FUJITSU

ETERNUS CD10000 V1.0 SP2

User Guide

English

English

FUJITSU

ETERNUS CD10000

User Guide

V1.0 SP2

Edition April 2015

Comments… Suggestions… Corrections…

The User Documentation Department would like to know your opinion of this manual. Your feedback helps us optimize our documentation to suit your individual needs.

Feel free to send us your comments by e-mail to manuals@ts.fujitsu.com.

Certified documentation according to DIN EN ISO 9001:2008

To ensure a consistently high quality standard and user-friendliness, this documentation was created to meet the regulations of a quality management system which complies with the requirements of the standard DIN EN ISO 9001:2008.

cognitas. Gesellschaft für Technik-Dokumentation mbH

www.cognitas.de

Copyright and trademarks

Delivery subject to availability; right of technical modifications reserved.

Contents

1 Introduction

1.1 Preface

This manual introduces the FUJITSU Storage ETERNUS CD10000 enhanced data protection system and provides information about:

- Features and hardware
- Installation
- Configuration and administration of Ceph storage cluster via Web interface and CLI
- Monitoring and logging
- Network configuration
- Software updates and upgrades
- Basic spare parts replacement procedures

Audience

This manual is written for ETERNUS CD10000 system operators and system administrators.

It is useful for the audience to have a basic understanding of Linux.

Document organization

Following is a brief description of chapter contents.

- Chapter "Overview " on page 21 provides an overview of the ETERNUS CD10000 storage cluster, its architecture and components and how they work together.
- Chapter "Networks" on page 33 describes existing network of the ETERNUS CD10000.
- Chapter "Getting Started" on page 39 describes all steps to install the initial software configuration.
- Chapter "Introduction to the User Interfaces" on page 57 describes introduction to the User Interfaces, Web-based tool **VSM** , the command line interface **`vsm_cli`** and Basking GUI.
- Chapter "Configuration and Administration of a Storage Cluster " on page 93 describes the creation, configuration and management of Ceph storage clusters with the Web-based tool **VSM** and the command line interface **`vsm_cli`**.
- Chapter "Backup and Restoration of Management Node " on page 133 describes the backup and restoration of the management node.

- Chapter "Monitoring" on page 139 describes the monitoring of the Ceph storage clusters and their instances
- Chapter "Replacing Disks Physically " on page 171 describes the replacing the disks without changing the overall storage cluster configuration. The storage cluster does not need to be stopped.
- Chapter "Software Configuration and Update/Upgrade System" on page 185 describes the software configuration and the process of software update/upgrade.
- Chapter "Log Manager" on page 197 describes the centralized log manager of ETERNUS CD10000.
- Chapter "SNMP" on page 203 describes how to monitor the storage cluster via SNMP V2.
- Chapter "Service Support" on page 219 describes two functions of the ETERNUS CD10000, which are implemented in order to support service engineers.
- Chapter "Preparing VSM Cluster Creation Manually" on page 233 describes how to prepare VSM cluster creation manually.
- Chapter "Command Line Interfaces" on page 247 describes the **`vsm_cli`** commands and the `cd10000 commands`.
- "Glossary" on page 365 provides definitions of terms used in this guide.

Notational conventions

This manual uses the following conventions:

A note emphasizes important information related to the main topic.

CAUTION!
A caution indicates potential hazards to equipment or data.

WARNING!
A warning indicates potential hazards to personal safety.

Data sizes are reported in base 1000 rather than base 1024.

For example: 1 GB = 1,000,000,000 bytes, 1 TB = 1,000,000,000,000 bytes

Related documents

The following documents are also available for ETERNUS CD10000 systems:

Order Number	Document Title	Document Description
A26361-F1452-Z253-7-8N19	Safety Notes and Regulations	Lists Fujitsu's general safety and regulatory information.

This User Guide requires the basic knowledge of the open-source software Ceph®. For detailed information see http://ceph.com/docs/master/.

1.1.1 Support and Help

Where to find information

Resource	Address
Product information and documentation	ETERNUS CD10000 product information on the internet.
ETERNUS CD10000 manuals	http://manuals.ts.fujitsu.com/index.php?id=9662-16783-16761
Release note	http://manuals.ts.fujitsu.com/index.php?cookie
Downloads and Software patches	http://support.ts.fujitsu.com/Index.asp?LNG=de

1.1.2 Technical Support

The ETERNUS CD10000 system includes AIS real-time support via remote access (see "Service Support" on page 219).

The Support section of the Fujitsu web site provides links with contact information and information about available support programs:http://www.fujitsu.com/fts/services/

If you have an active support agreement, you may contact your service provider.

If you do not have an active support agreement, contact your Fujitsu sales representative to purchase a service contract or updates.

The following error documents are needed for diagnostics:

- a detailed description of the error condition, indicating whether and how the error can be reproduced
- the ETERNUS CD10000 serial number

  The serial number can be obtained with the following command on the management node:

  `# cd10000 ip cluster show`

  Output:

```
+------------------+--------------------+--------------------+--------+
|  Cluster SN      | Management node SN | Administration net | ...    |
+------------------+--------------------+--------------------+--------+
| YM2D0010034      |     YLNT004583     |  192.168.20.0/23   | ...    |
+------------------+--------------------+--------------------+--------+
```

  The entry in column `Cluster SN` is the relevant serial number.

1.1.3 Customer Service

Resource	Address
Education and training	Fujitsu Mies-van-der-Rohe-Str. 8 80807 Munich, Germany training@ts.fujitsu.com
Support phone	+49-180-54040
Support fax	+49-180-5336779

Your comments

Your suggestions will help us continue to improve the accuracy, organization, and overall quality of the user publications. Please send your opinion of this document to: manuals@ts.fujitsu.com

If you have issues, comments, or questions about specific information or procedures, please include the title and, if available, the part number, the page numbers, and any other details that will help us locate the subject you are addressing.

Fujitsu home page

Visit the home page of Fujitsu at:

http://www.fujitsu.com/fts/

Fujitsu compliance link

See the compliance link of Fujitsu at:

https://globalsp.ts.fujitsu.com/sites/certificates/default.aspx

To get information about the storage component, select

"Storage Products > ETERNUS JX > ETERNUS JX40".

"Storage Products > ETERNUS JX > ETERNUS JX60".

To get information of the server component, select:

"Standard Industry Server > Rack Server > Primergy RX300 S8“.

Worldwide end-user product warranty

For information the Fujitsu worldwide end-user standard limited product warranty look at:
http://support.ts.fujitsu.com/warranty/index.asp?lng=COM&Level1=&LNID=1

1.2 Typographic Conventions

The following typographic conventions are used:

Convention	Explanation
	Indicates various types of risk, namely health risks, risk of data loss and risk of damage to devices.
	Indicates additional relevant information and tips.
bold	Indicates references to names of interface elements.
`monospace`	Indicates system output and system elements, e.g., file names and paths.
`monospace semibold`	Indicates statements that are to be entered using the keyboard.
blue continuous text	Indicates a link to a related topic.
pink continuous text	Indicates a link to a location you have already visited.
<abc>	Indicates variables which must be replaced with real values.
[abc]	Indicates options that can be specified (syntax).
[key]	Indicates a key on your keyboard. If you need to enter text in uppercase, the Shift key is specified, for example, [SHIFT] + [A] for A. If you need to press two keys at the same time, this is indicated by a plus sign between the two key symbols.

Screenshots

Some of the screenshots are system-dependent, so some of the details shown may differ from your system. There may also be system-specific differences in menu options and commands.

2 Overview

The FUJITSU Storage ETERNUS CD10000 is a hyperscale storage system, which provides scale out storage for up to 224 storage nodes.

It offers the following fundamental features, which permit auto-scaling, self-healing, and self-managing storage clusters:

Distributed, redundant storage of data

The overall design of the hardware and software components prevents a single point of failure by ensuring permanent data replication by means of replicas. The data is distributed randomly and uniformly across the storage nodes of the storage cluster. Therefore there is no hotspot which might be a bottleneck. Metadata are computed instead of being stored in a single database.

The placement of the data is based on the physical infrastructure by defining easy and flexible placement rules. For detailed information, see "CRUSH Data Placement" on page 29.

Scalability and flexibility

Changes in the infrastructure are dynamically adapted and have no significant impact on data mapping. For example, when you add a storage node to a storage cluster all cluster parts are automatically informed about the expansion. Data objects are dynamically stored, replicated, and rebalanced across all storage nodes of the storage cluster.

Fault tolerance

A failed disk or node is automatically detected and the affected data is copied to the remaining disk ensuring the defined rules.

Functionality based on Ceph

The functionality is provided by Ceph®, The Future of Storage™ (Ceph). Ceph is an open-source software-defined storage system also known as RADOS, that is a "reliable, autonomic, distributed, object store comprised of self-healing, self-managing, intelligent storage nodes" http://ceph.com/docs/master/architecture). A Ceph storage cluster can manage a huge amount of data and is therefore suitable for cloud applications (like OpenStack).

As in any classical distributed file system, the files put into the cluster are stripped and placed in the cluster nodes according to a pseudo-random, data-distribution algorithm known as the Ceph-controlled Replication Under Scalable Hashing (CRUSH). For detailed information, see "CRUSH Data Placement" on page 29.

For a detailed description of Ceph, see the web page http://ceph.com/docs/master/architecture.

For detailed information on the provided interfaces, see "Supported Interfaces" on page 26.

Hardware

ETERNUS CD10000 nodes are standard PRIMERGY servers and ETERNUS JX storage modules. Together with an InfiniBand (IB)-based cluster interconnect the components provide a reliable, consistent, and high-performance storage cluster supporting the Ceph functions.

Software

Ease of use is provided by the functionality of central software deployment, management, administration, and monitoring. These services run on the management node. All nodes - storage nodes and management nodes - are connected via an internal cluster network, which is used for all necessary communication and data transfer needed to setup the ease of use services.

2.1 ETERNUS CD10000 Features

The ETERNUS CD10000 storage clusters combine the following important features:

Scalability

ETERNUS CD10000 provides almost unlimited scalability. You can add storage nodes to the storage cluster at any time. Up to 224 storage nodes are supported.

Owing to the Ceph concept, the storage cluster offers the following advantages:

- Capacity expansions without negative impact to the overall performance.
- No hotspots because the data is equally distributed within the storage cluster even after adding storage nodes.
- Transparent phase out of old hardware and phase in of new hardware.
- Unified access for object, block, and file storage.

Redundancy and high availability

The Ceph concepts provides a self-managing and self-healing distributed storage architecture. It offers automatic failure detection and failover of single disks or storage nodes.

Fault tolerance is achieved with replicas that are copies of a data object. You can specify how many replicas shall be written. For example, if an object has three replicas there are one primary copy and two further copies of the object.

When a failure is detected, the data residing on the failed component is copied from one of the replicas to the remaining components (automatic recovery).

Ease of use

ETERNUS CD10000 decouples the management efforts from the storage growth rate. Adding capacity to the storage cluster is done without reconfiguration of the nodes or the network.

Tools to monitor and manage the appliance from the central management node are a basic part of the storage cluster.

These tools provide the following functions:

- A web-based graphical user interface Virtual Storage Manager (VSM) and command-line programs to manage and monitor the storage cluster.
- Monitoring and administrating with VSM from any computer with a web browser.
- Collection and management of all relevant system log files, Ceph log files, and SNMP messages on the management node to support central monitoring and maintenance.
- SNMP integration of all nodes and network components.
- A central software repository on the management node for an easy upgrade of the software on the storage nodes.

2.2 Components of a Storage Cluster

The ETERNUS CD10000 storage cluster consists of the following basic components:

Storage nodes

The storage nodes provide the physical disk storage to store and read the data that is written by the clients. All storage nodes are interconnected via the cluster network.

Three different types of storage nodes are available:

- **Basic storage node**
  A Primergy RX300 S8 server with 14 disks as data store (each disk of 900 GB) is used for this type of storage node. The total data storage capacity of one basic storage node is 12 TB.
- **Performance node**
  A basic storage node which has an ETERNUS JX40 disk system connected build this type of node. The ETERNUS JX40 system is equipped with 24 disks each with a capacity of 900 GB. The total storage capacity of one performance node is 33 TB.
- **Capacity node**
  A basic storage node which has an ETERNUS JX60 connected build this type of storage node. The ETERNUS JX60 is equipped with 60 disks each with a capacity of 4 TB. The total storage capacity of one capacity node is 252 TB.

It is possible to build the ETERNUS CD10000 appliance with different types of storage nodes. The minimum configuration consists of four storage nodes in one cluster. For example: either four basic storage nodes, four performance nodes or four capacity nodes or a two-to-two mixture of the type.

You can add additional storage nodes to the storage cluster at any time. This is possible either one by one or several nodes in one step. Please contact Fujitsu services for additional information.

Management node

The management node provides all functions to manage, administer, and monitor the storage cluster from a single point of access. The management node is connected to the cluster via the administration network.

Cluster network (private network)

The cluster interconnect network is a fast 40 Gb network for the communication between the storage nodes.

This InfiniBand (IB) based network is used for high speed transfer of the replicas according to the configured rules between the storage nodes.

Depending on the appropriate ETERNUS CD10000 option, the cluster network may be redundant to provide highly available data paths. With the redundant configuration up to 224 storage nodes are supported. 23 storage nodes maximum are supported in the non-redundant configuration.

Administration network

The 1GbE based administration network is used for all management, administration, and monitoring tasks of the storage cluster. Depending on the appropriate ETERNUS CD10000 option, the administration network may be redundant to provide highly available data paths. The management node is connected to the administration network.

Additional components

Public network

This 10GbE-based network connects the storage cluster to the customer network used to write and read data to and from the storage cluster. Each storage node has two 10 GbE ports for the connection to the public network.

Clients

A client is a host on which the Ceph client software is installed. All Ceph interfaces are available via the client. Clients are connected to the storage cluster via the public network.

2.3 Supported Interfaces

Ceph and therefore ETERNUS CD10000 offer the following interfaces with asynchronous communication capability for storing and retrieving data in the storage cluster:

- Object storage via the RADOS Gateway (RADOSGW).
- Block device storage via the RADOS Block Device (RBD). Virtual disks are provisioned by the RBD.
- Ceph file system (CEPH FS), on special release request only: CEPH FS is a POSIX-compliant file system that uses a Ceph storage cluster to store its data.

LIBRADOS is the basis of the three supported interfaces. It provides a low-level C-interface to the RADOS service that is a "reliable, autonomic, distributed, object store comprised of self-healing, self-managing, intelligent storage nodes" (http://ceph.com/docs/master/architecture)

You can use the different interfaces in parallel within one storage cluster. The interfaces are available on the clients.

The figure below shows the supported interfaces.

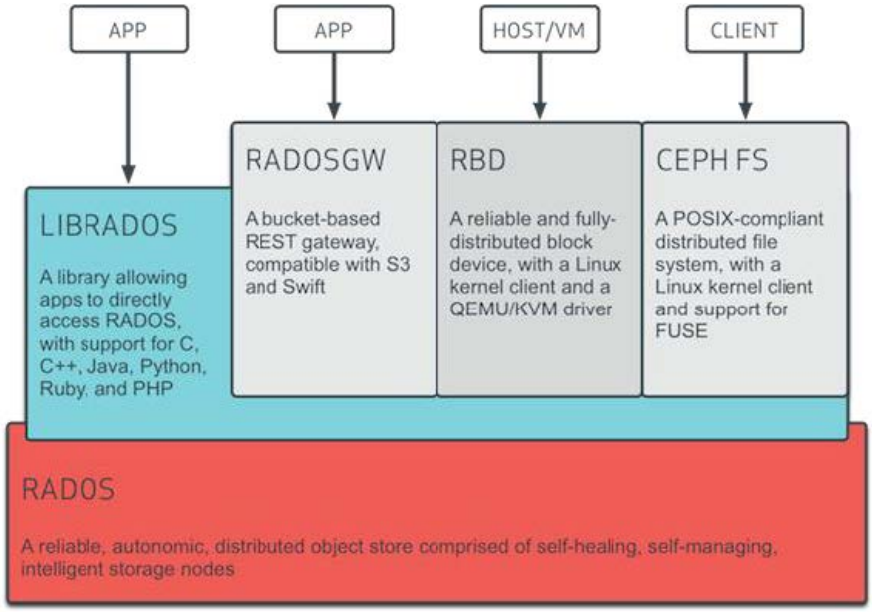


Figure 1: Ceph interfaces (http://ceph.com/docs/master/architecture)

The interfaces are used by different applications:

- The block device is supported by the hypervisors KVM and Xen and is integrated into OpenStack Cinder and Glance. You can use block devices as generic devices.
- The object storage supports the OpenStack Swift API and the Amazon S3 RESTful API.
- NFS and CIFS will be supported, as soon as the needed Ceph file system is supported. The Ceph file system uses the same Ceph storage cluster system as block device storage and object storage with its S3 and Swift APIs.

2.4 ETERNUS CD10000 Architecture

The ETERNUS CD10000 storage cluster offers cloud storage space to the clients. Ceph is used to provide the storage space. Therefore, the Ceph storage system software is running on the storage nodes of the storage cluster. A short overview of Ceph is given in chapter "Ceph Overview" on page 28 and for a detailed description of Ceph, see the web page http://ceph.com/docs/master/architecture.

In the narrower sense, the ETERNUS CD10000 storage cluster consists of multiple storage nodes, a management node, a cluster network based on InfiniBand, the internal administration network based on 1Gb ethernet, and the Ceph storage system software. Each storage node is equipped with several disks. To each disk an Object Storage Device (OSD) daemon is assigned. Furthermore, monitor daemons on the storage nodes maintain the cluster state.

The storage nodes are connected via a fast 40Gb InfiniBand cluster network. The storage nodes and the management node are connected to a separate 1Gb ethernet administration network which is used for administration and monitoring purposes. For a detailed description of the networks, see "Networks" on page 33.

For detailed information on the mentioned components of a storage cluster, see "Components of a Storage Cluster" on page 24.

To manage and monitor the storage cluster central management and monitoring is provided, including such features as SNMP integration, central collection of log files, central software repository, etc. For a more detailed description see "Configuration and Administration of a Storage Cluster " on page 93, "Monitoring" on page 139, and "Log Manager" on page 197.

The figure below illustrates the architecture of an ETERNUS CD10000 storage cluster.

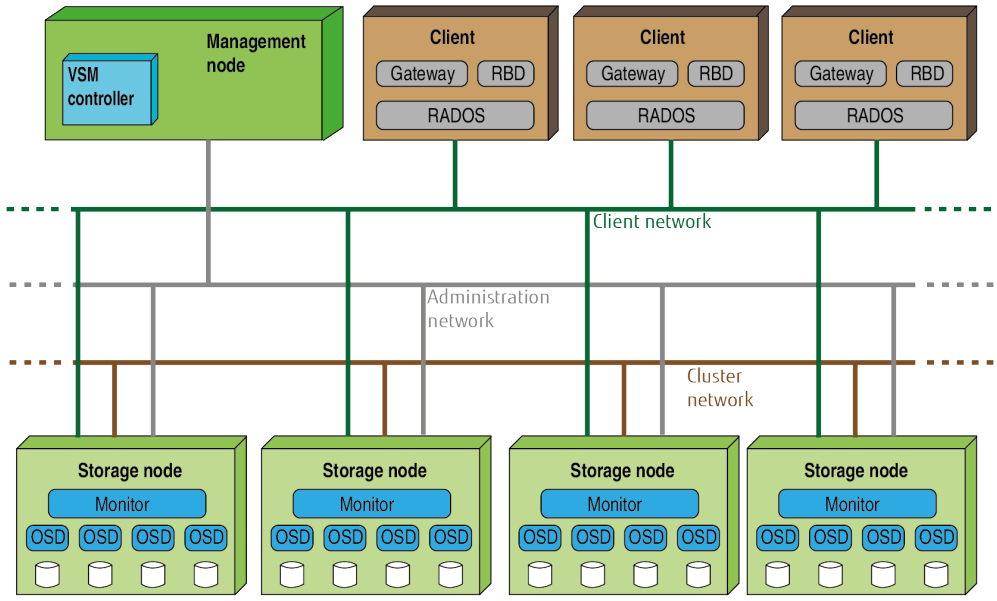


Figure 2: Architecture of an ETERNUS CD10000 storage cluster

2.5 Ceph Overview

Data storing and retrieving

The storage cluster stores or retrieves data as objects. "Each object corresponds to a file in a file system, which is stored on an Object Storage Device. Ceph OSD daemons handle the read/write operations on the storage disks." (http://ceph.com/docs/master/architecture/).

"Before Ceph clients can read or write data, they must contact a Ceph monitor to obtain the most recent copy of the cluster map." (http://ceph.com/docs/master/architecture/#crush-introduction).

An object consists of an object identifier and the object data. Both components are delivered by the user. The target storage node is calculated using the object identifier.Therefore no central directory is needed for storing metadata. This is the reason why the storage cluster can handle up to 224 storage nodes without a bottleneck.

According to the principles of storage tiering, frequently accessed data may be shifted to fast disks, so that it can be accessed quickly.

Monitors

At least three monitors are needed in one cluster.

The monitors maintain the cluster maps, these are the maps of the cluster state, including the monitor map, the OSD map, the placement group (PG) map, and the CRUSH map. The maps maintain a history of each operating state change in the Ceph monitors, Ceph OSD daemons, and PGs. These state changes are called epochs.

All monitors in a cluster synchronize with each other. Overall, a cluster must comprise an odd number of monitors, to be able to get a quorum. If one or more monitors fall behind the current state of the cluster, "Ceph always uses a majority of monitors (for example 1:2, 2:3, 3:5, etc.) " [...] "to establish a consensus among the monitors about the current state of the cluster" (http://ceph.com/docs/master/architecture/).

OSD daemons

An OSD daemon is assigned to each OSD in a storage node. It stores data, handles data replication, recovery, backfilling, rebalancing, and provides monitoring information to monitors by checking other OSD daemons. A storage cluster requires at least two OSD daemons to achieve an active and clean state when the cluster makes two copies of the data to be stored, as specified by default.

2.5.1 CRUSH Data Placement

Clients and OSD daemons use the CRUSH algorithm for on-demand data placement in the cluster.

CRUSH offers multiple advantages:

- Metadata computed on-demand: Therefore, no central directory is needed to determine the storage place of an object.
- No hotspots: Storage devices are assigned weights to control the amount of data they are responsible for storing. Data is uniformly distributed among these weighted devices.
- Dynamic adaption to infrastructure changes: Adding devices has no significant impact on data mapping.
- Infrastructure aware algorithm: Data placement is based on physical infrastructure,for example, devices, servers, etc.
- Easy and flexible placement rules
- Automatic and fast recovery of lost disks
- Composition of buckets to build a storage hierarchy

For detailed information on the CRUSH algorithm and its range of functions, see the web page http://ceph.com/docs/master/architecture/#crush-introduction.

The following concepts of CRUSH are elementary for the data placement in a storage cluster:

Levels of assignment

For storing objects Ceph uses the following three levels of assignment:

Pool

"Pools are logical partitions for storing objects." (http://ceph.com/docs/master/glossary/) A pool does not have a single physical location. It has the following attributes:

- Replicas: A pool determines the number of copies (replicas) of an object.
- Placement groups: A pool is defined by the number of its placement groups.
- CRUSH rules: When you store data in a pool, a CRUSH rule set mapped to the pool enables CRUSH to identify a rule for the placement of the object and its replicas in your cluster. By default three replicas of the object are written, one primary copy and two further copies. An expert user can modify the CRUSH rules.

Placement group (PG)

Each pool is logically divided into a number of placement groups. You can specify the number of placement groups before you create a pool. The assignment of an object to a placement group is computed by a hash function which is applied to the object identifier.

Object storage device (OSD)

A daemon is assigned to each OSD. The daemon is responsible for a specified data store, for example a disk. The mapping of an object and its replicas to OSDs is defined by the CRUSH map.

CRUSH maps

CRUSH maps contain two types of definitions:

Bucket types

A CRUSH map defines a hierarchical list of bucket types. The purpose of creating a bucket hierarchy is to segregate the leaf nodes by their failure domains, such as hosts, racks, rows, rooms, and data centers. Except the leaf nodes representing OSDs, the hierarchy is arbitrary.

Rules

CRUSH rules define the mapping of the replicas of an object from the associated placement group to OSDs. For example, you can define that the replicas of an object are stored in different racks or on different hosts.

Calculation of OSDs

The corresponding OSDs for the replicas of an object are calculated by the client using a hash function and the CRUSH rule that is defined for the assigned pool. The client can directly communicate with the OSD daemon storing the primary copy. Therefore a storage cluster can comprise hundreds of storage nodes without the risk of a bottleneck.

Example

The physical infrastructure consists of devices, servers, cabinets, rows in a data center, and different data center locations. A placement rule with 3 copies of each data object may state that the replicas must be placed in different cabinets in the same row.

2.5.2 Authentication

To avoid a centralized interface to the storage cluster, clients interact with OSDs directly. To protect data, Ceph provides its `cephx` authentication system, which serves for authentication of users operating clients. The `cephx` protocol operates similarly to Kerberos. Authorized users can read or write data and execute Ceph commands.

For details on cephx and how it differs from Kerberos, see the web page http://ceph.com/docs/v0.69/rados/operations/auth-intro/.

The connection between the client and the storage nodes is protected by this authentication. The authentication is not extended in excess of the client.

If you access the client from a remote host, Ceph authentication is not applied to the connection between your remote host and the client.

2.6 Administration and Monitoring Software

Additional software is installed for easily managing and monitoring the used hardware and software of the storage cluster. This chapter gives an overview of the software used for management and monitoring purposes. All functions are explained in more detail within this manual.

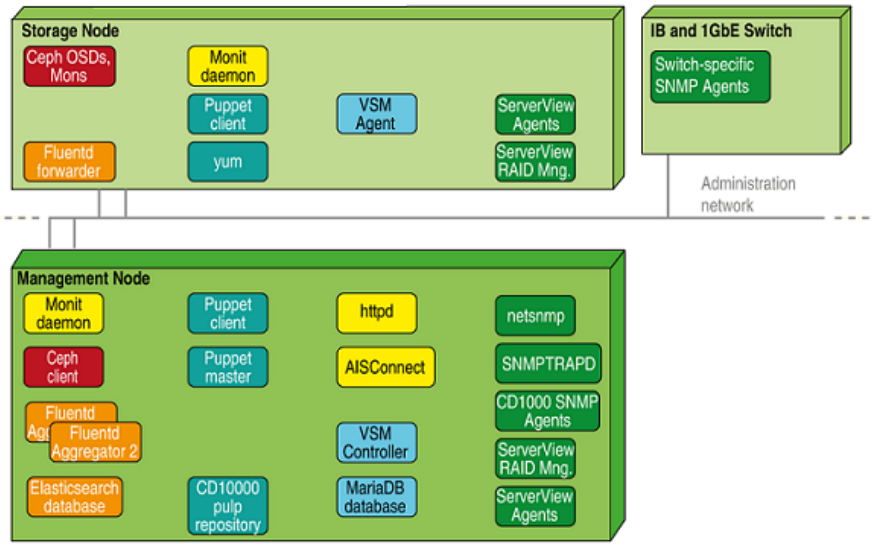


Figure 3: ETERNUS CD10000 software overview

The management node is the central point of administration and monitoring. On the storage nodes agents or client components are running that transfer information to the management node or that are used to configure the storage nodes.The used components and the corresponding functions that they cover are listed below:

Administration

- The configuration and update management system is provided for software installation and upgrade on all nodes of the storage cluster. For detailed information, see "Software Configuration and Update/Upgrade System" on page 185.
- The graphical user interface of the Virtual Storage Manager (VSM GUI) is provided for administration and monitoring purposes of the storage cluster. The administration tasks include cluster creation, adding nodes, creating pools, replacing disks, etc. Additionally, you can monitor the Ceph status of the OSDs, the nodes, the pools, etc. For detailed information, see "Virtual Storage Manager for Ceph" on page 57.

Monitoring

- SNMP monitoring: ETERNUS CD10000 is monitored via SNMP. For detailed information, see "Monitoring with SNMP" on page 169.
- Log file collection: ETERNUS CD10000 provides a centralized log manager. It collects configured log files from all storage nodes of a storage cluster in real-time and transfers them to the management node for supporting maintenance and support. For detailed information on the functions of the log manager, see "Functionality of the Log Manager"

on page 197.

- AISConnect: The storage cluster enables you to connect the management node to a Fujitsu Service Center. For detailed information on the functions of AISConnect, see "AISConnect" on page 222.

2.7 Networks

ETERNUS CD10000 needs different networks to work properly. They are separated into two types:

- the networks which connect the appliance to your existing infrastructure used to connect the clients (customer data network) and a possible administration network.
- ETERNUS CD10000 internal networks which are used as a cluster interconnect for data transfer and the internal administration/management network.

The figure below shows the architecture of a Ceph storage cluster and the corresponding networks.

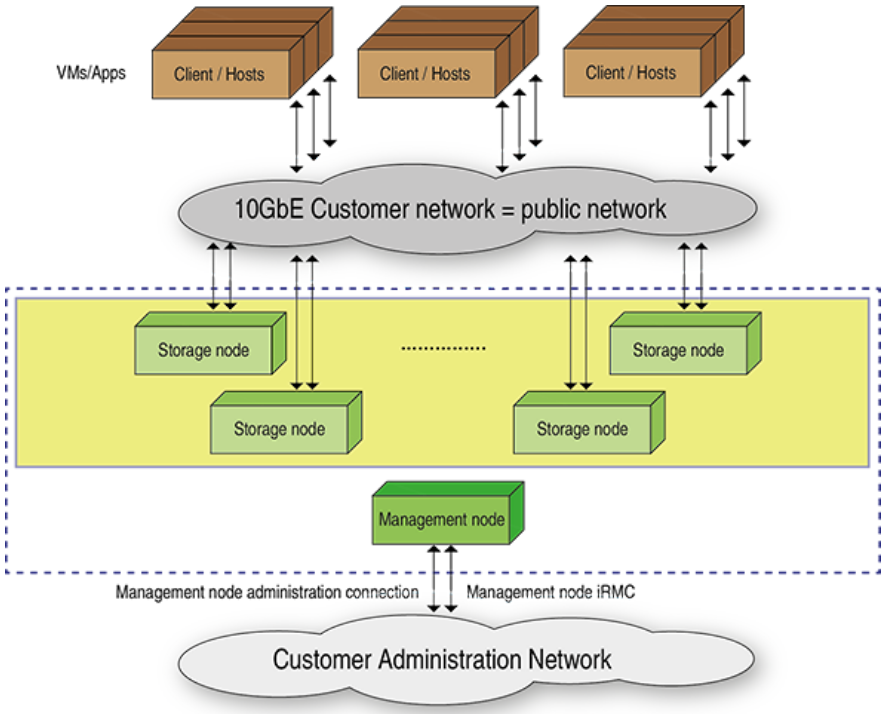


Figure 4: Architecture of a ETERNUS CD10000 network

2.7.1 Connection to Data and Administration Network

The most important network to operate in your environment is the connection to the clients. This network is called the "public network" in the Ceph context.

For this purpose each storage node has two 10GbE ports. These ports must be connected to your infrastructure. The way how the connection is done, depends on the needs in the existing infrastructure. Both ports might be connected to one 10GbE switch using bonding or the ports can be connected to different switches using bonding which ensures redundancy. The configuration of the public network is done via the management node, using the configuration management. The IP addresses used on the storage nodes must be provided by you. Only static IP addresses are allowed. The configuration of the public network must be done before the Ceph cluster is created. If an additional storage node is added, the node must be configured for the public network before it is added to the Ceph cluster.

The ETERNUS CD10000 management node is the central point of administration and monitoring. Therefore you must be able to access the management node from your existing network. For this purpose the management node is equipped with four 1GbE ports. After the management node is switched on the first time, you are asked to provide the IP address to be configured for the management node. The iRMC port of the management node can be connected to your network too.

2.7.2 ETERNUS CD10000 Internal Networks

There are different internal networks configured for the data transfer via the cluster interconnect and for administration and management purposes. These networks are completely separated from your environment. It is not allowed to connect the networks to your environment. The networks and the corresponding settings on the storage nodes and the management node are configured automatically during installation of the nodes.

The storage nodes are connected via a fast cluster interconnect based on InfiniBand. This network is called the "cluster network". With the redundancy configuration every storage node has two interfaces connected to the cluster network. The interfaces are bonded in an active-passive-mode. The non-redundant configuration has one port connected to the cluster network.

The addresses and names used in the cluster network are predefined and not changeable. The network 192.168.40.0/23 is used for this purpose.

The IP addresses for the storage nodes are ranged between 192.168.40.11 and 192.168.40.234. The names are "ibnode1" to "ibnode224".

Time synchronization is essential to ensure correct operation of the appliance. Therefore the ETERNUS CD10000 management node and three of the storage nodes are acting as NTP server. The management node should be configured in order to be able to access an appropriate NTP server in your network.

The storage nodes and the management node are connected to a separate 1GbE network which is used for administration and monitoring purposes. The iRMC ports of the storage nodes are connected to this 1GbE network too.

With the redundancy configuration every node has two interfaces connected to the 1GbE network. The interfaces are bonded using LACP. The non-redundant configuration has one port connected to the administration network.

The InfiniBand switches are connected to the internal administration network via the management port of each switch.

The figure below displays internal networks physical connections.

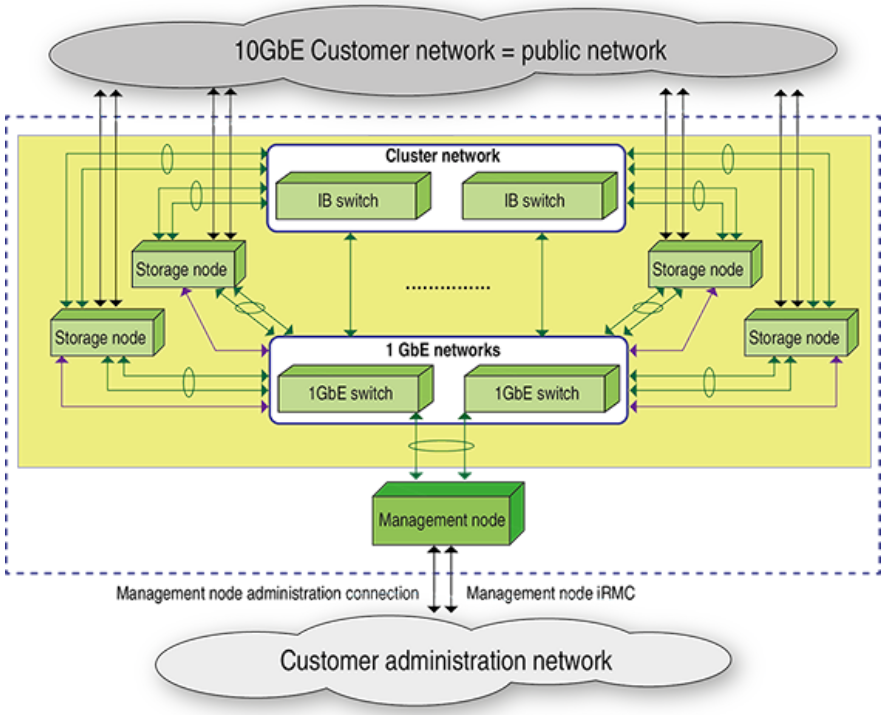


Figure 5: Internal networks physical connections

2.7.2.1 Network Configuration

Different VLANs are configured on top of the physical connections to provide separate networks for the needed services:

1. **cluster_admin_static network**

   This is the administration network between management node and storage nodes.

   a. Network: 192.168.20.0/23

   b. Management node: bond0.20

      IP address: 192.168.20.1

      Name: pmgmt

   c. Storage nodes: bond0

      IP address range: 192.168.20.11 - 192.168.20.234

      Name range: storage1 to storage224

2. **cluster_admin_dhcp network**k

   This is the DHCP network between management node and storage nodes. It is used for storage node installation with PXE and initial network setup.

   a. Network: 192.168.30.0/23

   b. Management node: bond0.20:0

      IP address: 192.168.30.1

      Name: pmgmt

   c. Storage nodes: when node is in preinst state, eth0 interface is used for connection, without bonding.

      IP address range: 192.168.30.11 - 192.168.30.234

      Name range: preinst1 to preinstl224

3. **cluster_irmc_static network**

   This is the network between the management node and the iRMC ports of storage nodes. It is used for iRMC connection to storage nodes from management node.

   a. Network: 192.168.10.0/23

   b. Management node: bond0.10

      IP address: 192.168.10.1

      Name: irpmgmt

   c. Storage nodes iRMC:

      IP address range: 192.168.10.11 - 192.168.10.234

      Name range: irstorage1 to irstorage224

4. **cluster_irmc_dhcp network**

   This is the DHCP network between the management node and the iRMC ports of the storage nodes. It is used for iRMC connection for storage nodes in cluster_admin_dhcp network.

   a. Network: 192.168.0.0/23

   b. Management node: bond0.10:0

      IP address 192.168.0.1

   c. Storage nodes: iRMC

      IP address range: 192.168.0.11 - 192.168.0.234

Whenever an additional storage node is added to the cluster the appropriate network settings for the internal networks are configured during first boot after the software installation.

2.8 Available User Interfaces for Administration and Monitoring

ETERNUS CD10000 consists of the hardware for the nodes and the network infrastructure. The hardware components are configured and monitored by the provided software stack. On top of this the Ceph software provides the storage cluster.

The web-based graphical interface Virtual Storage Manager (VSM) and the command line interface vsm_cli support all Ceph related aspects. This includes the ability of creating, configuring, administering and monitoring of Ceph storage clusters.

Another web-based graphical interface to manage and monitor the hardware components and the network configuration is provided.

The following tasks are provided by VSM and vsm_cli:

- creating and configuring a Ceph storage cluster
- managing storage groups
- creating and managing pools within the storage cluster
- managing the storage nodes and disks
- managing the storage cluster's daemons (OSDs and monitor daemons)
- monitoring of the cluster capacity and performance
- monitoring of the cluster health and the monitor status

For detailed information on the VSM see "Introduction to the User Interfaces" on page 57

The command line interface `vsm_cli` provides additional functions for the following tasks:

- deleting storage pools and modifying pool parameters
- stopping and restarting storage cluster daemons and the whole storage cluster
- replacing failed/retired disks
- changing the Crush weight of an OSD
- transferring storage nodes into maintenance mode

A complete list of all available options is given in "Appendix A: VSM Specific Technical References" on page 325

Beside these two interfaces which support all Ceph related aspects a web-based graphical interface to manage and monitor the ETERNUS CD10000 hardware and network components is provided.

3 Getting Started

Before the ETERNUS CD10000 is operational, you must configure the management node and the cluster networks through the initial configuration wizard.

After these steps the Ceph cluster can be created either with VSM or with vsm_cli. Cluster creation is a one time task.
Enhancing the cluster is done by adding storage nodes, creating pools, creating storage groups and adding additional Ceph monitors.

3.1 Initial Configuration of the Management Node

Before the ETERNUS CD10000 is operational, you must configure the management node through the initial configuration wizard.

This chapter describes the initial configuration of the ETERNUS CD10000. The steps are executed on the management node and include following:

- Licence approval
- Client networks configuration
- Users' password setup (for root on management node & storage nodes)
- Keyboard, language, date & time configuration
- SNMP initial configuration
- Enabling AISConnect

Requirements

1. Do necessary cabling before starting.
2. The management node must be connected to your administration network to work properly. Connect at least one of the 1GbE ports to this network. Before you start configure and connect the iRMC port to your network. This is needed to start the initial configuration.
3. Press the power button of the management node. Use **iRMC** or **KVM** console to display the GUI of CentOS.

3.1.1 Welcome

ETERNUS CD10000 Welcome page displays (see figure below).

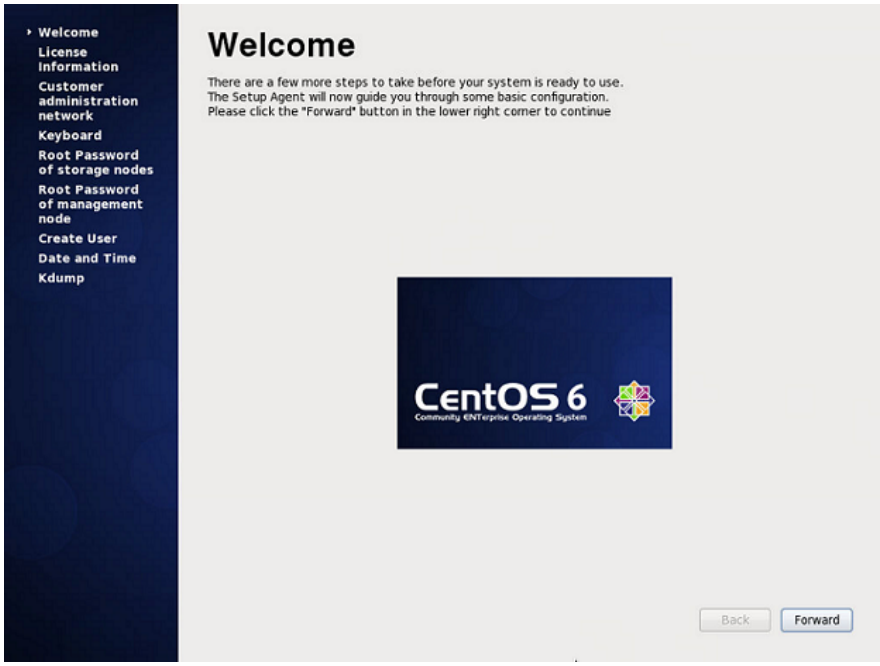


Figure 6: ETERNUS CD10000 welcome page

3.1.2 Approve Licence

Accept the licence of CentOS. Click **Yes, I agree to the License Agreement** (see figure below).

Figure 7: CentOS license agreement page

3.1.3 Customer Administration Network

1. Set up of the client administration network to which management node is connected.
2. Set DHCP or static configuration according to your network needs.

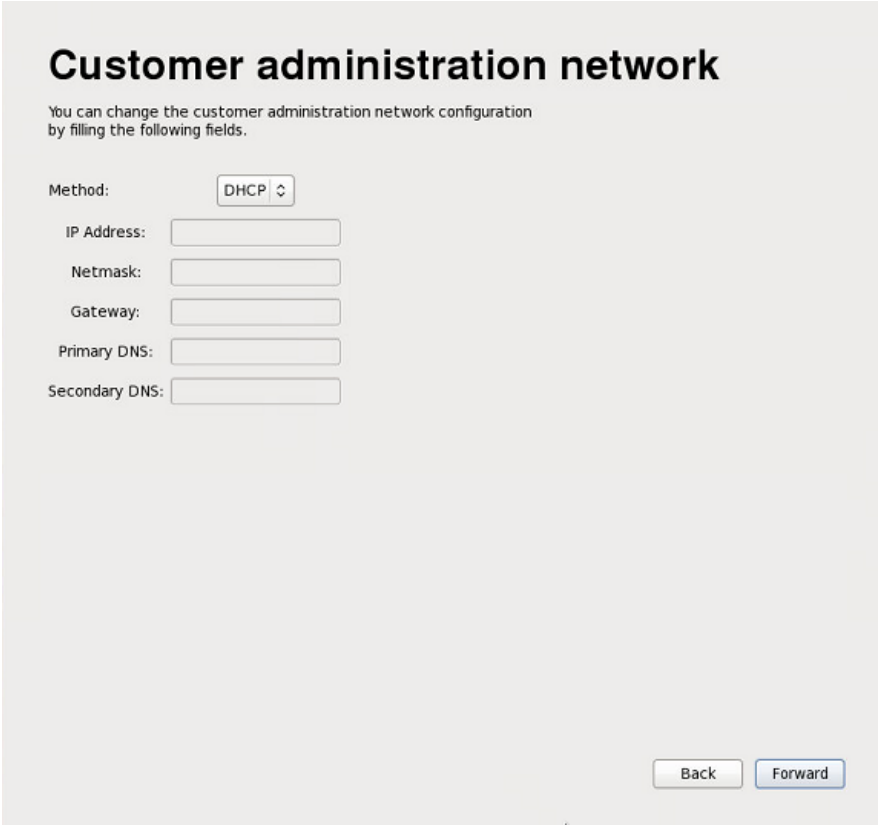


Figure 8: Customer administration network

3.1.4 Keyboard

Set up a keyboard language.

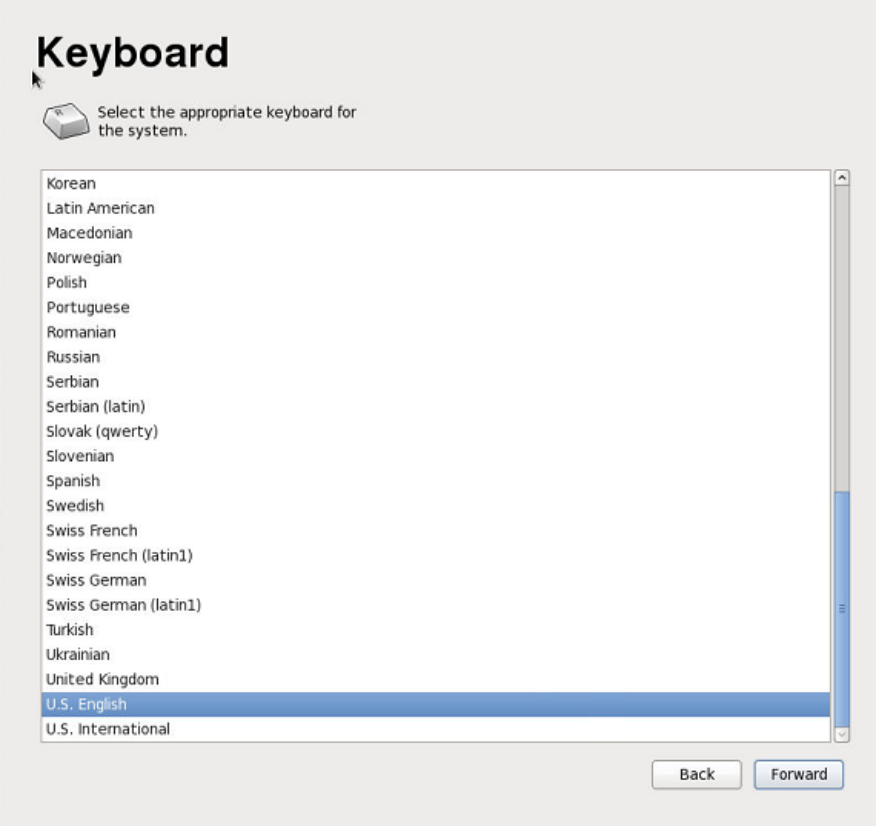


Figure 9: Keyboard language

3.1.5 Root Password

Provide a root password for the storage nodes. This password is copied to all storage nodes when the public network for the storage nodes is configured.

The root password for the storage nodes can be changed at any time with the following command:

```
cd10000 setup storage_root_password -p <new password>
update-manager update
```

Set a secure root password. (Secure password in terms of complexity and avoiding passwords based on dictionaries.) The password has to be kept secret.

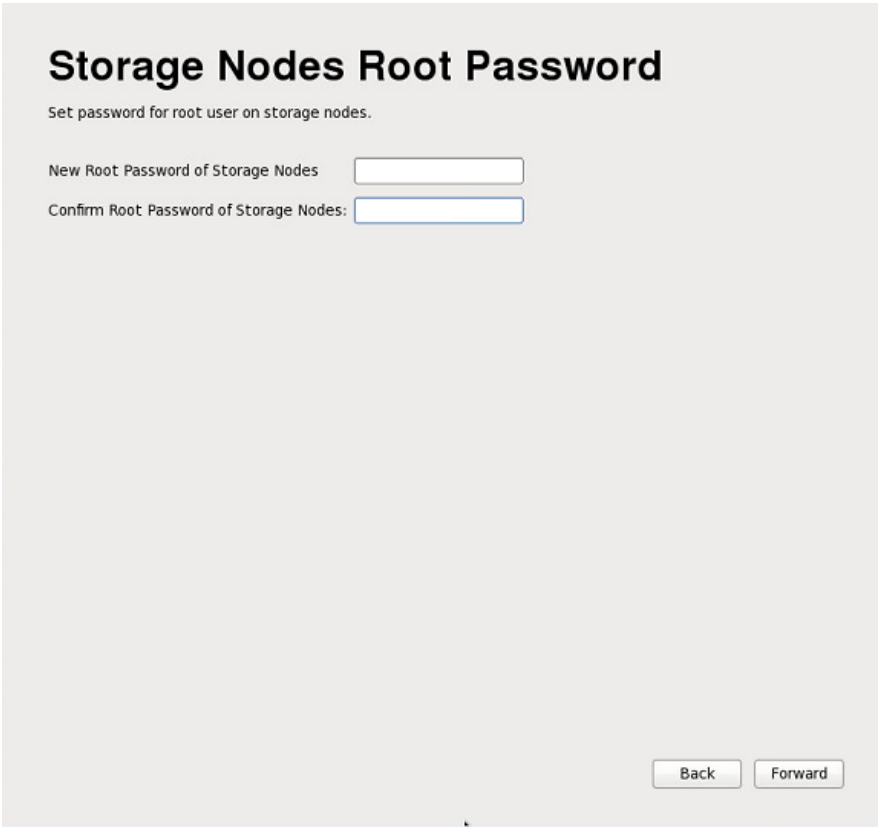


Figure 10: Storage node root password

Provide a root password for the management node.

You can change the management node root password at any time with the standard Linux command:

```
passwd
```

Set a secure root password. (Secure password in terms of complexity and avoiding passwords based on dictionaries.)
The password has to be kept secret.

Figure 11: Management node root password

3.1.6 Create Additional Users

If needed, create an additional user for the management node. You can add additional users at any time with the standard Linux commands if you create local users.

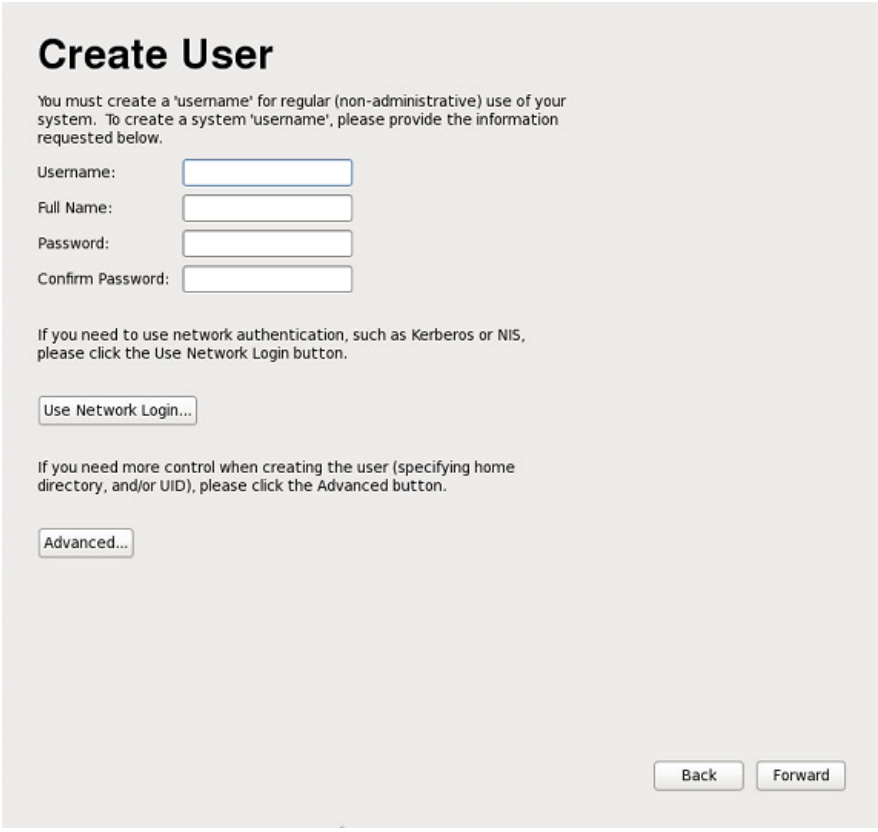


Figure 12: Create an additional user

3.1.7 Date and Time Settings

Set up date and time settings. It is recommended to use NTP server.

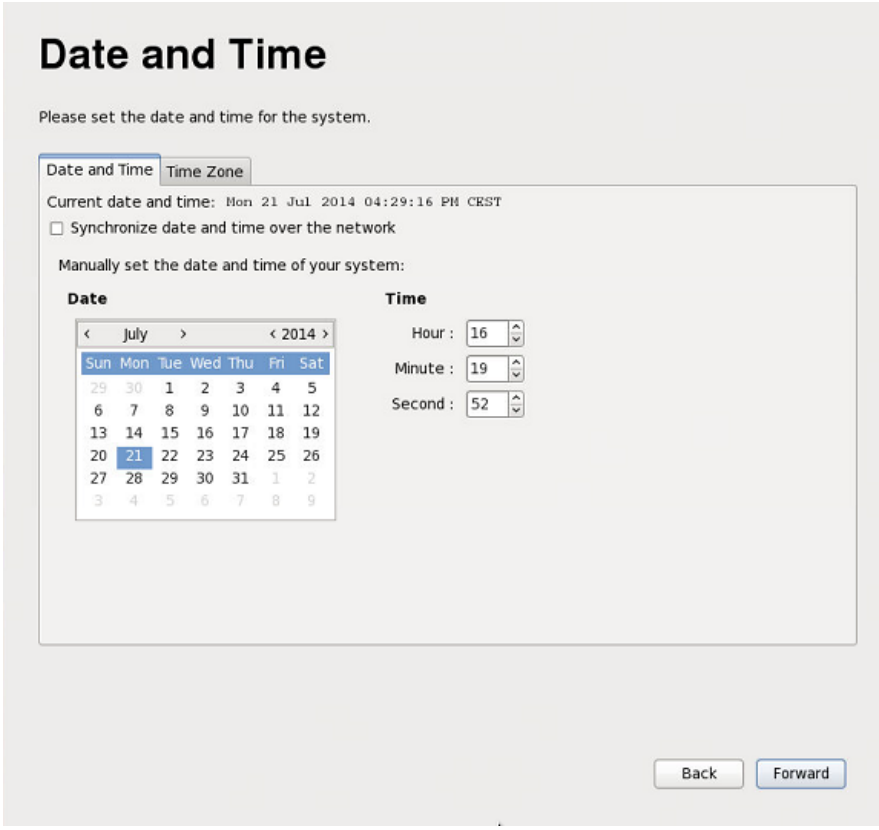


Figure 13: Date and time settings

3.1.8 Kdump

Enable or disable Kdump.

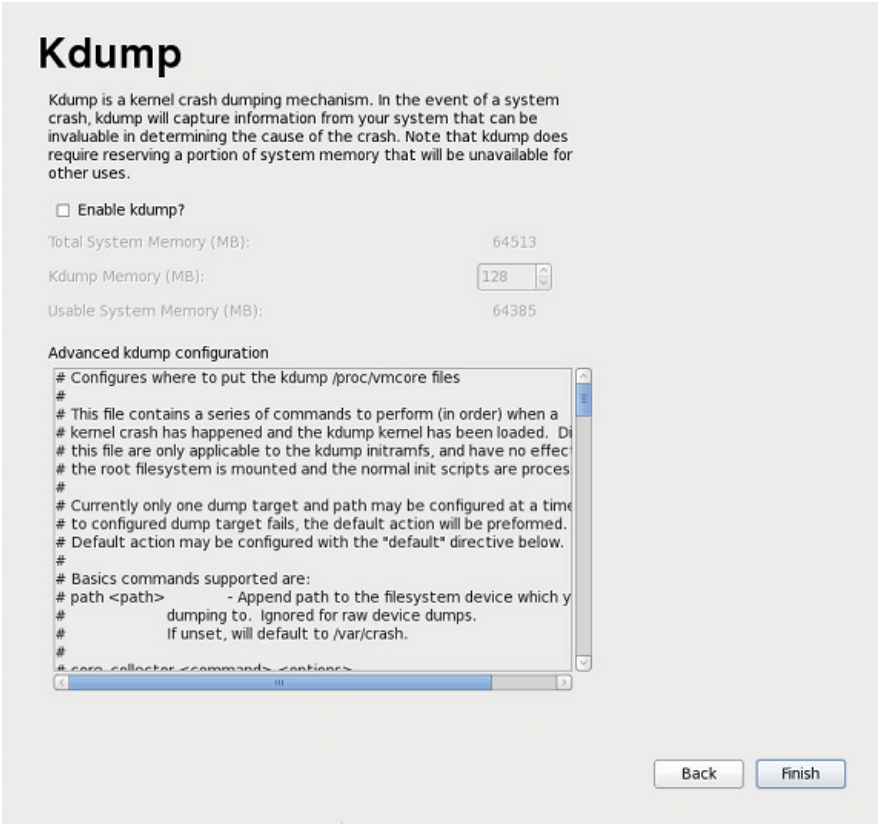


Figure 14: Kdump

3.2 Configure Backup of Management Node

ETERNUS CD10000 is equipped with one management node. If this management node must be replaced, a backup of the system should be available.

For this purpose the Fujitsu Backup and Restoration Tool (**`fbart`**) is created. It provides backup and restoration of the management node of the Fujitsu ETERNUS CD10000. Multiple backups according to a predefined schedule are created automatically.

The backups include all the configuration data of the management node which are needed for a complete restoration.

- Execute the following command to configure the backup of the management node:

```
[root@pmgmt ~]# fbart config

~~~~ Fujitsu Backup And Restoration Tool ~~~~
This script will guide you through the configuration process.
Current values are between [ ].
To keep them unchanged press Enter.
Backup Host N.1: [storage1]
Backup Host N.2: [storage2]
Backup Frequency in hours: [8]
Backup expiration delay in days: [7]
Destination folder on hosts: [/backup]
Source folders to back-up: [Defaults + []]
Excludes-multiple excludes should be separated with a comma:[*.tmp, *~]
[INFO] Configuration file wrote to disk.
[INFO] Updating cron ...
[INFO] Cron successfully updated.
[INFO] Configuration process finished!
```

The displayed values are the default values which are sufficient for a proper backup.

To change a value, override it and press the **Enter** key otherwise just press **Enter**.

For a detailed description of the **`fbart`** application, see chapter "Backup and Restoration of Management Node " on page 133.

All the VSM configuration and information on the management node is part of the backup. All automatically started backups will fail as long as the Ceph cluster is not configured.

3.3 Setting a Public Network

The storage nodes need a connection to your client network, in ETERNUS CD10000 called the public network. See chapter "Connection to Data and Administration Network" on page 34.

The public network is used by the Ceph monitors for internal communication. Therefore existing monitors are not supposed to change their IP addresses, i.e. change the addresses of the public network.

Execute the following steps to add the storage nodes to the public network:

1. Log in to the management node console as root user.
2. Execute the following command to setup the public network:

   ```
   cd10000 setup public_network -n <NETWORK> [-s <START>]
   [-l <LIMIT>] [-f] [-d <DNS1> <DNS2>] [-g <GATEWAY>] -t <TEMPLA
   TE>
   ```

   When adding new nodes later on these nodes will dynamically apply the same configuration there is no need to run the command for every new storage node.

   In the command above:

   NETWORK is a public network (e.g. 192.168.80.0/24),

   START (optional) is the starting address (default 11),

   LIMIT (optional) is the reserved count for storage nodes within the public network,

   the option **-f** overwrites an existing configuration,

   DNS1, DNS2 are additional parameters to configure DNS servers in a public network,

   GATEWAY (optional) is an additional parameter to configure the gateway in a public network.

   TEMPLATE is one of the following values:

bonding-active-backup	Only one slave in the bond is active. A different slave becomes active if, and only if, the active slave fails. The bond's MAC address is externally visible on only one port (network adapter) to avoid confusing the switch. This mode provides fault tolerance. The primary option affects the behavior of this mode.

bonding-balance-xor	Transmission is based on [(source MAC address XOR'd with destination MAC address) modulo slave count]. This selects the same slave for each destination MAC address. This mode provides load balancing and fault tolerance.
bonding-broadcast	Transmits everything on all slave interfaces. This mode provides fault tolerance.
802.3ad	IEEE 802.3ad Dynamic link aggregation. Creates aggregation groups that share the same speed and duplex settings. Utilizes all slaves in the active aggregator according to the 802.3ad specification.Prerequisites: • Ethtool support in the base drivers for retrieving the speed and duplex of each slave. • A switch that supports the IEEE 802.3ad Dynamic link aggregation. Most switches will require some type of configuration to enable the 802.3ad mode.
one-if-static	First interface is up and statically configured second is down (no fail over).

If step 2 was not performed the following has to be done:

1. Log in to management node console using root user.
2. Run command: **update-manager update** and wait until it completes.
3. Setup client data network manually on every storage node on interfaces eth2 & eth3. Networks 192.168.0.0/23, 192.168.10.0/23, 192.168.20.0/23, 192.168.30.0/23, 192.168.40.0/23 are forbidden to use since they are as internal CD10000 networks.
4. Setup routing on management node using the following command:
   ip route add 192.168.80.0/23 via 192.168.20.11 dev bond0.20 metric 0
   ip route add 192.168.80.0/23 via 192.168.20.12 dev bond0.20 metric 1
   ip route add 192.168.80.0/23 via 192.168.20.13 dev bond0.20 metric 2
   where address **192.168.80.0/23** is public network for CEPH and the "via" routes are the storage nodes designated to run the CEPH monitors. Step 2 will automatically add the routes to the first three storage nodes.

3.4 SNMP Initial Configuration

It is possible to monitor the ETERNUS CD10000 storage cluster via SNMP V2. All SNMP traps generated on the storage nodes or switches of the storage cluster are sent to the management node (see chapter "Monitoring with SNMP" on page 169). To forward the traps from the management node to your SNMP management station you have to execute the following steps:

1. Log in to management node console as root user.
2. Specify the IP address of your SNMP management station where traps will be sent by invoking the following command:

Syntax:

```
cd10000 setup snmp set-destination -i <ip-address>
```

Output:

After the command has been executed successfully the following is displayed:

```
Stopping snmptrapd: [ OK ]
Starting snmptrapd: [ OK ]
```

Example:

```
[root@pmgmt ~]# cd10000 setup snmp set-destination -i 10.172.44.97
Stopping snmptrapd: [  OK  ]
Starting snmptrapd: [  OK  ]
```

You can setup or change the SNMP configuration at a later time too.

3.5 AISConnect

To enable the connection to the AIS server in the Fujitsu Support Center to provide call-home and dial-in services, execute the following steps:

1. Log in to management node console using root user.
2. Follow the AISConnect Manual to set up AISConnect.

For detailed information, see "AISConnect" on page 222 and "Configuration of AISConnect" on page 226.

You can setup the AISConnect at a later time too.

3.6 Creating a Ceph Cluster

3.6.1 Preparing the Cluster Creation

Some information is needed about the Ceph storage clusters which shall be created. Therefore you must provide configuration files in which you describe the cluster configuration and the configuration of the particular nodes.

The configuration and administration of Ceph storage clusters via VSM or vsm_cli requires the following manifest files:

- Cluster manifest file on the management node (/etc/manifest/cluster.manifest)

  This file contains the global configuration parameters for the cluster.

  Such as the Storage Classes (performance, capacity) with its Storage Groups (the performance group is mandatory as first entry), the network configuration and other management details.
- Server manifest file on each storage node (/etc/manifest/server.manifest)

  This file identifies all storage devices and associated journal partitions on the node.

Templates of these files are created during the setup of the storage nodes. The following default values are configured:

- Cluster.manifest filled in with default cluster name (cd10000), administration, cluster and public network. Additionally storage groups will be organized in the following order:
  - Performance group
  - Capacity group
  - High performance group
- Server.manifest copied from /etc/fcephadm/server.manifest to /etc/manifest/server.manifest. The first 3 storage nodes will be enabled as Ceph monitor hosts in the manifest.

At this point the configuration files are ready for the next steps in the creation process. Whenever new nodes will be added they will be automatically updated with the similar configuration. At this point you can modify the configuration of the cluster.manifest and/or the server.manifest(s). For example configure monitors on other nodes than storage1, storage2, storage3.

VSM daemons must be registered and started on the management node and the storage nodes in the next step before you can create the Ceph cluster. To register nodes in VSM the following command can be executed:

- cd10000 setup cluster [-all] [-n NODES [NODES ...]]

The command executes the following tasks:

- VSM controller initialization (vsm-controller -r controller) on the management node. If the command was run for the second time it will check if vsm-controller is setup and progress with storage nodes registration
- Generate a token and update server.manifest on chosen storage nodes with this token.
- Run the vsm-node command on the specified storage nodes.
- At the end the tempory password for admin login to VSM is displayed. This password should be changed in VSM after first login.

At this point you can login to VSM and create a cluster if at least 4 storage nodes are available.

For a detailed description of the configuration files and how to register and setup the VSM daemons manually see chapter "Appendix C: Error Codes & Messages" on page 343.

3.6.2 Creating a Storage Cluster (VSM GUI)

To create a storage cluster, you must have executed the steps described in chapter "Setting a Public Network" on page 50 or must have done the manual steps described in chapter "Appendix C: Error Codes & Messages" on page 343

To create a storage cluster using VSM follow step-by-step the following instructions:

- Login to VSM with user account **admin.**
- In the navigation bar select from the main menu **Cluster Management** the **Create Cluster** menu command.

  If no cluster exists, you will be redirected to the **Create Cluster** page immediately.

The **Cluster Server List** will be displayed. It contains all available servers which are recognized by the VSM controller i.e. where the vsm-node was started.

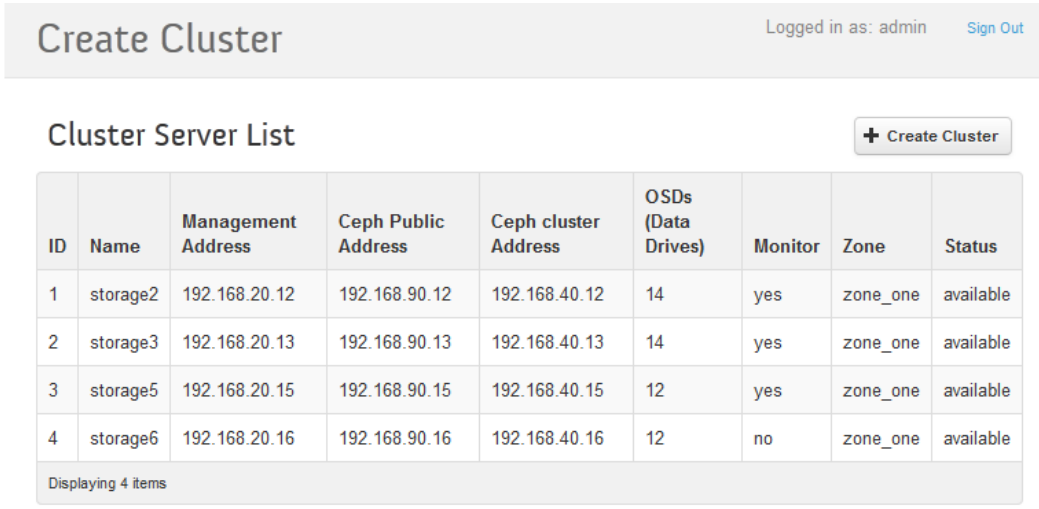


Figure 15: The Cluster Server List

The host names and the IP addresses of the servers in the different subnets (administration, public and cluster network), the number of disks within each server and the status of the servers are displayed.

Additionally the list shows the servers on which a monitor daemon shall be created (**Monitor** column). At least three monitor daemons must be created. By default the monitor daemons will be created on the first three servers in the list.

For detailed information about the displayed properties see "Managing Storage Nodes (VSM GUI)" on page 112.

After the cluster creation you can change the servers on which monitor daemons reside, furthermore you can enlarge the number of servers with monitor daemons (see "Adding monitors to the cluster " on page 115 and "Removing monitors from the cluster " on page 116).

- Check whether all servers, which shall be included in the cluster, are displayed and their properties are correct. The list contains the available nodes (connected and running) on which a VSM agent resides.
- If all servers are available (see **Status** column) and everything is ok, click the **Create Cluster** button.
- Confirm the action in the **Create Cluster** dialog box to create the cluster.

The **Status** column displays the progress of the creation operation.

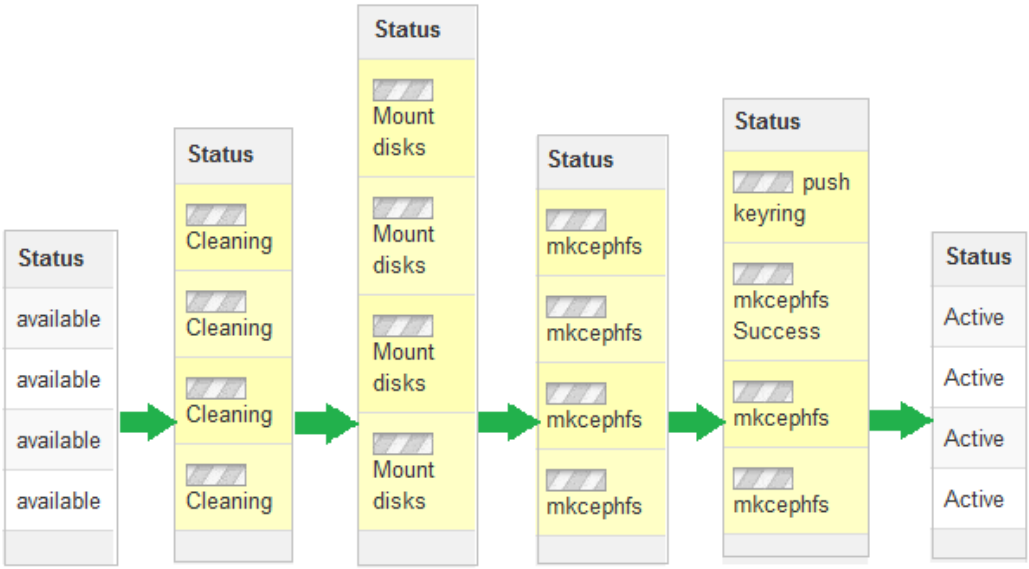


Figure 16: Servers Add Success

The storage cluster will be ready when **Status=Active** is displayed for all servers. After the successful creation the dashboard displays a summary of the cluster status.

For newly installed storage nodes the following command should be run to register this node to VSM:

```
cd10000 setup cluster -n storage_node_hostname
```

For a detailed description of the configuration files and how to register and setup the VSM daemons manually see "Appendix C: Error Codes & Messages" on page 343

4 Introduction to the User Interfaces

4.1 Virtual Storage Manager for Ceph

The ETERNUS CD10000 storage cluster provides the web-based user interface Virtual Storage Manager for Ceph (VSM) to manage and monitor the storage cluster. Via the comfortable VSM grafical user interface (VSM GUI) you can perform many configuration, administration, and monitoring functions.

Functionality

VSM offers the following fundamental functions:

- Organization and management of the storage nodes and disks
- Storage cluster creation and pool management
- Monitoring of the cluster capacity and performance, the underlying Ceph daemons, and data elements
- Automatic connection of RBD pools to OpenStack

For an overview of all VSM functions, see "Functions of the VSM GUI" on page 60.

Architecture

The VSM controller software is installed and running on the ETERNUS CD10000 management node. On each storage node a VSM agent is installed and running. The VSM controller is connected to the storage cluster via the VSM agents. The VSM agents relay server configuration and status to the controller.

4.1.1 VSM Main Window

After you have logged in to the Virtual Storage Manager for Ceph (VSM) or after you have successfully created the storage cluster, the main window, that is the **Dashboard** overview, is displayed in your browser.

It shows the **Cluster Status**, i.e. several summaries concerning the different cluster components:

- **Cluster Summary**,
- **Storage Group Summary**,
- **Monitor Summary**,
- **OSD Summary**,

- **PG Summary**,
- **Cluster Health Summary**.

If you click one of the **Status** buttons, the **Status** window of the appropriate cluster component is displayed.

Furthermore, **Warnings and Errors** are displayed for servers that are not synchronized with NTP.

The figure below shows the main window of the VSM GUI.

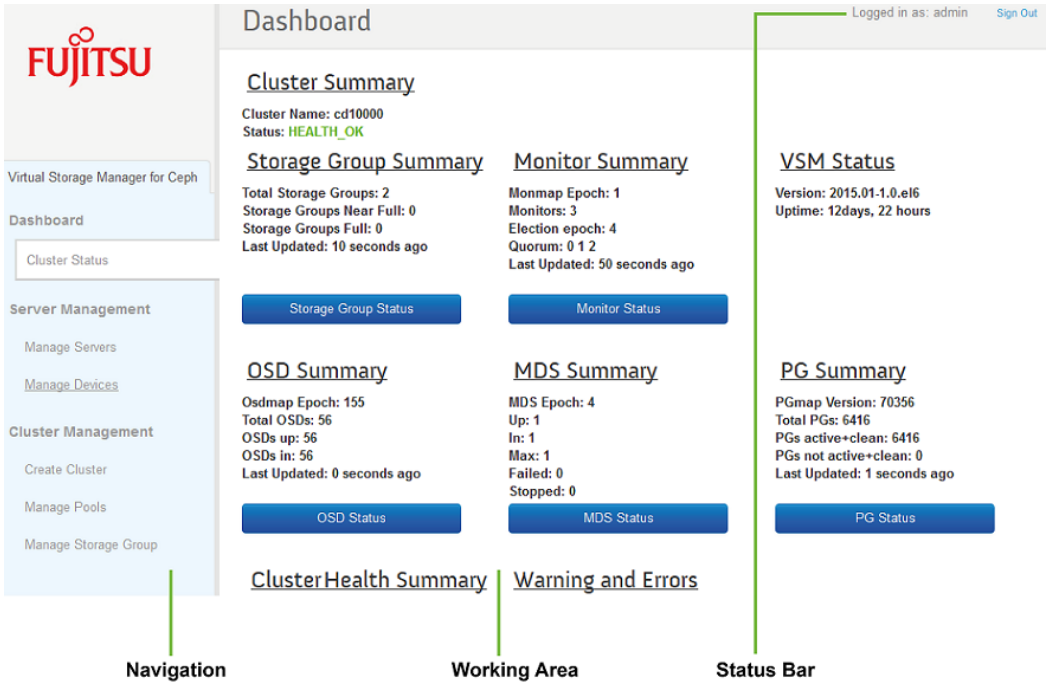


Figure 17: Main window of the VSM GUI

Structure of the VSM GUI

The VSM GUI contains different areas that are listed below:

Status bar

This area on top of the main window shows the following information:

Cluster	Name of the active storage cluster.
Logged in as	Login name of the currently active user.
Sign Out	Clicking on **Sign Out** logs you off the VSM GUI.

Navigation

The navigation on the left side of the main window provides access to all functions for configuring, administrating, and monitoring the storage cluster via six main menus and one or more appropriate menu commands. Triggering the individual functions as well as data output take place in the working area.

Working area

In the working area of the VSM GUI you can configure and administrate the storage cluster including server management and cluster management. Depending on the type of action, that you want to trigger, the GUI provides appropriate buttons or opens corresponding dialog boxes. Additionally you can display status data for cluster monitoring.

For a list of all functions that the VSM provides, see "Functions of the VSM GUI" on page 60.

4.1.2 Functions of the VSM GUI

In the navigation the commands are arranged according to the various tasks of configuration and administration.

Main menu	Menu command	Functions available in the working area
Dashboard	Cluster Status	Displays an overview of the cluster status
Server Management		Management of the cluster hardware
	Manage Servers	• Opens the cluster server list with all storage nodes that are recognized by the VSM controller. • Enables adding or removing servers or monitors, starting or stopping servers.
	Manage Devices	• Opens the device management list. • Enables restarting, restoring, or removing storage devices.
Cluster Management		Management of the resources of the storage cluster
	Create Cluster	• Displays the cluster server list with all present servers. • Enables creating a new storage cluster if a storage cluster does not yet exist.
	Manage Pools	• Displays a list of all storage pools in the storage cluster. • Enables creating a new storage pool.
	Manage Storage Group	• Displays all storage groups • Enables creating a new Storage Group
Monitor Cluster		Monitoring the storage cluster. For detailed information, see "Monitoring with VSM GUI" on page 139.
	Storage Group Status	• Visualizes used and available capacity of each configured storage group with a pie chart. • Lists the configured storage groups with their status.
	Pool Status	• Lists the existing pools with their assigned values.

Main menu	Menu command	Functions available in the working area
	OSD Status	• Displays an OSD summary with cluster-specific data. • Lists each individual OSD with its status.
	Monitor Status	• Displays a monitor summary with cluster-specific data. • Lists each individual monitor with its status.
	MDS Status	Status of MDS processes. Not applicable in ETERNUS CD10000 V1.0 SP2, as Ceph file system is not supported.
	PG Status	• Displays a PG summary with cluster-specific data. • Lists each individual PG with its status.
	RBD Status	Lists the existing configured Rados Block Devices (RBD), which are configured in the different pools, with their status.
Manage OpenStack		Managing OpenStack interconnection
	Manage RBD Pools	Enables attaching an RBD pool to an OpenStack configuration. For detailed information, see "Attaching an RBD Pool to OpenStack" on page 126.
	Openstack Access	• Lists the configured OpenStack access points. • Enables creating, changing, and deleting OpenStack access points. For detailed information, see "Managing OpenStack Access" on page 127.
Manage VSM	Add/Remove User	Enables creating or removing users and changing a password. For detailed information, see "Managing VSM Users " on page 128.
	Settings	Set the thresholds **full** and **near full** for storage groups.

For detailed information on the menu commands of the **Server Management** and **Cluster Management** main menus and their parameters, see the sections:

"Creating a Storage Cluster (VSM GUI)" on page 54,

"Managing Pools (VSM GUI)" on page 100,

"Managing Storage Groups (VSM-GUI)" on page 118

"Managing Storage Nodes" on page 112

For detailed information on the monitoring information output (**Status** commands of the **Monitor Cluster** main menu), see "Monitoring with VSM GUI" on page 139.

4.1.3 Accessing VSM

Prerequisites

Browser software requirements

VSM requires Google Chrome or Internet Explorer 10, or another modern Web-browser (tested with Google Chrome and Internet Explorer 10 with compatibility mode turned off).

Accessing VSM

- Start the browser on the VSM controller.
- Enter the VSM Url into the address field.
- In the **User Name** field, type in `admin` as the default user.
- In the **Password** field, type in the first time password.

  The first time password, ADMIN_PASSWORD, is present on management node in the file `/etc/vsmdeploy/deployrc`.

After that you can start to create a storage cluster or to manage and monitor the existing storage cluster.

First time login is possible after the VSM manifest files (cluster.manifest file and server.manifest file) are available and the VSM services (VSM controller and VSM agents) are running on the nodes.

For detailed information, see "Preparing the Cluster Creation" on page 53 and "Preparing VSM Cluster Creation Manually" on page 233.

Logging out from VSM

- Click **Sign out** in the upper right corner of the status bar.

4.2 Overview of the vsm_cli Commands

This overview shows all commands of the command line interface (CLI) **vsm_cli**. In the following lists the commands are arranged according to the various tasks of administering, maintaining and monitoring a storage cluster:

Administering

The administrative **vsm_cli** commands destroy or alter the configuration of a cluster. The following administrative **vsm_cli** commands are supported:

Command	Description
activate	Activating a new cluster that has been prepared by using the **prepare-disk** command.
add-disk	Adding new disk(s) to an existing storage cluster
prepare-disk	Preparing disks that shall be used to create a new cluster with the **activate** command.
remove-disk	Removing a disk from an existing storage cluster
add-mon	Creating a new monitor daemon on a storage node and adding this daemon to the storage cluster
remove-mon	Stopping a monitor daemon and removing it from the storage cluster.
create	Creating a new cluster. The **create** command combines the functionality of the **prepare-disk** and the **activate** command.
create-cache-tier	Converting two existing pools (a storage pool and a cache pool) into a tiered pool.
remove-cache-tier	Removing cache tiering associated to a specified cache pool.
create-ec-pool	Creating a pool that uses erasure coding instead of replication.
create-pool	Creating a new storage pool.
set-pool	Modifying the number of object replications within a pool and the pool quota size.
delete-pool	Deleting a pool from a storage cluster and destroying the data objects stored on that pool.

Command	Description
destroy	Destroying a storage cluster, i.e. all OSD and monitor daemons will be stopped and all disks associated with this storage cluster will be unmounted. If you execute this command all data stored in the cluster is lost. Do not use the command in productive configurations.

Maintaining

The **`vsm_cli`** maintenance commands allow to take down the whole cluster or parts of it safely, furthermore to replace disks, to adjust capacity utilization, etc. The following maintenance commands are supported:

Command	Description
crush-reweight	Adjusting capacity utilization by increasing or reducing the CRUSH weight of an OSD (disk)
maint-on	Switching the specified storage node into maintenance mode, i.e. all daemons on the storage node will be stopped and storage cluster rebalancing will be suppressed.
maint-off	Switching the maintenance mode off.
replace-disk-out	Substituting a new disk for a failed/retired disk keeping the overall storage cluster configuration intact
replace-disk-in	
stop	Stopping and restarting a whole storage cluster or a particular daemon (OSD or monitor) of a storage cluster
start	

Monitoring

Monitoring commands are informational and do not alter the cluster in any way. The following monitoring **vsm_cli** commands are supported:

Command	Description
list-cluster-names	Requesting the names and subnet addresses of all clusters which are configured
status	Requesting information on the storage cluster status and on critical situations at its instances
mon-status	Requesting information on the status of a storage cluster's monitor daemons
list	Requesting information on storage clusters and on the storage cluster instances

For detailed information about monitoring with **vsm_cli** see "Monitoring with vsm_cli" on page 156

Help

Command	Description
help	Displaying the list of all **vsm_cli** commands with their syntax. If you enter help together with the name of a vsm_cli command, then detailed help for that command is displayed.

For further administration, maintenance and monitoring functions, see section "Functions of the VSM GUI" on page 60.

The syntax of the **vsm_cli** commands is described in section "The Command Line Interface vsm_cli" on page 247

4.3 Basking GUI

4.3.1 Overview

Basking is a graphical interface for managing the ETERNUS CD10000 using the web browser. It provides functions of several command line applications for viewing data, and performing changes on cluster or nodes.

To access the Basking GUI enter the followig URL in the address field of your web browser:

`https://hostname/basking`

List of permissions for each role

- Monitor – basic role with read only permissions
  - Monitor can check versions of installed software. See chapter "Checking Version" on page 72
  - Monitor can perform diagnostic tests. See chapter "Checking Status" on page 73.
  - Monitor can see cluster settings and SNMP settings. See chapter "Managing Settings" on page 75
  - Monitor can check information about nodes. See chapter "Managing Nodes" on page 79.
  - Monitor can check information about switches. See chapter "Managing Switches" on page 84.
  - Monitor can see uninstalled nodes. See chapter "Managing Installation" on page 85.
  - Monitor can see history of user's activity. See chapter "Users Activity" on page 86.
- Appliance admin – administrator of cluster and nodes. The appliance admin can perform all operations on cluster and nodes, can see user's activity, but cannot manage users or delete user's activity.
  - Appliance admin can check versions of installed software. See chapter "Checking Version" on page 72
  - Appliance admin can perform diagnostic tests. See chapter "Checking Status" on page 73.
  - Appliance admin can see cluster settings and SNMP settings. See chapter "Managing Nodes" on page 79.
  - Appliance admin can check information about nodes. See chapter "Checking

Status" on page 73.

- Appliance admin can perform operations on nodes:
  - Power On. See chapter "Powering on Node" on page 80.
  - Power Off. See chapter "Powering off Node" on page 81.
  - Reboot. See chapter "Rebooting Node" on page 82.
  - Set Physical Location. See chapter "Changing Physical Location of Node" on page 83.
  - Set JX Serial Number. See chapter " Change JX Serial Number of Node" on page 83.
- Appliance admin can check information about switches. See chapter "Managing Switches" on page 84.
- Appliance admin can perform operations on switches:
  - Edit switch serial numbers.
  - Edit switch type.
- Appliance admin can see uninstalled nodes. See chapter "Managing Installation" on page 85.
- Appliance admin can install new nodes
- Appliance admin can see history of user's activity. See chapter "Users Activity" on page 86.

- Administrator – authorized role to perform operation on nodes and cluster. Administrator manages history of user's activity and manages user.
  - Administrator can check versions of installed software. See chapter "Checking Version" on page 72
  - Administrator can perform diagnostic tests. See chapter "Checking Status" on page 73.
  - Administrator can see cluster settings and SNMP settings. See chapter "Managing Settings" on page 75
  - Administrator can check information about nodes. See chapter "Managing Nodes" on page 79.
  - Administrator can perform operations on nodes:
    - Power On. See chapter "Powering on Node" on page 80.
    - Power Off. See chapter "Powering off Node" on page 81.
    - Reboot. See chapter "Rebooting Node" on page 82.
    - Set Physical Location. See chapter "Changing Physical Location of Node" on page 83.

        - Set JX Serial Number. See chapter " Change JX Serial Number of Node" on page 83.
    - Administrator can check information about switches. See chapter "Managing Switches" on page 84.
    - Administrator can perform operations on switches:
        - Edit switch serial numbers.
        - Edit switch type.

      Administrator can see uninstalled nodes. See chapter "Managing Installation" on page 85.
    - Administrator can install new nodes
    - Administrator can see history of user's activity. See chapter "Users Activity" on page 86.
    - Administrator can delete users activity. See chapter "Deleting Users Activity" on page 88.
    - Administrator can manage basking's users. See chapter "Users Activity" on page 86.
- Service user – special user for factory installation with instructions and permissions to initialize cluster in the factory.
    - Service user can check versions of installed software. See chapter "Checking Version" on page 72
    - Service user can perform diagnostic tests. See chapter "Checking Status" on page 73.
    - Service user can see cluster settings and SNMP settings. See chapter "Managing Settings" on page 75
    - Service user can check information about nodes. See chapter "Managing Nodes" on page 79.
    - Service user can perform operations on switches:
        - Edit switch serial numbers.
        - Edit switch type.
    - Service user can perform operations on nodes:
        - Power On. See chapter "Powering on Node" on page 80.
        - Power Off. See chapter "Powering off Node" on page 81.
        - Reboot. See chapter "Rebooting Node" on page 82.
        - Set Physical Location. See chapter "Changing Physical Location of Node" on page 83.

        - Set JX Serial Number. See chapter " Change JX Serial Number of Node" on page 83.
    - Service user can check information about switches. See chapter "Managing Switches" on page 84.
    - Service user can see uninstalled nodes. See chapter "Managing Installation" on page 85.
    - Service user can install new nodes
    - Service user can see history of user's activity. See chapter "Users Activity" on page 86.
    - Service user can manage basking's users.

Basking provides following functions:

- Checking version of the installed software
- Performing diagnostic tests
- Checking settings of cluster
- Checking settings of current public network
- Setup the public network
- Managing SNMP
    - Listing communities and destinations
    - Adding and removing communities and destinations
- Checking information about nodes
- Perform operations on nodes:
    - Power on
    - Power off
    - Reboot
    - Set physical location
    - Set JX serial number
- Checking information about switches
- Perform operation on switches:
    - Set switch type
    - Set switch serial numbers
- Checking information about not installed nodes
- Install raw nodes
- View history of action performed in basking graphical interface

- Manage basking's users
- Performing all steps needed for initialization for factory user

4.3.2 Handling Errors

If basking-api is not available, so every managing operation will not be executed.

External interface specification

Log entries, system messages on management node:

- /var/log/basking/
- /var/log/basking-api/

4.3.3 GUI Overview

4.3.3.1 Login Page

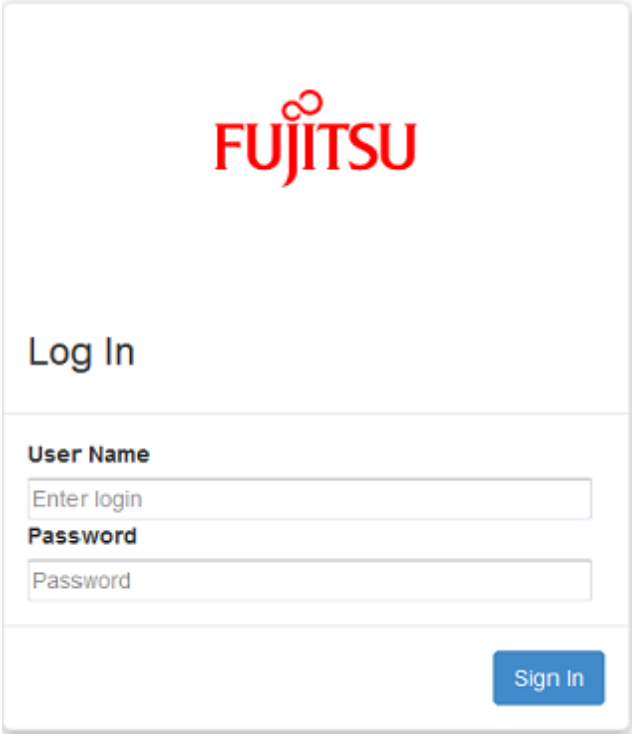


Figure 18: Login page

Enter the "User Name" and the "Password". Click on the button "Sign In" or press **Enter** to confirm.

4.3.3.2 Navigation Page

The navigation in Basking may differ depending on the role of the user account. The role affects the access to section, which requires the special permission.

The navigation is divided into two the section A and section B:

- Section A:
  - Left menu – this menu contains the list of all available sections for the role of the active user
- Section B:
  - Top section of the page:
    1. Title of the current section
    2. Access to ETERNUS CD10000 Manual
    3. Edit user account settings
    4. Logout

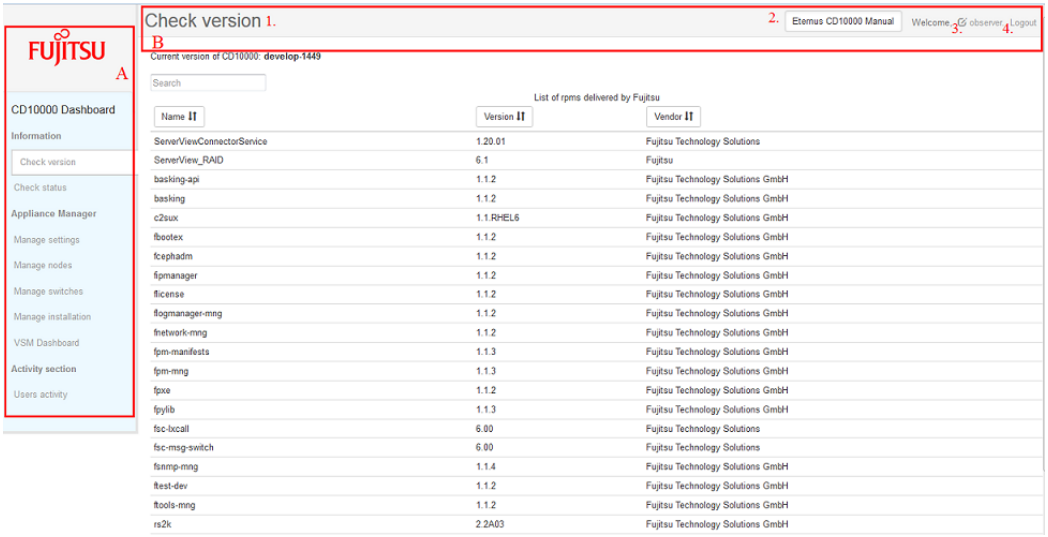


Figure 19: Navigation

4.3.4 Information

This chapter describes the basic information about the software, which is included in ETERNUS CD10000 and diagnostic information to identify possible issues.

4.3.4.1 Checking Version

This page contains information about ETERNUS CD10000 software.

1. Current version installed on the ETERNUS CD10000
2. Software list provided by Fujitsu
3. "Search Field" - exact RPM package can be found using this search field.

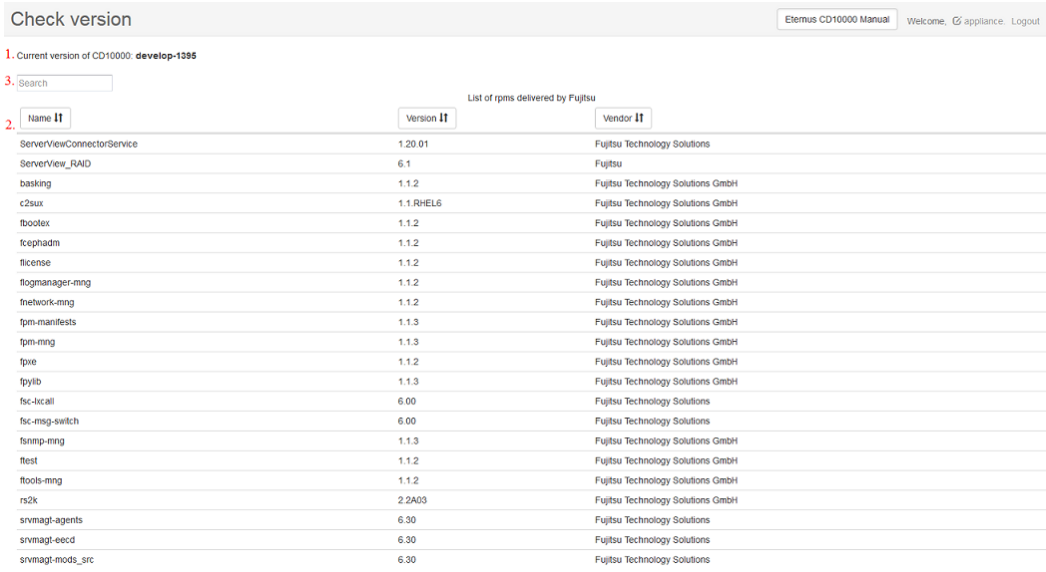


Figure 20: Checking version

4.3.4.2 Checking Status

This is a diagnostic page. The user can run different tests here. The tests are classified in categories, which may differ from the ETERNUS CD10000 version.

1. Available categories:
   - Factory – these tests give a basic information of the system (like kernel version or time zone). Also these tests give the information if the software installed version is correct and if the nodes are accessible in the network.
   - System – these tests give the information about the system configuration and the software installed version.
   - Network – these tests give the information about the administration network, the infiniband network and the irmc network.
   - All - these tests give the information about all above listed categories, except Factory.
2. By clicking the button "Show status information" the test of the currently selected category will be started.
3. By clicking the button "Show all", you can filter the tests results.
   Available options are: "Show all", "Succeeded", "Failed", "Skipped".

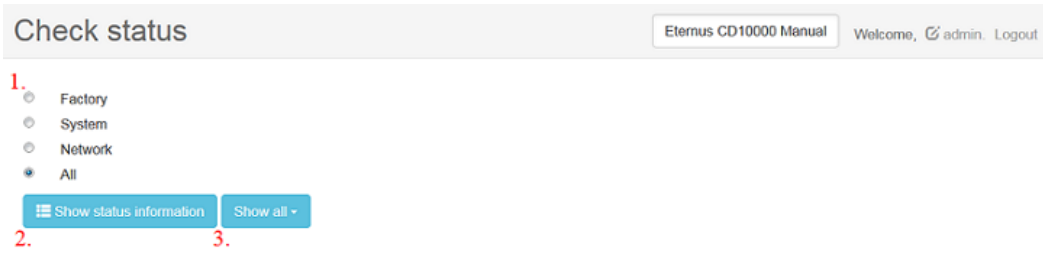
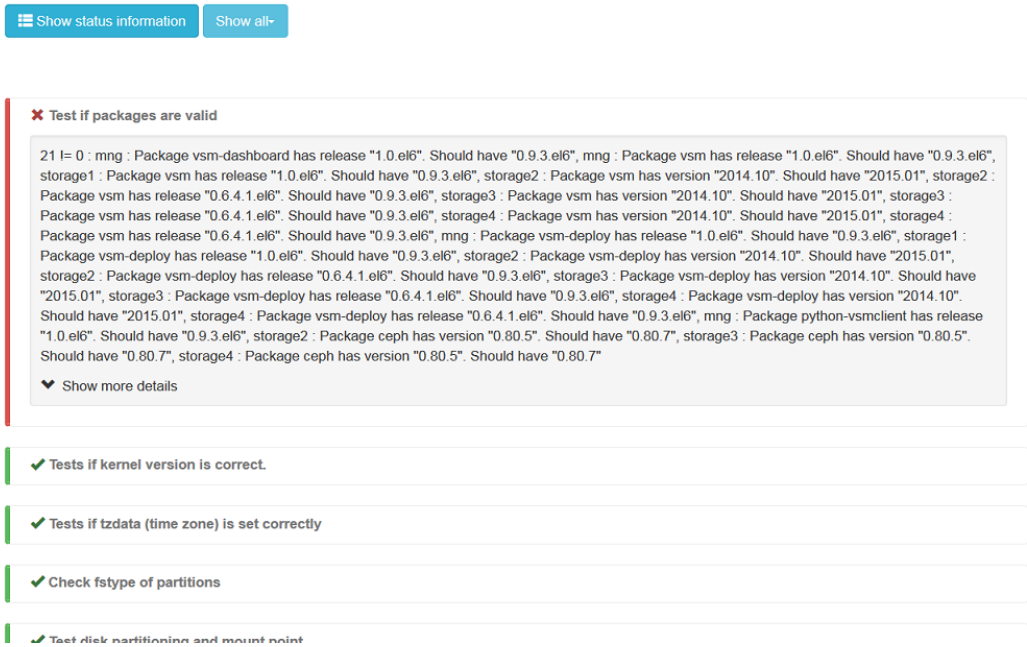


Figure 21: Checking status

4.3.4.3 Running Diagnostic Tests

In order to perform tests, it is required to choose group of tests. Click on the button "Show status information". This will start running test action in the background.

The test initialization is highlighted by appearance of the new label, which state that tests are already running. After finishing of the test performing, the results will appear in place of first label.



The failed test have additional description of the problem. By clicking on them, you are able to see the stack-trace of the error which cause the test failure.

4.3.5 Appliance Manager

This chapter describes the information and the settings for physical cluster of servers. There is also a link to the VSM Dashboard.

4.3.5.1 Managing Settings

1. The cluster settings section contains:
    1. The information about cluster and networks
    2. The cluster serial number – editable by clicking on the value
    3. The current public network settings
    4. The setup public network form
2. The SNMP settings section contains:
    1. A list of SNMP communities
    2. The button for creating a new SNMP community
    3. The button for removing one of existing SNMP community
    4. A list of SNMP destinations
    5. The button for creating new SNMP destination
    6. The button for removing one of existing SNMP destination

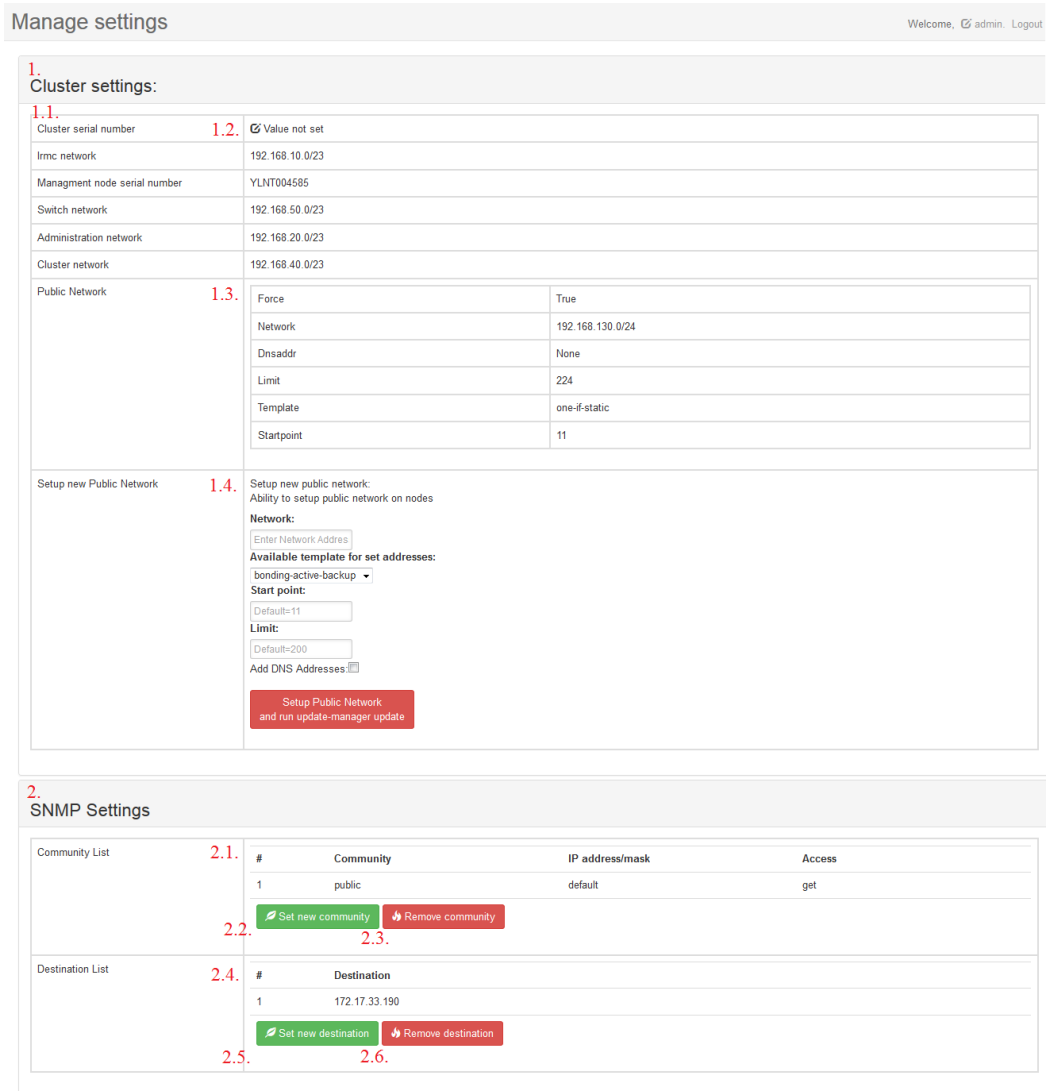


Figure 22: Manage settings

Changing the Cluster Serial Number

To change the serial number of a cluster, click on the icon "Cluster serial number". This field is editable.

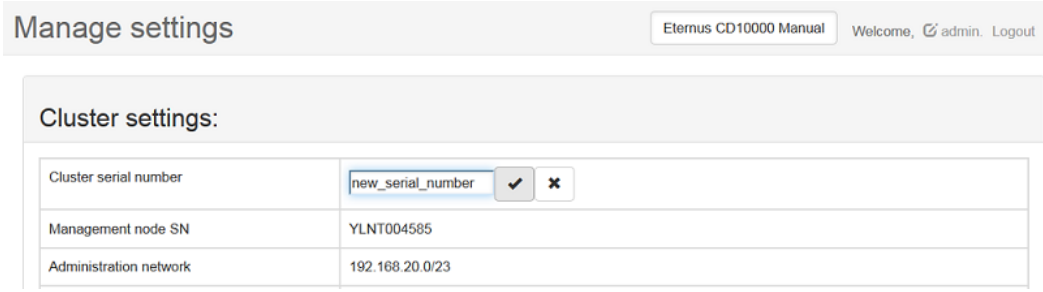


Figure 23: Changing the cluster serial number.

Confirm the change. For a short while, in place of input appears the text "Saving…". The database will be updated.

Setting New Community

Click on the button "Set new community". A new form for creating a new SNMP community appears.

For detailed description see "set-community - Defining a Community" on page 319.

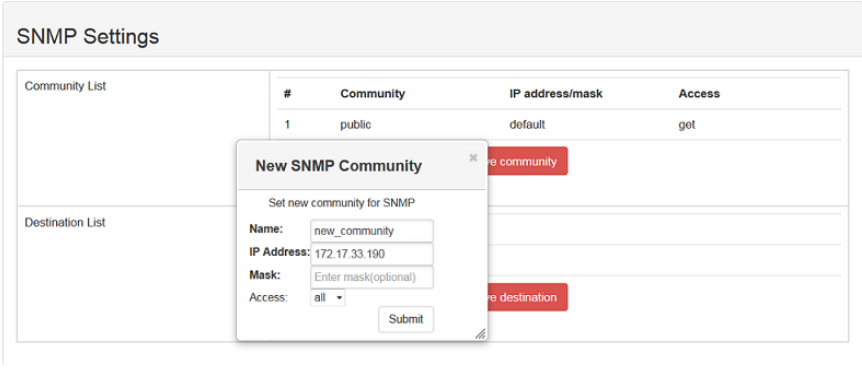


Figure 24: Setting new community

Removing Community

Click on the button "Remove community" to delete existing community. Enter the name of the community which is supposed to be deleted.

For detailed description see "remove-community - Removing a Community" on page 320.

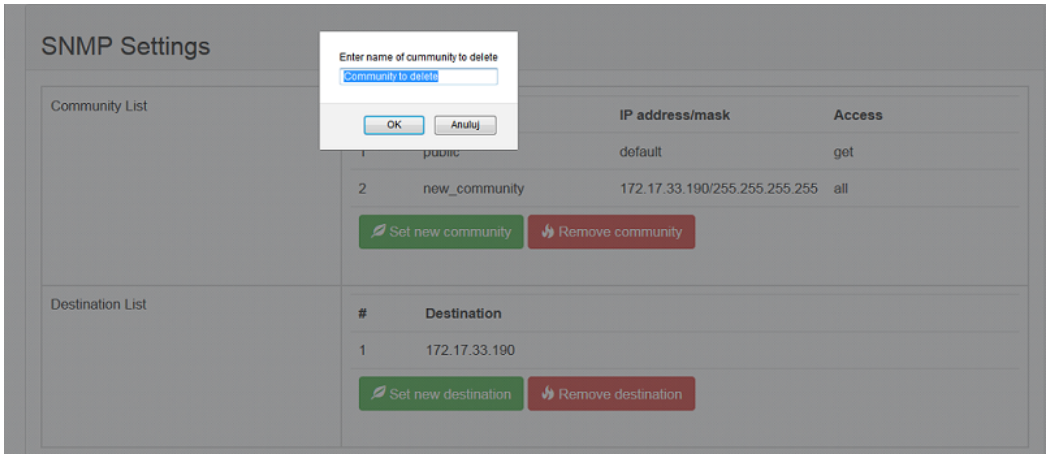


Figure 25: Removing community

Setting New Destination

Click on the button "Set new destination". A new form for creating a new SNMP destination appears.

For detailed description see "set-destination - Defining a Trap Destination" on page 322.

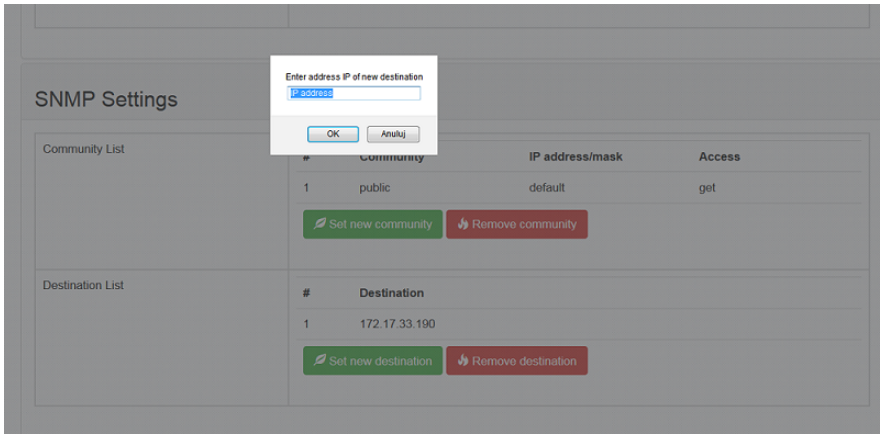


Figure 26: Setting new destination

Removing Destination

Click on the button "Remove destination" to delete existing destination.

Enter the IP address of the destination which is supposed to be deleted.

For detailed description see "remove-destination - Removing a Trap Destination" on page 323.

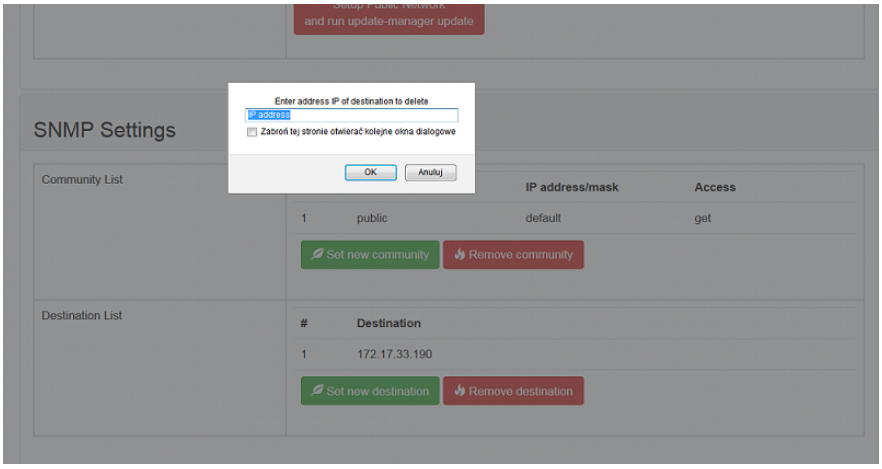


Figure 27: Removing destination

4.3.5.2 Managing Nodes

This chapter describes the detailed information about nodes installed in the cluster and how to manage them. The figure below shows following:

1. A table of currently installed nodes.
2. "Search field" - input for table of nodes (filter through all sortable columns).
3. An additional filter for "Status" of nodes, by clicking on it, a list will expand with all possible node statuses.
4. A list of columns available for nodes. You can configure information displayed by the table.
5. Auto-refreshing button - table will be refreshed every minute.
6. It shows the list of selected nodes on the page – on this selected node, actions will be performed.
7. All available operations for nodes.

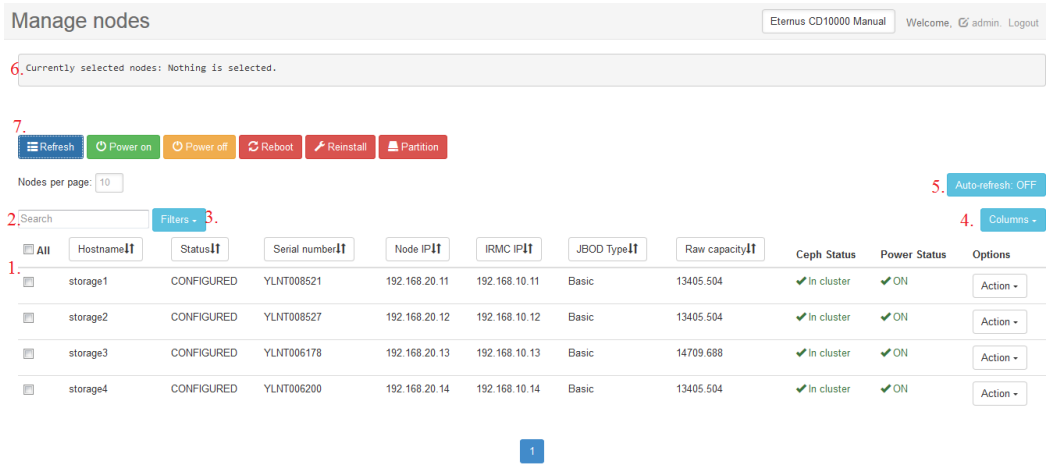


Figure 28: Managing nodes

Powering on Node

Powering on of the node requires first of all selecting a node from the table with "Power Status" equals "Off".

Next click on the button "Power on".

Running this action on the node, which is added to Ceph, can trigger re-balancing of data in the cluster. This command does not use maintenance mode.

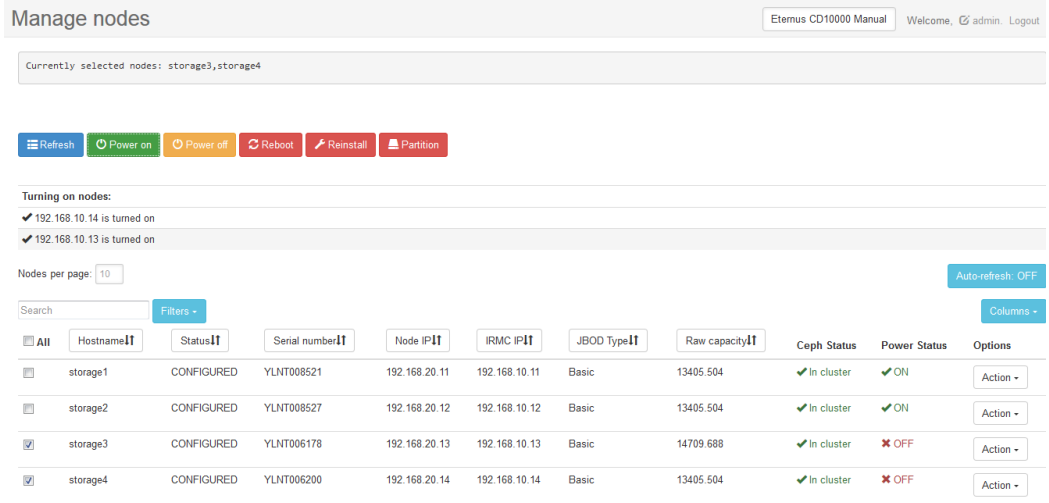


Figure 29: Powering on node

After a few seconds the "Power Status" of the node will be updated.

Powering off Node

Powering off of the node requires first of all selecting a node from the table with "Power Status" equals "On".

Next click on the button "Power off".

Running this action on the node, which is added to Ceph, can trigger re-balancing of data in the cluster. This command does not use maintenance mode.

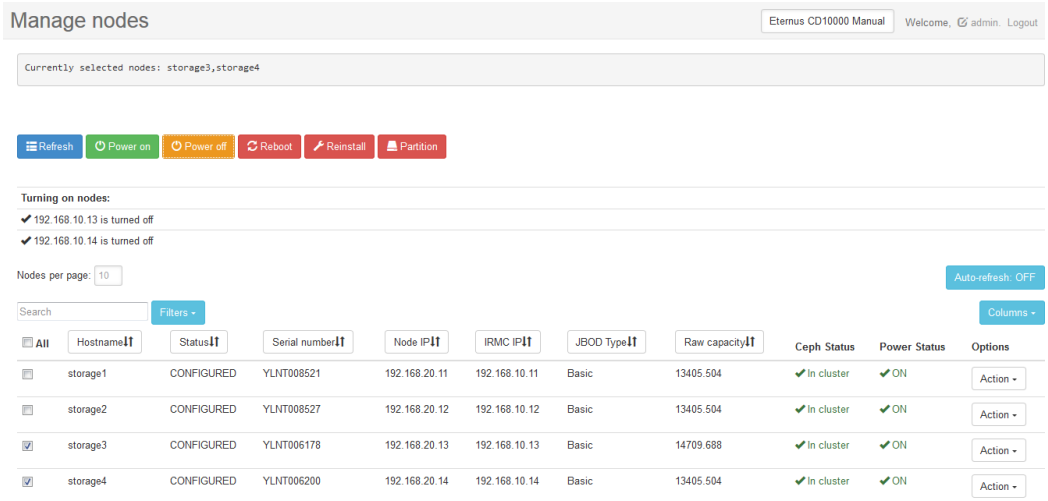


Figure 30: Powering off node

After a few seconds the "Power Status" of the node will be updated.

Rebooting Node

Rebooting of the node requires first of all selecting a node from the table with "Power Status" equals "On".

Next click on the button "Reboot

Running this action on the node, which is added to Ceph, can trigger re-balancing of data in the cluster. This command does not use maintenance mode.

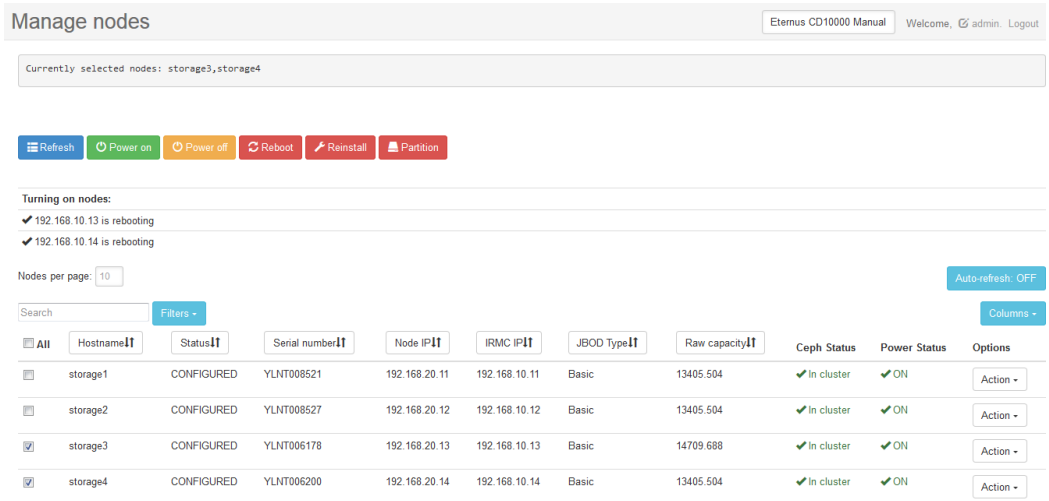


Figure 31: Rebooting node

After a few seconds the "Power Status" of the node will be updated.

Changing Physical Location of Node

To change the physical location of the node click on the icon "Physical Location". This field is editable.

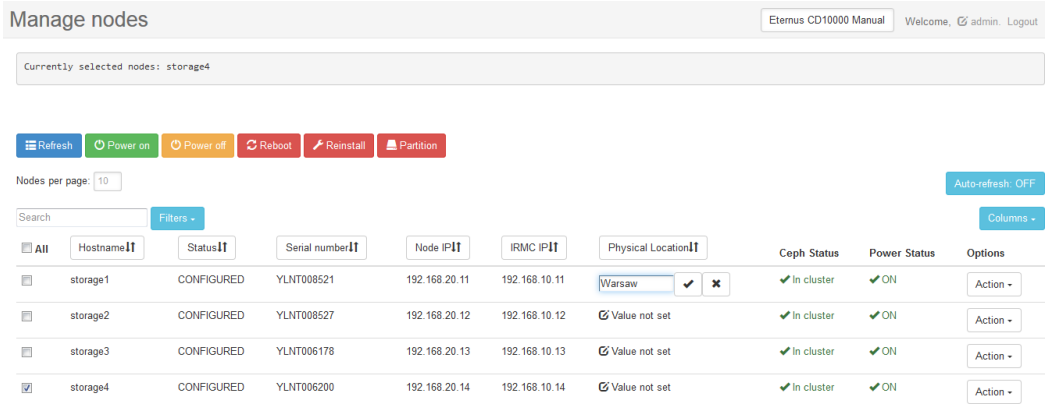


Figure 32: Changing physical location of node

Confirm the change. For a short while, in place of input appears the text "Saving…". The database will be updated.

Change JX Serial Number of Node

To change jx the serial number of the node, click on the icon" JX Serial number". This field is editable.

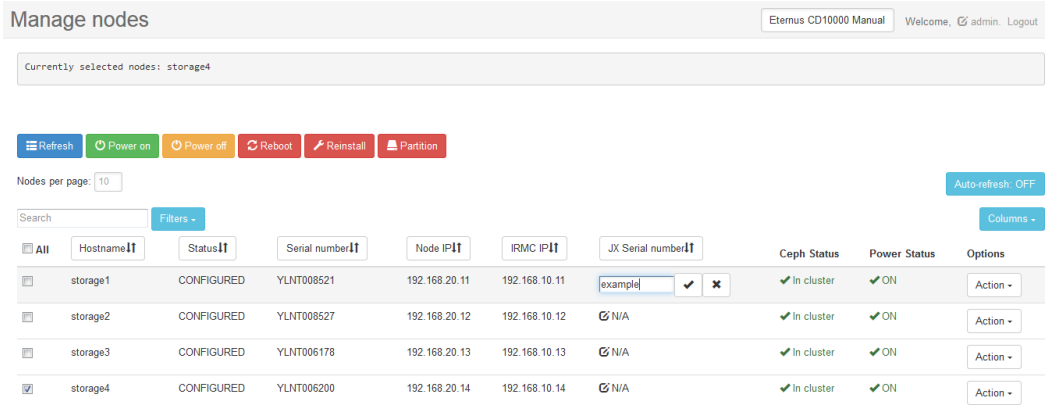


Figure 33: Change jx serial number of node

Confirm the change. For a short while, in place of input appears the text "Saving…". The database will be updated.

4.3.5.3 Managing Switches

This chapter describes the detailed information about switches.

1. A table of the presenting switches in the database.
2. "Search field" - input for table of switches (filter through all sortable columns).
3. Auto-refreshing button - table will be refreshed every minute.
4. The option "Edit" for changing type and serial number of switches.

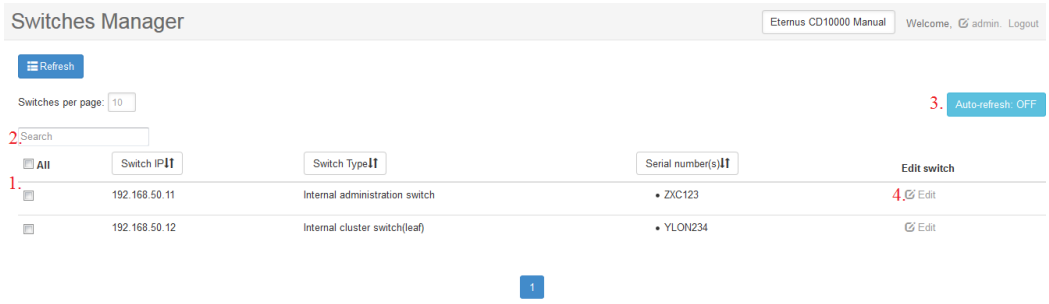


Figure 34: Managing switches

Changing Switch Type

To change the type of a switch click on the “Edit” option at the end of the of switch row in the"Switches Manager" window.

From the "Type" list select one of the available switch types. If you select “Internal cluster switch” from the "Type" list the Position list is displayed additionally.

Select one of the available values. Click on the “Submit” button to confirm.

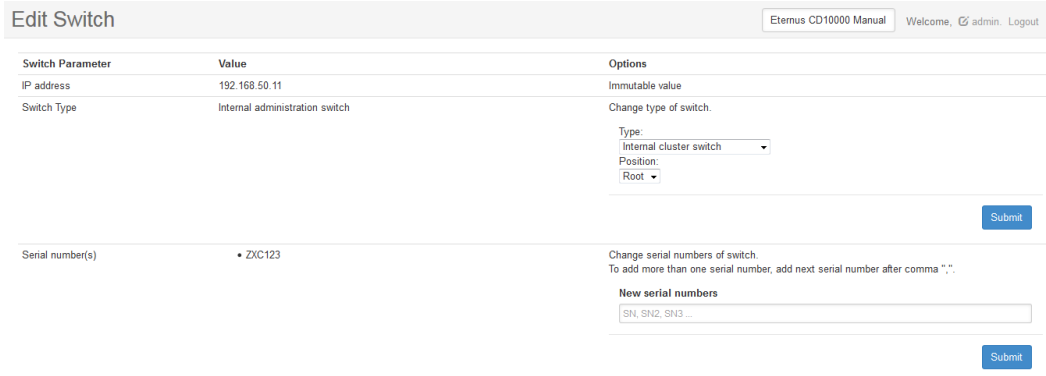


Figure 35: Edit switch type

Changing Serial Numbers of a Switch

To change the serial number of a switch click on the “Edit” option at the end of the of switch row in the"Switches Manager" window.

Enter the new serial numbers in the “New serial numbers” field. To add more than one serial number, enter the serial numbers separated by commas, e.g. “SERIAL_NUMBER_1, SERIAL_NUMBER_2, SERIAL_NUMBER_3”.

Click on the “Submit” button to confirm.

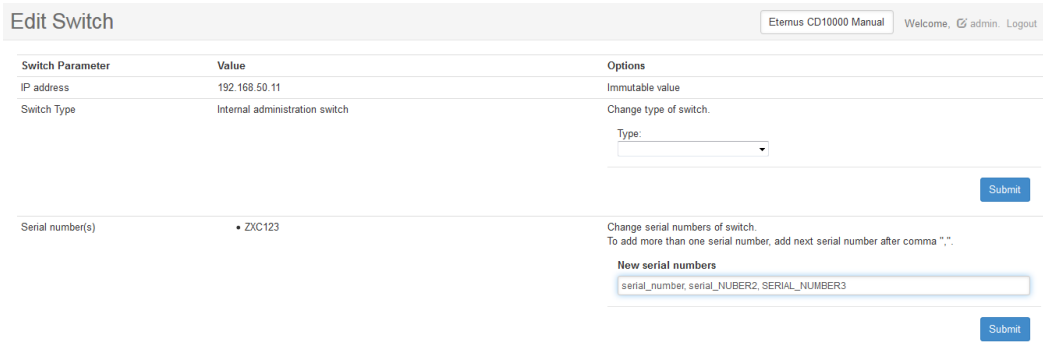


Figure 36: Edit serial numbers of a switch

4.3.5.4 Managing Installation

1. A table of the presenting uninstalled nodes.
2. "Search field" - input for table of nodes (filter through all sortable columns).
3. Auto-refreshing button - table will be refreshed every minute.
4. All available actions for uninstalled nodes.

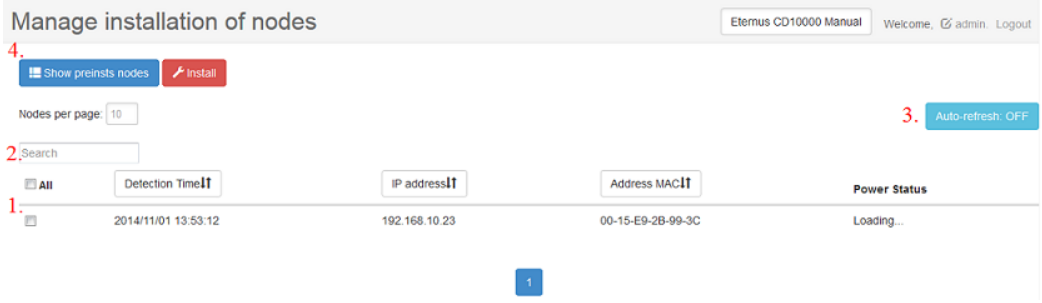


Figure 37: Managing installation

4.3.6 Activity Section

4.3.6.1 Users Activity

The users activity section contains history of all operation performed by any of basking users on cluster, nodes or SNMP. This is the history only of actions performed in Basking web gui. The history of operations performed by command line interface are not visible here.

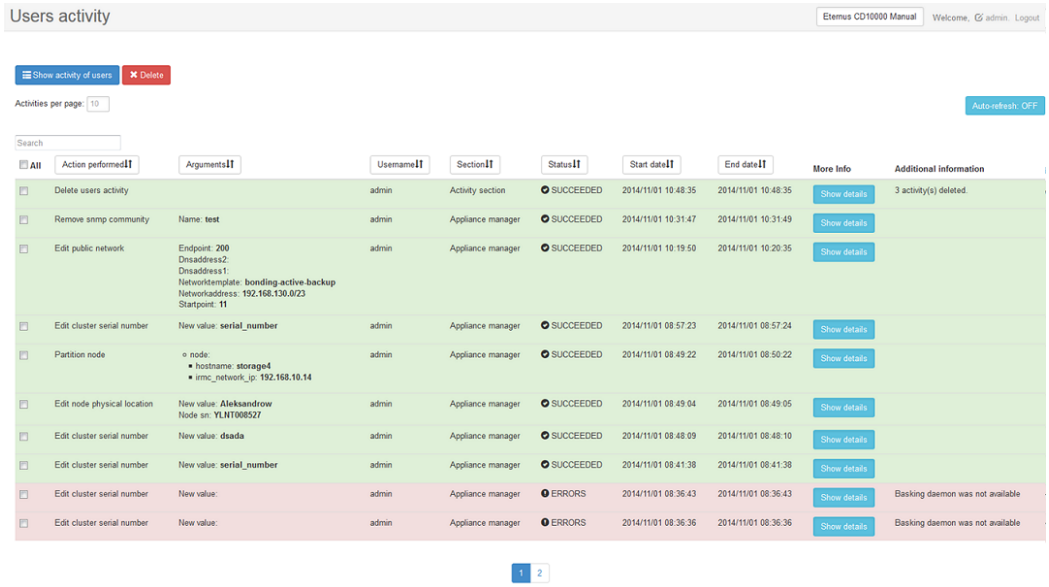


Figure 38: Users activity

There are three types of actions, you can find in the users activity:

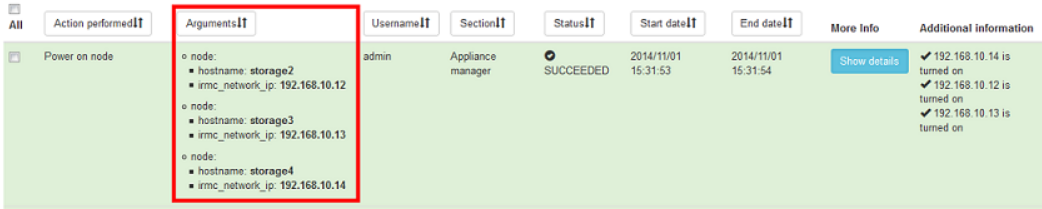
- SUCCEEDED (green background) – the action finished without any errors.

- EXECUTING (blue background) – the action has not finished yet.

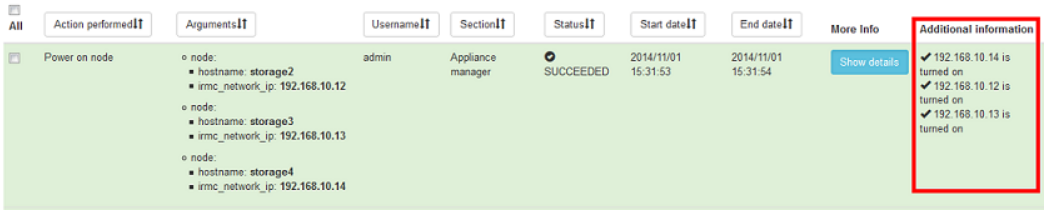
- ERRORS (red background) – the actions finished with errors or failed to finished.

Each activity is described by following fields:

- Action performed – a brief description of action like "Power on node" or "Edit public network".
- Arguments – it contains a list of the arguments connected to the performed action. For example, a list of the nodes which are powered on.



- Username – the name of user, who performs the current action.
- Section – a section in basking where the action was performed (like Appliance Manager, Activity Section or User Management).
- Start date – the accurate date of the starting action.
- End date – the accurate date of the finishing action.
- More info – a link button to "Activity log" page with detailed description of the action, described below.
- Additional information – a short information about the performance of the action. For example, in the case of the powering on of the nodes, it contains specified information which selected node was powered on successfully.



4.3.6.2 Deleting Users Activity

The administrator is able to delete the redundant user's activity. As administrator check all unwanted logs and click on the button "Delete".

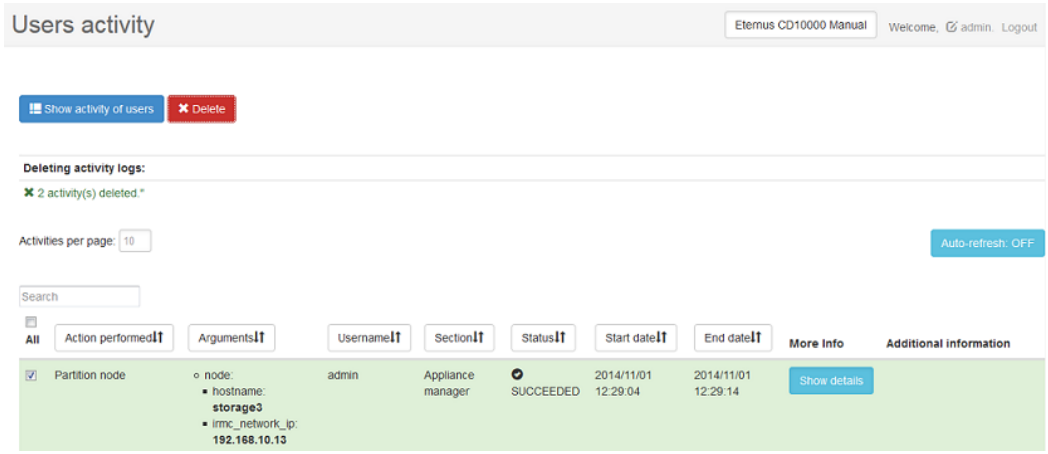


Figure 39: Deleting users activity

4.3.6.3 Activity Log - <<action_name>>

This chapter describes all basic information about the running action. Depending on your access rights, you are able to view some specific action additionally. For example logs of partitioning or setting up new public network.

If the action has "executing"status and the action is connected with logs, then logs will be visible on page in real-time. The figure below shows following:

1. An information overview of the activity logs.
2. The detailed logs generated in real-time.

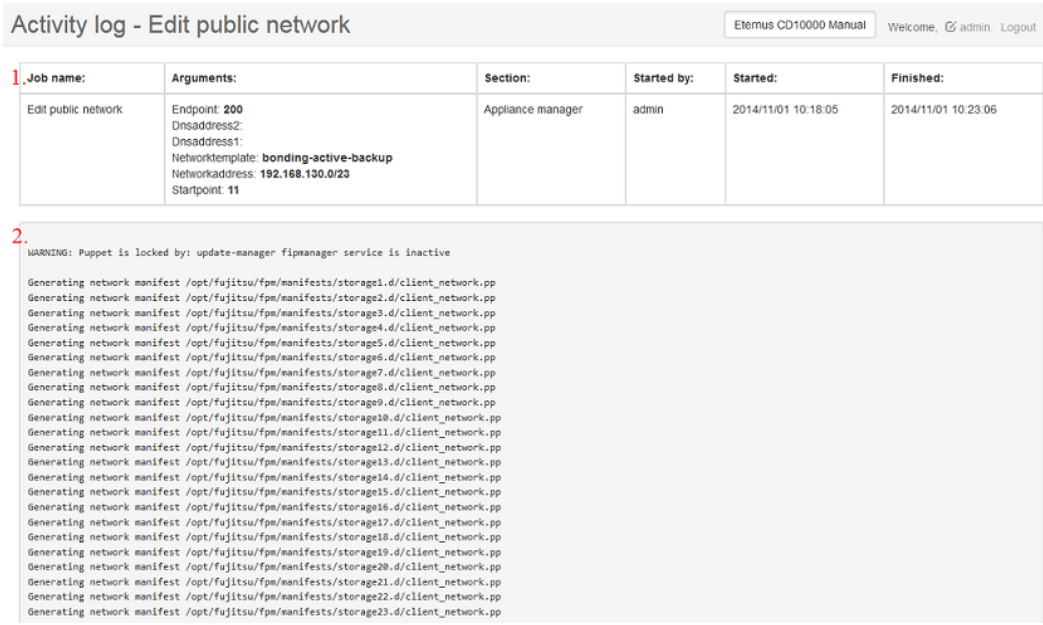


Figure 40: Activity log – edit public network

4.3.7 User Management

4.3.7.1 Managing Users

This page can only be accessed by administrator and service user.

1. Table presenting a list of Basking's users
2. "Search Field" - input for searching a concrete user.
3. Button for triggering a creation of a new user.
4. Button for deleting a selected user.
5. "Edit" option for editing users ("admin" and "service" are default users and cannot be edited).

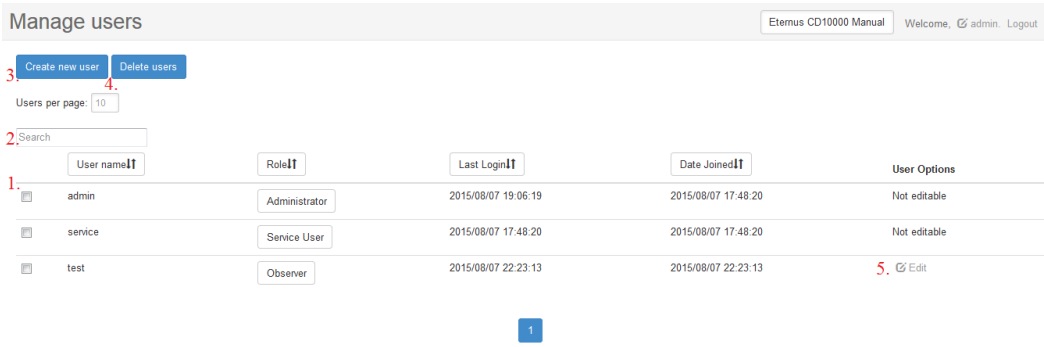


Figure 41: Managing users

4.3.7.2 Creating New User

You need the administrator access rights to perform this action.

To create a new user, go to "Manage users" tab and click on the button "Create new user" and enter the user name and the password of the new user.

This is the only way of creating a user via basking! There is no outside registration form!

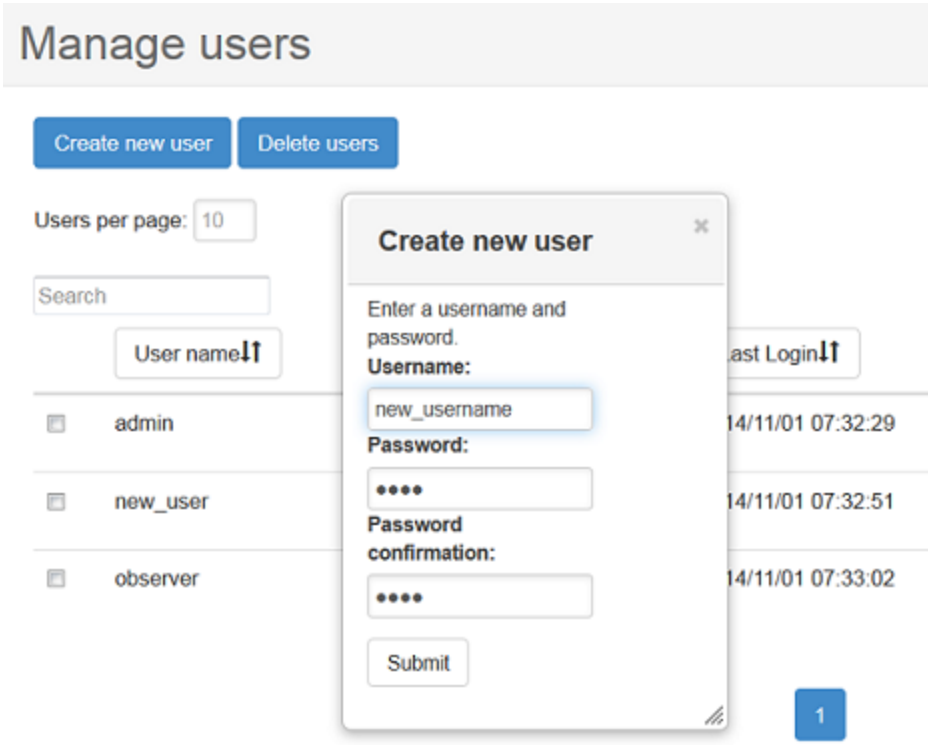


Figure 42: Creating of a New User

Click on the button "Submit" to confirm. The new user is automatically created and the table refreshes.

4.3.7.3 Deleting Users

You need the administrator access rights to perform this action.

To delete a user, select the checkboxes of each user who is supposed to be delete. Click on the button "Delete users". You will be able to check again selected users and you will be asked to confirm this action.

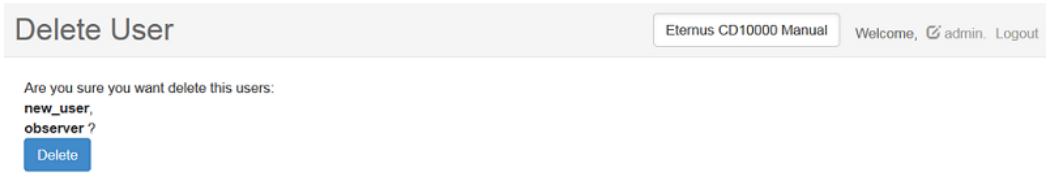


Figure 43: Deleting of users

Click on the button "Delete". The user is immediately deleted and there is no way to restore them.

5 Configuration and Administration of a Storage Cluster

ETERNUS CD10000 provides the web-based graphical interface Virtual Storage Manager (VSM) for creating, configuring, administering and monitoring of Ceph storage clusters.

Additionally the command line interface `vsm_cli` can be used to create, configure, administer and monitor the Ceph storage cluster.

VSM and vsm_cli provide functions for the following tasks:

- creating and configuring a Ceph storage cluster
- creating and managing pools within the storage cluster
- managing storage groups
- managing of the storage nodes and disks
- managing the storage cluster's daemons (OSDs and monitor daemons)
- monitoring of the cluster capacity and performance

For detailed information on the VSM see "Virtual Storage Manager for Ceph" on page 57

The Command line interface **`vsm_cli`** provides additional functions for the following tasks:

- deleting storage pools and modifying pool parameters
- stopping and restarting storage cluster daemons and the whole storage cluster
- replacing failed/retired disks
- changing the Crush weight of an OSD
- transfering storage nodes into maintenance mode
- monitoring of the cluster health and the monitor status

For detailed information on **`vsm_cli`** see

A list with all commands of the **`vsm_cli`** interface is displayed when you enter the command

```
vsm_cli help
```

5.1 Prerequisites

You can create, configure and administer a storage cluster if the following prerequisites are met:

- The management node, the node on which the VSM controller resides (VSM controller node) and all storage nodes are connected to the cluster internal administration network.
- The ETERNUS CD10000 Software is completely installed on each part of the storage cluster.
- The VSM cluster manifest file is available on the VSM controller node in /etc/manifest/cluster.manifest.
- On each storage node a VSM agent is running. A server manifest file describes the configuration of the storage node.

  During installation of VSM on the storage node a template for a server manifest file will be created automatically in `/etc/manifest/server.manifest`. You only must adapt it to the node configuration.

  You will find a template of the server.manifest file containing the generated mapping of journals to storage devices and the available disks at `/etc/fcephadm/server.manifest`.
- In addition VSM requires Google Chrome or Internet Explorer 10, or another modern Web-browser (tested with Google Chrome and Internet Explorer 10 with compatibility mode turned off).

5.2 Managing Storage Pools

VSM provides the following functions to manage storage pools:

- Retrieving information about the storage pools

  For created and active pools you get information about the type of pool (replicated or erasure coded), the number of data copies (replicas) respectively the number of placement groups and the cache tier status.
- Creating replicated or erasure coded pools within a storage cluster
- Converting two existing pools, a storage pool and a cache pool, into a tiered pool.
- Removing cache tiering.

`vsm_cli` additionally provides functions for the following tasks:

- Deleting pools from a storage cluster.
- Changing the number of copies from the data objects and the quota of storage size allocated to an existing storage pool.

5.2.1 Description about the Different Types of Pools

Ceph stores data within pools, which are logical groups for storing objects. Pools manage the number of placement groups, the number of replicas, and the ruleset for the pool. To store data in a pool, you must have an authenticated user with permissions for the pool. Ceph can snapshot pools.

Ceph stores, replicates and rebalances data objects across a cluster dynamically. With many different users storing objects in different pools for different purposes on countless OSDs, Ceph operations require some data placement planning about the type of used pools and the rules where to place the data. (Reference: http://ceph.com/docs/master/rados/operations/data-placement)

For more information, see:

http://ceph.com/docs/master/rados/operations/data-placement

5.2.1.1 Replicated Pools

The replicated pools store data according to the number of replicas specified. All replicas are either stored in one storage group or the non-primary replica data may optionally be placed in a separate storage group.

The last case can be used, for example, to place the primary copy in a Storage Group comprised of SSDs, while the replica copies are placed in a Storage Group comprised of 7200 RPM rotating disks.

In this configuration, data will be read with very high performance from the primary OSDs placed on SSDs, while data will be written with the usual performance of HDDs (because all copies must be written before the write operation is signaled as complete).

5.2.1.2 Erasure Coded Pool

(Reference: http://ceph.com/docs/master)

The erasure coded pool type can be used instead to save space.

Like replicated pools, in an erasure coded pool the primary OSD in the up set receives all write operations.

In replicated pools, Ceph makes a deep copy of each object in the placement group on the secondary OSD(s) in the set.

For erasure coding, the process is a bit different. An erasure coded pool stores each object as **K+M** chunks. Each object is divided into **K** data chunks and **M** coding chunks. The pool is configured to have a size of **K+M** so that each chunk is stored in an OSD in the acting set.

The object is first divided into K data chunks and the M coding chunks are computed.

The rank of the chunk is stored as an attribute of the object. The primary OSD is responsible for encoding the payload into K+M chunks and sending them to the other OSDs.

In an event of failure, to construct the original file, any K chunks out of these K+M chunks are needed to recover it.

For instance an erasure coded pool is created to use five OSDs (K+M = 5) and sustain the loss of two of them (M = 2).

The simplest erasure coded pool is equivalent to RAID5 and requires at least three hosts.

Erasure coded pools lack some functionality such as partial writes. Therefore the direct use of erasure coded pools is only supported with the rados gateway. It is not possible to create an RBD image on an erasure coded pool because it requires partial writes. It is however possible to create an RBD image on an erasure coded pool when a replicated pool tier sets a cache tier.

Erasure code profile

The default erasure code profile (which is created when the Ceph cluster is initialized) provides the same level of redundancy as two copies but requires 25% less disk space. It is described as a profile with K=2 and M=1, meaning the information is spread over three OSDs (K+M = 3) and one of them can be lost.

Choosing the right profile is important because it cannot be modified after the pool is created: a new pool with a different profile needs to be created and all objects from the previous pool moved to the new one.

The most important parameters of the profile are K, M and the ruleset-failure-domain because they define the storage overhead and the data durability.

Erasure code failure domain

The ruleset-failure-domain (erasure code failure domain) specifies in which only one (never more than one) segment of an erasure coded object is allowed to be placed.

- If **`erasure code failure domain = OSD`** is specified the erasure coded data are placed across OSDs.

  With this setting you may run into situation where you can lose data. It may happen that the number of chunks is greater than **M** and is written to one host. In this case the failure of the host will cause data loss.

- If **`erasure code failure domain = Zone`** is specified the erasure coded data are placed across failure zones.
- If **`erasure code failure domain = Host`** is specified erasure coded data are placed across servers.

The erasure coded failure domain must be equal to or larger than the K+M value of the selected erasure coded profile.

Example:

Loss of two chunks should be covered, i.e. M=2. The number of data chunks should be 3, i.e. K=3

For storing an object 5 OSDs are needed.

If you specify **`OSD`** as erasure coded failure domain the data is written to 5 separate OSDs but some of the OSDs might be in the same host (storage node). So in the worst case if 3 chunks are written to OSDs in the same host and this host is lost, the data is lost because it cannot be restored from the remaining OSDs.

If you specify host as erasure coded failure domain the data is written to 5 separate OSDs in different hosts. This implies that the Ceph cluster must be built of at least 5 hosts.

For more information, see:

http://ceph.com/docs/master/rados/operations/erasure-code

5.2.1.3 Cache Tier Pools

A cache tier provides Ceph Clients with better I/O performance for a subset of the data stored in a backing storage tier. Cache tiering involves creating a pool of relatively fast/expensive storage devices (e.g., solid state drives) configured to act as a cache tier, and a backing pool of either erasure-coded or relatively slower/cheaper devices configured to act as an economical storage tier. The Ceph objecter handles where to place the objects and the tiering agent determines when to flush objects from the cache to the backing storage tier. So the cache tier and the backing storage tier are completely transparent to Ceph clients.

The cache tiering agent handles the migration of data between the cache tier and the backing storage tier automatically. However, an administrator can configure how this migration takes place. There are two main scenarios:

- **Writeback Mode:**

  When an administrators has configured tiers with writeback mode, Ceph clients write data to the cache tier and receive an acknowledgment ACK from the cache tier.

  In time (in regular time intervals), the data written to the cache tier migrates to the storage tier and gets flushed from the cache tier. Conceptually, the cache tier is overlaid “in front” of the backing storage tier. When a Ceph client needs data that resides in the storage tier, the cache tiering agent migrates the data to the cache tier on read, then it is sent to the Ceph client. Thereafter, the Ceph client can perform I/O using the cache tier, until the data becomes inactive. This is ideal for mutable data (e.g., photo/video editing, transactional data, etc.).

- **Read-only Mode:**

  When an administrators has configured tiers with read-only mode, Ceph clients write data to the backing tier. On read, Ceph copies the requested object(s) from the backing tier to the cache tier. Objects get removed from the cache tier based on the defined policy. This approach is ideal for immutable data (e.g., presenting pictures/videos on a social network, DNA data, X-Ray imaging, etc.), because reading data from a cache pool that might contain out-of-date data provides weak consistency. Do not use read-only mode for mutable data.

Since all Ceph clients can use cache tiering, it has the potential to improve I/O performance for Ceph Block Devices, Ceph Object Storage, the Ceph Filesystem and native bindings.

To set up cache tiering, you must have two pools. One will act as the backing storage and the other will act as the cache. (Reference:
http://ceph.com/docs/master/rados/operations/cache-tiering/)

Setting up a backing storage pool

Setting up a backing storage pool typically involves one of two scenarios:

- Standard Storage:

  In this scenario, the pool stores multiple copies of an object in the Ceph Storage Cluster.

- Erasure Coding:

  In this scenario, the pool uses erasure coding to store data much more efficiently with a small performance tradeoff.

In the standard storage scenario, you can setup a CRUSH ruleset to establish the failure domain (e.g., osd, host, chassis, rack, row, etc.). Ceph OSD Daemons perform optimally when all storage drives in the ruleset are of the same size, speed (both RPMs and throughput) and type. See CRUSH Maps for details on creating a ruleset. Once you have created a ruleset, create a backing storage pool.

In the erasure coding scenario, the pool creation arguments will generate the appropriate ruleset automatically. See Create a Pool for details.

In subsequent examples, we will refer to the backing storage pool as cold-storage.

(Reference: http://ceph.com/docs/master/rados/operations/cache-tiering/)

Setting up a cache pool

Setting up a cache pool follows the same procedure as the standard storage scenario, but with the following difference:

- The drives for the cache tier are typically high performance drives that reside in their own servers and have their own ruleset.

When setting up a ruleset, it should take only account of the hosts that have the high performance drives but not of the hosts that do not. (Reference: http://ceph.com/docs/master/rados/operations/cache-tiering/)

For more information, see:

http://ceph.com/docs/master/rados/operations/cache-tiering

5.2.2 Managing Pools (VSM GUI)

To receive information about all pools within the storage cluster

- select from the **Cluster Management** main menu the function **Manage Pools**.

The **All Pools** window containing a list of all pools within the storage cluster is displayed.

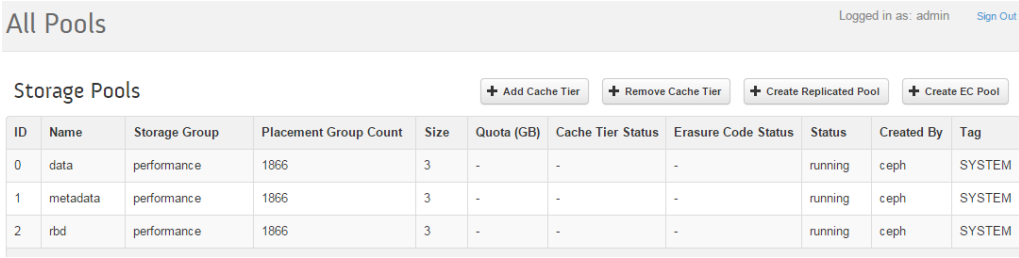


Figure 44: List of the currently existing storage pools

5.2.2.1 Information on the storage pools

In the columns the following is displayed for each pool:

ID — ID of the pool. Sequence number automatically assigned during the pool creation. The pools are listed ascending in order of creation.

Name — Name of the storage pool.

Storage Group — Storage group where the pool is placed.

From pools with separate primary and replica storage groups, the storage group is displayed where the primary copy is placed.

The storage groups are defined in the cluster manifest file (see "Cluster Manifest File" on page 234

Placement Group Count

The current number of placement groups assigned to the pool.

Size

For replicated pools:

The number of copies from the data objects (= primary + replicas).

Example: **Size** = 3 means that there will be one primary copy and two further copies of each data object.

For erasure coded pools:

The number of placement groups in the storage pool.

Quota	The pool quota of storage size in bytes. The quota was set when the pool was created. "-" indicates, that no quota was set when the pool was created.
Cache Tier Status	**For pools participating in a cache tier:** The pools tier role (**Cache** or **Storage**) and the storage group to which the pool is paired are displayed. **For pools not participating in a cache tier:** "–" is displayed.
Erasure Code Status	
	For erasure coded pools: The profile name used to create the pool is displayed. For replicated pools: "–" is displayed.
Status	Status of the pool. **Running** status indicates that the pool is currently available and operational.
Created by	Indicates where the pool was created (e.g. VSM means "created by VSM").
Tag	Informative tag that may be specified when pool is created.

5.2.2.2 Creating Replicated Storage Pools

To create a new replicated pool:

- Select from the **Cluster** Management main menu the **Manage Pools** function.

  The **All Pools** window will be displayed.
- Click the **Create Replicated Pool** button. The **Create Replicated Pool** dialog will be displayed.

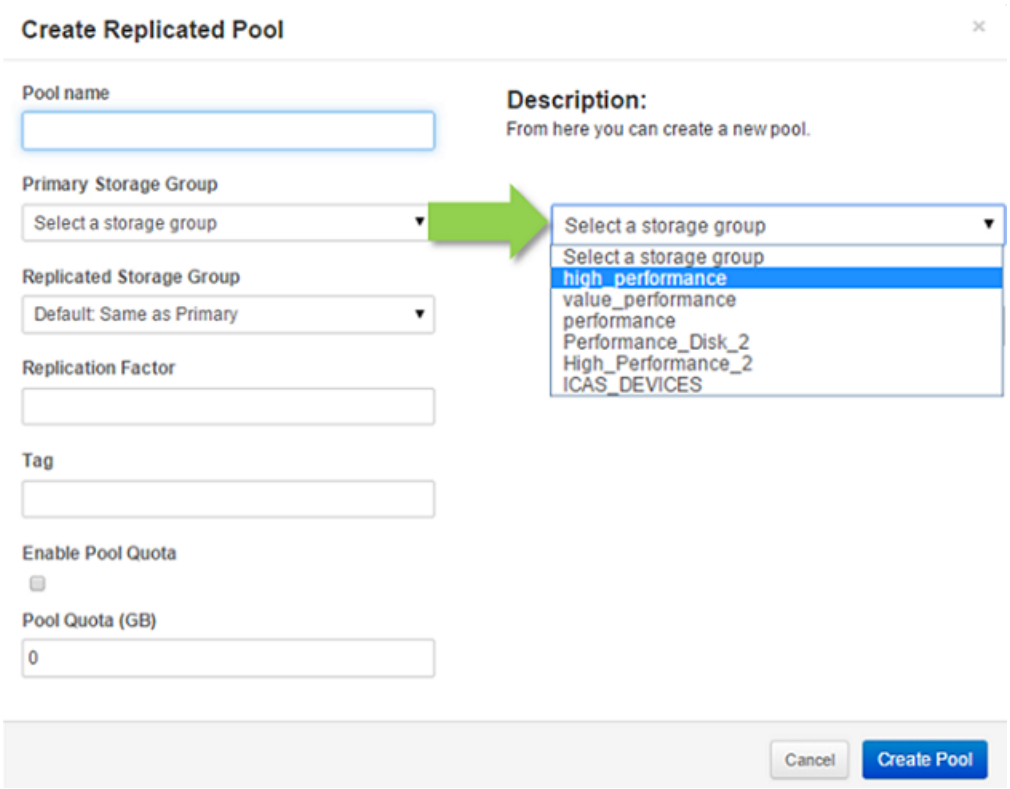


Figure 45: The Create Replicated Pool Dialog

- In the **Pool name** field enter a name for the new pool. The string may only contain ASCII characters and numbers.
- From the **Primary Storage Group** pull-down menu select a storage group where the pool's primary copy of data shall be stored.

  The **Storage Group** list contains all storage groups with IN status.

For detailed information see "Storage Group Status" on page 144.

- If you want the primary copies of data and the non-primary replica data to be placed in separate storage groups select from the **Replicated Storage Group** pull-down menu the storage group where the pool's replica data shall be stored.

  If the primary copies of data and the non-primary replicas are to be placed in the same storage group, set the **Replicated Storage Group** list to **Same as Primary** (default).

  Example:

  You can place the primary copy in a storage group comprised of SSDs, while non-primary replicas are stored in a storage group comprised of 7200 RPM rotating disks.

  In this configuration, data will be read with very high performance from the primary OSDs placed on SSDs, while the non-primary replicas will be written with the usual performance of HDDs (because all copies must be written before the write operation is signaled as complete).

  When using separate storage groups to store primary and non-primary copies, the disks belonging to the "primary storage group" and the disks belonging to the "non-primary storage group" must be located on separate servers in order to ensure that primary and replica data copies do not reside on the same server.

- In the **Replication Factor** field enter the number of copies from each data object that shall be stored on the pool. The number of copies must be ≥ 3.

  To ensure that the replica data reside on separate servers, the creation of the new storage pool will fail, if:

  the number of servers in each of the selected storage group (s) is **≤** the **Replication Factor**. A corresponding error message will be displayed.

- In the **Tag** field enter a descriptive string by which you can identify the pool.
- Optionally you can set a pool quota:

  Activate the **Enable Pool Quota** checkbox and enter the desired pool quota into the **Pool Quota** field.

  By default this field is not checked,therefore quota is not set by default.
- Click the **Create Pool** button.

VSM will choose default values for the further pool attributes, e.g. the number of placement groups. This number is derived from the calculation:

```
Target PG Count=[
pg_count_factor] * number of OSDs in the pool / Replication Factor
```

[pg_count_factor] is specified in the cluster manifest file. The default value is 100.

When additional OSDs are added to a Storage Group, VSM recalculates the target PG count for each affected replicated pool and updates the number of placement groups when it falls below one half of the value of **`Target PG Count`**.

The pool attributes set by VSM automatically cannot be modified by you via VSM.

5.2.2.3 Creating Erasure Coded Storage Pools

Erasure coded pools can be created as follows:

- Select the **Manage Pools** function from the **Cluster** Management main menu .

  The **All Pools** window will be displayed.
- Erasure coded pools can be created by clicking the Create Erasure Coded Pool button on the **Manage Pools** page. Clicking the **Create Erasure Coded Pool** button will open the **Create Erasure Coded Pool** dialog:.

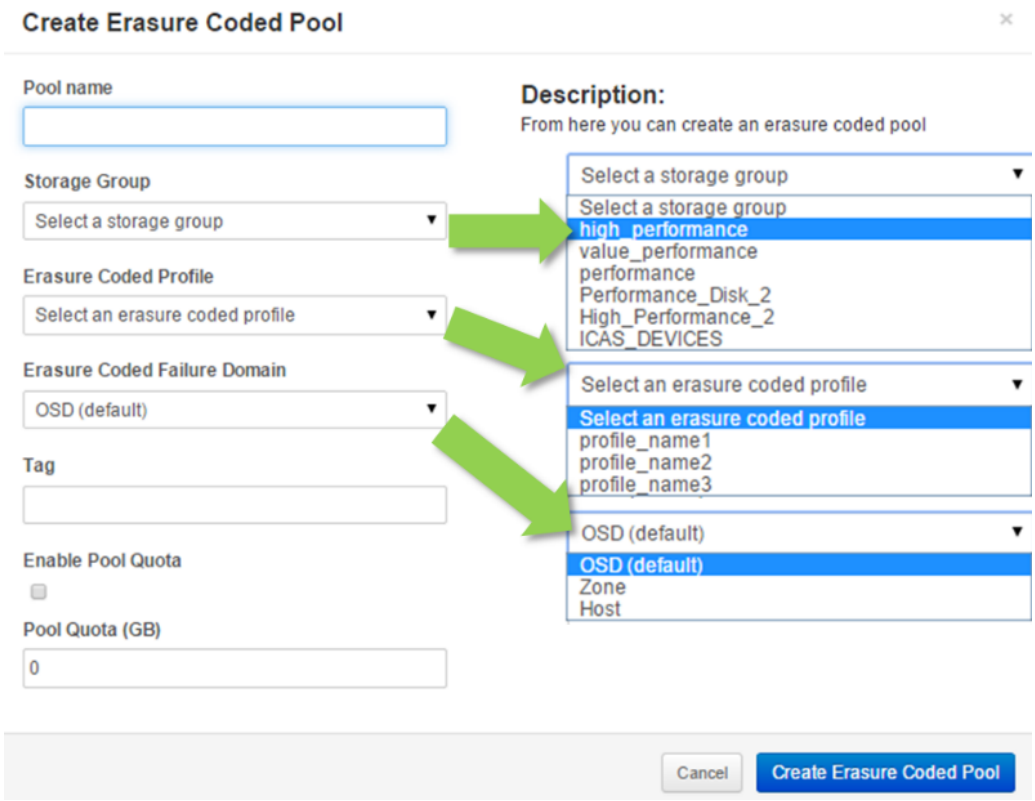


Figure 46: The Create Erasure Coded Pool Dialog

- To create a new Erasure Coded pool, enter the pool name, and select the storage group where the pool's erasure coded data will be placed from the **Storage Group** pull-down menu.
- Select the erasure code profile from the **Erasure Coded Profile** pull-down menu. The pull-down menu lists the storage profiles that are specified in the cluster manifest file.
- Select the erasure code failure domain from the **Erasure Coded Failure Domain** pull-down menu.

  Select **OSD** failure domain to place erasure coded across OSDs.

  Select **Zone** to place erasure coded data across failure zones.

  Select **Host** to place erasure coded data across servers.

  The erasure coded failure domain must be equal to or larger than the K+M value of the profile.

  For example, the failure domain for a K=3 and M=2 erasure code must be 5 or greater; for a cluster with three servers with each server containing 14 disks, a K=3, M=2 erasure code cannot be placed across three servers (**Host** failure domain or **Zone** failure domain), the only valid erasure code failure domain is **OSD**.
- Optionally specify an informative tag for the pool in the **Tag** field.
- Optionally set a quota for the pool by selecting the **Enable Pool Quota** checkbox and specifying the desired pool quota (**Pool Quota** field).
- Complete the operation by clicking on the **Create Erasure Coded Pool** button.

Configuring erasure code profiles

With the default configuration the erasure code profile with name **profile_name1** is available. This profile is described as a profile with K=2 and M=1.

Other profiles can be added to the configuration by excuting the command **ec-profile** on the management node:

```
# ec-profile -c --name profile32 --plugin jerasure
--path /usr/lib64/ceph/erasure-code
--kv-pair '{"k":3,"m":2,"technique":"reed_sol_van"}
```

with

`--name`	Erasure code profile's name.
`--plugin`	Plugin type, such as **jerasure.**
`--path`	The path of plugin.
`--kv-pair`	The key/value pair. This parameter must be JSON Format String. The key/value strings should not have spaces

To list all configured profiles execute the command:

```
# ceph osd erasure-code-profile ls
```

To display the parameters of a profile execute the command:

```
# ceph osd erasure-code-profile get {name}
```

5.2.2.4 Adding Cache Tiers

To create a cache tier:

- Select from the **Cluster** Management main menu the **Manage Pools** function.
  The **All Pools** window will be displayed.
- Click the **Create Cache Tier** button. The **Add Cache Tier** dialog will be displayed.

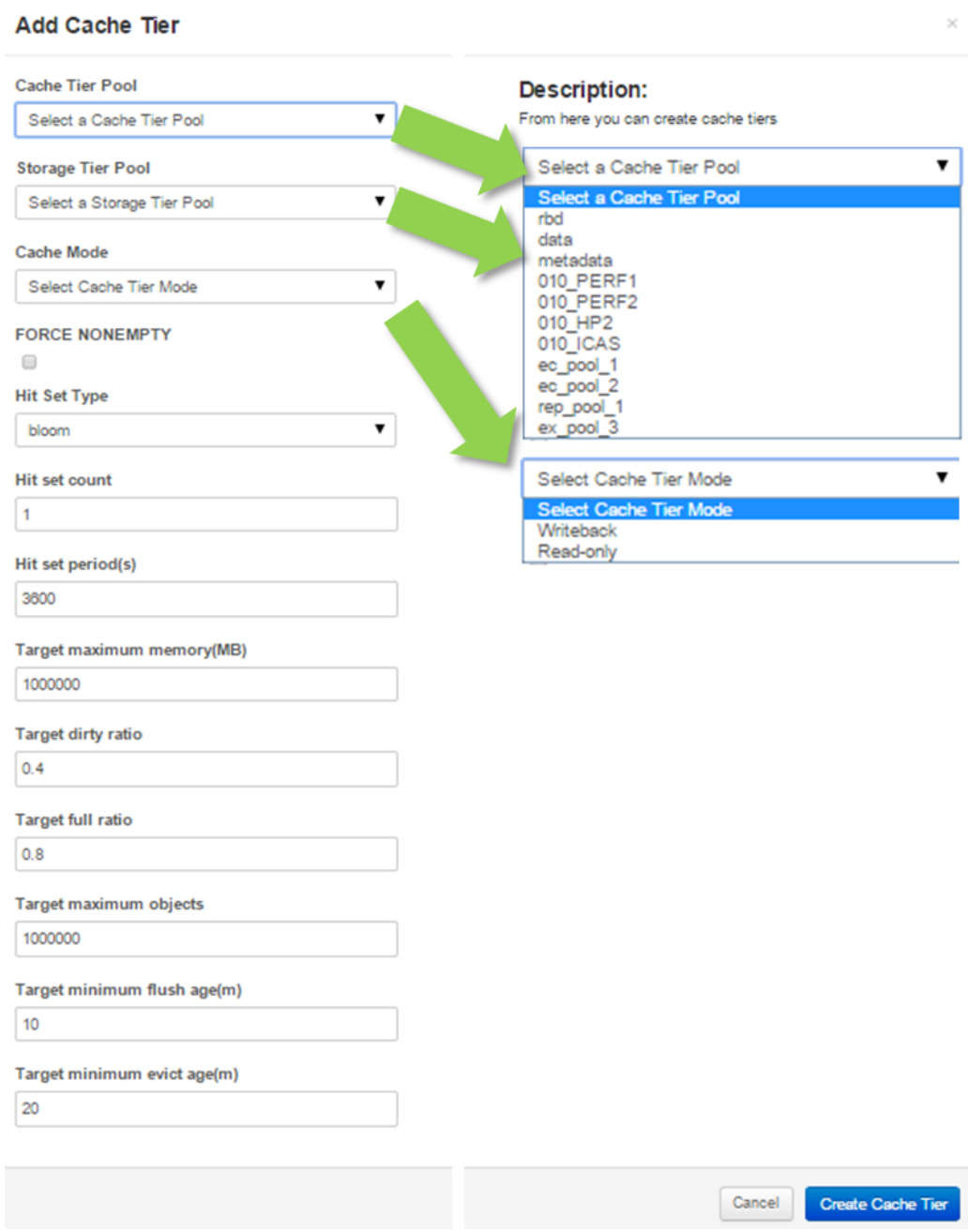


Figure 47: The **Add Cache Tier** Dialog

- From the **Cache Tier Pool** list select a pool that shall act as cache.

  The list contains all pools created in the cluster that are not currently participating in the existing cache tiers.
- From the **Storage Tier Pool** list select a pool that shall act as storage tier.

  The list contains all pools created in the cluster that are not currently participating in the existing cache tiers.

  The creation of a cache tier will fail, if you select the same pool for the cache and storage tier. The following message will be displayed:

  ```
  Error: Failed to add cache tier:
  cache_pool, storage_pool cannot be the same!
  ```
- From the **Cache Mode** list select the cache mode:

Writeback	The Ceph client writes new or modified data to the cache tier and the cache signals the write as complete (ACK). In time (in regular time intervals), the data written to the cache tier migrates to the storage tier and gets flushed from the cache tier. This mode is preferred for mutable data (e.g., RBD data, transactional data,).
Read-only	The Ceph client writes the modified data to the storage tier. When data is read, the corresponding object(s) are copied from the storage tier to the cache tier. This approach is suitable for immutable data (e.g., data that will not be modified, such as pictures/videos on a social network, DNA data, X-Ray imaging). Do not use read-only mode for mutable data.
- Optionally configure the remaining cache tier parameters as required for your application:

  Activate the **FORCE NONEMPTY** checkbox and enter the desired values into the parameter fields.

  For more information on the parameters see http://ceph.com/docs/master/rados/operations/pools/#set-pool-values
- Click the **Create Cache Tier** button.

5.2.2.5 Removing a Cache Tier

To remove the cache tiering associated with a cache pool:

- Select from the **Cluster** Management main menu the **Manage Pools** function.
  The **All Pools** window will be displayed.
- Click the **Remove Cache Tier** button.
  The **Remove Cache Tier** dialog will be displayed.

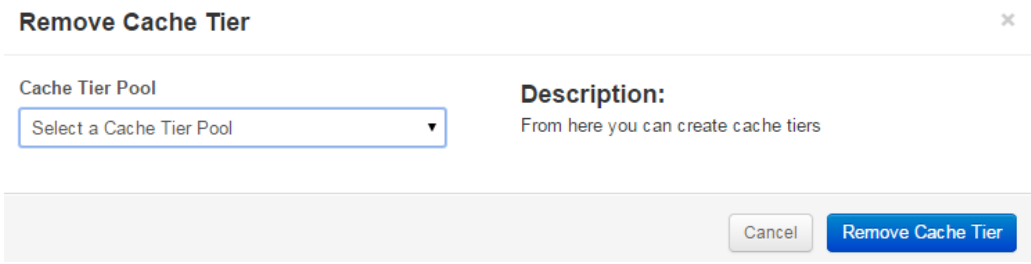


Figure 48: The **Remove Cache Tier** Dialog

- Select the cache pool you want to disable and remove from the **Cache Tier Pool** list.
  If cache mode = **Read-Only**:
  Since a read-only cache does not store the modified data, the cache pool can be disabled and removed immediately without losing any recent changes to the objects in the cache.
  If cache mode = **Writeback**:
  Before a writeback cache will be disabled and removed all modified data written to the cache tier will be flushed to the storage tier of the pool.
- Click on the **Remove Cache Tier** button.
- Confirm the message with **OK**.

Once the cache tiering has been removed the storage tier pool can continue to be used as a non-cached storage pool.
The cache pool will not be deleted.

5.2.3 Managing Pools (vsm_cli)

To manage the storage pools of a cluster the following `vsm_cli` commands are available:

- `set-pool` (see page 278)
- `delete-pool` (see page 262)

Additionally you can use the **vsm_cli** command **list** to check the result of command processing.

5.2.3.1 Modifying the Number of Data Replications on an Existing Storage Pool

When creating a Ceph storage pool it is necessary to define the number of copies from each data object to be stored on the pool. You can change this number of replications with the **vsm_cli** command **set-pool**.

- Enter the **set-pool** command with the option **-r <replication>**.

  Replace **<replication>** with the new number of replications.

  Minimum: 2, **Maximum:** Number of storage nodes in the storage cluster.

  Example:

  In the pool **POOL_1** of **CLUSTER_1** the number of data replicas shall be increased from two to three:

  vsm_cli CLUSTER_1 set-pool -r 3 POOL_1

To check the result of the command, see "Modifying the Number of Data Replications on an Existing Storage Pool " on page 110.

5.2.3.2 Modifying the Quota of an Existing Pool

VSM always creates Ceph storage pools with a default value for the quota of storage which may be allocated to the pool. You can modify this value with the **vsm_cli** command **set-pool**:

- Enter the **set-pool** command with the option `-q <max_bytes>`.
- Replace `<max_bytes>` with the new size of the pool.

```
vsm_cli CLUSTER_1 set-pool -q <max_bytes> POOL_1
```

You can specify the pool size `<max_bytes>` by using K,M,G,T, or P unit for KB, MB, GB, TB or PB (see page 278).

For example:

```
vsm_cli CLUSTER_1 set-pool -q 2.5T POOL_1
```

is equivalent to

```
vsm_cli CLUSTER_1 set-pool -q 2748779069440 POOL_1
```

(2,5 TB = 2748779069440bytes).

To check the result of the command, see "Modifying the Quota of an Existing Pool" on page 110

5.2.3.3 Deleting a Pool from a Storage Cluster

To delete a (replicated or erasure coded) pool from a storage cluster you use the **`delete-pool`** command. The associated storage will be deallocated.

When you delete a pool from a storage cluster all data stored on this pool will be lost!

The functionality of the **`delete-pool`** command is interactive, i.e. it requires your confirmation to delete the pool and the repetition of the pool name.

To delete a pool:

- Enter the command:
  `vsm_cli <cluster_name> delete-pool <pool_name>`
  The following will be displayed:
  `Do you really want to delete pool test? [y/n]`
- Enter `y`.
  The following will be displayed:
  `Type the pool name you want to delete again`
- Enter the pool name **`<pool_name>`** again.

The pool will be successfully deleted from the cluster configuration. You will not have access to the stored data any longer. All data will be lost!

5.2.3.4 Checking result of command processing

To check if the command processing was successful you can use the **`vsm_cli`** command **`list`** with option **`pools`**.

- For the sample cluster **`CLUSTER_1`** enter:
  `vsm_cli CLUSTER_1 list pools`

Output (example):

```
Pool       Type             Replicas   Objects    Data(KB)    Quota(KB)    Uti(%)
-------    ---------------- ---------  -------- ----------- ---------- --------
data       high_performance         2         0           0
metadata   high_performance         3        20           2
rbd        high_performance         2         0           0
Total data                                  20           2
```

For each pool the current number of copies is displayed in the column **`Replicas`**. The allocated quota of storage size is displayed in **`Quota`** `(KB)`.

If you have deleted a pool from the cluster this pool should be removed from the output list.

5.3 Managing Storage Nodes

VSM provides the following functions to manage storage nodes:

- Retrieving information about the storage nodes of a cluster (see page 112)

  You may inform on the management, cluster and public (client) side IP addresses, the number of storage devices and their capacity as well as the status/health of the storage nodes and the cluster.
- Stopping and (re)starting OSDs of a cluster (see page 114)

  You may stop and (re)start the OSDs of particular nodes.
- Adding and removing monitors (see page 115)

  You can remove a monitor from a server and create a new one on another server. Moreover it may be necessary to enlarge the number of monitors.

`vsm_cli` additionally provides functions for the following tasks:

- Stopping and (re)starting particular monitor daemons of a cluster
- Switching storage nodes of a cluster into maintenance mode

5.3.1 Managing Storage Nodes (VSM GUI)

5.3.1.1 Displaying information on storage nodes and their health status

To manage storage nodes

- select from the **Server Management** main menu the function **Manage Servers**.

The **Cluster Server List** is displayed.

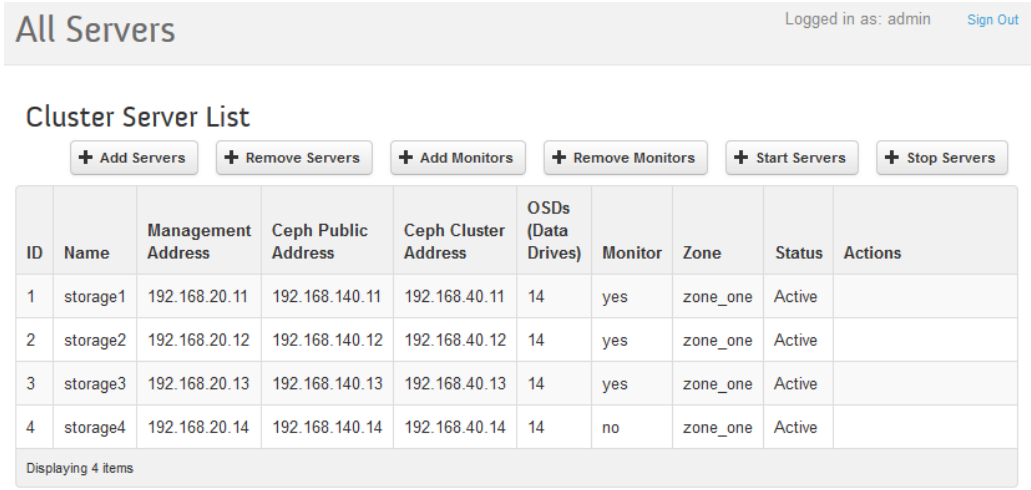


Figure 49: Managing storage nodes

In the **Cluster Server List** all storage nodes which are recognized by the VSM controller are listed.

For each storage node the following is displayed:

Management Address, Ceph Public Address, Ceph Cluster Address

The network configuration of nodes:

IP addresses of the storage node in the administration, public (client) and cluster network.

OSDs The number of disks (equal to the number of OSDs),

Monitor Information, if a monitor is configured for a node (**yes**) or not (**no**).

Zone Not applicable in this version of ETERNUS CD10000

Status The status of the storage node. Possible values are:

Available

The VSM agent is configured and running on this host, but the host is not added to the cluster.

Unavailable

The VSM agent is down. This implies that the host might be shutdown or switched off too.

Active

The storage node is added to the cluster and the OSD daemons are running.

Stopped

On the storage node the VSM agent is running. But all OSD daemons within the storage node are not running.

With the buttons above the list you can perform the following management operations on the storage nodes:

- Starting and stopping storage nodes
- Adding and removing monitors

Add Servers and **Remove Servers** should only be executed by service field engineers.

5.3.1.2 Stopping storage nodes

If you stop a storage node, all OSDs within the storage node will be stopped. Stopping a storage node is possible if its status is **Active**.

- Select from the **Server Management** main menu the function **Manage Servers**.
- Click on the **Stop Servers** button (see figure "Managing storage nodes " on page 112).

  The **Stop Servers** dialog box is displayed. In the dialog box all storage nodes with status **Active** are listed.
- Select the storage node(s) which you want to stop by activating the corresponding option (s) in the leftmost column.

  Please note that in a Ceph storage cluster three monitors must be running.

  VSM will reject stopping a server if one of the last three monitors is running on it. Therefore a minimum of three nodes will keep running in a Ceph storage cluster.
- Click on the **Stop Servers** button

The operation is started on the selected storage node(s) and the success or failure of starting the operation is displayed. The transition of the server status from **Active** to **stopping** to **Stopped** is displayed in the list of servers in the **Status** column.

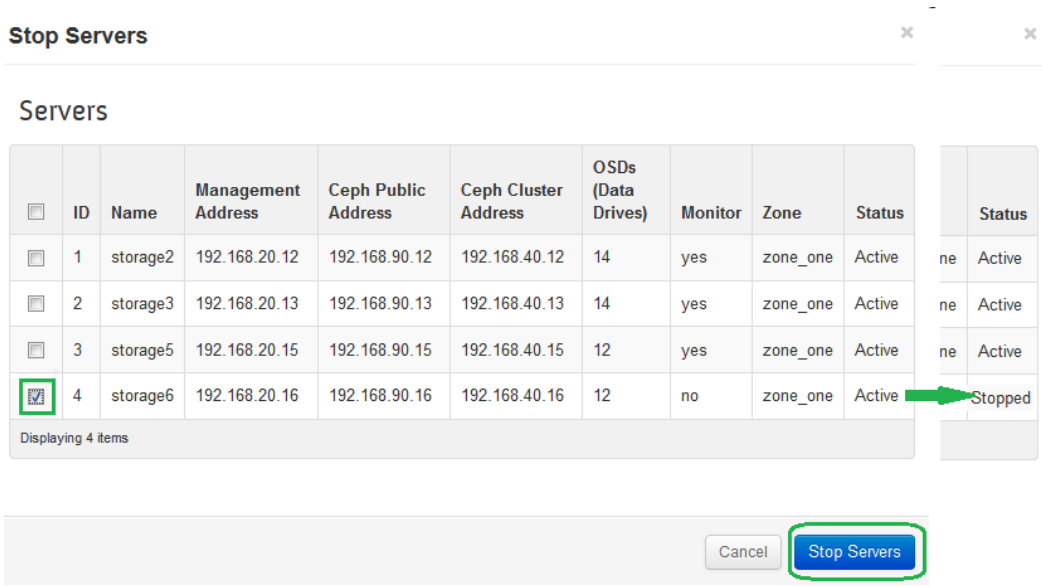


Figure 50: Server status transition

During this operation the OSDs within the storage nodes will be stopped. The "`Noout`" flag will be set up to prevent it from going to status "`out`". This keeps the OSDs in status "`down`"and avoids data rebalancing.

5.3.1.3 Starting storage nodes

Starting a storage node is possible if its status is **Stopped**. During this operation OSDs within the storage node will be started and the "`Noout`" flag will be removed. Data replication will be available again

- Select from the **Server Management** main menu the function **Manage Servers**.
- Click on the **Start Servers** button (see figure "Managing storage nodes " on page 112).

  The **Start Servers** dialog box is displayed. In the dialog box all storage nodes with the status **Stopped** are listed.

- Select the storage node(s) which you want to start by activating the corresponding option (s) in the leftmost column.

  To restart all stopped storage nodes activate the option in the column header,
- Click on the **Start Servers** button

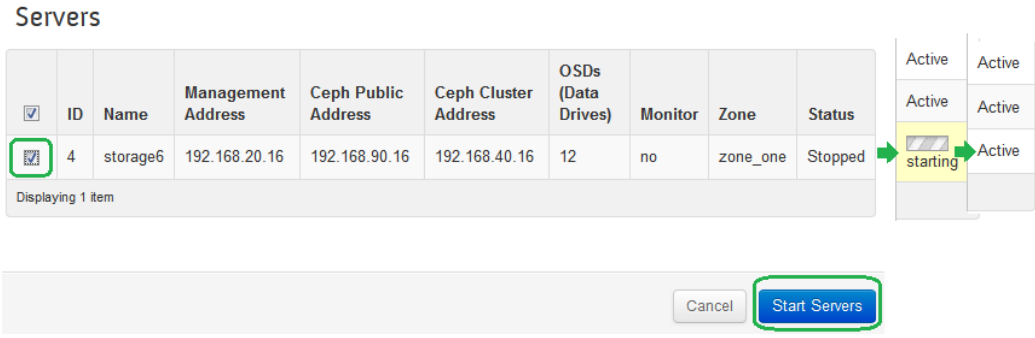


Figure 51: Server status transition

The operation is started on the selected storage node(s) and the success or failure of starting the operation is displayed. The transition of the server status from **Stopped** to **starting** to **Active** is displayed in the list of servers in the **Status** column.

5.3.1.4 Adding monitors to the cluster

You can add a monitor to each node on which currently no monitor daemon is running. Adding a monitor is possible if the status of the node is **Active** or **Available**.

At least three monitors must be active in the cluster. For highest reliability it is advisable to run an odd number of monitors evenly distributed to the physical configuration.

- Select from the **Server Management** main menu the function **Manage Servers**.
- Click on the **Add Monitors** button (see figure "Managing storage nodes " on page 112).
- The **Add Monitors** dialog box is displayed. In the dialog box all nodes with the status **Available** or **Active** and without monitor (**no** is displayed in the **Monitor** column) are listed.
- Select the node(s) on which you want to start a monitor by activating the corresponding option(s) in the leftmost column.
- Click on the **Add Monitors** button.

  If the total number of monitors will be even a warning message is displayed.
- Confirm the message with **OK**.

On each selected node a monitor is started during this operation.

5.3.1.5 Removing monitors from the cluster

Removing a monitor from the cluster is possible if the state of the node is **Active**.

At least three monitors must remain active in the cluster. For highest reliability it is advisable to run an odd number of monitors evenly distributed to the physical configuration.

- Select from the **Server Management** main menu the function **Manage Servers**.
- Click on the **Remove Monitors** button.

  The **Remove Monitors** dialog box is displayed. In the dialog box all nodes with monitor and status **Active** are listed.
- Select the node(s) on which you want to stop the monitor by activating the corresponding option(s) in the leftmost column.
- Click on the **Remove Monitors** button.

  When after executing the **Remove Monitors** command the total number of monitors will be even, a warning message will be displayed.

  If by executing the **Remove Monitors** command less than three monitors remain, a warning message will be displayed and VSM will reject the command.
- Confirm the message with **OK**.

5.3.2 Managing Storage Nodes (vsm_cli)

To manage the storage nodes of a cluster the following **vsm_cli** commands are available:

- **stop** and **start**,
- **maint-on** and **maint-off**

5.3.2.1 Stopping and (re)starting monitor daemons

To (re)start/stop monitor daemons of a cluster the commands **start** and **stop** are available on the **vsm_cli** interface.

- To stop/start a monitor (e.g. mon.2 on cluster CLUSTER_1) enter the command:

```
vsm_cli CLUSTER_1 stop mon.2
vsm_cli CLUSTER_1 start mon.2
```

Please note, that within a Ceph storage cluster three monitors must be running. Therefore VSM will reject stopping a monitor daemon if it is one of the last three monitors within the cluster.

The monitor names are generated when creating a storage cluster.

To figure out the name of a monitor you can use the following **vsm_cli** command:

```
vsm_cli <cluster_name> mon-status
```

5.3.2.2 Switching nodes into maintenance mode

For certain maintaining tasks it could be necessary that the storage node is in a maintenance mode. That means for example, that no I/O (write/read) request are directed to the node while it is in this mode.

The **vsm_cli** provides the **maint-on** command to transfer storage nodes of a cluster into maintenance mode. Switching on the maintenance mode transfers the OSDs into the status "down" (they will not change to "out"!) and suppresses the rebalancing.

Example:

- To transfer the storage node NODE_1 of the cluster CLUSTER_1 into maintenance mode enter the command

```
vsm_cli CLUSTER_1 maint-on NODE_1
```

After completion of the maintenance work use the **maint-off** command to switch off the maintenance mode.

- Enter

```
vsm_cli CLUSTER_1 maint-off NODE_1
```

All daemons will be restarted and the cluster rebalancing will be enabled.

5.4 Managing Storage Groups (VSM-GUI)

To receive basic information on the storage groups and to create new storage groups:

- select from the **Cluster Management** main menu the **Manage Storage Groups** function .

The **Storage Group List** window displays a list of all storage groups within the storage cluster .

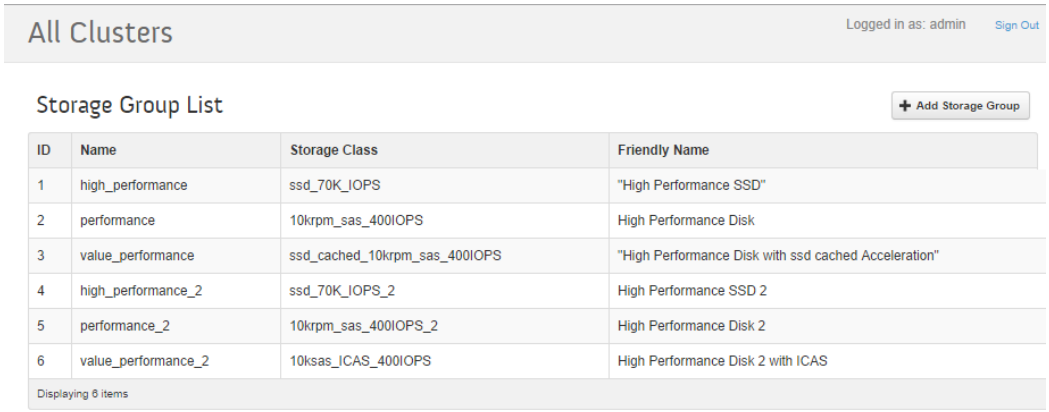


Figure 52: List of the currently defined storage groups

5.4.1 Information on the storage groups

In the columns of the **Storage Group List** the following is displayed for each storage group:

ID	ID of the storage group. Sequence number automatically assigned during creation.
Name	Name of the storage group.
Storage Class	Name of the associated storage class.
Friendly Name	Additional user friendly storage group name that is used at the user interface to identify the user interface.

5.4.2 Creating a new storage group

To create a storage group:

- Select from the **Cluster** Management main menu the **Manage Storage Groups** function.
- Click the **Add Storage Group** button. The **Create Storage Group** dialog will be displayed.

  From here you can create a new storage group.

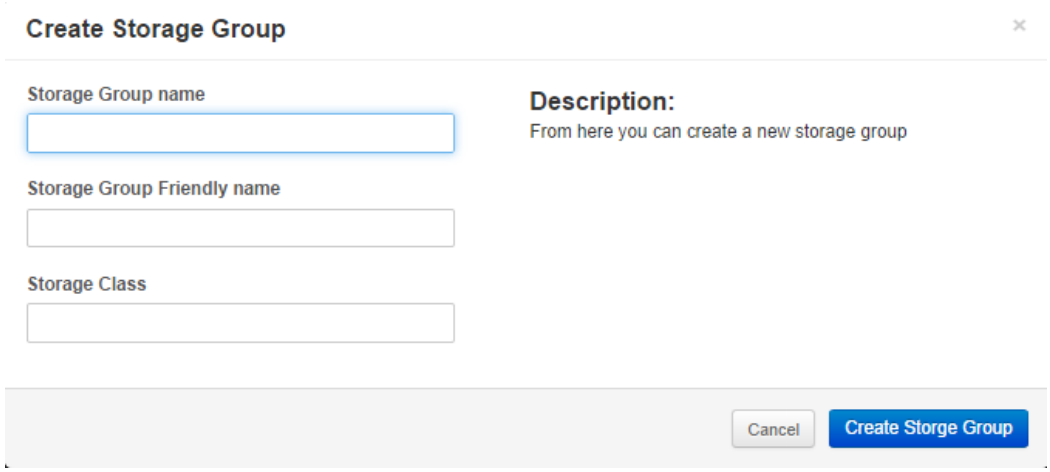


Figure 53: The Create Storage Group Dialog

- In the **Storage Group name** field enter a name for the new storage group (no restrictions on the length and the used characters).
- Additionally you can enter a user friendly storage group name in the **Storage Group Friendly name** field. (no restrictions on the length and the used characters). This name will be used at the user interface.
- In the **Storage Class** field enter the name of the associated storage class. The storage class must be placed in the server manifest file to identify disks that belong to this storage group.
- Click the **Create Storage Group** button.

Adding storage for a new storage class to the ETERNUS CD10000 storage system

You can add storage for a new storage class by adding new storage nodes to the ETERNUS CD10000 storage system.

Add Servers should only be executed by service field engineers.

After creating a new storage group proceed as follows:

- For new servers that will be added to the system, use the storage class name associated with the new storage group in the server manifest file to identify disks that belong to this storage group.

  For detailed information see "Server Manifest File" on page 243.
- Add the new servers to the system using the **Add Servers** button in the **Server Management** > **Manage Servers** window.

5.5 Starting/Stopping a Storage Cluster (vsm_cli)

To start/stop the storage cluster the commands **`start`** and **`stop`** are available on the **`vsm_cli`** interface.

If you stop the whole storage cluster, all OSDs and monitors within the cluster are stopped.

The status of all OSDs change to "`In - Down`" and then to "`Autoout`". No I/O requests will be processed by a stopped storage cluster.

- To stop the whole cluster CLUSTER_1 enter the command:

  `vsm_cli CLUSTER_1 stop`

Check the success of the command processing by calling the **`vsm_cli status`** command.

- To (re)start the cluster CLUSTER_1 enter the command:

  `vsm_cli CLUSTER_1 start`

  Before restarting the storage nodes of a cluster you should check whether all daemons are already down or deactivated.

Information on starting/stopping particular OSDs or monitors with **`vsm_cli`** see "Stopping and (re)starting OSDs" on page 125 and "Stopping and (re)starting monitor daemons" on page 117.

5.6 Managing Disks/OSDs

VSM provides the following functions to manage disks/OSDs:

- Retrieving information about the disks/OSDs of a cluster (see page 122)
- Stopping and (re)starting OSDs of a cluster (see page 124)
- Restoring disks/OSDs (see page 124)

`vsm_cli` additionally provides functions for the following tasks:

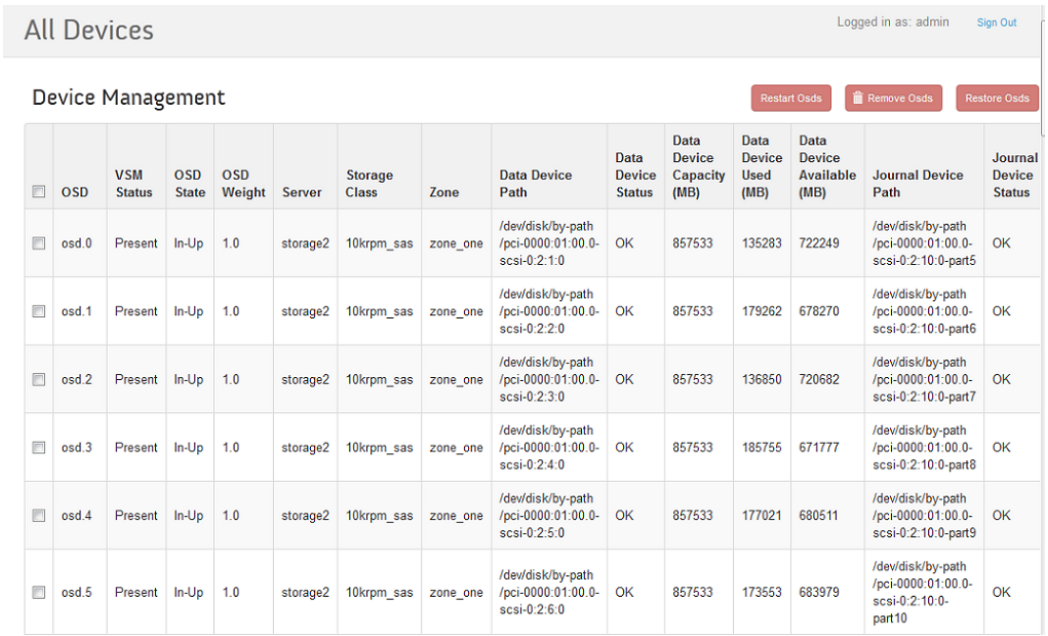
- Stopping and (re)starting particular OSDs of a storage cluster see page 125)
- Changing the crush weight of OSDs (see page 126)

5.6.1 Managing Disks/OSDs (VSM GUI)

To manage the storage devices within a storage cluster

- select from the **Server Management** main menu the function **Manage Devices**.

The **Device Management** window is displayed.



OSD	VSM Status	OSD State	OSD Weight	Server	Storage Class	Zone	Data Device Path	Data Device Status	Data Device Capacity (MB)	Data Device Used (MB)	Data Device Available (MB)	Journal Device Path	Journal Device Status
osd.0	Present	In-Up	1.0	storage2	10krpm_sas	zone_one	/dev/disk/by-path /pci-0000:01:00.0-scsi-0:2:1:0	OK	857533	135283	722249	/dev/disk/by-path /pci-0000:01:00.0-scsi-0:2:10:0-part5	OK
osd.1	Present	In-Up	1.0	storage2	10krpm_sas	zone_one	/dev/disk/by-path /pci-0000:01:00.0-scsi-0:2:2:0	OK	857533	179262	678270	/dev/disk/by-path /pci-0000:01:00.0-scsi-0:2:10:0-part6	OK
osd.2	Present	In-Up	1.0	storage2	10krpm_sas	zone_one	/dev/disk/by-path /pci-0000:01:00.0-scsi-0:2:3:0	OK	857533	136850	720682	/dev/disk/by-path /pci-0000:01:00.0-scsi-0:2:10:0-part7	OK
osd.3	Present	In-Up	1.0	storage2	10krpm_sas	zone_one	/dev/disk/by-path /pci-0000:01:00.0-scsi-0:2:4:0	OK	857533	185755	671777	/dev/disk/by-path /pci-0000:01:00.0-scsi-0:2:10:0-part8	OK
osd.4	Present	In-Up	1.0	storage2	10krpm_sas	zone_one	/dev/disk/by-path /pci-0000:01:00.0-scsi-0:2:5:0	OK	857533	177021	680511	/dev/disk/by-path /pci-0000:01:00.0-scsi-0:2:10:0-part9	OK
osd.5	Present	In-Up	1.0	storage2	10krpm_sas	zone_one	/dev/disk/by-path /pci-0000:01:00.0-scsi-0:2:6:0	OK	857533	173553	683979	/dev/disk/by-path /pci-0000:01:00.0-scsi-0:2:10:0-part10	OK

Figure 54: Managing storage devices

5.6.1.1 Displaying information on disks/OSDs and their status

In the **Cluster Server List** all disks/OSDs which are recognized by the VSM controller are listed.

For each disk/OSD the following is displayed:

OSD Name of the OSD.

When creating a storage cluster a name OSD.<number> is assigned to each OSD. The number reflects the order in which the OSDs are added to the cluster.

VSM Status VSM status. Possible values:

`Present`
: The OSD is present in the storage cluster

`Removed`
: The OSD has been removed.

OSD Status The current status of the disk/OSD. Possible values are:

`In-Up`

The OSD is an active part of the storage cluster (`in`). The OSD daemon is running. The OSD is fully functional and available for I/O (`up`).

`In-Down`

The OSD is part of the storage cluster, but no I/O is being directed to it.

Existing data on the disk will not be rebalanced. The disk may need to be restarted.

`Out-Up`

The OSD is fully functional and available for I/O (`up`), but no active part of the storage cluster (`out`).

All data on this disk will be relocated to other disks. CRUSH does not assign placement groups to the OSD.

`Out-Down`

The OSD is no active part of the storage cluster. No I/O is directed to it. CRUSH does not assign placement groups to the OSD.

Existing data on the disk will not be rebalanced.

`Out-Down-Autoout`

The OSD is currently no active part of the cluster. No OSD daemon is running. No I/O requests are directed to the disk.But VSM has recognized the disk.

To reconnect the OSD to the storage cluster the OSD "Restarting disks/OSDs" on page 124

OSD Weight Weight assigned to the OSD to control the amount of data the OSD is responsible for storing. For detailed information, see "CRUSH Data Placement" on page 29.

Server Host name of the storage node on which the OSD resides.

Storage Class Storage Class to which the disk belongs.

Zone Not applicable in this version of ETERNUS CD10000

Data Device Path OSD Drive path in by-path notation.

Data Device Status Result of OSD drive path check.

Data Device Capacity

The formatted capacity of the disk, specified in MB.

Data Device Used Currently used disk capacity, specified in MB.

Data Device Available

Currently available disk capacity, specified in MB.

Journal Device Path

Device path of the disk's journal device in by-path notation.

Journal Device Status

Result of journal device drive path check.

With the buttons above the list you can perform the following management operations on the disks/OSDs:

- Restart disks/OSDs.
- Restore disks/OSDs.

Remove Osds should only be executed by service field engineers.

5.6.1.2 Restarting disks/OSDs

Restarting an OSD is possible if its status is **Autoout**.

- Select from the **Server Management** main menu the function **Manage Devices**.
- Click the header of the **OSD Status** column to sort the list by the OSD status.
- Select the OSD(s) with status **Out - Down-Autoout** you want to restart by activating the corresponding option(s) in the leftmost column.
- Click the **Restart Osds** button. A warning will be displayed.
- Click the **Restart Osds** button to confirm the warning. The operation is started

The operation will be successful if the OSD status in the list changes to **in-up**.

5.6.1.3 Restoring disks/OSDs

Restoring an OSD adds an OSD which was removed before back to the storage cluster (**VSM status** = **Removed**).

Remove Osds should only be executed by service field engineers.

- Select from the **Server Management** main menu the function **Manage Devices**.
- Click the header of the **VSM Status** column to sort the list by the VSM status.

- Select the OSD(s) with VSM status **Removed** you want to restore by activating the corresponding option(s) in the leftmost column.
- Click the **Restore Osds** button. A warning will be displayed.
- Click the **Restore Osds** button to confirm the warning. The operation is started.
- Verify that the VSM status has changed to **Present**.

5.6.2 Managing Disks/OSDs (vsm_cli)

To manage the storage devices of a cluster the following **vsm_cli** commands are available:

- **list disks** and **list disks-of-nodes**,
- **stop** and **start**,
- **replace-disk-out** and **replace-disk-in**
- **crush-reweight**

5.6.2.1 Displaying information on disks/OSDs and their status

- Information on the storage nodes is displayed if you enter the command

  ```
  vsm_cli <cluster_name> list disks
  ```

- You can limit the output on the disks within particular storage nodes by using the `disks-of-nodes` option. For example:

  ```
  vsm_cli CLUSTER_1 list disks-of-nodes storage4
  ```

For details about the **list** command see page 266.

For information on the output "OSD Status and Utilization " on page 158.

5.6.2.2 Stopping and (re)starting OSDs

To stop/restart particular OSDs of a cluster the commands **start** and **stop** are available on the **vsm_cli** interface.

- To stop particular OSDs within CLUSTER_1 (e.g. osd.4 and osd.16) enter the command:

  ```
  vsm_cli CLUSTER_1 stop osd.4 osd.16
  ```

- To (re)start the OSDs enter the command:

  ```
  vsm_cli CLUSTER_1 start osd.4 osd.16
  ```

Before restarting the disk of a cluster you should check whether the daemon is already down or deactivated.

5.6.2.3 Changing OSDs Crush weight

Crush weights are assigned to OSDs to control the amount of data that the OSDs shall store. Data is uniformly distributed among these weighted devices.

The CRUSH weight affects the number of I/O operations and the storage utilization of a particular OSD. Generally, OSDs with the same weight will have similar capacity utilization. OSDs with a higher capacity may have a higher weight.

You can use the **`crush-reweight`** command to adjust the weight of an OSD up or down.

The current weight of an OSD can be ascertained from the **`list osd-tree`** command ("OSD Tree" on page 162).

Example:

To change the weight of the OSD "osd.03" that resides on the storage node "node001" from 1 to 1.2 enter the command

```
vsm_cli CLUSTER_1 crush-reweight node001 osd.03 1.2
```

5.7 Managing OpenStack Interconnection (VSM)

5.7.1 Attaching an RBD Pool to OpenStack

To attach an RBD pool to an OpenStack configuration, proceed as follows:

- In the navigation of the VSM GUI, select **Manage OpenStack** > **Manage RBD Pools**.
- Select the corresponding pool from the **Manage OpenStack RBD Pools** list.

  The attach status of the RBD pools is displayed in the **Attach Status** column of the list.

  In the **Created By** column, the creator (**VSM** or **Ceph** that is the RBD pool is created outside of VSM) is displayed.
- Above the **Manage OpenStack RBD Pools** list, click the **Present Pools** button. The **Present Pools** dialog box opens containing a list of all RBD pools with the attach status **no**.

  Only valid servers with the status **running** are listed.
- In the first column of the **Present RBD Pools** list, select the check box(es) belonging to the RBD pool(s) that you want to attach to the OpenStack configuration.
- Click the **Present Pools** button to confirm the attachment.

5.7.2 Managing OpenStack Access

VSM is able to communicate with the OpenStack nova controller to notify the nova controller about a new pool which can be used by OpenStack. To activate this function you must configure a connection between the VSM controller and the OpenStack nova controller.

Only RBD are supported, it is not possible to attach object storage to OpenStack.

Setting up the communication path

Prerequisites

The VSM controller must be able to access the Nova controller via network path, as the communication is done via a ssh connection without entering a password.

- Make sure that one of the storage nodes can connect to the OpenStack nova controller.
- On one of the storage nodes open the file `/root/.ssh/authorized_keys`.
- Make sure that this file contains all `id_rsa.pub` keys of all storage nodes. As the default, after the setup of VSM this file contains the keys. All storage node names must be listed in this file.
- Add the content of this file to the `/root/.ssh/authorized_keys` file on the OpenStack nova controller.
- Make sure that any storage node can connect to the OpenStack nova controller.

Activating the OpenStack access via VSM

Prerequisites

An established ssh connection is required.

- In the navigation of the VSM GUI, select **Manage OpenStack** > **OpenStack Access**.
- Above the **Manage OpenStack Access** list, click the **Add OpenStack Nova Controller** button.
- In the **Add OpenStack Nova Controller** dialog box, enter the IP address of the OpenStack nova controller.
- Click the **Add IP** button to confirm the action.

Changing an OpenStack access point

- In the navigation of the VSM GUI, select **Manage OpenStack** > **OpenStack Access**.
- In the **Manage OpenStack Access** list, click the **Edit** button of the appropriate access point.
- In the **Edit** dialog box, specify the new IP address in the appropriate field.
- Click the **Submit** button, to confirm the new access point.

Deleting an OpenStack access point

- In the navigation of the VSM GUI, select **Manage OpenStack** > **OpenStack Access**.
- In the **Manage OpenStack Access** list, click the **Delete** button of the appropriate access point.
- In the **Delete** dialog box, click the **Delete** button, to confirm the action.

5.8 Managing VSM Users

With VSM you can add or remove a user and change a password.

Creating a user

The default user `admin` is configured in VSM. You can locally add additional users. It is not possible to use your existing Active Directory or equivalent configuration.

- In the navigation of the VSM GUI, select **Manage VSM** > **Add/Remove User**.
- Above the **User List**, click the **Create User** button.

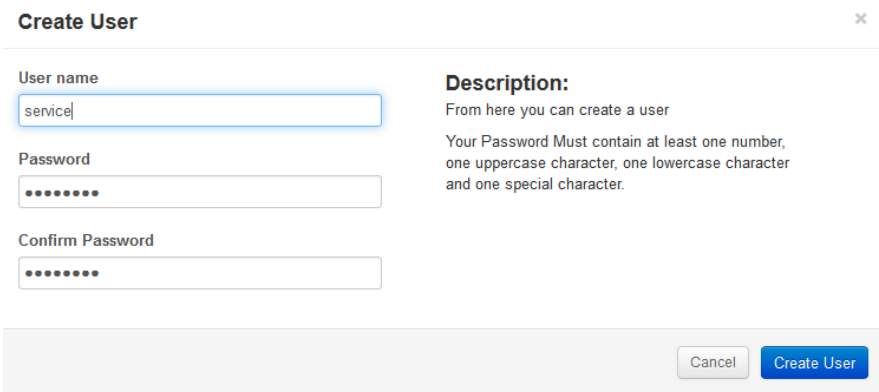


Figure 55: Creating a user

- In the **Create User** dialog box, specify the user name and a password in the appropriate fields. A password must consist of 8 or more characters. One numeric, one lower case character, one upper case character and one special character (!@#$%^&*().:;~\|[]{}) must be included.
- Click the **Create User** button, to confirm the new user.

Deleting a user

Figure 56: Deleting a user

- In the navigation of the VSM GUI, select **Manage VSM** > **Add/Remove User**.
- In the **User List**, click the **More** button of the appropriate user.
- Select the **Delete User** option, to remove the user.

The predefined user "admin" cannot be deleted.

Changing a password

- In the navigation of the VSM GUI, select **Manage VSM** > **Add/Remove User**.
- In the **User List**, click the **Change Password** button of the appropriate user.
- In the **Change Password** dialog, enter the new password in the **Password** field and in the **Confirm** field. A password must consist of eight or more characters. One numeric, one lower case character, one upper case character, and one special character (!@#$%^&*().:;~\|[]{}) must be included.
- Click the **Update User** button, to change the password.

5.9 Authentication Configuration for External User Account Database

CentOS give possibility to log-in to system not only for users created on operating system but is also possible to use LDAP, AD other user account database. To configure authentication method on management node, there are two possibility:

- by CLI command
- by CentOS GUI

To configure authentication method on management node by CLI command:

- setup LDAP client

Synopsis:

cd10000 setup authentication ldap -s LDAP_SERVER -b LDAP_BASED N

Parameters:

`<-s | --ldap-server>` LDAP server hostname

Example:

–ldapserver fji.server.world

`-b | --ldap-basedn - LDAP base DN`

Example:

`--ldapbasedn "dc=server,dc=world"`

- setup AD client

  Synopsis:

  cd10000 setup authentication ad -s SHORT_AD_DOMAIN -l LONG_AD_DOMAIN -u USER -p PASSWORDwhere:

  Parameters:

  `<-s | --short-ad-domain>` AD short domain name (all caps)

  Example:

  –short-ad-domain SERVER.WORLD

  `-l | --long-ad-domain - AD long domain name (all caps)`

  Example:

  –long-ad-domain FJI.SERVER.WORLD

  `-u | --user - user authorized to join computers to domain`

  Example:

  –user TEST

  `-p | --password - password for user`

  Example:

  `--password test`

To configure authentication method on management node by CentOS GUI see the external file: Authentication Configuration

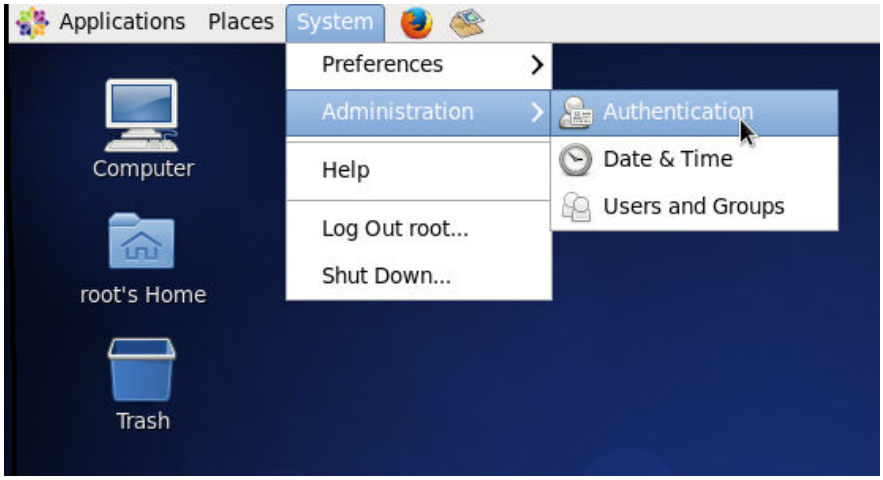


Figure 57: Desktop view

Figure 58: Authentication configuration - GUI application

6 Backup and Restoration of Management Node

ETERNUS CD10000 is equipped with one management node. If this management node must be replaced, a backup of the system should be available.

For this purpose the Fujitsu Backup and Restoration Tool (**`fbart`**) is created. It provides backup and restore of the management node of the Fujitsu ETERNUS CD10000. Multiple backups according to a predefined schedule are created automatically.

The restoration process in case of failure is also greatly simplified by **`fbart`**.

`fbart` uses multiple storage nodes of an ETERNUS CD10000 cluster to store two redundant copies of each backup, further reducing the chances of a data loss.

The backups include all the configuration data of the management node which are needed for a complete restoration.

After an ETERNUS CD10000 system is setup at the customer site the backup of the management node is automatically enabled with default values.

For detailed information on the default values, see "Configuration and Administration of a Storage Cluster " on page 93.

The default configuration can be changed with the command **`fbart config`**.

6.1 fbart call format

The backup and restoration tool is called by entering the **fbart** command as follows:

```
fbart [<options>] <operation>
```

Where:

<options> Specifies one or multiple of the following options:

--help | -h
: displays a list of all **fbart** operations with their syntax.

--verbose | -v
: prints the log messages on the screen

<operation> **fbart** operation to be executed. Supported operations are:

config
: starts the configuration tool, that can be used to configure the backup and restoration process

backup
: starts the backup process

restore
: restores the system to a previous state

Example:

Creating a backup and displaying the logs on the screen:

```
fbart –v backup
```

6.2 fbart functions

The `fbart` tool provides the following functions:

- Creating or editing the **`fbart`** configuration
- Creating a backup manually
- Restoring a backup
- Deleting expired backups automatically

6.2.1 Creating / Editing the fbart Configuration

`fbart` can create and edit the **`fbart`** configuration file where all parameters used for the backup and restoration operations are stored.

This configuration file should not be edited manually and can be personalized using the following fbart operation:

```
fbart config
```

This command runs the **`fbart`** configuration tool which will successively ask you to enter your individual settings for the recommended configuration parameters.

The current value of each parameter will be shown between brackets.

- To change a value, override it and press the **Enter** key.
- If you want to keep the currently displayed parameter value unchanged, do not write anything and just press the **Enter** key. This will leave the current value unchanged.

The following list shows the `fbart` configuration parameters and their recommended default values:

`Host1 = [Storage1]`

Replace **`Storage1`** with the name of the first host where the backups will be stored. This host must be a storage node on the ETERNUS CD10000 storage cluster.

`Host2 = [Storage2]`

Replace **`Storage2`** with the name of the second host where the backups will be stored. This host must be a storage node on the ETERNUS CD10000 storage cluster.

`Frequency = [8]` Enter the backup frequency in hours (integer >0).

Default: 8

Backups are created incrementally which allows them to fit in a much reduced disk space. This allows for a higher backup frequency to avoid any data loss.

`Expiration= [7]` Enter the number of days, after which the backups shall expire.

Default: 7

This default value creates a good balance of data security and storage costs.

`Destination_Folder = [/backup]`

Enter the name of a folder on the root file system of both hosts, on which the backups shall be created.

Default: `/backup`

`Source_folder = [/etc, /opt]`

List the folders that shall be included in the backup. The list items can be separated by either a comma, a space or a combination of both. The folder name(s) must not include spaces or commas.

Default:

The folders **`/etc`** and **`/opt`** will be included in the backups.

`Excludes = [*.tmp, *~]`

List the files, that shall be excluded from the backups. The list items can be separated by either a comma, a space or a combination of both. Wildcards can be used for text substitution.

Default: temporary files

6.2.2 Creating a Backup Manually

The command **fbart** allows not only to create backups automatically according to the configured frequency but also to create a backup manually anytime you want..

A backup is created by executing the following command:

```
fbart backup
```

fbart will use the available configuration to create a backup. The parameters that will be used include the following:

- the source folders that shall be backed up (**Source_Folder**),
- the destination hosts (**Host1**, **Host2**) and
- the backup folders on the hosts (**Destination_Folder**).

6.2.3 Restoring a Backup

When the setup of the cluster has been made without setting an external NTP server, it is required to set the correct date and timezone settings on the management node before executing **fbart restore**.

Example command: `date -s "Sat Nov 16 07:11:20 CET 2014"`

To restore your system to a previous state, the restoration tool can be used.

This tool will search the hosts listed on the configuration file for backups and let the users select one to restore.

Once a backup has been selected, it will overwrite the current folders, replacing their content.

The restoration tool can be called by the following command:

```
fbart  restore
```

Please be aware that the restoration process will also restore the old **fbart** parameters for configuration.

You have to reboot the management node after the restore has been completed.

6.2.4 Deleting of Expired Backup Files Automatically

Each time the **`fbart`** backup script is running, the tool checks if old backups are available on the host machines. Backups that are older than the expiration delay set in the configuration file will be deleted automatically .

7 Monitoring

Once you have a running storage cluster, you may use the VSM GUI to monitor the Ceph related parts of the storage cluster. See "Monitoring with VSM GUI" on page 139 for a detailed description. The command line interface vsm_cli provides monitoring functionality too, see "The Command Line Interface vsm_cli" on page 247.

Furthermore SNMP monitoring can be configured, see "SNMP" on page 203.

Log file generated by the systems are collected centrally on the management node and can be used for monitoring purposes as well. See "Log Manager" on page 197.

7.1 Monitoring with VSM GUI

Once you have a running storage cluster, you may use the VSM GUI to monitor the storage cluster.

VSM provides an overview of the health status of the storage cluster on the **Dashboard** overview.

For displaying a more detailed status of the single components of a storage cluster, for example, storage pools, OSDs, Monitors, placement groups (PG), select the corresponding **Status** option in the navigation of the VSM GUI under **Monitor Cluster**.

For detailed information on the VSM GUI, see "VSM Main Window" on page 57.

The displayed statuses are refreshed in intervals.

You will find a table containing the respective update intervals of the cluster information and the Ceph commands that the VSM pages use to collect this cluster information in "VSM Data Sources and Update Intervals" on page 140.

Monitoring a storage cluster involves checking the following issues:

Health status of the storage cluster	See "Health Status of the Storage Cluster (VSM GUI)" on page 141
Storage group status	See "Storage Group Status" on page 144
Storage pool status	See "Pool Status" on page 146.
OSD (disk) status	See "OSD Status" on page 148
Monitor status	See "Monitor Status" on page 150.

MDS status	MDSs are configured, when the Ceph file system is used. The Ceph file system is not supported in V1.0 SP2.
Placement group status	See "PG Status" on page 152.
RBD status	See "RBD Status" on page 155

7.1.1 VSM Data Sources and Update Intervals

The following table shows the Ceph commands that each of VSM pages use to collect and display cluster information, and their respective update intervals.

VSM Page	Source – Ceph Command	Update Intervals
Cluster Status		
Summaries	ceph status –f json-pretty	1 minute
Warnings and Errors	ceph status –f json-pretty>	1 minute
Health	ceph health	1 minute
Storage Group Status	ceph pg dump osds -f json-pretty	10 minutes
Pool Status		
Object and R/W stats	ceph pg dump pools -f json-pretty	1 minute
Size & PG Count	ceph osd dump -f json-pretty	1 minute
Client I/O rates	ceph osd pool stats –f json-pretty	1 minute
OSD Status		
Summary data	ceph status –f json-pretty	1 minute
OSD State	ceph osd dump -f json-pretty	10 minutes
CRUSH weight	ceph osd tree –f json-pretty	10 minutes
Capacity stats	ceph pg dump osds -f json-pretty	10 minutes
Monitor Status	ceph status –f json-pretty	1 minute
PG Status		
Summary data	ceph status –f json-pretty	1 minute
RBD Status	rbd ls -l {pool name} --format json --pretty-format	30 minutes
MDS Status	ceph mds dump -f json-pretty	1 minute

7.1.2 Health Status of the Storage Cluster (VSM GUI)

An overview of the health status of the storage cluster is displayed on the **Dashboard**.

After you have logged in to the Virtual Storage Manager for Ceph (VSM) or after you have successfully created the storage cluster, the main window, that is the **Dashboard** overview, is displayed in your browser.

It shows the **Cluster Status**, i.e. several summaries concerning the different cluster components:

- **Cluster Summary**,
- **Storage Group Summary**,
- **Monitor Summary**,
- **OSD Summary**,
- **PG Summary**,
- **Cluster Health Summary**.

If you click one of the **Status** buttons, the **Status** window of the appropriate cluster component is displayed.

Furthermore, **Warnings and Errors** are displayed for servers that are not synchronized with NTP.

To open the **Dashboard** overview, select **Dashboard** > **Cluster Status** in the navigation of the VSM GUI .

All displayed information is listed in detail below.

Cluster Summary

The name and the current health status of the storage cluster are displayed. Example:

```
Cluster Name: CD10000
Status: HEALTH_WARN
```

The possible values of the health status and their description, see "Monitor Summary" on page 142.

Storage Group Summary

The following information is displayed:

`Total Storage Groups`
: Number of storage groups

`Storage Groups Near Full`
: Number of storage groups that have the status `Near Full`

`Storage Groups Full`
: Number of storage groups that have the status `Full`

In a healthy storage cluster no storage groups have one of the status `Near Full` or `Full`.

Monitor Summary

The following information is displayed:

`Monmap Epoch`
: Number of state changes of the monitor map.

For detailed information on the monitor map, see "Ceph Overview" on page 28.

`Monitors`
: Odd number of monitors. Each storage cluster must comprise a minimum of three monitors.

`Election Epoch`
: Specifies the map version. This describes the version of monitor ranking.

`Quorum`
: For high availability, a Ceph storage cluster requires at least three monitors. Ceph uses the Paxos algorithm, which requires a consensus among the majority of monitors in a quorum. A majority of monitors must be counted as such: 1:1, 2:3, 3:4, 3:5, 4:6, etc.

`Overall Status`
: Overall status of the storage cluster. Possible values are:

HEALTH_OK
: The storage cluster is fully operational.

HEALTH_WARN
: One or more disks have the status `down`.

HEALTH_ERR
: One or more disks have the status `out`.

VSM Status

The following information is displayed:

`Uptime`	Period of time that VSM is running, specified in seconds

OSD Summary

The following information is displayed:

`Osdmap Epoch`	Number of state changes of the OSD map. For detailed information on the OSD map, see "ETERNUS CD10000 Architecture" on page 27.
`Total OSDs`	Number of OSDs
`OSDs up`	Number of OSDs with the status `up`
`OSDs in`	Number of OSDs with the status `in`

In a healthy storage cluster all OSDs are `up` and `in` (`Total OSDs` = `OSDs up` = `OSDs in`).

For detailed information on the mentioned OSD status, see "OSD Status" on page 148.

MDS Summary

MDSs are configured, when the Ceph file system is used. Ceph file system is not supported in V1.0 SP2.

PG Summary

The following information and a summary of the current PG states is displayed. For the possible values, see "PG states" on page 152.

`PGmap Version`	Version of the PG map. For detailed information on the PG map, see "ETERNUS CD10000 Architecture" on page 27.
`Total PGs`	Number of placement groups

In a healthy storage cluster the majority or the placement groups must have the status `PGs active+clean`.

Cluster Health Summary

The following information is displayed (example):

```
1558 pgs degraded
55 pgs down
3 pgs incomplete
19777 pgs peering
3 pgs stale
3891 pgs stuck inactive
3 pgs stuck stale
5698 pgs stuck unclean
recovery 1/60 objects degraded (1.667%)
```

For a detailed information on the possible values of the PG status, see "PG Status" on page 152 and "PG States" on page 337

7.1.3 Storage Group Status

An overview of the configured storage groups with the status of the available and used storage capacity is displayed in the **Storage Group Status** window.

To open the **Storage Group Status** window, select **Monitor Cluster** > **Storage Group Status** in the navigation of the VSM GUI. The window comprises the **Storage Groups** area and the **Storage Group List**.

Storage Groups area

The used storage capacity per storage group is visualized with a pie chart in each case. Furthermore, under **Capacity Warning Thresholds**, the **Storage Group Full** and **Storage Group Near Full** thresholds are displayed amongst others. You can configure these thresholds in the cluster manifest file. For detailed information on the manifest file, see "Cluster Manifest File" on page 234.

Storage Group List

The following information is displayed:

`Storage Group`	Name of the storage group
`Attached Pools`	Number of pools attached to the storage group
`Capacity Total (GB)`	Total storage capacity of all disks in the storage group, specified in GB
`Capacity Used (GB)`	Storage capacity that is used replicas included, specified in GB

`Capacity Available (GB)`	Remaining capacity, specified in GB
`Percent Used Capacity (%)`	Storage capacity that is used replicas included, specified in percent
`Largest Node Capacity Used (GB)`	Used storage capacity of the largest storage node, specified in GB If the `Largest Node Capacity Used` is bigger than the `Capacity Available` and the largest storage node fails, there isn't enough storage capacity available in the rest of the storage group to absorb the loss.
`Warning`	Warning message indicating that: • the storage group is full or near full. The appropriate threshold is exceeded. • the available capacity is insufficient to allow the Storage Group to rebalance in the event that the server with the largest amount of used capacity in the Storage Group fails
`Status`	Status of the storage group. Possible values are: `IN` Disks of three or more servers are assigned to the storage group. A storage group is in status `IN`, if a sufficient number of servers is assigned to the storage group servers to support 3-fold replication. `OUT` Less than three servers are assigned to the storage group.

To ensure the support of 3-fold replication within a pool, you can only locate a new pool on storage groups with IN status. Therefore the **Storage Group** list in the **Create Replicated Pool** or the **Create Erasure Coded Pool** dialog only contains storage groups with “IN” status

7.1.4 Pool Status

An overview of the status of the configured storage pools is displayed in the **Pool Status** window.

To open the **Pool Status** window, select **Monitor Cluster** > **Pool Status** in the navigation of the VSM GUI. The window consists of the **Pool List**.

Amongst others, the pool name, the storage group the pool belongs to, the number of replicas, information on placement groups, and information on the capacity per pool and read and write operations per pool are displayed. All displayed information is listed in detail below.

Pool List

The following information is displayed:

`ordinal` — ID of the pool. Sequence number automatically assigned during pool creation.

`Name` — Name of the storage pool

`Tag` — Optional: identifying tag string

`Storage Group` — Storage group where the pool is located. The storage groups are defined in the cluster manifest file.

For detailed information, see "Format of the cluster.manifest file" on page 234.

`Size` — Number of copies of the data object (primary copy and replicas)

`PG Count` — Number of placement groups in the storage pool. VSM derives the number from the calculation:

```
[pg_count_factor]*number of OSDs in storage group) /
replication factor
```

You can set the `[pg_count_factor]` during cluster creation in the cluster.manifest file (default: 100).

For detailed information, see "Format of the cluster.manifest file" on page 234.

`PgP Count` — Total number of placement groups automatically set by VSM:

`Status` — Status of the storage pool.

`Create By` — Indicates where the pool was created:

- `VSM` — The pool was created with VSM.
- `ceph` — The pool was created external to VSM.

`KB Used` — KB actually used by the pool

Object	Number of data objects in the pool
Clones	Number of cloned data objects
Degraded	Number of degraded data objects: Replicas are missing.
Unfound	Number of unfound data objects
	CAUTION! If unfound data objects exist, data is missing.
Read Ops	Total number of read operations
Read KB	Total read KB
Write Ops	Total number of write operations
Write KB	Total written KB
Client READ B/S	Amount of data read by the clients, specified in bytes/second
Client Write B/S	Amount of data written by the clients, specified in bytes/second
Client Ops/s	Number of I/O operations of all clients, specified per second

7.1.5 OSD Status

An overview of the status of the OSDs, that is the disks, is displayed in the **OSD Status** window.

To open the **OSD Status** window, select **Monitor Cluster** > **OSD Status** in the navigation of the VSM GUI. The window comprises an **OSD Summary** and the **OSD List**.

Amongst others, the OSD number, the status, the weight, the used and available capacity per disk, the server the OSD is running on, and the storage group are displayed.

All displayed information is listed in detail below.

OSD Summary

At a glance, the **OSD Summary** lists the total number of OSDs, the number of OSDs that are `up` and `in`.

OSD List

The following information is displayed:

`OSD Name`
: Name of the OSD

`VSM Status`
: VSM status. Possible values:

  `Present`
  : The OSD is present in the storage cluster

  `Removed`
  : The OSD has been removed.

`OSD Status`

The current status of the OSD.

The status of an OSD is either in the cluster (`in`) or out of the cluster (`out`). It is either up and running (`up`), or it is down and not running (`down`).

If an OSD is up, it may be either `in` the cluster (you can read and write data), or it is `out` of the cluster. If it was `in` the cluster and recently moved `out` of the cluster, Ceph will migrate placement groups to other OSDs.

The possible values are in detail:

`In-Up`

The OSD is an active part of the storage cluster (`in`). The OSD daemon is running. The OSD is fully functional and available for I/O (`up`).

`In-Down`

The OSD is part of the storage cluster, but no I/O is being directed to it.

Existing data on the disk will not be rebalanced. The disk may need to be restarted.

`Out-Up`

The OSD is fully functional and available for I/O (`up`), but no active part of the storage cluster (`out`).

All data on this disk will be relocated to other disks. CRUSH does not assign placement groups to the OSD.

`Out-Down`

The OSD is no active part of the storage cluster. No I/O is directed to it. CRUSH does not assign placement groups to the OSD.

Existing data on the disk will not be rebalanced.

`Out-Down-Autoout`

The OSD is currently no active part of the cluster. No OSD daemon is running. No I/O requests are directed to the disk.But VSM has recognized the disk.

To reconnect the OSD to the storage cluster the OSD "Restarting disks/OSDs" on page 124

`Crush Weight`	Weight assigned to the OSD to control the amount of data it is responsible for storing. For detailed information, see "CRUSH Data Placement" on page 29.
`Capacity Total (MB)`	The formatted capacity of the disk, specified in MB.
`Capacity Used (MB)`	Currently used disk capacity, specified in MB.
`Capacity Available (MB)`	Currently available disk capacity, specified in MB.
`Percent Used Capacity`	Currently used disk capacity, specified in percent.
`Server`	The host name of the storage node on which the disk is located.
`Storage Group`	Appropriate storage group
`Zone`	Not applicable in this version of ETERNUS CD10000.

7.1.6 Monitor Status

An overview of the status of the monitors, is displayed in the **Monitor Status** window.

To open the **Monitor Status** window, select **Monitor Cluster** > **Monitor Status** in the navigation of the VSM GUI. The window comprises a **Monitor Summary** and the **Monitor List**.

All displayed information is listed in detail below.

Monitor Summary

Amongst others, the **Monitor Summary** lists the total number of monitors and the overall status of the storage cluster.

Monitor List

The following information is displayed:

`Ordinal`	Sequence number automatically assigned to the monitor.
`Name`	Name of the monitor
`Address`	Address of the monitor daemon. The `Address` consists of the IP address of the storage node where the monitor is located and the port number of the monitor daemon.

`Health`	Health status of the storage cluster. Possible values are: HEALTH_OK The storage cluster is fully operational. HEALTH_WARN One or more disks have the status `down`. HEALTH_ERR One or more disks have the status `out`.
`Detail`	Information/warnings received from the Ceph monitor. Example: `clock skew 0.0515231s > max 0.05s` The current clock skew is greater than the maximum tolerable clock skew (Ceph default: 0.05 s).
`Skew`	Time difference between the storage nodes in seconds
`Latency`	Response time of the network in seconds
`MB Total (disk)`	Total storage capacity of all disks of the storage cluster, specified in MB
`MB Used (disk)`	Used storage capacity of all disks of the storage cluster, specified in MB
`MB Available(disk)`	Available storage capacity of all disks of the storage cluster, specified in MB
`Percent Available`	Available storage capacity of all disks of the storage cluster, specified in percent

7.1.7 PG Status

An overview of the status of the placement groups (PGs) is displayed in the **PG Status** window.

To open the **PG Status** window, select **Monitor Cluster** > **PG Status** in the navigation of the VSM GUI.

At the top of the window, Cluster-specific data is displayed, such as the summary of the placement groups, capacity values, and total write and read values.

Below, the **PG List** displays each individual PG with its status values.

All displayed information is listed in detail below.

PG states

The most important PG states are listed below:

`Active`	The PG is ready to accept request.
`Inactive`	Requests cannot be executed because an OSD is not up.
`Clean`	All data objects are replicated according to the configured number of replicas.
`Unclean`	Some data objects are not replicated according to the configured number of times. These data objects should be recovered.
`Down`	A replica with the necessary data is down, and the placement group is offline.
`Scrubbing`	A consistency check is running.
`Degraded`	Some data objects are not replicated the configured number of times.
`Inconsistent`	Inconsistencies were detected in one or more replicas of a data object. For example, a replica has a wrong size.
`Stale`	The PG is in an unknown state. The monitors have not received an update for it since the mapping changed.

For detailed information on the other available PG states, see "PG States" on page 337.

Object Summary

Information on the degraded and the unfound data objects is displayed

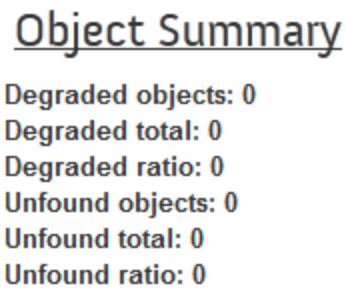


Figure 59: PG Status - Object Summary

In Ceph, the unit of data storage is called an object. Copies of objects are placed on different disks and servers in the cluster; the number of copies is determined by the replication factor of the storage pool that the object (data) is placed in.

`Degraded objects`

When a disk or a server fails or cannot be accessed by the cluster, copies of objects are not accessible; the affected objects are referred to as “degraded”.

Under normal operation, it is possible to have some small percentage of objects degraded. This may be either due to a disk or a server failure, or may occur if a server is stopped for maintenance and is temporarily inaccessible by the cluster.

`Unfound objects`

Under certain conditions, Ceph may be unable to find any valid copies of an object; when this occurs, the object is referred to as “unfound”. This condition may be temporary or may be permanent. I/O to lost objects will block and wait for the cluster to find a valid copy of the object.

Performance Summary

The following information is displayed:

`Total Reads`	Total amount of read data, specified in Bytes per second
`Total Writes`	Total amount of written data, specified in Bytes per second
`Total operations`	Number of operations, specified per second

PG Summary

The following information and a summary of the current PG states is displayed. For the possible values, see "PG states" on page 152.

PGmap Version	Version of the PG map. For detailed information on the PG map, see "ETERNUS CD10000 Architecture" on page 27.
Total PGs	Number of placement groups

Capacity Summary

The following information is displayed:

Data Capacity Used	Client data capacity used, specified in GB
Total Capacity Used	Total capacity, that is client data plus replicas, used, specified in GB
Capacity Available	Remaining cluster capacity, specified in GB
Capacity Total	Total cluster capacity, specified in GB

PG List

Each replica of a placement group is assigned to a different OSD. For example, if the pool requires three replicas of a placement group, the replicas are assigned to `osd.1`, `osd.2`, and `osd.3` respectively. The set of OSDs that should contain the replicas of a particular placement group is the acting set.

Ceph processes a client request using the up set, which is the set of OSDs that actually handles the requests. In most cases, the up set and the acting set are virtually identical. When they are not, it may indicate that Ceph is migrating data, that an OSD is recovering, or that there is a problem.

In detail the following information is displayed:

PG ID	Identifier of the PG. The PG ID is automatically assigned to the placement group.
State	Status of the PG. For the possible values, see "PG states" on page 152.
Up	Numbers of the OSDs that are `up`
Acting	Numbers of the OSDs that are acting

7.1.8 RBD Status

An overview of the RADOS Block Devices (RBDs), that are configured in the different storage pools, is displayed in the **RBD Status** window.

To open the **RBD Status** window, select **Monitor Cluster** > **RBD Status** in the navigation of the VSM GUI. The window consists of the **RBD List**.

All displayed information is listed in detail below.

RBD List

The following information is displayed:

`Pool`	Name of the storage pool
`Image Name`	Name of the image written by a user.
`Size`	Virtual disk size committed (not used in V1.0 SP2). This value comprises only the data, not the replicas.
`Objects`	Number of objects of this image
`Order`	Specifies the object size expressed as a number of bits, such that the object size is 1 << order. Default: 22 (4 MB).
`Format`	Specifies output formatting. Possible values are `plain`, `json`, `xml`. Default: `plain`

7.2 Monitoring with vsm_cli

vsm_cli provides the following monitoring commands:

- **status** (see page 277)

  the **status** command displays a brief summary of the cluster health status, the OSD status and the storage utilization

- **mon-status** (see page 269)

  the **mon-status** command displays information on the storage cluster's monitor daemons

- **list** (see page 266)

  the **list** command informs on the storage cluster and on the storage cluster components

Depending on the used options the **list** command output provides the following information:

- the status of the OSDs associated with the cluster (see page 158),
- information about the OSDs and the storage group they belong to displayed in an OSD tree (see page 162)
- information on the addresses and the capacity of the storage nodes (see page 164),
- information on the storage pools of the cluster and their utilization (see page 165),
- an overview of all cluster components (see page 166).

This section provides detailed information on these commands and their output.

7.2.1 Summary of the Cluster Status

The **vsm_cli status** command displays a brief summary of the cluster health status, the OSD status and the storage utilization .

The command

vsm_cli CLUSTER_1 status

provides a **vsm_cli** message with error flag N as result:

```
FCA1022:N:STATUS#health:HEALTH_OK #osd:12 osds,12 up,12 in#Storage utilization:0.013%
```

The message includes the following information:

- `#health: Health_OK`

  the health status of the storage cluster. Possible values are:

  HEALTH_OK

  The storage cluster is fully operational.

  HEALTH_WARN

  One or more disks have the status `down`.

  HEALTH_ERR

  One or more disks have the status `out`.

- `#osd: 12 osds, 12 up and 12 in`

  total number of OSDs and number of OSDs in the status `up` and `in`.

  If the number of OSDs in status `up` or `in` does not match the total number of OSDs, one or more OSDs of the storage cluster are `down` or `out`. Then the **status** command will display at least one additional message with error flag N (= notice) indicating that a disk (OSD) of the storage cluster is `down` or `out`.

- `#Storage utilization:0.013%`

  currently used storage capacity, specified in percent.

Detailed information about the syntax of the **status** command see "status - Information about the Cluster Status " on page 277

Timeout

The status command has a default 10 seconds timeout to collect the significant data for reporting the storage cluster status. The default timeout may be a too short period of time for a cluster with many nodes.Therefore, you can override the default timeout by passing a second parameter as timeout.

Example: Set the timeout to 12 sec

```
vsm_cli CLUSTER_1 status 12
```

7.2.2 Information on the Storage Cluster's Monitors

The output of the **vsm_cli** command informs about the quorum of monitors and the IP addresses of the hosts, on which the monitors reside.

For information on the monitors enter:

vsm_cli CLUSTER_1 mon-status

The command provides a **vsm_cli** message with error flag N as result:

```
FCA1026:N:MON_STATUS#monitor_quorum_rank_ids=[0,1,2] 0:192.168.0.104,
1:192.168.0.107, 2:192.168.0.109
```

The message includes the following information:

- `#monitor_quorum_rank_ids=[0,1,2]`
  The monitors `mon.0`, `mon.1`, `mon.2` exist in the cluster.
- `0:192.168.0.104 1:192.168.0.107, 2:192.168.0.109`
  the IP-addresses of the node the monitors reside on:
  the monitor `mon.0` resides on the node with the IP address 192.168.0.104,
  the monitor `mon.1` resides on the node with the IP address 192.168.0.107 and
  the monitor `mon.2` resides on the node with the IP address 192.168.0.109.

Detailed information about the syntax of the **mon-status** command see "mon-status - Displaying Monitor Status of a Storage Cluster" on page 269

7.2.3 OSD Status and Utilization

The **list** command of **vsm_cli** has the following options to request information on disks (OSDs) within a cluster:

- **list disks** povides information on all disks of a cluster
- **list disks-of-nodes** restricts information on the disks associated with specified storage nodes

The **list disks** command additionally provides options you can use to control the sort order of the output.

7.2.3.1 Status of all OSDs (disks)

When the `disks` option is used, a list with all disks associated with the storage cluster will be displayed.

`vsm_cli CLUSTER_1 list disks`

Osd No.	Size(MB)	Used(MB)	Avail(MB)	Util	Type	Node	Device name	Status
0	285429	50	285380	1%	ssd	node001	pci-**-0:2:2:0	up in
1	285429	35	285394	1%	sata	node001	pci-**-0:2:3:0	up in
2	285429	36	285394	1%	sas	node001	pci-**-0:2:4.0	up in
3	285429	36	285394	1%	sas	node001	pci-**-0:2:5.0	up in
4	285429	44	285385	1%	ssd	node002	pci-**-0:2:2.0	up in
5	285429	45	285384	1%	ssd	node002	pci-**-0:2:3.0	up in
6	285429	36	285394	1%	sata	node002	pci-**-0:2:4.0	up in
7	285429	35	285394	1%	sas	node002	pci-**-0:2:5.0	up in
12	285429	51	285378	1%	ssd	node003	pci-**-0:2:2.0	up in
13	285429	36	285394	1%	sata	node003	pci-**-0:2:3.0	up in
14	285429	36	285393	1%	sas	node003	pci-**-0:2:4.0	up in
15	285429	36	285393	1%	sas	node003	pci-**-0:2:5.0	up in

Where:

`OSD No.`	The virtual number assigned to the disk.
	OSD No.is equal to the OSD number, i.e. 01 specifies the OSD osd.1.
`Size (MB)`	The formatted capacity of the disk, specified in MB.
`Used (MB)`	Size of the currently used space, specified in MB.
`Avail (MB)`	Size of the currently available space, specified in MB.
`Util`	Utilization of space in percent.
`Type`	The type of disk, e.g. `SSD` and `SAS`.
`Node`	The host name of the storage node on which the disk resides.
`Device name`	The OSD's device pathname in by-path notation

`Status` The current status of the disk. Possible values are:

`In-Up`

The OSD is an active part of the storage cluster (`in`). The OSD daemon is running. The OSD is fully functional and available for I/O (`up`).

`In-Down`

The OSD is part of the storage cluster, but no I/O is being directed to it.

Existing data on the disk will not be rebalanced. The disk may need to be restarted.

`Out-Up`

The OSD is fully functional and available for I/O (`up`), but no active part of the storage cluster (`out`).

All data on this disk will be relocated to other disks. CRUSH does not assign placement groups to the OSD.

`Out-Down`

The OSD is no active part of the storage cluster. No I/O is directed to it. CRUSH does not assign placement groups to the OSD.

Existing data on the disk will not be rebalanced.

`Out-Down-Autoout`

The OSD is currently no active part of the cluster. No OSD daemon is running. No I/O requests are directed to the disk.But VSM has recognized the disk.

To reconnect the OSD to the storage cluster the OSD "Restarting disks/OSDs" on page 124

Controlling sort order of the output

The output of the **`list`** command with the option `disks` can be sorted by specifying a column name. Numeric columns will be sorted low-to-high.

For example:

- `vsm_cli cluster-name list disks avail`

  The output will be sorted based on the size of the available space (column `avail`).
- `vsm_cli cluster-name list disks used -r`

  The output will be sorted based on used capacity in reverse order (high-to-low.)

7.2.3.2 Status of OSDs (disks) on particular nodes

A subset of the output can be selected by using the `disks-of-nodes` option.

For example, the command

```
vsm_cli CLUSTER_1 list disks-of-nodes node001
```

will display only the disks within the storage node node001. Multiple host names can be entered.

Output:

```
Node     Public IP                      Cluster IP      # OSDs   Capacity(MB)
-------  -----------------------------  --------------  -------  ------------
node001   192.168.0.102,133.164.98.102   192.168.1.102        4       1/32104

Osd No. Size(MB)  Used(MB)  Avail(MB)  Util   Type    Node     Device name      Status
------  --------  --------  ----------  ------ ------  -------  ---------------  -------
     0   285429        50    285380     1%     ssd    node001   pci-**-0:2:2:0   up in
     1   285429        35    285394     1%     sata   node001   pci-**-0:2:3:0   up in
     2   285429        36    285394     1%     sas    node001   pci-**-0:2:4.0   up in
     3   285429        36    285394     1%     sas    node001   pci-**-0:2:5.0   up in
```

The first table of the output is described in section "Information on Storage Nodes" on page 164 and the second table in section "Status of all OSDs (disks) " on page 159.

7.2.4 OSD Tree

The **vsm_cli** command **list osd-tree** displays the OSD map as a tree with one line per OSD containing the crush weight and the status of the particular OSD.

The OSD tree is displayed when you enter the command (example for the cluster CLUSTER_1):

vsm_cli CLUSTER_1 list osd-tree

For detailed information on the **list** command see page 266

Output:

```
# id    weight  type name       up/down reweight
-19     12      root vsm
-16     5               storage_group performance
-13     5                       zone zone_one_performance
-1      2                               host node001_performance_zone_one
2       1                                       osd.2   up      1
3       1                                       osd.3   up      1
-4      1                               host node002_performance_zone_one
7       1                                       osd.7   up      1
-10     2                               host node003_performance_zone_one
14      1                                       osd.14  up      1
15      1                                       osd.15  up      1
-17     4               storage_group high_performance
-14     4                       zone zone_one_high_performance
-2      1                               host node001_high_performance_zone_one
0       1                                       osd.0   up      1
-5      2                               host node002_high_performance_zone_one
4       1                                       osd.4   up      1
5       1                                       osd.5   up      1
-11     1                               host node003_high_performance_zone_one
12      1                                       osd.12  up      1
-18     3               storage_group capacity
-15     3                       zone zone_one_capacity
-3      1                               host node001_capacity_zone_one
1       1                                       osd.1   up      1
-6      1                               host node002_capacity_zone_one
6       1                                       osd.6   up      1
-12     1                               host node003_capacity_zone_one
```

```
13      1                                       osd.13  up      1
```

The tree is arranged in the following levels (from left to right):

root level:	`vsm`	the tree contains all OSDs recognized by VSM
level 1:	`storage groups`	one leaf per storage group defined in the cluster.manifest-file
level 2:	`zones`	not applicable in this version of ETERNUS CD10000
level 3:	`storage nodes`	one leaf per storage node on which OSDs of the storage class (specified in level 1) reside.
level 4:	`OSDs`	one leaf per OSD that is assigned to the storage group (specified in level 1) and resides on the storage node (specified in level 3)

For each OSD the following data is displayed:

`#id`	The number of the OSD within the storage cluster. VSM assigns a number to each OSD when creating the storage cluster. The name osd.number uniquely identifies the OSD.
`weight`	Weight assigned to the OSD to control the amount of data the OSD is responsible for storing. For detailed information, see "CRUSH Data Placement" on page 29.

`up/down`	The OSD status:
	`up`
	The disk is fully functional and available for storing and retrieving data.
	`down`
	The disk is part of the storage cluster, but no i/o (read/write request) is directed to it.

7.2.5 Information on Storage Nodes

The **vsm_cli** command **list nodes** displays information on the IP addresses and the storage capacity of the storage nodes.

You can request the information on all storage nodes associated with the cluster:

```
vsm_cli CLUSTER_1 list nodes
```

or on particular nodes of the cluster:

```
vsm_cli CLUSTER_1 list nodes node001 node002 node003
```

Output (example):

```
Node     Public IP                      Cluster IP      # OSDs   Capacity(MB)
-------  -----------------------------  --------------  -------  ------------
node001   192.168.0.102,133.164.98.102   192.168.1.102        4       1/32104
node002   192.168.0.103,133.164.98.103   192.168.1.103        4       1/32104
node003   192.168.0.104,133.164.98.104   192.168.1.104        4       1/32104
```

Where:

Node	The host name of the storage node
Public IP	The node's IP address in the public network/subnet. The second address is the fall back if the first network is broken.
Cluster IP	The node's IP address used for peering between servers (cluster network). If this subnet is not displayed, then the public network is used for both client/server and peering traffic.
#OSDs	Number of OSDs on the storage node
Capacity (MB)	Used storage capacity / available storage capacity (in MB).

7.2.6 Pool Status

The **vsm_cli** command **list pools** displays a list of all pools created within a storage cluster.

For example, the command

```
vsm_cli CLUSTER_1 list pools
```

provides the following output:

```
Pool      Type             Replicas  Objects  Data(KB)     Quota(KB)  Uti(%)
--------- ---------------- --------  -------  -----------  ---------  ------
data      high_performance         3        0           0
metadata  high_performance         3       20           2
rbd       high_performance         3        0           0
Total                                     20           2
```

Where:

Pool	Name of the storage pool.
Type	Storage group where the pool is located. The storage groups are defined in the cluster manifest file (see "Cluster Manifest File" on page 234).
Replicas	Number of copies from the data objects (= primary object + its replicas). Example: Replication factor 3 means that there will be one primary copy and two further copies of each data object.
Objects	The number of data objects that are currently stored on the pool.
Data (KB)	The amount of data that is currently stored on the pool, specified in KB.
Quota (KB)	The quota of storage size allocated by the pool. The quota is specified in KB.
Uti (%)	Currently used storage capacity, specified in percent.
Total	This row consists the summations of stored objects and data over all pools.

7.2.7 Status of the Cluster and its Components (Overview)

An overview of the status of a storage cluster and all its components is provided by the command :

vsm_cli <cluster_name> list all

In the output of the command the following information is displayed among one another:

- health status of the storage cluster and a brief summary of the components
- information on the storage nodes associated to the cluster
- information on the disks (OSDs) associated to the storage cluster
- information on the pools associated to the cluster
- an overview of the components of the storage cluster and their utilization

Status of the storage cluster and a brief summary of its components

In the first area of the output the status of the storage cluster health, the OSD status, the monitors and the storage utilization are displayed

Example (<cluster_name> = CLUSTER_1):

```
cluster CLUSTER_1
health HEALTH_OK
3 mons at {0=192.168.0.102:6789/0,1=192.168.0.103:6789/0,3=192.168.0.104:6789/0}
12 osds: 12 up, 12 in
3072 pgs, 3 pools, 1884 bytes data, 20 objects
471 MB used, 3344 GB / 3344 GB avail
```

In the particular rows the following is displayed:

- the name of the storage cluster (e.g. CLUSTER_1)
- the health status of the storage cluster. Possible value are:

  HEALTH_OK

  The storage cluster is fully operational.

  HEALTH_WARN

  One or more disks have the status `down`.

  HEALTH_ERR

  One or more disks have the status `out`.
- Number of monitors and the IP address of the nodes on which the monitors are running.
- Total number of OSDs and number of OSDs in the status in, out, up and down.

- Number of placement groups (`pgs`), pools, the amount of stored data, the number of stored data objects.
- Size of the used storage, size of the still remaining storage / total size of the available storage.

Information on the storage nodes associated to the storage cluster

The output corresponds to the output of the command **`list nodes`** for all nodes of the cluster. For a detailed description of the output see "OSD Status and Utilization " on page 158.

Example:

```
Node      Public IP                        Cluster IP      # OSDs   Capacity(MB)
-------   ------------------------------   -------------   -------  ------------
node001   192.168.0.102,133.164.98.102   192.168.1.102        4       1/32104
node002   192.168.0.103,133.164.98.103   192.168.1.103        4       1/32104
node003   192.168.0.104,133.164.98.104   192.168.1.104        4       1/32104
```

Information on the disks (OSDs) associated to the storage cluster

The output corresponds to the output of the command **`list disks`**.

For a detailed description of the output see "Information on Storage Nodes" on page 164.

Example:

```
Osd No.  Size(MB)   Used(MB) Avail(MB) Util Type   Node    Device name      Status
-----  -----------   -------- --------- ---- ------ ------- --------------- -------
    0      285429       50    285380    1%  ssd    node001  pci-**-0:2:2:0   up in
    1      285429       35    285394    1%  sata   node001  pci-**-0:2:3:0   up in
    2      285429       36    285394    1%  sas    node001  pci-**-0:2:4.0   up in
    3      285429       36    285394    1%  sas    node001  pci-**-0:2:5.0   up in
    4      285429       44    285385    1%  ssd    node002  pci-**-0:2:2.0   up in
    5      285429       45    285384    1%  ssd    node002  pci-**-0:2:3.0   up in
    6      285429       36    285394    1%  sata   node002  pci-**-0:2:4.0   up in
    7      285429       35    285394    1%  sas    node002  pci-**-0:2:5.0   up in
   12      285429       51    285378    1%  ssd    node003  pci-**-0:2:2.0   up in
   13      285429       36    285394    1%  sata   node003  pci-**-0:2:3.0   up in
   14      285429       36    285393    1%  sas    node003  pci-**-0:2:4.0   up in
   15      285429       36    285393    1%  sas    node003  pci-**-0:2:5.0   up in
```

Information on the pools associated to the cluster

The output corresponds to the output of the command **`list pools`**. For a detailed description of the output see "Pool Status" on page 165.

```
Pool      Type              Replicas  Objects  Data(KB)     Quota(KB)  Uti(%)
--------- ----------------- -------- -------- ------------ ---------- ------
data      high_performance         3        0          0
metadata  high_performance         3       20          2
rbd       high_performance         3        0          0
Total                                     20          2
```

An overview of the components of the storage cluster and their utilization

Displayed are:

`Total nodes`
: number of nodes within the storage cluster

`Total disks`
: number of disks (OSDs) within the storage cluster

`Total pools`
: number of pools within the storage cluster

`Total space (MB)`
: size of the available storage in MB

`Total used  (MB)`
: size of the used storage in MB

`Total avail (MB)`
: size of the still remaining storage in MB

Example:

```
-----------------------------
Total nodes     :           3
Total disks     :          12
Total pools     :           3
Total space (MB):     3425145
Total used  (MB):         472
Total avail (MB):     3424673
Utilization     :       .013%
```

7.3 Monitoring with SNMP

You can monitor the ETERNUS CD10000 storage cluster via SNMP V2.

For monitoring the storage cluster via the SNMP management station, you must install an SNMP-capable monitoring system, for example ServerView Operations Manager, HP OpenView, or Nagios, on the SNMP management station.

All SNMP traps generated on the storage nodes or switches of the storage cluster are sent to the management node. The management node forwards these traps to the SNMP management station(s). Therefore the management node serves as a proxy for the storage nodes and switches.

The management node of the storage cluster must be connected to the administration network in such a way that the SNMP management station can receive the SNMP information send by the management node.

You have to configure the access to the managed objects and the destinations of traps via the `cd10000 setup snmp` CLI. For detailed information, see "cd10000 setup snmp - the command line interface for SNMP configuration.

The Net-SNMP agent is installed on the management node so that this node acts as a proxy.

To monitor the hardware of the management node and the storage nodes ServerView agents and ServerView RAID Manager are installed.

SNMP traps are sent whenever a hardware failure is detected or some other severe software issue has occurred. For detailed information on the SNMP traps see "SNMP Traps" on page 209.

Furthermore special ETERNUS CD10000 SNMP agents that act as subagents to Net-SNMP are installed (SNMP V2). These subagents implement managed objects defined in the FUJITSU-CD-MIB to monitor the cluster as a whole. For detailed information on these managed objects, see "Managed Objects " on page 205.

On the LAN and IB switches the special SNMP agent for the switch is installed. The SNMP agent is pre-installed by the switch dealer.

All traps are sent to the management node.

7.4 Log Manager

ETERNUS CD10000 provides a centralized log manager. It collects configured log files from all storage nodes of a storage cluster in real-time and transfers them to the management node for supporting maintenance and support.

A command line tool for collecting log files on demand is available too.

The log manager provides the following fundamental functions:

Collecting log files

Whenever new entries are added to a log file, the specified set of log files is collected automatically in real-time from all storage nodes in the storage cluster and is stored in the database on the management node. For log file collection and forwarding Fluentd agents (forwarder agents) are configured on all storage nodes. They send events to Fluentd agents (aggregator agents) on the management node, which write them into an Elasticsearch database.

If required you can add new log files to the log manager configuration for real-time collection.

On the management node you can collect log files that are not automatically collected using an application for one-demand gathering of log files. The command **`cd10000 support gather_logs`** is provided for this purpose.

Viewing log files

You can view the collected log files with a log viewer and export them from the database, for example for sending a dump to the support. Additionally, data can be imported into the database.

As a log viewer you can use any viewer that is compatible with the Elasticsearch database, for example Kibana, a single-page web application.

Automatic daemon restart after failure

On each storage node and on the management node of the storage cluster a monitoring application, the monit daemon, is running, which restarts the Fluentd agent or the Elasticsearch database after failure.

SNMP

Failures of the log manager trigger the generation of an SNMP trap. For detailed information, see "Monitoring with SNMP" on page 169

8 Replacing Disks Physically

If a disk (OSD) of a storage cluster fails/retires or if you want to substitute an old disk with a new one with a higher capacity you can replace the disks without changing the overall storage cluster configuration. The storage cluster does not need to be stopped.

- The new and the old disk must have the same storage class.
- The new disk must have the same capacity or higher than the replaced disk.

`vsm_cli` provides the commands `replace-disk-out` and `replace-disk-in` to replace disks.

Scenarios

There are two scenarios where disks can be replaced, unscheduled and scheduled maintenance:

- **Unscheduled maintenance:**

  The disk/OSD has failed in some way. The down OSD will likely time out and be marked "`out`" by Ceph. Then, Ceph will automatically rebalance the disk objects to other disks on the same node. If you replace the disk, Ceph will automatically rebalance the objects back to the new disk.

- **Scheduled maintenance:**

  The disk/OSD is likely to be up and in. **`replace-disk-out`** command will set the "`noout`" flag for that OSD. When the OSD is stopped and the disk is unmounted, the OSD is marked "`down`" and it will stay "`down`" until it is replaced. This avoids unnecessary rebalancing.

  If the OSD happened to be "`down`", the "`noout`" flag will prevent it from going "`out`".

  If the OSD was already in the "out" state, the "`noout`" flag will have no effect. When you successfully replace the disk, Ceph will reset the "`noout`" flag and will automatically copy the necessary objects to the new disk.

For both scenarios, upon successful replacement, the storage cluster configuration will not have changed and the object distribution will generally be the same as before.

8.1 Replacing Disk

For safety reasons, the `replace-disk` commands should only be executed when the storage cluster status is healthy. If the disk to be replaced is in the `"out"` state, you should wait for any rebalancing to finish before calling the `replace-disk-out` and the `replace-disk-in` commands.

If you want to replace a disk of CLUSTER_1 that resides on the storage node NODE_1 proceed as follows:

- Check whether the storage cluster status is healthy (HEALTH_OK) and balanced, i.e.any rebalancing has been finished.

- Enter the `replace-disk-out` command. The command stops the OSD daemon and unmounts the file system.The `replace-disk-out` command does not remove the OSD from the CRUSH map.

  ```
  vsm_cli CLUSTER_1 replace-disk-out NODE_1 pci-0000:01:00.0-scsi-0:2:1:0
  ```

  `pci-0000:01:00.0-scsi-0:2:1:0` is the device name of the disk in by-path notation found in /dev/disk/by-path directory.

  You can figure it out with the following `vsm_cli` commands:

  ```
  vsm_cli CLUSTER_1 list disks
  vsm_cli CLUSTER_1 list disks-of-nodes NODE_1
  ```

  The path of a disk is displayed in the column `Device name`.

- Locate the disk to be replaced following the instructions in section "Locating the Disk in a Server or Storage Subsystem" on page 173.

- Remove the old disk physically and replace it with a new one. For detailed information, see "Replacing Disks Physically " on page 171

- Enter the `replace-disk-in` command specifying the same device path name as in the `replace disk-out` command

  ```
  vsm_cli CLUSTER_1 replace-disk-in NODE_1 pci-0000:01:00.0-scsi-0:2:1:0
  ```

8.2 Replacing Disks Ejected by Mistake

If an incorrect disk was ejected accidentally the following steps must be executed:

- Reinsert the ejected disk. The disk is currently in the status **down**.
- To check the device name of this disk/OSD and the node where the disk/OSD resides on, enter the following **list** command:

  ```
  vsm_cli <cluster> list disks
  ```

- Enter the **replace-disk-out** and **-in** commands to reintegrate the disk into the Ceph storage cluster.

  ```
  vsm_cli <cluster> replace-disk-out <node> <device name>
  vsm_cli <cluster> replace-disk-in <node> <device name>
  ```

If more than one disk was ejected, you must carry out the procedure for one disk after the other.

8.3 Locating the Disk in a Server or Storage Subsystem

To replace a disk corresponding to a certain OSD, you have to find out where the disk is physically located.

- First find out the storage node the disk resides on and the device name of the disk using the following **vsm_cli** command:

  ```
  pmgmt# vsm_cli CLUSTER list disks
  ```

  The first column of the output contains the OSD numbers, the `Node` column contains the storage node name and the `Device name` column contains the device name of the disk in by-path notation.
  For detailed information about the output see "OSD Status and Utilization " on page 158.

- To locate the disk you need to find out the **RAID controller number**, the **RAID slot number** where the disk resides and the **enclosure ID**.

- You will get the RAID controller number, the RAID slot number from the controller-ID and the disk-group in the by-path notation of the previously found device name:

  ```
  pci-0000:<controller-ID>-scsi-0:2:<disk-group>:0
  ```

To find out the **RAID controller number** corresponding to the controller-ID enter the following command on the affected storage node:

```
# lspci | grep RAID
```

The first column of the output shows the controller-ID. The corresponding **RAID controller number** is:

0 for the first output line (controller for the server) and

1 for the second line (controller for the storage subsystem)

The first two RAID slots (**0** and **1**) are permanently assigned to the RAID controller for the server (**0**). They are reserved for the Operating System RAID1, which will represent disk-group **0**. Therefore, to get the **RAID slot number** for the RAID controller **0** you have only to increment the disk-group number found in the by-path notation of the device name by **1**. In a storage subsystem (RAID controller **1**) where each disk has its own RAID0 disk-group, the RAID slot number corresponds exactly to the disk-group from the by-path notation.

- To find the **enclosure ID** depending on the RAID controller, enter the command:

```
# /opt/MegaRAID/storcli/storcli64 /cX/d<disk-group> show all | grep HDD
```

In the output of this command, the first column represents the **enclosure ID** (EID) you need to start location, together with the **RAID slot number**: <EID>:<RAID slot>

- To locate the disk (turn on the location LED) enter the following command:

```
# /opt/MegaRAID/storcli/storcli64 /cX/eY/sZ start locate
```

Where:

X is the RAID controller number you have previously found

Y is the enclosure ID (EID)

Z is the RAID slot number determined from the disk-group

- To turn off the location LED, please use the following command:

```
# /opt/MegaRAID/storcli/storcli64 /cX/eY/sZ stop locate
```

Examples:

1. The device name of the searched disk is: `pci-0000:01:00.0-scsi-0:2:12:0`
   To locate the server disk from `disk-group 12` please enter:

   ```
   # /opt/MegaRAID/storcli/storcli64 /c0/d12 show all | grep HDD
   ```

   Output:

   ```
   14:13 17 Onln 12 837.844 GB SAS HDD N N 512B WD9001BKHG-50D22 U
   ```

   This means:

   - the enclosure ID (EID) is 14 and
   - the RAID slot number is 13 (12 incremented by 1)

   So you can locate the disk with the command:

   ```
   # /opt/MegaRAID/storcli/storcli64 /c0/e14/s13 start locate
   ```

2. The device name of the searched disk is: `pci-0000:03:00.0-scsi-0:2:12:0`
   To locate the storage subsystem disk from `disk-group 12` please use:

   ```
   # /opt/MegaRAID/storcli/storcli64 /c1/d12 show all | grep HDD
   ```

   Output:

   ```
   30:12 23 Onln 12 837.844 GB SAS HDD N N 512B WD9001BKHG-50D22 U
   ```

   This means:

   - the enclosure ID (EID) is 30 and
   - the RAID slot number is 12 (same as disk-group)

   So you can locate the disk with the command:

   ```
   # /opt/MegaRAID/storcli/storcli64 /c1/e30/s12 start locate
   ```

8.4 2.5'' Hard Drive Carriers Replacement by the Server

Unlocking the HDD module

- Release the locking mechanism by pressing the locking button (1).
- Push the handle of the HDD module fully in the direction of the arrow (2).
- The HDD module is now unlocked.
- Wait about 30 seconds to allow the hard disk drive to spin down.
- Pull the HDD module completely out of its bay (see figure below).

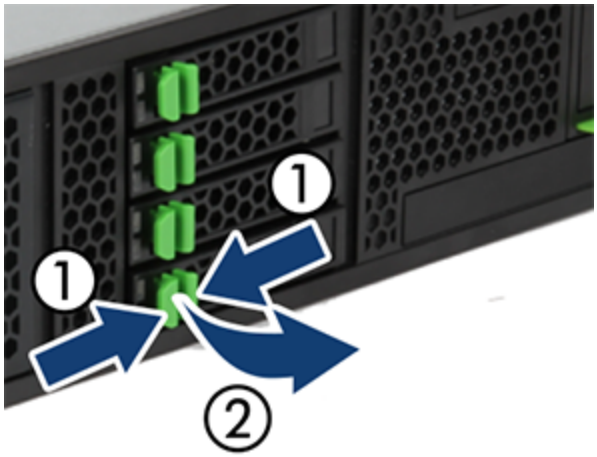


Figure 60: Disconnecting a 2.5-inch HDD

Figure 61: Removing a 2.5-inch HDD

Installing a 2.5-inch HDD

- Unlock the HDD module as described above.
- Insert the HDD module into a drive bay and carefully push back as far as it will go (1).
- Close the locking lever to lock the hard disk drive in place (2).
- If the status of the HDD remains "unknown" after rebuild has finished, contact Fujitsu Customer Support (see "Preface" on page 15)

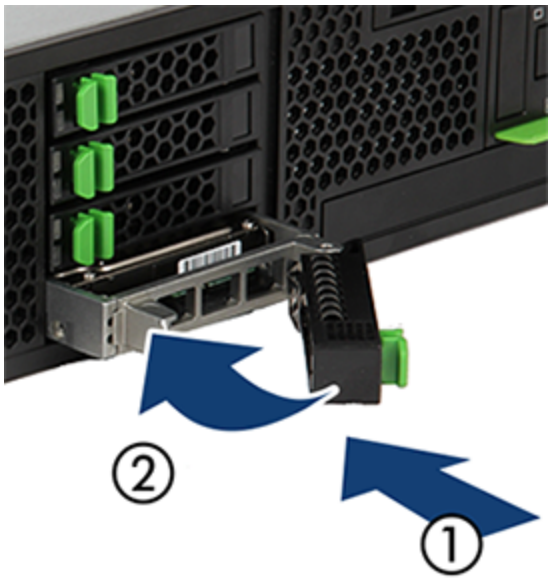


Figure 62: Inserting the 2.5-inch HDD

Removing the non-defective HDD modules

CAUTION!

If you need to remove non-defective HDD modules, consider the following:

- Do never remove a non-defective HDD module during operation of the ETERNUS CD10000.
- All HDD modules (drives) must be uniquely identified so that they can be reinstalled in their original mounting locations later. If this is not done, existing data can be lost.

8.5 2.5'' Hard Drive Carriers Replacement by the JX40 Storage Subsystem

- To remove disk drive press the latch of the disk drive and open the lock lever (1) (see figure below).
- Hold the lock lever and pull the disk drive partway.
- Take the hard disk drive out of the bay while supporting its bottom (2).

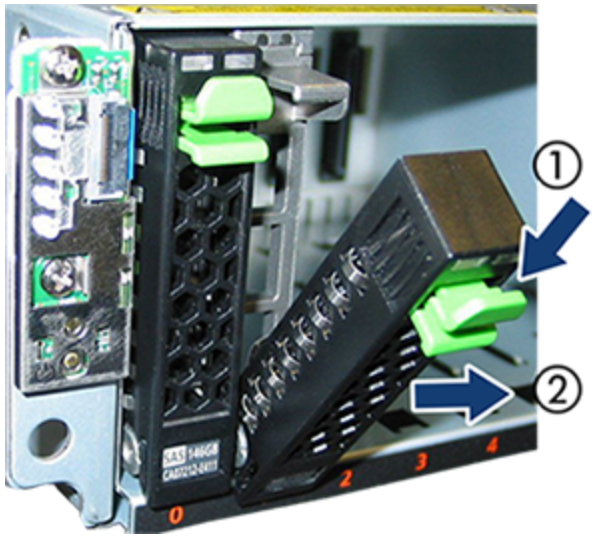


Figure 63: Removing/Installing hard disk drive

- To install disk drive insert the HDD module into a drive bay and carefully push back as far as it will go.
- Close the locking lever to lock the hard disk drive in place.

8.6 3.5'' Hard Drive Carriers Replacement by the JX60 Storage Subsystem

Inserting a hard disk drive

- Remove the front cover of the ETERNUS JX60.

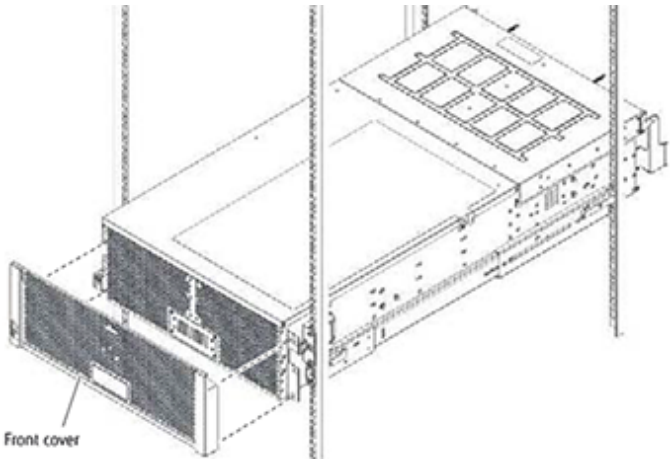


Figure 64: Removing the ETERNUS JX60 front cover

- Remove the four M5 screws that are used to fasten the ETERNUS JX60 in the rack pillar.

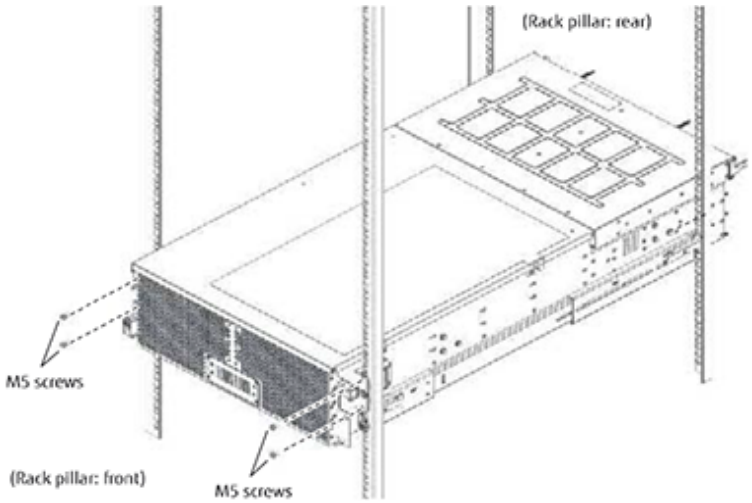


Figure 65: Unscrewing the ETERNUS JX60

- Pull the stopper catches on both sides of the ETERNUS JX60 forward to unlock it.
- Pull out the ETERNUS JX60 from the rack until the lock clicks open.

Make sure that there are no cables that are caught or stuck on the rear of the ETERNUS JX60 when pulling it out. In addition, pull it out slowly to avoid exposing it to physical shocks.

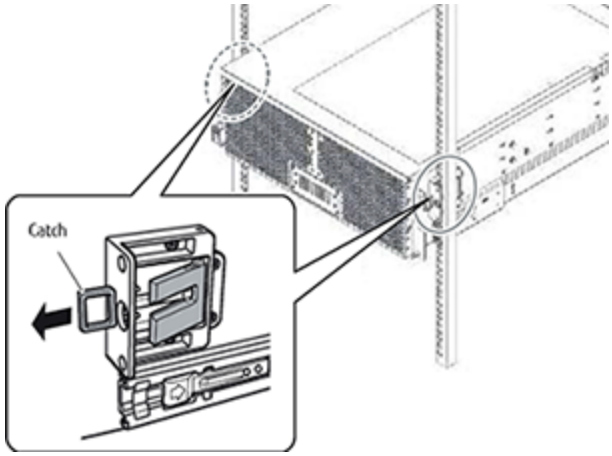


Figure 66: Pulling forward the stopper catches

- Push the two catches of the top cover in the direction of the arrows to unlock them.

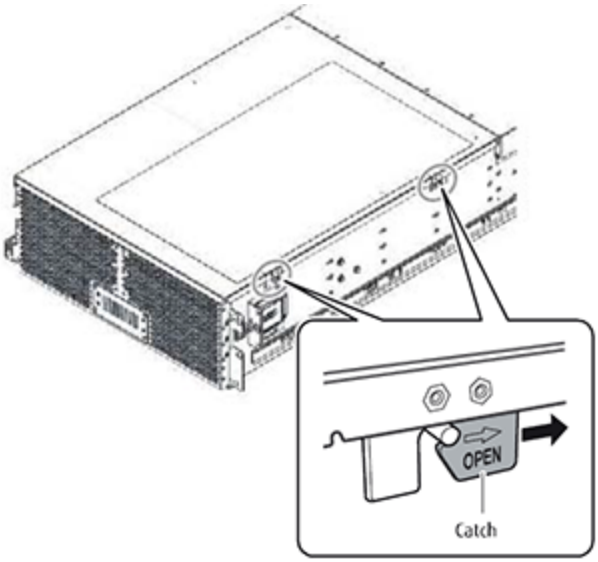


Figure 67: Unlocking the top cover

- Open the top cover.

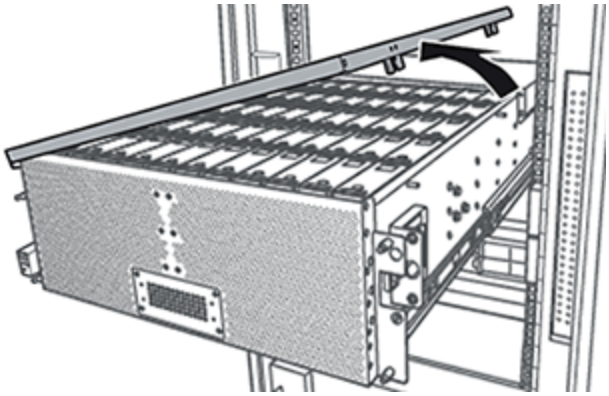


Figure 68: Opening the top cover

- Remove the dummy drive unit from the slot in which the drive is to be installed:

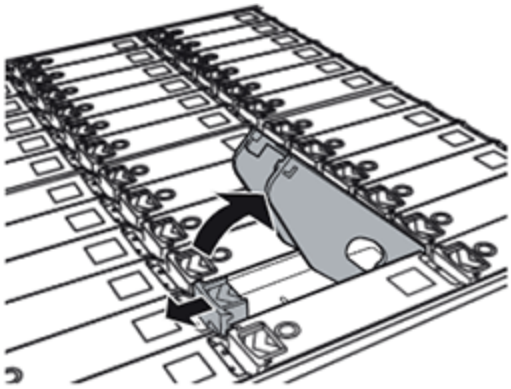


Figure 69: Removing the dummy drive (1)

- Pull the catch on the lock lever of the dummy drive unit to unlock the lock lever, and then open the lock lever upward.

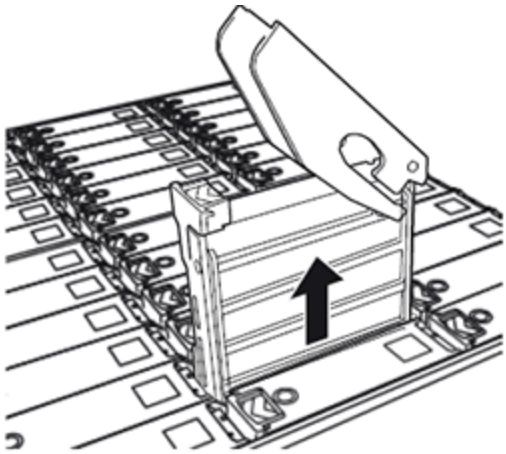


Figure 70: Removing the dummy drive (2)

- Pull up the dummy drive unit in a straight direction.

Install the new drive:

CAUTION!
When installing a drive, be careful not to pinch your fingers.

Hold the drive with both hands to protect against jarring.

- Pull the catch on the drive lock lever to unlock it.

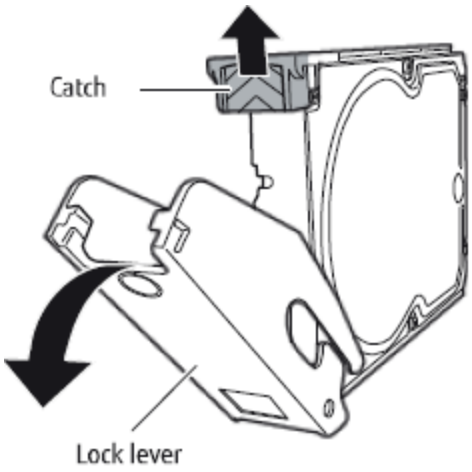


Figure 71: Installing the new drive (1)

- Insert the drive all the way into the slot, making sure that it is firmly seated and that the lock lever is still left open.

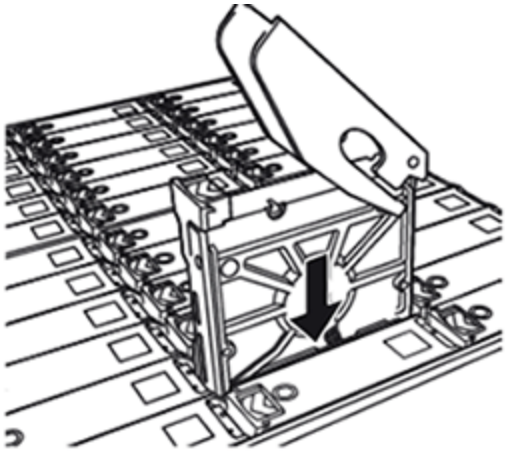


Figure 72: Installing the new drive (2)

9 Software Configuration and Update/Upgrade System

The ETERNUS CD10000 configuration and update management system consist of a RPM software repository and the Update Manager application. The RPM repository is configured and stored on the management node.

A RPM repository is accessible for all nodes in the cluster from the management node.

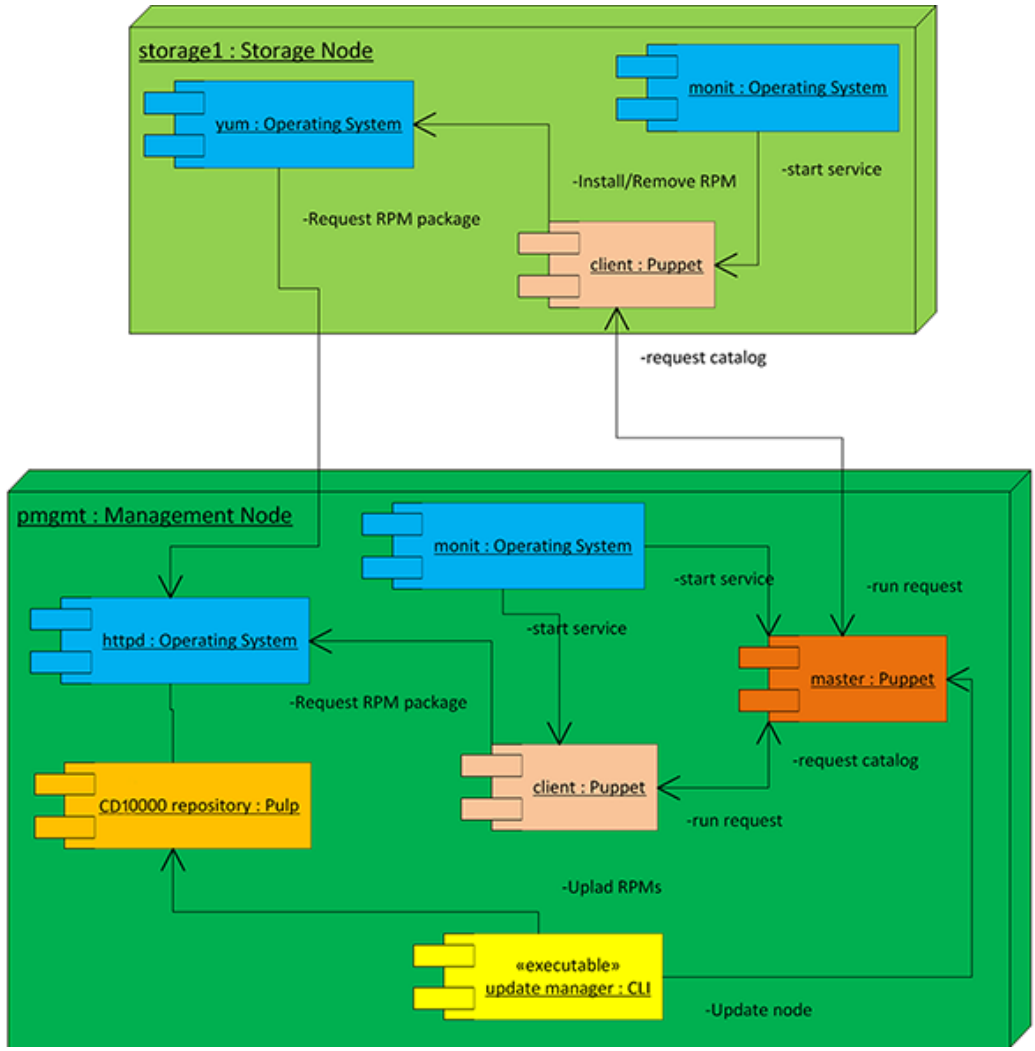
Software Updates provided on media are uploaded to the repository using the Update Manager. Update Manager is an application which manages the entire update/upgrade process in a cluster from the start to the end. This includes software updates and settings of the configuration, such as network.

An Updating/Upgrading process runs at a time on one node only to ensure the stability of the storage cluster as a whole. Puppet software is used for this purpose. Puppet software is an application which manages the configuration of the nodes. Puppet manifests define the resources and the state of a node or group of nodes and this is applied during update/upgrade process to the nodes.

On the management an own directory exists for each node of the cluster. In this directory multiple files, the partials of a manifest, are saved. Each node has its own directory, where are partials of manifest saved.

A manifest is a defined state of a node and it is generated from partials. They contain defined cluster state, e.g. set of packages to be installed on node, network configuration, user password, etc. The partials of a manifest are merged into a single manifest during update/upgrade process.

Customer specific partials can be defined and default partials can be modified, but this may also cause conflicts with default configuration provided within update/upgrade.


storage1 : Storage Node
yum : Operating System
monit : Operating System
-Install/Remove RPM
-start service
-Request RPM package
client : Puppet
-request catalog
pmgmt : Management Node
monit : Operating System
-start service
-run request
httpd : Operating System
-start service
master : Puppet
-Request RPM package
-request catalog
CD10000 repository : Pulp
client : Puppet
-run request
-Uplad RPMs
«executable»
update manager : CLI
-Update node

Component architecture:

- Management node
  - The Puppet master manages puppet clients on the storage nodes and performs configuration.
  - The RPM repository management is managed by Pulp and served through HTTP.
  - Update Manager is an application which manages the entire update/upgrade process in a whole cluster from the start to the end.
  - The monitoring process, monit, checks at intervals and restarts the Puppet master process in case it failed. After three unsuccessful attempts of restarting a SNMP trap is generated.
  - Puppet client application is also installed on management node.
    The management node itself is updated/upgraded by Update Manager too.
- Storage node
  - The Puppet client application is triggered by the puppet master on the management node and is responsible for update/upgrade of the storage nodes.
  - The monitoring process monit checks at intervals the Puppet client process and restarts it if not running. After three unsuccessful attempts of restarting a SNMP trap is generated.

9.1 Update/Upgrade Process Flow

This chapter describes a default update and upgrade process.

An Update or Upgrade process updates installed software and delivers an instruction with additional or changed steps to be executed. If the update of software packages or the other steps require a reboot of the system we use the term *upgrade*. An additional step during an upgrade sets the node into maintenance mode. With this steps it is ensured that Ceph rebalancing is not executed.

To ensure stability of Ceph cluster before upgrade of each node Update Manager waits until Ceph cluster is in HEALTH_OK state. Therefore an upgrade process of cluster needs more time than an update process.

Within each update/upgrade an instruction with additional or changed steps are delivered which are required to run successfully update/upgrade process. In that case provided instruction is superior to this description below.

The Update Manager marks an update/upgrade as successful if an update/upgrade report is send by the node within a given time period. If the report is not received within this period the update/upgrade is failed. The default update timeout is ten minutes and default upgrade timeout is fifteen minutes. The Timeout can be changed depending on the delivered instruction with the update/upgrade.

The Timeout depends also on update size and cluster load.

The time of the whole update/upgrade process can differ between versions. Also the load of a node has impact on it. It may be required to adjust timeout of executed update/upgrade.

When the timeout expires or the update/upgrade process of the node fails, the whole updating/upgrading process in the cluster will stop.

In the case that the updating/upgrading process fails, the cause has to be investigated. The failure could be caused by e.g. a network failure or inappropriately merged manifests.

Logs of the update/upgrade process are saved in `/var/log/fpm.log` and `/var/log/puppet/`, and reports for each node are stored in `/var/lib/puppet/reports/<hostname>/`.

Update Manager requires that all nodes are up and running connected to the administration network.

Default update process:

1. Uploading provided RPM packages to repository

   update-manager upload -d ABSOLUTE_PATH_TO_RPM_DIR
2. Updating manifest partials using Update Manager (manifest partials contain defined cluster state)

update-manager update_partials

3. Publish repository

   update-manager publish

4. Merging custom/changed manifests with Updates (this step can be omitted when there are no custom changes).
5. Updating the cluster using Update Manager (update the management node and update all storage nodes)

   update-manager update

Default upgrade process:

1. Uploading provided RPM packages to repository

   update-manager upload -d ABSOLUTE_PATH_TO_RPM_DIR

2. Updating cluster manifest partials using Update Manager (manifest partials contain defined cluster state)

   update-manager update_partials

3. Publish repository

   update-manager publish

4. Merging custom/changed manifests with Updated (this step can be omitted when there are no custom changes or no custom manifests).
5. Upgrading management node only using Update Manager.By applied manifests, management node will be rebooted. After rebooting wait about five minutes before going to next step.

   update-manager upgrade -n pmgmt

6. Execute

   vsm-update

7. Upgrading the cluster using Update Manager (this function upgrades only storage nodes).

   update-manager upgrade

8. Uploading given iso for future installation of storage nodes:

   update-manager upload -d ABSOLUTE_PATH_TO_MOUNTED_ISO --iso

mount -o loop path_to_iso ABSOLUTE_PATH_TO_MOUNTED_ISO

ABSOLUTE_PATH_TO_MOUNTED_ISO= /tmp/iso

ABSOLUTE_PATH_TO_RPM_DIR=/tmp/iso/Packages

Default update process of custom configuration

1. Adding custom manifests or changing existing.
2. Generating manifests using Update Manager.
3. Merging custom/changed manifests with the one provided.
4. Updating a cluster using Update Manager.

9.2 Custom Manifests Partials

It is recommended to contact Fujitsu Technical Support before any changes are implemented

Partials must be placed in an appropriate directory. The Update Manager should be executed in order to distribute the configuration in a cluster. All partials must have ".pp" extension.

Your Puppet modules may be required to add your configuration. Place the Puppet modules in the Puppet module directory, but they must not conflict with those provided by Fujitsu. Added partials must have ".pp" extension.

9.2.1 Merging the manifests partials

Default manifest partials are stored on the management node.

You are allowed to change manifest partials, but update/upgrade process will not start as long as possible conflicts are not merged.

Merge possibilities:

- If the default partial is not changed merging is not required.
- If you change a default partial, the new version provided with update/upgrade will have "rpmnew" extension.
  In this situation you need to merge both partials into one with the default name.
- If you remove a default partial, it may be added during update/upgrade again.
  In this situation you need to decide either to keep or to remove partial.
- If you change a default partial and in it is no longer available in the new version, the changed partial will be left.
  In this situation you need to decide if it should be removed.
- If you add own partials, they will not be removed during update.
  In this situation you need to decide whether your partials conflict with the default partials provided by Fujitsu.

9.2.2 Handling errors

- If reaching a node that is not available the update/upgrade process of that and next nodes will not start.
- If a node returns a report with status "failed" or when the specified timeout of waiting for report expires, the Update Manager will report a failure and it will not update/upgrade any other node.
- After the generation of manifests and before running update/upgrade process, the Update Manager will validate manifest files with a puppet parser for syntax errors. If the validation fails, the Update Manager will not start the update/upgrade process.
- When upgrading a node fails it will be left in maintenance mode and before any next attempt of upgrading that node the maintenance mode must be switched off manually.
- When update/upgrade fails usage of puppet will be blocked by Update Manager until update/upgrade will be finished successfully, so adding a new node of changing IP of a node will not be possible.

9.3 Limits

- For safety of Ceph cluster, Update Manager will update/upgrade only one node at a time.
- The default update timeout is 10 minutes. It should be adjusted depending on the update size and cluster load.
- The default upgrade timeout is 15 minutes. It should be adjusted depending on the upgrade size and cluster load.
- Each node in the cluster can have installed only software provided by Fujitsu. Any other installation of external software is prohibited.
- The end user is allowed to write the custom partials of manifests or to change partials provided with ETERNUS CD10000 only if they will not conflict. Within each update end user must check manifest partials for conflicts and resolve them if required.

9.4 Log Files

Location of log files and system messages

On the storage nodes

- /var/log/puppet/
- var/log/monit
- /var/log/messages

On the management node

- /var/log/puppet/
- /var/log/monit
- /var/log/messages
- /var/log/pulp/
- /var/lib/puppet/reports/
- /var/log/fpm.log

9.5 Configuration

Manifest is a defined state of a node and it is generated from partials. They contain defined cluster state, e.g. set of packages to be installed on node, network configuration, user password, etc.

Default partials are provided in the fpm-manifests RPM package which is delivered with each update/upgrade. The Installation of this package will the update default partials and the Puppet modules. If a provided partial is modified or you add own partial, manual merge may be required.

The following directories are used to store the manifests:

- Puppet modules `/etc/puppet/modules/`
- Puppet manifests `/opt/fujitsu/fpm/manifests/`

The tables below represent sets of directories on the management node, where configuration for types of nodes can be added.

Each node in the cluster has its own catalog. Nodes are also grouped into general types, e.g. the genericnode has a configuration which will be applied on all nodes in cluster.

To overview: the root is a genericnode which contains configuration for all nodes in the cluster, genericstorage inherits everything from genericnode and contains additionally configuration for all storage nodes in the cluster. Storage nodes inherit everything from genericstorage and has additionally configuration for exact node.

The following table describes the defined generic node types:

Directory	Inherits from	Description
genericnode.d/	-	Partials from this folder are applied on every node.
genericstorage.d/	genericnode	Partials from this folder are applied on every storage node.
management.d/	genericnode	Partials from this folder are applied on each management node.
backup.d/	genericstorage	Partials from this folder are applied on each backup node. Backup nodes are: storage1, storage2, storage3. Backup node in current configuration are nodes running NTP slave servers & ceph monitors.
storage.d/	genericstorage	Partials from this folder are applied on each storage node, except of the backup nodes.

Table 1: Defined generic node types

The following table describes that each physical node has its own directory with configuration:

Directory	Inherits from	Description
pmgmt.d/	management	Partials for pmgmt node.
storage1.d/	backup	Partials for storage1
storage2.d/	backup	Partials for storage2
storage3.d/	backup	Partials for storage3
storage4.d/	storage	Partials for storage4.
storage5.d/	storage	Partials for storage5
storageXXX.d/	storage	Partials for storageXXX

Table 2: Directories for nodes

9.6 Commands of the Update Manager

Commands of the Update Manager you can fine in the chapter "The Update Manager Commands" on page 313.

9.7 Firmware Update of the Hardware in the Cluster

The Firmware-manager is a python application to perform automatic firmware upgrades remotely or locally. An application is installed only on the management node. Sub-package containing ASP binaries are installed on the management node and storage nodes. Without passing list of nodes as parameter application will execute upgrade on all storage nodes, one by one, if upgrading will fail on one storage node, the upgrade process will stop and firmware manager will exit with exit message.

9.7.1 Functional Overview

Firmware update of the hardware in the cluster contains following steps:

- Upgrading one hardware component
- Upgrading all firmwares of the hardware in one node
- Upgrading all firmwares of the hardware in cluster
- Upgrading management node

9.7.2 Runtime Model

- Storage node:
  - Rpm package `firmware-manager-firmwares-<version>-<release>.noarch.rpm` – package with ASP firmwares for cd10000 servers
- Management Node:
  - Rpm package `firmware-manager-<version>-<release>.noarch.rpm` – main package with firmware upgrade application
  - Rpm package `firmware-manager-firmwares-<version>-<release>.noarch.rpm` – package with ASP firmwares for cd10000 servers

The default firmware upgrade process is described below. Upgrade process differs from update process. Then after upgrade process each node has to be rebooted (e.g. bios upgrade), so whole process of cluster actualization will last longer.

The Upgrade triggers reboot of a node in case it is need, so each node will be switched into the maintenance mode for the time of running upgrade process of that node.

The Firmware Manager waits for the upgrade firmware process of node with timeout. The default timeout is 15 minutes. When the timeout will elapse or the upgrade process of node will fail, the whole process of upgrading a cluster will stop.

If the upgrade process will fail, the cause of the fail has to be investigated. It may be e.g. a network failure or inappropriately ASP update.

Logs of Upgrade process are in `/var/log/firmware_manager.log`, and for each node from APS upgrade process are in `/tmp/sctmpdir[numbers]/tooltrace.txt`

9.7.2.1 Dependencies

- Management node:
  - Fpylib – internal cd10000 rpm package
  - Fpm – internal cd10000 internal rpm package

9.7.2.2 Handling errors

- On fail of upgrade specific error message with error code will be displayed.

Log file path:

- /var/log/firmware_manager.log

Log from ASP application on direct server path:

- /tmp/sctmpdir[numbers]/tooltrace.txt

9.7.3 CLI Commands

firmware-manager upgrade-firmwares – update bios and irmc on all storage nodes

This command updates bios and irmc on all storage nodes.

Syntax:

```
firmware-manager upgrade-firmwares
```

firmware-manager upgrade-irmc – update irmc on all storage nodes

This command updates irmc on all storage nodes.

Syntax:

```
firmware-manager upgrade-irmc
```

firmware-manager upgrade-bios – update bios on all storage nodes

This command updates bios on all storage nodes.

Syntax:

```
firmware-manager upgrade-bios
```

Parameters - optional:

`-h, --help`	This parameters shows help message and exit.
`n NODES [NODES ...], --nodes NODES [NODES ...]`	This is the node which firmware will be upgraded.
`-f, --force`	Unconditional installation of the firmware (suitable for update, overwrite and downgrade).
`-s, --skip`	Not interrupt upgrade firmware process on cluster if newer version is already installed.

The Upgrade of firmwares on the management node will be done separately by executing command with parameter: "**–n pmgmt**".

Example:

Upgrading management node bios with same version:

```
$ firmware-manager upgrade-bios –n pmgmt –f
```

Upgrading storage node irmc:

```
$ firmware-manager upgrade-irmc
```

10 Log Manager

ETERNUS CD10000 provides a centralized log manager. It collects configured log files from all storage nodes of a storage cluster in real-time and transfers them to the management node for supporting maintenance and support.

A command line tool for collecting log files on demand is available too.

10.1 Functionality of the Log Manager

The log manager provides the following fundamental functions:

Collecting log files

Whenever new entries are added to a log file, the specified set of log files is collected automatically in real-time from all storage nodes in the storage cluster and is stored in the database on the management node. For log file collection and forwarding Fluentd agents (forwarder agents) are configured on all storage nodes. They send events to Fluentd agents (aggregator agents) on the management node, which write them into an Elasticsearch database.

If required you can add new log files to the log manager configuration for real-time collection.

On the management node you can collect log files that are not automatically collected using an application for one-demand gathering of log files. The command **`cd10000 support gather_logs`** is provided for this purpose.

Viewing log files

You can view the collected log files with a log viewer and export them from the database, for example for sending a dump to the support. Additionally, data can be imported into the database.

As a log viewer you can use any viewer that is compatible with the Elasticsearch database, for example Kibana, a single-page web application.

Automatic daemon restart after failure

On each storage node and on the management node of the storage cluster a monitoring application, the monit daemon, is running, which restarts the Fluentd agent or the Elasticsearch database after failure.

SNMP

Failures of the log manager trigger the generation of an SNMP trap. For detailed information, see "Monitoring with SNMP" on page 169

10.2 Prerequisites for Using the Log Manager

The management node and all storage nodes are connected to a cluster network.

The management node is connected to a customer administration network where a log viewer must be deployed.

For detailed information on the networks, see "Networks" on page 33.

Software requirements

Log management station

Kibana requires Mozilla Firefox ≥ 17.0.10-1.

10.3 Architecture of the Log Manager

The figure below illustrates the architecture of the log manager and the cooperation of the involved components.

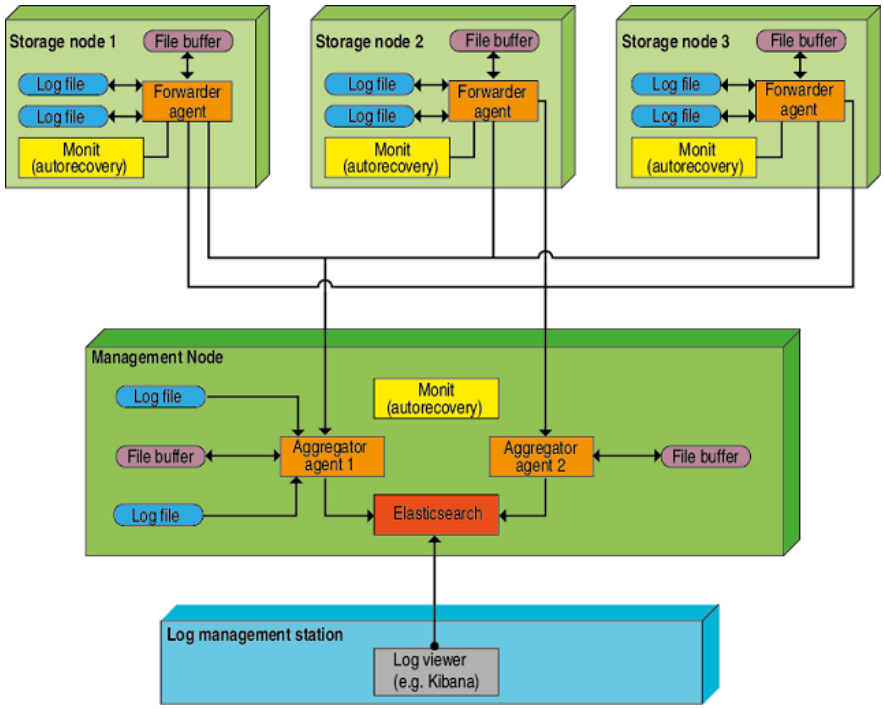


Figure 73: Architecture of the log manager

Runtime model

On the different nodes, the individual components of the log manager interact as follows:

On the storage nodes

The forwarder agent runs as a system daemon on each storage node. The forwarder agent is configured to send log files to the aggregator agents on the management node. For load balancing log files are split equally and each individual part is sent to a separate aggregator agent. If an aggregator agent is not available, the forwarder agent stores the log files in a local file buffer and resends them when possible.

The Monit daemon checks with intervals if the forwarder agent is still running. If it is not running, the Monit daemon restarts the forwarder agent. After three unsuccessful attempts of restarting the affected forwarder agent an SNMP trap is generated. For detailed information, see "Monitoring with SNMP" on page 169

On the management node

One instance of an Elasticsearch database serves as a storage for the collected log files. Each day the collected log files are added to a new index with the name: `prefix-year.month.day`, for example `logs-2014.01.30`.

Two aggregator agents are listening for log files from the forwarder agents. The first aggregator agent is also configured to collect log files from the management node. The aggregator agents write the log files into the Elasticsearch database.

If the database is not available the log files are stored in local file buffers.

The Monit daemon checks the Elasticsearch database and the aggregator agents with intervals and restarts them if they are not running. After three unsuccessful attempts of restarting an SNMP trap is generated.

Additionally, a gather application allows to gather log files from the storage cluster that are not collected in real-time.

On the log management station

Kibana is recommended for viewing the log files stored in the Elasticsearch database. It runs in a web browser, and with intervals it connects to the database via REST API and shows the log files and the corresponding statistics.

It is recommended to use only one instance of Kibana at a time, so that the database performance is not decreased.

Overall execution flow

The following components basically execute the data flow of the log manager:

Gather application

With the **`cd10000 support gather_logs`** command you can collect on-demand log files from one storage node, a set of storage nodes or all storage nodes.

The gather application copies the specified log files from each given storage node. The log files are transferred in series from the storage nodes and are stored in a separate subdirectory per storage node. The gathered files are compressed into one `tar.gz` file.

If the transfer from a storage node fails, an error is logged and the path from the concerned node is omitted.

Elasticsearch database

Each day, a new index consisting of the gathered log files is created in the Elasticsearch database. You can change the number of stored indexes in the configuration file of the log manager. For detailed information, see "Configuration of the Log Manager" on page 201.

Error handling

When the destination of the forwarder and/or aggregator agent(s) is not available, the log files are stored in a local buffer on the concerned storage node and/or on the management node for up to 24 hours. The log files are resent when the destination becomes available again. For detailed information, see "Specifications of the Log Manager" on page 201.

The Monit daemon restarts the forwarder and/or aggregator agents or the Elasticsearch database after a failure.

After three failed restart attempts an SNMP trap is generated.

10.4 Specifications of the Log Manager

The following specifications apply for the log manager:

Limits

- By default, the 30 last indexes are stored in the Elasticsearch database.
- It is recommended to use only one instance of a Log Viewer such as Kibana at a time, so that the database performance is not decreased.
- The log files are deleted from the local buffers on the management node and on the storage nodes after 24 hours.
- The local buffers on the management node and on the storage nodes can store up to 2 GBs of log files. When a buffer is full new log files are not stored.

Return codes of the cd10000 support gather_logs command

0 Successful.

1 Transfer error.

Locations of log entries and system messages

On the storage nodes

- /var/log/td-agent/td-agent1.log
- /var/log/monit

On the management node

- /var/log/td-agent/td-agent1.log
- /var/log/td-agent/td-agent2.log
- /var/log/monit
- /var/log/elasticsearch/
- /var/log/logmanager.log

10.5 Configuration of the Log Manager

The following parameters of the log manager can be modified:

Number of stored indexes

You can configure the number of stored indexes in the log manager configuration file:

```
/opt/fujitsu/flogmanager/etc/flogmanager.conf
```

`indexes_to_store` Number of indexes to be stored in the Elasticsearch database.
Old indexes are removed each day.
Value range: 0 to *
Default value: 30
Example:

```
indexes_to_store = 30
```

Real-time log files collection

By default, the log manager collects the following files.

From management node and storage nodes:

/var/log/ceph/ceph.log	Logs of the storage cluster
/var/log/fcephadm/fcephadm.log	Logs of vsm_cli and the ceph CLI
/var/log/messages	Logs of the operating system

From management node:

/var/log/fpm.log	Logs of the update-manager
/var/log/fcephadm/mkcephfs.log	Logs of the storage cluster deployment
/var/log/vsm/vsm-scheduler.log	Logs of VSM
/var/log/switches/br*.log	Logs of the internal administration switches
/var/log/switches/mlx*.log	Logs of the internal cluster switches

From storage nodes:

/var/log/vsm/vsm-agent.log	Logs of VSM

For changing the configuration of the collected log files, please contact the support of Fujitsu Technology Solutions.

10.6 Commands of the Log Manager

Commands of the Log Manager you can find in the chapter "Commands of the Log Manager" on page 285.

11 SNMP

You can monitor the ETERNUS CD10000 storage cluster via SNMP V2.

An SNMP management station must be available to monitor the activities. All SNMP traps generated on the storage nodes or switches of the storage cluster are sent to the management node. The management node forwards these traps to the SNMP management station(s) of the customer. Therefore the management node serves as a proxy for the storage nodes and switches.

Prerequisites

The SNMP management station must be connected to the customer's administration network. The management node of the storage cluster must be connected to the cluster internal administration network, as well as to the storage nodes, and the switches of the storage cluster.

For detailed information on the agents that deliver information on the state of all cluster components, see "Architecture of the SNMP Monitoring System" on page 204.

Network management software

For monitoring the storage cluster via the SNMP management station, you must install an SNMP-capable monitoring system, for example ServerView Operations Manager, HP OpenView, or Nagios, on the SNMP management station. For detailed information, see "Plugins for Network Management Software" on page 218.

MIBs

The following MIBSs are required:

`FUJITSU-CD-MIB.txt`

ETERNUS CD10000-specific MIB.

Required for the managed objects of the ETERNUS CD10000 SNMP agents.

`FSC-RAID-MIB.txt` Required for the managed objects of the ServerView RAID Manager

`SERVERVIEW-STATUS-MIB.txt` and `FSC-SERVERCONTROL2-MIB.txt`

Required for the managed objects of the ServerView agents.

Configuration of SNMP

You can configure the access to the managed objects and the destinations of traps via the `cd10000 setup snmp` CLI. For detailed information, see "SNMP Command Line Interface" on page 318.

11.1 Architecture of the SNMP Monitoring System

The figure below illustrates the architecture of the SNMP monitoring system and the cooperation of the involved components.

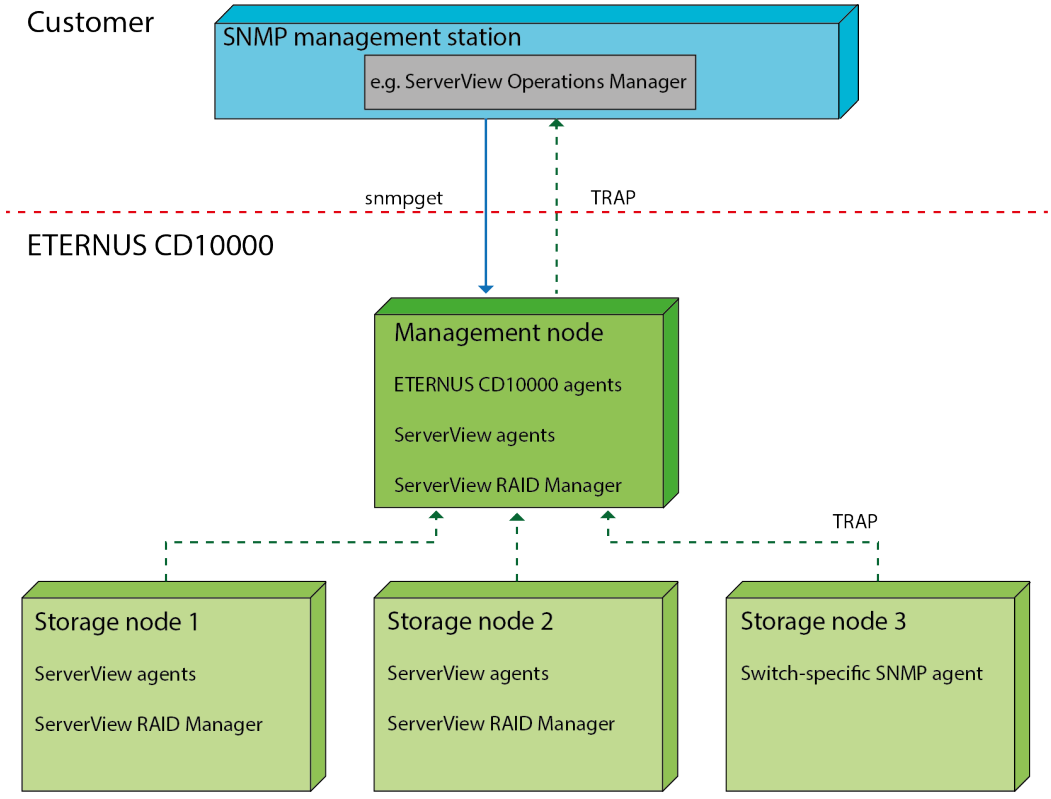


Figure 74: Architecture of the SNMP monitoring system

SNMP management station

You, the customer, are responsible for the SNMP management station. You have to install a network management software, which polls the management node and receives traps from the management node. For detailed information, see "Plugins for Network Management Software" on page 218.

Management node

The management node is connected to the customer administration network and to the storage nodes and switches. It serves as a proxy for the other hardware components (storage nodes, switches) of the storage cluster and sends all traps of all ETERNUS CD10000 components to one or more SNMP management stations. For this purpose the

Net-SNMP agent is installed on the management node. To monitor the hardware of the management node ServerView agents and ServerView RAID Manager are installed.

Furthermore special ETERNUS CD10000 SNMP agents that act as subagents to Net-SNMP are installed (SNMP V2). These subagents implement managed objects defined in the FUJITSU-CD-MIB to monitor the cluster as a whole. For detailed information on these managed objects, see "Managed Objects " on page 205.

Storage node

On the storage nodes the following software is installed for monitoring:

- Net-SNMP as SNMP agent.
- ServerView SNMP agents for monitoring the hardware (except the disks).
- ServerView RAID Manager for monitoring the disks.

Each storage node sends its traps to the management node.

Switch

On the switch the special SNMP agent for the switch is installed. The SNMP agent is pre-installed by the switch dealer.

All traps are sent to the management node.

11.2 Managed Objects

The following managed objects are defined in the MIB FUJITSU-CD-MIB for monitoring the storage cluster. All managed objects are accessable via the management node of the storage cluster. The access to the managed objects is configured via the ETERNUS CD10000 CLI.

cdClusterInfo

`cdClusterInfo` contains information on the storage cluster as a whole. It consists of the following entries:

Entry	Description
clusterState	Shows the overall state of the cluster. Possible values: ok, warning, critical
clusterUsed	Shows the totally used space of the cluster in percent. Possible values: 0 - 100
clusterOsds	Shows the number of all OSDs in the cluster. Possible values: > 0

Entry	Description
clusterOsdsUp	Shows the number of OSDs in the cluster which are up. Possible values: ≥ 0
clusterOsdsIn	Shows the number of OSDs in the cluster which are in. Possible values:≥ 0
clusterUsedGB	Shows amount of used space in gigabytes (1GB = 1073741824 bytes). Possible values: ≥ 0
clusterFreeGB	Shows amount of free space in gigabytes. Possible values: ≥ 0
clusterTotalGB	Shows total amount of space in gigabytes. Possible values: ≥ 0
clusterObjects	Shows the total amount of objects stored in the cluster. Possible values: ≥ 0

Table 3: Entries of cdClusterInfo

nodeTable

`nodeTable` contains information on each storage node. For each storage node it consists of the following entries:

Entry	Description
nodeDescr	Name of the node. Possible values: string
nodeStatus	Status of the node. Possible values: up, down

Table 4: Entries of nodeTable

monTable

monTable contains information on each monitor. For each monitor it consists of the following entries:

Entry	Description
monDescr	Name of the monitor. Possible values: mon.<n>
monStatus	Status of the monitor. Possible values: up, down
monNode	Node of the monitor. Possible values: string

Table 5: Entries of monTable

osdTable

osdTable contains information on each OSD. For each OSD it consists of the following entries:

Entry	Description
osdDescr	Name of the OSD. Possible values: osd.<n>
osdUpStatus	Indicates if the OSD is up or down. Possible values: up, down
osdInStatus	Indicates if the OSD is in or out. Possible values: in, out
osdNode	Node of the OSD. Possible values: string
osdUsed	Used space of the assigned disk in percent. Possible values: 0 - 100
osdDevice	Name of the device used for this device, for example "/dev/sdc"
osdUsedGB	Used space on this OSD, in GB
osdFreeGB	Free space on this OSD, in GB
Additionally, if a particular OSD is placed on an Intel PCI-E SSD drive, following entries are available:	

Entry	Description
ssdSerial	Serial of the SSD device
ssdTemperature	Temperature of the device (in celsius)
ssdAvailSpare	Percentage of available spare space (0-100)
ssdCriticalWarnings	String describing what critical warnings are flagged by the device, for example "Volatile Memory Backup System has failed"
ssdPowerOnHours	Cumulative time (described in hours) the device was powered-on and working
ssdMediaErrors	Amount of media (read and/or write) errors encountered

Table 6: Entries of osdTable

pgTable

`pgTable` contains information about each placement group in a cluster. Each placement group information contains the following :

Entry	Description
pgId	Identifier of the placement group
pgStatus	Status group, this placement group belongs to, used to quickly filter out the status of a placement group . Possible values of the status are: **`unknown`** **`OK`** **`warning`** **`critical`** **`fatal`**
pgStatusDesc	Placement group status as reported by Ceph.
pgObjects	Amount of objects stored in this placement group
pgSize	Actual size of this placement group in megabytes.

Figure 75: Entries of pgTable

For detailed information on possible traps, see "SNMP Traps" on page 209.

11.3 SNMP Traps

The following table shows the traps, the state transition which triggers the trap, the severity, and the action to be done if necessary. The destinations of traps are configured via ETERNUS CD10000 CLI. For detailed information on the actions to be done via CLI, see "SNMP Command Line Interface" on page 318.

Traps	State transition	Severity	Action
cdClusterOkTrap	Cluster state: !ok -> ok	OK	None.
cdClusterWarningTrap	Cluster state: ok -> warning	WARNING	Normally this is the result of a node, monitor, or OSD failure. ▶ Look at the action of the appropriate trap.
cdClusterCriticalTrap	Cluster state: !critical -> critical	CRITICAL	This can be the result of multiple node, monitor, or OSD failures. ▶ Look at the action of the appropriate trap. Further look for network failures.
cdNodeUpTrap	Node state: down -> up	OK	None.
cdNodeDownTrap	Node state: up -> down	CRITICAL	Check if the node is running. ▶ If the node isn't running, restart it. ▶ If the node is running, check for network errors.
cdMonUpTrap	Monitor state: down -> up	OK	None.

Traps	State transition	Severity	Action
cdMonDownTrap	Monitor state: up -> down	CRITICAL	▶ Restart monitor.
cdOsdUpTrap	OSD up state: down -> up	OK	None.
cdOsdDownTrap	OSD up state: up -> down	CRITICAL	Check if the associated disk is broken. ▶ If it is broken, replace the disk and restart the OSD. ▶ If it is not broken, restart the OSD.
cdOsdInTrap	OSD in state: out -> in	OK	None.
cdOsdOutTrap	OSD in state: in -> out	WARNING	First check if the OSD is down. ▶ If it is down, perform the action described for `cdOsdDownTrap`. Then call `ceph osd in <n>`, in which `<n>` is the number of the OSD. ▶ If it is not down, call `ceph osd in <n>`, in which `<n>` is the number of the OSD.

Traps	State transition	Severity	Action
cdClusterCapacityOkTrap	Cluster used capacity≥ 85% -> cluster used capacity < 85%	OK	N/A
cdClusterCapacityWarningTrap	Cluster used capacity <85% -> cluster used capacity @ <85%, 95%	WARNING	Free some data on nodes until capacity drops below 85%, or install additional disks/nodes.
cdClusterCapacityCriticalTrap	Cluster used capacity <95% -> cluster used capacity @ <95%, 100	CRITICAL	Free some data on nodes until capacity drops below 95%, or install additional disks/nodes.
cdSsdTempTrap	SSD temperature < 60'c -> SSD temperature ≥ 60'c	WARNING	Check the fans and assure proper airflow in that particular node, check the AC in your server room.
cdSsdSpareSpaceTrap	Previous SSD spare space > current SSD spare space	CRITICAL	Replace the drive as soon as possible.

Traps	State transition	Severity	Action
cdSsdCritWarningsTrap	Previous SSD warnings < current SSD warnings	CRITICAL	According to warnings occured, which include replacing the drive, lowering the drive temperature and calling service technician.
cdSsdMediaErrorsTrap	Previous SSD media errors < current SSD media errors	CRITICAL	Replace the drive as soon as possible.
cdMonitErrorTrap	N/A	CRITICAL	service mentioned in the trap description has failed and cannot be restarted automatically. Connect to specified host and manually check the service.
cdOsdNearfullTrap	OSD capacity ≥ 85%	WARNING	Trap sent when a particular OSD exceeds 85% of its capacity.
cdOsdFullTrap	OSD capacity ≥ 95%	CRITICAL	Trap sent when a particular OSD exceeds 95% of its capacity.
cdOsdNormalCapacityTrap	OSD capacity < 85%	NORMAL	Trap sent when a particular OSD capacity drops below 85%.
cdLinkDownTrap	One of Infiniband links goes down	WARNING	Trap sent when one of Infiniband links on specified node goes down. ▶ Check the cabling and switch.

Traps	State transition	Severity	Action
cdLinkLostTrap	Both or last Infiniband links goes down	CRITICAL	Trap sent when there is no Infiniband connectivity on specified node. ▶ Check cabling and switch.
cdLinkUpTrap	Any of Infiniband links goes up	NORMAL	Trap sent when at least one of Infiniband links goes back up.
cdSystemDiskCapacityOkTrap	System disk capacity < 85%	NORMAL	Trap sent when used capacity on system drive on specified node drops below 85%
cdSystemDiskCapacityCriticalTrap	System disk capacity ≥ 95%	CRITICAL	Trap sent when used capacity on system drive on specified node is at or exceeds 95%
cdSystemDiskCapacityWarningTrap	System disk capacity ≥ 85%	WARNING	Trap sent when used capacity on system drive on specified node is at or exceeds 85%
cdPgsDownTrap	At least one PG -> `down` state	CRITICAL	See http://ceph.com/docs/master/ for more information. Trap indicates disk/node/OSD/replication failure.
cdPgsInconsistentTrap,	At least one PG -> `inconsistent`state	CRITICAL	See http://ceph.com/docs/master/ for more information. Trap indicates disk/node/OSD/replication failure.

Traps	State transition	Severity	Action
cdPgsIncompleteTrap	At least one PG -> `incomplete` state	CRITICAL	See http://ceph.com/docs/master/ for more information. Trap indicates disk/node/OSD/replication failure.
cdPgsStaleTrap	At least one PG -> `stale` state	CRITICAL	See http://ceph.com/docs/master/ for more information. Trap indicates disk/node/OSD/replication failure.
cdPgsBackfillTooFullTrap	backfill of at least one pg is too full	CRITICAL	See http://ceph.com/docs/master/ for more information. Trap indicates disk/node/OSD/replication failure.
cdPgsDegradedTrap	At least one PG -> `degraded` state	WARNING	Check the placement groups. Trap indicates disk/OSD/node failure, but indication may also result from an increasing replication factor for one or more pools, in which case no action is required.
cdPgsCleanTrap	At least one PG -> `clean`	NORMAL	None. The trap indicates that placement groups were restored, finished replicating/peering, or otherwise returned to a healthy state.

Table 7: SNMP traps

11.4 Commands of the SNMP

Commands of the SNMP you can find in the chapter "SNMP Command Line Interface" on page 318.

11.5 Examples for an SNMP Data Output with snmpwalk

- You will get the table-formatted output using the **snmptable** command.
- For large outputs (like from **pgTable**), you might find the **snmpbulkwalk** command much faster than the **snmpwalk** command, used in the following examples.

Information on the storage cluster as a whole using the managed object cdClusterInfo

For detailed information on the entries of **cdClusterInfo**, see "Managed Objects " on page 205.

```
snmpwalk -v 2c -c vils_paryz paryz cdClusterInfo
FUJITSU-CD-MIB::clusterState.0 = INTEGER: ok(1)
FUJITSU-CD-MIB::clusterUsed.0 = INTEGER: 12
FUJITSU-CD-MIB::clusterOsds.0 = INTEGER: 4
FUJITSU-CD-MIB::clusterOsdsUp.0 = INTEGER: 3
FUJITSU-CD-MIB::clusterOsdsIn.0 = INTEGER: 3
```

Information on each storage node of the storage cluster using the managed object nodeTable

For detailed information on the entries **nodeTable**, see "Managed Objects " on page 205.

```
snmpwalk -v 2c -c vils_paryz paryz nodeTable
FUJITSU-CD-MIB::nodeIndex.1 = INTEGER: 1
FUJITSU-CD-MIB::nodeIndex.2 = INTEGER: 2
FUJITSU-CD-MIB::nodeIndex.3 = INTEGER: 3
FUJITSU-CD-MIB::nodeIndex.4 = INTEGER: 4
FUJITSU-CD-MIB::nodeDescr.1 = STRING: storage1
FUJITSU-CD-MIB::nodeDescr.2 = STRING: storage2
FUJITSU-CD-MIB::nodeDescr.3 = STRING: storage3
FUJITSU-CD-MIB::nodeDescr.4 = STRING: storage4
FUJITSU-CD-MIB::nodeStatus.1 = INTEGER: up(1)
FUJITSU-CD-MIB::nodeStatus.2 = INTEGER: up(1)
FUJITSU-CD-MIB::nodeStatus.3 = INTEGER: up(1)
FUJITSU-CD-MIB::nodeStatus.4 = INTEGER: up(1)
```

Information on each monitor of the storage cluster using the managed object monTable

For detailed information on the entries of **monTable**, see "Managed Objects " on page 205.

```
snmpwalk -v 2c -c vils_paryz paryz monTable
FUJITSU-CD-MIB::monIndex.1 = INTEGER: 1
FUJITSU-CD-MIB::monIndex.2 = INTEGER: 2
FUJITSU-CD-MIB::monIndex.3 = INTEGER: 3
FUJITSU-CD-MIB::monDescr.1 = STRING: mon.0
FUJITSU-CD-MIB::monDescr.2 = STRING: mon.1
FUJITSU-CD-MIB::monDescr.3 = STRING: mon.2
FUJITSU-CD-MIB::monStatus.1 = INTEGER: up(1)
FUJITSU-CD-MIB::monStatus.2 = INTEGER: up(1)
FUJITSU-CD-MIB::monStatus.3 = INTEGER: up(1)
FUJITSU-CD-MIB::monNode.1 = STRING: storage1
FUJITSU-CD-MIB::monNode.2 = STRING: storage2
FUJITSU-CD-MIB::monNode.3 = STRING: storage3
```

Information on each OSD of the storage cluster using the managed object osdTable

For detailed information on the entries of **osdTable**, see "Managed Objects " on page 205.

snmpwalk -c vils_paryz paryz osdTable

```
FUJITSU-CD-MIB::osdIndex.1 = INTEGER: 1
FUJITSU-CD-MIB::osdIndex.2 = INTEGER: 2
FUJITSU-CD-MIB::osdIndex.3 = INTEGER: 3
FUJITSU-CD-MIB::osdIndex.4 = INTEGER: 4
FUJITSU-CD-MIB::osdDescr.1 = STRING: osd.0
FUJITSU-CD-MIB::osdDescr.2 = STRING: osd.1
FUJITSU-CD-MIB::osdDescr.3 = STRING: osd.2
FUJITSU-CD-MIB::osdDescr.4 = STRING: osd.3
FUJITSU-CD-MIB::osdUpStatus.1 = INTEGER: up(1)
FUJITSU-CD-MIB::osdUpStatus.2 = INTEGER: down(2)
FUJITSU-CD-MIB::osdUpStatus.3 = INTEGER: up(1)
FUJITSU-CD-MIB::osdUpStatus.4 = INTEGER: up(1)
FUJITSU-CD-MIB::osdInStatus.1 = INTEGER: in(1)
FUJITSU-CD-MIB::osdInStatus.2 = INTEGER: out(2)
FUJITSU-CD-MIB::osdInStatus.3 = INTEGER: in(1)
FUJITSU-CD-MIB::osdInStatus.4 = INTEGER: in(1)
FUJITSU-CD-MIB::osdNode.1 = STRING: storage1
FUJITSU-CD-MIB::osdNode.2 = STRING: storage2
FUJITSU-CD-MIB::osdNode.3 = STRING: storage3
FUJITSU-CD-MIB::osdNode.4 = STRING: storage4
FUJITSU-CD-MIB::osdUsed.1 = INTEGER: 2
FUJITSU-CD-MIB::osdUsed.2 = INTEGER: 3
FUJITSU-CD-MIB::osdUsed.3 = INTEGER: 3
FUJITSU-CD-MIB::osdUsed.4 = INTEGER: 4
```

11.6 Plugins for Network Management Software

On the SNMP management station an SNMP-capable network management software has to be installed which polls the management node and receives traps from the management node.

Examples

- ServerView Operations Manager
- HP OpenView
- Nagios

For detailed information on the different software products refer to the particular product documentation in each case.

12 Service Support

This chapter describes how to enable the connection to the AIS server in the Fujitsu Support Center in order to provide call-home and dial-in services.

The support consists of two parts:

1. The first function is the possibility to get access to the ETERNUS CD10000 management node via SSH and VNC and transfer files to the ETERNUS CD10000 management node and vice versa. This is accomplished by installing and configuring AISConnect.
2. The second function is the possibility to send service calls to the service center. This is accomplished by a software stack consisting of rs2k, fsc-msg-switch and fsc-lxcall. fsc-lxcall monitors /var/log/messages (particularly messages generated by traps of ServerView agents, ServerView RAID Manager, ETERNUS CD10000 agents and traps send by switches) and delivers it via fsc-msg-switch, rs2k and AISConnect to the service center. Which messages trigger a service call can be configured in a file.

12.1 Architecture

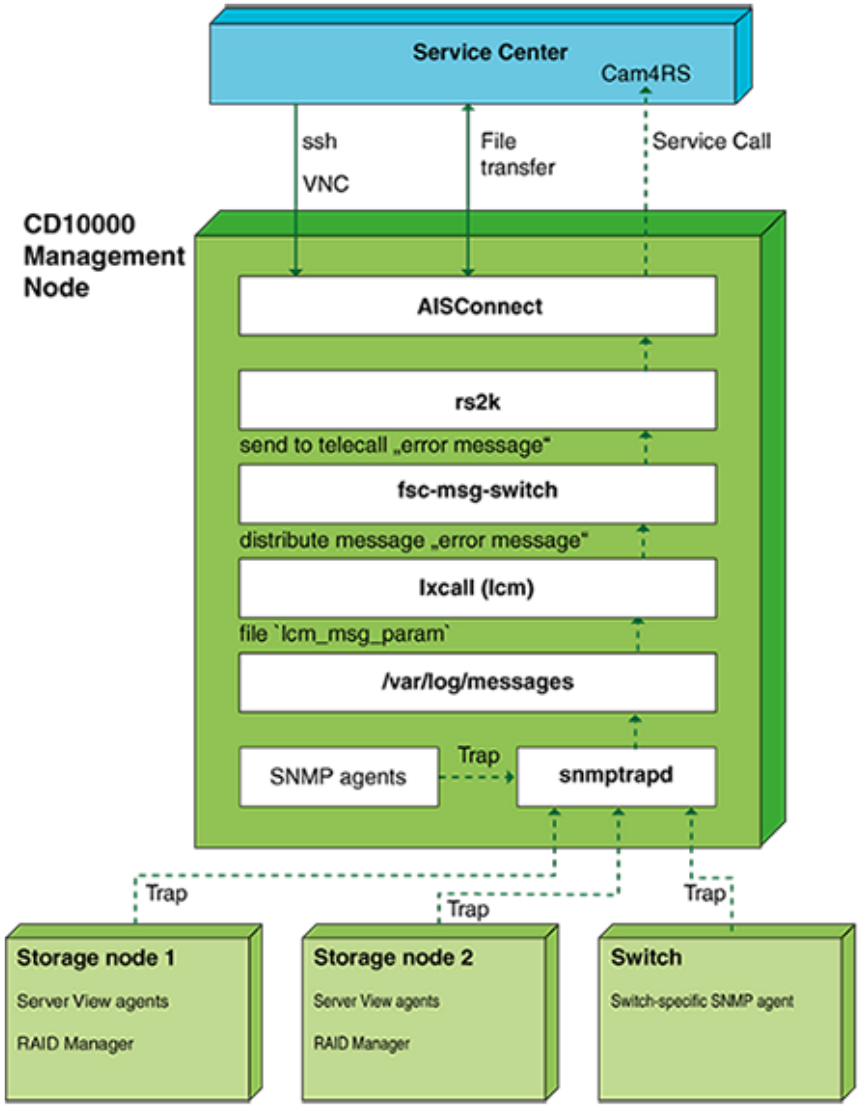


Figure 76: Architecture of management node

The ETERNUS CD10000 management node is connected to a Fujitsu service center.

The figure above shows the architecture of management node and the cooperation of the involved components inside.

12.2 Service Center

The service center is the basis of the service for the ETERNUS CD10000 cluster. Service engineers can access the ETERNUS CD10000 management node through the service center and transfer files to and from the management node. Furthermore, all service calls of the management node are sent to the service center.

cam4rs (call administration management for remote service) is a component which decides what to do with the call and forwards it to the appropriate component.

12.3 Components of the Management Node

Management node contains of several components. These are:

- "AISConnect" on page 222
- "rs2k" on page 223
- "fsc-msg-switch" on page 224
- "lxcall" on page 225
- "snmptrapd" on page 225

The management node serves as an interface to the service center. Service engineers can access the management node (via AISConnect) e.g. by opening a SSH shell. From there they can login to the storage nodes and the switches. They can transfer files from the service center to the management node and vice versa.

The management node also generates (via AISConnect) service calls which can trigger further actions.

The following chapters describe each component in detail.

12.3.1 AISConnect

Functions of the AISConnect are:

1. Enabling the access to the management node. There are several possibilities for the access to the management node e.g. SSH or VNC or the transfer of files from the management node to the service center and vice versa.
2. Sending service calls from the management node to the service center. The service calls are triggered through the following steps:
   a. A SNMP agent on the management node, a storage node or a switch sends a trap to the management node (see "Monitoring with SNMP" on page 169).
   b. The SNMP trap receiving daemon (**`snmptrapd`**) catches trap. **`snmptrapd`** writes a message to `/var/log/messages`.
   c. The daemon lcm of the component lxcall monitors `/var/log/messages`. **`lcm`** has a file (called `lcm_msg_param`) which defines the messages which will trigger a service call. If a message is specified in `lcm_mgs_param` **`1 cm`** will call the function **`distribute_message`** which is part of the component **`fsc-msg-switch`**.
   d. The function **`distribute_message`** will pass the message to **`rs2k`**.
   e. **`rs2k`** will pass the message to AISConnect
   f. AISConnect will send a service call to the service center.

VNC server

AISconnect contains a VNC server which is installed with the installation of AISconnect.

Adjust `/opt/c2sux/qsa/etc/vnc_usr` by adding:

```
c2suxadm tele
```

Adjust `/etc/gdm/custom.conf` to look like this:

```
# GDM configuration storage
[daemon]
[security]
DisallowTCP=false

[xdmcp]
Enable=true

[greeter]

[chooser]
IncludeAll=false
```

```
[debug]
```

Restart `gdm`: Execute the following commands:

```
telinit 3
telinit 5
```

Set password file comment field for user "`tele`" and "`c2suxadm`":

```
# usermod -c tele tele
# usermod -c c2suxadm -c c2suxadm
```

The VNC server is then started with `/opt/c2sux/qsa/bin/qsamon -p <password>` `/opt/c2sux/qsa/bin/qsamon`.

12.3.2 rs2k

rs2k is the link between fsc-msg-switch and AISConnect. It provides the interface */opt/rs2k/bin/send_to_telecall*.

`send_to_telecall "error message"`
sends an error message to AISConnect. The error message is afterwards forwarded to the service center.

The error message has the following layout:

`LIN <MessageKey> <MessageText>`
`<MessageKey>`
Unique identification for each error message. Each message key has to be registered for processing on the service center to guarantee uniqueness.

`<MessageText>`
Text of the error message.

Example:

LIN LXSM001 2014-04-29 15:15:18 paryz Serverview:[2014][CRITICAL] Fan 'FAN SYS5' in cabinet 0 of server paryz failed

12.3.3 fsc-msg-switch

fsc-msg-switch provides the following functions:

- Receiving error messages from various sources, e.g. lxcall.
  The function is provided by the interface distribute_message:

```
distribute_message "error message"
```

sends an error message to fsc-msg-switch. The format of the message is the same as for *send_to_telecall (see rs2k)*.

- Forwarding the error messages to various destinations, e.g. rs2k.

Logfile:

For diagnosis purposes all actions executed by *distribute_message* are logged in the following file, which is subject to the logrotate mechanism:

/var/log/msg-switch/distribute_message.log

Messages:

Each message handled by *distribute_message* is saved in a file. Messages from lxcall are stored in the following files:

Directory: /var/log/msg-switch/sas

Filename: msg_sas.%s.%N # %s: date/time in seconds.

fsc-msg-switch can receive messages from various sources and send messages to various destinations. In our case all messages came from lxcall and are sent to rs2k. Nevertheless fsc-msg-switch has two benefits:

1. It logs each message.
2. Further versions of ETERNUS CD10000 can implement more sources or more destinations for error messages.

12.3.4 lxcall

lxcall monitors the messages in */var/log/messages* and selects error messages according to defined criteria which are specified in a special file called *lcm_msg_param*. The process which does the monitoring is called lcm. If a message is found which hits the defined criteria, lcm calls the function *distribute_message* (see "fsc-msg-switch" on page 224).

Error message

The format of the error message is as follows:

```
LIN <MessageKey> <Timestamp> <hostname> <MessageText>
<MessageKey>
```

Unique identification for each error message. Each message key has to be registered for processing on the service center to guarantee uniqueness.

`<Timestamp>`

Timestamp consists of <date> and <time>. Basis is the timestamp of the selected message.

<date> yyyy-mm-dd

<time> hh:mm:ss

`<hostname>`

Name of the host.

Example:

/var/log/messages:

```
Jul 18 08:33:54 SAS1 Serverview: Battery voltage 1.5 predicted to fail
```

Error message for the function distribute_message:

```
LIN LXSM038 2014-07-18 08:33:54 SAS1 Serverview: Battery voltage 1.5 predic
ted to fail
```

Monitor process

There is a monitor process called *lcm_sv* which monitors *lcm*. If *lcm* is terminated with a signal unequal SIG_TERM then *lcm_sv* is terminated too. If *lcm* is terminated for a different reason it is restarted by *lcm_sv*.

12.3.5 snmptrapd

`snmptrapd` receives all traps sent to the management node. If no trap forwarding is specified (see "SNMP Initial Configuration" on page 52), a message is written to */var/log/messages* for each trap.

12.4 Configuration

This chapter describes configuration of the management node components.

12.4.1 Configuration of AISConnect

The installation of AISConnect creates the users *c2suxadm* and *tele*. Unfortunately the created rights of *tele* are not sufficient for the access via VNC. Therefore it is recommended to create the user *tele* in advance.

Configuration of AISConnect is done via c2suxadm tool. These are example answers for questions that cannot be left with default values.

Note that some characters (like underscore) are not permitted. In some cases, the qsaconfig tool might accept them, but AISConnect server can't, which will be resulting in connection errors.

Qsaconfig, when run on an empty management node, lets you use defaults. Once configured, previously set values are new defaults and can be left blank (which indicates no change to that parameter).

```
# su c2suxadm
# /opt/c2sux/qsa/bin/qsaconfig
Please, enter the name of the asset
->cd8-paryz                # This is the name which shows up at the Service
Center
Please, enter serial number
->CD10000PARYZ             # Cluster serial number
Read model number
Type
"1" for BladeFrame,
"2" for SparcServer,
"3" for ETERNUS CS,
"4" for Mainframes,
"5" for AIS Connect Gateway,
"6" for ETERNUS CD,
the appropriate string to identify the system,
e.g. "CentricStor V4"Default: "Allround Server"->6
```

```
ModelNumber is "ETERNUS CD"
```

```
->Storage                   # this parameter defines the service group which
is responsible
Please, enter ip address of local machine in quad dotted format
->172.17.33.165
Please, enter (1st) ip address of main enterprise server of the provider
in quad dotted or dns format followed by ":" and the port
->connect2service.ts.fujitsu.com.com:443  # this parameter defines the serv
ice center
Please, enter (2nd) ip address of total access enterprise server of the
provider in quad dotted or dns format followed by ":" and the port
->connect2agent.ts.fujitsu.com:443
Please, enter ip address of proxy to internet
in quad dotted or dns format followed by ":" and the port
->172.25.139.4:82    # this must be provided by customer
Configure agent features
If No use the defaults:
access policy:enabled,
session log:disabled
SSH:off,VNC:0,HTTP:80(access to '/'),
SoftWareDirector:N,FileTransfer:N,PropertyCheck:N
[Y|n]:Y
Please, enter if you want to permit Filetransfer [Y|N]
mandatory, from 1 to 1 character(s)
default is "N"
enter "q" to exit
->Y
Please, enter if you want to permit "SSH", give port number, else "0"
mandatory, from 1 to 5 character(s)
default is "0"
enter "q" to exit
->22
Please, enter if you want to permit "VNC", give port number, else "0"
mandatory, from 1 to 5 character(s)
default is "0"
enter "q" to exit
->5900
```

```
Read tty shadow mode (based on screen command)
=============================================================
Type    "0" or anything else for no shadow possible (default),
"1" for shadow is possible,
"2" for shadow is necessary
->1
/opt/c2sux/qsa/saf/vnccfg: WARNING: No fonts found in /usr/share/fonts!
/opt/c2sux/qsa/saf/vnccfg: WARNING: No windowmanager found
File /opt/c2sux/qsa/etc/vnc_param.dist successfully written
Ports in file /opt/c2sux/qsa/etc/vnc_param has been updated!
configuring ais connect begins
Start c2sux as service [Y|n] Y
Start c2sux under "c2suxadm" [Y|n] Y
generate configuration
Generating /opt/c2sux/qsa/etc/smartpollcfg.xml...
Generating /opt/c2sux/qsa/etc/tetherhscds.xml...
register configuration
configuring ais connect ends
Start c2sux now [y|N] y
```

12.4.1.1 Configuration file

The configuration is stored in */opt/c2sux/qsa/etc/.parameter_c2sux* file. Once set, it might be reused later (values from configuration file will be suggested as defaults).

12.4.2 Configuration of rs2k

```
# /opt/bin/rs2kadmin
(1) Call  - Call Configuration
(2) Ftbs2 - Filetransfer Configuration
(3) Help  - show help text
(4) End   - Exit
your choice ->1
```

The first time you are guided through a menu. After that you get an overview of all configured parameters.

```
+- CONFIGURE CALL
(1)  receive Call from other system    N
(2)  send Call      directly to service center via c2s
(6)  system type      Linux        (7)  serial no.      paryz
TS-Key
(10) Region  mch          (11) SKP-Kz  0: Other
(12) KS      paryz        (13) BS-Kz   0: Other
(15) add. info   00
(16) Call activated    Y
(17) Help              (18) Save              (19) Exit
Message: wrong choice
your choice ->
(2)  send Call      directly to service center via c2s
(7)  serial no.      paryz
(10) Region  mch
(12) KS      paryz
(16) Call activated    Y
(18) Save
(19) Exit
```

Generate test call

```
# /opt/bin/rs2_testcall
Nr.   Msgkey      Message                Class Destination  Devices
(1) - PSC0190     Remote-Service-Call    PSC   Cam4RS       all devices
(2) - SQ12345     Remote-Service-Call    LIN   Cam4RS       all devices
(3) - S123456789  Remote-Service-Call    FLA   Cam4RS       skp at S-Server
(4) - MER9999     Remote-Service-Call    SYS   Cam4RS       skp/Q-Server
(q) - quit without a call
--> 1
sending call, please wait
done
```

Logfiles

```
/var/opt/rs2k/prot/Calls_prot.<datetime>
```

12.4.3 Configuration of lxcall

The file *lcm_msg_param* defines the messages which are (via *distribute_message*) sent to fsc-msg-switch. Only messages defined in this file are sent.

lcm_msg_param

The lcm_msg_param starts with:

```
#ident "$Id: lcm_msg_param,v 1.13 2013/07/01 15:36:30 stc Exp $"
#-------------------------------------------------------------------------
 # $Copyright:
# Copyright (C) 2009-2013 Fujitsu Technology Solutions, All Rights
Reserved.
# $
#-------------------------------------------------------------------------
(INFO)
version=1.12
# Messages from navi_watch (monitoring of naviagent) ==================
[TSNAV01]
T+="navi_watch"
T+="Navisphere Agent is not running"
TSNAV02]
T+="navi_watch"
T+="Navisphere Agent"
T+="usage"
```

```
# Server Management Messages ==========================================

[LXSM001]
T+="Serverview"
T+="Fan"
T+="failed"

[LXSM002]
T+="Serverview"
T+="Temperature"
T+="warning"

[LXSM003]
T+="Serverview"
T+="Temperature"
T+="critical"

[LXSM004]
T+="Serverview"
T+="AC failure
```

Version:

`(INFO)` Indicator that this is a file info.

`version=<version>` Version of the file

Messages:

There exists an entry for each message which should be sent to the service center. Each entry consists of the following components:

`[<MessageKey>]`	<MessageKey> is a 7-digit key which is unique for each message. It is identical to the message key of the messages sent to fsc-msg-switch.
`T+="<string>"`	<string> defines a string which have to be contained in the message.
`T-="<string>"`	<string> defines a string which must not be contained in the message. This entry is optional. It has to be preceded by at least one T+ entry.

T+ entries have to be defined consecutively in the same order as they occur in the selected message.

T- entries have to be defined after T+ entries. The order of T- entries is arbitrary.

Double quotes inside <string> can be used without escape character.

13 Preparing VSM Cluster Creation Manually

VSM needs some information about the Ceph storage clusters which it shall create. Therefore you must provide configuration files, in which you describe the cluster configuration and the configuration of the particular nodes,

The configuration and administration of Ceph storage clusters via VSM requires the following manifest files:

- the `cluster.manifest` file on the VSM controller node.

  In a CD10000 configuration the VSM controller node is running on the management node.

  In the `cluster.manifest` file you must define the global configuration parameters for the cluster.

  A template for the `cluster.manifest` file is included in the ETERNUS CD10000 delivery. You will find it in the directory `/etc/manifest/`

- a server manifest file `/etc/manifest/server.manifest` on each node

  The server manifest file is used by the VSM agent to identify and authenticate with the VSM controller, and determines the node configuration.

  A template for the server manifest file of a node will be automatically created during the software installation of ETERNUS CD10000. From this you only need to configure the server.manifest file.

You will find a template of the server.manifest file containing the generated mapping of journals to storage devices and the available disks at `/etc/fcephadm/server.manifest`. The disks are grouped in storage groups according to their type. The file can be copied to the `/etc/manifest/server.manifest` file without changes.

13.1 Cluster Manifest File

To configure and administer a Ceph storage cluster with VSM a cluster manifest file `/etc/manifest/cluster.manifest` must exist on the VSM controller node. In this file you must define;

- Storage Classes (performance, capacity)
- Storage Groups (the performance group is mandatory as first entry)
- the network configuration
- other management details

After finishing the cluster.manifest file you can start the VSM-controller on the management node, see "Checking the cluster.manifest file, starting the VSM controller" on page 240.

13.1.1 Format of the cluster.manifest file

The cluster manifest file is splitted into the following sections:

`[storage_class]`

Defines the **Storage Classes** of the disks within the storage cluster.

Disks with similar performance characteristics are identified by **Storage Classes**. Examples of Storage Classes are:

7200_RPM_HDD, 10K_RPM_HDD, 15K_RPM_HDD

`[storage_group]`

Defines **Storage Groups** within the storage cluster.

Disks with the same Storage Class are grouped together in Storage Groups. Storage Groups are paired with Storage Classes. For example:

Storage Group	= **Storage Class**
Performance	= 10k_RPM_HDD
Capacity	= 7200_RPM_HDD
High Performance	= 15k_RPM_HDD

Additionally you can define a user friendly storage group name for each group. This name must be enclosed in quotes.

Input format:

Per storage group you must enter a row in the following format:

<storage group name> "<user friendly storage group name>" <storage class>

The performance group is mandatory as first entry.

[cluster]

Specifies the name of the Ceph storage cluster.

[file_system]

Specifies the file system which is used for the OSDs of the storage cluster.

Possible values are: `btrfs`, `xfs`, `ext4`

Default: xfs

With this software release only xfs is supported .

[management_addr]

Subnet address of the administration network. ETERNUS CD10000 uses 192.168.20.0/23 as administration network.

[ceph_public_addr]

Subnet address of the public (client) network.

[ceph_cluster_addr]

Subnet address of the cluster network. ETERNUS CD10000 uses 192.168.40.0/23 as cluster network.

[settings]

Specifies further global configuration parameters of the Ceph storage cluster. You can define the following parameters:

`storage_group_near_full_threshold`

Specifies the threshold for "near full" in percent. If the capacity utilization of a particular storage group exceeds this threshold, VSM warns: **Storage Group Near Full**.

A threshold "65" means 65% of the capacity.

Default: 65

`storage_group_full_threshold`

Specifies the threshold for "full". If the capacity utilization of a particular storage group exceeds this threshold, VSM warns: **Storage Group Full**.

A threshold "85" means 85% of the capacity.

Default: 85

`ceph_near_full_threshold`

Specifies the threshold for "storage near full" in percent. VSM displays a warning if the capacity utilization of the whole storage cluster exceeds this threshold.

A threshold "75" means 75% of the capacity.

Default: 75

`ceph_full_threshold`

Specifies the threshold for "storage full". Ceph will stop accepting writes if the capacity utilization of of the whole storage cluster exceeds this threshold until capacity is added to the storage cluster.

A threshold "90" means 90% of the capacity.

Default: 90

`pg_count_factor`

Factor to calculate the number of placement groups per storage pool.

The number of placement groups is derived from the following calculation:

`[pg_count_factor]` * number of OSDs/replication factor of the storage pool.

Default: 100

`heartbeat_interval`

Time interval in seconds that specifies how often the OSD daemon will ping a Ceph monitor daemon for the most recent copy of the cluster map in case the OSD daemon has no OSD daemon peers.

Default: 5

`osd_heartbeat_interval`

Time interval in seconds that specifies the regular time interval in which each OSD daemon checks the heartbeat of the other OSD daemons defined in the cluster map.

Default: 36

`osd_heartbeat_grace`

Specifies the grace time (ín seconds) an OSD daemon will wait without getting a heartbeat from a neighboring OSD daemon. After this period the OSD daemon will consider the neighboring OSD daemon to be down and reports this back to one of the cluster's monitor daemons. The monitor daemon will update the cluster map.

Default: 240

`mon_lease`

The length (in seconds) of the lease on the monitor’s versions.

Default: 20

`mon_lease_renew_interval`

The interval (in seconds) for the Leader to renew the other monitor’s leases, where "Leader" is a role of a monitor for the purpose to retrieve the most recent version of the cluster map.

For more information on Monitor Store Synchronization, see http://ceph.com/docs/master/rados/configuration/mon-config-ref/.

Default: 12

`mon_lease_ack_timeout`

The number of seconds the Leader will wait for the Providers to acknowledge the lease extension, where "Leader" and "Provider" are different roles of monitors for the purpose of synchronization.

For more information on Monitor Store Synchronization, see http://ceph.com/docs/master/rados/configuration/mon-config-ref/.

Default: 40

[ec_profiles]

Specifies the Ceph erasure code profiles.When the operator creates an erasure coded profile, the available profiles are displayed and can be selected. It is not possible to modify a profile. For more information on creating erasure coded pools, see "Creating Erasure Coded Storage Pools" on page 104.

The following settings are supported for each specified erasure code profile:

`[profile-name]`

User-visible name of the erasure code profile. When the operator creates an erasure coded pool, this name is displayed by VSM.

`[path-to-plugin]`

Specifies the name of the directory from which the erasure code plug-in is loaded.

`[plugin-name]`

Name of the erasure code plug-in which is used to compute coding chunks and recover missing chunks.

`[pg_num value]`

Specifies the value of `pg_num` used when creating an erasure coded pool with this profile.

`[json format key:value]`

Sequence of json-formatted **`key:value`** pairs used to configure the reassured code plug-in. The **`key:value`** pairs must match the semantics expected by the specified erasure code plugin. The **`key:value`** strings should not have spaces and must be included in quotation marks.

Input format:

For each erasure code profile you must enter a row in the following format:

```
<profile-name> <path-to-plugin> <plugin-name> <pg_num value> {<json format key1:value1>,<json format key2:value2>,...}
```

for example:

```
profile_name1 /usr/lib64/ceph/erasure-code jerasure  3 {"k":2,"m":1,"technique":"reed_sol_van"}
```

[cache_tier_defaults]

Specifies the cache tier default values. These values are displayed as the default values when the operator creates a cache tier, and can be modified by the operator to satisfy the intended use of the cache tier. Cache tier creation is described in section "Adding Cache Tiers" on page 106.

The following cache tier default settings are supported:

`ct_hit_set_count`

Specifies the time that each HitSet covers.

Default: 1

`ct_hit_set_period_s`

Specifies the number of HitSets to store

Default:3600

`ct_target_max_mem_mb`

Specifies the maximum number of bytes at which the cache tiering agent can flush or evict objects.

Default: 1000000

`ct_target_dirty_ratio`

Specifies the ratio of dirty objects to pool size. If this ratio is exceeded the cache tiering agent will flush dirty objects to the storage pool.

Default: 0.4

`ct_target_full_ratio`

Specifies the percentage of cache pool capacity at which the cache tiering agent will evict objects to maintain free capacity.

Default: 0.8

`ct_target_max_objects`

Specifies the maximum number of objects held in cache tier

Default: 1000000

`ct_target_min_flush_age_m`

Specifies the minimum age of an object before the cache tiering agent flushes a recently modified (i.e. dirty) object to the backing storage pool, in seconds

Default: 10

`ct_target_min_evict_age_m`

Specifies the minimum age of an object before it will be evicted from the cache tier, in seconds.

Default: 20

13.1.2 Checking the cluster.manifest file, starting the VSM controller

After you have changed the cluster.manifest file according to your needs, execute the following steps on the management node:

- Check the format of the cluster.manifest file by running the following command:

  `cluster_manifest`

  Make sure the result contains: `Check Success`
- After you have finished the previous step, start the VSM controller at the management node by running the following command:

  `vsm-controller -r controller`

  This step can take up to 10 minutes to finish.

13.1.3 Example of a cluster manifest file

```
[storage_class]
10krpm_sas
7200_rpm_sata
ssd
#ssd_cached_7200rpm_sata
#ssd_cached_10krpm_sas
#7200RPM_SATA_OBJECT
#7200RPM_SATA_BLOCK

[storage_group]
#format: [storage group name]  ["user friendly storage group name"] [storage class]
performance          High_Performance_Disk    10krpm_sas
capacity             Economy_Disk    7200_rpm_sata
high_performance     High_Performance_SSD    ssd
#value_performance "High_Performance_Disk_with_ssd_cached_Acceleration"    ssd_cached_1
0krpm_sas
#value_capacity    "Capacity_Disk_with_ssd_cached_Acceleration"    ssd_cached_7200rpm_s
ata
#capacity_object    Capacity_Object    7200RPM_SATA_OBJECT
#capacity_block     Capacity_Block     7200RPM_SATA_BLOCK

[cluster]
cd10000

[file_system]
xfs

[management_addr]
192.168.20.0/23

[ceph_public_addr]
192.168.110.0/24

[ceph_cluster_addr]
192.168.40.0/23
```

```
[settings]
storage_group_near_full_threshold  65
storage_group_full_threshold  85
ceph_near_full_threshold  75
ceph_full_threshold  90
pg_count_factor 100
heartbeat_interval 5
osd_heartbeat_interval 36
osd_heartbeat_grace 240
mon_lease 20
mon_lease_renew_interval 12
mon_lease_ack_timeout 40

[ec_profiles]
#format: [profile-name] [path-to-plugin] [plugin-name] [pg_num value] [json format key/
value]
##the key/value strings should not have spaces
profile_name1   /usr/lib64/ceph/erasure-code  jerasure  3  {"k":2,"m":1,"technique":"re
ed_sol_van"}
#profile_name2  /usr/lib64/ceph/erasure-code  jerasure  6  {"k":4,"m":2,"technique":"re
ed_sol_van"}

[cache_tier_defaults]
ct_hit_set_count 1
ct_hit_set_period_s 3600
ct_target_max_mem_mb 1000000
ct_target_dirty_ratio 0.4
ct_target_full_ratio 0.8
ct_target_max_capacity_gb 1000
ct_target_max_objects 1000000
ct_target_min_flush_age_m 10
ct_target_min_evict_age_m 20
```

13.2 Server Manifest File

A server manifest file must reside on each storage node and defines how the storage of the node is configured. It identifies all storage devices and associated journal partitions on the node. To add a storage node to the cluster the server manifest file must contain the IP address of the VSM controller, and a valid authentication key.

A valid authentication key must be generated from the VSM controller on the management node as follows:

- When the VSM controller is running, enter the command

  `agent-token`

When the VSM agent first runs, it contacts the VSM controller and provides the authentication key located in the server manifest file. Once validated, the VSM agent is always recognized by the VSM controller.

The authentication key generated by the VSM controller is only valid for 24 hours. After that, you will need to generate a new authentication key.

After you have finished the server.manifest file you can start the VSM agent on the storage node, see "Checking the server.manifest file, starting the VSM agent" on page 245.

13.2.1 Format of the server.manifest file

You will find a template of the server.manifest file containing the generated mapping of journals to storage devices and the available disks at `/etc/fcephadm/server.manifest`.

The server manifest file is splitted into the following sections:

`[vsm_controller_ip]`

Specifies the IP address of the VSM controller. The VSM agent needs this IP address to contact the VSM controller.

ETERNUS CD10000 uses always 192.168.20.1 as VSM controller address .

`[role]`

Specifies the role of a storage node. Possible roles are:

`storage`

Disks within the node shall be used as OSDs of the storage cluster.

`monitor`

A monitor daemon shall run on this node.

The node may have one of these roles or both.

You must activate the `monitor` role at three storage nodes at least.

`[auth_key]`

Specifies a valid authentication key. The VSM agent uses this key to authenticate with the VSM controller.

You have to generate the authentication key with the aid of the VSM controller. Therefore enter the command `agent-token` on the management node.

Notice: An authentication key is only valid for 24 hours after it has been generated. After that, you will need to generate a new key and to enter it here.

`[<storage_class>] (e.g. [7200_rpm_sata])`

Specifies the storage devices which shall be used as OSDs.

Storage devices are organized by their Storage Classes (as defined in Cluster Manifest). Devices and the corresponding journal devices are specified in by-path notation to ensure that paths remain constant in the event of a device removal or failure.

On each storage node you will find the file `/etc/fcephadm/server.manifest` that contains all required mapping of journals to storage devices.

Input format:

```
<storage_device> <journal_device>
```

Example:

```
[7200_rpm_sata]
/dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:1:0 /dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:3:0-part5
/dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:2:0 /dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:3:0-part6
```

13.2.2 Checking the server.manifest file, starting the VSM agent

Prerequisite: Before starting any VSM agent the VSM controller must be running on the management node.

After you have changed the server.manifest file according to your needs, execute the following steps on the storage node:

- Check the format of the server.manifest file by running the following command:

  `server_manifest`

  Make sure the result contains: `Check Success`
- After you have finished the previous step, start the VSM agent at the storage node by running the following command:

  `vsm-node`

  This step can take up to 5 minutes to finish.

Prepare all storage nodes (at least 4) in this way.

13.2.3 Example of a server manifest file

```
[vsm_controller_ip]
192.168.20.1

[role]
storage
monitor

[auth_key]
994ee70f1f3b44de953f090452d26355-5585985d26fc44cdb07e6b9448b824fb

[ssd]

[7200_rpm_sata]

[10krpm_sas]
/dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:1:0  /dev/disk/by-path/pci-0000:01:00.0-scs
i-0:2:3:0-part5
/dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:2:0  /dev/disk/by-path/pci-0000:01:00.0-scs
i-0:2:3:0-part6
/dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:5:0  /dev/disk/by-path/pci-0000:01:00.0-scs
i-0:2:3:0-part7
```

```
/dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:6:0  /dev/disk/by-path/pci-0000:01:00.0-scs
i-0:2:3:0-part8
/dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:7:0  /dev/disk/by-path/pci-0000:01:00.0-scs
i-0:2:3:0-part9
/dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:8:0  /dev/disk/by-path/pci-0000:01:00.0-scs
i-0:2:3:0-part10
/dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:9:0  /dev/disk/by-path/pci-0000:01:00.0-scs
i-0:2:4:0-part5
/dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:10:0 /dev/disk/by-path/pci-0000:01:00.0-scs
i-0:2:4:0-part6
/dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:11:0 /dev/disk/by-path/pci-0000:01:00.0-scs
i-0:2:4:0-part7
/dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:12:0 /dev/disk/by-path/pci-0000:01:00.0-scs
i-0:2:4:0-part8
/dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:13:0 /dev/disk/by-path/pci-0000:01:00.0-scs
i-0:2:4:0-part9
/dev/disk/by-path/pci-0000:01:00.0-scsi-0:2:14:0 /dev/disk/by-path/pci-0000:01:00.0-scs
i-0:2:4:0-part10

[ssd_cached_7200rpm_sata]

[ssd_cached_10krpm_sas]
```

13.3 VSM Dashboard

From the VSM dashboard you can create a ETERNUS CD10000 Ceph cluster.

The VSM dashboard can be accessed on the Web console using the URL of the management node:

`http://<management_node>/dashboard`

The initial credentials to access the VSM dashboard are:

- the login (**User Name**) **`admin`** and
- the **`ADMIN_PASSWORD`** entry from the file **`/etc/vsmdeploy/deployrc`** on the management node as "first time password"

At first login you will be redirected to the **Create Cluster** window automatically.

For detailed information on the VSM dashboard and how to access it, see "Virtual Storage Manager for Ceph" on page 57

14 Command Line Interfaces

14.1 The Command Line Interface vsm_cli

The ETERNUS CD10000 Command line interface **vsm_cli** provides commands for administering, maintaining and monitoring Ceph storage clusters.

An overview of the available vsm_cli commands is given in chapter "Overview of the vsm_cli Commands" on page 63.

14.1.1 Call Format

You must enter the **vsm_cli** commands as follows:

vsm_cli <cluster_name> <command> [<parameters>]

Where:

<cluster_name>	Specifies the name of the involved storage cluster.
<command>	Specifies the operation. You must enter one of the You can take the **vsm_cli** commands from the lists in "Overview of the vsm_cli Commands" on page 63.
<parameters>	Indicates command specific parameters like names of storage nodes, OSDs, monitors etc.

Output

All output is in plain text.

14.1.2 Messages and Common Return Codes

The following common return codes are supported:

Return Code	Description
0	Normal end. All actions were successful.
≥ 1	Error end. All or a part of actions failed.

The return codes will be defined as follows:

```
<msg-id> :<errflag>:<message>:<reason>
```

Where:

```
<msg-id> = <component><number>
```

`<msg-id>`	Message ID
`<component>`	Specifies the component which generates the return code FCA
`<number>`	The four-digit message number
`<errflag>`	The error flag specifies the message type. Possible values are: `E` = error, `N` = notice, `W` = warning
`<message>`	General description of the error/notice/warning,
`<reason>`	Additional information as to the cause of the error error/notice/warning

Example:

```
FCA1022:N:STATUS#health:HEALTH_OK #osd:12 osds,12 up,12 in#Storage utilization:0.013%
```

14.1.3 vsm_cli commands

This section describes the syntax of the **`vsm_cli`** commands.

activate - Activating a New Cluster

The **activate** command activates a new storage cluster.

Before activating a storage cluster you must prepare the disks of this cluster using the **prepare-disk** command.

Syntax:

```
vsm_cli <cluster_name> activate[ -f filesystem]
        [ -c caddr][ -p paddr] <node_1> <node_2> …
```

Parameters:

<cluster_name>	The name of the storage cluster.
-f <filesystem>	The file system used for the OSDs. Possible values are: xfs (default), btrfs, ext4. The specified file system must match the file system used in the **prepare-disk** command. The specified file system will override the setting in the cluster manifest file **/etc/manifest/cluster.manifest**. **Default:** The file system specified in **/etc/manifest/cluster.manifest**
-c <caddr>	The IP subnet address of the Ceph cluster network. The specified IP address will override the corresponding setting in the cluster manifest file **/etc/manifest/cluster.manifest**. **Default:** The IP address of the Ceph cluster network specified in **/etc/manifest/cluster.manifest.** "Cluster Manifest File" on page 234

`-p <paddr>`

The IP subnet address of the Ceph public network.

The specified IP address will override the corresponding setting in the cluster manifest file **`/etc/manifest/cluster.manifest`**.

Default:

The IP address of the Ceph public network specified in **`/etc/manifest/cluster.manifest.`**

`<node_1> <node2> … <node_n>`

The name(s) of the node(s) that will be used to create the storage cluster.

The disks on this nodes must be prepared using the **`prepare-disk`** command

The disks to be used as OSDs will be enumerated in the **`/etc/manifest/server.manifest`** for each node.

Output:

None

add-mon - Adding a Monitor Daemon to a Storage Cluster

The **add-mon** command creates a new monitor daemon and adds it to an existing storage cluster.

Please note:

- Per each **add-mon** command you can only create one new monitor daemon.
- The total number of monitor daemons in a storage cluster should be odd-numbered. Therefore you should add two new monitor daemons at least.
- The storage cluster the monitor daemon shall be added to must run.

If a monitor daemon shall be replaced you must remove it with the **remove-mon** command and add a new one with the **add-mon** command.

CD10000 storage cluster must have three monitor daemons at least. Three monitor daemons are sufficient for 100 storage nodes.

For a storage cluster with more than 100 storage nodes you should create two new monitor daemons and add them to the storage cluster.

Syntax:

```
vsm_cli <cluster_name> add-mon <new_mon_node>
```

Parameters:

<cluster_name>	The name of an existing storage cluster.
<new-mon-node>	The host name of the node where the monitor daemon shall run.

Output:

None

create - Creating a New Storage Cluster

The **create** command creates a Ceph storage cluster. This command combines the functionality of the commands **prepare-disk** and **activate**.

The **create** command requires the following manifest files:

- the cluster manifest file on the management node (**/etc/manifest/cluster.manifest**), which

  contains the global configuration parameters for the storage cluster.
- a server manifest file on each storage node (**/etc/manifest/server.manifest**), which identifies all storage devices and associated journal partitions on the node.

All options specified in the cluster manifest file will be overridden by the command line options.

Syntax:

```
vsm-cli <cluster_name> create[ -f <filesystem>][ -c <caddr>]
        [ -p <paddr>] <node_1> <node_2> ...
```

Parameters:

<cluster_name>	The name of the storage cluster. There are no restrictions on the length of the name and the used characters.
-f <file_system>	The file system used for the OSDs. Possible values are: xfs (default), btrfs, ext4. The specified file system will override the setting in the cluster manifest file **/etc/manifest/cluster.manifest**. **Default:** The file system specified in **/etc/manifest/cluster.manifest**
-c <caddr>	The subnet IP address of the Ceph cluster network. The specified IP address will override the corresponding setting in the cluster manifest file **/etc/manifest/cluster.manifest**. **Default:** The IP address of the Ceph cluster network specified in **/etc/manifest/cluster.manifest.** "Cluster Manifest File" on page 234

`-p <paddr>`

The subnet IP address of the Ceph public network.

The specified IP address will override the corresponding setting in the cluster manifest file **`/etc/manifest/cluster.manifest`**.

Default:

The subnet IP address of the Ceph public network specified in **`/etc/manifest/cluster.manifest.`**

`<node_1> <note_2> ...`

The host name(s) of all node(s) that will be used to create the storage cluster.

The disks to be used as OSDs will be enumerated in the server manifest file **`/etc/manifest/server.manifest`** for each node.

Output:

None

create-cache-tier - Converting Pools into Tiered Pools

The **create-cache-tier** command converts two existing pools into a tiered pool.

Before the **create-cache-tier** command can be executed, you must create two pools:

- a storage pool that acts as backend and will be used as the destination for data
- a cache pool that acts as frontend.

The cache_pool will manage objects based on the specified cache mode. Two cache modes are supported: **writeback** and **readonly**.

writeback mode acts as a hierarchical storage tier where all new data will first be written to the cache pool and migrated to the storage pool over time. Data in **writeback** mode will reside in either of the two pools.

readonly mode acts as a cache where all new data will be written to the storage pool and frequently read data will be copied to the cache pool. All data resides in the storage pool with the most frequently accessed data also residing in the cache pool.

Syntax:

```
vsm_cli <cluster_name> create-cache-tier -m <writeback|readonly>
        -s <storage_pool> -c <cache_pool> [-p <key_1>=<value_1>,…
        ]
```

Parameters:

<cluster_name>	The name of an existing storage cluster.

`-m <cache_mode>` Specifies the cache mode of the tiered pool. Possible values are:

`writeback`

Ceph clients write data to the cache tier and receive an acknowledgment (ACK) from it. In time, the data migrates to the storage tier and gets flushed from the cache tier.

When a Ceph client needs data from the storage tier, the cache tiering agent migrates the data to the cache tier on read, then it is sent to the Ceph client. Thereafter, the Ceph client can perform I/O using the cache tier.

Ideal for mutable data.

`readonly`

Ceph clients write data to the storage tier. On read, Ceph copies the requested object(s) from the storage tier to the cache tier. Objects get removed from the cache tier based on the defined policy. This approach is ideal for immutable data, because reading data from a cache pool that might contain out-of-date data provides weak consistency. Do not use read-only mode for mutable data.

`-s <storage_pool>` The name of the pool that shall act as storage tier pool (backend storage). You can specify any replicated or erasure coded pool that has been created in the storage cluster.

`-c <cache_pool>` The name of the pool that shall act as cache tier pool (frontend storage). You can enter the name of any pool that has been created in the storage cluster.

`-p <key_1>=<value_1>,<key_2>=<value_2>,…`

The specified values override the default parameters **<key_1>**, **<key_2>**... specified in the cluster manifest file **/etc/manifest/cluster.manifest**.

To change these settings after the cache pool creation you must remove this cache pool and create a new one with the different values defined in the **create-cache-tier** command.

Output:

None

create-ec-pool - Creating an Erasure Coded Pool

The **create-ec-pool** command creates a pool that uses erasure coding instead of replication.

Syntax:

```
vsm_cli <cluster_name> create-ec-pool -p <ec_profile>
        -r <ruleset_root> -f <ruleset_failure_domain>
        [ -q <pool_quota>] <pool_name>
```

Parameters:

`<cluster_name>` The name of an existing storage cluster.

`-p <ec_profile>` The Erasure Code profile as indicated in the cluster manifest file **/etc/manifest/cluster.manifest**.

`-r <ruleset_root>` Storage group that will be used to create the new ruleset for the pool.

Example: `-r capacity`

The storage groups are defined in the cluster.manifest file **/etc/manifest/cluster.manifest**."Cluster Manifest File" on page 234

`-f <ruleset_failure_domain>`

The domain in which only one (never more than one) segment of an erasure coded object is allowed to be placed.

Possible values are:

`host (default)`
Each segment of an erasure coded object will be placed on a different storage node.

`osd`
If OSD is specified than the erasure coded data are placed across OSDs.

With this setting you may run into situation where you can loose data. It may happen that the number chunks is greater than M are written to one host. In this case the failure of the host will cause data loss. prepare_disk

Example: `-f host`

The failure domain must be equal to or larger than K+M where K is the number of data chunks in which the original objects are divided and M is the number of the coding chunks.

-q <pool_quota>	The pool quota in bytes. You can specify the size <pool_quota> by using one of the following units: K for KB (Kilobyte) M for MB (Megabyte) G for GB (Gigabyte) T for TB(Terabyte) P for PB (Petabyte)
<pool-name>	Name of the new pool to be created.

Output:

None

create-pool - Creating a Replicated Storage Pool

The **`create-pool`** command creates a new replicated pool in an existing storage cluster.

Among others you can specify the number of copies from data objects and the number of placement groups.

In the storage pools created with the **`create-pool`** command primary copies and non-primary replicas of the data will be placed in the same storage group.

When creating a replicated pool with the functions of the VSM-GUI you can optionally place primary copies and non-primary replicas of data in separate storage groups.

Syntax:

```
vsm_cli <cluster> create-pool -d <disk_type> [-r replication]
 [-n pgnum] [-q pool-quota] [-o csv|json|plain|xml] <poolname>
```

Parameters:

`<cluster_name>` The name of an existing storage cluster.

`-d <disk_type>`

Type disk: high_performance, performance, capacity, value_ performance, or value_capacity

`-r <replication>` The number of replications.

For example: If you set the number of *replication*s to 3 (each object has 3 replicas) there will be one primary copy and two further copies of the object.

Maximum value of *replication:*

The number of nodes in the storage cluster.

Default: 2

`-n <pgnum>` Number of placement groups in the storage pool.

Default:

The number is derived from the calculation:

`[pg_count_factor]*number of OSDs in storage group) / <replication>`

`[pg_count_factor]` is set in the cluster.manifest file (default: 100).

`-q <pool_quota>`	The pool quota in bytes.
	You can specify the size `<pool_quota>` by using one of the following units:
	K for KB (Kilobyte)
	M for MB (Megabyte)
	G for GB (Gigabyte)
	T for TB(Terabyte)
	P for PB (Petabyte)
	For example:
	`vsm_cli CLUSTER_1 create-pool -q 2.5T POOL_1`
	is equivalent to
	`vsm_cli CLUSTER_1 create-pool -q 2748779069440 POOL_1`
	(2,5 TB = 2748779069440bytes).
	`o csv|json|plain|xml`
	output message in different formats.

`<pool-name>`	Name of the new pool to be created.

Output:

None

crush-reweight - Changing the Crush Weight of an OSD

With the **`crush-reweight`** command you can adjust the CRUSH weight of an OSD up and down.

The current weight of an OSD can be ascertained from the **`list osd-tree`** command ("OSD Tree" on page 162).

Syntax:

```
vsm_cli <cluster_name> crush-reweight <node_name>
        {<dev>|osd.<number>} <weight>
```

Parameters:

`<cluster_name>`	The name of an existing storage cluster.
`<node_name>`	The host name of the storage node where the OSD to be reweighted resides.
`<dev>|<osd_number>`	To specifiy the disk/OSD to be reweighted, enter the device name `<dev>` in by-path notation (e.g. pci-0000:01:00.0-scsi-0:2:2:0) or the OSD number `osd.<number>`.. To figure out the OSD number or the device name in by-path notation you can use the following **`vsm_cli`** command: `vsm_cli <cluster_name> list disks` In the output data: • `OSD No.` is equal to `<number>`, i.e. 01 specifies the OSD osd.1 (see page 159) • the device name in by-path notation is displayed in the `Device name` column
`<weight>`	The CRUSH weight that you want to assign to the disk specified in `<dev>`/`osd.<number>`. Disks with a larger capacity should have a greater weight. Default CRUSH weight: 1.0.

Output:

none

Example:

The current CRUSH weight of the disk osd.4 is 1. It shall be reduced by 10% . osd.4 resides on the storage node node002. The cluster name is CLUSTER_1.

- Specifying disk by osd.<number>:

```
vsm_cli CLUSTER_1 crush-reweight node002 osd.4 0.9
```

- Specifying disk by device path:

```
vsm_cli CLUSTER_1 crush-reweight node002 pci-0000:01:00.0-scsi-0:2:2:0 0.9
```

delete-pool - Deleting a Pool

The **`delete-pool`** command deletes a pool from a storage cluster.

When you delete a pool from a storage cluster all data stored on this pool will be lost!

Syntax:

`vsm_cli <cluster_name> delete-pool <pool_name>`

Parameters:

`<cluster_name>`	The name of an existing storage cluster.
`<pool_name>`	The name of the storage pool to be deleted..

Output:

```
Do you really want to delete pool <pool_name>? [y/n]
```

- If you enter `n` (no), the command is aborted.
- If you enter `y`, you will be asked :

  ```
  Type the pool name you want to delete again?
  ```

- Reenter the name of the storage pool to be deleted. This shall avoid the accidental deletion of data.

If the entered name matches input for `<pool_name>`, the pool will be deleted. Otherwise, the command is aborted.

maint-on/off - Switching Maintenance Mode on/off

The **maint-on** command places the specified storage node in maintenance mode i.e. on this node all Ceph daemons will be stopped.

While a storage node is in maintenance mode, all rebalancing for the storage cluster will be suppressed. Therefore, all OSDs on the specified storage node will be in status "down", but they will not change to the status "out" (unless they were "out" prior to the command.)

The **maint-off** command restarts all daemons and enables cluster rebalancing again.

Syntax:

```
vsm_cli <cluster_name> maint-on <node_name>
vsm_cli <cluster_name> maint-off <node_name>
```

Parameters:

<cluster_name>	Host name of an existing storage cluster.
<node_name>	Host name of the storage node where the maintenance mode shall be switched on/off.

Output:

None

help - Displaying the Syntax of vsm_cli Commands

The **help** command displays a list of all **vsm_cli** commands with their syntax. If you enter **help** together with the name of a **vsm_cli** command, then help displays detailed information on that command.

Syntax:

vsm_cli help [<vsm_cli-command>]

Parameters:

<vsm_cli-command>	The name of a **vsm_cli** command. Possible values are: crush-reweight, delete-pool, destroy, list, maint-on, maint-off, mon-status, replace-disk-in, replace-disk-out, set-pool, start, status, stop

Output:

If you enter help without any parameter, then general help is displayed:

```
vsm_cli help
Usage:
```

vsm_cli help vsm_cli help <subcommand>

```
 vsm_cli <cluster> activate [-f filesystem] [-c caddr] [-p paddr] <node1> <node2> <nod
e_n>
 vsm_cli <cluster> add-disk <node> <deviceType#device_name#journal_device_name>
 vsm_cli <cluster> add-mon <node>
 vsm_cli <cluster> create [-f filesystem] [-c caddr] [-p paddr] <node1> <node2> […] <no
de_n>
 vsm_cli <cluster> create-cache-tier -m <writeback|readonly> -s <storage_pool> -c <cach
e_pool> [-p <key_1>=<value_1>,...]
 vsm_cli <cluster> create-ec-pool -p <ec_profile> -r <ruleset_root> -f <ruleset_failur
e_domain> [-q <number>] <erasure_pool>
 vsm_cli <cluster> create-pool -d <storage_group> [-r replication] [-n pgnum] <poolnam
e>
 vsm_cli <cluster> crush-reweight <node dev|osd.number> <weight>
 vsm_cli <cluster> delete-pool <poolname>
 vsm_cli <cluster> destroy
 vsm_cli <cluster> list all|disks [sort] [-r]|disks-of-nodes <node_1> [node_2] ...|node
s [node_1] [node_2] ...|osd-tree|pools
 vsm_cli <cluster> maint-on nodename
 vsm_cli <cluster> maint-off nodename
```

```
vsm_cli <cluster> mon-status
vsm_cli <cluster> remove-cache-tier <cache_pool>

vsm_cli <cluster> remove-mon <node>

vsm_cli <cluster> replace-disk-out|replace-disk-in <node> <dev>
vsm_cli <cluster> set-pool [-r replica] [-q max_bytes] <pool>
vsm_cli <cluster> start [<osd.number|mon.number|mds.number>]
vsm_cli <cluster> stop [<osd.number|mon.number|mds.number>]
vsm_cli <cluster> status [ timeout ]
vsm_cli list-cluster-names
```

If you enter help together with the name of a **vsm_cli** command, then detailed help for that **vsm_cli** command is displayed. For example:

```
# vsm_cli help start
vsm_cli <cluster> start [<mds.number|mon.number|osd.number>]

       <cluster> = name of cluster
       If no daemon specified, start ceph processes on all nodes
```

list - Requesting Information about a Storage Cluster and the Utilization

The **list** command allows the monitoring of:

- the health, the status and the capacity utilization of a specified storage cluster
- the status of the OSDs (disks) of one storage node or the complete storage cluster
- the status of the pools in the specified storage cluster

Syntax:

```
vsm_cli <cluster_name> list
        all|
        disks [<sort>] [-r]|
        disks-of-nodes <node_1> [<node_2>]...[<node_n>]|
        nodes [<node_1>] [<node_2>]...[<node_n>]|
        osd-tree|
        pools
```

Parameters:

`<cluster_name>` The name of an existing storage cluster.

`all` Provides information on the health, monitors, all storage nodes, disks and pools of the specified storage cluster.

`disks [<sort>] [-r]`

Provides information about all OSDs of the storage cluster, their status and utilization (for an example see page 159).

`<sort>`:

The output of the list command with the disk option will be sorted by the column name specified in `<sort>`. Possible values are: `Size`, `Used Avail`, `Util`, `Type`, `node, Device name`, `Status` (for detailed information see page 159).

`-r`:

Displays information about the disks of a storage cluster in the reverse order.

`disks-of-nodes <node_1> [<node_2>]...[<node_n>]`

Provides information about all OSDs on the specified nodes (for detailed information see page 161). For `<node_1> <node_2> ... <node_n>` enter the host names of the storage nodes.

`nodes [<node_1>] [<node_2>]...[<node_n>]`

Provides information about the storage nodes specified in <node_1>, <node_2> ... If no node name is entered the `nodes` option provides information about all storage nodes of the cluster (see page 164).

`osd-tree` Displays the OSD tree (see page 162).

`pools` Provides information on the data pools in the storage cluster (for detailed information see page 165).

Output/Example:

Information on cluster resources, configuration, and capacity utilization.

list-cluster-names - Displaying a List of all Clusters

The **`list-cluster-names`** command displays a list of all clusters which are configured. For each cluster the command displays

- the cluster's name
- the file system which is used for the OSDs of each cluster
- the subnet addresses of the associated networks

Syntax:

`vsm_cli list-cluster-names`

Output (Example):

Per cluster the following information will be displayed:

```
Cluster name: cd10000
```

mon-status - Displaying Monitor Status of a Storage Cluster

You can use the **mon-status** command to monitor the status of the monitor daemons.

The output of the **mon-status** command informs about the quorum ranking and the IP addresses of the hosts, on which the monitors reside.

Syntax:

vsm_cli <cluster_name> mon-status

Parameter:

<cluster_name>	The name of an existing storage cluster.

Messages:

The **mon-status** command returns a message:

```
FCA1026:N:MON_STATUS#f_arg2
```

Example:

The command

```
vsm_cli CLUSTER_1 mon-status
```

displays the following information:

```
FCA1026:N:MON_STATUS#monitor_quorum_rank_ids=[0,1,2] 0:192.168.0.104,
1:192.168.0.107, 2:192.168.0.109
```

prepare-disk - Preparing disks

The **prepare-disk** command prepares disks on the specified nodes that shall be used to create a new cluster using the **activate** command.

Syntax:

```
vsm_cli <cluster_name> prepare-disk[ -f filesystem]
        [-t timeout] [-b num_of_bg_process] <node_1> <node_2> … <
        node_n>
```

Parameters:

<cluster_name>	The name of an existing storage cluster.
-f <filesystem>	The file system used for the OSDs. Possible values are: xfs (default), btrfs, ext4. The specified file system will override the setting in the cluster manifest file **/etc/manifest/cluster.manifest**. **Default:** The file system specified in **/etc/manifest/cluster.manifest**
-t <timeout>	The time, in minutes, the CLI will wait before abandoning the command due to unresponsive disks. **Default:** 10
-b <num_og_bg_processes>	The number of disks that can be prepared in parallel. Increasing this number can improve completion time. **Default:** 1
<node_1> <node2> … <node_n>	The name(s) of the node(s) on which the disks shall be prepared. The disks to be used as OSDs will be enumerated in the **/etc/manifest/server.manifest** for each node.

Output:

None

remove-cache-tier - Remove cache tiering

The **remove-cache-tier** command removes cache tiering associated with the specified cache pool.

Any data in the cache pool that is not already in its associated storage pool will be flushed. Once the cache tiering has been removed the storage pool can continue to be used as a non-cached pool.

The cache pool will not be deleted.

Syntax:

```
vsm_cli <cluster_name> remove-cache-tier <cache_pool>
```

Parameters:

<cluster_name>	The name of an existing storage cluster.
<cache_pool>	The name of the frontend pool to be removed.

Output:

None

remove-mon - Removing a monitor daemon from a storage cluster

This command stops a monitor daemon on the specified node and removes the daemon from the storage cluster.

The monitor daemons of a storage cluster must be able to establish a consensus about the cluster status. If a consensus is not possible you should add one new monitor daemon before removing another one.

CD10000 storage cluster must have three monitor daemons at least.

Syntax:

```
vsm-cli <cluster_name> remove-mon <node>
```

Parameters:

`<cluster_name>`	The name of an existing storage cluster.
`<node>`	The host name of the storage cluster node on which the monitor daemon resides.

Output:

None

replace-disk-out/in - Replacing a Disk of a Storage Cluster

The **replace-disk-out** and the **replace-disk-in** command allow you to substitute a new disk for a failed/retired disk keeping the overall storage cluster configuration intact.

Requirements:

- The disk type of the new and the old disk must be the same. The disks can have different capacities. But the new disk should have at least as much capacity as the replaced disk.
- For safety reasons, the **replace-disk** commands should only be executed when the storage cluster status is HEALTHY .

  If the disk to be replaced is in the “out” state,you should wait for any rebalancing to finish before calling the **replace-disk-out** and the **replace-disk-in** commands.

To replace a disk, call the **replace-disk-out** command. It stops the OSD daemon and unmounts the file system for the specified disk. Wait until the OSD is "down".

Replace the disk physically and after that call the **replace-disk-in** command. **replace-disk-in** will create a new logical volume, format and mount the file system, restart the OSD daemon and re-enable cluster rebalancing. Ceph will automatically copy the data objects (replicas) that were stored on the the old disk to the new disk.

Upon successful replacement, the storage cluster configuration will not have changed and the object distribution will generally be the same as before.

Syntax:

```
vsm_cli <cluster_name> replace-disk-out <node_name> <dev_by-path>

vsm_cli <cluster_name> replace-disk-in <node_name> <dev_by-path>
```

Parameters:

`<cluster_name>`	The name of an existing storage cluster.
`<node_name>`	Host name of the storage node.
`<dev_by-path>`	Disk to be replaced. You must specify the disk by its disk path in by-path notation. To figure out the disk path you can use the **vsm_cli** command **list** `disks` or `list disks-of-nodes`. The path of a disk is displayed in the column `Device name`.

Output:

None

Example:

You will find a detailed example in "Managing Disks/OSDs (vsm_cli)" on page 125.

start - (Re)Starting a Cluster Daemon / a Storage Cluster

The **`start`** command starts the specified OSD, MDS or a monitor daemon if it was stopped before. If no daemon is specified (you enter the `start` command without parameter), the storage cluster will be (re)started (i.e. all cluster daemons are started).

With the VSM GUI you can stop all daemons residing on particular storage nodes (function **Server Management** > **Manage Server**, button **Start servers**).

For detailed information see "Starting storage nodes" on page 115

Syntax:

`vsm_cli <cluster_name> start [osd.<number> | mon.<number>]`

Parameters:

`<cluster_name>`	The name of an existing storage cluster.
`osd.<number>`	OSD that shall be restarted. `<number>`: When creating a storage cluster VSM assigns a number to each OSD. The number reflects the order in which the OSDs are added to the storage cluster. To figure out the number of an OSD you can use for example the following **`vsm_cli`** command: `vsm_cli <cluster_name> list disks` In the output data `OSD No.` is equal to the OSD number, i.e. 01 specifies the OSD osd.1 (see page 159).
`mon.<number>`	Monitor daemon that shall be restarted. `<number>`: When creating a new cluster or adding new monitor daemons to the cluster VSM assigns a number to each monitor. To figure out the number of a monitor you can use the **`vsm_cli`** command: `vsm_cli <cluster_name> mon-status` (see page 158).

Output:

None

Example:

1. Start the monitor daemon number 1 of the storage cluster CLUSTER_1:

   ```
   vsm_cli CLUSTER_1 start mon.1
   ```

2. Start the monitor daemon that resides on the host with the IP address 192.168.0.104.

   First enter the **mon-status** command to verify the number of the monitor :

   ```
   vsm_cli CLUSTER_1 mon-status
   ```

   This command displays:

   ```
   FCA1026:N:MON_STATUS#monitor_quorum_rank_ids=[0,1,2] 0:192.168.0.104,
   1:192.168.0.107, 2:192.168.0.109
   ```

   "0:192.168.0.104" in the output means that the monitor daemon on the host 192.168.0.104 has the number 0.

   Enter the following command to start this monitor daemon:

   ```
   vsm_cli CLUSTER_1 start mon.0
   ```

status - Information about the Cluster Status

The **vsm_cli status** command displays the health status of the storage cluster and a brief summary of the OSD status and storage utilization.

Syntax:

vsm_cli <cluster_name> status [<timeout>]

Parameters:

<cluster_name>	The name of an existing storage cluster.
<timeout>	Specifies the time in seconds the **status** command shall collect the significant data for reporting the storage cluster status. Default: 10 seconds

Output/Example:

The command

```
vsm_cli CLUSTER_1 status
```

displays:

```
FCA1022:N:STATUS#health:HEALTH_OK #osd:12 osds,12 up,12 in#Storage utilization:0.013%
```

For detailed information on the output see section "Summary of the Cluster Status " on page 157

set-pool - Modify pool parameters

The **set-pool** command allows you to modify the number of data copies and the allocated quota of storage size.

Syntax:

```
vsm_cli <cluster_name> set-pool [-r <replication>]
        [-q <max_bytes>] <pool_name>
```

Parameters:

<cluster_name>	The name of an existing storage cluster.
-r <replication>	The number of replications. For example: If you set the number of replications to 3 (i.e.each object has three replicas) there will be one primary copy and two further copies of the object. Minimum value of <replication>: 2 Maximum value of <replication>: The number of nodes in the storage cluster.
-q <max_bytes>	The pool quota of storage size in bytes. You can specify the size <max_bytes> by using one of the following units: K for KB (Kilobyte) M for MB (Megabyte) G for GB (Gigabyte) T for TB(Terabyte) P for PB (Petabyte) For example: vsm_cli CLUSTER_1 set-pool -q 2.5T POOL_1 is equivalent to vsm_cli CLUSTER_1 set-pool -q 2748779069440 POOL_1 (2,5 TB = 2748779069440bytes).
<pool_name>	The name of the storage pool to be modified.

Output:

None

stop - Stopping a Cluster Daemon / a Storage Cluster

The **stop** command stops the specified OSD, MDS or a monitor daemon. If no daemon is specified (you enter the **stop** command without any parameter), the storage cluster will be stopped (i.e. all cluster daemons are stopped).

The stopped daemon or storage cluster may be restarted with the **vsm_cli start** command (see page 275).

With the VSM GUI you can stop all daemons residing on particular storage nodes (function **Server Management** > **Manage Server**, button **Stop servers**).

For detailed information see "Stopping storage nodes " on page 114.

Syntax:

```
vsm_cli <cluster_name> stop
        [osd.<number> | mon.<number>]
```

Parameters:

<cluster_name>	The name of an existing storage cluster.
osd.<number>	OSD that shall be stopped. <number>: When creating a storage cluster VSM assigns a number to each OSD. The number reflects the order in which the OSDs are added to the storage cluster. To figure out the number of an OSD you can use for example the following **vsm_cli** command: vsm_cli <cluster_name> list disks In the output data OSD No. is equal to the OSD number, i.e. 01 specifies the OSD osd.1 (see page 159).
mon.<number>	Monitor daemon that shall be stopped. <number>: When creating a new cluster or adding new monitor daemons to the cluster VSM assigns a number to each monitor. To figure out the number of a monitor you can use the **vsm_cli** command: vsm_cli <cluster_name> mon-status (see page 158).

Output:

None

Examples:

1. Stop the storage cluster (i.e. all cluster daemons) CLUSTER_1:

   ```
   vsm_cli CLUSTER_1 stop
   ```

2. Stop OSD number 4 of the storage cluster CLUSTER_1:

   ```
   vsm_cli CLUSTER_1 stop osd.4
   ```

14.2 cd10000 Commands

This chapter describes the overall usage of **cd10000** command line interface.

It contains most important components functionality to manage and to monitor RPM packages, nodes, switches, cluster and to perform diagnostic tests.

The **cd10000** command line interface collects functionality from different components.

Installed Fujitsu RPM packages can be listed.

The diagnose tests can be executed to check if specific parts of ETERNUS CD10000 are configured and work correctly. You can check following categories: System, Network, Factory and all.

Fipmanager is an application for IP and cluster management.

Storage nodes power status can be checked, turned on/off, rebooted or booted from pxe. They also can be partitioned.

Update Manager is an application for maintaining the Update/Upgrade process of a whole cluster. Updating/Upgrading one node at a time ensures stability Ceph cluster as a whole

Log Manager gathers on demand SOS reports, debug information and log files from the specified storage nodes and management node of a storage cluster.

Here is a functional overview of **cd10000** command line interface

- On the management node: there is an RPM repository available for all nodes
- Check of installed Fujitsu RPM packages on nodes
- Diagnose tests execution in order to inspect correctness of the configuration
- Listing, adding, removing, updating and setting parameters on the nodes
- Listing, removing setting and getting parameters from the switches
- Showing, setting and getting the cluster parameters
- Configuration of cluster, SNMP and redundant and non-redundant networks, initialization of the VSM
- Power management of the nodes
- Gathering of logs, SOS reports and debug information from the storage nodes
- On demand updates
- Broken update/configuration can break only on one node
- Upload updates to repository from media, e.g. DVD
- Possible configuration adjustments on all nodes

14.2.1 Runtime Model

The **cd10000** command line interface invokes the execution of the particular action in the responsible component.

Overall execution flow

- Typing particular command
- Responsible component will execute the command and return the output.

Dependencies

All **cd10000** components.

Error handling concept

Each component handles it's errors.

14.2.2 External Interface Specification

Return Values

- 0 - success
- 1 - error

Log entries, system messages

- /var/log/cd10000.log
- Particular component log file

14.2.3 Diagnose Command

cd10000 diagnose - Executing diagnose test

This command executes the diagnose tests of the set category with different output formats.

This command also prints output of the tests as text. This text contains describing status, summary and error message of executed tests.

By adding `--details` the command causes printing the traceback. Providing the command with `--format`, it gives output in set format with information about executed tests: status,summary,error and traceback.

Synopsis:

```
cd10000 diagnose [category] -d|--details -f|--format
[json,xml,csv,text]
```

Available categories:

1.
   - Factory – these tests give a basic information of the system (like kernel version or time zone). Also these tests give the information if the software installed version is correct and if the nodes are accessible in the network.
   - System – these tests give the information about the system configuration and the software installed version.
   - Network – these tests give the information about the administration network, the infiniband network and the irmc network.
   - All - these tests give the information about all above listed categories, except Factory.

Parameters:

Printed output with details:

`<category>`	The category of the tests which will be executed. This parameter is required.
`<--details>`	This parameter prints traceback if the output format is text. This parameter is optional.

Returned output in specified format:

`<category>`	The category of the tests which will be executed. This parameter is required.

<--format> This parameter returns the test results in specified format:

- xml
- json
- csv
- text(default)

This parameter is optional.

Output:

```
-----------------------------------------
NETWORK TESTS RESULTS
-----------------------------------------
[PASS]:Check if all hosts can be pinged
        in administration network from every storage node
[PASS]:Check if all storage nodes can be pinged
       in infiniband network from every storage node
[PASS]:Check if all storage nodes can be pinged
        in irmc network from mng node
[PASS]:Check if the offset of time between hosts
        is on correct range
[PASS]:Check if all storage nodes can be pinged
        in administration network from mng node
[PASS]:Check if all storage nodes can be pinged
        in public network from mng node
```

Example:

```
#cd10000 diagnose system -d
#cd10000 diagnose all -f xml
```

14.2.4 Commands of the Log Manager

The log manager provides the following commands for handling the log management:

Command	Description
cd10000 support gather_debug_info	On demand gathering debug information from the specified storage nodes and management node of a storage cluster. See "cd10000 support gather_debug_info - Gathering debug information" on page 286.
cd10000 support gather_logs	On demand gathering log files from the specified storage nodes and management node of a storage cluster. See "cd10000 support gather_logs - Gathering log Files " on page 287.
cd10000 support gather_sosreports	On demand gathering SOS reports from the specified storage nodes and management node of a storage cluster. See "cd10000 support gather_sosreports - Gathering SOS Reports" on page 288.
export	Exporting an index from the Elasticsearch database. See "export - Exporting an Index from the Database" on page 289.
import	Importing an index to the Elasticsearch database. See "import - Importing an Index to the Database" on page 290

cd10000 support gather_debug_info - Gathering debug information

This command gathers debug information from logs and the output of commands from the specified storage nodes and the management node.

You can access the **`cd10000 support gather_debug_info`** command as a user with root privileges.

Syntax:

```
cd10000 support gather_debug_info --output|-o path |
        [--nodes|-n host1 host2]
```

Parameters:

`output`	Path to the out directory where the gathered debug information will be stored. Content will be cleared.
`nodes`	Optional. List of the storage nodes from which debug information is to be gathered. If `nodes` is not specified, the debug information of all storage nodes and the management node is collected.

Example:

Gather the debug information from the storage nodes `storage1`, `storage3` and the management node `pmgmt` and store it in `/tmp/debug/`.

```
cd10000 support gather_debug_info -o /tmp/debug -n storage1 storage3 pmgmt
```

Output (tab format)

```
#cd10000 support gather_debug_info -o /tmp/debug -n storage1 storage3 pmgmt
Reading the xml configuration file:
/opt/fujitsu/ftools/etc/debug.xml
Copying from: storage1: /etc/vsm/
Copying from: storage1: /var/log/messages
Copying from: storage1: /var/log/puppet/
Copying from: storage1: /var/log/vsm/
...
Compressing the output. It can take some time...
Archive: /tmp/debug/gather-141117-043452.tar.gz
```

cd10000 support gather_logs - Gathering log Files

This command gathers the log files from the specified storage nodes and the management node of a storage cluster.

You can access the **`cd10000 support gather_logs`** command as a user with root privileges.

Syntax:

```
cd10000 support gather_logs --files|-f path1 path2 |
        --output|-o path |
        [--nodes|-n host1 host2]
```

Parameters:

`files`	List of the absolute paths to the log files on the storage nodes.
`output`	Path to the out directory where the gathered log files will be stored. Content will be cleared.
`nodes`	Optional. List of the storage nodes from which log files are to be gathered. If `nodes` is not specified, the log files of all storage nodes are collected.

Example:

Gather the log files from the `/var/log/puppet` directories on the storage nodes `storage13` and `storage14` and store them in `/tmp/logs/`.

```
cd10000 support gather_logs -f /var/log/puppet -o /tmp/logs -n storage13 st
orage14
```

Output (tab format)

```
# cd10000 support gather_logs -f /var/log/puppet -o /tmp/logs -n storage13
storage14
Gathering files from: storage13
Copying: /var/log/puppet
Gathering files from: storage14
Copying: /var/log/puppet
Archive: /tmp/logs/gather-140414-083837.tar.gz
```

cd10000 support gather_sosreports - Gathering SOS Reports

This command gathers SOS reports from the specified storage nodes and the management node.

You can access the **cd10000 support gather_sosreports** command as a user with root privileges.

Syntax:

```
cd10000 support gather_sosreports --output|-o path |
        [--nodes|-n host1 host2] [--verbose|-v]
```

Parameters:

output	Path to the out directory where the SOS reports will be stored. Content will be cleared.
nodes	Optional. List of the storage nodes from which SOS reports are to be gathered. If nodes is not specified, the SOS reports of all storage nodes and the management node are collected.
verbose	Optional. Increase the verbosity on the screen.

Example:

Gather the SOS reports from the storage nodes storage1, storage3 and the management node pmgmt and store them in /tmp/debug/.

```
cd10000 support gather_sosreports -o /tmp/sosreports -n storage1 storage3 p
mgmt
```

Output (tab format)

```
# cd10000 support gather_sosreports -o /tmp/sosreports -n storage1 storage3
pmgmt
Gathering from: storage1
Copying from: storage1: /tmp/sosreport-CD10000*/
Gathering from: storage3
Copying from: storage3: /tmp/sosreport-CD10000*/
Gathering from: pmgmt
Copying from: pmgmt: /tmp/sosreport-CD10000*/
Compressing the output. It can take some time...
Archive: /tmp/sosreports/gather-141117-052442.tar.gz
```

export - Exporting an Index from the Database

This command exports an index from the Elasticsearch database to a file.

Exporting an index to a file is needed if you have to provide the log files to Fujitsu Service for analyzing. An exported index can be imported in another Elasticsearch database easily.

The output format is a zipped tar file.

You can access the `export` command as a user with root privileges.

Syntax:

```
curl -XPOST pmgmt:9200/index_name/_export?path=output_path
```

Parameters:

`index_name`	Name of Elasticsearch index to be exported.
`output_path`	Path to the output file. The output file must have a `.tar.gz` extension. The Elasticsearch user must have write permissions for this path.

Example:

1. Export the `logs-2014.11.01` index from the Elasticsearch database to `/tmp/elasticsearch/logs.tar.gz`.

   ```
   curl -XPOST pmgmt:9200/logs-2014.11.01/_export?path=/tmp/elasticsearch/logs.tar.gz
   ```

   Output (tab format)

   ```
   {"running":true,"mode":"export","archive":"tar","path":"file:/tmp/elasticsearch/logs.tar.gz"}
   ```

2. * You can get the list of indices using the command:

   ```
   curl pmgmt:9200/_aliases
   ```

   Output (tab format)

   ```
   {"logs-2014.11.01":{"aliases":{}},"logs-2014.10.31":{"aliases":{}}}
   ```

   In this case the indices are: `2014.11.01` and `2014.10.31`

import - Importing an Index to the Database

This command imports a file into the Elasitcsearch database which was exported before from another database.

You can access the `import` command as a user with root privileges.

Syntax:

```
curl -XPOST pmgmt:9200/index_name/_import?path=input_path
```

Parameters:

`index_name`	Name of the Elasticsearch index to be imported.
`input_path`	Path to the exported index. The input file must have a `.tar.gz` extension with read permissions for Elasticsearch user. The index must not exist in the database.

Example:

Import the `logs-2014.11.01` index to the Elasticsearch database to `/tmp/elasticsearch/logs.tar.gz`.

```
curl -XPOST pmgmt:9200/logs-2014.11.01/_import?path=/tmp/elasticsearch/logs
.tar.gz
```

Output (tab format)

```
{"running":true,"mode":"import","type":"tar","path":"file:/tmp/elasticsearc
h/logs.tar.gz"}
```

14.2.5 Cluster Setup

The cluster setup provides the following commands:

Command	Description
cd10000 setup storage_root_password	Setting the root password for the storage nodes. See "cd10000 ip node update - Updating information about the storage node" on page 297
cd10000 setup snmp set-community	See "set-community - Defining a Community" on page 319
cd10000 setup snmp remove-community	See "remove-community - Removing a Community" on page 320
cd10000 setup snmp list-community	See "list-community - Displaying all Communities" on page 321
cd10000 setup snmp set-destination	See "set-destination - Defining a Trap Destination" on page 322
cd10000 setup snmp remove-destination	See "remove-destination - Removing a Trap Destination" on page 323
cd10000 setup snmp list-destination	See "list-destination - Displaying all Trap Destinations" on page 324
Initializing VSM GUI	See "Preparing the Cluster Creation" on page 53
cd10000 setup authentication ldap	Setting up LDAP authentication configuration. See "cd10000 nodes power off - Turning the power of node off" on page 310

cd10000 setup storage_root_password - Setting the root password

This command sets the root password for the storage nodes.

Synopsis:

cd10000 setup storage_root_password -p|--password [PASSWORD]

Parameters:

<--password>	The new root password for the storage node. This parameter is required.

Output:

```
------------------------------------------------------------
             Initializing Management Node
------------------------------------------------------------
2014-11-01 16:22:57,739 -INFO - Initializing Pulp repository...
2014-11-01 16:22:58,143 -INFO - Preparing Pulp Database...
Beginning database migrations.
Migration package pulp.server.db.migrations is up to date at version 6
Migration package pulp_rpm.migrations is up to date at version 14
Database migrations complete.
Loading content types.
Content types loaded.
```

Example:

```
cd10000 setup storage_root_password -p cd10000
```

Defining the community

See chapter "set-community - Defining a Community" on page 319

Removing the community

See chapter "remove-community - Removing a Community" on page 320

Listing all communities

See chapter"list-community - Displaying all Communities" on page 321

Defining the destination

See chapter "set-destination - Defining a Trap Destination" on page 322

Removing the destination

See chapter "remove-destination - Removing a Trap Destination" on page 323

Listing all destinations

See chapter "list-destination - Displaying all Trap Destinations" on page 324

Initializing VSM GUI

See chapter "Preparing the Cluster Creation" on page 53

cd10000 setup authentication ldap - Setting up LDAP authentication configuration

This command sets up the LDAP authentication configuration.

Synopsis:

```
cd10000 setup authentication ldap -s|
--ldap-server LDAP_SERVER -b|--ldap-basedn LDAP_BASEDN
```

Parameters:

<--ldap-server>	This parameter shows the LDAP server hostname or the IP. This parameter is required.
<--ldap-basedn>	LDAP own suffix. This parameter is required.

Output:

None

Example:

```
cd10000 setup authentication -s fji.server.world -b "dc=server,dc=world"
```

14.3 Nodes Management Command Line Interface

14.3.1 Fipmanager Commands

The Fipmanager provides the following commands for handling and managing monitor nodes, switches and cluster.

Command	Description
cd10000 ip node list	Listing all storage nodes with information about them. See "cd10000 ip node list - Listing storage nodes" on page 296.
cd10000 ip node update	Updating the information about the specified storage node to the database. See "cd10000 ip node update - Updating information about the storage node" on page 297
cd10000 ip node rm	Removing the storage node from the database. See "cd10000 ip node rm - Removing the storage node " on page 298
cd10000 ip node set	Setting the node's parameter and saving it in the database. See "cd10000 ip node set - Setting node's parameter" on page 299
cd10000 ip cluster show	Displaying the cluster and the management node serial number, the administration, the iRMC, the cluster and the switch networks. See "cd10000 ip cluster show - Showing the cluster network configuration" on page 300
cd10000 ip cluster get	Getting the serial number of the cluster. See "cd10000 ip cluster get - Getting the serial number of the cluster" on page 301
cd10000 ip cluster set	Setting the cluster's parameters. See "cd10000 ip cluster set - Setting the cluster's parameters" on page 302
cd10000 ip switch list	Showing the list of the switches with their parameters. See "cd10000 ip switch list - Showing the list of the switches" on page 303

cd10000 ip node list - Listing storage nodes

This command lists all storage nodes with information about them.

More information can be displayed if the `-detailed [1,2]` argument is chosen. The output format can be specified,otherwise the output format will be shown as text.

Synopsis:

cd10000 ip node list -d|--detailed [1,2] -f|--format [json,xml,csv,text]

Parameters:

<--detailed>	This parameter prints more detailed information about the nodes. The choices are {1,2}, and each of them displays more detailed information about the storage nodes. This parameter is optional.
<--format>	This parameter specifies output format, it can be one of the following: {json,xml,csv,text}. The default output format is text. This parameter is optional.

Output:

```
+-------------+------------+-------------+---------------+------------+
| Status      | Server SN  | Hostname    | IP adm net    | JX SN      |
+-------------+------------+-------------+---------------+------------+
| DISCONECTED | YLNT007081 |  storage1   | 192.168.20.11 | YL4N002372 |
| DISCONECTED | YLNT007082 |  storage2   | 192.168.20.12 | YL4N002373 |
| DISCONECTED | YLNT007080 |  storage3   | 192.168.20.13 | YL4N002371 |
+-------------+------------+-------------+---------------+------------+
```

Example:

```
#cd10000 ip node list
#cd10000 ip node list -d1
```

cd10000 ip node update - Updating information about the storage node

This command updates the information about the specified storage node to the database.

Synopsis:

cd10000 ip node update HOSTNAME -f

Parameters:

<HOSTNAME>	This parameter sets the hostname, which will be updated. This parameter is required.
<-f>	This parameter performs command without any prompt. This parameter is optional.

Output:

```
Updating host: storage2 ...
Retrieving facts ...
Storage node exist in data base !!!
Storage node updated successfully
```

Example:

```
#cd10000 ip node update storage2 –f
```

cd10000 ip node rm - Removing the storage node

This command removes the storage node from the database.

Synopsis:

cd10000 ip node rm NODE_SN -f

Parameters:

<NODE_SN>	The serial number of the storage node, which will be removed. This parameter is required.
<-f>	This parameter performs command without any prompt. This parameter is optional.

Output:

```
+------------+------------+-----------+-----------------+-------+
| Status     | Server SN  | Hostname  | IP adm net      | JX SN |
+------------+------------+-----------+-----+-----------+-------+
| CONFIGURED | YLNT006178 | storage4  |  192.168.20.14  | N/A   |
+------------+------------+-----------+-----------------+-------+
```

Example:

```
#cd10000 ip node rm storage2 -f
```

cd10000 ip node set - Setting node's parameter

This command sets the node's parameter and saves it in the database.

Synopsis:

cd10000 ip node set [PARAM] NODE_SN NEW_VALUE

Parameters:

<PARAM>	These parameters will be set on the storage node: • jxsn - JX serial number • physical_location - physical location • ip – IP address of the storage node in the management network These parameters are required.
<NODE_SN>	The serial number of the node, which parameters will be set.This parameter is required.
<NEW_VALUE>	This parameter sets the new value of the given parameter. This parameter is required.

Output:

```
JX serial number: TEST123 added to server: YLNT006200
```

Example:

```
# cd10000 ip node set jxsn YLNT006200 TEST123
```

cd10000 ip cluster show - Showing the cluster network configuration

This command displays the cluster and the management node serial number, the administration, the iRMC, the cluster and the switch networks.

Synopsis:

```
cd10000 ip cluster show -f|--format [json,xml,csv,text]
```

Parameters:

`<--format>` This parameter specifies the output format:

- xml
- json
- csv
- text(default)

This parameter is optional.

Output:

None

Example:

```
cd10000 ip cluster show
```

cd10000 ip cluster get - Getting the serial number of the cluster

This command returns the management node serial number or the cluster serial number.

Synopsis:

cd10000 ip cluster get [cluster_sn,mng_node_sn]

Parameters:

None

Output:

YLNT004585

Example:

```
cd10000 ip cluster get cluster_sn
cd10000 ip cluster get mng_node_sn
```

cd10000 ip cluster set - Setting the cluster's parameters

This command sets the parameters of the cluster. Following parameter are available:

- administration network
- iRMC network
- ceph private network
- switch network
- serial number of the cluster

Synopsis:

```
cd10000 ip cluster set PARAM NEW_VALUE
```

Parameters:

<PARAM>	Following cluster parameters will be changed: • ceph_priv_net • sn • irmc_net • switch_network • adm_net

Output:

```
New cluster serial number:TEST123
```

Example:

```
cd10000 ip cluster set sn TEST123
```

cd10000 ip switch list - Showing the list of the switches

This command displays the list of all switches with their parameters

The output format can be specified to: xml, json, csv or text (default).

Synopsis:

```
cd10000 ip switch list
```

Parameters:

`<--format>` This parameter specifies the output format:

- json
- xml
- csv
- text (default)

This parameter is optional.

Output:

```
+------------------+----------+--------------------------------------+
|     IP address   |    SN's  | Switch type                          |
+------------------+----------+--------------------------------------+
| 192.168.50.11    | 123456   |   Internal administration switch     |
+------------------+----------+--------------------------------------+
```

Example:

```
cd10000 ip switch list -f json
```

14.3.2 RPM Packages

cd10000 rpms –n - Displaying installed RPMs

This command displays the information about the installed RPMs on the specified node (default is the management node).

The output can be specified to one of: json, xml, csv, text (default).

Synopsis:

cd10000 rpms –n|--node [nodes] –f|--format [format]

Parameters:

<--node>	These parameters specifies which node will be checked for Fujitsu RPM. This parameter is required.
<--format>	This parameter specifies the output format; text is default.This parameter is optional.

Output:

```
basking       1.1.2          Fujitsu Technology Solutions GmbH
basking-api   1.1.2          Fujitsu Technology Solutions GmbH
c2sux         1.1.RHEL6      Fujitsu Technology Solutions GmbH
fbootex       1.1.2          Fujitsu Technology Solutions GmbH
fcephadm      1.1.2          Fujitsu Technology Solutions GmbH
```

Example:

```
cd10000 rpms --node storage1 --format xml
```

14.3.3 Nodes Management

The Nodes management provides the following commands:

Command	Description
cd10000 nodes list preinst	Listing the nodes with preinst state. See "cd10000 ip node update - Updating information about the storage node" on page 297
cd10000 nodes list static	Listing the nodes in static network. See "cd10000 nodes list static - Listing the nodes in static network" on page 307
cd10000 nodes power status	Displaying power status of the node. See "cd10000 nodes power status - Displaying power status" on page 308
cd10000 nodes power on	Turning the power of node on. See "cd10000 nodes power on - Turning the power of node on" on page 309
cd10000 nodes power off	Turning the power of node off. See "cd10000 nodes power off - Turning the power of node off" on page 310
cd10000 nodes power reboot	Rebooting the node. See "cd10000 nodes power reboot- Rebooting the node" on page 311
cd10000 nodes power pxe	Booting the node. See "cd10000 ip switch list - Showing the list of the switches" on page 303

cd10000 nodes list preinst - Listing the nodes with preinst state

This command display the nodes which are in "preinst" state and are reachable either from the admin or the irmc network. If the `--details` argument is set, more information will be shown.

Synopsis:

`cd10000 nodes list preinst [admin,irmc] --details`

Parameters:

`<NETWORK>`	This parameter prints the nodes reachable from: • admin • irmc This parameter is required.
`<--details>`	This parameter displays additional information about nodes in the results. This parameter is optional.

Output:

```
------------------Searching iRMC DHCP network------------------
Found hosts: 0
```

Example:

```
cd10000 nodes list preinst irmc
```

cd10000 nodes list static - Listing the nodes in static network

This command display the nodes which are in the static network and are reachable either from the admin or irmc network. If the `--details` argument is set, more information will be shown.

Synopsis:

```
cd10000 nodes list static [admin,irmc] --details
```

Parameters:

`<NETWORK>`	This parameter prints nodes reachable from: • admin • irmc This parameter is required.
`<--details>`	This parameter displays additional information about nodes in the results. This parameter is optional.

Output:

```
-----------------Searching ADMIN STATIC network-----------------
Found hosts: 4
-----------------Searching ADMIN STATIC network-----------------
[OK] 192.168.20.11                  Reachable
[OK] 192.168.20.12                  Reachable
[OK] 192.168.20.13                  Reachable
[OK] 192.168.20.14                  Reachable
```

Example:

```
cd10000 nodes list preinst irmc --details
```

cd10000 nodes power status - Displaying power status

This command displays the power status of the node.

Synopsis:

cd10000 nodes power status –ip [ip]

Parameters:

<--ip>	This parameter displays iRMC IP of the node, whose power status will be showed. This parameter is required.

Output:

```
INFO - Chassis Power is on
```

Example:

```
cd10000 nodes power status -ip 192.168.10.11
```

cd10000 nodes power on - Turning the power of node on

This command turns the power on the given node on.

Synopsis:

cd10000 nodes power on –ip [ip]

Parameters:

<--ip>	This parameter displays iRMC IP of the node, which will be powered on. This parameter is required.

Output:

```
INFO - Chassis Power Control: Up/On
```

Example:

```
cd10000 nodes power on -ip 192.168.10.11
```

cd10000 nodes power off - Turning the power of node off

This command turns the power on the given node off.

Synopsis:

cd10000 nodes power off –ip [ip]

Parameters:

<--ip>	This parameter displays iRMC IP of the node, which will be powered off. This parameter is required.

Output:

```
INFO - Chassis Power Control: Down/Off
```

Example:

```
cd10000 nodes power off -ip 192.168.10.11
```

cd10000 nodes power reboot- Rebooting the node

This command reboots the node.

Synopsis:

cd10000 nodes power reboot –ip [ip]

Parameters:

<--ip>	This parameter displays iRMC IP of the node, which will be rebooted. This parameter is required.

Output:

```
INFO - Chassis Power Control: Cycle
```

Example:

```
cd10000 nodes power reboot -ip 192.168.10.11
```

cd10000 nodes power pxe - Booting the node

This command sets the option to the `pxe` and then reboots the node.

Synopsis:

`cd10000 nodes power pxe –ip [ip]`

Parameters:

`<--ip>`	This parameter displays iRMC IP of the node, which will be rebooted from pxe. This parameter is required.

Output:

```
INFO - Set Boot Device to pxe
INFO - Chassis Power Control: Reset
```

Example:

```
cd10000 nodes power pxe -ip 192.168.10.11
```

14.4 The Update Manager Commands

The Update Manager provides the following commands:

Command	Description
update-manager upload	Uploading RPM Packages to the RPM Repository and copies ISO to the source installation directory. See "cd10000 ip node list - Listing storage nodes" on page 296
update-manager update_partials	Updating Default Manifest Partials. See "cd10000 ip node update - Updating information about the storage node" on page 297
update-manager update	Updating Nodes in a Cluster. See "cd10000 nodes list static - Listing the nodes in static network" on page 307
update-manager upgrade	Upgrading Nodes in a Cluster. See "upgrade - Upgrading Nodes in a Cluster" on page 317

upload - Uploading RPM Packages to the RPM Repository and copies ISO to the source installation directory

The `update-manager` command with option `upload -d|--directory` uploads RPM packages from a specified directory to the RPM repository on the management node. RPM packages must be available in this directory.

The `update-manager` command with option `--iso`copies ISO to the installation directory and it uploads RPM packages from ISO to the internal RPM repository.

A mounted ISO image can be used as upload directory.

Syntax:

`update-manager upload -d|--directory "path1" [-i|--iso]`

Parameters:

`<directory>`	upload RPM directory: Enter the absolute path to the directory which contains the RPM packages to be uploaded. The specification of `<path1>` is mandatory. upload ISO: Enter the absolute path to the directory mounted ISO.
`<iso>`	(Optional) Flag to upload ISO

Output:

After the command has been executed successfully the following is displayed:

"`Successfully uploaded`"

Example:

```
update-manager upload -d "/tmp/update2/"
update-manager upload -d "/tmp/mounted_iso/" --iso
```

update partials - Updating Default Manifest Partials

The `update-manager` command with option `update_partials` updates package with default partial configuration. When custom partials were defined or default partials were changed, merge is recommended.

Syntax:

`update-manager update_partials`

Output:

After the command has been executed successfully the following is displayed:

"`Successfully generated`"

Example:

```
update-manager update_partials
```

update - Updating Nodes in a Cluster

The `update-manager` command with option `update` updates the specified nodes of the cluster. The nodes are sequentially updated.

Syntax:

```
update-manager update [-n|--nodes "NODE1 NODE2 ..."]
        [-t|--timeout TIMEOUT]
```

Parameters:

`<node_1> ... <node_n>`

Enter the host names of the nodes which shall be updated. If you do not enter a node, all nodes in the cluster (the management node included) will be updated sequentially.

`<timeout>`

Enter the timeout in seconds for the update of one node. If the node does not send the report and timeout expires, the node fails.

Output:

After the command has been executed successfully the following is displayed:

"`Nodes updated successfully`"

Examples:

- Update the storage nodes "storage2" and storage40":

  ```
  update-manager update -n “storage2 storage40” -t 150
  ```

upgrade - Upgrading Nodes in a Cluster

The **update-manager** command with option `upgrade` loads a software upgrade to the specified nodes of the cluster. The nodes are sequentially upgraded and rebooted.

Before the upgrade each node is switched to the maintenance mode and after a successful upgrade maintenance mode is switched off.

If the node is not in the Ceph cluster, the maintenance mode will not be used.

Syntax:

```
update-manager upgrade [-n|--nodes "NODE1 NODE2 ..."]
        [-t|--timeout TIMEOUT]
```

Parameters:

`<node_1> ... <node_n>`
: Enter the host names of the nodes which shall be upgraded. If you do not enter a node all nodes in the cluster (without management node) will be upgraded sequentially.

`<timeout>`
: Enter the timeout in seconds for upgrade of one node. If a node does not send the report and timeout expires, the node update fails.

Output:

After the command has been executed successfully the following is displayed:

"`Nodes updated successfully`"

Examples:

- Upgrade the management node "pmgmt":

  ```
  update-manager upgrade -n “pmgmt” -t 250
  ```

14.5 SNMP Command Line Interface

cd10000 setup snmp is the command line interface (CLI) for configuring SNMP communities and trap destinations. The following commands are available:

Command	Function	Description
set-community	Specifying a new community	"set-community - Defining a Community" on page 319
list-community	Displaying all communities.	"list-community - Displaying all Communities" on page 321
remove-community	Removing a community.	"remove-community - Removing a Community" on page 320
set-destination	Specifying a new trap destination.	"set-destination - Defining a Trap Destination" on page 322
list-destination	Displaying all trap destinations.	"list-destination - Displaying all Trap Destinations" on page 324
remove-destination	Removing a trap destination.	"remove-destination - Removing a Trap Destination" on page 323
–help [<subcommand>]	Printing help.	<subcommand> is optional.

Call format

cd10000 setup snmp <command><parameters>

set-community - Defining a Community

This command defines a new SNMP community.

Syntax:

```
cd10000 setup snmp set-community -n <name> -i <ip-address>
        [-m mask] -a {get|set|all}
```

Parameters:

`-n <name>`	Name of the community.
`-i <ip-address>`	IP address of the host that is allowed to access the SNMP information.
`-m <mask>`	Optional. Network mask if the access is defined for a subnet.
`-a {get|set|all}`	Type of access. Possible values: • get: read access • set: write access • all: read and write access

Examples:

- Define the community `community-1` for the IP address 192.168.10.10 and read access.

  ```
  cd10000 setup snmp set-community -n community-1 -i 192.168.10.10 -a get
  ```

- If the community has to be defined for a subnet, add a network mask:
  Define the community `community-2` for the subnet 192.168.11.0/24 and read access.

  ```
  cd10000 setup snmp set-community -n community-2 -i 192.168.11.0 -m
  255.255.255.0 -a get
  ```

remove-community - Removing a Community

This command removes an SNMP community.

Syntax:

cd10000 setup snmp remove-community -n <name>

Parameter:

-n <name>	Name of the community.

Example:

Remove the community community-1.

```
cd10000 setup snmp remove-community -n community-1
```

list-community - Displaying all Communities

This command displays all defined SNMP communities.

Syntax:

```
cd10000 setup snmp list-community
```

Output:

```
community <ip-address>/<mask> <access>
```

<ip-address>	IP address of the host that is allowed to access the SNMP information.
<mask>	Optional. Network mask if the access is defined for a subnet.
<access>	Type of access. Possible values: • get: read access • set: write access • all: read and write access

Example:

Two communities with the IP address 192.168.10.10 and the subnet 192.168.11.0/24 respectively exist, both with read access.

```
community-1 192.168.10.10 get
community-2 192.168.11.0/24 get
```

set-destination - Defining a Trap Destination

This command defines a new SNMP trap destination.

Syntax:

cd10000 setup snmp set-destination -i <ip-address>

Parameter:

`-i <ip-address>` IP address of the host that receives the traps.

Example:

Define a trap destination with the IP address 192.168.10.10.

```
cd10000 setup snmp set-destination -i 192.168.10.10
```

remove-destination - Removing a Trap Destination

This command removes a trap destination.

Syntax:

cd10000 setup snmp remove-destination -i <ip-address>

Parameter:

`-i <ip-address>` IP address of the trap destination.

Example:

Remove a trap destination with the IP address 192.168.10.10.

```
cd10000 setup snmp remove-destination -i 192.168.10.10
```

list-destination - Displaying all Trap Destinations

This command displays all defined trap destinations.

Syntax:

```
cd10000 setup snmp list-destination
```

Output:

```
destinations --------------- <ip-address>
```

<ip-address>	IP address of the host that receives the traps.

Example:

One trap destination with the IP address 192.168.10.10 exists.

```
destinations --------------- 192.168.10.10
```

15 Appendix A: VSM Specific Technical References

This section provides low-level details of how VSM controls the Ceph cluster.

15.1 VSM-Generated CRUSH Map Generation

The following CRSUH map is for a Ceph cluster configured as follows:

- Five host servers
- Each host server contains 20 data disks, for a total of 100 data disks
    - 14 disks belong to the 7200RPM_SATA storage class
    - 6 disks belong to the 10KRPM_SAS storage class
- All host servers reside in zone 1

VSM does the following:

- Creates a `performance` storage group bucket for the drives belonging to the 7200RPM_SATA storage group
    - Creates a single `zone_one_performance` zone bu cket under the performance storage group
        - Creates a `hf1gpcephx_performance_zone_one` host bucket for each server in zone 1 containing drives belonging to the 7200RPM_SATA storage group
            - Creates and OSD leaf for each OSD on each server containing drives belonging to the 7200RPM_SATA storage group
- Creates a `high_performance` storage group for the drives belonging to the 10KRPM_SAS storage group
    - Creates a single `zone_one_high_performance` zone bucket under the performance storage group
        - Creates a `hf1gpcephx_high_performance_zone_one` host bucket for each server in zone 1 containing drives belonging to the 10KRPM_SATA storage group
            - Creates and OSD leaf for each OSD on each server containing drives belonging to the 10KRPM_SATA storage group
- Creates a ruleset for the `performance` storage group bucket
    - All pools created in the `performance` storage group will use this ruleset

- Creates a ruleset for the `high_performance` storage group buck
  - All pools created in the `high_performance` storage group will use this ruleset

Crush map

```
# begin crush map
tunable choose_local_tries 0
tunable choose_total_tries 50
tunable chooseleaf_descend_once 1
tunable chooseleaf_vary_r 1

# devices
device 0 osd.0
device 1 osd.1
device 2 osd.2
device 3 osd.3
device 4 osd.4
device 5 osd.5
device 6 osd.6
device 7 osd.7
device 8 osd.8
device 9 osd.9
device 93 osd.93
device 94 osd.94
device 95 osd.95
device 96 osd.96
device 97 osd.97
device 98 osd.98
device 99 osd.99
```

```
# types
type 0 osd
type 1 host
type 2 zone
type 3 storage_group
type 4 root

# buckets
host hf1gpceph1_performance_zone_one {
        id -1
        # weight 14.000
        alg straw
        hash 0  # rjenkins1
        item osd.6 weight 1.000
        item osd.7 weight 1.000
        item osd.8 weight 1.000
        item osd.9 weight 1.000
        item osd.10 weight 1.000
        item osd.11 weight 1.000
        item osd.12 weight 1.000
        item osd.13 weight 1.000
        item osd.14 weight 1.000
        item osd.15 weight 1.000
        item osd.16 weight 1.000
        item osd.17 weight 1.000
        item osd.18 weight 1.000
        item osd.19 weight 1.000
}
host hf1gpceph1_high_performance_zone_one {
        id -2
        # weight 6.000
        alg straw
        hash 0  # rjenkins1
        item osd.0 weight 1.000
        item osd.1 weight 1.000
        item osd.2 weight 1.000
        item osd.3 weight 1.000
        item osd.4 weight 1.000
        item osd.5 weight 1.000
}
```

```
host hf1gpceph2_performance_zone_one {
        id -3
        # weight 14.000
        alg straw
        hash 0  # rjenkins1
        item osd.26 weight 1.000
        item osd.27 weight 1.000
        item osd.28 weight 1.000
        item osd.29 weight 1.000
        item osd.30 weight 1.000
        item osd.31 weight 1.000
        item osd.32 weight 1.000
        item osd.33 weight 1.000
        item osd.34 weight 1.000
        item osd.35 weight 1.000
        item osd.36 weight 1.000
        item osd.37 weight 1.000
        item osd.38 weight 1.000
        item osd.39 weight 1.000
}
host hf1gpceph2_high_performance_zone_one {
        id -4
        # weight 6.000
        alg straw
        hash 0  # rjenkins1
        item osd.20 weight 1.000
        item osd.21 weight 1.000
        item osd.22 weight 1.000
        item osd.23 weight 1.000
        item osd.24 weight 1.000
        item osd.25 weight 1.000
}
```

```
host hf1gpceph3_performance_zone_one {
        id -5
        # weight 18.000
        alg straw
        hash 0  # rjenkins1
        item osd.42 weight 1.000
        item osd.43 weight 1.000
        item osd.44 weight 1.000
        item osd.45 weight 1.000
        item osd.46 weight 1.000
        item osd.47 weight 1.000
        item osd.48 weight 1.000
        item osd.49 weight 1.000
        item osd.50 weight 1.000
        item osd.51 weight 1.000
        item osd.52 weight 1.000
        item osd.53 weight 1.000
        item osd.54 weight 1.000
        item osd.55 weight 1.000
        item osd.56 weight 1.000
        item osd.57 weight 1.000
        item osd.58 weight 1.000
        item osd.59 weight 1.000
}
host hf1gpceph3_high_performance_zone_one {
        id -6
        # weight 2.000
        alg straw
        hash 0  # rjenkins1
        item osd.40 weight 1.000
        item osd.41 weight 1.000
}
```

```
host hf1gpceph4_performance_zone_one {
        id -7
        # weight 14.000
        alg straw
        hash 0  # rjenkins1
        item osd.66 weight 1.000
        item osd.67 weight 1.000
        item osd.68 weight 1.000
        item osd.69 weight 1.000
        item osd.70 weight 1.000
        item osd.71 weight 1.000
        item osd.72 weight 1.000
        item osd.73 weight 1.000
        item osd.74 weight 1.000
        item osd.75 weight 1.000
        item osd.76 weight 1.000
        item osd.77 weight 1.000
        item osd.78 weight 1.000
        item osd.79 weight 1.000
}
host hf1gpceph4_high_performance_zone_one {
        id -8
        # weight 6.000
        alg straw
        hash 0  # rjenkins1
        item osd.60 weight 1.000
        item osd.61 weight 1.000
       item osd.62 weight 1.000
        item osd.63 weight 1.000
        item osd.64 weight 1.000
        item osd.65 weight 1.000
}
```

```
host hf1gpceph6_performance_zone_one {
        id -9
        # weight 14.000
        alg straw
        hash 0  # rjenkins1
        item osd.86 weight 1.000
        item osd.87 weight 1.000
        item osd.88 weight 1.000
        item osd.89 weight 1.000
        item osd.90 weight 1.000
        item osd.91 weight 1.000
        item osd.92 weight 1.000
        item osd.93 weight 1.000
        item osd.94 weight 1.000
        item osd.95 weight 1.000
        item osd.96 weight 1.000
        item osd.97 weight 1.000
        item osd.98 weight 1.000
        item osd.99 weight 1.000
}

host hf1gpceph6_high_performance_zone_one {
        id -10
        # weight 6.000
        alg straw
        hash 0  # rjenkins1
        item osd.80 weight 1.000
        item osd.81 weight 1.000
        item osd.82 weight 1.000
        item osd.83 weight 1.000
        item osd.84 weight 1.000
        item osd.85 weight 1.000
}
```

```
zone zone_one_performance {
        id -11
        # weight 74.000
        alg straw
        hash 0  # rjenkins1
        item hf1gpceph1_performance_zone_one weight 14.000
        item hf1gpceph2_performance_zone_one weight 14.000
        item hf1gpceph3_performance_zone_one weight 18.000
        item hf1gpceph4_performance_zone_one weight 14.000
        item hf1gpceph6_performance_zone_one weight 14.000
}
zone zone_one_high_performance {
        id -12
        # weight 26.000
        alg straw
        hash 0  # rjenkins1
        item hf1gpceph1_high_performance_zone_one weight 6.000
        item hf1gpceph2_high_performance_zone_one weight 6.000
        item hf1gpceph3_high_performance_zone_one weight 2.000
        item hf1gpceph4_high_performance_zone_one weight 6.000
        item hf1gpceph6_high_performance_zone_one weight 6.000
}
storage_group performance {
        id -13
        # weight 74.000
        alg straw
        hash 0  # rjenkins1
        item zone_one_performance weight 74.000
}
storage_group high_performance {
        id -14
        # weight 26.000
        alg straw
        hash 0  # rjenkins1
        item zone_one_high_performance weight 26.000
}
```

```
root vsm {
        id -15
        # weight 100.000
        alg straw
        hash 0  # rjenkins1
        item performance weight 74.000
        item high_performance weight 26.000
}

# rules
rule high_performance {
        ruleset 0
        type replicated
        min_size 0
        max_size 10
        step take high_performance
        step chooseleaf firstn 0 type host
        step emit
}
rule performance {
        ruleset 1
        type replicated
        min_size 0
        max_size 10
        step take performance
        step chooseleaf firstn 0 type host
        step emit
}

# end crush map
```

15.2 VSM-Generated CRUSH Rulesets

In VSM each Storage Group has a single corresponding CRUSH ruleset; when a pool is created, selection of the Storage Group implicitly selects the corresponding ruleset.

The rulesets for each Storage Group are defined as follows:

Capacity

```
// Description: Primary copy and replicas in capacity (7200 RPM SAS/SATA) s
torage group
rule capacity{
    ruleset 3
    type replicated
    min_size 0
    max_size 10
    step take capacity
    step chooseleaf firstn 0 type zone
    step emit
}
```

Performance

```
// Description: Primary copy and replicas in performance (10K SAS) storage
group
rule performance{
    ruleset 4
    type replicated
    min_size 0
    max_size 10
    step take performance
    step chooseleaf firstn 0 type zone
    step emit
}
```

High Performance

```
// Description: Primary copy and replicas in high_performance (SSD) storage
group
rule high_performance{
    ruleset 5
    type replicated
    min_size 0
    max_size 10
    step take high_performance
    step chooseleaf firstn 0 type zone
    step emit
}
```

Value Capacity

```
// Description: Primary copy in value_capacity (ssd cached 7200RPM SATA) st
orage group
// replica copies in capacity storage group
rule value_capacity{
    ruleset 6
    type replicated
    min_size 0
    max_size 10
    step take value_capacity
    step chooseleaf firstn 1 type zone
    step emit
    step take capacity
    step chooseleaf firstn -1 type zone
    step emit
}
```

Value Performance

```
// Description: Primary copy in value_performance (ssd cached 10K SAS) stor
age group
// replica copies in performance (10K SAS) storage group
#rule value_performance{
    ruleset 7
    type replicated
    min_size 0
    max_size 10
    step take value_performance
    step chooseleaf firstn 1 type zone
    step emit
    step take performance
    step chooseleaf firstn -1 type zone
    step emit
}
```

15.3 PG States

The following table summarizes the possible Ceph PG states (contents was taken over from http://ceph.com/docs/master/rados/operations/pg-states/):

PG State	Description
Creating	Ceph is still creating the placement group.
Active	Ceph will process requests to the placement group.
Clean	Ceph replicated all objects in the placement group the correct number of times.
Down	A replica with necessary data is down, so the placement group is offline.
Replay	The placement group is waiting for clients to replay operations after an OSD crashed.
Splitting	Ceph is splitting the placement group into multiple placement groups.
Scrubbing	Ceph is checking the placement group for inconsistencies.
Degraded	Ceph has not replicated some objects in the placement group the correct number of times yet.
Inconsistent	Ceph detects inconsistencies in the one or more replicas of an object in the placement group (e.g. objects are the wrong size, objects are missing from one replica after recovery finished, etc.).
Peering	The placement group is undergoing the peering process
Repair	Ceph is checking the placement group and repairing any inconsistencies it finds (if possible).
Recovering	Ceph is migrating/synchronizing objects and their replicas.
Backfill	Ceph is scanning and synchronizing the entire contents of a placement group instead of inferring what contents need to be synchronized from the logs of recent operations. Backfill is a special case of recovery.
Wait-backfill	The placement group is waiting in line to start backfill.
Backfill-toofull	A backfill operation is waiting because the destination OSD is over its full ratio.
Incomplete	Ceph detects that a placement group is missing a necessary period of history from its log. If you see this state, report a bug, and try to start any failed OSDs that may contain the needed information.

PG State	Description
Stale	The placement group is in an unknown state - the monitors have not received an update for it since the placement group mapping changed.
Remapped	The placement group is temporarily mapped to a different set of OSDs from what CRUSH specified.

16 Appendix B: Troubleshooting

16.1 Stopping Storage Nodes without Rebalancing

Components of the system (e.g. storage nodes, networking) may periodically require maintenance. In addition you may need to resolve a problem that affects a failure domain (i.e. zone).

If you shut down a storage node in the Ceph storage cluster or disconnect a storage node from the cluster, Ceph will determine that OSDs have become unresponsive. Ceph will attempt to repair (rebalance) the storage cluster by creating new copies of the data that resided on the unresponsive OSDs. This may result in substantial data movement and may reduce cluster performance.

To prevent rebalancing in the case where a storage node disconnection is expected to be temporary, it is recommended to stop the storage node by running the VSM **Stop Servers** operation on each of the affected storage nodes.

- To run the operation click the **Stop Servers** button on the VSM page **Server Management** > **Manage Servers**.

  For detailed information, see "Stopping storage nodes " on page 114

  You can also use the **`vsm_cli maint-on`** command to stop a storage node without rebalancing (see page 263).

When the **Stop Servers** operation is executed, VSM sets Ceph to "`noout`", and then stops each OSD on the selected storage node(s).

Setting the cluster to "`noout`" tells Ceph not to rebalance the storage cluster if OSDs become unresponsive. Setting "noout" affects the entire cluster, and remains in effect while any storage node is in the "`stopped`" status.

When maintenance is completed, restart the stopped storage nodes using the **Start Servers** operation:

- To run the operation click the **Start Servers** button on the VSM page **Server Management** > **Manage Servers**.

  For detailed information, see "Starting storage nodes" on page 115

  You can also use the **`vsm_cli maint-off`** command to restart a storage node (see page 263).

VSM restarts the OSDs on the affected storage nodes and clears Ceph from "noout".

As "noout" affects the entire cluster, it is recommended to limit the time that storage nodes are stopped to the minimum required to performance maintenance or resolve other problems.

16.2 OSDs Not Up and In

OSDs may become unresponsive due to transient issues such as an overloaded OSD, or may become unresponsive because of the failure in an associated data or journal disk.

When Ceph detects that an OSD has become unresponsive, it will place the OSD to the status "down" and "out" – this is referred to as "Autoout".

You can identify OSDs that are "down" and "out" by inspecting the **OSD Summary** section (see the following figure) on the VSM dashboard.

- To display the **OSD Summary** select **Dashboard** > **Cluster Status** or **Monitor Cluster** > **OSD Status** on the VSM GUI.

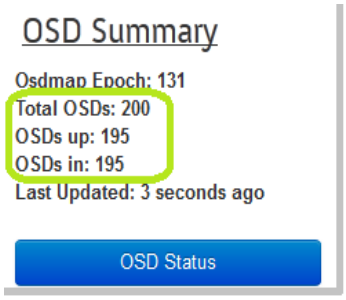


Figure 77: OSD Summary

The number of **OSDs up** and the number of **OSDs in** should be equal the number of **Total OSDs**. In the figure above, the number of **Total OSDs** is 200, while the number of **OSDs up** and **OSDs in** is 195, indicating that five OSDs are "down" and "out":

- To identify OSDs that have been placed to the status "Autoout" select **Server Management** > **Manage Devices** or **Monitor Cluster** > **OSD Status** on the VSM GUI.
- Click the header of the **OSD State column** to sort the list on the OSD status.
- You can attempt to restart OSDs in the status "Autoout" using the **Restart Osds** button on the **Server Management > Manage Devices** page.

  For detailed information, see "Restarting disks/OSDs" on page 124.

OSDs that will not restart or that repeated return to "Autoout" may indicate a failure in the associated data or journal disk.

A set of "Autoout" OSDs that share the same journal SSD may indicate that the journal disk is the common source of failure ((see the following figure)

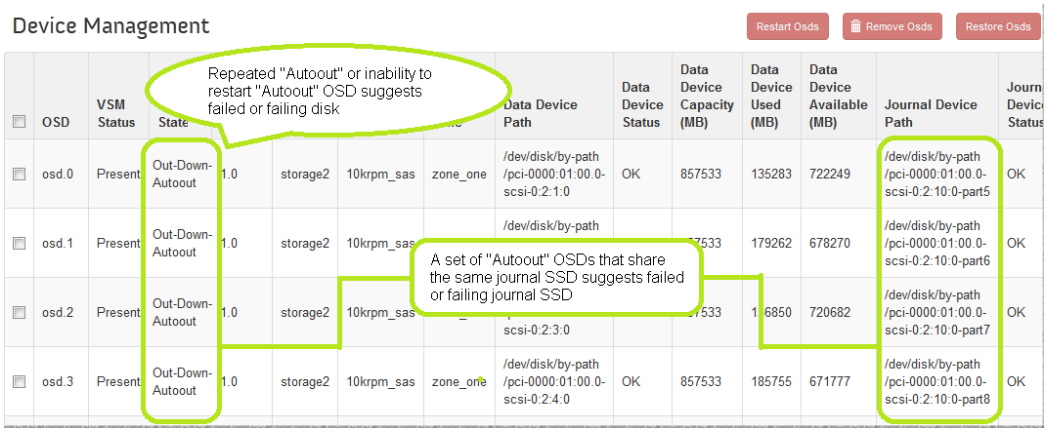


Figure 78: OSDs in the status "Autoout"

VSM periodically checks to verify the presence of the data and journal devices. An "Error" in the **Device Status** or **Journal Device Status** column indicates that the OS (Operating System) no longer thinks the drive is present, and may be indicative of a drive failure, cabling failure, or physical connection issue.

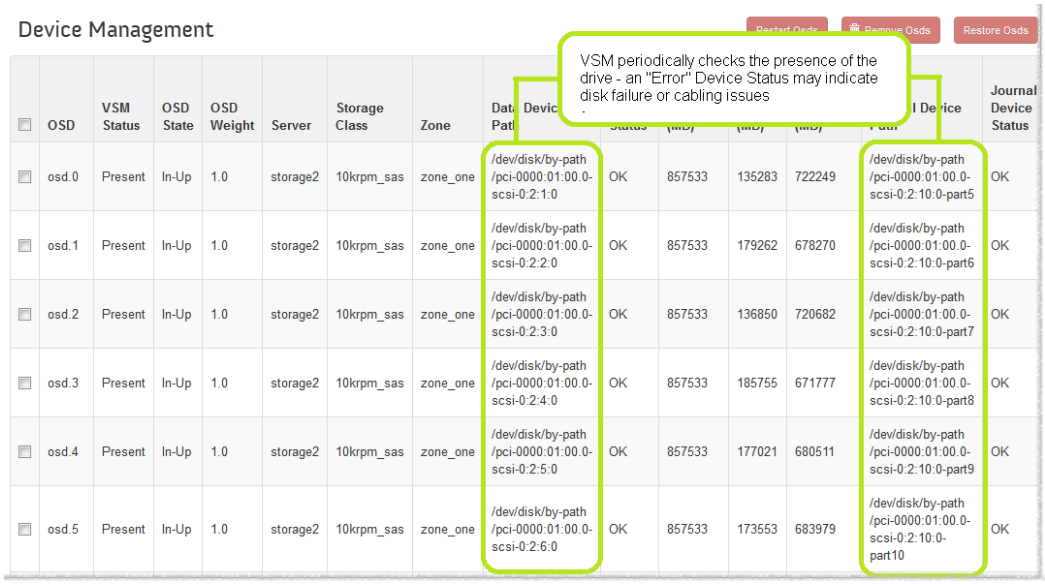


Figure 79: Checking Journal Device Status

16.3 Monitor Clock Skew

High skew indicates that the clocks of the servers hosting the monitors of your Ceph storage cluster are out of time synchronization.

- To identify high skew select **Monitor Cluster > Monitor Status** on the VSM GUI.

Clocks can be synchronized using the NTP (**N**etwork **T**ime **P**rotocol).

17 Appendix C: Error Codes & Messages

When unusual events occur, the individual software components of the ETERNUS CD10000 system issue messages.

In addition to the message number, the message text itself and a more detailed message provide additional information.

The different software functions are using error codes which are prefixed with an abbreviation, e.g. SNM for SNMP messages.

The return codes are defined as follows:

```
<msg-id> :<errflag>:<message>:<reason>
```

Where:

```
<msg-id> = <component><number>
```

`<msg-id>`	Message ID
`<component>`	Specifies the component which generates the return code, e.g. FCA for VSM CLI, SNM for SNMP, TLS for TOOLS
`<number>`	A four-digit message number
`<errflag>`	An error flag to specify the message type. Possible values are: `E` = error, `N` = notice, `W` = warning
`<message>`	General description of the error/notice/warning,
`<reason>`	Additional information as to the cause of the error/notice/warning

In the messages listed in this section the following placeholder will be used where applicable:

- the system error message (like "`No space left on device`") is `<reason>` and
- the name of the file that caused the error is `<file>`.

17.1 SNMP Related Messages (SNM)

17.1.1 Common SNMP Errors

Error Code	Message	Comment
SNM1000	SNM1000:E:Failed to write to pidfile <file>: <reason>	During startup, SNMP agent could not write to specified <filename>. The system has returned the error <reason>.
SNM1001	SNM1001:E:Failed to create pidfile <file>: <reason>	During startup the SNMP agent could not create the file <filename>. The system has returned the error <reason>.
SNM1002	SNM1002:E:cannot create a cluster handle: <reason>	The SNMP agent could not connect to the Ceph cluster. If <reason>= Success, Ceph has caught an internal error that is not passed to the SNMP agent.
SNM1003	SNM1003:E:cannot read config file through Ceph: <reason>	While parsing the Ceph configuration file the SNMP agent received an error from Ceph. If <reason> = Success, Ceph has either caught an internal error that is not passed to the SNMP agent, or has encountered malformed configuration.
SNM1004	SNM1004:E:cannot read config value <name>: <reason>	The SNMP agent cannot read the config value <name>. <reason> states the actual cause for this failure. If <reason> = Success, Ceph has either caught an internal error that was not passed to the SNMP agent, or has encountered malformed configuration.
SNM1005	SNM1005:E:bad or missing db config, check /etc/vsmdeploy/deployrc	The SNMP agent could not read /etc/vsmdeploy/deployrc file - at least the MYSQL_VSM_USER and the MYSQL_VSM_PASSWORD values are missing.

Error Code	Message	Comment
SNM1006	SNM1006:E:error connecting to db <DB>: <reason>	The SNMP agent could not connect to the database <DB>. <reason> provides more information on the actual cause.
SNM1007	SNM1007:E:mysql_init() failed	Invalid or missing MySQL C client libraries or fatal memory conditions.
SNM1008	SNM1008:W:lost connection to db <db>: <reason>	The connection to the MySQL server has been lost and reconnection is being attempted. If <reason> = MySQL server has gone away, the connection has been finished by timeout. If this error is followed by the error SNM1006, possibly the MySQL instance has been disabled or has crashed.
SNM1009	SNM1009:E:failed to receive results for query <query>: <reason>	The SNMP agent has failed to receive results for the query <query>. <reason> states the actual cause for the failure.
SNM1010	SNM1010:E:error executing query <query>: <reason>	The SNMP agent received the error <reason> while executing the MySQL query <query>. Possible reasons are: • an outdated VSM schema in the SNMP agents or • the SNMP agent is connected to the invalid MySQL database.
SNM1011	SNM1011:E:Can't connect directly to Ceph cluster: <reason>	The SNMP agent attempted to connect directly (using librados) to the cluster and failed with reason <reason>
SNM1012	SNM102:E:Cluster monitor command failed: <reason>	The SNMP agent issued a cluster monitor command to a Ceph cluster and got <reason> error response.

17.1.2 OSDTable Agent Errors

Error Code	Message	Comment
SNM1100	SNM1100:E:failed to update <file>: <reason>	The SNMP agent failed to update the file <file>. The cause is specified in <reason>.
SNM1101	SNM1101:E:failed to create <file>: <reason>	The SNMP agent failed to create the file <file>. The actual cause is specified in <reason>.
SNM1102	SNM1102:E:failed to copy admin keyring, every attempt has failed	While copying the configuration from VSM, the SNMP agent failed to copy the admin keyring from any node configured in VSM. Possibly, an invalid VSM setup or a network failure is given.
SNM1103	SNM1103:E:failed to locate admin keyring, no result from database	While copying the configuration from VSM, the SNMP agent did not find any nodes. The message indicates an invalid or empty VSM setup.
SNM1104	SNM1104:W:get_disk_usage: invalid osd index returned (<num>)	While parsing the OSD list, the SNMP agent has received data for the OSD that is out of the known range. The message indicates either a bug in Ceph or a foreign node (a node with OSDs that are not included in this cluster).
SNM1105	SNM1105:W:get_disk_usage: error scanning osd data (expected 5 args, got <num> on format <num>)	While parsing the OSD list, the SNMP agent has received garbage data. The message indicates either a bug in the SNMP agent or a node misbehavior.
SNM1106	SNM1106:E:read_osds: Ceph failure: <reason>	The SNMP agent failed to establish a connection to Ceph. The cause is specified in <reason>. Possibly, errors within the Ceph installation or a lack of system resources.
SNM1107	SNM1107:E:Possibly invalid routing table, can't locate monitors.	The SNMP agent can not connect to Ceph cluster, possibly because routing table for cluster is not valid (no routing between management node and storage nodes - monitor subnet not in routing table).

17.1.3 MonTable Agent Errors

Error Code	Message	Comment
SNM1200	SNM1200:E:failed to connect to ceph.	The SNMP agent could not access Ceph. This message indicates errors within the Ceph installation.
SNM1201	SNM1201:E:error scanning mon data (expected 1 arg, got <num>)	The SNMP agent has encountered an invalid [mon.<num>] section title. This message indicates errors in the /etc/ceph/ceph.conf file.

17.1.4 NodeTable Agent Errors

Error Code	Message	Comment
SNM1300	SNM1300:E:error resolving host <host>: <reason>	The SNMP agent failed to t resolve the name of the host <host>. The cause is specified in <reason>. If this error message is caused/signaled by the operating system, there will be another error message stating the reason on the operating system side (like "Out of memory").
SNM1301	SNM1301:E:error receiving ICMP datagram: <reason>	This error occurred during the reception of the ICMP ECHO REPLY packet.It indicates a sudden network failure or an operating system failure.
SNM1302	SNM1302:E:Global socket in err_fds: <reason>	The global ICMP send/receive socket for the SNMP agent was set by the operating system into error condition. The actual cause (if any) is specified in <reason>. Mostly this message indicates that the socket was forcibly closed on the OS side.

Error Code	Message	Comment
SNM1303	SNM1303:E:failed to create ICMP socket: <reason>	The SNMP agent failed to create the global ICMP socket. <reason> states the actual cause. In most cases, reason is: • the SNMP agent has insufficient privileges (the agent binary is not setuid, and the agent is not running as root) or • the limit for the open system file descriptors has been exceeded
SNM1304	SNM1304:E:failed to send ICMP packet: <reason>	The SNMP agent failed to send the ICMP packet. The cause is specified in <reason>. In most cases, the error is caused by a network stack malfunction or insufficient privileges (the agent must be running as root or the agent binary be setuid).

17.1.5 CDClusterInfo Agent Errors

Error Code	Message	Comment
SNM1400	SNM1400:W:unknown unit (<unit>) in units_to_gb	The SNMP agent failed to parse the <unit> data size unit returned by Ceph. Possible cause is a bug in the agent or in Ceph.
SNM1401	SNM1401:W:unit is null in units_to_gb	The SNMP agent received garbage data from Ceph. This error indicates a bug in Ceph.

17.2 INSTALLATION Related Messages (INS)

Error Code	Message	Comment
INS1001	INS1001:E:RAID controller configuration fail	RAID controller could not be configured.
INS1002	INS1002:E:Failed to prepare pxe on management node	Preparing pxe failed during configuration.
INS1003	INS1003:E:Failed to setup network on management node	Setup network failed on management node.
INS1004	INS1004:E:Failed to set LVM drive on management node	LVM drive creating fail on management node.
INS1005	INS1005:E:Failed partitioning disks	Partitioner could not partitioning disks.

17.3 Software PACKAGE MANAGER (FPM)

Error Code	Message	Comment
FPM1001	Directory <dir> does not exists	The directory <dir> specified in the upload option does not exist. <dir> should exist and contain the RPM packages to upload.
FPM1002	Repository <repository> does not exists	The update-manager failed to find the main RPM repository of a cluster. Probable solution: Restart the daemons on the management node: mongod, pulp-agent. If they do not work, contact support.
FPM1003	<node> is not reachable	The storage node <node> is not available in the network. Probably <node> is switched off. **Solution:** Switch <node> on.
FPM1004	VSM_CLI error	The executed VSM_CLI application to manage a Ceph cluster failed. For detailed information, see the VSM_CLI log file and documentation.

Error Code	Message	Comment
FPM1005	Ceph error. Failed to get cluster health	The command 'ceph health' on management node failed. Probably the Ceph configuration on the management node is missing or is not up to date.
FPM1101	Failed to create repository <repository>	The attempt to initialize the main RPM repository of a cluster failed. For detailed information, see /var/log/cd10000.log and /var/log/pulp/
FPM1102	Failed to upload RPMs from <directory> to repository <repository>	Probably given <directory> is empty. If <directory> is not empty, restart the daemons on the management node: mongod, pulp-agent. If it still does not work, contact support.
FPM1103	Failed to publish RPMs in repository <repository>	Pulp internal error, contact support.
FPM1104	Failed to initialize Pulp	Probable solution: Restart the daemons on the management node: mongod, pulp-agent. If they do not work, contact support.
FPM1105	Failed to clean yum cache	For the root of issue check yum log file.
FPM1106	Failed to install/update <packages>	Failed to update the RPM packages with the cluster defined state. Contact support
FPM1107	Failed to upgrade packages on <node>	Failed to upgrade <node> to new version. Contact support.
FPM1201	Manifest syntax is invalid <file>	The generated manifest in <file> is invalid due to unmerged partials or invalid partials. **Solution:** Fix broken partials.
FPM1202	Found unmerged partial file <file>	<file> is not proper partial. After an update or custom changes all partials cannot be duplicated and all must have .pp extension.
FPM1203	Failed to save manifest file <file>	The update-manager cannot write into <file>. Check the file permissions
FPM1204	Failed to open partial <file>	The update-manager failed to read <file>. Check if file exists.

Error Code	Message	Comment
FPM1205	Failed to trigger update/upgrade on <node>	The update-manager request was blocked by a client on <node> or <node>is already updating. Check the log files in /var/log/puppet/ on the storage node <node>.
FPM1206	Report waiting timeout in <directory>	The client node did not send the update/upgrade report to the management node. Probably, the timeout of the node update is too short. Check the log files in /var/log/puppet/ on the node.
FPM1207	Found error in report <report_path>	The update/upgrade of the storage node failed. Contact support.
FPM1208	<Failed to generate partial in <file>	The server <node> failed to write partial to a <file>. Check permissions of <file>.
FPM1209	<node> did not start reboot process	The server <node> failed to reboot after an update/upgrade. Manual reboot may be required.
FPM1210	<node> did not start after reboot	The server <node> did not boot. Issue must be checked manually.

17.4 IP MANAGER (FIP)

Error Code	Message	Comment
FIP1001	FIP1001:E:Can not reload puppetmaster service	The Fipmanager could not reload puppet service.
FIP1002	FIP1002:E:Can not remove old certificate: <certname>	The Fipmanager could not remove old certificate from puppet.
FIP1003	FIP1003:E:Can not kick node: <hostname>	The Fipmanager could not kick a storage node. See if puppet certificates are valid.
FIP1004	FIP1004:E:Can not get puppet facts from management node	The Fipmanager could not get facts from management node. See if puppet certificates are valid.
FIP1005	FIP1005:E:Can not get facts from host: <hostname>	The Fipmanager could not get facts from storage node. See if puppet certificates are valid.
FIP1006	FIP1006:E:Can not get puppet certificate list	The Fipmanager could not get certificate list from puppet.
FIP1007	FIP1007:E:Failed to prepare cluster.manifest	The Fipmanager could not prepare VSM cluster manifest. Check permissions in /etc/manifest/cluster.manifest.
FIP1008	FIP1008:E:Empty owner tag	Empty owner tag
FIP1009	FIP1009:E:Puppet locked by: <msg>	Puppet locked by: <msg>
FIP1010	FIP1010:E:Can not write puppet manifest	Fipmanager can not write manifest file to the file system.
FIP1101	FIP1101:E:Can not update node: <hostname>	The Storage node can not be updated. Check update manager logs: /var/log/fpm.log

Error Code	Message	Comment
FIP1102	FIP1102:E:Update manager error	The Storage node can not be updated. Check update manager logs: /var/log/fpm.log
FIP1201	FIP1201:E:Data Base connection error	The Fipmanager could not connect to posgresql data base. Check if postgresql service is running
FIP1301	FIP1301:E:Critical unhandled error	Critical unhandled error occurred. For more information /var/log/fipmanager/fipmanager.log
FIP1302	FIP1302:E:Storage node witch serial number: <sn> does not exist in data base	The Fipmanager could not find storage node in data base.
FIP1303	FIP1303:E:Switch witch ip: <sn> does not exist in data base	The Fipmanager could not find switch in data base.
FIP1304	FIP1304:E:IP value incorrect: <msg>	IP value is incorrect, in the message is detailed description.
FIP1305	FIP1305:E:Network value incorrect: <msg>	Network value incorrect, in the message is detailed description.
FIP1306	FIP1306:E:Change ip error. Timeout. Node unreachable.	The Fipmanager could not set new IP for storage node.
FIP1307	FIP1307:E:Unknown parameter	An Error occurs when the fipmanager can not recognize set parameter for storage node.
FIP1308	FIP1308:E:Hostname incorrect.	An Error occurs when the hostname of storage node is wrong, should start witch storageX.
FIP1309	FFIP1309:E:Unsupported Brocade's parameter	Brocade switch cannot be described with position
FIP1310	FIP1310:E:Incorrect limit value	Incorrect limit value (max = 200) in setup data network command.
FIP1311	FIP1311:E:Incorrect start point value	Incorrect start point value (default = 11) in setup data network command.

17.5 TOOLS (TLS)

17.5.1 Command cd10000 setup snmp

Error Code	Message	Comment
TLS1000	TLS1000:E:invalid network address <addr> (must be in A.B.C.D format)	The address provided as argument to cd10000 setup snmp is no valid Ipv4 address (format A.B.C.D, where A, B, C and D are integers from the range <0; 255>).
TLS1001	TLS1001:E:invalid network address <addr>, invalid IP.	The address provided as argument to cd10000 setup snmp is no valid Ipv4 address (format A.B.C.D, where A, B, C and D are integers from the range <0; 255>).
TLS1002	TLS1002:E:<binary> does not exist or not executable	The executable <binary> does not exist or does not have “execute” rights. You may need to reinstall the appropriate SNMP package.
TLS1003	TLS1003:E:Option <opt> not specified	The <opt> option is mandatory and is not given in the command line.
TLS1004	TLS1004:E:Community <comm> already exists	The message is shown after the attempt to add community that already exists in the SNMP configuration.
TLS1005	TLS1005:E:Option ‘-a’ has to be get, set or all	A wrong argument was entered for -a option. The argument needs to be get, set or all.
TLS1006	TLS1006:E:Destination <dest> already exists	The message is shown after attempt to add destination that already exists in the SNMP configuration.
TLS1007	TLS1007:E:Community <comm> does not exist	The message is shown after the attempt to remove a nonexistent community..
TLS1008	TLS1008:E:Destination <dest> does not exist	The message is shown after the attempt to remove a nonexisting destination

17.5.2 Command cd10000

Error Code	Message	Comment
TLS1051	TLS1051:E:Failed to generate partial with password	This error occurred when writing the partial file. Check the permissions for /opt/nforma/fpm/manifests/genericstorage.d/ root_password.pp
TLS1052	TLS1052:E:Firstboot error	The execution of /opt/nforma/ftools/scripts/ enable_firstboot.sh failed. For detailed information, see the /var/log/cd10000.log file.
TLS1053	TLS1053:E:Failed to list installed Fujitsu packages	Failed to query the installed packages with the rpm command. For detailed information, see the /var/log/cd10000.log file.

17.5.3 Command to Get Node Information (Node Manager)

Error Code	Message	Comment
TLS1101	TLS1101:E:Failed to scan network for storage nodes	Network failure. Check for network issues.
TLS1102	TLS1102:E:Host <hostname> is not reachable	The Server <hostname> is not available or the specified name <hostname> is wrong.
TLS1003	TLS1003:E:Ipmi error	Failed to execute ipmitool application. Look into /var/log/cd10000.log for more information.
TLS1004	TLS1004:E:Failed to execute fpartitioner on <node>	Partitioning of disks failed. For detailed information, see the /var/log/fparitioner.log file on node <node>.

17.5.4 Command cd10000 Support

Error Code	Message	Comment
TLS1151	TLS1151:E:Given destination directory is a file	The output destination is not a directory but a file.
TLS1152	TLS1152:E:Failed to create output directory	The output directory could not be created.
TLS1153	TLS1153:E:Failed to create the compress file	The output compressed file could not be created.
TLS1154	TLS1154:E:Failed to read the xml configuration file	The xml configuration file could not be read.
TLS1155	TLS1155:E:Host: <hostname> is not reachable	Server with <hostname> is not available or given hostname is wrong.
TLS1156	TLS1156:E:Failed to create output textfile	The output compressed file could not be created.
TLS1157	TLS1157:E:Failed to execute sosreport on <hostname>	The SOSReport could not be executed successfully on <hostname>.
TLS1158	TLS1158:E:Failed to remove sosreport on <hostname>	The temporary SOSReport's files could not be removed on <hostname>.
TLS1159	TLS1159:E:Failed to execute <command line> on <hostname>	The <command line> could not be executed successfully on <hostname>.
TLS1160	TLS1160:E:Cannot copy <file or directory> from <hostname>	The <file or directory> could not be copied from <hostname> to pmgmt.
TLS1161	TLS1153:E:Failed to create the compress file	The temporary gathered files could not be removed.

17.5.5 Command cd10000 Setup Cluster

Error Code	Message	Comment
TLS1251	TLS1251:E:<node> is already in Ceph Cluster	Given storage node cannot be registered in VSM because it is already in Ceph cluster.
TLS1252	TLS1252:E:Failed to execute vsm-controller	Command vsm-conroller -r controller failed. Look into VSM logs.
TLS1253	TLS1253:E:Failed to generate VSM token	Look into VSM and /var/log/cd10000 logs
TLS1254	TLS1254:E:Failed to set token on node <hostname>	Look into /var/log/cd10000 logs.
TLS1255	TLS1255:E:Failed to start vsm-node on <hostname>	Look into vsm logs on given stoarge node.
TLS1256	TLS1256:E:Service is not running: <service>	<service> is required for vsm.

17.5.6 Command cd10000 Backup and Restoration - fbart

Error Code	Message	Comment
TLS1301	TLS1301:E:Default configuration could not be restored	Default configuration could not be restored.
TLS1302	TLS1302:E:Configuration file could not be created	Unable to create configuration file, check permissions.
TLS1303	TLS1303:E:Configuration file could not be parsed	Invalid or nonexistent configuration file. The user probably manually changed or deleted the configuration file.
TLS1304	TLS1304:E:Directory could not be created	Unable to create a directory. Check permissions.
TLS1305	TLS1305:E:Crontab could not be updated	Unable to update crontab. Check if job is valid.

Error Code	Message	Comment
TLS1306	TLS1306:E:Could not connect to destination	Host name is not valid, or connection impossible. Check network, DNS, firewall ...
TLS1307	TLS1307:E:Synchronisation failed, rsync did not succeed	Rsync has encountered a problem. The exact problem will be in the log file.
TLS1308	TLS1308:E:No backup available for restoration	Host1 and Host2 as currently configured contain no successful backups.

17.5.7 Basking Errors

Error Code	Message	Comment
BSK1001	BSK1001:E: Basking api is not available	Unable to connect to basking api, check if service basking-api is running.
BSK1002	BSK1002:E: Fipmanager api is not available	Unable to connect to fipmanager rest api, check if fipmanager service is running.
BSK1003	BSK1003:E: Wrong parameters provided	Wrong parameters provided in input.
BSK1004	BSK1004:E: Action is not supported by basking api	Not supported action by basking api, check if version of basking api is the same as version of basking.

17.5.8 Firmware Manager Errors

Error Code	Message	Comment
FFM1001	FFM1001:E: <node> is not reachable	<node> is not available in the network. Probably is switched off. **Solution**: switch node on.
FFM1002	FFM1002:E: Upgrading Bios via ASP application fail	ASP package upgrade firmware fail. More information in log file: /tmp/sctmpdir [numbers]/tooltrace.txt
FFM1003	FFM1003:E: Upgrading IRMC via ASP application fail	ASP package upgrade firmware fail. More information in log file: /tmp/sctmpdir [numbers]/tooltrace.txt
FFM1004	FFM1004:E: RPM package firmware-manager-firmwares not installed on node: <node>	RPM package not properly installed.
FFM1005	FFM1005:E: No nodes in cluster	No nodes in fipmanager data base. Turn on all nodes.
FFM1007	FFM1007:E: Wrong firmware configuration file	Wrong configuration file: /opt/fujitsu/firmware-manager/firmwares/firmware.conf

17.6 VSM GUI Messages (E)

The VSM GUI functions do not return a return code specifying whether a function was executed successfully. If a function cannot be executed, then an exception is thrown. If necessary, such exceptions can be caught in a script and processed appropriately. The following exceptions are thrown:

Error Code	Exception	Comment
E1223261	VsmServiceUnavailable	This exception is raised when a VSM service is not responsive, for example it is shut down, or not reachable.
E1232141	VsmHostNotFound	This exception is raised when the host server cannot be found in the VSM DB. For example, you want to shut down host A, but host A is not found in the VSM database.
E1232231	VsmOsdNotFound	This exception is raised when VSM attempts to operate on an OSD that does not exist in VSM. For example, VSM tries to stop osd.100, but osd.100 is not included in the Ceph system.
E0000111	DBError	This exception is raised in response to a database access error. Possible causes include an incorrect privilege level or network error.
E0000122	BadHTTPResponseStatus	This exception is raised when a bad HTTP Response Status is returned. For example, the return of a 1234 response status code will result in an exception, while the return of a 502 or 200 is a legal response status and will not generate an exception.
E0400001	CephConfException	This exception is raised when writing to the Ceph configuration file results in an error, such as caused by a wrong content, wrong format or disk full etc.

Error Code	Exception	Comment
E0000123	DeviceException	General exception for device. Mainly, if an operation for a device failed.
E0000128	DuplicateControllerIPs	This exception is raised when a duplicated Controller IP address is found in the server manifest file - only one controller IP address is allowed in the server.manifest file.
E0300001	ExeCmdError	VSM issues commands to Ceph and various other systems. This exception is raised when an issued command fails.
E0500001	GetNoneError	Exception is raised when an unexpected null response is returned.
E0001225	MDSExists	This exception is raised if VSM attempts to add an already existing MDS. For example, when VSM attempts to add mds.0, but mds.0 is already running in the Ceph system.
E0001224	MonitorException	This exception is raised for operations on a non-valid monitor, for example VSM attempts to operate on mon.11, but mon.11 does not exist.
E0001229	NoValidHost	This exception is raised for operations on a non-valid host, for example VSM attempts to start host A, but host A does not exit in the VSM cluster.
E0400003	AddGlobalToCephConfFailed	This exception is raised when adding information to the `ceph.conf [global]` section failed.
E0400005	AddMdsToCephConfFailed	This exception is raised when adding information to the `ceph.conf [mds]` section failed.
E0400004	AddMonToCephConfFailed	This exception is raised when adding information to the `ceph.conf [mon]` section failed.

Error Code	Exception	Comment
E0400006	AddOsdToCephConfFailed	This exception is raised when adding information to the ceph.conf [osd] section failed.
E0001236	PGExists	This exception is raised when creating a PG in the DB but the PG already exits.
E0400002	LoadCephConfFailed	This exception is raised when reading ceph.conf fails. Possible causes include that the ceph.conf file does not exist, or that the ceph.conf file is incorrectly formatted.
E0100001	PathNotExist	This exception is raised when the path does not exist in the Linux system.
E0001232	PresentOpenStackException	Present pool to OpenStack results in error.
E0001238	RBDExists	This exception is raised when creating an already existing item in DB for RBD.
E0001231	DeviceAddFailed	This exception is raised when adding a device fails.
E0001235	DeviceRemoveFailed	This exception is raised when removing a device fails.
E0600001	StartCephFaild	This exception is raised when starting the Ceph service fails.
E0001233	DeviceStartFailed	This exception is raised when starting a device fails.
E0001234	DeviceStopFailed	This exception is raised when stopping a device fails.
E0001239	StorageGroupException	General exception for storage groups; this exception is raised when operations on a storage group fails.
E0001245	StoragePoolUsageFailure	This exception is raised when getting the storage group usage fails.
E0001241	StorageServerException	General exception for failed operations on a storage node, such as start/stop/remove storage node.

Error Code	Exception	Comment
E0200001	UpdateDBError	This exception is raised when updating DB fails.
E0012241	MonitorAddFailed	This exception is raised when adding a monitor fails.
E0012242	MonitorRemoveFailed	This exception is raised when removing a monitor fails.
E0012243	MonitorStartFailed	This exception is raised when starting a monitor fails.
E0012244	MonitorStopFailed	This exception is raised when stopping a monitor fails.
E0012311	AdminRequired	Admin privilege level is required to perform operation.
E0012324	ConfigNotFound	Cannot find the configuration files for VSM/Ceph.
E0012325	DiskNotFound	Cannot find disk.
E0012326	FileNotFound	Cannot find file.
E0012321	AddPoolToOpenStackFailed	This exception is raised when adding a pool to OpenStack fails.
E0012391	StorageGroupAddFailed	This exception is raised when adding a storage group fails.
E0012411	StorageServerAddFailed	This exception is raised when adding a storage node fails.
E0012412	StorageServerRemoveFailed	This exception is raised when removing a storage node fails.
E0012413	StorageServerStartFailed	This exception is raised when starting astorage node fails.
E0012414	StorageServerStopFailed	This exception is raised when stopping a storage node fails.
E0122326	ServiceUnavailable	Cannot get access to the service, for example, rpc-call to vsm-conductor service while vsm-conductor is shut down.

Error Code	Exception	Comment
E0122327	StoragePoolUsageInvalid	The usage of the storage pool is invalid.
E0122328	UnknownCmd	When VSM attempts to run command, but the command cannot be found in the Linux system.
E0123214	HostNotFound	Cannot find the host maybe storage-node/controller-node.
E0123218	MDSNotFound	Cannot find the mds service.
E0123220	MonitorNotFound	Cannot find the monitor service.
E0123223	OsdNotFound	Cannot find the osd.
E0123224	PGNotFound	This exception is raised when pg was not found.
E0123228	RBDNotFound	This exception is raised when rbd was not found.
E0123231	ServiceNotFound	This exception is raised when service was not found.
E0123235	StoragePoolUsageNotFound	This exception is raised when the usage of storage pool cannot be found in DB.
E0123236	SummaryNotFound	This exception is raised when the summary of the Ceph usage cannot be found in DB
E0123237	VsmStorageGroupNotFound	This exception is raised when a VSM Storage group cannot be found
E0123238	VsmappNotFound	This exception is raised when no OpenStack node is registered in the VSM cluster.

Glossary

A

AISconnect
Advanced Infrastructure Solutions Connect

E

Elasticsearch
Open-source, distributed, real-time search and analytics engine. Elasticsearch is written in Java and runs on OpenJDK.

F

Fluentd
Free open-source log management tool

I

iRMC
Integrated Remote Management Controller

K

KVM
KVM-Switch (Keyboard, Video, Mouse)

M

MAC
Media Access Control

N

NIC
Network Interface Card

NTP
(Network Time Protocol) Protocol to synchronize clocks over a network

O

OpenStack

Free open-source (Apache license) software governed by a non-profit foundation with a mission to produce the ubiquitous Open Source Cloud Computing platform that will meet the needs of public and private clouds regardless of size, by being simple to implement and massively scalable.

P

PowerBuilder (PB)

Devlelopment environment for building client/server applications

Puppet

Configuration Management System, it is used to apply on demand state of node defined in partials.

R

RADOS

"Reliable, autonomic, distributed, object store comprised of self-healing, self-managing, intelligent storage nodes" (http://ceph.com/docs/master/architecture)

RBD

RADOS Block Device

S

SNMP

SNMP is a set of protocols for managing complex networks.

SSH

Secure Shell

Index

D

E

S