SERVICE TEMPLATE

ZeroStack AI-as-a-Service

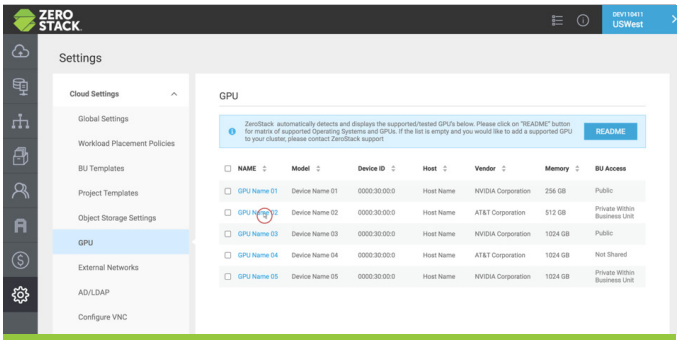
Artificial Intelligence and Machine Learning products and solutions are quickly becoming commonplace today and are shaping our experiences in computing like no other time in history. Interactive speech (e.g., Alexa, Google Home, etc.), Visual Search and recommendation engines are just a few of the consumer applications that are available today on our phones, websites and e-commerce platforms. **The impact of machine learning is getting broader** with enterprise applications in health sciences (e.g. Dr. Watson), finance, security, data centers and cyber surveillance.

These AI applications and solutions are now more viable than ever with the availability of modern machine learning and deep learning tools such as TensorFlow, Caffe, etc. and access to GPUs that are built specifically to perform parallel operations on large amounts of data, e.g., multiplying matrices of tens or hundreds of thousands of numbers. Processing large data sets through the same hypothesized algorithm for learning and for intelligent inference is a fairly common operation in machine learning and deep learning applications.

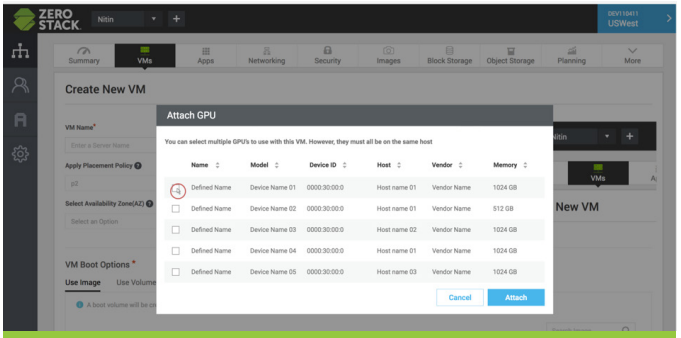
However, one significant challenge remains: deploying, configuring, and executing these complex tools and managing their interdependencies and versioning and compatibility with servers and GPUs. For example, in order to run TensorFlow, users need to make sure that they have the correct version of BIOS on their server, the compatible Windows or Linux drivers, and CUDA library for the specific GPU and server they want to run their AI workload on. **If any of these are not correctly configured with compatible versions, the AI application will not function correctly or will perform very poorly.**

ZeroStack provides single-click deployment of TensorFlow deep learning tool sets, taking care of all the OS and CUDA library dependencies. Furthermore, **users can enable GPU acceleration with dedicated access to multiple GPU resources** for an order-of-magnitude faster inference latency and user responsiveness.

ZeroStack's GPU capability gives customers powerful features to automatically detect GPUs and make them available for users to run their AI applications. In order to maximize utilization of this powerful resource, **cloud admins can configure, scale, and allow fine-grained access control of GPU resources to end users.**



ZeroStack automatically detects and displays supported GPUs available in the cluster.



Project member users can enable GPU acceleration and choose one or more GPUs to attach to their workload.

ZEROSTACK AI-AS-A-SERVICE:
Automated and Intuitive

Self-service ready with single-click deployment of TensorFlow tool set from the App Store	ZeroStack AI-as-a-service provides single click deployment of TensorFlow and other AI applications. The deployment templates enable users to select GPU acceleration and attach GPUs from the available list to their AI application. The automated deployment and configuration ensures that the correct Windows, Linux, and CUDA libraries are used for the GPUs that are selected for the workload. Users can just as easily detach the GPUs once they are done using them.
Automated detection of supported GPU cards available in the cluster	ZeroStack automatically performs PCI scans to retrieve GPU inventory in the system.
Built-in governance and fine-grained access control of GPU resources	Cloud administrators can control which business units have access to which GPU resources.
Built-in production operational capabilities	For field maintenance, administrators can add and remove GPUs on existing servers by following host evacuation workflow best practices. The ZeroStack cluster can be scaled up on demand by adding new physical nodes to the cluster with GPU resources.

ZEROSTACK AI-AS-A-SERVICE:
Use Cases

MSPs who want to provide an AI service to their end customers

Single-click, template-based automated deployment of TensorFlow including the CUDA and driver dependencies built-in as well as cost management of all resource usage make ZeroStack AI-as-a-service an attractive new revenue source for MSPs.

Provide GPUs on demand for AI applications

Multiple GPU capabilities, dedicated/full GPU access, and self-service make this attractive for companies that want to provide on-demand GPU access to their users.

Agile Operations

ZeroStack's self-healing features, software-defined networking and storage as well as seamless upgrades and scalability can now be used with GPU applications.

Hardware Requirements

ZeroStack recommends the following hardware specs for servers hosting the GPU cards.

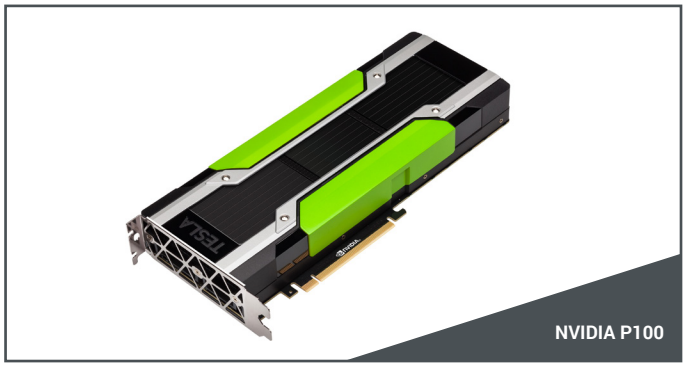
1. CPU: Intel Xeon E2630 v4 – 10 core processor w/ virtualization (VT-x) and IOMMU (VT-d) support or a similar AMD CPU with AMD-V and AMD-Vi support
2. RAM: Minimum 80 GB DDR4 2133 MHz (128 Gb recommended for Deep learning apps)
3. Motherboard: PCI 3.0 compliant motherboard, check for GPU compatibility and BIOS options
4. Storage: At least 2 TB HDD (7200 RPM) + 1TB SSD

Software Requirements

- Windows 2012, 2014 server Operating Systems
- RHEL, CentOS, Ubuntu 16.04 OS
- CUDA libraries

Supported GPUs

NVIDIA TESLA GPU Cards



NVIDIA P100

Tesla P100 delivers superior performance for HPC and hyperscale workloads. Based on PASCAL architecture, it supports more than 21 teraFLOPS of 16-bit floating-point (FP16) performance.

NVIDIA GTX 1080 Ti

The GeForce® GTX 1080 Ti is NVIDIA's new flagship gaming GPU, based on the NVIDIA Pascal™ architecture. It's equipped with next-gen 11 Gbps GDDR5X memory and a massive 11 GB frame buffer.

About ZeroStack

ZeroStack Intelligent Cloud Platform is a fully-integrated cloud solution that delivers the simplicity and agility of a public cloud, along with the performance and control of a private cloud, at a fraction of the cost.