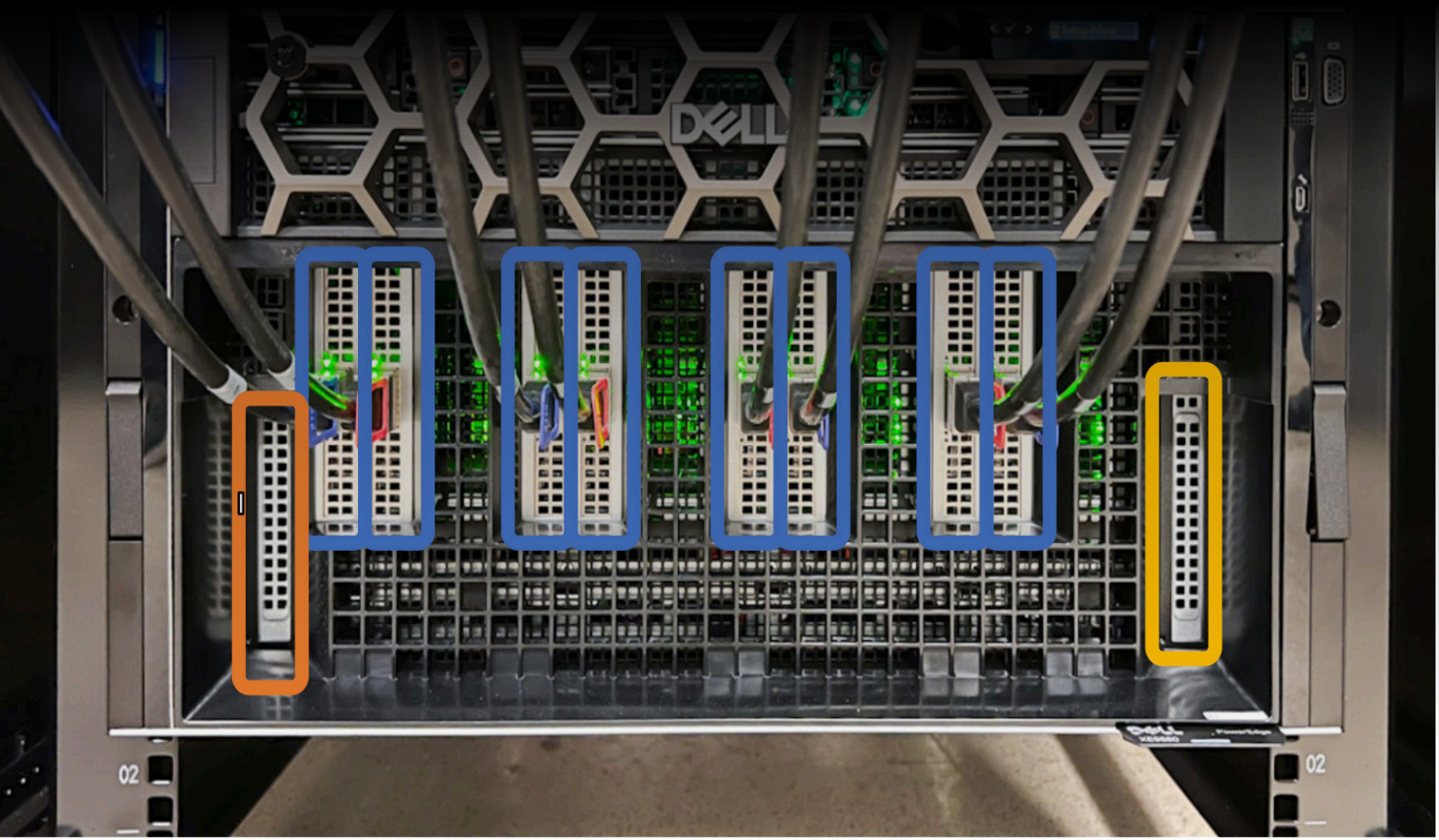
metalsoft®

SOLUTION BRIEF

Building GPU-as-a-Service Infrastructure with MetalSoft

Fully integrated, multi-tenant, multi-vendor server, switch and storage management for GenAI, GPU intensive analytics and big data.



Modern, GPU-as-a-Service with MetalSoft

Customer case study

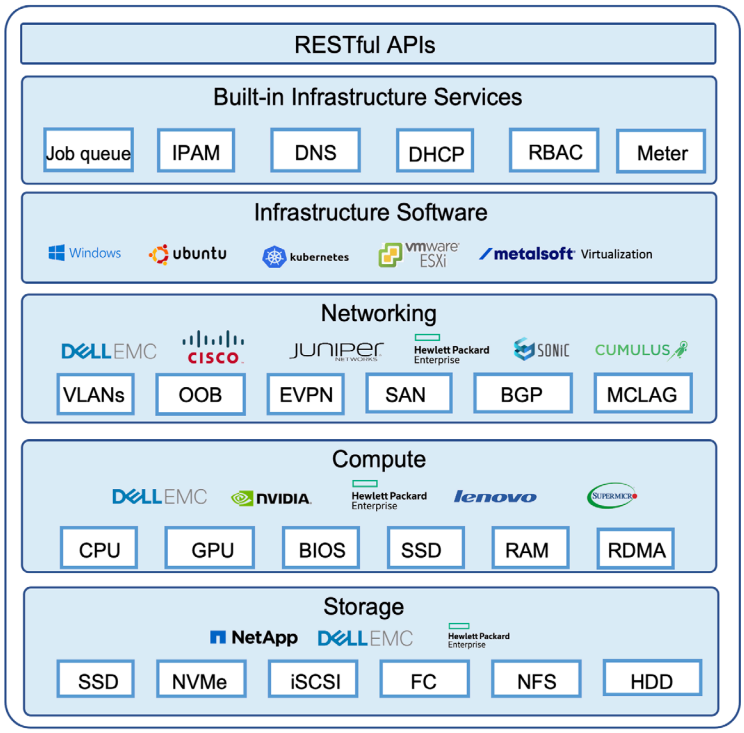
A leading US-based Telecom Service Provider is leveraging MetalSoft platform to provide GPU-compute-as-a-service from Edge locations to serve low-latency AI/ML use cases.

In 2023 the GPU-as-a-Service market size was valued at USD 3.16 billion. It is projected to grow to 25.53 billion by 2030, exhibiting a CAGR of 34.8% during the forecast period[1]

The Transformer model architecture has propelled the use of Large Language Models (LLMs) enabling numerous use cases like text summarization, Chatbots, translation, code assist and entity extraction. The growing adoption of machine learning and AI across various industries has increased the demand for robust computing resources.

GPU intensive infrastructures tend to be very expensive and thus many enterprises look for ways to maximize their investment by quickly allocating and de-allocating them as workloads begin and end. Many other companies rent these resources from general purpose public cloud providers or specialized GPU-as-a-Service providers.

MetalSoft is a software platform that enables fully automated, GPU-as-a-Service infrastructure. This allows end users to have a cloud like self-service-driven experience where they can consume GPU infrastructure in a matter of minutes. They can also use Terraform/ OpenTofu or APIs to consume these resources.



MetalSoft manages node discovery, GPU discovery, NVMe storage, CPUs generations etc, for multiple vendors across multiple data centers. It enables enterprises and service providers to build complex GPU-aware network architectures specialized for east-west traffic while also managing multi-tenancy and isolation for a heightened security posture.

[1]Fortune Business Insights - GPU as a Service Market -2023

Customer case study

An European Service Provider is leveraging MetalSoft platform to provide Big Data and HP BMaaS-services across multiple countries.

Benefits

GPU detection and management
MetalSoft detects GPUs and builds distinct server types based on type and count of GPUs in a host.

Server, network and switch management
Based on self-service requests MetalSoft performs all required configurations on all hardware.

Resource utilization metering
The system tracks hardware utilization to the second.

Enable Platform Engineering
Enable CI/CD integrations, Terraform and other forms of automated, self-service-driven consumption.

Deploy your software stack of choice
Use your platform of choice from simple PyTorch and TensorFlow to Spark on Kubernetes

Optimize your operations
Fully automatic life-cycle management, firmware management, secure deprovisioning, reduces the burden on IT.

MetalSoft is the perfect foundation for GPU services

Depending on your use case whether it is for Hyper-parameter search, large scale distributed training or Production inference, MetalSoft supports the unique requirements for each of them by providing following key capabilities:

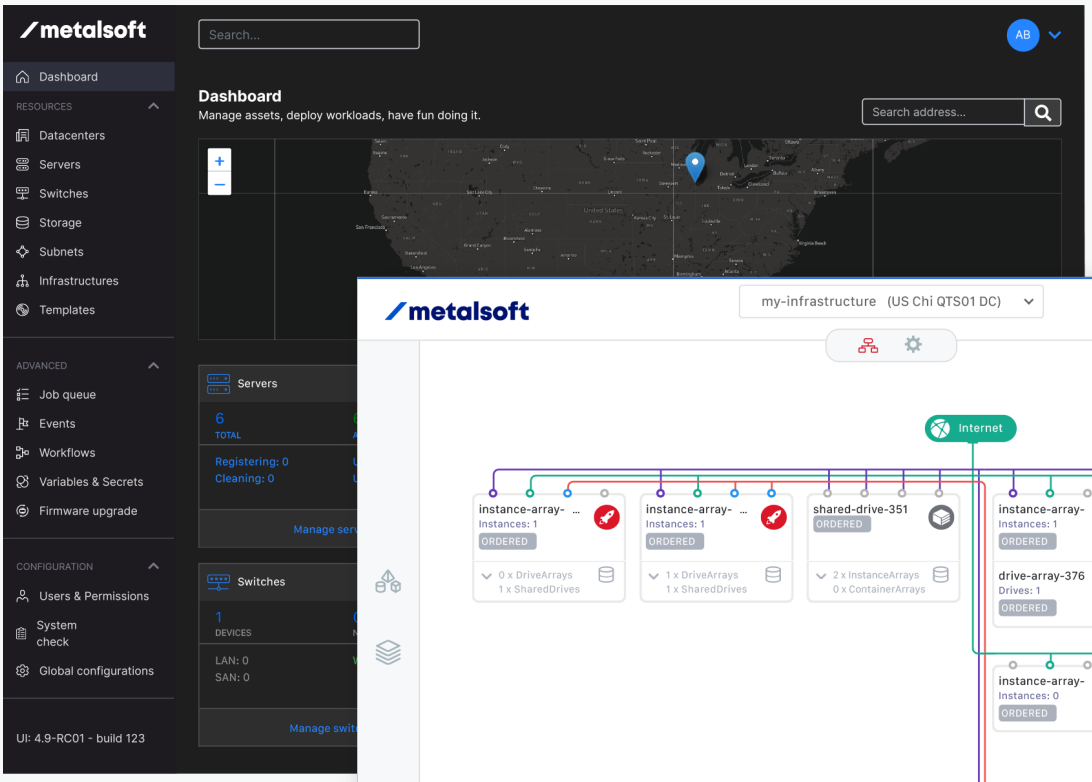
- Performs zero touch enrollment of servers
- Detects GPUs and creates specific server profiles
- Provides native multi-tenancy
- Supports any Hypervizor and Bare-metal
- Deploys operating systems
- Deploys tenant L2 and L3 networks
- Deploys storage volumes
- Performs secure cleanup when servers are returned the pool
- Provisions application clusters such as Kubernetes, VMWare VCF or native NN frameworks such as PyTorch
- Complete automated firmware management
- Performs on-demand metering and traffic accounting
- White-labeling capabilities
- Dedicated drag & drop self-service customer portal
- APIs and Terraform/OpenTofu support
- Pre-defined templates which have all configurations defined, or ability to create it manually, such as pre-installed NVidia CLI drivers.
- Ability to deploy endpoint with appropriate runtime and inference runtime can use CPUs or GPUs.
- Ability to deploy and manage ROCE-based Ethernet networks for high east-west traffic throughput.

Enable modern consumption capabilities with Intent-based automation

MetalSoft is "Intent based" which means the users or applications define what their intent is for the infrastructure and MetalSoft turns that into actual configurations on the equipment.

For Enterprises, this allows internal users and application developers to fully integrate the infrastructure into their CI/CD pipelines, enabling standardization and consumption agility improving Time-to-Value.

For Service Providers, these new capabilities open up new opportunities to extend your partner ecosystem with integration services, managed services, right-sizing and others helping your customers efficiently consume the GPU infrastructure.



MetalSoft Admin user Interface

MetalSoft end-user self-service user interface

Why MetalSoft

MetalSoft is the most comprehensive hardware orchestration solution on the market. It handles both servers, storage and more importantly network provisioning.

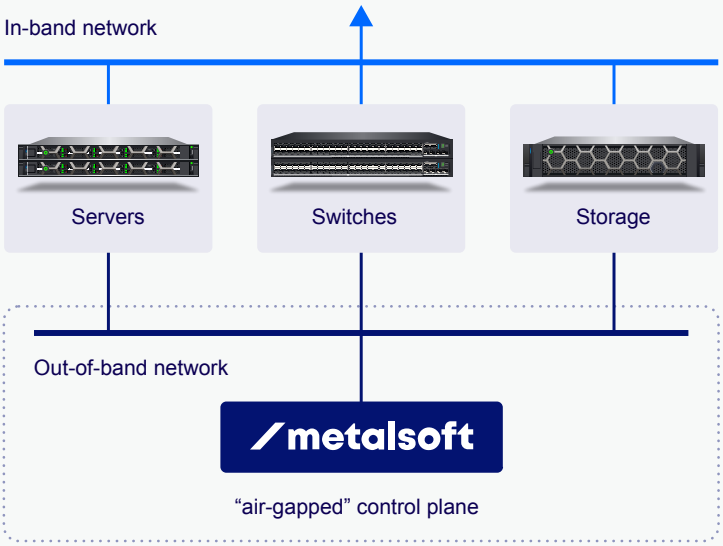
As our software was originally developed to run a HPC service provider, our team has many years of experience in building high performance and big data infrastructures for data analysis.

We also have a long history of operating our own data centers and all that experience is incorporated into our software.

MetalSoft also has presence across the globe in both APAC, EMEA and North America, offering 24/7 support and professional services.

Feature highlight:

MetalSoft is the **only** solution on the market, capable of operating hardware without the need for access to the in-band network. This greatly improves the security.



Evaluate or Purchase MetalSoft. Test drive MetalSoft today or contact us to find out more contact at: **sales@metalsoft.io** or visit **www.metalsoft.io**