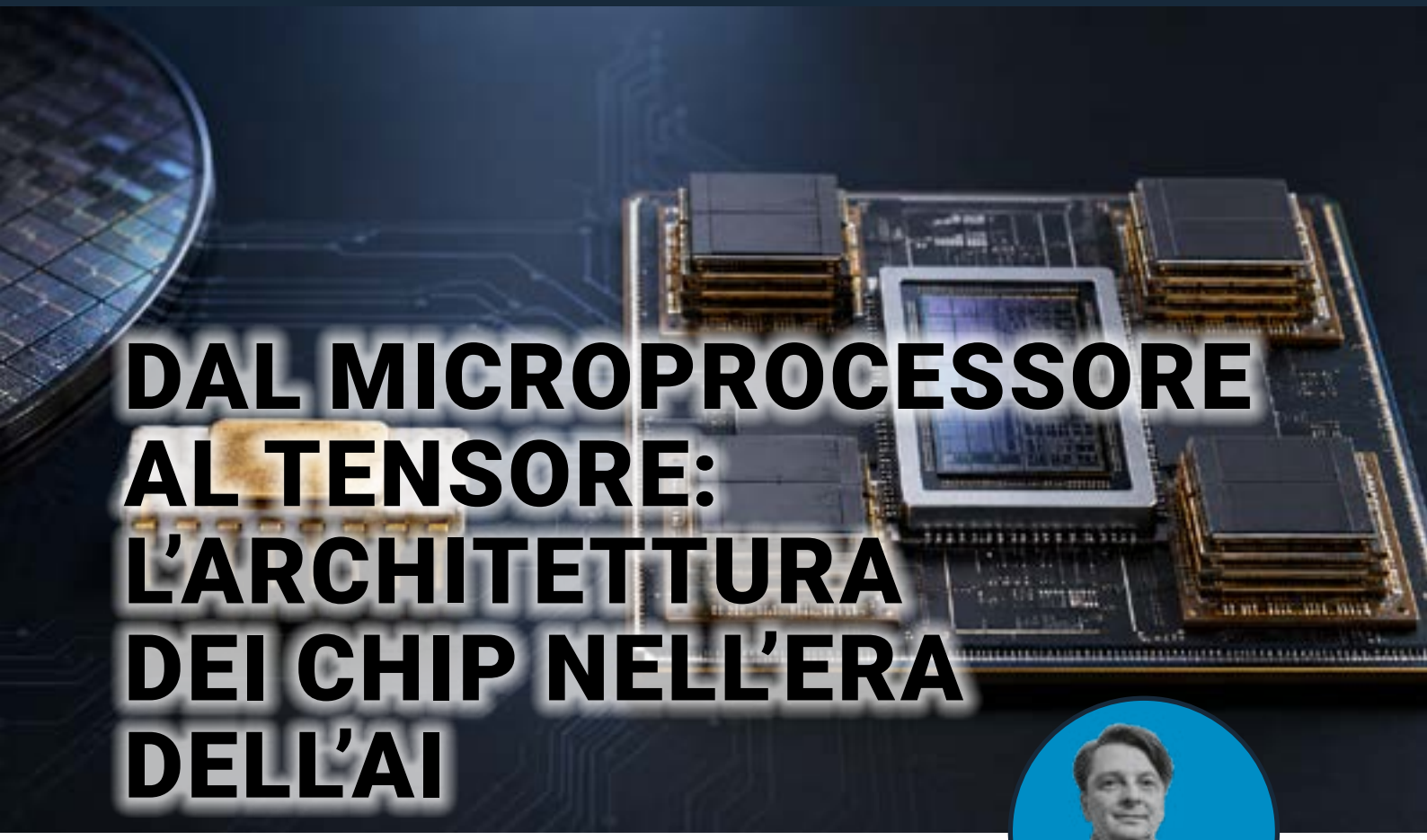


DAL MICROPROCESSORE AL TENSORE: L'ARCHITETTURA DEI CHIP NELL'ERA DELL'AI

a cura di Massimo Nannini (*)

Dalla CPU generalista agli acceleratori progettati per l'intelligenza artificiale, il silicio cambia architettura per sostenere il calcolo parallelo e tensoriale. L'articolo ripercorre l'evoluzione da GPU e CUDA a TPU, NPU e Tensor Core, spiegando il ruolo di array sistolici, precisione ridotta e memorie HBM. Un percorso che conduce ai sistemi eterogenei e all'Edge AI, dove CPU e unità specializzate collaborano per aumentare prestazioni, efficienza energetica e velocità di elaborazione.

SOMMARIO

- **Introduzione**
- **La CPU: il "*tuttofare*" e il collo di bottiglia di Von Neumann**
- **La Svolta del Calcolo Parallelo: Dalla Grafica 3D ad AlexNet**
- **L'Era degli Acceleratori AI Dedicati: TPU, NPU e Tensor Core**
- **Tensor Cores e la Precisione Ridotta**

Introduzione

Nel 1971, l'introduzione dell'Intel 4004 racchiuse 2 300 transistor su una singola piastrina di silicio, dando vita all'era del microprocessore commerciale. Per oltre quattro decenni, l'evoluzione del calcolo ha seguito il paradigma della CPU (Central Processing Unit): un cervello elettronico generalista progettato per eseguire qualsiasi algoritmo attraverso una sequenza logica di istruzioni. Tuttavia, l'esplosione dell'Intelligenza Artificiale e dei modelli di *Deep Learning* ha messo a nudo i limiti di questo approccio.

L'hardware per l'AI non richiede un'orchestra diretta da pochi esecutori altamente specializzati, bensì un esercito coordinato di milioni di calcolatori elementari che lavorano in parallelo. Comprendere la differenza tra una CPU e un chip specifico per l'AI (come GPU, TPU o NPU) significa esplorare una tran-

sizione fondamentale nell'ingegneria informatica: il passaggio dal calcolo sequenziale a quello tensoriale.

La CPU: il "*tuttofare*" e il collo di bottiglia di Von Neumann

La CPU moderna rappresenta il risultato di decenni di evoluzione dell'architettura di Von Neumann. Progettata per gestire il sistema operativo, elaborare gli input dell'utente ed eseguire software molto diversi tra loro, è ottimizzata soprattutto per ridurre la latenza nell'esecuzione di un singolo flusso di istruzioni. Per raggiungere questo obiettivo, una parte rilevante della superficie dei processori moderni, come quelli basati su architetture Intel x86 o ARM Cortex, non è occupata dalle unità aritmetico-logiche, ma da strutture di supporto. Tra queste rientrano le cache L1, L2 e L3, memorie estremamente veloci che mantengono i dati vicini ai core riducendo i tempi di accesso alla RAM, e le complesse unità di controllo. Tecniche come la previsione delle ramificazioni (*branch prediction*) e l'esecuzione fuori ordine (*out-of-order execution*) riorganizzano dinamicamente le istruzioni per sfruttare al meglio le risorse disponibili ed evitare che il processore rimanga inattivo.

Questa flessibilità rende la CPU ideale per le applicazioni general-purpose, ma meno efficiente nell'elaborazione delle reti neurali.

L'addestramento e l'inferenza dei modelli di intelligenza artificiale si basano infatti su grandi quantità di moltiplicazioni tra matrici, operazioni ad alta densità note come GEMM (*General Matrix Multiply*).

Il calcolo di un livello della rete può essere sintetizzato dalla relazione:

$$Y = f(W \cdot X + b)$$

dove W rappresenta la matrice dei pesi, X i dati di ingresso, b il termine di bias e f la funzione di attivazione che introduce la non linearità necessaria alla rete neurale. Per eseguire queste operazioni, una CPU deve trasferire continuamente grandi quantità di dati tra RAM, cache, registri e unità di calcolo attraverso il bus di sistema. Il cosiddetto collo di bottiglia di Von Neumann deriva proprio dalla separazione fisica tra memoria ed elaborazione: quando i dati non arrivano abbastanza rapidamente, le risorse computazionali restano in attesa e il throughput complessivo risulta limitato.

La Svolta del Calcolo Parallelo: dalla Grafica 3D ad AlexNet

La transizione verso l'hardware dedicato all'AI ha radici insospettabili nel settore del gaming. Le GPU (Graphics Processing Unit) nascevano per calcolare la posizione e l'illuminazione di milioni di pixel contemporaneamente, applicando trasformazioni geometriche a matrici tridimensionali.

Nel 2006, NVIDIA introdusse l'architettura CUDA, trasformando la GPU da processore a funzione fissa a piattaforma programmabile per scopi generali (GPGPU). La svolta storica avvenne nel 2012 con la competizione *ImageNet*: la rete neurale AlexNet, progettata da Alex Krizhevsky, Ilya Sutskever e Geoffrey Hinton, sbaragliò la concorrenza venendo addestrata non su server CPU tradizionali, ma su due GPU NVIDIA GeForce GTX 580.

Mentre una CPU high-end dispone di pochi core (da 8 a 128) operanti a frequenze elevate (fino a 5 GHz o più), una GPU integra migliaia di core più piccoli operanti a frequenze inferiori. L'architettura abbandona la complessa logica di controllo in favore della parallizzabilità massiva guidata dal modello SIMT (Single Instruction, Multiple Threads).

L'era degli acceleratori AI dedicati: TPU, NPU e Tensor Core

La necessità di elaborare modelli sempre più grandi ha spinto l'industria oltre il semplice adattamento delle GPU nate per la grafica, portando allo sviluppo di acceleratori specializzati e ASIC (Application-Specific Integrated Circuit) progettati per l'intelligenza artificiale. Nel 2016 Google presentò la prima TPU (Tensor Processing Unit), basata su un'architettura ad array sistolico (*systolic array*). A differenza della CPU, che trasferisce frequentemente dati e risultati intermedi tra unità di calcolo, registri e memoria, in un array sistolico i dati fluiscono in modo continuo attraverso una griglia di nodi di elaborazione costituiti da unità MAC (*Multiplier-Accumulate*), con un movimento ritmico che richiama quello del sangue spinto dal cuore. Le moltiplicazioni tra matrici possono così essere eseguite limitando gli accessi alla memoria principale, con una significativa riduzione dei consumi energetici e un aumento della densità di calcolo.

A partire dall'architettura Volta, introdotta nel 2017, NVIDIA ha integrato nelle proprie GPU blocchi hardwa-

re specializzati denominati Tensor Core. Questi componenti sono progettati per eseguire in modo estremamente parallelizzato operazioni di moltiplicazione e accumulo tra matrici, sintetizzabili nella relazione:

$$D = A \cdot B + C$$

Parallelamente, si è compreso che le reti neurali non richiedono sempre la precisione in virgola mobile a 64 bit, FP64, o a 32 bit, FP32, tipica di applicazioni come la simulazione fisica e alcuni calcoli scientifici o finanziari. L'addestramento e l'inferenza possono infatti utilizzare formati a precisione ridotta. FP16 e BF16 (Brain Floating Point) diminuiscono lo spazio occupato dai pesi e dai dati mantenendo un'adeguata gamma dinamica, mentre INT8, FP8 e FP4 consentono di ridurre ulteriormente la memoria necessaria e il carico computazionale attraverso tecniche di quantizzazione, soprattutto nell'inferenza. In questo modo è possibile moltiplicare le prestazioni e migliorare l'efficienza energetica a parità di superficie del chip.

La Banda di Memoria e le HBM

Nei moderni acceleratori AI, come NVIDIA Hopper e Blackwell o AMD Instinct, il limite non è solo la potenza di calcolo, ma anche la velocità con cui i dati raggiungono le unità di elaborazione. Le HBM (High Bandwidth Memory) rispondono a questa esigenza impilando verticalmente più strati di memoria DRAM, collocati accanto al chip su un *interposer* di silicio. Le generazioni HBM3 e HBM4 raggiungono bande di diversi terabyte al secondo, contro le decine o centinaia di gigabyte al secondo tipiche delle DDR5, aumentando il throughput e l'efficienza degli acceleratori AI.

L'Integrazione Eterogenea e l'AI al Margine

L'evoluzione recente risponde alla necessità di portare l'Intelligenza Artificiale direttamente sui dispositivi utente (*Edge AI*) per questioni di privacy, latenza ed efficienza energetica. Non si tratta più di scegliere tra CPU o chip AI, ma di integrarli nello stesso sistema. Processori moderni per PC e smartphone integrano architetture eterogenee su un unico SoC (System on

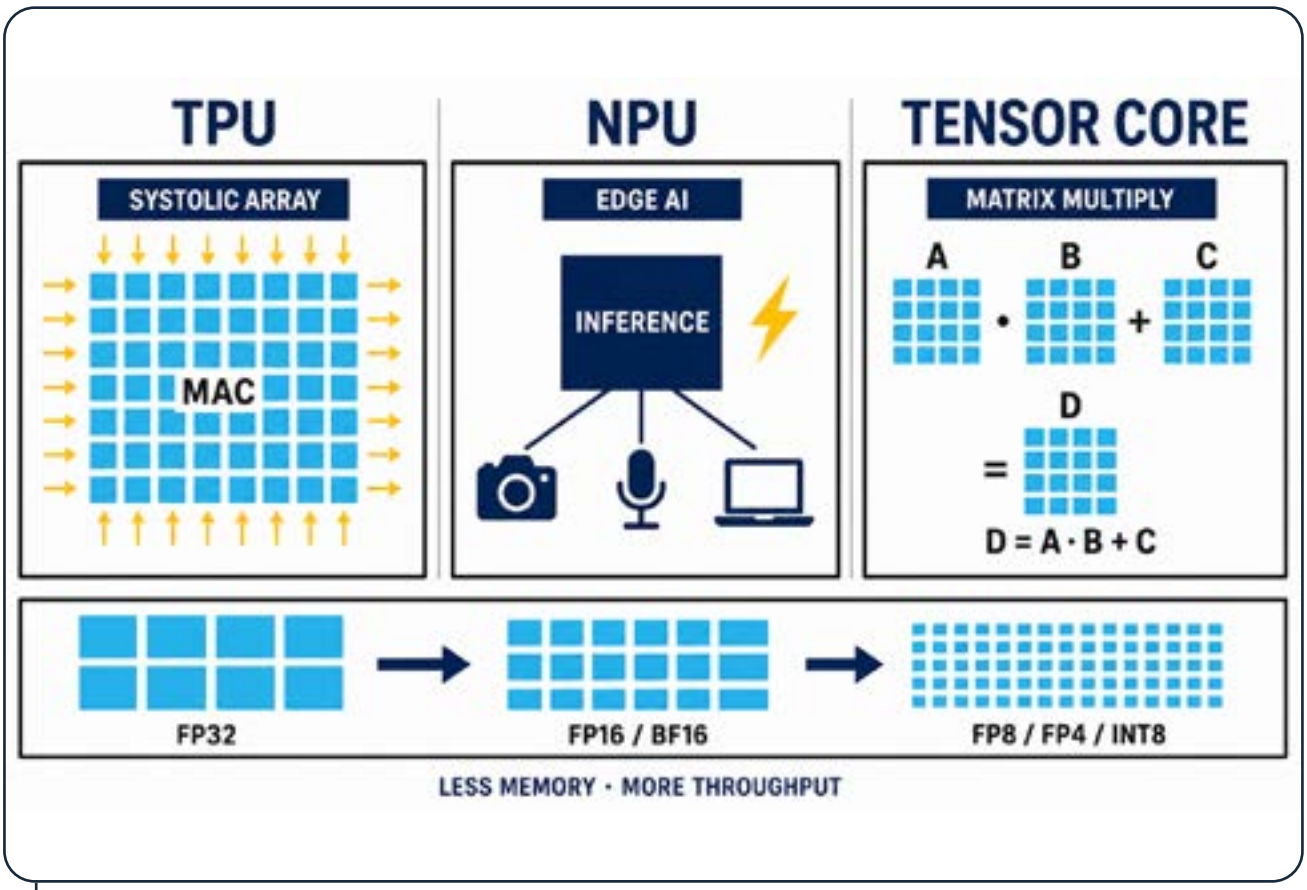


Figura 1 – TPU, NPU e Tensor Core: gli acceleratori specializzati per l'intelligenza artificiale.

Caratteristica	CPU (Central Processing Unit)	Chip AI (GPU / TPU / NPU)
Obiettivo Architetturale	Minima latenza su calcoli sequenziali	Massimo throughput su calcoli paralleli
Composizione del Die	Prevalenza di Cache e Logic Control	Prevalenza di Array ALU/MAC (Compute Units)
Modello di Esecuzione	MIMD / Out-of-Order / Branch Prediction	SIMD / SIMT / Dataflow Systolic
Precisione Nativa	FP32, FP64, INT64	FP16, BF16, FP8, FP4, INT8
Banda di Memoria Tipica	(DDR4 / DDR5)	(HBM3 / HBM4)
Casi d'Uso Ideali	Sistemi Operativi, Database, Logica condizionale	Addestramento e Inferenza di LLM, Visione Artificiale

Tabella 1 - CPU e chip AI: architetture e prestazioni a confronto.

Chip) o tramite *chiplet* interconnessi tramite packaging avanzato (come TSMC CoWoS):

- **CPU**: Gestisce il codice dell'applicazione e le chiamate di sistema.
- **GPU**: Gestisce il rendering grafico e calcoli paralleli ad hoc.
- **NPU (Neural Processing Unit)**: Blocco ad alta efficienza energetica progettato esclusivamente per eseguire modelli di inferenza in background (ad esempio il riconoscimento facciale, la cancellazione del rumore via software o la generazione locale di testo) con un impatto minimo sulla batteria.

L'avvenire della microelettronica non prevede la scomparsa della CPU, la cui flessibilità rimane insostituibile per l'infrastruttura di base dell'elaborazione dati. Assistiamo piuttosto alla fine del monopolio dell'architettura generalista. L'hardware dell'era dell'AI è dominato dalla specializzazione di dominio, in cui la struttura fisica del silicio ricalca direttamente la forma matematica degli algoritmi che deve eseguire.

Keywords: chip AI, microprocessori, CPU, GPU, TPU, NPU, Tensor Core, calcolo tensoriale, Edge AI, memorie HBM.

L'autore*
IT engineer and business consultant,
info@gemaxconsulting.it