GIGABYTE™

AI / DATA SCIENCE CLOUD PLATFORM

TURNKEY PACKAGE SOLUTION

INFINITIES

BIGTERA

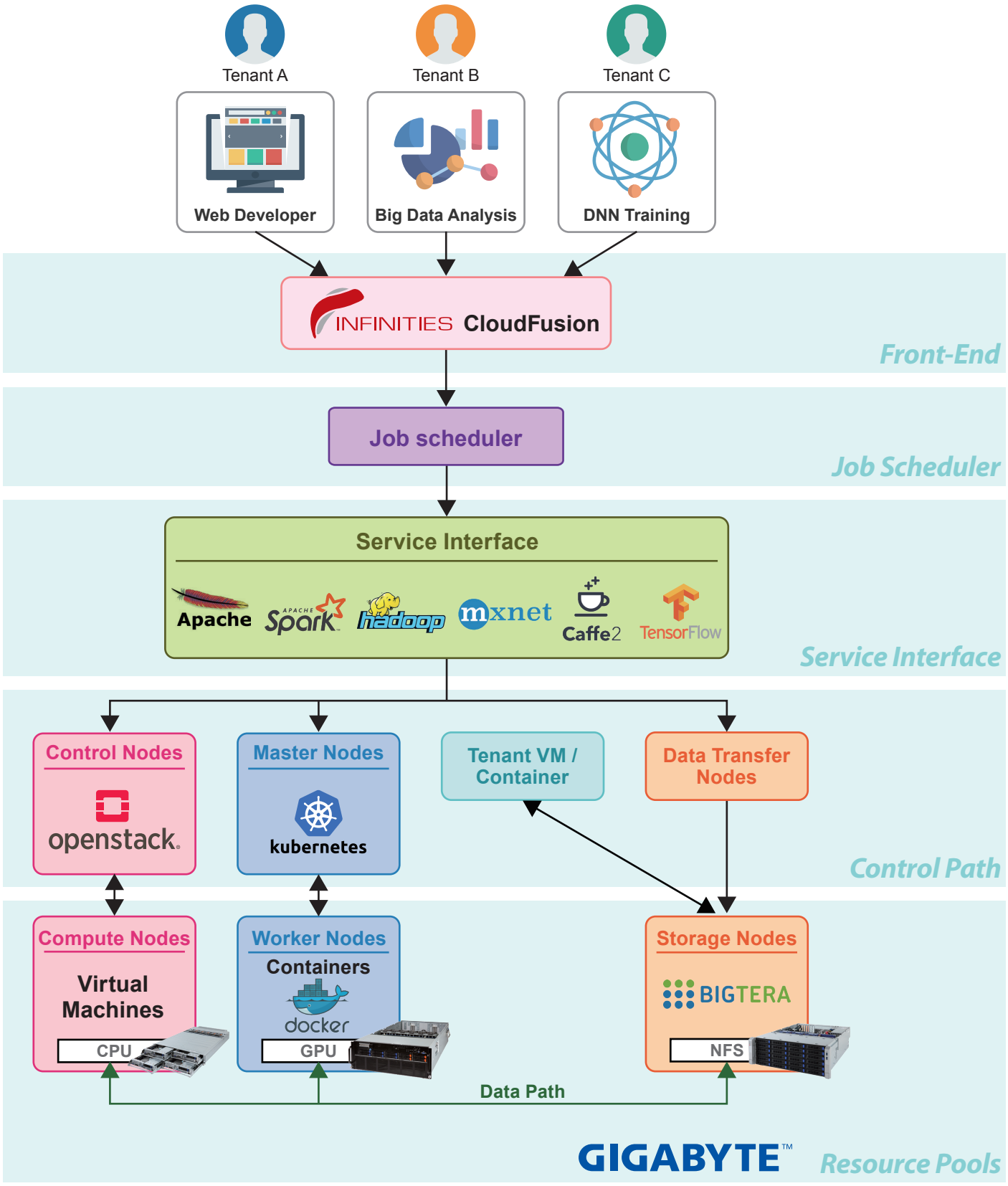
GIGABTYE has collaborated with software partners **InfinitiesSoft** & **Bigtera** to create an integrated **private / hybrid cloud platform turnkey package** to streamline data, tools and workflows in AI training & Big Data analysis. This cloud platform allows you to virtualize and share the GPU and CPU resources of your bare-metal hardware deployment, maximizing time and cost efficiency when running GPU-based AI / DNN training or CPU-based analysis workloads.

GIGABYTE's AI / Data Science Cloud turnkey package combines the following:

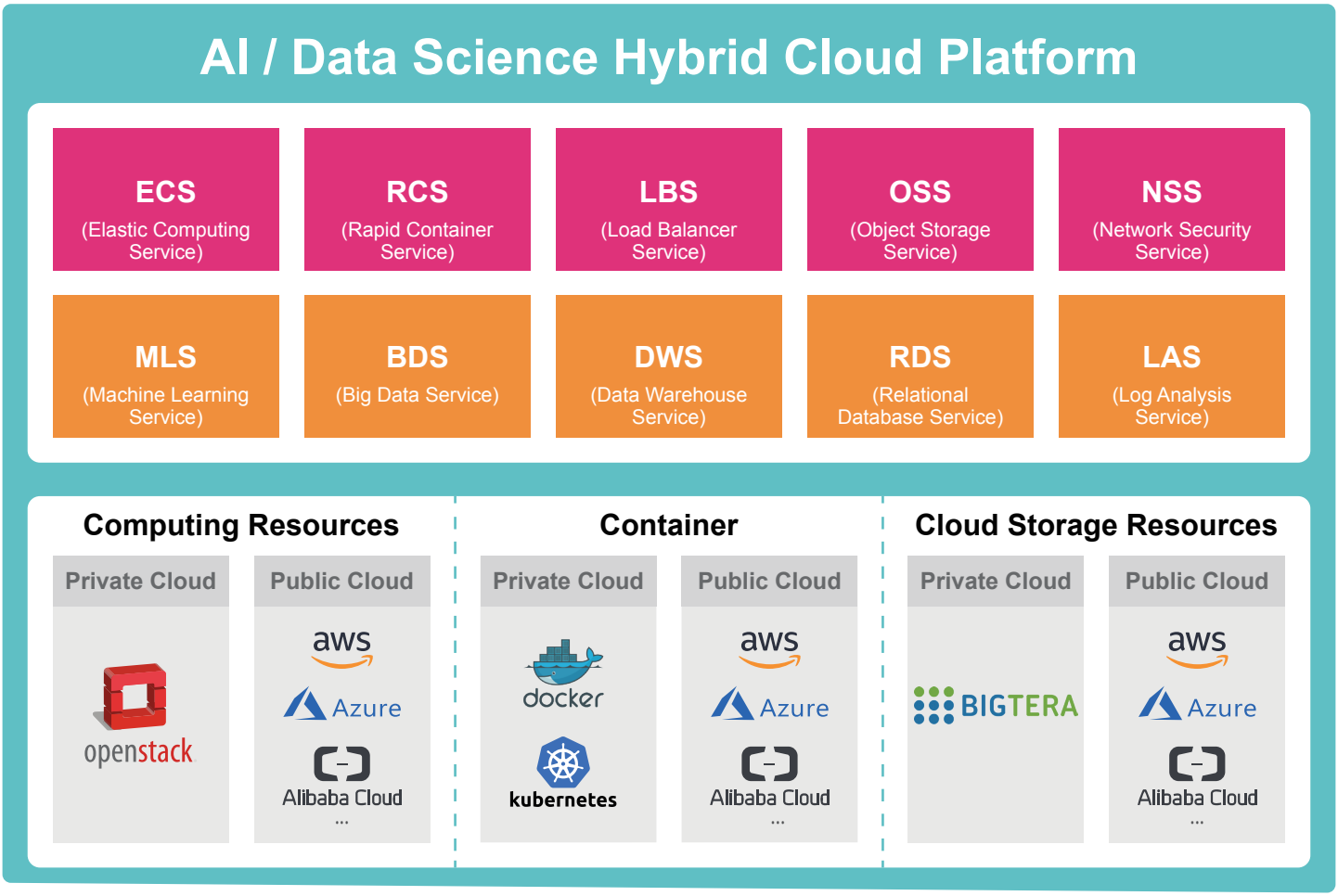
Management Layer: InfinitiesSoft CloudFusion cloud management platform to dynamically allocate virtualized resources and schedule workloads. **CloudFusion** also can pool on-premises physical resources with those from public cloud services (AWS, Azure, Google Cloud, Ali-Cloud etc.) to create cloud bursting functionality (hybrid cloud).

Virtualization Layer – Docker + Kubernetes for virtualization of GPU resources (containers), **OpenStack** for virtualization of CPU resources (virtual machines), and **BigTera VirtualStor Converger** or **Scaler** for software defined storage.

Hardware Layer – GIGABYTE server hardware for the underlying on-premises private cloud infrastructure



GIGABYTE's AI / Data Science Cloud is a holistic platform with a large range of features offered a single turnkey package:



InfinitiesSoft CloudFusion

InfinitiesSoft CloudFusion is used as the frontend platform layer for resource utilization, scheduling and management, and which can support and integrate the resources from over 30 different private and public clouds. This gives users the option to build a hybrid cloud by joining their on-premises private cloud to one or more public clouds. Furthermore, a highly elastic open API interface enables developers to connect and integrate new cloud options as they appear on the horizon, keeping your options open for future developments.

A **CloudFusion** deployment is designed both with users (i.e. AI and data scientists) and administrators in mind with comprehensive functionalities packaged in 2 portals designated for their distinctive roles:

CloudFusion User Portal: When AI and data scientists login to the **User Portal**, they can instantly view resource usage through the dashboard. The User Portal allows self-service by users for allocating virtual machine (CPU) and container (GPU) resources, selecting/mounting/loading their required CPU, GPU, Memory, AI Frameworks (e.g. Tensorflow, NVCaffe, Caffe2, PyTorch, MXNet, CNTK,… etc.) and accessing any other resource information relating to their work.

For use cases of interactive sessions, the system can automatically allocate data buckets to facilitate users to upload source training data for machine learning algorithms to produce post-training results (ML models). An object storage service is also provided to allow users to access bucket resources through the accesskeyid and accesskeysecret in S3 Tool.

A batch job mode is also supported to allow more advanced users to dispatch multiple model-training jobs without further human supervision. When it is found that computing resources needed for model-training are temporarily insufficient, a scheduling mechanism will initiate to automatically to put the jobs into a queue, so multiple jobs can be executed in parallel or when the next available computing resources become available, optimizing utilization for improved efficiency and to avoid leaving computing resources lying idle.

CloudFusion Administrator Portal: CloudFusion supports multi-tenancy. The administrator can define resource limits for each tenant and set user-accessible resource specifications, such as AI Framework, OpenStack Flavor configurations, and customizable pricing policies.

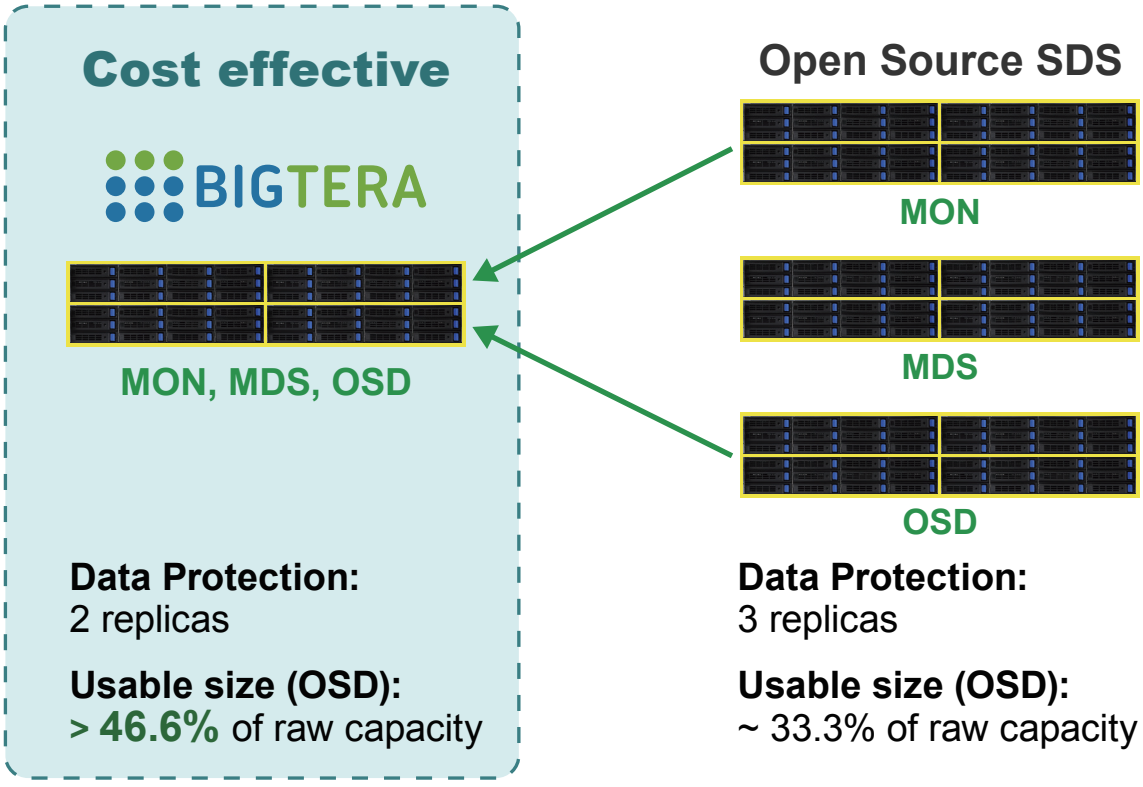
Kubernetes

Kubernetes is fast becoming essential to AI work and is a key feature of this cloud platform. It is the most popular container in machine learning workloads, as most scenarios are set up to run in Kubernetes containers due to its interactive mode capability. Containers make it easier, more secure, and faster for developers to develop, scale, and deliver AI applications. They also make it easier for data scientists to work with AI. Because Kubernetes containers can be scheduled and managed throughout the life cycle, it's also a favorite among developers and DevOps practitioners working with continuous release or continuous delivery application development processes. Machine learning developers also heavily favor Kubernetes for those same reasons.

Bigtera VirtualStor Converger or Scaler

Bigtera VirtualStor Converger or **Scaler** is used for a storage cluster. **VirtualStor Converger** is a hyper-converged storage solution that runs on top of the virtualization platform (KVM) and consolidates the internal HDD of the hypervisor into a single storage pool where it provides shared storage for the hypervisor. **VirtualStor Scaler** is a scale-out software defined storage solution that provides the flexibility to specify the storage type (NAS, SAN or Object Storage) and offers a cost-effective "all in one" architecture by combining together MON, MDS and OSD together in a single server.

Both products offer high performance features to significantly increase performance, with up to 10 x more IOPS than open source software defined storage system, as well as significantly reducing SSD write amplification to increase SSD lifespan endurance.



All in One Platform - Cost effective
(Bigtera SDS vs Open Source SDS)

The turnkey package is offered in the following configuration choices:

Light Version

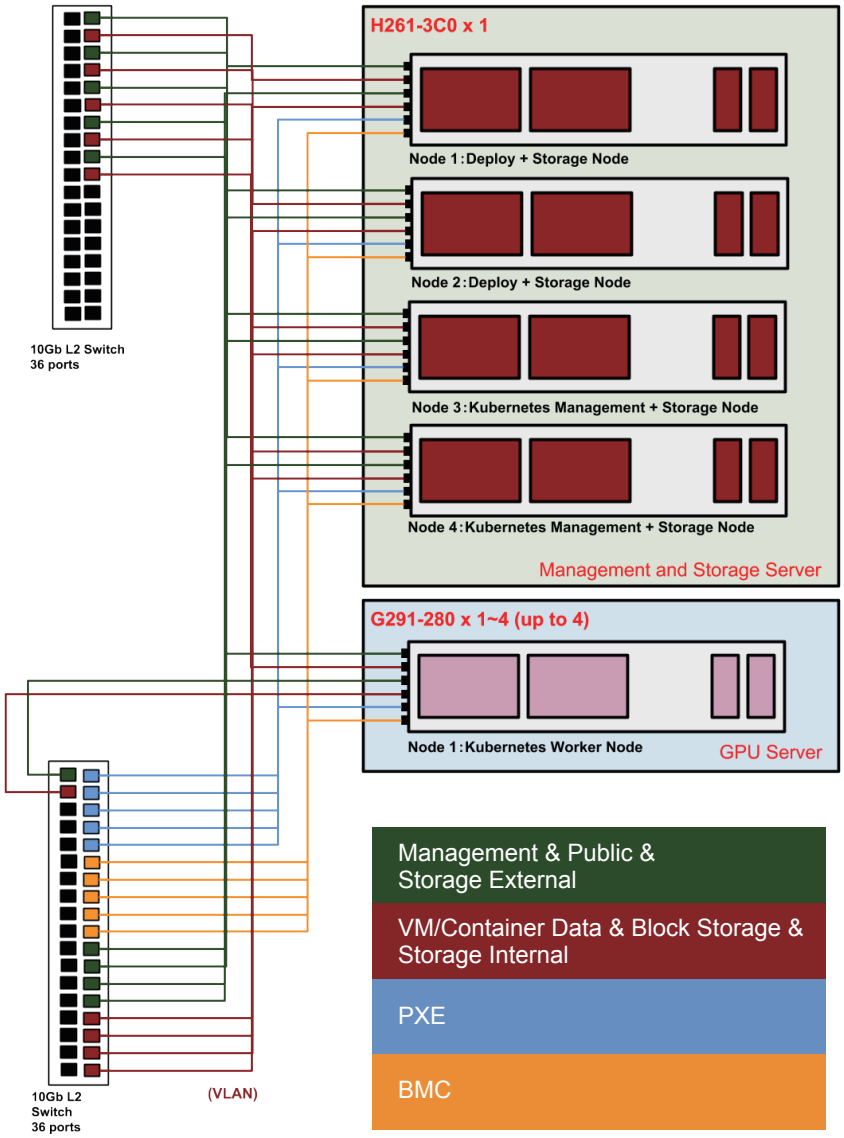
Hardware:

1 x H261-3C0, 1 x G291-280 (or up to 4 x G291-280)

Software:

InfinitiesSoft CloudFusion, Kubernetes, Bigtera VirtualStor Converger

Entry level package: GPU resources are virtualized with Kubernetes. DNN workloads can be scheduled and shared to run on these resources via CloudFusion. A virtual vSAN-like storage platform is provided with VirtualStor Converger.



Medium Version

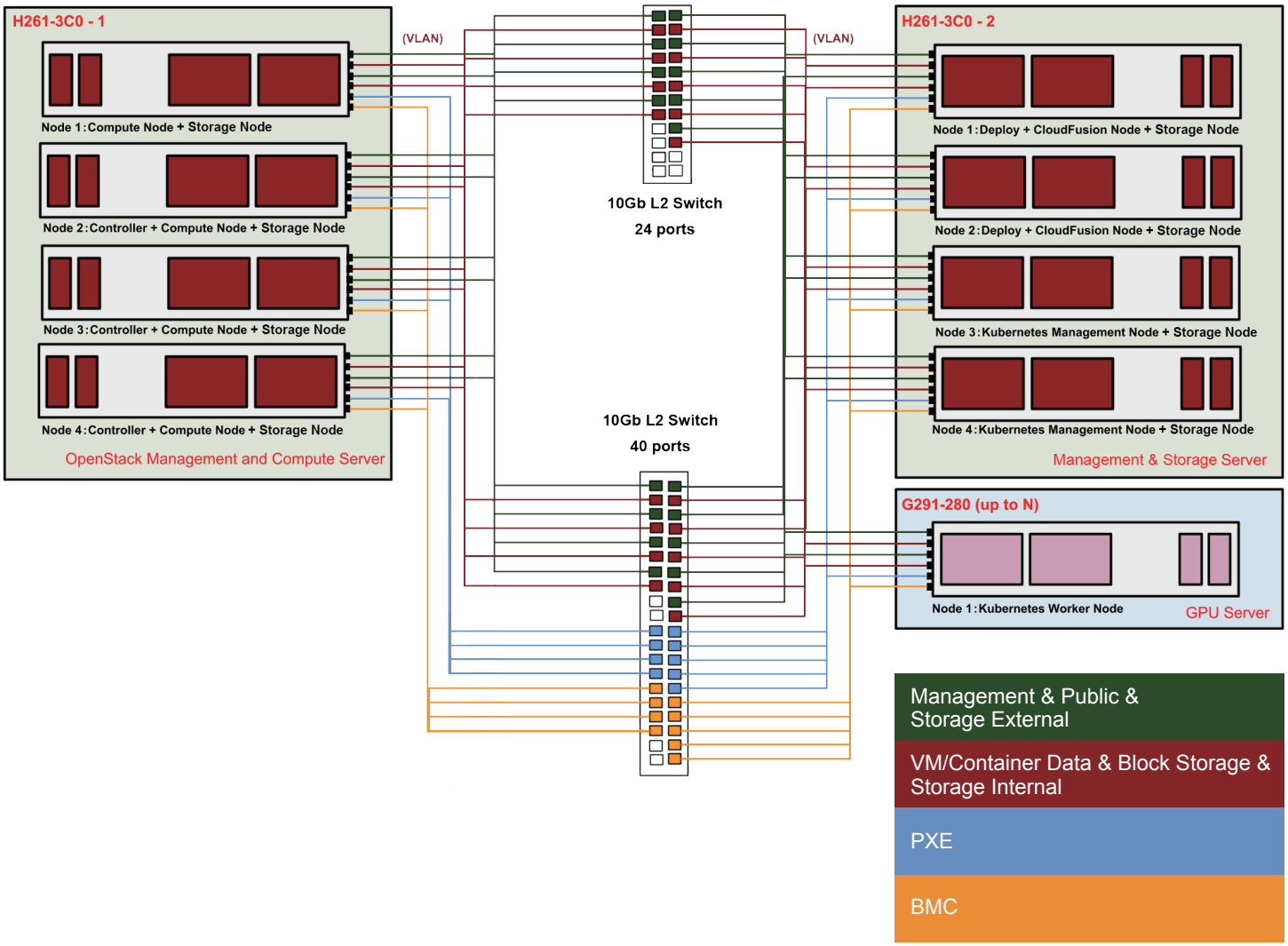
Hardware:

2 x H261-3C0, 1 x G291-280 (or up to 4 x G291-280)

Software:

InfinitiesSoft CloudFusion, Kubernetes, OpenStack, Bigtera VirtualStor Converger

Adds virtualized compute functionality. In addition to DNN workloads scheduled and run on GPU server via Kubernetes containers, compute workloads (such as big data analysis via Hadoop) can be run and managed on virtual machines via OpenStack. A virtual vSAN-like storage platform is provided with VirtualStor Converger.

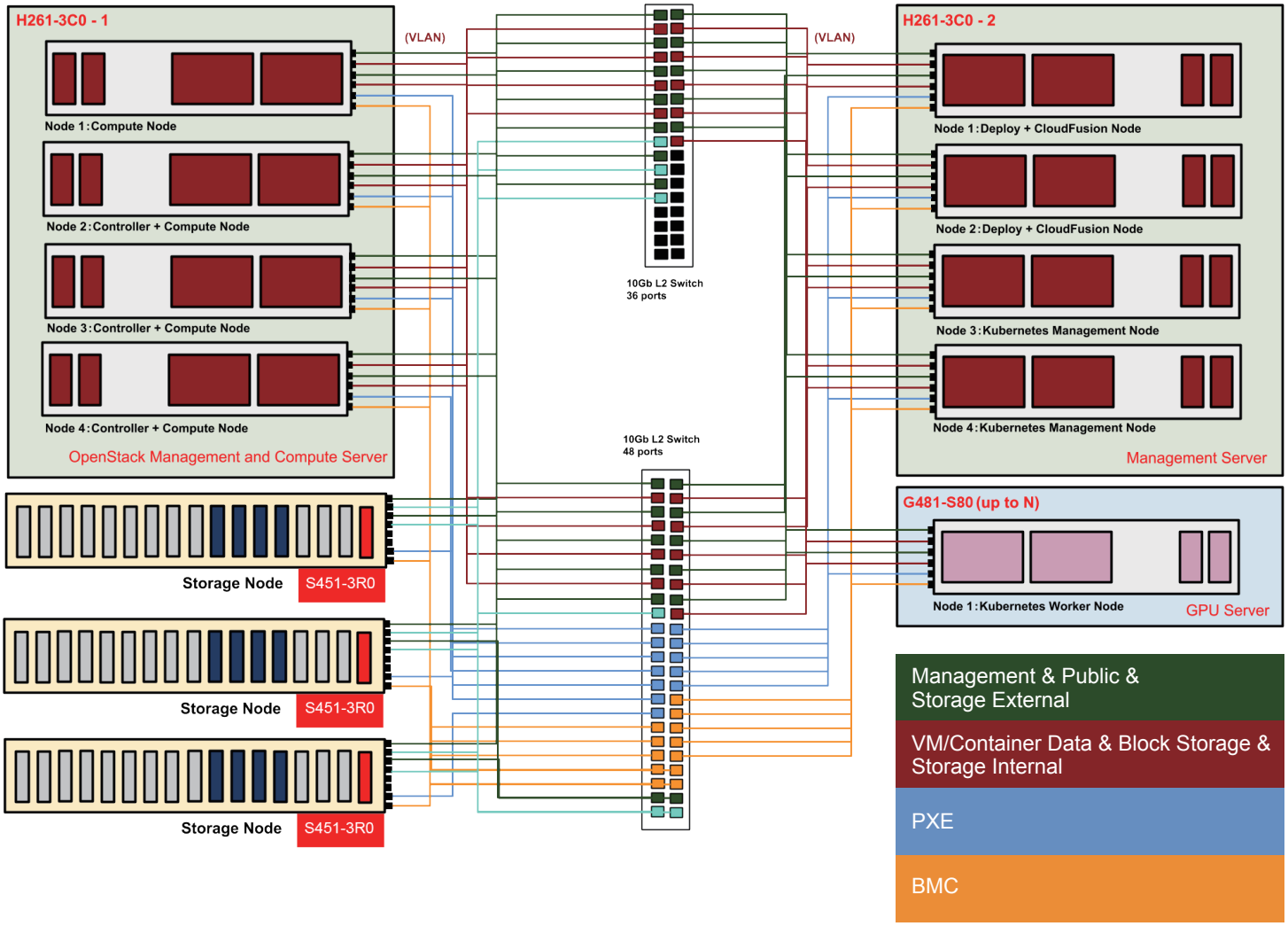


Heavy Version

Hardware:
2 x H261-3C0, 1 x G481-S80 (or up to 4 x G481-S80), 3 x S451-3R0

Software:
InfinitiesSoft CloudFusion, Kubernetes, OpenStack, Bigtera VirtualStor Scaler

The most powerful version, combining the virtual machine / containers functionality of the light / medium versions with a powerful 8 x SXM2 GPU server for maximum DNN training performance together with an independent software defined storage cluster (VirtualStor Scaler) for enhanced scale-out storage capacity.



Turnkey Package Configuration Specifications


H261-3C0


G291-280


G481-S80


S451-3R0

Solutions	Light	Medium	Heavy
Component	1 x H261-3C0 1 x G291-280 (or up to 4 x G291-280)	2 x H261-3C0 1 x G291-280 (or up to 4 x G291-280)	2 x H261-3C0 1 x G481-S80 (or up to 4 x G481-S80) 3 x S451-3R0
CPU	10 x Intel Xeon Silver 4114 processors	18 x Intel Xeon Silver 4114 processors	Customizable according to user's requirements
Memory	H261-3C0: 24 x 16GB DDR4-2666 RDIMM G291-280: 12 x 16GB DDR4-2666 RDIMM		
HDD	H261-3C0: 12 x 8TB SATA 6G HDD G291-280: 8 x 2TB SAS 6G HDD (w/ 1 x CRA4448 RAID card)		
SSD	H261-3C0: 8 x 480GB M.2 cards (w/ 4 x CMT4032 M.2 expander cards)		
Usable Storage Capacity	Up to 48TB	Up to 96TB	Up to 348TB
GPU	6 x NVIDIA Tesla V100 PCIe GPGPU		8 x NVIDIA Tesla V100 SXM2 GPU with NVLink
Networking	10GbE Base-T H261-3C0: 4 x CLN4222 G291-280: 1 x CLN4222		Customizable according to user's requirements
Warranty	3 years		