



# Virtuozzo Hybrid Infrastructure 3.5

## Installation Guide

July 21, 2020

Virtuozzo International GmbH

Vordergasse 59

8200 Schaffhausen

Switzerland

Tel: + 41 52 632 0411

Fax: + 41 52 672 2010

<https://virtuozzo.com>

Copyright ©2016-2020 Virtuozzo International GmbH. All rights reserved.

This product is protected by United States and international copyright laws. The product's underlying technology, patents, and trademarks are listed at <https://virtuozzo.com>.

Microsoft, Windows, Windows Server, Windows NT, Windows Vista, and MS-DOS are registered trademarks of Microsoft Corporation.

Apple, Mac, the Mac logo, Mac OS, iPad, iPhone, iPod touch, FaceTime HD camera and iSight are trademarks of Apple Inc., registered in the US and other countries.

Linux is a registered trademark of Linus Torvalds. All other marks and names mentioned herein may be trademarks of their respective owners.

# Contents

<b>1. Deployment Overview</b>	<b>1</b>
<b>2. Planning Infrastructure</b>	<b>2</b>
2.1 Storage Architecture Overview	2
2.1.1 Storage Role	3
2.1.2 Metadata Role	3
2.1.3 Supplementary Roles	3
2.2 Compute Architecture Overview	4
2.3 Planning Node Hardware Configurations	5
2.3.1 Hardware Limits	5
2.3.2 Hardware Requirements	5
2.3.3 Hardware Recommendations	7
2.3.3.1 Storage Cluster Composition Recommendations	8
2.3.3.2 General Hardware Recommendations	9
2.3.3.3 Storage Hardware Recommendations	9
2.3.3.4 Network Hardware Recommendations	11
2.3.4 Hardware and Software Limitations	12
2.3.5 Minimum Storage Configuration	13
2.3.6 Recommended Storage Configuration	15
2.3.6.1 HDD Only	15
2.3.6.2 HDD + System SSD (No Cache)	16
2.3.6.3 HDD + SSD	16
2.3.6.4 SSD Only	16
2.3.6.5 HDD + SSD (No Cache), 2 Tiers	17
2.3.6.6 HDD + SSD, 3 Tiers	18
2.3.7 Raw Disk Space Considerations	18
2.3.8 Checking Disk Data Flushing Capabilities	19

2.4	Planning Virtual Machine Configurations . . . . .	20
2.4.1	Running on VMware . . . . .	20
2.5	Planning Network . . . . .	21
2.5.1	General Network Requirements . . . . .	22
2.5.2	Network Limitations . . . . .	22
2.5.3	Per-Node Network Requirements and Recommendations . . . . .	22
2.5.4	Kubernetes Network Requirements . . . . .	24
2.5.5	Network Recommendations for Clients . . . . .	26
2.6	Understanding Data Redundancy . . . . .	26
2.6.1	Redundancy by Replication . . . . .	28
2.6.2	Redundancy by Erasure Coding . . . . .	29
2.6.3	No Redundancy . . . . .	30
2.7	Understanding Failure Domains . . . . .	30
2.8	Understanding Storage Tiers . . . . .	31
2.9	Understanding Cluster Rebuilding . . . . .	32
<b>3.</b>	<b>Installing Using GUI . . . . .</b>	<b>34</b>
3.1	Obtaining Distribution Image . . . . .	34
3.2	Preparing for Installation . . . . .	34
3.2.1	Preparing for Installation from USB Storage Drives . . . . .	35
3.3	Starting Installation . . . . .	35
3.4	Step 1: Accepting the User Agreement . . . . .	36
3.5	Step 2: Configuring the Network . . . . .	36
3.5.1	Creating Bonded Connections . . . . .	37
3.5.2	Creating VLAN Adapters . . . . .	39
3.6	Step 3: Choosing the Time Zone . . . . .	40
3.7	Step 4: Configuring the Storage Cluster . . . . .	41
3.7.1	Deploying the Primary Node . . . . .	42
3.7.2	Deploying Secondary Nodes . . . . .	42
3.8	Step 5: Selecting the System Partition . . . . .	43
3.9	Step 6: Setting the Root Password . . . . .	44
3.10	Finishing Installation . . . . .	44
<b>4.</b>	<b>Installing Using PXE . . . . .</b>	<b>46</b>
4.1	Preparing Environment . . . . .	46
4.1.1	Installing PXE Components . . . . .	46

4.1.2	Configuring TFTP Server . . . . .	47
4.1.3	Setting Up DHCP Server . . . . .	48
4.1.4	Setting Up HTTP Server . . . . .	48
4.2	Installing Over the Network . . . . .	49
4.3	Creating Kickstart File . . . . .	50
4.3.1	Kickstart Options . . . . .	50
4.3.2	Kickstart Scripts . . . . .	52
4.3.2.1	Installing Packages . . . . .	52
4.3.2.2	Installing Admin Panel and Storage . . . . .	52
4.3.2.3	Installing Storage Component Only . . . . .	53
4.3.3	Kickstart File Example . . . . .	53
4.3.3.1	Creating the System Partition on Software RAID1 . . . . .	55
4.4	Using Kickstart File . . . . .	56
<b>5.</b>	<b>Additional Installation Modes . . . . .</b>	<b>57</b>
5.1	Installing via VNC . . . . .	57
<b>6.</b>	<b>Troubleshooting Installation . . . . .</b>	<b>58</b>
6.1	Installing in Basic Graphics Mode . . . . .	58
6.2	Booting into Rescue Mode . . . . .	58

## CHAPTER 1

# Deployment Overview

To deploy Virtuozzo Hybrid Infrastructure for evaluation purposes or in production, you will need to do the following:

1. Plan the infrastructure.
2. Install and configure Virtuozzo Hybrid Infrastructure on required servers.
3. Create the storage cluster.
4. Create a compute cluster and/or set up data export services.

## CHAPTER 2

# Planning Infrastructure

To plan the infrastructure, you will need to decide on the hardware configuration of each server, plan networks, choose a redundancy method and mode, and decide which data will be kept on which storage tier.

Information in this chapter is meant to help you complete all of these tasks.

## 2.1 Storage Architecture Overview

The fundamental component of Virtuozzo Hybrid Infrastructure is a storage cluster: a group of physical servers interconnected by network. Each server in a cluster is assigned one or more roles and typically runs services that correspond to these roles:

- storage role: chunk service or CS
- metadata role: metadata service or MDS
- supplementary roles:
  - SSD cache,
  - system

Any server in the cluster can be assigned a combination of storage, metadata, and network roles. For example, a single server can be an S3 access point, an iSCSI access point, and a storage node at once.

Each cluster also requires that a web-based admin panel be installed on one (and only one) of the nodes. The panel enables administrators to manage the cluster.

## 2.1.1 Storage Role

Storage nodes run chunk services, store all data in the form of fixed-size chunks, and provide access to these chunks. All data chunks are replicated and the replicas are kept on different storage nodes to achieve high availability of data. If one of the storage nodes fails, the remaining healthy storage nodes continue providing the data chunks that were stored on the failed node.

The storage role can only be assigned to a server with disks of certain capacity.

## 2.1.2 Metadata Role

Metadata nodes run metadata services, store cluster metadata, and control how user files are split into chunks and where these chunks are located. Metadata nodes also ensure that chunks have the required amount of replicas. Finally, they log all important events that happen in the cluster.

To provide system reliability, Virtuozzo Hybrid Infrastructure uses the Paxos consensus algorithm. It guarantees fault-tolerance if the majority of nodes running metadata services are healthy.

To ensure high availability of metadata in a production environment, metadata services must be run on at least three cluster nodes. In this case, if one metadata service fails, the remaining two will still be controlling the cluster. However, it is recommended to have at least five metadata services to ensure that the cluster can survive simultaneous failure of two nodes without data loss.

## 2.1.3 Supplementary Roles

### SSD cache

Boosts chunk read/write performance by creating write caches on selected solid-state drives (SSDs). It is also recommended to use such SSDs for metadata, see *Metadata Role* (page 3). The use of write journals may more than double the write speed in the cluster.

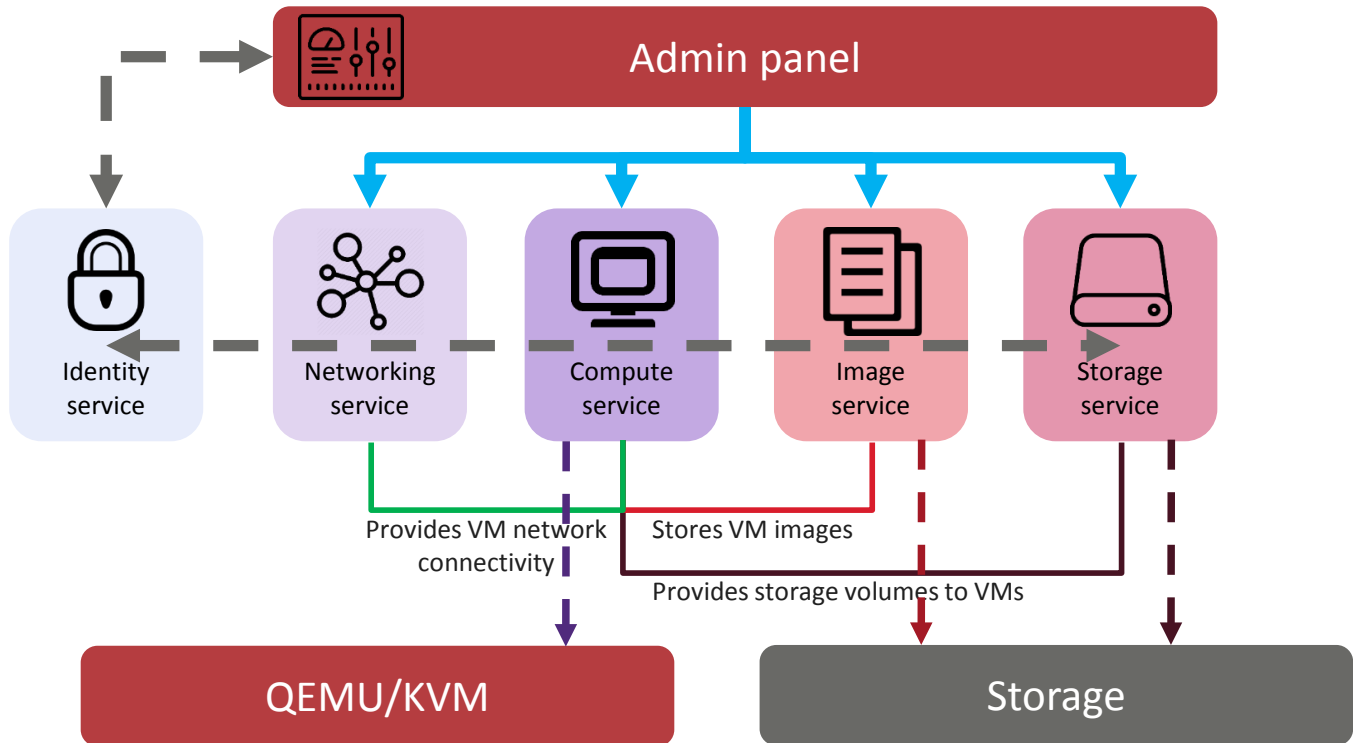
### System

One disk per node that is reserved for the operating system and unavailable for data storage.



## 2.2 Compute Architecture Overview

The following diagram shows the major compute components of Virtuozzo Hybrid Infrastructure.



- The storage service provides virtual disks to virtual machines. This service relies on the base storage cluster for data redundancy.
- The image service enables users to upload, store, and use images of supported guest operating systems and virtual disks. This service relies on the base storage cluster for data redundancy.
- The identity service provides authentication and authorization capabilities for Virtuozzo Hybrid Infrastructure.
- The compute service enables users to create, run, and manage virtual machines. This service relies on the custom QEMU/KVM hypervisor.
- The networking service provides private and public network capabilities for virtual machines.
- The admin panel service provides a convenient web user interface for managing the entire infrastructure.

## 2.3 Planning Node Hardware Configurations

Virtuozzo Hybrid Infrastructure works on top of commodity hardware, so you can create a cluster from regular servers, disks, and network cards. Still, to achieve the optimal performance, a number of requirements must be met and a number of recommendations should be followed.

---

**Note:** If you are unsure of what hardware to choose, consult your sales representative. You can also use the [online hardware calculator](#).

---

### 2.3.1 Hardware Limits

The following table lists the current hardware limits for Virtuozzo Hybrid Infrastructure servers:

Table 2.3.1.1: Server hardware limits

Hardware	Theoretical	Certified
RAM	64 TB	1 TB
CPU	5120 logical CPUs	384 logical CPUs

A logical CPU is a core (thread) in a multicore (multithreading) processor.

### 2.3.2 Hardware Requirements

The following table lists the minimum and recommended disk requirements according to the disk roles (refer to *Storage Architecture Overview* (page 2)):

Table 2.3.2.1: Disk requirements

Disk role	Quantity	Minimum	Recommended
System	One disk per node	100 GB SATA/SAS HDD	250 GB SATA/SAS SSD
Metadata	One disk per node Five disks recommended for one cluster	100 GB enterprise-grade SSD with power loss protection, 1 DWPD endurance minimum	

Continued on next page

Table 2.3.2.1 – continued from previous page

Disk role	Quantity	Minimum	Recommended
Cache	Optional One SSD disk per 4-12 HDDs	100+ GB enterprise-grade SSD with power loss protection and 75 MB/s sequential write performance per serviced HDD; 1 DWPD endurance minimum, 10 DWPD recommended	
Storage	Optional At least one per cluster	100 GB minimum, 16 TB maximum recommended SATA/SAS HDD or SATA/SAS/NVMe SSD (enterprise-grade with power loss protection, 1 DWPD endurance minimum)	

The following table lists the recommended amount of RAM and CPU cores for one node according to the services you will use:

Table 2.3.2.2: CPU and RAM requirements

Service		RAM	CPU cores*
System		6 GB	2 cores
Storage services: each disk with Storage role or Cache role (any size)**		1 GB	0.2 core
Compute		8 GB	3 cores
Load Balancer	Service	1 GB	1 core
	Each load balancer	1 GB	1 core
Kubernetes		2 GB	2 cores
S3		4.5 GB	3 cores
Backup Gateway***		1 GB	2 cores
NFS	Service	4 GB	2 cores
	Each share	0.5 GB	0.5 core
iSCSI	Service	1 GB	1 core
	Each volume	0.1 GB	0.5 core

\* 64-bit x86 AMD-V or Intel VT processors with hardware virtualization extensions enabled. For Intel processors, enable “unrestricted guest” and VT-x with Extended Page Tables in BIOS. It is recommended to have the same CPU models on each node to avoid VM live migration issues. A CPU core here is a physical core in a multicore processor (hyperthreading is not taken into account).

\*\* For clusters larger than 1 PB of physical space, please additionally add 0.5 GB of RAM per Metadata service.

\*\*\* When working with public clouds and NFS, Backup Gateway consumes as much RAM and CPU as with a local storage.

As for the networks, at least 2 x 10 GbE interfaces are recommended; 25 GbE, 40 GbE, and 100 GbE are even better. Bonding is recommended. However, you can start with 1 GbE links, but they can limit cluster throughput on modern loads.

Let us consider some examples and calculate the requirements for particular cases.

- If you have 1 node (1 system disk and 4 storage disks) and want to use it for Backup Gateway, the node should meet the following requirements: the system requirements (6 GB, 2 cores) + storage services for 4 disks (4 GB, 0.8 cores) + Backup Gateway (1 GB, 2 cores). All in all, that is 11 GB of RAM and 5 cores for the node.
- If you have 3 nodes (1 system disk and 4 storage disks) and want to use them for the Compute service, each cluster node should meet the following requirements: the system requirements (6 GB, 2 cores) + Compute (8 GB, 3 cores). All in all, that is 14 GB of RAM and 5 cores for each node. If you want to enable, for example, one Kubernetes VM and one load balancer, add the following requirements to the management node: Load Balancer (2 GB, 2 cores) + Kubernetes (2 GB, 2 cores). All in all, that is 18 GB of RAM and 9 cores for the management node.
- If you have 5 nodes (2 system disks and 10 storage disks) and want to use them for Backup Gateway, each cluster node should meet the following requirements: the system requirements (6 GB, 2 cores) + storage services for 10 disks (10 GB, 2 cores) + Backup Gateway (1 GB, 2 cores). All in all, that is 17 GB of RAM and 6 cores for each node.

In general, the more resources you provide for your cluster, the better it works. All extra RAM is used to cache disk reads. And extra CPU cores increase the performance and reduce latency.

### 2.3.3 Hardware Recommendations

In general, Virtuozzo Hybrid Infrastructure works on the same hardware that is recommended for Red Hat Enterprise Linux 7, including AMD EPYC processors: [servers](#), [components](#).

The following recommendations further explain the benefits added by specific hardware in the hardware requirements table. Use them to configure your cluster in an optimal way.

### 2.3.3.1 Storage Cluster Composition Recommendations

Designing an efficient storage cluster means finding a compromise between performance and cost that suits your purposes. When planning, keep in mind that a cluster with many nodes and few disks per node offers higher performance while a cluster with the minimum number of nodes (3) and a lot of disks per node is cheaper. See the following table for more details.

Table 2.3.3.1.1: Cluster composition recommendations

<b>Design considerations</b>	<b>Minimum nodes (3), many disks per node</b>	<b>Many nodes, few disks per node (all-flash configuration)</b>
Optimization	Lower cost.	Higher performance.
Free disk space to reserve	More space to reserve for cluster rebuilding as fewer healthy nodes will have to store the data from a failed node.	Less space to reserve for cluster rebuilding as more healthy nodes will have to store the data from a failed node.
Redundancy	Fewer erasure coding choices.	More erasure coding choices.
Cluster balance and rebuilding performance	Worse balance and slower rebuilding.	Better balance and faster rebuilding.
Network capacity	More network bandwidth required to maintain cluster performance during rebuilding.	Less network bandwidth required to maintain cluster performance during rebuilding.
Favorable data type	Cold data (e.g., backups).	Hot data (e.g., virtual environments).
Sample server configuration	Supermicro SSG-6047R-E1R36L (Intel Xeon E5-2620 v1/v2 CPU, 32GB RAM, 36 x 12TB HDDs, a 500GB system disk).	Supermicro SYS-2028TP-HC0R-SIOM (4 x Intel E5-2620 v4 CPUs, 4 x 16GB RAM, 24 x 1.9TB Samsung PM1643 SSDs).

Take note of the following:

1. These considerations only apply if failure domain is host.
2. The speed of rebuilding in the replication mode does not depend on the number of nodes in the cluster.
3. Virtuozzo Hybrid Infrastructure supports hundreds of disks per node. If you plan to use more than 36 disks per node, contact sales engineers who will help you design a more efficient cluster.

### 2.3.3.2 General Hardware Recommendations

- At least five nodes are required for a production environment. This is to ensure that the cluster can survive failure of two nodes without data loss.
- One of the strongest features of Virtuozzo Hybrid Infrastructure is scalability. The bigger the cluster, the better Virtuozzo Hybrid Infrastructure performs. It is recommended to create production clusters from at least ten nodes for improved resilience, performance, and fault tolerance in production scenarios.
- Even though a cluster can be created on top of varied hardware, using nodes with similar hardware in each node will yield better cluster performance, capacity, and overall balance.
- Any cluster infrastructure must be tested extensively before it is deployed to production. Such common points of failure as SSD drives and network adapter bonds must always be thoroughly verified.
- It is not recommended for production to run Virtuozzo Hybrid Infrastructure on top of SAN/NAS hardware that has its own redundancy mechanisms. Doing so may negatively affect performance and data availability.
- To achieve best performance, keep at least 20% of cluster capacity free.
- During disaster recovery, Virtuozzo Hybrid Infrastructure may need additional disk space for replication. Make sure to reserve at least as much space as any single storage node has.
- It is recommended to have the same CPU models on each node to avoid VM live migration issues. For more details, see the [Administrator's Command Line Guide](#).
- If you plan to use Backup Gateway to store backups in the cloud, make sure the local storage cluster has plenty of logical space for staging (keeping backups locally before sending them to the cloud). For example, if you perform backups daily, provide enough space for at least 1.5 days' worth of backups. For more details, see the [Administrator's Guide](#).
- It is recommended to use UEFI instead of BIOS if supported by hardware. In particular, if you use NVMe drives.

### 2.3.3.3 Storage Hardware Recommendations

- It is possible to use disks of different size in the same cluster. However, keep in mind that, given the same IOPS, smaller disks will offer higher performance per terabyte of data compared to bigger disks. It is recommended to group disks with the same IOPS per terabyte in the same tier.

- Using the recommended SSD models may help you avoid loss of data. Not all SSD drives can withstand enterprise workloads and may break down in the first months of operation, resulting in TCO spikes.
  - SSD memory cells can withstand a limited number of rewrites. An SSD drive should be viewed as a consumable that you will need to replace after a certain time. Consumer-grade SSD drives can withstand a very low number of rewrites (so low, in fact, that these numbers are not shown in their technical specifications). SSD drives intended for storage clusters must offer at least 1 DWPD endurance (10 DWPD is recommended). The higher the endurance, the less often SSDs will need to be replaced, improving TCO.
  - Many consumer-grade SSD drives can ignore disk flushes and falsely report to operating systems that data was written while it, in fact, was not. Examples of such drives include OCZ Vertex 3, Intel 520, Intel X25-E, and Intel X-25-M G2. These drives are known to be unsafe in terms of data commits, they should not be used with databases, and they may easily corrupt the file system in case of a power failure. For these reasons, use enterprise-grade SSD drives that obey the flush rules (for more information, see <http://www.postgresql.org/docs/current/static/wal-reliability.html>). Enterprise-grade SSD drives that operate correctly usually have the power loss protection property in their technical specification. Some of the market names for this technology are Enhanced Power Loss Data Protection (Intel), Cache Power Protection (Samsung), Power-Failure Support (Kingston), Complete Power Fail Protection (OCZ).
  - It is highly recommended to check the data flushing capabilities of your disks as explained in *Checking Disk Data Flushing Capabilities* (page 19).
  - Consumer-grade SSD drives usually have unstable performance and are not suited to withstand sustainable enterprise workloads. For this reason, pay attention to sustainable load tests when choosing SSDs. We recommend the following enterprise-grade SSD drives which are the best in terms of performance, endurance, and investments: Intel S3710, Intel P3700, Huawei ES3000 V2, Samsung SM1635, and Sandisk Lightning.
  - Performance of SSD disks may depend on their size. Lower-capacity drives (100 to 400 GB) may perform much slower (sometimes up to ten times slower) than higher-capacity ones (1.9 to 3.8 TB). Consult drive performance and endurance specifications before purchasing hardware.
- Using NVMe or SAS SSDs for write caching improves random I/O performance and is highly recommended for all workloads with heavy random access (e.g., iSCSI volumes). In turn, SATA disks are best suited for SSD-only configurations but not write caching.

- Using shingled magnetic recording (SMR) HDDs is strongly not recommended, even for backup scenarios. Such disks have unpredictable latency that may lead to unexpected temporary service outages and sudden performance degradations.
- Running metadata services on SSDs improves cluster performance. To also minimize CAPEX, the same SSDs can be used for write caching.
- If capacity is the main goal and you need to store non-frequently accessed data, choose SATA disks over SAS ones. If performance is the main goal, choose NVMe or SAS disks over SATA ones.
- The more disks per node the lower the CAPEX. As an example, a cluster created from ten nodes with two disks in each will be less expensive than a cluster created from twenty nodes with one disk in each.
- Using SATA HDDs with one SSD for caching is more cost effective than using only SAS HDDs without such an SSD.
- Create hardware or software RAID1 volumes for system disks using RAID or HBA controllers, respectively, to ensure its high performance and availability.
- Use HBA controllers as they are less expensive and easier to manage than RAID controllers.
- Disable all RAID controller caches for SSD drives. Modern SSDs have good performance that can be reduced by a RAID controller's write and read cache. It is recommend to disable caching for SSD drives and leave it enabled only for HDD drives.
- If you use RAID controllers, do not create RAID volumes from HDDs intended for storage. Each storage HDD needs to be recognized by Virtuozzo Hybrid Infrastructure as a separate device.
- If you use RAID controllers with caching, equip them with backup battery units (BBUs) to protect against cache loss during power outages.
- Disk block size (e.g., 512b or 4K) is not important and has no effect on performance.

#### 2.3.3.4 Network Hardware Recommendations

- Use separate networks (and, ideally albeit optionally, separate network adapters) for internal and public traffic. Doing so will prevent public traffic from affecting cluster I/O performance and also prevent possible denial-of-service attacks from the outside.
- Network latency dramatically reduces cluster performance. Use quality network equipment with low latency links. Do not use consumer-grade network switches.



- Do not use desktop network adapters like Intel EXPI9301CTBLK or Realtek 8129 as they are not designed for heavy load and may not support full-duplex links. Also use non-blocking Ethernet switches.
- To avoid intrusions, Virtuozzo Hybrid Infrastructure should be on a dedicated internal network inaccessible from outside.
- Use one 1 Gbit/s link per each two HDDs on the node (rounded up). For one or two HDDs on a node, two bonded network interfaces are still recommended for high network availability. The reason for this recommendation is that 1 Gbit/s Ethernet networks can deliver 110-120 MB/s of throughput, which is close to sequential I/O performance of a single disk. Since several disks on a server can deliver higher throughput than a single 1 Gbit/s Ethernet link, networking may become a bottleneck.
- For maximum sequential I/O performance, use one 1Gbit/s link per each hard drive, or one 10Gbit/s link per node. Even though I/O operations are most often random in real-life scenarios, sequential I/O is important in backup scenarios.
- For maximum overall performance, use one 10 Gbit/s link per node (or two bonded for high network availability).
- It is not recommended to configure 1 Gbit/s network adapters to use non-default MTUs (e.g., 9000-byte jumbo frames). Such settings require additional configuration of switches and often lead to human error. 10+ Gbit/s network adapters, on the other hand, need to be configured to use jumbo frames to achieve full performance.
- The currently supported Fibre Channel host bus adapters (HBAs) are QLogic QLE2562-CK and QLogic ISP2532.
- It is recommended to use Mellanox ConnectX-4 and ConnectX-5 InfiniBand adapters. Mellanox ConnectX-2 and ConnectX-3 cards are not supported.
- Adapters using the BN2X driver, such as Broadcom Limited BCM57840 NetXtreme II 10/20-Gigabit Ethernet / HPE FlexFabric 10Gb 2-port 536FLB Adapter, are not recommended. They limit MTU to 3616, which affects the cluster performance.

### 2.3.4 Hardware and Software Limitations

Hardware limitations:

- Each management node must have at least two disks (one system+metadata, one storage).
- Each compute or storage node must have at least three disks (one system, one metadata, one storage).

- Three servers are required to test all the features of the product.
- The system disk must have at least 100 GBs of space.
- Admin panel requires a Full HD monitor to be displayed correctly.
- The maximum supported physical partition size is 254 TiB.

Software limitations:

- One node can be a part of only one cluster.
- Only one S3 cluster can be created on top of a storage cluster.
- Only predefined redundancy modes are available in the admin panel.
- Thin provisioning is always enabled for all data and cannot be configured otherwise.
- Admin panel has been tested to work at resolutions 1280x720 and higher in the following web browsers: latest Firefox, Chrome, Safari.

For network limitations, see [Network Limitations](#) (page 22).

### 2.3.5 Minimum Storage Configuration

The minimum configuration described in the table will let you evaluate the features of the storage cluster. It is not meant for production.

Table 2.3.5.1: Minimum cluster configuration

Node #	1st disk role	2nd disk role	3rd+ disk roles	Access points
1	System	Metadata	Storage	iSCSI, S3 private, S3 public, NFS, Backup Gateway
2	System	Metadata	Storage	iSCSI, S3 private, S3 public, NFS, Backup Gateway
3	System	Metadata	Storage	iSCSI, S3 private, S3 public, NFS, Backup Gateway
3 nodes in total		3 MDSs in total	3+ CSs in total	Access point services run on three nodes in total.

**Note:** SSD disks can be assigned **System**, **Metadata**, and **Cache** roles at the same time, freeing up more disks for the storage role.

---

Even though three nodes are recommended even for the minimum configuration, you can start evaluating Virtuozzo Hybrid Infrastructure with just one node and add more nodes later. At the very least, a storage cluster must have one metadata service and one chunk service running. A single-node installation will let you evaluate services such as iSCSI, Backup Gateway, etc. However, such a configuration will have two key limitations:

1. Just one MDS will be a single point of failure. If it fails, the entire cluster will stop working.
2. Just one CS will be able to store just one chunk replica. If it fails, the data will be lost.

---

**Important:** If you deploy Virtuozzo Hybrid Infrastructure on a single node, you must take care of making its storage persistent and redundant to avoid data loss. If the node is physical, it must have multiple disks so you can replicate the data among them. If the node is a virtual machine, make sure that this VM is made highly available by the solution it runs on.

---

**Note:** Backup Gateway works with the local object storage in the staging mode. It means that the data to be replicated, migrated, or uploaded to a public cloud is first stored locally and only then sent to the destination. It is vital that the local object storage is persistent and redundant so the local data does not get lost. There are multiple ways to ensure the persistence and redundancy of the local storage. You can deploy your Backup Gateway on multiple nodes and select a good redundancy mode. If your gateway is deployed on a single node in Virtuozzo Hybrid Infrastructure, you can make its storage redundant by replicating it among multiple local disks. If your entire Virtuozzo Hybrid Infrastructure installation is deployed in a single virtual machine with the sole purpose of creating a gateway, make sure this VM is made highly available by the solution it runs on.

---

## 2.3.6 Recommended Storage Configuration

It is recommended to have at least five metadata services to ensure that the cluster can survive simultaneous failure of two nodes without data loss. The following configuration will help you create clusters for production environments:

Table 2.3.6.1: Recommended cluster configuration

Node #	1st disk role	2nd disk role	3rd+ disk roles	Access points
Nodes 1 to 5	System	SSD; metadata, cache	Storage	iSCSI, S3 private, S3 public, Backup Gateway
Nodes 6+	System	SSD; cache	Storage	iSCSI, S3 private, Backup Gateway
5+ nodes in total		5 MDSs in total	5+ CSs in total	All nodes run required access points.

A production-ready cluster can be created from just five nodes with recommended hardware. However, it is recommended to enter production with at least ten nodes if you are aiming to achieve significant performance advantages over direct-attached storage (DAS) or improved recovery times.

Following are a number of more specific configuration examples that can be used in production. Each configuration can be extended by adding chunk servers and nodes.

### 2.3.6.1 HDD Only

This basic configuration requires a dedicated disk for each metadata server.

Table 2.3.6.1.1: HDD only configuration

Nodes 1-5 (base)			Nodes 6+ (extension)		
Disk #	Disk type	Disk roles	Disk #	Disk type	Disk roles
1	HDD	System	1	HDD	System
2	HDD	MDS	2	HDD	CS
3	HDD	CS	3	HDD	CS
...	...	...	...	...	...
N	HDD	CS	N	HDD	CS

### 2.3.6.2 HDD + System SSD (No Cache)

This configuration is good for creating capacity-oriented clusters.

Table 2.3.6.2.1: HDD + system SSD (no cache) configuration

Nodes 1-5 (base)			Nodes 6+ (extension)		
Disk #	Disk type	Disk roles	Disk #	Disk type	Disk roles
1	SSD	System, MDS	1	SSD	System
2	HDD	CS	2	HDD	CS
3	HDD	CS	3	HDD	CS
...	...	...	...	...	...
N	HDD	CS	N	HDD	CS

### 2.3.6.3 HDD + SSD

This configuration is good for creating performance-oriented clusters.

Table 2.3.6.3.1: HDD + SSD configuration

Nodes 1-5 (base)			Nodes 6+ (extension)		
Disk #	Disk type	Disk roles	Disk #	Disk type	Disk roles
1	HDD	System	1	HDD	System
2	SSD	MDS, cache	2	SSD	Cache
3	HDD	CS	3	HDD	CS
...	...	...	...	...	...
N	HDD	CS	N	HDD	CS

### 2.3.6.4 SSD Only

This configuration does not require SSDs for cache.

When choosing hardware for this configuration, have in mind the following:

- Each Virtuozzo Hybrid Infrastructure client will be able to obtain up to about 40K sustainable IOPS (read + write) from the cluster.

- If you use the erasure coding redundancy scheme, each erasure coding file, e.g., a single VM HDD disk, will get up to 2K sustainable IOPS. That is, a user working inside a VM will have up to 2K sustainable IOPS per virtual HDD at their disposal. Multiple VMs on a node can utilize more IOPS, up to the client's limit.
- In this configuration, network latency defines more than half of overall performance, so make sure that the network latency is minimal. One recommendation is to have one 10Gbps switch between any two nodes in the cluster.

Table 2.3.6.4.1: SSD only configuration

Nodes 1-5 (base)			Nodes 6+ (extension)		
Disk #	Disk type	Disk roles	Disk #	Disk type	Disk roles
1	SSD	System, MDS	1	SSD	System
2	SSD	CS	2	SSD	CS
3	SSD	CS	3	SSD	CS
...	...	...	...	...	...
N	SSD	CS	N	SSD	CS

### 2.3.6.5 HDD + SSD (No Cache), 2 Tiers

In this configuration example, tier 1 is for HDDs without cache and tier 2 is for SSDs. Tier 1 can store cold data (e.g., backups), tier 2 can store hot data (e.g., high-performance virtual machines).

Table 2.3.6.5.1: HDD + SSD (no cache) 2-tier configuration

Nodes 1-5 (base)				Nodes 6+ (extension)			
Disk #	Disk type	Disk roles	Tier	Disk #	Disk type	Disk roles	Tier
1	SSD	System, MDS		1	SSD	System	
2	SSD	CS	2	2	SSD	CS	2
3	HDD	CS	1	3	HDD	CS	1
...	...	...	...	...	...	...	...
N	HDD/SSD	CS	1/2	N	HDD/SSD	CS	1/2

### 2.3.6.6 HDD + SSD, 3 Tiers

In this configuration example, tier 1 is for HDDs without cache, tier 2 is for HDDs with cache, and tier 3 is for SSDs. Tier 1 can store cold data (e.g., backups), tier 2 can store regular virtual machines, and tier 3 can store high-performance virtual machines.

Table 2.3.6.6.1: HDD + SSD 3-tier configuration

Nodes 1-5 (base)				Nodes 6+ (extension)			
Disk #	Disk type	Disk roles	Tier	Disk #	Disk type	Disk roles	Tier
1	HDD/SSD	System		1	HDD/SSD	System	
2	SSD	MDS, T2 cache		2	SSD	T2 cache	
3	HDD	CS	1	3	HDD	CS	1
4	HDD	CS	2	4	HDD	CS	2
5	SSD	CS	3	5	SSD	CS	3
...	...	...	...	...	...	...	...
N	HDD/SSD	CS	1/2/3	N	HDD/SSD	CS	1/2/3

### 2.3.7 Raw Disk Space Considerations

When planning the infrastructure, keep in mind the following to avoid confusion:

- The capacity of HDD and SSD is measured and specified with decimal, not binary prefixes, so “TB” in disk specifications usually means “terabyte”. The operating system, however, displays drive capacity using binary prefixes meaning that “TB” is “tebibyte” which is a noticeably larger number. As a result, disks may show capacity smaller than the one marketed by the vendor. For example, a disk with 6TB in specifications may be shown to have 5.45 TB of actual disk space in Virtuozzo Hybrid Infrastructure.
- 5% of disk space is reserved for emergency needs.

Therefore, if you add a 6TB disk to a cluster, the available physical space should increase by about 5.2 TB.

## 2.3.8 Checking Disk Data Flushing Capabilities

It is highly recommended to make sure that all storage devices you plan to include in your cluster can flush data from cache to disk if power goes out unexpectedly. Thus you will find devices that may lose data in a power failure.

Virtuozzo Hybrid Infrastructure ships with the `vstorage-hwflush-check` tool that checks how a storage device flushes data to disk in emergencies. The tool is implemented as a client/server utility:

- The client continuously writes blocks of data to the storage device. When a data block is written, the client increases a special counter and sends it to the server that keeps it.
- The server keeps track of counters incoming from the client and always knows the next counter number. If the server receives a counter smaller than the one it has (e.g., because the power has failed and the storage device has not flushed the cached data to disk), the server reports an error.

To check that a storage device can successfully flush data to disk when power fails, follow the procedure below:

1. On one node, run the server:

```
# vstorage-hwflush-check -l
```

2. On a different node that hosts the storage device you want to test, run the client, for example:

```
# vstorage-hwflush-check -s vstorage1.example.com -d /vstorage/stor1-ssd/test -t 50
```

where

- `vstorage1.example.com` is the host name of the server.
  - `/vstorage/stor1-ssd/test` is the directory to use for data flushing tests. During execution, the client creates a file in this directory and writes data blocks to it.
  - `50` is the number of threads for the client to write data to disk. Each thread has its own file and counter. You can increase the number of threads (max. 200) to test your system in more stressful conditions. You can also specify other options when running the client. For more information on available options, see the `vstorage-hwflush-check` man page.
3. Wait for at least 10-15 seconds, cut power from the client node (either press the **Power** button or pull the power cord out) and then power it on again.
  4. Restart the client:



```
# vstorage-hwflush-check -s vstorage1.example.com -d /vstorage/stor1-ssd/test -t 50
```

Once launched, the client will read all previously written data, determine the version of data on the disk, and restart the test from the last valid counter. It then will send this valid counter to the server and the server will compare it to the latest counter it has. You may see output like:

```
id<N>:<counter_on_disk> -> <counter_on_server>
```

which means one of the following:

- If the counter on the disk is lower than the counter on the server, the storage device has failed to flush the data to the disk. Avoid using this storage device in production, especially for CS or journals, as you risk losing data.
- If the counter on the disk is higher than the counter on the server, the storage device has flushed the data to the disk but the client has failed to report it to the server. The network may be too slow or the storage device may be too fast for the set number of load threads so consider increasing it. This storage device can be used in production.
- If both counters are equal, the storage device has flushed the data to the disk and the client has reported it to the server. This storage device can be used in production.

To be on the safe side, repeat the procedure several times. Once you have checked your first storage device, continue with all the remaining devices you plan to use in the cluster. You need to test all devices you plan to use in the cluster: SSD disks used for CS journaling, disks used for MDS journals and chunk servers.

## 2.4 Planning Virtual Machine Configurations

Even though Virtuozzo Hybrid Infrastructure performs best on bare metal, it can also run inside virtual machines. However, in this case only storage services will be available and you will not be able to create the compute cluster.

### 2.4.1 Running on VMware

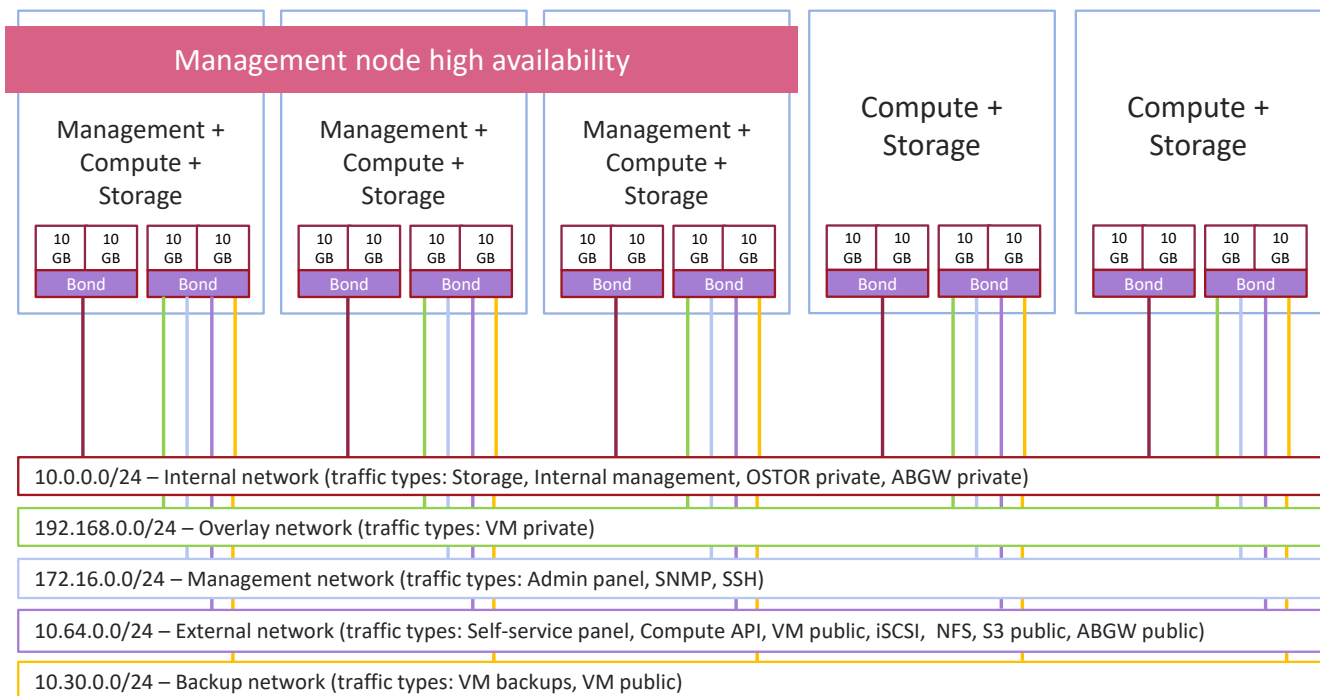
To be able to run the storage services on VMware, make sure the following requirements are met:

- Minimum ESXi version: ESXi 6.7.0 (build 8169922)
- VM version: 14

## 2.5 Planning Network

The recommended network configuration for Virtuozzo Hybrid Infrastructure is as follows:

- One bonded connection for internal management and storage traffic;
- One bonded connection with four VLANs over it:
  - for overlay network traffic between VMs
  - for management via the admin panel, SSH, and SNMP
  - for external access to and from the self-service panel, compute API, and VMs, as well as for public export of iSCSI, NFS, S3, and Backup Gateway data
  - for pulling VM backups by third-party backup management systems



## 2.5.1 General Network Requirements

- Internal storage traffic must be separated from other traffic types.

## 2.5.2 Network Limitations

- Nodes are added to clusters by their IP addresses, not FQDNs. Changing the IP address of a node in the cluster will remove that node from the cluster. If you plan to use DHCP in a cluster, make sure that IP addresses are bound to the MAC addresses of nodes' network interfaces.
- Each node must have Internet access so updates can be installed.
- The MTU value is set to 1500 by default. See *Step 2: Configuring the Network* (page 36) for information on setting an optimal MTU value.
- Network time synchronization (NTP) is required for correct statistics. It is enabled by default using the `chronyd` service. If you want to use `ntpd` or `ntpdate`, stop and disable `chronyd` first.
- The **Internal management** traffic type is assigned automatically during installation and cannot be changed in the admin panel later.
- Even though the management node can be accessed from a web browser by the hostname, you still need to specify its IP address, not the hostname, during installation.

## 2.5.3 Per-Node Network Requirements and Recommendations

- All network interfaces on a node must be connected to different subnets. A network interface can be a VLAN-tagged logical interface, an untagged bond, or an Ethernet link.
- Even though cluster nodes have the necessary `iptables` rules configured, we recommend to use an external firewall for untrusted public networks, such as the Internet.
- Ports that will be opened on cluster nodes depend on services that will run on the node and traffic types associated with them. Before enabling a specific service on a cluster node, you need to assign the respective traffic type to a network this node is connected to. Assigning a traffic type to a network configures a firewall on nodes connected to this network, opens specific ports on node network interfaces, and sets the necessary `iptables` rules.

The table below lists all the required ports and services associated with them:

Table 2.5.3.1: Open ports on cluster nodes

Service	Traffic type	Port	Description
Web control panel	Admin panel*	TCP 8888	External access to the admin panel.
	Self-service panel	TCP 8800	External access to the self-service panel.
Management	Internal management	any available port	Internal cluster management and transfers of node monitoring data to the admin panel.
Metadata service	Storage	any available port	Internal communication between MDS services, as well as with chunks services and clients.
Chunk service		any available port	Internal communication with MDS services and clients.
Client		any available port	Internal communication with MDS and chunk services.
Backup Gateway	ABGW public	TCP 44445	External data exchange with Acronis Backup agents and Acronis Backup Cloud.
	ABGW private	any available port	Internal management of and data exchange between multiple Backup Gateway services.
iSCSI	iSCSI	TCP 3260	External data exchange with the iSCSI access point.
S3	S3 public	TCP 80, 443	External data exchange with the S3 access point.
	OSTOR private	any available port	Internal data exchange between multiple S3 services.
NFS	NFS	TCP/UDP 111, 892, 2049	External data exchange with the NFS access point.
	OSTOR private	any available port	Internal data exchange between multiple NFS services.
Compute	Compute API*		External access to standard OpenStack API endpoints:
		TCP 5000	Identity API v3
		TCP 6080	noVNC Websocket Proxy
		TCP 8004	Orchestration Service API v1
		TCP 8041	Gnocchi API (billing metering service)
		TCP 8774	Compute API
		TCP 8776	Block Storage API v3

Continued on next page

Table 2.5.3.1 – continued from previous page

Service	Traffic type	Port	Description
		TCP 8780	Placement API
		TCP 9292	Image Service API v2
		TCP 9313	Key Manager API v1
		TCP 9513	Container Infrastructure Management API (Kubernetes service)
		TCP 9696	Networking API v2
		TCP 9888	Octavia API v2 (load balancer service)
	VM private	UDP 4789	Network traffic between VMs in private virtual networks.
		TCP 5900-5999	VNC console traffic.
	VM backups	TCP 49300-65535	External access to NBD endpoints.
SSH	SSH	TCP 22	Remote access to nodes via SSH.
SNMP	SNMP*	UDP 161	External access to storage cluster monitoring statistics via the SNMP protocol.

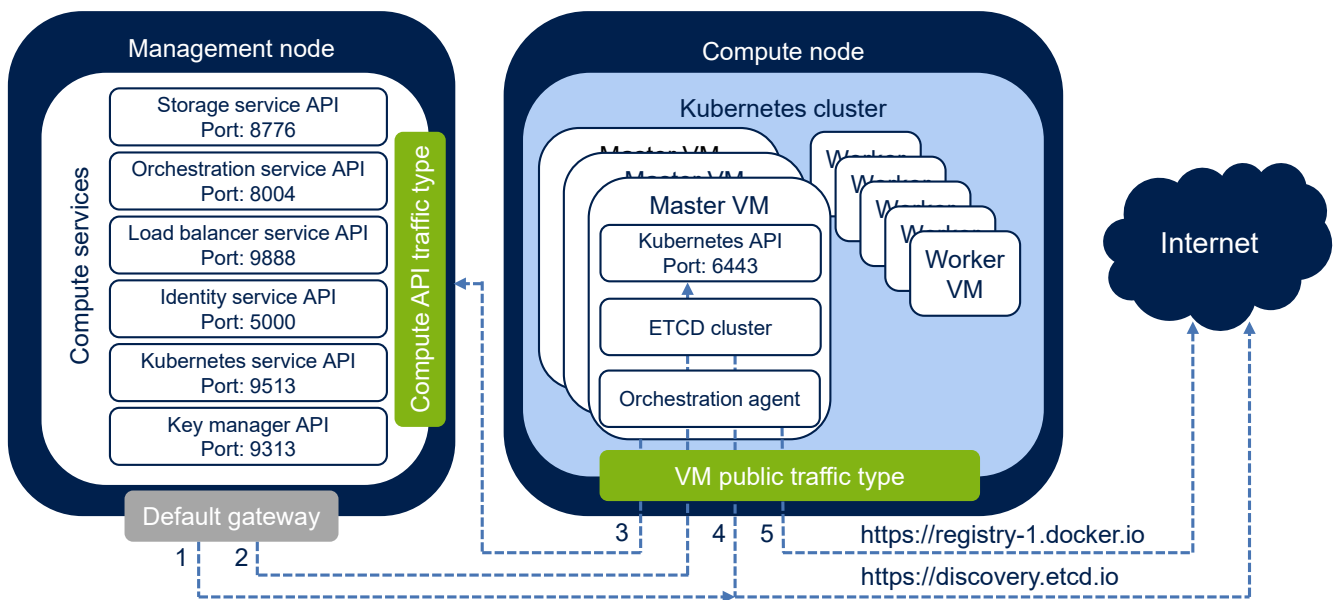
\* Ports for these traffic types must only be open on management nodes.

## 2.5.4 Kubernetes Network Requirements

To be able to deploy Kubernetes clusters in the compute cluster and work with them, make sure your network configuration allows the compute and Kubernetes services to send the following network requests:

1. The request to bootstrap the etcd cluster in the public discovery service—from all management nodes to `https://discovery.etcd.io` via the public network.
2. The request to obtain the “kubecfg” file—from all management nodes via the public network:
  - If HA for the master VM is enabled, the request is sent to the public or floating IP address of the load balancer VM associated with Kubernetes API on port 6443.
  - If HA for the master VM is disabled, the request is sent to the public or floating IP address of the Kubernetes master VM on port 6443.

3. Requests from Kubernetes master VMs to the compute APIs (the **Compute API** traffic type) via the network with the **VM public** traffic type (via a publicly available VM network interface or a virtual router with enabled SNAT). By default, the compute API is exposed via the IP address of the management node (or to its virtual IP address if high availability is enabled). But you can also access the compute API via a DNS name (refer to [Setting a DNS Name for the Compute API](#)).
4. The request to update etcd cluster member state in the public discovery service—from Kubernetes master VMs to <https://discovery.etcd.io> via the network with the **VM public** traffic type (via a publicly available VM network interface or a virtual router with enabled SNAT).
5. The request to download container images from the public Docker Hub repository—from Kubernetes master VMs to <https://registry-1.docker.io> via the network with the **VM public** traffic type (via a publicly available VM network interface or a virtual router with enabled SNAT).



It is also required that the network where you create a Kubernetes cluster does not overlap with these default networks:

- 10.100.0.0/24—used for pod-level networking
- 10.254.0.0/16—used for allocating Kubernetes cluster IP addresses from

## 2.5.5 Network Recommendations for Clients

The following table lists the maximum network performance a client can get with the specified network interface. The recommendation for clients is to use 10Gbps network hardware between any two cluster nodes and minimize network latencies, especially if SSD disks are used.

Table 2.5.5.1: Maximum client network performance

<b>Storage network interface</b>	<b>Node max. I/O</b>	<b>VM max. I/O (replication)</b>	<b>VM max. I/O (erasure coding)</b>
1 Gbps	100 MB/s	100 MB/s	70 MB/s
2 x 1 Gbps	~175 MB/s	100 MB/s	~130 MB/s
3 x 1 Gbps	~250 MB/s	100 MB/s	~180 MB/s
10 Gbps	1 GB/s	1 GB/s	700 MB/s
2 x 10 Gbps	1.75 GB/s	1 GB/s	1.3 GB/s

## 2.6 Understanding Data Redundancy

Virtuozzo Hybrid Infrastructure protects every piece of data by making it redundant. It means that copies of each piece of data are stored across different storage nodes to ensure that the data is available even if some of the storage nodes are inaccessible.

Virtuozzo Hybrid Infrastructure automatically maintains a required number of copies within the cluster and ensures that all the copies are up-to-date. If a storage node becomes inaccessible, copies from it are replaced by new ones that are distributed among healthy storage nodes. If a storage node becomes accessible again after downtime, out-of-date copies on it are updated.

The redundancy is achieved by one of two methods: replication or erasure coding (explained in more detail in the next section). The chosen method affects the size of one piece of data and the number of its copies that will be maintained in the cluster. In general, replication offers better performance while erasure coding leaves more storage space available for data (see table).

Virtuozzo Hybrid Infrastructure supports a number of modes for each redundancy method. The following table illustrates data overhead of various redundancy modes. The first three lines are replication and the rest are erasure coding.

Table 2.6.1: Redundancy mode comparison

Redundancy mode	Min. number of nodes required	How many nodes can fail without data loss	Storage overhead, %	Raw space needed to store 100GB of data
1 replica (no redundancy)	1	0	0	100GB
2 replicas	2	1	100	200GB
3 replicas	3	2	200	300GB
Encoding 1+0 (no redundancy)	1	0	0	100GB
Encoding 1+1	2	1	100	200GB
Encoding 1+2	3	2	200	300GB
Encoding 3+1	4	1	33	133GB
Encoding 3+2	5	2	67	167GB
Encoding 5+2	7	2	40	140GB
Encoding 7+2	9	2	29	129GB
Encoding 17+3	20	3	18	118GB

---

**Note:** The 1+0, 1+1, 1+2, and 3+1 encoding modes are meant for small clusters that have insufficient nodes for other erasure coding modes but will grow in the future. As redundancy type cannot be changed once chosen (from replication to erasure coding or vice versa), this mode allows one to choose erasure coding even if their cluster is smaller than recommended. Once the cluster has grown, more beneficial redundancy modes can be chosen.

---

You choose a data redundancy mode when configuring storage services and creating storage volumes for virtual machines. No matter what redundancy mode you choose, it is highly recommended to be protected against a simultaneous failure of two nodes as that happens often in real-life scenarios.

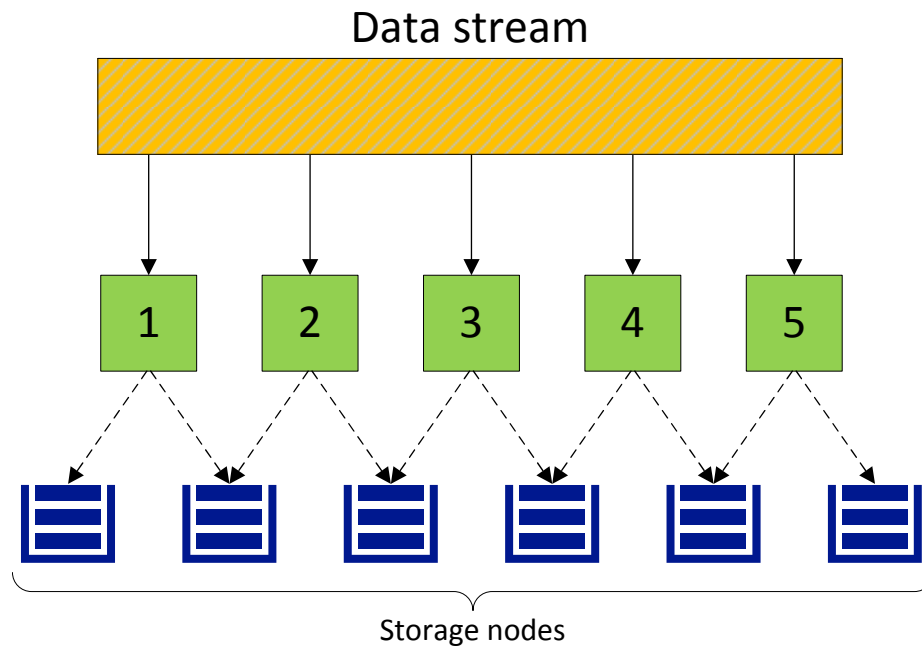
All redundancy modes allow write operations when one storage node is inaccessible. If two storage nodes are inaccessible, write operations may be frozen until the cluster heals itself.



## 2.6.1 Redundancy by Replication

With replication, Virtuozzo Hybrid Infrastructure breaks the incoming data stream into 256MB chunks. Each chunk is replicated and replicas are stored on different storage nodes, so that each node has only one replica of a given chunk.

The following diagram illustrates the 2 replicas redundancy mode.



Replication in Virtuozzo Hybrid Infrastructure is similar to the RAID rebuild process but has two key differences:

- Replication in Virtuozzo Hybrid Infrastructure is much faster than that of a typical online RAID 1/5/10 rebuild. The reason is that Virtuozzo Hybrid Infrastructure replicates chunks in parallel, to multiple storage nodes.
- The more storage nodes are in a cluster, the faster the cluster will recover from a disk or node failure.

High replication performance minimizes the periods of reduced redundancy for the cluster. Replication performance is affected by:

- The number of available storage nodes. As replication runs in parallel, the more available replication sources and destinations there are, the faster it is.
- Performance of storage node disks.

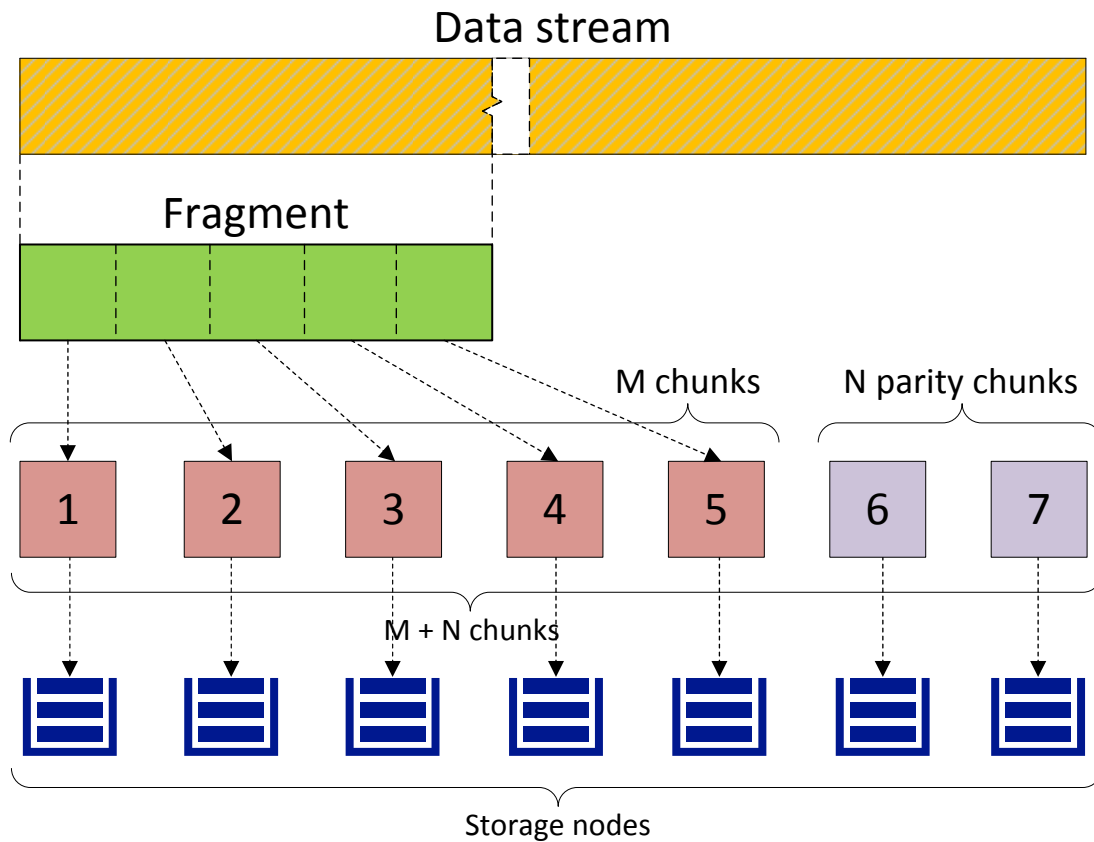
- Network performance. All replicas are transferred between storage nodes over network. For example, 1 Gbps throughput can be a bottleneck (see *Per-Node Network Requirements and Recommendations* (page 22)).
- Distribution of data in the cluster. Some storage nodes may have much more data to replicate than others and may become overloaded during replication.
- I/O activity in the cluster during replication.

## 2.6.2 Redundancy by Erasure Coding

With erasure coding, Virtuozzo Hybrid Infrastructure breaks the incoming data stream into fragments of certain size, then splits each fragment into a certain number (M) of 1-megabyte pieces and creates a certain number (N) of parity pieces for redundancy. All pieces are distributed among M+N storage nodes, that is, one piece per node. On storage nodes, pieces are stored in regular chunks of 256MB but such chunks are not replicated as redundancy is already achieved. The cluster can survive failure of any N storage nodes without data loss.

The values of M and N are indicated in the names of erasure coding redundancy modes. For example, in the 5+2 mode, the incoming data is broken into 5MB fragments, each fragment is split into five 1MB pieces and two more 1MB parity pieces are added for redundancy. In addition, if N is 2, the data is encoded using the RAID6 scheme, and if N is greater than 2, erasure codes are used.

The diagram below illustrates the 5+2 mode.



### 2.6.3 No Redundancy

**Warning:** Danger of data loss!

Without redundancy, singular chunks are stored on storage nodes, one per node. If the node fails, the data may be lost. Having no redundancy is highly not recommended no matter the scenario, unless you only want to evaluate Virtuozzo Hybrid Infrastructure on a single server.

## 2.7 Understanding Failure Domains

A failure domain is a set of services which can fail in a correlated manner. To provide high availability of data, Virtuozzo Hybrid Infrastructure spreads data replicas evenly across failure domains, according to a replica placement policy.

The following policies are available:

- Host as a failure domain (default). If a single host running multiple CS services fails (e.g., due to a power outage or network disconnect), all CS services on it become unavailable at once. To protect against data loss under this policy, Virtuozzo Hybrid Infrastructure never places more than one data replica per host. This policy is highly recommended for clusters of three nodes and more.
- Disk, the smallest possible failure domain. Under this policy, Virtuozzo Hybrid Infrastructure never places more than one data replica per disk or CS. While protecting against disk failure, this option may still result in data loss if data replicas happen to be on different disks of the same host and it fails. This policy can be used with small clusters of up to three nodes (down to a single node).

## 2.8 Understanding Storage Tiers

In Virtuozzo Hybrid Infrastructure terminology, tiers are disk groups that allow you to organize storage workloads based on your criteria. For example, you can use tiers to separate workloads produced by different tenants. Or you can have a tier of fast SSDs for service or virtual environment workloads and a tier of high-capacity HDDs for backup storage.

When assigning disks to tiers (which you can do at any time), have in mind that faster storage drives should be assigned to higher tiers. For example, you can use tier 0 for backups and other cold data (CS without SSD cache), tier 1 for virtual environments—a lot of cold data but fast random writes (CS with SSD cache), tier 2 for hot data (CS on SSD), caches, specific disks, and such.

This recommendation is related to how Virtuozzo Hybrid Infrastructure works with storage space. If a storage tier runs out of free space, Virtuozzo Hybrid Infrastructure will attempt to temporarily use the space of the lower tiers down to the lowest. If the lowest tier also becomes full, Virtuozzo Hybrid Infrastructure will attempt to use a higher one. If you add more storage to the original tier later, the data, temporarily stored elsewhere, will be moved to the tier where it should have been stored originally. For example, if you try to write data to the tier 2 and it is full, Virtuozzo Hybrid Infrastructure will attempt to write that data to tier 1, then to tier 0. If you add more storage to tier 2 later, the aforementioned data, now stored on the tier 1 or 0, will be moved back to the tier 2 where it was meant to be stored originally.

Inter-tier data allocation as well as the transfer of data to the original tier occurs in the background. You can disable such migration and keep tiers strict as described in the [Administrator's Command Line Guide](#).

**Note:** With the exception of out-of-space situations, automatic migration of data between tiers is not supported.

---

## 2.9 Understanding Cluster Rebuilding

The storage cluster is self-healing. If a node or disk fails, a cluster will automatically try to restore the lost data, i.e. rebuild itself.

The rebuild process consists of several steps. Every CS sends a heartbeat message to an MDS every 5 seconds. If a heartbeat is not sent, the CS is considered *inactive* and the MDS informs all cluster components that they stop requesting operations on its data. If no heartbeats are received from a CS for 15 minutes, the MDS considers that CS *offline* and starts cluster rebuilding (if the prerequisites below are met). In the process, the MDS finds CSes that do not have pieces (replicas) of the lost data and restores the data—one piece (replica) at a time—as follows:

- If replication is used, the existing replicas of a degraded chunk are locked (to make sure all replicas remain identical) and one is copied to the new CS. If at this time a client needs to read some data that has not been rebuilt yet, it reads any remaining replica of that data.
- If erasure coding is used, the new CS requests almost all the remaining data pieces to rebuild the missing ones. If at this time a client needs to read some data that has not been rebuilt yet, that data is rebuilt out of turn and then read.

---

**Note:** If a node or disk goes offline during maintenance, cluster self-healing is delayed to save cluster resources. The default delay is 30 minutes. You can adjust it by setting the `mds.wd.offline_tout_mnt` parameter, in milliseconds, with the `vstorage -c <cluster_name> set-config` command.

---

Self-healing requires more network traffic and CPU resources if replication is used. On the other hand, rebuilding with erasure coding is slower.

For a cluster to be able to rebuild itself, it must have at least:

1. As many healthy nodes as required by the redundancy mode
2. Enough free space to accommodate as much data as any one node can store

The first prerequisite can be explained on the following example. In a cluster that works in the 5+2 erasure coding mode and has seven nodes (i.e. the minimum), each piece of user data is distributed to 5+2 nodes for redundancy, i.e. each node is used. If one or two nodes fail, the user data will not be lost, but the cluster will become degraded and will not be able to rebuild itself until at least seven nodes are healthy again (that is, until you add the missing nodes). For comparison, in a cluster that works in the 5+2 erasure coding mode and has ten nodes, each piece of user data is distributed to the random 5+2 nodes out of ten to even out the load on CSes. If up to three nodes fail, such a cluster will still have enough nodes to rebuild itself.

The second prerequisite can be explained on the following example. In a cluster that has ten 10 TB nodes, at least 1 TB on each node should be kept free, so if a node fails, its 9 TB of data can be rebuilt on the remaining nine nodes. If, however, a cluster has ten 10 TB nodes and one 20 TB node, each smaller node should have at least 2 TB free in case the largest node fails (while the largest node should have 1 TB free).

Two recommendations that help smooth out rebuilding overhead:

- To simplify rebuilding, keep uniform disk counts and capacity sizes on all nodes.
- Rebuilding places additional load on the network and increases the latency of read and write operations. The more network bandwidth the cluster has, the faster rebuilding will be completed and bandwidth freed up.

## CHAPTER 3

# Installing Using GUI

After planning out the infrastructure, proceed to install the product on each server included in the plan.

---

**Important:** Time needs to be synchronized via NTP on all nodes in the same cluster. Make sure that the nodes can access the NTP server.

---

## 3.1 Obtaining Distribution Image

To obtain the distribution ISO image, visit the [product page](#) and submit a request for the trial version.

## 3.2 Preparing for Installation

Virtuozzo Hybrid Infrastructure can be installed from

- IPMI virtual drives
- PXE servers (in this case, time synchronization via NTP is enabled by default)
- USB drives

## 3.2.1 Preparing for Installation from USB Storage Drives

To install Virtuozzo Hybrid Infrastructure from a USB storage drive, you will need a 4 GB or higher-capacity USB drive and the Virtuozzo Hybrid Infrastructure distribution ISO image.

Make a bootable USB drive by transferring the distribution image to it with `dd`.

---

**Important:** Be careful to specify the correct drive to transfer the image to.

---

For example, on Linux:

```
# dd if=storage-image.iso of=/dev/sdb
```

And on Windows (with `dd` for Windows):

```
C:\>dd if=storage-image.iso of=\\?\Device\Harddisk1\Partition0
```

## 3.3 Starting Installation

The installation program requires a minimum screen resolution of 800x600. With 800x600, however, you may experience issues with the user interface. For example, some elements can be inaccessible. The recommended screen resolution is at least 1024x768.

To start the installation, do the following:

1. Configure the server to boot from the chosen media.
2. Boot the server and wait for the welcome screen.
3. On the welcome screen, do one of the following:
  - If you want to set installation options manually, choose **Install Virtuozzo Hybrid Infrastructure**.
  - If you want to install Virtuozzo Hybrid Infrastructure in the unattended mode, press **E** to edit the menu entry, append kickstart file location to the `linux` line, and press **Ctrl+X**. For example:

```
linux /images/pxeboot/vmlinuz inst.stage2=hd:LABEL=<ISO_img> quiet ip=dhcp \
logo.nologo=1 inst.ks=<URL>
```

For instructions on how to create and use a kickstart file, see [Creating Kickstart File](#) (page 50) and [Using Kickstart File](#) (page 56), respectively.



If you choose **Install Virtuozzo Hybrid Infrastructure**, you will be asked to complete these steps:

1. Read and accept the user agreement.
2. Set up the network.
3. Choose a time zone. The date and time will be configured via NTP.
4. Choose what storage cluster node you are installing: first or second/other. You can also choose to skip this step so you can add the node to the storage cluster manually later.
5. Choose the destination disk to install Virtuozzo Hybrid Infrastructure on.
6. Create the root password and start installation.

These steps are described in detail in the following sections.

## 3.4 Step 1: Accepting the User Agreement

On this step, please carefully read the End-User License Agreement. Accept it by ticking the **I accept the End-User License Agreement** checkbox and click **Next**.

## 3.5 Step 2: Configuring the Network

Virtuozzo Hybrid Infrastructure requires one network interface per server for management. You will need to specify a network interface to which to assign the network with the **Internal management** traffic type. After installation, you will not be able to remove this traffic type from the preconfigured network in the admin panel.

On the **Network and hostname** screen, you need to have at least one network card configured. Usually the network is configured automatically (via DHCP). If manual configuration is required, select a network card, click **Configure...**, and specify the necessary parameters.

In particular, consider setting a better MTU value. As mentioned in *Network Limitations* (page 22), MTU is set to 1500 by default, while 9000 is recommended. If you are integrating Virtuozzo Hybrid Infrastructure into an existing network, adjust the MTU value to that of the network. If you are deploying Virtuozzo Hybrid Infrastructure from scratch alongside a new network, set the MTU value to the recommended 9000.

---

**Important:** The MTU value must be the same across the entire network.

---

You will need to configure the same MTU value on:

- Each router and switch on the network (consult your network equipment manuals)
- Each node's network card as well as each bond or VLAN

It is also recommended to create two bonded connections as described in *Planning Network* (page 21) and create three VLAN interfaces on one of the bonds. One of the VLAN interfaces must be created in the installer and assigned to the admin panel network so that you can access the admin panel after the installation. The remaining VLAN interfaces can be more conveniently created and assigned to networks in the admin panel as described in the *Administrator's Guide*.

In addition, you need to provide a unique host name, either a fully qualified domain name (<hostname>.<domainname>) or a short name (<hostname>), in the **Host name** field.

---

**Important:** The only way to change the host name later is via the technical support.

---

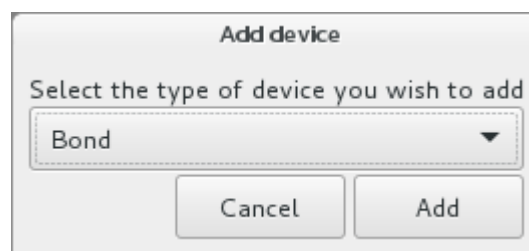
Having set up the network, click **Next**.

### 3.5.1 Creating Bonded Connections

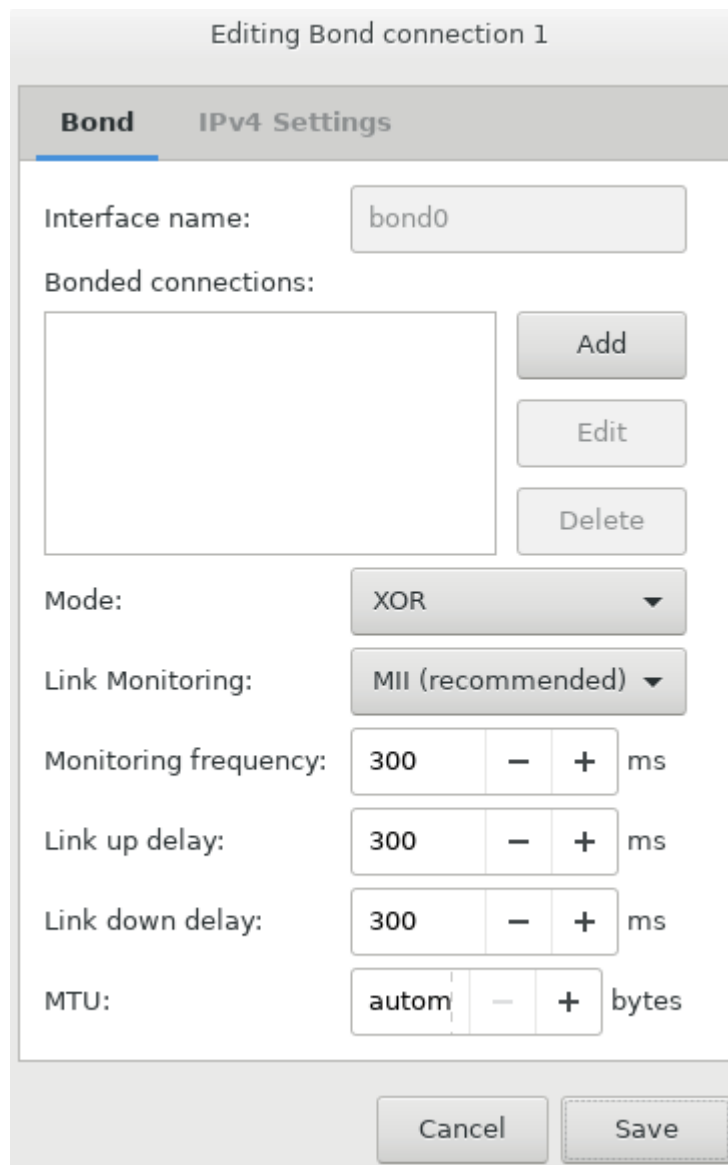
Bonded connections offer increased throughput beyond the capabilities of a single network card as well as improved redundancy.

You can create network bonds on the **Network and hostname** screen as described below.

1. To add a new bonded connection, click the plus button in the bottom, select **Bond** from the drop-down list, and click **Add**.



2. In the **Editing Bond connection...** window, set the following parameters for an Ethernet bonding interface:
  - 2.1. **Mode** to one required by your network
  - 2.2. **Link Monitoring** to MII (recommended)
  - 2.3. **Monitoring frequency**, **Link up delay**, and **Link down delay** to 300



The screenshot shows a window titled "Editing Bond connection 1" with two tabs: "Bond" (selected) and "IPv4 Settings". The "Bond" tab contains the following configuration options:

- Interface name:** bond0
- Bonded connections:** An empty list with "Add", "Edit", and "Delete" buttons.
- Mode:** XOR
- Link Monitoring:** MII (recommended)
- Monitoring frequency:** 300 ms
- Link up delay:** 300 ms
- Link down delay:** 300 ms
- MTU:** autom bytes

At the bottom of the window are "Cancel" and "Save" buttons.

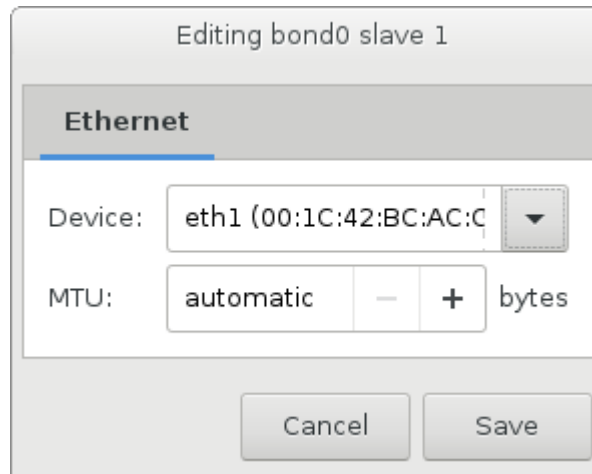
---

**Note:** It is also recommended to manually set `xmit_hash_policy` to `layer3+4` after the installation.

---

3. In the **Bonded connections** section on the **Bond** tab, click **Add**.

4. In the **Editing bond slave...** window, select a network interface to bond from the **Device** drop-down list.



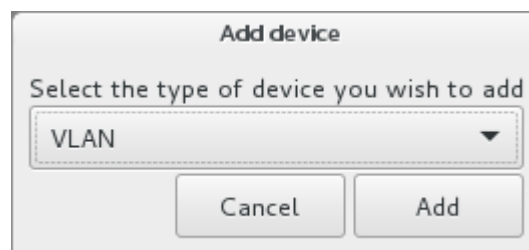
5. Configure MTU if required and click **Save**.
6. Repeat steps 3 to 5 for each network interface you need to add to the bonded connection.
7. Configure IPv4 settings if required and click **Save**.

The connection will appear in the list on the **Network and hostname** screen.

### 3.5.2 Creating VLAN Adapters

While installing Virtuozzo Hybrid Infrastructure, you can also create virtual local area network (VLAN) adapters on the basis of physical adapters or bonded connections on the **Network and hostname** screen as described below.

1. To add a new VLAN adapter, click the plus button in the bottom, select **VLAN** from the drop-down list, and click **Add**.



2. In the **Editing VLAN connection...** window:
  - 2.1. From the **Parent interface** drop-down list, select a physical adapter or bonded connection that the VLAN adapter will be based on.

2.2. Specify a VLAN adapter identifier in the **VLAN ID** field. The value must be in the 1-4094 range.

Editing VLAN connection 1

**VLAN** IPv4 Settings

Parent interface:  ▼

VLAN id:  - +

Cloned MAC address:  ▼

MTU:  - + bytes

Flags:  Reorder headers  GVRP  Loose binding  MVRP

Cancel Save

3. Configure IPv4 settings if required and click **Save**.

The VLAN adapter will appear in the list on the **Network and hostname** screen.

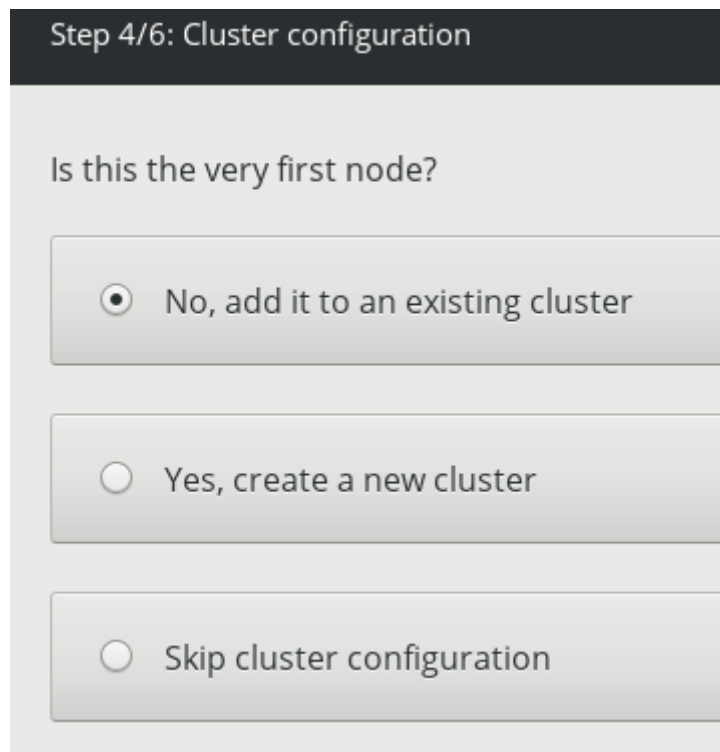
## 3.6 Step 3: Choosing the Time Zone

On this step, select your time zone. The date and time will be set via NTP. You will need an Internet connection for synchronization to complete.

## 3.7 Step 4: Configuring the Storage Cluster

On this step, you will need to specify what type of node you are installing:

- Choose **No, add it to an existing cluster** if this is a secondary node that you are adding to an existing storage cluster. Such nodes will run services related to data storage and will be added to the infrastructure during installation.
- Choose **Yes, create a new cluster** if you are just starting to set up Virtuozzo Hybrid Infrastructure and want to create a new storage cluster. This primary node, also called the management node, will host cluster management services and the admin panel. It will also serve as a storage node. Only one primary node is required.
- Choose **Skip cluster configuration** if you want to register the deployed node in the admin panel manually later on (see [Re-Adding Nodes to the Unassigned List](#)).



Step 4/6: Cluster configuration

Is this the very first node?

No, add it to an existing cluster

Yes, create a new cluster

Skip cluster configuration

Click **Next** to proceed to the next substep that depends on your choice.

### 3.7.1 Deploying the Primary Node

If you chose to deploy the primary node, do the following:

1. In the **Internal management network** drop-down list, select a network interface for internal management and configuration purposes.
2. In the **Admin panel network** drop-down list, select a network interface that will provide access to the admin panel.
3. Create and confirm a password for the superadmin account of the admin panel.
4. Click **Next**.

Create a new cluster  
This node will control the cluster. It will manage other nodes and host the web-based admin panel.

Internal management network  
eth1 - 10.37.130.183

This network is used to manage cluster nodes. It must be inaccessible from the outside.

Admin panel network  
eth0 - 10.94.94.17

The web-based admin panel will be available on this network. It should be inaccessible from the Internet and differ from the internal management network.

Create a password for the admin panel

●●●●●●●●

Strong

Confirm password

●●●●●●●●

The password must be at least 8 characters long, with at least one capital letter and one digit. The password can contain letters (a-z), numbers (0-9), dashes (-), underscores (\_), apostrophes ('), and periods (.).

### 3.7.2 Deploying Secondary Nodes

If you chose to deploy a secondary node, you will need to provide the IP address of the management node and the token that can only be obtained from the cluster admin panel. A single token can be used to deploy multiple secondary nodes in parallel.

To obtain the token and management node address:

1. Log in to the admin panel on port 8888. Panel's IP address is shown in the console after deploying the primary node. Use the default user name shown on the login screen and the primary node's root password.

If prompted, add the security certificate to browser's exceptions.

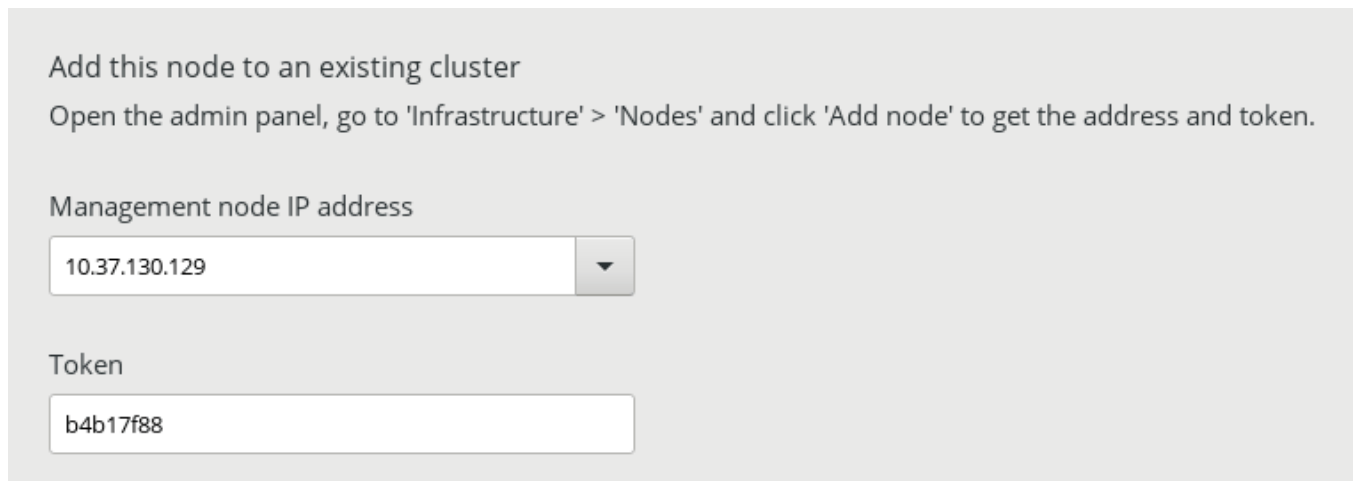
2. In the admin panel, open **INFRASTRUCTURE > Nodes** and click **ADD NODE** to invoke a screen with the management node address and the token.

---

**Note:** You can generate a new token if needed. Generating a new token invalidates the old one.

---

Back on the installation screen, enter the management node address and the token and click **Next**.



Add this node to an existing cluster

Open the admin panel, go to 'Infrastructure' > 'Nodes' and click 'Add node' to get the address and token.

Management node IP address

Token

The node may appear on the **INFRASTRUCTURE > Nodes** screen in the **UNASSIGNED** list as soon as token is validated. However, you will be able to join it to the storage cluster only after the installation is complete.

## 3.8 Step 5: Selecting the System Partition

On this step, you need to choose a disk for the operating system. This disk will have the supplementary role **System**, although you will still be able to set it up for data storage in the admin panel.

You can also create software RAID1 for the system disk to ensure its high performance and availability. To do this, tick the RAID1 checkbox and select at least two disks.



**Select system disk(s)**

Choose where to install the system: on a single disk or a software RAID volume. No data on disks will be touched until you click 'Start installation'.

Combine disks in a software RAID mirror. The resulting volume will be about as large as the smallest disk.

Disk	Type	Size	System	Purpose
sda / QEMU Vz HARDDISK0	HDD	256 GiB	<input checked="" type="radio"/>	Used by operating system
sdb / QEMU Vz HARDDISK2	HDD	1024 GiB	<input type="radio"/>	Available for storage
sdc / QEMU Vz HARDDISK3	HDD	1024 GiB	<input type="radio"/>	Available for storage

It is recommended to create RAID1 from disks of the same size as the volume equals the size of the smallest disk.

Click **Next**.

---

**Important:** All information on all disks recognized by the installer will be destroyed.

---

## 3.9 Step 6: Setting the Root Password

On the last step, enter and confirm the password for the root account and click **Start installation**.

Once the installation is complete, the node will reboot automatically. The admin panel IP address will be shown in the welcome prompt.

## 3.10 Finishing Installation

After deploying the primary node, proceed to deploy the required amount of secondary nodes as explained in [Deploying Secondary Nodes](#) (page 42). Make sure that all the nodes are shown in the admin panel: in the **UNASSIGNED** list on the **INFRASTRUCTURE > Nodes** screen.

If you skipped cluster configuration on step 4 and want to add the node to the **UNASSIGNED** list manually, consult [Re-Adding Nodes to the Unassigned List](#).

When all the required nodes are present in the **UNASSIGNED** list, proceed to create the storage cluster as described in the [Administrator's Guide](#).

## CHAPTER 4

# Installing Using PXE

This chapter explains how to install Virtuozzo Hybrid Infrastructure over network using a preboot execution environment (PXE) server.

You will need to do the following:

1. Get the distribution image as described in *Obtaining Distribution Image* (page 34).
2. Set up the TFTP, DHCP, and HTTP (or FTP) servers.
3. Boot the node where you will install Virtuozzo Hybrid Infrastructure from network and launch the Virtuozzo Hybrid Infrastructure installer.
4. Set installation options manually or supply them automatically by means of a kickstart file and complete installation.

## 4.1 Preparing Environment

This section explains how to set up the environment for installation over network.

### 4.1.1 Installing PXE Components

You will need these components to set up a PXE environment:

- TFTP server. This is a machine that allows your servers to boot and install Virtuozzo Hybrid Infrastructure over the network. Any machine that can run Linux and is accessible over network can be a TFTP server.
- DHCP server. This is a standard DHCP machine serving TCP/IP settings to computers on your network.

- HTTP server. This is a machine serving Virtuozzo Hybrid Infrastructure installation files over network.

You can also share Virtuozzo Hybrid Infrastructure distribution over network via FTP (e.g., with vsftpd) or NFS.

The easiest way is to set up all of these on the same physical machine:

```
# yum install tftp-server syslinux httpd dhcp
```

You can also use servers that already exist in your infrastructure. For example, skip httpd and dhcp if you already have the HTTP and DHCP servers.

## 4.1.2 Configuring TFTP Server

This section describes how to configure the TFTP server for BIOS-based systems. For information on how to configure it for installing Virtuozzo Hybrid Infrastructure on EFI-based systems, see the [Red Hat Enterprise Linux Installation Guide](#).

Do the following:

1. On the server, open the `/etc/xinetd.d/tftp` file, and edit it as follows:

```
service tftp
{
  disable          = no
  socket_type      = dgram
  protocol         = udp
  wait             = yes
  user             = root
  server           = /usr/sbin/in.tftpd
  server_args      = -v -s /tftpboot
  per_source       = 11
  cps              = 100 2
  flags            = IPv4
}
```

Once you are done, save the file.

2. Create the `/tftpboot` directory and copy the following files to it: `mlinuz`, `initrd.img`, `menu.c32`, `pxelinux.0`.

These files are necessary to start installation. You can find the first two in the `/images/pxeboot` directory of the Virtuozzo Hybrid Infrastructure distribution. The last two files are located in the `syslinux` directory (usually `/usr/share/syslinux` or `/usr/lib/syslinux`).

3. Create the `/tftpboot/pxelinux.cfg` directory and make the default file in it.

```
# mkdir /tftpboot/pxelinux.cfg
# touch /tftpboot/pxelinux.cfg/default
```

4. Add the following lines to default:

```
default menu.c32
prompt 0
timeout 100
ontimeout INSTALL
menu title Boot Menu
label INSTALL
    menu label Install
    kernel vmlinuz
    append initrd=initrd.img ip=dhcp
```

For detailed information on parameters you can specify in this file, see the documentation for `syslinux`.

5. Restart the `xinetd` service:

```
# /etc/init.d/xinetd restart
```

6. If necessary, configure firewall to allow access to the TFTP server (on port 69 by default).

When running the TFTP server, you might get the “Permission denied” error. In this case, you may try to fix the problem by running the following command: `# restorecon -Rv /tftpboot/`.

### 4.1.3 Setting Up DHCP Server

To set up a DHCP server for installing Virtuozzo Hybrid Infrastructure over network, add the following strings to the `dhcpd.conf` file, which is usually located in the `/etc` or `/etc/dhcp` directory:

```
next-server <PXE_server_IP_address>;
filename "/pxelinux.0";
```

To configure a DHCP server for installation on EFI-based systems, specify filename `"/bootx64.efi"` instead of filename `"/pxelinux.0"` in the `dhcpd.conf` file, where `/bootx64.efi` is the directory to which you copied the EFI boot images when setting up the TFTP server.

### 4.1.4 Setting Up HTTP Server

Now that you have set up the TFTP and DHCP servers, you need to make the Virtuozzo Hybrid Infrastructure distribution files available for installation over the network. To do this:

1. Set up an HTTP server (or configure the one you already have).

2. Copy the contents of your Virtuozzo Hybrid Infrastructure installation DVD to some directory on the HTTP server (e.g., /var/www/html/distrib).
3. On the PXE server, specify the path to the Virtuozzo Hybrid Infrastructure installation files in the append line of the /tftpboot/pxelinux.cfg/default file.

For EFI-based systems, the file you need to edit has the name of /tftpboot/pxelinux.cfg/efidefault or /tftpboot/pxelinux.cfg/<PXE\_server\_IP\_address>.

Assuming that the HTTP server is at 198.123.123.198, the installation files are in /var/www/html/distrib/, and DocumentRoot is set to /var/www/html, the default file may look like this:

```
default menu.c32
prompt 0
timeout 100
ontimeout INSTALL
menu title Boot Menu
label INSTALL
    menu label Install
    kernel vmlinuz
    append initrd=initrd.img ip=dhcp inst.repo=http://198.123.123.198/distrib
```

## 4.2 Installing Over the Network

Now that you have prepared all the servers, you can install Virtuozzo Hybrid Infrastructure over the network:

1. Boot the Virtuozzo Hybrid Infrastructure server from the network. You should see the **Boot Menu** that you have created.
2. In the boot menu, choose **Install Virtuozzo Hybrid Infrastructure**.
3. On the main installer screen, set installation options as described in *Installing Using GUI* (page 34) and click **Begin Installation**.

If you want to install Virtuozzo Hybrid Infrastructure in the unattended mode, you will need to do the following:

1. Create a kickstart file as described in *Creating Kickstart File* (page 50).
2. Add the kickstart file location to the boot menu as explained in *Using Kickstart File* (page 56).
3. Boot the node from network and choose **Install Virtuozzo Hybrid Infrastructure** in the boot menu.

Installation should proceed automatically.

## 4.3 Creating Kickstart File

If you plan to perform an unattended installation of Virtuozzo Hybrid Infrastructure, you can use a kickstart file. It will automatically supply to the Virtuozzo Hybrid Infrastructure installer the options you would normally choose by hand. Virtuozzo Hybrid Infrastructure uses the same kickstart file syntax as Red Hat Enterprise Linux.

The following sections describe the options and scripts you will need to include in your kickstart file, provide an example you can start from, and explain how to use the kickstart file you have created.

### 4.3.1 Kickstart Options

Even though your kickstart file may include any of the standard options, it is recommended to only use the ones listed in this section. They are mandatory and must be included in your kickstart file.

`auth --enablshadow --passalgo=sha512`

Specifies authentication options for the Virtuozzo Hybrid Infrastructure physical server.

`autopart --type=lvm`

Automatically partitions the system disk, which is `sda`. This option must follow `clearpart --all`.

Other disks will be partitioned automatically during cluster creation.

`bootloader`

Specifies how the boot loader should be installed.

`clearpart --all`

Removes all partitions from all recognized disks.

**Warning:** This option will destroy data on all the disks that the installer can reach!

`keyboard <layout>`

Sets the system keyboard type.

`lang <lang>`

Sets the language to use during installation and the default language to use on the installed system.

logvol

Creates a logical volume for a Logical Volume Management (LVM) group.

network <options>

Configures network devices and creates bonds and VLANs.

raid Creates a software RAID volume.

part Creates a partition on the server.

---

**Note:** The size of the /boot partition must be at least 1 GB.

---

rootpw --iscrypted <passwd>

Sets the root password for the server. The value is your password's hash obtained with the algorithm specified in the --passalgo parameter. For example, to create a SHA-512 hash of your password, run `python -c 'import crypt; print(crypt.crypt("yourpassword"))'`.

selinux --disabled

Disables SELinux, because it prevents virtualization from working correctly.

services --enabled="chronyd"

Enables time synchronization via NTP.

timezone <timezone>

Sets the system time zone. For a list of time zones, run `timedatectl list-timezones`.

volgroup

Creates a Logical Volume Management (LVM) group.

zerombr

Initializes disks with invalid partition tables.

**Warning:** This option will destroy data on all the disks that the installer can reach!



## 4.3.2 Kickstart Scripts

After setting the options, add scripts to the kickstart file that will install the mandatory package group and Storage components.

### 4.3.2.1 Installing Packages

In the body of the `%packages` script, specify the package group `hci` to be installed on the server:

```
%packages
@^hci
%end
```

### 4.3.2.2 Installing Admin Panel and Storage

Only one admin panel is required, install it on the first node only. To deploy all other nodes, you will need to obtain a token from a working admin panel. For more information, see the [Deploying Secondary Nodes](#) (page 42).

To install the admin panel and storage components on the node without exposing the superadmin password and storage token in the kickstart file, do the following:

1. Add the `%addon com_vstorage` script to the kickstart file:

```
%addon com_vstorage --management --bare
%end
```

2. Once the installation is complete, execute the following command on the node to configure the admin panel component:

```
echo <superadmin_password> | /usr/libexec/vstorage-ui-backend/bin/configure-backend.sh \
-i <private_iface> -x <public_iface>
```

where

- `<superadmin_password>` is the password of the superadmin account of admin panel.
- `<private_iface>` is the name of the private network interface (the one you would choose for the management network during attended installation).
- `<public_iface>` is the name of the public network interface (the one you would choose for the admin panel network during attended installation).

3. Start the admin panel service:

```
# systemctl start vstorage-ui-backend
```

4. If you also installed the storage component on the node, execute the following command:

```
# /usr/libexec/vstorage-ui-agent/bin/register-storage-node.sh -m <management_IP_address>
```

To install the components without running scripts afterwards at the expense of exposing the password and token, specify the interfaces for the public (external) and private (internal) networks and the password for the superadmin account of the admin panel in the kickstart file. For example:

```
%addon com_vstorage --management --internal-iface=<private_iface> \
--external-iface=<public_iface> --password=<password>
%end
```

### 4.3.2.3 Installing Storage Component Only

The storage component alone, without the admin panel, is installed by default and does not require any scripts in the kickstart file unless you want to specify the token.

If you do not want to expose the token in the kickstart file, run the following command on the node after the installation to register the node in the admin panel:

```
# /usr/libexec/vstorage-ui-agent/bin/register-storage-node.sh -m <MN_IP_address> -t <token>
```

where

- <token> is the token that can be obtained in the admin panel.
- <MN\_IP\_address> is the IP address of the private network interface on the node with the admin panel.

To install the storage component without running scripts afterwards at the expense of exposing the token, specify the token and the IP address of the node with the admin panel in the kickstart file. For example:

```
%addon com_vstorage --storage --token=<token> --mgmt-node-addr=<MN_IP_address>
%end
```

## 4.3.3 Kickstart File Example

Below is an example of kickstart file that you can use to install and configure Virtuozzo Hybrid Infrastructure in the unattended mode. You can use this file as the basis for creating your own kickstart files.

**Important:** This kickstart file instructs the installer to erase and automatically partition every disk that it recognizes. Make sure to disconnect any disks with useful data prior to installation.

```
# Use the SHA-512 encryption for user passwords and enable shadow passwords.
auth --enablshadow --passalgo=sha512
# Use the US English keyboard.
keyboard --vckeymap=us --xlayouts='us'
# Use English as the installer language and the default system language.
lang en_US.UTF-8
# Specify the encrypted root password for the node.
rootpw --iscrypted xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
# Disable SELinux.
selinux --disabled
# Enable time synchronization via NTP.
services --enabled="chronyd"
# Set the system time zone.
timezone America/New_York

# Specify a hostname for the node.
# NOTE: The only way to change the host name later is via the technical support.
network --hostname=<hostname>

# Configure network interfaces via DHCP.
network --device=<iface1> --activate
network --device=<iface2> --activate
# Alternatively, assign static addresses to network interfaces.
#network --device=<iface1> --activate --bootproto=static --ip=<IP_addr> \
#--netmask=<mask> --gateway=<gw> --nameserver=<ns1>[,<ns2>,...]
#network --device=<iface2> --activate --bootproto=static --ip=<IP_addr> \
#--netmask=<mask> --gateway=<gw> --nameserver=<ns1>[,<ns2>,...]

# If needed, uncomment and specify network interfaces to create a bond.
#network --device=bond0 --bondslaves=<iface1>,<iface2> \
#--bondopts=mode=balance-xor,miimon=100,xmit_hash_policy=layer3+4

# Erase all partitions from all recognized disks.
# WARNING: Destroys data on all disks that the installer can reach!
clearpart --all --initlabel
zerombr
# Automatically partition the system disk, which is 'sda'.
autopart --type=lvm

# Install the required packages on the node.
%packages
@^hci
%end

# Uncomment to install the admin panel and storage components.
# Specify an internal interface for the management network and
```

```
# an external interface for the admin panel network.
#%addon com_vstorage --management --internal-iface=eth0 \
--external-iface=eth1 --password=xxxxxxxx
#%end

# Uncomment to install the storage component. To register the node,
# specify the token as well as the IP address of the admin panel.
#%addon com_vstorage --storage --token=xxxxxxxx --mgmt-node-addr=10.37.130.1
#%end
```

### 4.3.3.1 Creating the System Partition on Software RAID1

To create a system partition on a software RAID1 volume, you will need to do the following instead of using `autopart`:

1. Partition the disks.
2. Create a RAID1 volume.
3. Create swap and root LVM volumes.

It is recommended to create RAID1 from disks of the same size as the volume equals the size of the smallest disk.

The following example for a BIOS-based server partitions the disks `sda` and `sdb`, assembles the software RAID1 array, and creates expandable swap and root LVM volumes:

```
# Create partitions on sda.
part biosboot --size=1 --ondisk=sda --fstype=biosboot
part raid.sda1 --size=1024 --ondisk=sda --fstype=ext4
part raid.sda2 --size=101376 --ondisk=sda --grow
# Create partitions on sdb.
part biosboot --size=1 --ondisk=sdb --fstype=biosboot
part raid.sdb1 --size=1024 --ondisk=sdb --fstype=ext4
part raid.sdb2 --size=101376 --ondisk=sdb --grow
# Create software RAID1 from sda and sdb.
raid /boot --level=RAID1 --device=md0 --fstype=ext4 raid.sda1 raid.sdb1
raid pv.01 --level=RAID1 --device=md1 --fstype=ext4 raid.sda2 raid.sdb2
# Make LVM volumes for swap and root partitions.
volgroup vgsys pv.01
logvol swap --fstype=swap --name=swap --vgname=vgsys --recommended
logvol / --fstype=ext4 --name=root --vgname=vgsys --size=10240 --grow
# Set the RAID device md0 as the first drive in the BIOS boot order.
bootloader --location=mbr --boot-drive=sda --driveorder=md0
bootloader --location=mbr --boot-drive=sdb --driveorder=md0
```

For installation on EFI-based servers, specify the `/boot/efi` partition instead of `biosboot`.

```
part /boot/efi --size=200 --ondisk={sda|sdb} --fstype=efi
```

## 4.4 Using Kickstart File

To install Virtuozzo Hybrid Infrastructure using a kickstart file, you first need to make the kickstart file accessible over the network. To do this:

1. Copy the kickstart file to the same directory on the HTTP server where the Virtuozzo Hybrid Infrastructure installation files are stored (e.g., to `/var/www/html/astor`).
2. Add the following string to the `/tftpboot/pxelinux.cfg/default` file on the PXE server:

```
inst.ks=<HTTP_server_address>/<path_to_kickstart_file>
```

For EFI-based systems, the file you need to edit has the name of `/tftpboot/pxelinux.cfg/efidefault` or `/tftpboot/pxelinux.cfg/<PXE_server_IP_address>`.

Assuming that the HTTP server has the IP address of 198.123.123.198, the DocumentRoot directory is set to `/var/www/html`, and the full path to your kickstart file on this server is `/var/www/html/astor/ks.cfg`, your default file may look like the following:

```
default menu.c32
prompt 0
timeout 100
ontimeout ASTOR
menu title Boot Menu
label ASTOR
    menu label Install
        kernel vmlinuz
        append initrd=initrd.img ip=dhcp inst.repo=http://198.123.123.198/astor \
inst.ks=http://198.123.123.198/astor/ks.cfg
```

## CHAPTER 5

# Additional Installation Modes

This chapter describes additional installation modes that may be of help depending on your needs.

## 5.1 Installing via VNC

To install Virtuozzo Hybrid Infrastructure via VNC, boot to the welcome screen and do the following:

1. Select the main installation option and press **E** to start editing it.
2. Add `text` at the end of the line starting with `linux /images/pxeboot/vmlinuz`. For example:

```
linux /images/pxeboot/vmlinuz inst.stage2=hd:LABEL=<ISO_img> quiet ip=dhcp logo.nologo=1 text
```

3. Press **Ctrl+X** to start booting the chosen installation option.
4. When presented with a choice of starting VNC or proceeding to the text mode, press **1**.
5. Enter a VNC password when offered.
6. In the output that follows, look up the hostname or IP address and VNC port to connect to, e.g.,  
192.168.0.10:1.
7. Connect to the address in a VNC client. You will see the usual **Installation Summary** screen.

The installation process itself is the same as in the default graphics mode (see *Installing Using GUI* (page 34)).

## CHAPTER 6

# Troubleshooting Installation

This chapter describes ways to troubleshoot installation of Virtuozzo Hybrid Infrastructure.

## 6.1 Installing in Basic Graphics Mode

If the installer cannot load the correct driver for your graphics card, you can try to install Virtuozzo Hybrid Infrastructure in the basic graphics mode. To select this mode, on the welcome screen, choose **Troubleshooting->**, then **Install in basic graphics mode**.

In this mode, however, you may experience issues with the user interface. For example, some of its elements may not fit the screen.

The installation process itself is the same as in the default graphics mode (see *Installing Using GUI* (page 34)).

## 6.2 Booting into Rescue Mode

If you experience problems with your system, you can boot into the rescue mode to troubleshoot these problems. Once you are in the rescue mode, your Virtuozzo Hybrid Infrastructure installation is mounted under `/mnt/sysimage`. You can go to this directory and make the necessary changes to your system.

To enter the rescue mode, do the following:

1. Boot your system from the Virtuozzo Hybrid Infrastructure distribution image.
2. On the welcome screen, click **Troubleshooting->**, then **Rescue system**.

3. Once Virtuozzo Hybrid Infrastructure boots into the emergency mode, press **Ctrl+D** to load the rescue environment.
4. In the rescue environment, you can choose one of the following options:
  - Continue (press **1**): mount the Virtuozzo Hybrid Infrastructure installation in read and write mode under `/mnt/sysimage`.
  - Read-only mount (press **2**): mount the Virtuozzo Hybrid Infrastructure installation in read-only mode under `/mnt/sysimage`.
  - Skip to shell (press **3**): load shell, if your file system cannot be mounted; for example, when it is corrupted.
  - Quit (Reboot) (press **4**): reboot the server.
5. Unless you press **4**, a shell prompt will appear. In it, run `chroot /mnt/sysimage` to make the Virtuozzo Hybrid Infrastructure installation the root environment. Now you can run commands and try to fix the problems you are experiencing.
6. After you fix the problem, run `exit` to exit the chrooted environment, then `reboot` to restart the system.