HPE GreenLake

Launching a new age of intelligence with accelerated computing innovation

Enhance HPC/AI performance 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 tight coupling compute power with industry-leading GPUs for unbeatable job throughput and performance

- A fully tested and configured HPE solution that is NVIDIA® A100 Tensor Core GPU and AMD Instinct MI100 Accelerator–ready for demanding HPC, AI, ML, and DL workloads

- Comprehensive system security and management to help you confidently work and innovate

- 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 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 enable managing rising data parameters, running complex modeling and simulation applications, and running massive training and inference jobs at breakneck speeds.

Figure 1. HPE Apollo 6500 Gen10 Plus System

Figure 2. HPE XL645d Gen10 Plus Modular Accelerator Tray single processors server with AMD EPYC 7002 or 7003 Series processor

Figure 3. NVIDIA A100 Tensor Core GPUs

Figure 4. AMD Instinct MI100 PCIe accelerator

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 an extreme scale. Highly sophisticated workloads will demand maximum durability, greater bandwidth, and high-speed interconnects 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 such as machine learning (ML) and deep learning (DL), 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 vastly exceeds human knowledge and capabilities.

Yet modern AI workloads are already pushing the boundaries of accelerated computing. Savvy organizations are looking for cutting-edge solutions to unleash the full power of AI, gain a 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 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 3), AMD EPYC™ processors (Figure 2) and AMD Instinct™ MI100 accelerator (Figure 4) to help eliminate the strain of legacy infrastructure and make way for exascale.

New A100 GPUs help 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. Helping maximize GPU utilization drives revolutionary performance gains of up to 20x from Volta to Ampere architecture.

The A100 offers up to 6x out-of-the-box performance for training large models, plus up to 7x performance with MIG 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 (High Bandwidth Memory [HBM]) is a highly cost-efficient option for mainstream AI. A100 80 GB (HBM2) has doubled the high-bandwidth memory of these powerhouse accelerators.[1] A100 80 GB delivers bandwidth at over 2 TB per second, making it the ideal choice for use cases that require large, memory-intensive datasets or models.[2]

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 the 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 requires.

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

[1, 2] nvidianews.nvidia.com/news/nvidia- doubles-down-announces-a100-80gb- gpu-supercharging-worlds-most- powerful-gpu-for-ai-supercomputing

AMD EPYC Series Processors offer tremendous bandwidth and high per-core performance to continuously feed information to data-hungry GPUs. High-frequency processors integrated with HDR InfiniBand 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 and 2P AMD, enterprises can harness up to 16 PCIe GPUs.

The HPE Apollo 6500 Gen10 Plus System 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 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.

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 revolutionizing the industry with its accelerated computing capabilities, unleashing the potential of HPC and AI for the future HPE secures your deployments in firmware protection, malware detection, and firmware recovery down to the silicon. The silicon root of trust creates a digital fingerprint in the silicon that helps ensure HPE systems do not boot with compromised firmware.

HPE iLO server management software enables you 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 you with end-to-end solutions that are thoroughly tested, secured, and backed by a variety of financial and professional services.

HPE Services provide the support and expertise to accelerate innovation and achieve your desired outcomes. HPE professionals collaborate and work with you 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 operating costs to actual usage. Cloud services from HPE GreenLake deliver business outcomes faster with an as-a-service model. Now, you 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—everything within the control of your on-premises environment. HPE GreenLake also provides a support team to help you create a road map from your needs to your ideal solution.

* May be subject to minimums or reserve capacity may apply

Table 1. HPE Apollo 6500 Gen10 Plus System with the HPE XL645d Gen10 Plus Modular Accelerator Tray single processor

Feature	HPE XL645d Gen10 Plus Modular Accelerator Tray single processor
GPUs	HGX A100 4 GPU—4 double-wide PCIe or 8 single-wide PCIe with AMD CPUs and AMD Infinity Fabric links AMD Instinct MI100 4 GPUs—4 double-wide PCIe GPUs in 4 GPU Hives with direct P2P connectivity All on Gen4 per node, two nodes per chassis
Computer	2nd and 3rd Gen AMD EPYC series processors
Memory	Up to 8 3200 MT/s DDR4 SmartMemory
Storage	HPE Smart Array S100i Software RAID (now with up to 2 drive NVMe RAID 0/1) HPE Smart Array including new Tri-mode NVMe RAID controllers
Drives	8 drives per node—HDD, SSD, or maximum 3 NVMe
PCIe / I/O slots	Up to 2 LP + OCP Mezz + M.2
Networking	Up to 3 Ethernet, InfiniBand, or HPE Slingshot high-speed adapters
Power supplies	Integrated power supplies to provide fully redundant power and cooling

Resources

hpe.com/us/en/compute/hpc/apollo-systems.html

hpe.com/us/en/services/overview.html

hpe.com/us/en/greenlake.html

Conclusion

HPE is bringing the power of accelerated computing to every organization. Across the market, HPE offers a deep set of solutions from data center, to edge, and 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 you 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

buy.HPE.com/us/en/compute/Apollo-systems/Apollo-6500-System/Apollo-6500-System/HPE-Apollo-6500-Gen10-Plus-System/p/1013092236

Visit **HPE GreenLake**

Chat now (sales)

Hewlett Packard Enterprise