



Hewlett Packard
Enterprise

vmware®



Reference Configuration

HPE Reference Configuration for Enterprise AI Infrastructure on HPE ProLiant DL380 GEN10 server

Flexible and Democratized Enterprise AI infrastructure with
HPE ProLiant DL380 Gen10 Servers Using NVIDIA A100 Tensor Core
GPU

CONTENTS

Executive summary.....	3
Introduction.....	4
Solution overview.....	5
Solution components.....	7
Hardware.....	7
Software components.....	10
High-level solution workflow.....	14
Solution configuration guidance.....	15
BIOS configuration.....	15
ESXi Host GPU settings.....	16
NVIDIA A100 configuration on the ESXi host.....	16
Assign vGPU profile to a VM.....	17
VMware vSphere Cluster configuration.....	18
Ubuntu VM configuration.....	19
AI/ML workload.....	19
Summary.....	20
Resources and additional links.....	21



EXECUTIVE SUMMARY

Artificial intelligence is powering businesses across all industries and verticals becoming an inevitable general-purpose need like electricity. We are on the cusp of the fourth industrial revolution where newer experiences are made available bridging the human–cyber interface. An explosive amount of data is being generated and consumed under hybrid cloud consumption models yielding superior speed, cost, and agility to open up newer and better ways of doing business. More than 75% of organizations are embracing the hybrid cloud to exploit the cloud-like ease of use while maintaining control of their on-premises infrastructure. AI-powered solutions are an essential avenue to such businesses in analyzing continuously arriving data (Data pipelines) and interpreting them using task-specific AI models (Machine Learning/ Natural Language Processing pipelines) to gain insights and power real-time decision making. These pipelines are operationalized along with traditional software and solutions.

Real-time inference on the input data by such trained models requires robust infrastructure to meet application latency and bandwidth requirements, as some models use millions to billions of parameters. The need for more accurate models and faster inference is driving rapid growth in GPU accelerated computing. GPU accelerated servers like the HPE ProLiant DL380 Gen10 are exceptionally well-suited to satisfy these requirements. The NVIDIA-Certified HPE ProLiant DL380 server can be configured to accelerate AI training and inference – while still providing the resources necessary to address traditional IT workloads.

VMware® and NVIDIA® have partnered together to unlock the power of AI for every business, by delivering an end-to-end enterprise platform optimized for AI workloads. This integrated platform delivers best-in-class AI software, the NVIDIA Enterprise AI suite, optimized and exclusively certified to run on the industry's leading virtualization platform, VMware vSphere®, with industry-leading accelerated servers such as the HPE ProLiant DL380 server that have been NVIDIA-Certified. This platform accelerates the speed at which developers can build AI and high-performance data analytics for their business, enable organizations to scale modern workloads on the same VMware vSphere infrastructure they've already invested in, and deliver enterprise-class manageability, security, and availability.

NVIDIA AI Enterprise is an end-to-end, cloud-native suite of AI and data analytics software - optimized, certified, and supported by NVIDIA to run on VMware vSphere with NVIDIA-Certified Systems. AI Enterprise offers key enabling technologies from NVIDIA for rapid deployment, management, and scaling of AI workloads in the modern hybrid cloud. Product announcement of the NVIDIA AI Enterprise platform on VMware vSphere 7 Update 2 can be found [here](#).

VMware® and NVIDIA® partnership maximizes GPU performance with VMware's vSphere® 7 Update 2 and later - delivering near bare-metal performance. This release enables support for NVIDIA's A100 Tensor Core GPUs with Multi-Instance GPUs (MIG) capability. MIG is a method of fractionalizing the NVIDIA A100 Tensor Core GPU into as many as seven instances, each fully isolated with its own high-bandwidth memory, cache, and compute cores.

Given the accelerated momentum of AI becoming a mainstream technology, enterprises need an easy-to-adopt and easy-to-integrate hardware-software stack to move faster. It is timely that familiar and widely adopted VMware virtualization technology flexibly combines its power with GPU accelerators, allowing GPU profiles to be assigned to a virtual machine via the VMware vCenter® while making vSphere vMotion®, DRS, and HA available to the vGPU backed VMs.

Given that GPU accelerators can have a sizable impact on the cost of a server node, it is imperative to make the right GPU selection and ensure high utilization of these resources to reduce the Total Cost of Ownership (TCO). Data Scientists and AI developers require different quantum of GPU resources at different points in times like training, testing, and inference.

NVIDIA MIG capability and VMware's vSphere 7 Update 2 allows IT Admins to size each VM with the appropriate GPU resources for the job, delivering the performance needed and optimizing utilization.

Target audience: The target audience for this Reference Configuration are Chief Information Officers (CIOs), Chief Technology Officers (CTOs), IT decision-makers who have a directive to invest in the AI/ML container workloads using GPUs as well as data scientists, software developers, and architects interested in accelerating AI workflows. The intent is to assist in defining and implementing accelerated AI/ML workloads on HPE Rack-based DL380 Gen10 servers with NVIDIA A100 GPUs extending the reach of these compute intense resources to every user driving the AI/ML productivity.

Document purpose: This Reference Configuration provides a deployment and configuration overview of the HPE ProLiant DL380 Gen10 server with NVIDIA A100 Tensor Core GPU and VMware ESXi 7 Update 2 and later for AI/ML workloads. Readers can use this document to achieve the following goals:



- Learn the configuration overview of the NVIDIA A100 GPU on HPE ProLiant DL380 Gen10 server for AI workloads and use Multi-Instance GPU to deliver up to 7 GPU instances and raise utilization rates.
- Delivering VMware vSphere based virtualized environment for enabling productivity, performance, and flexibility for AI/ML workloads.
- Accelerate AI/ML development and deployment with GPU-accelerated tools and frameworks from the NVIDIA AI Enterprise Suite.

This Reference Configuration describes solution testing performed in June 2021.

INTRODUCTION

Artificial Intelligence (AI) technologies are a crucial aspect of the digital transformation of enterprises. The emergence of AI applications resulted in a paradigm shift of data also being a fundamental component in building applications in addition to code. The enterprise data is not just meant to be acquired, analyzed, and acted upon anymore. AI applications are built using both data and code. Furthermore, this fundamental shift has driven revolutionary changes in development, deployment, and operational infrastructure. Despite the technological advancement and growing recognition of the value that AI can bring to businesses, complexity in deployments and effective operations are major hindrances in adopting AI in enterprises.

Evolution due to AI: For AI applications, the industry organized itself around the data by establishing Data Pipelines and DevOps for management. The abstractions of Continuous Integration/Continuous Deployment (CI/CD) evolved to also include Continuous Training (CT). Successful AI applications go through multiple painful iterations of preparing the data, refining the AI model, retraining the model, deploying, operating, and observing. The nature of the data and the focus of the AI application can change over time resulting in inaccuracies called Model Drift and Concept Drift. Even world-class intelligent algorithms trained on a golden dataset experience these unavoidable drifts, making rework necessary. If enterprises do not have a development and deployment environment that is resilient, one can easily get lost in the technical debt, evidenced by 80% of models that haven't seen the light of the day.

Complexity in AI: Training, testing, and inference are the core tasks in developing and running an AI application. Typically, the training is very resource-intensive and when the cost is prohibitive, these tasks become frustratingly time-consuming. Under these conditions, a small change in the dataset may result in a very large retraining exercise.

Acceleration: GPU acceleration for AI workloads maximize performance, productivity, and ROI. Many of the AI frameworks natively make use of the GPU on the system and significantly reduce training time and inference latency.

Underutilization: These GPU accelerated systems with the specialized Compiler toolchain ecosystem could become an economical challenge if an enterprise wants to use them at scale. There are times when core AI tasks may starve while there is an abundance of underutilized GPU-equipped servers. History suggests that server virtualization emerged due to very similar constraints with CPU-based computation. AI workloads in virtualized GPUs are gaining traction as there is a strong desire to improve GPU utilization and Return of Investment (ROI).

The key challenges enterprises face to make AI readily available to their customers are:

- Affordable, efficient, and flexible enterprise-grade AI infrastructure at scale.
- Slow procurement, deployment, and availability of enterprise-ready AI infrastructure.
- Lack of support from traditional IT operations to data scientists, ML practitioners.

Democratizing AI: The point we want to drive home here is that an affordable and efficient AI infrastructure is a must for enterprises to meet the need for explosive growth in AI application development and deployment. Hewlett Packard Enterprise/VMware/NVIDIA have come together to create avenues to build cost effective-high performance infrastructure with flexibility built-in. Their complementing core capabilities lead to democratizing the hardware and software capabilities in a way that more users within the enterprise can now access them with ease.

The approach Hewlett Packard Enterprise, VMware, and NVIDIA together are taking to make such an AI infrastructure is through:

1. HPE ProLiant DL380 Gen10 server, the industry's most trusted compute platform with world-class performance, expandability, and simplified manageability with a wide range of operating environments.
2. NVIDIA A100 family of GPU based on Ampere architecture that accelerates AI workloads.
3. VMware's integration of A100 GPU with its virtualization and MIG capabilities both at the hypervisor level and virtual machine level to make them available as part of core virtualization (VM and Containers).



4. NVIDIA AI Enterprise software suite for AI workload acceleration on VMware vSphere 7 Update 2 and later.

This powerful integration lets IT Administrators readily use familiar VMware virtualization products and provides data scientists with the self-serve capability to efficiently use the compute resources resulting in lower complexity and a higher ROI. In addition pre-built containers significantly reduce complexity and improve ease of use, enabling data scientists to work with images, text, and audio using AI, ML, DL, and NLP toolchains and datasets.

This Reference Configuration highlights the design and deployment guidelines for a virtualized AI platform built on best-in-class products from Hewlett Packard Enterprise, VMware, and NVIDIA. The solution combines the industry's most secure server, the HPE ProLiant DL380, spatial partitioning based NVIDIA multi-instance GPUs on NVIDIA A100 Tensor Core GPU, and Industry's widely adopted virtualized platform VMware vSphere. This solution aims to simplify the process of deploying the platform that can help accelerate AI/ML workloads to enable business agility and faster innovation.

SOLUTION OVERVIEW

The solution as described in this Reference Configuration is based on HPE ProLiant DL380 Gen10 server and VMware vSphere 7 Update 2 and later in a two-node VMware vSphere cluster configuration. Each HPE ProLiant DL380 Gen10 server has one NVIDIA A100 40GB PCIe GPU Accelerator card on the PCI-E slot 2 of the Primary Riser card.

The vSphere cluster is HA and DRS enabled with shared storage through HPE Nimble AF20 storage connected over the iSCSI network. The two HPE ProLiant DL380 Gen10 Servers are wired to two Aruba 8325 ToR switches connected through VSX links. Two iSCSI paths are established to Nimble to achieve iSCSI multipathing. iSCSI multipathing provides HA and failover to the alternate path.

VMware vSphere distributed switch (vDS) is set up with port groups for management, production, and iSCSI storage. The uplinks from each server are through HPE Ethernet 10/25GB 2-port 640FLR-SFP28 Adapter.

Infrastructure services such as Active Directory, DNS server, NTP server, NVIDIA license server are hosted on a different server and have network connectivity to the solution that is set up.



Figure 1 shows a high-level solution architecture diagram of the solution.

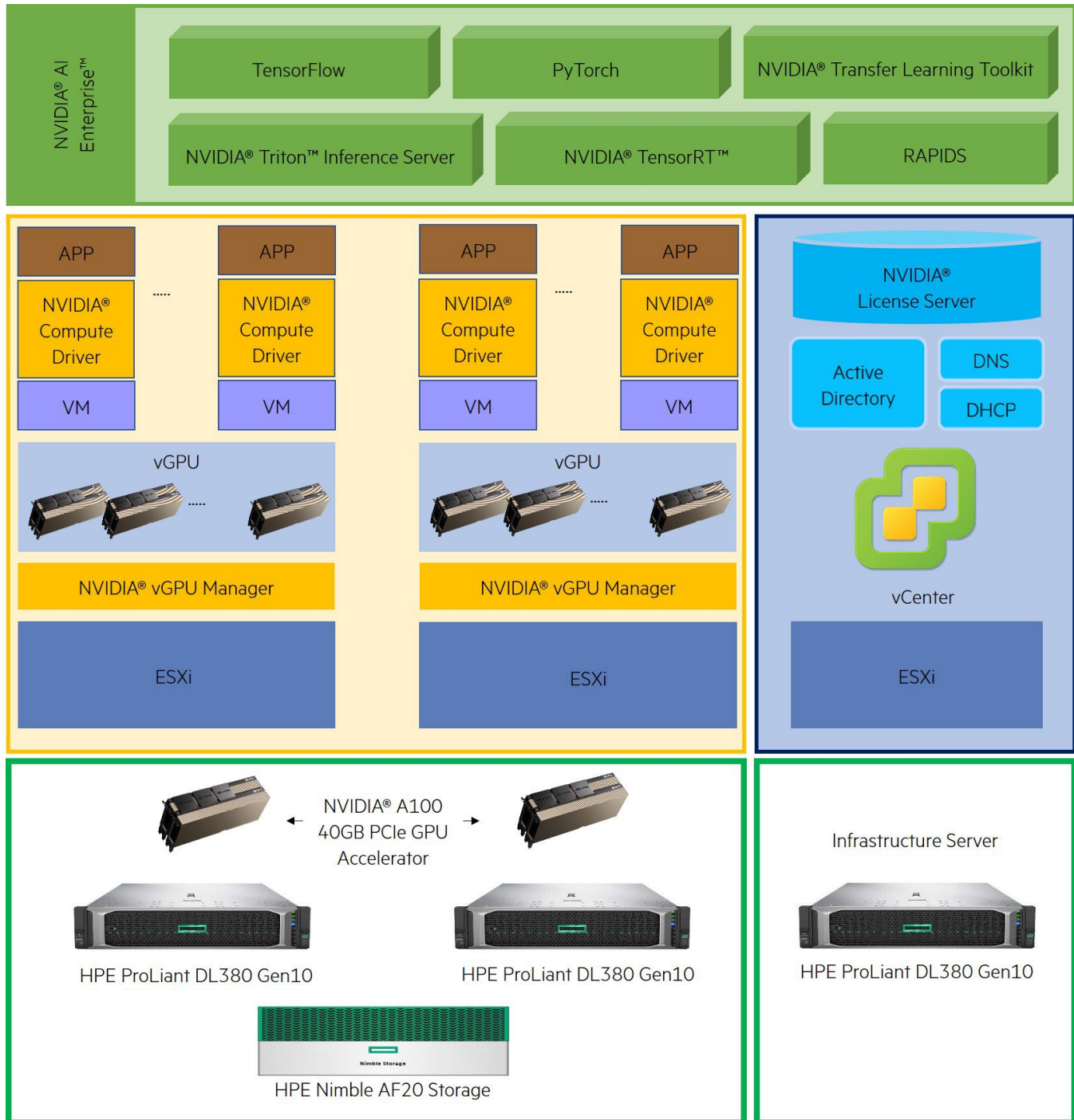


FIGURE 1. High-level solution architecture



SOLUTION COMPONENTS

This section describes the key hardware and software components used in the solution design.

Hardware

HPE ProLiant DL380 Gen10 server

The HPE ProLiant DL380 Gen10 server delivers the latest in security, performance, and expandability and offers the ultimate flexibility for end-user computing workloads. The HPE ProLiant DL380 Gen10 server supports up to two NVIDIA A100 GPUs for workload acceleration and offers the ultimate flexibility for end-user computing workloads. The combination of the HPE ProLiant DL380 Gen10 server and NVIDIA A100 GPU creates a powerful platform for AI and container workloads.

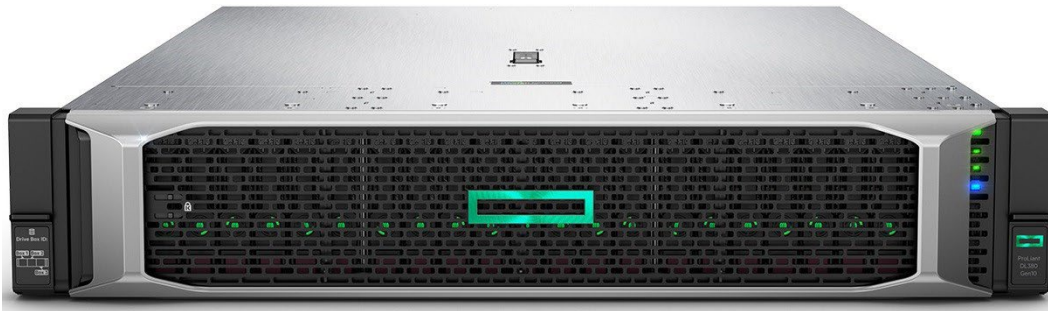


FIGURE 2. HPE ProLiant DL380 Gen10 server

Table 1 describes the HPE ProLiant DL380 Gen10 server hardware configuration.

TABLE 1. HPE ProLiant DL380 Gen10 server components

Component	Quantity	Description
Processor	2	Intel® Xeon®-Gold 5220 (2.20 GHz/18-core/125 W)
Memory	12	HPE 32GB (1x32GB) DDR4-2933 CAS-21-21-21 Registered Smart Memory Kit (384GB total)
GPU	1	NVIDIA A100 40GB GPU
Network Adapter	1	HPE Eth 10/25Gb 2p 640FLR-SFP28 Adapter
Storage Controller	1	HPE Smart Array P816i-a SR Gen10 controller (16 Internal Lanes/4GB Cache) 12G SAS Modular Controller
Disks	14	12 x HPE 800GB SSD 12G Read Mixed use SFF SSD 2 x HPE 3.75GB NVMe X4 lanes write intensive SFF SSD



NVIDIA A100 40GB PCIe GPU Accelerator

NVIDIA A100 Tensor Core GPU is based on the NVIDIA Ampere architecture and accelerates compute workloads such as artificial intelligence (AI), data analytics, and HPC in the data center. Multi-Instance GPU (MIG) expands the performance and value of each NVIDIA A100 Tensor Core GPU. MIG can partition the NVIDIA A100 GPU into as many as seven instances, each fully isolated with dedicated high-bandwidth memory, cache, and compute cores. This enables IT operators to maximize GPU utilization while ensuring predictable performance with quality of service based on the right sizing for the workload and fault isolation. Refer to [NVIDIA A100 PCIe 40GB vGPU](#) types for supported vGPU modes.



FIGURE 3. NVIDIA A100 GPU

Table 3 provides information about NVIDIA A100 GPU specifications.

TABLE 3. NVIDIA A100 GPU specifications

Specifications	Description
GPU Architecture	NVIDIA Ampere
Form Factor	PCIe
Multi-instance GPU(MIG)	Up to 7 GPU instance
Memory size	40GB HBM2
Peak Memory bandwidth	Up to 1555GB/ss
Single-Precision performance	FP32: 19.5 TFLOPS Tensor Float 32 (TF32): 156 TFLOPS 312 TFLOPS*
Double-Precision Performance	FP64: 9.7 TFLOPS FP64 Tensor Core: 19.5 TFLOP
Half-Precision Performance	312 TFLOPS 624 TFLOPS*
Integer Performance	INT8: 624 TOPS 1,248 TOPS* INT4: 1,248 TOPS 2,496 TOPS*

* Structural sparsity enabled



HPE Nimble Storage All Flash AF20 Array

The HPE Nimble Storage solution provides a complete data storage architecture that includes primary storage, intelligent caching, instant application-aware backups, and replication. HPE Nimble Storage All Flash AF20 Array is a predictive flash storage solution that is simple to deploy and easy to manage. With a single Dual-Port 10GbaseT add-on card, the HPE Nimble storage provides a cost-efficient shared iSCSI storage solution for the VMware cluster. HPE Nimble Storage arrays present iSCSI or Fibre Channel (FC) target volumes to VMware hosts and iSCSI target volumes to guest virtual machines (VMs). The volumes created on HPE Nimble Storage arrays are highly optimized for VMs. They offer inline compression, inline deduplication, thin provisioning, snapshot backups, zero-copy cloning, and WAN-optimized replication.

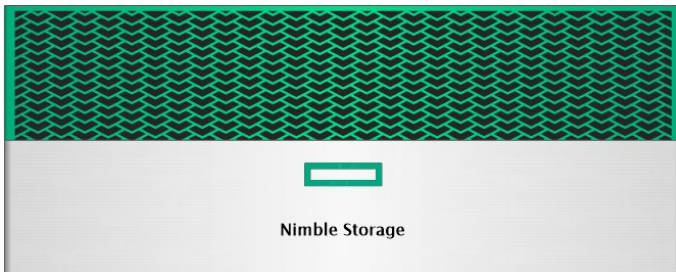


FIGURE 4. HPE Nimble Storage All Flash AF20 Array

With its built-in availability, scalability, manageability, and business continuity, vSphere provides a solid foundation for a virtualization environment for enterprise applications. To take advantage of vSphere features such as vSphere high availability (HA), Storage Distributed Resource Scheduler (DRS), vMotion, and Storage vMotion, shared storage is a requirement. In this solution, HPE Nimble Storage AF20 is used to present iSCSI target volumes to VMware hosts and guest virtual machines (VMs) to provide persistent, block storage to provide persistent volume for applications.

Aruba Switches

Aruba CX 6300M and 8325 switch series provide management and data center network connectivity to the servers and storage in this solution. The Aruba CX 6300 Switch Series is a modern, flexible, and intelligent family of stackable switches ideal for enterprise network access, aggregation, core, top of rack (ToR), and out-of-band-management (OOBM) data center deployments. The 8325 series includes industry-leading line rate ports 1/10/25GbE (SFP/SFP+/SFP28) and 40/100GbE (QSFP+/QSFP28) with connectivity in a compact 1U form factor. These switches offer a fantastic investment for customers wanting to migrate from older 1GbE/10GbE to faster 25GbE, or 10GbE/40GbE to 100GbE ports.

Aruba CX 6300M is used as an OOBM switch and Aruba 8325 switch is used as a ToR switch in this solution. The ToR switch provides connectivity from the virtual machines to the NVIDIA NGC repository to pull required container images.

Figure 5 shows the front view of the Aruba 6300M Switch.

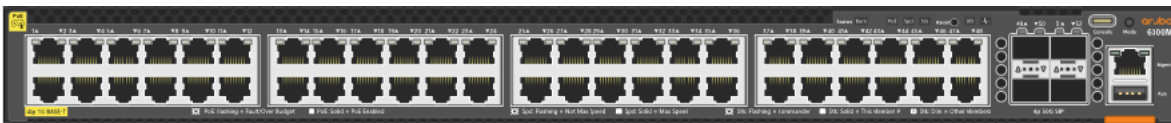


FIGURE 5. Aruba 6300M Switch



Figure 6 shows the front view of the Aruba 8325 32Y8C Switch.

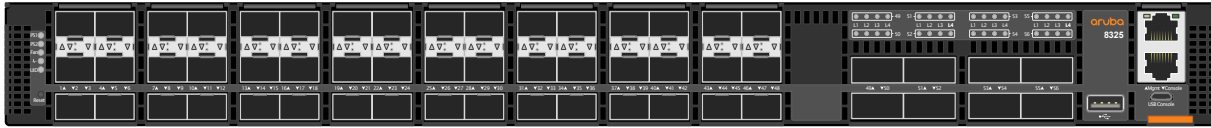


FIGURE 6. Aruba 8325 32Y8C Switch

Software components

Table 4 shows various software components and versions used in the solution.

TABLE 4. Software components and versions

Specifications	Description
VMware vSphere 7.0 U2a	VMware-ESXi-7.0.2-17867351-HPE-702.0.0.10.7.0.52-May2021.iso
VMware vCenter 7.0 U2a	VMware-VCSA-all-7.0.2-17920168.iso
HPE VMware Upgrade Pack (VUP) for ProLiant Note: Used for systems software and firmware updates	1.4.2 (P45553_001_VUP142-SPP-VUP142.2021_0427.26.ISO)
NVIDIA Software Bundle	12.1
NVIDIA A100 Driver Package	NVIDIA-GRID-vSphere-7.0-460.32.04-460.32.03-461.33
ESXi Host Driver	460.32.04 - NVIDIA_bootbank_NVIDIA-VMware_ESXi_7.0_Host_Driver_460.32.04-10EM.700.0.0.15525992.vib
Linux Driver	460.32.03 - NVIDIA-Linux-x86_64-460.32.03-grid.run
NVIDIA License Server	2020.05.0.28406365
Ubuntu	ubuntu-20.04.1.0-desktop-amd64.iso
Kernel	5.4.0-42

NOTE

The versions listed in the above table are for the early access version of NVIDIA AI Enterprise and are subject to change with general availability.

VMware vSphere 7 Update 2

VMware vSphere 7 Update 2 and later provides key enhancements to bring AI to the on-premises data center:

- The NVIDIA AI Enterprise software suite of software is certified on vSphere 7 Update 2 and later. NVIDIA AI Enterprise software suite needs to be purchased separately from NVIDIA and is not part of vSphere.
- Support for the latest generation of GPUs from NVIDIA based on the NVIDIA Ampere architecture, NVIDIA A100 GPU with support for Multi-Instance GPUs.
- Optimizations for device-to-device peer communication over the PCIe bus, enabling increased performance with NVIDIA's GPUDirect RDMA.

VMware Sphere 7 Update 2 and later supports the NVIDIA A100 GPU in traditional time-sliced vGPU mode and the new MIG mode. This solution covers MIG-backed vGPU mode and the ease of managing virtual machines with MIG resources via vSphere Client. Organizations can also choose to use time-sliced virtual GPUs if they wish. For more information on time-sliced vGPU, see <https://docs.nvidia.com/grid/latest/grid-vgpu-user-guide/index.html>.

MIG mode on the GPU is supported at the vSphere host level. When MIG is enabled, a specific MIG-related set of vGPU profiles are available to be allocated to virtual machines. From the vSphere Client, the administrator can choose one vGPU profile to be allocated from the available



options. The vGPU profile chosen in the vSphere Client maps directly to a GPU instance encapsulating a specific combination of compute and memory slices.



NVIDIA AI Enterprise Suite

NVIDIA® AI Enterprise is a software suite that enables AI workload acceleration on VMware vSphere ESXi 7 Update 2 and later hypervisor. The software stack is optimized, certified, and supported by NVIDIA® to run on VMware vSphere. It includes infrastructure software components and data science and AI frameworks and tools. Containerized software can be run directly with a tool such as Docker. NVIDIA AI Enterprise is currently available for early access. For more information, refer to <https://www.nvidia.com/en-us/data-center/products/ai-enterprise-suite/>.

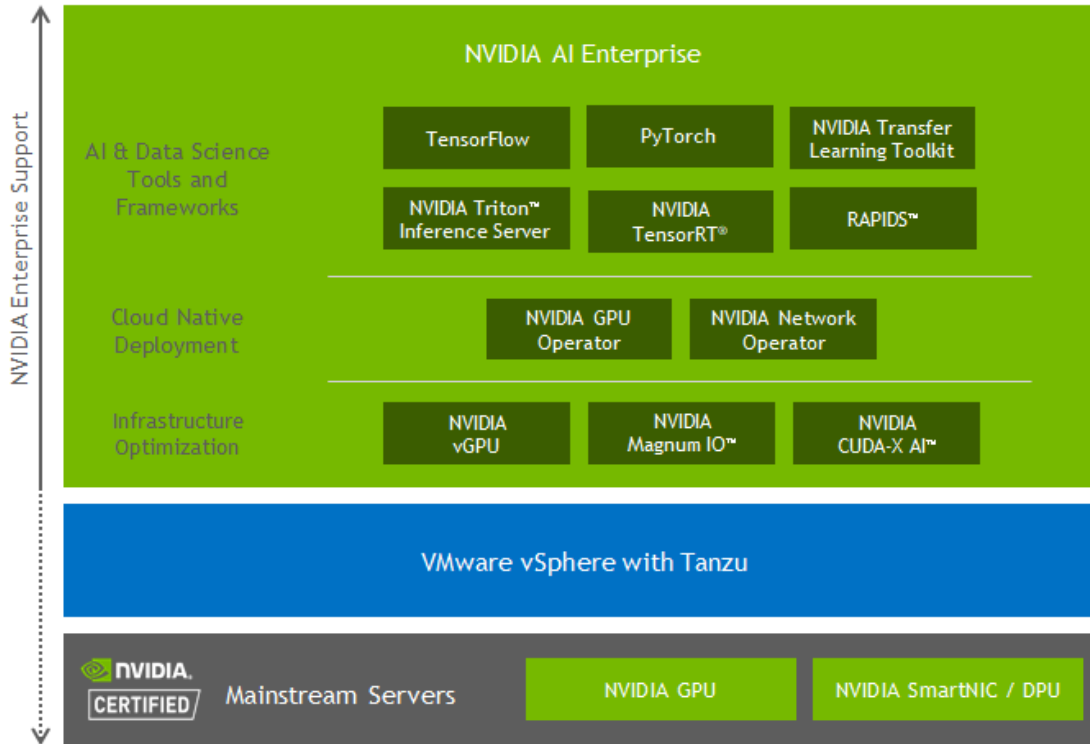
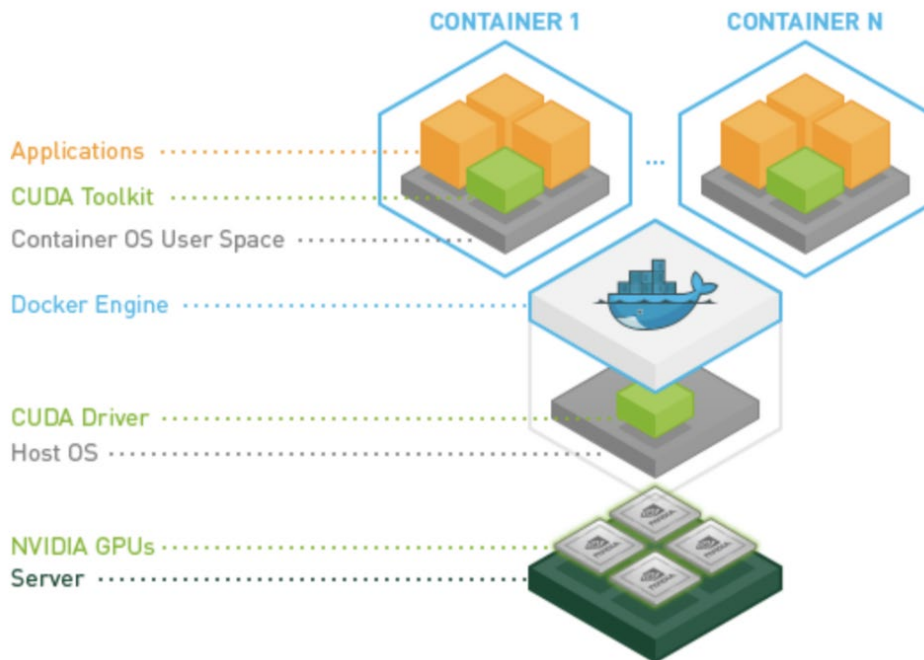


FIGURE 7. NVIDIA AI Enterprise Suite

The enterprise-grade AI and data science tools and frameworks are distributed through an NGC private registry. Each container image contains the entire user-space software stack that is required to run the application or framework, namely, the CUDA libraries, NVIDIA CUDA® Deep Neural Network library (cuDNN), any required Magnum IO components, TensorRT, and the framework. NVIDIA AI Enterprise Getting Started Guide is available as part of the early access program.

The NVIDIA Container Toolkit allows users to build and run GPU accelerated Docker containers. The toolkit includes a container runtime [library](#) and utilities to configure containers to leverage NVIDIA GPUs automatically. Complete documentation and frequently asked questions are available on the repository [wiki](#).



NVIDIA Container Toolkit**FIGURE 8.** NVIDIA Container Toolkit

The NVIDIA Container Toolkit as represented by Figure 8 is architected so that it can be targeted to support any container runtime in the ecosystem. For Docker, the NVIDIA Container Toolkit is comprised of `nvidia-docker2`, `nvidia-container-runtime`, `nvidia-container-toolkit`, and `libnvidia-container`. The control flow of this would originate from a Docker container talking via a shim layer i.e., NVIDIA Container Runtime, and eventually to the NVIDIA guest VM driver via the toolkit prestart-hook and NVIDIA container library.

Infrastructure software primarily contains the NVIDIA vGPU software and CUDA toolkit. Data science and AI frameworks bring in the power of TensorFlow, PyTorch, and RAPIDS to be readily consumable. In our experiments, it was easy to install and use it with minimal effort to get started with data science and AI projects. However, you need to get the Linux kernel version right and lock it down, which may limit your ability to install security updates.

Transfer Learning Toolkit (TLT) speeds up AI training by over 10x and creates highly accurate and efficient domain-specific AI models. Creating an AI/ML model from scratch to solve a business problem is capital intensive and time-consuming. Transfer learning is a popular technique that can be used to extract learned features from an existing neural network model to a new one. The NVIDIA Transfer Learning Toolkit (TLT) is the AI toolkit that abstracts away the AI/DL framework complexity and enables you to build production-quality pre-trained models faster with no coding required. A toolkit for anyone building AI apps and services, TLT helps reduce costs associated with large-scale data collection, labeling, and eliminates the burden of training AI/ML models ground up. With TLT, you can use NVIDIA's production quality pre-trained models and deploy as is or apply minimal fine-tuning for various computer vision and conversational AI use-cases.

RAPIDS is a suite of open-source software libraries and APIs for executing data science pipelines entirely on GPUs—and can reduce training times from days to minutes. Built on NVIDIA® CUDA-X AI™, RAPIDS unites years of development in graphics, machine learning, deep learning, high-performance computing (HPC), and more. It runs entire data science workflows with high-speed GPU compute and parallelize data loading, data manipulation, and machine learning for 50X faster end-to-end data science pipelines.



HIGH-LEVEL SOLUTION WORKFLOW

The high-level solution workflow described in Figure 9 will allow readers to quickly go through the prerequisites, required configurations, and assess the amount of time it might take to set up the HPE ProLiant DL380 Gen10 and NVIDIA A100 solution stack.

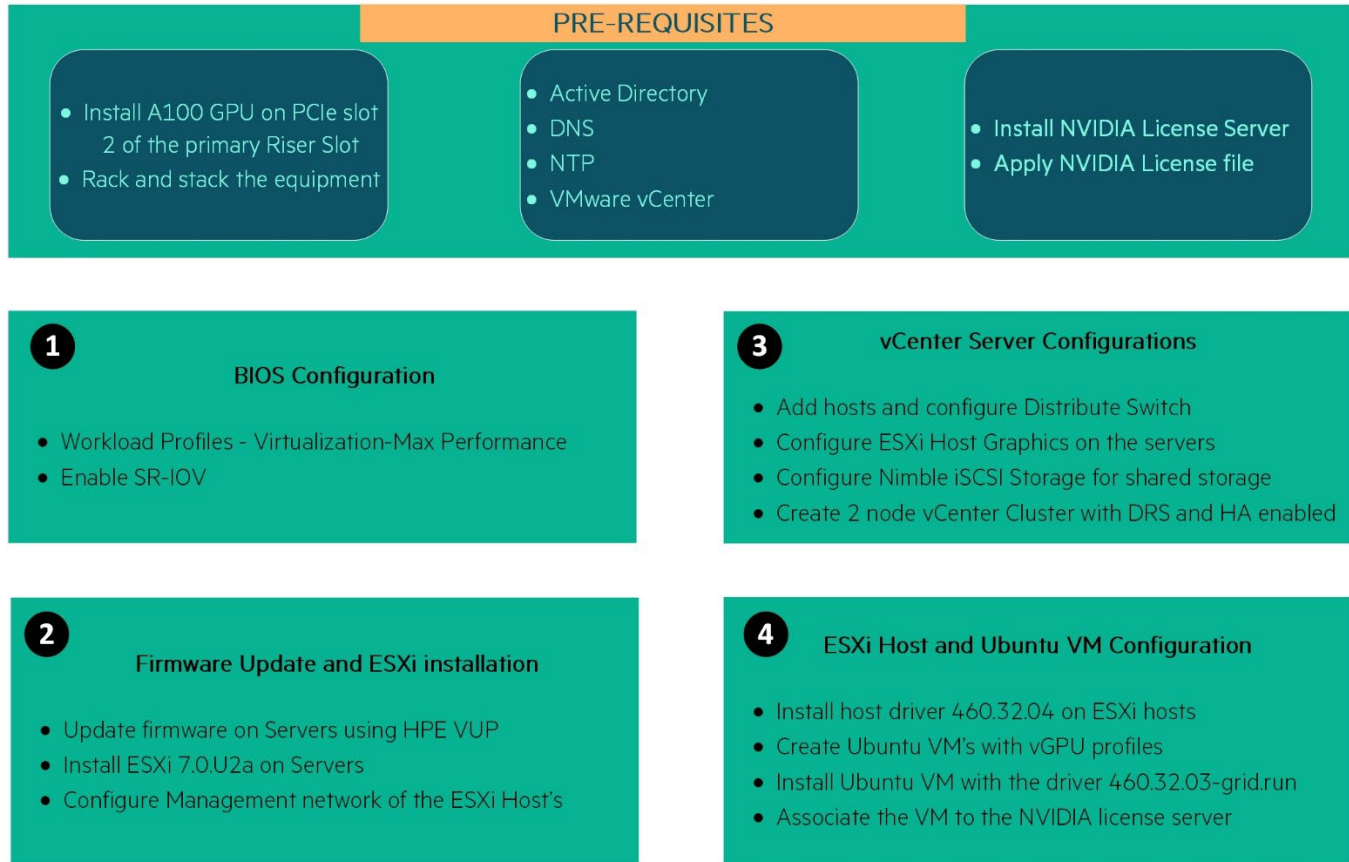


FIGURE 9. High-level solution workflow



Figure 10 further allows the readers to gain an in-depth understanding of how the underlying infrastructure is set up to enable the HPE NVIDIA AI solution stack on HPE ProLiant DL380 Gen10 servers. The infrastructure server hosts vCenter server appliance, NVIDIA license server, Active Directory, DNS, and NTP servers. The HPE Integrated Lights-Out or iLO is an embedded server management technology that provides out-of-band management capability for the server. The iLO port for managing the servers is connected to a pair of Aruba 6300 VSF stacks for out-of-band management connectivity. The uplinks providing data center connectivity and iSCSI connectivity to the HPE Nimble storage from HPE Ethernet 10/25Gb 2-port 640FLR adapter are wired to Aruba 8325 VSX fabric.

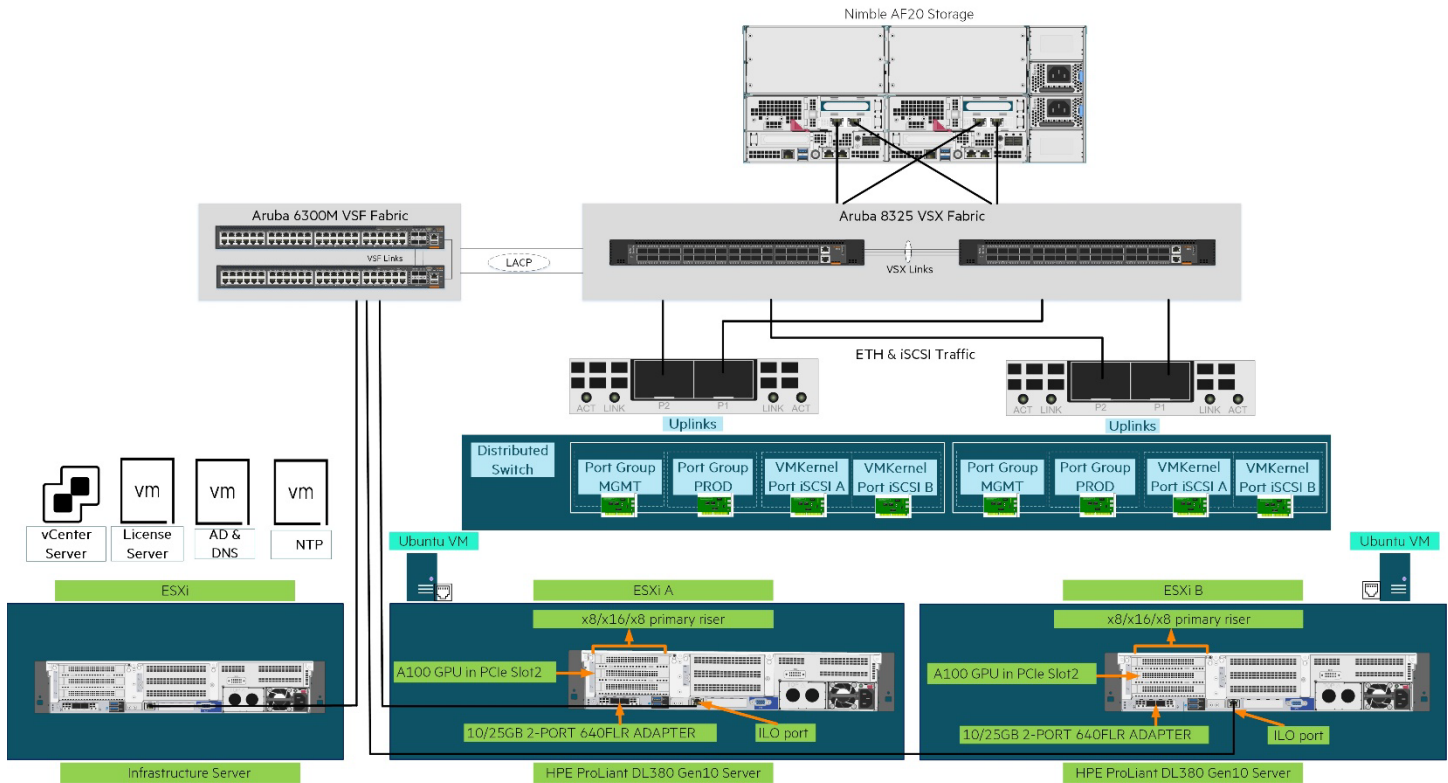


FIGURE 10. HPE Infrastructure solution setup

SOLUTION CONFIGURATION GUIDANCE

Hewlett Packard Enterprise has validated the capabilities of this solution as described in this paper. In this section, we focus on configurations that are used in the Reference Configuration. These configurations and choice of platforms vary as per installation factors in a real-world production scenario. The actual choices should be driven as per the business needs of an organization.

BIOS configuration

The following BIOS settings must be performed on each of the HPE ProLiant DL380 Gen10 servers:

1. Enable SR-IOV, this setting enables a hypervisor to create virtual instances of a PCIe device.
2. Enable VT-d/IOMMU to enable PCI passthrough.

HPE ProLiant DL380 Gen10 servers simplify BIOS configuration and allow system administrators to choose the most appropriate Workload Profile for the applications and set BIOS parameters accordingly. To enable the above settings, apply the "Virtualization –Max Performance" workload profile in the BIOS/Platform Configuration (RBSU) section.

The following BIOS setting is optional:

1. Enable Maximum cooling under **Advanced Options > Fan and Thermal Options > Thermal Configuration > Maximum Cooling**.



ESXi Host GPU settings

For the GPU to be shared among multiple virtual machines the GPU must be configured in “Shared Direct” Mode. To make these changes, navigate to **Host > Configure > Graphics** and edit the graphic devices in the vCenter Server.

1. Choose Shared Direct.
2. Choose Spread VMs across GPUs.

NVIDIA A100 configuration on the ESXi host

1. After the NVIDIA A100 cards are physically installed on the servers, place the host server into maintenance mode, install the NVIDIA Virtual GPU Manager Package VIB, reboot the server, and exit maintenance mode.

```
# esxcli software vib install -d /vmfs/volumes/<volume containing the vib>/NVIDIA-ESXi-<vib name>.zip
```

2. NVIDIA vGPU software supports MIG only with NVIDIA Virtual Compute Server and Linux® guest operating systems. To support GPU instances with NVIDIA vGPU, a GPU must be configured with MIG mode enabled. Enable Multi-Instance GPU (MIG). Run the command `nvidia-smi -I 0 -mig 1` to enable the MIG. MIG partitions a single NVIDIA A100 GPU into as many as seven independent GPU instances.
3. Verify the installation of the NVIDIA vGPU software package by checking for the NVIDIA kernel driver in the list of kernels loaded modules.

To ensure that the NVIDIA vib is installed correctly on the ESXi host and is communicating with the GPU, run `nvidia-smi` command to list the GPUs on the HPE ProLiant DL380 Gen10 server. Figure 11 shows details of the NVIDIA A100 GPU installed on the Server.

```
[root@saroesx:~] nvidia-smi
Wed May 19 04:03:16 2021

+-----+
| NVIDIA-SMI 460.32.04      Driver Version: 460.32.04      CUDA Version: N/A      |
+-----+
| GPU   Name               Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf    Pwr:Usage/Cap|      Memory-Usage | GPU-Util  Compute M. |
|                                           MIG M.         |
+-----+-----+
|  0   A100-PCIE-40GB      On          | 00000000:37:00.0 Off  |      N/A   Default |
| N/A   42C    P0      85W / 250W | 24704MiB / 40536MiB |           Enabled |
+-----+-----+

+-----+
| MIG devices:              |
+-----+-----+
| GPU  GI  CI  MIG |      Memory-Usage | Vol|      Shared |
| ID   ID  ID  Dev |      BAR1-Usage | SM  Unc| CE  ENC  DEC  OFA  JPG |
|                                           ECC |
+-----+-----+
|  0   1   0   0   | 19968MiB / 20096MiB | 42  0 | 3   0   2   0   0 |
|                                           0MiB / 32767MiB |
+-----+-----+
|  0  11   0   1   | 4736MiB / 4864MiB | 14  0 | 1   0   0   0   0 |
|                                           0MiB / 8191MiB |
+-----+-----+

+-----+
| Processes:                |
| GPU   GI  CI          PID   Type   Process name          GPU Memory |
| ID   ID  ID                |                   |          Usage     |
+-----+-----+
|  0   11   0    2342209   C+G   ubu-srv-VM2           4736MiB |
|  0    1   0    2345276   C+G   ubu-srv-vm3           19968MiB |
+-----+-----+
```

FIGURE 11. NVIDIA A100 GPU details



Assign vGPU profile to a VM

The process of creating GPU instances and the compute instances is more automated with VMware vSphere 7 Update 2 and later. With vSphere 7 Update 2 and later, you do not need to manually create GPU instances to use MIG-backed vGPU profiles. The set of A100-specific vGPU profiles that represent the required GPU instance profiles for MIG, are made available to you in the vSphere Client at the time of configuring a new PCIe device on a virtual machine as shown in Figure 14. Follow the steps below to customize the VM

1. It is important to choose EFI as the boot option when creating a virtual machine.
2. Two advanced options should be added to the advanced configuration section of the virtual machine. `pciPassthru.use64bitMMIO` and `pciPassthru.64bitMMIOSizeGB` values should be added to configuration parameters depending on the type of the vGPU instance that the virtual machine will use. vGPU profile that uses more than 16GB of memory would require this additional step.
3. In the vSphere Client, choose the virtual machine and choose **Edit Settings > VM Options > Advanced > Configuration Parameters > Edit Configuration** to get to the list of PCI-related options. Add the following two parameters as shown in Figure 12.

```
pciPassthru.use64bitMMIO="TRUE"
```

```
pciPassthru.64bitMMIOSizeGB=<n>
```

Configuration Parameters	
pciBridge6.virtualDev	pcieRootPort
pciBridge6.functions	8
pciBridge7.present	TRUE
pciBridge7.virtualDev	pcieRootPort
pciBridge7.functions	8
hpet0.present	TRUE
sched.cpu.latencySensitivity	normal
vmware.tools.internalversion	0
vmware.tools.requiredversion	11333
migrate.hostLogState	none
migrate.migrationId	0
migrate.hostLog	test-32ee4b6d.hlog
viv.moid	925f59f7-122c-4880-a4bb
pciPassthru0.fbSizeMB	40960
pciPassthru.use64bitMMIO	TRUE
pciPassthru.64bitMMIOSizeGB	80

FIGURE 12. Advanced configuration parameters of VM



The value of the second parameter in the dialog above is adjusted to suit your specific GPU requirements. In this example, the vGPU profile we chose is `grid_a100-7-40c` which would use 40GB of memory as shown in Figure 13 and we would set `pciPassthru.64bitMMIOSizeGB` to "80". For more information, follow the VMware documentation blog at <https://blogs.vmware.com/apps/2018/09/using-gpus-with-virtual-machines-on-vmware-part-2-vmdirectpath-i-o.html>.



FIGURE 13. vGPU Profile for VM

If the NVIDIA A100 drivers are correctly installed on a host machine, you will see a list of vGPU profiles by default when creating a virtual machine. Figure 14 shows the NVIDIA A100 MIG backed vGPU profiles.

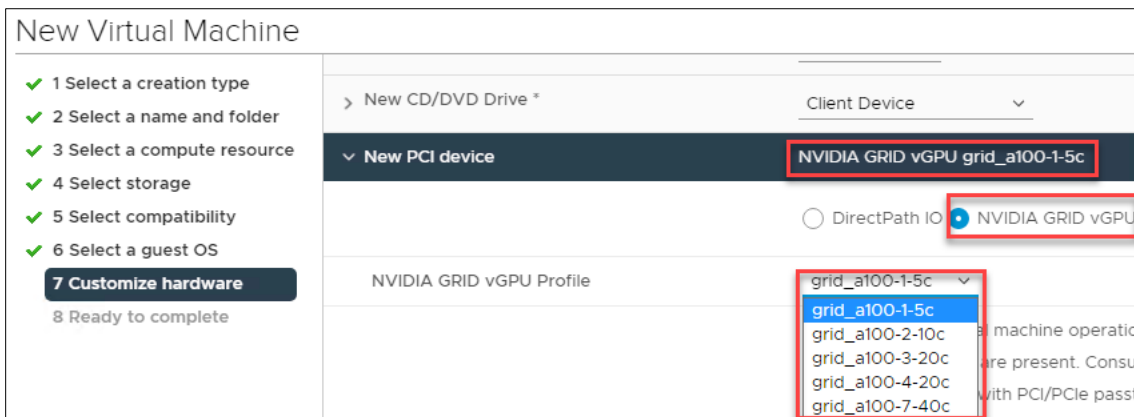


FIGURE 14. NVIDIA vGPU Profiles available for a Virtual Machine

Total and available instances can also be verified by running the command `nvidia-smi mig -lgip` on the ESXi host.

VMware vSphere Cluster configuration

1. VMware vSphere supports vMotion to perform a live migration of NVIDIA MIG-backed vGPU-powered virtual machines without causing data loss. vMotion of the virtual machine with vGPU between two hosts is supported without any downtime provided `vgpu.hotmigrate.enabled` is enabled and the target host has sufficient vGPU resources. Once the migration is completed, access to the VM resumes and all applications continue from their previous state.
2. Cluster enabled with High availability (HA) restarts the VM with vGPU to a different host in case of failure provided the failover host has sufficient vGPU resources. HA is a VMware vSphere feature that is enabled on the cluster in the vCenter. HA ensures VM's are up and running as soon as possible in the event of host failure. The high availability feature works flawlessly with vGPU enabled VM as well. The VM will automatically restart as soon as possible on another host with sufficient resources. It is important to note that the fail-over host should have matching resources in terms of vGPU instances, CPU, memory, and other resources required by the virtual machine to restart.

NOTE

Workloads will experience downtime as the virtual machine restarts, the administrator must manually intervene and restart the workloads.

3. Cluster enabled with Distributed Resource Scheduler (DRS) automatically executes only the initial placement of VM's with NVIDIA vGPU on suitable host servers. When a user sets up a bunch of VM's with vGPU's and powers them on, DRS will kick in and executes the initial placement of VMs depending on the available GPU resources on the hosts. The load balancing feature of DRS is not currently supported on

VMs with vGPU's. This means that if the host is running out of GPU resources DRS will not automatically migrate VMs with vGPU's to other hosts however the administrator can perform manual vMotion.

Ubuntu VM configuration

Follow the software components table to install the appropriate Ubuntu image, kernel, and driver for the Ubuntu VM. The beta release of NVIDIA Enterprise AI requires a specific release of the Ubuntu operating system with a specific Linux kernel version, namely Ubuntu 20.04.01 with Linux kernel 5.4.0-42. To ensure this requirement is met, you must prevent Linux kernel updates in the VM. Refer to NVIDIA AI Enterprise Getting Started Guide for details. The document is currently only available to early access customers. NVIDIA is working on providing the public documents and will provide the links when they go live.

1. Install the Ubuntu Linux compiler toolchain and kernel headers.
2. Install the vGPU software graphics driver downloaded from the NVIDIA Licensing Portal.
3. License the NVIDIA vGPU. Refer to the document [here](#).
 - a. NVIDIA vGPU software deployments require an appropriate license to be applied. vCS license was applied in this configuration.
4. Run the command `nvidia-smi` to verify if the virtual machine is communicating with the GPU.

```
thunder@ubuntuusrv-2004:~$ nvidia-smi
Wed May 19 04:28:37 2021

+-----+
| NVIDIA-SMI 460.32.03   Driver Version: 460.32.03   CUDA Version: 11.2   |
+-----+-----+
| GPU  Name           Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf    Pwr:Usage/Cap|      Memory-Usage | GPU-Util  Compute M. |
|                                           MIG M.         |
+-----+-----+
|   0   GRID A100-2-10C      On          | 00000000:02:00:0 Off |           On         |
| N/A   N/A     P0     N/A /  N/A | 1058MiB / 10235MiB |    N/A    Default   |
|                                           Enabled       |
+-----+-----+
```

FIGURE 15. Output of nvidia-smi command

5. Install NVIDIA Container Toolkit.
 - a. NVIDIA Container Toolkit includes the required version of Docker.

AI/ML workload

NVIDIA AI Enterprise includes GPU-optimized artificial intelligence (AI), machine learning (ML), tools, and frameworks that simplify and accelerate end-to-end workflows. NVIDIA AI Enterprise empowers researchers, data scientists, and developers with performance-engineered containers featuring AI software like TensorFlow, PyTorch, NVIDIA TensorRT™, and RAPIDS. Also included are NVIDIA® CUDA® Toolkit, NVIDIA deep learning libraries that enable data scientists, developers, and DevOps teams to build and deploy AI solutions faster. Deep Learning is sweeping across industries. It simplified lots of engineering efforts earlier required as ML feature engineering, an approach to spoon-feed “what to look for” when training an AI model. Image classification, video analytics, speech recognition, Natural Language Processing, and Understanding are a few popular internet content-driven use cases. In the Medicine field Drug discovery, Cancer cell detection, and Diabetic retinopathy are few sample use-cases. Autonomous Machines, Security and Defense, Media and Entertainment industries have also directly deployed their relevant use cases in a similar fashion and getting benefited.

NVIDIA A100 GPU can be partitioned into different-sized Multiple Instance GPU (MIG) instances. Each MIG instance can be allocated to a VMware virtual machine to run different types of workloads – model development, deep learning training, AI inference. The workloads run simultaneously on different MIG instances, with dedicated compute, memory, and memory bandwidth.



NVIDIA A100 in MIG mode can run a mix of up to seven AI/ML workloads of different sizes. For example, IT administrators can create development environments for model development and low latency inference with a single GPU Instance allocated to each virtual machine. Up to seven data scientists can simultaneously work to develop and fine-tune deep learning models. MIG instances can be reconfigured (VM would need to be shut down for the MIG instance reconfiguration), enabling administrators to reallocate GPU resources based on changing workload demands. The GPU can be reconfigured to create three MIG instances with 10GB memory or two MIG instances with 20GB of memory each providing the ability to scale the development environment with the size of the model and datasets used.

NVIDIA AI Enterprise - Sample application deployment

NVIDIA AI Enterprise provides GPU-optimized pre-integrated containers that include pre-trained models, SDKs, and frameworks. NVIDIA AI Enterprise with built-in support for common ML frameworks such as TensorFlow to develop and training ML models can be deployed on Ubuntu 20.04 virtual machines.

Example for using NGC container for TensorFlow based workloads:

```
$ docker pull nvcr.io/nvidia/tensorflow:20.08-tf1-py3
$ docker run -it --shm-size=1g --ulimit memlock=-1 --ulimit stack=67108864 --gpus all -p 8888:8888 -v
$PWD:/projects --network=host nvcr.io/nvidia/tensorflow:20.08-tf1-py3
```

We have not covered these interesting capabilities within the scope of this Reference Configuration - VMware DirectPath I/O, Leveraging RDMA capabilities in conjunction with GPU, Multi-Node training, and Distributed training of AI models. We have also used a simple AI model in a Cocker container to assess how quickly we can build a readymade AI development and deployment setup instead of rigorous benchmarking with various core capabilities. In summary, this configuration will give a quick start to your AI applications.

SUMMARY

Hewlett Packard Enterprise, VMware, and NVIDIA together developed this Reference Configuration as a quick start for AI researchers, data scientists, and developers on AI development to develop and deliver successful AI projects faster. IT operations and administrators now can have the ability to support AI development and deployment using a familiar toolchain. The power of the platform, agility, and ease of familiar VMware virtualization, sharing GPU resources for the acceleration of AI training in a cost-effective manner, and ready to use pre-built container ecosystem from NVIDIA with Proprietary and open source options make this Reference Configuration a must-have for enterprises looking to delve into AI more seriously. We have just scratched the surface of the overall capabilities and deferring more important capabilities built over RDMA enabling multi-node and distributed training of AI workloads.



Reference Configuration

RESOURCES AND ADDITIONAL LINKS

HPE Reference Architectures, <https://www.hpe.com/docs/reference-architecture>

HPE Pointnext Services, <https://www.hpe.com/us/en/services/pointnext.html>

HPE Servers, [hpe.com/servers](https://www.hpe.com/servers)

HPE Server Storage, [hpe.com/us/en/servers/server-storage.html](https://www.hpe.com/us/en/servers/server-storage.html)

HPE Server Networking, [hpe.com/us/en/servers/networking.html](https://www.hpe.com/us/en/servers/networking.html)

NVIDIA GPU Catalog, <https://www.nvidia.com/en-us/gpu-cloud/>

NVIDIA AI Enterprise suite, <https://www.nvidia.com/en-us/data-center/products/ai-enterprise-suite/>

HPE GreenLake Advisory and Professional Services, <https://www.hpe.com/us/en/services/consulting.html>

To help us improve our documents, please provide feedback at [hpe.com/contact/feedback](https://www.hpe.com/contact/feedback).

© Copyright 2021-2024 Hewlett Packard Enterprise Development LP. The information contained herein is subject to change without notice. The only warranties for Hewlett Packard Enterprise products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. Hewlett Packard Enterprise shall not be liable for technical or editorial errors or omissions contained herein.

VMware, VMware vSphere, VMware NSX are registered trademarks of VMware, Inc. in the United States and/or other jurisdictions. VMware vCenter, VMware ESXi, VMware vSAN are the trademark of VMware, Inc. in the United States and/or other jurisdictions. Microsoft and Windows Server are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries. Intel, Xeon are trademarks of Intel Corporation in the U.S. and other countries. NVIDIA is registered trademark of NVIDIA Corporation in the U.S. and other countries. Linux is the registered trademark of Linus Torvalds in the U.S. and other countries. All third-party marks are property of their respective owners.