Hewlett Packard Enterprise

Solution brief

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

ENABLING GPU AS A SERVICE

A cloud-like experience for GPU infrastructure using containers

The ability to leverage data in today's computationally intensive business environment is essential for a business's success. As 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, HPE offers a new GPU-as-a-service (GPUaaS) solution for on-premises deployments.

GPU-ACCELERATED WORKLOADS FOR ENTERPRISE AI DEPLOYMENTS

The development of ML and DL predictive models is compute intensive. The use of accelerators such as graphics processing units (GPUs) provides a performance boost that significantly speeds up development as compared to CPU-only systems making GPUs a common infrastructure choice for ML and DL.

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.

Furthermore, the complexity in standing up the right software components together with the underlying infrastructure is very time consuming and the process has to be repeated each time a new ML/DL application is requested.

Once the infrastructure is provisioned, IT has very little visibility into utilization to be able to reassign infrastructure to a different application. This lack of visibility also makes it difficult to implement more robust cost chargeback models.

There are public cloud services that offer the ability to deploy virtualized GPU resources on demand (that is, GPU as a service), but public cloud is not the only solution and, in some cases, may not be an option. Many organizations have workload requirements that require on-premises deployments due to considerations involving, security, performance, or data gravity.

ON-DEMAND AND ELASTIC PROVISIONING OF GPU RESOURCES

Now, there is a GPUaaS solution that combines best-in-class HPE infrastructure and HPE Ezmeral Container Platform, together with HPE Pointnext Services to ensure a successful deployment. This new HPE 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.

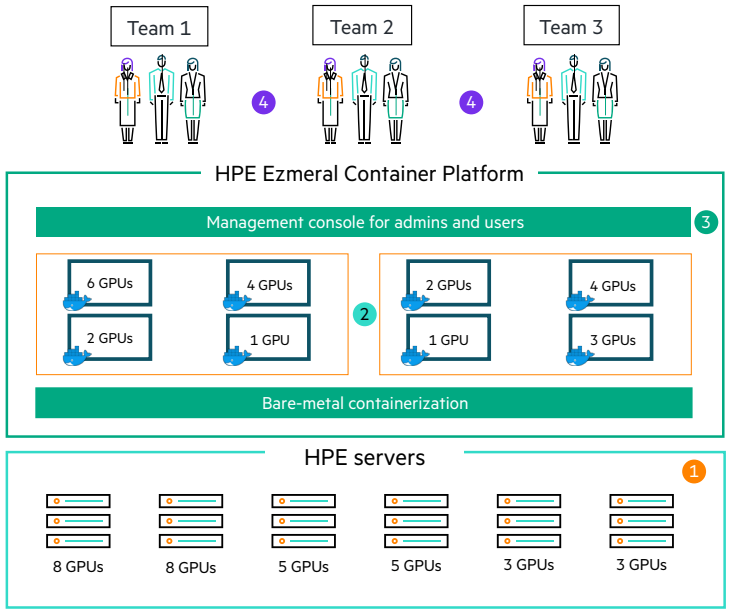
Using the container-based HPE Ezmeral Container Platform, GPUs from multiple heterogeneous servers can be consolidated and shared across multiple applications—for on-demand and elastic provisioning of containerized GPU resources, with just a few mouse clicks. To enable GPUaaS, HPE Ezmeral Container Platform can be deployed with GPU-enabled servers including HPE Apollo and HPE ProLiant servers with NVIDIA® Tesla or Quadro GPUs.

Furthermore, using the 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.

KEY FEATURES

- **Unified management console:** A GUI for administrators provides 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.
- **On-demand, elastic provisioning of GPU resources:** 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 deprovisioned (releasing the GPUs) when no longer needed.
- **Pause and restart applications:** Provides the ability to pause an application and release the attached GPUs while preserving the current state of the application. This allows IT admins to monitor usage and reassign the GPU when the GPU-specific code is executed.
- **Enterprise-grade security and multitenancy:** Provides multitenancy 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.
- **Out-of-the-box GPU-enabled application images:** Includes 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.

- **Bare-metal performance with containers:** Patented innovations that deliver the agility benefits of containers for ML/DL workloads while ensuring performance comparable to that of bare-metal servers.
- **External storage connectivity and data access control:** 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.



1. Consolidate and manage pool of heterogeneous GPU servers
2. Spin up rightsized containers on demand to meet user needs
3. Maximize GPU utilization by pausing/stopping containers
4. Allocate and manage resource quotas for different teams. Integrate with AD/LDAP for security isolation between teams

KEY BENEFITS

- **Reduce costs:** Achieve cost savings by improving GPU utilization, controlling usage, eliminating cluster sprawl, and minimizing data duplication.
- **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.

With HPE GPUaaS solution, enterprises can get their data science teams up and running quickly with GPU-accelerated applications in multinode containerized environments on state-of-art HPE infrastructure. Fully configured environments with GPUs can be provisioned in minutes—allowing data science teams to rapidly build prototypes, iterate, and experiment with their preferred ML/DL tools and frameworks. Enterprise IT teams can ensure enterprise-grade security, data protection, and performance in a multitenant architecture while increasing the ROI and reducing overall TCO for their GPU infrastructure.

LEARN MORE AT

hpe.com/info/container-platform

Make the right purchase decision.
Contact our presales specialists.

Chat **Email** **Call**

Get updates

Hewlett Packard Enterprise

a00075067ENW, September 2020, Rev. 2