NVIDIA DGX-2
IL SISTEMA DI DEEP LEARNING PIÙ POTENTE DEL MONDO PER LE SFIDE IA PIÙ COMPLESSE

La sfida della scalabilità per soddisfare le moderne esigenze di IA e deep learning

Le reti neurali profonde stanno crescendo rapidamente in dimensioni e complessità, in risposta alle sfide più esigenti in termini di business e ricerca. La capacità computazionale necessaria per supportare i carichi di lavoro IA moderni è di gran lunga superiore alle capacità dei data center tradizionali. Le tecniche moderne che sfruttano sempre di più l'uso del parallelismo nei modelli collidono con i limiti della larghezza di banda intra-GPU, dato che gli sviluppatori creano cluster di elaborazione accelerati sempre più voluminosi, spingendo al limite i data center. Serve un nuovo approccio che garantisca scalabilità di elaborazione IA quasi illimitata, al fine di abbattere le barriere e ottenere informazioni fruibili più rapidamente per trasformare il mondo.

Prestazioni per ottenere risultati prima ritenuti impossibili con il training

L'IA diventa sempre più complessa e richiede livelli elaborazione sempre maggiori. NVIDIA® DGX-2™ è il primo sistema a 2 petaFLOPS al mondo in grado di contenere la potenza di 16 tra le GPU più avanzate, accelerando i più recenti modelli di deep learning prima impossibili da addestrare. Con la scalabilità innovativa delle GPU, è possibile addestrare modelli 4 volte più grandi su un singolo nodo. Rispetto alle architetture x86 tradizionali, la capacità del sistema DGX-2 di addestrare ResNet-50 richiederebbe l'equivalente di 300 server con doppia GPU Intel Xeon Gold per un costo complessivo di oltre 2,7 milioni di dollari.

NVIDIA NVSwitch: un fabric di rete IA rivoluzionario

La ricerca all'avanguardia richiede la libertà di sfruttare il parallelismo dei modelli e necessità di livelli di larghezza di banda intra-GPU mai immaginati prima. NVIDIA ha creato NVSwitch per affrontare questa esigenza. Come per l'evoluzione dal cavo alla banda larga ultra veloce, NVSwitch fornisce un fabric di rete per il futuro, ma lo fa oggi. Con NVIDIA DGX-2, la complessità e le dimensioni del modello non sono più un problema perché i limiti delle architetture tradizionali vengono finalmente superati. Adotta una metodologia di training dei modelli in parallelo con un fabric di rete in DGX-2 che garantisce 2,4 TB/s di larghezza di banda bisezione per una capacità 24 volte superiore rispetto alle precedenti generazioni. Questa nuova super autostrada di interconnessione apre le porte a possibilità infinite per quanto riguarda le tipologie di modelli con cui sfruttare la potenza del training distribuito su 16 GPU contemporaneamente.

SPECIFICHE DEL SISTEMA

GPU	**16 NVIDIA® Tesla V100**
Memoria della GPU	**512 GB totali**
Prestazioni	**2 petaFLOPS**
Core NVIDIA CUDA®	**81920**
NVIDIA Tensor Core	**10240**
NVSwitch	**12**
Consumo energetico massimo	**10 kW**
CPU	**Dual Intel Xeon Platinum 8168, 2,7 GHz, 24 core**
Memoria di sistema	**1,5 TB**
Rete	**8 Infiniband 100 Gb/sec/ Ethernet 100 GigE Dual 10/25 Gb/sec**
Spazio di archiviazione	**SO: 2 SSD NVME 960 GB Memoria interna: SSD NVME 30 TB (8 x 3,84 TB)**
Software	**Ubuntu Linux OS Vedi lo stack software per i dettagli**
Peso del sistema	**154,2 kg**
Dimensioni del sistema	**Altezza: 440 mm Larghezza: 482,3 Lunghezza: 795,4 mm - Senza Bezel anteriore 834 mm - Con Bezel anteriore**
Temperatura di funzionamento	**da 5°C a 35°C**

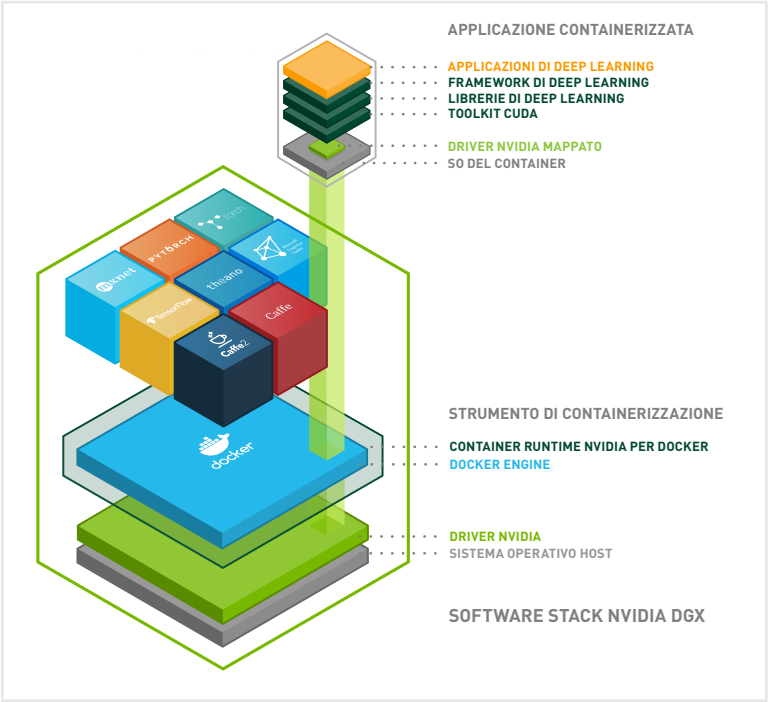
Nuovi livelli di scalabilità IA

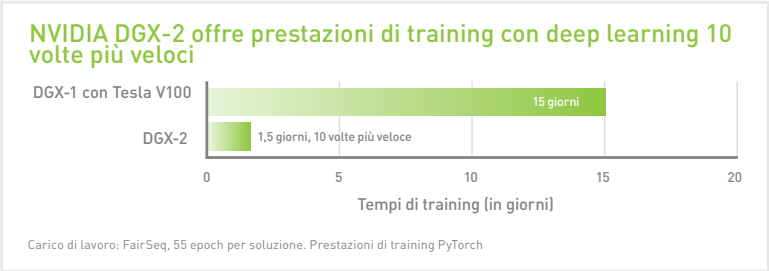
Le imprese moderne devono essere in grado di distribuire rapidamente la potenza dell'IA in risposta agli imperativi aziendali e riuscire a scalare i sistemi senza aumentare i costi e le complessità. Abbiamo creato DGX-2 e lo abbiamo alimentato con il software DGX che permette distribuzioni accelerate e operazioni semplificate con la massima scalabilità. DGX-2 offre una soluzione pronta all'uso con il percorso più rapido per la scalabilità IA, unitamente al supporto per la virtualizzazione, per consentire la creazione di un cloud IA di livello aziendale totalmente personalizzato. Le nuove imprese possono sfruttare la potenza illimitata dell'IA in una soluzione che scala senza fatica con un'infrastruttura di rete molto più agile e flessibile per integrare e accelerare tutte le risorse di elaborazione. Con un modello di distribuzione accelerata e un'architettura progettata appositamente per semplificare la scalabilità, il tuo team potrà dedicare più tempo all'analisi e meno alla manutenzione dell'infrastruttura.

Infrastruttura IA di livello aziendale

Se la piattaforma IA è critica per la tua impresa, ti occorre un'infrastruttura affidabile, disponibile e utilizzabile (RAS). DGX-2 è una soluzione di livello aziendale progettata per garantire una disponibilità illimitata delle operazioni IA, garantire il RAS e ridurre i tempi di fermo non pianificati, semplificare l'usabilità e assicurare la continuità operativa.

Dedica meno tempo all'ottimizzazione e all'adeguamento e concentra le tue risorse sulla ricerca. Il supporto di classe aziendale offerto da NVIDIA ti aiuta a risparmiare i tempi finora dedicati alla risoluzione di problemi dell'hardware e del software open source. Con ogni sistema DG, puoi avviare velocemente i tuoi progetti, eseguire training in tempi rapidi e mantenere ritmi elevati con una soluzione integrata che include software, strumenti e competenze NVIDIA.





Per ulteriori informazioni, visita **nvidia.it/data-center/dgx-2**

NVIDIA DGX-2
IL SISTEMA DI DEEP LEARNING PIÙ POTENTE DEL MONDO PER LE SFIDE IA PIÙ COMPLESSE

La sfida della scalabilità per soddisfare le moderne esigenze di IA e deep learning

Le reti neurali profonde stanno crescendo rapidamente in dimensioni e complessità, in risposta alle sfide più esigenti in termini di business e ricerca. La capacità computazionale necessaria per supportare i carichi di lavoro IA moderni è di gran lunga superiore alle capacità dei data center tradizionali. Le tecniche moderne che sfruttano sempre di più l'uso del parallelismo nei modelli collidono con i limiti della larghezza di banda intra-GPU, dato che gli sviluppatori creano cluster di elaborazione accelerati sempre più voluminosi, spingendo al limite i data center. Serve un nuovo approccio che garantisca scalabilità di elaborazione IA quasi illimitata, al fine di abbattere le barriere e ottenere informazioni fruibili più rapidamente per trasformare il mondo.

Prestazioni per ottenere risultati prima ritenuti impossibili con il training

L'IA diventa sempre più complessa e richiede livelli elaborazione sempre maggiori. NVIDIA® DGX-2™ è il primo sistema a 2 petaFLOPS al mondo in grado di contenere la potenza di 16 tra le GPU più avanzate, accelerando i più recenti modelli di deep learning prima impossibili da addestrare. Con la scalabilità innovativa delle GPU, è possibile addestrare modelli 4 volte più grandi su un singolo nodo. Rispetto alle architetture x86 tradizionali, la capacità del sistema DGX-2 di addestrare ResNet-50 richiederebbe l'equivalente di 300 server con doppia GPU Intel Xeon Gold per un costo complessivo di oltre 2,7 milioni di dollari.

NVIDIA NVSwitch: un fabric di rete IA rivoluzionario

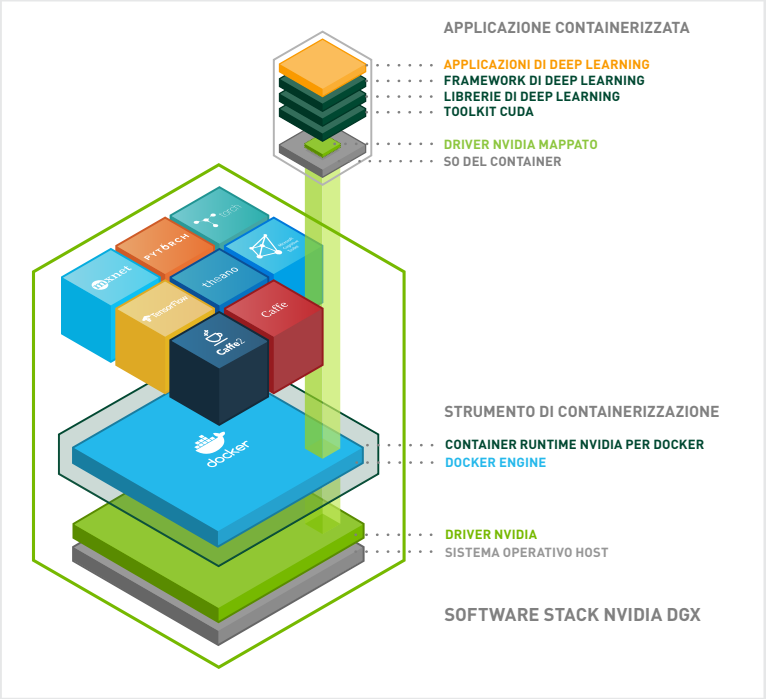
La ricerca all'avanguardia richiede la libertà di sfruttare il parallelismo dei modelli e necessità di livelli di larghezza di banda intra-GPU mai immaginati prima. NVIDIA ha creato NVSwitch per affrontare questa esigenza. Come per l'evoluzione dal cavo alla banda larga ultra veloce, NVSwitch fornisce un fabric di rete per il futuro, ma lo fa oggi. Con NVIDIA DGX-2, la complessità e le dimensioni del modello non sono più un problema perché i limiti delle architetture tradizionali vengono finalmente superati. Adotta una metodologia di training dei modelli in parallelo con un fabric di rete in DGX-2 che garantisce 2,4 TB/s di larghezza di banda bisezione per una capacità 24 volte superiore rispetto alle precedenti generazioni. Questa nuova super autostrada di interconnessione apre le porte a possibilità infinite per quanto riguarda le tipologie di modelli con cui sfruttare la potenza del training distribuito su 16 GPU contemporaneamente.

SPECIFICHE DEL SISTEMA

GPU	**16 NVIDIA® Tesla V100**
Memoria della GPU	**512 GB totali**
Prestazioni	**2 petaFLOPS**
Core NVIDIA CUDA®	**81920**
NVIDIA Tensor Core	**10240**
NVSwitch	**12**
Consumo energetico massimo	**10 kW**
CPU	**Dual Intel Xeon Platinum 8168, 2,7 GHz, 24 core**
Memoria di sistema	**1,5 TB**
Rete	**8 Infiniband 100 Gb/sec/ Ethernet 100 GigE Dual 10/25 Gb/sec**
Spazio di archiviazione	**SO: 2 SSD NVME 960 GB Memoria interna: SSD NVME 30 TB (8 x 3,84 TB)**
Software	**Ubuntu Linux OS Vedi lo stack software per i dettagli**
Peso del sistema	**154,2 kg**
Dimensioni del sistema	**Altezza: 440 mm Larghezza: 482,3 Lunghezza: 795,4 mm - Senza Bezel anteriore 834 mm - Con Bezel anteriore**
Temperatura di funzionamento	**da 5°C a 35°C**

Nuovi livelli di scalabilità IA

Le imprese moderne devono essere in grado di distribuire rapidamente la potenza dell'IA in risposta agli imperativi aziendali e riuscire a scalare i sistemi senza aumentare i costi e le complessità. Abbiamo creato DGX-2 e lo abbiamo alimentato con il software DGX che permette distribuzioni accelerate e operazioni semplificate con la massima scalabilità. DGX-2 offre una soluzione pronta all'uso con il percorso più rapido per la scalabilità IA, unitamente al supporto per la virtualizzazione, per consentire la creazione di un cloud IA di livello aziendale totalmente personalizzato. Le nuove imprese possono sfruttare la potenza illimitata dell'IA in una soluzione che scala senza fatica con un'infrastruttura di rete molto più agile e flessibile per integrare e accelerare tutte le risorse di elaborazione. Con un modello di distribuzione accelerata e un'architettura progettata appositamente per semplificare la scalabilità, il tuo team potrà dedicare più tempo all'analisi e meno alla manutenzione dell'infrastruttura.



Infrastruttura IA di livello aziendale

Se la piattaforma IA è critica per la tua impresa, ti occorre un'infrastruttura affidabile, disponibile e utilizzabile (RAS). DGX-2 è una soluzione di livello aziendale progettata per garantire una disponibilità illimitata delle operazioni IA, garantire il RAS e ridurre i tempi di fermo non pianificati, semplificare l'usabilità e assicurare la continuità operativa.

Dedica meno tempo all'ottimizzazione e all'adeguamento e concentra le tue risorse sulla ricerca. Il supporto di classe aziendale offerto da NVIDIA ti aiuta a risparmiare i tempi finora dedicati alla risoluzione di problemi dell'hardware e del software open source. Con ogni sistema DG, puoi avviare velocemente i tuoi progetti, eseguire training in tempi rapidi e mantenere ritmi elevati con una soluzione integrata che include software, strumenti e competenze NVIDIA.

Per ulteriori informazioni, visita **nvidia.it/data-center/dgx-2**