



# GIGABYTE™



## InfinitesSoft AI-Stack



GIGABYTE has collaborated with InfinitesSoft to create an integrated private / hybrid cloud platform 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.

# InfinitesSoft AI-Stack combines the following:

## Management Layer:

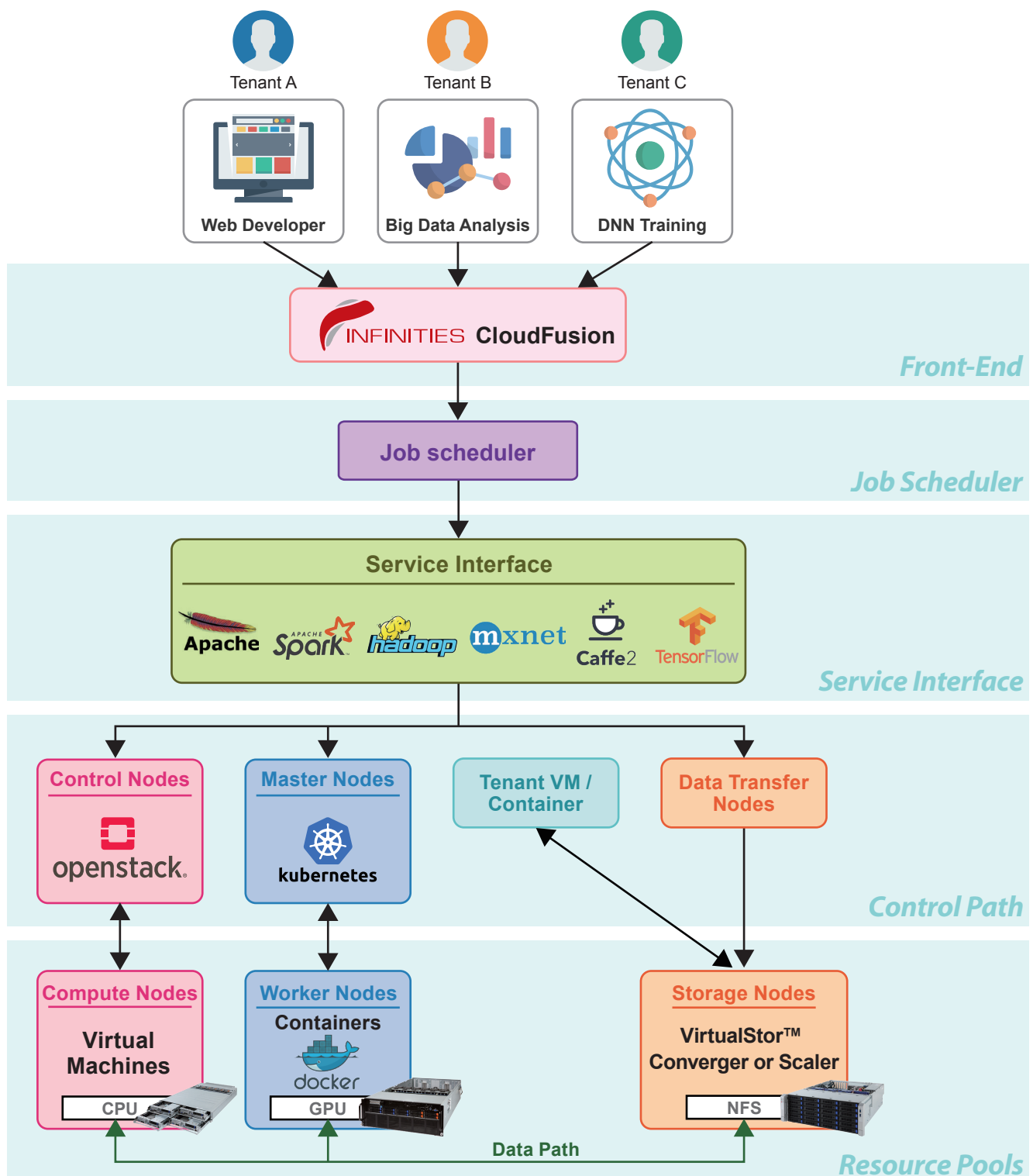
InfinitesSoft CloudFusion cloud management platform to dynamically allocate virtualized resources and schedule workloads. CloudFusion also can pool on-premises physical resources with those from public cloud services (AWS, Azure, Google Cloud, Ali-Cloud etc.) to create cloud bursting functionality (hybrid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

## Virtualization Layer:

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

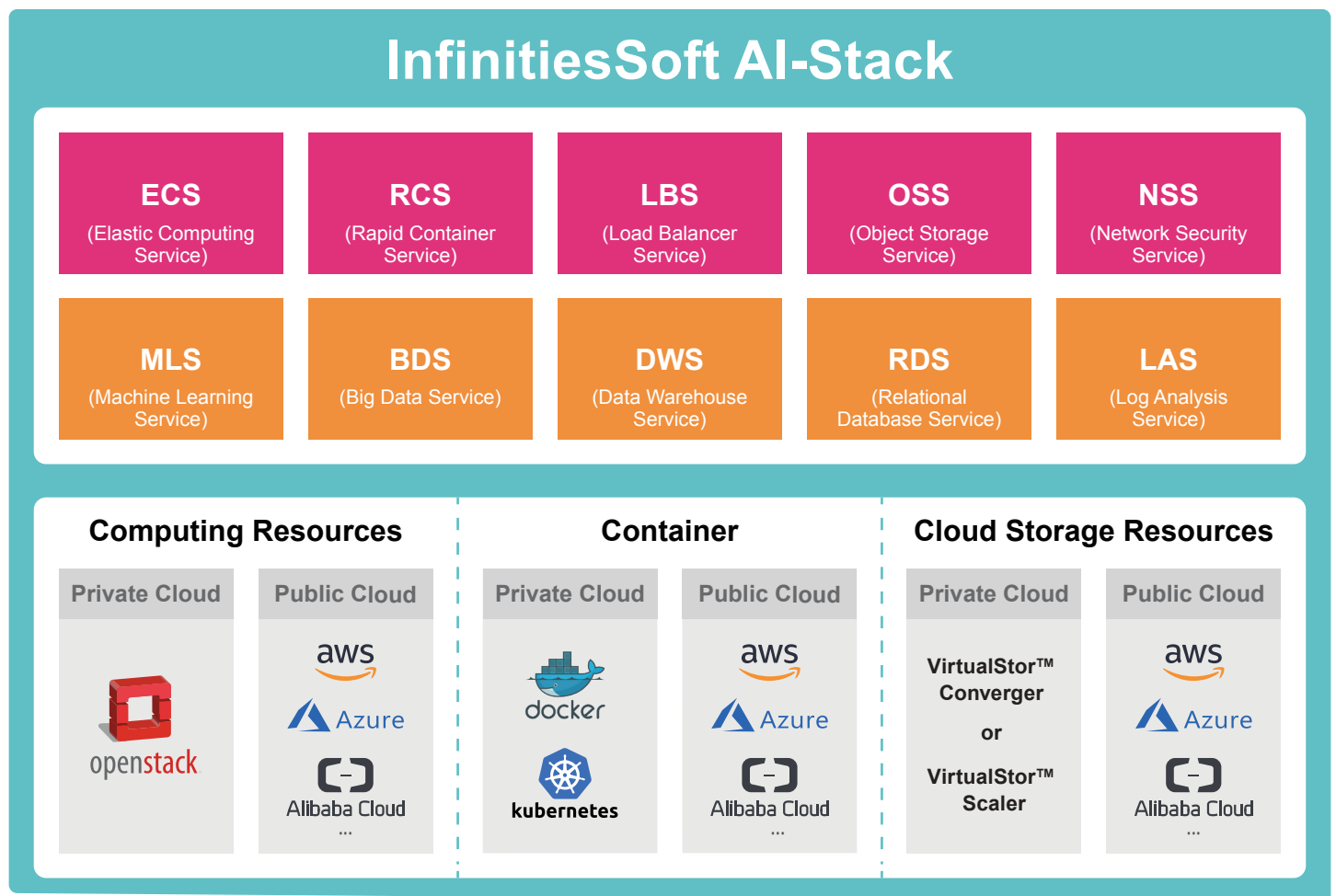
## Hardware Layer:

GIGABYTE server hardware for the underlying on-premises private cloud infrastructure.



# AI-STACK FEATURES

The AI-Stack enables data science teams, developers, and IT teams to simplify and streamline workloads through a single system.



AI/ML capabilities are already integrated into this stack so that users can focus on AI/ML workloads and not on system maintenance, adjustment and deployment scheduling. The stack reduces complexity and the learning curve for users to adopt and master Tensorflow, Caffe, and other deep learning tools.

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. Both Docker + Kubernetes and Singularity are containers that can be used in this system. Singularity is lightweight and non-IP (HPC) based which is designed for a single user, and ideal for non-interactive batch jobs. By comparison, Kubernetes is heavyweight, IP-based and therefore allows multiple user connections, and ideally suited for interactive jobs.

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. 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. Open source tools are increasingly becoming available on the market and further add appeal to using Kubernetes for data science work. For example, the Kubeflow open source tool enables teams to easily attach existing machine learning jobs to a cluster without having to do much in the way of adaptations or integrations.

# HOW IT IS BUILT: Management Layer

## InfinitesSoft CloudFusion Cloud Management Platform

The front-end management platform layer of the AI-Stack is provided by InfinitesSoft CloudFusion, which can support and integrate over 30 different private and public clouds. This gives users the option to build a hybrid cloud wherein they can join their private cloud to one or more public clouds and reap the benefits of all those cloud options.

Users can easily add, drop or change any of their clouds. They can also use the easy-to-understand visualizations on the dashboard to:

- 1) Allocate resources and manage access,
- 2) Evaluate and manage cloud data center CPU, memory,
- 3) Manage storage resource utilization.

Further, a highly elastic open API interface enables developers to connect and integrate new cloud options as they appear on the horizon thus keeping your options open for future developments.

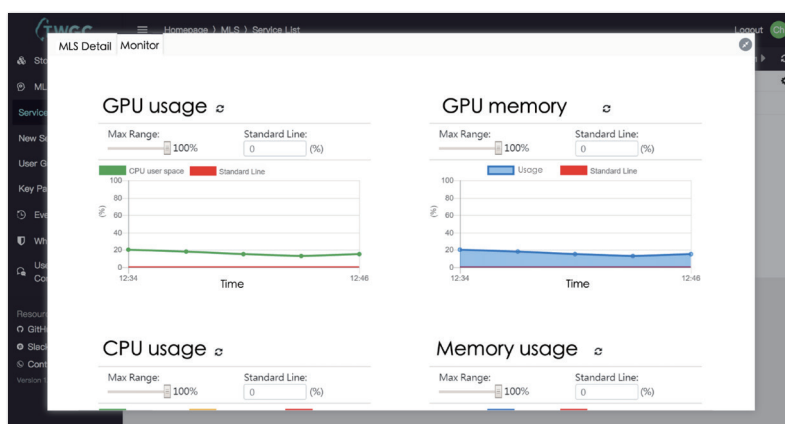
A CloudFusion deployment for AI-Stack is designed with users (i.e. AI and data scientists) and administrators in mind with comprehensive functionalities packaged in 2 portals designated for their distinctive roles:

## User Portal

When AI and data scientists (as users) login to the User Portal, they can instantly view resource usage through the dashboard. 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.

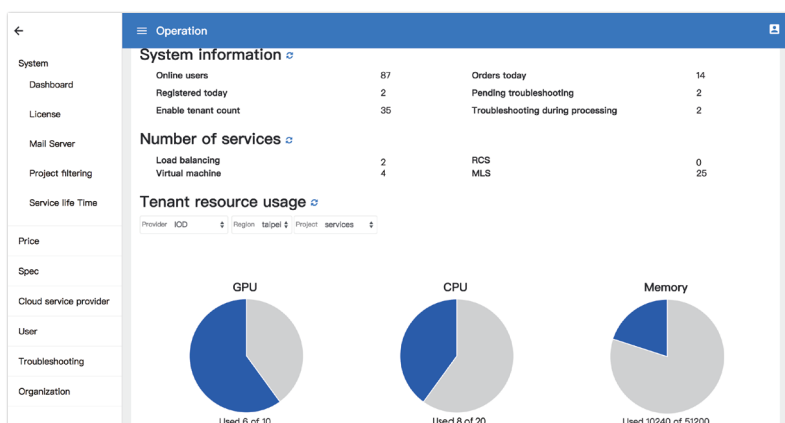
A batch job mode is also supported to allow more advanced users to dispatch multiple model-training jobs without further human supervision. When it is found that computing resources needed for model-training are temporarily insufficient, a scheduling mechanism will initiate to automatically to put the 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 without minute-by-minute human interventions.



CloudFusion User Portal Interface Screenshot

## 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. Besides the private cloud platform incorporated within the AI-Stack additional cloud platform resources can be integrated and managed under the hybrid/multi-cloud management capabilities of CloudFusion, including, but not limited to, resources from public clouds (e.g. AWS, Alicloud), and/or private clouds (e.g. OpenStack, VMware, Kubernetes), etc.

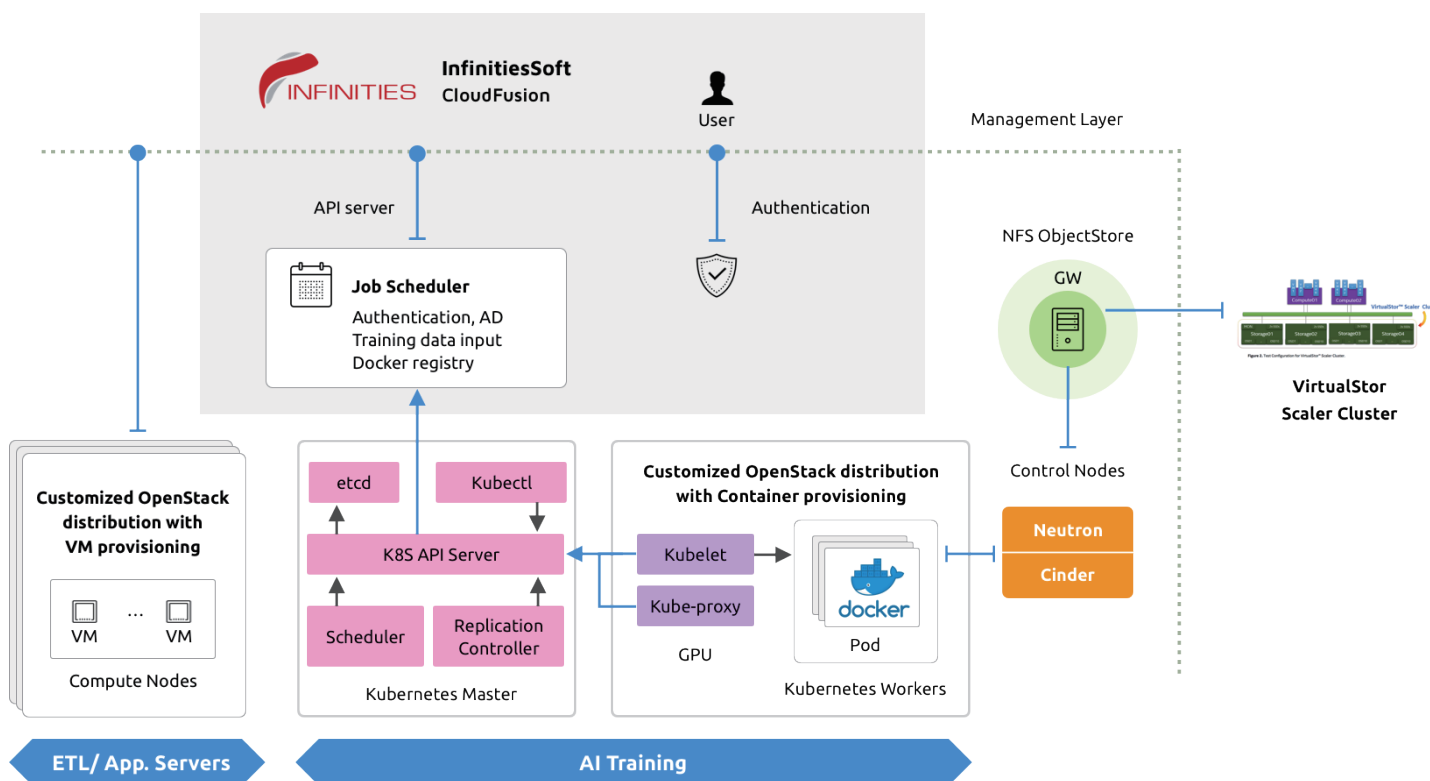


CloudFusion Administrator Portal Interface Screenshot

# HOW IT IS BUILT: Virtualization Layer

## Customized OpenStack Distribution for Machine Learning

The virtualization layer is used to control nodes and resource pools of the on-premises hardware infrastructure of the private cloud, and is delivered by a customized OpenStack distribution which includes capabilities for machine learning, by featuring integration with Kubernetes for automatic deployment of machine learning containers onto GPU servers for AI training.



This customized OpenStack distribution is administered and managed by InfinitiesSoft CloudFusion cloud management platform for resource allocation and scheduling, and is integrated with a software defined storage cluster using Bigtera VirtualStor™ (Scaler / Converger / Extreme).

On top of standard OpenStack features, this customized distribution also includes the following additions:





- **OpenStack and Kubernetes integration:** for automatic policy-based deployment of VMs and containers onto any compute node or Kubernetes worker node, through the user-friendly CloudFusion User Portal
- **Tenant based isolation:** OpenStack has an inherent architectural concept of tenant which is completely missing from Kubernetes
- **Kubernetes master node clustering:** to provide full HA and load balancing capability for Kubernetes
- **NFS – Object Storage Gateway:** to ease the migration of legacy software based on the NFS semantics/syntax towards the adoption of an object based storage system
- **Automatic deployment of DNN development environment:** specifically, the customized OpenStack distribution automates the deployment of a DNN development environment (TensorFlow, Tensorboard, Caffe, Jupyter, DIGITS etc.) in the form of containers onto GPU enabled servers
- **User authentication and authorization of DNN IDEs (Integrated Development Environment):** the nature of DNN development/training is, unlike traditional HPC, interactive. For the security and integrity of the system, the customized OpenStack distribution provides mandatory authentication and authorization for users
- **Integration with HPC job schedulers:** unlike in the context of VMs and generic containers, containers for DNN training occupy GPU-enabled servers for acceleration and quicker iterations. Even so, each submitted job will still takes days or weeks for completing one iteration. The use of GPU servers comes at exorbitant costs. Thus, traditional job schedulers like SLURM or Univa Grid Engine are used for resource (storage and GPUs) management and schedulers to contain that expense.

In summary, these features free data scientists and data science teams from resource allocation, environment setup, container preparation duties or other integration and security woes, giving them more time instead to focus on the actual work of training machine learning algorithms.



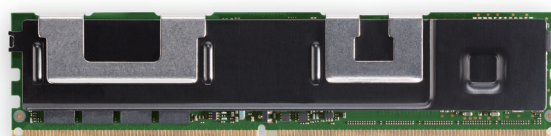
## HOW IT IS BUILT: Hardware Layer

The AI-Stack is designed and optimized to be used with GIGABYTE's 2nd Generation Intel® Xeon® Scalable Family server systems for the underlying hardware layer.

R-Series	H-Series	S-Series	G-Series
			
All-Flash Storage Support for a High-Speed File System	2U 4 Nodes for a High Density Virtualized Compute Cluster	Dense HDD Capacity for Scale-Out Block or Object Storage	Dense GPGPU Capacity for a GPU-Compute Cluster

GIGABYTE has a rich product family of server systems designed for Intel's 2nd Generation Xeon® Scalable Family platform, and are engineered to support the full family of different Xeon® Scalable SKUs that are workload optimized to support different applications, making your GIGABYTE server ideal for a myriad of use cases, from enterprise IT and database, cloud and storage to the most high-demand HPC workloads.

GIGABYTE's 2nd Generation Intel® Xeon® Scalable Family processor servers also come ready to support Intel® Optane™ DC Persistent Memory, a revolutionary new product that re-defines traditional memory & storage architectures by filling the gap between DRAM, which is ultra-fast but low capacity and expensive, and SSDs, which are higher capacity and more affordable but slower, by enabling a large persistent memory tier between DRAM and SSDs that's fast and affordable.



Intel® Optane™ DC Persistent Memory modules integrate seamlessly into existing DDR4 DIMM slots on each GIGABYTE server, and use Intel's 3D Xpoint™ memory technology, which unlike DRAM retains data after power is removed.

The high capacity, persistence and affordability compared to DRAM of Intel® Optane™ DC Persistent Memory will allow users to bring more data closer to the CPU, for faster time to insight. When paired with the 2nd Generation of Intel® Xeon® Scalable Family processors on GIGABYTE servers, users will see real performance, throughput and persistent advantages on many of the most memory bound workloads.



[www.gigabyte.com](http://www.gigabyte.com)



[gigabyteserver](https://www.facebook.com/gigabyteserver)



[@GIGABYTESERVER](https://twitter.com/GIGABYTESERVER)



[GIGABYTE](https://www.linkedin.com/company/gigabyte)



### GIGABYTE TECHNOLOGY CO., LTD.

- \* All intellectual property rights, including without limitation to copyright and trademark of this work and its derivative works are the property of, or are licensed to, GIGA-BYTE TECHNOLOGY CO., LTD. Any unauthorized use is strictly prohibited.
- \* The entire materials provided herein are for reference only. GIGABYTE reserves the right to modify or revise the content at anytime without prior notice.
- \* All other brands, logos and names are property of their respective owners.