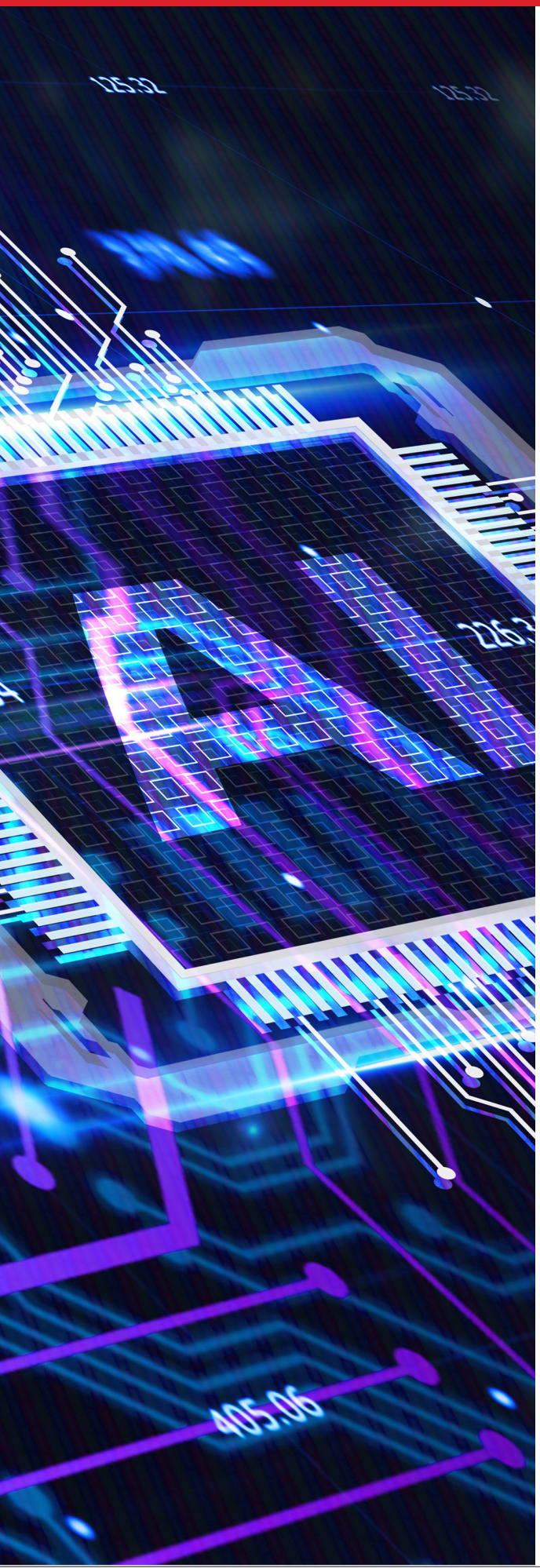


Affrontare le sfide dei carichi di lavoro di AI con il portafoglio AI Dell

Confronto tra il portafoglio AI Dell e le offerte analoghe di HPE

L'intelligenza artificiale (AI) ha senza dubbio trasformato il panorama aziendale e ha fornito a settori e organizzazioni di tutte le dimensioni gli strumenti per ottenere informazioni più approfondite dai propri dati, automatizzare i processi aziendali, offrire esperienze personalizzate per clienti e utenti ed essere più competitivi nel proprio settore. Per sfruttare efficacemente la potenza dell'AI, le organizzazioni hanno bisogno di un provider di infrastrutture in grado di offrire una soluzione integrata e completa che copra l'intero ciclo di vita dell'AI.

Grazie ai vendor di infrastrutture come Dell Technologies e Hewlett Packard Enterprise (HPE), i clienti possono far fronte alle crescenti esigenze dell'AI e gestirne le complessità intrinseche. Presentando portafogli predisposti per l'AI, tali vendor offrono diversi livelli di soluzioni AI che combinano soluzioni di infrastruttura on-premise e cloud a prestazioni elevate con partnership strategiche e una serie di servizi di supporto e consulenza.

Il presente report esamina informazioni pubblicamente disponibili sui portafogli* di AI Dell e HPE con l'obiettivo di evidenziare i vantaggi specifici dell'architettura, delle prestazioni e del supporto da cui i clienti possono trarre vantaggio scegliendo Dell Technologies per le proprie esigenze legate all'AI. Confrontiamo i dettagli dei server creati da Dell per supportare i deployment dell'AI e facciamo riferimento ai risultati dei test di benchmark del settore di ML Commons®. Analizziamo inoltre ulteriori offerte di software e servizi che supportano i clienti in ogni fase del loro percorso verso l'AI.

*Nota: PT ha completato tutte le ricerche entro il 5 dicembre 2023, pertanto questo documento non rispecchia le offerte o le modifiche delle versioni Dell o HPE successive a tale data.

Le sfide dell'adozione dell'AI

L'adozione di una strategia AI presenta molte nuove sfide per i data center e il personale IT che li gestisce, tra cui:

- Colmare le lacune di competenze esistenti del personale attuale tramite formazione interna sull'AI o assunzioni esterne.
- Comprendere le esigenze in termini di preparazione dei dati dell'AI, tra cui la qualità, la quantità, la posizione e lo stato attuale dei dati aziendali.
- Valutare gli specifici obiettivi aziendali relativi all'AI per stabilire al meglio quali modelli e implementazioni di AI offrono dei vantaggi.
- Valutare i requisiti computazionali, di rete e di storage dei sistemi AI pianificati ed elaborare un piano di acquisizione.

Questi sono solo alcuni esempi delle molteplici sfide, spesso significative, che un'azienda deve affrontare quando cerca di trarre il massimo dall'implementazione dell'AI nei propri data center.

Il portafoglio AI Dell mira a supportare i clienti nell'affrontare queste sfide attraverso servizi professionali e di consulenza, con cui i clienti possano creare roadmap di implementazione e preparare i propri dati per i modelli di AI.[1] Il portafoglio include anche corsi di formazione che coprono concetti di apprendimento automatico (ML) e altri argomenti e offre progettazioni convalidate per l'AI per garantire il successo dell'implementazione.[2] Dell collabora inoltre con terze parti per offrire ai clienti ulteriori strumenti di AI, come un portale Dell personalizzato all'interno della community Hugging Face con container e script dedicati per il deployment di modelli di AI open source[3] e il deployment semplificato del modello linguistico di grandi dimensioni (LLM) Meta Llama 2.[4] Insieme a un'ampia selezione di offerte di elaborazione e PC, dalle workstation portatili ai server che supportano fino a 8 GPU NVIDIA di fascia alta, Dell offre anche lo storage dei dati non strutturati richiesto dall'AI con un portafoglio di array di file e storage a oggetti a prestazioni elevate. Tali offerte di storage, tra cui Dell PowerScale, ObjectScale, ECS e storage integrato, sono in grado di gestire i dati non strutturati che vengono frequentemente utilizzati nei carichi di lavoro di AI.[5] Dell ha inoltre avviato una partnership con Snowflake per fornire una soluzione di storage su hybrid cloud per i clienti Dell.[6] Secondo un'analisi Dell, al mese di agosto 2023, l'azienda offre il "più ampio portafoglio di AI generativa", che va oltre i soli server e lo storage, fornendo risorse per l'intero percorso di implementazione dell'AI.[7]

Prestazioni dell'AI e opzioni di elaborazione accelerate: confronto tra Dell e HPE

I carichi di lavoro di AI possono utilizzare CPU, GPU o entrambe le opzioni come risorse computazionali a seconda delle dimensioni o del tipo di carico di lavoro. Alcune CPU forniscono acceleratori specifici per l'AI, come Intel Advanced Matrix Extensions (Intel AMX) nei più recenti processori scalabili Intel Xeon.[8] Le GPU sono spesso migliori per i carichi di lavoro più grandi e/o più complessi, ma il fattore di forma della GPU può influire sui livelli delle prestazioni. Ad esempio, alcune GPU NVIDIA A100 e H100 sono disponibili in fattori di forma PCIe universali o SXM proprietari; questi ultimi utilizzano l'architettura NVIDIA SXM a prestazioni più elevate.[9] Anche le grandi capacità di memoria e le funzionalità di progettazione dei server, come l'architettura di raffreddamento e l'efficienza energetica, influiscono sulle prestazioni. La maggior parte dei data center utilizza ancora il raffreddamento ad aria e ciò significa che i carichi di lavoro di elaborazione a prestazioni elevate (HPC) necessitano di server progettati per raffreddare con aria nel modo più efficace possibile. Di seguito, mettiamo in evidenza le offerte di server PowerEdge in termini di componenti, opzioni di raffreddamento e altro ancora, insieme ai punteggi MLPerf ® di MLCommons® pubblicati.

Prestazioni di benchmark dei modelli di AI: confronto dei risultati di MLPerf

MLPerf® è una suite di benchmark che testa le prestazioni dell'AI sia per l'addestramento che per l'inferenza. Affinché un'organizzazione pubblichi i risultati di MLPerf® ufficiali, questi devono essere conformi alle condizioni specifiche stabilite dallo sviluppatore di benchmark, MLCommons®.[10] Queste linee guida di conformità forniscono standard che semplificano il confronto delle prestazioni. Per i test di inferenza, MLPerf® utilizza dataset di Datacenter, Edge, Mobile e Tiny e indica i punteggi dell'AI e i watt di energia consumati durante i test. La suite di benchmark per l'inferenza include il test di molti modelli diffusi di AI, ML e DL; vedere la Tabella 1.

Tabella 1: Modelli di AI, ML e DL inclusi nei test di MLPerf® e casi d'uso tipici per ciascuno di essi. Fonte: Principled Technologies.

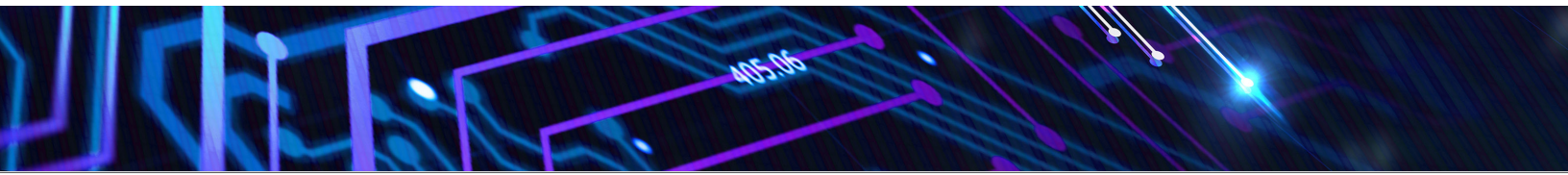
Modelli di AI comunemente utilizzati	Use case tipici
ResNet	Un modello di classificazione delle immagini che aiuta i computer a imparare, ricordare e identificare immagini diverse per casi d'uso come l'imaging medicale, la moderazione dei contenuti sui social media e il riconoscimento facciale
RetinaNet	Un tipo di rilevamento degli oggetti in grado di gestire ulteriori complessità rispetto a ResNet. I computer sono così in grado di identificare e localizzare gli oggetti all'interno di immagini o frame video e di classificarli in base all'importanza. Utilizzato per applicazioni come la guida autonoma, la tecnologia di assistenza automatica dei veicoli, la sorveglianza e il riconoscimento facciale
3D-UNet	Specifico per la segmentazione delle immagini per uso medicale
RNN-T	Riconoscimento vocale per casi d'uso come la traduzione linguistica automatizzata
BERT	Elaborazione del linguaggio naturale per casi d'uso come sintesi dei testi, traduzione linguistica e completamento automatico delle attività
DLRM-v2-99.9	Modello di raccomandazione per casi d'uso come annunci mirati e consigli personalizzati sui prodotti
GPTJ-99 e 99.9	LLM per l'elaborazione del linguaggio naturale che eccelle nella generazione di testo per casi d'uso come chatbot e strumenti di AI basati su chat

MLPerf

I risultati di MLPerf® includono diversi parametri oltre ai modelli di AI stessi e ciò può consentire l'analisi di molti dati in un singolo grafico o tabella. Di seguito è riportato un riferimento rapido a questi parametri:

- 99,0 e 99,9: questi numeri si riferiscono alla precisione per la quale il modello è stato addestrato. Maggiore è la precisione necessaria per l'output, maggiori sono la complessità del modello e il tempo necessario per elaborare i dati.
- Esempi offline/sec: modalità in cui il benchmark invia tutte le query all'inizio del test simulando i dati già presenti nel sistema.
- Query del server/sec: modalità in cui il benchmark invia query per tutta la durata del test simulando l'analisi di un flusso di dati in tempo reale.

Per ulteriori informazioni sui risultati MLCommons® e MLPerf®, visitare https://mlcommons.org/benchmarks/inference-datacenter/.



I risultati di questo report derivano dai risultati di MLPerf® v3.1 Inference Datacenter pubblicati sul sito MLCommons® da novembre 2023.[11] Questi risultati includono gli invii di produttori di tecnologia e Cloud Service Provider e coprono una vasta gamma di configurazioni. Rispetto agli invii pubblicamente disponibili di HPE, i server Dell hanno prodotto risultati migliori in alcuni modelli di AI. (Nota: configurazioni della GPU diverse tra i server possono rendere difficili i confronti diretti.) Vedere la Tabella 2 per i dettagli.

Tabella 2: Server Dell e HPE inclusi nei risultati di MLCommons® MLPerf® 3.1 pubblicati in data 29/11/2023.
Fonte: Principled Technologies.

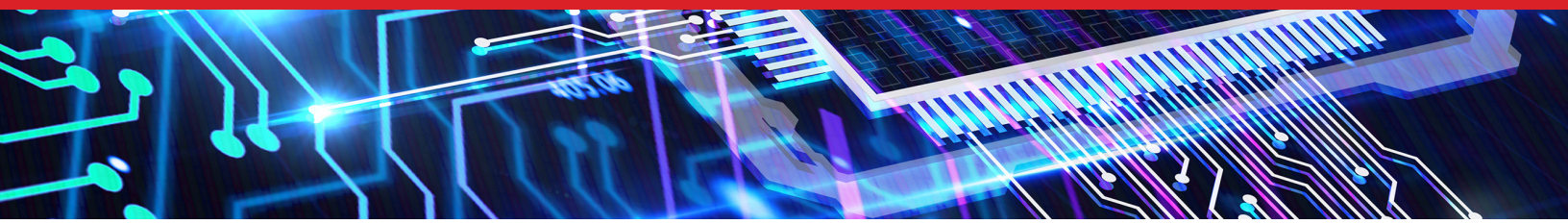
Mittente	Modello server	Numero e modello delle GPU	Description
Dell[12]	PowerEdge XE9680	8 NVIDIA H100 SXM	Per l'addestramento e l'inferenza dell'AI con carichi di lavoro elevati, ad esempio modelli linguistici di grandi dimensioni
	PowerEdge XE9640	4 NVIDIA H100 SXM	Per l'addestramento di modelli di AI di grandi dimensioni in data center ad alta densità e con raffreddamento a liquido
	PowerEdge XE8640	4 NVIDIA H100 SXM	Per promuovere applicazioni tradizionali di addestramento dell'AI, HPC e analisi dei dati in un fattore di forma 4U per data center con raffreddamento ad aria
	PowerEdge R760xa	4 NVIDIA H100 PCIe	Per un'ampia gamma di carichi di lavoro a elaborazione elevata, tra cui l'addestramento e l'inferenza di AI-ML/DL che non richiedono GPU ad alte prestazioni
HPE[13,14]	ProLiant XL675d Gen10 Plus	8 NVIDIA A100 SXM	Per high performance computing e AI
	ProLiant DL380a Gen11	4 NVIDIA H100 PCIe	Server 2U per carichi di lavoro di AI di media intensità

Confronto diretto tra i server Dell e HPE

Sebbene la strategia AI completa includa molto di più del solo hardware, garantire le prestazioni hardware più potenti è uno dei fattori essenziali per il successo dei carichi di lavoro di AI. Con la disponibilità di GPU e altre tecnologie sempre più nuove, si evolvono anche le capacità dei carichi di lavoro di AI. Al momento della pubblicazione dei risultati di MLPerf® v3.1, la migliore GPU NVIDIA disponibile era H100 Tensor Core con cui Dell ha pubblicato i risultati di MLPerf® in diversi server propri nei fattori di forma PCIe e SXM5.[15] I risultati HPE pubblicati includevano un solo invio di H100, soltanto con il fattore di forma PCIe. La nostra ricerca ha dimostrato che nessuno dei server HPE ProLiant e pochi server HPE compatibili con le GPU disponibili supportano il fattore di forma SXM5 H100 per le migliori prestazioni della GPU NVIDIA.[16] Come mostrato di seguito, avere GPU più potenti migliora in genere le prestazioni dei carichi di lavoro di AI.

Risultati di MLPerf con otto GPU

Dell PowerEdge XE9680 offre supporto per un massimo di otto GPU NVIDIA H100 SXM5 per l'accelerazione dell'AI e fino a due processori scalabili Intel® Xeon® di quarta generazione. La famiglia di prodotti PowerEdge XE dispone di un'architettura modulare che supporta le GPU NVIDIA SXM4 o SXM5 o gruppi GPU Open Compute Project Accelerator Module (OAM) di AMD, che possono migliorare le prestazioni rispetto a una GPU PCIe standard.[17] Occupando solo 6U di spazio su rack, PowerEdge XE9680 è un server compatto NVIDIA H100 SXM5 a otto vie. Attualmente, i server HPE ProLiant Gen11 più recenti non supportano il fattore di forma H100 SXM,[18] a differenza di alcuni server HPE Cray Supercomputing.[19] Poiché HPE non ha inviato alcun risultato di MLPerf® con i server Cray e mette in evidenza soltanto i server ProLiant nella pagina del proprio portafoglio AI, per il presente documento ci concentreremo sui server ProLiant. (vedere la Figura 1).



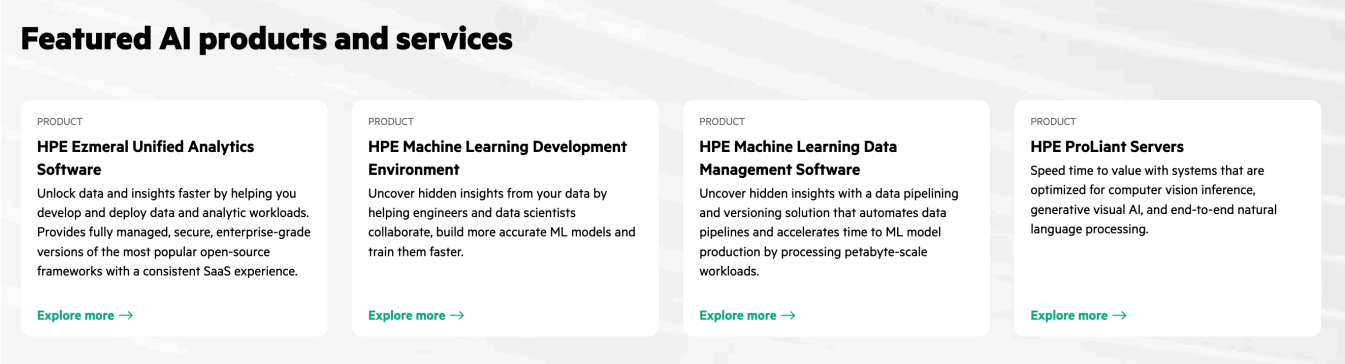


Figura 1. Screenshot dei prodotti e dei servizi di AI in primo piano nella pagina https://www.hpe.com/us/en/solutions/ai-artificial-intelligence.html che evidenzia i server HPE ProLiant al 5/12/2023.

Nei risultati di MLPerf® v3.1 pubblicati per la prima volta a novembre 2023 per server a otto GPU, le prestazioni di Dell PowerEdge XE9680 con GPU NVIDIA SXM5 H100 sono 4,25 volte superiori rispetto a quelle di HPE ProLiant XL675d Gen10 Plus con GPU NVIDIA SXM4 A100 (v. la Figura 2).

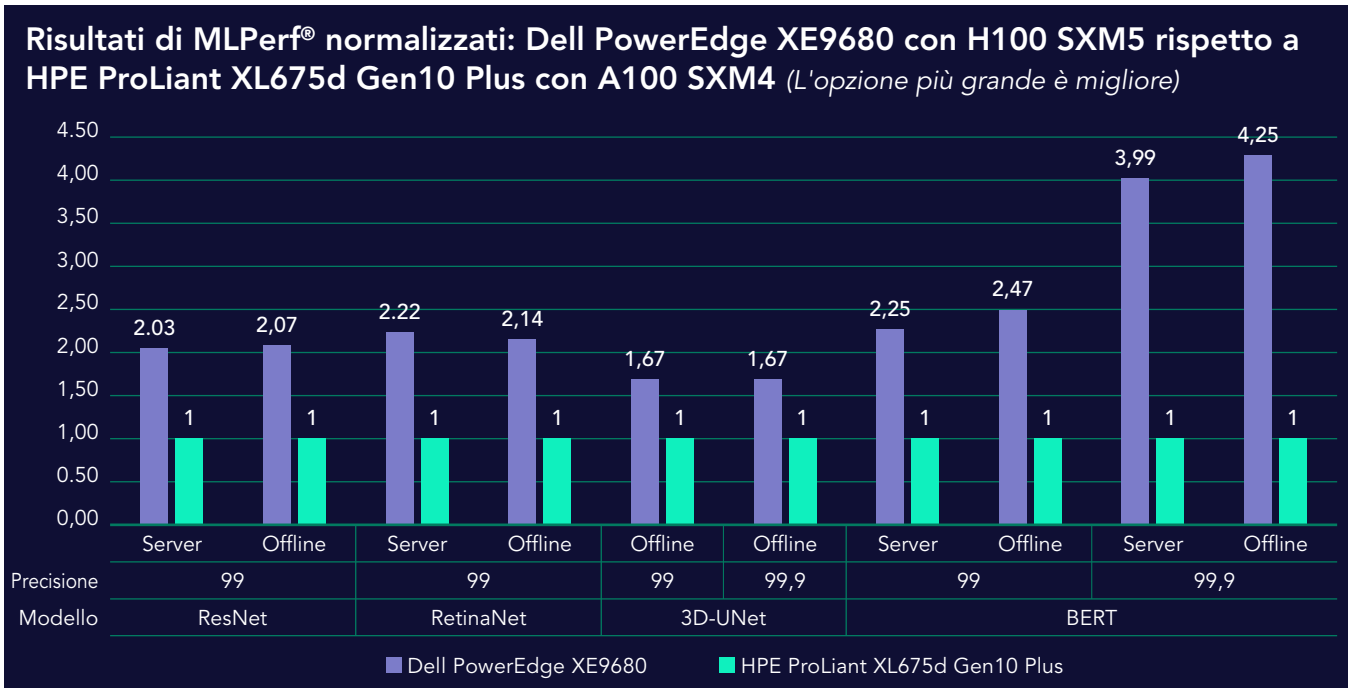


Figura 2. Risultati di MLPerf® pubblicati per Dell PowerEdge XE9680 e HPE ProLiant XL675d Gen10 Plus al 29/11/2023. Il sistema Dell utilizza la GPU NVIDIA H100, mentre le GPU nel sistema HPE sono di una generazione precedente. Fonte: Principled Technologies con i dati di MLCommons®.[20,21]

Per facilitare il confronto, abbiamo normalizzato i risultati dei test nelle Figure da 2 a 5. Ciò significa che assegniamo il valore di 1 a ciascun risultato di HPE ProLiant DL380a Gen 11 e mostriamo il risultato corrispondente di Dell PowerEdge R760xa a esso correlato. Come dimostrano questi risultati, anche la differenza di una sola generazione tra i modelli di GPU può influire in modo significativo sulle prestazioni che ci si aspetta di vedere in molteplici carichi di lavoro di AI.

Risultati di MLPerf con 4 GPU

Quando il risparmio di energia o spazio nel data center rappresenta un fattore chiave, Dell PowerEdge XE9640 2U potrebbe essere la soluzione giusta. Con un massimo di quattro GPU NVIDIA H100 SXM, PowerEdgeXE9640 offre la metà della potenza di elaborazione della GPU di XE9680, in due terzi di spazio in meno.[22] Dell PowerEdge XE9640, ad alta densità, integra la tecnologia Dell Smart Cooling, che fornisce una serie di tecnologie termiche, tra cui il raffreddamento a liquido diretto per CPU e GPU.[23]

Lo chassis 2U di PowerEdge XE9640 supporta meccanismi di circolazione dell'aria migliorati, tra cui ventole e dissipatori di calore più grandi, per raffreddare gli altri componenti essenziali, come schede PCIe e memoria.[24] PowerEdge XE9640 è attualmente l'unica offerta di Dell o HPE fornita con GPU HGX H100 2U a quattro vie. Il portafoglio AI di HPE offre server ProLiant Gen11 1U e 2U, ma questi sono limitati alle GPU con fattore di forma PCIe.[25]

Il server Dell PowerEdge XE9640 supporta anche GPU Intel Max Series 1550 OAM, che offrono una GPU a bassa potenza e ad alta densità che include una scheda PCIe e un OpenCompute Accelerator Module (OAM).[26] Non è stato possibile accertare se, a partire dal 5/12/2023, HPE offrisse un server con queste GPU, sebbene offra server HPE ProLiant DL380 Gen11 e DL380a Gen11 con GPU Intel Data Center Max 1100.[27] Ciò significa che Dell PowerEdge XE9640 potrebbe essere l'unica offerta attuale con quattro GPU Intel Max 1550 OAM in un server 2U. Per le aziende che si preoccupano per lo spazio nel data center e l'efficienza energetica, un server 2U con quattro GPU Intel Max 1550 offre una soluzione che coniuga high performance computing ed efficienza energetica senza sacrificare lo spazio del data center.

Nei risultati pubblicati di MLPerf® 3.1, le prestazioni di Dell PowerEdge XE9640 con quattro GPU HGX H100 sono risultate fino a 1,99 volte superiori rispetto a quelle di HPE ProLiant DL380a con quattro GPU PCIe H100 (v. la Figura 3).

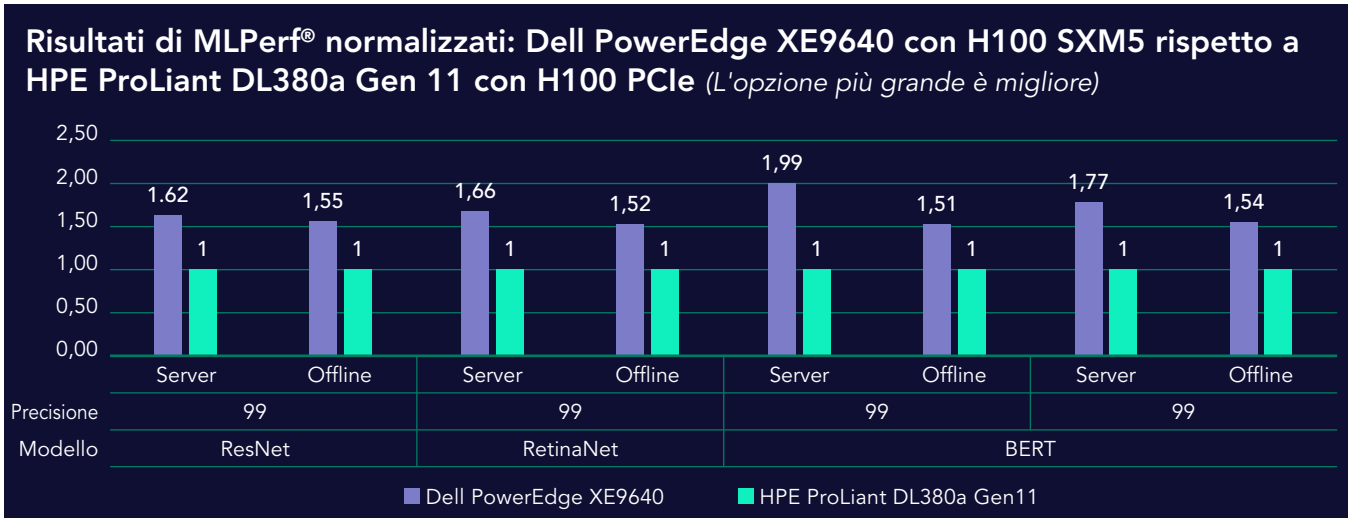


Figura 3. Risultati di MLPerf® pubblicati per Dell PowerEdge XE9680 e HPE ProLiant XL675d Gen10 Plus al 29/11/2023. Il sistema Dell utilizza la GPU NVIDIA H100, mentre le GPU nel sistema HPE sono di una generazione precedente. Fonte: Principled Technologies con i dati di MLCommons®.[28,29]

PowerEdge XE8640 offre una configurazione GPU a quattro vie con raffreddamento ad aria per i processori e un radiatore con raffreddamento ad aria assistita a liquido per le GPU, che non richiede la disponibilità di alcuna struttura di approvvigionamento di acqua per la rack. Per chi non utilizza o non può utilizzare liquidi di raffreddamento esterni,[30] Dell PowerEdge XE8640 4U supporta quattro GPU NVIDIA H100 SXM5 che forniscono la stessa potenza di elaborazione di PowerEdge XE9640 senza la necessità di raffreddamento a liquido diretto.[31]

Dell PowerEdge XE8640 include i più recenti processori scalabili Intel Xeon di quarta generazione e fino a 4 TB di memoria[32] per gestire dataset di grandi dimensioni e calcoli complessi diffusi nell'AI e nell'analisi dei dati. Anche in questo caso, HPE offre le GPU NVIDIA H100 SXM5 nei sistemi HPE Cray, ma i server abilitati per GPU HPE ProLiant non le supportano.

Rispetto ai dati di MLPerf® pubblicati a novembre 2023, il server PowerEdge XE8640 con quattro GPU NVIDIA H100 SXM5 ha raggiunto il throughput di AI più elevato tra tutte le quattro GPU inviate in nove diverse categorie. Come mostra la Figura 4, ha un punteggio fino a 2,07 volte superiore rispetto al server HPE ProLiant DL380a.

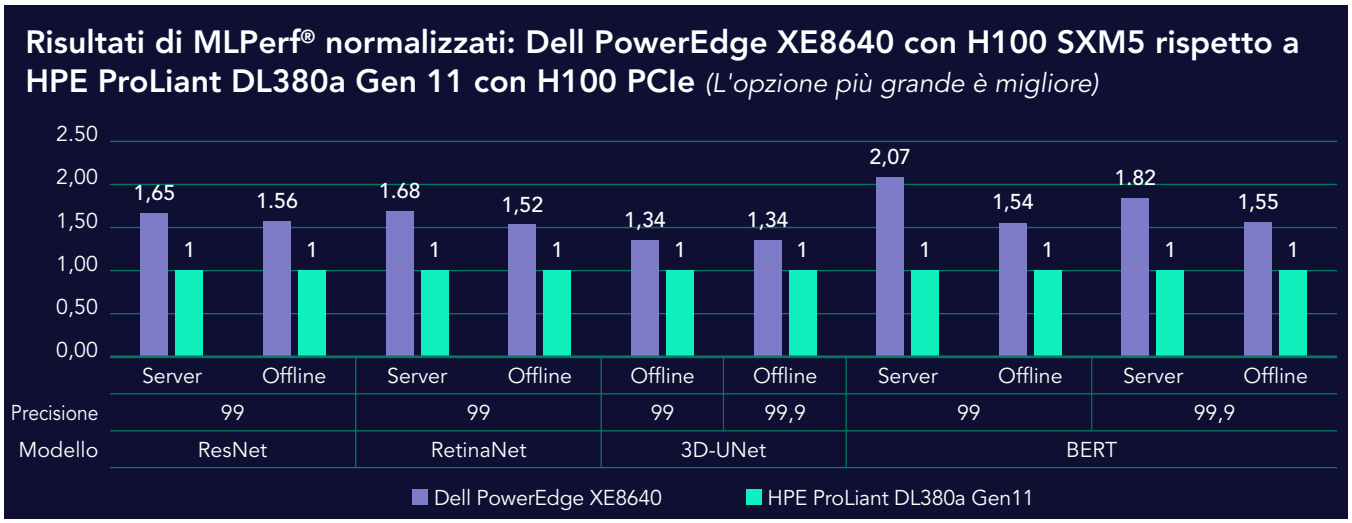


Figura 4. Risultati di MLPerf® pubblicati per Dell PowerEdge XE8640 e HPE ProLiant DL380a Gen11 al 29/11/2023. Il sistema Dell utilizza il fattore di forma NVIDIA H100 SXM, mentre il sistema HPE utilizza il fattore di forma PCIe meno potente. Fonte: Principled Technologies con i dati di MLCommons®.[33,34]

Infine, per le organizzazioni che desiderano iniziare con una soluzione più piccola e aumentarla in base alle esigenze, il server 2U Dell PowerEdge R760xa supporta una gamma di GPU NVIDIA, AMD e Intel, con supporto per un massimo di quattro GPU PCIe Gen 5 double-width o 12 GPU PCIe single-width.[35] È dotato di 32 slot DIMM, un alloggiamento per otto unità per dischi da 2,5 pollici e 12 slot PCIe, che forniscono storage scalabile che può aumentare con l'incremento dei requisiti dei dati AI e supporto per un massimo di 12 GPU PCIe single-width o quattro GPU PCIe double-width come NVIDIA H100 o L40S.[36] Tale scalabilità consente al server di adattarsi alle attività di AI in continua evoluzione, dall'addestramento dei modelli di apprendimento automatico all'elaborazione avanzata dei dati.

Il sistema di raffreddamento ad aria di PowerEdge R760xa supporta ambienti di elaborazione ad alta densità e può ospitare acceleratori TDP (Thermal Design Power) fino a 350 W,[37] una capacità grazie alla quale l'IT può mantenere le prestazioni in presenza di carichi di elaborazione intensivi. Nei risultati dei test di MLPerf® su ResNet, RetinaNet e BERT Server pubblicati a novembre 2023 utilizzando la modalità "server", le prestazioni di PowerEdge R760xa con quattro GPU PCIe NVIDIA H100 hanno superato quelle dell'HPE ProLiant DL380a Gen 11 dotato anch'esso di quattro GPU PCIe H100 (v. la Figura 5).

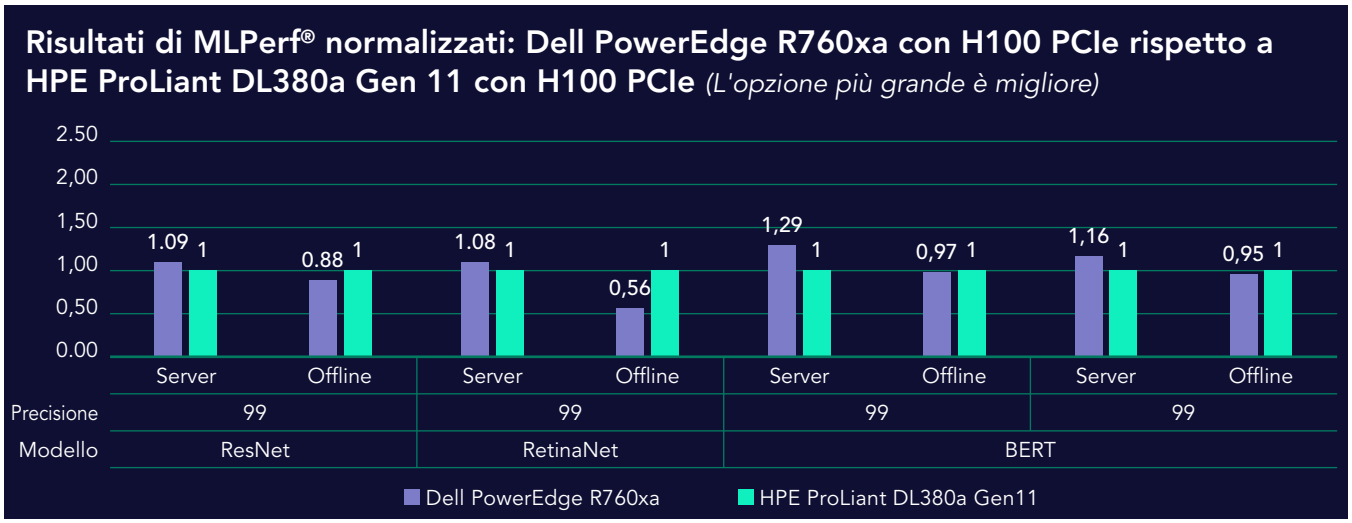


Figura 5. Risultati di MLPerf® pubblicati per Dell PowerEdge R760xa e HPE ProLiant DL380a Gen11 al 29/11/2023. Entrambi i sistemi utilizzano il fattore di forma PCIe delle GPU NVIDIA H100. Fonte: Principled Technologies con i dati di MLCommons®.[38,39]

Nel complesso, i risultati di MLPerf® mostrano che le prestazioni variano ampiamente tra server e componenti, pertanto risulta fondamentale la scelta delle opzioni giuste per supportare i carichi di lavoro e le relative esigenze in termini di prestazioni. I server Dell PowerEdge per i carichi di lavoro di AI offrono diverse opzioni di raffreddamento e densità per soddisfare qualsiasi esigenza di un'azienda in merito ai data center, fornendo al contempo solide prestazioni di MLPerf®.

Copertura più dettagliata del portafoglio AI Dell

Anche se cruciale, le prestazioni di elaborazione sono solo un aspetto da considerare quando si pianificano i carichi di lavoro di AI. Quando si avvia l'implementazione dell'AI, è necessario considerare anche il resto del portafoglio AI offerto da un vendor. Di seguito vengono illustrate altre categorie fondamentali per questi portafogli AI, tra cui workstation client, prodotti nativi per il cloud, storage e altro ancora. Vengono inoltre evidenziate le aree in cui le offerte Dell possono offrire un vantaggio rispetto a HPE.

Workstation

Per gli sviluppatori dell'AI e i data scientist, le workstation Dell Precision Data Science offrono GPU NVIDIA RTX™ e CPU Intel Xeon®, insieme a una suite di strumenti di Data Science.[40] Questi sistemi sfruttano opzioni di elaborazione di livello professionale con GPU NVIDIA certificate per oltre 100 applicazioni professionali[41] e acceleratori di processori scalabili Intel Xeon come Intel DL Boost.[42] Le workstation Precision sono disponibili in formati portatile, tower e rack per soddisfare esigenze che vanno dall'analisi dei dati stazionari di grandi dimensioni alla modellazione scientifica durante gli spostamenti.

Le offerte di workstation HPE sono più ristrette, principalmente con singole tower per workstation dotate di NVIDIA L4s; HPE non offre opzioni per workstation portatili.[43] Pur essendo adeguate per molte attività, le offerte di tower per workstation non offrono la stessa flessibilità e copertura dei carichi di lavoro della gamma più ampia offerta da Dell. La varietà di opzioni in termini di dimensioni e portabilità delle workstation Dell Precision rende possibili soluzioni più personalizzate, in grado di soddisfare esigenze diverse in ambienti quali laboratori, uffici e operazioni sul campo.

Storage

Lo storage può essere fondamentale tanto quanto l'elaborazione per l'esecuzione dei carichi di lavoro di AI. Un numero maggiore di dati migliora la precisione dei modelli AI, ma lo storage e la gestione di enormi dataset possono rappresentare una sfida per le capacità di molti data center. Inoltre, poiché i modelli sono generalmente addestrati utilizzando dati non strutturati, i sistemi di storage predisposti per l'AI devono gestire con facilità molti tipi di dati diversi.[44] Per offrire capacità e dimensionamento per i dataset di AI, ML e DL, Dell offre la serie PowerScale™ per lo storage su file e l'Elastic Cloud Storage (ECS) o ObjectScale software-defined per lo storage a oggetti.

Il portafoglio All-Flash NAS PowerScale offre opzioni di capacità che vanno da 3,84 TB fino a 720 TB di capacità raw per nodo, con capacità All-Flash in cluster che raggiungono i 186 PB di capacità raw. La flessibilità e la scalabilità di PowerScale possono supportare un'ampia varietà di clienti e casi d'uso di AI.[45] Quando in cluster, PowerScale F900 può raggiungere fino a 186 PB di storage raw totale.[46] Tutti e tre i modelli All-Flash PowerScale, F200, F600 e F900, includono la compressione e la deduplica dei dati in linea per migliorare l'efficienza dello storage.[47] Ogni modello di storage PowerScale impiega il file system Dell OneFS™, che utilizza policy per definire il tier dello storage per assegnare la priorità ai dati più importanti sui tier più veloci per l'ottimizzazione dei carichi di lavoro.[48] Dell offre il software OneFS nel Marketplace AWS, insieme a Dell APEX File Storage for AWS. I clienti possono utilizzare OneFS con le istanze di elaborazione AWS per un'esperienza utente coerente con le stesse funzioni disponibili negli array OneFS on-premise.[49] Sebbene HPE offra l'integrazione del public cloud per le soluzioni di storage ibrido, non abbiamo trovato tra le sue offerte un'opzione nativa per il cloud come Dell APEX File Storage for AWS.

Le opzioni di storage a oggetti di Dell includono Dell Enterprise Object Storage (ECS), che è "progettato in maniera specifica per archiviare i dati non strutturati sul public cloud".[50] Insieme alla compatibilità integrata con lo storage a oggetti Amazon S3 per la funzionalità di hybrid cloud, gli storage node ECS offrono capacità fino a 14 PB per rack.[51] HPE offre anche storage non strutturato con opzioni di storage a oggetti e su file, anche se la sua offerta di storage a oggetti avviene tramite una partnership con Scality. I clienti possono acquistare le soluzioni HPE per Scality da HPE.[52]

Professional services

Dell offre una vasta gamma di servizi professionali, tra cui consulenza, preparazione dei dati, deployment, supporto, e servizi gestiti per supportare i deployment dell'AI. Per le organizzazioni che sono alla ricerca di architetture e soluzioni convalidate, Dell offre Validated Design for AI, mirato a casi d'uso specifici per eliminare le incertezze dalla progettazione e dal deployment delle risorse AI. Queste soluzioni di AI convalidate da Dell includono pacchetti hardware e software, modelli di AI conversazionale, operazioni di apprendimento automatico e altro ancora. Dell offre una soluzione completa per tutte le esigenze di AI, combinando soluzioni pre-configurate e appositamente progettate con servizi correlati all'AI. Queste offerte potrebbero fornire un percorso più rapido e semplice verso il successo dell'AI rispetto alla creazione di soluzioni ad hoc.

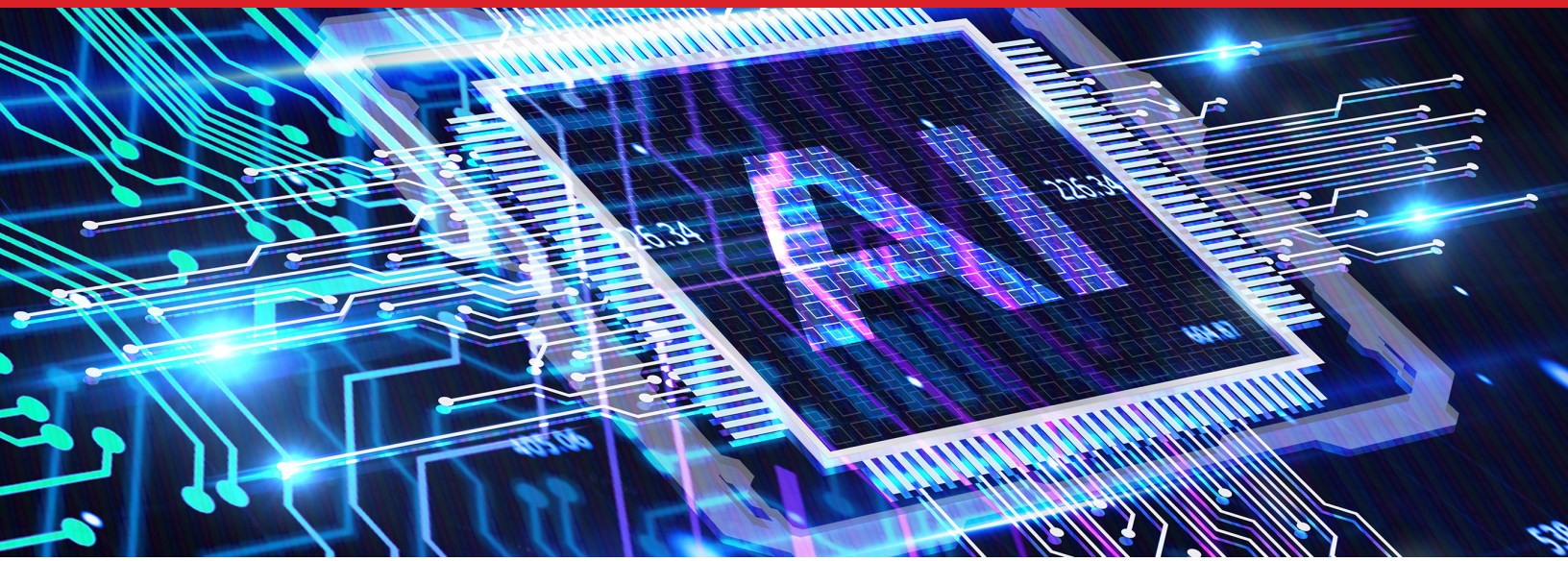
I servizi Dell possono anche guidare il percorso verso il deployment dell'AI, dalla consulenza all'implementazione. Con Dell ProConsult Advisory Services, i clienti possono individuare i vantaggi che gli utenti possono trarre dall'adozione dei processi di GenAI e creare una roadmap che comprenda le soluzioni e le competenze IT richieste. I servizi Dell sono in grado di preparare i dati per l'integrazione di modelli linguistici di grandi dimensioni e formare i team IT sulle conoscenze relative all'AI. Per un'adozione completa della GenAI, i team di Dell esaminano i casi d'uso specifici e determinano, implementano e configurano il modello di AI migliore in base alle esigenze specifiche del cliente. HPE offre servizi professionali per supportare le aziende nelle loro attività relative all'AI.[53,54]

Considerazioni sulla gestione

I server richiedono una gestione continua che sottrae tempo di amministrazione. Firmware, software e driver necessitano di aggiornamenti periodici e il personale IT deve ottimizzare e mantenere prestazioni, temperature e altro ancora. In test di Principled Technologies (PT) precedenti, abbiamo valutato le funzionalità di gestione dei server Dell con Integrated Dell Remote Access Controller 9 (iDRAC9).[55] Gli amministratori possono contare su aggiornamenti online automatizzati grazie a iDRAC9 OpenManage™ Enterprise (OME) con pianificazione configurabile per mantenere aggiornati i server e utilizzare i profili per integrare rapidamente e facilmente nuovi server con l'aumento dei carichi di lavoro. Con iDRAC e OME, i clienti Dell possono accedere a più funzioni di gestione remota, implementare i server con maggiore facilità e aggiornare il firmware più facilmente di quanto potrebbero fare con HPE OneView e HPE iLO. Con i server Dell PowerEdge vengono forniti gestione e servizi Dell che possono assistere le organizzazioni "riducendo il tempo e l'impegno necessari per attività come il monitoraggio dell'integrità del sistema o l'aggiornamento del firmware", in modo che i cicli IT possano essere dedicati all'innovazione e altre attività.[56]

Tabella 3: Riepilogo del confronto tra gli strumenti di gestione Dell e HPE da un report di PT risalente a novembre 2022.[57]
Fonte: Principled Technologies.

	Quali sono le differenze con gli strumenti di gestione Dell	Quanto sono migliori
Più funzioni di gestione remota Confronto tra iDRAC e iLO	Più funzioni di configurazione della console HTML5 e del BIOS per una maggiore funzionalità remota in iDRAC	Un numero di funzioni della console HTML5 2,5 volte superiore e un numero di funzionalità del BIOS 13 volte superiore
Deployment dei server più semplice Confronto tra OME e OneView	Deployment del profilo one-to-many con OME	52% di tempo in meno per implementare un server rispetto a OneView
Aggiornamenti più semplici del firmware Confronto tra OME e OneView	Aggiornamenti online automatizzati con OME	Aggiornamento di più server tramite connessione a Dell.com, risparmiando il tempo richiesto per aggiornare i server caricando manualmente i pacchetti con OneView
Invio degli avvisi più semplice Confronto tra OME e OneView	Configurazione di policy di avviso in OME ed esecuzione di azioni automatizzate basate sugli avvisi	L'automazione di questo processo riduce il tempo necessario e i potenziali errori rispetto all'esecuzione manuale di azioni ogni volta che si riceve un avviso in OneView
Funzioni di sicurezza più semplici da utilizzare (blocco del sistema e porte USB dinamiche) Confronto tra iDRAC e iLO	Meno passaggi, meno tempo, nessun riavvio con iDRAC	¼ dei passaggi, il 91% di tempo in meno per il blocco del sistema
Analisi più solide Confronto tra CloudIQ per PowerEdge e InfoSight	Report personalizzabili, più metriche relative all'integrità per un migliore controllo amministrativo con CloudIQ for PowerEdge	Un numero di metriche oltre 15 volte superiore tra cui scegliere rispetto a InfoSight



Conclusioni

Sfruttare la potenza dell'AI per semplificare e migliorare le operazioni aziendali può rivelarsi un'attività impegnativa, con implicazioni significative per l'impresa. Con il progresso sempre più rapido della tecnologia, la collaborazione con il giusto vendor per l'AI è fondamentale. Scegliendo un'azienda come Dell che non solo offre un portafoglio AI completo, ma può anche fornire servizi di pianificazione, preparazione, implementazione e gestione, i clienti possono affrontare serenamente queste sfide. I test di benchmark di MLPerf® dimostrano che le offerte del portafoglio AI Dell offrono prestazioni coerenti e solide per i carichi di lavoro di AI. Con opzioni di server flessibili e a prestazioni elevate, oltre a diverse opzioni di storage, soluzioni convalidate e servizi professionali progettati in maniera specifica per l'AI, Dell sostiene le aziende nell'adozione dell'AI e dei suoi vantaggi.

---

1. Dell, "Increase Your Data Value with Dell Generative AI Solutions", consultato il 19 dicembre 2023, https://www.dell.com/en-us/blog/increasing-your-data-value-with-dell-generative-ai-solutions/.
2. Dell, "Dell AI Solutions", consultato il 12 dicembre 2023, https://www.dell.com/en-us/dt/solutions/artificial-intelligence/index.htm#accordion0&tab0=0.
3. Dell, "Dell Technologies and Hugging Face to Simplify Generative AI with On-Premises IT", consultato il 12 dicembre 2023, https://www.dell.com/en-us/dt/corporate/newsroom/announcements/detailpage.press-releases~usa~2023~11~20231114-dell-technologies-and-hugging-face-to-simplify-generative-ai-with-on-premises-it.htm#/filter-on/Country:en-us.
4. Dell, "Dell and Meta Collaborate to Drive Generative AI Innovation", consultato il 12 dicembre 2023, https://www.dell.com/en-us/blog/dell-and-meta-collaborate-to-drive-generative-ai-innovation/.
5. Dell, "Dell AI-Ready Data Platform", consultato il 12 dicembre 2023, https://www.dell.com/en-us/dt/solutions/artificial-intelligence/storage-for-ai.htm?hve=explore+unstructured+storage#tab0=0.
6. Dell, "Snowflake and Dell Partnership Gains Momentum", consultato il 19 dicembre 2023, https://www.dell.com/en-us/blog/snowflake-and-dell-partnership-gains-momentum/.
7. Robert McNeal, "Dell, VMware and NVIDIA Bring AI to Your Data", consultato il 17 gennaio 2024, https://www.dell.com/en-us/blog/dell-vmware-and-nvidia-bring-ai-to-your-data/. Come da link riportato sopra: "Dati basati su analisi Dell, agosto 2023. Dell Technologies offre soluzioni progettate per supportare i carichi di lavoro di AI dai PC workstation (portatili e fissi) ai server per High Performance Computing, storage dei dati, infrastruttura software-defined nativa per il cloud, switch di rete, protezione dei dati, HCI e servizi."
8. Intel, "Accelerate Artificial Intelligence (AI) Workloads with Intel Advanced Matrix Extensions (Intel AMX)", consultato il 12 dicembre 2023, https://www.intel.com/content/www/us/en/content-details/785250/accelerate-artificial-intelligence-ai-workloads-with-intel-advanced-matrix-extensions-intel-amx.html.

9. Vipera, "NVIDIA's H100 and A100 GPU Cards: Exploring the Intricacies of SXM and PCI-E Connections", consultato il 12 dicembre 2023, https://www.viperatech.com/unraveling-the-mysteries-sxm-vs-pci-e-connections-in-nvidias-high-end-h100-and-a100-gpus/.
10. GitHub, "MLPerf® Results Messaging Guidelines", consultato il 16 gennaio 2024, https://github.com/mlcommons/policies/blob/master/MLPerf_Results_Messaging_Guidelines.adoc.
11. MLCommons®, "MLPerf® Inference: Datacenter Benchmark Suite Results", consultato il 12 dicembre 2023, https://mlcommons.org/en/inference-datacenter-31/.
12. Dell, "PowerEdge XE Servers", consultato il 12 dicembre 2023, https://www.dell.com/en-us/dt/servers/specialty-servers/poweredge-xe-servers.htm?hve=explore+poweredge+xe#tab0=0.
13. HPE, "HPE ProLiant XL675d Gen10 Plus Configure-to-order Server", consultato il 12 dicembre 2023, https://www.hpe.com/us/en/product-catalog/compute/proliant-servers/pip.1013142988.html.
14. HPE, "HPE ProLiant DL380a Gen11", consultato il 12 dicembre 2023, https://www.hpe.com/us/en/product-catalog/compute/proliant-servers/pip.proliant-dl380-server.1014696168.html.
15. MLCommons®, "MLPerf® Inference: Datacenter Benchmark Suite Results v 3.1", consultato il 12 dicembre 2023, https://mlcommons.org/benchmarks/inference-datacenter/.
16. HPE, "NVIDIA Accelerators for HPE ProLiant Servers", consultato il 12 dicembre 2023, https://www.hpe.com/psnow/doc/c04123180.html?jumpid=in_pdp-psnow-qs.
17. Dell, "PowerEdge XE9680 Specification Sheet", consultato il 19 gennaio 2024, https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe9680-spec-sheet.pdf.
18. HPE, "HPE & NVIDIA financial services solution sets new records in performance", consultato il 12 dicembre 2023, https://community.hpe.com/t5/alliances/hpe-amp-nvidia-financial-services-solution-sets-new-records-in/ba-p/7197388.
19. HPE, "QuickSpecs: HPE Cray Supercomputing XD670", consultato il 12 dicembre 2023, https://www.hpe.com/psnow/doc/a50004292enw.
20. Punteggio MLPerf® verificato di v3.1 Inference Closed. Consultato il 5 dicembre 2023 da https://mlcommons.org/benchmarks/inference-datacenter/, voce 3.1-0069. Il nome e il logo MLPerf® sono marchi registrati e non registrati di MLCommons® Association negli Stati Uniti e in altri Paesi. Tutti i diritti riservati. L'uso non autorizzato è severamente vietato. Consultare www.mlcommons.org per maggiori informazioni.
21. Punteggio MLPerf® verificato di v3.1 Inference Closed. Consultato il 5 dicembre 2023 da https://mlcommons.org/benchmarks/inference-datacenter/, voce 3.1-0085. Il nome e il logo MLPerf® sono marchi registrati e non registrati di MLCommons® Association negli Stati Uniti e in altri Paesi. Tutti i diritti riservati. L'uso non autorizzato è severamente vietato. Consultare www.mlcommons.org per maggiori informazioni.
22. Dell, "PowerEdge XE9640 Rack Server", consultato il 12 dicembre 2023, https://www.dell.com/en-us/shop/ipovw/poweredge-xe9640.
23. Accelsius, "Enabling the AI Revolution with Liquid Cooling", consultato il 12 dicembre 2023, https://www.accelsius.com/blog/enabling-the-ai-revolution-with-liquid-cooling.
24. Dell, "Dell PowerEdge XE9640 Technical Guide", consultato il 12 dicembre 2023, https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe9640-technical-guide.pdf.
25. HPE, "HPE ProLiant DL380a Gen11", consultato il 12 dicembre 2023, https://www.hpe.com/psnow/doc/PSN1014696168WWEN.pdf?jumpid=in_pdp-psnow-dds.
26. Intel, "Intel® Data Center GPU Max Series Technical Overview", consultato il 12 dicembre 2023, https://www.intel.com/content/www/us/en/developer/articles/technical/intel-data-center-gpu-max-series-overview.html#gs.08874l.
27. HPE, "Intel Data Center GPU Max 1100 48GB Accelerator for HPE Data sheet", consultato il 12 dicembre 2023, https://www.hpe.com/psnow/doc/PSN1014779728WWEN.
28. Punteggio MLPerf® verificato di v3.1 Inference Closed. Consultato il 5 dicembre 2023 da https://mlcommons.org/benchmarks/inference-datacenter/, voce 3.1-0066. Il nome e il logo MLPerf® sono marchi registrati e non registrati di MLCommons® Association negli Stati Uniti e in altri Paesi. Tutti i diritti riservati. L'uso non autorizzato è severamente vietato. Consultare www.mlcommons.org per maggiori informazioni.
29. Punteggio MLPerf® verificato di v3.1 Inference Closed. Consultato il 5 dicembre 2023 da https://mlcommons.org/benchmarks/inference-datacenter/, voce 3.1-0084. Il nome e il logo MLPerf® sono marchi registrati e non registrati di MLCommons® Association negli Stati Uniti e in altri Paesi. Tutti i diritti riservati. L'uso non autorizzato è severamente vietato. Consultare www.mlcommons.org per maggiori informazioni.

30. Dell, "AI and HPC —with Air or Liquid Cooling", consultato il 12 dicembre 2023, https://www.delltechnologies.com/asset/en-us/products/servers/briefs-summaries/poweredge-xe9640-and-xe8640-infographic.pdf.
31. Dell, "PowerEdge XE8640: Drive AI, HPC modeling and simulation workloads with superior performance", consultato il 12 dicembre 2023, https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe8640-spec-sheet.pdf.
32. Dell, "PowerEdge XE8640 Rack Server", consultato il 12 dicembre 2023, https://www.dell.com/en-us/shop/ipovw/poweredge-xe8640.
33. Punteggio MLPerf® verificato di v3.1 Inference Closed. Consultato il 5 dicembre 2023 da https://mlcommons.org/benchmarks/inference-datacenter/, voce 3.1-0067. Il nome e il logo MLPerf® sono marchi registrati e non registrati di MLCommons® Association negli Stati Uniti e in altri Paesi. Tutti i diritti riservati. L'uso non autorizzato è severamente vietato. Consultare www.mlcommons.org per maggiori informazioni.
34. Punteggio MLPerf® verificato di v3.1 Inference Closed. Consultato il 5 dicembre 2023 da https://mlcommons.org/benchmarks/inference-datacenter/, voce 3.1-0084. Il nome e il logo MLPerf® sono marchi registrati e non registrati di MLCommons® Association negli Stati Uniti e in altri Paesi. Tutti i diritti riservati. L'uso non autorizzato è severamente vietato. Consultare www.mlcommons.org per maggiori informazioni.
35. Dell, "PowerEdge R760xa Rack Server", consultato il 12 dicembre 2023, https://www.dell.com/en-us/shop/dell-poweredge-servers/poweredge-r760xa-rack-server/spd/poweredge-r760xa/pe_r760xa_16902_vi_vp#features_section.
36. SANStorageWorks, "Dell EMC PowerEdge R760xa: Powerful and scalable for GPU workloads", consultato il 12 dicembre 2023, https://www.sanstorageworks.com/PowerEdge-R760xa.asp.
37. Dell, "Dell PowerEdge Servers and NVIDIA GPUs", consultato il 12 dicembre 2023, https://infohub.delltechnologies.com/l/design-guide-generative-ai-in-the-enterprise-inferencing/dell-poweredge-servers-and-nvidia-gpus-1/.
38. Punteggio MLPerf® verificato di v3.1 Inference Closed. Consultato il 5 dicembre 2023 da https://mlcommons.org/benchmarks/inference-datacenter/, voce 3.1-0064. Il nome e il logo MLPerf® sono marchi registrati e non registrati di MLCommons® Association negli Stati Uniti e in altri Paesi. Tutti i diritti riservati. L'uso non autorizzato è severamente vietato. Consultare www.mlcommons.org per maggiori informazioni.
39. Punteggio MLPerf® verificato di v3.1 Inference Closed. Consultato il 5 dicembre 2023 da https://mlcommons.org/benchmarks/inference-datacenter/, voce 3.1-0084. Il nome e il logo MLPerf® sono marchi registrati e non registrati di MLCommons® Association negli Stati Uniti e in altri Paesi. Tutti i diritti riservati. L'uso non autorizzato è severamente vietato. Consultare www.mlcommons.org per maggiori informazioni.
40. Dell, "Workstations for AI", consultato il 12 dicembre 2023, https://www.dell.com/en-us/dt/ai-technologies/index.htm?hve=explore+dell+precision+for+ai#pdf-overlay=//www.delltechnologies.com/asset/en-us/products/workstations/briefs-summaries/ai-industry-brochure.pdf.
41. NVIDIA, "NVIDIA RTX in Professional Workstations", consultato il 12 dicembre 2023, https://www.nvidia.com/en-us/design-visualization/desktop-graphics/.
42. Intel, "Intel® Deep Learning Boost (Intel® DL Boost)", consultato il 12 dicembre 2023, https://www.intel.com/content/www/us/en/artificial-intelligence/deep-learning-boost.html.
43. HPE, "HPE ProLiant ML350 Gen11", consultato il 12 dicembre 2023, https://buy.hpe.com/us/en/compute/tower-servers/proliant-ml300-servers/proliant-ml350-server/hpe-proliant-ml350-gen11/p/1014696172.
44. ComputerWeekly.com, "Storage requirements for AI, ML and analytics in 2022", consultato il 12 dicembre 2023, https://www.computerweekly.com/feature/Storage-requirements-for-AI-ML-and-analytics-in-2022.
45. Dell, "PowerScale AI-Ready Data Platform", consultato il 12 dicembre 2023, https://www.dell.com/en-us/shop/powerscale-family/sf/powerscale.
46. Dell, "Compare PowerScale", consultato il 12 dicembre 2023, https://www.dell.com/en-us/shop/powerscale-family/sf/powerscale#compare-module.
47. Dell, "Dell PowerScale All-Flash", consultato il 12 dicembre 2023, https://www.delltechnologies.com/asset/en-us/products/storage/technical-support/h15963-ss-powerscale-all-flash-nodes.pdf.
48. Dell, "Dell PowerScale OneFS Software Features", consultato il 12 dicembre 2023, https://www.delltechnologies.com/asset/en-us/products/storage/technical-support/h18275-onefs-software-features-data-sheet.pdf.
49. Dell, "Dell APEX File Storage for AWS", consultato il 12 dicembre 2023, https://www.delltechnologies.com/asset/en-us/products/storage/briefs-summaries/h19575-so-apex-file-storage-for-aws.pdf.

50. Dell, "Dell ECS Enterprise Object Storage", consultato il 12 dicembre 2023, https://www.dell.com/en-us/dt/storage/ecs/index.htm?hve=explore+ecs#tab0=0&tab1=0.
51. Dell, "Dell ECS Enterprise Object Storage", consultato il 12 dicembre 2023, https://www.dell.com/en-us/dt/storage/ecs/index.htm#tab0=0&tab1=0&accordion0.
52. HPE, "Storage Solutions for Scality", consultato il 12 dicembre 2023, https://www.hpe.com/us/en/storage/file-object/scality.html.
53. HPE, "Make AI Work for You", consultato il 16 gennaio 2024, https://www.hpe.com/us/en/solutions/ai-artificial-intelligence.html.
54. HPE, "HPE AI Services – Generative AI Implementation", consultato il 16 gennaio 2024, https://www.hpe.com/us/en/services/generative-ai-implementation-service.html.
55. Principled Technologies, "Simplify administrator tasks and improve security and health monitoring with tools from the Dell management portfolio vs. comparable tools from HPE", consultato il 12 dicembre 2023, https://www.principledtechnologies.com/Dell/Management-tools-vs-HPE-1122.pdf.
56. Principled Technologies, "Simplify administrator tasks and improve security and health monitoring with tools from the Dell management portfolio vs. comparable tools from HPE", consultato il 12 dicembre 2023, https://www.principledtechnologies.com/Dell/Management-tools-vs-HPE-1122.pdf.
57. Principled Technologies, "Simplify administrator tasks and improve security and health monitoring with tools from the Dell management portfolio vs. comparable tools from HPE".

La versione originale in inglese di questo report è disponibile all'indirizzo https://facts.pt/zPmSx4c

Questo progetto è stato commissionato da Dell Technologies.

Principled Technologies®

Facts matter.®