Analizziamo i limiti della miniaturizzazione dei chip e le nuove tecnologie per l'intelligenza artificiale: GPU, TPU, chip neuromorfici e computer quantistici.

Nel corso degli ultimi decenni i chip elettronici hanno visto una crescita esponenziale della loro capacità di calcolo. La **Legge di Moore**, formulata da Gordon Moore nel 1965, osservava che il numero di transistor per chip raddoppiava circa ogni due anni. Questo trend ha garantito una crescente potenza di calcolo, come mostra la **Fig. 1**: su scala semilogaritmica, il numero di transistor nei microprocessori è cresciuto esponenzialmente dal 1970 al 2020, seguendo da vicino la previsione di Moore.

Parallelamente, l'architettura di **von Neumann** – la struttura classica di CPU con memoria separata, proposta negli anni '40 – si è affermata come modello di riferimento (**Fig. 2**).

In questo schema i dati e i programmi risiedono in memoria mentre la CPU li elabora e per molti decenni questa soluzione

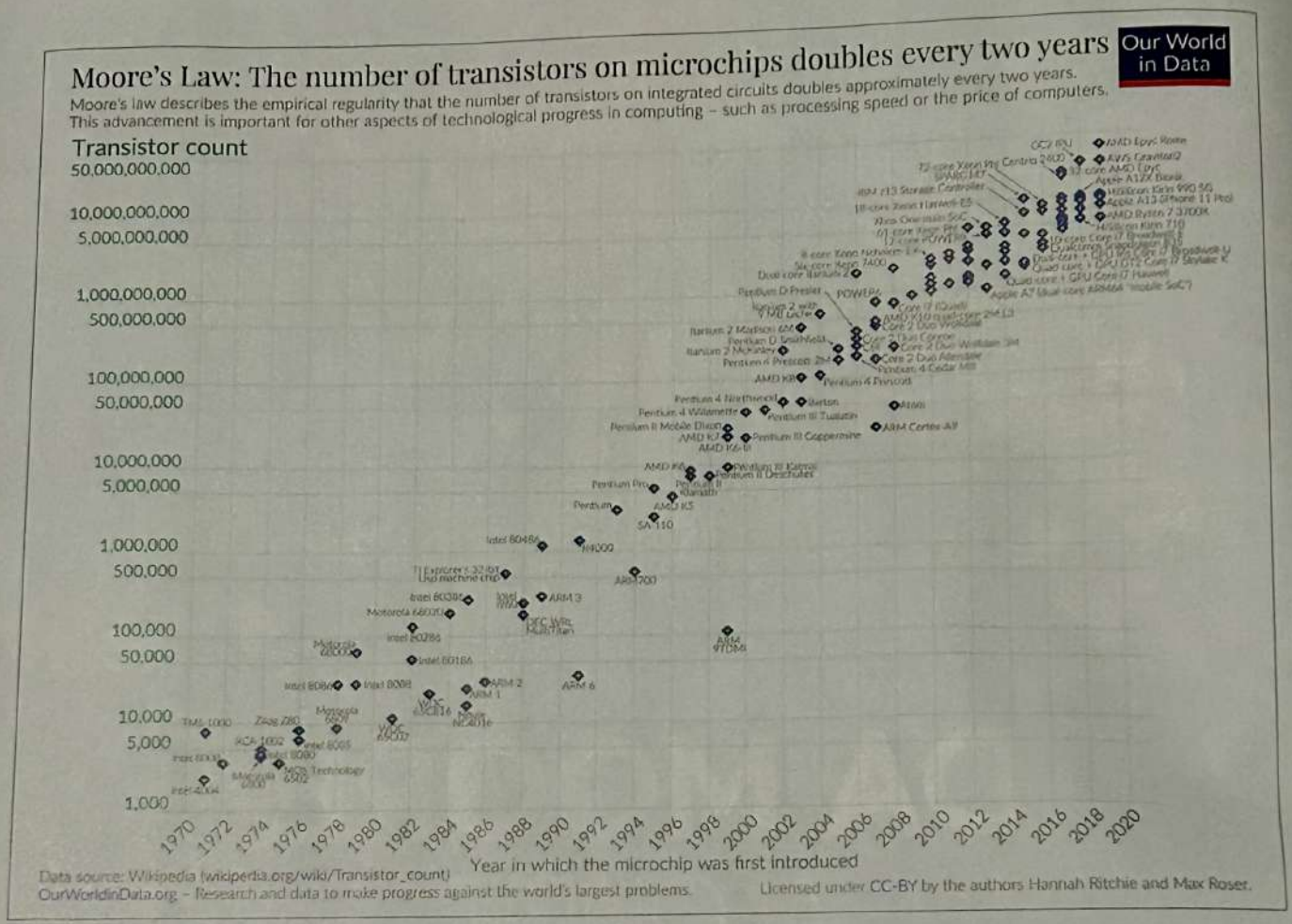

Fig. 1 *Andamento della Legge di Moore: il numero di transistor nei microprocessori raddoppia circa ogni due anni, ma il ritmo sta rallentando con l'avvicinarsi ai limiti fisici della miniaturizzazione.*

ha funzionato bene, perché permetteva flessibilità e semplicità di progettazione. Tuttavia, *per i carichi di lavoro moderni, e soprattutto per l'intelligenza artificiale*, il modello di von Neumann si è rivelato un collo di bottiglia. Con calcoli molto ripetitivi e matriciali (ad es. deep learning) la CPU resta spesso in attesa che i dati viaggino tra memoria e unità di calcolo, causandone un forte sottoutilizzo. In altre parole, il processore veloce finisce per "sedersi in panchina" mentre aspetta i dati, fenomeno noto come **von Neumann bottleneck (Fig. 3)**.

Nel frattempo, la progressiva miniaturizzazione dei transistor si è avvicinata a limiti fisici estremi. Oggi i transistor di punta hanno dimensioni dell'ordine di pochi nanometri (dati per dieci o venti atomi di larghezza) e si avvicinano ai limiti della meccanica quantistica.

Effetti come il tunneling quantistico (elettroni che "sgusciano" attraverso i sottilissimi isolanti) rendono difficile spegnere il transistor completamente e mantengono alte correnti di perdita. Inoltre, la legge di **Dennard** (che fino ai primi anni 2000 garantiva che potenza e densità di potenza rimanessero costanti al ridursi delle dimensioni) non è più valida: continuando a ridurre la geometria i chip generano troppo calore per unità di potenza, senza poterlo dissipare facilmente. In pratica, *stiamo toccando il "fondo del barile" della fisica del silicio*: non è più possibile ridurre indefinitamente le dimensioni senza introdurre nuovi materiali o tecnologie radicalmente diverse.

A tutto ciò si aggiunge un problema economico: le attuali tecnologie a 2 - 3 nm (**Fig. 4**) richiedono impianti di produzione talmente sofisticati che una nuova fabbrica di chip costa ormai oltre 10 miliardi di dollari. Grazie alla cosiddetta **seconda legge di Moore** (o legge di Rock) si stima infatti che il costo di un nuovo stabilimento di produzione raddoppi ogni 4 anni. Oggi pochissime aziende – come TSMC, Intel, Samsung – possono permettersi in-

vestimenti simili. In sintesi, i chip tradizionali basati sullo scaling planare hanno raggiunto **limiti pratici** sia fisici che economici: non possono più diventare molto più piccoli o veloci a costi sostenibili con le tecniche attuali.

STRATEGIE ATTUALI: PARALLELISMO E CHIP SPECIALIZZATI

Davanti a questi vincoli, l'industria del silicio ha cercato di mantenere il ritmo di crescita delle prestazioni con approcci diversi dal solo "mettere più transistor sul chip". In particolare, si è puntato sul parallelismo e sulla specializzazione hardware. Da un lato, anziché un singolo core sempre più

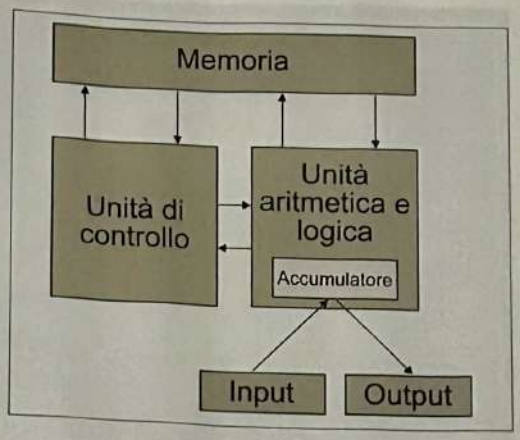

Fig. 2 Architettura di von Neumann: il modello classico di calcolatore in cui CPU e memoria comunicano attraverso un bus condiviso.

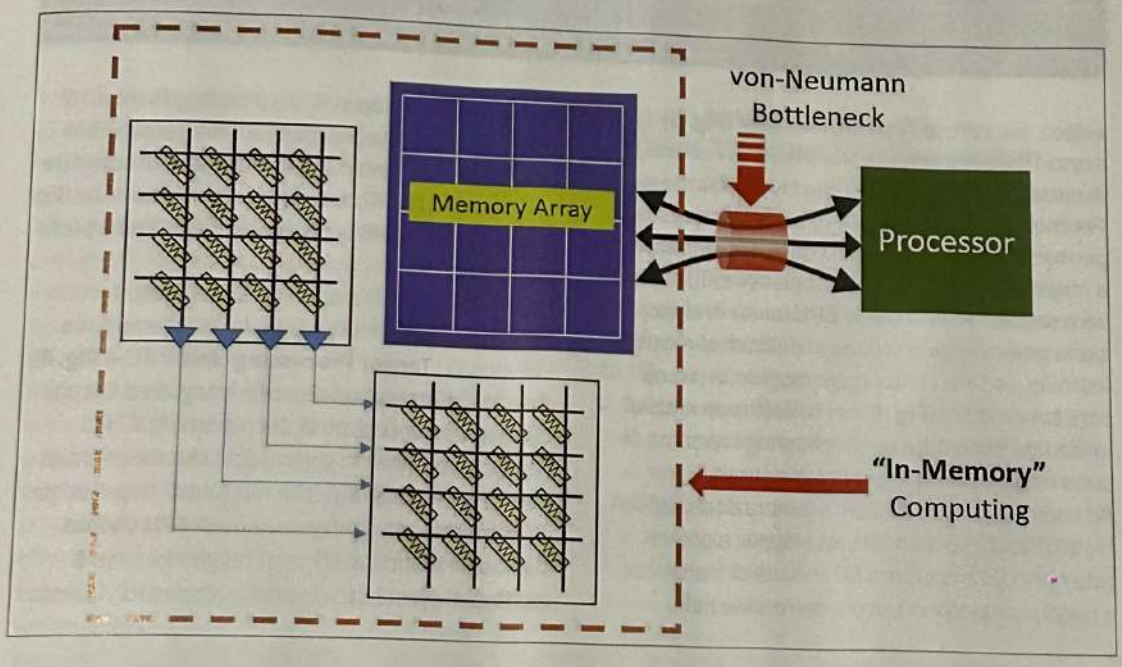

Fig. 3 Rappresentazione del collo di bottiglia di von Neumann: la separazione tra CPU e memoria genera rallentamenti dovuti alla banda limitata del bus.

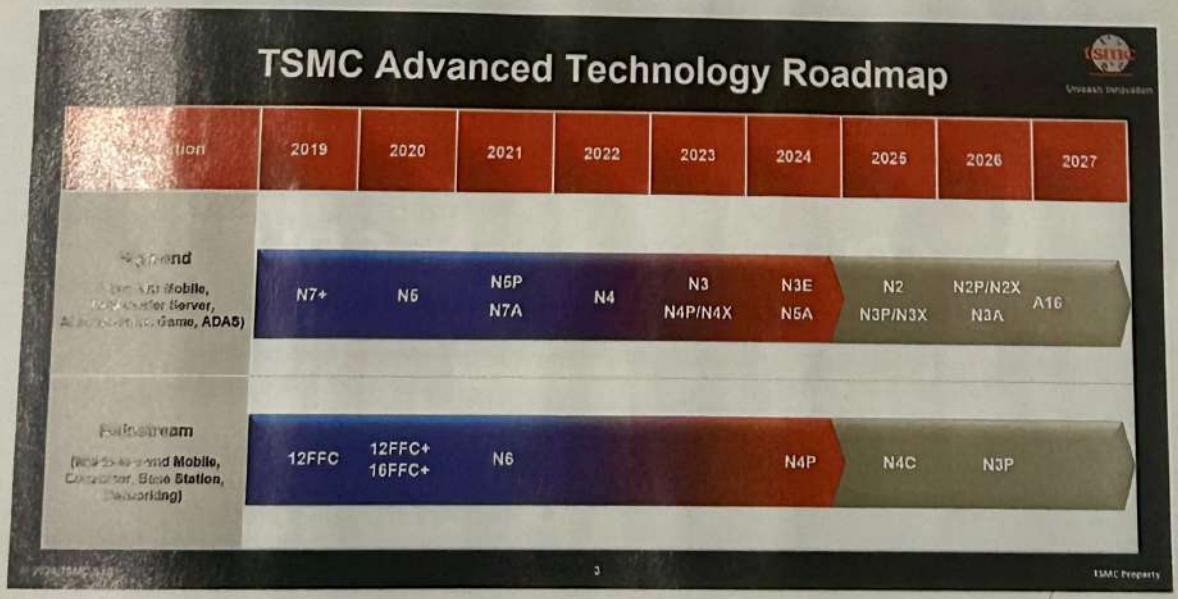

Fig. 4 Roadmap TSMC: evoluzione delle tecnologie di processo in nanometri dal 65nm al 2nm, con indicazione delle sfide crescenti legate a costi, fisica e resa.

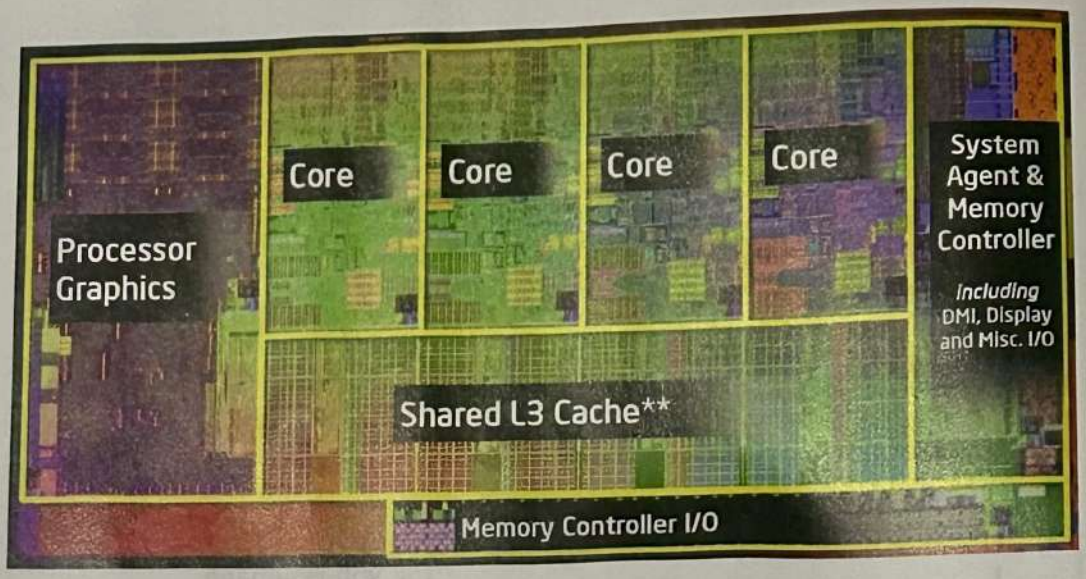

Fig. 5
Architettura interna di una moderna CPU multicore: ogni core è un'unità di calcolo indipendente che condivide cache e interconnessioni.

veloce, si è diffuso l'uso di multi-core **(Fig. 5)**: i processori moderni contengono decine (o centinaia) di nuclei di calcolo paralleli che lavorano insieme. Per molti software e calcoli tipici dell'IA, questo permette di processare dati contemporaneamente e migliorare le prestazioni complessive. Tuttavia, il vero salto è arrivato con le **GPU** (unità di elaborazione grafica) e gli acceleratori dedicati al machine learning. Le GPU contengono migliaia di piccoli core specializzati (**Fig. 6**) nel calcolo matriciale e vettoriale: sono nate per il rendering video ma si sono rivelate perfette per le reti neurali.
Ad esempio, la GPU di ultima generazione NVIDIA H100 (basata su architettura Hopper e presentata nel 2022) incorpora 80 miliardi di transistor e raggiunge picchi di calcolo dell'ordine delle centinaia di teraflop in FP16 (Floating Point a 16 bit), oltre 3 volte le prestazioni della generazione precedente. In termini pratici, l'H100 può superare i 500 teraflop in FP16 e raggiungere circa 1 exaflop (10^18 flop) in FP8 – cifre impensabili fino a pochi anni fa (**Fig. 7**).
In parallelo alle GPU, sono nati acceleratori hardware dedicati all'IA. Google, ad esempio, ha sviluppato i **Tensor Processing Unit** (TPU – **Fig. 8**), chip ASIC (Application-Specific Integrated Circuit) ottimizzati per i calcoli di deep learning. I TPU vengono connessi in giganteschi cluster chiamati "pod" che mettono insieme migliaia di chip. Google afferma che una configurazione di **TPU v4 Pod** su Google Cloud (4096 chip) raggiunge circa **9 exaflop** di prestazioni teoriche aggregate. Questa

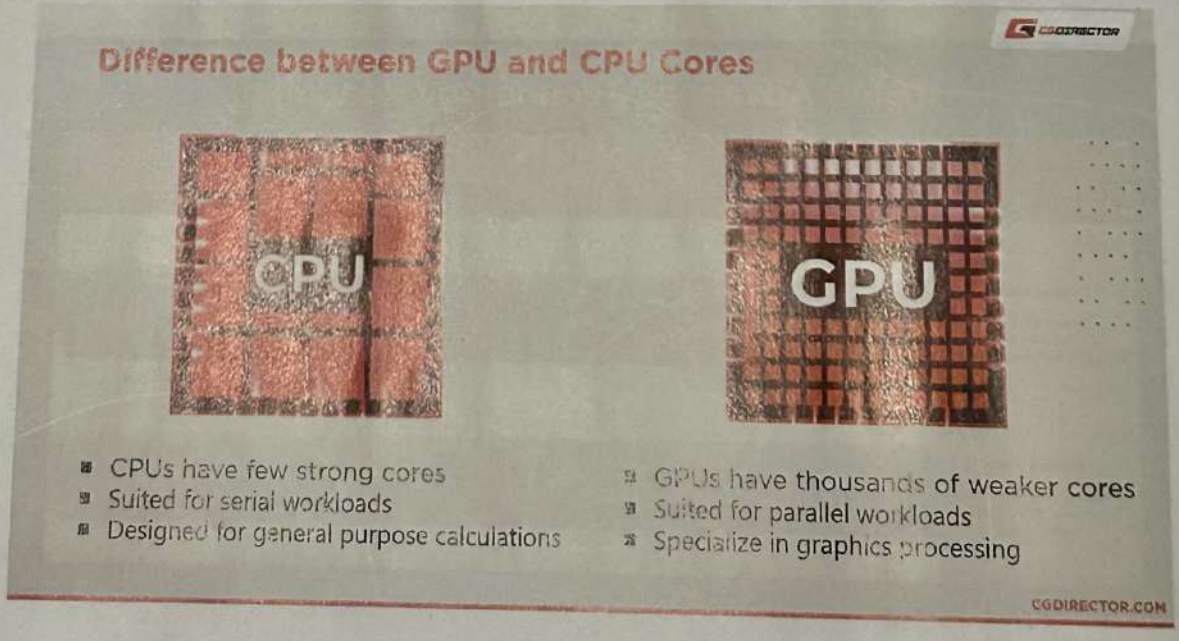

Fig. 6
Differenza tra CPU e GPU: le GPU contengono centinaia o migliaia di core più semplici, ideali per calcoli paralleli come quelli del deep learning.

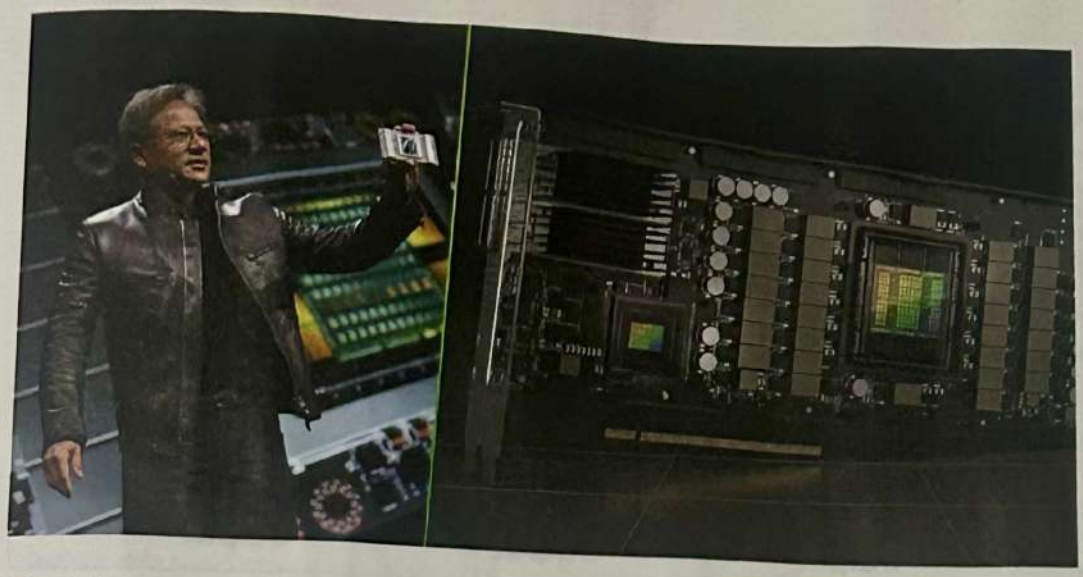

Fig. 7
NVIDIA H100 Hopper: una delle GPU più potenti per l'intelligenza artificiale, progettata con architettura avanzata e memoria ad alta banda.

infrastruttura consente di addestrare modelli di AI estremamente complessi (come i moderni modelli di linguaggio) in tempi sostenibili. In pratica, con GPU e TPU oggi si sfruttano linee di calcolo fortemente parallele e ottimizzate: piuttosto che cercare un nuovo monocore più veloce, si mette in campo un esercito di core specializzati che lavorano insieme a problemi di machine learning con una efficienza straordinaria.

Altre soluzioni parallele includono unità acceleratrici neurali (**NPU**) integrate direttamente in telefoni o PC, i tradizionali FPGA (chip riconfigurabili) e architetture miste CPU+FPGA (**Fig. 9**). Vi è infine un caso estremo di parallelismo e dimensioni: il chip su wafer intero. Un esempio è il Wafer-Scale Engine 2 di Cerebras (**Fig. 10**), che occupa un intero wafer di silicio (~46.000 mm²) con 2.6 trilioni di transistor e 850.000 core dedicati all'IA.

Questo enorme chip (alimentato da 15 kW) sostituisce migliaia di GPU usate in rack e fornisce enormi capacità di calcolo parallelo. Tali soluzioni di "chip gigante" sono però molto costose e specializzate: puntano a dominare nicchie di supercalcolo AI piuttosto che il mercato generale dei microprocessori.

Nonostante questi progressi, anche le architetture specializzate hanno limiti. Sebbene aumentino la densità di calcolo per watt, restano costose da sviluppare e ottimizzate per task specifici. Inoltre, non risolvono del tutto il problema dell'efficienza energetica e dei colli di bottiglia interni (come le strozzature nella memoria condivisa). In sostanza, si è riusciti a migliorare le prestazioni aggirando il fabbisogno di miniaturizzazione estrema, ma si va incontro a complessità progettuali e consumi elevati.

ARCHITETTURE EMERGENTI: OLTRE IL TRADIZIONALE TRANSISTOR

Poiché le soluzioni "a silicio convenzionale" si avvicinano ai loro limiti, da più anni vengono esplorate architetture di calcolo alternative radicalmente diverse. Queste non puntano tanto sulla miniaturizzazione dei transistor, quanto sul cambiare il modo di rappresentare e elaborare l'informazione. Di seguito alcuni esempi promettenti:

- **Calcolo quantistico**. A differenza dei bit classici (0 o 1), i **qubit** sfruttano i fenomeni quantistici di *sovrapposizione* ed *entanglement* (**Fig. 11**). Un singolo qubit può trovarsi contemporaneamente in sovrapposizione di 0 e 1, e due qubit possono essere correlati in modo tale che la misura di uno influisca istantaneamente sull'altro. Ciò significa che 100 qubit possono codificare 2^{100} valori contemporaneamente, creando spazi computazionali esponenzialmente grandi. I computer quantistici (**Fig. 12**) lavorano manipolando tali sovrapposizioni tramite porte quantistiche; solo alla fine si misura il risultato, e l'interferenza delle ampiezze di probabilità annulla tutte le so-

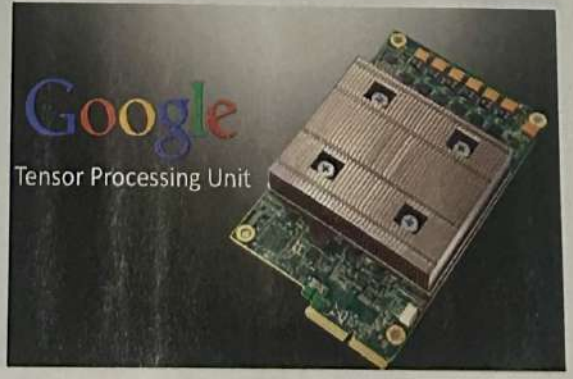

Fig. 8
Google TPU (Tensor Processing Unit): chip specializzati per l'accelerazione di modelli IA basati su tensori, sviluppati per i data center.

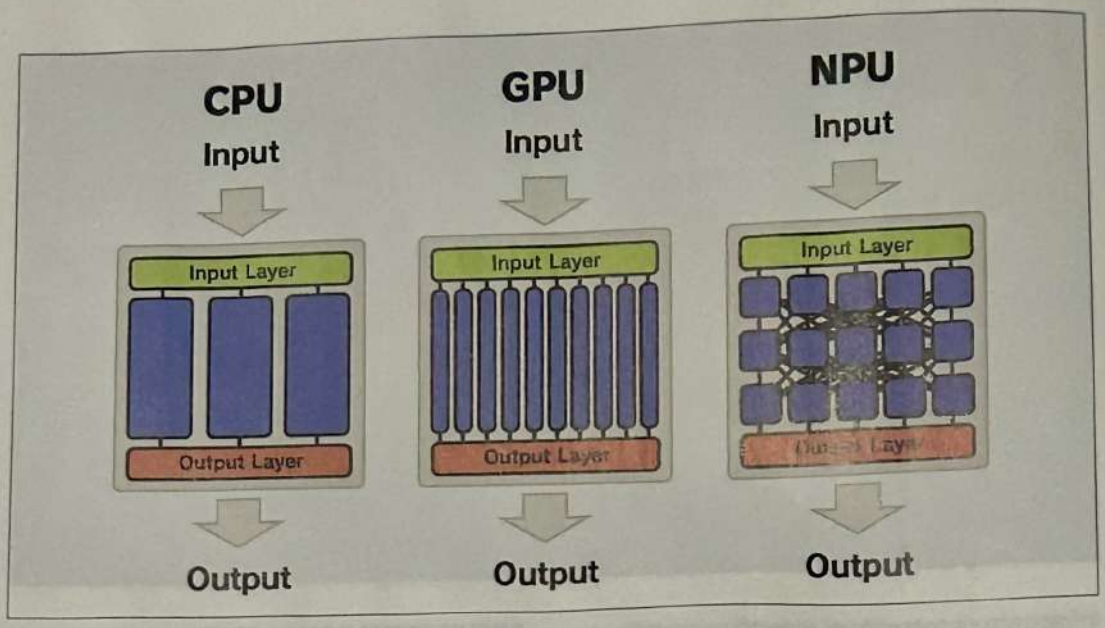

Fig. 9
Confronto visivo tra CPU, GPU e NPU: architetture pensate rispettivamente per elaborazione generale, grafica/AI, e reti neurali.

luzioni non desiderate. Questo permette accelerazioni esponenziali per alcuni problemi specifici (per esempio la fattorizzazione con l'algoritmo di Shor o la ricerca di strutture in dati combinatorici). Oggi esistono prototipi quantistici con decine o centinaia di qubit reali (ad es. i processori di IBM, Google, Rigetti, IonQ, ecc.), ma sono ancora afflitti da errori e decoerenza. In ogni caso, molte grandi aziende e centri di ricerca investono massicciamente: IBM, Google, Microsoft, Intel e startup dedicate puntano a far diventare il calcolo quantistico una tecnologia di largo impiego entro un decennio o due. Alcuni analisti stimano che il mercato del quantum computing possa raggiungere l'ordine di *migliaia di miliardi di dollari* entro il 2035, se si troveranno le chiavi per aumentare affidabilità e scala. In sintesi, la forza dei qubit è di offrire un modo totalmente nuovo di calcolo (sfruttando la natura quantistica della materia), che promette soluzioni drasticamente più veloci per certi calcoli, sebbene oggi sia ancora agli inizi.

- **Calcolo neuromorfico**. Ispirandosi al cervello umano, queste architetture implementano reti di neuroni e sinapsi sintetici sul chip. A differenza dei tradizionali transistor sincroni, i neuroni spingono segnali discreti (spike) in modo *event-driven*: elaborano informazioni solo quando

Fig. 10
Wafer Scale Engine (WSE-2) di Cerebras: il chip più grande mai realizzato, costruito su un intero wafer per il training di modelli IA su larga scala.

ricevono uno spike in ingresso, proprio come nel sistema nervoso biologico. In pratica, un chip neuromorfico memorizza le informazioni (sinapsi) e le elabora nei medesimi elementi: ciò riduce drasticamente i trasferimenti di dati tra memoria e calcolo (rispetto all'architettura von Neumann) e permette consumi estremamente bassi. Per esempio, chip come lo TrueNorth di IBM o il Loihi di Intel (**Fig. 13**) possono simulare milioni di neuroni in tempo reale consumando pochi watt. Grazie a questa efficienza energetica, i calcoli neuromorfici sono ordini di grandezza più economici in termini di energia per operazione rispetto alle CPU/GPU tradizionali quando si tratta di riconoscimento di pattern o inferenza neurale. Inoltre, essendo estremamente paralleli, i neuroni artificiali possono lavorare in asincrono: teoricamente si può attivare ogni neurone contemporaneamente, permettendo enormi gradi di parallelismo. I principali vantaggi dei sistemi neuromorfici sono quindi adattività (possono imparare in tempo reale modificando le "sinapsi"), efficienza energetica e velocità su compiti neurali. Tra le sfide c'è la complessità di programmare reti di neuroni (necessarie nuove librerie e algoritmi), nonché una certa perdita di precisione rispetto ai modelli tradizionali. Nonostante ciò, grandi aziende e università (ad es. IBM con il chip TrueNorth e il successore NorthPole, Intel con Loihi, l'università di Heidelberg con BrainScaleS) stanno sviluppando queste architetture per potenziare applicazioni di IA e robotica.

- **Calcolo fotonico (ottico)**. Un altro approccio è impiegare la luce al posto degli elettroni per trasportare e perfino elaborare informazioni. I circuiti fotonici integrati utilizzano guide d'onda ottiche e modulatori per compiere operazioni come somme e moltiplicazioni di segnali luminosi, offrendo potenzialmente altissime velocità e bassissimo consumo energetico (la luce ha dissipazione minima rispetto ai metalli). Recenti esperimenti, come quelli del MIT, hanno dimostrato prototipi di chip fotonici in grado di eseguire i calcoli fondamentali di una rete neurale convoluzionale direttamente con dispositivi ottici, ottenendo accuratezza comparabile ai chip elettronici tradizionali ma in un regime di densità di potenza molto più basso. In particolare, il nuovo chip fotonico di MIT completa gli strati chiave di un modello di deep learning in meno di mezzo nanosecondo, con consumi energetici trascurabili rispetto ai transistor convenzionali. In futuro, processori ibridi optoelettronici potrebbero permettere calcoli massicci per l'IA (ad es. elaborazione in tempo reale di video, riconoscimento vocale) superando i limiti di larghezza di banda e latenza dei fili elettrici. Anche qui, però, la tecnologia è emergente e ha ostacoli: alcuni tipi di funzioni (non-lineari) sono difficili da ottenere con soli componenti ottici, quindi servono comunque parti elettroniche complementari.

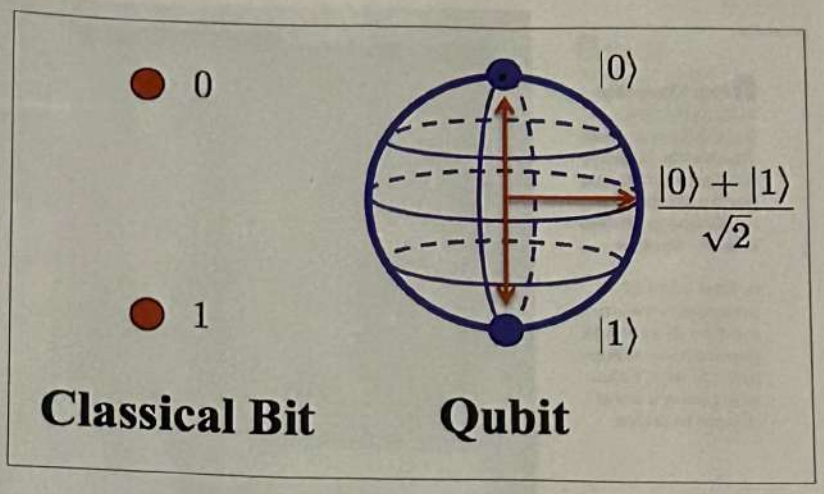

Fig. 11
Un singolo qubit può trovarsi contemporaneamente in sovrapposizione di 0 e 1.

Oltre a questi grandi paradigmi, si studiano altre vie come i **memristori (Fig. 14)** (componente elettronico in grado di "ricordare" la corrente passata) per memorie non volatili ultra-dense, e materiali

Fig. 12
IBM Quantum System One: sistema criogenico che ospita un processore quantistico operante a temperature prossime allo zero assoluto.

a.

b.

Fig. 13a e 13b
a. Scheda IBM SyNAPSE con chip TrueNorth: piattaforma sperimentale per il calcolo neuromorfico in ambito edge e robotica.

b. Intel Loihi 2: processore neuromorfico di seconda generazione sviluppato da Intel Labs, energetica e parallelismo massivo.

quantistici alternativi (grafene, nanofili) per nuovi tipi di transistor. In generale, la ricerca "oltre il silicio" include materiali a **bandgap largo** (SiC, GaN) per elettronica di potenza più efficiente, semiconduttori fotonici per guide d'onda, spintronica, e così via. Queste innovazioni mirano tutte a creare dispositivi più veloci, meno rumorosi e meno energivori, anche se spesso richiedono infrastrutture di fabbricazione completamente nuove.

APPROCCI STRUTTURALI E AVANZAMENTI NELL'ASSEMBLAGGIO

Un filone particolarmente attivo riguarda infine il design fisico dei chip, sfruttando lo spazio tridimensionale o combinando tecnologie. Ad esempio, il 3D stacking (**Fig. 15**) prevede di impilare più strati di circuiti uno sopra l'altro, anziché sullo stesso piano. Questo approccio "a grattacielo" può aumentare esponenzialmente la densità di transistor per area, riducendo le distanze elettriche tra strati. MIT e altre istituzioni hanno realizzato prototipi di chip 3D in cui strati attivi sono collegati verticalmente con vie conduttive (TSV): si possono raggiungere prestazioni maggiori e consumi inferiori rispetto a chip planar di pari numero di transistor. La sfida è gestire i calori e le tensioni su più livelli, oltre alla complessità di produzione.

Un altro approccio di packaging avanzato è l'uso di **chiplet** e sistemi eterogenei (**Fig. 16**): invece di costruire tutto un chip monolitico, la logica viene suddivisa in più piccoli chip ("chiplet") che vengono uniti sullo stesso package (a volte con un interposer di silicio 2.5D). Questo rende la produzione più flessibile, perché ogni chiplet può essere realizzato con la tecnologia più adatta (in alcuni casi più vecchia), mentre l'interconnessione tra chiplet (anche ottica o avanzata) mantiene alte velocità di comunicazione. Ad esempio, molti processori moderni dividono la CPU tra die separati o integrano pacchetti "system-in-package" con CPU, GPU e memoria uniti, ottenendo vantaggi in prestazioni/consumo.

Infine, vale richiamare un concetto già visto: la wafer-scale integration (WSE), come nel caso di Cerebras. Invece di ritagliare numerosi die da un wafer, l'intero wafer funge da un unico chip gigantesco. Il Wafer-Scale Engine 2 di Cerebras misura oltre 46.000 mm² e contiene 2,6 trilioni di transistor con 850.000 core ottimizzati per l'intelligenza artificiale. Questo chip offre prestazioni paragonabili a migliaia di GPU convenzionali, occupando uno spazio simile a un singolo rack. È un esempio estremo di come, restando nei transistor convenzionali, si cerchi di aumentare la scala fisica per superare i limiti di densità superficiale.

Fig. 14
Matrice di memristor su silicio per applicazioni di calcolo neuromorfico.

Fig. 15
Strategie di impilamento 3D (3D stacking) secondo AMD: integrazione verticale di chip e blocchi funzionali per superare i limiti dell'integrazione monolitica.

CONCLUSIONI

La storia del calcolo digitale dimostra che ogni volta che una tecnologia raggiunge i suoi limiti, nuovi paradigmi emergono per proseguire l'avanzamento.

Oggi ci troviamo proprio in questo punto di svolta: gli approcci tradizionali basati sullo scaling planare di transistor (Legge di Moore, architettura von Neumann) stanno esaurendo i margini. Per continuare a migliorare le prestazioni, l'industria sta dunque percorrendo molte strade contemporanee. Da un lato, si sfruttano al massimo le soluzioni "ibride" e parallele già mature, come processori multicore, GPU e acceleratori AI altamente specializzati. Dall'altro, la ricerca punta a innovazioni radicali: computer quantistici, chip neuromorfici, processori ottici e nuovi materiali.

Allo stesso tempo, tecniche di integrazione avanzate (3D stacking, chiplet, wafer-scale) consentono di aumentare ulteriormente densità e velocità senza cambiare la tecnologia base dei transistor. In definitiva, non esiste una sola soluzione miracolosa, bensì un ecosistema di tecnologie che si integrano e si bilanciano a vicenda.

Per carichi di lavoro generici forse continueremo ancora per qualche anno con CPU/GPU ottimizzate, ma per applicazioni d'avanguardia – dall'IA alla simulazione molecolare – verrà adottato un mix di architetture specializzate.

I computer del futuro probabilmente saranno "eterogenei": conterranno classici microprocessori accanto a moduli quantistici, memorie avanzate, reti neurali hardware e interconnessioni ottiche.

A livello didattico, il messaggio chiave è che non possiamo più affidare il progresso del calcolo all'aumento della densità di transistor fine a se stesso. Invece, impariamo dall'elettronica, dalla fisica quantistica e persino dalla biologia per **reinventare** come elaboriamo l'informazione. Conoscere bene il percorso storico (Moore, Neumann, etc.) aiuta a capire perché oggi occorrono queste soluzioni alternative.

In questo momento di transizione, la comunità scientifica e industriale sta sperimentando in parallelo molte di queste strade – e solo il tempo ci dirà quali diventeranno dominanti nei prossimi decenni. □

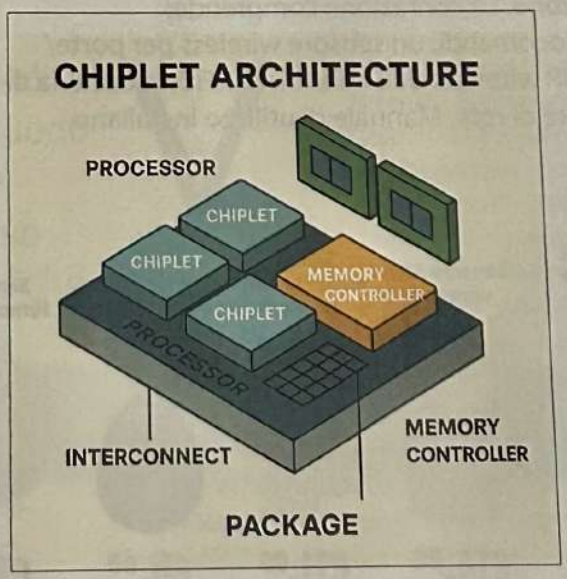

Fig. 16
Architettura a chiplet: suddivisione del processore in più die interconnessi all'interno dello stesso package.