Hewlett Packard Enterprise

Reference Architecture

HPE Reference Configuration for enabling GPU-as-a-Service for Enterprise AI deployments with BlueData

Providing a cloudlike experience for GPU infrastructure using containers for on-premises deployment

Contents

Executive summary 3
Introduction 3
Solution overview 7
Solution components 7
BlueData EPIC software platform 7
HPE Elastic Platform for Analytics (EPA) 9
HPE EPA system configuration for BlueData EPIC software 11
HPE EPA with BlueData EPIC deployments 12
Summary 15
Resources and additional links 16

Executive summary

GPU-as-a-Service (GPUaaS) is a solution for on-premises GPU deployments powered by HPE infrastructure and BlueData software. It provides the ability for enterprise IT organizations to offer a cloudlike experience for their on-premises GPU infrastructure using containers.

This new Hewlett Packard Enterprise solution enables on-demand and elastic provisioning for GPU-accelerated applications while sharing and allocating GPU infrastructure resources across multiple applications. With GPUaaS from Hewlett Packard Enterprise, our customers can increase business agility, optimize GPU utilization, and reduce overall total cost of ownership (TCO) for their GPU infrastructure.

Hewlett Packard Enterprise recently announced the integration of HPE infrastructure and BlueData[1] software; this is one of the use cases for this powerful new combination of HPE hardware and software.

Target audience: This paper is intended to assist Heads of IT to help and simplify the deployment of their GPUaaS and improve their utilization; Chief Data Officers to accelerate their machine learning projects by allowing their team greater flexibility to innovate; and Data Scientists to spin up their GPU environments in minutes and start exploring their data and building models.

Document purpose: This Reference Configuration provides an overview of the deployment of the HPE GPUaaS solution. In addition to outlining the opportunity and key solution components, this Reference Configuration provides guidelines for configuring and deploying the GPUaaS solution.

Introduction

The ability to leverage data in today's computationally intensive business environment is essential for a business's success. As Artificial Intelligence (AI) adoption in the enterprise grows, Hewlett Packard Enterprise delivers the compute and storage power needed to meet the challenges posed by machine learning (ML), deep learning (DL), and advanced data analytics. Now Hewlett Packard Enterprise offers a new GPUaaS solution for on-premises deployments.

GPU-accelerated workloads for enterprise AI Deployments

Success with ML/DL requires a lot of experimentation. Data scientists typically explore a wide variety of data sources, experimenting and iterating with multiple different ML/DL frameworks before they find the best-fit model for the application.

The development of machine learning and deep learning predictive models is also compute-intensive; the use of accelerators such as Graphics Processing Units (GPUs) provides a performance boost that significantly speeds up development. As a result, GPUs are a common infrastructure choice for ML/DL applications.

However, in most enterprises today IT teams find it challenging to meet the growing demand for GPUs from multiple data science teams for multiple different ML/DL applications and use cases.

Standing up the right software components together with the underlying infrastructure for ML/DL applications is a very complex and time-consuming process. It can take several days or often weeks to provision and deploy a new GPU-enabled server for each application; and this process has to be repeated each time a new ML/DL application is requested. It's a major impediment to the rapid experimentation and business agility that data science teams require; these teams are often waiting in queues for access to GPU resources.

Furthermore, once the application has been deployed on a GPU-enabled server, IT has very little visibility into GPU utilization. If the GPUs are underutilized, they can't reallocate or reassign those GPU resources to a different application. This lack of visibility also makes it difficult to implement showback or chargeback models for on-premises deployments.

There are public cloud services that offer the ability to deploy virtualized GPU resources on-demand (i.e., GPU-as-a-Service), but the public cloud isn't a panacea. Many organizations have workload requirements that require on-premises deployments – due to considerations involving security, performance, and data gravity.

[1] Refer to https://www.hpe.com/us/en/newsroom/press-release/2019/05/hewlett-packard-enterprise-integrates-bluedata-to-accelerate-ai-and-data-driven-innovation-in-the-enterprise.html to learn about how Hewlett Packard Enterprise integrates BlueData to accelerate Artificial Intelligence and data-driven innovation in the enterprise.

On-demand and elastic provisioning of GPU resources

Now, there is a GPUaaS solution that combines best-in-class infrastructure from Hewlett Packard Enterprise and software from BlueData, recently acquired by Hewlett Packard Enterprise, together with Hewlett Packard Enterprise professional services to ensure a successful deployment. This new Hewlett Packard Enterprise solution enables enterprise IT organizations to deliver GPUaaS in an on-premises deployment to increase business agility, optimize GPU utilization, and reduce overall TCO for GPUs. Using the container-based BlueData software platform, GPUs from multiple heterogeneous servers can be consolidated and shared across multiple applications—for on-demand and elastic provisioning of containerized GPU resources, with just a few mouse clicks. To enable GPUaaS, BlueData software can be deployed with GPU-enabled servers including HPE Apollo and HPE ProLiant servers with NVIDIA® Tesla or Quadro® GPUs. Furthermore, using BlueData's unique ability to pause containers (where GPU, CPU, and memory resources are released while the overall application state is persisted), data science teams can run multiple different ML/DL applications on shared GPU infrastructure without recreating or reinstalling their applications and libraries.

On-premises infrastructure for large-scale distributed analytics is evolving away from the traditional model of a single dedicated bare metal cluster for analytics using direct-attached storage, serving many different users and use cases. Analytics and processing engines have evolved from only MapReduce to a very broad set of options, now including advanced streaming engines such as Spark, Flink and Storm. More recently, tools such as TensorFlow™, H2O, Caffe2, Keras, PyTorch, and others are being adopted for a wide range of AI/ML use cases in the enterprise. The overall AI/ML ecosystem offers an almost unlimited palette of tools to choose from – depending on your specific use case, processing requirements (from batch to real-time), the preferences of your data science teams, and the existing systems that need to be integrated. The richness and complexity of the AI/ML continues to increase as new frameworks, new versions, and new innovations are constantly being introduced.

As companies grow their on-premises implementations for these various tools, they often find themselves deploying multiple clusters to support the diverse and growing needs of their data scientists, analysts, and other business users. This could be to evaluate various commercial and open source AI/ML tools such as DataRobot, Dataiku, H2O, and TensorFlow, etc., to support different analytics environments such as Cloudera, Spark, Kafka, NiFi, NoSQL databases, etc., to support workload partitioning for departmental or line of business requirements, or simply as a byproduct of multi-generational hardware. Enterprises typically deploy multiple dedicated bare-metal servers for each of these clusters.

This traditional approach can be complex, expensive, and time-consuming and often leads to isolated data ponds and cluster sprawl. It often takes several weeks to provision the systems and infrastructure required for each new cluster and each new environment for various use cases, data science teams, and projects across the business. There are typically multiple clusters containing mostly the same data. There is often massive data duplication and the need to copy large amounts of data between systems. It's extremely inefficient and leads to additional cost over time.

This approach leads to management challenges on multiple fronts – data management and governance, infrastructure deployment, software maintenance, and lastly, the availability of scarce resources with the skills necessary to manage these technologies.

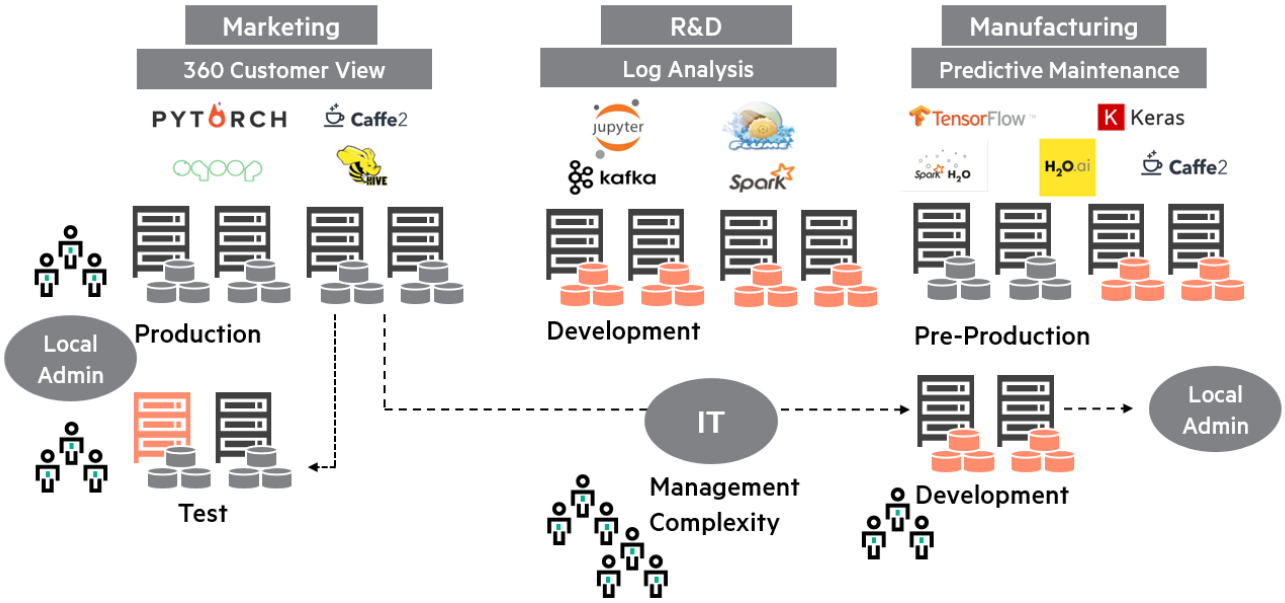

Figure 1. The complexity and challenges of AI/ML deployments

Many enterprise customers are searching for a way to recapture some of the traditional benefits of shared infrastructure such as the ability to easily share data between different applications running on different platforms, the ability to scale compute and storage separately, and the ability to rapidly provision new compute clusters without repartitioning data to achieve optimal performance. While the public cloud is attractive for some AI/ML workloads, in many cases organization's need to retain their data and AI/ML models on-premises due to security, regulatory, and data gravity considerations.

To address these needs, Hewlett Packard Enterprise provides cost-effective and flexible on-premises infrastructure to optimize compute and storage resources in response to these ever-changing requirements in the evolving AI/ML ecosystem. BlueData EPIC provides a software platform to run AI/ML workloads such as TensorFlow and H2O on Docker containers – enabling an on-premises deployment model. By leveraging the rapidly changing technology advances that have occurred since the inception of the original Hadoop architecture in 2005, both Hewlett Packard Enterprise and BlueData have challenged the traditional architectural assumptions that rely on having compute and data elements co-located in the same server.

By combining BlueData's software innovations with HPE server and Storage infrastructure, as well as HPE Pointnext services, enterprise AI/ML and advanced analytics deployments that may have taken months can now be completed within a few days. Leveraging BlueData software and the power of Docker containers, data scientists and analysts can create their own workload defined clusters, on-demand, within minutes. They can spin up clusters for their AI/ML and analytic tools of choice, with the ability to access common pools of data stored on local or remote systems. They can easily try out new versions, new applications, and new AI/ML tools – without waiting for additional infrastructure. All of these have a complete visibility and control, in a multi-tenant environment with secure data isolation.

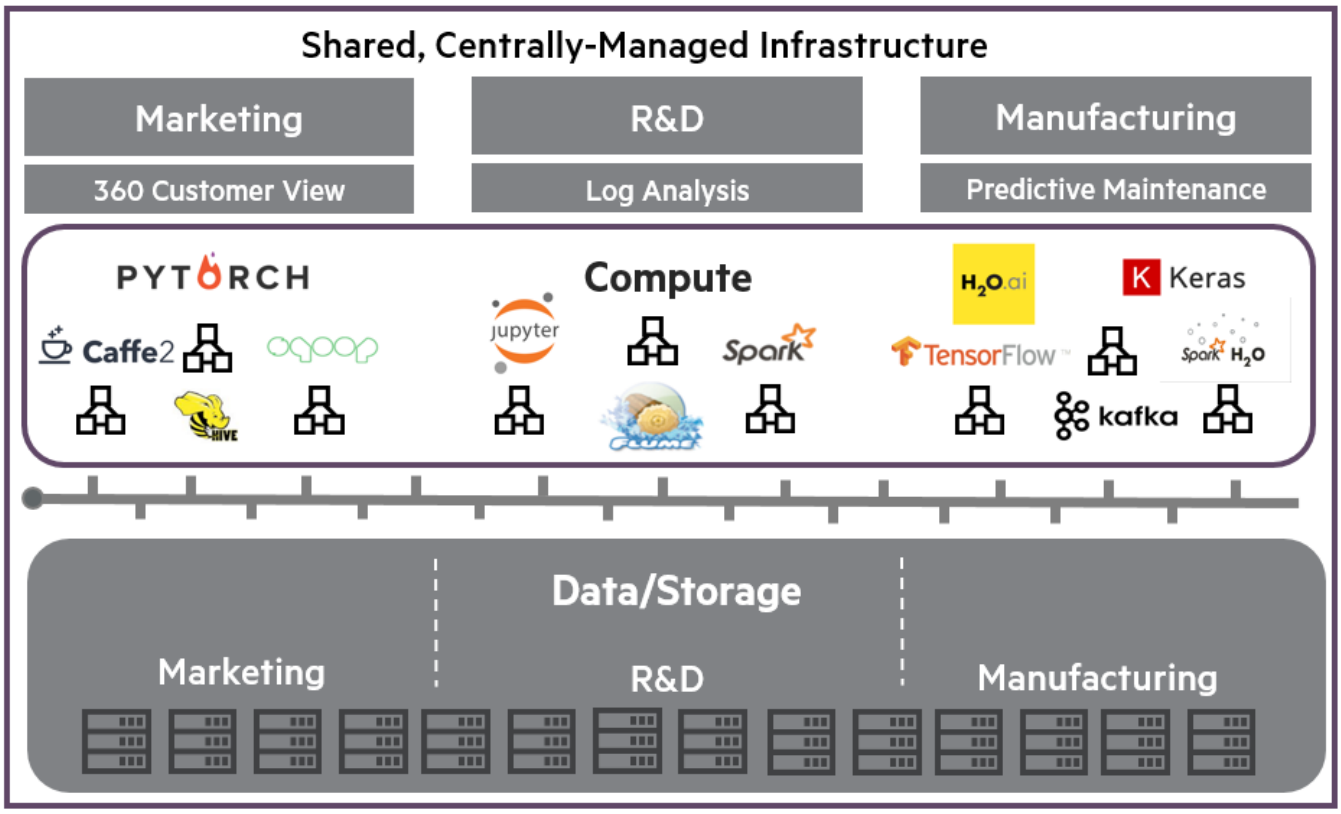

Figure 2. A new multi-tenant deployment model for AI/ML using shared infrastructure

The result is a more flexible, agile, and cost-effective approach to AI/ML models and advanced analytics. Enterprise customers can leverage all the benefits of the cloud operating model (self-service, agility, and elasticity) for AI/ML, Spark and other Analytic workloads – while keeping their data on-premises or in the public cloud. Key benefits include:

- Simplify on-premises deployments for analytics serving up AI/ML models, while also supporting advanced analytics and distributed data processing frameworks such as Spark, Kafka, and a wide range of AI/ML frameworks.
- Increase business agility by empowering data scientists and analysts to quickly create AI/ML environments running in Docker containers, in a matter of minutes with just a few mouse clicks.
- Deliver faster time-to-insights with the ability to rapidly deploy new Docker images for a wide range of different business intelligence, analytics, visualization, and data preparation tools.
- Minimize the need to copy data by separating compute and storage – enabling AI/ML models and analytics on data stored in NFS, HDFS and other storage systems.

- Maintain security and control in a multi-tenant environment, integrated with enterprise-class security models (e.g., LDAP, Active Directory, and Kerberos).
- Lower cost of operation by improving hardware utilization, eliminating cluster sprawl, and minimizing data duplication.

Solution overview

BlueData's mission is to streamline and simplify AI/ML and advanced analytics deployments, eliminating complexity as a barrier to adoption. The BlueData EPIC software platform reduces complexity, accelerates provisioning and reduces the cost for enterprises to deploy AI/ML infrastructure and applications. When combined with the HPE's Elastic Platform for analytics architecture, BlueData EPIC™ provides the flexibility and agility to support the rapidly changing requirements and uses cases for AI/ML and advanced analytics.

The BlueData EPIC software platform can be used to deploy distributed AI/Machine Learning/Deep Learning (AI/ML/DL) environments such as TensorFlow, Caffe2, H2O, BigDL, and SparkMLlib. This allows organizations embarking on AI initiatives to quickly spin up multi-node ML/DL sandbox environments for their data science teams. It leverages the power of Docker containers, while maintaining near bare metal performance. It can run with any shared storage environment, so enterprises don't have to move their data. It delivers the high-grade security and governance that enterprise IT teams require for a multi-tenant deployment.

HPE's Elastic Platform for Analytics harnesses the power of faster Ethernet networks and the HPE Apollo and ProLiant density optimized servers, providing a scalable and elastic multi-tenant architecture that assists in delivering a self-service experience when combined with BlueData's EPIC software platform. The HPE Elastic Platform for Analytics provides modular building blocks for compute, storage and networking that can be combined to build a density optimized platform. HPE EPA also provides "accelerator" building blocks for optimizing workload performance, storage efficiency, or accelerating deployment. Additionally, Hewlett Packard Enterprise provides baseline Reference Architectures (e.g., for Cloudera, Hortonworks, and MapR) as well as use case-based Reference Architectures (e.g., based on workload or application) to help customers understand how to build solutions with the EPA building blocks[2].

With BlueData EPIC and the HPE Elastic Platform for Analytics, enterprises can now derive value from AI/ML within days instead of months and with significantly lower Total Cost of Ownership (TCO) compared to traditional approaches. They can deliver the self-service agility, elasticity, and flexibility in an on-premises deployment model.

Solution components

BlueData EPIC software platform

The BlueData EPIC software platform leverages Docker container technology and patented innovations to deliver self-service, speed, and efficiency for AI/ML and analytics environments:

- ElasticPlane enables users to spin up virtual clusters on-demand with a sophisticated policy engine for resource quotas and SLA management in a secure, multi-tenant environment.
- IOBoost ensures performance on par with bare metal, with the agility and simplicity of Docker containers.

[2] Please visit the HPE RA library for a complete listing of HPE References Architectures for Hadoop.

- DataTap accelerates time-to-value for AI/ML and analytics by eliminating time-consuming data movement.

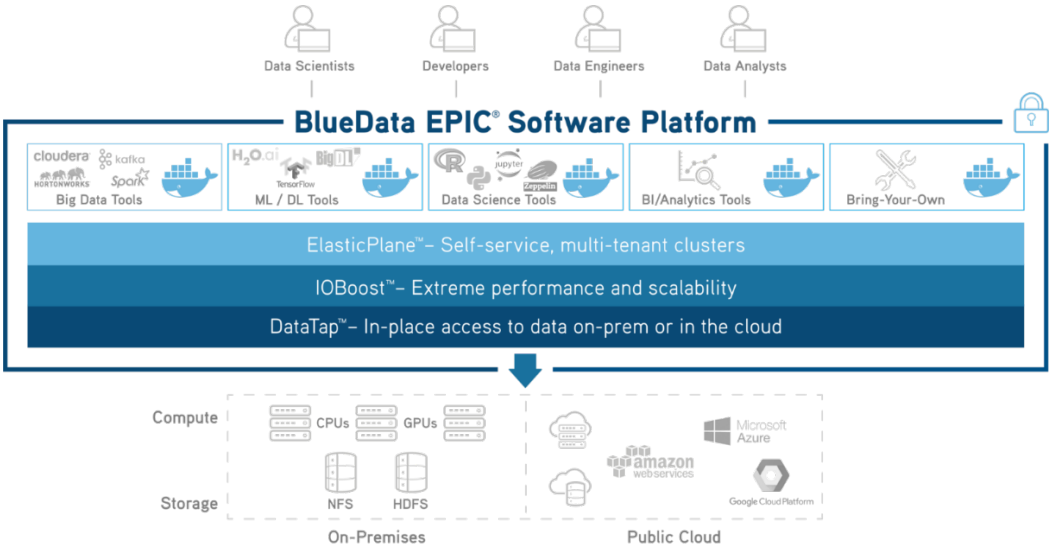

Figure 3. The BlueData EPIC software platform

BlueData installs as a software layer between the underlying server infrastructure and the AI/ML libraries and applications. The use of Docker is completely transparent, and BlueData customers benefit from greater agility and performance due to the lightweight nature of containers. They can leverage the flexibility of Docker to simplify development for AI/ML applications, and the portability of containers to support both on-premises and public cloud deployments.

The key features of the GPUaaS solution include:

- **On-demand, elastic provisioning of GPU resources:** With BlueData, applications can be quickly and easily deployed with access to one or more GPUs. New containerized environments with GPUs can be provisioned on-demand and then de-provisioned (releasing the GPUs) when no longer needed.
- **Pause applications and reassign GPUs:** BlueData provides the ability to pause an application and release the attached GPUs while preserving the current state of the application. This allows IT administrators to monitor usage and reassign the GPU when the GPU-specific code has executed.
- **Out-of-the-box GPU-enabled application images:** BlueData comes with pre-integrated container images for common GPU-enabled applications and ML/DL tools such as TensorFlow, H2O, Caffe2, and JupyterHub – as well as utility images for Ubuntu and CentOS – including NVIDIA CUDA drivers. IT can quickly upgrade these images to new versions and add new tools as needed.
- **Enterprise-grade security and multi-tenancy:** BlueData provides multi-tenancy and data isolation between multiple users and project teams that share the same infrastructure including GPUs. This includes integration with security and authentication such as LDAP, Active Directory, and Kerberos.
- **Unified management console for tracking GPU utilization:** A graphical UI for administrators provide the ability to monitor and manage a shared pool of GPU resources, with complete visibility and usage reporting for GPU utilization across multiple servers and multiple user groups.
- **Bare-metal performance with containers:** BlueData has developed patented innovations to deliver the agility and efficiency benefits of containerization for ML/DL workloads, while ensuring performance comparable to that of bare-metal deployments.
- External storage connectivity and data access control: BlueData's ability to separate compute from data storage eliminates the need to copy or move data. Sensitive data can stay in your secure storage system with enterprise-grade data governance, without the cost and risks of creating and maintaining multiple copies or moving large-scale data.

The key benefits of the GPUaaS solution include:

- **Simplify deployments:** Provide rightsized GPU environments for every workload and give your data scientists the right number of GPUs they need with just a few mouse clicks.
- **Accelerate ML/DL development:** Provision and deprovision GPU resources, within minutes (instead of days). Enable rapid prototyping and increased experimentation with pre-integrated container images for common ML/DL applications, data science tools, and data frameworks.
- **Maintain security and control:** Integrate with your enterprise's security and authentication systems to provide built-in governance and fine-grained access controls for GPUs and other resources.
- **Reduce costs:** Achieve cost savings by improving GPU utilization, controlling usage, eliminating cluster sprawl, and minimizing data duplication.

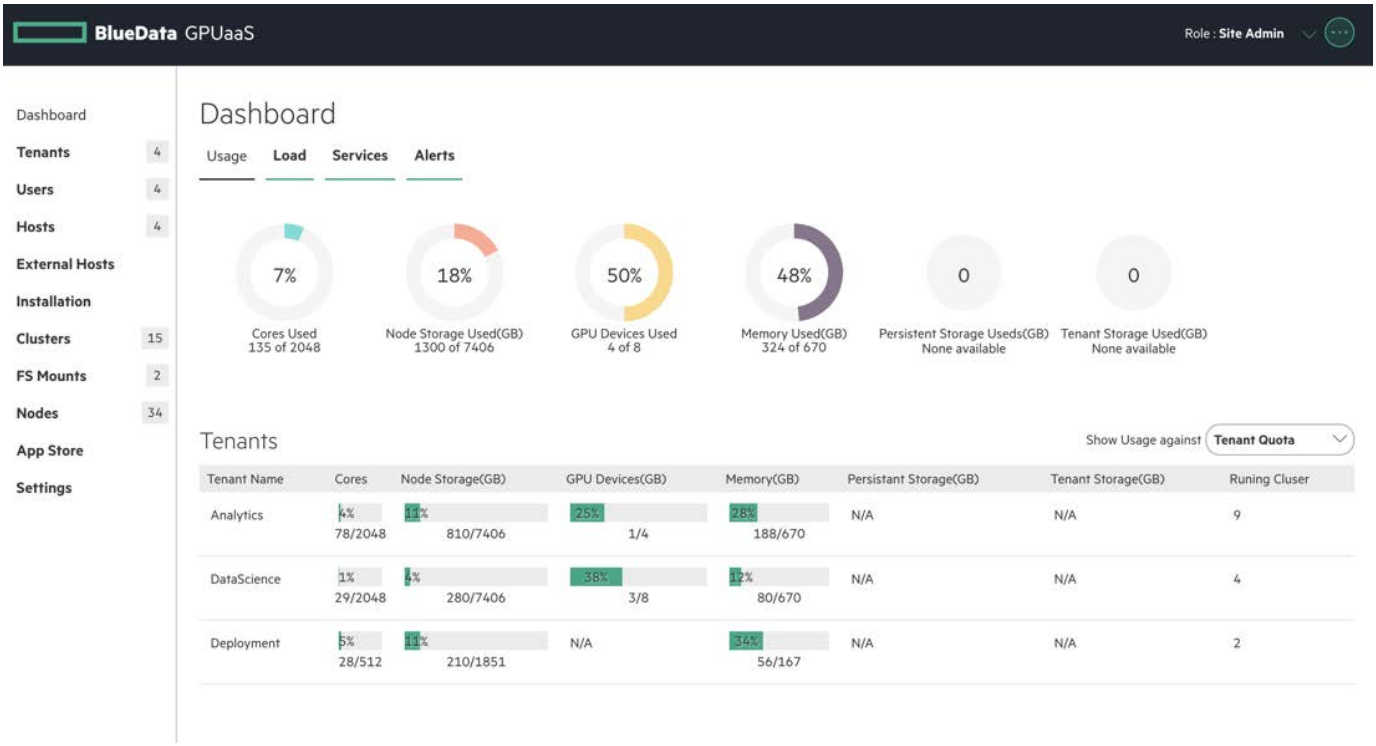

Figure 4. Web-based interface of the BlueData dashboard showing consolidation of GPU, CPU, memory and Storage resources from underlying infrastructure

HPE Elastic Platform for Analytics (EPA)

The HPE Elastic Platform for Analytics is a premier modular infrastructure foundation to accelerate business insights, enabling organizations to rapidly deploy, efficiently scale and securely manage the explosive growth in volume, speed and variety of AI/ML Big Data workloads. EPA harnesses the power of faster Ethernet networks which enables a building block approach to independently scale compute and storage and lets you consolidate your data and workloads growing at different rates. The base HPE EPA system uses the HPE Apollo 4200 as a storage block and the HPE Apollo 2600 with the HPE ProLiant XL190r Gen10 server, along with the Apollo 6500, and the HPE ProLiant DL380 as a compute blocks offering GPU acceleration. By leveraging a building block approach, customers can simplify the underlying infrastructure needed to address a myriad of different business initiatives around Data Warehouse modernization, analytics and BI, and building large-scale data lakes with diverse sets of data. As the workloads and data storage requirements change (often uncorrelated to each other) the HPE EPA system allows customers to easily scale by adding compute and storage blocks independently from each other, maximizing the infrastructure requirements for the workload demands.

Figure 5 below highlights the different building blocks that are part of the HPE Elastic Architecture.

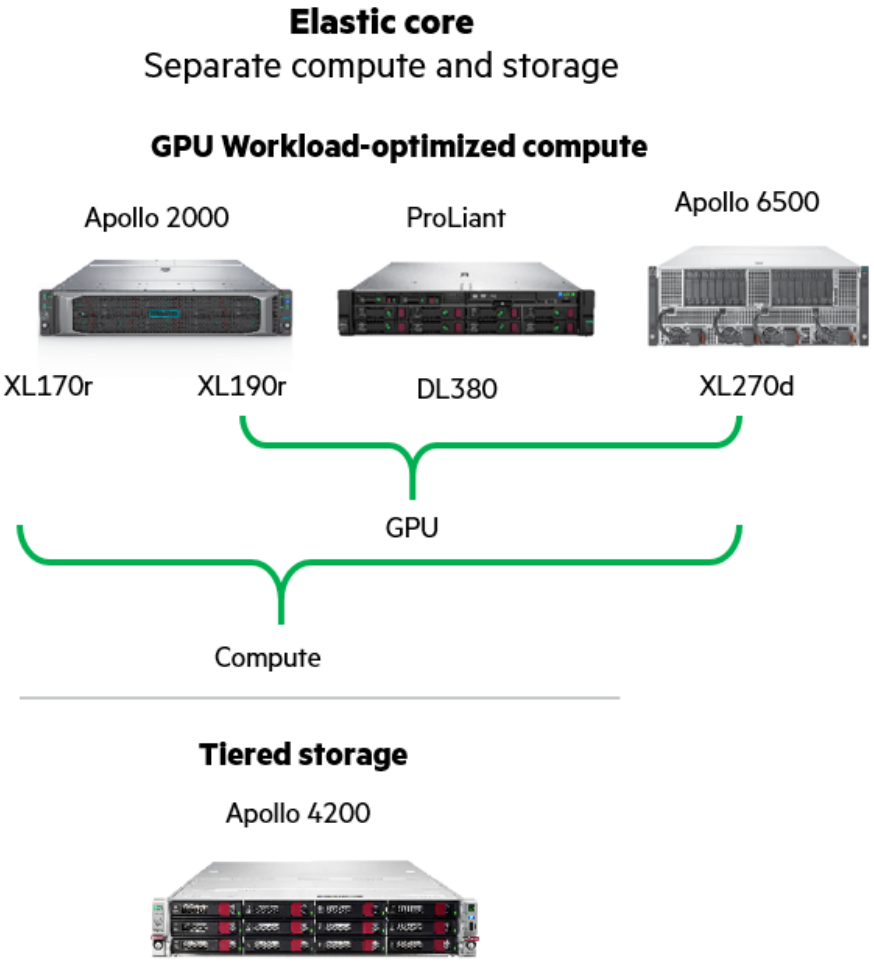

Figure 5. HPE Elastic Platform for analytics (EPA) building blocks

Figure 6 below shows a logical architecture of the GPUaaS solution. Using the container-based BlueData software platform, IT administrators and users can allocate and provision GPUs dynamically with GPUaaS.

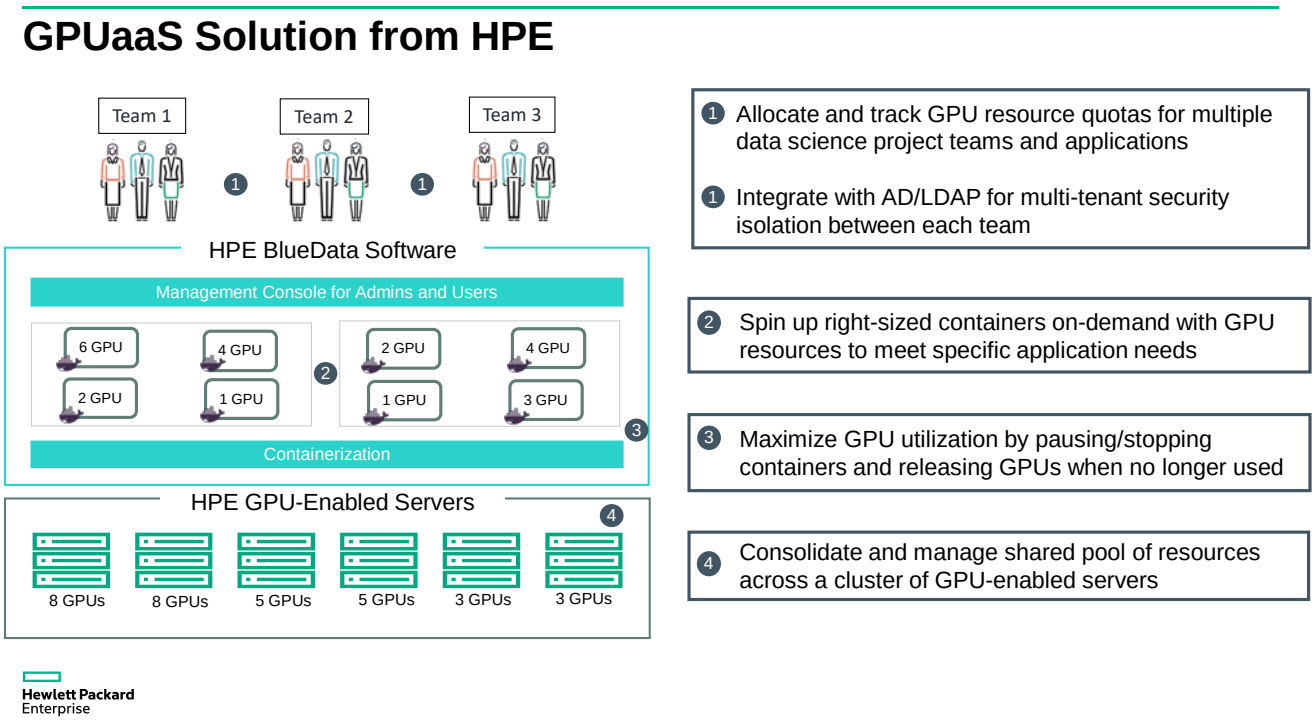

Figure 6. HPE GPUaaS solution – Providing a cloud-like experience for GPU infrastructure using containers

HPE EPA system configuration for BlueData EPIC software

The HPE EPA system provides great flexibility in deploying your workloads and managing your resource growth, by decoupling storage from compute. This allows you to add compute or storage capacity as needed without having to add both in lock-step. The architecture preserves the performance and availability benefits achieved through rack locality, while eliminating the need for node locality by leveraging high-speed networks for storage I/O performance and intelligent placement of analytic compute services on servers optimized for running specific components

In this HPE EPA example, the storage tier is an external HDFS cluster with the BlueData EPIC software running on the compute tier. This allows the BlueData clusters to be deployed on servers optimized for running containerized services. Furthermore, compute capacity can be easily scaled by adding more compute servers as new tenants and services are brought online without incurring the cost of scaling storage capacity unnecessarily.

HPE Apollo 2000 Compute servers

HPE Apollo 2000 systems deliver a scalable, high-density workload-optimized compute module, supporting HPE ProLiant XL190r Gen10 servers.

HPE ProLiant XL190r Gen10 server: For GPU workloads the HPE ProLiant XL190r Gen10 server delivers two servers in a single 2U chassis. Each HPE ProLiant XL190r Gen10 server is serviced individually without impacting the operation of the other server sharing the same chassis to provide increased server uptime. Each server supports two NVIDIA graphic cards and harnesses the performance of up to 2666MHz memory (16 DIMM slots per server) and two Intel processors in a very efficient solution that shares both power and cooling infrastructure.

HPE ProLiant XL170r Gen10 server: For compute intensive workloads the HPE ProLiant XL170r Gen10 server delivers four servers in a single 2U chassis. The HPE ProLiant XL170r Gen10 server has the same configurations options as the HPE ProLiant XL190r Gen10 server for CPU and memory.

For more information on HPE Apollo 2000 servers, visit https://buy.hpe.com/b2c/us/en/servers/apollo-systems/apollo-2000-system/apollo-2000-system/hpe-apollo-2000-system/p/1010192759

HPE Apollo 6500 GPU System (High performance and density optimized GPU server)

The HPE Apollo 6500 Gen10 System is an ideal high performance computing (HPC) and deep learning platform providing unprecedented performance with industry leading GPUs, fast GPU interconnect, high bandwidth fabric and a configurable GPU topology to match your workloads. The ability of computers to autonomously learn, predict, and adapt using massive data sets is driving innovation and competitive advantage across many industries and applications are driving these requirements. This system with rock-solid reliability, availability, and serviceability (RAS) features includes up to eight GPUs per server, NVLink for fast GPU-to-GPU communication, Intel® Xeon® Scalable processors support, choice of high-speed / low latency fabric, and is workload enhanced using flexible configuration capabilities. While aimed at deep learning workloads, the system is suitable for complex simulation and modeling workloads.

For more detailed information, visit https://buy.hpe.com/b2c/us/en/servers/apollo-systems/apollo-6500-system/apollo-6500-system/hpe-apollo-6500-gen10-system/p/1010742495.

HPE ProLiant DL380 server (Entry level GPU server)

The HPE ProLiant DL380 Gen10 server delivers the latest in security, performance and expandability, backed by a comprehensive warranty. Standardize on the industry's most trusted compute platform. The HPE ProLiant DL380 Gen10 server is securely designed to reduce costs and complexity, featuring the First and Second Generation Intel Xeon processor scalable family, plus the HPE 2933 MT/s DDR4 SmartMemory supporting 3.0 TB. It supports 12 Gb/s SAS, and up to 20 NVMe drive plus a broad range of compute options. HPE persistent memory offers unprecedented levels of performance for databases and analytic workloads. Run everything from the most basic to mission-critical applications and deploy with confidence. It supports up to four NVIDIA® Tesla or Quadro GPUs, which makes it a powerful deep learning platform in a 2U chassis.

For more detailed information, visit https://buy.hpe.com/b2c/us/en/servers/rack-servers/proliant-dl300-servers/proliant-dl380-server/hpe-proliant-dl380-gen10-server/p/1010026818.

HPE EPA with BlueData EPIC deployments

A BlueData EPIC system uses four distinct host types. The table below lists the host types and the recommended HPE server model to use.

Table 1. BlueData EPIC host types

Host Type	HPE server Model
EPIC - Controller/Shadow Controller/Arbiter	Apollo 2000 with up to 4 x HPE ProLiant XL170r Gen10 server
	Apollo 2000 with up to 2 x HPE ProLiant XL190r Gen10 server each with up to 2 GPUs
	HPE ProLiant DL380 with up to 4 GPUs
EPIC – Compute with GPUs	Apollo 6500 with HPE ProLiant XL270d and 4 or 8 GPUs
EPIC - Gateway	Apollo 2000 with up to 4 x HPE ProLiant XL170r Gen10 server

Figure 7 below shows a virtual rack diagram depicting the EPIC host node types and the corresponding HPE hardware.

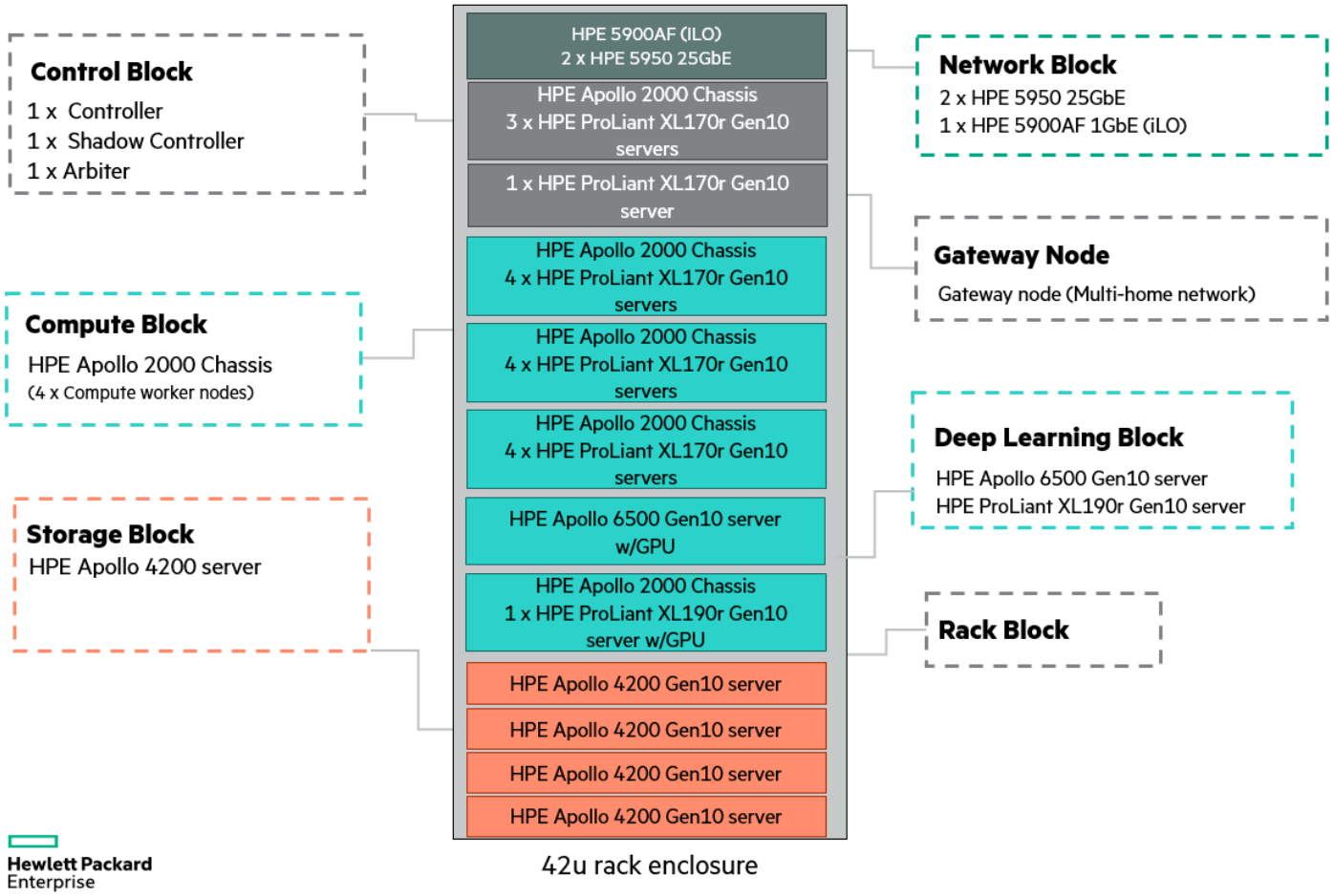

Figure 7. BlueData System - Virtual rack with host node types and recommended HPE servers

Listed below are three sample hardware deployments to support GPU intensive workloads; small, medium production, and large production. Each of the deployments has been rated for the number of worker nodes it supports with the BlueData EPIC CPU allocation ratio set as 1. The table below lists the virtual CPU, virtual memory, BlueData node storage capacity, and GPUs per node required for each sample node size.

Table 2. Sample deployment node sizes with GPUs

Node Size	Virtual CPUs	Virtual Memory GB	BlueData EPIC Node Storage GB	GPUs per Node
Small	4	16	250	1
Medium	8	32	250	2
Large	10	64	250	4

Small deployment: A lab deployment for prototyping, evaluating, and exploring AI/ML tools for initial evaluation or development, testing and quality assurance (QA). With BlueData and Hewlett Packard Enterprise they can accelerate their deployment in a multi-tenant lab environment for dev/test/QA, evaluate different AI/ML libraries and apps, or quickly prototype multiple data centric AI/ML pipelines. They can achieve faster

time-to-results with the ability to quickly and easily try out multiple libraries, and tools on shared infrastructure. They can increase business agility by empowering their data scientists and analysts to spin up new clusters in a matter of minutes, with just a few mouse clicks.

A BlueData EPIC cluster of four HPE ProLiant XL190r Gen10 servers works well for this type of deployment. This hardware configuration will support up to eight small and four medium nodes. Large nodes require a server that supports at least four GPUs.

Table 3. Starter deployment hardware configuration

Compute servers	Description
Model	HPE Apollo r2600 chassis - each holds 2 HPE ProLiant XL190r Gen10 servers
Server	HPE ProLiant XL190r Gen10 server
Processor	2 x Intel Cascade Lake 5215 - 10C 2.5GHz or Intel Skylake 5115 - 10C 2.4GHz per HPE ProLiant XL190r Gen10 server
Memory	384GB per HPE ProLiant XL190r Gen10 server
OS	2 x 960GB SFF SATA MU SSD per HPE ProLiant XL190r Gen10 server
BlueData EPIC Storage	3 x 1.64TB SFF SATA MU SSD per HPE ProLiant XL190r Gen10 server
Controller	HPE Smart Array E208i-p SR Gen10 per HPE ProLiant XL190r Gen10 server
Network card	HPE Ethernet 25Gb 2P 640FLR-SFP28 per HPE ProLiant XL190r Gen10 server
GPU cards	2 x HPE NVIDIA Tesla V100-32GB Module per HPE ProLiant XL190r Gen10 server

Medium production deployment: A system designed for use with multiple AI/ML departmental deployments that leverages a shared infrastructure. AI/ML initiatives usually begin in small isolated groups which often result in a proliferation of clusters, with each team managing their own environment. With BlueData and Hewlett Packard Enterprise, enterprises can leverage shared GPU infrastructure for multiple departments and centrally managed user groups. This multi-tenant architecture allows enterprises to consolidate and simplify GPU management for greater efficiency, enabling the sharing of resources for cost savings as well as the sharing of data – to eliminate the hassles and security risks of having to duplicate and store the same data for different user groups.

A medium production BlueData EPIC cluster of 6 HPE ProLiant DL380 Gen10 servers works well for this type of deployment. This hardware configuration will support up to 24 small, 12 medium, and 6 large nodes.

Table 4. Medium production deployment hardware configuration

Compute servers	Description
Server	HPE ProLiant DL380 Gen10
Processor	2 x Intel Cascade Lake 6246 - 12C 3.3Ghz or Intel Skylake 6146 - 12C 3.2GHz
Memory	384GB
OS	2 x 960GB SFF SATA MU SSD
BlueData EPIC Storage	3 x 1.64TB SFF SATA MU SSD
Controller	HPE Smart Array E208i-p SR Gen10
Network card	HPE Ethernet 25Gb 2P 640FLR-SFP28
GPU cards	4 x HPE NVIDIA Tesla T4 16GB

Large production deployment: A system designed for enterprise-wide, mission-critical AI/ML and analytics implementations in production. As AI and analytics initiatives within an organization evolve, there are a variety of different operational challenges, including high availability, security, data backup/recovery, software upgrades and patches, and more. In many enterprises, there are strict regulatory and compliance controls on internal data that need to be addressed as these deployments move from dev/test, through QA/UAT, and finally into production.

Performance becomes increasingly important, as AI/ML implementations move from "nice to have" science projects to "must have" mission-critical enterprise-wide deployments. As more and more users are on boarded for a variety of AI/ML projects and using a variety of tools (e.g., data wrangling, machine learning, real-time analytics), IT needs to support these new environments and applications – with limited resources.

BlueData enables faster software development and QA cycles, and non-disruptive testing and upgrades of the software components platform. As applications move from development and testing to production for their AI/ML initiatives, customers can leverage the same shared infrastructure to quickly provision different development and testing environments as needed, and have access to relevant datasets for comprehensive testing, without having to duplicate data. Data scientists and analysts can spin up instant clusters with their preferred applications and tools, to evaluate new technologies and analytic models, which reduce time-to-insight for analytics. A large production BlueData EPIC cluster of eight Apollo XL270d servers works well for this type of deployment. This hardware configuration will support up to 64 small, 32 medium, and 16 large nodes.

Table 5. Large production deployment hardware configuration

Compute servers	Description
Model	HPE Apollo 6500 chassis
Server	HPE ProLiant XL270d Gen10
Processor	2 x Intel Cascade Lake 6240 - 18C 2.6Ghz or Intel Skylake 6140 - 18C 2.3GHz
Memory	384GB
OS	2 x 960GB SFF SATA MU SSD
BlueData EPIC Storage	3 x 1.64TB SFF SATA MU SSD
Controller	HPE Smart Array P408i-a SR Gen10
GPU	8 x HPE NVIDIA Tesla V100-32GB Module
Network card	HPE Ethernet 10/25Gb 2P 640FLR-SFP28

Summary

The ability to leverage data in today's computationally intensive business environment is a key indicator of a business's success. As AI adoption in the enterprise grows, Hewlett Packard Enterprise delivers the compute and storage power to meet the challenges posed by machine learning (ML), deep learning (DL), and data analytics.

Hewlett Packard Enterprise offers a variety of products and services founded on industry-leading servers, with support for GPU acceleration and an ecosystem of software partners. By providing enterprises the critical set of fundamental accelerated computing hardware and software to choose from, Hewlett Packard Enterprise and its industry-leading partners enable machine learning, deep learning, data science, and other compute-intensive AI workloads in our customer's data centers, in hybrid cloud deployments, and at the edge.

With Hewlett Packard Enterprise's unique GPU-as-a-Service solution, enterprises can get started with their AI transformation journey and quickly scale, saving time and resources. Each customer's AI transformation journey is unique, requiring different environments and architectures for AI and advanced analytics. Hewlett Packard Enterprise delivers software and hardware solutions to accelerate enterprise computing in the way that works best for their infrastructure and application choices.

Furthermore, this GPUaaS solution enables IT organizations to increase business agility, simplify their GPU infrastructures, and deliver a cloud-like experience for the internal user's on-premises. The sharing of GPUs improves utilization and delivers a better return on investment.

Resources and additional links

BlueData, bluedata.com, and www.hpe.com/info/bluedata

HPE 'Seismic' sales briefcase for BlueData, https://hpe.seismic.com/X5/DocCenter.aspx?ContentId=d9c46d88-e08b-400c-9eba-9cc6ae456a3a#/search?appType=All&keyword=GPUaaS&contentType=All%20Documents&selectedProperties=&folderId=&folderName=&fromAppType=&currentTeamSiteId=&sharedTeamSiteId=&pageIndex=0

HPE Sizing Tool for the Elastic Platform for Analytics, https://solutionsizers.ext.hpe.com/EPASizer/

HPE AI Solutions, https://www.hpe.com/us/en/solutions/artificial-intelligence.html

HPE Big Data Solutions, hpe.com/bigdata

HPE Reference Architectures, hpe.com/info/ra

HPE Servers, hpe.com/servers

HPE Storage, hpe.com/storage

HPE Apollo 2000 systems, https://buy.hpe.com/b2c/us/en/servers/apollo-systems/apollo-2000-system/apollo-2000-system/hpe-apollo-2000-system/p/1010192759

HPE ProLiant DL380 systems, https://buy.hpe.com/b2c/us/en/servers/rack-servers/proliant-dl300-servers/proliant-dl380-server/hpe-proliant-dl380-gen10-server/p/1010026818

HPE Apollo 6500 systems, https://buy.hpe.com/b2c/us/en/servers/apollo-systems/apollo-6500-system/apollo-6500-system/hpe-apollo-6500-gen10-system/p/1010742495

HPE Networking, hpe.com/networking

HPE Technology Consulting Services, hpe.com/us/en/services/consulting.html

To help us improve our documents, please provide feedback at hpe.com/contact/feedback.

Make the right purchase decision. Click here to chat with our presales specialists.

Share

Sign up for updates

© Copyright 2019 Hewlett Packard Enterprise Development LP. The information contained herein is subject to change without notice. The only warranties for Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

Intel and Xeon are trademarks of Intel Corporation in the U.S. and other countries. Microsoft is a registered trademark of Microsoft Corporation in the United States and/or other countries. Google is a registered trademark of Google Inc. NVIDIA and NVIDIA Quadro are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. TensorFlow is trademark of Google Inc. BlueData Epic is trademark of BlueData Software.

a00078919enw, July 2019