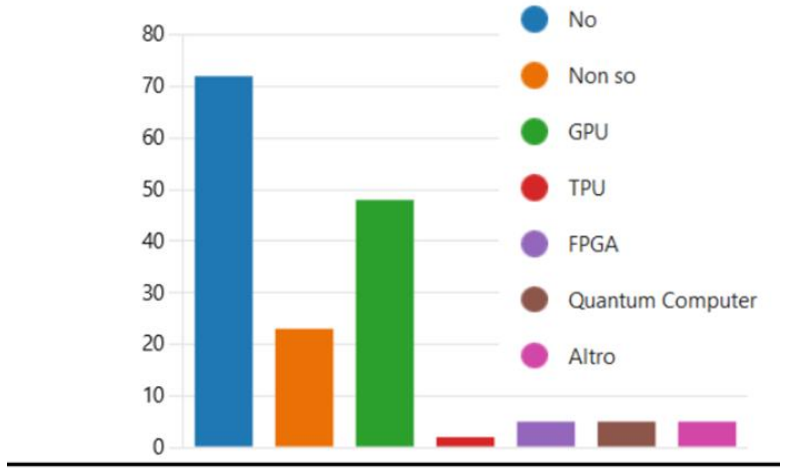
UniNuvola-GPU

Fondo Ricerca di Ateneo, edizione 2022

Sondaggio Ateneo marzo-aprile 2023

Emerge l’esigenzadi calcolo ad alte prestazioni, potenzialmente basato su GPU

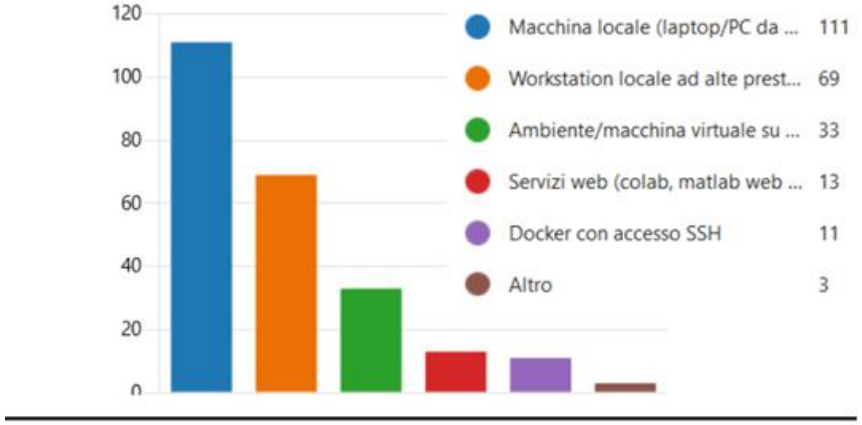
Utilizzi macchine dotate di architetture di calcolo specifiche?

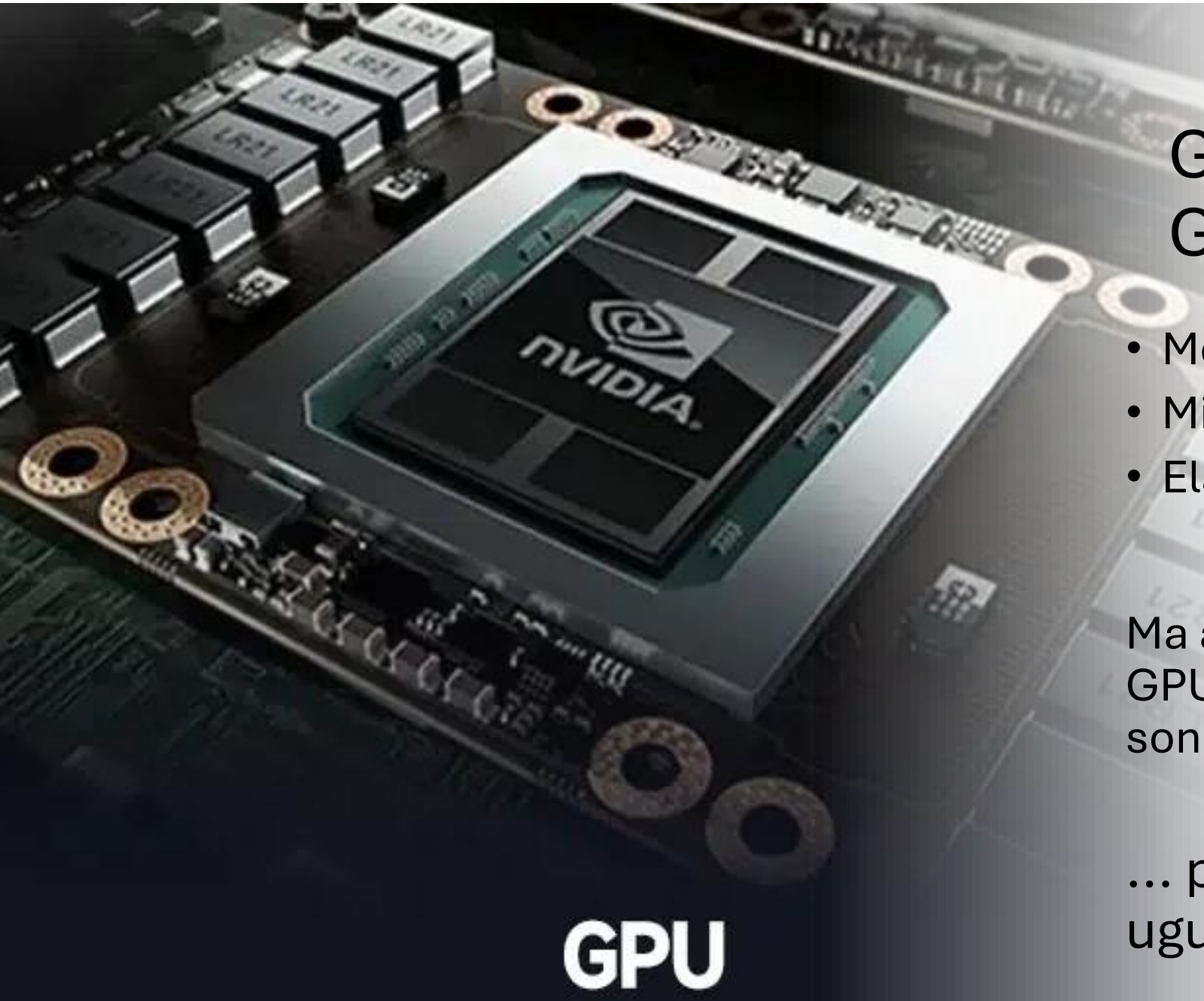


Quanta RAM/VRAM è necessaria per le tue sessioni di calcolo in media?

meno di 1 GB	10
Tra 1 GB e 10 GB	49
Più di 10 GB	63
Non so	25

Che tipo di dispositivo/servizio utilizzi per eseguire le tue sessioni di calcolo



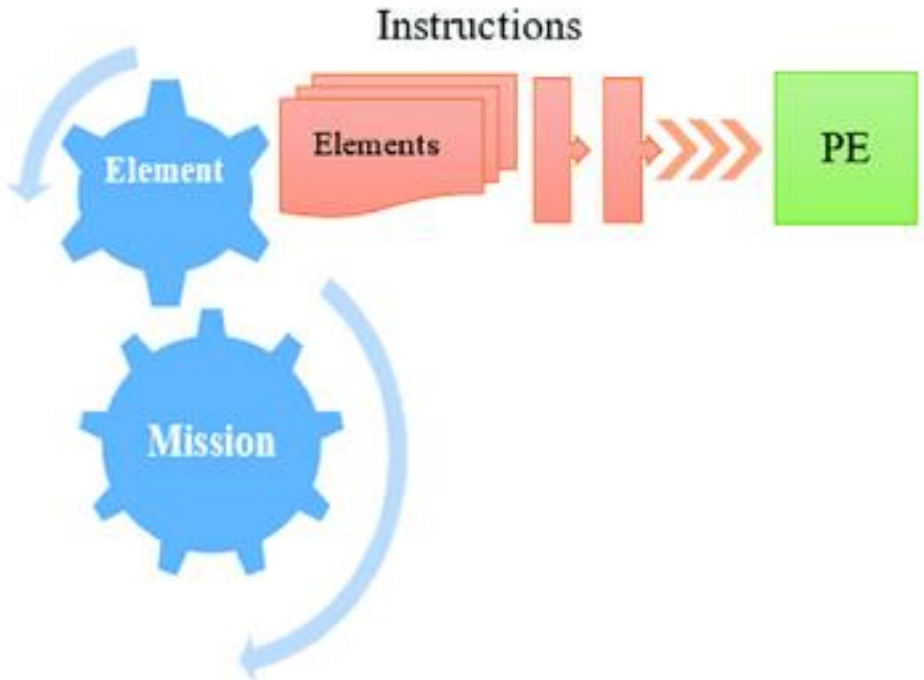


GPU
Graphic Processing Unit

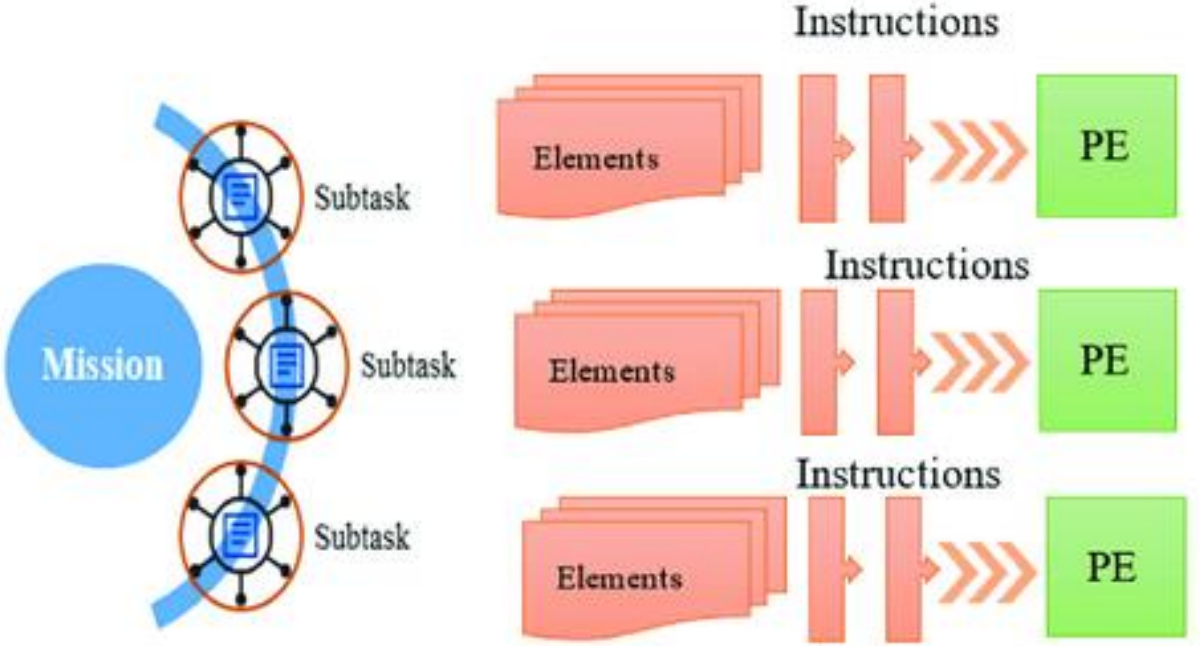
- Motore di calcolo.
- Microprocessore basato su chip.
- Elabora dati.

Ma a differenza di CPU standard, le GPU hanno una architettura diversa e sono progettate per scopi diversi...

... perché i calcoli non sono tutti uguali !



a. **Serial computing**



b. **Parallel computing**

$$f_0 = 1$$
$$f_1 = 1$$
$$f_n = f_{n-1} + f_{n-2}$$

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & a_{i3} & \cdots & a_{in} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}}_{m \times n \text{ matrix } \mathbf{A}} \underbrace{\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1j} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2j} & \cdots & b_{2p} \\ b_{31} & b_{32} & \cdots & b_{3j} & \cdots & b_{3p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nj} & \cdots & b_{np} \end{bmatrix}}_{n \times p \text{ matrix } \mathbf{B}} = \underbrace{\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1j} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2j} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{i1} & c_{i2} & \cdots & c_{ij} & \cdots & c_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mj} & \cdots & c_{mp} \end{bmatrix}}_{m \times p \text{ matrix } \mathbf{C}}$$

GPU (Graphics Processing Unit):

- **Parallel Processing**: Le GPU sono progettate per eseguire operazioni di calcolo massicciamente parallele. Sono originariamente progettate per l'elaborazione grafica, che richiede calcoli paralleli su grandi volumi di dati.
- **Cores**: Le GPU **hanno migliaia di core**, molto semplici e specializzati nel calcolo parallelo.
- **Clock Speed**: Le GPU hanno velocità di clock inferiori rispetto alle CPU, ma compensano con un numero molto maggiore di core.
- **Memory Bandwidth**: Le GPU dispongono di una maggiore larghezza di banda della memoria per gestire grandi quantità di dati necessari per le operazioni parallele.

GPU

- **Parallel Processing**: Le GPU sono progettate per eseguire operazioni di calcolo massicciamente parallele. Sono originariamente progettate per l'elaborazione grafica, che richiede calcoli paralleli su grandi volumi di dati.
- **Cores**: Le GPU **hanno migliaia di core**, molto specializzati per il calcolo parallelo.
- **Clock Speed**: Le GPU hanno velocità di clock inferiori rispetto alle CPU, ma compensano con un numero molto maggiore di core.
- **Memory Bandwidth**: Le GPU dispongono di una grande larghezza di banda della memoria per gestire grandi quantità di dati in parallelo durante le operazioni parallele.

CPU

General Purpose Computation

Cores: qualche decina

- **Clock Speed**: elevate velocità ottimizzate per calcolo a thread singolo.
- **Cache**: cache di grandi dimensioni e gerarchiche (L1, L2, L3) per ridurre la latenza di accesso alla memoria.

Throughput vs Latency

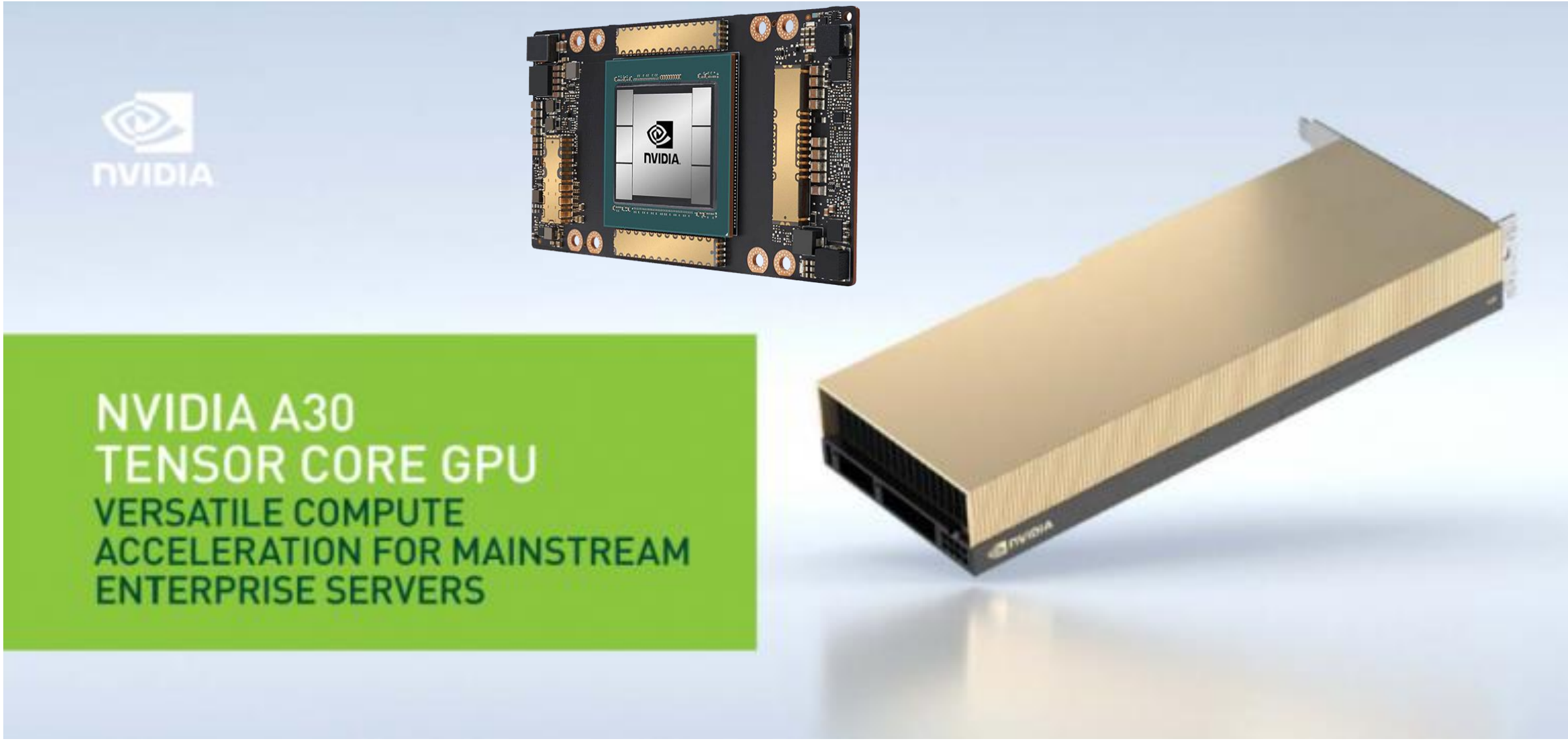
CPU

- **Latenza**: Le CPU sono progettate per avere una bassa latenza, rendendole ideali per operazioni che richiedono decisioni rapide e logica sequenziale.
- **Sequencial Operations**: Le CPU sono ottimizzate per gestire operazioni di input/output e per eseguire operazioni con dipendenze sequenziali.

GPU

- **Throughput**: Le GPU offrono un throughput molto elevato per operazioni che possono essere eseguite in parallelo, come le operazioni di matrice in machine learning e le trasformazioni grafiche.
- **Data Parallelism**: Le GPU sono estremamente efficienti per applicazioni che coinvolgono l'elaborazione di grandi blocchi di dati in parallelo.

Segmentabilità Ottimizzazione dei consumi Server Costo/prestazioni

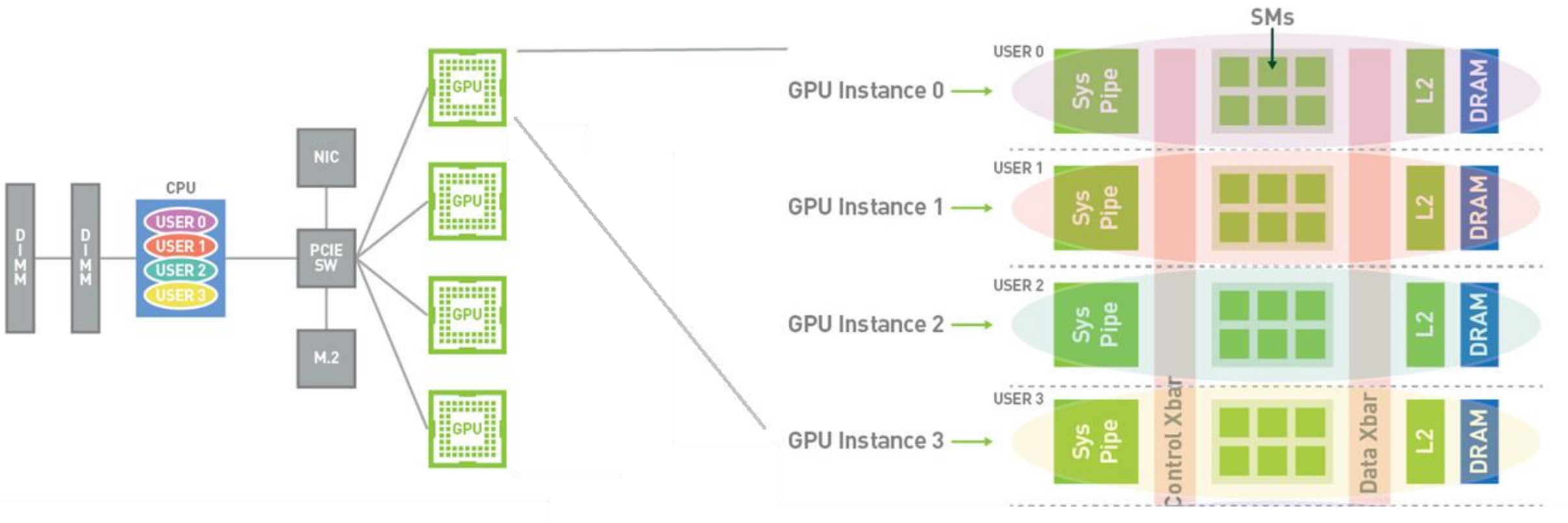


	NVIDIA A30
Power	165W
Form Factor	x16 PCIe Gen4 2 Slot FHFL 1 NVLink bridge
Memory	24GB HBM2
Memory Bandwidth	933 GB/s
Multi-Instance GPU	Up to 4
Media Acceleration	1 JPEG Decoder 4 Video Decoder
Fast FP64	Yes

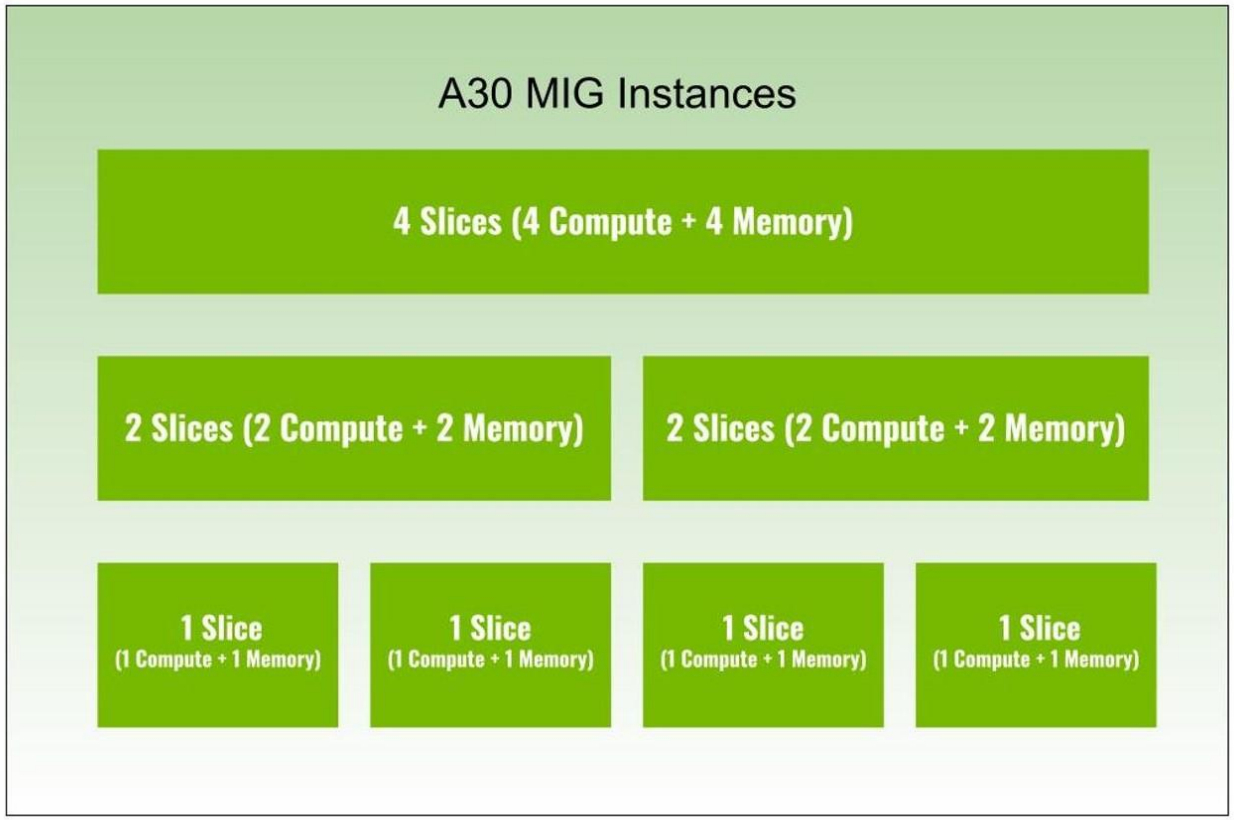
A30 supports MIG

- MIG = Multi-Instance GPU
- Permette di suddividere in modo sicuro le GPU in istanze GPU separate per le applicazioni CUDA, fornendo a più utenti risorse GPU separate per un'ottimale utilizzo della GPU.
- Particolarmente utile per i carichi di lavoro che non saturano la capacità di calcolo della GPU e quindi gli utenti possono voler eseguire diversi carichi di lavoro in parallelo per massimizzare l'utilizzo.

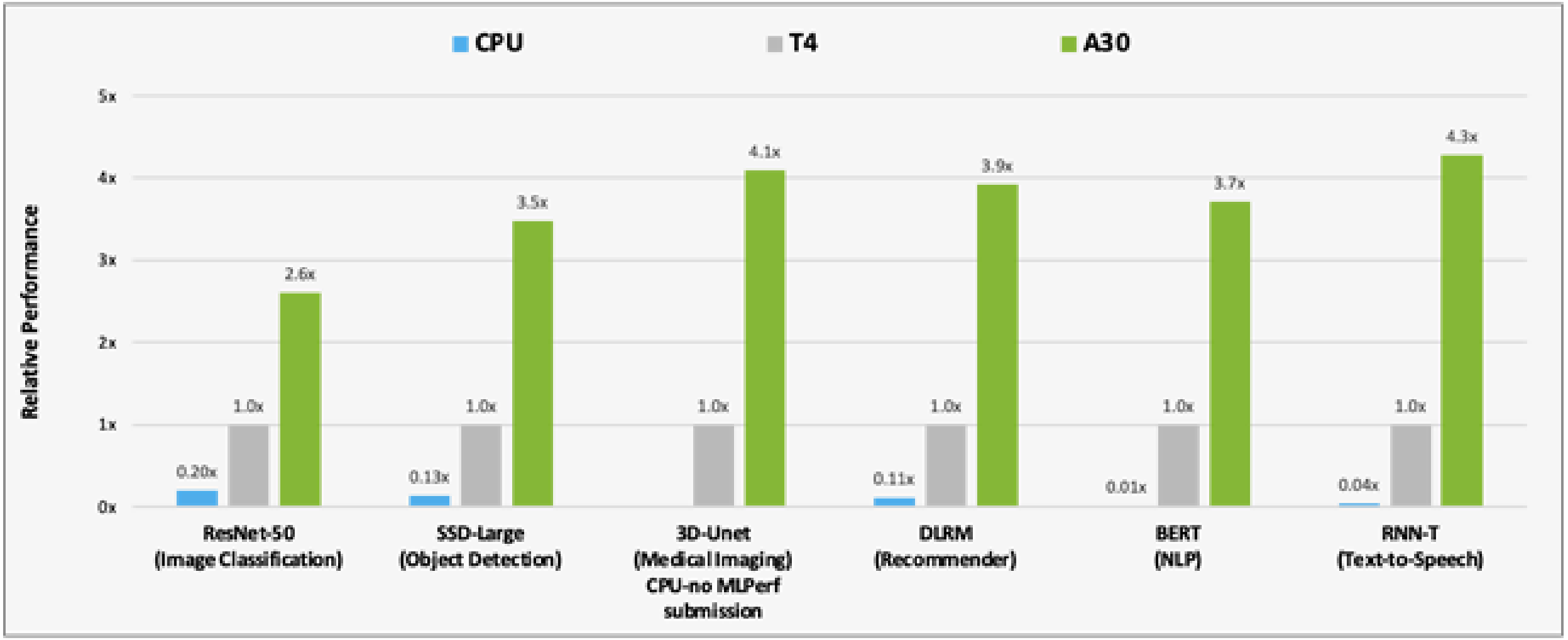
MULTI-INSTANCE GPU ("MIG")



Uninuvola-GPU = 6 x



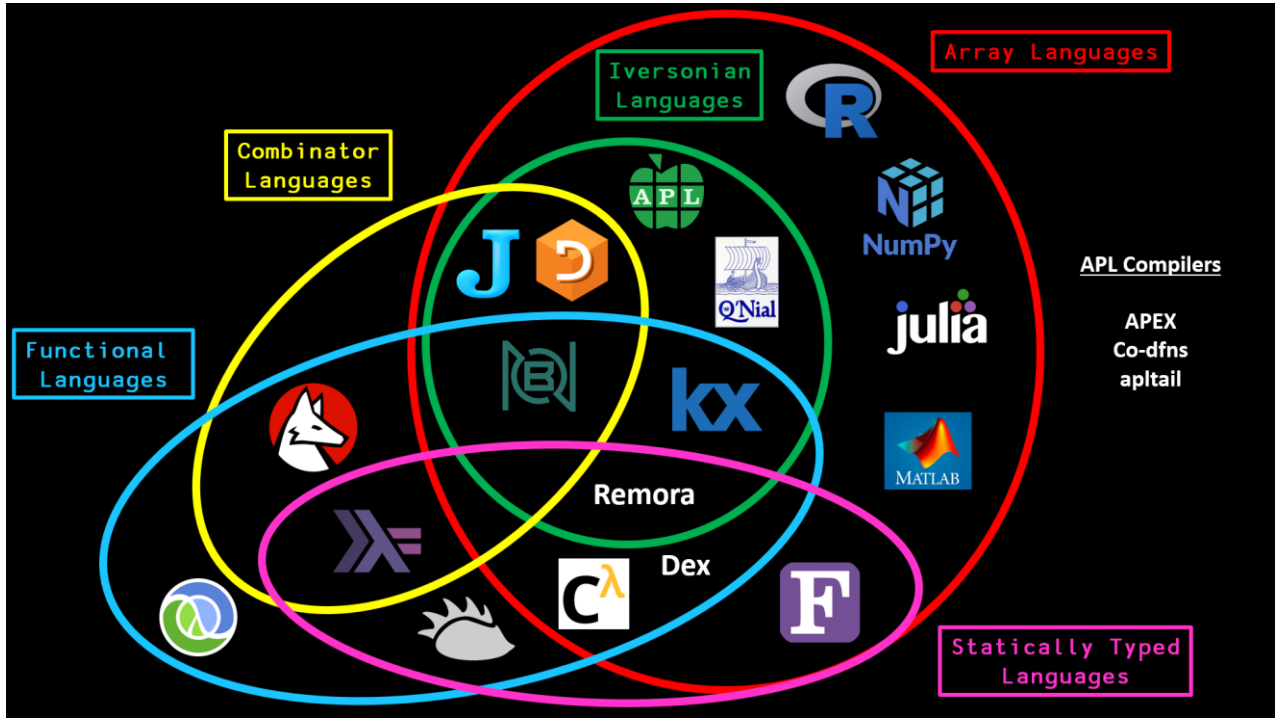
GPU Nvidia A30 performance increase



Caratteristiche ideali del codice

- **Parallelismo Elevato:** compiti indipendenti eseguibili indipendentemente
- **Elevata Computazione rispetto alla Comunicazione :** molte elaborazioni rispetto al volume di dati trasferiti tra la memoria e i core di calcolo.
- **Poca Dipendenza Sequenziale:** Le operazioni che dipendono fortemente dai risultati delle precedenti non si prestano bene all'elaborazione parallela.
- **Utilizzo Efficiente della Memoria:** Evitare accessi concorrenti agli stessi indirizzi di memoria (bank conflicts).
- **Divisione in Blocchi e Thread:** Il codice deve essere organizzato in blocchi e thread.
- **Registri e Memoria Condivisa:** Utilizzo efficiente dei registri e della memoria condivisa della GPU per minimizzare l'accesso alla memoria globale più lenta.
- **Riduzione del Trasferimento di Dati tra CPU e GPU:** può essere molto lento

Programming languages and libraries



Microsoft Cognitive Toolkit

MATLAB

mxnet

TensorFlow

torch

Wolfram Language

Caffe

PyTorch

Chainer

PaddlePaddle

Caffe2