SOLUTION OVERVIEW

GPU as a Service on the F5 Distributed Cloud Platform

Smart Computing for the Cloud and the Edge

F5 Distributed Cloud Services bring the power of graphics processing units (GPUs) to your cloud and edge locations with GPU as a Service, providing a cloud-like experience using containers and virtual machines.

For decades, the use of GPUs helped companies render 3D graphics and other graphics intensive application workloads. Enterprises today are leveraging the computing power of GPU for exponentially more—real-time data processing, data analysis from AI and machine learning applications, and deep learning capabilities.

GPU advancements combined with the growing demand for artificial intelligence (AI) and high-performance computing have inspired an increasing number of industries—from healthcare to construction to automotive and energy—to begin realizing more value from their data. GPUs are helping companies accelerate their digital transformation.

Moreover, GPU is now available in the cloud and at the edge—closer to the user and closer to where applications and data reside—so companies can take full advantage of its vast potential. However, the GPU must be configurable and operational independent of the hardware or platform used. Also, latency must be managed and minimized, sufficient bandwidth made available, and data governance maintained. F5 is bringing GPU to your cloud and edge locations with GPU as a Service (GPUaaS) on the F5 Distributed Cloud Platform, supported by NVIDIA GPU (P1000 and Tesla T4).

Run it in multi-cloud, edge, and on-premises environments

With F5® Distributed Cloud GPUaaS, you get a cloud-like experience using containers and virtual machines. It can be deployed in multi-cloud, edge, and on-premises environments. With a single-pane-of-glass dashboard, you can accelerate AI and machine learning workloads, expedite deep learning model and application deployment, and boost insights and analysis by processing data in real time.

Distributed Cloud GPUaaS enables massive deployment at scale and parallel data processing across a fleet of GPUs, reducing computing tasks and decreasing total cost of ownership. Users leverage a fully managed platform with pre-installed services and built-in GPU enablement and automations, all controlled from a single dashboard.

Multi-tenancy capabilities allow you to share GPU services with trusted users and organizations. You reap resource utilization benefits from time-slicing deployment to 1 GPU, plus system multi-GPUs that improve operational efficiencies.

Using a single-pane-of-glass dashboard, you can deploy, process, scale, and manage the lifecycle of your unique distributed AI/ML and deep learning workloads and use cases.

DELIVER THE FULL POTENTIAL OF GPU BY SERVICING USERS AND APPLICATIONS IN THE CLOUD AND AT THE EDGE.

Key features

- **SaaS-based service:** Leverage GPU as a Service for faster, scalable AI-based applications. Distributed Cloud GPUaaS runs on any cloud, edge, and on-premises environments, providing a cloud-like experience.
- **Efficiently share GPU resources:** Manage a shared pool of resources and autoscale GPUs as needed for maximum utilization. Minimize costly underutilized CPUs caused by inefficient resource allocation, data bottlenecks, and complicated DevOps.
- **Accelerate AI/ML deployment and inferencing:** With F5 Distributed Cloud Console, you can provision, configure, deploy, and operate AI/ML workloads and applications on GPUs based on resource availability. Get end-to-end visibility and manage inferencing and continuous training.
- **Multi-tenancy support across any platform:** Take advantage of multi-tenancy capabilities, with choice of shared mode (multiple container time-slicing with 1 GPU) and pass-through mode (1 container to 1 GPU).

Key benefits

- **Zero-touch provisioning:** Onboard, provision, configure, and deploy GPUs based on policy across multi-cloud, edge, and on-premises sites—with minimal human touch and effort.
- **Premium servers and dedicated access:** Benefit from F5 premium servers that are embedded with GPUs and preconfigured for plug and play. Zero virtualization is required. Get direct, unfettered access to the system, applying resources based on role-based access control (RBAC).
- **Maximum GPU utilization and cost** efficiency: Understand and leverage GPU resources for the right applications, based on consumption, and maximize the benefit of your GPU purchase by subdividing the resources and unblocking the queue.
- **Security and compliance for your data:** F5 encrypts data to reduce man-in-the-middle (MITM) attacks and eliminate data exposure. Utilize world-class security features at the edge and in the cloud. Fully comply with regulations such as GDPR, HIPAA, and others depending on your requirements and geographical locations.

F5 Distributed Cloud GPUaaS vs. traditional mechanisms

Features	Traditional	Distributed Cloud GPU
GPU as a Service	Yes	Yes
Multi-tenancy support with choice of share mode	No	Yes
Deploy at the edge and to any cloud	No	Yes
Hardware pre-installed with GPU	Yes	Yes
Resource-based deployment	No	Yes
Shared pool of resources	No	Yes
Inferencing at the edge	Yes	Yes
Lifecycle management	Manual	SaaS

Figure 1: View how F5 Distributed Cloud GPUaaS compares to traditional GPU service offerings available today.

Use Case	Deployment Location
GPU Cloud as a Service inference or training workloads	Private cloud—Customer Edge
Multi-tenancy support with choice of share mode	Private cloud
Training workloads on the edge: Video analytics, smart retail, data analytics	Edge: Retail stores, smart building, manufacturing plants, manufacturing line edge
Inference on the edge: Video analytics, smart retail, data analytics	Edge: Retail stores, smart building, manufacturing plants, manufacturing line edge

Figure 2: Use cases for GPU as a Service span multiple industries and deployment locations.

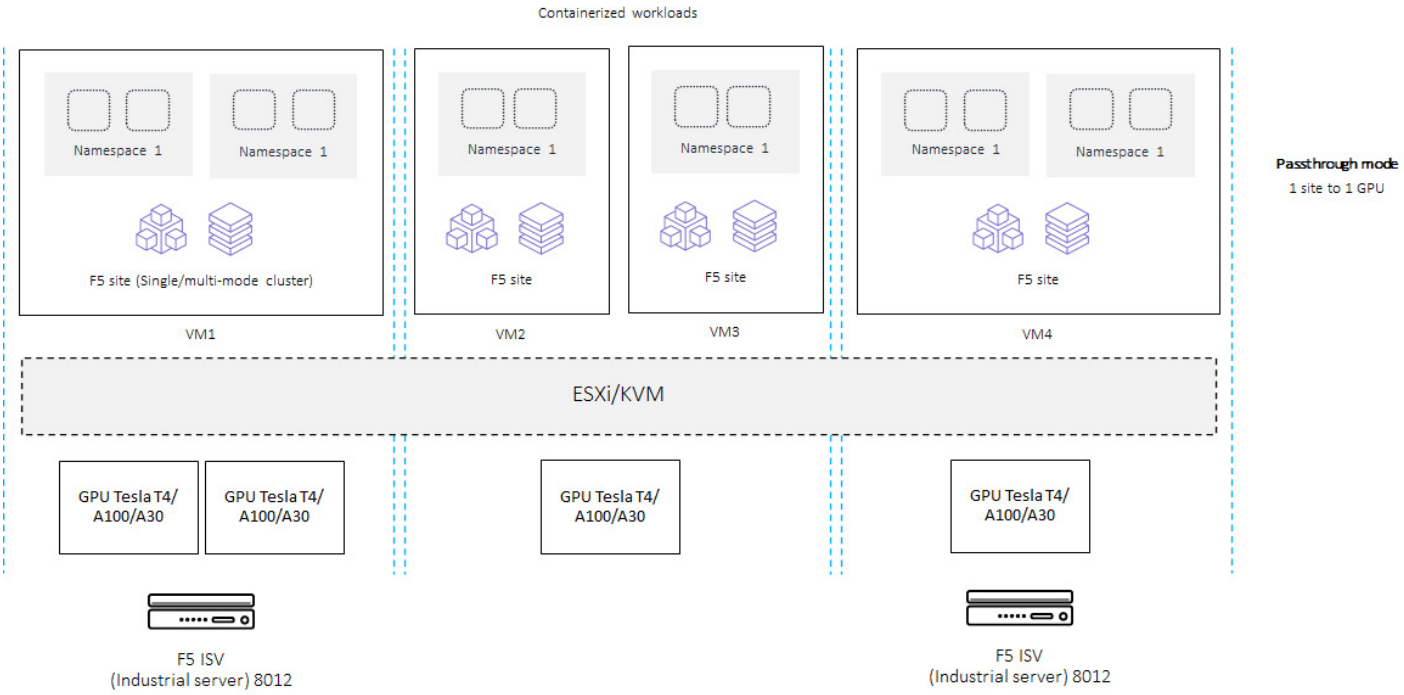

Figure 3: This architectural view depicts how GPU as a Service is deployed across the different environments.

How to enable GPUaaS on the F5 Distributed Cloud Platform

Distributed Cloud GPUaaS can be enabled can be enabled in the following ways:

- **F5 Distributed Cloud App Stack site configuration:** Distributed Cloud App Stack sites can be enabled with GPU on a per-site basis via configuration, provided that the site hardware includes a GPU.
- **Fleet configuration:** Multiple sites on F5 Distributed Cloud Services can be bulk enabled as a fleet. The sites are applied with the GPU capability, provided that the site hardware includes a GPU.

The GPU applications are then deployed using the F5 virtual Kubernetes (vK8s) associated with the same virtual site as that of the fleet of sites.

Use the instructions provided in our online guide to enable GPU capabilities on F5 sites using Distributed Cloud App Stack site or fleet configuration.

For more information

F5 Distributed Cloud Services are SaaS-based security, networking, and application management services that can be deployed across multi-cloud, on-premises, and edge locations.

GPU as a Service is available as part of Distributed Cloud App Stack. F5 offers both GPU as a Service in the cloud and GPU with hardware.

Interested? Contact sales@f5.com today.