Deployment guide

VMware vSphere Bitfusion on Dell EMC PowerEdge servers

Abstract

VMware vSphere Bitfusion is a software solution that you can deploy on Dell EMC PowerEdge R740xd and C4140 servers. The solution virtualizes hardware resources to provide a pool of shared resources that are accessible to any virtual machine in the network.

May 2021

Revisions

Date	Description
September 2020	Initial release
March 2021	Document updated with support information for VMware vSphere Bitfusion 2.5
May 2021	Document updated with support information for VMware vSphere Bitfusion 3.0

Acknowledgements

Authors: Jay Engh, Chris Gully

Support: Gurupreet Kaushik, Sherry Keller

Table of contents

Revisions....2
Acknowledgements....2
Table of contents....3
Executive summary....5
1 New and enhanced in VMware vSphere Bitfusion 3.0....6
2 Audience and scope....7
3 Overview....8
4 Component overview....10
4.1 DELL EMC PowerEdge R740xd server....10
4.2 DELL EMC PowerEdge C4140 server....10
4.3 Dell EMC vSAN Ready Node R740xd....11
4.4 Dell EMC Networking S5248F-ON Switch....12
4.5 NVIDIA T4 Datacenter GPU....12
4.6 NVIDIA V100 for NVLINK and PCIe, Datacenter GPU....12
4.7 Mellanox ConnextX-5 Dual Port 10/25GbE Adapter....13
5 Pre-deployment requirements and introduction to new features....14
5.1 GPU hosts....14
5.2 Client cluster....14
5.3 Introduction to new features....15
5.3.1 Remote clients....15
5.3.2 Bare-metal server clients....15
5.3.3 Improved health checks....16
5.4 Bitfusion server and client software....16
5.5 vCenter....16
5.6 Client virtual machine....17
5.7 Connectivity....17
5.8 Network services....17
6 Solution overview....18
6.1 Architecture....18
6.2 Component information....19
6.3 VLANs and IP subnet information.....19
7 Deployment and configuration....20
7.1 Verify the GPU host hardware configuration....20
7.2 Add the GPU hosts and client cluster to the vCenter inventory....20

7.3 Prepare and configure the GPU hosts and client cluster for PVRDMA 20
7.3.1 Tag a VMkernel adapter for PVRDMA 21
7.3.2 Add and manage hosts on the Bitfusion distributed switch 22
7.4 Deploy the Open Virtual Appliance (OVA) to create the bitfusion-server-1 virtual machine 23
7.5 Edit the bitfusion-server-1 hardware settings 27
7.6 Deploy the Open Virtual Appliance (OVA) to create the bitfusion-server-2 virtual machine 32
7.7 Edit the bitfusion-server-2 virtual machine hardware settings 33
7.8 Deploy the OVA to create the bitfusion-server-3 virtual machine 35
7.9 Edit the bitfusion-server-3 virtual machine hardware settings 36
7.10 Alternate method of installing an additional Bitfusion server 38
7.11 Improved support for multiple networks 39
7.12 Provide client cluster access to GPU resources 39
7.13 Support for remote clients and bare-metal server 41
7.14 Support for backup and restore 42
8 Getting help 44
8.1 Contacting Dell EMC 44
8.2 Documentation resources 44
8.3 VMware Hands-On-Labs (HOLs) 44
8.4 Documentation feedback 44

Executive summary

Modern computational requirements have evolved and diverged with the wide acceptance and use of containers and virtualization. Today’s workloads are presenting challenges to the CPU-centric paradigm which has been a key determinant of customers' server performance and utility needs. General purpose CPUs no longer adequately handle these new workloads, whether they are artificial intelligence, machine learning or virtual desktops. However, CPU power can be enhanced with add-in graphics processing units (GPUs). GPUs leverage thousands of computing cores instead of the tens of cores that general purpose CPUs can address. While the cost of putting dedicated GPU hardware into every server is uneconomical, VMware vSphere Bitfusion on PowerEdge servers enables network delivery of these pooled resources to any properly configured client node. With Bitfusion built into the well-known vSphere infrastructure, administrators can optimize their use and increase the utilization of these expensive resources. This delivers immense value to those networked nodes which require powerful GPUs to execute massively parallel workloads.

1 New and enhanced in VMware vSphere Bitfusion 3.0

Following are some of the important features that are included with the release of VMware vSphere Bitfusion 3.0:

- New installation experience for upcoming VMware vSphere Bitfusion servers.
- Improved support for VMware vSphere Bitfusion servers with multiple networks.
- Seamless installation of the NVIDIA System Management Interface (nvidia-smi) application on VMware vSphere Bitfusion client servers.

The new features are described in the following sections of the deployment guide:

- Alternate method of installing an additional Bitfusion server.
- Improved support for multiple networks.

2 Audience and scope

This deployment guide includes step-by-step instructions for deployment and configuration of the VMware vSphere Bitfusion appliance on Dell EMC PowerEdge R740xd and C4140 rack servers.

This deployment guide makes certain assumptions about the prerequisite knowledge of the deployment personnel and the hardware they are using. This includes:

- Use of Dell EMC servers and switches including the location of buttons, cables, and components in the hardware
- Functional knowledge of the items in the Dell EMC owner's manuals for the products being used
- Use of VMware products and the components or features of VMware vSphere
- Data center infrastructure best practices in the areas of server, storage, networking, and environmental considerations such as power and cooling
- Installation, configuration and package management familiarity of CentOS
- Familiarity with NVIDIA CUDA toolkit

The scope of this document excludes existing infrastructure components outside of the specific hardware and software that is mentioned in this guide. VMware vSphere Bitfusion support is not limited to the hardware models, configuration values, and software components versions used in this document. Dell EMC takes no responsibility for any issues that may be caused to existing infrastructure during deployment.

3 Overview

With the new VMware vSphere Bitfusion software, graphics processing units (GPU) are no longer isolated from other resources. GPUs are now shared in a virtualized pool of resources and you can access them through any virtual machine in the infrastructure as shown in Figure 1. Similar to processors and storage resources, GPU deployments can now benefit from optimized utilization, reduced Capex and Opex, and accelerated development and deployment of R&D resources. Data scientists and AI developers can benefit from how Bitfusion supports monitoring workloads of higher volume.

Bitfusion offers the following key features:

- Dynamic GPU attach anywhere

  Bitfusion disaggregates your GPU compute and dynamically attaches GPUs anywhere in the datacenter, just like attaching storage.

- Fractional GPUs for efficiency

  Bitfusion enables use of any arbitrary fractions of GPUs. Support more users in the test and development phase.

- Standards based accelerator access

  Leverage GPUs across an infrastructure plus integrate evolving technologies as standards emerge.

- Application run time virtualization

  Bitfusion attaches GPUs based on CUDA calls at run-time, maximizing utilization of GPU servers anywhere in the network.

- Any application

  Bitfusion is a transparent layer and runs with any workload in a Tensorflow or Pytorch ecosystem.

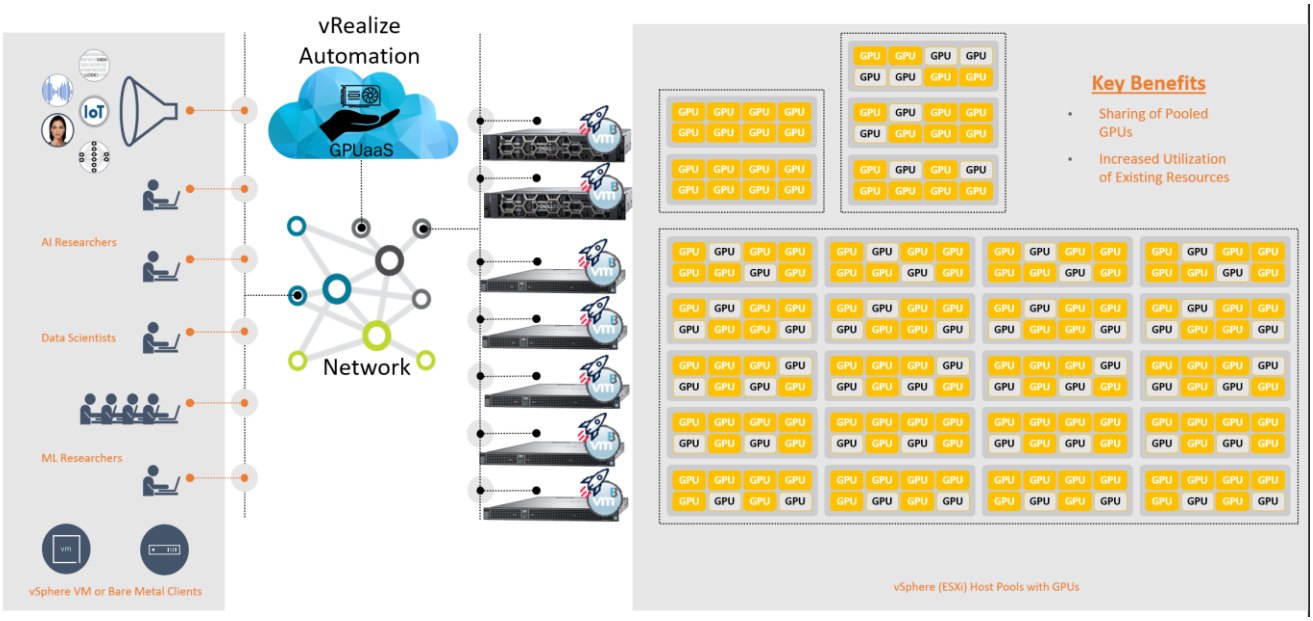

Figure 1 Bitfusion GPU sharing model

Deployment of VMware Bitfusion on the Dell EMC PowerEdge servers provide an infrastructure solution incorporating the best-in class hardware from Dell EMC with core VMware products. Virtualization of computation, storage, networking and accelerators is delivered on a cluster of PowerEdge servers. The combination of VMware vSphere Bitfusion software on the Dell EMC PowerEdge hardware described in this document has been validated in Dell EMC labs.

4 Component overview

This section briefly describes the components that support VMware vSphere Bitfusion and their key capabilities to help you deploy the software.

4.1 DELL EMC PowerEdge R740xd server

The PowerEdge R740xd server provides the benefit of scalable storage performance and data set processing. This 2U, 2-socket platform brings you scalability and performance to adapt to a variety of applications. This platform could be configured with up to 3x V100 GPUs or 6x NVIDIA T4 GPUs, but also offers the flexibility to support additional configurations such as 24x 2.5” NVMe drives and two NVIDIA GPUs. As you scale your deployments, scale your productivity with embedded intelligence and automation from iDRAC9 and the entire Open Manage portfolio that is designed to simplify the IT lifecycle from deployment to retirement.

Key capabilities:

- 24 DIMM slots of DDR4 memory (RDIMM or LRDIMM),
- Up to 24 SAS or SATA SSD or hard drive and NVMe PCIe SSDs
- Boot device options such as BOSS
- Double wide GPUs, up to 300W each, or single wide GPUs, up to 150W each

Figure 2 Front view of a Dell EMC PowerEdge R740xd

Figure 3 Rear view of a Dell EMC PowerEdge R740xd

4.2 DELL EMC PowerEdge C4140 server

PowerEdge C4140 is an incredibly dense purpose-built rack server designed to handle the most demanding technical computing workloads. With the 2nd Generation Intel® Xeon® Scalable processors and NVIDIA® Volta® technologies, the C4140 fills a key gap as a leading GPU-accelerated platform in the PowerEdge server portfolio to enable a scalable business architecture in a heterogeneous data center environment. With four double-width accelerators in just 1U of space, the C4140 delivers outstanding performance and maximum density while reducing your space, cost and management requirements.

Key capabilities:

- Unthrottled performance and superior thermal efficiency with patent-pending interleaved GPU system design*
- No-compromise (CPU + GPU) acceleration technology up to 500 TFLOPS / U+ using the NVIDIA® Tesla™V100 with NVLink™
- 2.4KW PSUs help future-proof for next generation GPUs

Figure 4 Front view of a Dell EMC PowerEdge C4140

Figure 5 Rear view of a Dell EMC PowerEdge C4140

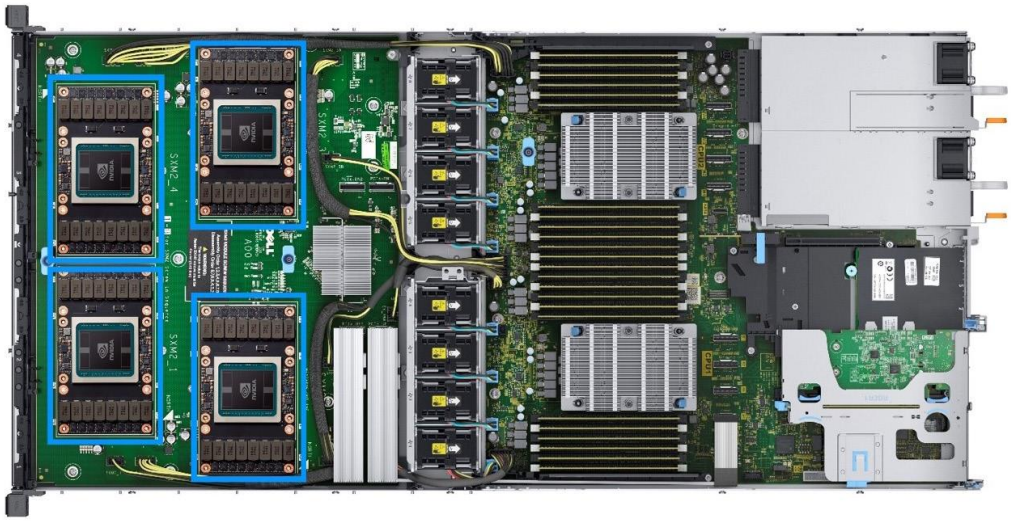

Figure 6 Internal view of a Dell EMC PowerEdge R740xd displaying a NVIDIA T4 GPU card

4.3 Dell EMC vSAN Ready Node R740xd

Dell EMC vSAN Ready Nodes are pre-configured building blocks that reduce deployment risks with certified configurations, improve storage efficiency by up to 50%, and can help you build or scale your vSAN cluster faster. Whether you're just getting started, and/or expanding your existing VMware environment, Dell EMC is here for you every step of the way with consulting, education, deployment and support services for the entire solution.

The Dell EMC vSAN Ready Node R740xd is a two socket, 2U rack servers designed to run complex workloads using highly scalable memory, I/O capacity and network options. The vSAN RN R740xd is available in All-Flash and Hybrid configurations and features the latest Generation Intel® Xeon® Scalable processor family. Being a vSAN RN all approve configurations can be access in the VMware Compatibility Guide. The vSAN RN R740xd adds extraordinary storage capacity options, making it well-suited for data-intensive applications that require greater storage, while not sacrificing I/O performance. Being that these are ready nodes, they take the guesswork and hassle out of procurement, deployment and management.

4.4 Dell EMC Networking S5248F-ON Switch

The S5200F-ON series introduces optimized 25GbE and 100GbE open networking connectivity for servers/storage in demanding web and cloud environments. Innovative next-generation top-of-rack family of 25GbE switches providing optimized performance both in-rack and between-racks, cost-effective 50/100GbE leaf/spine fabric, and migration capabilities for future connectivity needs.

Key capabilities:

- 48 port 10/25GbE SFP28 auto negotiating ports
- 4 port 100GbE QSFP28
- 2 port 2 x 100 QSFPDD-28

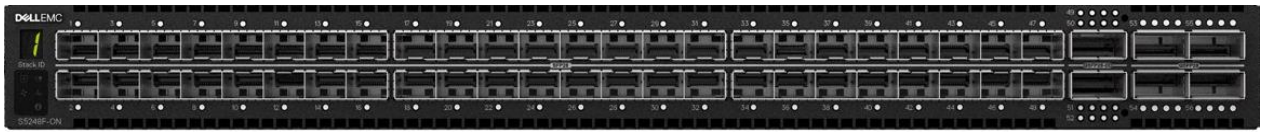

Figure 7 Front view of Dell PowerSwitch S5248F-ON

4.5 NVIDIA T4 Datacenter GPU

The NVIDIA® T4 is a single-slot, low-profile, 6.6-inch PCI Express Gen3 Universal Deep Learning Accelerator based on the TU104 NVIDIA graphics processing unit (GPU). The T4 has 16 GB GDDR6 memory and a 70 W maximum power limit. The T4 is offered as a passively cooled board that requires system air flow to operate the card within its thermal limits.

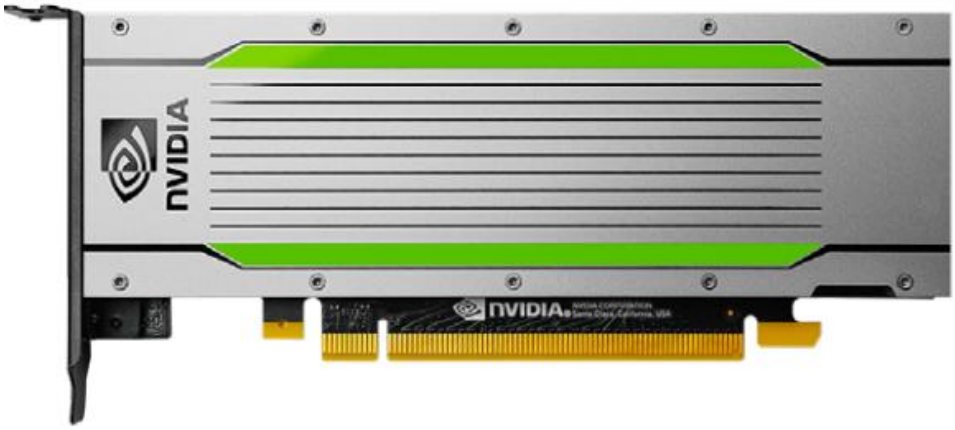

Figure 8 NVIDIA T4 GPU

4.6 NVIDIA V100 for NVLINK and PCIe, Datacenter GPU

The NVIDIA V100 GPU powered by NVIDIA Volta architecture is the most widely used accelerator for scientific computing and artificial intelligence. NVIDIA® V100 Tensor Core is the most advanced data center GPU ever built to accelerate AI, data science. It's powered by NVIDIA Volta architecture, comes in 16 and 32GB configurations, and offers the performance of up to 32 CPUs in a single GPU. Deep Learning training workloads can leverage NVLink capability of the V100 SXM2 GPUs on the C4140 with NVLink. Using the V100 SXM2 GPU with the NVLink capabilities enables direct communication between GPUs with bandwidth of up to 300GB/s; further increasing performance of AI training workloads.

Figure 9 NVIDIA V100 for PCIE

Figure 10 NVIDIA V100 for NVLink

4.7 Mellanox ConnextX-5 Dual Port 10/25GbE Adapter

ConnectX-5 EN supports two ports of 25Gb Ethernet connectivity, sub-600 ns latency, and very high message rate, plus PCIe switch and NVMe over Fabric offloads, providing the highest performance and most flexible solution for the most demanding applications and markets: Machine Learning, Data Analytics, and more.

Key capabilities:

- Up to 25 Gb/s connectivity per port
- Industry-leading throughput, low latency, low CPU utilization and high message rate
- RoCE for Overlay Networks

Figure 11 Mellanox ConnextX-5 Dual Port 10/25GbE Adapter

5 Pre-deployment requirements and introduction to new features

This section describes the pre-deployment requirements for configuring VMware vSphere Bitfusion. Figure 13 shows an illustration of the components involved in creating the Bitfusion client-server cluster.

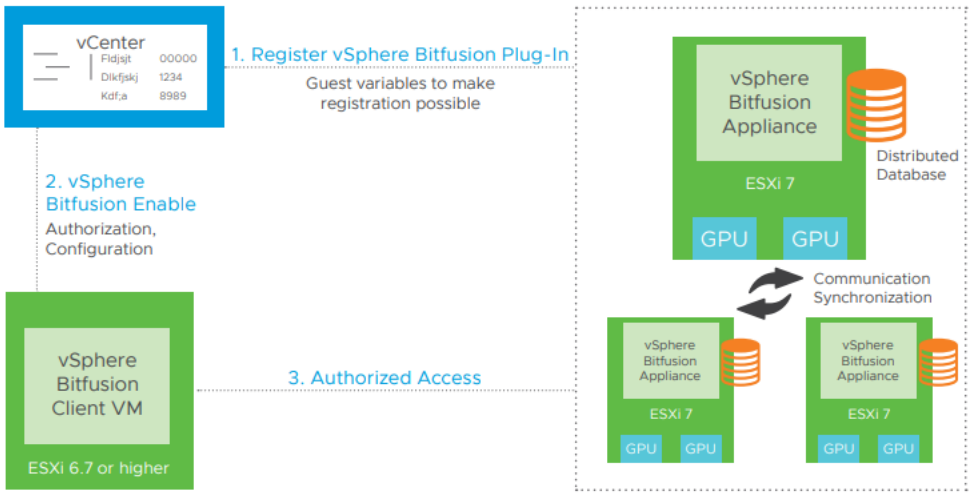

Figure 12 Bitfusion sample client-server cluster

5.1 GPU hosts

The Bitfusion OVA is deployed on the GPU hosts. Below are the Dell EMC customized images for VMware ESXi 7.0 that support Bitfusion OVA deployed on the GPU hosts:

Dell EMC customized image of VMware ESXi 7.0 installed on a PowerEdge R740xd server with NVIDIA T4, NVIDIA V100, Mellanox RoCE v2 compatible network adapter.

Dell EMC customized image of VMware ESXi 7.0 installed on a PowerEdge C4140 sxm2 chassis server with four NVIDIA V100, Mellanox RoCE v2 compatible network adapter. The GPU chassis configuration for the PowerEdge C4140 does not accommodate any redundant local storage. You can create a shared Network File System (NFS) and mount it as a datastore on this host.

For information about the network cards that are supported for PVRDMA configuration, see the VMware Compatibility Guide page. Select **RoCE v1** and **RoCE v2** options from the **Features** tab and then select the **DELL** option from the **Brand Name** tab. For more information on how to download Dell EMC customized ESXi images, see Dell Knowledge Base article SLN288152.

5.2 Client cluster

vSAN cluster comprises of virtual machines that use the GPU for data analytics, training models or for running an inference. However, a vSAN cluster is not required for the client virtual machine.

PVRDMA is supported on VMware vSAN 7.0 if deployed on four vSAN ReadyNode nodes of PowerEdge R740xd server with a compatible network adapter. With a dual port network adapter for management, vSAN

and vMotion traffic can be enabled. For information on certified and supported vSAN ReadyNode, see the VMware Compatibility Guide.

5.3 Introduction to new features

5.3.1 Remote clients

VMware vSphere Bitfusion 2.0 allows you to run client servers that are a part of the same vCenter software as of the Bitfusion server. With VMware vSphere Bitfusion 2.5, you can generate tokens and install the clients on virtual machines or server containers that are a part of adjacent vCenter clusters. This option provides for more flexibility and reach that is necessary to allow any virtualized workload to have access to one or many GPUs. For more information on token generation, see Support for remote clients and bare-metal server.

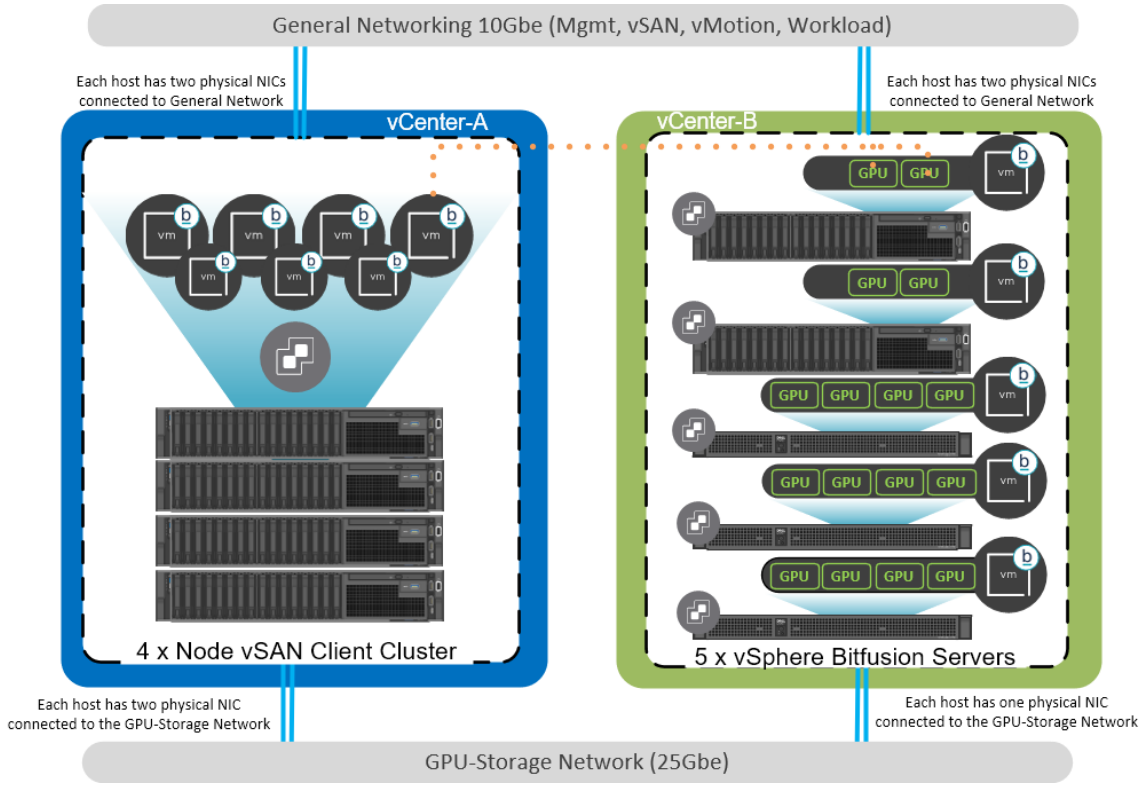

Figure 13 Clients instaled on VMs that are a part of adjacent vCenter clusters

5.3.2 Bare-metal server clients

Remote clients that can generate Bitfusion tokens from the vCenter plug-in has encouraged a more flexible implementation model for resources that require GPU access. Applications and workloads that are a part of bare-metal servers must be converted to virtual machines, an obvious performance drawback.

VMware vSphere Bitfusion 2.5 directly addresses this issue and provides an easier path to make use of GPU resources for many servers that cannot be retrofitted to handle a native GPU due to thermal and power constraints and client enablement allows bare-metal workloads to maintain their current support and keep operating while providing the workloads access to greater application performance. This results in better outcomes in advanced computing workloads such as artificial intelligence or machine learning. For more information on bare-metal server clients, see Support for remote clients and bare-metal server.

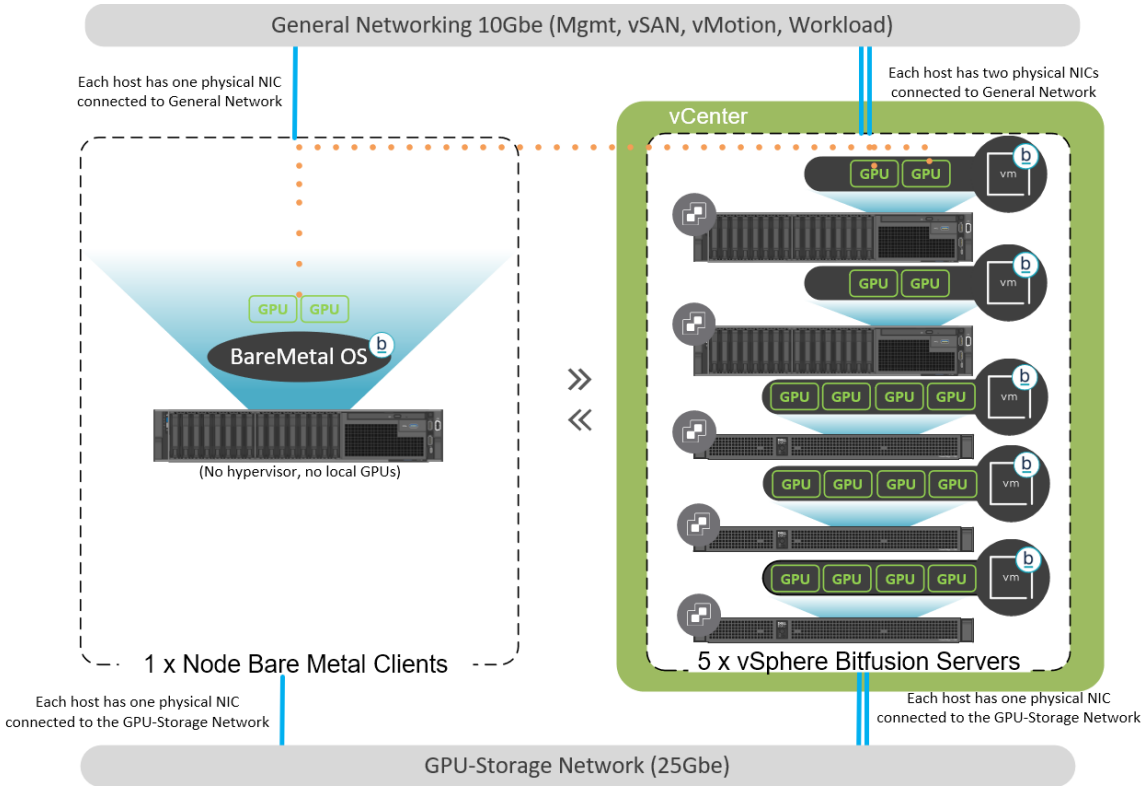

Figure 14 Improved management of bare-metal server workloads

5.3.3 Improved health checks

In earlier releases of vSphere Bitfusion, health checks are performed with a simple command-line parameter. With the integration of vCenter, a new graphical user interface is introduced to display the command-line outputs within a dialog box. A new method is introduced for suppressing the health checks when it is not compatible with your server environment configuration or if the health checks do not reflect the server environment's current state. For more information on the improvements made to health checks, see VMware vSphere Bitfusion 2.5 release article.

5.4 Bitfusion server and client software

Bitfusion OVA is a VMware appliance prepackaged with GPU software and services. Bitfusion client package runs on the virtual machines where, the applications make use of the GPU resources. To download the OVA and client package, see the Download VMware vSphere Bitfusion page after logging into My VMware account.

5.5 vCenter

Once the Bitfusion server OVA is deployed, select **Bitfusion** from the **vCenter** menu. vCenter Server 7.0 lists the components connected to the server once the GPU hosts and client clusters are connected. By doing this, embedded platform services are installed on the client cluster. To download the vCenter Server Appliance, see the Download VMware vSphere page.

Note: You can download the vCenter Server Appliance from the Download VMware vSphere page after logging in to My VMware account.

5.6 Client virtual machine

The client cluster has a virtual machine with CentOS installed with the required NVIDIA tools and drivers. You can use this virtual machine to access the GPUs remotely.

Install the following components on the CentOS virtual machine to set up the client cluster:

- Python 3 and pip3 package manager
- Compute Unified Device Architecture 10.0 Toolkit (CUDA) for Red Hat Enterprise Linux 7
- cuDNN 7 python library
- TensorFlow v1.13.1 GPU framework
- TensorFlow benchmark toolkit compatible with TensorFlow v1.13 framework

This client virtual machine is connected to the management and RDMA network through PVRDMA.

Note: For instructions to deploy the above pre-requisites on a client virtual machine, see the Running TensorFlow on vSphere Bitfusion vSphere Bitfusion guide.

5.7 Connectivity

The Bitfusion server appliances, client virtual machine with the remote GPUs and vCenter are connected over a dedicated management network. In addition to this, vSAN, vMotion and Hardware Acceleration communication are all required to be connected to the client cluster.

To monitor the GPU traffic, a dedicated RDMA (RoCE) connection is established between the GPU hosts and the client cluster hosts.

The Dell EMC PowerSwitch ToR is configured for VLANs to accommodate vSAN, vMotion and GPU data traffic management. Two switches are set up with Virtual Link Trucking (VLT) for redundancy.

Route the Bitfusion Appliance management network subnet to access the internet and then download the NVIDIA driver.

5.8 Network services

Domain Name Service (DNS) is required to fetch both forward and reverse name resolution. The IP addresses of name servers, search domains, and hostnames of all the Bitfusion appliance virtual machines should be tested and verified for both forward and reverse lookups. Test the DNS entries using their Fully Qualified Domain Name (FQDN) and their short name or hostname.

Time synchronization is critical to the Bitfusion server appliances. All the GPU hosts, client clusters and Bitfusion appliance virtual machines are synchronized to a reference time source. Network Time Protocol (NTP) traffic is routed from client to source or it can travel over the same L2 network.

6 Solution overview

Following is the solution overview for the deployment instructions provided in the rest of the document.

6.1 Architecture

The GPU hosts and client cluster architecture shown in Figure 14 is the reference architecture for the use case described in the forthcoming section.

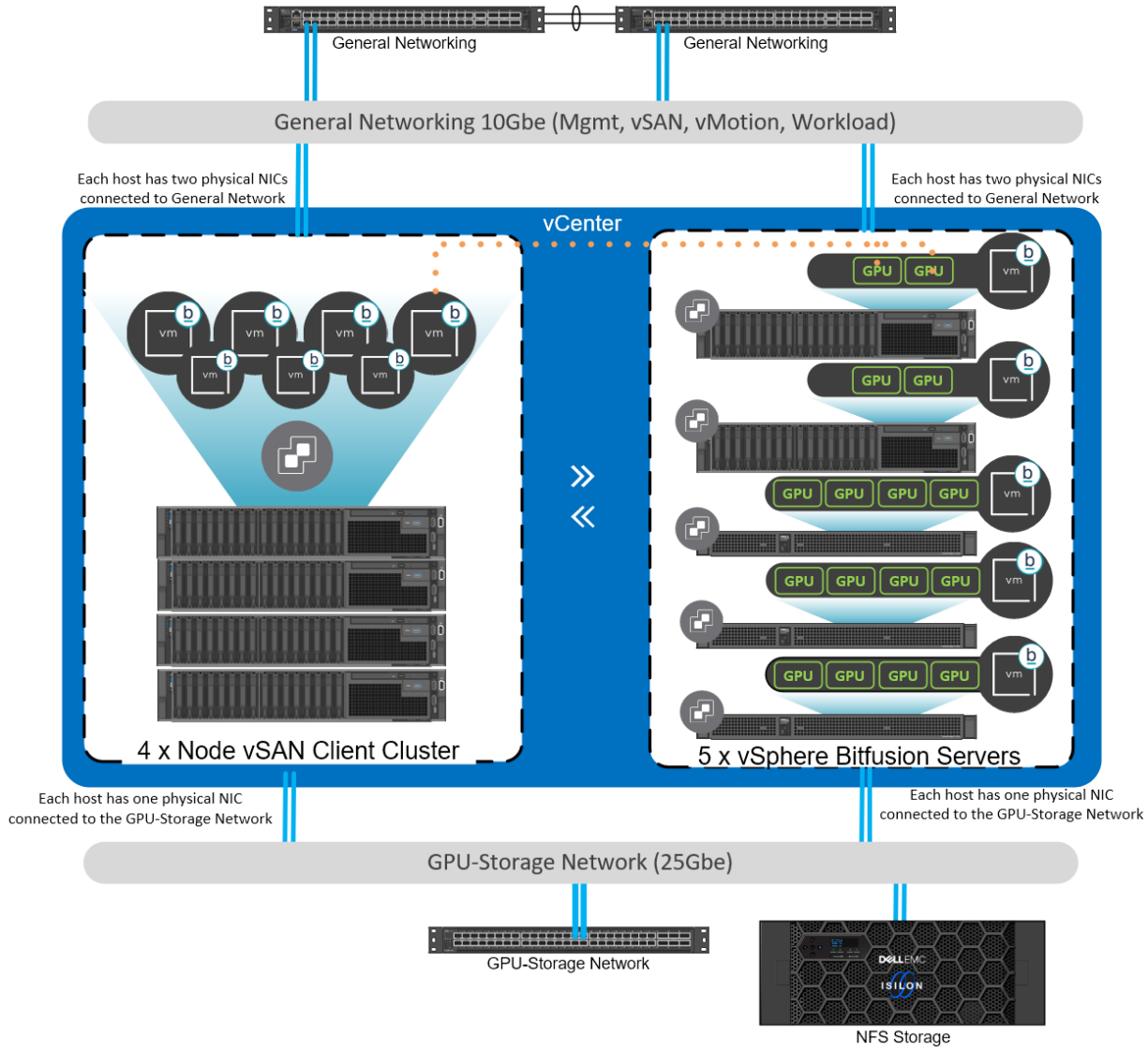

Figure 15 GPU hosts and client cluster architecture

6.2 Component information

Table 1 Hardware information that is used in this deployment guide.

Component	Description	Specification
PowerEdge 740xd	GPU Host	2 * Intel(R) Xeon(R) Gold 6242 CPU @ 2.80GHz RAM 256GB Mellanox ConnectX-4 LX 25GbE SFP Boot Device Dell BOSS-S1 2*120GB SATA SSD
PowerEdge C4140	GPU Host	2 * Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz RAM 384GB Mellanox ConnectX-5 LX 25GbE SFP Boot Device Dell BOSS-S1 2*120GB SATA SSD
vSAN RN 740xd	Client Cluster	4 Identical Nodes in a cluster 2 * Intel(R) Xeon(R) Gold 6242 CPU @ 2.80GHz Boot Device Dell BOSS-S1 2*120GB SATA SSD Intel(R) 10GbE 4P X710 rNDC Mellanox ConnectX-5 LX 25GbE SFP 2 x 1.8TB GB SAS SSD 6 x 1.2 TB SATA SSD
NVIDIA T4	Accelerator	NVIDIA Turing Tensor cores: 320 GPU Memory: 16GB GDDR6 300 GB/sec System Interface: x16 PCIe Gen3
NVIDIA V100 PCIe	Accelerator	NVIDIA Volta Tensor cores: 640 GPU Memory: 16GB HBM2 900 GB/sec System Interface: x16 PCIe Gen3
NVIDIA V100 SXM2	Accelerator	NVIDIA Volta Tensor cores: 640 GPU Memory: 16GB HBM2 900 GB/sec System Interface: NVIDIA NVLink
S5248F	Network Switch	2 * ToR configured with VLT Dell Networking OS10
Client VM	Client	CentOS 7 (64bit) 4 vCPUs, 64GB RAM 2* network adpaters
Hypervisor	GPU Host	VMware ESXi, 7.0.0, 15843807 or later vSphere Enterprise Plus with Add-on for Bit fusion
Hypervisor	Client host	VMware ESXi, 7.0.0, 15843807 or later VSphere Enterprise Plus

6.3 VLANs and IP subnet information.

Table 2 VLANs and IP subnet information

VLAN ID	Function	Subnet	Gateway
96	Management	100.71.x.x/21	100.71.x.x
90	PVRDMA	172.16.6.0/24	
10	vSAN	172.16.4.0/24	
20	vMotion	172.16.5.0/24	

7 Deployment and configuration

The following section describes the step-by-step instructions to deploy Bitfusion appliances on the GPU hosts and a quick test to showcase the GPU accessibility by a remote VM assuming all the pre-requisites are met.

7.1 Verify the GPU host hardware configuration

Follow the steps:

1. Login to iDRAC and browse to the Configuration > BIOS settings > System Profile Settings and verify that the System Profile is set to Performance.
2. Browse to the Configuration > BIOS Settings > Integrated Devices and verify that the Memory Mapped I/O above 4GB is set to Enabled.

7.2 Add the GPU hosts and client cluster to the vCenter inventory

Follow the steps:

1. Create folders for GPU Hosts and Client Cluster (TKG).
2. Add the hosts with GPUs to the GPU Hosts folder and the PowerEdge R740xd vSAN Ready Nodes to the TKG folder.
3. Configure DRS, Hardware Acceleration and vSAN on the client cluster using Quickstart under the configuration menu.

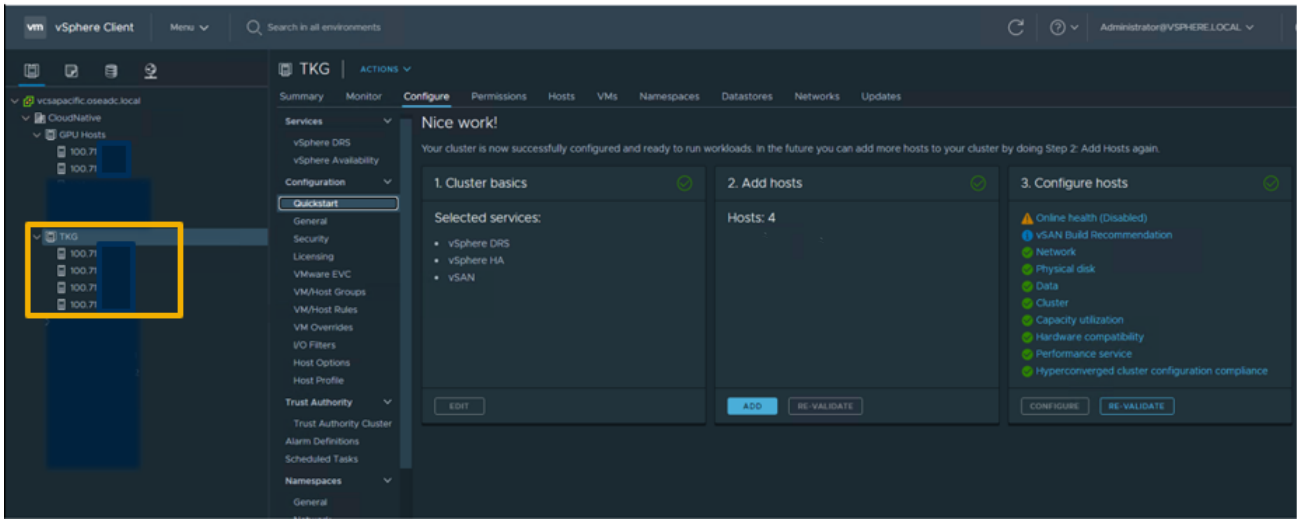

Figure 16 Adding GPU hosts and client cluster to the vCenter inventory.

7.3 Prepare and configure the GPU hosts and client cluster for PVRDMA

Follow the steps:

1. Verify the listing of the RoCE v2 compatible RDMA network adapters under the RDMA Adapters menu of each host. To do this, click on **Configure** > **Networking** > **RDMA Adapters**.
2. Create a distributed switch with the name **Bitfusion** using the **New Distributed Switch** wizard. Set the **switch version** to **v7.0.0** and configure two uplink ports.

3. Create **pvrdma** as a new distributed portgroup pvrdma using the **New Distributed Port Group** wizard under the Bitfusion distributed switch created in the previous step. Set the **Port Binding** to **ephemeral** and **VLAN** to **90**.
4. Create VMKernel port for PVRDMA on all hosts by clicking on **Configure** > **Networking** > **VMkernel Adapters** > **Add Networking** wizard. Create **vmk1** on the **GPU hosts** and then create **vmk3** on the client cluster.
5. Assign the created Bitfusion distributed switch to the **pvrdma** port group.
6. Set **MTU** to **9000** and set the **TCP/IP stack** to **Default**.
7. Assign the IPv4 address to static IP in the 172.16.6.x/24 subnet.

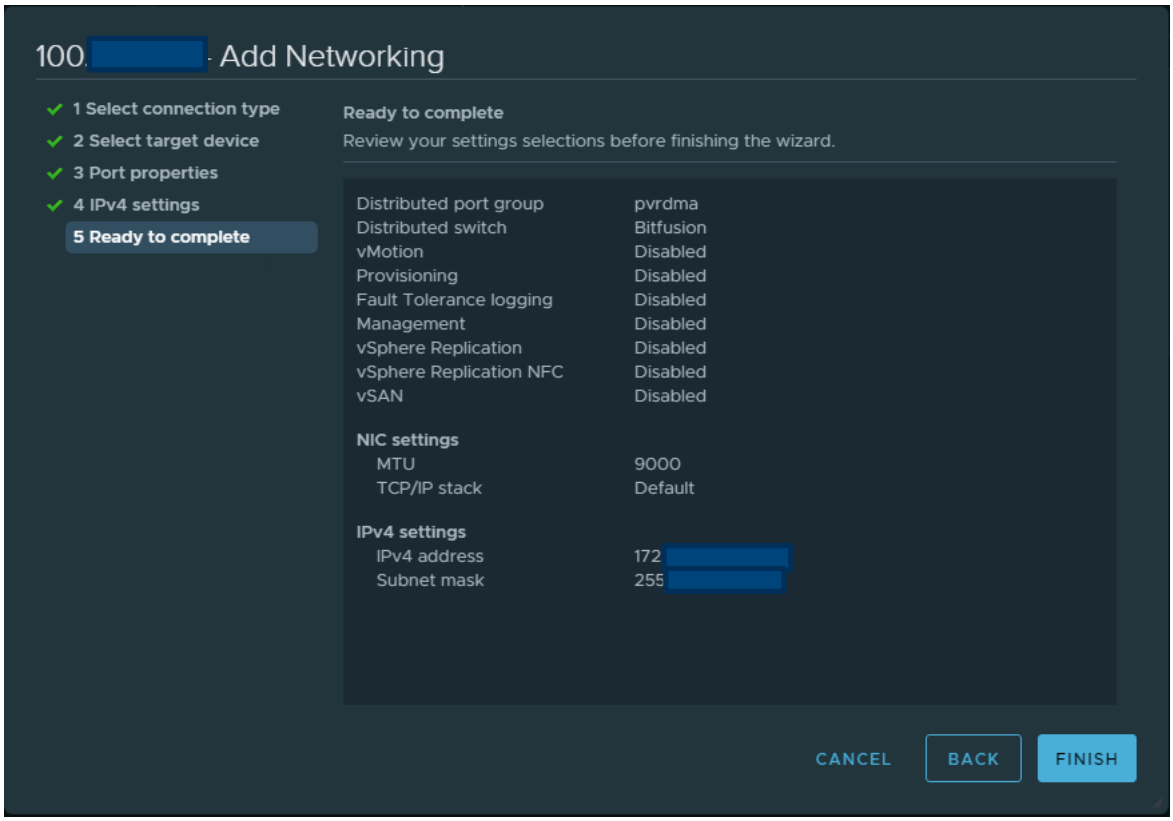

Figure 17 Creating the VMKernel port for PVRDMA on all hosts

7.3.1 Tag a VMkernel adapter for PVRDMA

Follow the steps:

1. Navigate to the host on the vSphere web client.
2. Click on Configure > System > Advanced System Settings > Edit
3. Locate **Net.PVRDMAvmknic** from the list and click on **Edit**
4. Enter the value **vmk1** on the GPU hosts and **vmk3** on Client cluster hosts and click **OK** to complete tagging the VMkernel adapter.

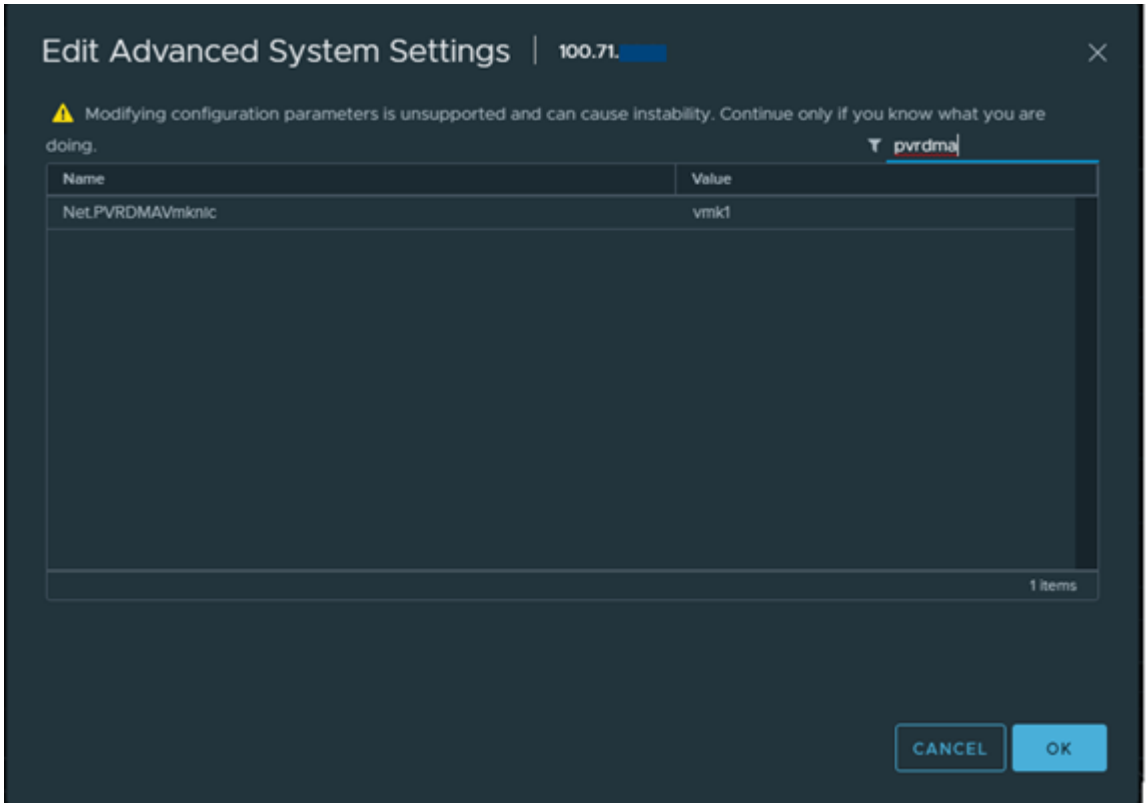

Figure 18 Tagging the VMkernel adapter

7.3.2 Add and manage hosts on the Bitfusion distributed switch

Follow the steps:

1. Attach the GPU hosts and client cluster hosts.
2. Assign the two RDMA NICs (vmnic6 & vmnic7) from each host.
3. Assign the vmkernel port (vmk3 on client cluster hosts and vmk1 on GPU hosts) on each host created in the earlier step to **pvrdma** port group.

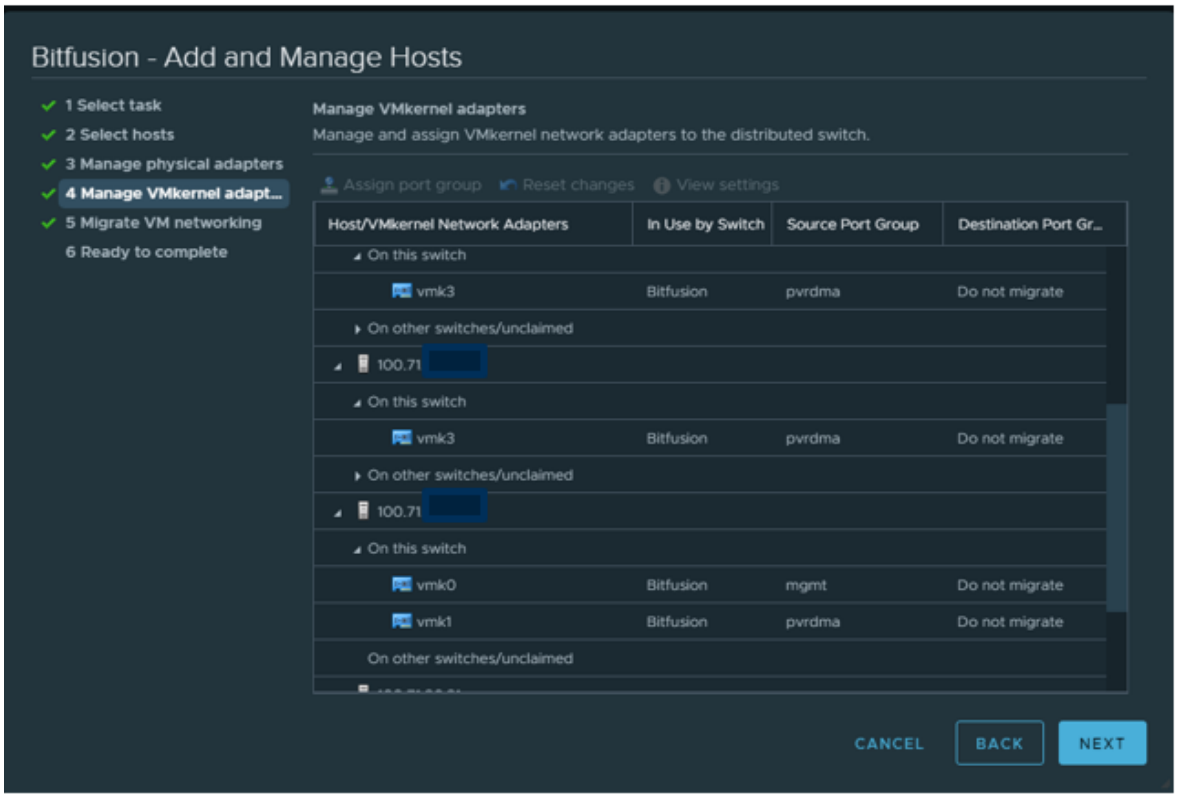

Figure 19 Adding and managing hosts on the Bitfusion distributed switch.

7.4 Deploy the Open Virtual Appliance (OVA) to create the bitfusion-server-1 virtual machine

Follow the steps:

1. Use **Deploy OVF Template** action on the C4140 to deploy the first appliance.
2. Select the appliance OVA file, **bitfusion-server-2.0.0-11.ova**.
3. Select the folder and provide the name for the first appliance **bitfusion-server-1**.
4. Select **the GPU Host C4140**, review the details and proceed to select the **NFS datastore** for appliance storage.
5. Select **port group (mgmt)** from the **dropdown menu** for the management traffic that connects to vCenter.

Note: The OVA has an option to configure a single destination network during the OVF deployment. We need to configure the additional network for pvrdma when customizing the template later.

6. Provide the hostname **bitfusion1** that has an entry in the DNS server for name resolution.
7. Extract the vCenter GUID from the vSphere client URL found in the navigation bar of the browser.
8. Enter the vCenter URL, https://100.71.x.x, vCenter username administrator@vsphere.local and password for the administrator account.

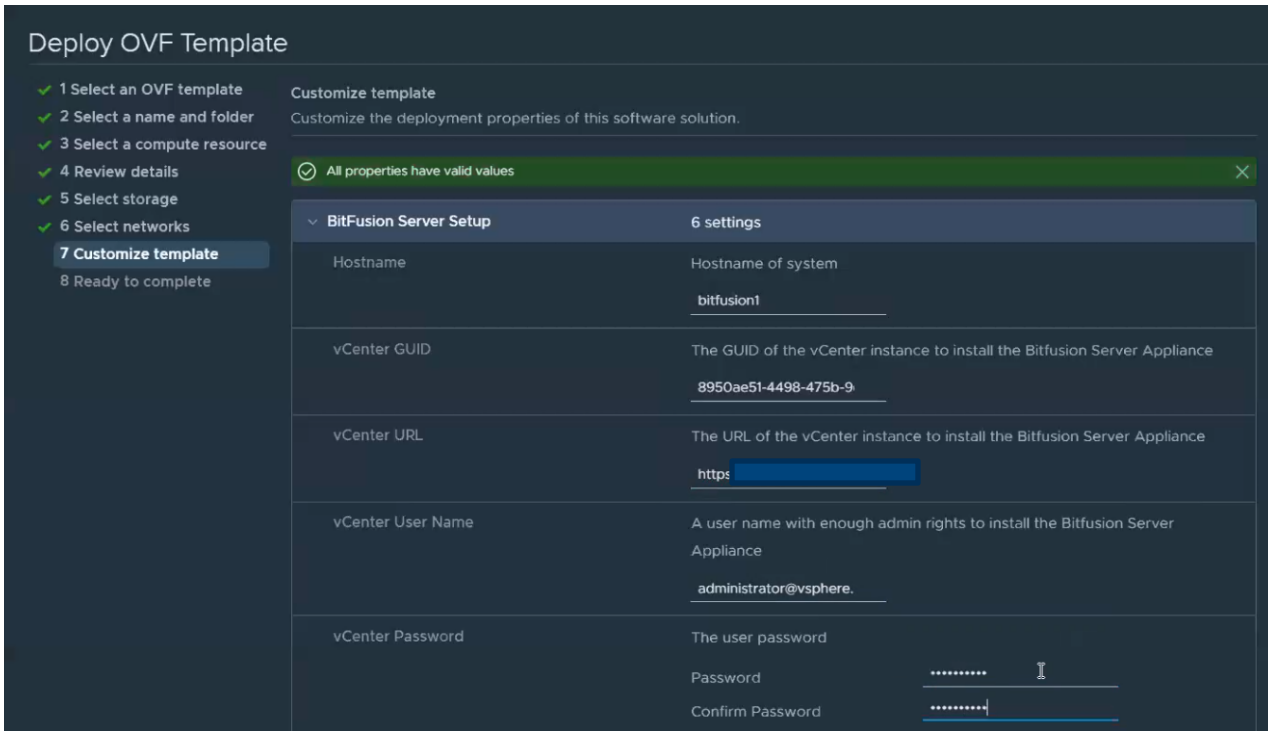

Figure 20 Enter the vCenter URL and the vCenter administrator account credentials

9. Extract the TLS certificate from the browser navigation pane. The hexadecimal is case sensitive.

    a. For Google Chrome: click on the **lock** icon or **not secure** icon to the left of the URL bar in the browser and then click **Certificate** > **Details** > **Thumbprint**.
    b. For Mozilla Firefox: click on the **lock** icon to the left of the URL bar in the browser and then expand **Connection (secure or not secure)** > **More Information** > **View Certificate > scroll to fingerprints** and select **SHA-1**.

Note: Ensure that the thumbprint information captured using chrome browser has a delimiter ':' after every two characters and the alphabets in the hexadecimal code are in capital case E.g., FE:C8:F0:05:1D:C0:69:E5:BE:6C:42:78:8D:BE:8A:32:C3:7D:37:6D

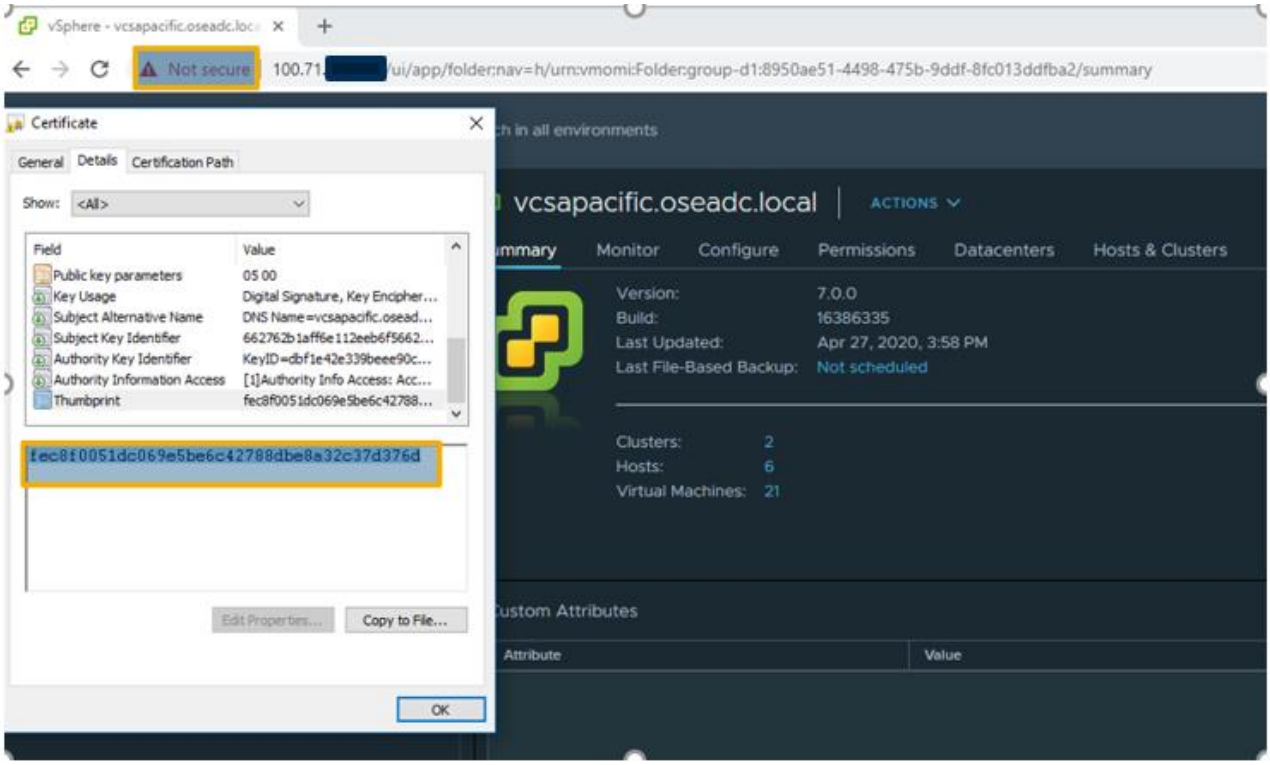

Figure 21 Extracting the TLS certificate from Google Chrome web browser navigation pane

10. Provide credentials for user customer. This account is used to login to the appliance for any troubleshooting.
11. **Select the checkbox** for **NVIDIA driver license agreement**. The appliance has connectivity to the internet to download the NVIDIA software.

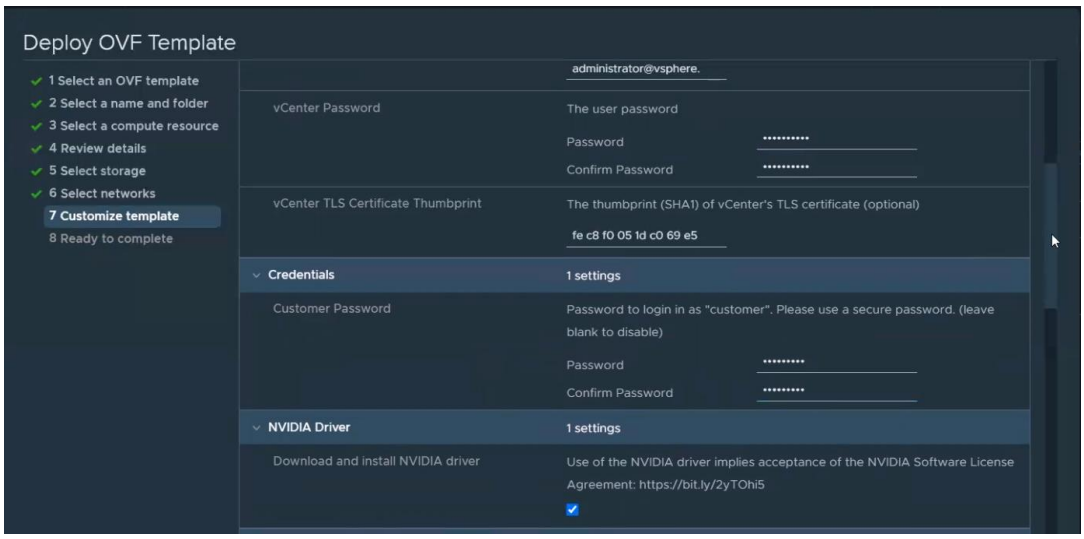

Figure 22 Select the checkbox for NVIDIA driver license agreement

12. Provide information for Network Adapter 1 settings. This provides the appliance access to the management plane (vCenter) and access to internet to download the NVIDIA driver.

- IPv4 address – 100.71.x.x
- CIDR – 21

- MTU – 9000
- Gateway – 100.71.x.x
- DNS – 100.71.x.x
- Search domain – oseadc.local
- NTP – 100.71.x.x

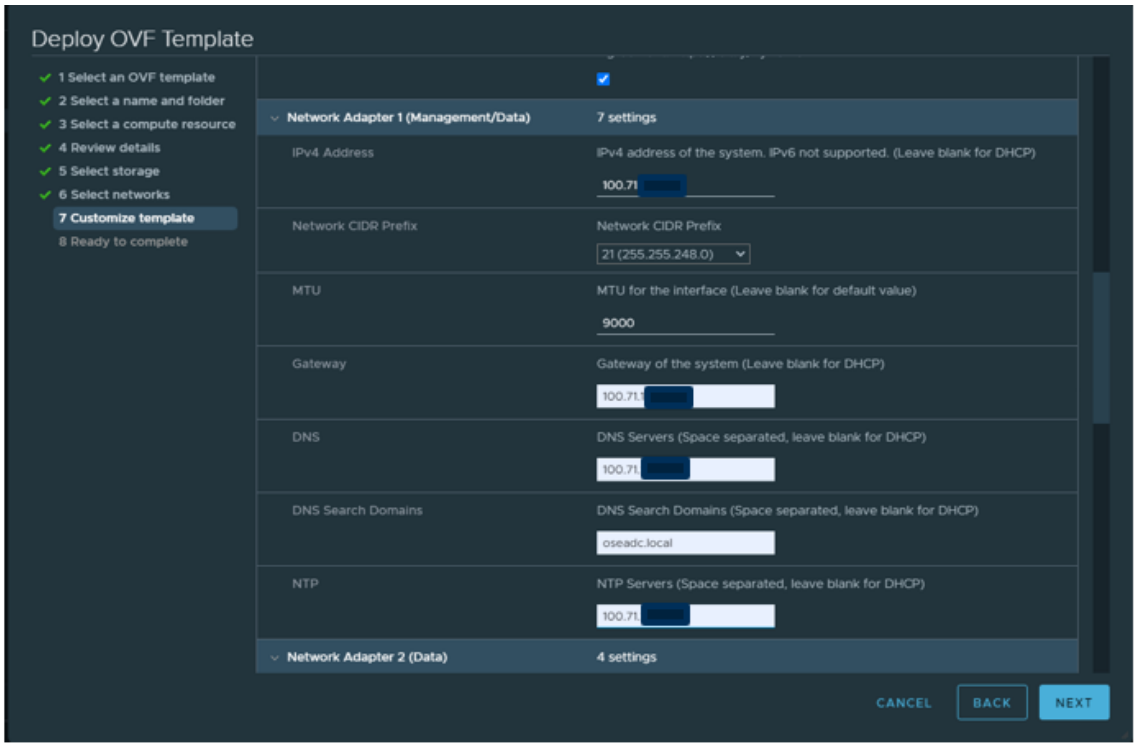

Figure 23 Configuring the network adapter 1 settings

13. Provide information for Network Adapter 2 settings. This provides the appliance access to the data plane (pvrdma) for GPU traffic. Select the **checkbox** for configuring network adapter 2 settings.

- IPv4 address – 172.[redacted]
- CIDR – 24
- MTU – 9000

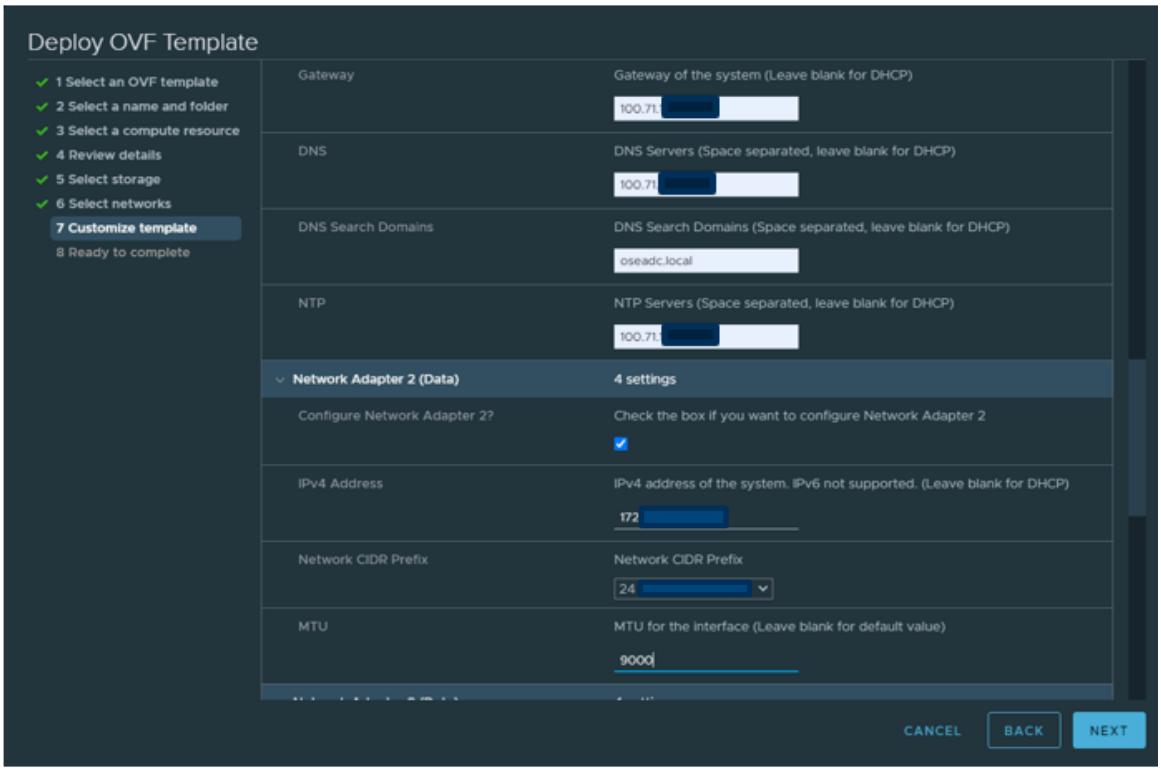

Figure 24 Configure the network adapter 2 settings and select the checkbox

14. Click **Next** to complete the deployment configuration and wait for the task to complete. Refrain from powering on the virtual machine.

7.5 Edit the bitfusion-server-1 hardware settings

Follow the steps:

1. Under the **Virtual Hardware tab**, verify if the number of **vCPUs** is **8**

- Minimum No. of vCPUs = 4x No. of GPU devices attached to the appliance. In this case, 4 x 2 GPUs i.e. 8

2. Verify if the **memory** is set to **48GB** and select the checkbox **Reserve all guest memory**.

- Minimum GB of memory = 1.5x aggregate total of GPU memory on all GPU cards passed through. In this case, 1.5x 32GB i.e. 48GB

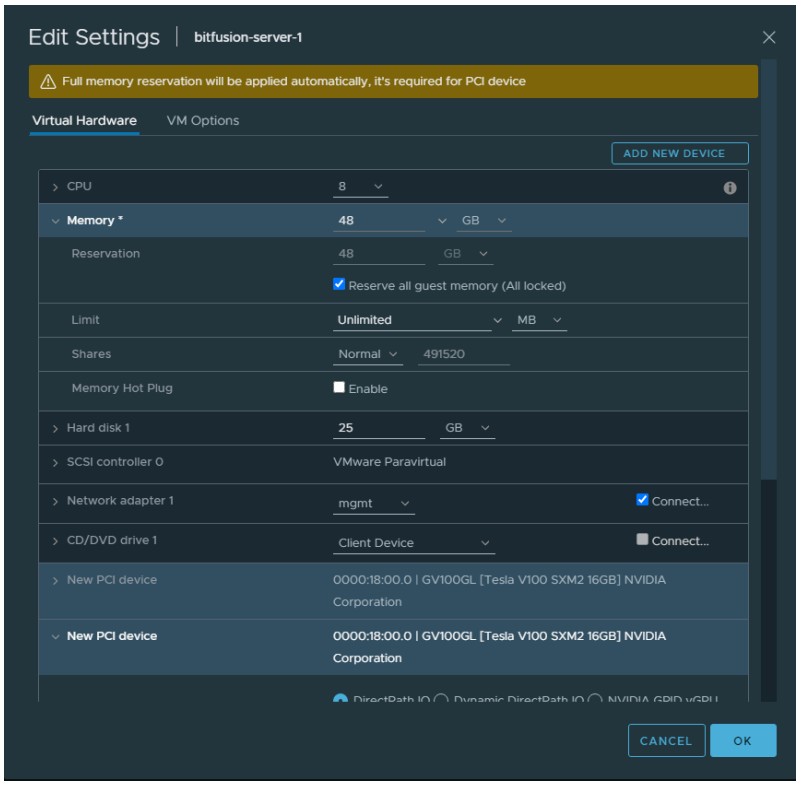

Figure 25 Editing the bitfusion-server-1 hardware settings

3. Click on **Add New Device** and add two PCIe devices. Select the **PCI devices** from the **drop-down menu**.

- 0000:18:00.0 | GV100GL V100 SXM2 16GB
- 0000:3b:00.0 | GV100GL V100 SXM2 16GB

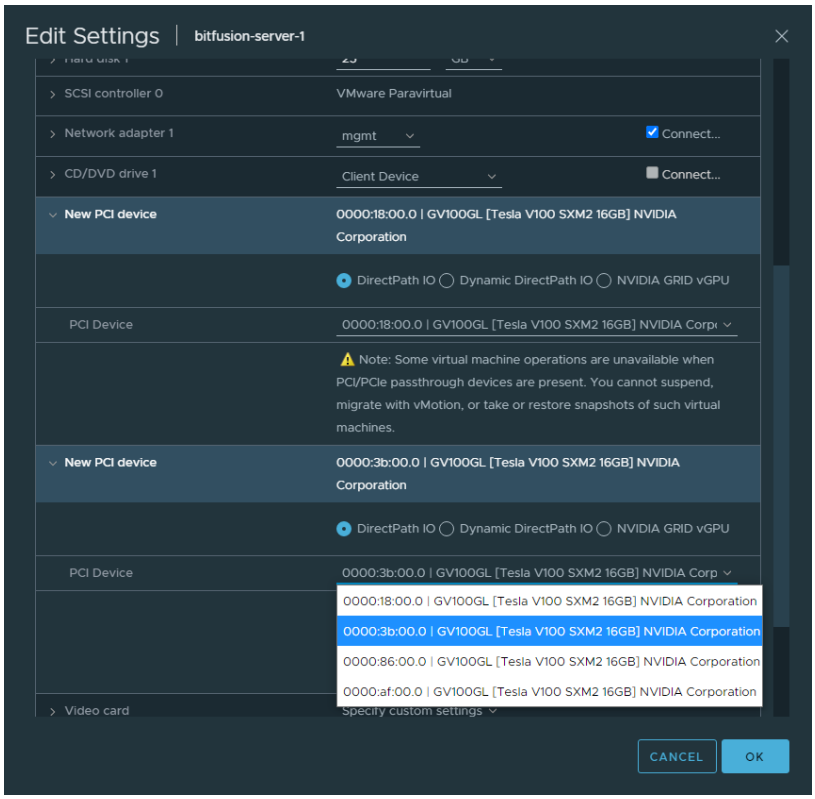

Figure 26 Adding PCI devices from the drop-down menu

4. Click on Add New Device and select Network adapter.

- Browse and select the **network pvrdma** that is created on the **Bitfusion distributed switch** earlier.
- Select the **Adapter Type PVRDMA** and **Device Protocol RoCE v2**

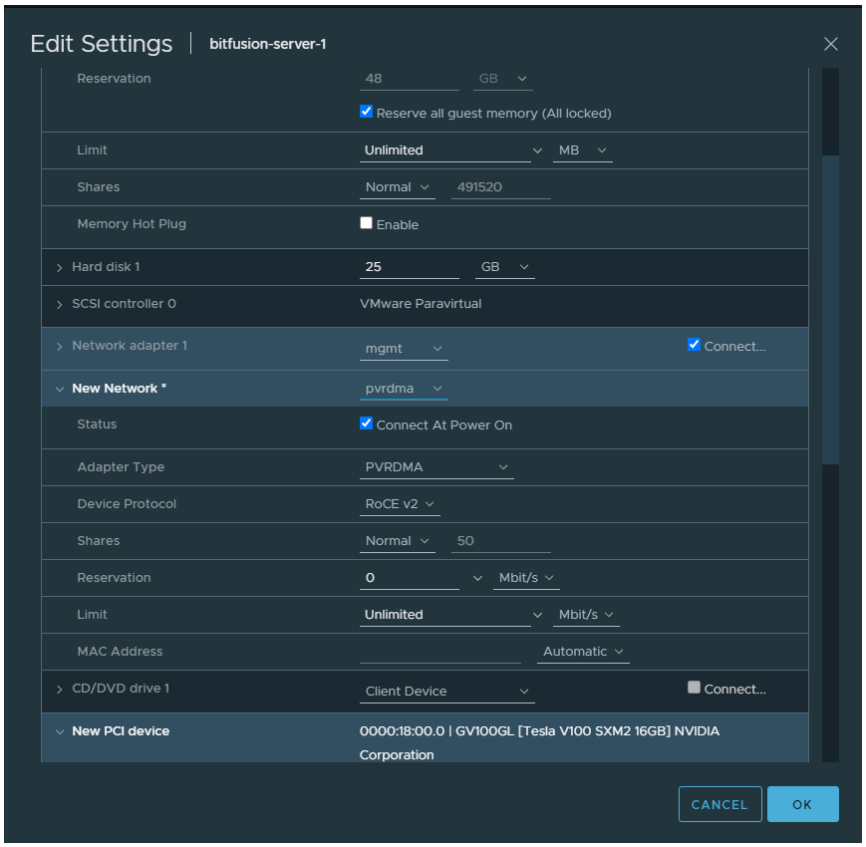

Figure 27 Setting the adapter type as PVRDMA the device protocol to RoCE v2

5. Under the Virtual Machine Options tab, select Advanced > Configuration parameters and then select Edit Configuration. Edit the pciPassthru.64bitMMIOSizeGB parameter to 64.

- pciPassthru.64bitMMIOSizeGB= <n>, where n equals (num-cards * size-of-card-in-GB) rounded up to NEXT power of 2. In this case, 2x 16 → 32, rounded to next power of 2 i.e. 64

6. Finish the configuration changes and power on the virtual machine. Wait for the virtual machine to boot up and run through the initial configuration. This might take 10-15 minutes. A blue alert bar along the top of vCenter (version 7) will appear indicating the Bitfusion plugin is successfully deployed.

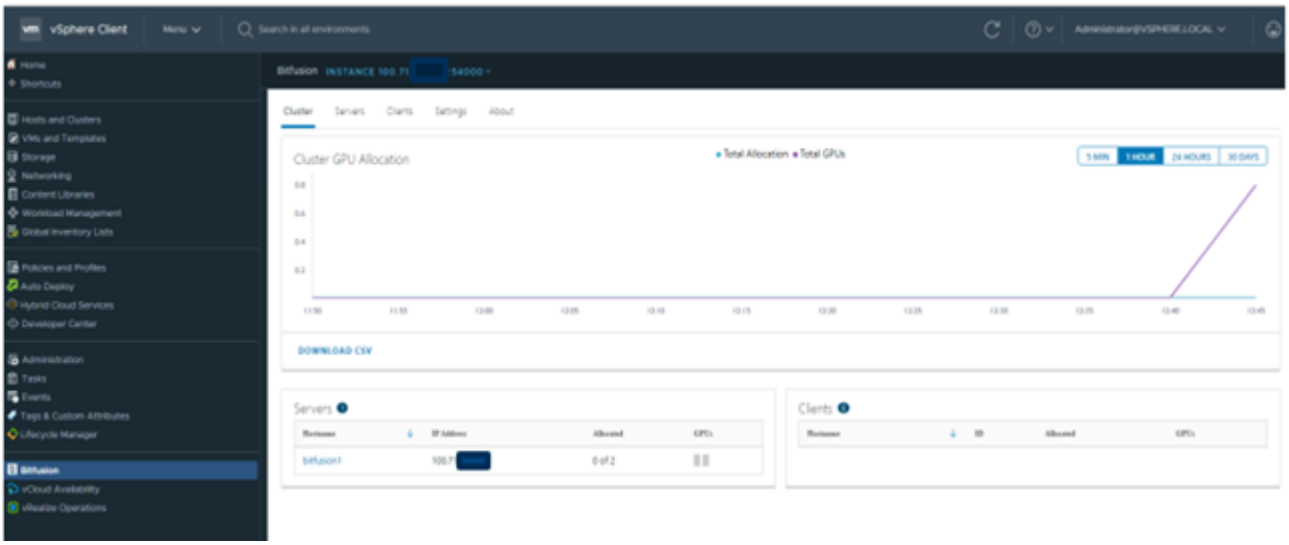

Figure 28 Notification on the top of the window indicating successful deployment of Bitfusion

7. Select **Menu** > **Bitfusion** and wait for the Bitfusion plug-in GUI to load.

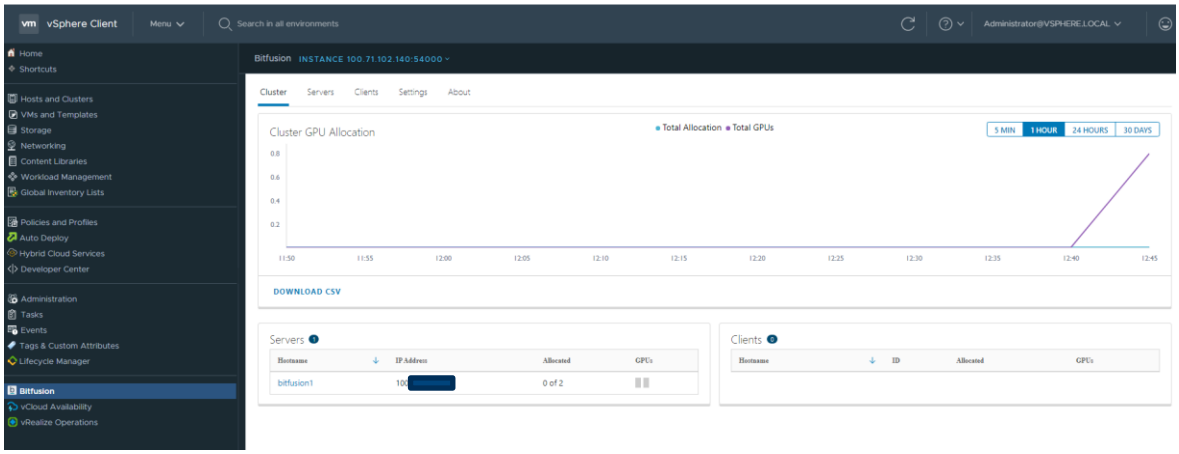

Figure 29 Bitfusion plug-in GUI bitfusion-server-1 virtual machine

8. Login to the bitfusion-1 appliance using the same credentials used during deploying the appliance. Run the following command to reload, restart and edit the bitfusion-manager service file:

```
sudo vi /usr/lib/systemd/system/bitfusion-manager.service
```

Add the line **Environment=BF_IB_GID_INDEX=1** at the end of the **[Service]** section

```
[Unit]
Description=Bitfusion Manager
After=network.target

[Service]
Type=simple
ExecStartPre=/bin/mkdir -p /var/run/bitfusion
ExecStart=/usr/bin/bitfusion manager --auto_start
User=root
RestartSec=5
Restart=always
KillMode=process
Environment=BF_IB_GID_INDEX=1

[Install]
WantedBy=multi-user.target
Alias=bitfusion.service
~
~
```

Figure 30 Editing the bitfusion-manager service file

9. Use the following commands to first save the file, and then restart the bitfusion-manager service for the changes to take effect:

```
sudo systemctl daemon-reload
sudo systemctl restart bitfusion-manager
```

7.6 Deploy the Open Virtual Appliance (OVA) to create the bitfusion-server-2 virtual machine

Follow the steps:

1. Select the appliance OVA file, **bitfusion-server-2.0.0-11.ova**.
2. Select the folder and provide the name for the first appliance **bitfusion-server-2**.
3. Select the **GPU host C4140**, review the details and proceed to select the **NFS datastore** for appliance storage.
4. Select the **port group (mgmt)** from the **dropdown** for the management traffic that connects to vCenter.
5. Verify the network adapter 1 settings:

- Hostname–bitfusion2
- IPv4 address–100.71.x.x141
- CIDR–21
- MTU–9000
- Gateway–100.71.x.x
- DNS–100.71.x.x
- Search domain–oseadc.local
- NTP–100.71.x.x

6. Verify the network adapter 2 settings:

- IPv4 address–172[redacted]

- CIDR–24
- MTU–9000

7.7 Edit the bitfusion-server-2 virtual machine hardware settings

Follow the steps:

1. Under the **Virtual Hardware tab**, Verify the number of **vCPUs** is **8**

- Minimum No. of vCPUs = 4x No. of GPU devices attached to the appliance. In this case, 4 x 2 GPUs i.e. 8

2. Verify the **memory** is set to **48GB** and select the checkbox **Reserve all guest memory**.

- Minimum GB of memory = 1.5x aggregate total of GPU memory on all GPU cards passed through. In this case, 1.5x 32GB i.e. 48GB

3. Click on **Add New Device** and add two PCIe devices. Select the PCI devices from the drop-down menu.

- 0000:86:00.0 | GV100GL V100 SXM2 16GB
- 0000:af:00.0 | GV100GL V100 SXM2 16GB

4. **Add New Device**, **Network adapter**. Browse and select the network **pvrdma** that is created on the **Bitfusion distributed switch** earlier.
5. Select the Adapter Type as PVRDMA and the Device Protocol as RoCE v2
6. Under the Virtual Machine Options tab, select Advanced > Configuration parameters and then select Edit Configuration. Edit the pciPassthru.64bitMMIOSizeGB parameter to 64.

- pciPassthru.64bitMMIOSizeGB= <n>, where n equals (num-cards * size-of-card-in-GB) rounded up to NEXT power of 2. In this case, 2x 16 → 32, rounded to next power of 2 i.e. 64

7. Refrain from powering on the virtual machine. Click on **Enable Bitfusion** from the **bitfusion-server-2** and then select **Actions** > **Bitfusion menu**.

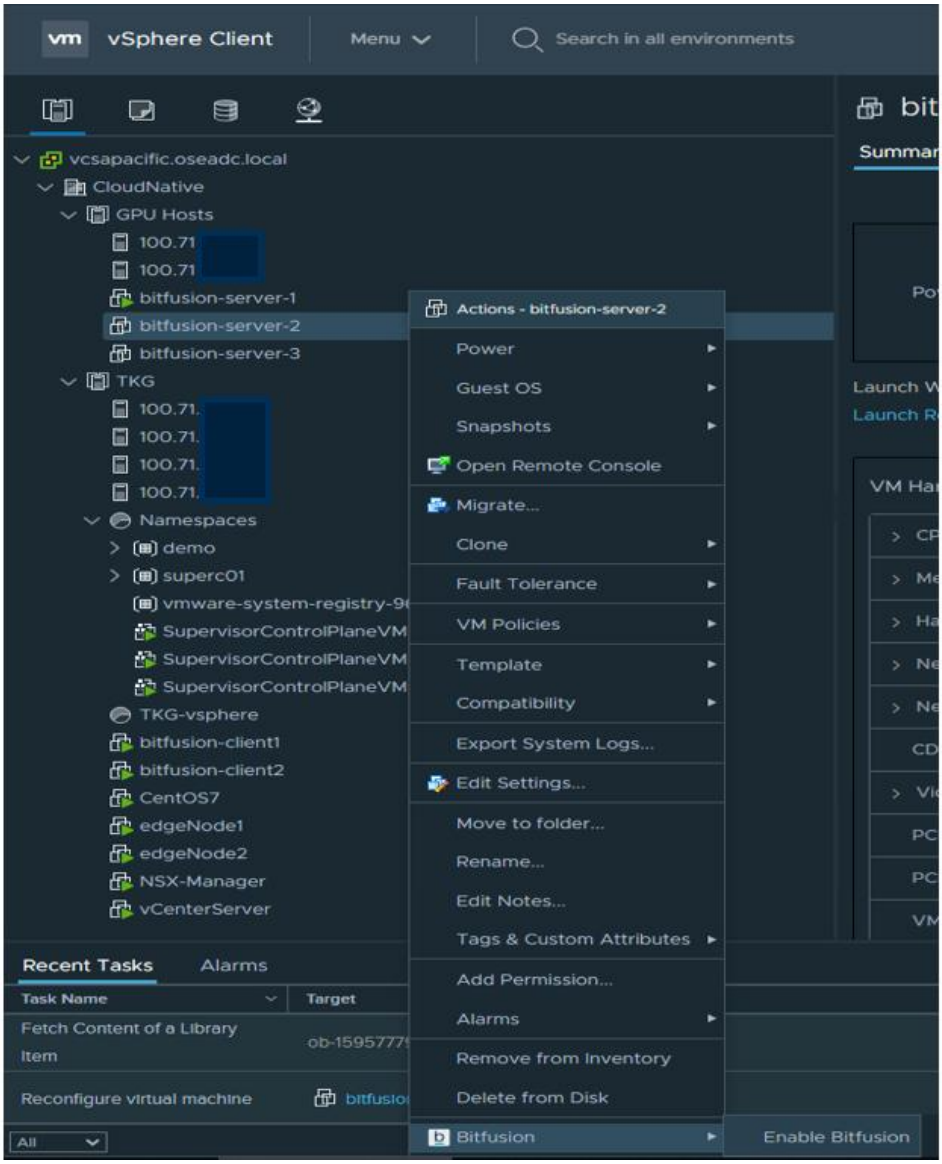

Figure 31 Select Enable Bitfusion from the Actions menu

8. A window pops up listing the options to enable as a client or server. Select the **For a server, this will allow it to be used as a GPU server** radio button and click on **ENABLE**. This adds guest variables informing the server it is not the first GPU server in the Bitfusion cluster.
9. Power on the virtual machine and wait for the Bitfusion plugin UI to show the additional appliance and additional GPUs added to the cluster.

Note: The hostname is shown as 'Unreachable' on the Bitfusion plugin GUI initially before showing the name of the host that is part of the cluster.

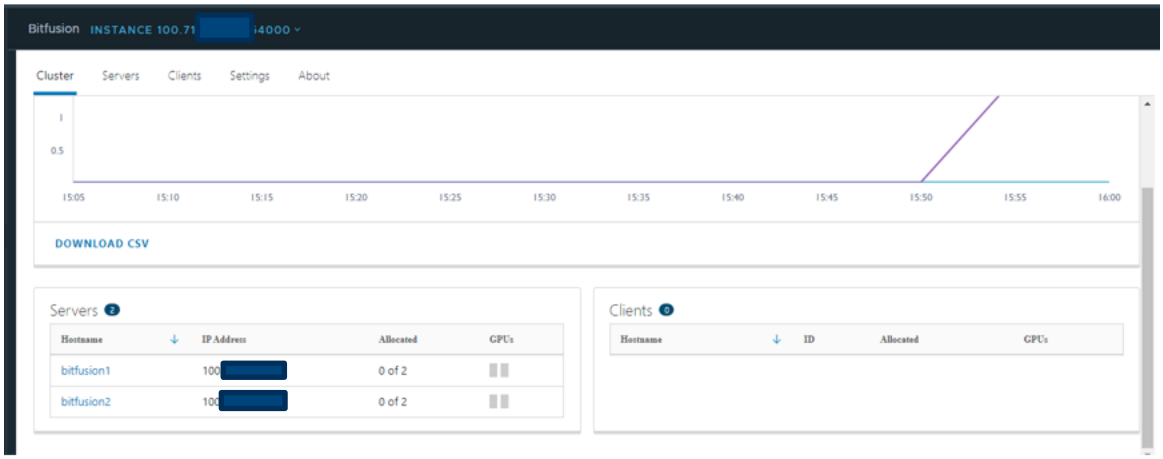

Figure 32 Bitfusion plug-in GUI for bitfusion-server-2 virtual machine

10. Login to the bitfusion-2 appliance using the same credentials used during deploying the appliance. Run the following command to reload, restart and edit the bitfusion-manager service file:

```
sudo vi /usr/lib/systemd/system/bitfusion-manager.service
```

Add the line **Environment=BF_IB_GID_INDEX=1** at the end of the **[Service]** section

11. Use the following commands to first save the file, and then restart the bitfusion-manager service for the changes to take effect:

```
sudo systemctl daemon-reload
sudo systemctl restart bitfusion-manager
```

7.8 Deploy the OVA to create the bitfusion-server-3 virtual machine

Follow the steps:

1. Select the appliance OVA file, bitfusion-server-2.0.0-11.ova.
2. Select the folder and provide the name for the first appliance **bitfusion-server-3**.
3. Select the GPU host **PowerEdge R740xd**. Review the details and proceed to select the local NVMe based datastore.
4. Select the **port group (mgmt)** from the **dropdown** for the management traffic that connects to vCenter.
5. Verify the network adapter 1 settings:

- Hostname–bitfusion3
- IPv4 address–100.71.x.x
- CIDR–21
- MTU–9000
- Gateway–100.71.x.x
- DNS–100.71.x.x
- Search domain–oseadc.local
- NTP–100.71.x.x

6. Verify the network adapter 2 settings:

   - IPv4 address–172.[redacted]
   - CIDR–24
   - MTU–9000

7.9 Edit the bitfusion-server-3 virtual machine hardware settings

Follow the steps:

1. Under the **Virtual Hardware tab**, verify the number of **vCPUs** is **8**

   - Minimum No. of vCPUs = 4x No. of GPU devices attached to the appliance. In this case, 4 x 2 GPUs i.e. 8

2. Verify that the **memory** is set to **48GB** and select the checkbox **Reserve all guest memory**.

   - Minimum GB of memory = 1.5x aggregate total of GPU memory on all GPU cards passed through. In this case, 1.5x 32GB i.e. 48GB

3. Click on **Add New Device** and add two PCIe devices. Select the PCI devices from the drop-down menu.

   - 0000:3b:00.0 | GV100GL Tesla V100 PCIe 16GB
   - 0000:d8:00.0 | TU104GL Tesla T4

4. Click on **Add New Device** and select **Network adapter**. Browse and select the network pvrdma that is created on the Bitfusion distributed switch earlier.
5. Select the Adapter Type as PVRDMA and the Device Protocol as RoCE v2
6. Under the Virtual Machine Options tab, select Advanced > Configuration parameters and then select Edit Configuration. Edit the pciPassthru.64bitMMIOSizeGB parameter to 64.

   - pciPassthru.64bitMMIOSizeGB= <n>, where n equals (num-cards * size-of-card-in-GB) rounded up to NEXT power of 2. In this case, 2x 16 → 32, rounded to next power of 2 i.e. 64

7. Refrain from powering on the virtual machine. Click on **Enable Bitfusion** from **the bitfusion-server-3** option and then select **Actions** > **Bitfusion menu**.
8. A window pops up listing the options to enable as a client or server. Select the **For a server, this will allow it to be used as a GPU server** radio button and click on **ENABLE**. This adds guest variables informing the server it is not the first GPU server in the Bitfusion cluster.
9. Power on the virtual machine and wait for the Bitfusion plugin UI to show the additional appliance and additional GPUs added to the cluster.

Note: The hostname is shown as 'Unreachable' on the Bitfusion plugin GUI initially before showing the name of the host that is part of the cluster.

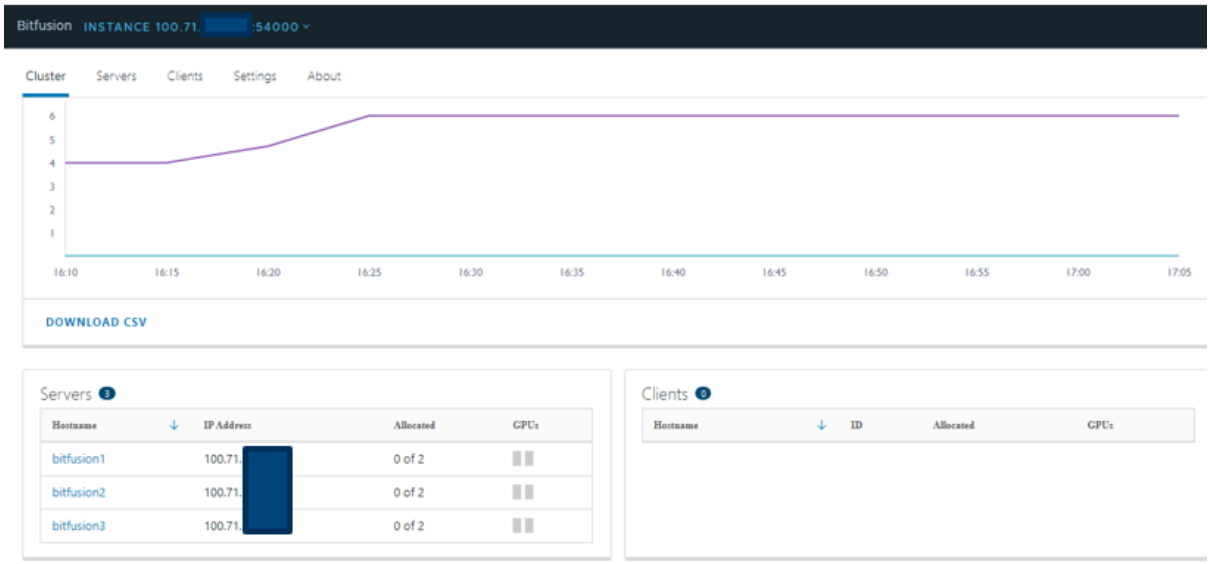

Figure 33 Bitfusion plug-in GUI for bitfusion-server-3 virtual machine

10. In the Bitfusion user interface, verify that the **total available GPUs** is **6** and the **allocation** is **0** in the **Cluster GPU Allocation graph**.

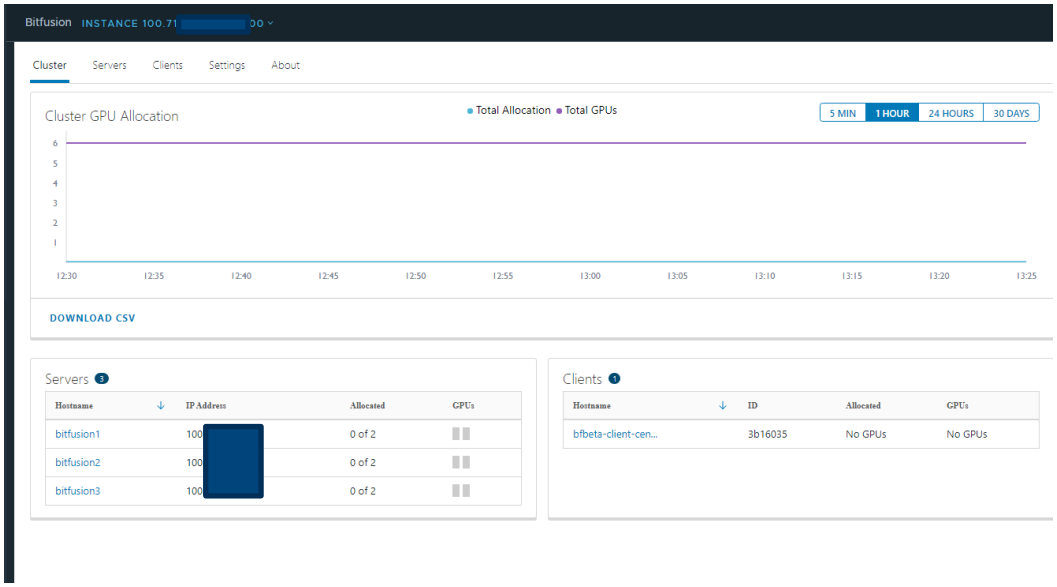

Figure 34 All six available GPUs in the cluster

11. Login to the bitfusion-3 appliance using the same credentials used during deploying the appliance. Run the following command to reload, restart and edit the bitfusion-manager service file:

```
sudo vi /usr/lib/systemd/system/bitfusion-manager.service
```

Add the line **Environment=BF_IB_GID_INDEX=1** at the end of the **[Service]** section

12. Use the following commands to first save the file, and then restart the bitfusion-manager service for the changes to take effect:

```
sudo systemctl daemon-reload
sudo systemctl restart bitfusion-manager
```

7.10 Alternate method of installing an additional Bitfusion server

Alternatively, you can add a new server to your vSphere Bitfusion cluster using the deploy procedure on the primary server. To do this, deploy the vSphere Bitfusion appliance on a virtual machine (VM), customize the vSphere Bitfusion OVF template and then pass the GPUs to the vSphere Bitfusion server virtual machine to enable the virtual machine as a vSphere Bitfusion server.

Follow the steps:

1. From the **Hosts and Clusters** view in vCenter Server, right-click on an ESXi host, and select **Bitfusion** > **Install Bitfusion server**. This opens the **Install Bitfusion server** dialog box.
2. On the **Select an OVA image** page, enter the URL of the vSphere Bitfusion OVA file, or browse to the file, and click **Next**.
3. On the **Verify template details** page, review the OVA template details, and click **Next**.
4. On the **Select a name and hostname** page, enter a name for the virtual machine and a hostname for the vSphere Bitfusion server, and click **Next**. Optionally, you can specify a host ID for the vSphere Bitfusion server, for example, when you upgrade your vSphere Bitfusion server. If you skip this step, a host ID is generated and assigned automatically.
5. On the **Select storage** page, define where and how to store the files of the deployed virtual machine, and click **Next.**
6. On the **Select networks** page, specify the networking configuration for Network Adapter 1 and click **Next.**
7. On the **Select GPUs** page, add GPUs to the subsequent server and click **Next.** Follow:

    a. Click **Add GPU**.
    b. Select a GPU from the **GPU Device drop-down menu.**
    c. Specify the total memory of the GPU. (Optional).
    d. To accept the NVIDIA license, select the **Download and Install NVIDIA Driver** check box. (Optional).

8. On the **Customize server** page, specify the vSphere Bitfusion server details, and click **Next**. Follow:

    a. Specify the number of CPUs for the virtual machine.
    b. Specify the total virtual machine memory in giga-bytes (GB).
    c. Enter a password for the customer account (Optional). After the deployment is complete, use the customer user account to log in to the vSphere Bitfusion server through the console shell or SSH. If this step is skipped, you cannot log in to the subsequent server.
    d. Select the **Power On VM After Create** check box. (Optional). You can deselect the check box, if you make changes to the virtual machine before powering it on.

9. On the **Summary** page, review the deployment details and click **Finish.**

7.11 Improved support for multiple networks

During the deployment process of a vSphere Bitfusion server, you must configure at least Network Adapter 1 as it is used for management and data traffic. Network Adapter 2, Network Adapter 3 and Network Adapter 4 are all optional and are only used for data traffic. With vSphere Bitfusion 3.0, you can now add network interfaces for data traffic management after the deployment process of a server.

Follow the steps:

1. In the vSphere Client, right-click on the virtual machine of a vSphere Bitfusion server and select **Edit Settings.**
2. On the **Virtual Hardware** tab, click the **Add New Device** button.
3. Under **Network**, select **Network Adapter.**
4. From the **New Network** drop-down menu, select a network to which the virtual machine can be connected.
5. Expand the **New Network** section and from the **Adapter Type** drop-down menu, select the network adapter to which the virtual machine can be assigned and click **OK**.

Note: vSphere Bitfusion supports VMXNET3 and PVRDMA adapters.

7.12 Provide client cluster access to GPU resources

Follow the steps:

1. Enable bitfusion on the powered off bitfusion client virtual machine. To do this, select **Actions** > **Bitfusion** > **Enable**. Select the option **For a client, this will allow users to run Bitfusion workloads** radio button and click on **ENABLE**
2. Power on the virtual machine, create a user group with the name **bitfusion** and add all users that need access to GPU resources to this user group.
3. Install the InfiniBand packages and reload pvrdma driver using the following commands:

```
yum groupinstall "Infiniban Support" -y
rmmod vmw_pvrdma
modprobe vmw_pvrdma
ibv_devinfo
```

4. Export the environment variable **BF_IB_GID_INDEX** and add the following line to the bash profile file for this setting to persist across reboots:

```
export BF_IB_GID_INDEX=1
```

5. Install the bitfusion client rpm, bitfusion-client-centos7-2.0.0-11.x86_64.rpm

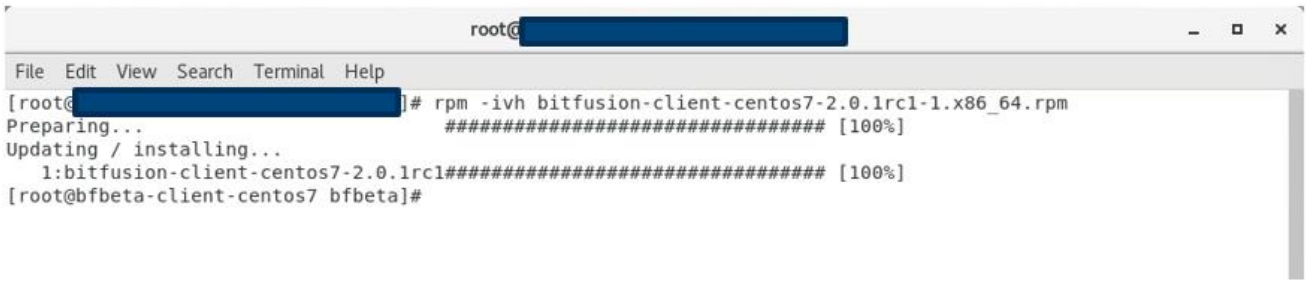

Figure 35 Install the bitfusion-client-centos7-2.0.0-11.x86_64.rpm file

6. Run Bitfusion commands to list all the GPUs available on the GPU server cluster by using the command:

```
bitfusion – list_gpus
```

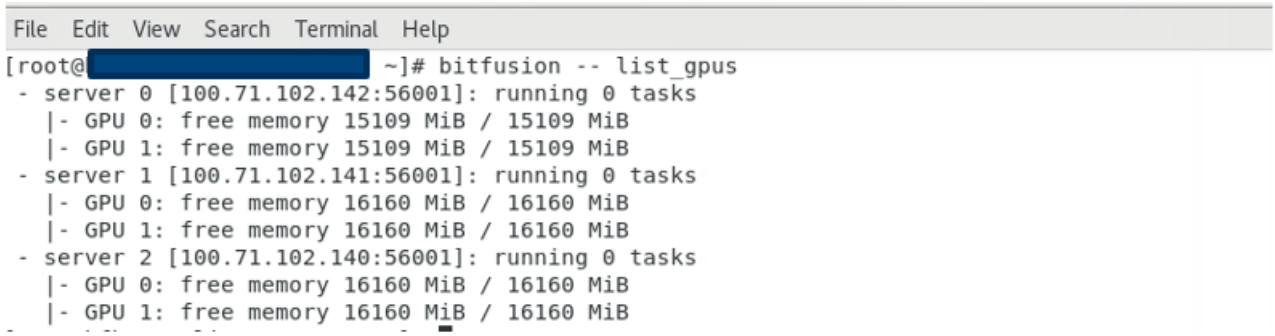

Figure 36 List all available GPUs in the cluster

7. Run Bitfusion commands over the PVRDMA network targeting each GPU server and verify that the available resources are listed. Use the following command:

```
bitfusion list_gpus -l 172.16.6.x
```

8. Verify the GPU allocation in the Bitfusion user interface by running the **TensorFlow benchmark** and assign two GPUs using the **Bitfusion command-line interface**.

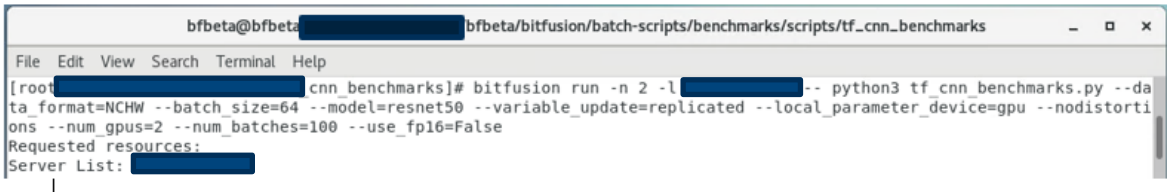

Figure 37 Verify the GPU allocation in the Bitfusion user interface by running the TensorFlow benchmark

The client section of the Bitfusion UI lists two GPUs allocated to the client virtual machine.

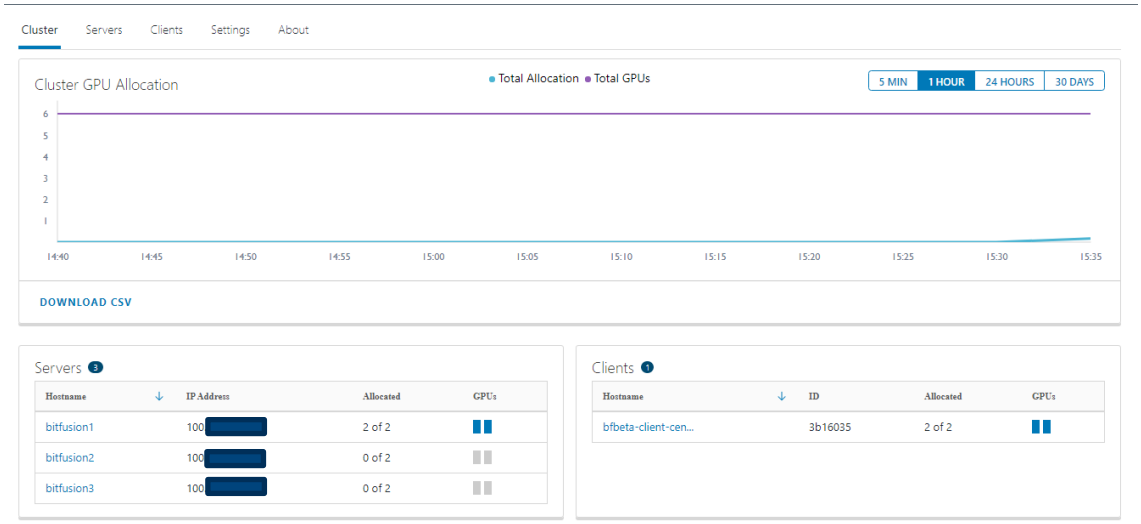

Figure 38 Client cluster access is successfully provided to the GPU resources

7.13 Support for remote clients and bare-metal server

VMware vSphere Bitfusion 2.5 allows you to create tokens to enable clients that are installed on different vCenter server instances or on a Tanzu Kubernetes Grid (TKG) container or on a bare-metal client server.

Follow the steps:

1. From the vSphere Client software **Menu** tab, select **Bitfusion.**
2. From the **Tokens** tab, select **New Token.**
3. In the **Create New token** dialog box**,** enter a description **of** the token and then click **Create.**
4. Select the token created and click **Download**. Save the TAR file in the local filesystem.
5. Copy the TAR file to the filesystem of the client machine(s).
6. Extract the contents of the TAR file to the following folders:

   a. Copy ca.crt to /etc/bitfusion/tls
   b. Copy client.yaml to ~/.bitfusion
   c. Copy servers.conf to ~/.bitfusion

7. Open the terminal to add users to the Bitfusion group by running the following command:

```
sudo usermod -aG bitfusion <username>
```

8. Optionally, you can verify that the users were successfully added to the vSphere Bitfusion group by doing the following:

   a. Log out and log in to the terminal of the server.
   b. Run the `groups` command. The users and their associated groups are listed.

9. Optionally, you can verify that the VMware vSphere Bitfusion client is working by listing the available GPUs in the vSphere Bitfusion deployment. To do this, use the `bitfusion list_gpus` command.

7.14 Support for backup and restore

Note: To upgrade your cluster, back up the VMware environment, deploy new server virtual machines with vSphere Bitfusion 2.5 and restore the backup. To upgrade the operating system of a client, install the latest version of CentOS, Red Hat Enterprise Linux or Ubuntu Server package.

To upgrade the VMware vSphere Bitfusion software from version 2.0 to 2.5, follow the steps:

1. From the **Settings** tab of the vSphere Bitfusion software, choose **Backup/Restore**.

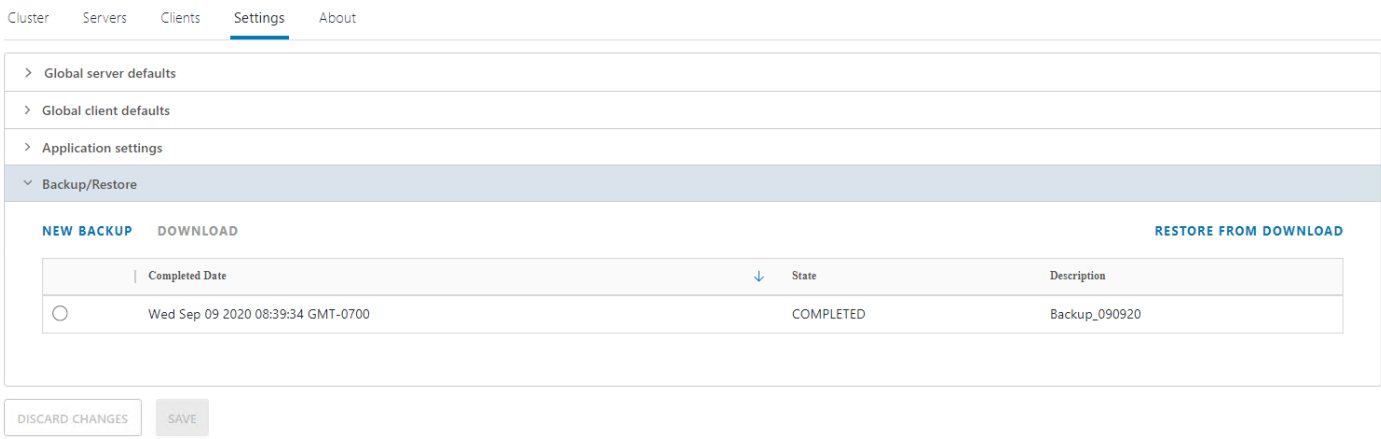

Figure 39 Create and download a backup of your vSphere Bitfusion 2.0 cluster

2. Using the vSphere Bitfusion plugin, create and download a backup of your vSphere Bitfusion 2.0 cluster.
3. Open the `manifest.json` file from the downloaded `bitfusionbackup.tar.gz` archive file to find the servers section that has information about the servers in your vSphere Bitfusion cluster at the time of the backup.

Note: The host ID, hostname, and number of GPUs for each server must be noted.

4. Power off all the vSphere Bitfusion servers in the cluster.
5. Install a new primary vSphere Bitfusion server:
   a. Deploy a new primary vSphere Bitfusion server virtual machine using a vSphere Bitfusion 2.5 Appliance OVF Template.
   b. During deployment, enter the same hostname as of the primary vSphere Bitfusion 2.0 server.
   c. While setting up the new virtual machine, add the same number of GPUs as of the vSphere Bitfusion 2.0 server.
   d. Add a guestinfo.bitfusion.server.host-id configuration parameter. The value of the parameter should match the host ID of your primary server with vSphere Bitfusion 2.0 server that was listed in the `manifest.json` file. For more information, see the Edit Configuration File Parameters section in the vSphere Virtual Machine Administration document.
   e. Power on the server and wait for the vSphere Bitfusion plugin to register itself with the vCenter server.
6. Restore the backup of the vSphere Bitfusion 2.0 cluster created using the vSphere Bitfusion plugin and follow the steps to power on the virtual machines:

a. Deploy a new server virtual machine using a vSphere Bitfusion 2.5 OVF Template.
b. During deployment, enter the same hostname as of the primary vSphere Bitfusion 2.0 server.
c. While setting up the new virtual machine, add the same number of GPUs as of the vSphere Bitfusion 2.0 server.
d. Add a guestinfo.bitfusion.server.host-id configuration parameter. The value of the parameter should match the host ID of your primary server with vSphere Bitfusion 2.0 server that was listed in the manifest.json file.
e. Enable the virtual machine as a vSphere Bitfusion server. For more information, see Add Additional vSphere Bitfusion Servers.
f. Power on the virtual machine. Multiple virtual machines will be powered on in sequential order.

7. Delete the vSphere Bitfusion 2.0 server virtual machines. The servers in the cluster are now successfully upgraded to vSphere Bitfusion 2.5.

8 Getting help

8.1 Contacting Dell EMC

Dell EMC provides several online and telephone-based support and service options. Availability varies by country, region, and product, and some services may not be available in your area.

To contact Dell EMC for sales, technical assistance, or customer service issues, see https://www.dell.com/contactdell.

If you do not have an active Internet connection, you can find contact information on your purchase invoice, packing slip, bill, or the product catalog.

For information about proof of concept and demonstration engagements, see Customer Solution Centers.

8.2 Documentation resources

- VMware vSphere Bitfusion Documentation
- VMware vSphere Bitfusion 3.0 Installation Guide
- VMware vSphere Bitfusion 2.0 Installation Guide
- PVRDMA Deployment and Configuration of QLogic CNA devices in VMware ESXi
- VMware vSphere Bitfusion Offers the Best ROI for Machine Learning Dev-Test Workloads
- VMware vSphere Bitfusion Release 2.5.0 Delivers the Feature Mix You Wanted – Romeo and Juliet Sing Simon and Garfunkel

8.3 VMware Hands-On-Labs (HOLs)

- HOL-2147-91-ISM - Using Bitfusion GPU virtualization in vSphere - Lightning Lab
- HOL-2147-02-ISM - Using Bitfusion GPU virtualization in vSphere

8.4 Documentation feedback

If you have feedback for this document, write to **documentation_feedback@dell.com**. Alternatively, you can click on the Feedback link in any of the Dell documentation pages, fill out the form, and click Submit to send your feedback.