VMware vSphere Bitfusion Installation Guide

11 MAY 2021

Updated to include VMware vSphere Bitfusion 3.5

VMware vSphere Bitfusion 3.0

You can find the most up-to-date technical documentation on the VMware website at:

https://docs.vmware.com/

VMware, Inc.
3401 Hillview Ave.
Palo Alto, CA 94304
www.vmware.com

Contents

About *VMware vSphere Bitfusion Installation Guide*

The *VMware vSphere Bitfusion Installation Guide* describes how to install and configure VMware vSphere Bitfusion in your VMware® vSphere environment.

At VMware, we value inclusion. To foster this principle within our customer, partner, and internal community, we create content using inclusive language.

VMware vSphere Bitfusion Installation Guide is intended for experienced vSphere administrators who want to install and configure vSphere Bitfusion.

Intended Audience

This information is intended for anyone who wants to install vSphere Bitfusion. The information is written for experienced Linux system administrators who are familiar with virtual machine technology and data center operations using VMware vSphere and vCenter Server.

Understanding VMware vSphere Bitfusion

1

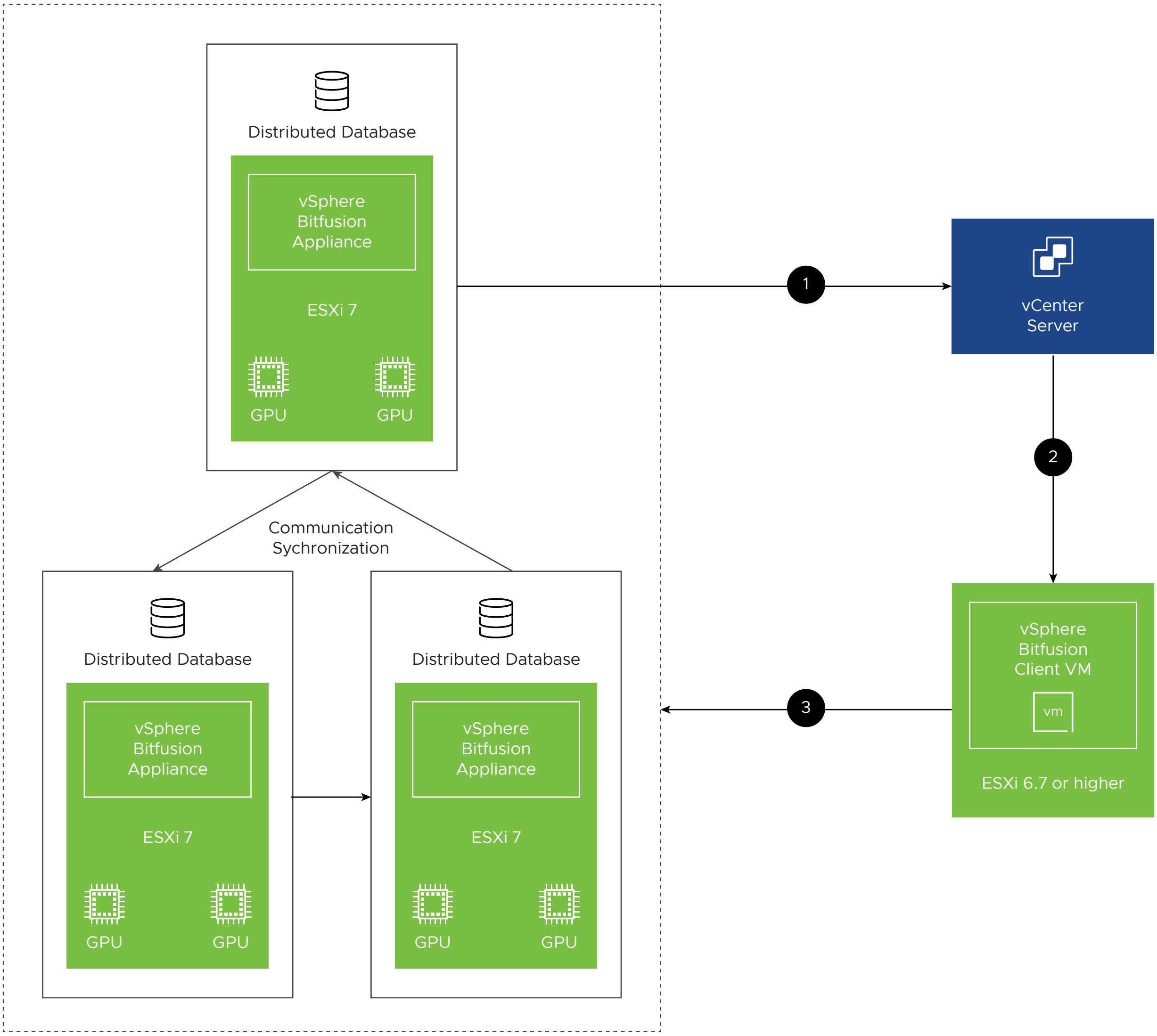
VMware vSphere Bitfusion virtualizes hardware accelerators such as graphical processing units (GPUs) to provide a pool of shared, network-accessible resources that support artificial intelligence (AI) and machine learning (ML) workloads.

vSphere Bitfusion Architecture

vSphere Bitfusion has a client-server architecture. The product allows multiple client virtual machines (VMs) running artificial intelligence (AI) and machine learning (ML) applications to share access to remote GPUs on virtual machines running vSphere Bitfusion server software. You run the applications on the vSphere Bitfusion client machines, while the GPUs that provide acceleration are installed on the vSphere Bitfusion server machines across a network. The applications can open files, allocate memory, and call CUDA as if operating on a machine with local GPUs.

The following figure is an example of a small vSphere Bitfusion cluster, such as a set of vSphere Bitfusion server-client machines and vCenter Server on a switched network. A minimal vSphere Bitfusion cluster configuration is one client, one server, and one vCenter Server. You can create large clusters with multiple clients and multiple servers.

Figure 1-1. Example of a small vSphere Bitfusion cluster


Distributed Database
vSphere Bitfusion Appliance
ESXi 7
GPU
GPU
1
vCenter Server
2
Communication Sychronization
Distributed Database
vSphere Bitfusion Appliance
ESXi 7
GPU
GPU
Distributed Database
vSphere Bitfusion Appliance
ESXi 7
GPU
GPU
3
vSphere Bitfusion Client VM
vm
ESXi 6.7 or higher

1 The primary vSphere Bitfusion server registers a vSphere Bitfusion Plug-in with vCenter Server.

2 The vSphere Bitfusion Plug-in enables a vSphere Bitfusion client VM.

3 The vSphere Bitfusion Client has authorized access to all vSphere Bitfusion servers in the vSphere Bitfusion cluster.

Note Before using VMware vSphere Bitfusion, you must deploy a vSphere Bitfusion server, and install and enable a vSphere Bitfusion client. For more information, see the *VMware vSphere Bitfusion Installation Guide*.

vSphere Bitfusion Functionality

When you start an AI or ML application on the vSphere Bitfusion client, vSphere Bitfusion intercepts the CUDA calls of the application, and sees the data and data pointers of the calls. The vSphere Bitfusion server does not require a connection to the data, but to the vSphere Bitfusion client only. The client transfers the data and the rest of the CUDA calls to the server. The vSphere Bitfusion server processes the calls and returns the results back to the client.

When you run AI and ML applications, vSphere Bitfusion can perform the following operations.

- Dynamically allocate and access GPU resources from vSphere Bitfusion servers.

  Applications can share GPU resources that are not dedicated to individual machines and you can run each application on a configured machine, container, and environment. Applications consume GPU acceleration services from a pool of vSphere Bitfusion servers across a network, and consume the resources only for the length of time that an application or session runs. GPUs return to the pool when applications or sessions complete.

- Access partitions of GPU resources for concurrent sharing with other applications.

  Another option to share GPUs is by partitioning the GPUs. The memory of a physical GPU can be divided into fractions of an arbitrary size and allocated to different applications at the same time. vSphere Bitfusion performs sharing with an interposition technology. vSphere Bitfusion intercepts API calls that normally address a local accelerator on a PCIe host bus and sends the API calls and related data across a network. vSphere Bitfusion provides sharing services for AI and ML applications, and supports the CUDA API to target NVIDIA GPUs.

vSphere Bitfusion Components

vSphere Bitfusion Server

vSphere Bitfusion server runs on an ESXi host with locally installed GPUs as a VMware appliance, which is a preconfigured virtual machine (VM) with prepackaged software and services. The server requires access to the local GPUs, usually through VMware vSphere® DirectPath I/O™.

vSphere Bitfusion Client

vSphere Bitfusion client runs on VMs which run the AI and ML applications.

vSphere Bitfusion Plug-In

The vSphere Bitfusion servers register a vSphere Bitfusion Plug-in with VMware vCenter Server. The plug-in provides monitoring and management of vSphere Bitfusion clients and servers.

vSphere Bitfusion Cluster

vSphere Bitfusion cluster is the set of all vSphere Bitfusion servers and clients in a vCenter Server instance.

vSphere Bitfusion Group

The vSphere Bitfusion client creates a vSphere Bitfusion group during the installation process. Only the members of the group can use vSphere Bitfusion. Certain configuration files are set up with appropriate permissions and the members of the group inherit appropriate limits to work effectively with vSphere Bitfusion.

vSphere Client

The vSphere Client lets you connect to vCenter Server instances by using a Web browser, so that you can manage your vSphere infrastructure. You access the vSphere Bitfusion Plug-in through the vSphere Client.

Command-Line Interface (CLI)

You can manage vSphere Bitfusion servers and clients by using command-line interface (CLI) commands.

vCenter Server

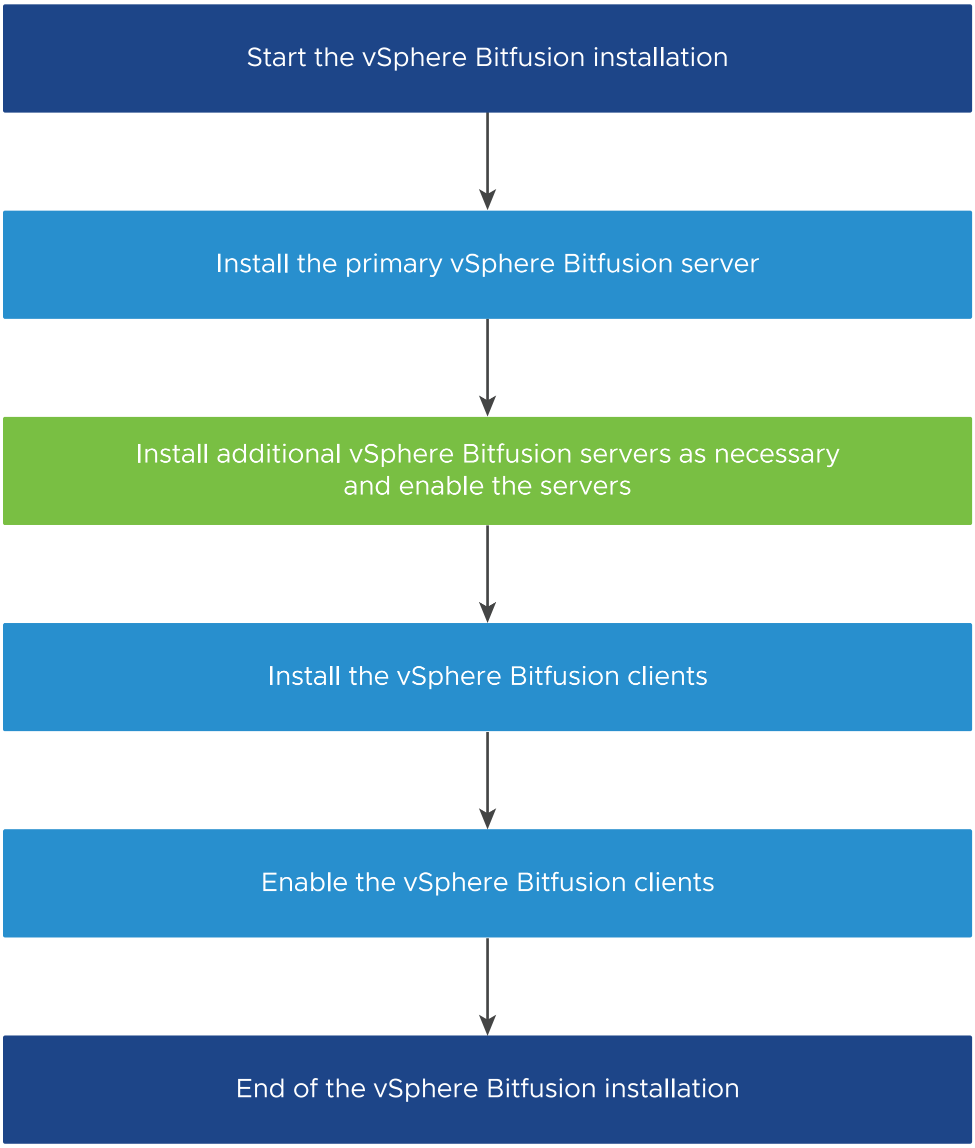
vCenter Server is the server management software that provides a centralized platform for controlling your vSphere environment.

Overview of the vSphere Bitfusion Installation Process

2

VMware vSphere Bitfusion is a sophisticated product with multiple components to install and set up. To ensure a successful vSphere Bitfusion deployment, understand the sequence of tasks required.

Figure 2-1. vSphere Bitfusion Installation Workflow



Start the vSphere Bitfusion installation
Install the primary vSphere Bitfusion server
Install additional vSphere Bitfusion servers as necessary and enable the servers
Install the vSphere Bitfusion clients
Enable the vSphere Bitfusion clients
End of the vSphere Bitfusion installation

The steps to successfully install vSphere Bitfusion are as follows.

1 Read the vSphere Bitfusion release notes.

2 Ensure that your environment meets the minimum system requirements and any additional resources required to run the artificial intelligence and machine learning workloads you plan to run. See Chapter 3 System Requirements for vSphere Bitfusion Server and Chapter 5 Installing the vSphere Bitfusion Client.

3 Install the primary vSphere Bitfusion server. See Chapter 4 Deploying the vSphere Bitfusion Appliance.

4 Install additional vSphere Bitfusion servers as necessary. See Add Subsequent vSphere Bitfusion Servers.

5 Install vSphere Bitfusion clients. See Chapter 5 Installing the vSphere Bitfusion Client.

6 Enable the vSphere Bitfusion clients. See Chapter 6 Enabling the vSphere Bitfusion Client.

System Requirements for vSphere Bitfusion Server

3

vSphere Bitfusion requires an ESXi host on which to install the vSphere Bitfusion server.

System Requirements for vSphere Bitfusion Server

The vSphere Bitfusion server must run on a vSphere deployment with the following system requirements.

- The minimum disk space requirement for a vSphere Bitfusion server appliance is 50 GB.
- The ESXi host version, on which a vSphere Bitfusion server runs, must be 7.0 or later.
- The minimum memory requirement for a vSphere Bitfusion server is 32 GB or 150% of the total GPU memory that is installed on the server, whichever is higher.
- The minimum virtual CPU (vCPU) requirement for a vSphere Bitfusion server is the number of GPU cards multiplied by 4.
- Network supporting TCP/IP or RoCE (PVRDMA adapters).
- A minimum of 10 Gbps of bandwidth for any machine that accesses two or more GPUs.
- The latency between a client machine and a server virtual machine must be 50 microseconds or less. This is not a strict requirement, but the performance of the vSphere Bitfusion deployment is better with low latency.
- All vSphere Bitfusion servers must be connected to the same set of valid NTP servers.

Required Ports for vSphere Bitfusion Server

Verify that these ports are not blocked by using a denylist or firewall rules. They are required for communication.

Since vSphere Bitfusion 3.5, due to security concerns, the TLSv1.0 and TLSv1.1 protocols are disabled by default. For more information, see Knowledge Base article 2145796.

Port	Description
443	This port is used for the communication between vSphere Bitfusion and the vSphere Client.
7000 and 7001	This port is used by Apache Cassandra for the communication with the Cassandra cluster.

Port	Description
9042	This port is used by Apache Cassandra to communicate with the native protocol clients.
9142	This port is used by Apache Cassandra for the Cassandra Thrift API.
9160	This port is used by Apache Cassandra. The CQL native transport listens for vSphere Bitfusion clients on this port and the port is used when encrypted and unencrypted connections are required.
45201 - 46225	These ports are used by vSphere Bitfusion clients to communicate with the vSphere Bitfusion server processes that handle CUDA requests.
55001 - 55201	These ports are used by vSphere Bitfusion clients to communicate with a task specific dispatcher process that runs on the vSphere Bitfusion server. The process requests a session and starts the service workers of the vSphere Bitfusion server that handle the workload.
56001	This port is used for vSphere Bitfusion intercommunication. vSphere Bitfusion servers communicate with each other on this port and vSphere Bitfusion clients use this port to start a task on a vSphere Bitfusion server.

Web Browser Requirements for vCenter Server

To use vSphere Bitfusion, you require a web browser version that is supported by vCenter Server. For more information, see vSphere Client Software Requirements.

vSphere Bitfusion Compatibility and Interoperability

For a list of versions, models, and products that are compatible with vSphere Bitfusion, see the VMware vSphere Bitfusion Compatibility and Interoperability page.

Deploying the vSphere Bitfusion Appliance 4

The vSphere Bitfusion OVA file contains the compressed open virtualization format files that compose the vSphere Bitfusion server. After setting up your vSphere environment, download the vSphere Bitfusion OVA file and deploy it as a vSphere appliance.

You can deploy virtual machines, virtual appliances, and vApps in Open Virtual Format (OVF) and Open Virtual Appliance (OVA). Deploying an OVF or OVA template allows you to add pre-configured virtual machines or vApps to your vCenter Server or ESXi inventory. Deploying an OVF or OVA template is similar to deploying a virtual machine from a template. However, you can deploy an OVF or OVA template from any local file system accessible from the vSphere Client, or from a remote web server.

vSphere Bitfusion is distributed as a OVA file. To learn more about OVA or OVF files, see the *vSphere Virtual Machine Administration* documentation.

This chapter includes the following topics:

Preparing for the vSphere Bitfusion vSphere Bitfusion Deployment

Before you start the vSphere Bitfusion deployment, you must perform several tasks. The outcomes of these tasks are prerequisites for the deployment process.

Locate the vCenter Server TLS Certificate Thumbprint

The vCenter Server TLS Certificate Thumbprint is the secure hash algorithm (SHA1) signature of the vCenter Server TLS certificate.

Perform the steps in the following procedure to locate the TLS certificate thumbprint for your environment. You must copy the thumbprint and add it later in the deployment properties of the OVF template.

Procedure

1. Open a Web browser and enter the URL for your vSphere Client: **`https://vcenter_server_ip_address_or_fqdn/ui`**
2. Enter the credentials of a user who has permissions on vCenter Server, and click **Login**.
3. Locate the TLS certificate thumbprint.
   - Locate the TLS certificate thumbprint in Google Chrome.
     a. Click the **Secure** icon to the left of the web address, and select **Certificate**.
     b. In the **Certificate** dialog box, click the **Details** tab.
     c. On the **Details** tab, scroll down the list, and in the **Field** column, select **Thumbprint**.
     d. The vCenter Server TLS Certificate Thumbprint is shown in the text box below the list box.
   - Locate the TLS certificate thumbprint in Mozilla Firefox.
     a. Click the **Secure** icon to the left of the web address, select the arrow to the right of the connection status, and click **More Information**.
     b. In the **Page Info** dialog box, on the **Security** tab, select **View Certificate**.
     c. On the **Certificates** browser tab, the fingerprints are shown in the **Fingerprints** section.

What to do next

Customize the vSphere Bitfusion Appliance OVF Template

Enable a GPU for Passthrough

To use a GPU in a vSphere Bitfusion server, you must enable the device in passthrough mode. The operation allows the GPU to be accessed directly by the server, bypassing the ESXi hypervisor, which provides a level of performance that is similar to the performance of the GPU on a native system.

When using passthrough mode, each GPU device is dedicated to the virtual machine (VM) of the vSphere Bitfusion server. You can use multiple physical GPUs in passthrough mode. The following procedure must be performed for all GPU devices, that you plan to use in a vSphere Bitfusion server.

Prerequisites

- Verify that your GPU device is supported by your server vendor.
- Verify that your GPU can be used in passthrough mode.
- Verify you have created a virtual machine for the vSphere Bitfusion server.

- Verify whether your GPU device maps memory regions with a total size of 16 GB or more.

  Note Typically, high-end GPU cards need high amounts of memory mapping. These memory mappings are specified in the PCI Base Address Registers (BARs) for the device. You can find relevant information in the vendor documentation of your GPU.

Procedure

1 If your GPU requires 16 GB or more of memory mapping, in the BIOS settings of the ESXi host, enable the GPU for passthrough.

   Typically, the name of the setting is *Above 4G decoding*, *Memory mapped I/O above 4GB*, or *PCI 64-bit resource handing above 4G*.

2 Enable the GPU for passthrough on the ESXi host.

   a In the vSphere Client, right-click on the ESXi host and select **Settings**.

   b On the **Configure** tab, select **Hardware > PCI Devices**, and click **Configure Passthrough**.

   c In the **Edit PCI Device Availability** dialog box, in the ID column, select the check box for the GPU device.

   d Click **OK**.

      The GPU is displayed on the **Passthrough-enabled devices** tab.

   e Reboot the ESXi host.

3 Enable UEFI or EFI in the boot options the virtual machine.

   The VM of the vSphere Bitfusion server must boot in EFI or UEFI mode for correct GPU use.

   a In the vSphere Client, right-click on the VM.

   b Select **Edit Settings > VM Options > Boot Options**.

   c From the **Firmware** drop-down menu, select UEFI or EFI.

   d Click **OK**.

During the deployment process of the vSphere Bitfusion appliance, you can pass through the GPU to the VM of the vSphere Bitfusion server.

Start the vSphere Bitfusion Appliance Deployment

To begin the vSphere Bitfusion deployment, you specify information about the product, including its name, location, and storage within your vSphere environment.

When installing additional vSphere Bitfusion servers, you must register the subsequent servers with the first, or primary, vSphere Bitfusion server you installed. You must perform this extra step before powering on the vSphere Bitfusion virtual machine. See Add Subsequent vSphere Bitfusion Servers.

Prerequisites

- Download the vSphere Bitfusion OVA file from https://my.vmware.com/downloads/.
- Verify that the vSphere environment on which you are deploying the vSphere Bitfusion appliance meets the minimum system requirements. See Chapter 3 System Requirements for vSphere Bitfusion Server.
- Verify that you can log into the vSphere Client as an administrator.
- Verify that the ESXi hosts on which you want to deploy the vSphere Bitfusion appliances are running.

Procedure

1. Log in to vSphere Client as an administrator.
2. In the vSphere Client, right-click the ESXi host on which to deploy the vSphere Bitfusion appliance and select **Deploy OVF Template**.
3. On the **Select an OVF template** page, enter the URL of the `OVA` file or browse to the file, and click **Next**.
4. On the **Select a name and folder** page, enter a name for the vSphere Bitfusion virtual machine and select a location for your deployment, and click **Next**.
5. On the **Select a compute resource** page, select a resource on which to run the deployed VM template and click **Next**.
6. On the **Review details** page, verify the OVF template details and click **Next**.

   The **Review details** page displays a warning cautioning that the vSphere Bitfusion OVF uses advanced configuration values that might pose a security risk. The configuration values that trigger the alert are `pciPassthru.use64bitMMIO = true` and `pciPassthru.64bitMMIOSizeGB = 256`. The first parameter enables PCI passthrough for GPU devices, that require 16 GB or more of memory mapping, and the second parameter configures a memory mapped I/O (MMIO) size of 256 GB. You can adjust this value later in the settings of the vSphere Bitfusion virtual machine.
7. On the **Select storage** page, define where and how to store the files for the deployed OVF template, and click **Next**.
8. On the **Select networks** page, select the network to use with Network Adapter 1 and click **Next**.

   The network you use with Network Adapter 1 carries management traffic. You can use the same network concurrently for data traffic or add additional network adapters later.

What to do next

The vSphere Bitfusion appliance requires several custom parameters. Complete the **Customize template** page to customize the deployment properties of the OVF template. See Customize the vSphere Bitfusion Appliance OVF Template.

Customize the vSphere Bitfusion Appliance OVF Template

As part of the vSphere Bitfusion appliance deployment, you must specify several custom parameters in the **Deploy OVF Template** dialog box.

On the **Customize template** page of the **Deploy OVF Template** dialog box, you customize the deployment properties of the OVF template.

Prerequisites

- Verify that you have the vCenter Server TLS Certificate Thumbprint. See Locate the vCenter Server TLS Certificate Thumbprint.
- Verify that the DNS and NTP services you use are set up properly in your environment.
  - If you use DHCP, verify that it provides DNS and NTP addresses.
  - If you do not use DHCP, specify the DNS and NTP server addresses in the OVF template.

  Note Clock synchronization is important for the functionality of vSphere Bitfusion.

Procedure

1. In the **Bitfusion Server Setup** section, specify a hostname for the server.

   Valid characters for hostnames are the ASCII characters A through Z (both upper- and lower-case), the digits 0 to 9, and the hyphen (-). A hostname cannot start with a hyphen. The hostname is stored in `/etc/hostname`.

2. In the **Bitfusion Server Setup** section, enter the vCenter Server GUID and URL as displayed in the address bar of your web browser.

   For example, if the navigation bar displays the URL `https://example.vslab.local/ui/app/vm;nav=h/urn:vmomi:VirtualMachine:vm-4450:612d27ff-d297-4573-bdc0-2c0dac8589a5/summary`, the vCenter Server URL is `https://example.vslab.local` and the GUID is `612d27ff-d297-4573-bdc0-2c0dac8589a5`.

3. In the **Bitfusion Server Setup** section, enter the user name and password for the vCenter Server instance on which you are deploying the vSphere Bitfusion OVF template.

4. In the **Bitfusion Server Setup** section, enter the vCenter Server TLS Certificate Thumbprint.

5. In the **Credentials** section, specify a customer password.

   After the deployment is complete, you use the customer user account to log into the vCenter Server appliance using the console shell or SSH.

6 (Optional) In the **NVIDIA Driver** section, select the **Download and Install NVIDIA Driver** check box to accept the NVIDIA license.

By accepting the NVIDIA license, vSphere Bitfusion downloads and installs the NVIDIA driver, CUDA libraries, and NVIDIA Fabric Manager during the first boot of the virtual machine.

Note If you are operating vSphere Bitfusion in an environment without access to the Internet, for example, by using an air-gapped network, do not select the check box. You must manually download and install the NVIDIA software after deploying the vSphere Bitfusion appliance.

7 In the **Network Adapter** section, specify the networking configuration for your environment.

You must specify the configuration for Network Adapter 1 that is used for management and data traffic. Network Adapter 1 must be connected to a network that communicates with the vCenter Server instance.

Network Adapters 2, 3, and 4 are optional and are used for data traffic only. Each network adapter must be connected to a separate network. vSphere Bitfusion chooses the network that is most efficient for data transfers to the vSphere Bitfusion server.

Option	Description
IPv4 Address	Enter the IPv4 address of the network adapter. If you are using DHCP, leave this text box blank. **Note** IPv6 is not supported.
Network CIDR Prefix	Enter the network Classless Inter-Domain Routing (CIDR) settings. For example, if your network uses a `/24` netmask, choose `24 (255.255.255.0)` from the drop-down menu.
MTU	Enter an MTU size. The default value is 1500. For optimal performance, specify an MTU size that is equal to the maximum MTU size supported by your network hardware. **Note** If you set an MTU size greater than 1500, verify that the network switches in your data center are enabled for jumbo frames.
Gateway	Enter the network gateway address to use with the appliance. If you are using DHCP, leave this text box blank.
DNS	Enter the DNS server address to use with the appliance. If you are using DHCP, leave this text box blank.
DNS Search Domains	Enter the DNS search domain address to use with the appliance. If you are using DHCP, leave this text box blank.
NTP	Enter the NTP server address to use with the appliance. If you are using DHCP and the DHCP server supports sending NTP server information, leave this text box blank.
Configure Network Adapter	Select the check box if you want to configure Network Adapter 2. Repeat for each subsequent network adapter.

8 Click **Next**.

9 On the **Ready to complete** page, review the vSphere Bitfusion server configuration and click **Finish**.

Results

A new task for creating the vSphere Bitfusion appliance appears in the Recent Tasks pane. After the task finishes, the new appliance is created on the selected resource.

What to do next

- Pass through the GPUs to the vSphere Bitfusion virtual machine. See Pass Through a GPU to the vSphere Bitfusion Virtual Machine.
- If you chose not to download and install the NVIDIA driver, CUDA libraries, and NVIDIA Fabric Manager during first boot, you must install the software manually. See Chapter 7 Installing NVIDIA Software for Use with vSphere Bitfusion.
- You can add more network adapters for data traffic. See *Modifying the Network Settings of a vSphere Bitfusion Server* in the *VMware vSphere Bitfusion User Guide*.

Pass Through a GPU to the vSphere Bitfusion Virtual Machine

VMDirectPath I/O allows the guest operating system to access the GPU directly, bypassing the ESXi hypervisor. By using passthrough devices, you can use resources more efficiently and improve the performance of your vSphere Bitfusion environment. Enabling passthrough of the GPU provides a level of performance on vSphere close to that of its native system.

Prerequisites

- Verify that you have the privileges that you need for the task that you plan to perform.
  - Verify that you have the **Virtual machine.Configuration.Add or remove device** privilege.
  - Verify that you have the **Virtual machine.Configuration.Advanced configuration** privilege.
  - If you plan to increase the memory reservation when you edit a virtual machine, verify that you have the **Virtual machine.Configuration.Change resource** privilege.
  - Verify that you have the **Virtual machine.Configuration.Change Memory** privilege.
- Verify that the virtual machine of the vSphere Bitfusion server is powered off.
- To use DirectPath, verify that Intel Virtualization Technology for Directed I/O (VT-d) or AMD I/O Virtualization Technology (IOMMU) is enabled in the BIOS of the ESXi host.
- Verify that the GPU PCI devices are connected to the host and marked as available for passthrough. See Enable a GPU for Passthrough.

- If your ESXi host is configured to boot from a USB device, or if the active coredump partition is configured to be on a USB device or SD cards connected through USB channels, disable the USB controller for passthrough.

  Note VMware does not support a USB controller passthrough for ESXi hosts that boot from USB devices or SD cards connected through USB channels. A configuration in which the active coredump partition is configured to be on a USB device or SD card connected through USB channels is also not supported. For information, see http://kb.vmware.com/kb/1021345.

Procedure

1 Add a GPU device.

   a In the vSphere Client, right-click the vSphere Bitfusion virtual machine in the inventory and select **Edit Settings**.

   b On the **Virtual Hardware** tab, click the **Add New Device** button.

   c From the drop-down menu, under **Other Devices**, select **PCI Device**.

   d Expand the **New PCI device** section and select the access type.

   e In the **New PCI device** section, select a GPU from the **PCI device** drop-down menu.

   Note By default, the same PCI device address is listed for each new GPU. When adding multiple devices, you must select the PCI addresses of each individual device.

   f Click **OK**.

2 Configure the CPU and memory resources of the ESXi host.

   If the ESXi host is a dedicated vSphere Bitfusion server, set the CPU and memory to their maximum values. If the host machine is not dedicated to vSphere Bitfusion, specify the minimum CPU value as the number of GPUs multiplied by 4, and the minimum memory as 1.5 times that of the aggregated GPU card memory or 32 GB, whichever is higher.

   a In the vSphere Client, right-click the vSphere Bitfusion virtual machine and select **Edit Settings**.

   b Expand the **CPU** section and edit the resources.

   c Expand the **Memory** section and edit the resources.

   d Under **Memory**, select the **Reserve all guest memory (All locked)** check box.

   e Click **OK**.

3 Adjust the memory mapped I/O (MMIO) size.

   By default, the vSphere Bitfusion installer configures an MMIO size of 256 GB. To calculate how much actual memory you must reserve for MMIO, consider the following MMIO memory calculations for two and three cards, each with 16 GB of memory.

   - `2 x 16 GB = 32`. Round 32 GB to the next power of 2, and the memory mapped I/O size that is needed is 64 GB.

- `3 x 16 GB = 48`. Round 48 GB to the next power of 2, and memory mapped I/O size that is needed is 64 GB.

a In the vSphere Client, right-click the vSphere Bitfusion virtual machine and select **Edit Settings**.

b Click **VM Options**, and expand the **Advanced** section.

c Under **Configuration Parameters**, click **Edit Configuration**.

d In the **Configuration Parameters** dialog box, locate the parameter `pciPassthru.64bitMMIOSizeGB`, and enter the MMIO size in gigabytes.

e Click **OK**.

4 (Optional) Take a snapshot of the virtual machine.

Snapshots capture the state of the virtual machine at the time you take the snapshot. If an error occurs when you start the virtual machine, you can recover your vSphere Bitfusion installation from the snapshot.

a In the vSphere Client, right-click the vSphere Bitfusion virtual machine and select **Snapshots > Take Snapshot**.

b Enter a name and description for the snapshot.

c Click **Create**.

5 If you are deploying a subsequent vSphere Bitfusion server, enable it.

Note When you enable an additional vSphere Bitfusion server, the server recognizes the primary vSphere Bitfusion server you previously deployed. If you do not perform this step before you power on the vSphere Bitfusion virtual machine, your subsequent server becomes primary and the configuration of the vSphere Bitfusion cluster is overwritten.

a In the vSphere Client, right-click the virtual machine in the inventory and select **Bitfusion > Enable Bitfusion.**.

b In the **Enable Bitfusion** dialog box, select the **For a server, this will allow if to used used as a GPU server** radio button and click **Enable**.

6 In the vSphere Client, right-click the vSphere Bitfusion virtual machine and select **Power > Power On**.

If you are powering on multiple vSphere Bitfusion servers, power them on one at a time. Wait three or more minutes between powering on each server.

Results

After the virtual machine powers on, allow it to run for ten or more minutes before performing any further configuration tasks or operations. During this time the virtual machine registers with vCenter Server and, downloads and installs the NVIDIA driver if you choose this installation option.

What to do next

- If you chose not to download and install the NVIDIA driver during first boot, you can now manually install the driver. See Install the NVIDIA Software for vSphere Bitfusion from the Internet.
- Verify that the vSphere Bitfusion plug-in registers with vCenter Server. See Verify That the vSphere Bitfusion Plug-In Registers with vCenter Server.

Verify That the vSphere Bitfusion Plug-In Registers with vCenter Server

After deploying the vSphere Bitfusion server, and installing the NVIDIA driver, verify that the vSphere Bitfusion plug-in appears in the vSphere Client.

Prerequisites

- Restart the vSphere Bitfusion virtual machine before verifying that vSphere Bitfusion plug-in appears in the vSphere Client.

  > **Note** If you chose to install the NVIDIA driver as part of the vSphere Bitfusion deployment, the server restarts itself.

- Allow the virtual machine to run for ten or more minutes before performing any further configuration tasks or operations. During this time, the virtual machine registers with vCenter Server.

Procedure

1. Open a web browser and enter the URL for your vCenter Server instance: `https://`*`vcenter_server_ip_address_or_fqdn`*.
2. Select **Launch vSphere Client (HTML5)**.
3. Enter the credentials of a user who has permissions on vCenter Server, and click **Login**.
4. (Optional) To update all data in the current vSphere Client view, click the refresh icon (↻).
5. In the vSphere Client, select **Menu > Bitfusion**.

   The vSphere Bitfusion plug-in loads.

What to do next

If the vSphere Bitfusion plug-in is registered properly, you can deploy additional vSphere Bitfusion servers and clients.

If the vSphere Bitfusion plug-in is not working or is not available, verify that the NTP, DNS, GUID, and SHA1 settings are configured properly in the OVF template. See Customize the vSphere Bitfusion Appliance OVF Template.

Add Subsequent vSphere Bitfusion Servers

You can add more servers to your vSphere Bitfusion cluster when you require more GPU resources.

After the primary vSphere Bitfusion server starts, vSphere Bitfusion registers a vSphere Bitfusion plug-in in the vCenter Server, resulting in a single vSphere Bitfusion cluster containing one vSphere Bitfusion server. After the vSphere Bitfusion plug-in is registered, you can add subsequent servers by following the steps in this procedure. The vSphere Bitfusion plug-in uses the primary server's configuration data, which allows faster deployment of the subsequent servers.

Alternatively, you can add a new server in your vSphere Bitfusion cluster by following the deploy procedure for the primary server. You deploy the vSphere Bitfusion appliance on a virtual machine (VM), customize the vSphere Bitfusion OVF template, pass through the GPUs to the vSphere Bitfusion server VM, and enable the VM as a vSphere Bitfusion server.

Additional vSphere Bitfusion servers must be part of the same vCenter Server instance as the first vSphere Bitfusion server.

Prerequisites

- Verify you have installed a primary vSphere Bitfusion server.
- Verify that the vSphere Bitfusion is registered with vCenter Server server.

Procedure

1. From the **Hosts and Clusters** view in vCenter Server, right-click an ESXi host, and select **Bitfusion > Install Bitfusion server**.

   The **Install Bitfusion server** dialog box appears.

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 VM, and click **Next**.
6. On the **Select networks** page, specify the networking configuration for Network Adapter 1 and click **Next**.

   You must specify the configuration for Network Adapter 1 that is used for management and data traffic. Network Adapter 1 must be connected to a network that communicates with the vCenter Server instance.

If your vSphere Bitfusion server requires additional network adapters for data traffic, you can click **Add Network Adapter** and specify the network configuration for the additional adapter.

Option	Description
Network Adapter	Select a network from the drop-down menu.
Adapter Type	Select a network adapter to assign to the virtual machine. **Note** vSphere Bitfusion supports VMXNET3 and PVRDMA adapters.
DHCP/Fixed IP	Specify whether a DHCP server assigns the address of the network adapter or you use a fixed IPv4 address.
IPv4 Address	Enter the IPv4 address of the network adapter. If you are using DHCP, leave this text box blank. **Note** IPv6 is not supported.
Netmask	Select a netmask from the drop-down menu. For example, if your network uses a `/24` netmask, select `24 (255.255.255.0).` .
Gateway	Enter the network gateway address to use with the appliance. If you are using DHCP, leave this text box blank.
MTU	Enter an MTU size. The default value is 1500. For optimal performance, specify an MTU size that is equal to the maximum MTU size supported by your network hardware. **Note** If you set an MTU size greater than 1500, verify that the network switches in your data center are enabled for jumbo frames.
DNS Servers	Enter the DNS server address to use with the appliance. If you are using DHCP, leave this text box blank.
DNS Search Domains	Enter the DNS search domain address to use with the appliance. If you are using DHCP, leave this text box blank.
NTP	Enter the NTP server address to use with the appliance. If you are using DHCP and the DHCP server supports sending NTP server information, leave this text box blank.

7 On the **Select GPUs** page, add GPUs to the subsequent server and click **Next**.

   a Click **Add GPU**.

   b Select a GPU from the **GPU Device** drop-down menu.

c (Optional) Specify the total memory of the GPU.

The vSphere Bitfusion plug-in uses the aggregated GPU memory of all GPUs you add on the **Select GPUs** page to calculate the values for the minimum memory and the recommended memory mapped I/O size of the virtual machine of your vSphere Bitfusion server.

d (Optional) To accept the NVIDIA license, select the **Download and Install NVIDIA Driver** check box.

By accepting the NVIDIA license, vSphere Bitfusion downloads and installs the NVIDIA driver, CUDA libraries, and NVIDIA Fabric Manager during the first boot of the virtual machine.

Note If you are operating vSphere Bitfusion in an environment without access to the Internet, for example, by using an air-gapped network, do not select the check box. You must manually download and install the NVIDIA software after deploying the vSphere Bitfusion appliance.

If your vSphere Bitfusion server requires additional GPUs, you can click **Add GPU Device** again and specify the settings for the GPU.

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

a Specify the number of CPUs for the virtual machine.

b Specify the memory mapped I/O (MMIO) size of the virtual machine in GB.

c (Optional) Enter a password for the customer account.

After the deployment is complete, you use the customer user account to log into the vSphere Bitfusion server by using the console shell or SSH. If you skip this step, you cannot log into the subsequent server.

d (Optional) Select the **Power On VM After Create** check box.

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**.

Results

A new task for installing the vSphere Bitfusion server appears in the Recent Tasks pane. After the task finishes, the new appliance is created on the selected resources.

When a new vSphere Bitfusion server joins the cluster, vCenter Server supplies a token, a certificate, and a configuration to access the vSphere Bitfusion cluster.

Installing the vSphere Bitfusion Client

5

You run the AI and ML applications on a vSphere Bitfusion client. Since vSphere Bitfusion 2.5, you can install and enable a vSphere Bitfusion client on any machine.

System Requirements for vSphere Bitfusion Client

- The minimum disk space requirement for a vSphere Bitfusion client is 2 GB.
- The minimum memory requirement for a vSphere Bitfusion client is at least 150% of the GPU memory that applications request to use.
- The minimum virtual CPU (vCPU) requirement for a vSphere Bitfusion client is the same as the requirement for running applications with dedicated, local GPUs.
- The vSphere Bitfusion client must be installed on a machine that uses one of the following operating systems.
  - CentOS 7
  - CentOS 8
  - Red Hat Linux 7.4 or later
  - Ubuntu 16.04
  - Ubuntu 18.04
  - Ubuntu 20.04

Prerequisite for VMs

If the vSphere Bitfusion client runs on a virtual machine (VM), all VMware Tools Scripts must be enabled. When creating a new virtual machine, the scripts are enabled in the default configuration.

Additional Prerequisites for VMs

If the vSphere Bitfusion client runs on a VM that is part of the same vCenter Server instance as the vSphere Bitfusion servers, additional system requirements apply.

- vSphere Bitfusion client VM must run on a vSphere deployment managed by vCenter Server 7.0.
- The vSphere Bitfusion client must be installed on an ESXi host with version 6.7 or later.
- All VMware Tools Scripts must be enabled. When creating a new virtual machine, the scripts are enabled in the default configuration.

vSphere Bitfusion Client Enablement

You can enable a client in one of the following ways.

- If the client does not run on a VM that is part of the same vCenter Server instance as the vSphere Bitfusion servers, see Generate a Client Authentication Token.
- If the client runs on a VM that is part of the same vCenter Server instance as the vSphere Bitfusion servers, see Enable the vSphere Bitfusion Client .

  The vSphere Bitfusion plug-in must be registered with vCenter Server. See Verify That the vSphere Bitfusion Plug-In Registers with vCenter Server.

Required Ports for vSphere Bitfusion Client

Verify that the following ports are not blocked by using a denylist or firewall rules. A vSphere Bitfusion client communicates with a vSphere Bitfusion server on the following ports.

Since vSphere Bitfusion 3.5, due to security concerns, the TLSv1.0 and TLSv1.1 protocols are disabled by default. For more information, see Knowledge Base article 2145796.

Port	Description
45201 - 46225	These ports are used by vSphere Bitfusion clients to communicate with the vSphere Bitfusion processes that handle CUDA requests.
55001 - 55201	These ports are used by vSphere Bitfusion clients to communicate with a task specific dispatcher process that runs on the vSphere Bitfusion server. The process requests a session and starts the service workers of the vSphere Bitfusion server that handle the workload.
56001	This port is used for vSphere Bitfusion intercommunication. vSphere Bitfusion servers communicate with each other on this port and vSphere Bitfusion clients use this port to start a task on a vSphere Bitfusion server.

Web Browser Requirements for vCenter Server

To use vSphere Bitfusion, you require a web browser version that is supported by vCenter Server. For more information, see vSphere Client Software Requirements.

vSphere Bitfusion Compatibility and Interoperability

For a list of versions, models, and products that are compatible with vSphere Bitfusion, see the VMware vSphere Bitfusion Compatibility and Interoperability page.

This chapter includes the following topics:

- Install the vSphere Bitfusion Client on CentOS and Red Hat
- Install the vSphere Bitfusion Client on Ubuntu

Install the vSphere Bitfusion Client on CentOS and Red Hat

You can install the vSphere Bitfusion client on CentOS and Red Hat.

Prerequisites

- Verify that the version of your CentOS or Red Hat operating system is supported. See Chapter 5 Installing the vSphere Bitfusion Client.
- Verify that the version of the vSphere Bitfusion client you install is the same as the version of your vSphere Bitfusion servers or earlier. See Chapter 9 Upgrading vSphere Bitfusion.

Procedure

1. For CentOS, install the Extra Packages for Enterprise Linux (EPEL), an additional package repository that provides access to install packages for commonly used software.

   ```
   sudo yum install -y epel-release
   ```

2. Add the public key for VMware Bitfusion to the GNU Privacy Guard (GPG).

   ```
   sudo rpm --import https://packages.vmware.com/bitfusion/vmware.bitfusion.key
   ```

3. Download your vSphere Bitfusion client version from the VMware website at https://packages.vmware.com/bitfusion/centos/.

   For example, run `wget https://packages.vmware.com/bitfusion/centos/8/bitfusion-client-centos8-3.5.0-5.x86_64.rpm`

4. Install the client package by running the `sudo yum install -y ./`**`bitfusion_client_version`** command, where **`bitfusion_client_version`** is the filename of the vSphere Bitfusion client.

   For example, `sudo yum install -y ./bitfusion-client-centos8-3.5.0-5.x86_64.rpm`.

5 (Optional) Verify the version of the vSphere Bitfusion client.

```
bitfusion version
Bitfusion version 3.5.0
```

What to do next

Enable the vSphere Bitfusion client on the client virtual machine (VM). See Enable the vSphere Bitfusion Client . If the client is not installed on a VM that is a part of the same vCenter Server instance as the servers, see Generate a Client Authentication Token.

Install the vSphere Bitfusion Client on Ubuntu

You can install the vSphere Bitfusion client on Ubuntu.

Prerequisites

- Verify that the version of your Ubuntu operating system is supported. See Chapter 5 Installing the vSphere Bitfusion Client.
- Verify that the version of the vSphere Bitfusion client you install is the same as the version of your vSphere Bitfusion servers or earlier. See Chapter 9 Upgrading vSphere Bitfusion.

Procedure

1 Download the vSphere Bitfusion client for your Linux distribution from the VMware website at https://packages.vmware.com/bitfusion/ubuntu/.

For example, run `wget https://packages.vmware.com/bitfusion/ubuntu/20.04/bitfusion-client-ubuntu2004_3.5.0-5_amd64.deb`.

2 Update the package by running the `apt-get update` command.

```
sudo apt-get update
```

3 Install the package by running the `sudo apt-get install -y ./`**`bitfusion_client_version`** command, where **`bitfusion_client_version`** is the filename of the vSphere Bitfusion client.

For example, `sudo apt-get install -y ./bitfusion-client-ubuntu2004_3.5.0-5_amd64.deb`

4 Verify the version of the vSphere Bitfusion client.

```
bitfusion version
Bitfusion version 3.5.0
```

What to do next

Enable the vSphere Bitfusion client on the client virtual machine (VM). See Enable the vSphere Bitfusion Client . If the client is not installed on a VM that is a part of the same vCenter Server instance as the servers, see Generate a Client Authentication Token.

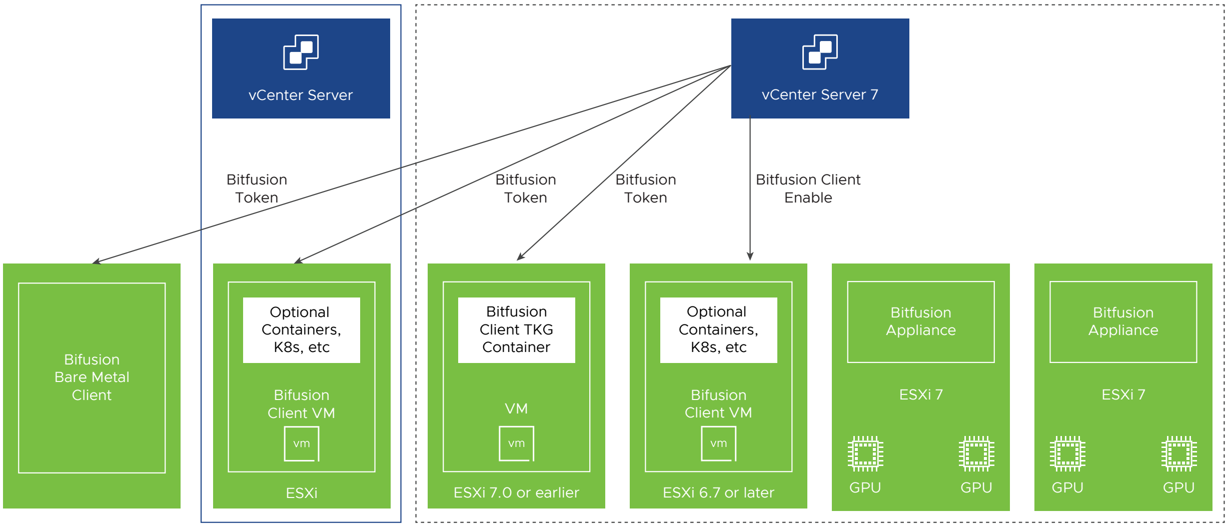
Enabling the vSphere Bitfusion Client 6

Since vSphere Bitfusion 2.5 you can install and enable a vSphere Bitfusion client on multiple platforms.

There are two ways of enabling a vSphere Bitfusion client.

- For clients in the same vCenter Server instance as your servers, you can enable the client from the vSphere Bitfusion Plug-in.
- In vSphere Bitfusion 2.5 you can enable a client on Tanzu Kubernetes Grid (TKG) containers, different vCenter Server instances, and bare metal machines. By using the vSphere Bitfusion Plug-in, you can generate an authorization token and use it to enable a single or multiple clients. You can create multiple tokens to enable groups of clients. To manage the clients or client groups, you can enable or disable a specific token.

The following figure displays the available enablement options for a vSphere Bitfusion client on multiple platforms.



This chapter includes the following topics:

- Enable the vSphere Bitfusion Client
- Generate a Client Authentication Token

Enable the vSphere Bitfusion Client

You enable the vSphere Bitfusion client on the client virtual machine (VM).

Note Bitfusion clients must be part of the same vCenter Server instance as the Bitfusion servers. To add a client that is installed outside of the vCenter Server instance, see Generate a Client Authentication Token.

Prerequisites

- Install the vSphere Bitfusion client for your Linux distribution. See Install the vSphere Bitfusion Client on CentOS and Red Hat and Install the vSphere Bitfusion Client on Ubuntu.
- Power off the vSphere Bitfusion client VM.

Procedure

1. In the vCenter Server inventory, right-click the vSphere Bitfusion client VM and select **Bitfusion > Enable Bitfusion**.
2. In the **Bitfusion Enablement** dialog box, select the **For a client, this will allow users to run Bitfusion workloads** radio button, and click **Enable**.
3. Power on the client VM.
4. In the virtual machine terminal, add users to the vSphere Bitfusion group by using the `sudo usermod -aG bitfusion` *`username`* command.
5. (Optional) Verify that the users were successfully added to the vSphere Bitfusion group.

   a. Log out and log in the vSphere Bitfusion virtual machine terminal.

   Note If you do not log out and back in to the virtual machine terminal, the new users and their group assignments do not register.

   b. In the virtual machine terminal, run the `groups` command to list users and their associated groups.

   ```
   groups
   testuser bitfusion
   ```

6. (Optional) Verify that the vSphere Bitfusion client is working by listing the available GPUs in the vSphere Bitfusion deployment by running the `bitfusion list_gpus` command.

   ```
   /home/bitfusion$ bitfusion list_gpus
   - server 0 [10.202.8.185:56001]: running 0 tasks
   |- GPU 0: free memory 16160 MiB / 16160 MiB
   |- GPU 1: free memory 16160 MiB / 16160 MiB
   |- GPU 2: free memory 16160 MiB / 16160 MiB
   |- GPU 3: free memory 16160 MiB / 16160 MiB
   ```

Results

You have successfully enabled the vSphere Bitfusion client.

What to do next

Start an application in the vSphere Bitfusion client. After the first run, the vSphere Bitfusion client joins the cluster.

Generate a Client Authentication Token

Enable a vSphere Bitfusion client that is installed on a Tanzu Kubernetes Grid (TKG) container, a different vCenter Server instance, or a bare metal machine.

To enable a vSphere Bitfusion client that is not part of the same vCenter Server instance as your servers, follow the procedure. You must generate an authorization token, download the related `tar` file, and extract the contents of the file in the filesystem of your client.

To enable a client that is part of the same vCenter Server instance as your servers, see Enable the vSphere Bitfusion Client .

Prerequisites

- Verify that you have installed a vSphere Bitfusion 2.5 server or later.
- Verify that the version of your vSphere Bitfusion client is the same version as your vSphere Bitfusion servers or earlier. See Chapter 9 Upgrading vSphere Bitfusion.
- Verify that the vSphere Bitfusion client has network access to the servers in your cluster.

Procedure

1. In the vSphere Client, select **Menu > Bitfusion**.
2. On the **Tokens** tab, select **New Token**.
3. In the **Create Token** dialog box, enter a description, and click **Create**.
4. Select the token from the list, click **Download**, and save the `tar` file to your local machine.
5. Copy the `tar` file to the filesystem of your client machine or machines.
6. Extract the contents of the `tar` file and copy them 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. In the terminal of the machine, add users to the Bitfusion group by running the `sudo usermod -aG bitfusion` *`username`* command.

8 (Optional) Verify that the users were successfully added to the vSphere Bitfusion group.

   a Log out and log in the terminal of the machine.

   b In the terminal, run the `groups` command.

   The users and their associated groups are listed.

9 (Optional) Verify that the vSphere Bitfusion client is working by listing the available GPUs in the vSphere Bitfusion deployment by running the `bitfusion list_gpus` command.

Results

You have successfully enabled the vSphere Bitfusion client.

What to do next

Start an application in the vSphere Bitfusion client. After the first run, the vSphere Bitfusion client joins the cluster.

Installing NVIDIA Software for Use with vSphere Bitfusion

7

If you chose not to download and install the NVIDIA driver, CUDA library, and the NVIDIA Fabric Manager during the initial boot of the vSphere Bitfusion server virtual machine, you must install the software manually.

There are three different installation methods of the NVIDIA software depending on your vSphere Bitfusion cluster environment.

- Installation directly from the Internet.
- Installation in an air-gapped network environment with a local web server.
- Installation in an air-gapped network environment without a local web server.

This chapter includes the following topics:

Install the NVIDIA Software for vSphere Bitfusion from the Internet

You can manually install the NVIDIA software for your vSphere Bitfusion deployment. Follow this procedure if you chose not to download and install the NVIDIA driver, CUDA library, and NVIDIA Fabric Manager during the initial boot of the vSphere Bitfusion server virtual machine (VM) and your vSphere Bitfusion has access to the Internet.

You can skip this procedure if you chose to download and install the NVIDIA software during the initial boot of the vSphere Bitfusion server VM.

Prerequisites

- The use of the NVIDIA driver implies acceptance of the NVIDIA Software License Agreement. See License For Customer Use of NVIDIA Software.
- The NVIDIA driver certified for use with vSphere Bitfusion is `NVIDIA-Linux-x86_64-460.32.03.run`.
- The CUDA library that is necessary for NCCL operations and certified for use with vSphere Bitfusion is `cuda_11.2.0_460.27.04_linux.run`.

- The NVIDIA Fabric Manager package certified for use with vSphere Bitfusion is `nvidia-fabricmanager-460-460.32.03-1.x86_64.rpm`.

Procedure

1. Log in to the appliance shell of the vSphere Bitfusion server VM.

   ```
   ssh customer@bitfusion_server_IP_address
   ```

2. To install the NVIDIA driver, CUDA library, and NVIDIA Fabric Manager, run the `sudo install-nvidia-packages --defaults --yes` command.

3. Restart the VM.

Results

As the vSphere Bitfusion server VM powers on, allow the VM to run for 10 minutes or longer before performing any further configuration tasks or operations. During this time, the vSphere Bitfusion server registers with vCenter Server.

What to do next

Verify That the vSphere Bitfusion Plug-In Registers with vCenter Server

Install the NVIDIA Software in an Air Gapped Network Environment

You can manually install the NVIDIA software in an environment with an air-gapped network. Follow this procedure task if you chose not to download and install the NVIDIA driver, CUDA library, and NVIDIA Fabric Manager during the initial boot of the vSphere Bitfusion server virtual machine (VM) and your vSphere Bitfusion does not have access to the Internet.

You can skip this procedure if you chose to download and install the NVIDIA software during the initial boot of the vSphere Bitfusion server VM.

Prerequisites

- The use of the NVIDIA driver implies acceptance of the NVIDIA Software License Agreement. See License For Customer Use of NVIDIA Software.
- The NVIDIA driver certified for use with vSphere Bitfusion is `NVIDIA-Linux-x86_64-460.32.03.run`. You can download the driver software from the NVIDIA's website: http://us.download.nvidia.com/tesla/460.32.03/NVIDIA-Linux-x86_64-460.32.03.run
- The CUDA library that is necessary for NCCL operations and certified for use with vSphere Bitfusion is `cuda_11.2.0_460.27.04_linux.run`. You can download the library from the NVIDIA's website: https://developer.download.nvidia.com/compute/cuda/11.2.0/local_installers/cuda_11.2.0_460.27.04_linux.run

- The NVIDIA Fabric Manager package certified for use with vSphere Bitfusion is `nvidia-fabricmanager-460-460.32.03-1.x86_64.rpm`. You can download the library from the NVIDIA's website: http://developer.download.nvidia.com/compute/cuda/repos/rhel7/x86_64/nvidia-fabricmanager-460-460.32.03-1.x86_64.rpm

Procedure

1 On a machine with access to the Internet, create and navigate to the `nvidia-packages` folder.

```
mkdir ~/nvidia-packages
cd ~/nvidia-packages
```

2 Download the NVIDIA driver, CUDA library, and NVIDIA Fabric Manager.

```
wget http://us.download.nvidia.com/tesla/460.32.03/NVIDIA-Linux-x86_64-460.32.03.run
wget https://developer.download.nvidia.com/compute/cuda/11.2.0/local_installers/
cuda_11.2.0_460.27.04_linux.run
wget http://developer.download.nvidia.com/compute/cuda/repos/rhel7/x86_64/nvidia-
fabricmanager-460-460.32.03-1.x86_64.rpm
```

3 Move and install the NVIDIA software.

Follow the procedure to install the NVIDIA driver, CUDA library, and NVIDIA Fabric Manager with by using either a local web server, or no web server, as appropriate for your vSphere Bitfusion network environment.

Option	Description
Option	Action

With a local web server	a To copy the NVIDIA software folder to the root directory or a similar directory on the local web server, run the `scp` following command.
	`scp ~/nvidia-packages/*`
	`mylogin@mylocalwebserver:/var/www/html/`
	b To log into the local web server, run the `mylogin` command.
	`ssh mylogin@mylocalwebserver@mylocalwebserver`
	c To give read permission to the NVIDIA driver, run the `chmod` command.
	`chmod +r /var/www/html/*`
	d To log into the vSphere Bitfusion server, run `ssh customer@bitfusion_server_ip_address`.
	e To install the NVIDIA software from the local web server, run the `install-nvidia-packages` command.
	`sudo install-nvidia-packages --yes --driver http://`
	`mylocalwebserver/NVIDIA-Linux-x86_64-460.32.03.run \`
	`    --cuda http://mylocalwebserver/`
	`cuda_11.2.0_460.27.04_linux.run \`
	`    --fm http://mylocalwebserver/nvidia-`
	`fabricmanager-460-460.32.03-1.x86_64.rpm`

No web server	a To copy the NVIDIA software to the vSphere Bitfusion server, run the `scp` command.
	`scp NVIDIA-Linux-x86_64-460.32.03.run`
	`customer@bitfusion_server_ip_address:~/`
	`scp cuda_11.2.0_460.27.04_linux.run`
	`customer@bitfusion_server_ip_address:~/`
	`scp nvidia-fabricmanager-460-460.32.03-1.x86_64.rpm`
	`customer@bitfusion_server_ip_address:~/`
	b To log into the vSphere Bitfusion server, run `ssh customer@bitfusion_server_ip_address`.
	c To install the NVIDIA software from the local file, run the `install-nvidia-packages` command.
	`sudo install-nvidia-packages --yes --driver NVIDIA-`
	`Linux-x86_64-460.32.03.run \`
	`    --cuda cuda_11.2.0_460.27.04_linux.run \`
	`    --fm nvidia-`
	`fabricmanager-460-460.32.03-1.x86_64.rpm`

4 Restart the VM.

Results

As the vSphere Bitfusion server VM powers on, allow the VM to run for 10 minutes or longer before performing any further configuration tasks or operations. During this time, the vSphere Bitfusion server registers with vCenter Server.

What to do next

Verify That the vSphere Bitfusion Plug-In Registers with vCenter Server

Use a Paravirtual RDMA Network Adapter with vSphere Bitfusion

8

You can use a Paravirtual RDMA (PVRDMA) adapter to improve the performance of your vSphere Bitfusion deployment.

RDMA provides direct memory access from the memory of one computer to the memory of another computer without involving the operating system or CPU. The transfer of memory is offloaded to the RDMA-capable Host Channel Adapters (HCA). A PVRDMA network adapter provides remote direct memory access in a virtual environment.

Prerequisites

- Your vSphere environment must have PVRDMA set up before you configure vSphere Bitfusion to use PVRDMA. To learn more, see the *vSphere Networking* documentation.
- vSphere Bitfusion servers and clients must be configured with two network adapters. Use the first network adapter for a management traffic using a default adapter type such as VMXNET3. Use the second network adapter for PVRDMA traffic.
- You must power off the vSphere Bitfusion server and client virtual machines prior toconfiguring them to use PVRDMA adapters.

Procedure

1. Locate the virtual machines hosting the vSphere Bitfusion servers and clients in the vSphere Client.
2. Right-click a virtual machine in the inventory and select **Edit Settings**.
3. From the **Add new device** drop-down menu, select **Network Adapter 2**.

   The New Network section is added to the list in the **Virtual Hardware** tab.
4. Select a PVRDMA network.
5. Expand the New Network section and connect the virtual machine to a distributed port group.
6. Change the **Status** setting to **Connect at power on**.
7. From the **Adapter type** drop-down menu, select PVRDMA.
8. Power on the virtual machine.

9. If you powered on a virtual machine that is hosting the vSphere Bitfusion clients, install the RDMA drivers.

   In addition to the RDMA drivers, diagnostic tools are installed.

   - For CentOS and Red Hat Linux, run the following command.

     ```
     yum install -y open-vm-tools rdma-core libibverbs libibverbs-utils infiniband-diags
     ```

   - For Ubuntu Linux, run the following command.

     ```
     sudo apt-get install -y rdma-core libmlx4-1 infiniband-diags ibutils ibverbs-utils
     rdmacm-utils perftest
     ```

Results

You have successfully enabled vSphere Bitfusion to use PVRDMA network adapters.

You can test the connection between the vSphere Bitfusion server and client by using the `ib_send_bw` (InfinBand send bandwidth) command. For example, if the IP addresses of the vSphere Bitfusion server and client are 192.168.10.10 and 192.16.10.11, run the following commands.

```
#From the server 192.16.10.10
ib_send_bw

#From the client 192.16.10.11 - connects to the server
ib_send_bw 192.168.10.10
```

The vSphere Bitfusion client writes a bandwidth report to standard output (stdout).

Upgrading vSphere Bitfusion 9

Since vSphere Bitfusion 2.5 you can perform an upgrade of your vSphere Bitfusion environment. By upgrading your cluster, you keep the current configuration data and monitoring statistics.

vSphere Bitfusion supports a multi-version server and client environment. All servers must run the same version of the vSphere Bitfusion OVA, while the client versions can be mixed. The version of a vSphere Bitfusion client must be the same as the version of your vSphere Bitfusion servers or earlier. When upgrading your vSphere Bitfusion environment, first upgrade your servers and then the clients.

vSphere Bitfusion version	Server version	Client version
2.0.0	2.0.0	2.0.0
2.0.1	2.0.1	2.0.1
2.0.2	2.0.2	2.0.2
2.5.0	2.5.0	2.0.0 2.0.1 2.0.2 2.5.0
2.5.1	2.5.1	2.0.0 2.0.1 2.0.2 2.5.0 2.5.1
3.0.0	3.0.0	2.0.0 2.0.1 2.0.2 2.5.0 2.5.1 3.0.0

vSphere Bitfusion version	Server version	Client version
3.0.1	3.0.1	2.0.0 2.0.1 2.0.2 2.5.0 2.5.1 3.0.0 3.0.1
3.5.0	3.5.0	2.0.0 2.0.1 2.0.2 2.5.0 2.5.1 3.0.0 3.0.1 3.5.0

This chapter includes the following topics:

- Upgrade a vSphere Bitfusion Cluster from 2.0 to 2.5
- Upgrade a vSphere Bitfusion Cluster from 2.5 to 3.0
- Upgrade a vSphere Bitfusion Cluster from 3.0 to 3.5

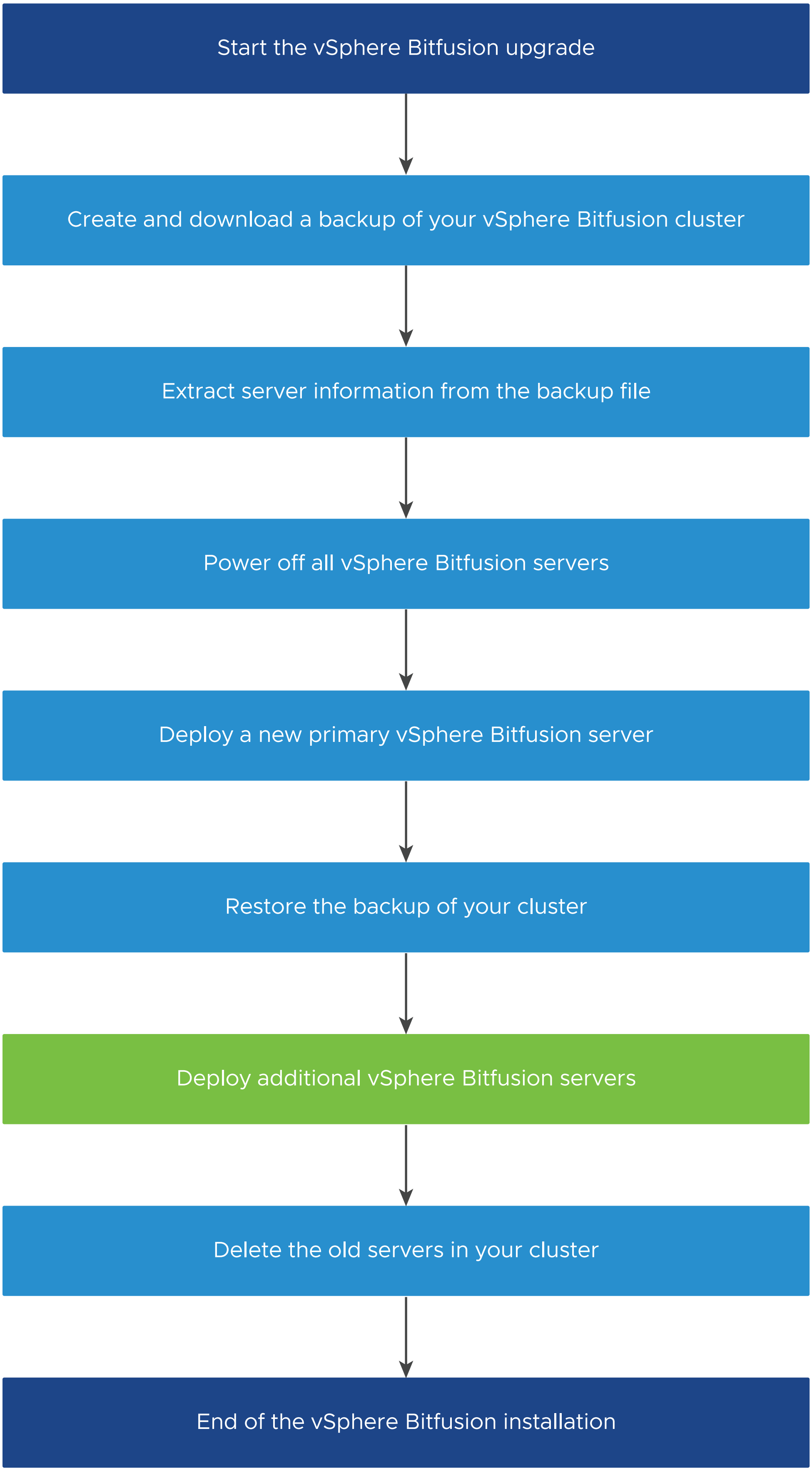
Upgrade a vSphere Bitfusion Cluster from 2.0 to 2.5

You can run artificial intelligence (AI) and machine learning (ML) workloads on vSphere Bitfusion 2.5 without losing your current cluster configuration and monitoring data.

To upgrade your cluster, you must back up the environment, deploy new server virtual machines (VMs) with version 2.5 of the vSphere Bitfusion appliance, and restore the backup.

You can use your current vSphere Bitfusion 2.0 clients or upgrade the clients to version 2.5. To upgrade a client, you must install the latest CentOS, Red Hat, or Ubuntu package. For more information, see Chapter 5 Installing the vSphere Bitfusion Client.

Figure 9-1. vSphere Bitfusion Upgrade Workflow


Start the vSphere Bitfusion upgrade
Create and download a backup of your vSphere Bitfusion cluster
Extract server information from the backup file
Power off all vSphere Bitfusion servers
Deploy a new primary vSphere Bitfusion server
Restore the backup of your cluster
Deploy additional vSphere Bitfusion servers
Delete the old servers in your cluster
End of the vSphere Bitfusion installation

1 By using the vSphere Bitfusion Plug-in, create and download a backup of your vSphere

Bitfusion 2.0 cluster.

For more information, see *Back up a Bitfusion Cluster* in the *VMware vSphere Bitfusion User Guide*.

2 From the downloaded `bitfusionbackup.tar.gz` archive file, open the `manifest.json` file, and find the servers section. The section includes information about the servers in your vSphere Bitfusion cluster at the time of the backup.

   Take note of the host ID, hostname, and number of GPUs for each server.

   For example, the host ID of the following server is `6a2f4e80-70d8-4c51-bf10-00284f3ed2c6`, the hostname is `bitfusion-server-2.0.1-3-1`, and the server has one GPU installed.

```
"servers": [
        {
            "id": "6a2f4e80-70d8-4c51-bf10-00284f3ed2c6",
            "hostname": "bitfusion-server-2.0.1-3-1",
            "ip": "10.202.8.209",
            "port": "56001",
            "address": "10.202.8.209:56001",
            "mode": "manager",
            "health": "PASS",
            "num_devices": 1,
            "lastseen": "2020-10-14T21:29:38Z",
            "license": {
                "type": "vcenter-license",
                "name": "vSphere 7 Enterprise Plus",
                "license-id": "example",
                "key": "example",
                "expiry": "2025-09-30T00:00:00Z"
            }
```

3 Power off all vSphere Bitfusion servers in the cluster.

4 Install a new primary vSphere Bitfusion server.

   a Deploy a new primary vSphere Bitfusion server VM by using a vSphere Bitfusion 2.5 Appliance OVF Template.

      For more information, see Chapter 4 Deploying the vSphere Bitfusion Appliance.

      During the deployment process, enter the same hostname as your primary vSphere Bitfusion 2.0 server uses.

   b In the settings of the new VM, add the same number of GPUs as your primary vSphere Bitfusion 2.0 server uses.

   c In the settings of the new VM, add a `guestinfo.bitfusion.server.host-id` configuration parameter. The parameter value must match the host ID of your primary server with version 2.0, that is listed in the `manifest.json` file.

      For more information, see *Edit Configuration File Parameters* in the *vSphere Virtual Machine Administration* documentation.

   d Power on the server and wait until the vSphere Bitfusion Plug-in is registered with vCenter Server.

5 By using the vSphere Bitfusion Plug-in, restore the backup of your vSphere Bitfusion 2.0 cluster.

   For more information, see *Restore a Bitfusion Cluster* in the *VMware vSphere Bitfusion User Guide*.

6 For each subsequent vSphere Bitfusion server in your cluster, perform the following steps.

   a Deploy a new server VM by using a vSphere Bitfusion 2.5 Appliance OVF Template.

      During the deployment process, enter the same hostname as the corresponding vSphere Bitfusion 2.0 server uses.

   b In the settings of the new VM, add the same number of GPUs as the corresponding vSphere Bitfusion 2.0 server uses.

   c In the settings of the new VM, add a `guestinfo.bitfusion.server.host-id` configuration parameter. The parameter value must match the host ID of the corresponding server with version 2.0, that is listed in the `manifest.json` file.

   d Enable the VM as a vSphere Bitfusion server.

      For more information, see Add Subsequent vSphere Bitfusion Servers.

   e Power on the VM. Multiple VMs must be powered on in a sequential order.

7 Delete the vSphere Bitfusion 2.0 server VMs.

The servers in your cluster are upgraded to version 2.5.

Upgrade a vSphere Bitfusion Cluster from 2.5 to 3.0

You can upgrade your vSphere Bitfusion cluster to version 3.0 and retain your current cluster configuration and monitoring data.

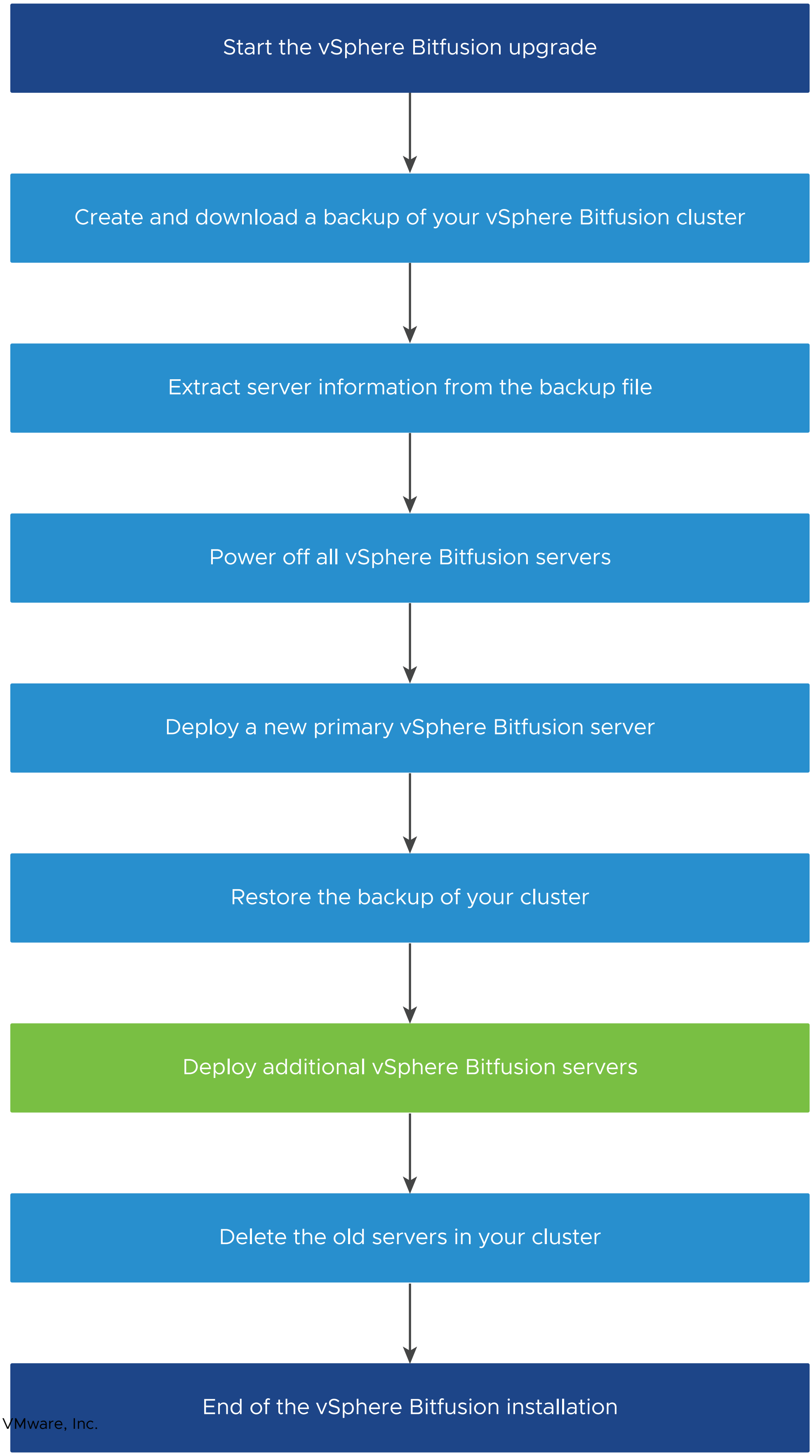
To upgrade your cluster, you must upgrade the servers in your vSphere Bitfusion environent. You must back up the environment, deploy new server virtual machines (VMs) with the latest version of the vSphere Bitfusion appliance, and restore the backup.

vSphere Bitfusion Clients Upgrade

You can use your current vSphere Bitfusion 2.x clients or upgrade the clients to version 3.0. To upgrade a client, you must install the latest package on your Ubuntu, CentOS, or Red Hat Linux operating system. The client version can be the same as the version of your vSphere Bitfusion servers or earlier. For more information, see Chapter 5 Installing the vSphere Bitfusion Client.

vSphere Bitfusion Servers Upgrade

Figure 9-2. vSphere Bitfusion Upgrade Workflow


Start the vSphere Bitfusion upgrade
Create and download a backup of your vSphere Bitfusion cluster
Extract server information from the backup file
Power off all vSphere Bitfusion servers
Deploy a new primary vSphere Bitfusion server
Restore the backup of your cluster
Deploy additional vSphere Bitfusion servers
Delete the old servers in your cluster
End of the vSphere Bitfusion installation

Bitfusion 2.5 cluster.

For more information, see *Back up a Bitfusion Cluster* in the *VMware vSphere Bitfusion User Guide*.

2. From the downloaded `bitfusionbackup.tar.gz` archive file, open the `manifest.json` file, and find the servers section. The section includes information about the servers in your vSphere Bitfusion cluster at the time of the backup.

   Take note of the host ID, hostname, and number of GPUs for each server.

   For example, the host ID of the following server is `6a2f4e80-70d8-4c51-bf10-00284f3ed2c6`, the hostname is `bitfusion-server-2.5.1-3-1`, and the server has one GPU installed.

   ```
   "servers": [
           {
               "id": "6a2f4e80-70d8-4c51-bf10-00284f3ed2c6",
               "hostname": "bitfusion-server-2.5.1-3-1",
               "ip": "10.202.8.209",
               "port": "56001",
               "address": "10.202.8.209:56001",
               "mode": "manager",
               "health": "PASS",
               "num_devices": 1,
               "lastseen": "2020-10-14T21:29:38Z",
               "license": {
                   "type": "vcenter-license",
                   "name": "vSphere 7 Enterprise Plus",
                   "license-id": "example",
                   "key": "example",
                   "expiry": "2025-09-30T00:00:00Z"
               }
   ```

3. Power off all vSphere Bitfusion servers in the cluster.

4. Install a new primary vSphere Bitfusion server.

   a. Deploy a new primary vSphere Bitfusion server VM by using a vSphere Bitfusion 3.0 Appliance OVF Template.

      For more information, see Chapter 4 Deploying the vSphere Bitfusion Appliance.

      During the deployment process, enter the same hostname as your primary vSphere Bitfusion 2.5 server uses.

   b. In the settings of the new VM, add the same number of GPUs as your primary vSphere Bitfusion 2.5 server uses.

   c. In the settings of the new VM, add a `guestinfo.bitfusion.server.host-id` configuration parameter. The parameter value must match the host ID of your primary server with version 2.5, that is listed in the `manifest.json` file.

      For more information, see *Edit Configuration File Parameters* in the *vSphere Virtual Machine Administration* documentation.

   d Power on the server and wait until the vSphere Bitfusion Plug-in is registered with vCenter Server.

5 By using the vSphere Bitfusion plug-in, restore the backup of your vSphere Bitfusion 2.5 cluster.

   For more information, see *Restore a Bitfusion Cluster* in the *VMware vSphere Bitfusion User Guide*.

6 For each subsequent vSphere Bitfusion server in your cluster, perform the following steps.

   a Deploy a new server VM by using the vSphere Bitfusion plug-in.

      During the deployment process, enter the hostname and host ID that are listed in the `manifest.json` for the corresponding vSphere Bitfusion 2.5 server. For more information, see Add Subsequent vSphere Bitfusion Servers.

   b In the settings of the new VM, add the same number of GPUs as the corresponding vSphere Bitfusion 2.5 server uses.

   c In the settings of the new VM, add a `guestinfo.bitfusion.server.host-id` configuration parameter. The parameter value must match the host ID of the corresponding server with version 2.5, file.

   d Power on the VM. Multiple VMs must be powered on in a sequential order.

7 Delete the vSphere Bitfusion 2.5 server VMs.

Upgrade a vSphere Bitfusion Cluster from 3.0 to 3.5

To use the latest vSphere Bitfusion version and retain your current cluster configuration and monitoring data, you can upgrade your vSphere Bitfusion cluster.

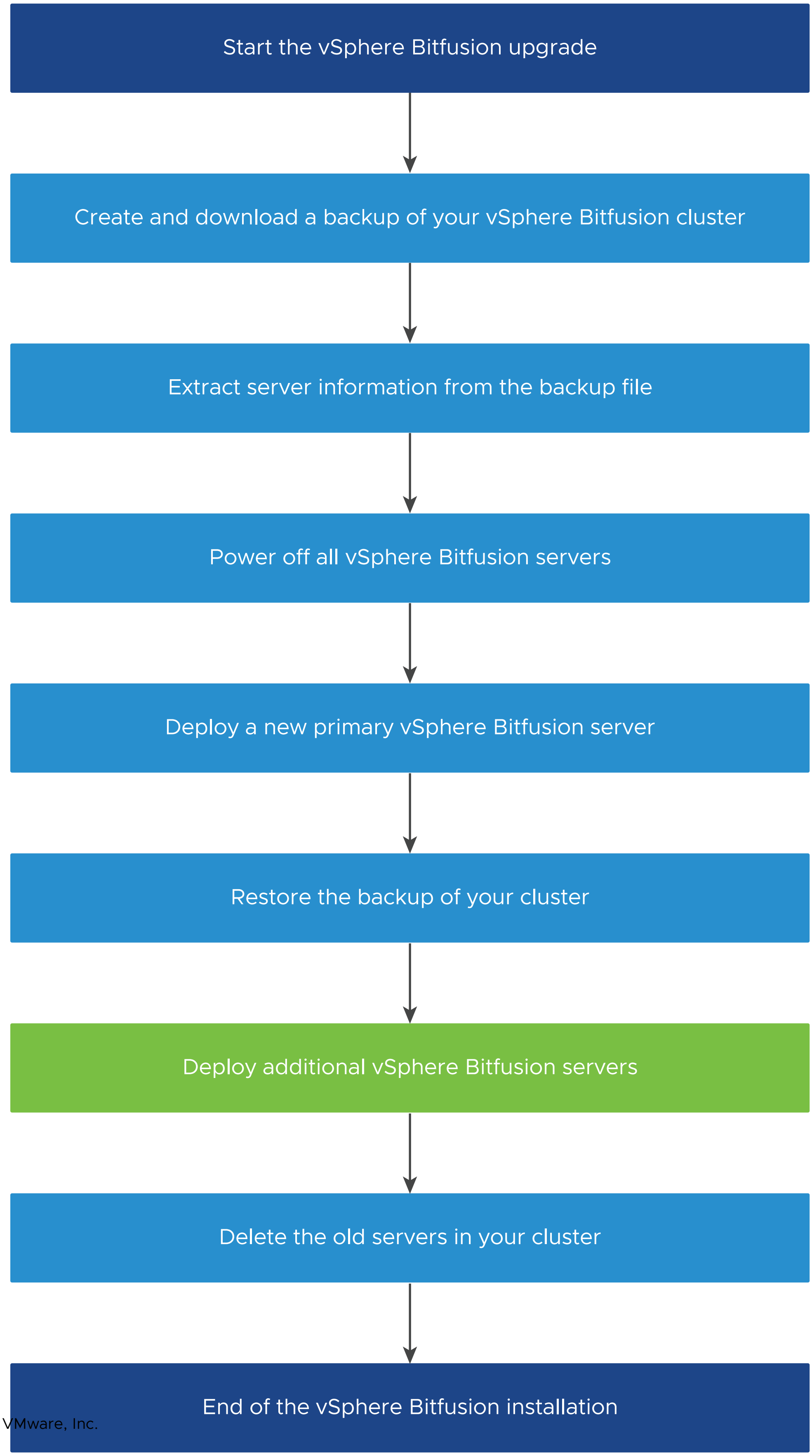
To upgrade your cluster, you must upgrade the servers in your vSphere Bitfusion environment. You must back up the environment, deploy new server virtual machines (VMs) with the latest version of the vSphere Bitfusion appliance, and restore the backup.

vSphere Bitfusion Clients Upgrade

You can use your current vSphere Bitfusion 2.x clients or upgrade the clients to version 3.5. To upgrade a client, you must install the latest package on your Ubuntu, CentOS, or Red Hat Linux operating system. The client version can be the same as the version of your vSphere Bitfusion servers or earlier. For more information, see Chapter 5 Installing the vSphere Bitfusion Client.

vSphere Bitfusion Servers Upgrade

Figure 9-3. vSphere Bitfusion Upgrade Workflow



Start the vSphere Bitfusion upgrade
Create and download a backup of your vSphere Bitfusion cluster
Extract server information from the backup file
Power off all vSphere Bitfusion servers
Deploy a new primary vSphere Bitfusion server
Restore the backup of your cluster
Deploy additional vSphere Bitfusion servers
Delete the old servers in your cluster
End of the vSphere Bitfusion installation

Bitfusion 3.0 cluster.

For more information, see *Back up a Bitfusion Cluster* in the *VMware vSphere Bitfusion User Guide*.

2 From the downloaded `bitfusionbackup.tar.gz` archive file, open the `manifest.json` file, and find the servers section. The section includes information about the servers in your vSphere Bitfusion cluster at the time of the backup.

Take a note of the host ID, hostname, and number of GPUs for each server.

For example, the host ID of the following server is `6a2f4e80-70d8-4c51-bf10-00284f3ed2c6`, the hostname is `bitfusion-server-3.0.1-4`, and the server has one GPU installed.

```
"servers": [
        {
            "id": "6a2f4e80-70d8-4c51-bf10-00284f3ed2c6",
            "hostname": "bitfusion-server-3.0.1-4",
            "ip": "10.202.8.209",
            "port": "56001",
            "address": "10.202.8.209:56001",
            "mode": "manager",
            "health": "PASS",
            "num_devices": 1,
            "lastseen": "2020-10-14T21:29:38Z",
            "license": {
                "type": "vcenter-license",
                "name": "vSphere 7 Enterprise Plus",
                "license-id": "example",
                "key": "example",
                "expiry": "2025-09-30T00:00:00Z"
            }
```

3 Power off all vSphere Bitfusion servers in the cluster.

4 Install a new primary vSphere Bitfusion server.

a Deploy a new primary vSphere Bitfusion server VM by using a vSphere Bitfusion 3.5 Appliance OVA Template.

For more information, see Chapter 4 Deploying the vSphere Bitfusion Appliance.

During the deployment process, enter the same hostname as your primary vSphere Bitfusion 3.0 server uses.

b In the settings of the new VM, add the same number of GPUs as your primary vSphere Bitfusion 3.0 server uses.

c In the settings of the new VM, add a `guestinfo.bitfusion.server.host-id` configuration parameter. The parameter value must match the host ID of your primary server with version 3.0, that is listed in the `manifest.json` file.

For more information, see *Edit Configuration File Parameters* in the *vSphere Virtual Machine Administration* documentation.

   d Power on the server and wait until the vSphere Bitfusion Plug-in is registered with vCenter Server.

5 By using the vSphere Bitfusion plug-in, restore the backup of your vSphere Bitfusion 3.0 cluster to your new cluster.

   For more information, see *Restore a Bitfusion Cluster* in the *VMware vSphere Bitfusion User Guide*.

6 For each subsequent vSphere Bitfusion server in your cluster, perform the following steps.

   a Deploy a new server VM by using the vSphere Bitfusion plug-in.

      During the deployment process, enter the hostname and host ID that are listed in the `manifest.json` for the corresponding vSphere Bitfusion 3.0 server. For more information, see Add Subsequent vSphere Bitfusion Servers.

   b In the settings of the new VM, add the same number of GPUs as the corresponding vSphere Bitfusion 3.0 server uses.

   c In the settings of the new VM, add a `guestinfo.bitfusion.server.host-id` configuration parameter. The parameter value must match the host ID of the corresponding server with version 3.0, file.

   d Power on the VM. You must power on multiple VMs in a sequential order.

7 Delete the vSphere Bitfusion 3.0 server VMs.