TECHNICAL WHITE PAPER - MAY 2020

SHARING GPUS IN MACHINE LEARNING ENVIRONMENTS

REFERENCE ARCHITECTURE

1 Executive Summary

This Reference Architecture describes a VMware Cloud Foundation based solution for Machine Learning environments with GPUs. The solution combines VMware virtualization and container orchestration with the latest hardware innovations to provide robust infrastructure for machine learning applications.

Two modes of sharing GPUs available in vSphere environments are leveraged in the solution. The NVIDIA vGPU based solution is used for use cases requiring local access by the VM to the GPUs (i.e. the use of remote GPUs is not allowed). vSphere Bitfusion is used for distributed use cases that access the GPUs over the network. VMware Cloud Foundation introduces the concept of a “workload domain”, which is a set of VMs and resources that contain a particular workload. A machine learning and GPU application workload domain is used in this solution to leverage all VMware capabilities, while optimizing GPU usage. High speed PVRDMA based networking is leveraged to reduce network impact with Bitfusion based workloads.

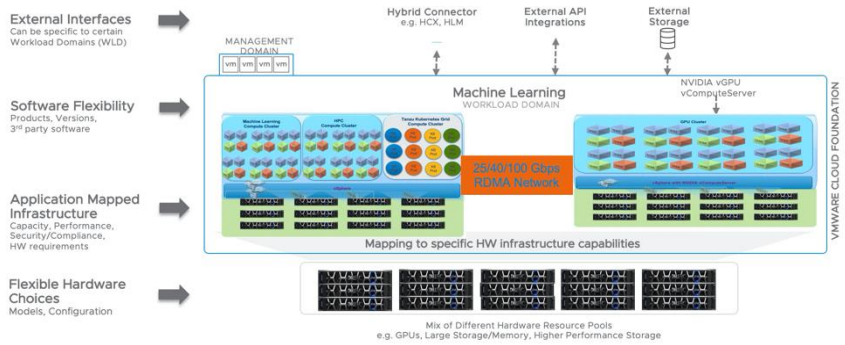

Figure 1: GPU Optimized Workload Domain

2 Challenges with GPU Usage

GPUs are getting increasingly faster but not all ML and GPU applications are currently using them. GPUs in the physical bare-metal world are stuck in silos and grossly underutilized. A survey of most enterprises has shown that GPUs are utilized only 25-30% of the time. Even though some GPUs are shared in servers, the majority of GPUs are bound to workstations that are dedicated to individual users. In a typical workflow, a researcher will set up a large number of experiments, wait for them to finish and on completion work to digest the results, while the GPUs sit idling. Due to the disaggregated nature of existing GPU infrastructures

there is minimal automation and ability to effectively utilize the GPU capabilities. The physical nature of the infrastructure for GPUs does not allow for secure access and sharing across teams with multi-tenancy.

This leads to major inefficiencies in GPU utilization and user productivity. The disaggregation of the resources makes it hard to automate the solutions and reduces the overall potential of the infrastructure.

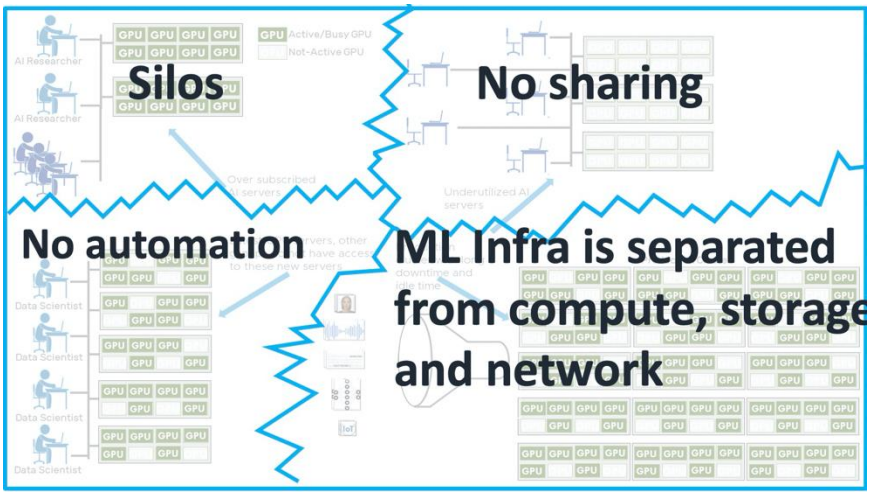

Figure 2: Current challenges with GPU Utilization

2.1 Modern Developers & Data Scientist challenges

Modern applications require infrastructures that support all of their components. Without an adaptive infrastructure, developers and IT operators are often at odds and struggle to provide the services that applications require.

- Lack of modern developer services within organization
- Access to the modern coding tools and backend databases is slow
- Deployment, day 2 operations and lifecycle management are painful

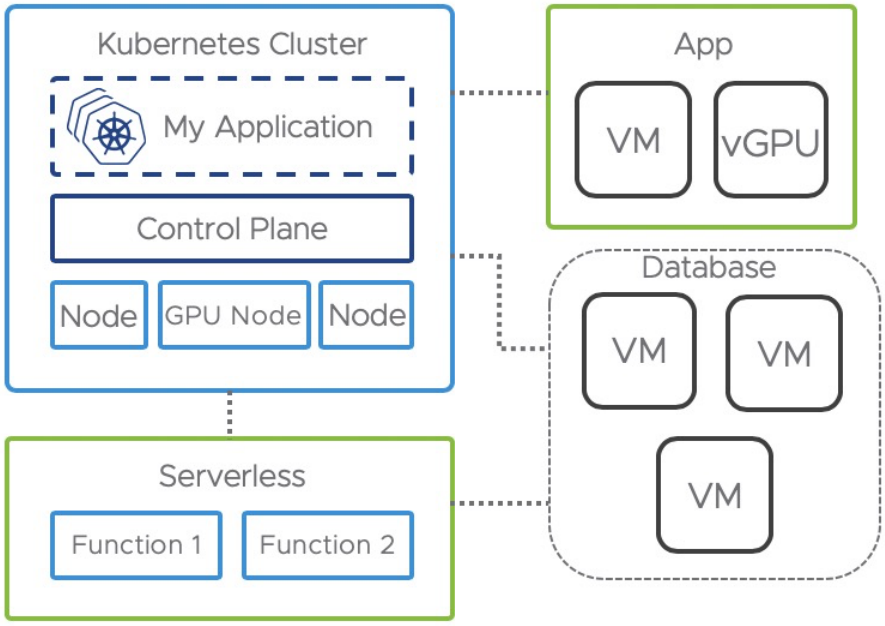

Figure 3: Hybrid components of modern IT landscape

2.2 **IT Operators challenges**

- Deploying and configuring right infrastructure defined by Modern App Interfaces
- Infrastructure silos exist as provisioning resources for developers is a nightmare
- Security isolation of modern apps and sensitive databases is difficult
- Inconsistent operations and cross-functional workflows remains a concern

3 Why Share GPUs for Machine Learning?

Machine learning is a subset of the broader field of artificial intelligence, which uses statistical techniques to allow programs to learn from experiences or from existing data.

Deep learning is a machine learning technique that enables computers to learn from example data. Deep Learning models support applications from autonomously driven vehicles, medical image recognition, voice

processing systems and many others. The rise of deep learning platforms has led to an exponential growth in these workloads, both in both data centers and in cloud environments. GPUs provide the computing power needed to run deep learning and other machine learning programs efficiently, reliably and quickly. These GPU-based workloads are even more versatile, flexible and efficient when they run in virtual machines on VMware vSphere. Sharing GPUs provides the following advantages:

- Achieve higher utilization of each individual GPU
- Gain greater efficiency through sharing of GPUs across users and applications
- Allow users to make use of partial or multiple GPUs on a case-by-case basis as their applications need them

4 Features that can be leveraged for machine learning on vSphere

Increasingly, data scientists and machine learning developers are asking their systems administrators to provide them with a GPU-capable machine setup so they can execute workloads that need GPU power. The data scientist often describes these workloads as machine "training," "inference," or "development."

GPUs are commonly used today for highly graphical applications on the desktop. Organizations already using desktop VMs on vSphere can also use their existing GPUs with vSphere for applications other than this virtual desktop infrastructure (VDI) scenario. This non-graphical use case is known as a "Compute" workload in vSphere, and it enables end users to consume GPUs in VMs in the same way they do in any GPU-enabled public cloud instance or on bare metal, but with more flexibility. Through collaboration with VMware technology partners, vSphere allows flexible consumption and multiple GPU utilization models that can increase the ROI of this infrastructure, while providing end users with exactly what they need.

4.1 NVIDIA GPUs for Machine Learning

With the impending end to Moore's law, the spark that is fueling the current revolution in deep learning is having enough compute horsepower to train neural-network based models in a reasonable amount of time

The needed compute horsepower is derived largely from GPUs, which NVIDIA began optimizing for deep learning in 2012. One of the latest in this family of GPU processors is the NVIDIA v100 Tensor Core GPU.

Figure 4: The NVIDIA V100 GPU

NVIDIA® V100 line of GPUs is currently the most advanced data center GPU built to accelerate AI, high performance computing (HPC), data science and graphics. It's powered by NVIDIA Volta architecture, comes in different memory configurations, and offers the performance of many CPUs in a single GPU. Data scientists, researchers, and engineers can now spend less time optimizing memory usage and more time designing the next AI breakthrough. (Source: NVIDIA)

The NVIDIA vGPU software family enables GPU virtualization for any workload and is available through licensed products such as the NVIDIA vComputeServer. NVIDIA vComputeServer software, enables virtualize NVIDIA GPUs to power the more than 60 GPU accelerated applications for AI, deep machine learning, and high-performance computing (HPC) as well as the NGC GPU optimized containers. With GPU sharing, multiple VMs can be powered by a single GPU, maximizing utilization and affordability, or a single VM can be powered by multiple virtual GPUs, making even the most compute-intensive workloads possible. With vSphere integration, GPU clusters for compute can be managed by vCenter, maximizing GPU utilization and ensuring security.

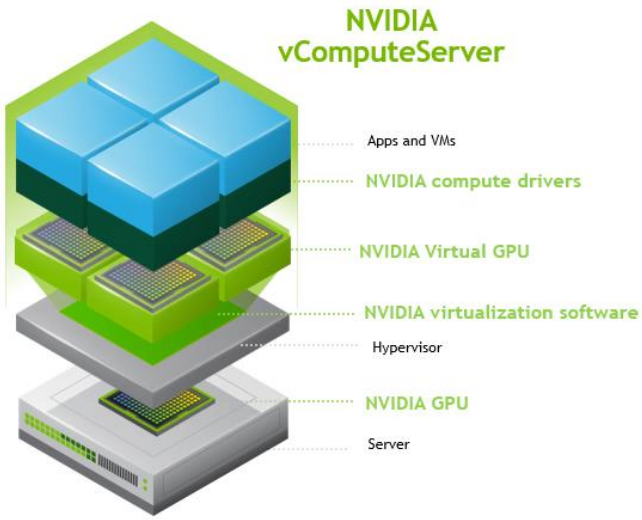

Figure 5: Layered model showing NVIDIA vGPU components

4.2 High Speed Networking with PVRDMA & RoCE

Remote Direct Memory Access (RDMA) provides direct memory access from the memory between hosts bypassing the Operating System and CPU. This can boost network and host performance with reduced latency & CPU load while providing higher bandwidth. RDMA compares favorably to TCP/IP, which adds latency and consumes significant CPU and memory resources.

Benchmarks demonstrate that the NVIDIA® vComputeServer or virtualized GPUs achieve two times better efficiency by using VMware's paravirtualized RDMA (PVRDMA) technology than when using traditional networking protocols. Give a reference here to the benchmark. (yes, reference is needed here)

4.3 RDMA over Converged Ethernet

RDMA over Converged Ethernet (**RoCE**) is a network protocol that allows remote direct memory access (RDMA) over an Ethernet network. There are two RoCE versions, RoCE v1 and RoCE v2. RoCE v1 is an Ethernet link layer protocol and hence allows communication between any two hosts in the same Ethernet broadcast domain. RoCE v2 is an internet layer protocol which means that RoCE v2 packets can be routed. Although the RoCE protocol benefits from the characteristics of a converged

Ethernet network, the protocol can also be used on a traditional or non-converged Ethernet network. (Source: Wikipedia)

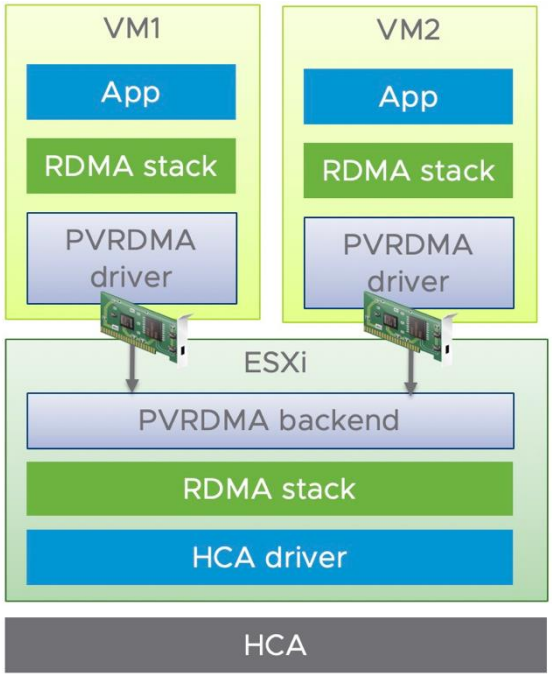

Figure 6: The end to end PVRDMA stack

4.3.1 PVRDMA in vSphere

In vSphere, a virtual machine can use a PVRDMA network adapter to communicate with other virtual machines that have PVRDMA devices. The PVRDMA device automatically selects the method of communication between the virtual machines. vMotion is supported when PVRDMA is used as the transport mechanism.

4.4 vSphere Bitfusion

vSphere Bitfusion extends the power of VMware vSphere technology to enable dynamic sharing of GPUs. vSphere Bitfusion helps enterprises disaggregate the GPU compute and dynamically attach GPUs anywhere in the datacenter just like attaching storage. Bitfusion enables use of any arbitrary fractions of GPUs. Support more users in test and development phase. vSphere Bitfusion supports the CUDA API and demonstrates

virtualization and remote attach for all hardware. GPUs are attached based on CUDA calls at run-time, maximizing utilization of GPU servers anywhere in the network.

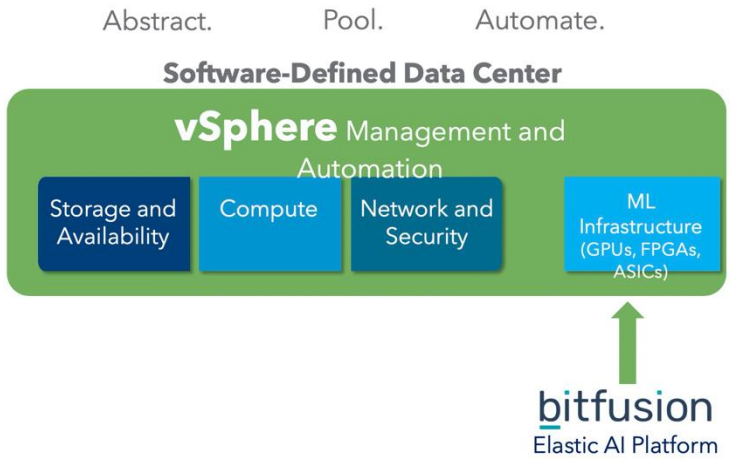

Figure 7: Elastic AI Platform with vSphere Bitfusion

Bitfusion is now part of vSphere and NVIDIA GPU accelerators can now be part of a common infrastructure resource pool, available for use by any virtual machine in the data center in full or partial configurations, attached over the network. The solution works with any type of GPU server and any networking configuration such as TCP, RoCE or InfiniBand. GPU infrastructure can now be pooled together to offer an elastic GPU as a service, enabling dynamic assignment of GPU resources based on an organization's business needs and priorities. Bitfusion runs in the user space of a guest operating system and doesn't require any changes to the OS, drivers, kernel modules or AI frameworks.

5 Common GPU Use Cases:

GPU Configuration	Use Cases
A Full GPU dedicated to a data science workstation user	• Data science workstation for development and training of machine models
Multiple GPUs used by a standalone application	• High-end machine model training • High-performance computing (genomic sequencing, Monte-Carlo analysis) • GPU-enabled databases

Partial GPU allocated per application or user	• Development and testing of machine learning applications • Small data science workstations • The inference phase of machine learning
Full or partial GPUs across multiple virtual machines used for distributed processing	• Horovod based distributed ML • Distributed TensorFlow

Table 1: Common GPU use cases

6 Cloud Infrastructure Components Leveraged for Machine Learning Infrastructure

6.1 VMware Tanzu Kubernetes Grid (TKG)

VMware Tanzu Kubernetes Grid is a container services solution that enables Kubernetes to operate in multi-cloud environments. VMware TKG simplifies the deployment and management of Kubernetes clusters with Day 1 and Day 2 operations support. VMware TKG manages container deployment from the application layer all the way to the infrastructure layer, according to the requirements VMware TKG supports high availability, autoscaling, health-checks, and self-repairing of underlying VMs and rolling upgrades for the Kubernetes clusters.

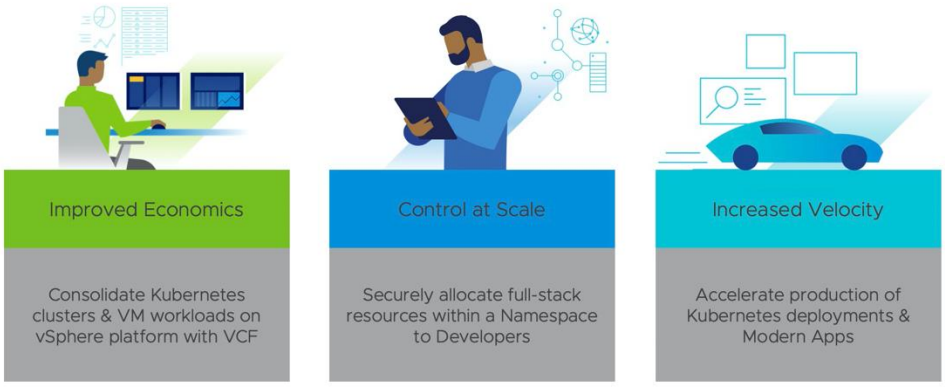

Figure 8: Tanzu Kubernetes Grid Benefits

6.2 VMware Cloud Foundation with SDDC Manager

VMware Cloud Foundation is a unified SDDC platform that brings together a hypervisor platform, software-defined services for compute, storage, network, and security and network virtualization into an integrated stack whose resources are managed through a single

administrative tool. VMware Cloud Foundation provides an easy path to hybrid cloud through a simple, security-enabled, and agile cloud infrastructure on premises and as-a-service public cloud environments. VMware SDDC Manager manages the start-up of the Cloud Foundation system. It enables the user to create and manage workload domains, and perform lifecycle management to ensure the software components remain up-to-to date. SDDC Manager also monitors the logical and physical resources of VMware Cloud Foundation.

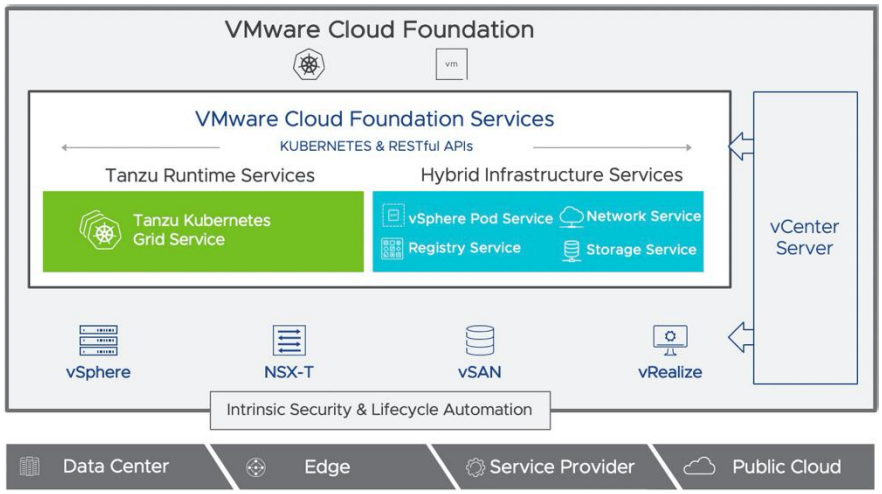

Figure 9: VMware Cloud Foundation with Tanzu Runtime and Hybrid Infrastructure services

6.3 VMware vSphere

VMware vSphere extends virtualization to storage and network services and adds automated, policy-based provisioning and management. As the foundation for VMware's complete SDDC platform, vSphere is the starting point for your virtualization infrastructure providing wide ranging support for the latest hardware and accelerators used for Machine Learning. VMware vSphere with Kubernetes (formerly called "Project Pacific") empowers IT Operators and Application Developers to accelerate innovation by converging Kubernetes, containers and VMs into VMware's vSphere platform.

VMware has leveraged Kubernetes to rearchitect vSphere and extend its capabilities to all modern and traditional applications.

6.3.1 vSphere with Kubernetes innovations:

- Unite vSphere and Kubernetes by embedding Kubernetes into the control plane of vSphere, unifying control of compute, network and storage resources. Converging VMs and containers

using the new native container service that creates high performing, secure and easy to consume containers

- App-focused management with App level control for applying policies, quota and role-based access to Developers. Ability apply vSphere features (HA, vMotion, DRS) at the app level and to the containers. Unified visibility in vCenter for Kubernetes clusters, containers and existing VMs
- Enable Dev & IT Ops collaboration. Developers use Kubernetes APIs to access the datacenter infrastructure (compute, storage and networking). IT operators use vSphere tools to deliver Kubernetes clusters to developers. Consistent view between Dev and Ops via Kubernetes constructs in vSphere
- Enterprises can expect to get improved economics due to the convergence of vSphere, Kubernetes and containers. Operators get control at scale by focusing on managing apps versus managing VMs. Developers & operators collaborate to gain increased velocity due to familiar tools (vSphere tools for Operators and Kubernetes service for Developers).

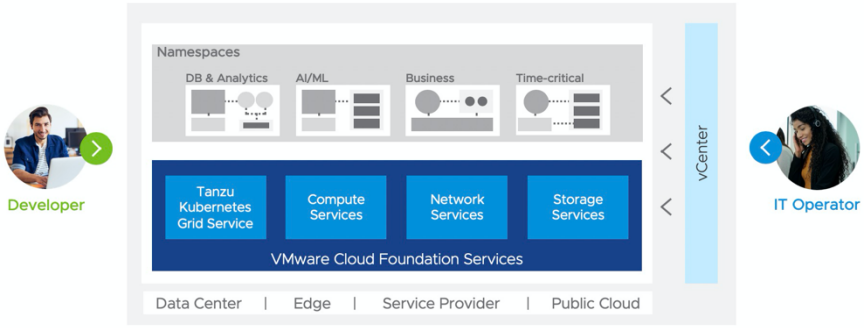

Figure 10: vSphere 7 with Kubernetes innovations

6.4 **VMWare NSX-T**

VMware NSX-T provides an agile software-defined infrastructure to build cloud-native application environments. NSX-T is focused on providing networking, security, automation, and operational simplicity for emerging application frameworks and architectures that have heterogeneous endpoint environments and technology stacks. NSX-T supports cloud-native applications, bare metal workloads, multi-hypervisor

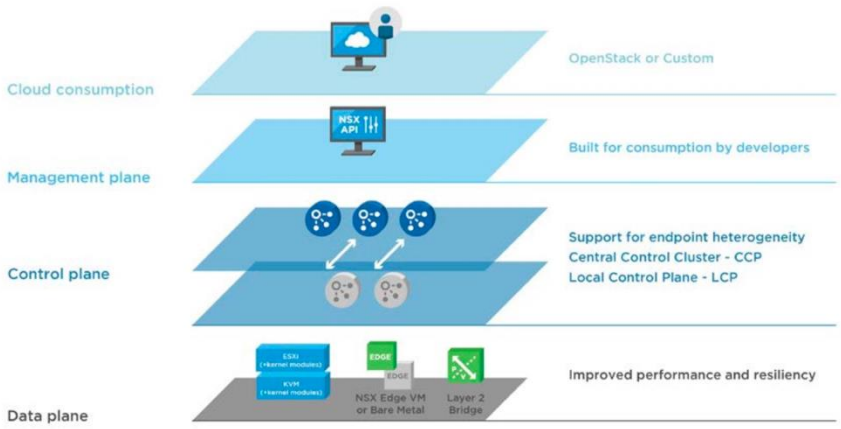

Figure 11: NSX-T Components

environments, public clouds, and multiple clouds. NSX-T is designed for management, operation, and consumption by development organizations. NSX-T Data Center allows IT and development teams to select the technologies best suited for their applications.

7 Solution Overview

This reference architecture presents a solution that takes all the above virtualization features available to build an optimal machine learning infrastructure

The reference architecture consists of the components shown in Figure 1.

- Choice of different types of servers and hardware
- VMware Cloud Foundation with
    1. Traditional Compute
    2. Kubernetes with TKG
    3. Aggregated GPU Clusters
    4. Dedicated Workload Domain
    5. VMWare NSX-T network virtualization which enables software defined ne
- GPU as a Service
    1. GPU based compute acceleration
    2. RDMA based network connectivity
    3. vSphere Bitfusion for remote full or partial GPU access
    4. NVIDIA vComputeServer

VMware Cloud Foundation deployed with VMware TKG offers a simple solution for securing and supporting containers within existing

environments that already support VMs based on VMware vSphere, without requiring any retooling or rearchitecting of the network.

This comprehensive solution can help enterprises share and operationalize GPU based acceleration for machine learning and other HPC applications, to meet the use cases required by the users. The hybrid cloud capability provides flexibility in workload placement as well as business agility.

7.1 Solution Components

The solution is layered on top of the VMware Cloud Foundation platform and it provides the following levels of technical flexibility:

1. Aggregated GPU resources in dedicated clusters
2. CPU clusters
3. RDMA based high-speed networking

7.2 GPU Clusters

7.2.1 Requirements

All GPU resources are consolidated into the GPU Cluster. A requirement analysis of all the potential GPU workloads in the organization should be done. This analysis of the use cases and the potential need for GPUs will provide an estimate of the type of GPUs and the number of GPUs per server. Once this is finalized, servers with the appropriate capacity can be chosen as the building block for the cluster. Figure 12 shows the GPU hardware that is presented through a cluster of specifically-designed virtual machines, called the GPU cluster.

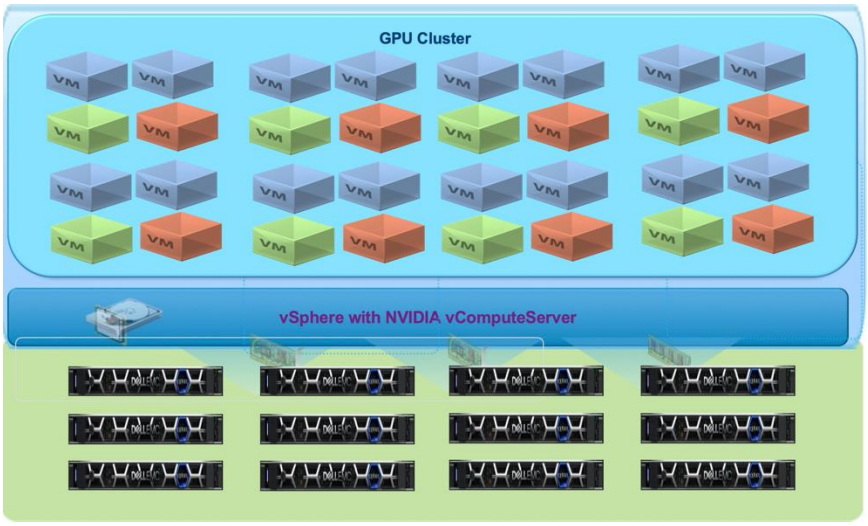

Figure 12: Logical GPU Cluster Architecture

The Dell components recommended for the GPU cluster are shown in Table 2. Depending on the capacity assessment and the need for the GPU compute, multiple Dell servers with the appropriate number of GPUs would be brought together in a VMware cluster.

Dell	Components
Servers	C4140, R740-XD
Accelerators	NVIDIA v100 NVIDIA T4 Mellanox Connect X-5
Software	VMware VCF with vSphere, TKG, NSX vSphere Bitfusion NVIDIA vComputeServer Automation with HashiCorp Terraform vSphere Provider
Ethernet	25, 40, 100 GbE

Storage	Isilon F800 Scale-out NAS
Frameworks **Libraries**	Caffe2, MXNET, TensorFlow, NVIDIA CUDA® Deep Neural, Horovod, Distributed TensorFlow, NVIDIA CUDA Deep Neural Network Library (cuDNN), NVIDIA CUDA basic linear algebra subroutines (cuBLAS)

Table 2: Dell GPU Cluster Components

7.3 Compute Clusters

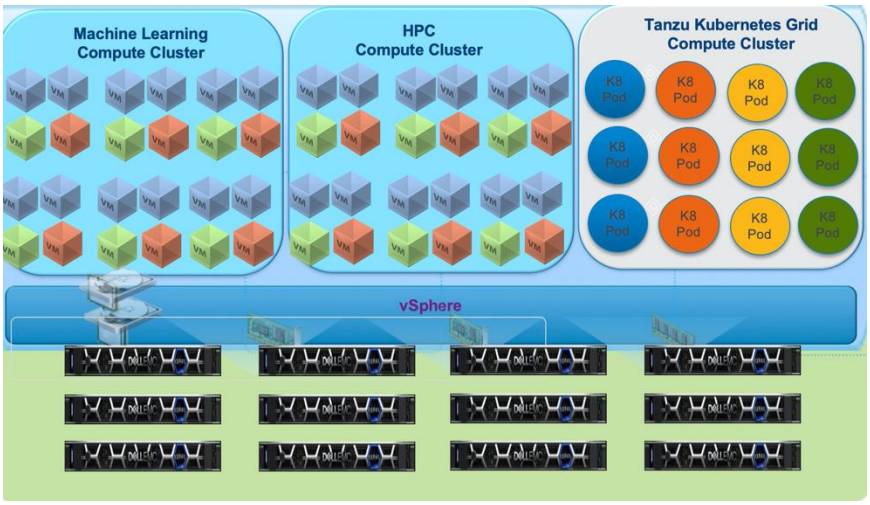

Figure 13: Compute Cluster Logical Architecture

The compute clusters are dedicated to particular application types, such as machine learning. These clusters are the clients of the GPU cluster and make use of the latter's services.

7.4 VMware Cloud Foundation Workload Domain

A workload domain in vCloud Foundation is a collection of compute, storage and networking power that is designed specifically for a particular application type. It has its own vCenter instance and therefore it lives independently of any other Workload Domain on the same hardware. It is independent of vSphere clusters for example and one Workload Domain can contain many vSphere clusters. A workload

domain exists within a software defined data center (SDDC) for one region.

Each Workload domain contains the following components:

- One vCenter Server instance connected to a pair of Platform Services Controller instances in the same or another workload domain.
- At least one vSphere cluster with vSphere HA and vSphere DRS enabled.
- One vSphere Distributed Switch for management traffic and NSX logical switching.
- NSX components that connect the workloads in the cluster for logical switching, logical dynamic routing, and load balancing.
- One or more shared storage allocations.

7.5 VCF Dedicated Workload Domain for Machine Learning and other GPU Applications

The model behind Cloud Foundation Workload Domains aligns to a concept of “Application Ready Infrastructure” where workload domains can be aligned to specific platform stacks to support different application workloads. This enables quick infrastructure and platform management components to support running a suite of different applications ranging from enterprise applications running in VMs, to Virtual Desktops, to containerized cloud native apps. Cloud Foundation becomes the common cloud platform to run both traditional and cloud native applications - all leveraging a common suite of management tools and driven through the use of automation.

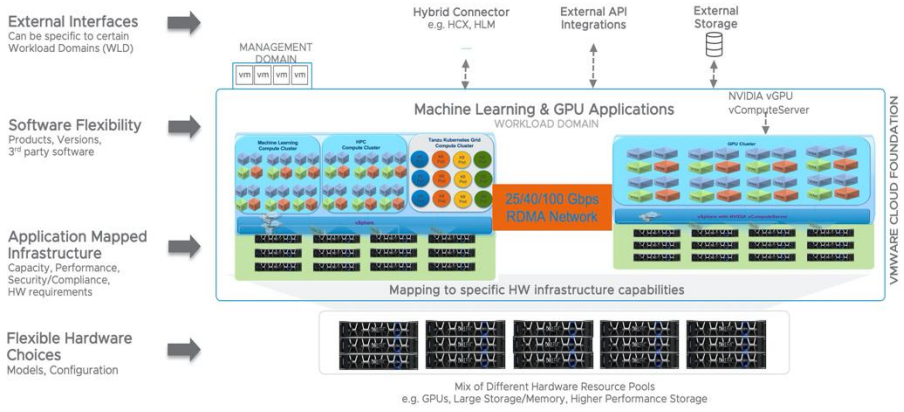

Figure 14: Consolidated GPU Applications Workload Domain

8 Starter GPU Cluster Design Example:

The scope of this sizing example is confined to the GPU requirements only. Sizing of all other infrastructure components such as CPU, Memory, Storage and Networking can be sized based on standard capacity analysis and design

8.1 GPU Specific requirements:

On performing capacity analysis, the following details emerged on the need for GPU compute capacity.

Customer has a total capacity requirement of 14 GPUs

8.2 Customer Use Cases:

- Data Scientists needing full or partial GPUs to perform analysis
- Multiple Developers requiring full or partial GPUs in their development activities

8.3 Workload Design:

8.3.1 Data Science, & Developer Users

These users require a single or partial GPU to meet their computing needs. To be able to have these users access these resources in a flexible manner from anywhere in the datacenter, these users should use Bitfusion to access their GPUs over the network.

These users multiple partial or full GPUs distributed across many worker nodes. Since each worker node gets its own full or partial GPU, Bitfusion based access over the network would be ideal for this use case.

8.4 Cluster Design:

The GPU cluster would be deployed in a VCF Workload domain. The design will incorporate the requirements of the above use cases along with some allowance for high availability and maintenance.

8.4.1 Design Decisions:

(1) Each physical node will have 2 GPUs each
(2) To provide for a capacity of 14 GPUs, a minimum of 7 hosts with 2 GPU each are required.
(3) An additional host is required for high availability and maintenance. Any failure during maintenance is assumed to be allowed to avoid adding additional hosts as overhead.
(4) The cluster would use eight nodes with 4 GPUs each providing a total capacity of 16 GPUs and 14 GPUs in the case of a node failure.

(5) Bitfusion servers will be configured with GPUs in pass through mode. All GPUs and compute resources will be fully allocated to the Bitfusion servers.
(6) Regular 10 GBPS networking between compute and GPU clusters for Bitfusion access.

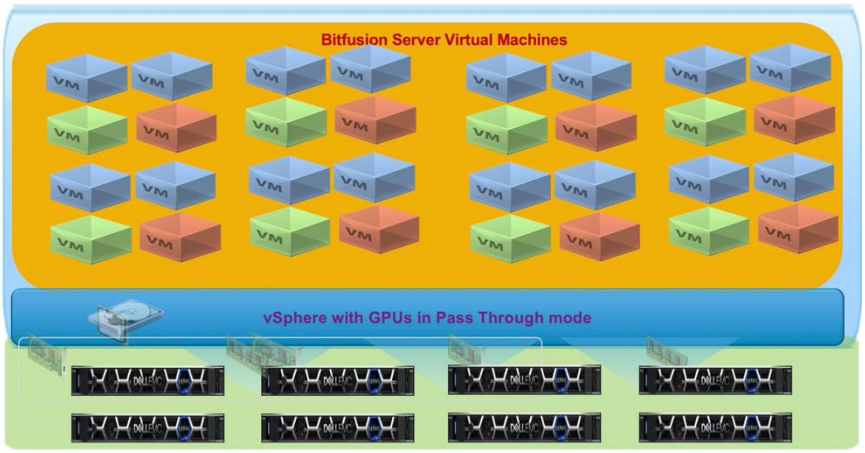

Figure 15: Sample Starter GPU Cluster Logical Architecture

9 Enterprise GPU Cluster Design Example:

The scope of this sizing example is confined to the GPU requirements only. Sizing of all other infrastructure components such as CPU, Memory, Storage and Networking can be sized based on standard capacity analysis and design

9.1 GPU Specific requirements:

On performing capacity analysis, the following details emerged on the need for GPU compute capacity.

Customer has a total capacity requirement of 28 GPUs

9.2 Customer Use Cases:

- High performance researchers with per user need of a maximum of 4 GPUs
- Data Scientists needing full or partial GPUs to perform analysis
- HPC users needing full or partial GPUs for their applications
- Multiple Developers requiring full or partial GPUs in their development activities
- Distributed Machine Learning users with need for multiple GPU resources across many worker nodes

9.3 Workload Design:

9.3.1 High Performance GPU users:

Some data scientists and researchers need more than one GPU at a time and to ensure best performance all these GPUs should exist in a single physical node. Certain HPC applications need GPU access with minimum latency over prolonged periods of time. The virtual machine used by these researchers and specialized HPC applications would run on the GPU cluster itself and leverage NVIDIA vComputeServer for GPU allocation.

9.3.2 Data Science, HPC & Developer Users

These users require a single or partial GPU to meet their computing needs. To be able to have these users access these resources in a flexible manner from anywhere in the datacenter, these users should use Bitfusion to access their GPUs over the high speed RDMA network.

9.3.3 Distributed Machine Learning

These users require multiple partial or full GPUs distributed across many worker nodes. Since each worker node gets its own full or partial GPU, Bitfusion based access over the high speed RDMA network would be ideal for this use case.

9.4 Cluster Design:

The GPU cluster would be deployed in a VCF Workload domain. The design will incorporate the requirements of the above use cases along with some allowance for high availability and maintenance.

9.4.1 Design Decisions:

(1) Each physical node will have 4 GPUs each to meet the need for high performance researchers
(2) To provide for a capacity of 28 GPUs, a minimum of 7 hosts with 4 GPU each are required.
(3) An additional host is required for high availability and maintenance. Any failure during maintenance is assumed to be allowed to avoid adding additional hosts as overhead.
(4) The cluster would use eight nodes with 4 GPUs each providing a total capacity of 32 GPUs. 28 GPUs will be available in the case of a node failure.
(5) All GPU allocations in the cluster will be through NVIDIA vComputeServer. Virtual machines using vGPU allocations can be vMotioned from one node to another which is supported. This makes maintenance of the GPU cluster feasible.
(6) Bitfusion servers are also like other consumers of GPUs allocated by NVIDIA vComputeServer that controls all allocations. A subset of the GPUs are reserved and allocated by vComputeServer for Bitfusion based on user requirements

(7) High Speed networking with PVRDMA for minimal latencies for Bitfusion access to GPUs via the network

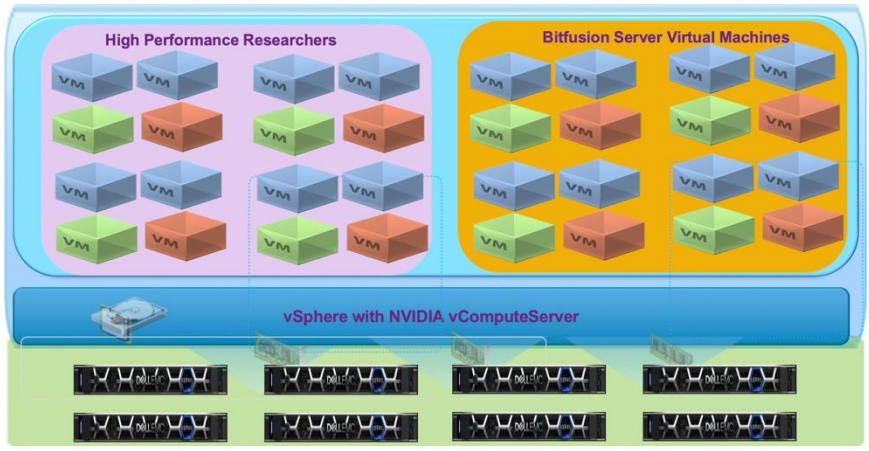

Figure 16: Sample Enterprise GPU Cluster Logical Architecture

10 Future of Machine Learning is Distributed

The quality of the prediction for complex applications require the use of a substantial amount of training data. Even though smaller machine learning models can be trained with modest amounts of data, the data and its memory requirements for training of larger models such as neural networks exponentially grow with the number of parameters. More and more GPUs are packed into physical servers to provide processing capabilities for these larger models. But his proposition is very expensive and not optimal. The paradigm for machine learning is shifting to provide the ability to scale out the processing and distributing the workload across multiple machines. Distribution Machine Learning methods like Horovod and distributed TensorFlow will dominate the future landscape with the tremendous growth of data and the complexity of the models.

11 Summary

With the rapid increase in the need for GPU computing, enterprises are seeking flexible solutions to meet the needs of data scientists, developers and other HPC users. The shared GPU workload domain solution combines the best of VMware virtualization software, Tanzu Kubernetes Grid, vSphere Bitfusion and NVIDIA vComputeServer to provide a robust yet flexible solution for GPU users. This end-to-end solution provides a

reference framework to deploy a GPU workload domain to meet the common use cases for machine learning and HPC applications.

VMWare Cloud Foundation provides a solid framework with software defined compute, storage and networking. VMware's support for accelerators such as NVIDIA GPUs combined with the vComputeServer software provides vMotion and DRS capabilities for virtual machines that have associated vGPUs in use. vSphere Bitfusion facilitates the access of GPU resources over the network providing remote users to avail of centralized GPU resources. The sample GPU cluster design discussed shows the flexibility of the solution to accommodate all common GPU use cases.