UNIVERSITÀ DI PARMA

DIPARTIMENTO DI SCIENZE MATEMATICHE, FISICHE ED INFORMATICHE
Corso di Laurea in Informatica

Performance dell'hardware

Programmazione parallela e HPC - a.a. 2022/2023

Roberto Alfieri

Programmazione Parallela e HPC: sommario

Performance

La performance complessiva di un sistema di calcolo è la quantità di lavoro utile in relazione al tempo e alle risorse disponibili.

La performance è misurata principalmente nel **tempo** impiegato per terminare il lavoro.
Un'altra metrica importante è il **consumo di risorse computazionali e di energia.**

Sulle prestazioni incidono le caratteristiche tecnologiche dell'**hardware** utilizzato e la qualità del **software.**

Le risorse hardware che incidono sulle performance sono le unità di processamento (e.g. CPU, GPU), la memoria, lo storage e la comunicazione in rete.

I fattori che incidono sulle prestazioni del **Software** sono gli algoritmi e l'organizzazione dei dati dell'applicazione, l'hardware exploitation ovvero la capacità di sfruttare le risorse hardware disponibili, le caratteristiche del software di base utilizzato (sistema operativo, librerie e compilatori).

Definizioni:

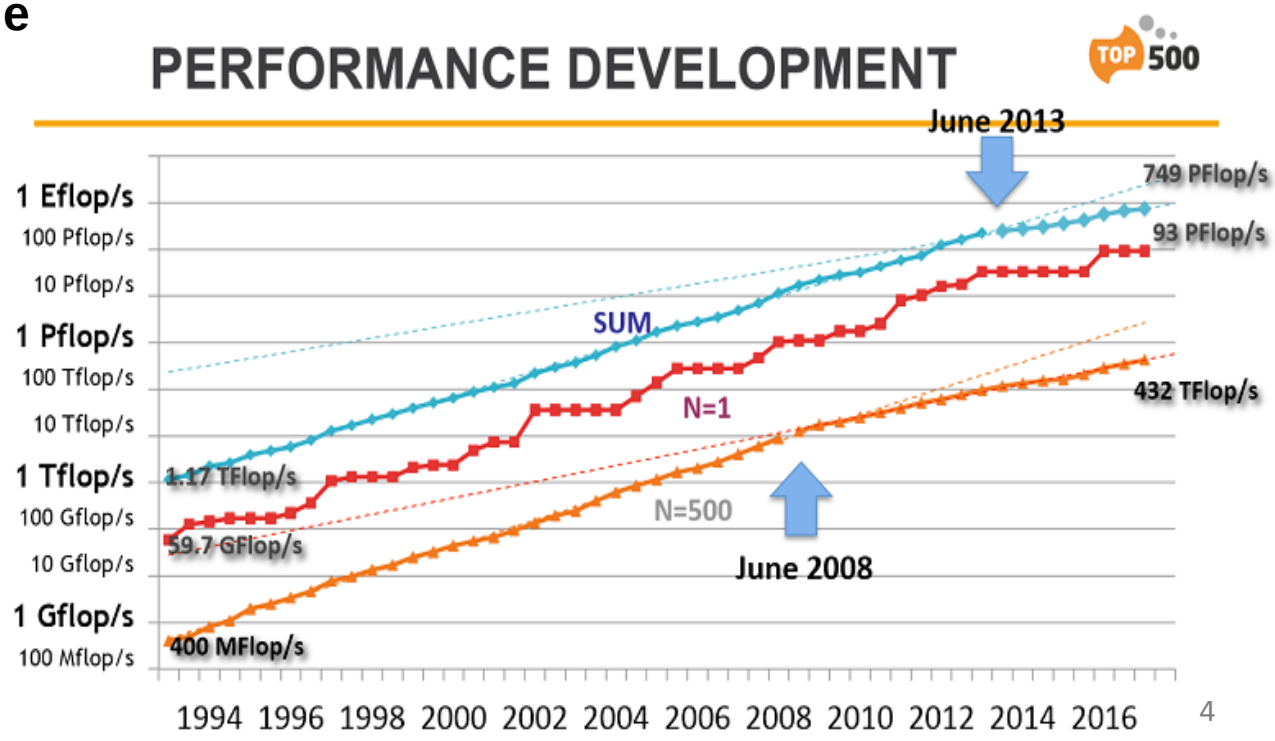
- **Theoretical Peak Performance** è una stima della performance di un componente Hardware (unità di calcolo, memoria, rete, storage) in base alle caratteristiche tecnologiche.
- **Sustained Performance** (Throughput): Prestazioni effettive misurate, di un componente hardware o di un sistema di calcolo tramite l'esecuzione di specifici programmi detti **Benchmark.**

CPU

La performance di una CPU è data principalmente dalla quantità di operazioni svolte nell'unità di tempo. In passato si consideravano tutte le istruzioni del processore e si misurava in MIPS (milioni di istruzioni eseguire per secondo) ora vengono valutate solo le operazioni in virgola mobile e si misura in MFLOPS o GFLOPS (operazioni in virgola mobile per secondo)

Le prestazioni di un core di calcolo (operazioni in virgola mobile al secondo, FLOPS) sono determinate dal numero di cicli di clock per secondo (Clock) e dal numero di operazioni f.p. per ciclo di clock (FLOPs/cycle) che il core può eseguire:
FLOPS=Clock x FLOPs/cycle

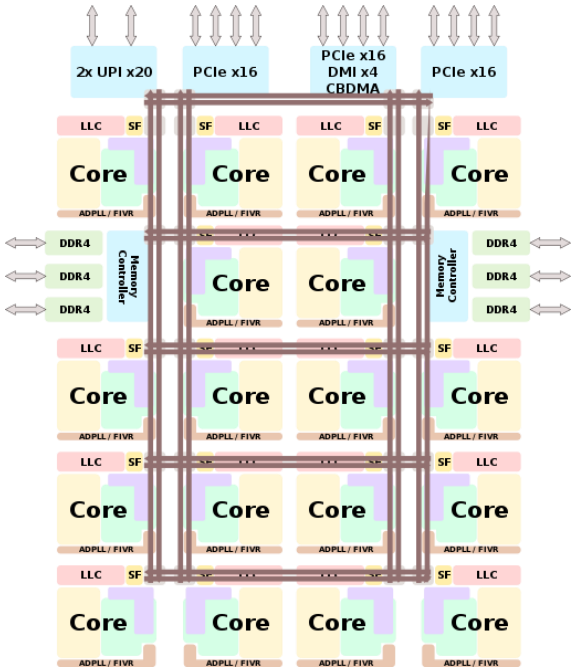
Il numero di operazioni per ciclo dipende anche dalla **precisione** del dato in virgola mobile che può essere:
- Singola (SP, 32bit)
- Doppia (DP, 64bit)
- Mezza (HP, 16bit)



Esempio CPU: Intel Xeon E5-6140

Questo processore è utilizzato all'interno di diversi nodi di calcolo del cluster HPC.
[Scheda tecnica Xeon E5-6140]

- Numero core: 18
- Clock: 2.3 GHz (3.7 GHz turbo mode)
- TDP: 140 W



Le operazioni Floating Point richiedono diversi cicli di clock ma possono generare un risultato per ciclo di clock grazie all'architettura pipeline
https://it.wikipedia.org/wiki/Pipeline_dati
FLOPs/cycle = 1
Theoretical peak performance:
Clock x FLOPs/cycle = 2.3 -> 3.7 GFLOPS

Peak Performance per processore:
Da 2.3 x 18 GFLOPS a 3.7 x 18 GFLOPS

Istruzioni vettoriali (SIMD) e FMA

Le principali applicazioni con elevato carico computazionale sono «Data Parallel» ovvero richiedono l'esecuzione di singole operazioni su diversi dati (SIMD - Single Instruction Multiple Data) organizzati in array multidimensionali come ad esempio il rendering grafico, il calcolo scientifico e calcolo tensoriale nelle reti neurali.
Per accelerare queste operazioni i processori Intel e AMD hanno inserito nei processori un set di istruzioni aggiuntive che si appoggiano su registri dedicati possono eseguire una istruzione floating point su N dati contemporaneamente.
Nei processori INTEL la dimensione di questi registri era inizialmente di 128 bit (SSE – Streaming SIMD Extensions) per poi aumentare fino agli attuali 512 bit (AVX-512).
Un core con AVX-512 può eseguire contemporaneamente 16 Flops in Singola precisione o 8 in doppia.

Inoltre, poiché le operazioni sono frequentemente delle moltiplicazioni vettoriali, è stata aggiunta una ulteriore istruzione dedicata, **FMA (Fused Multiply-Add)** che in un solo ciclo di clock esegue 2 operazioni, somma e moltiplicazione.

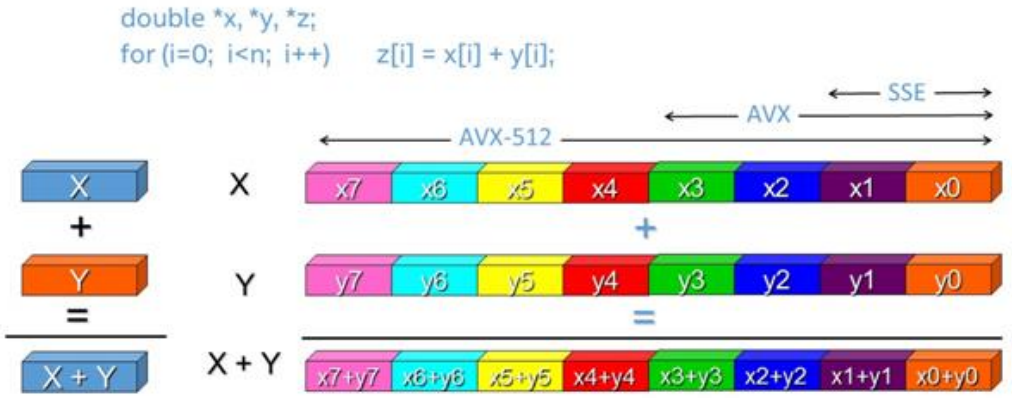


Figure 1 Scalar and vectorized loop versions with Intel® SSE, AVX and AVX-512.

Intel Xeon E5-6140 con SIMD e FMA

Tenendo conto di queste estensioni le prestazioni di picco di un core Intel Xeon E5 sono:

FMA (Fused Multiply-Add) : 2 flops
SIMD AVX-512: 16 flops s.p. - 8 flops d.p.

Peak performance in doppia precisione per core
2.3*2*8 GFlops = 36 Gflops (base)
3.7*2*8 GFlops = 59 Gflops (max)

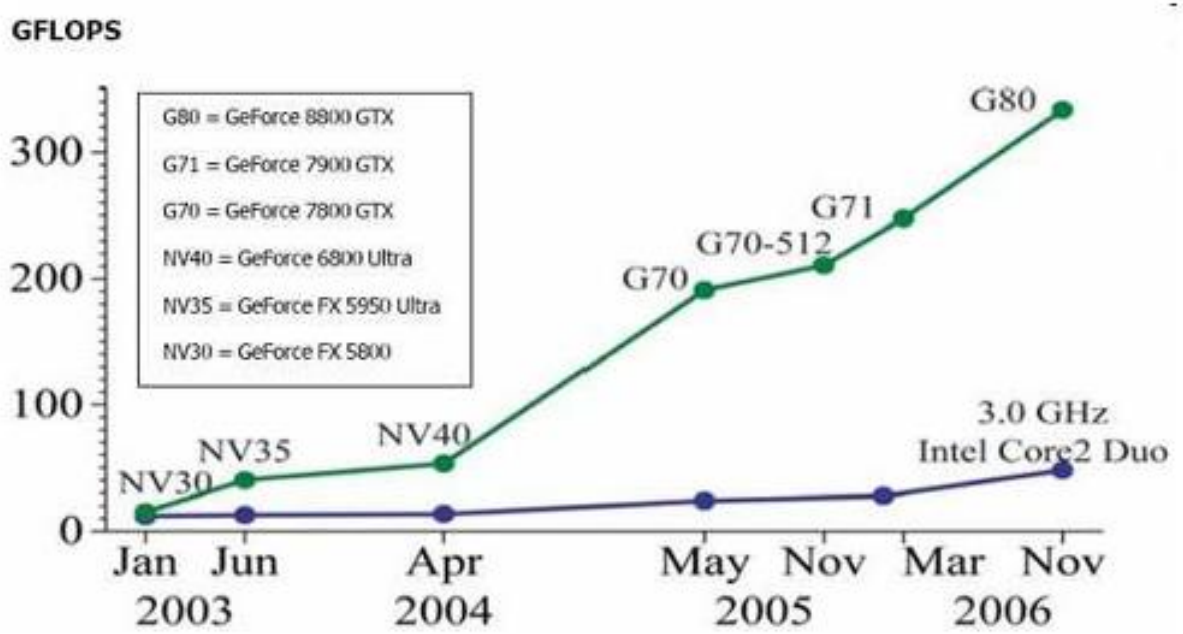
Peak Perf per processore (18 cores):
1 TFlops in doppia precisione
2 TFlops in singola precisione

GPU

Le GPU sono nate come processori specializzate per la grafica 3D sotto la spinta delle esigenze dei Videogame.

La grande diffusione ha permesso una drastica riduzione dei costi e incremento delle prestazioni.

Il costruttore di GPU NVIDIA ha realizzato diversi modelli di GPU che possono essere utilizzate come acceleratore del calcolo per applicazioni parallele.

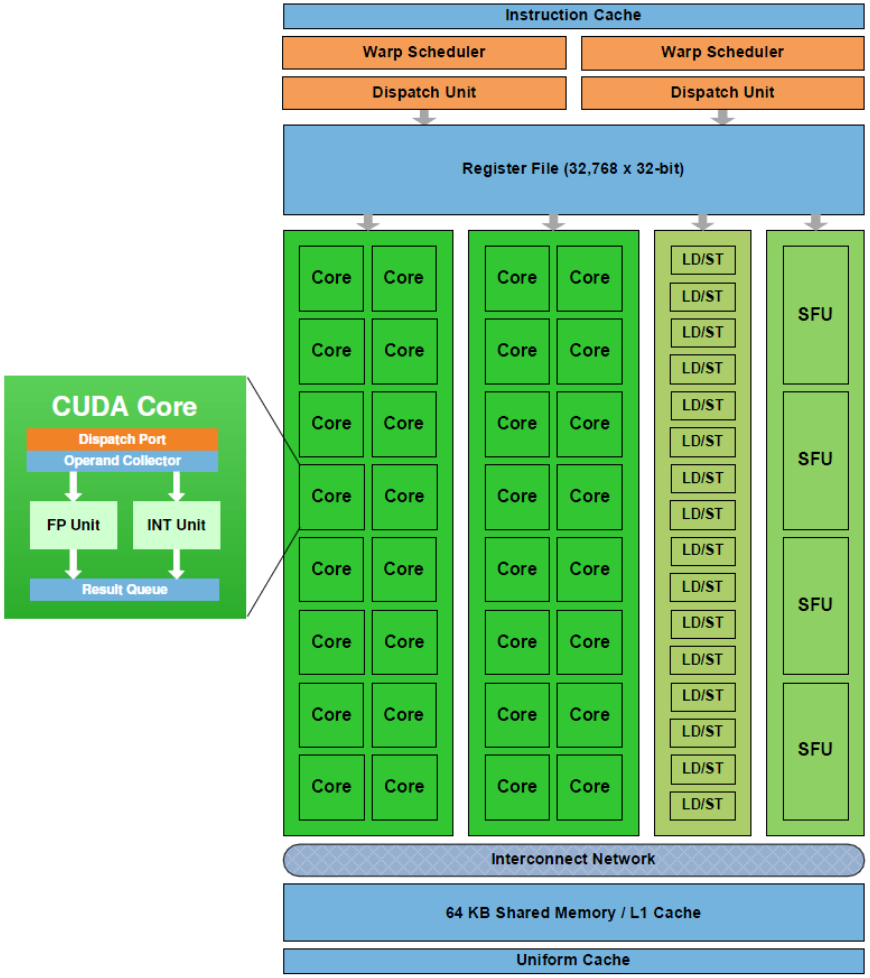


Esempio GPU: NVIDIA A100

[Scheda tecnica A100]

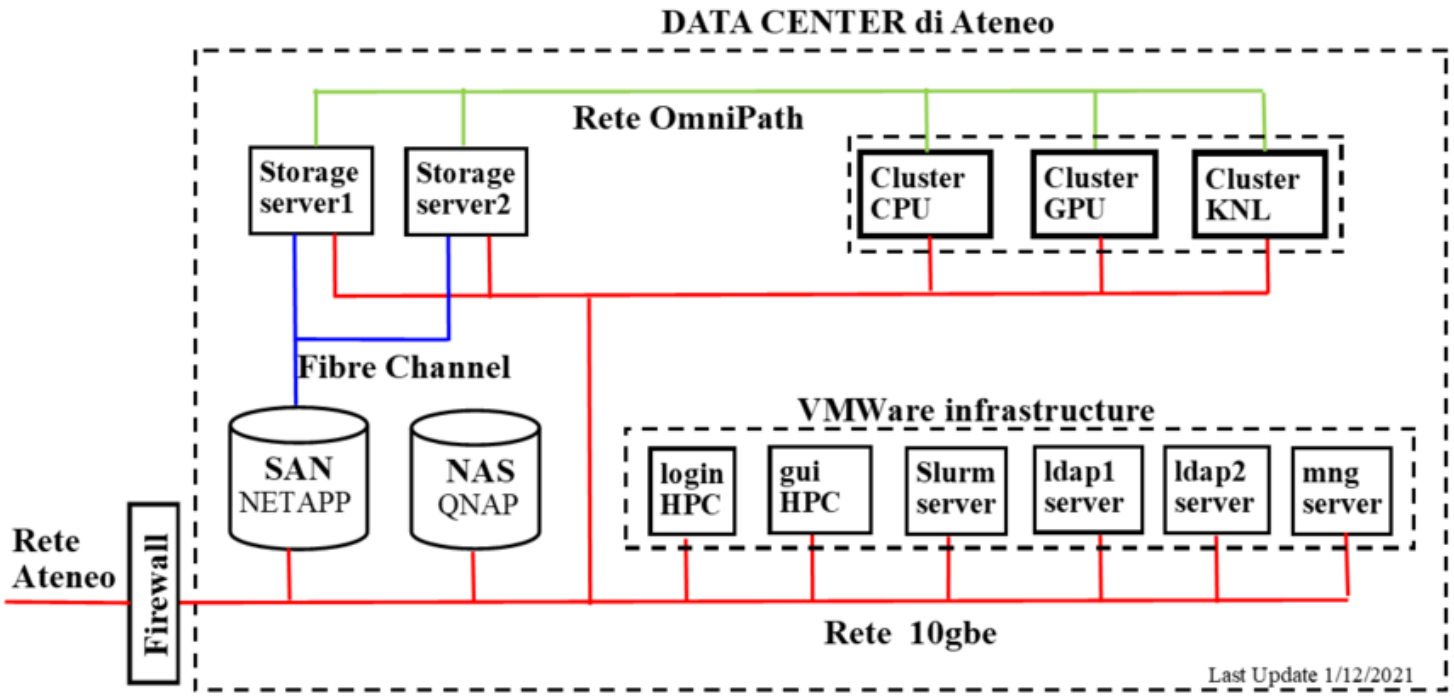
6912 CUDA Cores (s.p.) per GPU
Frequenza Base: 765 MHz - Boost: 1410 MHz
TDP 400 watt

Prestazioni di picco (s.p.)
6912 core * 2 ops (FMA) * 1410 MHz =
19.5 TFlops



Confronto P100 V100 A100:
https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/

CPU e GPU del cluster HPC.unipr.it



nodi CPU

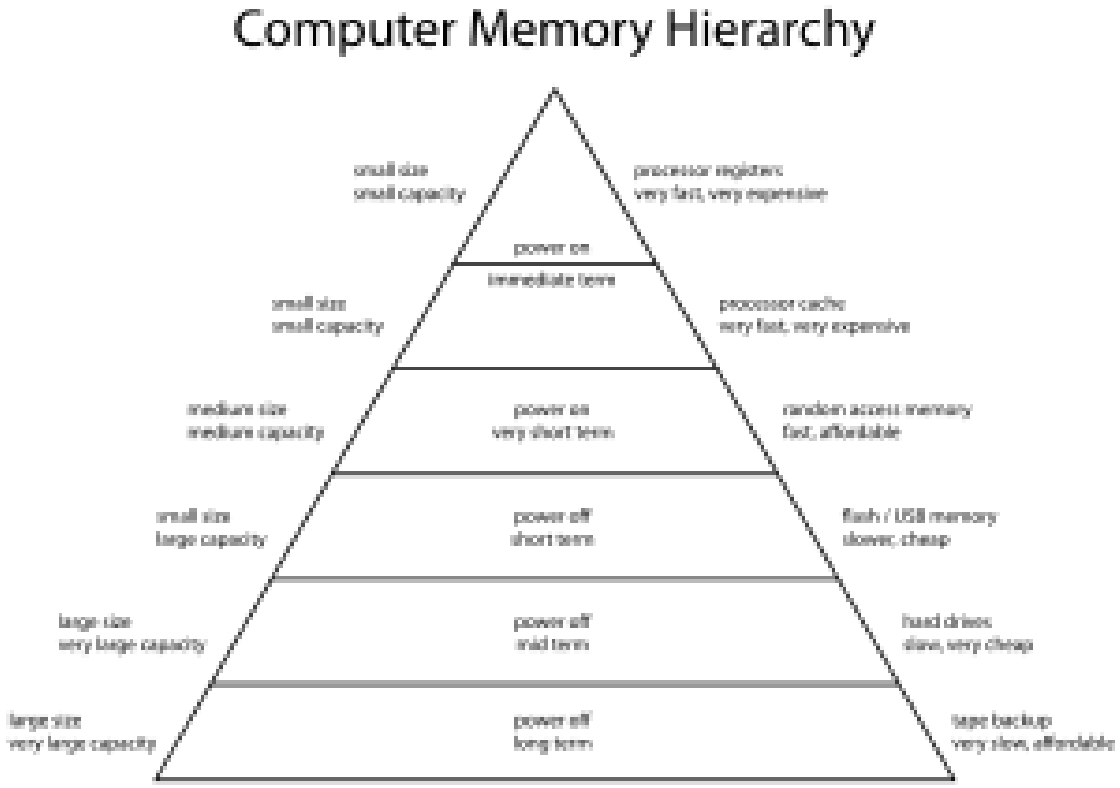
22 DualBDW: 11 Tflops
1 QuadBDW 1 TFlops
8 QuadSKL: 22 Tflops
4 KNL 6 Tflops
40 Tflops d.p.

schede GPU

14 NVIDIA P100: 66 TFlops
2 NVIDIA V100: 14 Tflops
4 NVIDIA A100:39 Tflops
119 Tflops d.p.

Memoria

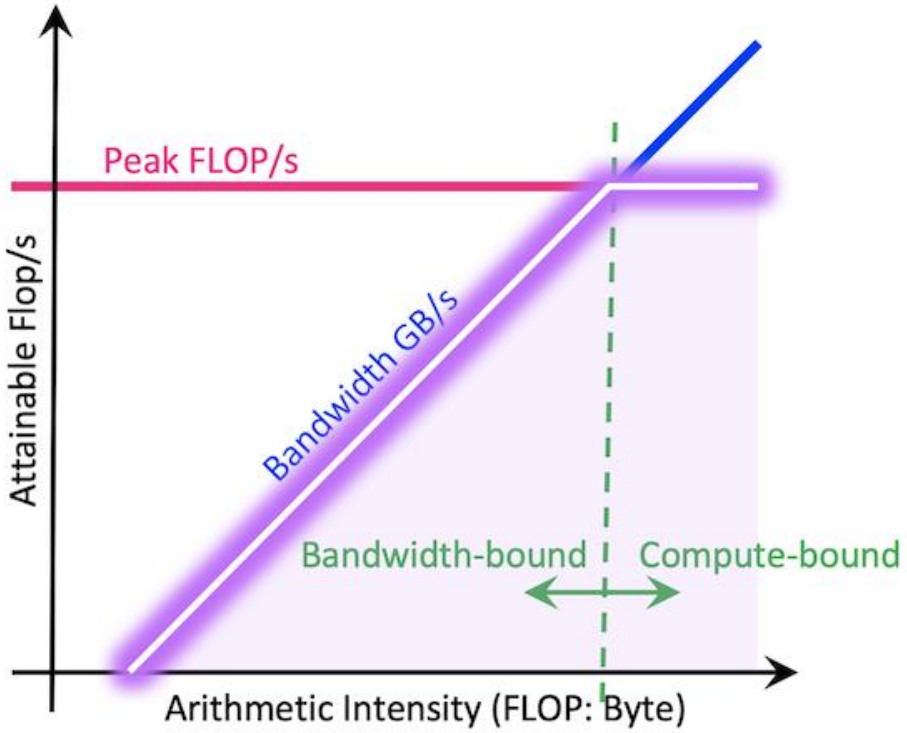
La gerarchia della memoria determina tempi di accesso differenti a seconda della localizzazione del dato.

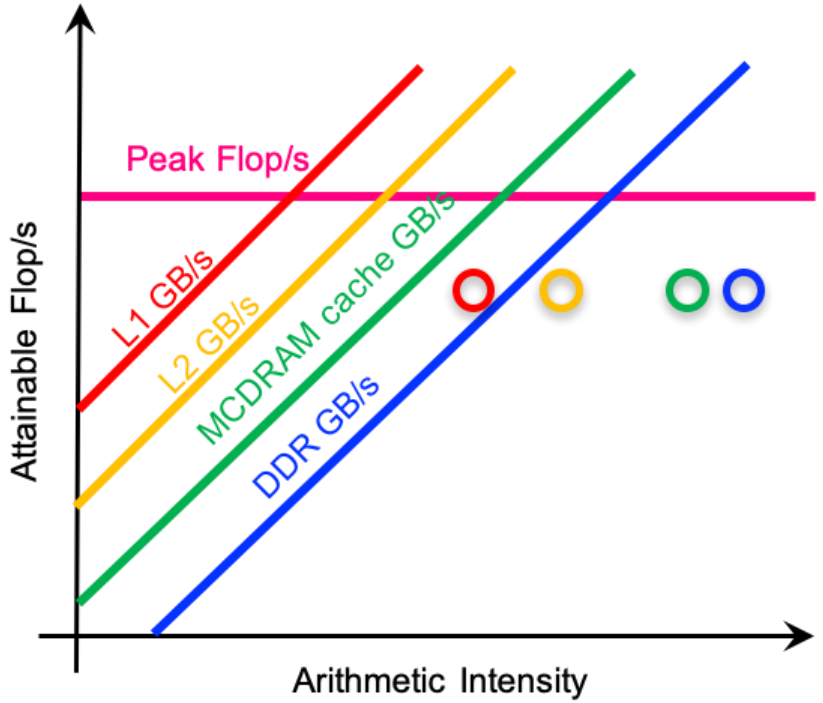


La memoria può diventare un collo di bottiglia nelle prestazioni se non è in grado di fornire dati con il ritmo richiesto dal processore.

RoofLine Analysis

Il modello Roofline consente in maniera intuitiva di stimare le performance di un kernel computazionale mostrando graficamente le limitazioni inerenti CPU/GPU e memoria.
FLOP/s = min(peak FLOP/s, Peak Memory bandwidth x Aritmetic Intensity)
Dove Arithmetic Intensity = Total FLOPS/Total Bytes



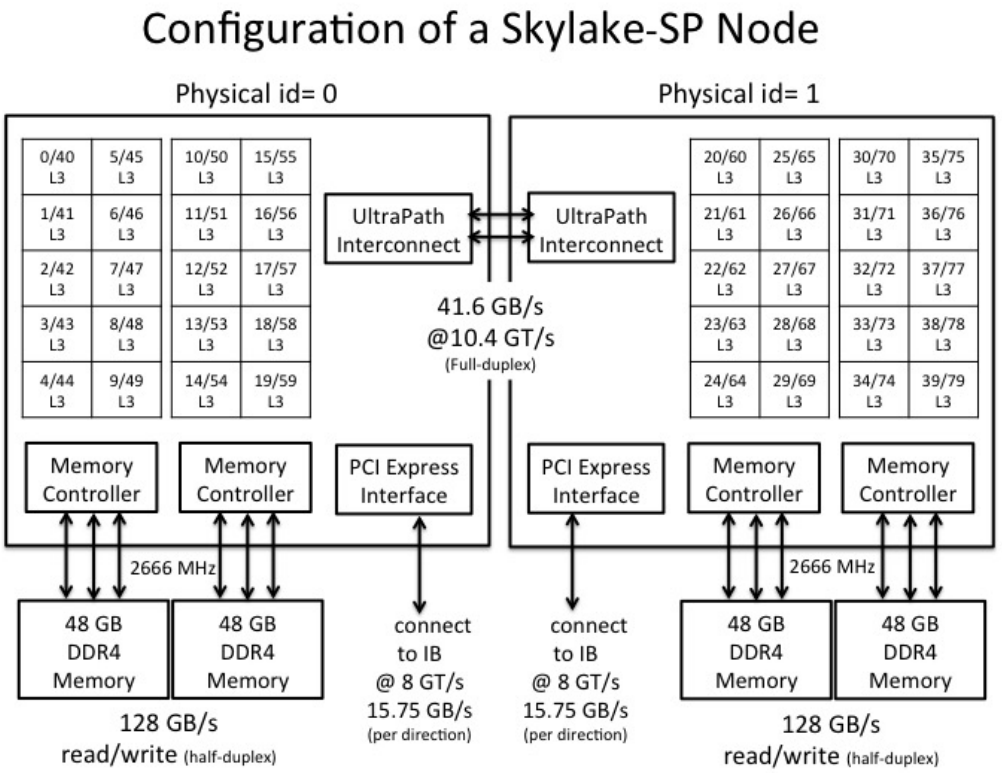


https://docs.nersc.gov/tools/performance/roofline/

Esempio CPU Xeon E5-6140

[Skylake Processors]

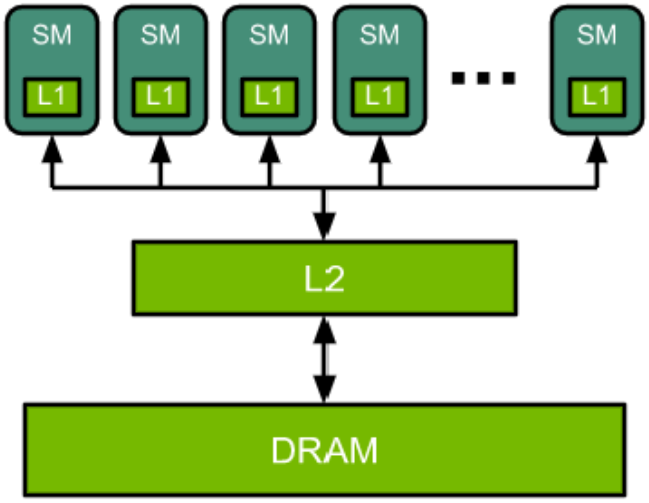
Cache L1	32+32KB (per core)
Cache L2	1 MB (per core)
Cache L3	24,75 MB (per socket)
RAM DDR	384 GB (per node)

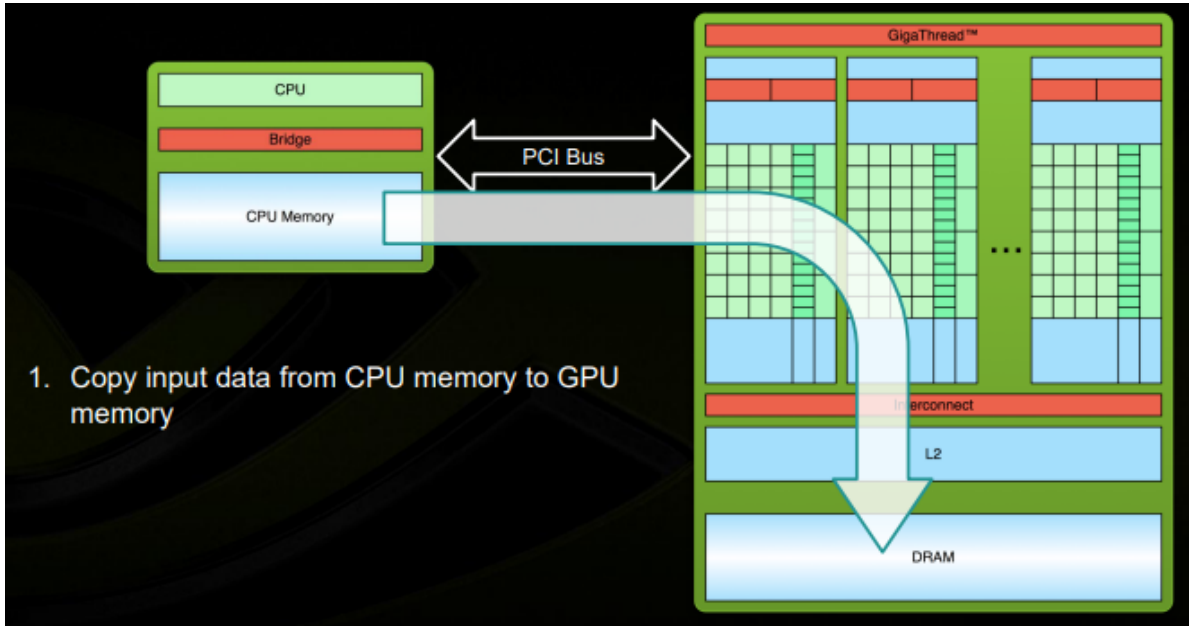


Esempio GPU A100

La GPU possiede una gerarchia interna di memoria a cui si aggiunge la comunicazione con l'host.

Esempio NVIDIA A100
L1 Cache 192 KB/SM
L2 Cache 40 MB
DRAM 40 GB (1.5 TB/s)

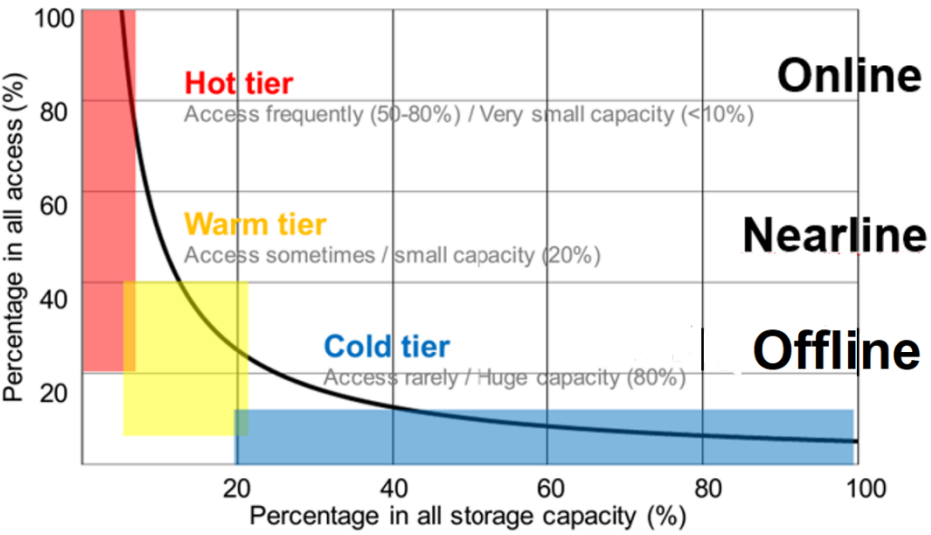




Confronto P100 V100 A100:
https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/

Storage

Nei sistemi HPC esistono diverse tipologie di esigenze di storage (online, nearline , offline) a cui corrispondono dispositivi con diverse caratteristiche (prestazioni , capacità e costi)



Tecnologie storage Esempi	Capacità TB	Prestazioni MB/s (tipico)	Costo K€
SSD	7.6	600/800	2.7
HDD	20	280	0.6
TAPE Cartridge	30	3.6TB/hour	0.1

valori tipici , costi indicativi

NAS (Network Attached Storage)

Un **Network Attached Storage (NAS)** è un dispositivo collegato alla rete la cui funzione è quella di consentire agli utenti di accedere e condividere una memoria di massa, in pratica costituita da uno o più dischi rigidi, all'interno della propria rete o dall'esterno. (Wikipedia)

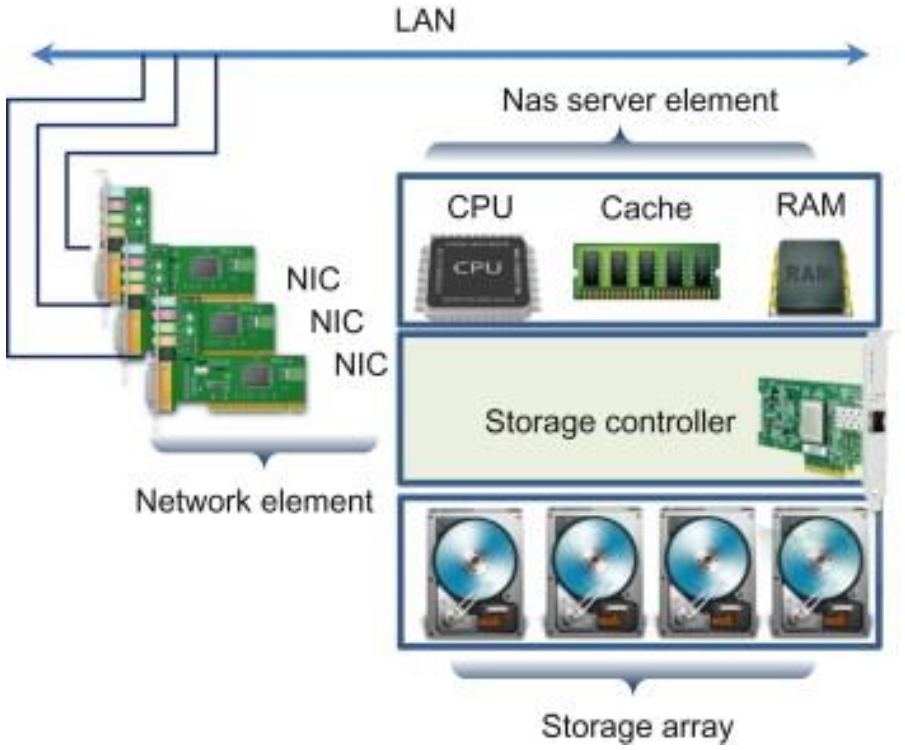
L'accesso ai file avviene tramite specifici protocolli di rete come NFS, SMB e iSCSI.

Vantaggi:

- Condivisione dati
- Gestione centralizzata
- Basso costo

Svantaggi:

- Basse prestazioni
- Risorse limitate

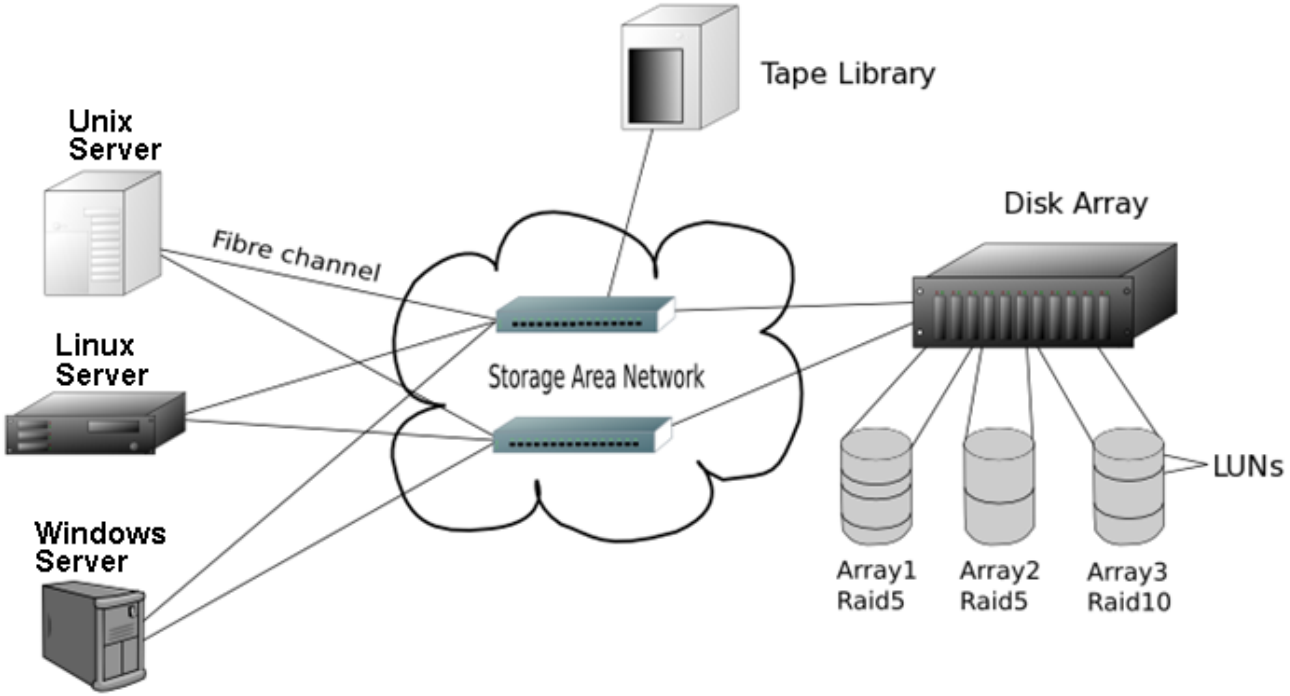


SAN (Storage Area Network)

Una SAN è una rete ad alta velocità di trasmissione costituita esclusivamente da dispositivi di memorizzazione di massa, in alcuni casi anche di tipi e tecnologie differenti. Il suo scopo è quello di rendere tali risorse di immagazzinamento (storage) disponibili per qualsiasi computer connesso ad essa. (wikipedia)

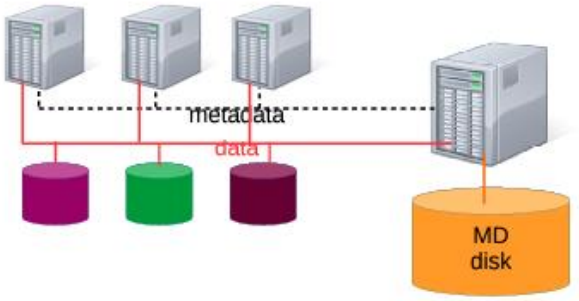
Vantaggi
- prestazioni
- Scalabilità
- Ridondanza

Svantaggi
- gestione complessa
- costo



File System per cluster HPC

Un file system per un cluster HPC deve avere caratteristiche avanzate quali:



Shared filesystem: Separazione dati e metadati.
Gestione centralizzata dei metadati.
Diversi nodi hanno accesso diretto ai dischi condivisi.

Clustered filesystem: tutti i dischi vengono utilizzati contemporaneamente da tutti i nodi

Parallel filesystem: il singolo file viene suddiviso in blocchi che vengono distribuiti su tutti i dischi del file system (striping)

Byte range locking: accesso concomitante di più utenti allo stesso file

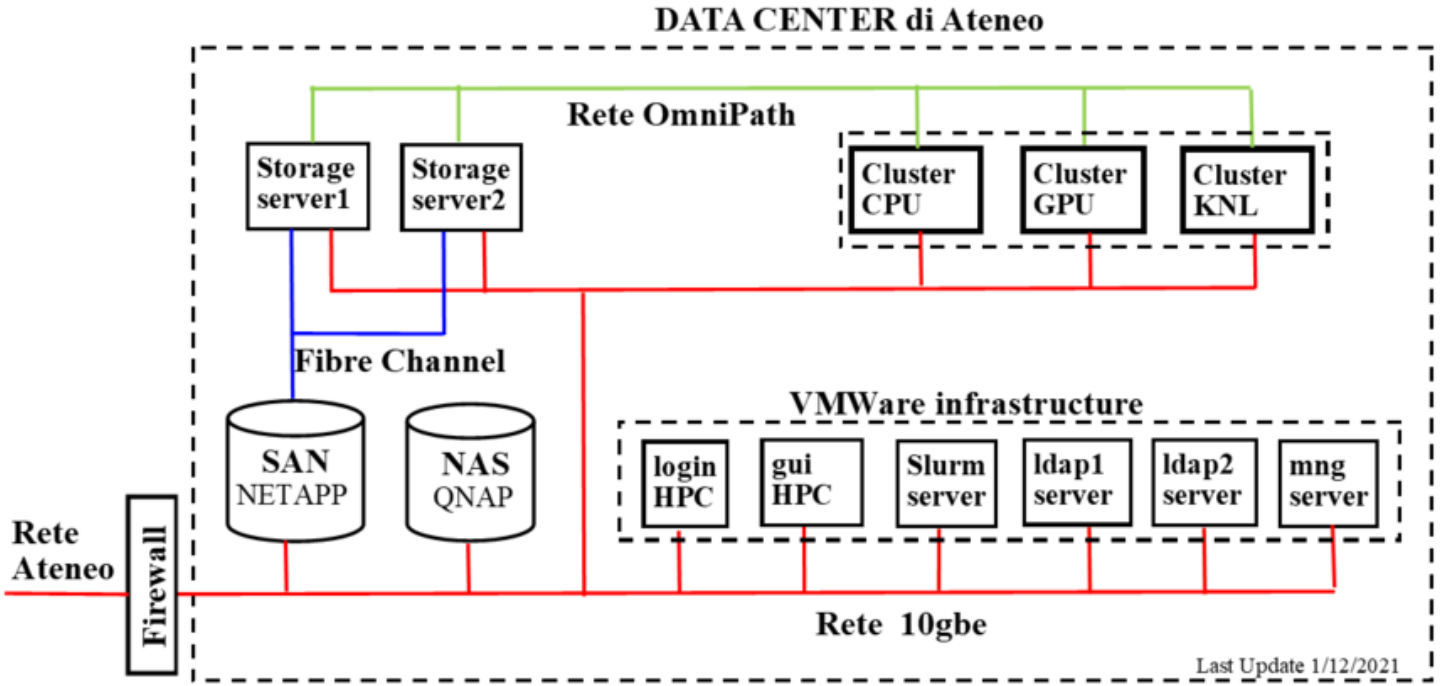
Tiering: permette di definire gerarchie di storage con diverse prestazioni

I file system più diffusi per cluster HPC sono

- GPFS (IBM Spectrum Scale) https://en.wikipedia.org/wiki/GPFS
- Lustre https://en.wikipedia.org/wiki/Lustre_(file_system)

Vedi IBM-Spectrum-Scale-Concepts-and-features.pdf nel materiale didattico.

Lo storage di HPC.unipr.it



/hpc/home	**20 TB**	**SAN GPFS HDD**
/hpc/group	**50 TB**	**SAN GPFS HDD**
/hpc/archive	**176 TB**	**SAN GPFS HDD nearline**
/hpc/scatch	**46 TB**	**SAN GPFS HDD + SSD**

High Speed Networks

	Bitrate (Gb/s)	Bandwidth (GB/s)	Latency (microsec.)	Costo scheda (K€)
GbEthernet (tcp/ip)	1	0.1	47	0.03
10GbEthernet (tcp/ip)	10	0.9	13	0.1
Intel OmniPath	100	12	1	1
Infiniband Mellanox EDR	100	12	1	1

valori tipici, costi indicativi

Referenze:

- https://www.hpcadvisorycouncil.com/pdf/IB_and_10GigE_in_HPC.pdf
- https://agenda.infn.it/event/13040/contributions/17299/attachments/12506/14064/INFN_CCR_2017_-_DELLEMC_v2.pdf

Benchmarks

Con il termine benchmark si intende un insieme di test software volti a fornire una misura delle prestazioni reali (sustained performance) di un computer per quanto riguarda diverse operazioni.

SPEC (Standard Performance Evaluation Corporation) è una organizzazione no-profit che produce e mantiene performance benchmark per computers (Wikipedia)
Il Benchmark più recente per le CPU è SPEC CPU2017
https://www.spec.org/cpu2017/

I **Benchmark LINPACK** sono utilizzati per misurare le prestazioni dei computer nelle operazioni in virgola mobile. LINPACK è una libreria software sviluppata per eseguire operazioni di algebra lineare. Vedi Wikipedia
Nell'HPC viene utilizzato l' **High Performance Linpack**, una versione portabile del Benchmark LINPACK che viene utilizzato per stilare la classifica TOP500.

time, gprof e clock_gettime()

Il comando **time** ritorna i tempi di esecuzione di un programma. Esempio:
> time **sleep 1**
real 0m1.003s # tempo reale di esecuzione (wall clock time)
user 0m0.000s # tempo di utilizzo della CPU nello stato User
sys 0m0.003s # tempo di utilizzo della CPU nello stato Kernel

gprof è il profiler del progetto GNU. Un profiler consente di determinare quali parti del programma consumano più tempo.
Per utilizzarlo occorre compilare con l'opzione –pg
Al momento dell'esecuzione viene generato il file gmon.out che potrà poi essere analizzato con il comando gprof
Riferimenti: https://users.cs.duke.edu/~ola/courses/programming/gprof.html

La funzione **clock_gettime()** consente di determinare i tempi di esecuzione all'interno di un programma.
E' possibile determinare il wall clock time (CLOCK_REALTIME) oppure il tempo di utilizzo della CPU (CLOCK_PROCESS_CPUTIME_ID)
Riferimenti: https://people.cs.rutgers.edu/~pxk/416/notes/c-tutorials/gettime.html

pandas e matplotlib

pandas https://pandas.pydata.org/ è una libreria software scritta per il linguaggio di programmazione Python per la manipolazione e l'analisi dei dati.

matplotlib https://matplotlib.org/ è una libreria per la creazione di grafici per il linguaggio di programmazione Python progettata per assomigliare a quella di MATLAB.

L'utilizzo congiunto di **pandas e matplotlib** consente di creare semplici script python per la gestione e la visualizzazione dei dati prodotti dall'esecuzione dei programmi di calcolo.
Riferimenti: https://ourcodingclub.github.io/tutorials/pandas-python-intro/