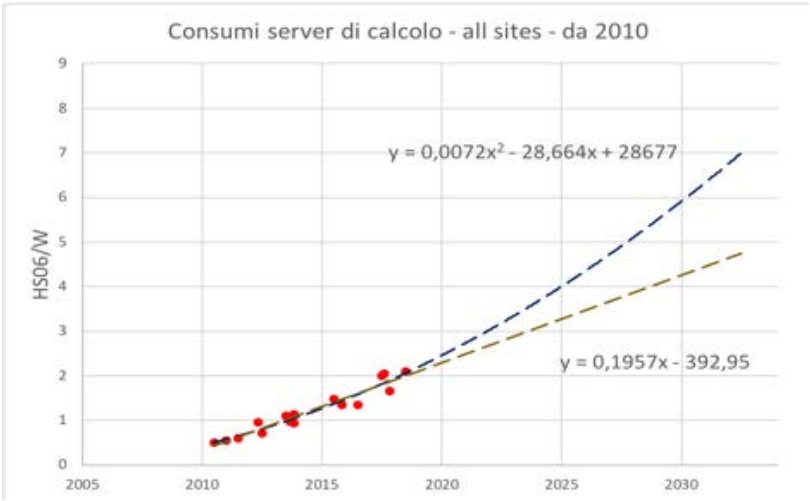


Fig. 10: HepSpec2006 per Watt (tutti i siti)

Per avere conferma sulla affidabilità dell'analisi fatta, abbiamo utilizzato dati sui consumi rilevati da altri centri di calcolo HEP (Gridka, BNL, INFN-PD). I dati accorpati (vedi Fig. 10) mostrano valori compatibili, rafforzando la nostra analisi. Il peggioramento del coefficiente di regressione è dovuto probabilmente alla disomogeneità nelle metodologie di misura di consumi e prestazioni.

In virtù della analisi fatta, è ragionevolmente conservativo attendersi una crescita di HepSpec06/W dell'ordine do 0.2 HS06/W per anno.

Va tuttavia considerato che questa estrapolazione non può tenere conto di evoluzioni tecnologiche nella architettura dei processori, che potrebbero invalidare alcune previsioni.

In base alle informazioni sulle roadmap dei processori oggi disponibili è possibile considerare questa analisi ragionevolmente affidabile solo fino al 2021-2022.

2.6 Considerazioni conclusive

Intel attualmente ricopre una posizione dominante, AMD appare diventare nuovamente competitivo, ARM da verificare seguendo gli sviluppi effettivi e le roadmap che saranno disponibili in futuro.

Per lo scopo della nostra analisi, l'evoluzione dei processori appare qualitativamente costante, con un miglioramento del rapporto tra potenza di calcolo e consumi di un fattore pari al 20-30% a generazione, almeno fino allo sfruttamento delle potenzialità della tecnologia a 7 nm. Questo aumento non è apparentemente in grado di supportare le esigenze di potenza di calcolo e di sostenibilità degli esperimenti nel prossimo decennio.

La potenza di calcolo del singolo core sembra invece essere più o meno costante, attorno ai 10-12 HS06 per core, nonostante i miglioramenti delle cache e delle istruzioni vettoriali. Questo andamento è mostrato in Fig. 11, tratta da uno studio fatto da AMD e presentato alla Salishan Conference on High-speed Computing ([9, 10]), in cui si evidenzia come l'evoluzione intorno al 2005 ha smesso di puntare all'aumento della frequenza del clock, aumentando invece il numero di transistor e conseguentemente dei core logici. La blanda crescita della potenza del singolo core si riscontra in modo accentuato, quasi un appiattimento, nei test fatti sui codici tipici di HEP, forse per via di una certa difficoltà che queste applicazioni incontrano nello sfruttare le caratteristiche evolutive dei processori.

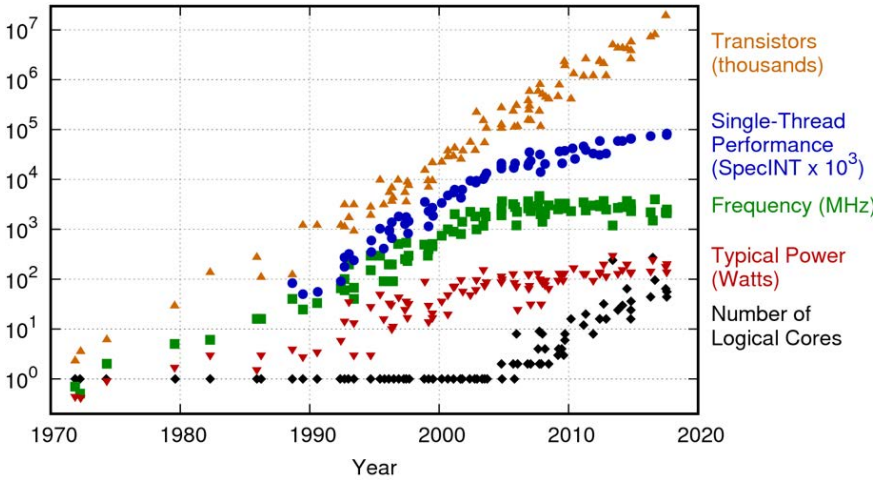


Fig. 11: Evoluzione dei consumi

Infine, una considerazione va fatta sulla quantità di energia assorbita e dissipata dai server: ad ogni evoluzione del processore si ha un aumento del numero di core, ed un conseguente aumento della potenza richiesta dal processore, in un rapporto che favorisce l'aumento della capacità di calcolo per Watt, ma che costituisce comunque un aumento significativo dei consumi per singolo server. Questo comporterà, a parità di densità di server per rack-unit, un aumento della potenza assorbita e dissipata dal singolo rack e un conseguente aumento della richiesta energetica.

3 ACCELERATORI/GPU

3.1 Nvidia

Nvidia ha inventato il GPU computing e resta indiscusso leader del mercato. Grazie alle API CUDA, molto performanti e di facile utilizzo, ha di fatto creato un lock-in che scoraggia gli sviluppatori a cambiare acceleratori, poiché dovrebbero riprogrammare buona parte del codice per adattarlo alle API concorrenti. La roadmap (vedi Fig. 12) sembra, come nel caso di Intel molto rallentata rispetto al passato, a causa della mancanza di un vero concorrente. Siamo ora alla sesta generazione architetturale, denominata Volta, che porta con sé il solito incremento prestazionale mantenendo un costo competitivo rispetto alla passata generazione. Una novità che ha scontentato molti clienti è la nuova politica di Nvidia volta a scoraggiare in modo deciso l'utilizzo degli acceleratori grafici sviluppati per il mercato *gaming and entertainment* (economici) come acceleratori per il calcolo HPC; per tale uso sono supportate solo le GPU della famiglia Tesla o successive, appositamente sviluppate e decisamente più costose.

È tecnicamente possibile destinare le GPU più economiche al calcolo scientifico, ma si perde il diritto al supporto tecnico di Nvidia, hardware e software.

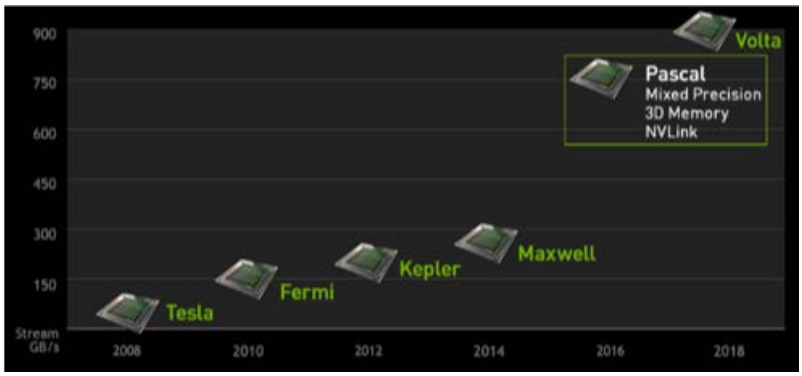


Fig. 12: Roadmap di Nvidia

L'oggetto attualmente più interessante proposto sul mercato è denominato DGX-2 (vedi Fig. 13). Si tratta di uno chassis dalle specifiche ben definite, venduto da OEM ma interamente ingegnerizzato da Nvidia, che contiene fino a 16 GPU, ottimizzando la gestione e il raffreddamento. Un oggetto di questo tipo è ovviamente estremamente costoso, ma in un contesto specifico, è sicuramente una soluzione ottimale.

Una cosa da rammentare nell'adozione di GPU, che vale da almeno due generazioni, è che i server che devono ospitare questo tipo di acceleratori devono essere progettati per ospitarle, dato che le GPU non sono più provviste di raffreddamento attivo ma fanno affidamento unicamente sul raffreddamento fornito dagli chassis in cui vengono inserite. È quindi spesso necessario prevedere l'acquisto dello chassis unitamente alla scheda, perché chassis generici non sono quasi mai predisposti per l'alloggiamento di una o più GPU (almeno non quelli ad alta densità con cui abbiamo a che fare generalmente nei nostri centri di calcolo).

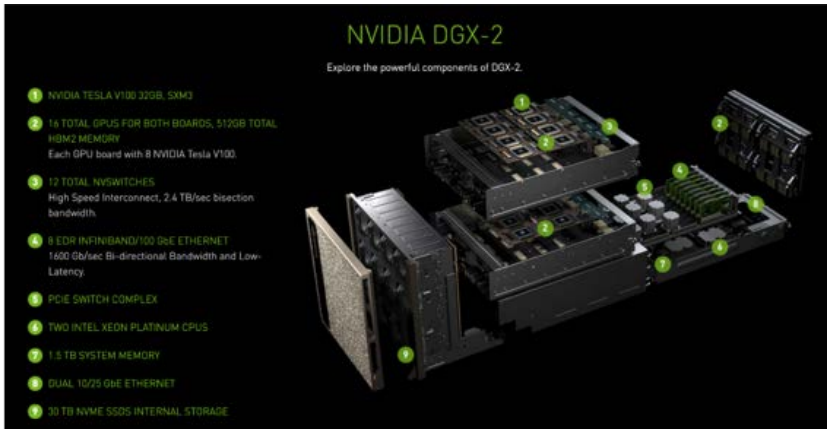


Fig. 13: NVidia DGX-2

3.2 ATI/AMD

Come per il mondo CPU, anche su GPU AMD sta cercando un rilancio proponendo una tecnologia più economica della controparte Nvidia e puntando tutto sulla apertura verso gli sviluppatori. Il loro prodotto di punta, denominato Vega 20, fornisce 1TB di banda passante, è disponibile nei tagli da 16 o 32 GB, dotato di PCI-X gen4 e in grado di supportare la virtualizzazione della GPU tramite SR-IOV[7] (MxGPU); il processo produttivo è a 14nm. A seguire saranno realizzati, nel corso del 2019, acceleratori denominati Navi, che dovrebbero trarre significativi benefici dal punto di vista dei consumi grazie al processo a 7nm, mentre nel 2020 si dovrebbe arrivare alla nuova generazione di microarchitettura, non ancora ben definita, in grado di sfruttare ulteriori miglioramenti dal punto di vista del processo produttivo, grazie al processo denominato 7nm+, di cui si sa ben poco. È evidente come fino ad ora il mondo GPU sia stato monopolizzato da Nvidia che grazie alle proprie API proprietarie, altamente ottimizzate e semplici da utilizzare, ha fidelizzato molti sviluppatori.

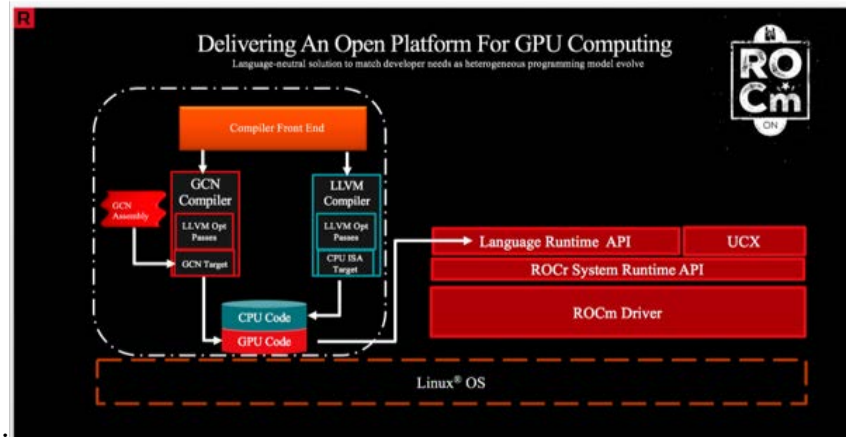


Fig. 14: AMD Open Platform

[7] Single Root I/O Virtualization e' una specifica di protocollo che permette di condividere una risorsa PCI Express tra diverse VM

Per convincere sempre più persone ad adottare le proprie soluzioni, AMD ha reso disponibile un compilatore in grado di tradurre codice dal mondo Nvidia al mondo open proposto da AMD, (vedi Fig. 14) che dovrebbe coniugarsi all'interno della piattaforma ROCm ([11]) al momento disponibile solo per S.O. Linux.

Il vendor lock-in creato di fatto da Nvidia è molto preoccupante dal nostro punto di vista, poiché limita totalmente la concorrenza, con il risultato di mantenere prezzi molto alti e riducendo la spinta innovativa. È auspicabile una maggiore generalizzazione dei paradigmi di programmazione, al fine di rimuovere questo vincolo e stimolare la competizione.

3.3 Intel

Avendo abbandonato il progetto Xeon PHI, ora Intel sembra in difficoltà dal punto di vista delle GPU, non avendo mai offerto un prodotto competitivo. Notizie di cronaca riportano l'assunzione del capo progetto ATI, ora in forza ad Intel, per dare spinta alla creazione di un prodotto destinato al mondo HPC, con la produzione di una scheda discreta: i primi annunci non sono attesi però prima del 2020 e attualmente è difficile fare previsioni. Sarà sicuramente interessante analizzare il prodotto finale, ma così come per il caso di AMD/ATI, il problema del lock-in creato da Nvidia rischia di rendere questo sforzo inutile.

3.4 Considerazioni conclusive

Il panorama GPU appare abbastanza consolidato e la sfida dovrebbe, come in molti altri ambiti, spostarsi sul supporto software. Le librerie OpenCL, un tempo miraggio della compatibilità a 360 gradi per tutti gli acceleratori, sono in fase calante di interesse. Al momento nessun produttore di hardware fa del supporto ad OpenCL un punto di forza per il proprio marketing. Le librerie CUDA, al contrario, continuano ad evolvere e supportano caratteristiche sempre più avanzate. Nel panorama attuale sembra possa valere la pena di puntare, come alternativa, su OpenMP: recenti estensioni a questo standard ne stanno espandendo gli orizzonti in direzioni che appaiono interessanti e i maggiori produttori annunciano il supporto. Può ad esempio essere esteso per lavorare con le GPU e con un po' di sforzo anche con altri acceleratori quali gli FPGA. Diverse ottimizzazioni sono state introdotte nei compilatori più diffusi quali GCC, Intel e LLVM: il supporto appare ampio. Le istruzioni su cui tutti gli acceleratori si stanno maggiormente concentrando si stanno spostando dall'aritmetica a bassa precisione verso le reti neurali: se il mondo HEP riesce a sfruttare questa novità, avremo grossi vantaggi.

4 STORAGE - DISCO

Lo storage su disco attualmente viene realizzato utilizzando due tecnologie: disco a stato solido (SSD) e disco magnetico (HDD). Il primo, più costoso e performante, viene utilizzato per motivi prestazionali, il secondo, economico e capiente, per esigenze di capacità.

4.1 SSD: trend tecnologico

La tendenza nel mondo SSD è orientata verso il calo dei prezzi e l'aumento delle capacità: il mercato ha iniziato la transizione verso la memoria NAND QLC (quad-level-cell) (si veda [12] per una esaustiva rivisitazione delle tecnologie relative alla *flash memory*).

NAND QLC è meno costosa da produrre della TLC (triple-level cell) attualmente usata nella maggior parte degli SSD odierni, ed offre una maggiore densità, ma è più lenta e garantisce

una durata minore per via del minore numero di cicli di *programming/erase* supportati dalle celle. Per contrastare questo problema si fa uso di tecniche di *wear leveling* ([13]), finalizzate a distribuire le scritture in modo uniforme su tutte le celle. Spesso il *wear leveling* è affiancato all'*over-provisioning* sulla disponibilità di memoria, per poter sostituire in modo trasparente le celle danneggiate e prolungare ulteriormente la vita del device. Queste tecniche permettono di creare unità sufficientemente veloci e resistenti nel tempo da soddisfare la stragrande maggioranza degli utenti.

L'altro aspetto che contribuisce ad aumentare la capacità dei device riducendo i costi di produzione è l'utilizzo della tecnologia 3D NAND ([12]), che consiste nell'impilare più celle di archiviazione una sopra l'altra durante il processo produttivo, in un numero di layer sempre crescente, aumentando ulteriormente la capacità del device.

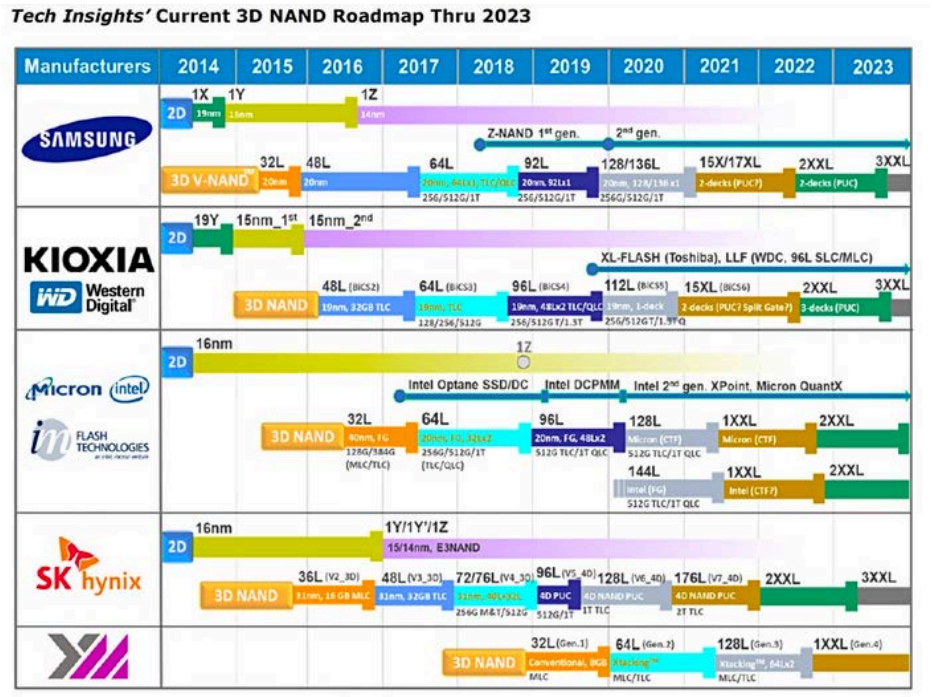


Fig. 15: roadmap per la tecnologia 3D NAND

Tutti i principali produttori di SSD attualmente utilizzano NAND a 92 o 96 layer per i device a capacità più elevata, e sono pronti ad introdurre le soluzioni oltre i 100 layer entro la fine del 2020, in una situazione che vede i diversi competitori quasi allo stesso livello di sviluppo (vedi Fig. 15, tratta da [14]).

Sul versante del bus di interconnessione l'NVMe[8] è destinato a soppiantare il meno performante bus SATA: in base al trend dei costi già nel corso del 2020 gli SSD/NVMe diventeranno economicamente più convenienti degli SSD/SATA ([15]).

Per quanto riguarda le prestazioni dei prodotti attualmente disponibili la memoria 3D XPoint di Intel ([16]) si è rivelata attualmente la più veloce sul mercato. Samsung ha rapidamente risposto con una tecnologia concorrente chiamata Z-NAND ([17]). L'azienda sudcoreana ha illustrato il potenziale della Z-NAND per anni, e dopo un primo approdo l'anno

[8] NVMe: Non-Volatile Memory Express è una interfaccia di accesso alla memoria sviluppata per sfruttare parallelismo e bassa latenza specificatamente per gli SSD ([30])

passato nel mercato enterprise in capacità limitate, ora la nuova tecnologia è disponibile anche sui computer tradizionali.

È molto difficile prevedere l'andamento dei prezzi nel prossimo futuro, ed impossibile a lungo termine. Dopo tre anni di trend in decrescita, a fine 2019 si è verificato un leggero aumento dei costi delle memorie NAND e dei dischi SSD. In questo particolare momento, infatti, la diminuzione dei costi legati alle innovazioni tecnologiche viene controbilanciata da circostanze contingenti quali un grande aumento della richiesta di device ed un incremento dei costi nella realizzazione dei wafer. Stime recentissime di osservatori di mercato (si veda ad esempio TrendForce[9], attraverso la sua divisione specializzata DRAMeXchange[10]), prevedono che nel corso del 2020 i prezzi dei dischi SSD potrebbero anche subire un incremento fino al 15% ([18]).

Si deve però considerare che queste previsioni hanno valore a breve termine e non possono essere proiettate troppo avanti nel futuro.

4.2 HDD ad alta capacità: tecnologie attuali

Il disco ad alta capacità di categoria enterprise a sua volta si divide in disco ad elevate prestazioni (15 krpm, con interfaccia SAS o FC, di capacità dell'ordine del TB) e ad elevata capacità (7.2 krpm, interfaccia SAS, capacità fino a 12 TB, noto come disco NL-SAS).

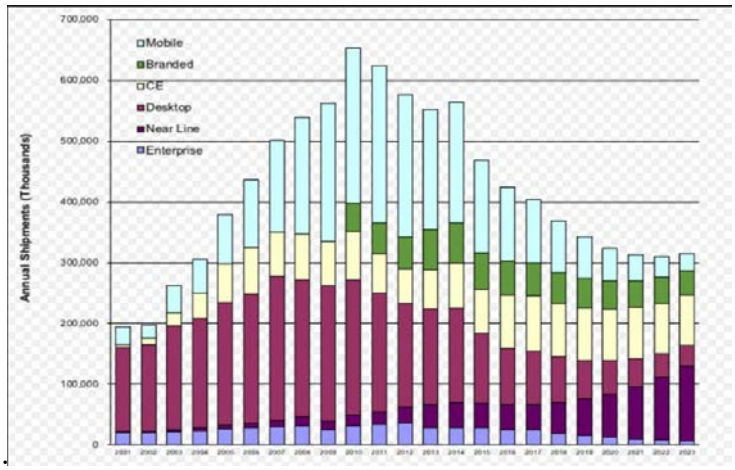


Fig. 16: Unita' di HDD venduti per anno

Come riportato in un recente articolo [10, 19], il trend del mercato per i prossimi anni (vedi Fig. 16) mostra come il disco NL-SAS sia l'unico che aumenterà il proprio volume di vendite, in quanto le altre tipologie verranno sostituite gradualmente dal disco SSD o da altre soluzioni basate su flash. Questo trend è già in atto, anche in ambito HEP.

L'evoluzione tecnologica di nostro interesse è quindi quella volta ad aumentare la capacità del singolo device. Attualmente la tecnologia PMR (Perpendicular Magnetic Recording) è quella che garantisce la massima densità; accanto a questa, si fa uso di dischi immersi in He, tecnologia che permette di aumentare il numero di piatti (fino a 9) nello stesso volume del device a 3.5". Queste tecnologie assieme hanno permesso di ottenere dischi

[9] TrendForce è una nota compagnia specializzata in ricerche di mercato in vari settori della Information Technology

[10] DRAMeXchange è la divisione di TrendForce dedicata specificatamente alla tecnologia DRAM

NL-SAS fino a 10-12 TB.

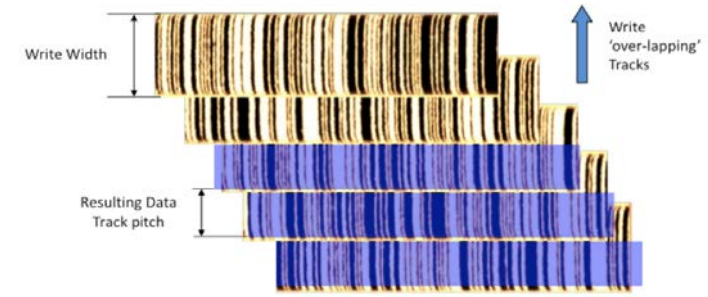


Fig. 17: Tecnologia SMR

Un ulteriore aumento di capacità si è ottenuto grazie alla tecnica del SMR (Shingled Magnetic Recording): questa tecnica implementa scritture di tracce parzialmente sovrapposte (Fig. 17), cosa possibile in quanto la lettura richiede una traccia più sottile della scrittura. Tramite SMR si ottengono oggi capacità fino a 14 TB (senza He), che potrebbero crescere fino a 20 TB nel prossimo futuro.
Questo tipo di disco tuttavia ha pessime prestazioni in scrittura, in quanto la scrittura di una porzione di traccia richiede la successiva riscrittura della porzione di traccia adiacente, che è stata parzialmente sovrascritta, e così via fino all'ultima traccia del disco. Il disco di questo tipo è da intendersi idoneo come archivio: singola scrittura.

4.3 HDD ad alta capacità: evoluzione

Seagate ha annunciato per fine 2020 ([20]) la produzione di HDD che utilizzano la tecnologia HAMR (Heat Assisted Magnetic Recording): facendo uso di un laser che aumenta ad oltre 400 gradi la temperatura della superficie magnetica, è possibile utilizzare un campo magnetico meno intenso per scrivere, occupando quindi una superficie inferiore (vedi Fig. 18).

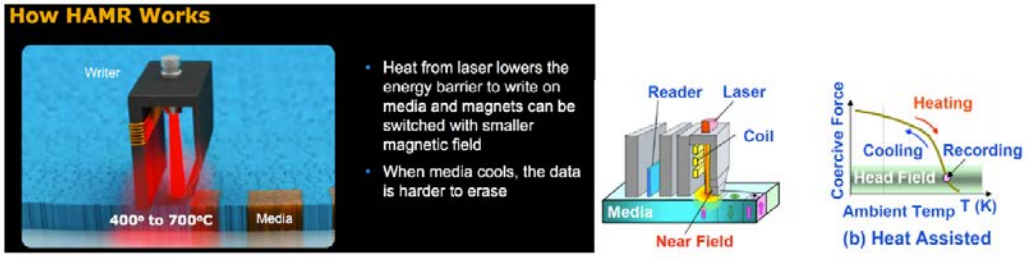


Fig. 18: La tecnologia HAMR di Seagate

In precedenza Western Digital ([19]) aveva annunciato già per il 2019 la produzione di dischi a tecnologia MAMR (Microwave Assisted Magnetic Recording), mostrata in Fig. 19: tramite l'utilizzo di una componente a microonde in scrittura, si riesce ad occupare una porzione di spazio inferiore per la scrittura perpendicolare della superficie magnetica.

Fig. 19: La tecnologia MAMR di Western Digital

Western Digital ha recentemente modificato la sua roadmap introducendo una tecnologia mista, ePMR (energy-assisted PMR) ([21]) che vedrà i primi HDD prodotti per il 2021 e che costituisce il primo passo verso la realizzazione dei dischi in tecnologia MAMR, ora pianificati per il 2022-2023.

Le tecnologie HAMR e MAMR, simili nella efficacia, permettono di aumentare la capacità del singolo piatto. L'aumento di capacità a parità di meccaniche comporta però una diminuzione delle prestazioni relative, sia il throughput per TB che il numero di IOpS per TB. Per migliorare le prestazioni Seagate ha annunciato una soluzione, Multi-Actuator HDD ([22]), che consiste nel dotare l'HDD di due motori di testine indipendenti: raddoppiando le testine e rendendole indipendenti di fatto si raddoppia la prestazione dell'HDD.

Tramite queste tecnologie la capacità degli HDD crescerà nei prossimi anni (vedi Fig. 20), con un trend che per entrambi i produttori promette di avere dischi da 40 TB per il 2023-2025.

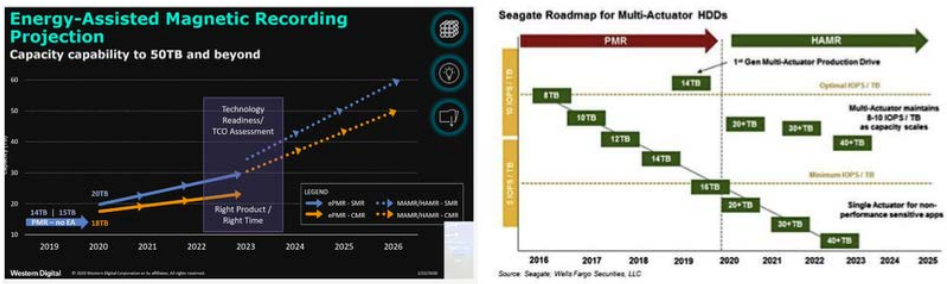


Fig. 20: Roadmap per gli HDD rispettivamente per WD e Seagate

4.4 Protezione dei dati

4.4.1 Protezione dei dati con RAID tradizionale

L'utilizzo di HDD di enormi dimensioni pone un problema sulle tecnologie per garantire l'affidabilità dei dati.

La tecnologia RAID è nata con l'obiettivo di ottenere elevate prestazioni e affidabilità e dispone di diverse soluzioni per proteggersi dalla rottura di un HDD: stripe (RAID0), mirror (RAID1) o utilizzo di striping con aggiunta di uno o più bit di parità (RAID5, RAID6). È possibile anche combinare le diverse tecniche.

Il mirror consiste nell'utilizzare un pool di dischi in cui la metà viene utilizzata come spazio utile, l'altra metà è usata per replicare il dato.

Nella configurazione RAID5 un pool di N+1 HDD viene utilizzato per offrire uno spazio

utile equivalente ad N HDD, mentre lo spazio rimanente viene utilizzato per memorizzare bit di parità. Lo spazio disco viene suddiviso in stripe, costituite ciascuna da N+1 blocchi, uno per ciascun disco. Lo stripe viene usato per ospitare N blocchi di dati, ed utilizza il blocco rimanente per registrare i bit di parità. In caso di rottura di un HDD si utilizzeranno le informazioni degli N blocchi rimanenti di ciascuna stripe per ricostruire il blocco mancante in base alla parità.

Nella configurazione RAID6 la tecnologia è la stessa, solo che si utilizzano due bit di parità, realizzando pool N+2. Questa tecnologia è più costosa in termini di spazio utile, e leggermente meno efficiente in termini di prestazioni, ma offre una protezione anche sulla rottura di due HDD, aumentando il livello di affidabilità del dato. Il RAID6 ha sostituito quasi completamente il RAID5 data la maggiore affidabilità.

Il RAID1 risulta in generale più costoso delle soluzioni RAID5/6 in quanto richiede un raddoppio dello spazio necessario ad ospitare i dati rispetto ai fattori (N+1)/N del RAID5, (N+2)/N del RAID6; le configurazioni tipiche sono 4+1 o 8+2, con un costo del 20% nello spazio utile.

Il grosso limite della tecnologia RAID hardware consiste nei lunghi tempi di ricostruzione dei volumi in occasione della sostituzione di HDD guasti. La ricostruzione è una procedura estremamente intensiva e incide sulle prestazioni del volume coinvolto e del RAID controller che gestisce le operazioni di ricostruzione.

4.4.2 Tecnologie alternative per la protezione dei dati

Nel caso dei dischi ad alta capacità, l'aumento delle dimensioni degli HDD comporta un aumento dei tempi di ricostruzione, in misura proporzionale. Già con le dimensioni attuali la ricostruzione dei volumi comporta più di un giorno, durante il quale le prestazioni del controller sono degradate. L'utilizzo della stessa tecnologia su dischi che aumenteranno di due/quattro volte la dimensione attuale non è pensabile: si arriverebbe ad una situazione in cui il sistema di storage non sarebbe mai in condizioni stabili.

In base alle considerazioni precedenti, l'affidabilità dei dati richiederà soluzioni diverse.

Una potrà essere l'utilizzo di configurazioni di storage più semplici e quindi di costo inferiore, su cui configurare un mirror con replica a fattore 2 o superiore, ma applicato ai soli dati. Questa soluzione, tipicamente utilizzata dalle tecnologie attuali di object storage, limita il raddoppio dei costi dovuto alla ridondanza con un minore costo dell'hardware non avendo necessità di sofisticati controller RAID per la gestione dei dischi. Va tuttavia considerato l'impatto delle rotture dei dischi, che comportano la rigenerazione delle repliche dei dati per il ripristino della affidabilità, che tipicamente vanno ad occupare banda sulla rete di produzione.

Una diversa soluzione è implementata da diversi vendor, chiamata "RAID Distribuito", "declustered RAID", o "RAID 2.0". Questa soluzione implementa la stessa logica dei bit di parità, tramite stripe costituite da blocchi di bit fisicamente corrispondenti su ciascun HDD, ma a livello software, applicata a blocchi di dati. Ogni blocco di dati appartiene ad un set di blocchi a cui vengono associati due blocchi di parità. La collocazione di questi blocchi ha il solo requisito di garantire la protezione dalla rottura di uno o due HDD.

In questa soluzione, la ricostruzione della parità riguarda unicamente i blocchi di dati perduti col disco rotto, ed il carico di lavoro viene suddiviso tra tutti i dischi del pool, riducendo così i tempi di esecuzione e l'impatto sulle prestazioni.

Ai controller verranno quindi richieste capacità software più efficaci e meno complicazione hardware.

4.5 Densità e consumi

Il consumo di un HDD ha un valore compreso tra i 10 ed i 14W, sostanzialmente indipendente dalla capacità. L'evoluzione tecnologica mostrata nei precedenti paragrafi non modificherà le caratteristiche fisiche dei dischi (3.5", stesso consumo per HDD), e non cambierà in modo sostanziale il consumo per HDD, quindi l'aumento di capacità si rifletterà proporzionalmente nella corrispondente diminuzione del consumo espresso in termini di TB, oggi circa 1.3W/TB con HDD da 12TB.

Tutti i rivenditori offrono soluzioni ad alta densità, con cassetti da 60-85 HDD in 4 rack unite alcuni fino a 90 HDD in 4 rack unit. Queste informazioni devono essere tenute in debito conto in sede di definizione del formato dei rack (profondità fino a 120 cm, larghezza fino a 90 cm) e delle caratteristiche dell'eventuale pavimento flottante (per il peso del rack).

4.6 HDD vs SSD

Il contesto attuale vede una rapida evoluzione delle tecnologie di storage ad alte prestazioni, sia dal punto di vista tecnologico che da quello prettamente commerciale.

Come mostrato in Fig. 21, negli ultimi anni l'evoluzione dei costi per TB dello storage delle due soluzioni tecnologiche ha mostrato un calo costante dei prezzi. Il calo è stato contenuto per gli HDD e mediamente più accentuato per gli SSD, al punto da portare il rapporto di costi sotto il fattore 10 ([23]).

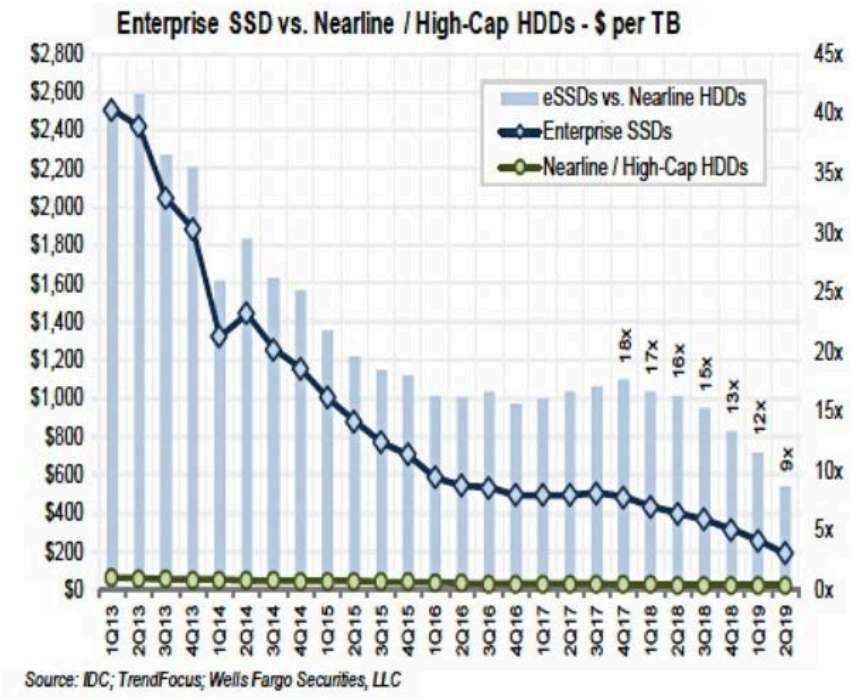


Fig. 21: Andamento dei costi SSD e Nearline HDD

In un contesto di questo genere è probabile che il fattore costo determinerà la composizione dello storage per i dati HEP dei prossimi anni, con la gran parte dello spazio costituito da HDD capacitivi, decisamente più economici, mentre lo storage SSD potrà essere impiegato solo per volumi limitati e per scopi particolari (metadati, cache, primo livello di *tiering*, database), capaci di sfruttare in modo efficace le migliori performance. Sarà importante verificare le esigenze di performance relative (MB/s per TB e IOpS per TB) necessarie alle applicazioni, e la capacità di selezionare efficacemente i dati caldi da posizionare su storage

performante sarà critica.

Tuttavia, queste considerazioni potrebbero dover essere riviste in base alla evoluzione dei costi, che hanno un peso significativo.

Le innovazioni tecnologiche per l'aumento della densità (ePMR, HAMR, MAMR) e le soluzioni come il multi-actuator che manterranno le prestazioni su valori soddisfacenti, avranno un impatto favorevole sui costi per TB degli HDD, ed i *vendor* propongono roadmap ben definite e proiettate verso soluzioni ad oltre 50 TB per HDD. Il calo di costi per TB dovuto all'incremento della densità andrà comunque verificato a fronte degli alti costi di produzione delle tecnologie sviluppate e dalla loro complessità, fattori che hanno già provocato ritardi sull'ingresso nel mercato.

D'altra parte, il prezzo dello storage SSD è ormai da anni in una continua fase di decrescita, anche se attenuata negli ultimissimi mesi, ed è plausibile che riprenda a scendere grazie alla combinazione di ampio spazio di evoluzione tecnologica, forte concorrenza, aumento dei volumi di mercato.

Previsioni affidabili non sono possibili: è presumibile che per il futuro più prossimo la differenza dei costi non si discosterà troppo dai valori attuali, ma non è impossibile che il rapporto del costo SSD/HDD possa presto riprendere a scendere, fino a raggiungere quel valore critico (5x) che per alcuni analisti costituisce il momento in cui inizierà ad essere conveniente sostituire l'HDD con l'SSD, almeno per alcune tipologie di utilizzo ([14]).

In questa eventualità, ovviamente, le considerazioni precedenti potrebbero cambiare radicalmente, soprattutto in un'ottica decennale, spostando la tecnologia utilizzata verso il disco SSD e relegando il disco HDD allo storage di archiviazione in competizione con il nastro, in misura tanto più accentuata quanto minore sarà il rapporto di costo unitario tra le due soluzioni.

5 STORAGE - NASTRO

5.1 Evoluzione della tecnologia

I tape drive operano oggi a densità che sono inferiori di due ordini di grandezza rispetto agli HDD, mantenendo comunque un vantaggio in termini di costi e di capacità grazie ad un supporto con molta più superficie disponibile.

Questo offre margine per continuare a migliorare la crescita di densità con andamento uguale al passato senza dover cercare tecnologie radicalmente diverse. Mantenere una crescita di un fattore 2 ogni due/tre anni permetterà al nastro di mantenere, e probabilmente migliorare ulteriormente, il vantaggio economico rispetto agli HDD.

I supporti magnetici per i tape drive sono oggi prodotti da due sole aziende: Sony e Fujifilm. Pur essendo entrambe attive nello sviluppo tecnologico, la mancanza di altri competitor può generare qualche problema. Ad esempio, la distribuzione dell'ultima tecnologia di nastro LTO è attualmente in ritardo per una causa relativa a royalties tra le due aziende.

L'immagine (Fig. 22), tratta da uno studio pubblicato dall'Insic[11], mostra l'andamento dell'aumento di densità per nastri e HDD degli ultimi 30 anni, ed include la tendenza stimata per i prossimi 10 ([24]).

[11] INSIC: consorzio di aziende, istituti di ricerca e università volto a stimolare e analizzare l'evoluzione tecnologica in ambito di *Information Storage Technology*

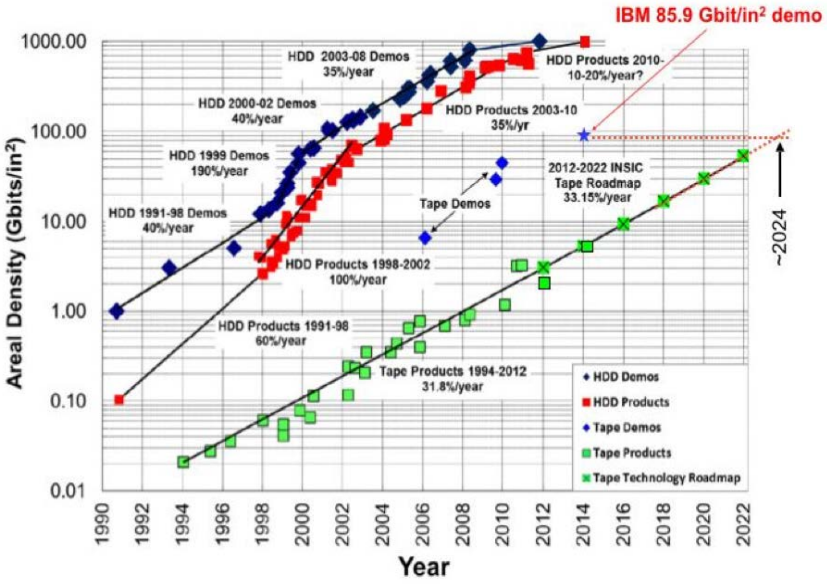


Fig. 22: Evoluzione della densità di scrittura su film

La tecnologia attuale consiste in supporto magnetico costituto da particelle di *Barium-Ferrite* (BaFe), che ha sostituito la tecnologia basata su *Metal Particles*. Questa, unitamente alla tecnologia di registrazione GMR (Giant Magneto Resistance), ha permesso negli ultimi 6 anni di migliorare la densità di circa il 33% per anno, fino a raggiungere il valore attuale di 85.9Gb/in^2. L'introduzione della seconda generazione di particelle di BaFe, più piccole, permetterà di mantenere, nei prossimi anni, la stessa progressione nell'aumento della densità di registrazione delle informazioni.

IBM, in collaborazione con Fujifilm, ha realizzato una dimostrazione per cui, utilizzando una nuova tecnologia di magnetizzazione perpendicolare su nastri in BaFe, è riuscita ad ottenere una densità pari a 123 Gb/in^2, corrispondente ad un tape di 220 TB lordi ([25]).

In base alla tendenza prevista, come si può vedere in figura, tale soluzione sarà disponibile su prodotti in commercio presumibilmente intorno al 2025-2026.

L'evoluzione tecnologica procede anche in altre direzioni: sempre IBM, questa volta con Sony, ha dimostrato l'efficacia di una nuova tecnologia nella produzione del supporto magnetico, detta *sputtered tape*, che utilizzando grani di CoPtCr, ha permesso, sempre come evento dimostrativo, di raggiungere nel 2017 densità pari a 201 GB/in^2 ([26]).

Queste ed altre evoluzioni, come la registrazione con tecnologia TMR (Tunnel Magneto Resistive) e l'utilizzo di grani di SrFe garantisce per il prossimo decennio ed oltre il mantenimento del trend attuale di crescita di densità dei dati, e conseguente aumento di capacità per tape e riduzione del costo per TB dello storage su nastro.

5.2 Nastro

Oracle è recentemente uscita dal mercato dei tape drive, in cui era presente con i modelli della serie T100000D della StorageTek, lasciando IBM Enterprise e LTO Ultrium Generation come uniche soluzioni sviluppate. Il mercato (vedi Fig. 23) vede prevalere largamente la soluzione LTO, come mostrato nella seguente figura, che evidenzia il rapporto tra i diversi tipi di cassette vendute negli ultimi anni e la previsione per il prossimo futuro.

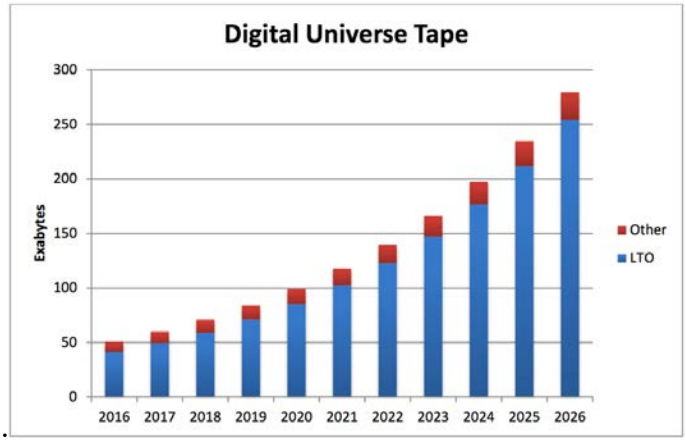


Fig. 23: Evoluzione e previsioni di diffusione delle tecnologie di nastro

Questo dominio dovrebbe rimanere inalterato nei prossimi anni, garantendo sicuramente la permanenza della tecnologia LTO sul mercato.

La roadmap delle tecnologie di nastro per i prossimi anni è illustrata in Fig. 24.

IBM Enterprise è prodotta da IBM e presenta caratteristiche tecniche superiori rispetto a LTO, sia nella capacità che nelle prestazioni.

Il modello attuale, TS1160, ospita fino a 20 TB di spazio non compresso, 400 MB/s di banda in lettura, ed un valore di UBER[12] pari a 10^{-20}.

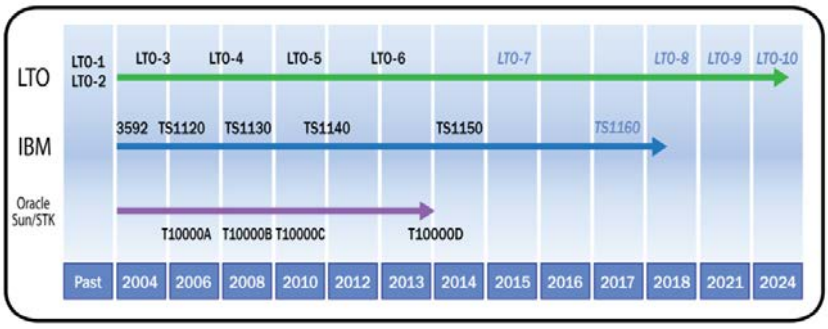


Fig. 24: Roadmap per le tecnologie di nastro

L'incremento rispetto alla generazione precedente (TS1155, 2017) è del 33% in spazio e 11% in velocità. Per la prossima generazione ci si aspettano incrementi anche superiori (30 TB, 500 MB/s). Questa soluzione dispone della tecnologia RAO (Recommended Access Order[13]), che migliora notevolmente l'efficienza nella lettura di diversi file da un nastro, ottimizzandone l'ordine di accesso e riducendo i tempi di posizionamento, aumentando in questo modo anche la durata del nastro (vedi Fig. 25)

---

[12] UBER (Uncorrectable Bit Error Rate): la probabilita' di errore nella lettura di un bit dopo aver applicato gli eventuali meccanismi di autocorrezione

[13] Tecnologia proprietaria IBM, vedi https://www.ibm.com/support/knowledgecenter/STAKKZ/dd_pr_kc/con_a89p4_rao.html

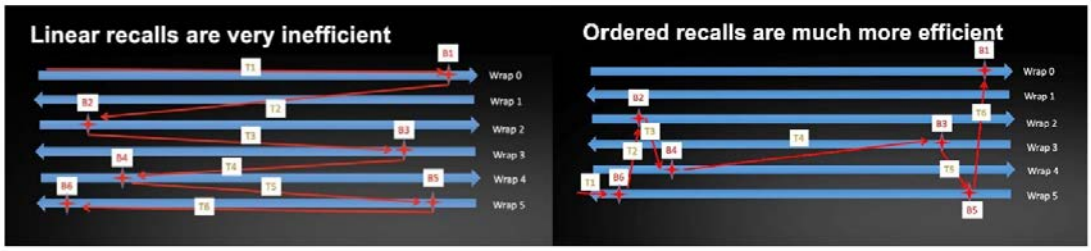


Fig. 25: Tecnologia Recommended Access Order

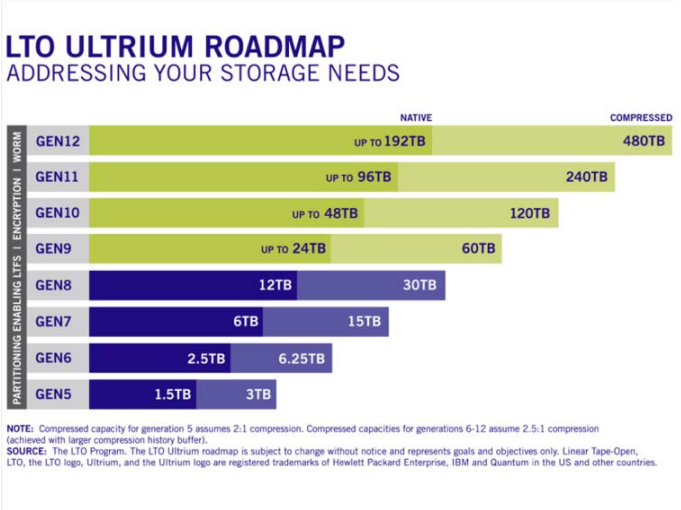
La roadmap per questa tecnologia prevede l'evoluzione a 30 TB per il 2019-2020, i 50-60 TB per il 2021-2022.

LTO è il tipo di tape drive più diffuso, prodotto da IBM, HP e Quantum. La generazione attuale, LTO-8, ha capacità di 12 TB non compressi, velocità di 360 MB/s, un valore di UBER di 10^{-19}. La tecnologia supporta opzionalmente il WORM (Write Once Read Many), mentre manca di meccanismi integrati per l'ottimizzazione dell'accesso, che alcuni produttori di tape library implementano a livello software.

La roadmap per i nastri LTO è ben definita e mostrata in Fig. 26, in cui sono riportate le previsioni di crescita per capacità nativa e compressa. L'evoluzione tecnologica prevede sostanzialmente un raddoppio di capacità ed un aumento del 20% in velocità di accesso ad ogni generazione. Il periodo di rinnovo tecnologico è previsto essere di tre anni per generazione. La roadmap sembra quindi essere definita per tutto il prossimo decennio e oltre. Entrambe le soluzioni sembrano quindi avere una roadmap definita per il periodo di interesse di questa analisi.

La tecnologia LTO è più diffusa e non è prodotta da una singola azienda, questo dovrebbe garantire più protezione per disponibilità e costi.

Le differenti caratteristiche possono essere tradotte in una diversa tipologia di utilizzo, in cui IBM Enterprise sembra più adatta a accessi random e riscritture, mentre LTO si adatta meglio per accessi in streaming o come area di cold storage.



.

Fig. 26: Evoluzione della tecnologia LTO

5.3 Librerie di nastri

Attualmente i produttori di librerie di nastri di grandi dimensioni sono quattro:

- IBM, il cui modello attuale è la libreria IBM TS4500, fino a 24000 slot, 1.0 EB, 128 drive e supporta nastri IBM Enterprise e LTO.
- Oracle, che offre la libreria SL8500, 2000 slot native, ma espandibile fino a 100000 slot per un totale di 1.2 EB e oltre 500 TB/hr grazie a 640 drive installabili, con supporto per nastri LTO e T100000D (Oracle/StorageTek).
- Quantum offre il modello Scalar i6000, fino a 12000 slot, 192 drive, 360 TB (nativi) di spazio disponibile. Supporta unicamente nastri LTO.
- Spectra Logic dispone della libreria T-Finity, capace di ospitare fino a 40000 slot LTO, IBM Enterprise o Oracle/StorageTek serie T100000, per un totale di circa 2 EB (compressi) e 207 TB/hr.

Spectra Logic offre soluzioni software a corredo che implementano ottimizzazioni di accesso, come il TAOS (Time-based Access Order System) che permette di sopperire alla mancanza di RAO per i nastri LTO.

Le librerie di nastri non hanno un impatto rilevante dal punto di vista dei consumi. La capacità delle librerie cresce con il crescere delle capacità dei nastri. Fattori da considerare sono la capacità di far evolvere l'oggetto acquistato per supportare le nuove tecnologie senza dover perdere l'investimento fatto sulla vecchia. Qui si devono verificare i piani di upgrade che i produttori offrono ai loro clienti.

6 TECNOLOGIE DI COLLEGAMENTO ALLA RETE

L'evoluzione delle tecnologie costruttive dei processori ed il relativo aumento delle capacità di elaborazione dei server comporta anche una sempre maggiore necessità di capacità di I/O. Le tecnologie di interconnessione alla rete locale stanno evolvendo di pari passo.

6.1 Protocolli ethernet

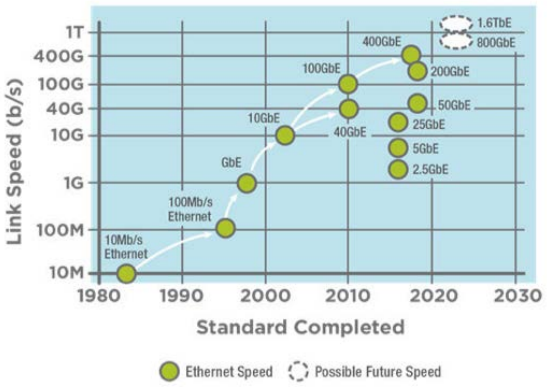


Fig. 27: Evoluzione protocolli Ethernet

6.1.1 10 Gigabit Ethernet

La tecnologia Ethernet completamente "Commodity" alla data di redazione di questo documento sta diventando il 10 Gigabit Ethernet (10GE) che vede numerosissime implementazioni e che ormai si trova di default nelle schede madri dei server di tutti i produttori nella versione 10GE Base-T (IEEE 802.3an).

6.1.2 25 Gigabit Ethernet

La prossima tecnologia ethernet che verosimilmente diverrà *commodity* nei prossimi 2 anni (2019-2020) è il 25 Gigabit Ethernet (IEEE 802.3by) che nasce dalla linea di sviluppo del 100 Gbit Ethernet.

6.1.3 40 Gigabit Ethernet

Una menzione particolare la richiede il 40 Gigabit Ethernet che è disponibile già da parecchi anni e che sembra destinato a scomparire per via dello sviluppo del 100 Gigabit e dei suoi sottomultipli (25 e 50 Gigabit Ethernet).

Ad oggi molte implementazioni basate sul 40 Gigabit Ethernet sono in uso per via della disponibilità di transceiver in grado di sfruttare cablaggi esistenti grazie alle implementazioni 40GE- BiDi in grado di sfruttare i cablaggi basati su coppie di fibre Multimodali (LC).

6.1.4 100 Gigabit Ethernet

È disponibile nelle primissime versioni già dal 2009 ma solo negli ultimi 2 anni è diventata una tecnologia veramente acquistabile anche se ancora un pò lontana dal diventare "Commodity", soprattutto per via dei costi alti dei transceiver.

6.1.5 100GBase-SR4

I più diffusi ed economici transceiver 100Gb Ethernet sono i 100GBase-SR4 che utilizzando cavi in fibra ottica multimodale multifibra di tipo MPO (Multifiber Push-On) per collegamenti su brevi distanze (100 metri).

Questo tipo di connessioni ha un costo decisamente superiore ai semplici cavi in fibra multimodale "Duplex" di cui spesso sono costituiti gli attuali cablaggi dei datacenter. Per rispondere ai requisiti di costo e compatibilità con i cablaggi esistenti, sono disponibili da qualche anno anche transceiver 100 Gbase-SR-BD (basati su tecnologia BiDi) e quindi in grado di sfruttare i cablaggi esistenti basati su coppie fibre multimodali con normale connessione LC.

Fig. 28: 100 Gbps basati su tecnologia BiDi

I maggiori produttori rendono ad oggi disponibili HBA (Host Bus Adapter) su bus PCIe Gen3 16 lane, da installare sui server con una o 2 interfacce a 100Gb/s da popolare con gli opportuni transceiver.

Sono disponibili sul mercato anche HBA dotate di 2 porte a 200G Ethernet predisposti all'offloading di funzionalità orientate all'accesso ai dati come NVMe over fabric e di tecnologie di overlay come VXLAN e NVGRE.

6.1.6 400Gigabit Ethernet

L'ultima tecnologia Ethernet disponibile ad oggi è il 400 Gigabit Ethernet (802.3bs). Sono stati sviluppati vari fattori di forma per i transceiver (vedi Fig. 29) fra i quali sembra prevalere il QSFP-DD.

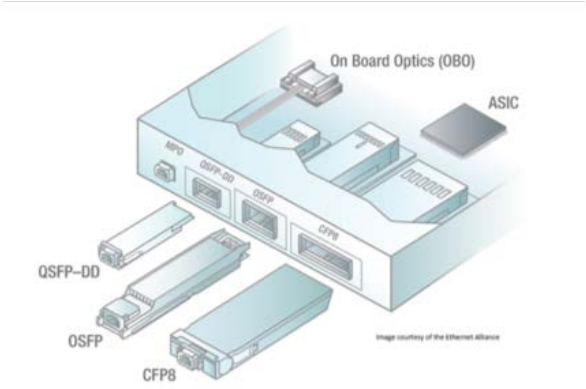


Fig. 29: Transceiver per i 400 Gbps

Per questa tecnologia sono già disponibili ottiche per brevi e lunghe distanze. Le ottiche compatibili con le QSFP-DD saranno compatibili con i QSFP 28.

Per esempio, le ottiche 400 GBASE-SR 16 saranno le ottiche da impiegare per collegamenti in LAN su fibre ottiche multimodali (MPO).

L'evoluzione dell'Ethernet sembra portarci nel giro di pochi anni 2021-2023 agli 800 Gb Ethernet ed al 1,6 Tb Ethernet. Difficile prevedere i tempi di adozione su larga scala di queste nuove tecnologie in quanto sono legate molto ai costi, nonché dai limiti di distanze percorribili dai cavi di interconnessione.

6.2 Apparati di rete per data-center e architetture

6.2.1 Architetture

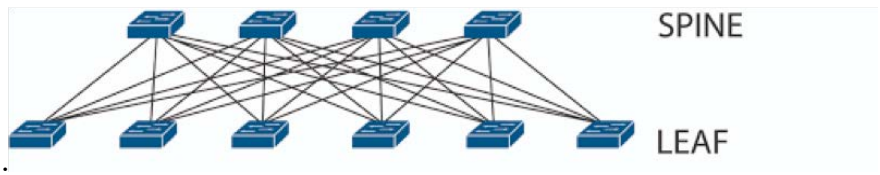


Fig. 30: Architettura Spine-Leaf

Tutti i produttori stanno sviluppando apparati modulari e stand alone per la realizzazione di data center con topologie classiche (a stella) ma anche con topologie di tipo *"Spine Leaf"* (Fig. 30) in cui il traffico non è concentrato tutto su uno o due grossi switch modulari con più livelli di aggregazione ma si costruisce un mesh di switch di cui alcuni destinati a collegare i nodi di calcolo (Nodi LEAF) ed altri destinati a collegare gli switch di leaf (SPINE).

L'approccio Spine Leaf si propone di potere crescere orizzontalmente con maggiore flessibilità rispetto alle architetture a stella e si prestano ad una gestione molto più "software defined". Non banale però è il dimensionamento dei link fra gli switch per potere scalare ad elevati numeri di nodi connessi ad alta velocità con fattori di "oversubscription" nulli o molto bassi. Per data-center di grandi dimensioni il ricorso ad un ulteriore livello di concentrazione "Super Spine" è quasi obbligatorio per indirizzare le esigenze di scalabilità.

La gestione del cablaggio in caso di datacenter molto estesi potrebbe diventare critica.

Un approccio di tipo Software-Defined per la gestione di complesse topologie e di "IP Fabric" diventerebbe praticamente una necessità.

6.2.2 Apparati di rete

Tutti i principali produttori di apparati per data center oltre agli switch modulari con matrici di switching che devono scalare a decine di Tb/s (per matrice) hanno a listino switch stand alone da 1 Unità Rack (1U) o 2U con elevata densità di porte ad alta velocità da utilizzare come building block per architetture classiche come switch ToR o di edge oppure come Spine o Leaf:

- Switch 1U: 32-48 porte 100 GbE QSFP
- Switch 2U: 64 porte 100Gb in 2U
- Switch 1U: 32 porte da 400 GbE (QSFP-DD)

Gli switch 1U più densi disponibili ad oggi gestiscono un traffico aggregato dell'ordine dei 12,8 Tb/s.

Lo sviluppo prosegue anche sui modelli modulari da impiegare sia come livelli di "Super Spine" sia come grossi apparati di aggregazione o centro stella e si sono raggiunte densità di porte elevatissime su chassis dalla capienza dai 10 ai 18 slot. Consideriamo che la densità per scheda sia assimilabile alla densità di porte che si raggiunge per gli apparati 1U. La differenza è che il traffico fra un modulo e l'altro è gestito da matrici di switching ad altissima capacità spesso senza nemmeno un mid-plane di interconnessione fra moduli e matrici.

Nel 2020 si vedranno i primi apparati ad alta densità di porte 800 Gigabit Ethernet.

Quasi tutti i produttori hanno sviluppato switch, e più in generale apparati di rete, in versione virtuale da gestire con i principali sistemi di virtualizzazione ed orchestrazione con le stesse caratteristiche degli switch fisici per una migliore integrazione ed automazione in ottica NFV (Network Function Virtualization).

6.3 Tecnologie per rete geografica

Per quanto riguarda le principali evoluzioni a livello di rete geografica, c'è senza dubbio da citare l'introduzione di apparati trasmissivi di tipo ROADM (Reconfigurable Optical Add/Drop Multiplexer) programmabili.

Questi apparati consentono grazie anche allo sviluppo di specifici chip fotonici di riuscire a gestire su di una singola coppia di fibre, bande dell'ordine dei 30 Tb/s.

Tramite l'utilizzo di apparati chiamati "Transponder" con interfacce di tipo Ethernet per il collegamento alla LAN ed in grado di collegarsi direttamente ai ROADM lato linea, è

possibile arrivare a realizzare soluzioni di trasporto ottico end-to-end realizzando collegamenti di tipo DCI (Data Center Interconnect) ad altissima capacità e con una elevata dinamicità.

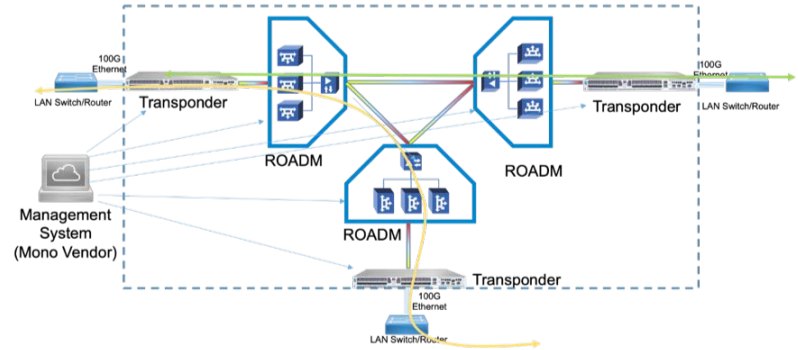


Fig. 31: Trasporto ottico end-to-end basato su apparati ROADM

Senza ricorrere ad una articolata struttura di ROADM trasmissivi, oggi è possibile realizzare semplici configurazioni di DCI (Data center Interconnect) ad altissima capacità (ordine del Terabit/s) fra due o più centri di calcolo distanti non più di 100 Km utilizzando apparati *Transponder* collegati con una semplice coppia di fibre ottiche.

Per dare un'idea della tipologia di transponder disponibile oggi sul mercato, si parla di apparati di 1U di altezza con una decina di interfacce 100Gb Ethernet lato LAN (in genere da 8 a 12) ed una interfaccia lato "Linea" in grado di illuminare una fibra ottica e gestire la capacità totale necessaria.

7 CONCLUSIONI

Nei paragrafi precedenti si è proceduto ad una analisi sulla evoluzione che le tecnologie di interesse per il calcolo dell'INFN potranno avere nei prossimi 10 anni, per quanto sia possibile fare basandosi su dati del recente passato e su roadmap che hanno generalmente una affidabilità di soli due-tre anni.

7.1 CPU e Server

L'evoluzione delle CPU x86 appare procedere ancora per qualche anno nella stessa direzione, con una riduzione delle dimensioni dei processi produttivi a 10, 7 e forse 5 nm, aumento del numero di core per socket, aumento del numero di thread per core. Accanto a queste, ci saranno evoluzioni tecnologiche che gli attuali codici utilizzati in HEP non sono oggi in grado di sfruttare al meglio, come l'aumento della profondità vettoriale e l'evoluzione degli Instruction-Set verso un maggiore supporto al Machine Learning.

Dal punto di vista pratico è interessante l'analisi della capacità di calcolo per potenza elettrica. Le roadmap dei prossimi due-tre anni prevedono processori con aumento di consumi e di potenza di calcolo, in proporzioni tali da favorire il rapporto potenza di calcolo/Watt. Questa previsione conforta l'estrapolazione di tale rapporto che è possibile fare in base ai dati raccolti dai principali centri di calcolo HEP negli ultimi anni: è possibile valutare una forchetta compresa tra i 3 ed i 5 HS06/Watt oltre il 2025.

Accanto a questo va considerato, per valutare la densità di calore prodotto, l'aumento di consumo che si avrà sul singolo server di calcolo. Questo valore avrà un impatto sulle

tecnologie di raffreddamento necessarie, e potrebbe costituire un fattore importante nella valutazione del dimensionamento dei centri di calcolo in termini di occupazione di spazi.

Appare chiaramente interessante l'evoluzione dei coprocessori, che oggi costituiscono uno strumento molto efficace per il calcolo HPC e per le tecnologie di ML e DL.

Non c'è dubbio che un'evoluzione dei codici in direzione tale da saper sfruttare tali tecnologie potrebbe aiutare molto a supportare le esigenze degli esperimenti INFN del prossimo decennio. In questo caso un'analisi quantitativa richiederà la definizione di nuove metriche per valutare le esigenze di calcolo, superando il concetto di HS06 oggi universamente utilizzato.

7.2 Disco

La roadmap promette già per il 2025 la disponibilità di dischi rotazionali capacitivi da 40 TB (corrispondente ad un incremento di circa il 26%/anno)

Il fattore di forma non cambierà, permanendo a 3.5", e quindi rimarrà sostanzialmente invariata l'occupazione degli spazi per HDD ospitati in scatole che arriveranno a 90 HDD in 4 rack unit. Su questo andranno fatte opportune considerazioni sulle caratteristiche di profondità dei rack e sulla capacità di carico degli eventuali pavimenti flottanti.

Dovrebbe restare sostanzialmente invariato anche il consumo elettrico per singolo HDD. Per lo storage la densità di potenza da raffreddare sarà, come oggi, meno importante rispetto ai rack con server di calcolo.

In relazione a questo, è molto importante considerare che, pur con l'apporto di migliorie tecnologiche come il Multi-Actuator HDD, il throughput per TB del disco diminuirà, e questo potrebbe portare ad una insufficiente velocità di accesso allo storage. Inoltre, le tecnologie di ridondanza a protezione dei dati sui dischi capacitivi dovranno essere necessariamente più efficienti di quelle tradizionali per evitare tempi di ricostruzione troppo lunghi, durante i quali c'è degrado prestazionale, che implicano anche un aumento della probabilità di perdere dati.

Il disco SSD mostra prestazioni eccellenti ed in crescita, anche rapida, con l'evoluzione tecnologica. I costi eccessivi ne fanno uno storage di uso necessariamente specifico, per compiti particolari e su volumi limitati.

Queste considerazioni andranno probabilmente rivalutate in considerazione dell'evoluzione tecnologica e soprattutto dell'evoluzione dei costi del disco SSD, che potrebbe comportare una riduzione del rapporto di costi tra le due tecnologie.

Il recente trend, se confermato nei prossimi anni, potrebbe portare ad un consistente ridimensionamento dell'utilizzo del disco magnetico rispetto alle più performanti soluzioni SSD, con vantaggi significativi in termini di IOpS e di throughput per TB.

7.3 Nastro

La tecnologia su nastro ha margini per continuare la sua evoluzione aumentando la densità di scrittura per superficie di nastro e sembra garantire una crescita di capacità equivalente a quella del recente passato. L'evoluzione prevista è di un raddoppio della capacità ogni 3 anni.

All'aumento della densità corrisponderà un aumento delle prestazioni, sia in velocità che in termini di algoritmi di ricerca dei dati più efficienti.

Come per il disco, non cambiano fattore di forma e consumo, che per il nastro è irrilevante.

7.4 Rete

La tecnologia Ethernet appare essere molto viva nell'evoluzione futura e ad oggi sembra improbabile che venga sostituita da altre tecnologie se non per use case specifici e limitati.

Il 100 Gbps è disponibile oggi sul singolo server, e già nei prossimi 2-3 anni saranno disponibili switch con porte a 400 a 800 Gbps, per arrivare a 1.6 Tbps per il 2023-2024.

L'hardware dovrà essere affiancato da una corretta analisi del throughput desiderato, per definire le corrette topologie (stella, leaf-spine, leaf-spine-superspine).

In funzione della complessità del centro si dovranno prendere in considerazioni soluzioni flessibili quali Software Defined Network e Network Function Virtualization.

Gli apparati di rete non avranno un impatto rilevante sui consumi elettrici, ma non dovrà esser trascurato l'impatto sul costo, sia per gli apparati che per il cablaggio, che in un centro ad elevata densità dovrà essere estremamente accurato.

8 RIFERIMENTI

[1] J. L. Hennessy and D. A. Patterson, "A New Golden Age for Computer Architecture: Domain-Specific Hardware/Software Co-Design, Enhanced Security, Open Instruction Sets, and Agile Chip Development - International Symposium on Computer Architecture - ISCA 2018," 4 6 2018. [Online]. Available: https://iscaconf.org/isca2018/turing_lecture.html. [Accessed 14 4 2020].

[2] W. M. Mitchell, "The chips are down for Moore's law," *Nature,* vol. 530, pp. 144-147, 2016.

[3] G. Moore, "Gordon Moore: The Man Whose Name Means Progress, The visionary engineer reflects on 50 years of Moore's Law - IEEE Spectrum: Special Report: 50," 30 3 2015. [Online]. Available: https://spectrum.ieee.org/computing/hardware/gordon-moore-the-man-whose-name-means-progress. [Accessed 14 4 2020].

[4] Hep Software Foundation, "A Roadmap for HEP Software and Computing R&D for the 2020s (Community White Paper)," 15 12 2017. [Online]. Available: https://arxiv.org/abs/1712.06982. [Accessed 14 4 2020].

[5] H. Meinhard, "Proposal for Technology Watch WG," 26 03 2018. [Online]. Available: https://indico.cern.ch/event/658060/contributions/2889030/attachments/1622879/2583193/2018-03-26-WLCGWorkshop-TechnologyWGProposal.pdf. [Accessed 14 4 2020].

[6] Intel Corporation, "Intel Tick-Tock Model," [Online]. Available: https://www.intel.com/content/www/us/en/silicon-innovations/intel-tick-tock-model-general.html. [Accessed 14 4 2020].

[7] "Process, Architecture, Optimization," [Online]. Available: https://it.wikipedia.org/wiki/Process,_Architecture,_Optimization. [Accessed 14 4 2020].

[8] "Cosa Project," [Online]. Available: http://www.cosa-project.it/. [Accessed 14 4 2020].

[9] Chuck Moore, AMD Corporate Fellow, "Data Processing in Exascale-Class Computing Systems," 27 4 2011. [Online]. Available: https://www.lanl.gov/conferences/salishan/salishan2011/3moore.pdf. [Accessed 14 4 2020].

[10] T. Coughlin, "HDD Growth In Nearline Markets," 05 02 2018. [Online]. Available: https://www.forbes.com/sites/tomcoughlin/2018/02/05/hdd-growth-in-nearline-markets/#4d67fe629979. [Accessed 14 4 2020].

[11] AMD Corporation, "ROCm, a New Era in Open GPU Computing," 2020. [Online]. Available: https://rocm.github.io/. [Accessed 14 4 2020].

[12] Wikipedia, "Flash memory," 14 04 2020. [Online]. Available: https://en.wikipedia.org/wiki/Flash_memory. [Accessed 19 04 2020].

[13] Wikipedia, "Wear leveling," 10 02 2020. [Online]. Available: https://en.wikipedia.org/wiki/Wear_leveling. [Accessed 19 04 2020].

[14] Chris Mellor - Blocks and Files, "Hard disk drives will disappear from your data centre – unless you work for a hyperscaler," 7 2 2020. [Online]. Available: https://blocksandfiles.com/2020/02/07/hard-disks-disappear-small-data-centres/. [Accessed 19 4 2020].

[15] Chris Mellor - Blocks and Files, "The irresistible rise of NVMe means SATA SSD days are numbered," 3 3 2019. [Online]. Available: https://blocksandfiles.com/2019/03/03/the-death-of-ssd-sata/. [Accessed 19 4 2020].

[16] Rick Coulson - Intel Corporation, "Stanford University Department of Electrical Engineering Computer Systems Colloquium: The quest for low storage latency changes everything," 02 03 2016. [Online]. Available: http://web.stanford.edu/class/ee380/Abstracts/160302.html. [Accessed 14 4 2020].

[17] Samsung Electronics Co., "Ultra-Low Latency with Samsung Z-NAND SSD," 07 2017. [Online]. Available: https://www.samsung.com/semiconductor/global.semi.static/Ultra-Low_Latency_with_Samsung_Z-NAND_SSD-0.pdf. [Accessed 19 4 2020].

[18] TrendForce, "Owing to Growing Impact of COVID-19 Pandemic, NAND Flash ASP May Tumble in 2H20 Ahead of Expectations, Says TrendForce," 26 3 2020. [Online]. Available: https://press.trendforce.com/press/20200326-3345.html. [Accessed 19 4 2020].

[19] Western Digital Corporation, "Western Digital Unveils Next-Generation Technology To Preserve And Access The Next Decade Of Big Data," 11 10 2017. [Online]. Available: https://www.westerndigital.com/company/newsroom/press-releases/2017/2017-10-11-western-digital-unveils-next-generation-technology-to-preserve-and-access-the-next-decade-of-big-data. [Accessed 14 4 2020].

[20] John Paulsen - Seagate, "HAMR Milestone: Seagate Achieves 16TB Capacity on Internal HAMR Test Units," 2 12 2018. [Online]. Available: https://blog.seagate.com/craftsman-ship/hamr-milestone-seagate-achieves-16tb-capacity-on-internal-hamr-test-units/. [Accessed 14 4 2020].

[21] Billy Tallis - AnandTech, "Western Digital Roadmap Updates: Energy Assisted Recording, Multi-Stage Actuators, Zoned Storage," 31 1 2020. [Online]. Available: https://www.anandtech.com/show/15457/western-digital-roadmap-updates-energy-assisted-recording-multistage-actuators-zoned-storage. [Accessed 14 4 2020].

[22] Jason Feist - Seagate, "Multi Actuator Technology: A New Performance Breakthrough," 2018. [Online]. Available: https://blog.seagate.com/craftsman-ship/multi-actuator-technology-a-new-performance-breakthrough/. [Accessed 14 4 2020].

[23] Chris Mellor - Blocks and Files, "How long before SSDs replace nearline disk drives?," 28 8 2019. [Online]. Available: https://blocksandfiles.com/2019/08/28/nearline-disk-drives-ssd-attack/. [Accessed 19 4 2020].

[24] R. Raymond and et al., "2019 INSIC Technology Roadmap," 07 2019. [Online]. Available: http://www.insic.org/wp-content/uploads/2019/07/INSIC-Technology-Roadmap-2019.pdf. [Accessed 14 4 2020].

[25] e. a. Mark A. Lantz, "123 Gbit/in2 Recording Areal Density on Barium Ferrite Tape," *IEEE Transactions on Magnetics,* vol. 51, no. 11, 2015.

[26] e. a. S.Furrer, "201 Gb/in2 Recording Areal Density on Sputtered Magnetic Tape," *Magnetics IEEE Transactions,* vol. 54, no. 2, 2018.

[27] Wikipedia, "NVM Express," 18 4 2020. [Online]. Available: https://en.wikipedia.org/wiki/NVM_Express. [Accessed 19 4 2020].