SERVICE TEMPLATE

ZeroStack GPU-as-a-Service

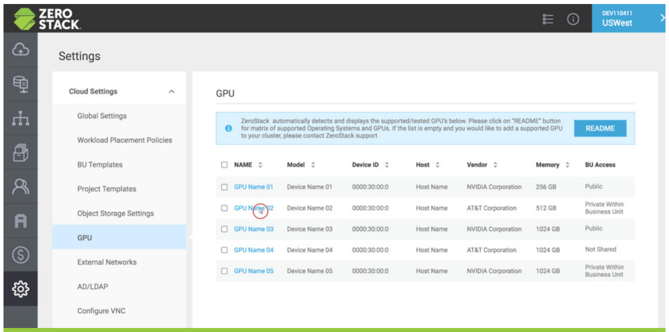
Deliver GPU-Accelerated AI and Machine Learning Workloads on Demand

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.

General-purpose CPUs cannot deliver the user responsiveness and inference latency required by complex deep learning and AI workloads. That's because – unlike GPUs built for this purpose – general purpose CPUs are not designed to rapidly 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.

ZeroStack's GPU-as-a-Service capability gives customers powerful features to automatically detect GPUs and make them available in the ZeroStack environment. 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.

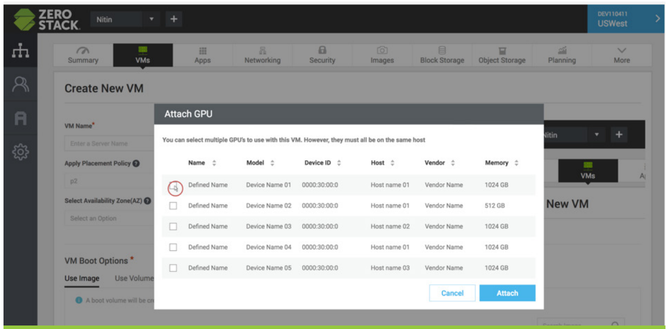
Users can enable GPU acceleration, deploy new machine learning and deep learning workloads with tools such as TensorFlow, Caffe, etc., and provide the apps dedicated access to multiple GPU resources for an order of magnitude, faster inference latency and user responsiveness.



ZeroStack automatically detects and displays the supported/tested GPU's below. Please click on "README" button for matrix of supported Operating Systems and GPUs. If the list is empty and you would like to add a supported GPU to your cluster, please contact ZeroStack support

□	NAME	Model	Device ID	Host	Vendor	Memory	BU Access
□	GPU Name 01	Device Name 01	0000:30:00:0	Host Name	NVIDIA Corporation	256 GB	Public
□	GPU Name 02	Device Name 02	0000:30:00:0	Host Name	AT&T Corporation	512 GB	Private Within Business Unit
□	GPU Name 03	Device Name 03	0000:30:00:0	Host Name	NVIDIA Corporation	1024 GB	Public
□	GPU Name 04	Device Name 04	0000:30:00:0	Host Name	AT&T Corporation	1024 GB	Not Shared
□	GPU Name 05	Device Name 05	0000:30:00:0	Host Name	NVIDIA Corporation	1024 GB	Private Within Business Unit

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



You can select multiple GPU's to use with this VM. However, they must all be on the same host

	Name	Model	Device ID	Host	Vendor	Memory
□	Defined Name	Device Name 01	0000:30:00:0	Host name 01	Vendor Name	1024 GB
□	Defined Name	Device Name 02	0000:30:00:0	Host name 01	Vendor Name	512 GB
□	Defined Name	Device Name 03	0000:30:00:0	Host name 02	Vendor Name	1024 GB
□	Defined Name	Device Name 04	0000:30:00:0	Host name 01	Vendor Name	1024 GB
□	Defined Name	Device Name 05	0000:30:00:0	Host name 03	Vendor Name	1024 GB

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

ZEROSTACK GPU-AS-A-SERVICE:
Powerful and Simple

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.
Self-service Ready	ZeroStack GPU-as-a-service is easy to use once access is provided by the cloud admins. Users enable GPU acceleration and attach GPUs from the available list to their workloads including Windows, Linux, and CUDA library support. Users can just as easily detach the GPUs once they are done using them.
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 GPU-AS-A-SERVICE:
Use Cases

MSPs who want to provide a GPU service to their end customers	Deep learning and Machine Learning applications	Agile Operations
Automated detection of GPU resources, self-service option for end users, controlling customer access to GPUs, as well as cost management make ZeroStack GPU-as-a-service an attractive new revenue source for MSPs.	Multiple GPU capabilities, dedicated/full GPU access, and self-service make this attractive for enterprises and universities who want to provide on-demand access to their users.	ZeroStack's self-healing features, software-defined networking and storage as well as seamless upgrades and scalability can now be used with GPU applications as well.

Hardware Requirements

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

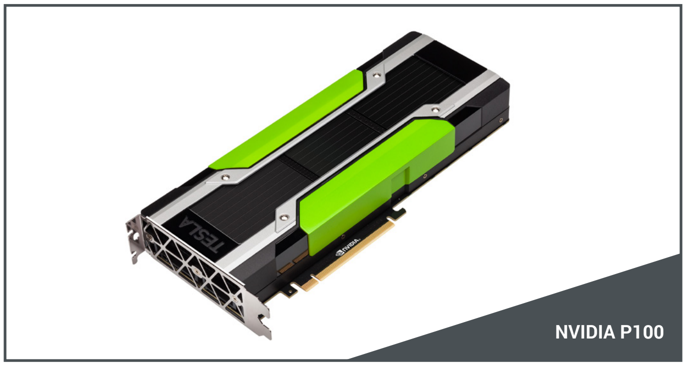
1. CPU: Intel Xeon E2630 v4 – 10 core processor w/ virtualization (VT-x) and IOMMU (VT-d) support or a similar AMD CPU with AMD-V and AMD-Vi support
2. RAM: Minimum 80 GB DDR4 2133 MHz (128 Gb recommended for Deep learning apps)
3. Motherboard: PCI 3.0 compliant motherboard, check for GPU compatibility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

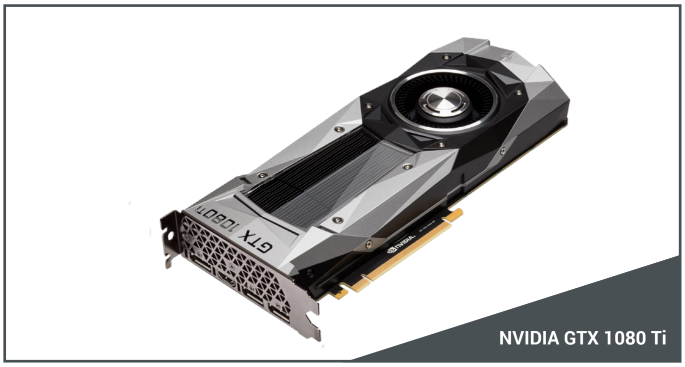
Supported GPUs

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.