HPE GreenLake



# Launching a new age of intelligence with accelerated computing innovation

## Unleash revolutionary HPC/AI with the HPE Apollo 6500 Gen10 Plus system



**The HPE Apollo 6500 Gen10 Plus System is enhanced for high-performance results by enabling:**

- Faster time-to-insight for competitive advantage, allowing for better use of data.
- Superior performance by tightly coupling compute power with industry-leading GPUs for unbeatable job throughput and performance.
- A fully tested and configured HPE solution that is A100 and MI100-ready for demanding HPC, AI, machine learning, and deep learning workloads.
- Comprehensive system security and management to help you work and innovate with confidence.
- Air-cooled version with nodes of HPE Apollo 6500 Gen10 Plus System in HPE Cray Supercomputer architecture—coupled with HPE Slingshot interconnect and supported by HPE Cray OS.

### Escalating hpc and AI workloads

In this dynamic global economy, success begins in the data center. Today's organizations rely on the latest technology developments to unlock the value of their data. The ability to innovate is key to unleashing exceptional performance, achieving greater intelligence, and delivering the outcomes that will take businesses to the next level.

Data centers are changing dramatically as the demand for high-performance computing (HPC) and artificial intelligence (AI) skyrockets across a wide range of industries. HPC and AI needs are fueling major improvements in data processing and computation, and driving ongoing progress in a variety of scientific, industrial, and societal challenges.

HPC and AI workloads continue to escalate in size and complexity, quickly exhausting the capacity of traditional infrastructure. As a result, many organizations have deployed accelerated computing solutions that provide greater power and memory bandwidth to handle their most data-intensive workloads. Accelerated computing supports HPC, AI, and data analytics at scale by enhancing overall speed and performance. These robust platforms make it possible to manage rising data parameters, run complex modeling and simulation applications, and run massive training and inference jobs at breakneck speeds.

### Preparing for the exascale era

AI is evolving at a pace never seen before, and the exascale revolution is fast approaching. The onset of this era will drive a dramatic shift toward data-centric computing in the enterprise space. Exascale is expected to place rigorous demands on IT infrastructure to digest massive troves of data for AI at extreme scale. Highly sophisticated workloads will demand maximum durability, greater bandwidth, and high-speed interconnects in order to avoid data bottlenecks. Organizations are racing to prepare for exascale by developing compatible technologies that will ease digital transformation and enable more efficient, cost-effective solutions.

Accelerated computing is the ideal foundation for AI techniques like machine learning (ML) and deep learning, which are transforming entire industries with unmatched speed, precision, and insight. A number of vertical markets are reaping the benefits of these game-changing advancements, including Healthcare and Life Sciences, Energy, Manufacturing, Government, and Financial Services. Whether organizations utilize AI to run millions of genome sequences to uncover genetic mutations or monitor smart factories in different geographies, AI augments human expertise, dramatically accelerates anomaly detection and discovery.

(AI and human knowledge are different, complimentary. The decisions are made to make capabilities better, not to replace)

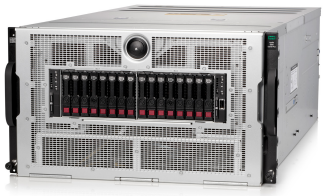

**Figure 1.** HPE Apollo 6500 Gen10 Plus System

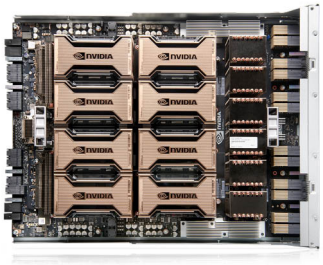

**Figure 2.** NVIDIA A100 Tensor Core GPUs

**Figure 3.** HPE ProLiant XL675d Gen10 plus dual processor server with AMD EPYC 7003 Series processor

**Figure 4.** AMD Instinct MI100 GPU with Infinity Fabric™ Link

[1] NVIDIA Ampere Architecture

[2, 3] NVIDIA A100 TENSOR CORE GPU

[4, 5] nvidianews.nvidia.com/news/nvidia-doubles-down-announces-a100-80gb-gpu-supercharging-worlds-most-powerful-gpu-for-ai-supercomputing

Yet modern AI workloads are already pushing the boundaries of accelerated computing. Now, savvy organizations are looking for cutting-edge solutions to unleash the full power AI, gain competitive advantage, and solve some of the world's biggest problems.

With so much intelligence at stake, the next phase of accelerated computing is essential to harness the expansion of AI. These breakthrough innovations will not only meet existing standards for I/O, security, and manageability, but they will also provide superior processing at scale, with faster, more reliable data communication to optimize demanding workloads.

## A new breed of accelerated computing

Hewlett Packard Enterprise (HPE) is welcoming the coming wave of exascale, with accelerated computing innovations to empower businesses on their AI journeys.

At HPE, our mission is to deploy end-to-end solutions that can tackle any data center workload—from edge to cloud. To achieve this, we are introducing the HPE Apollo 6500 Gen10 Plus System (Figure 1) engineered with performance and density in mind. Backed by the most effective accelerators on the market, HPE systems leverage CPU-GPU heterogeneous compute for a broad range of mission-critical HPC/AI applications. We employ the unparalleled processing capacity of the latest NVIDIA® A100 Tensor Core GPUs (Figure 2), AMD EPYC™ processors (Figure 3) and AMD Instinct™ MI100 accelerator (Figure 4) to eliminate the strain of legacy infrastructure and make way for exascale.

A100 GPUs ensure low latency at high throughput to enhance these powerful accelerated computing solutions. With third-generation tensor cores, the A100 can efficiently scale up to thousands of GPUs, or, with NVIDIA Multi-Instance GPU (MIG) technology, can be divided into seven isolated GPU instances to accelerate diverse workloads. Maximizing GPU utilization drives revolutionary performance gains of up to 20x[1] from Volta to Ampere architecture. The A100 offers up to 6x out-of-the-box performance[2] for training large models, plus up to 7x performance with MIG[3] for inference to drastically reduce time-to-insight.

Each A100 GPU is available in 40 GB or 80 GB configurations, so users can tailor deployments to fit their specific requirements. Businesses can achieve extraordinary performance with standard 40 GB—the original A100 GPU configuration—as well as 80 GB of memory. A100 40 GB (HBM) is a highly cost-efficient option for mainstream AI. A100 80 GB (HBM2e) has doubled the high-bandwidth memory of these powerhouse accelerators while increasing GPU memory bandwidth by 30% redundant.[4] A100 80 GB delivers the world's fastest memory bandwidth at over 2 TB per second, making it the ideal choice for use cases that require large, memory intensive datasets or models.[5]

## High-density compute for data-driven innovation

The HPE Apollo 6500 Gen10 Plus System provides vast computational power, leveraging superior memory bandwidth, throughput, and data communication. Now, businesses can run multiple iterations in less time, quickly deploy AI models into production, and expedite time-to-solution.

The HPE Apollo 6500 Gen10 Plus System is purpose-built to deliver unbeatable value:

- Accelerated performance for the most complex HPC and AI applications
- Flexible to meet your diverse workload and data center requirements
- Customized design for reduced costs, improved reliability, and leading serviceability
- Energy-efficient computing with air cooling and liquid cooling system options
- Comprehensive server security and management

HPE enables peak accelerated computing performance to meet the rising complexities of AI. The A100 is a central feature of the HPE Apollo 6500 Gen10 Plus System, with up to 16 GPUs per server to tackle next-level challenges—from deep recommendation engines to conversational AI. These enterprise systems harness the power, frequency, and processing capacity to rapidly capture, analyze, and operationalize intelligence, whatever your workload requirements.

NVIDIA NVLink establishes a seamless connection between GPUs, so they can work together as a single robust accelerator. NVLink interconnects provide dedicated communication which enables memory to migrate from GPU to GPU. A single A100 supports up to 12 NVLink connections for a total bandwidth of 600 gigabytes per second.

AMD EPYC Series Processors offer tremendous bandwidth and a high per-core performance to continuously feed information to data-hungry GPUs. High-frequency processors integrated with HDR InfiniBand[6] add up to 150 gigabytes per second of bandwidth for every two GPUs, so even businesses operating at the cluster level can communicate at twice the speed.

HPE's first optimized 4 GPU server delivers better price performance than ever for HPC. Also, with a refreshed 8 GPU offering with 2P AMD CPUs, enterprises can harness up to 16 PCIe GPUs.

The HPE Apollo 6500 Gen10 Plus platform also supports the newly announced AMD Instinct MI100 GPU with Infinity Fabric Link. MI100 GPUs are expertly engineered for the next wave of HPC and AI, enhancing accelerated computing so that enterprises can propel world-changing discoveries. Powered by the first AMD Compute DNA architecture (AMD CDNA), MI100 GPUs deliver a giant leap in compute and connectivity, offering[7] nearly 3.5x the performance for HPC (FP32 Matrix) and nearly 7x the performance for AI (FP16) workloads, compared to AMD prior generation accelerators.

In addition, the HPE Apollo 6500 Gen10 Plus System offers extensive storage options, with up to 16 storage drives and choices of SAS, SATA, or NVMe to meet your workload requirements. HPE has plans to roll out new solutions through Q1 2021 which will feature a staggering 16 NVMe drives for almost 8x greater bandwidth than in our Gen10 servers.

## Choosing a trusted partner

HPE is uniquely positioned to help organizations meet today's requirements and evolve for tomorrow's challenges. The HPE Apollo 6500 Gen10 Plus System is disrupting the market with accelerated computing capabilities to unleash the value of integrated HPC and AI.

HPE secures your deployments in firmware protection, malware detection, and firmware recovery down to the silicon. The Silicon Root of Trust from HPE creates a digital fingerprint in the silicon that ensures HPE systems will never boot with compromised firmware. HPE iLO server management software enables customers to securely configure, monitor, and update HPE systems seamlessly, from anywhere in the world, so you can operate with confidence.

HPE provides total flexibility that other vendors on the market cannot, equipping customers with end-to-end solutions that are thoroughly tested, secured, and backed by a variety of financial and professional services.

HPE Pointnext Services provide the support and expertise to accelerate innovation and achieve your desired outcomes. HPE professionals collaborate and work with customers to design and implement technology solutions, optimize processes, smooth skill gaps, and determine the right financial model for your needs.

Key areas of expertise include:

- **Cloud services:** Bring agility and manageability to your technology environment with hybrid cloud.
- **Edge services:** Harness the power of data at the edge to achieve better insights and automation.
- **IT modernization services:** Modernize IT and data centers with automation and container technologies.
- **AI and data-driven services:** Streamline AI adoption, and migrate to unified and secured data platforms with a carefully curated ecosystem of partners.

HPE GreenLake is a consumption-based payment model that aligns cash to actual usage. Cloud services from HPE GreenLake deliver business outcomes faster with an as-a-service model. Now, customers can achieve the cloud experience in just a few clicks, without the cost, risk, and time to move data or refactor applications. Cloud services unify your data, centralize operations, boost operational efficiencies, and free up capital with pay-per-use economics—all within the control of your on-premises environment. HPE GreenLake also provides a support team to help customers create a road map from your needs to your ideal solution.

[6] Introducing 200G HDR InfiniBand Solutions

[7] Introducing AMD Instinct™ MI100 Accelerator—Accelerate Your Discoveries

**Table 1.** HPE Apollo 6500 Gen10 Plus system offering a choice of server trays to meet your most demanding AI and HPC workloads

| Feature | HPE ProLiant XL645d Gen10 Plus<br>Single Processor | HPE ProLiant XL675d Gen10 Plus<br>Dual Processor |
|---|---|---|
| **GPUs** | HGX A100 4 GPU—4 double-wide PCIe or 8 single-wide PCIe<br>With AMD CPUs and AMD Infinity Fabric links<br>AMD Instinct MI100 4 GPUs—4 double-wide PCIe GPUs in 4 GPU Hives with direct P2P connectivity<br>All on Gen4 per node, two nodes per chassis | HGX A100 8 GPU—Up to 8, 10 double-wide PCIe or up to 16 8 single-wide PCIe<br>With AMD CPUs and AMD Infinity Fabric links<br>AMD Instinct MI100 8 GPUs—8 double-wide PCIe GPUs in 4 GPU Hives with direct P2P connectivity through AMD Infinity Fabric link<br>All on Gen4 per node |
| **Computer** | 3rd Gen AMD EPYC 7003 series processors, 3 GHz, up to 280W, 64 Cores top bin<br>Half width system board—1 processor 8 DIMMs | 3rd Gen AMD EPYC 7003 series processors, 3 GHz, up to 280W, 64 Cores top bin |
| **Memory** | Up to 8 3200 MT/s DDR4 SmartMemory | Up to 32 3200 MT/s DDR4 SmartMemory |
| **Storage** | HPE Smart Array S100i Software RAID (now with up to 2 drive NVMe RAID 0/1)<br>HPE Smart Array including new Tri-mode NVMe RAID controllers | Broad selection of HPE Smart Array SR Gen10 Controllers |
| **Drives** | 8 drives per node—HDD, SSD, or 3 Max NVMe | 16 total drives—HDD, SSD, or 6 Max NVMe |
| **PCIe/I/O slots** | Up to 2 LP + OCP Mezz + M.2 | Up to 6 LP + Smart Array |
| **Networking** | Up to 3 Ethernet, InfiniBand, or Slingshot high speed adapters | Up to 6 Ethernet, InfiniBand, or Slingshot high speed adapters |
| **Power supplies** | Integrated power supplies to provide fully redundant power and cooling | Integrated power supplies to provide fully redundant power and cooling |

## Resources

HPE Apollo Systems

Explore HPE Pointnext Services

HPE GreenLake

## Conclusion

HPE is bringing the power of accelerated computing to every organization. Across the market, HPE offers the deepest set of solutions from data center, to edge, to cloud. Our comprehensive Gen10 portfolio is tailored for HPC/AI on an increasing scale, and we are continuously working to collaborate, build, validate, and deliver leading-edge technologies and services to suit your workloads and economic requirements. HPE's holistic approach provides best-in-class technologies, an extensive partner ecosystem, management services, and support from experts around the globe to help you succeed. We are committed to being the long-term partner that customers trust to make innovation fast and simple.

Whatever your goals for HPC and AI, HPE can help. To learn more about accelerated computing, visit us online today.

## Learn more at

hpe.com/apollo

**Make the right purchase decision. Contact our presales specialists.**

**Chat now (sales)**

**Call now**

**Get updates**

Explore **HPE GreenLake**

Hewlett Packard Enterprise



a50003152ENW, Rev. 3