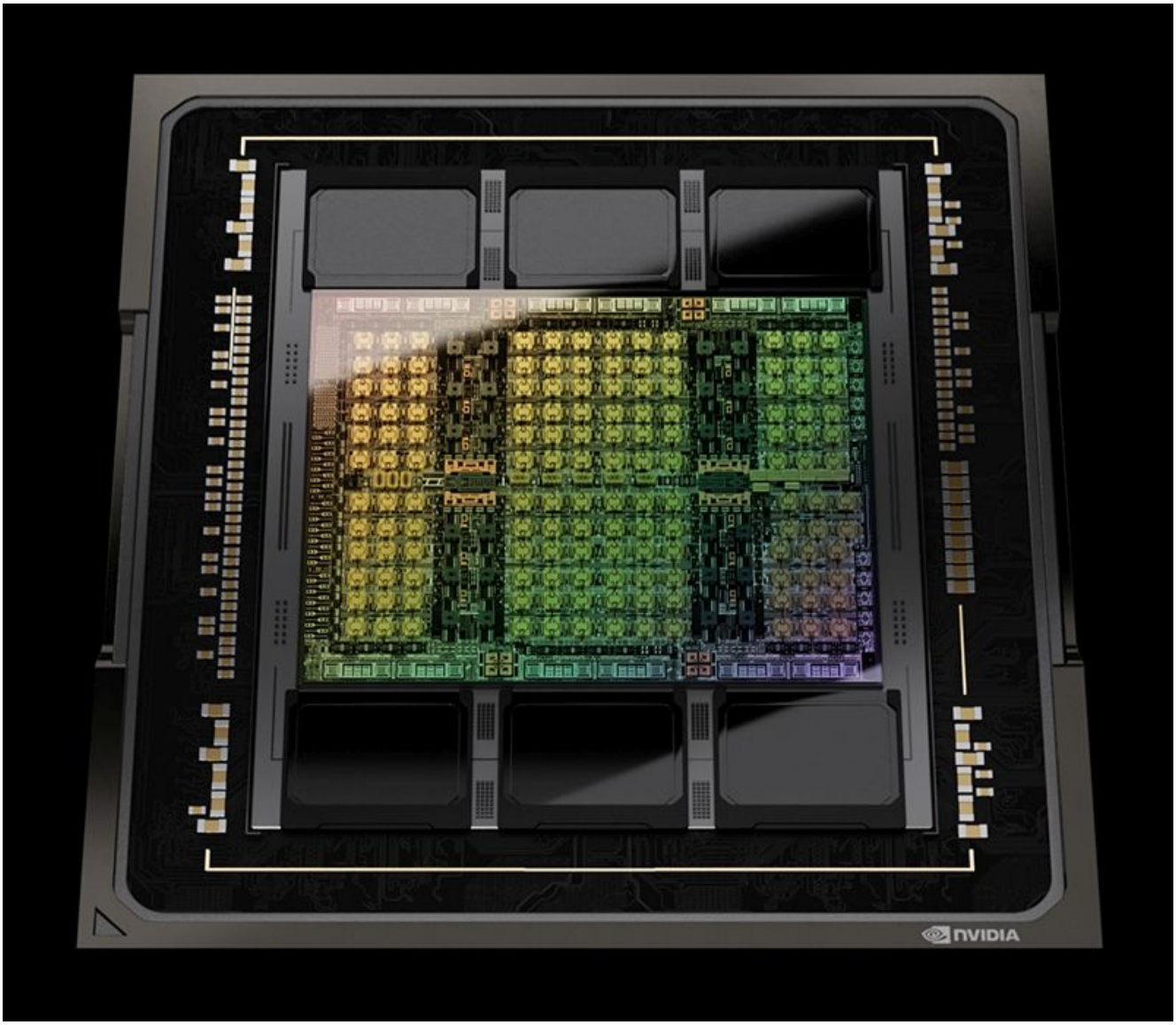
NVIDIA Hopper: la nuova mostruosa architettura da 80 miliardi di transistor

NVIDIA ha presentato l’architettura Hopper, alla base dei futuri acceleratori H100 in formato SXM e PCI Express 5.0. Forte di una GPU con 80 miliardi di transistor e fino a 18432 CUDA core, le soluzioni NVIDIA H100 promettono di far mangiare la polvere alla concorrenza e alla precedente generazione A100.

di Manolo De Agostini pubblicata il **22 Marzo 2022**, alle **18:22** nel canale Schede Video

HopperNVIDIA

NVIDIA ha annunciato **Hopper**, la nuova **architettura** che succede ad Ampere nel campo degli acceleratori destinati ai datacenter per calcoli ad alte prestazioni (HPC) e intelligenza artificiale. Il nome è **un omaggio a Grace Hopper**, pioniera statunitense della programmazione informatica, nota per il suo lavoro su uno dei primi compilatori, un impegno di fondamentale importanza che ha portato allo sviluppo del linguaggio COBOL.



NVIDIA GH100, la prima GPU basata su architettura Hopper, si presenta con **80 miliardi di transistor** (+48% rispetto ai 54,2 miliardi del GA100 di NVIDIA A100) prodotti con **processo 4N di TSMC**, una versione personalizzata per NVIDIA dell'N4

dell’azienda taiwanese. NVIDIA GH100 **è la prima GPU con supporto PCI Express 5.0 e la prima abbinata a memoria HBM3** per una bandwidth di memoria di **3 TB/s**.

All’interno di una GPU Hopper GH100 completamente attiva ci sono 144 Streaming Multiprocessor per un totale di **18432 CUDA core FP32** (un GA100 completo prevede 8192 CUDA core, quindi l’aumento è del 125%), accompagnati da 576 Tensor core e 60 MB di cache L2. 12 controller di memoria a 512 bit e 6 stack di HBM3 per un totale di 96 GB compongono la parte relativa alla memoria. A queste specifiche si aggiunge NVLink di quarta generazione.

	NVIDIA A100	**NVIDIA H100 SXM5**	**NVIDIA H100 PCIe**
Architettura	NVIDIA Ampere	NVIDIA Hopper	NVIDIA Hopper
Formato	SXM4	SXM5	PCIe Gen 5
SM	108	132	114

TPC	54	62	57
FP32 Core / SM	64	128	128
FP32 Core / GPU	6912	16896	14592
FP64 Core / SM (esclusi Tensor)	32	64	64
FP64 Core / GPU (esclusi Tensor)	3456	8448	7296
INT32 Core / SM	64	64	64
INT32 Core / GPU	6912	8448	7296
Tensor Core / SM	4	4	4
Tensor Core / GPU	432	528	456
GPU Boost Clock (Non finalizzati per H100)	1410 MHz	Non dichiarati	Non dichiarati
FP8 Tensor TFLOPS di picco con FP16 Accumulate	N/A	2000/4000	1600/3200
FP8 Tensor TFLOPS di picco con FP32	N/A	2000/4000	1600/3200

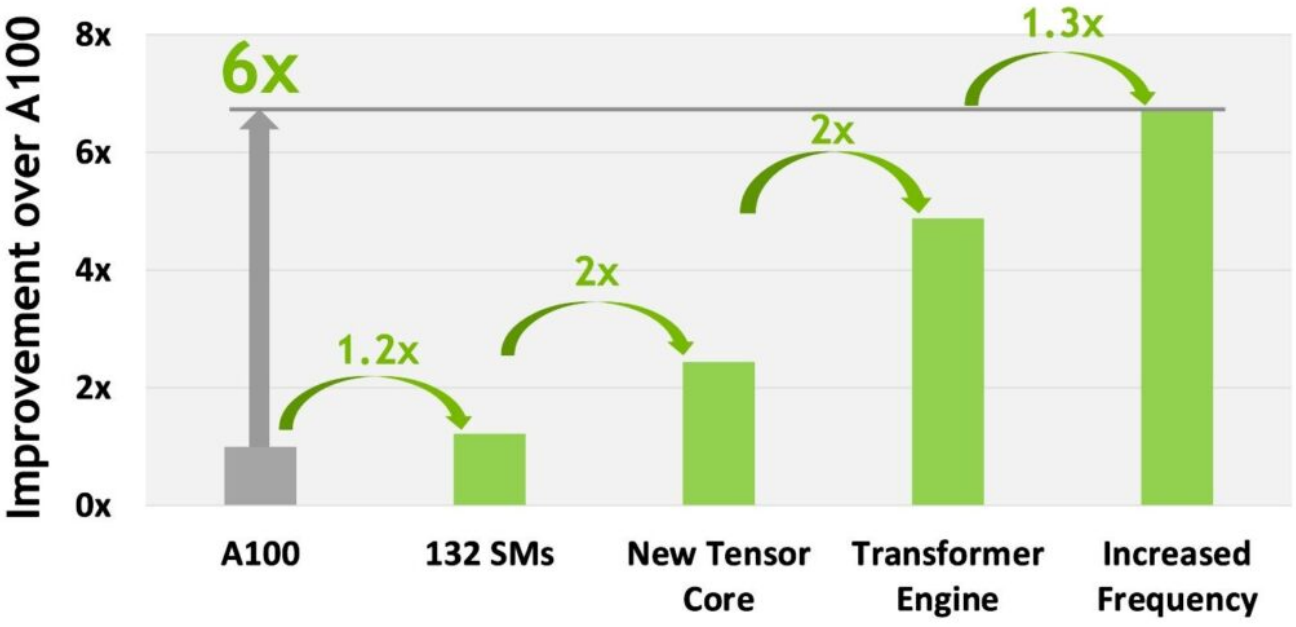
FP16 Tensor TFLOPS di picco con FP16 Accumulate	312/624	1000/2000	800/1600
FP16 Tensor TFLOPS di picco con FP32 Accumulate	312/624	1000/2000	800/1600
BF16 Tensor TFLOPS di picco con FP32 Accumulate	312/624	1000/2000	800/1600
TF32 Tensor TFLOPS picco	156/312	500/1000	400/800
FP64 Tensor TFLOPS picco	19.5	60	48
INT8 Tensor TOPS picco	624/1248	2000/4000	1600/3200
FP16 TFLOPS picco (non-Tensor)	78	120	96
BF16 TFLOPS picco (non-Tensor)	39	120	96
FP32 TFLOPS picco (non-Tensor)	19.5	60	48
FP64 TFLOPS picco (non-Tensor)	9.7	30	24
INT32 TOPS picco	19.5	30	24

Texture Unit	432	528	456
Interfaccia memoria	5120-bit HBM2	5120-bit HBM3	5120-bit HBM2e
Memoria	40 GB	80 GB	80 GB
Data Rate memoria (Non finalizzato per H100)	1215 MHz DDR	Non dichiarato	Non dichiarato
Bandwidth memoria	1555 GB/s	3000 GB/s	2000 GB/s
Cache L2	40 MB	50 MB	50 MB
Dimensione memoria condivisa / SM	Configurabile fino a 164 KB	Configurabile fino a 228 KB	Configurabile fino a 228 KB
Dimensione file Register / SM	256 KB	256 KB	256 KB
Dimensione file Register / GPU	27648 KB	33792 KB	29184 KB
TDP	400W	700W	350W
Transistor	54,2 miliardi	80 miliardi	80 miliardi
Dimensione die GPU	826 mm2	814 mm2	814 mm2

Processo produttivo TSMC	7 nm N7	4N personalizzato per NVIDIA	4N personalizzato per NVIDIA

La versione SXM di NVIDIA H100 non ha una GPU totalmente abilitata, ci sono infatti 132 SM per un totale di **16986 CUDA core**, 528 Tensor core e 50 MB di cache L2. Accanto alla GPU troviamo 80 GB memoria HBM3 gestiti tramite 10 controller a 512 bit. **La versione in formato PCI Express 5.0** prevede 114 SM per **14592 CUDA core**, 456 Tensor core e 50 MB di cache L2. Anche in questo caso abbiamo 80 GB di memoria, stavolta HBM2E.

“**Venti GPU H100 possono sostenere l’equivalente del traffico Internet mondiale**, consentendo ai clienti di fornire sistemi di raccomandazione avanzati e modelli linguistici di grandi dimensioni che eseguono inferenze sui dati in tempo reale”, dichiara NVIDIA annunciando che **il chip è in produzione e debutterà nel terzo trimestre**.



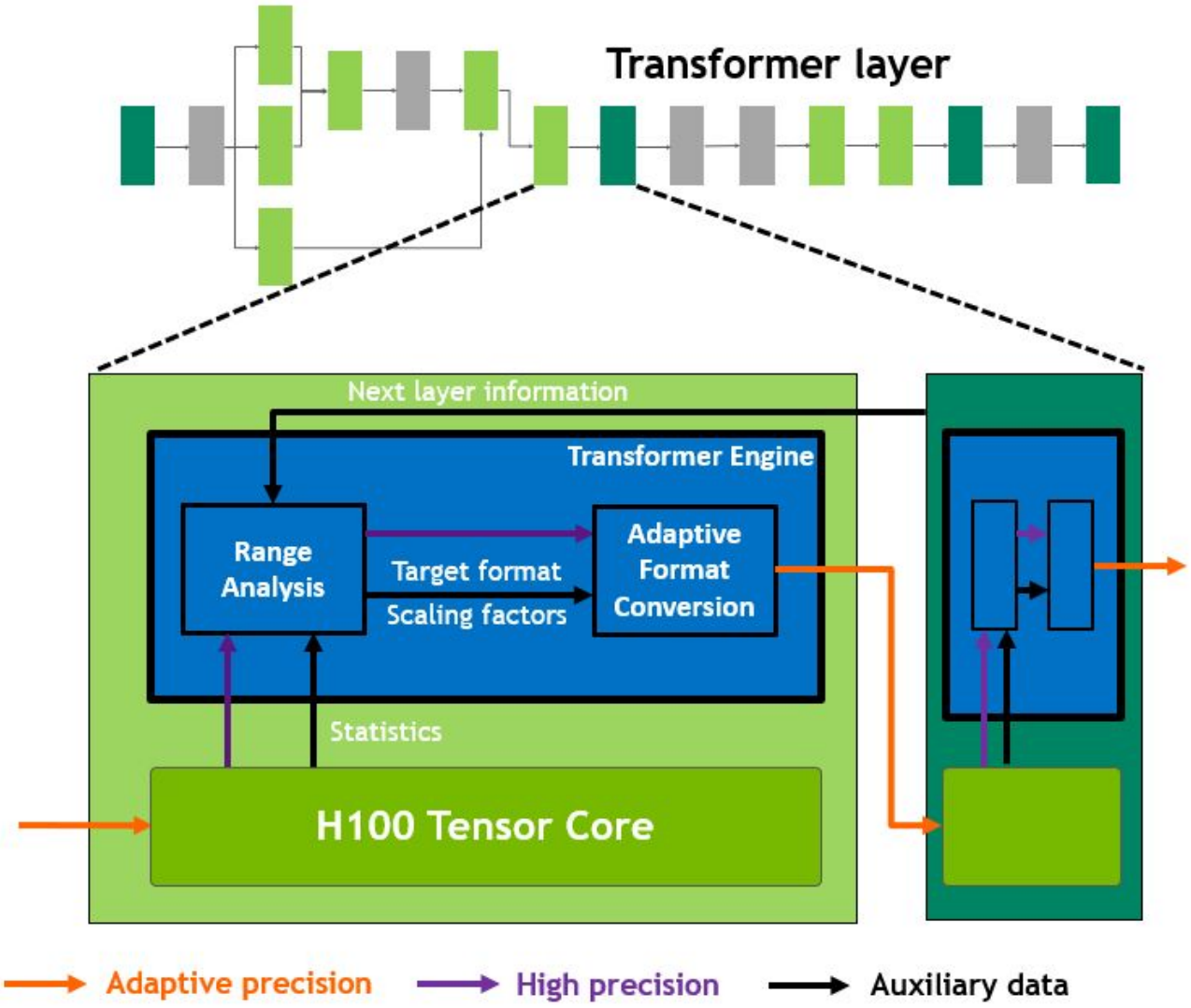
L’azienda parla di **sei novità fondamentali per l’architettura**

Hopper, che insieme permettono di offrire prestazioni senza precedenti. La novità più importante di questo acceleratore prende il nome di **Transformer Engine**, ovvero una **soluzione hardware dedicata (sfrutta i ben noti Tensor core) al processo di linguaggio naturale e altri compiti** grazie alla sua duttilità. Il Transformer Engine è stato pensato per **velocizzare queste reti fino a sei volte rispetto alla generazione precedente** senza perdere accuratezza.

NVIDIA Transformer Engine: cos'è e perché è importante

"**I modelli AI più grandi possono richiedere mesi** per l'addestramento sulle piattaforme di calcolo odierne. Il **Transformer Engine**, parte della nuova architettura Hopper, velocizzerà significativamente le prestazioni e le capacità di IA, e **aiuterà ad addestrare grandi modelli in giorni oppure ore**", spiega NVIDIA.

I **modelli della famiglia Transformer** sono la spina dorsale dei modelli linguistici usati ampiamente oggi, come **asBERT e GPT-3**. Inizialmente sviluppati per l'elaborazione del linguaggio naturale, la loro versatilità permette di applicarli ad altri compiti come la visione artificiale, la scoperta di farmaci e molto altro.

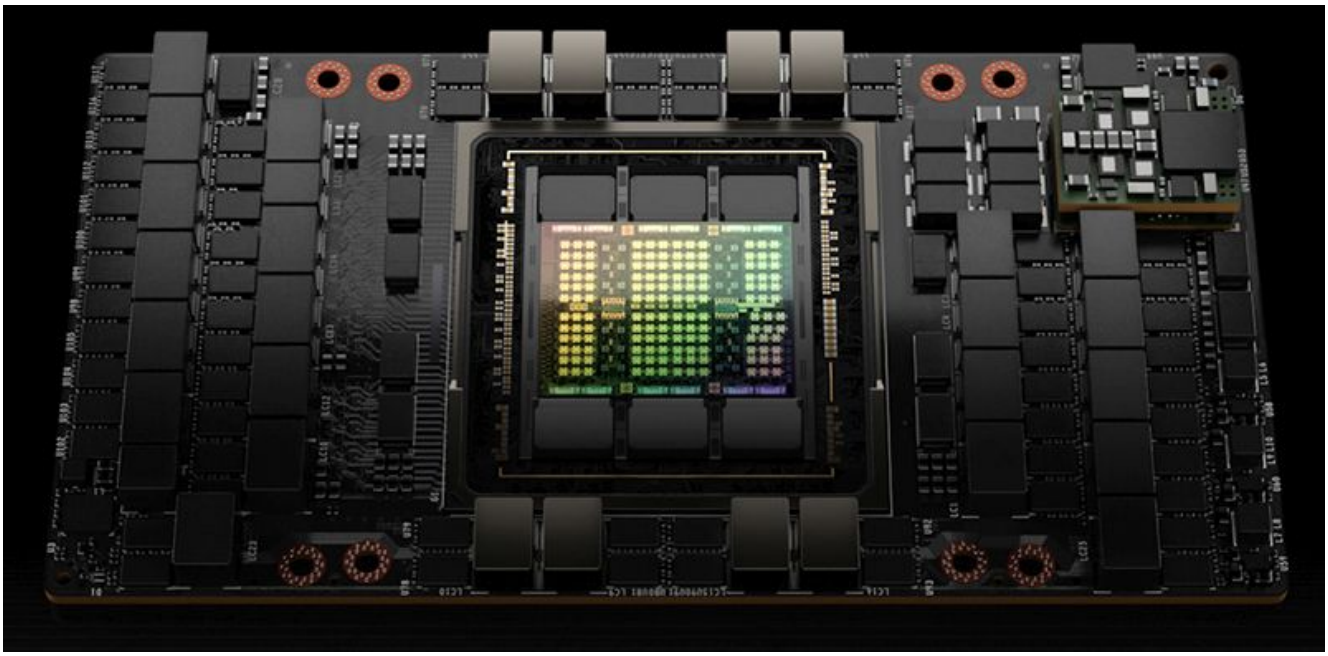


Le dimensioni dei modelli continuano a crescere in modo esponenziale, raggiungendo mille miliardi di parametri e questo sta causando un allungamento dei tempi di allenamento in mesi, intervallo che non si sposa bene con i tempi di sviluppo di un'azienda. Il Transformer Engine **usa la precisione in virgola mobile a 16 bit e il nuovo formato dati a 8 bit combinati con algoritmi software avanzati** per tagliare questi tempi in modo drastico.

L'addestramento di intelligenze artificiali si basa su numeri in virgola mobile, e la maggior parte dei calcoli è svolta usando l'half precision a 16 bit (FP16), la singola precisione a 32 bit (FP32) e, per operazioni specializzate, la doppia precisione a 64 bit (FP64). **Riducendo i calcoli a soli 8 bit,**

Transformer Engine consente di allenare reti di grandi dimensioni più rapidamente.

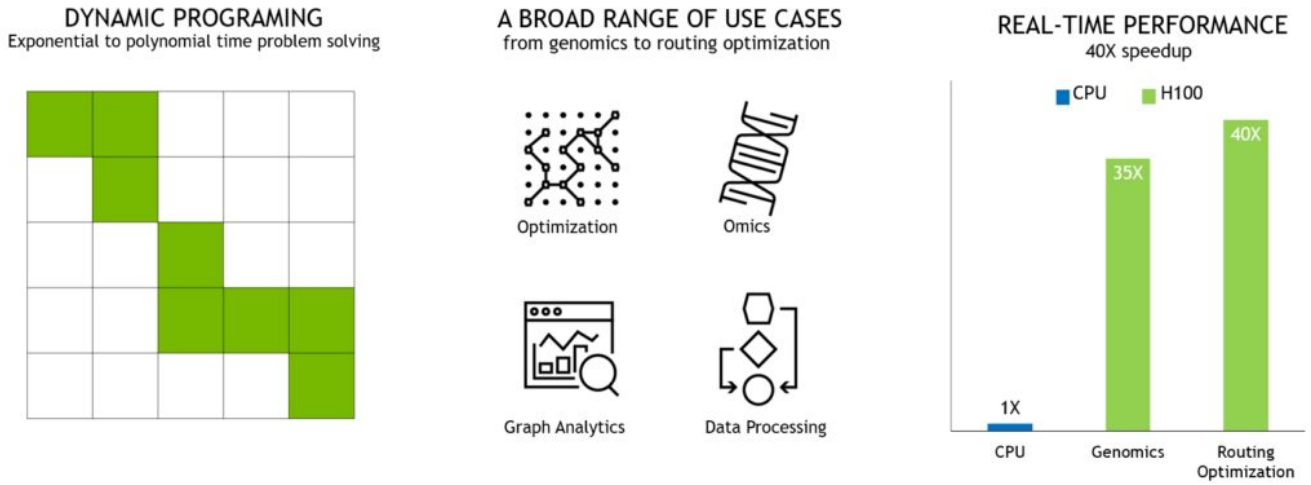
Il Transformer Engine usa software e i Tensor Core di quarta generazione di Hopper per applicare formati misti FP8 e FP16 per accelerare drasticamente i calcoli di intelligenza artificiale. Il problema è garantire un'elevata accuratezza preservando al tempo stesso le prestazioni di formati numerici più piccoli e veloci. Il Transformer Engine consente tutto questo con un'euristica messa a punto da NVIDIA che sceglie dinamicamente tra calcoli FP8 e FP16 e gestisce automaticamente il recasting e il ridimensionamento tra queste precisioni in ogni layer.



L'architettura NVIDIA Hopper migliora anche i Tensor Core di quarta generazione triplicando le operazioni in virgola mobile al secondo rispetto alle precisioni TF32, FP64, FP16 e INT8 di generazione precedente. Combinando tutto, le innovazioni garantiscono maggiore produttività e una **riduzione del tempo di allenamento di 9 volte, passando sette giorni a sole 20 ore**. Inoltre, il Transformer Engine può essere usato anche per l'inferenza senza alcuna conversione di formato dati.

NVIDIA: le istruzioni DPX per accelerare la programmazione dinamica

Con le istruzioni DPX NVIDIA Hopper accelera la programmazione dinamica – una tecnica di problem solving usata in algoritmi destinati a genomica, quantum computing e altro – **fino a 40 volte rispetto alle CPU e fino a 7 volte rispetto alle GPU di precedente generazione**.



La programmazione dinamica si basa su due tecniche chiamate **ricorsione e memoizzazione** e contempla un algoritmo come quello di Floyd-Warshall per trovare percorsi ottimali per flotte di robot autonomi in ambienti dinamici e l'algoritmo Smith-Waterman usato nell'allineamento delle sequenze per la classificazione e il ripiegamento del DNA e delle proteine.

Le altre novità di NVIDIA H100 e Hopper

La **seconda generazione della Multi-Instance GPU** permette a una

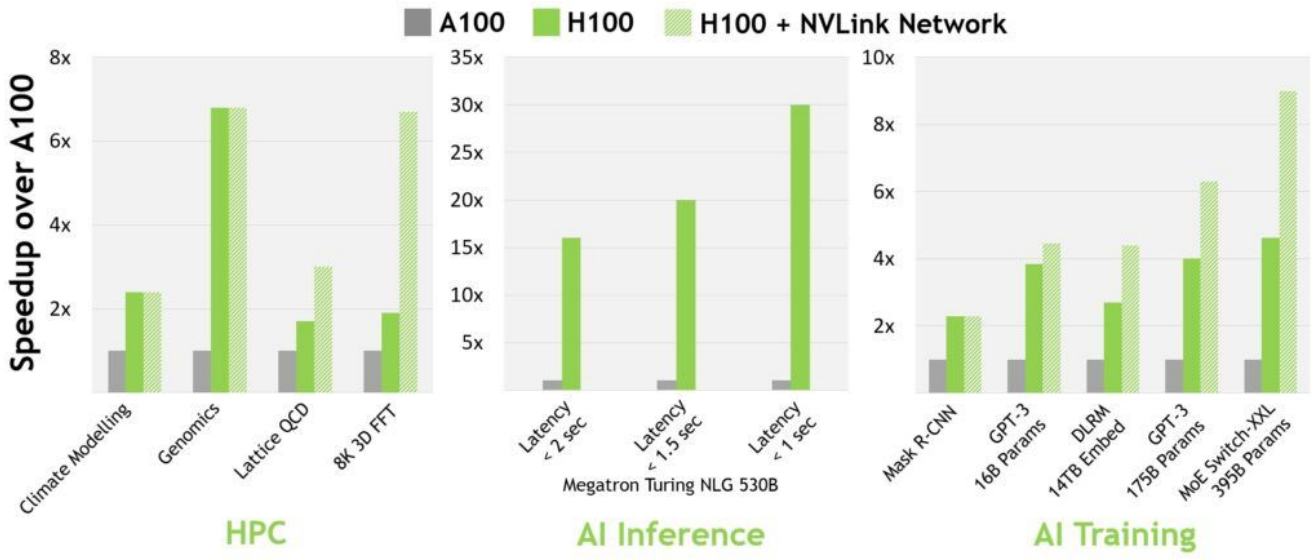
singola GPU di essere partizionata in sette istanze più piccole e completamente isolate per gestire differenti tipi di lavoro. L'architettura Hopper **estende le capacità di MIG fino a 7 volte rispetto alla precedente generazione** offrendo configurazioni multitenant sicure in ambienti cloud lungo ogni istanza di GPU.

H100 è anche **il primo acceleratore al mondo con capacità di confidential computing** per proteggere i modelli di IA e i dati mentre vengono processati. I clienti possono applicare il confidential computing all'apprendimento federato per industrie dove la privacy è sensibile come i servizi di carattere sanitario e finanziario, nonché su infrastrutture cloud condivise.

L'altra novità è la **quarta generazione dell'interconnessione NVIDIA NVLink**. Per accelerare i modelli di IA più grandi, NVLink si combina a un nuovo NVLink Switch esterno per stendere NVLink oltre il server, **connettendo fino a 256 GPU H100 con una bandwidth 9 volte maggiore** rispetto alla precedente generazione usando NVIDIA HDR Quantum InfiniBand.

Prestazioni senza precedenti per l'intelligenza artificiale

Tutte queste innovazioni permetteranno a **NVIDIA H100** di portare l'intelligenza artificiale a nuovo livello. H100 consentirà ai chatbot di usare il modello Megatron 530B con **un throughput fino a 30 volte superiore rispetto alla precedente generazione**, soddisfacendo al contempo il requisito di una latenza inferiore al secondo richiesta per l'IA conversazionale in tempo reale.



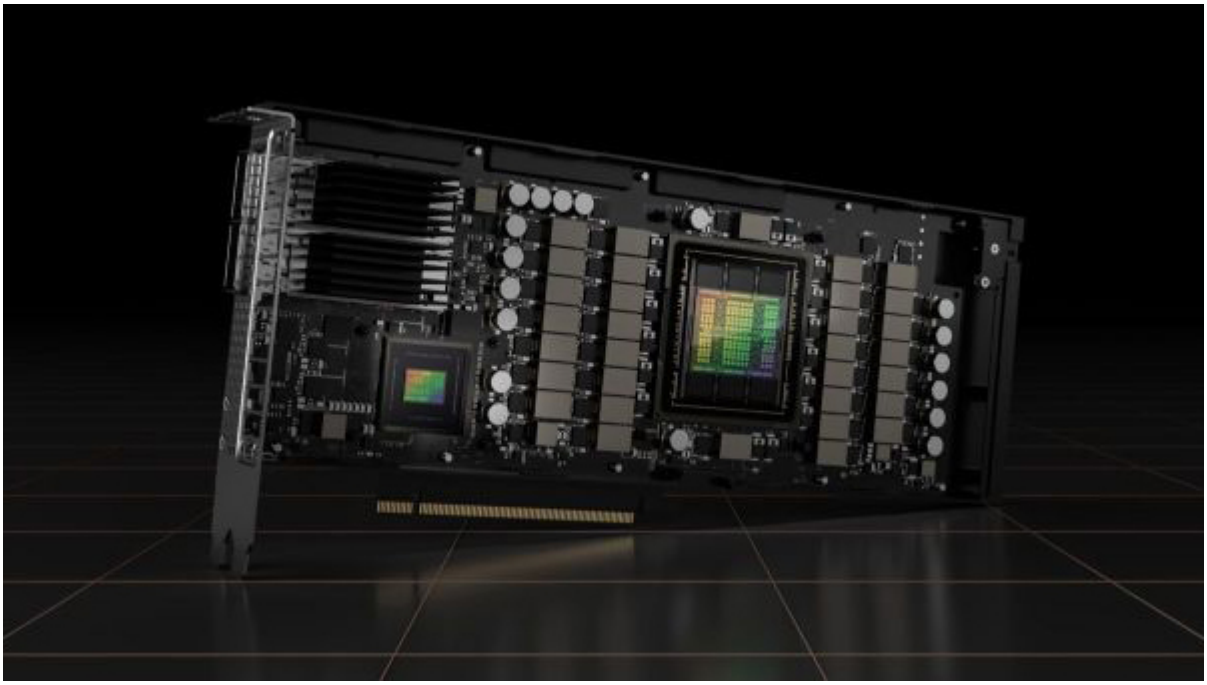
H100 consentirà inoltre a ricercatori e sviluppatori di **addestrare modelli massicci come Mixture of Experts, con 395 miliardi di parametri, fino a 9 volte più velocemente**, riducendo il tempo di formazione da settimane a giorni.

NVIDIA H100 Hopper in tutte le salse dal terzo trimestre

NVIDIA prevede un'ampia adozione di NVIDIA H100 grazie a molteplici versioni pronte ad adattarsi a ogni server. Il debutto è previsto a partire dal terzo trimestre.

Il sistema DGX di quarta generazione di NVIDIA, **DGX H100**, prevede **otto GPU H100 per fornire 32 petaflop** con la nuova precisione FP8. Ogni GPU nei sistemi DGX H100 è collegata da NVLink di quarta generazione per una connettività di 900 GB/s, 1,5 volte superiore rispetto alla generazione precedente. NVSwitch permette a tutte e otto le GPU H100 di connettersi tramite NVLink e **un NVLink Switch esterno consente di collegare in rete fino a 32 nodi DGX H100** nei supercomputer NVIDIA DGX SuperPOD.

Alibaba Cloud, Amazon Web Services, Baidu AI Cloud, Google Cloud, Microsoft Azure, Oracle Cloud e Tencent Cloud prevedono di offrire **istanze basate su H100**, inoltre gli OEM Atos, BOXX Technologies, Cisco, Dell Technologies, Fujitsu, GIGABYTE, H3C, Hewlett Packard Enterprise, Inspur, Lenovo, Nettrix e Supermicro prevedono di realizzare un'ampia gamma di server basati su questo acceleratore.



NVIDIA H100 CNX

H100 sarà disponibile nei fattori di forma SXM e PCI Express. La versione SXM di NVIDIA H100 permetterà di creare configurazioni a quattro e otto vie, mentre quella PCIe permetterà di collegare due GPU tramite NVLink. Inoltre, arriverà un "**acceleratore convergente**" chiamato **H100 CNX** che accoppia una GPU H100 con NVIDIA ConnectX-7 400 Gb/s InfiniBand e Ethernet SmartNIC.

Le GPU basate sull'architettura NVIDIA Hopper potranno anche essere **accoppiate con le CPU NVIDIA Grace tramite un'interconnessione NVLink-C2C ultraveloce** per una comunicazione oltre 7 volte più veloce tra CPU e GPU rispetto al PCI Express 5.0.

Read More