Hewlett Packard Enterprise

Solution brief

Check if the document is available in the language of your choice.

MAXIMUM EFFICIENCY FOR INFERENCING WITH YOUR AI WORKLOADS ON HPE PROLIANT AND NVIDIA GPUS

A POWERFUL INFERENCING SOLUTION FOR ARTIFICIAL INTELLIGENCE

The HPE ProLiant DL380 server is engineered to meet AI data and workload requirements for taking in new information and inferring insights based on trained models. The HPE ProLiant DL380 is a simplified server engineered to run inferencing, unlike the powerhouse required for training, it provides excellent throughput and latency for quick responses. AI applications require powerful processing capabilities outside the reach of standard CPUs; NVIDIA® GPUs provide accelerated compute.

HPE PROLIANT AI INFERENCING BUNDLE WITH NVIDIA GPUS

The HPE ProLiant DL380 server with NVIDIA T4 GPUs is the ideal platform for AI inference, providing unprecedented performance, scalability, and energy efficiency. The T4 GPU, powered by NVIDIA Turing™ Tensor Cores, delivers markedly improved inference performance over predecessors and CPUs.[1] Designed with size and energy efficiency in mind, the NVIDIA T4 GPU is ideal for the scale-out flexibility of the HPE ProLiant DL380, with each allowing up to 7 T4 GPUs per server at a low power consumption. As an NVIDIA GPU Cloud (NGC) ready server, the HPE ProLiant DL380 has passed an extensive suite of tests that validates its ability to deliver high performance running NGC containers.

A BETTER SOLUTION BUILT FOR AI INFERENCING

AI is constantly challenged to keep up with exploding volumes of data yet still deliver fast responses. Not only has AI taken over traditional methods of computing, but it has also changed the way industries perform. From modernizing healthcare and finance streams to research and manufacturing, everything has changed in the blink of an eye. AI is demanding a new breed of performance accelerated machines that can solve highly complex problems quickly, while simplifying IT management and reducing time to insight. Start accelerating your AI workloads today with the HPE ProLiant DL380 server. The following are some key use cases where the HPE ProLiant DL380 system excels.

TABLE 1. Key use cases for the HPE ProLiant DL380 bundle

Facial recognition	**Quality inspection**
Anomaly detection	**Heat mapping**
Object recognition	**Speech recognition**

HPE PROLIANT DL380 SERVER: IDEAL FOR AI INFERENCING

The HPE ProLiant DL380 Gen10 server has an adaptable chassis, including new HPE modular drive bay configuration options with up to 30 small form factor (SFF), up to 19 large form factor (LFF), or up to 20 NVMe drive options along with support for up to three double-wide GPU options.

[1] nvidia.com/en-us/gpu-cloud/

Along with an embedded 4x1GbE, there is a choice of HPE FlexibleLOM or PCIe standup adapters, which offer a choice of networking bandwidth (1GbE to 40GbE) and fabric that adapt and grow to changing business needs. By applying this unique configuration for inferencing, you benefit from:

- HPE's industry-leading server technology that delivers unprecedented performance with reliability, availability, and services (RAS) features
- Selecting a complete AI/ML solution that delivers better price performance
- Access to worldwide services (Advisory, Professional, and HPE GreenLake consumption model) and support
- 3-year parts, labor, and on-site support warranty

FIGURE 1. HPE ProLiant DL380

- Simple system management with optional HPE Performance Cluster Manager
- Essential firmware is anchored by the HPE iLO 5 chip and HPE Silicon Root of Trust to create an immutable fingerprint that verifies the firmware code is valid, so the server won't boot with compromised firmware
- Superior performance per dollar, ease of serviceability, and zoned cooling with GPUs
- Second-generation Intel® Xeon® Scalable processors for highest levels of performance and reliability
- Broad choice of tier-one OS supported
- Added security and smart remote functionality with the HPE iLO 5 Advanced License
- Customization of compute, storage, and memory to fit your unique needs

VALIDATED SOLUTION ELEMENTS FOR AI INFERENCING WITH HPE PROLIANT DL380

- Servers: (1) HPE ProLiant DL380 server
- CPU: (1) Intel® Xeon® Gold 6254 Scalable processors
- Memory: (12) HPE 32GB Dual Rank x4 DDR4-2933 CAS-21-21-21 Registered Smart Memory Kit
- Storage: (6) HPE 400GB SAS 12G Write Intensive SFF (2.5in) SC 3yr Wty Digitally Signed Firmware SSD
- Smart storage controller: (1) HPE Smart Array P408i-a SR Gen10 (8 Internal Lanes/2GB Cache) 12G SAS Modular Controller
- NICs: (1) HPE Ethernet 10/25Gb 2-port 621SFP28 Adapter
- GPUs: (2) NVIDIA T4 GPU with 16 GB Computation Accelerator
- Service and support: 3/3/3—Server Warranty, HPE 3 year Proactive Care Next Business Day Service, worldwide HPE Pointnext Services and HPE iLO Advanced License

DESIGNED TO SCALE

The chassis with up to seven GPUs helps you meet the demanding data read/write requirements on the storage and data management components of AI environments is built to scale. Also, to accommodate your unique requirements, bundles provide the flexibility to configure the CPU, storage, and memory capacity as needed. As your needs grow, you can easily scale these solutions with additional HPE ProLiant server nodes to support your increasing needs.

A POWERFUL SOLUTION SUPPORTED BY GLOBAL SERVICES

HPE Pointnext Services leverages our strength in infrastructure, partner ecosystems, and end-to-end lifecycle experience to accelerate powerful, scalable IT solutions to provide you the assistance for faster time to value. HPE Pointnext Services provides a comprehensive portfolio including Advisory and Transformational, Professional, and Operational Services to help accelerate your digital transformation.

LET'S MAKE IT HAPPEN TOGETHER

The HPE ProLiant DL380 for AI/ML inferencing is easy to configure, deploy, and manage. Get started today by sending questions or inquiries to ai_madeeasy@hpe.com or contact your local HPC/AI specialist to take advantage of these bundles at special pricing.

LEARN MORE AT

hpe.com/us/en/servers/proliant-dl-servers.html

Get updates

Hewlett Packard Enterprise

a50001587ENW, May 2020