FUJITSU

Fujitsu Software
Technical Computing Suite V4.0L20

Job Operation Software Setup Guide

J2UL-2454-02ENZ0(11)
September 2024

Preface

Purpose of This Manual

This manual describes procedures for installing Technical Computing Suite and changing the system configuration after installation.

Intended Readers

This manual is intended for system administrators who operate and manage systems where the Job Operation Software is installed.

The manual assumes readers have the following knowledge:

- Basic Linux knowledge
- Overall understanding of the Job Operation Software, obtained from the "Job Operation Software Overview."
  For details on functions used in the manual, see the "Job Operation Software Administrator's Guide for System Management."

Organization of This Manual

This manual is organized as follows.

Chapter 1 Introduction

This chapter provides an overview of installation work using the Job Operation Software.

Chapter 2 New System Installation

This chapter describes the work of building a new system.

Chapter 3 Adding a Cluster

This chapter describes the work of adding a cluster to an existing system.

Chapter 4 Deleting a Cluster

This chapter describes the work of deleting a cluster from the system.

Chapter 5 Adding a Node

This chapter describes the work of adding a node to an existing cluster.

Chapter 6 Deleting a Node

This chapter describes the work of deleting a node from a cluster.

Chapter 7 Changing Configuration Information

This chapter describes the work of changing various information managed by the Job Operation Software.

Appendix A Configuring GPUs

This appendix describes the work to use the nodes equipped with GPUs.

Appendix B Configuring Singularity

This appendix describes the work to use Singularity.

Appendix C Preparing Subscripted Servers

This appendix describes the servers that are required for Red Hat Enterprise Linux 8 package updates.

Notation Used in This Manual

Representation of Units

The following table lists the prefixes representing units in this manual. Basically, disk size is represented as a power of 10, and memory size is represented as a power of 2. Be careful about specifying sizes when displaying or entering commands.

Prefix	Value	Prefix	Value
K (kilo)	10^3	Ki (kibi)	2^{10}
M (mega)	10^6	Mi (mebi)	2^{20}

Prefix	Value	Prefix	Value
G (giga)	10^9	Gi (gibi)	2^{30}
T (tera)	10^{12}	Ti (tebi)	2^{40}
P (peta)	10^{15}	Pi (pebi)	2^{50}

Notation of Model Names

In this manual, the computer that based on Fujitsu A64FX CPU is abbreviated as "FX server", and FUJITSU server PRIMERGY as "PRIMERGY server" (or simply "PRIMERGY").

Also, specifications of some of the functions described in the manual are different depending on the target model. In the description of such a function, the target model is represented by its abbreviation as follows:

[FX]: The description applies to FX servers.
[PG]: The description applies to PRIMERGY servers.

Administrators

The Job Operation Software has different types of administrator: system administrator, cluster administrator, and job operation administrator. Unless otherwise noted, the descriptions in this manual apply to functions for system administrators and cluster administrators. For this reason, the term "administrator" usually means a system administrator and a cluster administrator.

Prompts in Command Input Examples

means that the command must be executed with administrator privileges (superuser) for the OS.

Path Names of Commands

In the examples of operations, commands in the /bin, /usr/bin, /sbin, or /usr/sbin directory may not be represented by absolute paths.

Symbols in This Manual

This manual uses the following symbols.

Note

The Note symbol indicates an item requiring special care. Be sure to read these items.

See

The See symbol indicates the reference source of detailed information.

Information

The Information symbol indicates a reference note related to the Job Operation Software.

Trademarks

- Linux(R) is the registered trademark of Linus Torvalds in the U.S. and other countries.
- Red Hat and Red Hat Enterprise Linux are registered trademarks of Red Hat, Inc. in the U.S. and other countries.
- Microsoft, Windows are either registered trademarks or trademarks of Microsoft Corporation in the U.S. and other countries.
- Other company and product names in this manual are trademarks or registered trademarks of their respective owners.

Export Controls

Exportation/release of this document may require necessary procedures in accordance with the regulations of your resident country and/or US export control laws.

Date of Publication and Version

Version	Manual Code
September 2024, Version 2.11	J2UL-2454-02ENZ0(11)
March 2023, Version 2.10	J2UL-2454-02ENZ0(10)
September 2022, Version 2.9	J2UL-2454-02ENZ0(09)
March 2022, Version 2.8	J2UL-2454-02ENZ0(08)
March 2022, Version 2.8	J2UL-2454-02ENZ0(08)
November 2021, Version 2.7	J2UL-2454-02ENZ0(07)
August 2021, Version 2.6	J2UL-2454-02ENZ0(06)
March 2021, Version 2.5	J2UL-2454-02ENZ0(05)
January 2021, Version 2.4	J2UL-2454-02ENZ0(04)
December 2020, Version 2.3	J2UL-2454-02ENZ0(03)
September 2020, Version 2.2	J2UL-2454-02ENZ0(02)
June 2020, Version 2.1	J2UL-2454-02ENZ0(01)
March 2020, Second version	J2UL-2454-02ENZ0(00)
January 2020, First version	J2UL-2454-01ENZ0(00)

Copyright

Update history

Changes	Location	Version
Added description of NVIDIA H100 for the PRIMERGY compute node.	2.3.12.6 Appendix A	2.11
Added a procedure for recovering from a configuration file synchronization failure.	2.3.10.5	2.10
Added a procedure for executing synchronization processing with the pxinst command after executing the command for applying cluster configuration.	2.3.11 4.8 5.8 6.8 7.1.4	
Improved procedure for adding nodes.	5.8 5.11	
Added procedures for creating and modifying PRIMERGY server resource units.	2.1.1.8 Figure 2.3 in 2.3 2.3.3 2.3.5 Figure 4.1 in Chapter 4 4.4 4.5 4.7 Figure 5.1 in Chapter 5 5.3 5.5	2.9

Changes	Location	Version
	Figure 6.1 in Chapter 6 6.4 6.5 6.7 7.1.1	
Added a note registering cluster configuration information.	2.3.4 2.3.11	
Added example deletes nodes.	6.6	
Corrected errors.	-	
Added a procedure to restart services on the system management node and the compute cluster sub management node, when changing the system configuration.	2.3.5 5.9.2 5.9.3 6.6.2	2.8
Corrected errors.	-	
Added resource unit description.	2.1.1.8	2.7
Improved description of procedure to prepare rootfs images for building FX servers.	2.3.7.2.4	
Corrected errors.	2.4.2.2	
Added a note about instructions for installing the FX server in step-by-step.	2.3.8.1	2.6
Added description of NVIDIA A100 for the PRIMERGY compute node.	2.3.10.6 Appendix A	2.5
Added description of servers required for Red Hat Enterprise Linux 8 package updates.	Appendix C	
Added the instructions for using a system management node, a compute cluster management node, and a login node on one node.	2.1.1.2 2.1.1.3 2.1.1.7 2.1.2.1 2.2.2.3 2.2.3 2.3.5 2.4.3.2 2.4.4.4 2.4.5 5.1.2 5.9.1 5.9.2 6.1.2 6.6.1 6.6.2 Chapter 7	2.4
Added instructions for using OS versions RHEL8 for peripherals.	2.2.1.2 2.2.2.1 2.4.2.1 2.4.3.1 2.4.3.2	2.3
Improved description of procedure to prepare rootfs images for building FX servers.	2.3.7.2	
Added a note to the OS installation completion confirmation procedure when building a cluster.	2.3.8.5	
Added a package to remove for system noise reduction.	2.4.1.4	

Changes	Location	Version
Added a required package for the integrated development environment for the Technical Computing Suite Development Studio.	2.4.2.1	
Added a procedure for package removal for system noise reduction.	2.4.1.4	2.2
Changed a procedure for removing a cluster.	4.5	
Added package required for installation function.	2.2.2.1	2.1
Improved description of procedure to prepare rootfs images for building FX servers.	2.3.7.2.4	
Added a description of MariaDB autostart configuration.	2.4.3.2	
Added the stopping ranges of operation when adding or deleting multiuse nodes.	Chapter 5 Chapter 6	
Added a procedure for starting up a diskless node before reflecting the cluster configuration information when adding an FX server.	5.5.4	
Added methods of configuring settings to use GPUs.	2.3.1 2.3.10.6 Appendix A	2
Added methods for installing Singularity.	2.3.1 2.3.10.7 Appendix B	
Changed the look according to product upgrades.	-	

Contents

Chapter 1 Introduction

This manual describes procedures for new installation of a system using the Technical Computing Suite and for configuration changes to the system while in operation.

1.1 Usage Scenarios in This Manual

The procedures described in this manual assume the following scenarios.
In this manual, a system means the cluster groups as a whole managed by a set of system management nodes.
For each usage scenario described in the following table, see the chapters shown under "Chapter Reference."

Table 1.1 Usage Scenarios

Category	Category Details	Description	Chapter Reference
New system installation (Installation phase)	Batch installation	Install a new system. Build all the nodes together during the installation.	Chapter 2 New System Installation
	Step-by-step installation	Build nodes in a step-by-step manner according to the hardware delivery schedule, etc. The new system installation procedures cover only those nodes that are ready to be built. Build other nodes upon delivery by following the procedures in "Chapter 5 Adding a Node".(*)	Chapter 2 New System Installation
System configuration change (Maintenance phase)	Adding a cluster	Add a cluster to an existing system.	Chapter 3 Adding a Cluster
	Deleting a cluster	Delete a cluster from an existing system.	Chapter 4 Deleting a Cluster
	Adding a node	Add a node to an existing cluster.	Chapter 5 Adding a Node
	Deleting a node	Delete a node from an existing cluster.	Chapter 6 Deleting a Node
	Changing configuration information	Change configuration information such as the IP address of a node belonging to an existing cluster.	Chapter 7 Changing Configuration Information

(*) During new installation, the following conditions must be satisfied for step-by-step system installation.

Table 1.2 Step-by-Step System Installation Conditions

Use Case	Condition
Initial system installation	Can build the active system management node
New installation on a cluster containing a compute cluster	Can build the active compute cluster management node
New installation on a cluster containing a storage cluster	Can build the active storage cluster management node
New installation targeting nodes (e.g., FX server and PRIMERGY server compute nodes) belonging to the same node group as the compute cluster sub management node	Can build the active compute cluster sub management node

1.2 Notes

Abbreviations such as "[SMM]" in the work procedures in this manual mean the nodes where the work will be done.

Table 1.3 Nodes Used for Work

Abbreviation	Meaning
[SMM]	Work on the system management node. Unless otherwise stated, this means the active system management node.
[CCM]	Work on the compute cluster management node. Unless otherwise stated, this means the active compute cluster management node.
[PC]	Work on a PC running Windows.
[LN]	Work on the login node.

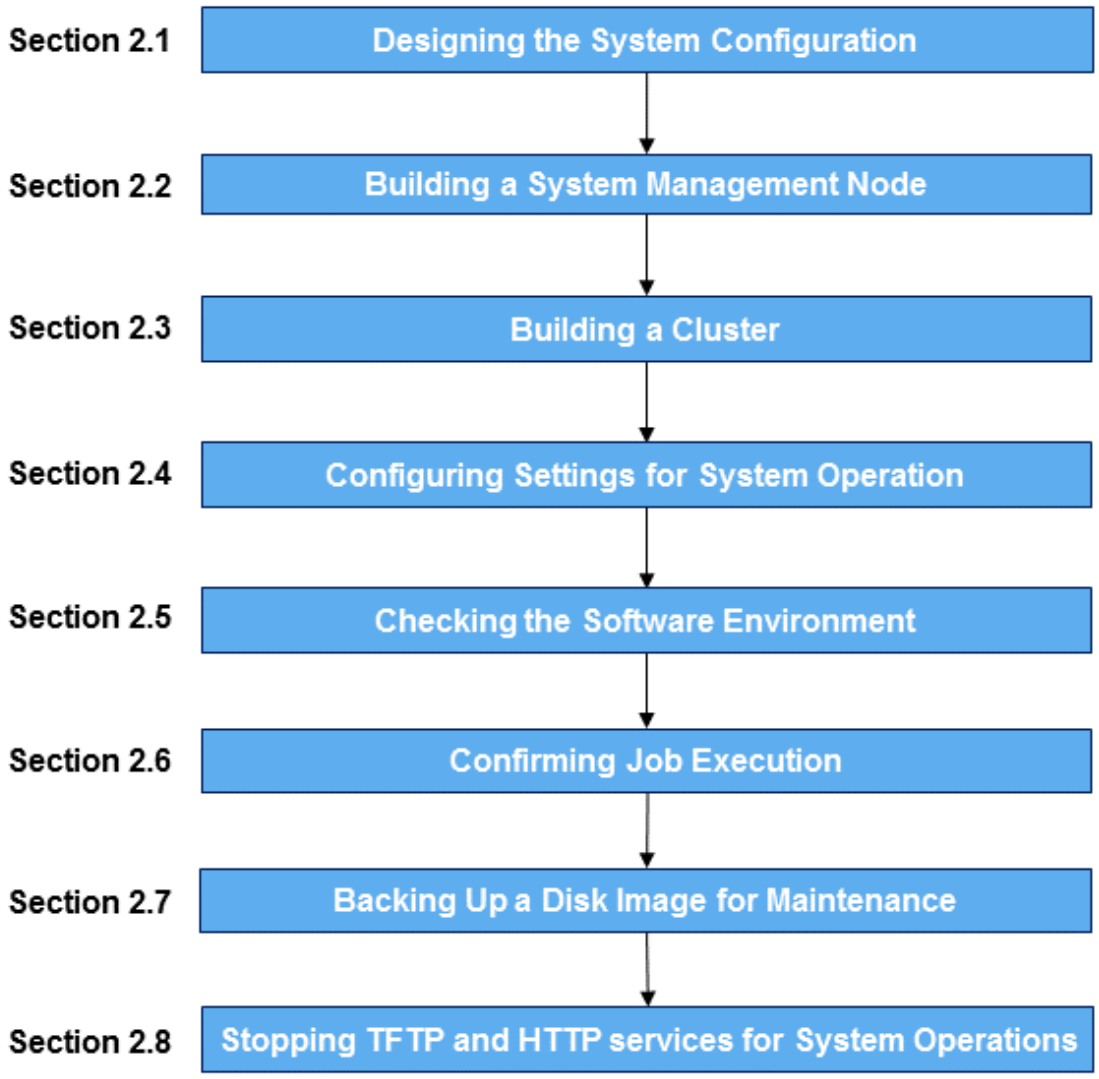
Chapter 2 New System Installation

This chapter describes procedures to install a system using the Job Operation Software.

Detailed descriptions about installing or connecting hardware such as the main unit and peripherals are beyond the scope of this manual. To perform that work, see the hardware manuals. The settings for all required hardware must be done prior to the procedures in "2.3 Building a Cluster." For details on settings for the required hardware, see "2.3.1 Prerequisites to Installation."

New system installation uses the following workflow.

Figure 2.1 Workflow for New System Installation



2.1 Designing the System Configuration

2.1.1 Considering Cluster Configurations

Consider the roles of nodes and the node grouping configurations in the system.

2.1.1.1 Components

The Job Operation Software manages the system in units of clusters. For details on cluster types and the roles of nodes composing a cluster, see the "Job Operation Software Overview."

2.1.1.2 Redundant Configuration

The Job Operation Software supports node redundancy for higher availability against node failures and problems. Consider node redundancy as required. The Job Operation Software supports the following redundant configurations of nodes.

Also, a node serving as all of system management node, compute cluster management node, and login node cannot be redundant.

Table 2.1 Node Redundancy

Node Type	Supported Redundant Configuration
System management node	Active and standby node configuration Nodes of each node type are paired for redundancy. Normally, one of them is running as the active node. The detection of a fault in the active node causes a switchover to the standby node to continue services.
Compute cluster management node	
Compute cluster sub management node	
Storage cluster management node	
MGS node	Mutual standby configuration Paired nodes work independently, with each running as a node where the Job Operation Software services are operating, instead of only one node operating and the other node on standby. As a result, even when one of the nodes becomes faulty, the services can continue on only the other node.
MDS node	
OSS node	
Compute node also serving as global I/O node (FX server)	When a compute node that also serves as a global I/O node fails, operation continues on another compute node that also serves as a global I/O node within the same rack.

The following table lists effects of non-redundancy of the above-described nodes.

Table 2.2 Effects of Non-Redundancy of Nodes

Node Type	Effect of Non-Redundancy
System management node	If a fault is detected, it will no longer be possible to control (turn on or off) system power and check the run state of nodes in the system.
Compute cluster management node	If a fault is detected, it will no longer be possible to execute new jobs. However, running jobs continue.
Compute cluster sub management node	If a fault is detected, it will no longer be possible to use compute nodes in the node group to execute new jobs. However, running jobs continue.
Storage cluster management node	If a fault is detected, it will no longer be possible to check the run state of nodes in the storage cluster.
MGS node	If a fault is detected, it will no longer be possible to use the shared file system (FEFS: Fujitsu Exabyte File System). For details on redundancy for these nodes, see the *FEFS User's Guide*.
MDS node	
OSS node	
Compute node also serving as global I/O node (FX server)	

For details on the behavior when a node is faulty, see "Details of the System Monitoring Function" in the "Job Operation Software Administrator's Guide for System Management."

2.1.1.3 Node Types for Multiple Purposes

A single node can serve multiple purposes in the following combinations of node types in the Job Operation Software:

- System management node + compute cluster management node + login node

- System management node + storage cluster management node
- MGS node + MDS node + OSS node
- MGS node + MDS node
- Compute node + boot I/O node (FX server)
- Compute node + storage I/O node (FX server)
- Compute node + global I/O node (FX server)

Information

With the FX server, the boot I/O node, storage I/O node, and global I/O node also always serve as compute nodes.

2.1.1.4 Node OS

The Job Operation Software supports configurations where a mix of nodes in a cluster has different OS versions.
The Job Operation Software supports a range of OS versions. You can freely integrate any of them in your configurations.
For details on the supported OS versions, see the *Software Description*.

2.1.1.5 Node Model Name

The Job Operation Software can control the power of a node in the system. However, the method of power control may vary depending on the type of machine used as the node. To manage nodes that have varying power control methods, the system administrator uses the configuration management function to manage the nodes under different model names.

The Job Operation Software uses the following model names.

Table 2.3 Model Names

Model Name	Meaning
PG	PRIMERGY RX Series
FT	FX server
Other	Arbitrary model (*)

(*) To use a model other than the PRIMERGY Series and the FX server, the system administrator defines a model name using two uppercase alphabetic characters or two single-byte numbers. The model is called an arbitrary model.

Information

- Model names are written in the configuration file of the cluster configuration management function as described in "2.3.4 Registering Cluster Configuration Information [SMM]." You can control the power of nodes of arbitrary models by configuring the system control function. For details on configuring the system control function, see "Configuring the System Control Function" in the "Job Operation Software Administrator's Guide for System Management."
- Nodes having the following functions can be defined with the model name "PG" if they are not included in the PRIMERGY RX Series:
  - Architecture: x86_64
  - OS: Red Hat Enterprise Linux or CentOS
  - IPMI 2.0 compliant
    It must be possible to use an NMI interrupt complaint with IPMI to collect a dump file as an investigation materials for a detected faulty node. For the availability of this function, see the hardware manual or contact your hardware vendor.
- The Job Operation Software supports virtualization of a login node using a KVM, etc. To use a login node in a virtualized environment, the system administrator defines the target virtualized node as a node of an arbitrary model.

2.1.1.6 Criteria for Cluster Configuration Estimates

The Job Operation Software manages a cluster configuration by organizing the nodes into a hierarchy or grouping the nodes. The following table shows guidelines on the number of nodes for each hierarchy or group.

Table 2.4 Criteria for Cluster Configuration Estimates

Additional Node	Node Installation Criteria
System management node	To divide a network by cluster for the purpose of enhancing security, install a system management node in each cluster.
Compute cluster sub management node	With the FX server, if the number of BoBs exceeds 252, install a compute cluster sub management node to distribute load. If the number of BoBs is equal to or less than 252, a compute cluster sub management node does not need to be installed. If the number of PRIMERGY server is 1,024 or more, a compute cluster sub management node is required. For a redundant configuration of the compute cluster sub management node, the pair of this node for each node group must have one active and one on standby. If your FX server has more than 252 BoBs or a PRIMERGY server has more than 1024 compute nodes, install a compute cluster sub management node and define node groups to balance the load on the nodes efficiently. A node group is defined by 1 (2 for redundancy) compute cluster sub management node and up to 252 BoB or 1024 PRIMERGY servers.
Compute node	When you build a compute cluster, you must have a compute node for the FX server or the PRIMERGY server. Be sure to define compute nodes in the configuration design, even if the compute node does not exist at the time of the build, such as in a phased installation.

When installing a compute cluster sub management node, use the following rules to define node groups by dividing the system so that the number of compute nodes under one compute cluster sub management node is evenly divided.

- The model names of the compute nodes in each node group must be the same (see "2.1.1.5 Node Model Name").
- The compute nodes with the same model name in a compute cluster cannot consist of both compute nodes with a node group and compute nodes without a node group.

FX server

If the number of BoBs exceeds 252, define node groups. You will need to consider the following when defining an FX server node group.

- Under a configuration with the Tofu interconnect D as a communication unit made up of 3 BoBs, the number of BoBs composing one node group must be a multiple of three. Also, the number of BoBs that can be installed in a single node group is 252 or less.
- Considering the connections for network switches and various wiring, set node groups in the following units depending on the Z-axis width of the system.

Table 2.5 Criteria for BoB Quantity Estimates

Z-Axis Width	Unit	Description	Maximum Number of BoBs per Node Group
4	-	Do not define node groups for a configuration with a Z-axis width of 4, because the quantity does not fill 1 rack (less than 252 BoBs).	-
8	1 rack	Set up to 252 BoBs, in multiples of 24 per node group since the unit used is 1 rack (24 BoBs).	240
16	2 racks	Set up to 252 BoBs, in multiples of 48 per node group since the unit used is 2 racks (48 BoBs).	240
24	3 racks	Set up to 252 BoBs, in multiples of 72 per node group since the unit used is 3 racks (72 BoBs).	216

The method of calculation to determine the number of node groups in a system using one of the units shown above is as follows.

1. Let z be the maximum number of BoBs per node group, and S be the total number of BoBs in the system.
2. Assuming the total number of node groups is N, then calculate N by dividing S by z. Round up the remainder.
   If S is divisible, define N node groups having z boot groups. Further calculation is not required.
   If S is not divisible, proceed to step 3.
3. Calculate n by dividing S by N.
   If n is a multiple of 24, define N node groups having n boot groups. Further calculation is not required.
   If n is not a multiple of 24, proceed to step 4.
4. Let X be the unit for the number of BoBs composing a node group, depending on the Z-axis width of the system.
5. Align n to a multiple of X. Assuming s0 is the nearest multiple of X less than n and s1 is the nearest multiple of X greater than n, then let s0 and s1 each be the number of BoBs in a node group.
6. Use the following formula to find n0, the number of node groups where the number of BoBs is s0:
   n0 = (s1 x N - S)/X
7. Use the following formula to find n1, the number of node groups where the number of BoBs is s1:
   n1 = N - n0

The following table shows examples that apply the following three patterns to the above calculations:

- X,Y,Z is 12,12,8 (36 racks)
- X,Y,0 is 10,4,16 (20 racks)
- X,Y,Z is 14,14,24 (147 racks)

Table 2.6 Examples of Node Group Quantity Calculations

Variable	Description	(X,Y,Z)=(12,12,8)	(X,Y,Z)=(10,4,16)	(X,Y,Z)=(14,14,24)
S	Total number of BoBs in the system	864	480	3528
z	Maximum number of BoBs per node group	240	240	216
N	Total number of node groups S/z (Remainder rounded up)	4 (Calculation ended with divisible number)	2 (Calculation ended with divisible number)	17
n	S/N	216 (Calculation ended with divisible number)	-	207.52...
X	Unit for the number of BoBs composing a node group	-	-	72 (3 racks)
s0	Nearest multiple of X less than n, where the number of BoBs in a node group is 1	-	-	144
s1	Nearest multiple of X greater than n, where the number of BoBs in a node group is 2	-	-	216
n0	Out of N groups, the number of node groups having s0 BoBs is (s1 x N - S)/X	-	-	2
n1	Out of N groups, the number of node groups having s1 BoBs is N - n0	-	-	15

PRIMERGY server

For a system containing 1,024 or more PRIMERGY server, set node groups.
To distribute the network load in the system, set the number of compute nodes as evenly as possible within each node group. The following method of calculation shows how to evenly distribute PRIMERGY server compute nodes to node groups in the system.

1. Let z be the maximum number of nodes that can be installed in a single node group, and S be the total number of PRIMERGY server in the system.
2. Assuming the total number of node groups is N, then calculate N by dividing S by z. Round up the remainder.
   If S is divisible, define N node groups having z nodes. Further calculation is not required.
   If S is not divisible, proceed to step 3.
3. Calculate n by dividing S by N, where the remainder is p.
   If there is no remainder, define N node groups having n nodes. Further calculation is not required.
4. In p node groups, the number of nodes per node group is n+1. In N - p node groups, the number of nodes per node group is n.

2.1.1.7 Node serving as all of system management node, compute cluster management node, and login node considerations

- About configurations

  You can only build single compute cluster configurations with the following node configurations:

  - Node serving as all of system management node, compute cluster management node, and login node (Not Redundant)
  - PRIMERGY compute node (up to 128)
  - Storage cluster cannot be created, and shared file system used in job execution support NFS only.

- About monitoring

  For details service monitoring of job operation software on the node serving as all of system management node, compute cluster management node, and login node, see "Software Abnormality Monitoring" in the "Job Operation Software Administrator's Guide for System Management."

- About rebooting the node

  For details on rebooting the node serving as all of system management node, compute cluster management node, and login node, see "How to Restart the Node Serving as All of System Management Node, Compute Cluster Management Node, and Login Node" in the "Job Operation Software Troubleshooting."

- About job operation

  The KVM mode is not available for the system with a node serving as all of system management node, compute cluster management node, and login node.

2.1.1.8 Resource units

Resource units are created automatically.

You can confirm resource units by specifying the --rscunit option to pashowclst on the system management node.

```
# pashowclst -c clstname --rscunit
[ CLST: clstname ]
RSCUNIT           RUNNING  STOPPED  ERROR    DISABLE
rscunit_ft01      48       0        0        0
```

clstname: Cluster name

- Normally, do not modify resource units.
  Resource unit changes are:

  - Renaming Resource Unit

  - Changing Resource Unit Scope
  - Dividing Resource Units
- If you want to add a compute node with a different configuration (Model, number of CPU cores, CPU frequency, memory capacity, etc.) on the PRIMERGY server, follow the procedure to divide the resource unit for each compute node with a different configuration.
  You cannot change the FX server resource unit.
  If you change a resource unit, you must stop all cluster operations when you change the system configuration.
  System configuration changes include:
  - 2.3 Building a Cluster
  - Chapter 3 Adding a Cluster
  - Chapter 4 Deleting a Cluster
  - Chapter 5 Adding a Node
  - Chapter 6 Deleting a Node
  - 7.1 Changing Cluster Configuration Information
- The "Registering Cluster Configuration Information" operation automatically creates the resource unit and returns to the default configuration.
  If you have changed the resource unit, check the parscunit.conf configuration file for the resource unit.
- If the resource unit configuration after changing the resource unit differs from the settings in the parscunit.conf file, restore and edit the parscunit.conf file from a backup of the resource unit information.
  If you perform the "Registering Cluster Configuration Information" operation without backing up and restoring resource unit information, the submitted job information is deleted.

2.1.2 Considering File System Configurations

Considering the file system configuration.

2.1.2.1 Design of a Shared File System Used in Job Execution

With the Job Operation Software, the storage cluster has the role of providing a shared file system. The compute cluster mounts the shared file system provided by the storage cluster so that compute nodes, login nodes, and management nodes can have the necessary access to the files (such as programs) required for job execution.

The following configuration is required when using the Job Operation Software.

- The compute cluster management node and login node mount all the file systems shared in the cluster.
- The global I/O nodes (FX server) and compute nodes mount all the shared file systems used in the resource unit.

See the "FEFS User's Guide," and design a shared file system.

Information

If the Job Operation System you are building uses the KVM mode of the job execution environment, the home directory of the user executing the job must be created on a shared file system. This should be taken into account when designing shared file systems for use in job execution.
The KVM mode is not available for the system with a node serving as all of system management node, compute cluster management node, and login node.

2.1.2.2 Design of a Shared File System for a Redundant Configuration

During node switching of a node redundantly configured with an active node and standby node, specific file systems may have to be remounted on the active node. The Job Operation Software uses the directory /var/opt/FJSVtcs/shared_disk for shared file systems on redundant system managemant nodes and redundant compute cluster management nodes.

The area under the /var/opt/FJSVtcs/shared_disk directory is referred to as "shared_disk" below.

- System Management Node

  If you want the system management nodes to be redundant, you can set shared_disk to be the area where logs and dump files are stored so that you can see the shared_disk information on the active node when nodes failover.

  However, shared_disk is not strictly required for redundant system management nodes, and should be provided if you want to use a common file system for storing logs and dump files.

- Compute Cluster Management Node

  The compute cluster management node uses shared_disk as an area to store management information about job operations.

  When the compute cluster management node is redundant, shared_disk is required to locate internal information about job operation.

Information

For the system management node or compute cluster management node in a redundant configuration, the information shared in redundant nodes must be accessible only to the active node. Therefore, place the information in a shared_disk that is mounted and used only on the active node. Also, during node switching, it must be possible to switch the mounted shared_disk from the old active node to the new active node.

The system management function supports that operation by automatically switching the mounted shared_disk during node switching. Set the required mounting information in advance for each of the nodes in a redundant configuration as described in "2.3.4 Registering Cluster Configuration Information [SMM]." Then, when a node becomes the active node, this function automatically mounts the shared_disk on the node.

When designing a system, determine the device names of disks to mount to the /var/opt/FJSVtcs/shared_disk mount points of the system management node and compute cluster management node.

You can also use this function where considerations about an area mounted only on the active node are necessary for a system-specific operation tool. In this case too, determine the disk device names and the mount points in advance.

Note

A shared_disk is a file system prepared separately from the file system shared by the storage cluster described in the previous section. Only the nodes in a redundant configuration can share the shared_disk.

2.1.2.3 Designing a File System for the Job Execution Environment Customization Function

The Docker mode of the job execution environment-customizing function in the Job Operation Software creates containers with container images stored (docker import/pull) in the system volume (/var/lib/docker) of each compute node when the job is executed. Therefore, you must provide enough disk space to store the container images.
In an operation using the SDI specification, the container images are pre-stored on the system volume of each compute node, so the total size of the container images should be considered as the size of the system volume for always use.
In an operation using the UDI specification, the container image stored in a dedicated disk area separate from the system volume is stored in the system volume only when the job is executed (When the job is finished, the container image stored on the system volume is deleted.). Therefore, the UDI specification must consider the size of the system volume in anticipation of a momentary increase in system volume usage.
When running multiple UDI specification jobs in a node or UDI specification jobs in multiple nodes in the same BOB in Docker mode, if it is assumed that there is not enough space in the system volume, take measures such as allowing only node-exclusive jobs to be executed using the job ACL function.

Note

The system volume size is fixed on the FX server and cannot be expanded. Therefore, the total size of available container images is limited to 10 GB per node. The size of the container image is the sum of the size of the container image that is stored in advance by SDI specification and the size of the container image that is stored only at job execution by UDI specification. (For example, if the total size of the pre-stored container images in the SDI specification is 7 GB, the total size of the container images available in the UDI specification is up to 3 GB.)

See

For details on the job execution environment customization function, see "Job Operation Software Administrator's Guide for Job Management."

2.1.3 Considering Network Configurations

2.1.3.1 Network Types

The Job Operation Software uses the following networks to connect nodes.

Table 2.7 Network Types

Term	Description
Control network	Network for node control to power on/off nodes, reset nodes, etc. The IP addresses that you configure as the control network include: - Control network IP address (ControlNet) - Control network (software communication) IP address (ControlNetSrv)
Management network	Network used from the OS on each node, in service status monitoring, and in communications for the operation of each function of the Job Operation Software. It is also used as a provisioning network at system build. (The network does not include the NFS diskless nodes of the FX server.) The IP addresses that you configure as the management network include: - Management network IP address (ManageNet) - Representative IP address for the management network (ManageNetMaster)
I/O network	High-speed network for exchanging input/output data with a shared file system between nodes
Compute network	High-speed network used by parallel programs like MPI programs, which are user programs, to communicate between nodes. For the FX server, this applies to the Tofu interconnect D.

Each node type determines the type of network to be assigned.
When considering and deciding the network configuration, check the following to see if there are any omissions in the network information that needs to be assigned for each node type.

Table 2.8 Network type and corresponding node type

Name	Displaying with the pashowclst Command	MAC Address Preparation	Node Type
Representative IP address for the management network	ManageNetMaster	Not Required	SMM,CCM,CCS,SCM Required only if the above nodes are redundant (*1)
Management network IP address	ManageNet	Required	All nodes
Control network (software communication) IP address	ControlNetSrv	Required	SMM
Control network IP address	ControlNet	Not Required	All nodes (*2)
I/O network	-	Not Required	SMM,CCM,GIO,CN(PG),LN, SCM,MGS,MDS,OSS,Multiuse node (*3)
Compute network (tofu0)	TofuNet	Not Required	BIO,SIO,GIO,CN(FX)
Compute network (tofu1)	-	Not Required	BIO,SIO,GIO,CN(FX)

(*1)Provide one IP address for each redundant pair of nodes.
(*2)For BIO, SIO, GIO, CN(FX), this is the IP address of the BMC that manages this BoB.
(*3)On a multiuse node, it is only required when using FEFS.

See

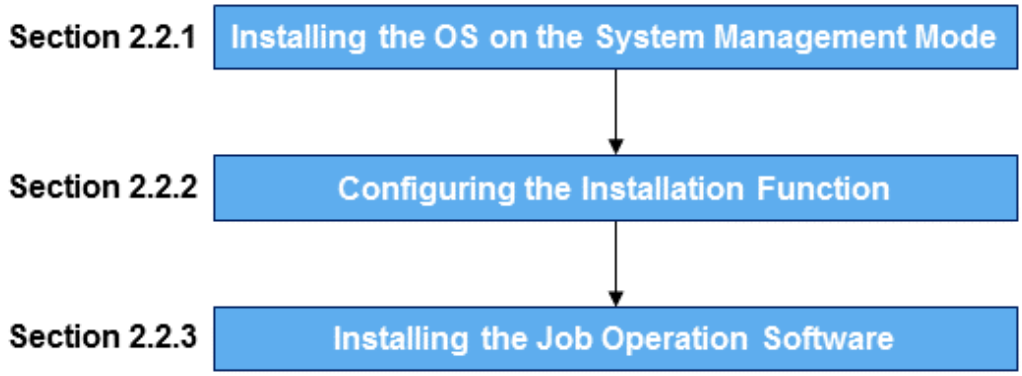
"Provisioning" in terms of the installation function means installing the OS and configuring environment settings on a node. For details on the method of provisioning by the installation function, see "Provisioning Method" in the "Job Operation Software Administrator's Guide for System Management."

2.2 Building a System Management Node

Perform these procedures to build a new system management node. The node built as described in this section is the active system management node. The standby system management node in a redundant configuration is built separately at the same time as nodes of types other than system management node.

The workflow for building a system management node is shown below.

Figure 2.2 Workflow for Building a System Management Node



2.2.1 Installing the OS on the System Management Mode

Install the OS on the system management node. Use ServerView Installation Manager (called SVIM, below), which is the installation function provided by ServerView Suite, to install the OS.

Note

Use of SVIM is only necessary in OS installation on the PRIMERGY server. This is because a driver included in SVIM is required for OS installation on the PRIMERGY server. When installing the OS on third-party hardware, see the manuals provided by the hardware vendor.

2.2.1.1 Installation Using SVIM [SMM]

For the procedure using SVIM to install the OS, see the *ServerView Suite Installation Manager* manual supplied with ServerView Suite.

Note

- Specify and set guide mode since customization of the disk partition configuration, etc. is required.
- In network settings, set the IP addresses of all the network devices used.
- If the system clock does not use UTC, uncheck the [System clock uses UTC] check box.
- For a Japanese language environment, click [Select Applications] -> [Software Packages for JAPAN] on the screen for SVIM application settings. You will need to uncheck the [Fujitsu Linux Support Package] check box.

- Check or uncheck the check boxes of the supplied software, such as ServerView Agent, as required.
- Select the minimum installation for the package configuration.

The following table shows the conditions of a required OS setting for using the installation function.

Table 2.9 OS Setting Conditions

Item	Setting Condition
Disk capacity	Reserve the following disk space on the system management node serving as an installer node, in order to allocate repository and installation resources: at least 10 GB x number of OS versions used in the entire system. Set the resource allocation location separately.

2.2.1.2 Performing the Work After OS Installation [SMM]

After completing OS installation, perform the following work.

1. Confirming linkup of network devices

   Confirm that the network devices used for the management network, control network, and other networks are enabled.
   At this time, configure any network that was not configured during OS installation. Also configure settings for all the network devices used. For details on the setting method, see the manual of the OS used.

2. Enabling shadow passwords and setting a password encryption method

   If the system management node is using RHEL8 or later OS version, this procedure is unnecessary.

   a. Checking settings

      Execute the following command to check settings.

```
# authconfig --test
...
pam_unix is always enabled
 shadow passwords are enabled                    (*1)
 password hashing algorithm is sha512            (*2)
...
```

      (*1) Confirm that shadow passwords are enabled.
      (*2) Check the encryption method.

      If shadow passwords are disabled, or if you need to change the encryption method, perform step b for these settings.

   b. Changing settings

      Enable shadow passwords and set an encryption method as required.
      The following example enables shadow passwords and sets the encryption method to sha512.

```
# authconfig --update --enableshadow --passalgo=sha512
```

      See

      For details on the setting method, see the manual of the OS used.

3. Disabling SELinux

   If SELinux is enabled, edit /etc/sysconfig/selinux to disable it.
   The setting method may vary with the OS version used.

4. Stopping a firewall

   If a firewall (such as firewalld) is running, stop its service to disable it.
   The setting method may vary with the OS version used.

5. Setting the default gateway

   Edit the NetworkManager function, /etc/sysconfig/network, and /etc/sysconfig/network-scripts/route-device as required to set the default gateway.
   The setting method may vary with the OS version used.

6. UDP Receive Buffer Settings

   1. Creating a Configuration File

      Create a kernel parameter configuration file. You can use arbitrary name for the file.
      The following example creates a file with the file name udp.conf:

      ```
      # vi /etc/sysctl.d/udp.conf
      net.core.rmem_default=8388608
      net.core.rmem_max=8388608
      ```

   2. Reflecting settings

      ```
      # sysctl -p /etc/sysctl.d/udp.conf
      net.core.rmem_default = 8388608
      net.core.rmem_max = 8388608
      ```

   3. Checking the result of reflecting settings

      ```
      # cat /proc/sys/net/core/rmem_default
      8388608
      # cat /proc/sys/net/core/rmem_max
      8388608
      ```

2.2.2 Configuring the Installation Function

2.2.2.1 Installing the Installation Function [SMM]

Apply the installation function packages included on the Technical Computing Suite media to the system management node. To apply the installation function packages, the dependent OS package must have been applied in advance.

Information

This procedure uses the yum command as an example. To use another package manager, apply the required package according to the instructions for the package manager used.

1. Applying the dependent OS package

   Apply the OS package to the system management node in the following procedure.

   a. Mounting the OS DVD or ISO image file

      Mount the OS DVD or ISO image file.

      [Using the DVD]

      ```
      # mount /dev/cdrom /mnt
      ```

      [Using the ISO image file]

      ```
      # mount -o loop filepath /mnt
      ```

      filepath: Path to the OS ISO image file

   b. Setting the yum repository

      Set the yum repository.

      [The system management node is using RHEL7]

```
# vi /etc/yum.repos.d/rhel-dvd.repo
[rhel-dvd]
name=Install Media
baseurl=file:///mnt
enabled=0
gpgcheck=0
```

[The system management node is using RHEL8]

```
# vi /etc/yum.repos.d/rhel-dvd.repo
[rhel-AppStream]
name=Install Media
baseurl=file:///mnt/AppStream
enabled=0
gpgcheck=0

[rhel-BaseOS]
name=Install Media
baseurl=file:///mnt/BaseOS
enabled=0
gpgcheck=0
```

c. Deleting cached information

Delete the cached yum information in advance.

[The system management node is using RHEL7]

```
# yum --disablerepo=\* --enablerepo=rhel-dvd clean all
```

[The system management node is using RHEL8]

```
# yum --disablerepo=\* --enablerepo=rhel-AppStream,rhel-BaseOS clean all
```

d. Installing the package

The packages that must be installed depend on the OS version of the system managed node.
Use the yum command to install the package.

[The system management node is using RHEL7]

```
# yum --disablerepo=\* --enablerepo=rhel-dvd install \
createrepo dhcp httpd xinetd tftp-server nfs-utils \
python PyYAML python-jinja2 python-six \
openssl openssh cpio ipmitool rsync chrony python-netaddr \
expect
```

[The system management node is using RHEL8]

```
# yum --disablerepo=\* --enablerepo==rhel-AppStream,rhel-BaseOS install \
createrepo_c createrepo dhcp-server httpd xinetd tftp-server nfs-utils \
python36 python3-pyyaml python3-jinja2 python3-six \
tar openssl openssh cpio ipmitool rsync chrony expect
```

Information

Skipping the packages already installed at the time of OS installation is not a problem.

e. Unmounting the DVD or ISO image file

```
# umount /mnt
```

f. Removing the DVD

Remove the DVD, if any, used for OS installation.

```
# eject -rv
```

2. Installing the installation function

   a. Mounting the Technical Computing Suite DVD

   ```
   # mount /dev/cdrom /mnt
   ```

   b. Applying the FJSVpxinst package

   ```
   # rpm -ivh /mnt/MODULES2/JOS/FX/PG/FJSVpxinst-version-release.noarch.rpm
   ```

   c. Unmounting the Technical Computing Suite DVD

   ```
   # umount /mnt
   ```

   d. Removing the DVD

   ```
   # eject -rv
   ```

2.2.2.2 Setting the Initial Values for the Installation Function [SMM]

Set the initial values for the installation function so that it works.
Perform the steps in this section only for new systems.

1. Creating a configuration file

   Create the pxinst.conf configuration file in the following path.

   ```
   /etc/opt/FJSVtcs/pxinst.conf
   ```

   Information

   The pxinst.conf file does not yet exist at this point in this procedure. Copy the sample file (/etc/opt/FJSVtcs/sample/pxinst.conf) and place it in the path shown above.

   Edit the configuration file, changing items as required according to the environment. For details on setting items, see "Initial Settings of the Installation Function" in "Details of the Installation Function" in the "Job Operation Software Administrator's Guide for System Management."

2. Initializing the installation function

   Execute the init subcommand in the pxinst command to set the initial values for the installation function. With "all" specified in the init subcommand, it sets all the values in batch.

   ```
   # pxinst init all
   dhcp:    Configuring dhcp
   dhcp:      Creating /etc/dhcp/dhcpd.conf  [OK]
   [INFO] INST 0801 pxinst Created a configuration backup.(/tmp/pxinst/dhcpd.conf_date)        (*)
   dhcp:      systemctl enable dhcpd         [OK]
   dhcp:      systemctl restart dhcpd        [OK]
   tftp:    Configuring tftp
   tftp:      Creating /etc/xinetd.d/tftp    [OK]
   [INFO] INST 0801 pxinst Created a configuration backup.(/tmp/pxinst/tftp_date)             (*)
   tftp:      systemctl enable xinetd        [OK]
   ...
   [INFO] INST 1801 pxinst Initialization completed successfully.(all)
   ```

   date: Command execution date and time (output in the YYYYMMDDhhmmss format)
   (*) The settings from before command execution are saved as a backup.

   For details on what is set in the items, see "Initial Setting of the Installation Function" in "Details of the Installation Function" in the "Job Operation Software Administrator's Guide for System Management."

Information

If you execute pxinst init all on the system management node with the kernel parameter net.core.somaxconn less than 4096, it changes to 4096. Also, the kernel parameter net.core.somaxconn is set to 4096 for standby system management node and compute cluster sub management node which built by using the Installation function.

2.2.2.3 Creating a Repository for the System Management Node [SMM]

Create a repository for reference by the system management node itself. By creating the repository, you can apply packages by using the yum command. For details on creating the repository, see "Managing a Repository" in "Overview of the Installation Function" in the "Job Operation Software Administrator's Guide for System Management."

1. Creating a repository

   [Using the DVD]

   a. Mounting the OS DVD

   ```
   # mount /dev/cdrom mntpath
   ```

   mntpath: Path to an arbitrary directory for mounting the OS DVD

   b. Creating a repository

   ```
   # pxinst repo create -R reponame --enable
   [INFO] INST 0800 pxinst repo completed successfully.(create repository)
   ```

   reponame: Arbitrary repository name

   [Using the ISO image file]

   a. Creating a repository

   ```
   # pxinst repo create -R reponame -S filepath --enable
   [INFO] INST 0800 pxinst repo completed successfully.(create repository)
   ```

   reponame: Arbitrary repository name
   filepath: Path to the OS ISO image file

Note

Do not create a duplicate repository name with an existing one.

2. Registering the RPM package file of Technical Computing Suite

   Register the Technical Computing Suite packages with the created repository.

```
# pxinst repo pkgkit -R reponame --kit packagekit_path
[INFO] INST 0800 pxinst repo completed successfully.(add package kit)
```

reponame:: Name of the repository created in step 1
packagekit_path: Path to the Technical Computing Suite package kit

Note

The Technical Computing suite package kit is a tar file containing the RPM package file of each component supplied on the Technical Computing Suite DVD. The file name is packagekit.tar.gz. For *packagekit_path*, directly specify the DVD, or specify the package kit file (packagekit.tar.gz) copied from the DVD. The registered packages in the Technical Computing Suite, except for the node performing this procedure, are automatically applied to the node at the time of "2.3.11 Reflecting Cluster Configuration Information [SMM]."

3. Updating the repository

   Update the repository.

```
# pxinst repo update -R reponame
[INFO] INST 0800 pxinst repo completed successfully. (update repository)
```

   reponame: Name of the repository created in step 1

Note

In an environment where both a system management node and a login node are used, use the ACL function of the OS to restrict access to the directories provided by the installation function.

It is possible to carry out the operation without this procedure, but note that the user who logged in as a login node can refer to the files placed under each directory.

1. ACL configuration of the directory to be used by the installation feature

```
# chmod 700 /var/opt/FJSVtcs/inst
# setfacl -m user:apache:r-x /var/opt/FJSVtcs/inst
```

2. ACL configuration of the TFTP public directory

```
# chmod 700 tftpdir
# setfacl -m user:apache:r-x tftpdir
# setfacl -m user:nobody:r-x tftpdir
```

   tftpdir: Directory in which to place the TFTP material (Directory specified tftpdir in /etc/opt/FJSVtcs/pxinst.conf)

3. ACL Configuration of Repository Directory

```
# chmod 700 repodir
# setfacl -m user:apache:r-x repodir
```

   repodir: Directory in which to deploy the repository (Directory specified repodir in /etc/opt/FJSVtcs/pxinst.conf)

2.2.3 Installing the Job Operation Software [SMM]

Apply the Job Operation Software packages to the system management node.

1. Checking a package list

   To identify the packages required for installation on the system management node, see the "package list" file for a description, by node type, of the packages that must be applied. This package list is the packagekit.yaml file stored in the Technical Computing Suite package kit.

   The following example extracts the Technical Computing Suite package kit to the /tmp/pkgkit directory.

```
# mkdir -p /tmp/pkgkit
# cd /tmp/pkgkit
# tar xvzf pkgkit packagekit/packagekit.yaml
```

   pkgkit: Path to the package kit

   packagekit.yaml is written in the following format.

```
node:
  SMM:
    FEFS: FJSVfefs-client-modules,FJSVfefs-client,FJSVfefs-client-debuginfo,FJSVfefs-iokit
    LANG: null
    TCS: FJSVpxinst,FJSVpxpsm,FJSVpxprm,...
  CCM:
    FEFS: FJSVfefs-client-modules,FJSVfefs-client,FJSVfefs-client-debuginfo,FJSVfefs-iokit
    LANG: null
```

```
    TCS: FJSVpxinst,FJSVpxpsm,FJSVpxprm,...
...
```

Apply the packages described in the SMM section to the system management node.

For node serving as all of system management node, compute cluster management node, and login node, apply the packages described in the SMM, CCM, and LN section to the system management node.

2. Applying packages

   Apply the packages confirmed in step 1 to the system management node.
   The following example applies packages by using the yum command.

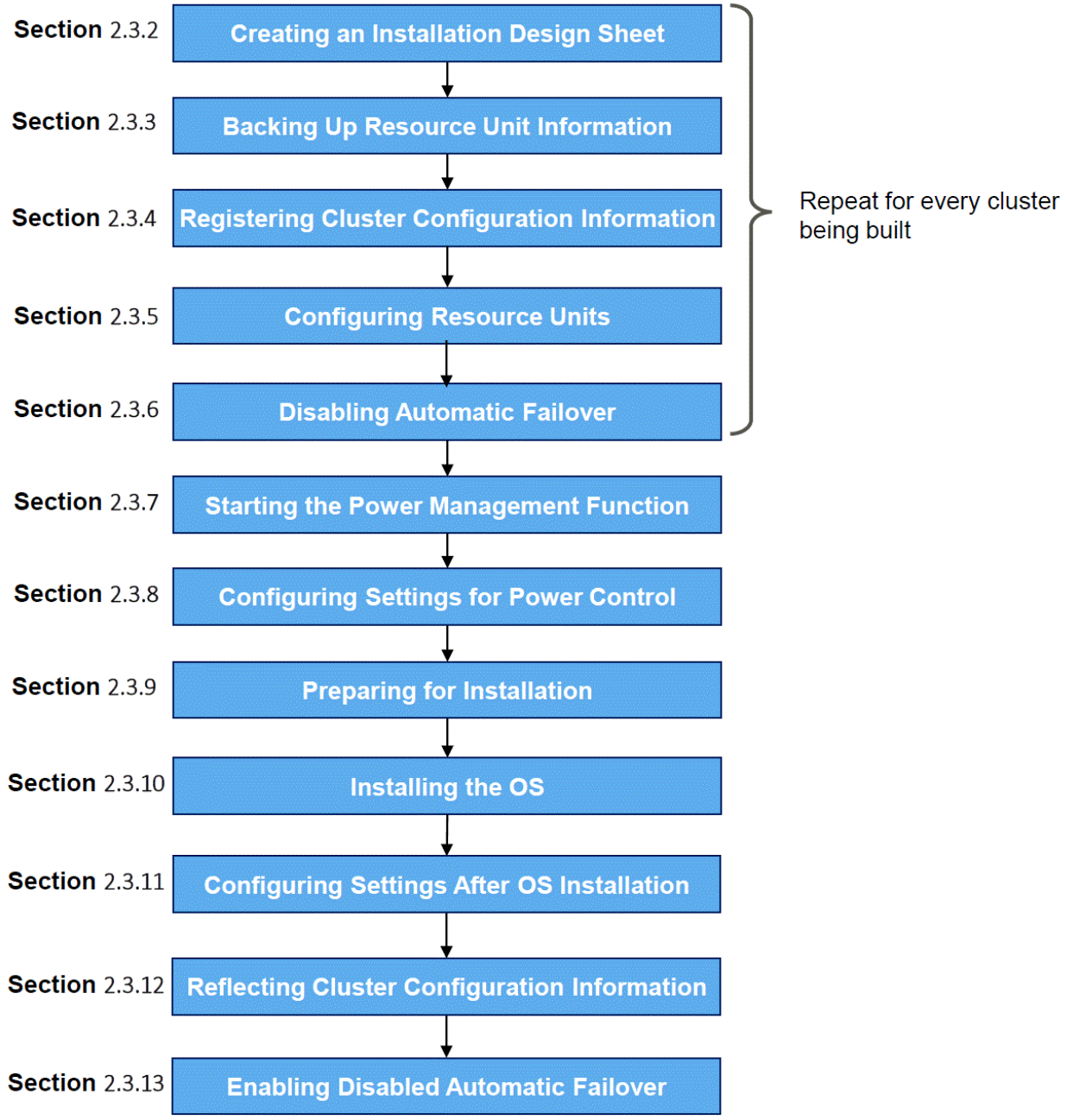
```
# yum install FJSVfefs-client-modules FJSVfefs-client FJSVfefs-client-debuginfo FJSVfefs-iokit
FJSVpxinst FJSVpxpsm FJSVpxprm ...
```

2.3 Building a Cluster

This section describes the cluster building workflow. For new system installation or the addition of a cluster to an existing system, see these procedures. The work must be done by the users having root user privileges to the OS.

The following figure shows the workflow for building a cluster.

Figure 2.3 Workflow for Building a Cluster



Information

The two methods of registering cluster configuration information are as follows:

- Creating a template for the cluster configuration definition file by using the output information from an installation design sheet (Recommended)
- Manually creating a cluster configuration definition file from scratch

With the recommended method, you can easily create a cluster configuration definition file. The descriptions in the subsequent procedures assume that you are using the recommended method. In this method, you need to create an installation design sheet for each cluster.

2.3.1 Prerequisites to Installation

To perform the subsequent work, the prerequisites shown below must be satisfied:

- Completion of a review of the information described in "2.1 Designing the System Configuration"
- Preparation of media with the software to be installed
  - OS for the PRIMERGY server
    - Red Hat Enterprise Linux (Server/Scientific Computing) DVD or ISO image file (prepared with every OS to be used)
    - Red Hat Enterprise Linux Server-Extras Repository (Using Red Hat Enterprise Linux Scientific Computing)
  - ServerView Suite DVD disc1 supplied with the PRIMERGY server
  - ETERNUS multipath driver DVD (if used)
  - Technical Computing Suite DVD
  - OS for the FX server (For the FX server in a cluster)
    - OS ISO image file for the FX server
  - CD containing the frequency table file rpm that is attached to the FX server (For the FX server in a cluster)
  - ISO image file of Mellanox OFED

    Prepare the OFED version supporting the IB HCA cards mounted in the PRIMERGY server.
    From https://support.ts.fujitsu.com/IndexDownload.asp, download the appropriate OFED version for the hardware and OS version used.
  - DVD with Fujitsu Linux support packages or ISO image file (if you have a support desk contract)
  - Logstash package, filebeat package (if used the periodical log collection function)
  - GPU driver (for any existing GPU-equipped PRIMERGY compute nodes)
    NVIDIA CUDA Toolkit (for any existing GPU-equipped PRIMERGY compute nodes)
    NVIDIA Container Toolkit (for using GPU from a Docker container)
  - Singularity package (for using Singularity)
- Completion of settings for the required hardware (PRIMERGY server, FX server, and peripherals) shown below

Table 2.10 Required Hardware Settings

Target Hardware	Setting Description
PRIMERGY server	LAN cable and other wiring
	RAID settings
	IPMI IP address settings (to remotely control power during installation)
	InfiniBand wiring (if used)
FX server	LAN cable, Tofu cable, and other wiring
	BMC settings (Tofu coordinates, IPMI IP address settings, etc.)
Network switch (L2/L3)	LAN cable wiring
	vlan settings (if used)
	IP routing settings (if used)
	DHCP relay settings (for a system with a network configuration divided into multiple subnets)
External storage device	Settings for control from the OS (LAN settings, etc.)
	FC cable wiring
	RAID settings

2.3.2 Creating an Installation Design Sheet [PC]

Based on the designed system configuration, enter the required input information for OS installation into an installation design sheet. An installation design sheet must be created for each cluster because it is used together with the information subsequently described in "2.3.4 Registering Cluster Configuration Information [SMM]." A folder showing the creation date and time is created in the same location as the folder with the design sheet, then the information entered in the design sheet for installation is output into the folder as an installation configuration file.(*)

Place the installation configuration files at any location on the system management node for use in the subsequent procedures "2.3.4 Registering Cluster Configuration Information [SMM]" and "2.3.9 Preparing for Installation." For details on how to use an installation design sheet, see "Creating Settings of the Installation Target Node" in " Details of the System Management Function" in the "Job Operation Software Administrator's Guide for System Management."

(*) According to the entered configuration, a node information definition file, disk partition information definition file, and FX server node information definition file are output.

Note

The design sheet for subsequent clusters does not require the definition of a system management node.

2.3.3 Backing Up Resource Unit Information [SMM]

If you are adding a cluster, the "2.3.4 Registering Cluster Configuration Information [SMM]" operation automatically creates the resource unit and returns to the default configuration.

1. Back up the following directory:
   /etc/opt/FJSVtcs/psm/cluster_config/Rscunit.d
2. Back up the following directory:
   /etc/opt/FJSVtcs/Rscunit.d
3. Back up the following file if it exists.
   /etc/opt/FJSVtcs/parscunit.conf

2.3.4 Registering Cluster Configuration Information [SMM]

Create the cluster configuration definition file pxsystem.conf to define a configuration for the Job Operation Software. Use the --create option in the pxsystemadm command to set the file. Perform this procedure in units of clusters to build.

Note

Even if any compute node does not exist, be sure to include the compute node information in the compute cluster entry (Cluster which has ClusterType = "compute") in pxsystem.conf.

If the compute node information is not included, the STATUS column of the active compute cluster management node displayed by the pashowclst command does not change from "Stopped" in "2. Confirming the completion of settings" of "2.3.11 Reflecting Cluster Configuration Information [SMM]."

See

For details on the cluster configuration definitions, see "Registering Cluster Configuration Information" in the "Job Operation Software Administrator's Guide for System Management."

1. Creating a template for the cluster configuration definition file pxsystem.conf

   You can create a template for the pxsystem.conf file used in this procedure together with the node information definition file that is output from an installation design sheet.

   Perform the subsequent steps to create a template for the pxsystem.conf file.

a. Editing the node information definition file

   Each time that the command is executed for a template for pxsystem.conf, it creates information for a single cluster. As a result, information on nodes in multiple clusters may be written in the node information definition file. If so, split the file so that each file contains only the node information for one cluster.

b. Creating a template for the pxsystem.conf file

   Execute the pxsystemadm command with the --convert option.

```
# pxsystemadm --convert filename
[INFO] PSM 1081 pxsystemadm The file was output. (outfile)
```

   filename: Path to the node information definition file
   outfile: File name of the output template

   The template for the pxsystem.conf file is output under the current directory at the command execution time. The file name is pxsystem.conf_*command execution date and time*. The administrator enters items where not enough items exist.

2. Creating the cluster configuration definition file pxsystem.conf

Create the cluster configuration definition file pxsystem.conf.
If you created a template for pxsystem.conf in step 1, enter items where not enough items exist in the template. For details on the template, see "Creating a Template of the Cluster Configuration Definition File" in the "Job Operation Software Administrator's Guide for System Management." For information about pxsystem.conf setting items, see "pxsystem.conf File" in the "Job Operation Software Command Reference."

- Since the representative IP address of the management network and the management network IP address set in pxsystem.conf are also used as the provisioning network at the time of system construction, if incorrect information is registered, the OS installation will fail at the timing of "2.3.10 Installing the OS".
- If you change the contents of this item after "2.3.10 Installing the OS", you may need to reinstall the OS on nodes throughout the system, not just the target node.
- If you already have other cluster configuration registered by the time you perform the procedures in this section, the pxsystem.conf file that is provided in this section should contain information for the existing system management nodes.
  The following example defines information for an existing system managed node (redundant configuration).

  [How to Confirm Existing System Management Node Information]
  Specify the -v and -l options in the pashowclst command, and confirm the node information for the existing system management nodes.

```
# pashowclst -v -l --nodetype SMM
[ CLST: storage ]
[ NODETYPE: SMM ]
NODE        MODEL  MNG_NET_MASTER   MNG_NET          CTRL_NET_SRV     CTRL_NET         HOSTNAME
0x00000001  PG     192.0.2.10       192.0.2.11       198.51.100.1     198.51.100.11    smm01
0x00000002  PG     192.0.2.10       192.0.2.12       198.51.100.2     198.51.100.12    smm02
```

  [How to Append to the pxsystem.conf File]
  Add information about existing system management nodes to the pxsystem.conf file generated by the --convert option of the pxsystemadm command.

```
# cat pxsystem.conf
Operation {
    Type = add-clst
}
Cluster {
    ClusterName = cluster
    ClusterType = compute
    ...
```

```
    SMMGroup {
        ManageNetMaster = 192.0.2.10
        SMM {
            ControlNetSrv = 198.51.100.1
            ControlNet = 198.51.100.11
            ManageNet = 192.0.2.11
            Model = PG
            CombinedType = SCM
        }
        SMM {
            ControlNetSrv = 198.51.100.2
            ControlNet = 198.51.100.12
            ManageNet = 192.0.2.12
            Model = PG
            CombinedType = SCM
        }
    }
    ...
}
```

Information

- The active and standby states of redundant nodes are automatically set when "2.3.11 Reflecting Cluster Configuration Information [SMM]" is implemented. However, for system management node, it becomes active from the time of this configuration.
- The HA section of pxsystem.conf sets the shared_disk information that system management nodes and compute cluster management nodes mount. If you know the by-id of the shared_disk to be mounted at the time of this setup, the setup can be described at the time of this procedure so that the shared_disk is automatically mounted on the active node when the "2.3.11 Reflecting Cluster Configuration Information [SMM]" is executed. If you do not set it at this procedure, set it at the timing of "2.4.1.3 Configuring a shared_disk [SMM]".

3. Registering the pxsystem.conf file contents

Register the pxsystem.conf file contents with the system management node.

- Adding nodes that include the FX server

  Use the --create option in the pxsystemadm command to get the system management node to reflect the cluster configuration definitions.
  Specify the --ft option to specify the path to the storage directory of the FX server node information definition file for the installation function.

```
# pxsystemadm --create filename --ft node_ft
Operation/Type is "add-clst".                                   (*1)
Do you really want to continue (y/n)? y                         (*2)
[INFO] PSM 1082 pxsystemadm The setting was completed.
```

  (*1) Shows the type of operations to be performed.
  (*2) This prompts for confirmation before registration.

  filename: Path to the created cluster configuration definition file
  node_ft: Path to the storage directory of the FX server node information definition file

See

Use an installation design sheet to create the FX server node information definition file.
For details on the installation design sheet, see "Creating Settings of the Installation Target Node" in the "Job Operation Software Administrator's Guide for System Management."

- Adding nodes that do not include the FX server

  Use the --create option in the pxsystemadm command to get the system management node to reflect the cluster configuration definitions.

```
# pxsystemadm --create filename
Operation/Type is "add-clst".                                    (*1)
Do you really want to continue (y/n)? y                          (*2)
[INFO] PSM 1082 pxsystemadm The setting was completed.
```

  (*1) Shows the type of operations to be performed.
  (*2) This prompts for confirmation before registration.

  filename: Path to the created cluster configuration definition file

Note

If you are installing a new system and the system management nodes are redundant, this procedure starts the representative IP address of the management network.
If the management network is restarted or the IP address is changed after the procedure is performed, the representative IP address of the management network will be stopped. Restart the node to start the representative IP address of the management network.

Information

At times such as during step-by-step FX server installation, the node information written in the pxsystem.conf and FX server node information definition files may include some hardware that does not exist in the system. Even in that case, the nodes are registered with the system management node.
The nodes registered with the system management node while some hardware does not exist do not need to be registered again. Instead, after OS installation in "2.3.10 Installing the OS," they are incorporated into job operations in "2.3.12 Configuring Settings After OS Installation" and subsequent procedures.

4. Confirming the registered information

   Specify the -a option in the pashowclst command, and confirm that the set contents are registered.

```
# pashowclst -a
...
```

   Confirm that the following information is correct.

   - The unique information for the added nodes does not have an error.
   - The mount information displayed when the shared_disk is set does not have an error.

   At this point, a hyphen is displayed for the service run status (SRV_STATUS column) because the functions of the Job Operation Software do not work on newly added nodes. Also, "Stopped" is displayed for the node run status (STATUS column) because the system monitoring function is not working.

   Furthermore, in a situation such as step-by-step system installation, if hardware is missing, "Unknown" is displayed for the power status (PWR_STATUS column) of the target node.

See

For details on what is displayed by the pashowclst command, see "Displaying System Configuration Information" in the "Job Operation Software Administrator's Guide for System Management."

2.3.5 Configuring Resource Units [SMM]

If you are adding a compute cluster in the system for the first time, the following steps 1-5 are not required.
When you are adding a cluster, the "2.3.4 Registering Cluster Configuration Information [SMM]" operation automatically creates the

resource unit and returns to the default configuration.
If you have changed the resource unit, check the parscunit.conf configuration file for the resource unit.
If the resource unit configuration after changing the resource unit differs from the settings in the parscunit.conf file, restore and edit the parscunit.conf file from a backup of resource unit information in the following procedure.

1. Deleting directories and files
   Delete the following directories and files:

   - /etc/opt/FJSVtcs/psm/cluster_config/Rscunit.d
   - /etc/opt/FJSVtcs/Rscunit.d
   - /etc/opt/FJSVtcs/parscunit.conf

2. Placing backed up directories and files
   Copy the directories and files that you backed up in "2.3.3 Backing Up Resource Unit Information [SMM]" and locate them as follows:

   a. Directory in step 1 of "2.3.3 Backing Up Resource Unit Information [SMM]"
      Place it under the directory /etc/opt/FJSVtcs/psm/cluster_config/.

   b. Directory in step 2 of "2.3.3 Backing Up Resource Unit Information [SMM]"
      Place it under the directory /etc/opt/FJSVtcs/.

   c. The file in step 3 of "2.3.3 Backing Up Resource Unit Information [SMM]"
      Place it under the directory /etc/opt/FJSVtcs/.

3. Editing the resource unit configuration file
   Edit /etc/opt/FJSVtcs/parscunit.conf. For details on the setting items, see "parscunit.conf file" in the "Job Operation Software Command Reference".

4. Reflecting settings
   Execute the parscunitadm command with the options --set, --no-dist, and --force to reflect the settings.

```
# parscunitadm --set --no-dist --force
[INFO] PSM 1402 parscunitadm The setting of resource unit completed
```

5. Checking the result of reflecting settings
   Execute the pashowclst command with the --rscunit option to verify that the resource unit information is:

   - The resource unit name is correct.
   - The sum of RUNNING, STOPPED, ERROR, and DISABLE for each resource unit matches the number of compute nodes.

```
# pashowclst -c clstname --rscunit
[ CLST: clstname ]
RSCUNIT           RUNNING  STOPPED  ERROR    DISABLE
unit1             0        48       0        0
unit2             0         2       0        1
unit3             0         3       0        0
```

   clstname: Cluster name

2.3.6 Disabling Automatic Failover [SMM]

To prevent unnecessary failover during installation work, use the --set-failover option in the paclstmgr command disable failover of the cluster to build. Perform this procedure for all the clusters to build.

For the following nodes in redundant configurations, execute the paclstmgr command to disable automatic failover:

- System management node
- Compute cluster management node
- Compute cluster sub management node
- Storage cluster management node

```
# paclstmgr -c clstname  -n nodeid --set-failover disable
```

clstname: Cluster name
nodeid: Node ID of a redundantly configured node (Either of the active and standby nodes would be okay.)

Use the paclstmgr command to check failover setting information.
Confirm that the FAILOVER column shows "disable" for the nodes in the cluster.

```
# paclstmgr -c clstname -n nodeid --set-failover list
[ CLST: clstname ]
NODE        NODETYPE      FAILOVER
0x00000001  SMM           disable
0x00000002  SMM           disable
```

clstname: Cluster name
nodeid: Node ID of a redundantly configured node (Either of the active and standby nodes would be okay.)

2.3.7 Starting the Power Management Function [SMM]

Start the power management function service on the system management node.

For the node serving as all of system management node, compute cluster management node, and login node, this step is unnecessary.

Perform this work on the system management node.

- For a new system installation

```
# paclstmgr -c clstname  -n nodeid --service start
```

clstname: Cluster name
nodeid: Node ID of the active system management node

- For adding a Cluster

```
# systemctl restart pxsyspwrd
```

Confirm that "PWRD(o)" is shown by the pashowclst command.

```
# pashowclst -c clstname -n nodeid
[ CLST: clstname ]
[ NODE: nodeid ]
NODE        NODETYPE      STATUS        REASON             PWR_STATUS        ARCH_STATUS       SRV_STATUS
nodeid      SMM,SCM       Running       -                  on                -                 IC(o),PWRD(o)
```

clstname: Cluster name
nodeid: Node ID of the active system management node

2.3.8 Configuring Settings for Power Control [SMM]

Perform the following procedure to configure settings for system management power control as required, in order to control the power of nodes composing the cluster. Perform this procedure in units of clusters to build.

1. Creating the configuration file /etc/opt/FJSVtcs/papwr.conf

   Create the configuration file /etc/opt/FJSVtcs/papwr.conf.

   [Copying a sample file]

   ```
   # cp /etc/opt/FJSVtcs/sample/papwr.conf /etc/opt/FJSVtcs/papwr.conf
   ```

   [Editing]

   ```
   # vi /etc/opt/FJSVtcs/papwr.conf
   Cluster {
       ClusterName = clstname
           Model {
   ```

```
            Model = PG
            User = username
            Password = password
        }
...
```

clstname: Name of the cluster being built
username: IPMI user name
password: IPMI user password

Note

The password entered in/etc/opt/FJSVtcs/papwr.conf is encrypted when the setting is applied. To update, re-enter the password.

Information

When defining models other than "PG (PRIMERGY server)" and "FT (FX Server)" as "Arbitrary model", create a configuration file referring to "Power Control Plugin" in "Job Operation Software Administrator's Guide for System Management."

2. Reflecting settings

   Use the --set option in the papwradm command to reflect settings.

```
# papwradm --set
[WARNING]
Do you really want to continue (y/n)y
[INFO] PSM 0082 papwradm The processing of the configuration file was completed.
```

Note

In a new installation of a system with a standby system management node, the papwradm command prints a warning message because the OS for the standby system managed node is not installed at this time.

```
# papwradm --set
[WARNING]
Do you really want to continue (y/n)? y
[WARN] PSM 0051 papwradm There are nodes that failed in the setting.(filename)
[INFO] PSM 0082 papwradm The processing of the configuration file was completed.
```

filename: File name
If the command display is as shown above, since the configuration to the active system management node is completed, perform the procedure from "2.3.9 Preparing for Installation". After performing the procedure in "2.3.11 Reflecting Cluster Configuration Information [SMM]", execute the papwradm command again to apply the settings to the standby system management node.

3. Confirm Power Status

   Use the pashowclst command to confirm that the power state has been obtained.

```
# pashowclst -c clstname -v
...
```

clstname: Cluster name

Note

If the node you want to add has a power state (PWR_STATUS column) of "on" or "os-running", you must power off the node.
If you use the papwrctl command, run "force-off" because the OS may not be installed.

2.3.9 Preparing for Installation

Configure the required environment and create a configuration file for OS installation.

2.3.9.1 Preparing for Installation on the PRIMERGY server

Perform this work when the installation destination nodes include the PRIMERGY server.

You can select from three methods of OS installation on the PRIMERGY server: package-based OS installation, image-based OS installation (recommended), and OS installation from a backup image. The choice of OS installation from a backup image requires an already created backup image on the built system beforehand. For this reason, this method is not available in new system installation.

The descriptions in the subsequent procedures assume image-based OS installation, which is the recommended method because the installation can be done quickly and even has fewer factors causing faults.

- For details on the installation methods, see "Provisioning Method" in "OS Installation on PRIMERGY server" in the "Job Operation Software Administrator's Guide for System Management."
- For details on OS installation from a backup image, see "Restoring a Disk Image" in the "Job Operation Software Administrator's Guide for System Management."

2.3.9.1.1 Placing the Installation Configuration Files

Place the configuration files that are output from an installation design sheet in any directory on the system management node. Installation on the PRIMERGY server requires both a node information definition file and a disk partition information definition file.

2.3.9.1.2 Creating a Repository [SMM]

Create a repository for reference on the installation destination node. Create a repository for each OS version used. Also, prepare the OS DVD or the OS ISO image file in advance to create a repository. If a repository has already been created for the OS version to be used on the installation destination node, this procedure is unnecessary.

1. Creating a repository

   - Using the DVD

     Mount the OS DVD.

     ```
     # mount /dev/cdrom mntpath
     ```

     mntpath: Path to an arbitrary directory for mounting the OS DVD

     Create a repository.

     ```
     # pxinst repo create -R reponame
     [INFO] INST 0800 pxinst repo completed successfully.(create repository)
     ```

     reponame: Arbitrary repository name

   - Using the ISO image file

     Create a repository.

     ```
     # pxinst repo create -R reponame -S filepath
     [INFO] INST 0800 pxinst repo completed successfully.(create repository)
     ```

     reponame: Arbitrary repository name
     filepath: Path to the OS ISO image file

2. Registering the Technical Computing Suite packages

   Register the Technical Computing Suite packages with the created repository.

```
# pxinst repo pkgkit --kit packagekit_path -R reponame
[INFO] INST 0800 pxinst repo completed successfully.(add package kit)
```

packagekit_path: Path to the Technical Computing Suite package kit
reponame: Name of the repository created in step 1

3. Registering more packages if not sufficient
   If the OS is Red Hat Enterprise Linux for Scientific Computing, some of the packages required for running the Job Operation Software are not included. For this OS, you need to copy the same version of packages from the Red Hat Enterprise Linux Server repository or from the ISO image file (DVD).
   The required packages are python-jinja2, PyYAML, python-babel, hwloc, protobuf-c, lzop, and container-selinux.
   Packages other than container-selinux are included with Red Hat Enterprise Linux Server, and container-selinux is included with Red Hat Enterprise Linux Server-Extras. Copy them.
   The following example copies the packages from the repository for Red Hat Enterprise Linux Server in order to use Red Hat Enterprise Linux for Scientific Computing.

   ```
   # cp repopath/reponame_server/Packages/package_file_name repopath/reponame/Packages/
   ```

   repopath: Specified value in the repodir item in the pxinst.conf file

   reponame_server: Repository name for Red Hat Enterprise Linux Server
   package_file_name: File name of the package to copy
   reponame: Repository name created in step 1

4. Updating the repository

   Update the repository.

   ```
   # pxinst repo update -R reponame
   [INFO] INST 0800 pxinst repo completed successfully.(update repository)
   ```

   reponame: Name of the repository created in step 1

Information

To install the OS on a node where the OS update package program (containing the kernel update package) has been applied, create a repository and then register the required update packages with the repository. By registering the update packages and then performing the subsequent procedures, you can install the OS while the update packages are automatically applied to the initrd and rootfs. For details on how to add the packages to the repository, see "Managing a Repository" in " Details of the System Management Function" in the "Job Operation Software Administrator's Guide for System Management."

2.3.9.1.3 Creating the initrd [SMM]

Create the initrd (initial RAM disk for OS installation) required for OS installation. Create the initrd for each OS version to install. Also, to create the initrd, you need to have created a repository for the corresponding OS version in advance. If the initrd corresponding to the OS version used has already been created, this procedure is unnecessary.

```
# pxinst initrd create -R reponame
[INFO] INST 0800 pxinst initrd completed successfully.(create initrd)
```

reponame: Name of the repository for the created initrd

Information

If integrating a special kernel module in the initrd is necessary, use the kernel module integration function. Special kernel modules are modules other than the OS standard kernel modules and kernel modules provided on the ServerView Suite DVD. For details on integrating the kernel module, see "Installing a Kernel Module" in the "Job Operation Software Administrator's Guide for System Management."

Integrating kernel modules is required in the following case examples:

- To use hardware other than the PRIMERGY server, and a kernel module that supports the hardware is needed

- To integrate an updated version of a kernel module, which is needed because the one provided on the ServerView Suite DVD is old

2.3.9.1.4 Configuring ServerView Suite [SMM]

ServerView Suite provides the kernel modules required for the PRIMERGY server. Import the kernel modules and supplied software for the PRIMERGY server. If they have already been imported, this procedure is unnecessary. If the installation destination nodes include the PRIMERGY server, this work is required.

Note

Suppose you are adding a new model different from the existing nodes when adding a cluster, for example. In this case, if you have performed this procedure on the relevant repository, use the ServerView Suite DVD corresponding to the new model, and configure ServerView Suite again.

1. Preparation

   Insert ServerView Suite DVD 1 supplied with the hardware into the active system management node. Alternatively, place an ISO image file created from the DVD beforehand in any path on the active system management node.

2. Integrating kernel modules and supplied software

   Extract and integrate the kernel modules and supplied software from the ServerView DVD or ISO image file.

   [Using the DVD]

   ```
   # pxinst svs setup -R reponame
   Copying the ServerView Suite software and kernel modules
   Updating the initial ram disk
   [INFO] INST 0800 pxinst svs completed successfully.(setup)
   ```

   reponame: Name of the repository for importing the kernel modules and supplied software

   [Using the ISO image file]

   ```
   # pxinst svs setup -R reponame -S isoimage
   Copying the ServerView Suite software and kernel modules
   Updating the initial ram disk
   [INFO] INST 0800 pxinst svs completed successfully.(setup)
   ```

   reponame: Name of the repository for importing the kernel modules and supplied software
   isoimage: ISO image file path of the ServerView DVD

2.3.9.1.5 Creating a rootfs Image [SMM]

Create a rootfs image required for OS installation. If the image has already been created, this procedure is unnecessary. Create a rootfs image for each OS version to install (for each repository). Also, to create the rootfs image, you need to have created a repository for the corresponding OS version in advance. Only the basic OS packages and the minimum packages for running the Job Operation Software are installed in the rootfs. If you need to install additional packages, complete installation on nodes and then install the packages. Alternatively, use the rootfs creation function to import them in advance.

Execute the following command to create a rootfs image.

```
# pxinst rootfs create -R reponame
[INFO] INST 0800 pxinst rootfs completed successfully.(create rootfs)
```

reponame: Name of the repository for the created rootfs image

2.3.9.1.6 Registering Node Information [SMM]

Use the node subcommand in the pxinst command to register the installation destination node. If there will be multiple OS versions used, execute the command multiple times so that it is executed separately for each version. If the OS versions are grouped using the "group" item on an installation design sheet, as many node information definition files as OS versions are output. In that case, execute the command to

register the node information as many times as the number of configuration files for the OS versions. For details on registering the node information, see "Registering Node Information" in the "Job Operation Software Administrator's Guide for System Management."

Information

- To boot the installation destination node using BIOS mode, add the --bios option when executing the node subcommand in the pxinst command.
- To reference the remote console by using the paconsole command or ipmitool during OS installation, specify the --bootopt option (kernel startup option) when executing the node subcommand in the pxinst command. This can enable the remote console. (Example of specification: --bootopt="console=ttyS0")
- For package-based OS installation, specify "pkg" in the -T option of the node subcommand in the pxinst command.

Note

- If the remote console is not enabled by the --bootopt option when the node subcommand in the pxinst command is executed, the paconsole and ipmitool commands cannot be used to connect the OS console until the target node restarts after OS installation.
- Applying the supplied software for ServerView

  The supplied software for ServerView Suite may not be required for system operation. We recommend not installing any unnecessary software since it would not be appropriate for HPC system operations in terms of increased installation time, conflicts between packages with dependencies on the supplied software, and consumption of CPU, memory, and other resources. We recommend not applying all the software or applying only ServerView RAID Manager, which is required when a disk fails.

  Even when the supplied software is not installed, the kernel modules required for hardware are automatically applied. For details on the roles of the supplied software and possible combinations for installation, see the ServerView Suite documentation.

- Specifying the supplied software to apply
  To apply the supplied software for ServerView Suite, specify the --svs option in the node subcommand in the pxinst command. If this option is omitted, the supplied software is not applied.

  Specify supplied software numbers as comma-separated values in an argument of the --svs option. Use the pxinst svs list command to output supplied software numbers (displayed in the NO column). The following example shows that the applicable software may vary depending on the ServerView Suite version used.

```
# pxinst svs list -R reponame
NO   SOFTWARE
--   --------
1    ServerView Operations Manager
2    ServerView Update Agent (online flash)
3    ServerView CIM Provider
4    ServerView Agents
5    ServerView Agentless Service
6    ServerView RAID Manager
```

  reponame: Name of the repository for applying the supplied software for ServerView

Perform the following procedure to register node information.

1. Registering node information for OS installation

```
# pxinst node add -R reponame -N nodefile -P diskfile --svs no
[INFO] INST 1803 pxinst Adding nodes has been completed.(nodenum=xxx)
```

   reponame: Repository name
   nodefile: Path to the node information definition file
   diskfile: Path to the disk partition information definition file
   no: Number of the ServerView Suite supplied software to apply
   xxx: Number of nodes

Note

The node information definition file and disk partition definition file are required when making system configuration changes.
Be sure to take a backup.
The following example creates a backup in /root/packagekit/.

```
# mkdir -p /root/packagekit
# cp nodefile /root/packagekit/
# cp diskfile /root/packagekit/
```

nodefile: Path of node information definition file
diskfile: Path of disk partition definition file

2. Confirming that the node information is registered
   OPERATION varies according to what is specified in the -T option. The following example shows the -T option not specified or "img" specified.
   Confirm that the registered node information is displayed. STATUS shows "NotStarted" and STATE shows "Waiting" for a node before OS installation on the node.

```
# pxinst status show
HOSTNAME       OPERATION      STATUS       STATE        DETAIL
--------       ---------      ------       -----        ------
hostname1      ImgProvision   NotStarted   Waiting      -
hostname2      ImgProvision   NotStarted   Waiting      -
hostname3      ImgProvision   NotStarted   Waiting      -
hostname4      ImgProvision   NotStarted   Waiting      -
...
```

3. Restarting the DHCP service

   Restart the DHCP service.

```
# systemctl restart dhcpd
```

4. Synchronizing the setting

   Synchronize the settings for the standby installer nodes and relay nodes.
   Perform this procedure if a standby installer node or relay node exists in the system and you have already installed the OS.

```
# pxinst sync all
...
[INFO] INST 1805 pxinst File replication is complete.
```

2.3.9.2 Preparing for Installation on the FX server

Perform this work when the installation destination nodes include the FX server.

2.3.9.2.1 Placing the Installation Configuration Files

Place the installation configuration files that are output from an installation design sheet in any directory on the system management node.
Installation on the FX server requires the FX server node information definition file.

2.3.9.2.2 Configuring Hardware Control [SMM]

Configure the required settings for installation on FX server hardware.

Perform this procedure on every FX server that is a build target.

Use the --set-hardware option in the pxsystemadm command to configure FX server control settings.

```
# pxsystemadm --set-hardware -c clstname
Do you really want to continue (y/n)? y
[INFO] PSM 1082 pxsystemadm The setting was completed.
```

clstname: Cluster name

In step-by-step FX server installation, for example, the system may not have some of the hardware. In this case, use the scope option to execute the command in only the range of existing hardware.

The following example specifies a boot group and executes the pxsystemadm command.

```
# pxsystemadm --set-hardware -c clstname --bootgrp bootgid
Do you really want to continue (y/n)? y
[INFO] PSM 1082 pxsystemadm The setting was completed.
```

clstname: Cluster name
bootgid: Boot group ID

2.3.9.2.3 Checking Hardware Control Settings [SMM]

Confirm that hardware settings are correct.

Perform this procedure on every FX server that is a build target.

Use the --check-hardware option in the pxsystemadm command to check the FX server control settings.

```
# pxsystemadm --check-hardware -c clstname
[INFO] PSM 1083 pxsystemadm The setting check succeeded.
```

clstname: Cluster name

If the settings are not correct for any FX server, a message appears as shown below. Take action based on the output file.

```
# pxsystemadm --check-hardware -c clstname
[ERR.] PSM 1001 pxsystemadm There are nodes which has not setting of the hardware.(filename)
```

filename: Name of the output file

If some hardware in the system is missing, such as a phased FX server installation, use the scope option to run the command only where the hardware is present.
The following example runs the pxsystemadm command with a boot group specification.

```
# pxsystemadm --check-hardware -c clstname --bootgrp bootgid
[INFO] PSM 1083 pxsystemadm The setting check succeeded.
```

clstname: Cluster name
bootgid: Boot group ID

2.3.9.2.4 Preparing a rootfs Image [SMM]

Perform a network installation on any boot I/O node and create a rootfs image from the disk image of the network-installed node.

Note

Since the purpose of the network installation performed in this procedure is to create a rootfs image that is used to build an FX server, building an FX server is not completed even after this procedure is completed.

1. Backing Up the TFTP Configuration File

   Back up the TFTP configuration file.
   Place the backup file outside of /etc/xinetd.d/ .

   ```
   # cp -p /etc/xinetd.d/tftp dirname/
   ```

   dirname: Any directory name

2. Editing the TFTP Configuration File

   Configure the TFTP configuration file to use the directory for the TFTP service used temporarily in this procedure.
   Edit /etc/xinetd.d/tftp file.
   Change the "server_args" entry as follows.

```
# vi /etc/xinetd.d/tftp
...
        server_args             = -s /var/lib/tftpboot
...
```

3. Reflect the TFTP configuration file

   Restart xinetd.service with the systemctl command.

```
# systemctl restart xinetd.service
```

4. Backing up the DHCP configuration file

   Back up the DHCP configuration file.

```
# cp -p /etc/dhcp/dhcpd.conf /etc/dhcp/dhcpd.conf.bak
```

5. Editing the DHCP Configuration File

   Edit /etc/dhcp/dhcpd.conf file.

   Add the "subnet" and "group" sections for any boot I/O nodes that you want to install (hereinafter referred to as the node to be installed).

Note

- If the DHCP configuration file already contains a "subnet" section with the same subnet IP address as the node being installed, you do not need to add the "subnet" section.
- If the target node does not have a gateway IP address, "option routers" is not required.

```
# vi /etc/dhcp/dhcpd.conf
...
subnet subnet_ipaddr netmask subnet_mask {
    default-lease-time 600;
    max-lease-time 3600;
    option PXE.mtftp-ip 0.0.0.0;

    option routers bio_gateway;
}
...
group {
    next-server smm_ipaddr;
    host bio_hostname {
        option host-name "bio_hostname";
        hardware ethernet bio_macaddr;
        fixed-address bioip_addr;
    }
}
...
```

subnet_ipaddr: Subnet IP address of the node to be installed
subnet_mask: Subnet mask of the node to be installed
bio_gateway: Gateway IP address of the node to be installed
smm_ipaddr: Representative IP address of the management network of the system management node
bio_hostname: Hostname of the node to be installed

bio_macaddr: MAC address of the node to be installed
bio_ipaddr: IP address of the management network of the node to be installed

6. Reflection of DHCP configuration files

   Restart dhcpd.service with the systemctl command.

```
# systemctl restart dhcpd.service
```

7. Preparing the Repository

   Mount the repository to be used during a network installation.
   Prepare the ISO image file of the OS beforehand.

```
# mkdir /var/www/html/FXREPO
# mount -t iso9660 -o loop filepath /var/www/html/FXREPO
```

   filepath: OS ISO image file path

8. Copying Modules

   Copy the modules to use during a network installation.

```
# mkdir -p /var/lib/tftpboot/uefi/
# cp /var/www/html/FXREPO//EFI/BOOT/grubaa64.efi /var/lib/tftpboot/uefi/
# cp /var/www/html/FXREPO/images/pxeboot/initrd.img /var/lib/tftpboot/uefi/
# cp /var/www/html/FXREPO/images/pxeboot/vmlinuz /var/lib/tftpboot/uefi/
```

9. Creating the kickstart File

   Create a kickstart file to use during a network installation.
   Create a kickstart file in /var/www/html/ with the contents shown in "Contents of the kickstart File for the FX Server".
   The following is an example of creation using the file name fx1000_ks.cfg.

```
# vi /var/www/html/fx1000_ks.cfg
...
```

   [Contents of the kickstart File for the FX Server]

```
# Redhat kickstart template
#
# Copyright (C) Fujitsu limited 2016
#
# Licensed under GPL version 2; See LICENSE file for details.
#

# System authorization information
auth --useshadow --passalgo=sha512

# Install OS instead of upgrade
install
# Reboot after installation
reboot

# Use text mode install
text
# Firewall configuration
firewall --disabled
firstboot --disable
# Keyboard layouts
keyboard jp106
# System language
lang C

# SELinux configuration
selinux --disabled
```

```
# configure the X Window System
skipx
# System timezone
timezone Asia/Tokyo --utc

# System bootloader configuration
ignoredisk --only-use=nvme0n1

# Partition clearing information
# clearpart --all --drives=nvme0n1

# Clear the Master Boot Record
# zerombr

# TODO: Disk partitioning information
part /boot/efi --fstype=fat32 --onpart=/dev/nvme0n1p1 --size=105 --asprimary
part /boot --fstype=xfs --onpart=/dev/nvme0n1p2 --size=419 --asprimary
part / --fstype=xfs --onpart=/dev/nvme0n1p3 --size=460010 --asprimary
part /var/crash --fstype=xfs --onpart=/dev/nvme0n1p4 --size=51500 --grow --asprimary

%packages --ignoremissing
@core
@base
@development
@debugging
@ftp-server
@large-systems
dhcp-server
httpd
ksh
nfs-utils
nmap
nss-pam-ldapd
papi
tftp-server
zsh
perf
ncurses-term
rpm-apidocs
audit-libs-devel
binutils-devel
bzip2-devel
dbus-devel
e2fsprogs-devel
elfutils-devel
elfutils-libelf-devel
environment-modules
expat-devel
expect
fftw
ftp
gamin
gdbm-devel
glib2-devel
glibc-utils
iptables-devel
iso-codes
keyutils-libs-devel
krb5-devel
libacl-devel
libaio-devel
libattr-devel
```

```
libcap-devel
libcgroup
libcom_err-devel
libqb
libsepol-devel
libtirpc-devel
lm_sensors
ncurses-devel
newt-devel
nspr-devel
nss-devel
nss-softokn-devel
nss-softokn-freebl-devel
nss-util-devel
numactl-devel
OpenIPMI
OpenIPMI-libs
openldap-clients
openssl-devel
pam-devel
papi-devel
pciutils-devel
pcre-devel
popt-devel
portreserve
readline-devel
sqlite-devel
tcl
tcl-devel
telnet
telnet-server
tftp
zlib-devel
java-1.8.0-openjdk-devel
libevent-devel
jansson-devel
qemu-kvm
chrpath
lm_sensors-devel
libnl3-devel
fuse
fuse-libs
ipmitool
ruby
ruby-irb
perl
perl-core
emacs
acpica-tools
gcc-gfortran
http-parser
libpfm-devel
protobuf
protobuf-c
psmisc
rpm-devel
swig
sysstat
systemd-devel
tcsh
bison-runtime
buildah
python3-jinja2
```

```
python3-netaddr
skopeo
glibc-all-langpacks
compat-openssl10
libxml2-devel
libffi-devel
python2-devel
python2-Cython
python36-devel
golang
libcurl-devel
python2-jinja2
gmp-c++
gmp-devel
gnutls-c++
gnutls-dane
gnutls-devel
iperf3
libtasn1-devel
nettle-devel
p11-kit-devel
xinetd
%end

%pre
# TODO: Disk partitioning information
LANG=C
dd if=/dev/zero of=/dev/nvme0n1 count=1 bs=512
/usr/sbin/parted -s /dev/nvme0n1 mklabel gpt
disk_size=`/usr/sbin/parted /dev/nvme0n1 unit MiB print | grep "^Disk /dev/nvme0n1" | (*)
awk '{print $3}'`
/usr/sbin/parted -s -a optimal /dev/nvme0n1 -- mkpart primary 1MiB 101MiB
/usr/sbin/parted -s /dev/nvme0n1 -- set 1 boot on
/usr/sbin/parted -s -a optimal /dev/nvme0n1 -- mkpart primary 101MiB 501MiB
/usr/sbin/parted -s -a optimal /dev/nvme0n1 -- mkpart primary 501MiB 438700MiB
/usr/sbin/parted -s -a optimal /dev/nvme0n1 -- mkpart primary 438700MiB ${disk_size}
/usr/sbin/parted /dev/nvme0n1 u MiB print
udevadm settle
%end

# Network information
network --device=bio_mac --bootproto=static --ip=bio_ipaddr --noipv6 --onboot=yes (*)
--netmask=bio_netmask --hostname=bio_hostname --activate

# Root password
rootpw --iscrypted root_passwd

# SSH configuration
sshpw --username=root --iscrypted root_passwd
```

bio_mac: MAC address of the node to be installed
bio_ipaddr: IP address of the management network of the node to be installed
bio_netmask: subnet mask for bio_ipaddr
bio_hostname: Hostname of the node to be installed
root_passwd: Root password in /etc/shadow (Encrypted String)

A new line starts just after the (*) part because of the space limitation of this document, but you actually create a file, write it on one line.

10. Creating a PXE configuration file

    Create the PXE configuration file named grub.cfg.

```
# vi /var/lib/tftpboot/uefi/grub.cfg
set default=install
set timeout=1
menuentry 'localdisk' {
  set root=(hd0,1)
  chainloader /EFI/redhat/grubaa64.efi
}
menuentry 'install' {
  linux uefi/vmlinuz ip=dhcp inst.repo=http://smm_ipaddr/FXREPO (*)
inst.ks=http://smm_ipaddr/fx1000_ks.cfg inst.text inst.sshd earlycon=pl011,0x1c050000 (*)
console=ttyAMA0
  initrd uefi/initrd.img
}
```

smm_ipaddr: Representative IP address of the management network of the system management node

A new line starts just after the (*) part because of the space limitation of this document, but you actually create a file, write it on one line.

11. Network Installation Start

    Start the network installation by powering up the node to be installed.
    Use the papwrctl command to power up.

```
# papwrctl -c clstname -n nodeid --boottype pxe on
```

clstname: Cluster Name
nodeid: Node ID of the node to be installed

12. Check the status of the network installation

    Open a terminal separate from the "Step 11" and use the paconsole command to check the installation.
    Ensure that no errors occur and that the login prompt is output.

```
# paconsole -c clstname -n nodeid
```

clstname: Cluster Name
nodeid: Node ID of the node to be installed

13. Creating a rootfs Image

    Transfer the rootfs image to the system management node with the following command.

```
# ssh root@ipaddr "tar czSf - --exclude=/dev/* --exclude=/proc/* --exclude=/run/* --exclude=/sys/
* \
--exclude=/tmp/* --exclude=/etc/ssh/ssh_h* /" > rootfs
```

ipaddr: IP address of the node to be installed
rootfs: rootfs image name for FX server

14. Restoring Configuration Files

    Restore the prepared file to its original state.

    [Restoring TFTP Configuration File]

```
# cp -p dirname/tftp /etc/xinetd.d/
# systemctl restart xinetd.service
```

dirname: Backup directory for tftp files

[Restoring DHCP Configuration Files]

```
# cp -p /etc/dhcp/dhcpd.conf.bak /etc/dhcp/dhcpd.conf
# systemctl restart dhcpd.service
```

[Unmount Repository]

```
# umount /var/www/html/FXREPO
# rm -fr /var/www/html/FXREPO
```

[Deleting PXE-related files]

```
# rm -fr /var/lib/tftpboot/uefi/
```

[Removing the kickstart File]

```
# rm -f /var/www/html/fx1000_ks.cfg
```

15. Powering down the node to be installed.

    Power down the target to be installed.
    Use the papwrctl command to power down.

    ```
    # papwrctl -c clstname -n nodeid --jobnowait force-off
    ```

    clstname: Cluster Name
    nodeid: Node ID of the node to be installed

2.3.9.2.5 Creating a Repository [SMM]

Create a repository for reference on the installation destination node. To create the repository, prepare the OS ISO image file in advance. If a repository has already been created for the OS version used by the installation destination node, this procedure is unnecessary.

1. Creating a repository

   Create a repository.

   ```
   # pxinst repo create -R reponame -S filepath
   [INFO] INST 0800 pxinst repo completed successfully.(create repository)
   ```

   reponame: Name for the created repository

   filepath: Path to the OS ISO image file

2. Registering the Technical Computing Suite packages

   Register the Technical Computing Suite packages with the created repository.

   ```
   # pxinst repo pkgkit  -R reponame --kit packagekit_path
   [INFO] INST 0800 pxinst repo completed successfully.(add package kit)
   ```

   packagekit_path: Path to the Technical Computing Suite package kit
   reponame: Name of the repository created in step 1

3. Updating the repository

   Update the repository.

   ```
   # pxinst repo update -R reponame
   [INFO] INST 0800 pxinst repo completed successfully.(update repository)
   ```

   reponame: Name of the repository created in step 1

2.3.9.2.6 Creating the initrd [SMM]

Create the initrd (initial RAM disk for OS installation) required for OS installation. To create the initrd, you need to have created a repository for the corresponding OS version in advance. If the initrd corresponding the OS version used has already been created, this procedure is unnecessary.

```
# pxinst initrd create -R reponame
[INFO] INST 0800 pxinst initrd completed successfully.(create initrd)
```

reponame: Name of the repository for the created initrd

2.3.9.2.7 Editing a rootfs Image [SMM]

Import the installation function settings to the FX server rootfs. If they have already been imported, this procedure is unnecessary. Prepare the rootfs created in advance to configure the settings.

Execute the following command to edit the rootfs image.

```
# pxinst rootfs create -R reponame --use-image rootfs
[INFO] INST 0800 pxinst rootfs completed successfully.(create rootfs)
```

reponame: Name of the repository created in "2.3.9.2.5 Creating a Repository [SMM]"
rootfs: FX server rootfs image prepared in "2.3.9.2.4 Preparing a rootfs Image [SMM]"

2.3.9.2.8 Registering Node Information

1. Registering node information

   Specify the --ft option to mean the FX server node when registering FX server node information. The result is the registration of the node information.

   #### Information

   Only a compute node that also serves as a boot I/O node is targeted in node information registration. At the point that OS installation is completed on the compute node that also serves as a boot I/O node, installation is completed on all the nodes in a BoB.

   The registration of node information differs depending on whether dynamic IP addresses are used. The respective steps are described below.
   For details on installation using dynamic IP addresses, see "Assigning a Dynamic IP Address when Installing a BoB" in the "Job Operation Software Administrator's Guide for System Management."

   [Using dynamic IP addresses]
   Specify the range of an IP address pool in the --pool option to dynamically allocate IP addresses. With comma-separated values, you can specify multiple ranges for an IP address pool.
   The following example shows a dynamic IP address pool with a range from 192.0.2.10/24 to 192.0.2.225/24.

   ```
   # pxinst node add -R reponame -N ft_file --ft --pool 192.0.2.10/24-192.0.2.225/24
   [INFO] INST 1803 pxinst Adding nodes completed.(nodenum=xxx)
   ```

   reponame: Name of the repository created in "2.3.9.2.5 Creating a Repository [SMM]"
   ft_file: Path to the FX server node information definition file
   xxx: Number of nodes

   [Without using dynamic IP addresses]

   ```
   # pxinst node add -R reponame -N ft_file --ft
   [INFO] INST 1803 pxinst Adding nodes completed.(nodenum=xxx)
   ```

   reponame: Name of the repository created in "2.3.9.2.5 Creating a Repository [SMM]"
   ft_file: Path to the FX server node information definition file
   xxx: Number of nodes

   #### Note

   The FX server node information definition file is required when making system configuration changes.
   Be sure to take a backup.
   The following example creates a backup in /root/packagekit/.

   ```
   # mkdir -p /root/packagekit
   # cp ft_file /root/packagekit/
   ```

   ft_file: Path to the FX server node information definition file

2. Restarting the service

   Restart the DHCP service so that the settings take effect.

```
# systemctl restart dhcpd
```

3. Synchronizing the setting

   Synchronize the settings for the standby installer nodes and relay nodes.
   Perform this procedure if a standby installer node or relay node exists in the system and you have already installed the OS.

```
# pxinst sync all
...
[INFO] INST 1805 pxinst File replication is complete.
```

2.3.10 Installing the OS

2.3.10.1 Order of Installation

If the installation destination nodes include the compute cluster sub management node (including node groups), you need to install the OS in the following order (*). If they do not include the compute cluster sub management node, you can install the OS on all nodes at the same time, irrespective of whether they are the PRIMERGY server or FX server.

1. Installation on nodes other than nodes under the compute cluster sub management node (PRIMERGY server compute node, FX server)
2. Installation on nodes under the compute cluster sub management node (PRIMERGY server compute node, FX server)

(*) Install the OS in this order because the nodes under the compute cluster sub management node use the sub management node as a relay node (installation sub server for distributing load). For details on the organized hierarchy for OS installation, see "Basic Configuration of Installation Function" in the "Job Operation Software Administrator's Guide for System Management."

Note

If you installed the FX server in step-by-step, you must configure the FX server hardware for installation. Follow the steps in "2.3.9.2.2 Configuring Hardware Control [SMM]" and "2.3.9.2.3 Checking Hardware Control Settings [SMM]".

2.3.10.2 Checking the States of Services Required for OS Installation [SMM]

Use the check subcommand in the pxinst command to check whether the services required for the installation function to work are running. For new system installation, specify the -m option to check only the active system management node.

[Example of checking the active installer nodes in new system installation]

```
# pxinst check -m ipaddr
Installer node[s]           1/1         [OK]
[INFO] INST 1806 pxinst Service state check is complete.
```

ipaddr: IP address of the active system management node (management network)

[Example of checking all the installer nodes, relay nodes, and boot I/O nodes]

```
# pxinst check
Installer node[s]          2/2         [OK]
Sub-Installer node[s]    10/10         [OK]
Boot IO node[s]        100/100         [OK]
[INFO] INST 1806 pxinst Service state check is complete.
```

2.3.10.3 Installing the OS on the PRIMERGY server [SMM]

Power on the target node to start OS installation. Use the papwrctl command, ipmitool, etc. of the system management function to power on the node, or power it on manually. To remotely power on the node, you need to set the IPMI IP address on the BIOS beforehand. Select an appropriate way to power on the node according to the BIOS settings.

If PXE boot is not set as a priority in the BIOS boot order, execute the command for PXE boot of the node.
For BIOS boot, use ipmitool because you may not be able to PXE boot with the papwrctl command.

[Example of powering on using the papwrctl command]

Specify --boottype pxe to PXE boot the node.

```
# papwrctl -c clstname --nodetype SMM,CCM,CCS,LN --model PG -a --boottype pxe on
```

clstname: Cluster name

[Example of powering on using ipmitool]

Set the action for the next start time to PXE boot.

- BIOS boot

```
# ipmitool -I lanplus -U user_name -P password -H ipaddr chassis bootdev pxe
```

user_name: IPMI user name
password: IPMI password
ipaddr: IPMI IP address

- UEFI boot

```
# ipmitool -I lanplus -U user_name -P password -H ipaddr chassis bootdev pxe options=valid,efiboot
```

Power on the node.

```
# ipmitool -I lanplus -U user_name -P password -H ipaddr chassis power on
```

2.3.10.4 Installing the OS on the FX server [SMM]

Power on the boot I/O node to install the OS on the FX server in units of BoBs. Use the papwrctl command to power on the node.

The following examples show powering on the FX server.

[Example of powering on every FX server in a cluster]

```
# papwrctl -c clstname --nodetype BIO --model FT -a --boottype pxe on
```

clstname: Cluster name

[Example of powering on only a specific boot group]

```
# papwrctl -c clstname --nodetype BIO --model FT --bootgrp bootgid --boottype pxe on
```

clstname: Cluster name
bootgid: Boot group ID

2.3.10.5 Confirming the Completion of Installation [SMM]

You can see the progress of installation with the pxinst status command. Installation has completed where OPERATION shows "ImgProvision" (based on an image), STATUS shows "Completed," and STATE shows "Succeeded." If STATE shows "Failed," installation failed. In this case, you need to see the "Job Operation Software Troubleshooting" and take action based on the message displayed at DETAIL. For the FX server, HOSTNAME shows the host name of a compute node that also serves as a boot I/O node.

The following example displays the status for installation completion.

```
# pxinst status show
HOSTNAME       OPERATION      STATUS        STATE        DETAIL
```

```
--------        ---------       ------          -----         ------
host01          ImgProvision    Completed       Succeeded     -
host02          ImgProvision    InProgress      Installing    -
host03          ImgProvision    Completed       Failed        detail
```

detail: Detailed information

Note

If you have the standby system management node or the standby compute cluster sub management node, note the following.

If the standby system management node or the standby compute cluster sub management node is not started (include during OS installation), the following message may be output to DETAIL as the state of the target node to be installed.

```
# pxinst status show
HOSTNAME        OPERATION       STATUS        STATE         DETAIL
--------        ---------       ------        -----         ------
hostname        ImgProvision    Completed     Failed        Update PXE configuration file failed. (*)
(detail) Please change default label to "localdisk" manually.
```

detail: Detailed error message
A new line starts just after the (*) part because of the space limitation of this document, but only one line is actually displayed.

If STATUS is InProgress, the OS installation is still in progress and wait for Completed.
If STATUS is Completed, STATE displays Failed because the configuration files failed to synchronize after a successful OS installation. After confirming that the standby system management node and the standby compute cluster sub management node are started, execute "pxinst sync all" command to synchronize the settings. Also, even if this command is successful, the Failed display of STATE does not disappear, but recovery is complete. To clear the Failed display above, execute "pxinst status clear" with the hostname to clear the display of the status for installation complete.

See

For details on the function for checking installation completion, including the items displayed at the command execution time, see "Checking the Installation Status" in the "Job Operation Software Administrator's Guide for System Management."

2.3.10.6 Starting Up a Diskless Node [SMM]

If the installation destination nodes include the FX server, confirm that installation on the boot I/O node has completed before using the papwrctl command to power on diskless nodes (global I/O node, storage I/O node, and compute node). The OS can run when the nodes start up because installation on the diskless nodes has already been done, at the same time as installation on the boot I/O node.

[Example of powering on all the diskless nodes in a cluster]

```
# papwrctl -c clstname --nodetype CN --model FT --excludetype BIO -a on
```

clstname: Cluster name

[Example of powering on only a specific boot group]

```
# papwrctl -c clstname --nodetype CN --model FT --excludetype BIO --bootgrp bootgid on
```

clstname: Cluster name
bootgid: Boot group ID

2.3.11 Reflecting Cluster Configuration Information [SMM]

Apply the Job Operation Software packages and set the initial values on a node where the OS is already installed.
For these settings, execute the pxsystemadm command with the --set option. The cluster configuration information must be reflected in the following order (*):

1. Compute cluster management node (for compute clusters), storage cluster management node (for storage clusters)

2. Compute cluster sub management node (for a cluster with node groups), nodes that do not belong to node groups
3. Node under the compute cluster sub management node

You can reflect cluster configuration information simultaneously for sequentially numbered targets.

(*) Compute cluster management nodes and compute cluster sub management nodes are responsible for propagating information to the upper levels of the system monitoring software abnormality monitoring hierarchy. Therefore, if the cluster configuration information is not reflected in the above order, the software abnormality monitoring is not performed on the target node. For details of software error monitoring, refer to "Software Abnormality Monitoring" in "Job Operation Software Administrator's Guide System Management."

1. Executing the pxsystemadm command

```
# pxsystemadm --set -c clstname
Do you really want to continue (y/n)? y
[INFO] PSM 1082 pxsystemadm The setting was completed.
```

clstname: Cluster name
Perform this step on all clusters.

Note

- While the pxsystemadm command is running, yum is run remotely on the target node to apply the Job Operation Software packages. While the pxsystemadm command is running, do not run commands that manipulate rpmdb, such as the rpm or yum commands. Also, do not run the pachecker command because it internally references rpmdb.
- If the nodes other than the system management nodes are redundant, this procedure starts the representative IP address of the management network.
  If the management network is restarted or the IP address is changed after the procedure is performed, the representative IP address of the management network will be stopped. Restart the node to start the representative IP address of the management network.

Information

In step-by-step FX server installation, for example, the system may not have some of the hardware. In this case, use the scope option to execute the command in only the range of existing hardware.

The following example executes the pxsystemadm command on nodes whose boot group IDs range from 0x0101 to 0x0103 in a compute cluster.

```
# pxsystemadm --set -c cluster1 --bootgrp 0x0101-0x0103
Do you really want to continue (y/n)? y
[INFO] PSM 1082 pxsystemadm The setting was completed.
```

After installing a node, execute the pxsystemadm again on the installed node.

See

Any node that is already powered off, for example, may fail to reflect the information. In such cases, an error list file is output to the current directory.

```
[ERR.] PSM 0020 pxsystemadm There are nodes that failed in the setting.(filename)
```

filename: Name of the error list file output to the current directory
If an error list file is output, check the file contents, and eliminate the cause of the error. Then, re-execute the command on the node that failed to reflect the information.
For details on troubleshooting based on an error message, see "Failed Reflection of Cluster Configuration Definitions" in the "Job Operation Software Troubleshooting."

2. Confirming the completion of settings

   Use the -v option in the pashowclst command to confirm the completion of settings for nodes and the start of services.

   Confirm that the STATUS column for the nodes show "Running" or "Init." If the STATUS column of the active compute cluster management node does not change from "Stopped," check and, if necessary, revise the setting files (pxsystem.conf, papwr.conf) in "2.3 Building a Cluster."

```
# pashowclst -v
[ CLST: compute ]
NODE        NODETYPE  STATUS   REASON  PWR_STATUS  ARCH_STATUS  SRV_STATUS
0x00000001  SMM       Running  -       on          -            PWRD(o)
0x00000002  SMM       Running  -       on          -            PWRD(o)
...
[ CLST: storage ]
NODE        NODETYPE  STATUS   REASON  PWR_STATUS  ARCH_STATUS  SRV_STATUS
...
```

   Note

   - If you install a new system that uses FEFS, the STATUS column of the node is not "Running" until you configure FEFS as described in "2.4.4 Performing Job Operation-related Work". When you check that system, confirm the operational statuses of non-FEFS services in the SRV_STATUS column of the node are "(o)" at this point
     The following example displays compute nodes. The FEFS service is shown as "(b)".

```
0x01010004  CN        Init     -       os-running  ICC_Running  PLE(o),NRD(o),FEFS(b),PWRD(o)
```

   - If you add a compute cluster, the STATUS column of the active compute cluster management node is not "Running" until you configure MariaDB as described in "2.4.3 Performing MariaDB-related Work for Job Operations". Also, when that uses FEFS, the STATUS column is not "Running" until you configure FEFS as described in "2.4.4 Performing Job Operation-related Work". When you check that system, confirm the operational statuses of PJM and FEFS services in the SRV_STATUS column of the active compute cluster management node are "(b)" and the other services are "(o)" at this point.
     The following example displays the active compute cluster management node.

```
0xFFFF0001  CCM       Init     -       on          -
PJM(b),PLE(o),MRD(o),SRD(o),FEFS(b),PWRD(o)
```

3. Executing the pxinst command

   Execute the sync subcommand of the pxinst command to synchronize the configuration files created in "1. Executing the pxsystemadm command" to the standby system management node and the compute cluster sub management node.

```
# pxinst sync all
...
[INFO] INST 1805 pxinst File replication is complete.
```

4. Configuring Settings for Power Control

   The --set option of the papwradm command reflects the power control settings on the standby system management nodes. This step is not necessary if you are not installing a new system with a standby system management node. The password entered in the configuration file /etc/opt/FJSVtcs/papwr.conf is encrypted, so before performing any of the following operations, reenter the password referring to "Creating the configuration file /etc/opt/FJSVtcs/papwr.conf" in "2.3.8 Configuring Settings for Power Control [SMM]".

```
# papwradm --set
[WARNING]
Do you really want to continue (y/n)y
[INFO] PSM 0082 papwradm The processing of the configuration file was completed.
```

5. Disabling Automatic Failover

   The --set-failover option of the paclstmgr command reflects disabling automatic failover on the standby system management node. If the following nodes are redundant, run the paclstmgr command to disable automatic failover:

- System management node
- Compute cluster management node
- Compute cluster sub management node
- Storage cluster management node

```
# paclstmgr -c clstname -n nodeid --set-failover disable
```

clstname: Cluster name
nodeid: Node ID of a redundantly configured node (Either of the active and standby nodes would be okay.)

Confirm failover configuration information by using the paclstmgr command.
Confirm that the FAILOVER column for the nodes in the cluster is "disable".

```
# paclstmgr -c clstname -n nodeid --set-failover list
[ CLST: clstname ]
NODE        NODETYPE  FAILOVER
0x00000001  SMM       disable
0x00000002  SMM       disable
```

clstname: Cluster name
nodeid: Node ID of a redundantly configured node (Either of the active and standby nodes would be okay.)

2.3.12 Configuring Settings After OS Installation

Configure the following settings for a node after completing OS installation.

2.3.12.1 Configuring kdump [SMM/Target Node]

Configure kdump on a node other than the FX server.
If you have a support desk contract, you can get a dump assist tool (FJSVdumptools) and configure kdump.
If a Support Desk is not available or if you are not using the above tools, configure the settings according to your OS documentation.

Set the system operation to halt after collecting memory dump.

- If you use a dump support tool (FJSVdumptools) to configure kdump, unmount the dump area before configuring. Also, remove the dump save area from /etc/fstab.
  - Unmounting the dump save area

    ```
    # pmexe -c clstname --model PG "umount dumpsavearea"
    ```

    clstname: Cluster name
    dumpsavearea: Mount point of dump save area
  - Removing dump save area from /etc/fstab

    ```
    # pmexe -c clstname --model PG "sed -i '/dumpsavearea/d' /etc/fstab"
    ```

    clstname: Cluster name
    dumpsavearea: Mount point of dump save area
- By setting this procedure, even if memory dump collection is completed, the node will not be restarted or powered down. Check the condition of the panicked node before attempting maintenance or incorporation into operation.

2.3.12.2 Applying the System Information Gathering Tool [SMM/Target Node]

Apply the Gather System Information tool (FJSVsnap) to a node other than the FX server.

If you have a support desk contract, obtain the system information collection tool and follow the documentation of the system information collection tool to apply.

2.3.12.3 Setting the Core Dump File Destination

Sets the destination for the core dump file.
This configuration is not required for RHEL 7 or earlier nodes.

1. Creating a Configuration File

   Create a kernel parameter configuration file. You can use arbitrary name for the file.
   The following example creates a file with the file name core.conf:

```
# vi /etc/sysctl.d/core.conf
kernel.core_pattern=core
```

2. Reflecting settings

```
# sysctl -p /etc/sysctl.d/core.conf
kernel.core_pattern = core
```

   Information

   Step 1 and Step 2 can also be set in batch using the collective execution of commands. In batch configuration, specify scope options to target RHEL 8 or later nodes.

3. Creating a Check Definition File for Checking the Software Environment

   Create a check definition file in "2.5 Checking the Software Environment [SMM]" to ensure that the core dump file destinations match on all RHEL installed nodes.

   The following is an example of creating a check definition file to ensure that the core dump file destinations match.
   When this item is set, the output destination setting of the core dump matches the setting of RHEL 7 or earlier, so the whole check is performed.

   The check definition file is written in yaml format.

```
Command:
-
  Name: Core_configuration
  Type: Diff
  ExecCmd: cat /proc/sys/kernel/core_pattern
  NodeType: ALL
```

   The check definition file you create is enabled by placing it in /etc/opt/FJSVtcs/plugin/psm/check/ on the active system management node.
   The following example copies the check definition file check_coredump_setting created under the directory /tmp to /etc/opt/FJSVtcs/plugin/psm/check/.

```
# cp /tmp/check_coredump_setting /etc/opt/FJSVtcs/plugin/psm/check/
```

2.3.12.4 CPU Frequency Table Settings [SMM]

Perform the package installation to configure the CPU frequency table for the FX server node.

1. Isolation FX server from operation

   Isolate the FX server with the --disable option of the paclstmgr command.

```
# paclstmgr -c clstname -a --model FT --disable
```

   clstname: Cluster name

2. Confirming the status of the operation isolation

   Confirm the state of the node with the -v option of the pashowclst command.
   Confirm that the STATUS column of the FX server transitioned to "Disable" and the REASON column of the FX server transitioned to "Manual".

```
# pashowclst -v -c clstname
[ CLST: clstname ]
NODE        NODETYPE STATUS   REASON           PWR_STATUS   ARCH_STATUS   SRV_STATUS
...
0x01010001  BIO,CN   Disable  Manual           os-running   ICC_Running
PLE(o),SRD(o),NRD(o),FEFS(b),PWRD(o)
0x01010002  SIO,CN   Disable  Manual           os-running   ICC_Running
PLE(o),NRD(o),FEFSSR(b),FEFS(b),PWRD(o)
0x01010003  GIO,CN   Disable  Manual           os-running   ICC_Running
PLE(o),NRD(o),FEFSSR(b),FEFS(b),PWRD(o)
0x01010004  CN       Disable  Manual           os-running   ICC_Running
PLE(o),NRD(o),FEFS(b),PWRD(o)
0x01010005  CN       Disable  Manual           os-running   ICC_Running
PLE(o),NRD(o),FEFS(b),PWRD(o)
...
```

   clstname: Cluster name

3. Transitioning to software maintenance mode

   Switch the FX server into software maintenance mode with the --soft-mainte option of the paclstmgr command. Specify the --service-stop option to stop the service of the Job Operation Software.

```
# paclstmgr -c clstname -a --model FT --soft-mainte --service-stop
```

   clstname: Cluster name

4. Confirming software maintenance mode

   Use the -v option in the pashowclst command to check the node status.
   Confirm that the REASON column of the FX server is transitioned to "SoftMaintenance" and that the SRV_STATUS column of the FX server is transitioned to a hyphen (-).

```
# pashowclst -v -c clstname
[ CLST: clstname ]
NODE        NODETYPE STATUS   REASON           PWR_STATUS   ARCH_STATUS   SRV_STATUS
...
0x01010001  BIO,CN   Disable  SoftMaintenance  os-running   ICC_Running   -
0x01010002  SIO,CN   Disable  SoftMaintenance  os-running   ICC_Running   -
0x01010003  GIO,CN   Disable  SoftMaintenance  os-running   ICC_Running   -
0x01010004  CN       Disable  SoftMaintenance  os-running   ICC_Running   -
0x01010005  CN       Disable  SoftMaintenance  os-running   ICC_Running   -
...
```

   clstname: Cluster name

5. Applying the frequency table file rpm package

   Apply the frequency table file rpm package to the FX server.

Information

The rpm package of the frequency table files is included on the CD media as a hardware attachment.

   a. Confirm the path of the directory where the package will be stored

      Use the repo subcommand of the pxinst command to confirm the FX server repository to which you want to register packages.

```
# pxinst repo list
REPONAME      PATH                              PKGDIR     OS      VERSION    ARCH
--------      ----                              ------     --      -------    ----
reponame      repopath                          dirname    os      version    arch
...
```

reponame: Repository name
repopath: Repository path
dirname: Package deployment directory in the repository
os: OS type
version: OS version
arch: Architecture

Confirm the repository row used by the node to which you are applying the package.

"repopath/dirname" is the path where the package will be stored.

Information

Depending on the OS version, PKGDIR displays multiple directories separated by commas.

```
# pxinst repo list
REPONAME  PATH            PKGDIR                                OS       VERSION  ARCH
--------  ----            ------                                --       -------  ----
rhel80    /repos/rhel80   AppStream/Packages,BaseOS/Packages   rhel     8.0      aarch64
```

In the example output above, "/repos/rhel80/AppStream/Packages" or "/repos/rhel80/BaseOS/Packages" is the path where the packages will be stored.

If you want to store packages of the Job Operation Software or packages that the user wants to add, store them in the directory shown at the top "/repos/rhel80/AppStream/Packages".

When storing OS update packages, be sure to store them like the distributor's public repository.

b. Adding the Package

Store the package in the directory you confirmed.

```
# cp pkgname repopath/dirname
```

pkgname: Package name
repopath: Repository path
dirname: Package deployment directory in the repository

c. Updating the Repository

Update the repository using the repo subcommand of the pxinst command.
This operation refreshes the metadata in the repository.

```
# pxinst repo update -R reponame
[INFO] INST 0800 pxinst repo completed successfully.(update repository)
```

reponame: Repository name

d. Synchronizing the Repository

Reflect the repository changes to the standby system management node and the compute cluster sub management nodes.
If a standby system management node or a compute cluster sub management node does not exist in the system, this procedure is not required.

```
# pxinst sync repo
...
[INFO] INST 1805 pxinst File replication completed.
```

e. Deleting Cached Information

Use the pmexe command to delete the cached information in advance on the nodes to which the package is applied.
The following is an example of running the yum command with the pmexe command on all FX servers in a cluster:

```
# pmexe --chroot -c clstname --model FT "yum clean all"
```

clstname: Cluster name

f. Applying the Package

Use the pmexe command to apply the package to the target node.

The following example shows how to use the yum command to apply the package to all FX servers in a cluster.

```
# pmexe --chroot -c clstname --model FT "yum -y install pwrdrv_conf-typename"
```

clstname: Cluster name
typename: Device type name

6. Restarting the Power Driver

Restart the FX server power driver using the pmexe command.

```
# pmexe -c clstname --model FT "systemctl restart FJSVxoshpcpwr"
```

clstname: Cluster name

7. Recovering From Software Maintenance Mode

Use the --recover option of the paclstmgr command to recover the FX server from software maintenance mode.
At this time, specify --service-restart to restart the service of the Job Operation Software.

```
# paclstmgr -c clstname -a --model FT --recover --service-restart
```

clstname: Cluster name

8. Confirming the recovery from software maintenance mode

Use the -v option in the pashowclst command to check the status of the FX server.

Confirm that the node status is the same as that confirmed in step 2 in "2.3.11 Reflecting Cluster Configuration Information [SMM]."

9. Creating a Check Definition File for Checking the Software Environment

Create a check definition for checking that the frequency table file rpm package is applied on all FX servers at the procedure "2.5 Checking the Software Environment [SMM]".

The following is an example of creating a check definition file to perform a check of the frequency table file rpm package.

```
Package:
-
   Package: pwrdrv_conf-typename
   Model: FT
   Action: "Install pwrdrv_conf package"
```

typename: Device type name

The check definition file you create is enabled by placing it in /etc/opt/FJSVtcs/plugin/psm/check/ on the active system management node.
The following example copies the check definition file check_pwrdrvconf created under the directory /tmp to /etc/opt/FJSVtcs/plugin/psm/check/.

```
# cp /tmp/check_pwrdrvconf /etc/opt/FJSVtcs/plugin/psm/check/
```

2.3.12.5 Configuring OFED [SMM/Target Node]

Apply OFED to the nodes that use InfiniBand. For details on how to configure OFED, see the OFED installation guide.

2.3.12.6 Configuring GPUs [SMM/Target Node]

Configure settings on compute nodes equipped with GPUs (NVIDIA V100, A100 or H100). For details on how to configure GPUs, see "Appendix A Configuring GPUs."

2.3.12.7 Configuring Singularity [SMM/Target Node]

Singularity is container virtualization software for HPC.
Configure settings on compute nodes that use Singularity. For details on how to configure Singularity, see "Appendix B Configuring Singularity."

If Singularity is not used, this procedure is unnecessary.

2.3.12.8 UDP Receive Buffer Settings [Standby SMM]

If you installed the OS on the standby system management node, configure the UDP receive buffer on the standby system management node. The active system management node is not required configuration in this procedure because it have already been configured in "2.2.1.2 Performing the Work After OS Installation [SMM]."

1. Creating a Configuration File

   Create a kernel parameter configuration file. You can use arbitrary name for the file.
   The following example creates a file with the file name udp.conf:

```
# vi /etc/sysctl.d/udp.conf
net.core.rmem_default=8388608
net.core.rmem_max=8388608
```

2. Reflecting settings

```
# sysctl -p /etc/sysctl.d/udp.conf
net.core.rmem_default = 8388608
net.core.rmem_max = 8388608
```

3. Checking the result of reflecting settings

```
# cat /proc/sys/net/core/rmem_default
8388608
# cat /proc/sys/net/core/rmem_max
8388608
```

2.3.13 Enabling Disabled Automatic Failover [SMM]

To enable automatic failover on the nodes where automatic failover has been disabled in "2.3.6 Disabling Automatic Failover [SMM]," execute the paclstmgr command.

```
# paclstmgr -c clstname -n nodeid --set-failover enable
```

clstname: Cluster name
nodeid: Node ID

Use the paclstmgr command to check failover setting information.
Confirm that the FAILOVER column shows "enable" for the nodes in the cluster.

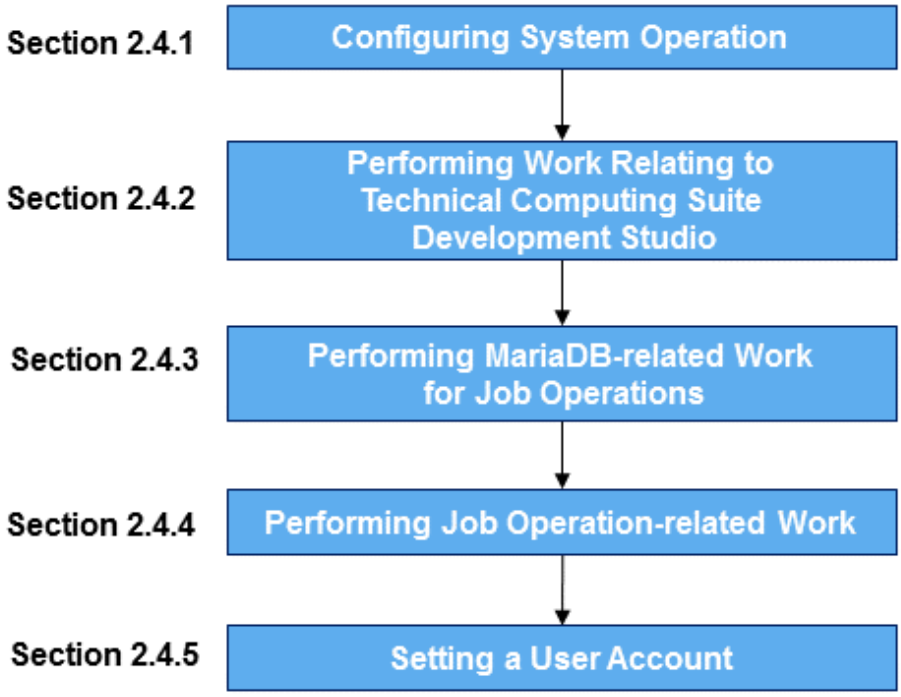
```
# paclstmgr -c clstname -n nodeid --set-failover list
[ CLST: clstname ]
NODE        NODETYPE      FAILOVER
0x00000001  SMM           enable
0x00000002  SMM           enable
...
```

clstname: Cluster name
nodeid: Node ID (Either of the active and standby nodes would be okay.)

2.4 Configuring Settings for System Operation

This section describes procedures for configuring various settings for system operation.

Figure 2.4 Workflow for Settings for System Operation



2.4.1 Configuring System Operation

Configure the monitoring function and log management function for the system as required.

2.4.1.1 Configuring the System Monitoring Function [SMM]

To change the monitoring interval and fault detection timeout value of the system monitoring function, configure the system monitoring function. For details, see "Setting for the System Monitoring" in the "Job Operation Software Administrator's Guide for System Management."

2.4.1.2 Configuring the Log Management Function [SMM]

To monitor the log files of each node or collect them at a single location, configure the log management function. For details, see "Overview of Log Management Function" and "Details of the Log Management Function" in the "Job Operation Software Administrator's Guide for System Management."

2.4.1.3 Configuring a shared_disk [SMM]

For the use of a shared_disk on the system management node, or for the compute cluster management node in a redundant configuration, perform the following procedure to configure a shared_disk.
If the shared_disk has already been defined in "2.3.4 Registering Cluster Configuration Information [SMM]," this procedure is unnecessary.

1. Backing up a shared_disk

   Back up the /var/opt/FJSVtcs/shared_disk directory on the target node.
   The following example uses the pmexe command to back up the /var/opt/FJSVtcs/shared_disk directory on the target node to the /var/tmp directory.

```
# pmexe -c clstname --nodetype nodetype --active "cp -rp /var/opt/FJSVtcs/shared_disk /var/tmp/"
```

2. Editing and registering the cluster configuration definition file [SMM]

   Configure the shared_disk with the cluster configuration management function. Check the by-id of the shared_disk, and create a cluster configuration definition file at any location on the active system management node. The following example configures a shared_disk on both the system management node and compute cluster management node.
   For details on the sections and items in a cluster configuration definition file, see "pxsystem.conf file" in the "Job Operation Software Command Reference."

```
# vi /var/tmp/pxsystem.conf
Operation {
    Type = "add-node"
}
Cluster {
    ClusterName = "cluster1"                                    # (1)
    HA {
        Node {
            NodeType = "SMM"                                    # (2)
            Mount {
                Device = "/dev/disk/by-id/xxxxxxxxxxxxxxxxx"    # (3)
                MountPoint = "/var/opt/FJSVtcs/shared_disk"     # (4)
                FSType = "ext4"                                 # (5)
                OptList = "rw"                                  # (6)
            }
        }
        Node {
            NodeType = "CCM"
            Mount {
                Device = "/dev/disk/by-id/xxxxxxxxxxxxxxxxx"
                MountPoint = "/var/opt/FJSVtcs/shared_disk"
                FSType = "ext4"
                OptList = "rw"
            }
        }
    }
}
```

   (1) Cluster name
   (2) Node type
   (3) by-id of a shared_disk
   (4) Mount point
   (5) File system type (Example: ext4)
   (6) Option to specify for mounting on the shared_disk

3. Registering cluster configuration information [SMM]
   Use the --create option in the pxsystemadm command to get the system management node to reflect the cluster configuration definition file contents.

```
# pxsystemadm --create filename
Operation/Type is "add-node".
Do you really want to continue (y/n)? y                  (*)
[INFO] PSM 1082 pxsystemadm The setting was completed.
```

   (*) This prompts for confirmation before execution.

   filename: Path to the created cluster configuration definition file

4. Checking settings [SMM]
   Use the -a option in the pashowclst command to confirm that the configuration is reflected.
   Confirm that the [HA] column displays the contents written in the cluster configuration definition file in step 2.

```
# pashowclst -a
...
```

5. Reflecting cluster configuration information [SMM]

   In the same way as step 1 in "2.3.11 Reflecting Cluster Configuration Information [SMM]," get nodes in the system to reflect the cluster configuration information registered with the system management node.

6. Mounting the shared_disk [SMM, CCM]

```
# mount -t fstype /dev/disk/by-id/xxxxxxxxxxxxxxxx  /var/opt/FJSVtcs/shared_disk
```

   fstype: File system type

   Use the pmexe command from the system management node to execute the mount command on the target node.

```
# pmexe -c clstname --nodetype nodetype --active \
"mount -t fstype /dev/disk/by-id/xxxxxxxxxxxxxxxx  /var/opt/FJSVtcs/shared_disk"
```

   clstname: Cluster name
   nodetype: Node type
   fstype: File system type

7. Restoring the shared_disk

   Restore the data that was backed up in step 1 to the original location. The following example uses the pmexe command to restore the data backed up in the example shown in step 1.

```
# pmexe -c clstname --nodetype nodetype --active "cp -rp /var/tmp/shared_disk/* /var/opt/FJSVtcs/
shared_disk/"
```

2.4.1.4 Package Removal for System Noise Reduction [SMM]

Remove Red Hat Subscription Manager (RHSM) packages that cause system noise on the FX server.

Note

If you remove these packages, RHSM will be unusable. Do not remove when managing with RHSM on the FX server.

1. Isolation FX Server from Operation

   Isolate the FX server with the --disable option of the paclstmgr command.

```
# paclstmgr -c clstname -a --model FT --disable
```

   clstname: Cluster name

2. Confirming the Status of the Operation Isolation

   Confirm the state of the node with the -v option of the pashowclst command. Confirm that the STATUS column of the FX server transitioned to "Disable" and the REASON column of the FX server transitioned to "Manual".

```
# pashowclst -v -c clstname
[ CLST: clstname ]
NODE        NODETYPE STATUS   REASON           PWR_STATUS   ARCH_STATUS   SRV_STATUS
...
0x01010001  BIO,CN   Disable  Manual           os-running   ICC_Running
PLE(o),SRD(o),NRD(o),FEFS(b),PWRD(o)
0x01010002  SIO,CN   Disable  Manual           os-running   ICC_Running
PLE(o),NRD(o),FEFSSR(b),FEFS(b),PWRD(o)
0x01010003  GIO,CN   Disable  Manual           os-running   ICC_Running
```

```
PLE(o),NRD(o),FEFSSR(b),FEFS(b),PWRD(o)
0x01010004  CN       Disable  Manual           os-running   ICC_Running
PLE(o),NRD(o),FEFS(b),PWRD(o)
0x01010005  CN       Disable  Manual           os-running   ICC_Running
PLE(o),NRD(o),FEFS(b),PWRD(o)
...
```

clstname: Cluster name

3. Transitioning to Software Maintenance Mode

   Switch the FX server into software maintenance mode with the --soft-mainte option of the paclstmgr command. Specify the --servicestop option to stop the service of the Job Operation Software.

```
# paclstmgr -c clstname -a --model FT --soft-mainte --service-stop
```

   clstname: Cluster name

4. Confirming Software Maintenance Mode

   Use the -v option in the pashowclst command to check the node status. Confirm that the REASON column of the FX server is transitioned to "SoftMaintenance" and that the SRV_STATUS column of the FX server is transitioned to a hyphen (-).

```
# pashowclst -v -c clstname
[ CLST: clstname ]
NODE        NODETYPE STATUS   REASON           PWR_STATUS   ARCH_STATUS   SRV_STATUS
...
0x01010001  BIO,CN   Disable  SoftMaintenance  os-running   ICC_Running   -
0x01010002  SIO,CN   Disable  SoftMaintenance  os-running   ICC_Running   -
0x01010003  GIO,CN   Disable  SoftMaintenance  os-running   ICC_Running   -
0x01010004  CN       Disable  SoftMaintenance  os-running   ICC_Running   -
0x01010005  CN       Disable  SoftMaintenance  os-running   ICC_Running   -
...
```

   clstname: Cluster name

5. Deleting Cached Information

   Use the pmexe command to delete the cached information in advance on the nodes to which the package is applied.

```
# pmexe --chroot -c clstname --model FT "yum clean all"
```

   clstname: Cluster name

6. Removing RHSM Packages

   Use the pmexe command to remove RHSM packages. Package dependencies may also remove non-RHSM packages, but this is not a problem.

```
# pmexe --chroot -c clstname --model FT \
"yum -y remove subscription-manager subscription-manager-cockpit \
dnf-plugin-subscription-manager"
```

   clstname: Cluster name

7. Recovering From Software Maintenance Mode

   Use the --recover option of the paclstmgr command to recover the FX server from software maintenance mode. At this time, specify --service-restart to restart the service of the Job Operation Software.

```
# paclstmgr -c clstname -a --model FT --recover --service-restart
```

   clstname: Cluster name

8. Confirming the Recovery from Software Maintenance Mode

   Use the -v option in the pashowclst command to check the status of the FX server. Confirm that the STATUS column for the nodes show "Running" or "Init."

2.4.2 Performing Work Relating to Technical Computing Suite Development Studio [SMM]

2.4.2.1 Performing Work on the Login Node [SMM]

To use the Technical Computing Suite Development Studio on the login node, the following packages must be applied:

- gcc-c++
- libstdc++-devel
- elfutils-devel
- elfutils-libelf-devel
- zlib-devel
- openssl-devel
- libatomic
- ncurses-compat-libs (If the login node is using RHEL8)

In addition, if you use the integrated development environment in the Technical Computing Suite Development Studio, you must apply the following packages to the login node:

- git
- perl-Data-Dumper

Perform this work on the system management node.

1. Deleting cached information

   Before applying the packages to the target node, use the pmexe command to delete cached information on the node.
   To delete the cached information, follow the instructions for the package manager used.

   The following example uses the pmexe command to execute the yum command on the login node.

   ```
   # pmexe -c clstname --nodetype LN "yum clean all"
   [INFO] PSM 5082 pmexe The execution result file was output.(resultfile)
   ```

   clstname: Cluster name
   resultfile: Execution results file

   Check the execution result file and confirm that no errors are output in the command results for each node.

2. Applying the packages

   Use the pmexe command to apply the packages on the target node.
   To apply the packages, follow the instructions for the package manager used.
   The following example applies a package with the yum command to all the login nodes in a cluster.

   ```
   # pmexe -c clstname --nodetype LN "yum -y install pkgname"
   [INFO] PSM 5082 pmexe The execution result file was output.(resultfile)
   ```

   clstname: Cluster name
   pkgname: Name of the package to apply
   resultfile: Execution results file

   Check the execution result file and confirm that no errors are output in the command results for each node.

2.4.2.2 Creating a Cross-Compilation Environment [SMM]

Application programs that run on the FX server are created using Technical Computing Suite Development Studio and cross-compiled on the login node. To cross-compile a program, a Devkit including the required headers and libraries must be installed in advance on the login node.

Perform the following procedure on the active system management node.

1. Obtaining a Devkit build script

   Obtain a Devkit build script and place it in any directory on the active system management node.
   The Devkit build script is included in the Technical Computing Suite DVD. The file name is xosmakedevkit.tar.gz.

   Information

   A Devkit build script creates a Devkit in RPM package format, based on the contents of the root file system of an FX server node.

2. Transferring the Devkit build script

   The Devkit build script runs on an FX server node. For this reason, use the pmscatter command to transfer the archive of the Devkit build script to any boot I/O node. The following example transfers the archive (file name: xosmakedevkit.tar.gz) of a Devkit build script located in the /tmp directory on the active system management node to the /root directory on an arbitrary boot I/O node.

   ```
   # pmscatter -c clstname -n nodeid /tmp/xosmakedevkit.tar.gz /root
   ```

   clstname: Cluster name
   nodeid: Node ID of an arbitrary boot I/O node

3. Extracting the Devkit build script

   Use the pmexe command to extract the archive.

   ```
   # pmexe -c clstname -n nodeid "tar xzf /root/xosmakedevkit.tar.gz -C /root"
   ```

   clstname: Cluster name
   nodeid: Node ID of the boot I/O node with the archive transferred in step 2

4. Executing the Devkit build script

   Use the pmexe command to execute the Devkit build script.
   This command will take approximately 30 minutes to complete.

   ```
   # pmexe -c clstname -n nodeid "/root/xosmakedevkit/xosmakedevkit.sh"
   ```

   clstname: Cluster name
   nodeid: Node ID of the boot I/O node with the archive transferred in step 2

   This operation creates the rpm package of the Devkit in the same directory as the build script.

   ```
   /root/xosmakedevkit/FJSVxosdevkit-date-1.el8.x86_64.rpm
   ```

   date: Execution date and time of the Devkit build script. The output is in the YYYYMMDD format.

5. Registering in a repository

   Collect the rpm package created on the FX server in step 4 at the active system management node, and register it with a repository.

   a. Collecting the rpm package

      Use the pmgather command to collect the rpm package at the system management node.
      The following example collects the rpm package created in step 4 at the /tmp directory on the active system management node.

      ```
      # pmgather -c clstname -n nodeid /root/xosmakedevkit/FJSVxosdevkit-
      date-1.el8.x86_64.rpm /tmp
      ```

      clstname: Cluster name
      nodeid: Node ID of the boot I/O node that has the archive transferred in step 2.
      date: Execution date and time of the Devkit build script. The output is in the YYYYMMDD format.

   b. Confirming the path to the package storage directory

      Use the repo subcommand in the pxinst command to confirm the repository for registering the package for the login node.

      ```
      # pxinst repo list
      REPONAME      PATH                              PKGDIR     OS      VERSION    ARCH
      ```

```
--------      ----                          ------     --      -------    ----
reponame      repopath                      dirname    os      version    arch
...
```

reponame: Name of the repository for registering the package
repopath: Path to the repository
dirname: Directory where the package is located in the repository
os: OS type
version: OS version
arch: Architecture

"*repopath*/*dirname*" is the storage path of the package.

Information

Depending on the OS version, PKGDIR displays multiple directories separated by commas.

```
# pxinst repo list
REPONAME  PATH            PKGDIR                                OS       VERSION  ARCH
--------  ----            ------                                --       -------  ----
rhel80    /repos/rhel80   AppStream/Packages,BaseOS/Packages   rhel     8.0      aarch64
```

In the example output above, "/repos/rhel80/AppStream/Packages" is the path where the packages will be stored.

c. Adding the package

Place the package in the directory confirmed in step b.

```
# cp /tmp/FJSVxosdevkit-date-1.el8.x86_64.rpm repopath/dirname
```

date: Execution date and time of the Devkit build script. The output is in the YYYYMMDD format.
repopath: Path to the repository
dirname: Directory where the package is located in the repository

d. Updating the repository

Use the repo subcommand in the pxinst command to update the repository.
This operation updates the metadata of the repository.

```
# pxinst repo update -R reponame
[INFO] INST 0800 pxinst repo completed successfully.(update repository)
```

reponame: Repository name

e. Synchronizing the repository

Reflect the changes in the repository to the standby system management node and compute cluster sub management node.
If the system does not have the standby system management node and compute cluster sub management node, this operation is unnecessary.

```
# pxinst sync repo
...
[INFO] INST 1805 pxinst File replication is complete.
```

6. Installing the Devkit

Install the Devkit on the login node.

a. Deleting cached information

Before applying the package to the target node, use the pmexe command to delete cached information on the node.
To delete the cached information, follow the instructions for the package manager used.

The following example uses the pmexe command to execute the yum command on all the login nodes in a cluster.

```
# pmexe -c clstname --nodetype LN "yum clean all"
[INFO] PSM 5082 pmexe The execution result file was output.(resultfile)
```

clstname: Cluster name
resultfile: Execution results file

Check the execution result file and confirm that no errors are output in the command results for each node.

b. Applying the package

Use the pmexe command to apply the package on the target node.
To apply the package, follow the instructions for the package manager used.

The following example applies a package with the yum command to all the login nodes in a cluster.

```
# pmexe -c clstname --nodetype LN "yum -y install FJSVxosdevkit"
[INFO] PSM 5082 pmexe The execution result file was output.(resultfile)
```

clstname: Cluster name
resultfile: Execution results file

Check the execution result file and confirm that no errors are output in the command results for each node.

2.4.2.3 Performing Work for Compute Nodes That Use Red Hat Enterprise Linux for Scientific Computing [SMM]

To use Technical Computing Suite Development Studio on compute nodes whose OS is Red Hat Enterprise Linux for Scientific Computing, the following packages must be applied to the compute nodes:

- elfutils-devel
- elfutils-libelf-devel
- gcc-c++
- glusterfs-cli
- gperftools-libs
- ipxe-roms-qemu
- libatomic
- libiscsi
- libstdc++-devel
- libvirt
- libvirt-daemon
- libvirt-daemon-config-network
- libvirt-daemon-config-nwfilter
- libvirt-daemon-driver-interface
- libvirt-daemon-driver-lxc
- libvirt-daemon-driver-network
- libvirt-daemon-driver-nodedev
- libvirt-daemon-driver-nwfilter
- libvirt-daemon-driver-qemu
- libvirt-daemon-driver-secret
- libvirt-daemon-driver-storage

- netcf-libs
- openssl-devel
- qemu-img
- qemu-kvm
- qemu-kvm-common
- radvd
- seabios-bin
- seavgabios-bin
- sgabios-bin
- spice-server
- usbredir
- zlib-devel

Copy them from the Red Hat Enterprise Linux Server repository and apply them in the following procedure.

1. Confirming the path to the package storage directory

   Use the repo subcommand in the pxinst command to confirm the repository for registering the Red Hat Enterprise Linux Server packages and Red Hat Enterprise Linux for Scientific Computing packages.

```
# pxinst repo list
REPONAME      PATH                          PKGDIR     OS      VERSION    ARCH
--------      ----                          ------     --      -------    ----
reponame      repopath                      dirname    os      version    arch
...
```

   reponame: Name of the repository for registering the packages
   repopath: Path to the repository
   dirname: Directory where the packages are located in the repository
   os: OS type
   version: OS version
   arch: Architecture

   "*repopath*/*dirname*" is the storage path of the packages.

   ### Information

   Depending on the OS version, PKGDIR displays multiple directories separated by commas.

```
# pxinst repo list
REPONAME       PATH            PKGDIR                                OS       VERSION  ARCH
--------       ----            ------                                --       -------  ----
rhel80         /repos/rhel80   AppStream/Packages,BaseOS/Packages     rhel     8.0      aarch64
```

   In the example output above, "/repos/rhel80/AppStream/Packages" is the path where the packages will be stored.

2. Adding the packages

   Place the packages in the directory confirmed in step 1.

```
# cp repopath_server/dirname_server/pkg-name* repopath_SC/dirname_SC
```

   repopath_server: Repository path for Red Hat Enterprise Linux Server
   dirname_server: Directory where the package is located in the repository for Red Hat Enterprise Linux Server
   pkg-name: Package name

repopath_SC: Repository path for Red Hat Enterprise Linux for Scientific Computing
dirname_SC: Directory where the package is located in the repository for Red Hat Enterprise Linux for Scientific Computing

(The above execution example adds an asterisk to correspond to the ending (version, etc.) of the package name.)

3. Updating the repository

   Use the repo subcommand in the pxinst command to update the repository.
   This operation updates the metadata of the repository.

```
# pxinst repo update -R reponame
[INFO] INST 0800 pxinst repo completed successfully.(update repository)
```

   reponame: Repository name

4. Synchronizing the repository

   Reflect the changes in the repository to the standby system management node and compute cluster sub management node.
   If the system does not have the standby system management node and compute cluster sub management node, this operation is unnecessary.

```
# pxinst sync repo
...
[INFO] INST 1805 pxinst File replication is complete.
```

5. Deleting cached information

   Before applying the packages to the target node, use the pmexe command to delete cached information on the node.
   To delete the cached information, follow the instructions for the package manager used.

   The following example uses the pmexe command to execute the yum command on all the compute nodes in a cluster.

```
# pmexe -c clstname --nodetype CN --model PG "yum clean all"
[INFO] PSM 5082 pmexe The execution result file was output.(resultfile)
```

   clstname: Cluster name
   resultfile: Execution results file

   Check the execution result file and confirm that no errors are output in the command results for each node.

6. Applying the packages

   Use the pmexe command to apply the packages on the target node.
   To apply the packages, follow the instructions for the package manager used.
   The following example applies a package with the yum command to all the compute nodes in a cluster.

```
# pmexe -c clstname --nodetype CN --model PG "yum -y install pkgname"
[INFO] PSM 5082 pmexe The execution result file was output.(resultfile)
```

   clstname: Cluster name
   pkgname: Name of the package to apply
   resultfile: Execution results file

   Check the execution result file and confirm that no errors are output in the command results for each node.

2.4.3 Performing MariaDB-related Work for Job Operations

Configure MariaDB for use by the job operation management function and power management function.

2.4.3.1 Applying MariaDB [SMM]

Apply the following packages to the system management node and login node:

[The system management node or login node is using RHEL7]

- mysql-connector-odbc

- unixODBC

[The system management node or login node is using RHEL8]

- mariadb-connector-odbc
- unixODBC

Also apply the following packages to the compute cluster management node:

[The compute cluster management node is using RHEL7]

- mariadb-devel
- mariadb-libs
- mariadb
- mariadb-server
- mysql-connector-odbc
- unixODBC

[The compute cluster management node is using RHEL8]

- mariadb
- mariadb-server
- mariadb-connector-odbc
- unixODBC

Obtain these packages from the Red Hat Enterprise Linux or other OS media, and apply them in the following procedure.
Perform this work on the system management node.

1. Deleting cached information

   Before applying the packages to the target node, use the pmexe command to delete cached information on the node.
   To delete the cached information, follow the instructions for the package manager used.

   The following example uses the pmexe command to execute the yum command on the compute cluster management node in a cluster.

   ```
   # pmexe -c clstname --nodetype CCM "yum clean all"
   [INFO] PSM 5082 pmexe The execution result file was output.(resultfile)
   ```

   clstname: Cluster name
   resultfile: Execution results file

   Check the execution result file and confirm that no errors are output in the command results for each node.

2. Applying the packages

   Use the pmexe command to apply the packages on the target node.
   To apply the packages, follow the instructions for the package manager used.
   The following example applies a package with the yum command to the compute cluster management node in a cluster.

   ```
   # pmexe -c clstname --nodetype CCM "yum -y install pkgname"
   [INFO] PSM 5082 pmexe The execution result file was output.(resultfile)
   ```

   clstname: Cluster name
   pkgname: Name of the package to apply
   resultfile: Execution results file

   Check the execution result file and confirm that no errors are output in the command results for each node.

2.4.3.2 Configuring MariaDB [CCM]

Configure MariaDB in the following procedure. Perform this procedure as an administrator who has root privileges for the compute cluster management node. Also, for redundancy of the compute cluster management node, the settings must be done on both the active and standby nodes. Configure the active compute cluster management node, and then configure the standby compute cluster management node.

1. Creating /var/opt/FJSVtcs/shared_disk/mysql

   Create the /var/opt/FJSVtcs/shared_disk/mysql directory, which will be the MariaDB database directory.

```
# mkdir /var/opt/FJSVtcs/shared_disk/mysql
# chown mysql:mysql /var/opt/FJSVtcs/shared_disk/mysql
```

2. Editing the /etc/my.cnf configuration file

   [The compute cluster management node is using RHEL7]
   Make the following settings in the MariaDB configuration file /etc/mycnf:

```
[mysqld]
datadir=/var/opt/FJSVtcs/shared_disk/mysql
socket=/var/lib/mysql/mysql.sock
max_allowed_packet=1G
innodb_buffer_pool_size=12G
innodb_log_file_size=3G
innodb_file_per_table
innodb_file_format=Barracuda
```

   [The compute cluster management node is using RHEL8]
   Make the following settings in the MariaDB configuration file /etc/my.cnf.d/mariadb-server.cnf:

```
[mysqld]
datadir=/var/opt/FJSVtcs/shared_disk/mysql
socket=/var/lib/mysql/mysql.sock
max_allowed_packet=1G
innodb_buffer_pool_size=12G
innodb_log_file_size=3G
innodb_file_per_table
innodb_file_format=Barracuda
```

   Information

   You can tune innodb_buffer_pool_size and innodb_log_file_size to any value to improve performance. However, to avoid affecting the operation of the job operation software, specify values for these items that are greater than or equal to the above example.

3. Starting mariadb.service
   Start mariadb.service. After you start the service, confirm that it started successfully.
   mariadb.service autostart configuration can be either "enabled" or "disabled".

```
# systemctl enable mariadb.service
Created symlink from /etc/systemd/system/multi-user.target.wants/mariadb.service to /usr/lib/
systemd/system/mariadb.service.
# systemctl start mariadb
Redirecting to /bin/systemctl start  mariadb.service
# systemctl status mariadb
* mariadb.service - MariaDB database server
   Loaded: loaded (/usr/lib/systemd/system/mariadb.service; enabled; vendor preset: disabled)
   Active: active (running) since Thursday 2018-09-06 21:46:13 JST; 4s ago
  Process: 36401 ExecStartPost=/usr/libexec/mariadb-wait-ready $MAINPID (code=exited, status=0/
SUCCESS)
  Process: 36319 ExecStartPre=/usr/libexec/mariadb-prepare-db-dir %n (code=exited, status=0/
SUCCESS)
```

```
 Main PID: 36400 (mysqld_safe)
   CGroup: /system.slice/mariadb.service
           tq36400 /bin/sh /usr/bin/mysqld_safe --basedir=/usr
           mq36573 /usr/libexec/mysqld --basedir=/usr --datadir=/var/opt/FJSVtcs/shared_disk/
mysql...

...
```

4. Setting the MariaDB administrator password
   Set the MariaDB administrator password.

```
# mysql_secure_installation

NOTE: RUNNING ALL PARTS OF THIS SCRIPT IS RECOMMENDED FOR ALL MariaDB
      SERVERS IN PRODUCTION USE!  PLEASE READ EACH STEP CAREFULLY!

In order to log into MariaDB to secure it, we'll need the current
password for the root user.  If you've just installed MariaDB, and
you haven't set the root password yet, the password will be blank,
so you should just press enter here.

Enter current password for root (enter for none): [LF]
OK, successfully used password, moving on...

Setting the root password ensures that nobody can log into the MariaDB
root user without the proper authorisation.

Set root password? [Y/n]  [LF]
New password: [root password for MariaDB]
Re-enter new password: [root password for MariaDB]
Password updated successfully!
Reloading privilege tables..
 ... Success!


By default, a MariaDB installation has an anonymous user, allowing anyone
to log into MariaDB without having to have a user account created for
them.  This is intended only for testing, and to make the installation
go a bit smoother.  You should remove them before moving into a
production environment.

Remove anonymous users? [Y/n] [LF]
 ... Success!

Normally, root should only be allowed to connect from 'localhost'.  This
ensures that someone cannot guess at the root password from the network.

Disallow root login remotely? [Y/n] [LF]
 ... Success!

By default, MariaDB comes with a database named 'test' that anyone can
access.  This is also intended only for testing, and should be removed
before moving into a production environment.

Remove test database and access to it? [Y/n] [LF]
 - Dropping test database...
 ... Success!
 - Removing privileges on test database...
 ... Success!

Reloading the privilege tables will ensure that all changes made so far
will take effect immediately.
```

```
Reload privilege tables now? [Y/n] [LF]
 ... Success!

Cleaning up...

All done!  If you've completed all of the above steps, your MariaDB
installation should now be secure.

Thanks for using MariaDB!
```

5. Configuring the standby compute cluster management node
   Perform this step only if the compute cluster management node is redundantly configured.
   Perform step 2 on the standby compute cluster management node.

6. Configuring the job operation management function and power management function

   Additionally configure the job operation management function and power management function to use the MariaDB functions.

   - Configuring the job operation management function

     See "MariaDB Settings" in " Job Operation Management Function Settings" in the "Job Operation Software Administrator's Guide for Job Management," and then configure the function.

   - Configuring the power management function

     If you use the System Power Database or the Job Power Estimate Database of the Power Management function, see the following respectively to perform the configuration.

     - Using the System Power Database: "Settings for the System Power Database" in "Power Management Function Settings" in "Job Operation Software Administrator's Guide for Power Management"
     - Using the Job Power Estimate Database: "Settings for the job power estimate database" in "Power Management Function Settings" in "Job Operation Software Administrator's Guide for Power Management"

7. Starting the job manager function
   Perform the following step to start the job manager function service on the active and standby compute cluster management nodes.
   Perform this work on the active system management node.

   - For the system management node does not have a configuration serving for multiple node types

     ```
     # paclstmgr -c clstname -n nodeid1,nodeid2 --service restart
     ```

     clstname: Cluster name
     nodeid1: Node ID of the active compute cluster management node
     nodeid2: Node ID of the standby compute cluster management node

   - For the node serving as all of system management node, compute cluster management node, and login node

     For the node serving as all of system management node, compute cluster management node, and login node, specify start instead of restart in the paclstmgr command --service option.

     ```
     # paclstmgr -c clstname --nodetype CCM -a --service start
     ```

     clstname: Cluster name

Information

In a system that uses FEFS, the job manager function waits for the FEFS service to start. Therefore, if you check the status of the active compute cluster management node with the pashowclst command at this point, the PJM service is shown as "(s)".

```
0xFFFF0001  CCM      Init      -           on            -
PJM(s),PLE(o),MRD(o),SRD(o),FEFS(b),PWRD(o)
```

2.4.4 Performing Job Operation-related Work

2.4.4.1 Configuring the Job Operation Management Function

Configure the job operation management function based on the assumed job operation model. For details, see "Job Operation Management Function Settings" in the "Job Operation Software Administrator's Guide for Job Management."

Information

A default resource unit is automatically created.
You can confirm resource units by specifying the --rscunit option to pashowclst on the system management node.

```
# pashowclst -c clstname --rscunit
[ CLST: clstname ]
RSCUNIT           RUNNING  STOPPED  ERROR    DISABLE
rscunit_ft01      48       0        0        0
```

clstname: Cluster name

2.4.4.2 Building FEFS

See the *FEFS User's Guide*, an FEFS manual, perform the work for building an FEFS file system.

2.4.4.3 Configuring the Power Management Function

To use the power management function, configure the function. For details on how to configure the power management function, see the "Job Operation Software Administrator's Guide for Power Management."

2.4.4.4 Setting NFS

If a node serving as all of system management node, compute cluster management node, and login node is used, set NFS as the shared file system used for job execution.

2.4.5 Setting a User Account

Set a user account for the OS as required. To link the system to work with an external server via LDAP, etc., set the account according to the operation model.

Note

This setting is required to use the parallel execution debugger.

See

For instructions on creating a user, see "Adding a User" in "Job Operation Software Administrator's Guide for Maintenance."

Information

If the Job Operation System you are building uses the KVM mode of the job execution environment, the home directory of the user executing the job must be created on a shared file system. This should be taken into account when designing shared file systems for use in job execution.
The KVM mode is not available for the system with a node serving as all of system management node, compute cluster management node, and login node.

2.5 Checking the Software Environment [SMM]

From the software perspective, check whether the installed nodes are ready to be incorporated into operation.

Execute the pachecker command, and confirm that settings are as expected.
For details on the pachecker command, see "Details of the Software Environment Check Function" in the "Job Operation Software Administrator's Guide for System Management."

2.6 Confirming Job Execution

2.6.1 Checking Resources [SMM]

Execute the pjshowrsc command, and confirm that the added compute nodes are available for job operations.
Perform this operation on all of the added compute clusters.

```
# pjshowrsc -c clstname
[ CLST: clstname ]
CLUSTER          NODE
                 TOTAL   FREE  ALLOC
clstname             7      7      0
```

clstname: Cluster name

2.6.2 Executing a Test Job [LN]

Run a test job, and confirm that the job can be executed.
The following example of a test job script (script name: job.sh) executes the hostname command on each node.

For details on how to write a job script and how to execute a job, see the "Job Operation Software End-user's Guide."

```
#!/bin/bash
#PJM -L node=nodenum
#PJM -L "rscunit=rscuname"
#PJM -L "rscgrp=rscgname"

/usr/bin/hostname
```

nodenum: Number of nodes
rscuname: Resource unit name
rscgname: Resource group name

Use the pjsub command to run the created job.

```
# pjsub job.sh
[INFO] PJM 0000 pjsub Job jobid submitted.
```

jobid: Job ID allocated to the job that was run

The standard output and standard error output are created as respective files in the current directory at job execution.

Standard output: *Job name.job ID.out*
Standard error output: *Job name.job ID.err*

Check the above files to confirm that the created job was executed as expected.

2.7 Backing Up a Disk Image for Maintenance

Execute the paclone command as required to back up a built disk image by using the backup/restore function. You can use the backed-up image for recovery in the event of a hardware failure (such as a disk failure).

For details on the paclone command, see "Backup of the Disk Image" in the "Job Operation Software Administrator's Guide for System Management."

2.8 Stopping TFTP and HTTP services for System Operations

Stop TFTP and HTTP services during system operation, except for performing maintenance tasks using the installation function (Apply or remove packages, back up or restore nodes, rebuild nodes).

```
# pmexe -c clstname --nodetype SMM,CCS "systemctl stop httpd"
# pmexe -c clstname --nodetype SMM,CCS "systemctl stop xinetd"
# pmexe -c clstname --nodetype SMM,CCS "systemctl disable httpd"
# pmexe -c clstname --nodetype SMM,CCS "systemctl disable xinetd"
```

clstname: Compute cluster name

Chapter 3 Adding a Cluster

This chapter describes procedures to add a cluster.

Since the workflow for adding a cluster is the same as in new system installation, follow the procedures in "2.3 Building a Cluster" and subsequent sections.

Information

You can add a cluster while job operations continue as is on the already installed clusters.

Chapter 4 Deleting a Cluster

This chapter describes procedures to delete a cluster.

Note

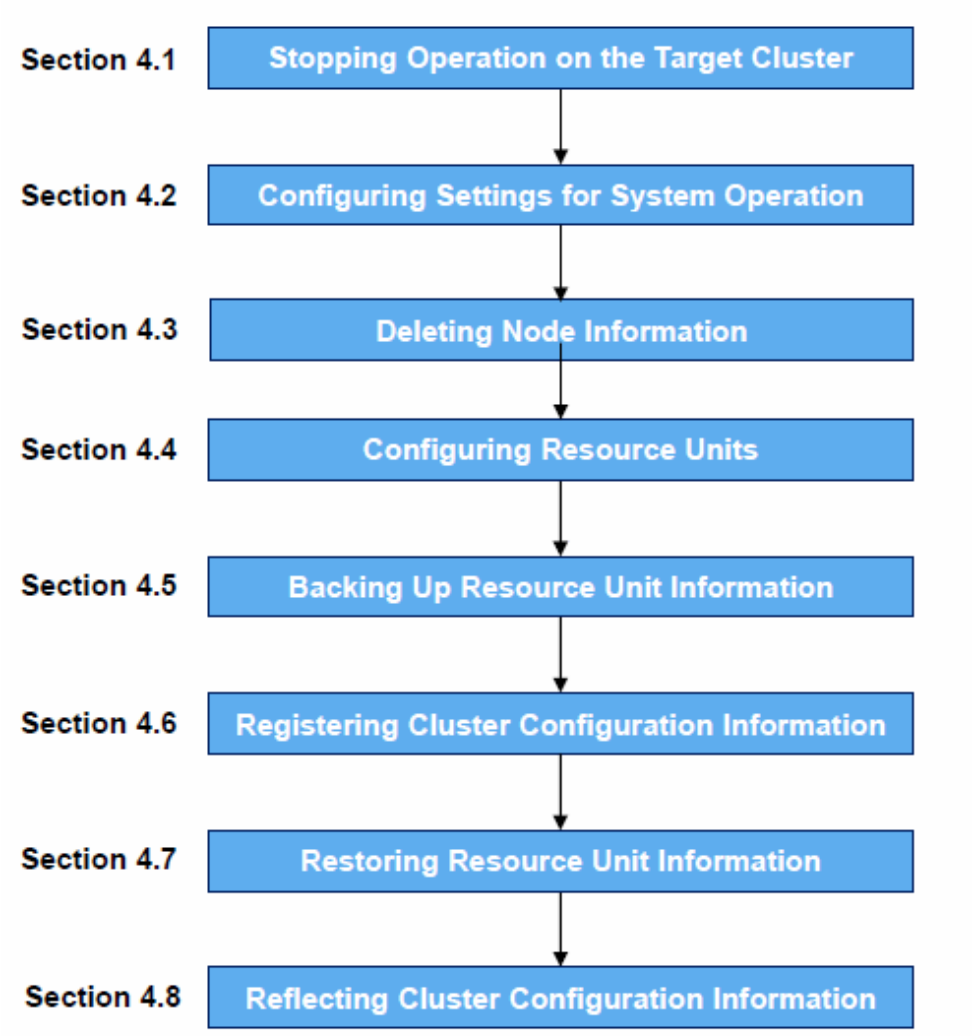
- To reincorporate a deleted node into the system, you need to reinstall the OS on this target node.
- A storage cluster used by a compute cluster or multiuse cluster cannot be deleted. To delete the storage cluster, first delete the compute cluster and multiuse cluster using it.

Information

- You can delete a cluster while job operations continue on the other clusters, which are not going to be deleted.
- If not even a single cluster remains in the system after you delete a cluster, the system management node is also deleted from configuration information. For this reason, to build a new cluster, perform the procedures in "Chapter 2 New System Installation."

To remove a cluster, use the following flow.

Figure 4.1 Workflow for deleting clusters



4.1 Stopping Operation on the Target Cluster [SMM]

To delete a compute cluster, stop job operations on the cluster, and isolate the compute nodes from operation.

```
# paclstmgr -c clstname --disable --ic -a
```

clstname: Cluster name

Information

When executed from the command line, the command shown above waits for jobs to complete and then isolates the nodes from operation. If you want the nodes isolated from operation without waiting for jobs to complete, specify the -k option.

Use the pashowclst command to confirm that the target nodes are isolated from operation.
Confirm that the STATUS column shows "Disable" and the REASON column shows "Manual" for of the compute nodes.

```
# pashowclst -c clstname -v --nodetype CN
[ CLST: clstname ]
[ NODETYPE: CN ]
NODE        NODETYPE  STATUS   REASON         PWR_STATUS    ARCH_STATUS    SRV_STATUS
nodeid1     CN        Disable  Manual        os-running    ICC_Disable    PLE(o),NRD(o),FEFS(o),PWRD(o)
nodeid2     CN        Disable  Manual        os-running    ICC_Disable    PLE(o),NRD(o),FEFS(o),PWRD(o)
...
```

clstname: Cluster name

4.2 Configuring Settings for System Operation

Remove information about the target cluster from the Job Operation Software settings that you have set for the system.

4.2.1 Configuring the System Monitoring Function [SMM]

Remove the cluster definition that you want to remove from the system monitoring function settings.
For details, see "Configuring the System Monitoring" in "Job Operation Software Administrator's Guide for System Management."

4.2.2 Configuring the Job Operation Management Function [SMM]

If you are removing a compute cluster, remove the job operation management function settings that are set for the target cluster.
For details, see "Job Operation Software Administrator's Guide for Job Management."

4.2.3 Configuring the Power Management Function [SMM]

Remove the cluster definition that you want to remove from the power management function settings.
For details, see "Job Operation Software Administrator's Guide for Power Management."

4.3 Deleting Node Information [SMM]

Before deleting nodes, delete their information from installation settings. This prevents installation processing from running unintentionally even if a deleted node is started.

1. Deleting node information

   ```
   # pxinst node del -f hostfile
   [INFO] INST 0801 pxinst Created a configuration backup.(/tmp/pxinst/node_date/)    (*)
   [INFO] INST 1804 pxinst Deleting nodes completed.(nodenum=xxx)
   ```

   hostfile: File listing the host names of the nodes to be deleted
   xxx: Number of nodes

(*) The /tmp/pxinst/node_*date* directory contains information saved from before pxinst command execution.

Information

Example of writing in *hostfile*

```
# cat hostfile
host001
host002
host003
host004
...
```

2. Synchronizing settings

   Synchronize settings with the standby installer node and relay node. Perform this work only in a system that has the standby installer node or relay node.

```
# pxinst sync all
[INFO] INST 1805 pxinst File replication is complete.
```

4.4 Configuring Resource Units [SMM]

If there are no more compute clusters in the system, the following steps 1-3 are not required.
If you have changed the resource unit, delete the resource units for the cluster that you are deleting in the following procedure.

1. Editing the resource unit configuration file

   Edit /etc/opt/FJSVtcs/parscunit.conf. For details on the setting items, see "parscunit.conf file" in the "Job Operation Software Command Reference".

2. Reflecting settings

   Execute the parscunitadm command with the options --set, --no-dist, and --force to reflect the settings.

```
# parscunitadm --set --no-dist --force
[INFO] PSM 1402 parscunitadm The setting of resource unit completed
```

3. Checking the result of reflecting settings

   Execute the pashowclst command with the --rscunit option to verify that the resource unit information is:

   - The resource unit name is correct.
   - The sum of RUNNING, STOPPED, ERROR, and DISABLE for each resource unit matches the number of compute nodes.

```
# pashowclst -c clstname --rscunit
[ CLST: clstname ]
RSCUNIT           RUNNING  STOPPED  ERROR    DISABLE
unit1             0        48       0        0
unit2             0        2        0        1
unit3             0        3        0        0
```

   clstname: Cluster name

4.5 Backing Up Resource Unit Information [SMM]

The "4.6 Registering Cluster Configuration Information [SMM]" operation automatically creates the resource unit and returns to the default configuration.

1. Back up the following directory:
   /etc/opt/FJSVtcs/psm/cluster_config/Rscunit.d

2. Back up the following directory:
   /etc/opt/FJSVtcs/Rscunit.d

3. Back up the following file if it exists.
   /etc/opt/FJSVtcs/parscunit.conf

4.6 Registering Cluster Configuration Information [SMM]

Delete the target cluster from registered cluster configuration information.

Note

If you are removing a storage cluster, or a compute cluster or a multipurpose cluster that utilizes a storage cluster, you must modify the FEFS configuration before performing this procedure. For more information, see "FEFS User's Guide."

1. Creating the cluster configuration definition file pxsystem.conf

   Create the cluster configuration definition file pxsystem.conf.
   The following example deletes the cluster cluster1.

```
Operation {
    Type = "del-clst"
}
Cluster {
    ClusterName = "cluster1"
}
```

   For details on the setting items in pxsystem.conf, see "pxsystem.conf file" in the "Job Operation Software Command Reference".

2. Registering the pxsystem.conf file contents

Information

Performing this procedure will stop the various services of the Job Operation Software on the target nodes, and none of the functions will work even when the nodes are restarted. However, the settings for deleting nodes cannot be applied to nodes already stopped at the time of the procedure. Consequently, the Job Operation Software services run when these particular nodes restart. For this reason, do not start a node while connected to the network in operation. Otherwise, the system in operation may be affected.

   Use the --create option in the pxsystemadm command to register the pxsystem.conf file contents with the system management node.

```
# pxsystemadm --create filename
Operation/Type is "del-clst".                                   (*1)
Do you really want to continue (y/n)? y                         (*2)
[INFO] PSM 1082 pxsystemadm The setting was completed.
```

   (*1) The displayed type of operation will be performed.
   (*2) This prompts for confirmation before registration.

   filename: Path to the created cluster configuration definition file

3. Confirming the registered information

   Specify the -a option in the pashowclst command to confirm that the clusters written in the pxsystem.conf file have been deleted.

```
# pashowclst -a
...
```

   Confirm that the clusters written in the pxsystem.conf file in step 1 have been deleted.

4.7 Restoring Resource Unit Information [SMM]

If there are no more compute clusters in the system, the following steps 1-2 are not required.
The "4.6 Registering Cluster Configuration Information [SMM]" operation automatically creates the resource unit and returns to the default configuration.
If you have changed the resource unit, check the parscunit.conf configuration file for the resource unit.
If the resource unit configuration after changing the resource unit differs from the settings in the parscunit.conf file, restore the parscunit.conf file from a backup of resource unit information in the following procedure.

1. Deleting directories and files

   Delete the following directories and files:

   - /etc/opt/FJSVtcs/psm/cluster_config/Rscunit.d
   - /etc/opt/FJSVtcs/Rscunit.d
   - /etc/opt/FJSVtcs/parscunit.conf

2. Placing backed up directories and files

   Copy the directories and files that you backed up in "4.5 Backing Up Resource Unit Information [SMM]" and locate them as follows:

   a. Directory in step 1 of "4.5 Backing Up Resource Unit Information [SMM]"

      Place it under the directory /etc/opt/FJSVtcs/psm/cluster_config/.

   b. Directory in step 2 of "4.5 Backing Up Resource Unit Information [SMM]"

      Place it under the directory /etc/opt/FJSVtcs/.

   c. The file in step 3 of "4.5 Backing Up Resource Unit Information [SMM]"

      Place it under the directory /etc/opt/FJSVtcs/.

4.8 Reflecting Cluster Configuration Information [SMM]

For systems that do not have standby system management node, no further action is required.

1. Verify the existing cluster name and node ID of the standby system management node.

```
# pashowclst -m
[ CLST: clstname ]
NODEGRP    NODETYPE      NODE        STATUS  NODE        STATUS
SMMGRP     SMM           0x00000001  (A)     0x00000002  (S)
MASTERGRP  CCM           0xFFFF0001  (A)     0xFFFF0002  (S)
```

2. Reflecting Cluster Configuration Information

```
# pxsystemadm --set -c clstname -n nodeid
```

   clstname: Cluster name

   nodeid: Node ID of the standby system management node

3. Service restart of standby system management node

```
# paclstmgr -c clstname -n nodeid --service restart
```

   clstname: Cluster name

   nodeid: Node ID of the standby system management node

4. Synchronizing the configuration file

   Execute the sync subcommand of the pxinst command to synchronize the configuration file created in "2. Reflecting Cluster Configuration Information" to the standby system management node.

```
# pxinst sync all
...
[INFO] INST 1805 pxinst File replication is complete.
```

Chapter 5 Adding a Node

This chapter describes procedures to add a node.
Perform these procedures when adding a shared file system (shared_disk) for a redundant configuration.

When adding a node to a cluster, you may need to stop cluster operation, depending on the type of node added.
In the following table, the patterns show which nodes can be added (node adding pattern), and the ranges show which clusters have to stop operating (operation stopping range).

Table 5.1 Node Adding Patterns and Operation Stopping Ranges

Node Adding Pattern	Possible to Add?	Operation Stopping Range
System management node System management node also serving as storage cluster management node	Yes (*1)	All clusters
Storage cluster management node	Yes (*1)	Target clusters
MGS node MGS node also serving as MDS node and OSS node MGS node also serving as MDS node	Yes	None
MDS node	Yes	None
OSS node	Yes	None
Compute cluster management node	Yes (*1)	Target clusters
Login node	Yes	None
Multiuse node	Yes	None
Compute cluster sub management node	Yes (*2)	Target clusters
Compute node also serving as boot I/O node (FX server)	Yes (*3)	Target clusters
Compute node also serving as storage I/O node (FX server)	Yes (*3)	Target clusters
Compute node also serving as global I/O node (FX server)	Yes (*3)	Target clusters
Compute node (FX server)	Yes (*3)	Target clusters
Compute node (PRIMERGY server)	Yes	Target clusters
Boot group (FX server)	Yes	Target clusters
Node group	Yes	Target clusters
Shared file system (shared_disk)	Yes	None

(*1) The node can be added only when expanding a standalone configuration into a redundant configuration.
(*2) The compute cluster sub management node cannot be added alone. Add it together with compute nodes under it as a node group at the same time.
(*3) A standalone FX server node cannot be added. Add it in the unit of a boot group.

According to "Operation Stopping Range" shown above, you may need to isolate a node from operation, and you may need to restart services after adding a node. This chapter describes actions for each pattern of an operation stopping range.

Operation stopping range: "All clusters"

To add a system management node, all clusters must be restarted in order for each function of the Job Operation Software to reflect information for the added node. To prevent unnecessary work like failover and automatic dump collection on the nodes in a cluster during the operation to add the node, stop operation on all clusters before the add operation. Also transition the nodes to software maintenance mode beforehand.

Operation stopping range: "Target clusters"

- Adding a storage cluster management node and a compute cluster management node

  To add a storage cluster management node and a compute cluster management node, the target clusters must be restarted in order for each function of the Job Operation Software to reflect information for the added nodes. To prevent unnecessary work like failover and automatic dump collection on the nodes in the clusters during the operation to add the nodes, stop operation on the target clusters before the add operation. Also transition the nodes to software maintenance mode beforehand.

- Adding a compute node, boot group, and node group

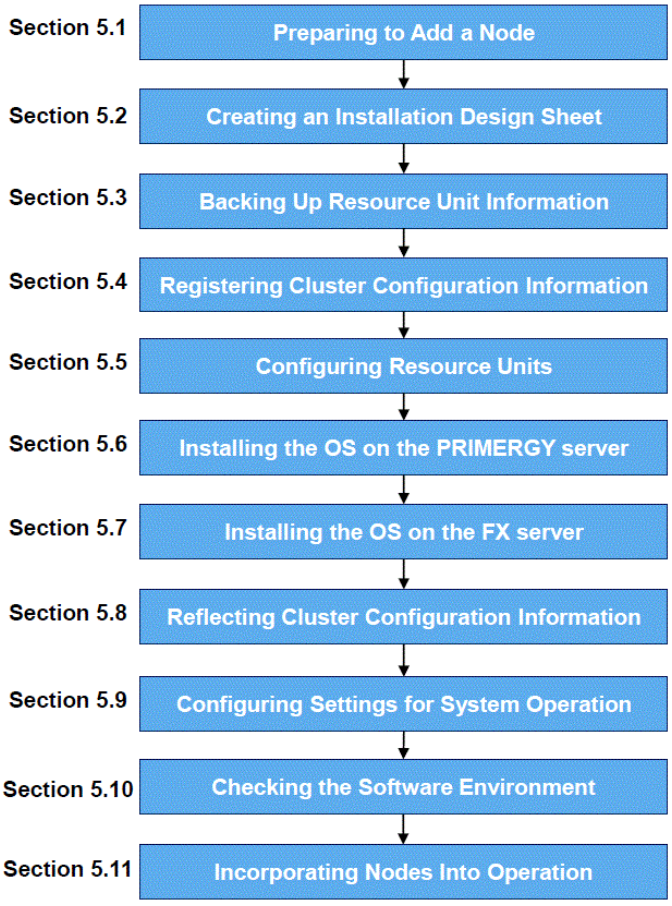
  When adding a compute node, a boot group, and a node group change the resource unit at the same time. After adding a node, you need to restart the compute cluster management node in order for the job operation management function to reflect information on the resource unit.
  Therefore, to add a compute node, a boot group, and a node group, stop operation on the target compute cluster, and transition its nodes to software maintenance mode.

Operation stopping range: "None"

You can add any node other than the above while system operation continues as is.

To adding a node, use the following flow.

Figure 5.1 Workflow for adding nodes



5.1 Preparing to Add a Node

Out of the patterns shown in "Table 5.1 Node Adding Patterns and Operation Stopping Ranges," some patterns require operation to stop. For these patterns, perform the subsequent procedures.
The procedures in this section are unnecessary for patterns where the operation stopping range is "None."

5.1.1 Stopping Cluster Operation [SMM]

To prevent unnecessary work like failover and automatic dump collection during the operation to add a node, transition the nodes in the target cluster to software maintenance mode.

When adding a system management node, perform the subsequent steps on all clusters because the node belongs to all of the clusters in the system.

1. Isolating nodes from operation

   Isolate the nodes in the target cluster from operation.

   ```
   # paclstmgr -c clstname -a --disable
   ```

   clstname: Cluster name

   The above command waits for jobs to complete. If you want the cluster isolated from operation without waiting for jobs to complete, specify the -k option.

2. Confirming that the nodes are isolated from operation

   Use the pashowclst command to confirm that compute nodes are isolated from operation.
   Confirm that the STATUS column shows "Disable" and the REASON column shows "Manual" for the nodes.

   ```
   # pashowclst -c clstname -v
   ...
   ```

   clstname: Cluster name

3. Transitioning to software maintenance mode

   Transition the nodes in the target clusters to software maintenance mode.

   ```
   # paclstmgr -c clstname -a --soft-mainte --no-service-stop
   ```

   clstname: Cluster name

4. Confirming the transition to software maintenance mode

   Use the --soft-mainte option in the pashowclst command to confirm the transition to software maintenance mode.

   ```
   # pashowclst -c clstname --soft-mainte
   ...
   ```

   clstname: Cluster name

5.1.2 Stopping Services on the Compute Cluster Management Node [SMM]

With the addition of nodes including a compute node, stop services on the compute cluster management node in order to change resource unit information. The addition of nodes including a compute node means the following addition of nodes:

- Adding a compute node (PRIMERGY server)
- Adding a boot group (FX server)
- Adding a node group

Perform this procedure when the work in "5.1.1 Stopping Cluster Operation [SMM]" is already done.

For the node serving as all of system management node, compute cluster management node, and login node, this step is unnecessary.

1. Stopping services on the compute cluster management node

   Stop the Job Operation Software services running on the compute cluster management node.
   Specify the --service option in the paclstmgr command, and execute the command.

```
# paclstmgr -c clstname --nodetype CCM -a --service stop
```

clstname: Cluster name

2. Confirming that the services have stopped on the compute cluster management node

   Use the pashowclst command to confirm that the services have stopped on the compute cluster management node.
   Confirm that the SRV_STATUS column shows a hyphen ("-") for compute cluster management nodes.

```
# pashowclst -c clstname -v --nodetype CCM
[ CLST: clstname ]
[ NODETYPE: CCM ]
NODE        NODETYPE  STATUS     REASON           PWR_STATUS    ARCH_STATUS    SRV_STATUS
nodeid1     CCM       Disable    SoftMaintenance  on            -              -
nodeid2     CCM       Disable    SoftMaintenance  on            -              -
```

clstname: Cluster name

5.2 Creating an Installation Design Sheet [PC]

Enter the required configuration information for adding a node into an installation design sheet in the same way as described in "2.3.2 Creating an Installation Design Sheet [PC]."

5.3 Backing Up Resource Unit Information [SMM]

The "5.4 Registering Cluster Configuration Information [SMM]" operation automatically creates the resource unit and returns you to the default configuration.

1. Back up the following directory:
   /etc/opt/FJSVtcs/psm/cluster_config/Rscunit.d
2. Back up the following directory:
   /etc/opt/FJSVtcs/Rscunit.d
3. Back up the following file if it exists.
   /etc/opt/FJSVtcs/parscunit.conf

5.4 Registering Cluster Configuration Information [SMM]

Create the cluster configuration definition file pxsystem.conf to define a configuration for the Job Operation Software. Use the --create option in the pxsystemadm command to set the file.

See

For details on cluster configuration definitions, see "Registering Cluster Configuration Information" in the "Job Operation Software Administrator's Guide for System Management."

1. Creating a template for the cluster configuration definition file pxsystem.conf

   You can create a template for the pxsystem.conf file used in this procedure together with the node information definition file that is output from an installation design sheet.

   To add only FX server or shared disk (shared_disk) settings, this step is unnecessary. Perform step 2 and subsequent steps.

   Perform the following steps to create a template for the pxsystem.conf file.

   a. Editing the node information definition file

      Each time that the command is executed for a template for pxsystem.conf, it creates information for a single cluster. As a result, information on nodes in multiple clusters may be written in the node information definition file. If so, split the file so that each file contains only the node information for one cluster.

b. Creating a template for the pxsystem.conf file

Execute the pxsystemadm command with the --convert option.

```
# pxsystemadm --convert filename
[INFO] PSM 1081 pxsystemadm The file was output.(outfile)
```

filename: Path to the node information definition file
outfile: File name of the output template

The template for the pxsystem.conf file is output under the current directory at the command execution time. The file name is pxsystem.conf_*command execution date and time*. The administrator enters items where not enough items exist.

2. Creating the cluster configuration definition file pxsystem.conf

Create the cluster configuration definition file pxsystem.conf.

```
Operation {
    Type = "add-node"
}
Cluster {
    ClusterName = "cluster1"
    MasterGroup {
...
}
```

For details on the setting items in pxsystem.conf, see "pxsystem.conf file" in the "Job Operation Software Command Reference."

Information

When adding the FX server only, write the number of added boot groups in the pxsystem.conf file.

The following example of the pxsystem.conf file adds 72 BoBs under the CCM.
Write the BootGroup section in the MasterGroup section.

```
Operation {
    Type = "add-node"
}
Cluster {
    ClusterName = "cluster1"
    MasterGroup {
        BootGroup {
            Model = "FT"
            NumberOfBootGroups = "72"
        }
    }
}
```

When adding only shared file system (shared_disk) settings, write the HA section in pxsystem.conf.

```
Operation {
    Type = "add-node"
}
Cluster {
    ClusterName = "cluster1"
    HA {
        Node {
            NodeType = "CCM"
            Mount {
                Device = "/dev/disk/by-id/wwn-xxxxxxxxxxxxxxxxxxxxxxxxxxxx"
                MountPoint = "/var/opt/FJSVtcs/shared_disk"
                FSType = "ext4"
                OptList = "rw"
            }
        }
```

```
    }
}
```

3. Registering the pxsystem.conf file contents

   Use the --create option in the pxsystemadm command to set the pxsystem.conf file contents on the system management node.

   - Adding nodes that include the FX server

     Specify the --ft option to specify the path to the storage directory of the FX server node information definition file for the installation function.

     ```
     # pxsystemadm --create filename --ft node_ft
     Operation/Type is "add-clst".                                  (*1)
     Do you really want to continue (y/n)? y                        (*2)
     [INFO] PSM 1082 pxsystemadm The setting was completed.
     ```

     (*1) The displayed type of operation will be performed.
     (*2) This prompts for confirmation before registration.

     filename: Path to the created cluster configuration definition file
     node_ft: Path to the storage directory of the FX server node information definition file

   - Adding nodes that do not include the FX server

     ```
     # pxsystemadm --create filename
     Operation/Type is "add-node".                                  (*1)
     Do you really want to continue (y/n)? y                        (*2)
     [INFO] PSM 1082 pxsystemadm The setting was completed.
     ```

     (*1) The displayed type of operation will be performed.
     (*2) This prompts for confirmation before registration.

     filename: Path to the created cluster configuration definition file

4. Confirming the registered information

   Specify the -a option in the pashowclst command to confirm that the set contents are registered.

   ```
   # pashowclst -a
   ...
   ```

   Confirm that the following information is correct.

   - The unique information for the added nodes does not have an error.
   - The mount information displayed when the shared_disk is set does not have an error.

   At this point, a hyphen is displayed for the service run status (SRV_STATUS column) because the functions of the Job Operation Software do not work on newly added nodes. Also, "Stopped" is displayed for the node run status (STATUS column) because the system monitoring function is not working.

   Furthermore, in a situation such as step-by-step system installation, if hardware is missing, "Unknown" is displayed for the power status (PWR_STATUS column) of the target node.

Note

If the node you want to add has a power state (PWR_STATUS column) of "on" or "os-running", you must power off the node.

If you use the papwrctl command, run "force-off" because the OS may not be installed.

See

For details on what is displayed by the pashowclst command, see "Displaying System Configuration Information" in the "Job Operation Software Administrator's Guide for System Management."

5.5 Configuring Resource Units [SMM]

If you are adding a compute node in the system for the first time, the following steps 1-5 are not required.
The "5.4 Registering Cluster Configuration Information [SMM]" operation automatically creates the resource unit and returns you to the default configuration.
If you have changed the resource unit, check the parscunit.conf configuration file for the resource unit.
If the resource unit configuration after changing the resource unit differs from the settings in the parscunit.conf file, restore and edit the parscunit.conf file from a backup of resource unit information in the following procedure.

1. Deleting directories and files
   Delete the following directories and files:

   - /etc/opt/FJSVtcs/psm/cluster_config/Rscunit.d
   - /etc/opt/FJSVtcs/Rscunit.d
   - /etc/opt/FJSVtcs/parscunit.conf

2. Placing backed up directories and files
   Copy the directories and files that you backed up in "5.3 Backing Up Resource Unit Information [SMM]" and locate them as follows:

   a. Directory in step 1 of "5.3 Backing Up Resource Unit Information [SMM]"
      Place it under the directory /etc/opt/FJSVtcs/psm/cluster_config/.

   b. Directory in step 2 of "5.3 Backing Up Resource Unit Information [SMM]"
      Place it under the directory /etc/opt/FJSVtcs/..

   c. The file in step 3 of "5.3 Backing Up Resource Unit Information [SMM]"
      Place it under the directory /etc/opt/FJSVtcs/.

3. Editing the resource unit configuration file
   Edit /etc/opt/FJSVtcs/parscunit.conf. For details on the setting items, see "parscunit.conf file" in the "Job Operation Software Command Reference".

4. Reflecting settings
   Execute the parscunitadm command with the options --set, --no-dist, and --force to reflect the settings.

```
# parscunitadm --set --no-dist --force
[INFO] PSM 1402 parscunitadm The setting of resource unit completed
```

5. Checking the result of reflecting settings
   Execute the pashowclst command with the --rscunit option to verify that the resource unit information is:

   - The resource unit name is correct.
   - The sum of RUNNING, STOPPED, ERROR, and DISABLE for each resource unit matches the number of compute nodes.

```
# pashowclst -c clstname --rscunit
[ CLST: clstname ]
RSCUNIT           RUNNING  STOPPED  ERROR    DISABLE
unit1             0        48       0        0
unit2             0         2       0        1
unit3             0         3       0        0
```

clstname: Cluster name

5.6 Installing the OS on the PRIMERGY server [SMM]

5.6.1 Preparing for Installation on the PRIMERGY server

Prepare to install the OS on an added PRIMERGY server in the same way as described in "2.3.9.1 Preparing for Installation on the PRIMERGY server."

5.6.2 Checking the States of Services Required for OS Installation

Prepare to install the OS on an added PRIMERGY server in the same way as described in "2.3.10.2 Checking the States of Services Required for OS Installation [SMM]."

5.6.3 Installing the OS on the PRIMERGY server

Install the OS on an added PRIMERGY server in the same way as described in "2.3.10.3 Installing the OS on the PRIMERGY server [SMM]."

5.7 Installing the OS on the FX server [SMM]

5.7.1 Preparing for Installation on the FX server

Prepare to install the OS on an added FX server in the same way as described in "2.3.9.2 Preparing for Installation on the FX server."

5.7.2 Checking the States of Services Required for OS Installation

Prepare to install the OS on an added FX server in the same way as described in "2.3.10.2 Checking the States of Services Required for OS Installation [SMM]."

5.7.3 Installing the OS on the FX server

Install the OS on an added FX server in the same way as described in "2.3.10.4 Installing the OS on the FX server [SMM]."

5.7.4 Starting Up a Diskless Node

When you reflect the cluster configuration information in the following steps, diskless nodes must be running as well as a boot I/O node.

Start up the diskless nodes on an added FX server in the same way as described in "2.3.10.6 Starting Up a Diskless Node [SMM]"

5.8 Reflecting Cluster Configuration Information [SMM]

Apply the Job Operation Software packages and set the initial values on a node where the OS is already installed. Also, reflect the set cluster configuration information on the active system management node to the nodes in the system.
For these settings, execute the pxsystemadm command with the --set option.

5.8.1 Applying the Cluster Configuration Definition

1. Backing up the area used as a shared file system

   When adding a shared_disk, back up the mounting directory as required. If no shared_disk is set, this step is unnecessary.
   The following example uses the pmexe command to back up the /var/opt/FJSVtcs/shared_disk directory on the target node to the /var/tmp directory.

   ```
   # pmexe -c clstname -n nodeid "cp -rp /var/opt/FJSVtcs/shared_disk /var/tmp/"
   ```

   clstname: Cluster name
   nodeid: Node ID (Specify the active node.)

2. Executing the pxsystemadm command

```
# pxsystemadm --set -c clstname
Do you really want to continue (y/n)? y
[INFO] PSM 1082 pxsystemadm The setting was completed.
```

clstname: Cluster name
Perform this step on all clusters.

Note

While the pxsystemadm command is running, yum is run remotely on the target node to apply the Job Operation Software packages. While the pxsystemadm command is running, do not run commands that manipulate rpmdb, such as the rpm or yum commands. Also, do not run the pachecker command because it internally references rpmdb.

See

Any node that is already powered off, for example, may fail to reflect the information. In such cases, an error list file is output to the current directory.

```
[ERR.] PSM 0020 pxsystemadm There are nodes that failed in the setting.(filename)
```

filename: Name of the error list file output to the current directory
If an error list file is output, check the file contents, and eliminate the cause of the error. Then, re-execute the command on the node that failed to reflect the information.
For details on troubleshooting based on an error message, see "Failed Reflection of Cluster Configuration Definitions" in the "Job Operation Software Troubleshooting."

3. Mounting a shared_disk

When adding a shared_disk, mount the added shared_disk. To not add a shared_disk, this step is unnecessary.
The following example uses the pmexe command to mount the shared disk in the /var/opt/FJSVtcs/shared_disk directory on the target node.

```
# pmexe -c clstname -n nodeid --active \
"mount -t ext4 /dev/disk/by-id/xxxxxxxxxxxxxxxxx /var/opt/FJSVtcs/shared_disk"
```

clstname: Cluster name
nodeid: Node ID (Specify the active node.)

4. Restoring the shared_disk

If you backed up the area in step 1 before mounting, restore the directory. If you did not perform step 1, this step is unnecessary.
The following example uses the pmexe command to restore the data backed up in step 1.

```
# pmexe -c clstname -n nodeid  "cp -rp /var/tmp/shared_disk/* /var/opt/FJSVtcs/shared_disk/"
```

clstname: Cluster name
nodeid: Node ID (Specify the active node.)

5.8.2 Restarting the service

5.8.2.1 Restarting Services on the Compute Cluster Management Node [SMM]

If you worked in "5.1.2 Stopping Services on the Compute Cluster Management Node [SMM]," or worked on the node serving as all of system management node, compute cluster management node, and login node, recover the compute cluster management node in the target cluster by restarting the services.

1. Recovering a node

   Recover the compute cluster management node in the target cluster by restarting the services.
   The options you specify depend on the configuration of the system managed node.

   - For the system management node does not have a configuration serving for multiple node types

     Specify the --recover and --service-restart options in the paclstmgr command, and execute the command.

     ```
     # paclstmgr -c clstname --nodetype CCM -a --recover --service-restart
     ```

     clstname: Cluster name

   - For the node serving as all of system management node, compute cluster management node, and login node

     Specify the --service option in the paclstmgr command, and execute the command.

     ```
     # paclstmgr -c clstname --nodetype CCM -a --service restart
     ```

     clstname: Cluster name

2. Confirming node recovery

   Use the pashowclst command to confirm the recovery of each compute cluster management node.
   Confirm that the STATUS column shows "Running" for compute cluster management nodes.

   ```
   # pashowclst -c clstname -v --nodetype CCM
   [ CLST: clstname ]
   [ NODETYPE: CCM ]
   NODE        NODETYPE  STATUS     REASON    PWR_STATUS  ARCH_STATUS  SRV_STATUS
   nodeid1     CCM       Running    -         on          -
   PJM(o),PLE(o),MRD(o),SRD(o),FEFS(o),PWRD(o)
   nodeid2     CCM       Running    -         on          -
   PJM(o),PLE(o),MRD(o),SRD(o),FEFS(o),PWRD(o)
   ```

   clstname: Cluster name

5.8.2.2 Restarting Services on the Compute Cluster Sub Management Node [SMM]

If you added a node that includes a compute node under the compute cluster sub management node, restart the services on the compute cluster sub management node. The addition of nodes including a compute node means the following addition of nodes:

- Adding a compute node (PRIMERGY server)
- Adding a boot group (FX server)
- Adding a node group

1. Recovering a node

   Recover the compute cluster sub management node in the target cluster by restarting the services.
   Specify the --recover and --service-restart options in the paclstmgr command, and execute the command.

   ```
   # paclstmgr -c clstname --nodetype CCS -a --recover --service-restart
   ```

   clstname: Cluster name

2. Confirming node recovery

   Use the pashowclst command to confirm the recovery of each compute cluster sub management node.
   Confirm that the STATUS column shows "Running" for compute cluster sub management nodes.

   ```
   # pashowclst -c clstname -v --nodetype CCS
   [ CLST: clstname ]
   [ NODETYPE: CCS ]
   NODE        NODETYPE  STATUS     REASON    PWR_STATUS  ARCH_STATUS  SRV_STATUS
   nodeid1     CCS       Running    -         on          -            PLE(o),SRD(o)
   nodeid2     CCS       Running    -         on          -            PLE(o),SRD(o)
   ```

clstname: Cluster name

5.8.2.3 Restarting the Power Management Function [SMM]

If you worked in "5.1.2 Stopping Services on the Compute Cluster Management Node [SMM]," restart the power management function service on the system management node. For the node serving as all of system management node, compute cluster management node, and login node, this step is unnecessary.

1. Disabling Automatic Failover

   For the system management node in redundant configurations, execute the paclstmgr command to disable automatic failover. Perform this procedure for all clusters.

```
# paclstmgr -c clstname  -n nodeid --set-failover disable
```

   clstname: Cluster name
   nodeid: Node ID of the system management node (Either of the active and standby nodes would be okay.)

   Use the paclstmgr command to check failover setting information.
   Confirm that the FAILOVER column shows "disable" for the nodes in the cluster.

```
# paclstmgr -c clstname -n nodeid --set-failover list
[ CLST: clstname ]
NODE        NODETYPE      FAILOVER
0x00000001  SMM           disable
0x00000002  SMM           disable
```

   clstname: Cluster name
   nodeid: Node ID of the system management node (Either of the active and standby nodes would be okay.)

2. Restarting the Power Management Function

```
# systemctl restart pxsyspwrd
```

   Use the pashowclst command to confirm that the SRV_STATUS column shows "PWRD(o)".

```
# pashowclst -c clstname -n nodeid
[ CLST: clstname ]
[ NODE: nodeid ]
NODE        NODETYPE    STATUS      REASON        PWR_STATUS     ARCH_STATUS     SRV_STATUS
nodeid      SMM,SCM     Running     -             on             -               IC(o),PWRD(o)
```

   clstname: Cluster name
   nodeid: Node ID of the active system management node

3. Enabling Disabled Automatic Failover

   To enable automatic failover on the nodes where automatic failover has been disabled in step 1, execute the paclstmgr command.

```
# paclstmgr -c clstname -n nodeid --set-failover enable
```

   clstname: Cluster name
   nodeid: Node ID of the system management node (Either of the active and standby nodes would be okay.)

   Use the paclstmgr command to check failover setting information.
   Confirm that the FAILOVER column shows "enable" for the nodes in the cluster.

```
# paclstmgr -c clstname -n nodeid --set-failover list
[ CLST: clstname ]
NODE        NODETYPE      FAILOVER
0x00000001  SMM           enable
0x00000002  SMM           enable
```

   clstname: Cluster name
   nodeid: Node ID of the system management node (Either of the active and standby nodes would be okay.)

5.8.3 Checking Cluster Configuration Information

1. Confirming the completion of settings

   Use the -v option in the pashowclst command to confirm the completion of settings for added nodes and the start of services. Confirm that the STATUS column shows "Running" for the nodes.

```
# pashowclst -v
...
```

Note

If you add a node that uses FEFS, the STATUS column is not "Running" until you configure FEFS for the node that uses FEFS. Therefore, when you check the system that uses FEFS, confirm that the operational statuses of non-FEFS services in the SRV_STATUS column of the node are "(o)" at this point.
The following example displays compute nodes. The FEFS service is shown as "(b)".

```
NODE        NODETYPE  STATUS   REASON  PWR_STATUS  ARCH_STATUS  SRV_STATUS
0x01010004  CN        Init     -       os-running  ICC_Running  PLE(o),NRD(o),FEFS(b),PWRD(o)
```

2. Executing the pxinst command

   Execute the sync subcommand of the pxinst command to synchronize the configuration files created in "5.8.1 Applying the Cluster Configuration Definition" to the standby system management node and the compute cluster sub management node.

```
# pxinst sync all
...
[INFO] INST 1805 pxinst File replication is complete.
```

5.9 Configuring Settings for System Operation

Configure the various settings required for system operation in the same way as described in "2.4 Configuring Settings for System Operation."

5.10 Checking the Software Environment

From the software perspective, check whether the installed nodes are ready to be incorporated into operation.

Execute the pachecker command, and confirm that settings are as expected.
For details on the pachecker command, see "Checking the Software Environment" in the "Job Operation Software Administrator's Guide for System Management."

5.11 Incorporating Nodes Into Operation

Incorporate nodes into operation from the range of operations stopped in "5.1 Preparing to Add a Node."

5.11.1 Incorporating a Cluster Into Operation [SMM]

If you worked in "5.1.1 Stopping Cluster Operation [SMM]," restart the Job Operation Software services on the target cluster.

Note

When adding a system management node, perform the subsequent steps on all clusters because the node belongs to all of the clusters in the system.

1. Incorporating nodes

   Use the --recover option in the paclstmgr command to incorporate nodes into operation.
   The option that is specified varies depending on the type of node added.

   - For the system management node, storage cluster management node, or compute cluster management node, or adding a compute node to the configuration of a node serving as all of system management node, compute cluster management node, and login node

     Restart the Job Operation Software services to reflect information for adding the node to the nodes in the cluster.

     ```
     # paclstmgr -c clstname -a --recover --service-restart
     ```

     clstname: Cluster name

   - For a node other than the above

     Specify the --no-restart option to recover the node.

     ```
     # paclstmgr -c clstname -a --recover --no-restart
     ```

     clstname: Cluster name

2. Confirming node incorporation

   Use the pashowclst command to confirm that the nodes are incorporated.
   Confirm that the STATUS column shows "Running" for each node.

   ```
   # pashowclst -v -c clstname
   ...
   ```

   clstname: Cluster name

Chapter 6 Deleting a Node

This chapter describes procedures to delete a node.
Use this procedure even if you are removing a shared file system for redundancy (shared_disk).

When deleting a node from a cluster, you may need to restart the node or cluster, depending on the type of node deleted.
In the following table, the patterns show which nodes can be deleted (node deleting pattern), and the ranges show which clusters have to stop operating (operation stopping range).

Table 6.1 Node Deleting Patterns and Operation Stopping Ranges

Node Deleting Pattern	Possible to Delete?	Operation Stopping Range
System management node System management node also serving as storage cluster management node	No (*1)	-
Storage cluster management node	No (*1)	-
MGS node MGS node also serving as MDS node and OSS node MGS node also serving as MDS node	Yes	None
MDS node	Yes	None
OSS node	Yes	None
Compute cluster management node	No (*1)	-
Login node	Yes	None
Multiuse node	Yes	None
Compute cluster sub management node	No (*1)(*2)	-
Compute node also serving as boot I/O node (FX server)	No (*3)	-
Compute node also serving as storage I/O node (FX server)	No (*3)	-
Compute node also serving as global I/O node (FX server)	No (*3)	-
Compute node (FX server)	No (*3)	-
Compute node (PRIMERGY server)	Yes	Target clusters
Boot group (FX server)	Yes	Target clusters
Node group	Yes	Target clusters
Shared file system (shared_disk)	Yes	None

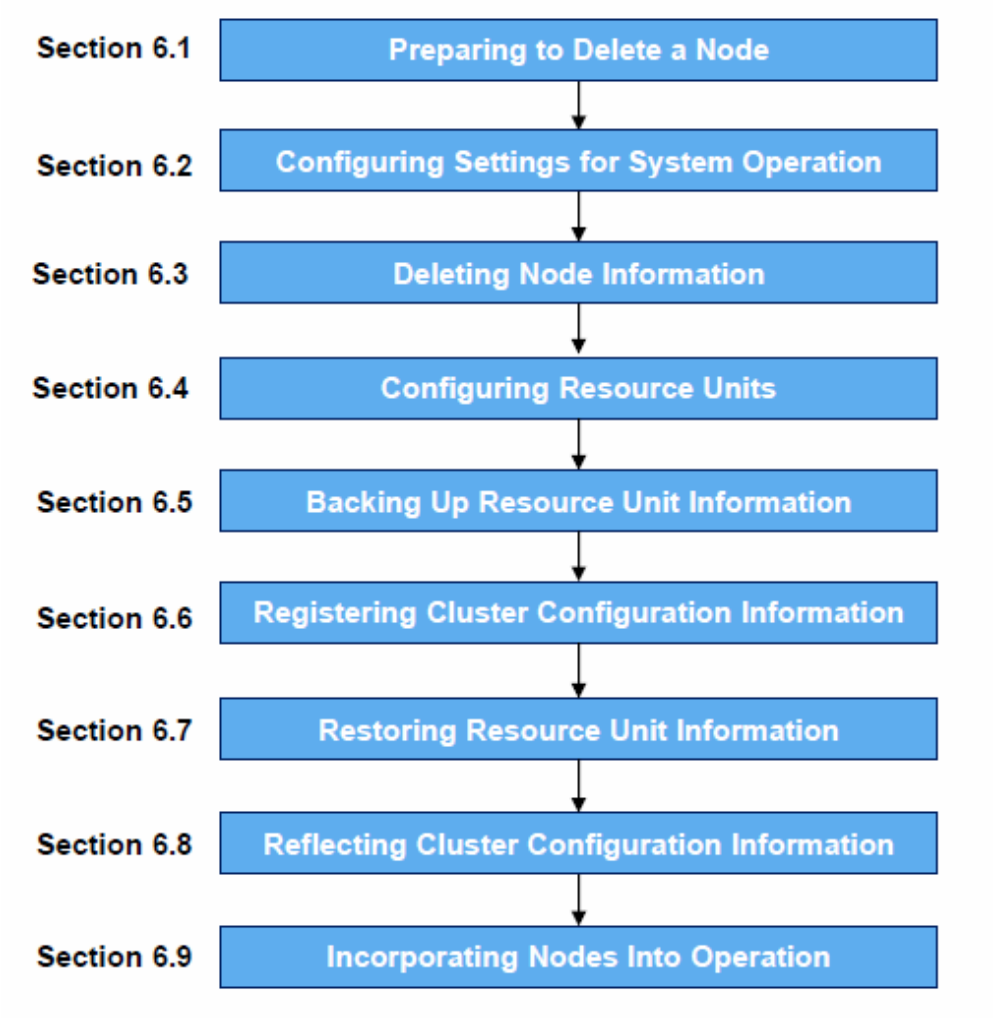
(*1) Neither of the nodes in a redundant configuration can be deleted alone to leave a standalone configuration.
(*2) The compute cluster sub management node cannot be deleted alone. Delete it together with the compute nodes under it as a node group at the same time.
(*3) A standalone FX server node cannot be deleted. Delete it in the unit of a boot group.

Note

- To reincorporate a deleted node into the system, you need to reinstall the OS on this target node.
- If deleting a node leaves only "System management node" or "System management node also serving as storage cluster management node" in the cluster, delete the cluster, not the node.
  For details on deleting clusters, see "Chapter 4 Deleting a Cluster".

To delete a node, use the following flow.

Figure 6.1 Workflow for deleting nodes



6.1 Preparing to Delete a Node [SMM]

Perform these procedures when deleting the following nodes:

- Compute node (PRIMERGY server)
- Boot group (FX server)
- Node group

The procedures in this section are unnecessary for patterns where the operation stopping range is "None."

6.1.1 Stopping Cluster Operation [SMM]

To prevent unnecessary work like failover and automatic dump collection during the operation to delete a node, transition the nodes in the target cluster to software maintenance mode.

Note

When deleting the system management node, perform the subsequent steps on all clusters because the node belongs to all of the clusters in the system.

1. Isolating nodes from operation

   Isolate the nodes in the target cluster from operation.

```
# paclstmgr -c clstname -a --soft-mainte --no-service-stop
```

   clstname: Cluster name

   The above command waits for jobs to complete. If you want the cluster isolated from operation without waiting for jobs to complete, specify the -k option.

2. Confirming that the nodes are isolated from operation

   Use the pashowclst command to confirm that compute nodes are isolated from operation.
   Confirm that the STATUS column shows "Disable" and the REASON column shows "SoftMaintenance" for the compute nodes.

```
# pashowclst -c clstname -v --nodetype CN
...
```

   clstname: Cluster name

6.1.2 Stopping Services on the Compute Cluster Management Node [SMM]

For the node serving as all of system management node, compute cluster management node, and login node, this step is unnecessary.

1. Transitioning to software maintenance mode

   Transition the compute cluster management node to software maintenance mode in order to stop the Job Operation Software services on the node.
   Specify the --soft-mainte and --service-stop options in the paclstmgr command, and execute the command.

```
# paclstmgr -c clstname --nodetype CCM -a --soft-mainte --service-stop
```

   clstname: Cluster name

2. Confirming the transition to software maintenance mode

   Use the pashowclst command to confirm that each compute cluster management node is isolated from operation.
   Confirm that the STATUS column shows "Disable" and the REASON column shows "SoftMaintenance" for the compute cluster management nodes.

```
# pashowclst -c clstname -v --nodetype CCM
[ CLST: clstname ]
[ NODETYPE: CCM ]
NODE        NODETYPE    STATUS     REASON             PWR_STATUS    ARCH_STATUS    SRV_STATUS
nodeid1     CCM         Disable    SoftMaintenance    on            -              -
nodeid2     CCM         Disable    SoftMaintenance    on            -              -
```

   clstname: Cluster name

6.2 Configuring Settings for System Operation

Delete the configuration information of the target node from the job operation software configuration set on the system.
Configure the various settings required for system operation in the same way as described in "2.4 Configuring Settings for System Operation."

6.3 Deleting Node Information [SMM]

Before deleting nodes, delete their information from installation settings. This prevents installation processing from running unintentionally even if a deleted node is started.

1. Deleting node information

```
# pxinst node del -f hostfile
[INFO] INST 0801 pxinst Created a configuration backup.(/tmp/pxinst/node_date/)    (*)
[INFO] INST 1804 pxinst Deleting nodes completed.(nodenum=xxx)
```

hostfile: File listing the host names of the nodes to be deleted
xxx: Number of nodes

(*) The /tmp/pxinst/node_*date* directory contains information saved from before pxinst command execution.

Example of writing in hostfile

```
# cat hostfile
host001
host002
host003
host004
...
```

2. Synchronizing settings

   Synchronize settings with the standby installer node and relay node. Perform this work if the system has the standby installer node or relay node.

```
# pxinst sync all
[INFO] INST 1805 pxinst File replication is complete.
```

6.4 Configuring Resource Units [SMM]

If there are no more compute nodes in the system, the following steps 1-3 are not required.
If you have changed the resource unit, delete the node you are deleting from the resource unit range in the following procedure.

1. Editing the resource unit configuration file

   Edit /etc/opt/FJSVtcs/parscunit.conf. For details on the setting items, see "parscunit.conf file" in the "Job Operation Software Command Reference".

2. Reflecting settings

   Execute the parscunitadm command with the options --set, --no-dist, and --force to reflect the settings.

```
# parscunitadm --set --no-dist --force
[INFO] PSM 1402 parscunitadm The setting of resource unit completed
```

3. Checking the result of reflecting settings

   Execute the pashowclst command with the --rscunit option to verify that the resource unit information is:

   - The resource unit name is correct.
   - The sum of RUNNING, STOPPED, ERROR, and DISABLE for each resource unit matches the number of compute nodes.

```
# pashowclst -c clstname --rscunit
[ CLST: clstname ]
RSCUNIT           RUNNING  STOPPED  ERROR    DISABLE
unit1             0        48       0        0
unit2             0        2        0        1
unit3             0        3        0        0
```

clstname: Cluster name

6.5 Backing Up Resource Unit Information [SMM]

The "6.6 Registering Cluster Configuration Information [SMM]" operation automatically creates the resource unit and returns you to the default configuration.

1. Back up the following directory:
   /etc/opt/FJSVtcs/psm/cluster_config/Rscunit.d
2. Back up the following directory:
   /etc/opt/FJSVtcs/Rscunit.d
3. Back up the following file if it exists.
   /etc/opt/FJSVtcs/parscunit.conf

6.6 Registering Cluster Configuration Information [SMM]

Delete the target nodes from cluster configuration information so that none of the functions of the Job Operation Software work on the nodes to be deleted.

Note

If you are removing a node of a storage cluster, or a compute or multipurpose node that utilize the storage cluster, you must modify the FEFS configuration before performing this procedure. For details, see the "FEFS User's Guide."

1. Creating the cluster configuration definition file pxsystem.conf

   Create the cluster configuration definition file pxsystem.conf.
   The following example deletes the node group whose node group ID is 0x01 in the cluster cluster1, and then deletes the set shared file system (shared_disk) of the compute cluster management node.

```
Operation {
    Type = "del-node"
}
Cluster {
    ClusterName = "cluster1"
    NodeGroup {
        NodeGroupID = "0x01"
    }
    HA {
        Node {
            NodeType = "CCM"
            Mount {
                Device = "/dev/disk/by-id/wwn-xxxxxxxxxxxxxxxxxxxxxxxxxxxx"
            }
        }
    }
}
```

   The following example deletes the node with node ID 0xFFFF0004 and the node with node ID 0xFFFF0005 in the cluster1.

   If you want to delete multiple nodes, multiple node groups, or multiple boot groups, write the Node section, NodeGroup section, and BootGroup section for each node or group you want to delete.

```
Operation {
    Type = "del-node"
}
Cluster {
    ClusterName = "cluster1"
    Node {
        NodeID = "0xFFFF0004"
    }
    Node {
```

```
        NodeID = "0xFFFF0005"
    }
}
```

For details on the setting items in pxsystem.conf, see "pxsystem.conf file" in the "Job Operation Software Command Reference."

2. Registering the pxsystem.conf file contents

Information

Performing this procedure will stop the various services of the Job Operation Software on the target nodes, and none of the functions will work if the nodes are restarted. However, the settings for deleting nodes cannot be applied to nodes already stopped at the time of the procedure. Consequently, the Job Operation Software services run when these particular nodes restart. For this reason, do not start a node while connected to the network in operation. Otherwise, the system in operation may be affected.

Use the --create option in the pxsystemadm command to set the pxsystem.conf file contents on the system management node.

```
# pxsystemadm --create filename
Operation/Type is "del-node".                                   (*1)
Do you really want to continue (y/n)? y                         (*2)
[INFO] PSM 1082 pxsystemadm The setting was completed.
```

(*1) The displayed type of operation will be performed.
(*2) This prompts for confirmation before registration.

filename: Path to the created cluster configuration definition file

3. Confirming the registered information

Specify the -a option in the pashowclst command to confirm that the set contents are registered.

```
# pashowclst -a
...
```

Confirm that the clusters written in the pxsystem.conf file in step 1 have been deleted.

6.7 Restoring Resource Unit Information [SMM]

If there are no more compute nodes in the system, the following steps 1-2 are not required.
The "6.6 Registering Cluster Configuration Information [SMM]" operation automatically creates the resource unit and returns you to the default configuration.
If you have changed the resource unit, check the parscunit.conf configuration file for the resource unit.
If the resource unit configuration after changing the resource unit differs from the settings in the parscunit.conf file, restore the parscunit.conf file from a backup of resource unit information in the following procedure.

1. Deleting directories and files

   Delete the following directories and files:

   - /etc/opt/FJSVtcs/psm/cluster_config/Rscunit.d
   - /etc/opt/FJSVtcs/Rscunit.d
   - /etc/opt/FJSVtcs/parscunit.conf

2. Placing backed up directories and files

   Copy the directories and files that you backed up in "6.5 Backing Up Resource Unit Information [SMM]" and locate them as follows:

   a. Directory in step 1 of "6.5 Backing Up Resource Unit Information [SMM]"

      Place it under the directory /etc/opt/FJSVtcs/psm/cluster_config/.

   b. Directory in step 2 of "6.5 Backing Up Resource Unit Information [SMM]"

      Place it under the directory /etc/opt/FJSVtcs/.

c. The file in step 3 of "6.5 Backing Up Resource Unit Information [SMM]"

   Place it under the directory /etc/opt/FJSVtcs/.

6.8 Reflecting Cluster Configuration Information [SMM]

Use the --set option in the pxsystemadm command to get the system to reflect the registered information.

```
# pxsystemadm --set -c clstname
Do you really want to continue (y/n)? y
[INFO] PSM 1082 pxsystemadm The setting was completed.
```

clstname: Cluster name

Perform this step on all clusters.

Then execute the sync subcommand of the pxinst command to synchronize the configuration files created to the standby system management node and the compute cluster sub management node.

```
# pxinst sync all
...
[INFO] INST 1805 pxinst File replication is complete.
```

See

Any node that is already powered off, for example, may fail to reflect the information. In such cases, an error list file is output in the current directory.

```
[ERR.] PSM 0020 pxsystemadm There are nodes that failed in the setting.(filename)
```

filename: Name of the error list file output to the current directory

If an error list file is output, check the file contents, and eliminate the cause of the error. Then, re-execute the command on the node that failed to reflect the information.
For details on troubleshooting based on an error message, see "Failed Reflection of Cluster Configuration Definitions" in the "Job Operation Software Troubleshooting."

Note

To reincorporate a deleted node into the system, you need to reinstall the OS on this target node.

6.9 Incorporating Nodes Into Operation

Incorporate nodes into operation from the range of operations stopped in "Preparing to Delete a Node."

6.9.1 Restarting Services on the Compute Cluster Management Node [SMM]

If you worked in "6.1.2 Stopping Services on the Compute Cluster Management Node [SMM]," or worked on the node serving as all of system management node, compute cluster management node, and login node, recover the compute cluster management node in the target cluster by restarting the services.

1. Recovering a node

   Recover the compute cluster management node in the target cluster by restarting the services.

   The options you specify depend on the configuration of the system managed node.

- For the system management node does not have a configuration serving for multiple node types

  Specify the --recover and --service-restart options in the paclstmgr command, and execute the command.

```
# paclstmgr -c clstname -n nodeid --recover --service-restart
```

  clstname: Cluster name
  nodeid: Node ID

- For the node serving as all of system management node, compute cluster management node, and login node

  Specify the --service option in the paclstmgr command, and execute the command.

```
# paclstmgr -c clstname --nodetype CCM -a --service restart
```

  clstname: Cluster name

2. Confirming node recovery

   Use the pashowclst command to confirm the recovery of each compute cluster management node.
   Confirm that the STATUS column shows "Running" for compute cluster management nodes.

```
# pashowclst -c clstname -v --nodetype CCM
[ CLST: clstname ]
[ NODETYPE: CCM ]
NODE        NODETYPE  STATUS     REASON    PWR_STATUS  ARCH_STATUS  SRV_STATUS
nodeid1     CCM       Running    -         on          -
PJM(o),PLE(o),MRD(o),SRD(o),FEFS(o),PWRD(o)
nodeid2     CCM       Running    -         on          -
PJM(o),PLE(o),MRD(o),SRD(o),FEFS(o),PWRD(o)
```

   clstname: Cluster name

6.9.2 Restarting Services on the Compute Cluster Sub Management Node [SMM]

If you delete a node that contains a compute node under a compute cluster sub management node, restart the services on the compute cluster sub management node. Perform these procedures when deleting the following nodes:

- Compute node (PRIMERGY server)
- Boot group (FX server)
- Node group

1. Recovering a node

   Recover the compute cluster sub management node in the target cluster by restarting the services.

   Specify the --recover and --service-restart options in the paclstmgr command, and execute the command.

```
# paclstmgr -c clstname --nodetype CCS -a --recover --service-restart
```

   clstname: Cluster name

2. Confirming node recovery

   Use the pashowclst command to confirm the recovery of each compute cluster sub management node.

   Confirm that the STATUS column shows "Running" for compute cluster sub management nodes.

```
# pashowclst -c clstname -v --nodetype CCS
[ CLST: clstname ]
[ NODETYPE: CCS ]
NODE        NODETYPE  STATUS     REASON    PWR_STATUS  ARCH_STATUS  SRV_STATUS
```

```
nodeid1     CCS       Running    -         on          -           PLE(o),SRD(o)
nodeid2     CCS       Running    -         on          -           PLE(o),SRD(o)
```

clstname: Cluster name

6.9.3 Restarting Services on the Target Cluster [SMM]

If you worked in "6.1.1 Stopping Cluster Operation [SMM]" for the target cluster, restart the Job Operation Software services on the cluster.

1. Restarting the Job Operation Software services

   Use the --recover option in the paclstmgr command to incorporate nodes into operation.
   The option that is specified varies depending on the type of node deleted.

   - For the system management node, storage cluster management node, or compute cluster management node, or removing a compute node from the configuration of a node serving as all of system management node, compute cluster management node, and login node

     Restart the Job Operation Software services to reflect information for deleting the node to the nodes in the cluster.

     ```
     # paclstmgr -c clstname -a --recover --service-restart
     ```

     clstname: Cluster name

   - For a node other than the above

     Specify the --no-restart option to recover the node.

     ```
     # paclstmgr -c clstname -a --recover --no-restart
     ```

     clstname: Cluster name

2. Confirming the start of the Job Operation Software services

   Use the pashowclst command to confirm that the Job Operation Software services have started.
   Confirm that the SRV_STATUS column shows "o" for each node.

   ```
   # pashowclst -v -c clstname
   ...
   ```

   clstname: Cluster name

Chapter 7 Changing Configuration Information

This chapter describes procedures to rewrite configuration information, such as a node ID address.

For small, node serving as all of system management node, compute cluster management node, and login node, do not change configuration information and perform a node rebuild.

The information that can be changed varies by node type.

The procedures also vary depending on the information changed. See "Change Procedure" in the following table, and follow the appropriate procedure.

Table 7.1 Configuration Information That Can be Changed

Node Type	Information That Can be Changed	Change Procedure
System management node System management node also serving as storage cluster management node	IP address of control network (for communication)	7.1 Changing Cluster Configuration Information
	IP address of control network	
	IP address of management network (Redundant configuration only)	
	IP address other than those above	7.2 Changing an IP Address
	Host name	7.3 Changing a Host Name
	MAC address	7.4 Changing a MAC Address
	Disk partition configuration	7.5 Changing a Disk Partition
Storage cluster management node Compute cluster management node Compute cluster sub management node	IP address of control network	7.1 Changing Cluster Configuration Information
	IP address of management network (Redundant configuration only)	
	IP address other than those above	7.2 Changing an IP Address
	Host name	7.3 Changing a Host Name
	MAC address	7.4 Changing a MAC Address
	Disk partition configuration	7.5 Changing a Disk Partition
MGS node MGS node also serving as MDS node and OSS node MGS node also serving as MDS node MDS node OSS node Login node Compute node (PRIMERGY server)	IP address of control network	7.1 Changing Cluster Configuration Information
	IP address of management network	
	IP address other than those above	7.2 Changing an IP Address
	Host name	7.3 Changing a Host Name
	MAC address	7.4 Changing a MAC Address
	Disk partition configuration	7.5 Changing a Disk Partition
Compute node also serving as boot I/O node (FX server)	IP address of management network	7.1 Changing Cluster Configuration Information
	Host name	7.3 Changing a Host Name
	MAC address	7.4 Changing a MAC Address
Compute node also serving as storage I/O node (FX server) Compute node also serving as global I/O node (FX server) Compute node (FX server)	Host name	7.3 Changing a Host Name
	MAC address	7.4 Changing a MAC Address
Multi-use node	IP address of control network	7.1 Changing Cluster Configuration Information
	IP address of management network	

Node Type	Information That Can be Changed	Change Procedure
	Enable setting except for control functions	
	IP address other than those above	7.2 Changing an IP Address
	Host name	7.3 Changing a Host Name
	MAC address	7.4 Changing a MAC Address
	Disk partition configuration	7.5 Changing a Disk Partition

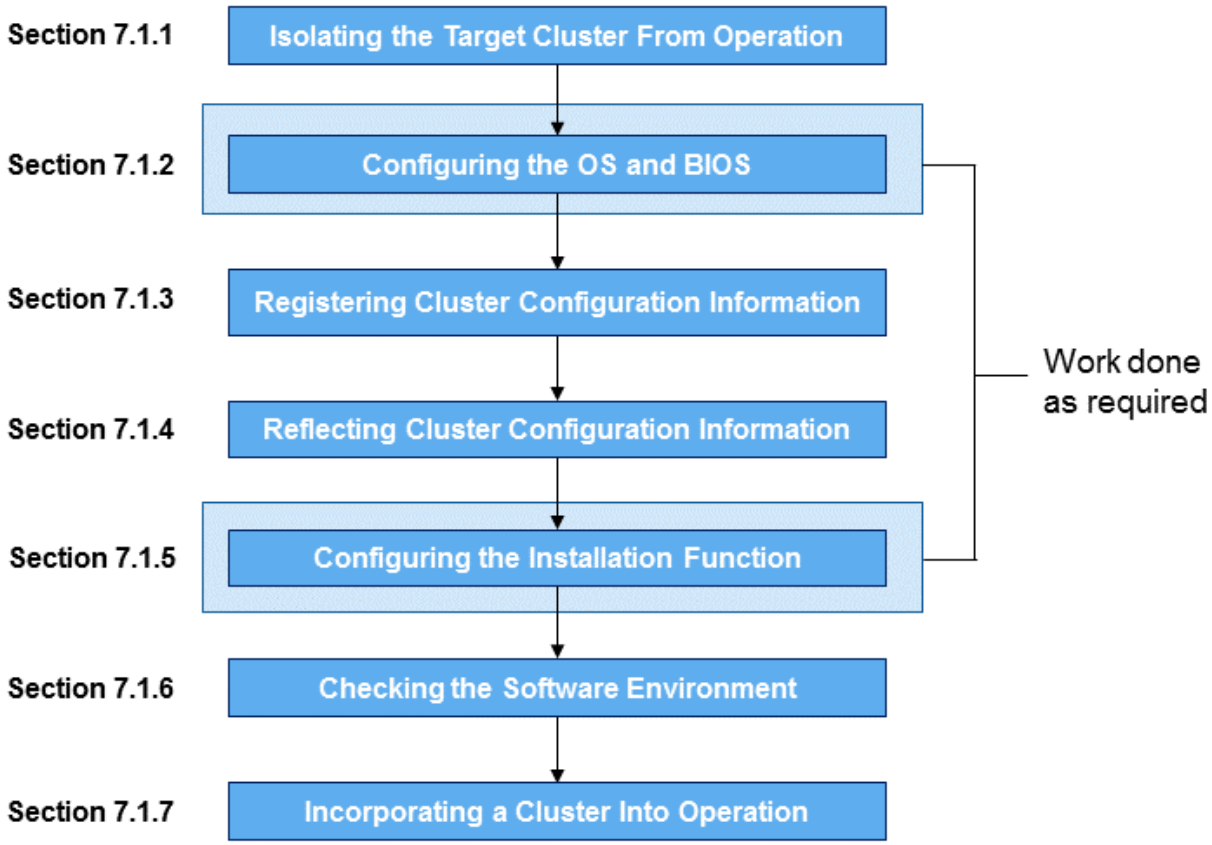
7.1 Changing Cluster Configuration Information

Changes to cluster configuration information require a restart of the target cluster in order to get each function of the Job Operation Software to reflect the changed information.

So that the configuration information can be changed safely, stop the Job Operation Software services running on the nodes in the target cluster before changing the information, and restart it after changing the information.

Follow the workflow below to change cluster configuration information.

Figure 7.1 Workflow for Changing Cluster Configuration Information



7.1.1 Detaching the Target Cluster From Operation [SMM]

Stop job operations on the target cluster to isolate it from operation.
To update node information in the compute cluster, to update node information in the storage cluster, or to update information of the system management node, perform isolating in the following order.

[To update node information in the compute cluster]
Isolate the target compute cluster and the standby system management node from operation.

1. Compute cluster
2. Standby system management node

[To update node information in the storage cluster or to update information of the system management node]
Isolate all clusters from operation.

1. Compute cluster
2. Storage cluster
3. Standby system management node

7.1.1.1 Detaching a Calculation Cluster

Detaches the compute cluster from the job operation.

```
# paclstmgr -c clstname -a --disable
```

clstname: Cluster name

Perform the -v option of the pashowclst command to verify that the STATUS column of all compute nodes in the specified cluster has transitioned to Disable.

```
# pashowclst -c clstname --nodetype CN -v
[ CLST: clstname ]
[ NODETYPE: CN ]
NODE        NODETYPE  STATUS   REASON       PWR_STATUS    ARCH_STATUS    SRV_STATUS
0x01010001  BIO,CN    Disable  Manual       os-running    ICC_Running
PLE(o),SRD(o),NRD(o),FEFS(o),PWRD(o)
0x01010002  CN        Disable  Manual       os-running    ICC_Running    PLE(o),NRD(o),FEFS(o),PWRD(o)
0x01010003  GIO,CN    Disable  Manual       os-running    ICC_Running
PLE(o),NRD(o),FEFSSR(o),FEFS(o),PWRD(o)
0x01010004  CN        Disable  Manual       os-running    ICC_Running    PLE(o),NRD(o),FEFS(o),PWRD(o)
```

clstname: Cluster name

Information

If you want to change the IP address of management network of a compute cluster sub management node, you must switch the target compute cluster sub management node to a standby state. If the target node is active, switch with the --failover command of the paclstmgr command.

```
# paclstmgr -c clstname --failover nodeid1
Do you really want to continue (y/n) ? y
[INFO] PSM 4084 paclstmgr Failover process starts now. nodeid1 is going to failover to nodeid2.
...
[INFO] PSM 4085 paclstmgr Failover process finished normally. nodeid1 was failovered to nodeid2.
```

nodeid1: The node ID of the active node (Node ID of the node to standby by performing the paclstmgr command)

nodeid2: The node ID of the standby node (Node ID of the node to activate by performing the paclstmgr command)

Transition to software maintenance mode to stop servicing the Job Operation Software on all nodes of the specified cluster except system management node.

```
# paclstmgr -c clstname -a --soft-mainte --service-stop
```

clstname: Cluster name

Perform the -v option of the pashowclst command to verify that the STATUS column is Disable, the REASON column is SoftMaintenance, and the SRV_STATUS column is - for all nodes in the specified cluster except system management node.

```
# pashowclst -c clstname -v
[ CLST: clstname ]
```

```
NODE       NODETYPE     STATUS       REASON             PWR_STATUS       ARCH_STATUS      SRV_STATUS
0x00000001  SMM          Running      -                 on               -                IC(o),PWRD(o)
0x00000002  SMM          Running      -                 on               -                IC(o),PWRD(o)
0xFFFF0001  CCM           Disable       SoftMaintenance     on               -                -
0xFFFF0002  CCM           Disable       SoftMaintenance     on               -                -
...
0x01010001  BIO,CN        Disable       SoftMaintenance     os-running       ICC_Running      -
0x01010002  CN           Disable       SoftMaintenance     os-running       ICC_Running      -
0x01010003  GIO,CN        Disable       SoftMaintenance     os-running       ICC_Running      -
0x01010004  CN           Disable       SoftMaintenance     os-running       ICC_Running      -
```

clstname: Cluster name

7.1.1.2 Detaching a Storage Cluster

Transition the storage cluster node except system management node to software maintenance mode.

```
# paclstmgr -c clstname -a --soft-mainte --service-stop
```

clstname: Cluster name

Perform the -v option of the pashowclst command to verify that the STATUS column is Disable, the REASON column is SoftMaintenance, and the SRV_STATUS column is - for all nodes of the storage cluster except the system management node.

```
# pashowclst -c clstname -v
[ CLST: clstname ]
NODE       NODETYPE     STATUS       REASON             PWR_STATUS       ARCH_STATUS      SRV_STATUS
0x00000001  SMM,SCM      Running      -                 on               -                IC(o),PWRD(o)
0x00000002  SMM,SCM      Running      -                 on               -                IC(o),PWRD(o)
0x00000003  MGS           Disable       SoftMaintenance     on               -                -
0x00000004  MGS           Disable       SoftMaintenance     on               -                -
0x00000005  MDS           Disable       SoftMaintenance     on               -                -
0x00000006  MDS           Disable       SoftMaintenance     on               -                -
0x00000007  OSS           Disable       SoftMaintenance     on               -                -
0x00000008  OSS           Disable       SoftMaintenance     on               -                -
```

7.1.1.3 Detaching a System Management Node

If you change the IP address of management network of a system management node, you must switch the system management node to the standby state. If the target node is active, switch with the --failover command of the paclstmgr command.

```
# paclstmgr -c clstname --failover nodeid1
Do you really want to continue (y/n) ? y
[INFO] PSM 4084 paclstmgr Failover process starts now. nodeid1 is going to failover to nodeid2.
...
[INFO] PSM 4085 paclstmgr Failover process finished normally. nodeid1 was failovered to nodeid2.
```

clstname: Cluster name
nodeid1: The node ID of the active node (Node ID of the node to standby by performing the paclstmgr command)
nodeid2: The node ID of the standby node (Node ID of the node to activate by performing the paclstmgr command)

Transition the standby system management node to the software maintenance mode.

```
# paclstmgr -c clstname -n nodeid --soft-mainte --service-stop
```

clstname: Cluster name
nodeid: the node ID of the standby system management node

Perform the -v option of the pashowclst command to verify that the STATUS column is Disable, the REASON column is SoftMaintenance, and the SRV_STATUS column is - for the standby system management node.

```
# pashowclst -c clstname --nodetype SMM -v
[ CLST: clstname ]
[ NODETYPE: SMM ]
```

```
NODE        NODETYPE      STATUS        REASON              PWR_STATUS        ARCH_STATUS       SRV_STATUS
0x00000001  SMM,SCM       Running       -                   on                -                 IC(o),PWRD(o)
0x00000002  SMM,SCM        Disable        SoftMaintenance      on                  -                   -
```

clstname: Cluster name

7.1.2 Configuring the OS and BIOS

Configure the OS and BIOS as required on the target node.
Perform the procedure for this work according to the OS or BIOS specifications.

- To change the set IP address for the management network in the OS, change the settings on the OS.
- To change the set IP address for the control network in the BIOS, change the settings on the BIOS.

7.1.3 Registering Cluster Configuration Information [SMM]

Create the cluster configuration definition file pxsystem.conf to define a configuration for the Job Operation Software. Use the --create option in the pxsystemadm command to set the file.

1. Creating the cluster configuration definition file pxsystem.conf

   Create the cluster configuration definition file pxsystem.conf.
   The following example changes the IP address of the management network for a node (node ID: 0xFFFF0005) in the cluster cluster1.

```
Operation {
    Type = "update"
}
Cluster {
    ClusterName = "cluster1"
    Node {
        NodeID = "0xFFFF0005"
        ManageNet = "192.0.2.254"
    }
}
```

   For details on the setting items in pxsystem.conf, see "pxsystem.conf file" in the "Job Operation Software Command Reference".

2. Registering the pxsystem.conf file contents

   Use the --create option in the pxsystemadm command to set the pxsystem.conf file contents on the system management node.

```
# pxsystemadm --create filename
Operation/Type is "update".                                  (*1)
Do you really want to continue (y/n)? y                      (*2)
[INFO] PSM 1082 pxsystemadm The setting was completed.
```

   (*1) The displayed type of operation will be performed.
   (*2) This prompts for confirmation before registration.

   filename: Path to the created cluster configuration definition file

3. Confirming the registered information

   Specify the -a option in the pashowclst command to confirm that the set contents are registered.

```
# pashowclst -a
...
```

7.1.4 Reflecting Cluster Configuration Information [SMM]

1. Reflecting cluster configuration information

   Use the --set option in the pxsystemadm command to get the system to reflect the registered information.

   Perform this step on all clusters.

```
# pxsystemadm --set -c clstname
Do you really want to continue (y/n)? y
[INFO] PSM 1082 pxsystemadm The setting was completed.
```

clstname: Cluster name

Information

This operation may take a while to process because communication takes place with all the nodes in the cluster.

See

Any node that is already powered off, for example, may fail to reflect the information. In such cases, an error list file is output to the current directory.

```
[ERR.] PSM 0020 pxsystemadm There are nodes that failed in the setting.(filename)
```

filename: Name of the error list file output to the current directory

If an error list file is output, check the file contents, and eliminate the cause of the error. Then, re-execute the command on the node that failed to reflect the information.
For details on troubleshooting based on an error message, see "Failed Reflection of Cluster Configuration Definitions" in the "Job Operation Software Troubleshooting".

2. Synchronizing Node Information

   To reflect the node information changes, execute the sync subcommand of the pxinst command to synchronize the configuration files created in "1. Reflecting cluster configuration information" to the standby system management node and the compute cluster sub management node.

```
# pxinst sync all
...
[INFO] INST 1805 pxinst File replication is complete
```

7.1.5 Configuring the Installation Function [SMM]

If the IP address allocated to a node has changed, the configuration file information held by the installation function must be updated. See "7.6 Registering Node Information Again," and change the installation function settings. Update the node information definition file to reflect the changed IP address.

7.1.6 Checking the Software Environment

From the software perspective, check whether the nodes where information was changed are ready to be incorporated into operation.

Execute the pachecker command, and confirm that settings are as expected.
For details on the pachecker command, see "Checking the Software Environment" in the "Job Operation Software Administrator's Guide for System Management".

7.1.7 Incorporating a Cluster Into Operation [SMM]

1. Restarting an active system management node

   Restart the active system management node. Perform the following on the active system management node.

```
# shutdown -r now
```

2. Incorporating a standby system management node

   A standby system management node is incorporated into operation by a cold reboot with the --cold-reboot option to the --recover option of the paclstmgr command.

For example, if the standby system management node is 0x00000002.

```
# paclstmgr -c clstname -n 0x00000002 --recover --cold-reboot
```

clstname: Cluster name

3. Incorporation confirming

   Verify that the STATUS column of the system management node is Running with the pashowclst command.

```
# pashowclst -c clstname  -v --nodetype SMM
[ CLST: clstname ]
[ NODETYPE: SMM ]
NODE       NODETYPE  STATUS       REASON            PWR_STATUS       ARCH_STATUS      SRV_STATUS
0x00000001  SMM,SCM   Running      -                 on               -                IC(o),PWRD(o)
0x00000002  SMM,SCM   Running      -                 on               -                IC(o),PWRD(o)
```

4. Incorporating a Cluster Into Operation

   Specify the --cold-reboot option for the --recover option in the paclstmgr command to incorporate the target cluster into operation with a cold reboot.

```
# paclstmgr -c clstname -a --recover --cold-reboot
```

clstname: Cluster name

For a multiple-cluster configuration, perform the paclstmgr command for storage cluster, and then for compute cluster.

Check the RUNNING column of the pashowclst command to ensure that all nodes are incorporated.

```
# pashowclst -c clstname
CLUSTER          CLSTTYPE  RUNNING  STOPPED  ERROR    DISABLE
clustname        COMPUTE   48       0        0        0
```

clstname: Cluster name

7.2 Changing an IP Address

This section describes the work for changing the IP address allocated to a node.

7.2.1 Changing Settings on the OS

Change an IP address on the OS on the target node.
Change it in a way appropriate to the specifications of the OS used.

The following example edits ifcfg-eth0 to change an IP address from 192.0.2.1 to 192.0.2.10.

1. Changing network settings

   Change and save IPADDR in /etc/sysconfig/network-scripts/ifcfg-eth0.

```
# vi /etc/sysconfig/network-scripts/ifcfg-eth0
DEVICE=eth0
IPADDR=192.0.2.1 => Change to 192.0.2.10
NETMASK=255.255.255.0
BOOTPROTO=none
TYPE=Ethernet
NAME=eth0
IPV4_FAILURE_FATAL=no
DEFROUTE=yes
ONBOOT=yes
IPV6INIT=no
```

2. Restarting the network service

   Restart the network service. If you are working with an ssh connection using the old IP address, operation hangs at the service restart time. If so, re-establish and confirm a connection using the new IP address.

```
# systemctl restart network
```

3. Confirming the change

```
# ip a
...
3: eth0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP qlen 1000
    link/ether 00:00:00:00:00:20 brd ff:ff:ff:ff:ff:ff
    inet 192.0.2.10/24 brd 192.168.122.255 scope global eth0 => Confirm changed IP address
       valid_lft forever preferred_lft forever
    inet6 fe80::200:ff:fe00:20/64 scope link
       valid_lft forever preferred_lft forever
...
```

7.2.2 Changing the Installation Function Settings

See "7.6 Registering Node Information Again," and change the installation function settings. Update the node information definition file to reflect the changed IP address.

Note

If the target IP address is not automatically set during OS reinstallation or restoring backup files, this procedure is unnecessary. However, if the IP address of the management network has changed, be sure to perform the procedure.
The installation function or backup/restore function uses the management network in OS reinstallation or backup/restore. Unless this procedure is performed, the function may use the old IP address and not be able to install the OS or perform a backup/restore.

7.3 Changing a Host Name

7.3.1 Changing Settings on the OS

Change a host name on the OS on the target node.

Change it in a way appropriate to the specifications of the OS used.

The following example changes a host name from node01 to node02.

1. Changing a host name

```
# hostname node02
```

2. Reflecting in a configuration file
   Since the change by the hostname command is temporary, edit a configuration file so that it reflects the host name change permanently even after a node restart.

```
# vi /etc/hostname
node01
```

   Change "node01" to "node02" as shown above.

3. Confirming the change

```
# hostname
node02
```

7.3.2 Changing the Installation Function Settings

See "7.6 Registering Node Information Again," and change the installation function settings. Update the node information definition file to reflect the changed host name.

Note

If you reinstall the OS or restore backup files without performing this procedure, the old host name is used.

7.4 Changing a MAC Address

If a MAC address has changed due to a network interface card failure, for example, change the installation function settings in preparation for OS reinstallation or restoring backup files.

7.4.1 Checking the MAC address

Confirm the changed MAC address. Check the MAC address obtained during repair or at OS startup.
The following example uses the ip command to check the MAC address.

```
# ip a
...
3: eth0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP qlen 1000
    link/ether 00:00:00:00:00:20 brd ff:ff:ff:ff:ff:ff        => Check MAC address
    inet 192.0.2.10/24 brd 192.168.122.255 scope global eth0
       valid_lft forever preferred_lft forever
    inet6 fe80::200:ff:fe00:20/64 scope link
       valid_lft forever preferred_lft forever
...
```

7.4.2 Changing the Installation Function Settings

See "7.6 Registering Node Information Again," and change the installation function settings. Update the node information definition file to reflect the changed MAC address.

7.5 Changing a Disk Partition

This section describes the work for changing a disk partition configuration assigned to a node.

7.5.1 Changing a Disk Partition Configuration

Change a disk partition configuration on the OS on the target node.
Change it in a way appropriate to the specifications of the OS used.

7.5.2 Registering Node Information Again

See "7.6 Registering Node Information Again," and registering node information again.
When registering node information again, Update the disk partition definition file to reflect the changed disk partition configuration.

Note

If you reinstall the OS without performing this procedure, will install the previous disk partition configuration.
If you attempt to restore the backup of a node whose disk partition configuration has changed without performing this procedure, you will get an error during the restore.

7.6 Registering Node Information Again

7.6.1 Registering PRIMERGY server Node Information Again [SMM]

If PRIMERGY server node configuration information has changed, update the information that was set by the installation function in preparation for OS reinstallation or backup file restoration due to a failure or the like. Register the node information again in the following procedure to update the information.

1. Deleting node information

   Delete the registered node information.

   ```
   # pxinst node del -m hostname
   [INFO] INST 1804 pxinst Deleting nodes completed.(nodenum=1)
   ```

   hostname: Node name to register again

2. Preparing a node information definition file

   You can create a node information definition file by extracting and using only the target node definitions from the file used at build or by using an installation design file.If you are changing configuration information, rewrite the definitions as required.

   Information

   Provides a sample file of a node information definition file for manual creation.
   /etc/opt/FJSVtcs/sample/pxinst_nodedef.conf

   See

   For details on the node information definition file format, see "Node Information Definition File" in the "Job Operation Software Administrator's Guide for System Management."

3. Preparing a disk partition definition file

   You can create a disk partition definition file by extracting and using only the target node definitions from the file used at build or by using an installation design file. If you are changing configuration information, rewrite the definitions as required.

   Information

   Provides a sample file of a disk partition information file for manual creation.
   /etc/opt/FJSVtcs/sample/pxinst_partdef.conf

   See

   For details on the disk partition definition file format, see "Disk Partition Definition File" in the "Job Operation Software Administrator's Guide for System Management".

4. Registering node information

   Execute the node subcommand of the pxinst command to register the node information.

   ```
   # pxinst node add -R reponame -N nodefile -P diskfile -T disk
   [INFO] INST 1803 pxinst Adding nodes completed.(nodenum=1)
   ```

   reponame: Repository name for node reregistration
   nodefile: Path to the node information definition file
   diskfile: Path to the disk partition information definition file

Note

- Be sure to specify "disk" in the -T option. If the -T option is omitted or the specified value is not "disk", OS installation processing may run at the next node restart time.
- The node information definition file and disk partition information definition file are required when making system configuration changes.
  Please make sure to take a backup.
  The following example creates a backup in /root/packagekit/.

```
# mkdir -p /root/packagekit
# cp nodefile /root/packagekit/
# cp diskfile /root/packagekit/
```

  nodefile: Path to the node information definition file
  diskfile: Path to the disk partition information definition file

5. Synchronizing Node Information

   To reflect the node information changes, perform the sync subcommand of the pxinst command to the standby system management node and the compute cluster sub management nodes.
   However, if the standby system management node and the compute cluster sub management nodes do not exist in the system, this operation is not necessary.

```
# pxinst sync srv
...
[INFO] INST 1805 pxinst File replication is complete.
```

7.6.2 Registering FX server Node Information Again [SMM]

If FX server node configuration information has changed, update the information that was set by the installation function in preparation for OS reinstallation or backup file restoration due to a failure or the like. Register the node information again in the following procedure to update the information.

Note

If you installed the FX server using a dynamic IP address, be sure to save the DHCP service configuration file (/etc/dhcp/dhcpd.conf). If you have not saved the DHCP service configuration file, you will need to reinstall the OS in order to backup and restore the node.

1. Saving DHCP Service Configuration Files

   If you installed the FX server using a dynamic IP address, the information registered before the deletion of node information is required. Save the DHCP service configuration file (/etc/dhcp/dhcpd.conf) in advance.
   If the FX server is not installed using a dynamic IP address, this operation is not necessary.

```
# cp -p /etc/dhcp/dhcpd.conf /etc/dhcp/dhcpd.conf.bak
```

2. Deleting node information

   Delete the registered node information.

```
# pxinst node del -m hostname
[INFO] INST 1804 pxinst Deleting nodes completed.(nodenum=1)
```

   hostname: Node name to register again

3. Preparing a node information definition file for the FX Server

   You can create a node information definition file by extracting and using only the target BoB definitions from the file used at build or by using an installation design file. If you are changing configuration information, rewrite the definitions as required.

Information

Provides a sample file of a node information definition file for the FX Server for manual creation.
/etc/opt/FJSVtcs/sample/pxinst_nodedef_ft.conf

See

For details on the node information definition file format for the FX server, see "Node Information Definition File for the FX Server" in the "Job Operation Software Administrator's Guide for System Management."

4. Registering node information

   Execute the node subcommand of the pxinst command to register the node information.

   When you register node information, you register the FX server node information by specifying the --ft option, which means FX server.
   Register without the --pool option, even if you use dynamic IP addresses when you reregister node information.

```
# pxinst node add -R reponame -N ft_file --ft
[INFO] INST 1803 pxinst Adding nodes completed.(nodenum=xxx)
```

   reponame: Repository name for node reregistration
   ft_file: Path to the node information definition file for the FX Server
   xxx: Number of nodes

Note

The node information definition file for the FX Server is required when making system configuration changes.
Please make sure to take a backup.
The following example creates a backup in /root/packagekit/.

```
# mkdir -p /root/packagekit
# cp ft_file /root/packagekit/
```

ft_file: Path to the node information definition file for the FX Server

Information

If you specify a range that is already registered as an IP address pool (IP address range to assign) with the --pool option when registering node information, an error occurs. For node reregistration, you do not need to specify the --pool option because the IP address pool is already registered.

5. Updating Node Information

   Immediately after registering the FX server node information, the OS reinstallation is configured to operate the next PXE boot. Therefore, run the following command on the node that reregistered the node so that it will boot to disk the next time you PXE boot. However, if reinstall the operating system, this operation is not necessary.

```
# pxinst node update -m hostname -T disk
[INFO] INST 0800 pxinst node completed successfully.(update)
```

   hostname: Node name to register again

   Clear status information after changing node information.

```
# pxinst status clear -m hostname
[INFO] INST 0800 pxinst status completed successfully.(clear)
```

   hostname: Node name to register again

Ensure that no status information is displayed for nodes that have re-registered node information.

```
# pxinst status show
```

6. Synchronizing Node Information

   To reflect the node information changes, perform the sync subcommand of the pxinst command to the standby system management node and the compute cluster sub management nodes.
   However, if the standby system management node and the compute cluster sub management nodes do not exist in the system, this operation is not necessary.

```
# pxinst sync srv
...
[INFO] INST 1805 pxinst File replication is complete.
```

7. Modifying DHCP Service Settings

   If you installed the FX server using a dynamic IP address, the information about the node being reregistered has been removed from the DHCP service configuration file (/etc/dhcp/dhcpd.conf) and must be added.
   If you do not add the information, you cannot takeor restore the backup, so please check the difference between the dhcpd.conf that has been saved with the deleted information and the current /etc/dhcp/dhcpd.conf.

```
# diff /etc/dhcp/dhcpd.conf.bak /etc/dhcp/dhcpd.conf
```

   Add the removed information to /etc/dhcp/dhcpd.conf.
   Examples of the information to add are.

```
group {
    next-server 192.0.2.10;                          (1)
    host hostname-bio01 {                            (2)
        option host-name "hostname-bio01";           (3)
        hardware ethernet 00:00:00:00:00:20;         (4)
        fixed-address 192.0.2.101;                   (5)
    }
}
```

   If you have changed the following configuration information, modify /etc/dhcp/dhcpd.conf to reflect the changes when you add the information.

Table 7.2 Correspondence between configuration information to be changed and changes in dhcpd.conf

Node type to be changed	Configuration information you want to change	Changes to dhcpd.conf
System management node (For a single unit configuration)	IP address of management network	IP address listed as the next-server ((1) Information)
System management node (For redundant configuration)	Representative IP address of management network	
Compute cluster sub management node (For a single unit configuration)	IP address of management network	
Compute cluster sub management node (For redundant configuration)	Representative IP address of management network	
Compute node also serving as boot I/O node (FX server)	Host name	Host name listed as host ((2) Information) Host name listed as option host-name ((3) Information)
	IP address of management network	IP address listed as fixed-address ((5) Information)

Node type to be changed	Configuration information you want to change	Changes to dhcpd.conf
	MAC address of management network	MAC address listed as hardware ethernet ((4) Information)

8. Restarting the DHCP Service

   If you make any configuration changes to the DHCP service, restart the DHCP service.

   However, if you have not made any configuration changes to the DHCP service, you do not need to perform this operation.

```
# systemctl restart dhcpd
```

Appendix A Configuring GPUs

The contents of this appendix cover the NVIDIA V100, A100 or H100. Here, the term "GPU" refers to the NVIDIA V100, A100 or H100.
This appendix describes the procedures for installing a driver and installing and configuring the NVIDIA Container Toolkit on nodes equipped with GPUs (NVIDIA V100, A100 or H100).
The NVIDIA Container Toolkit provides tools for using GPUs from Docker containers.

See

For details on the NVIDIA Container Toolkit, see https://github.com/nvidia/nvidia-docker or other sites.

A.1 Preparing Executable Files and Packages

To use GPUs, obtain and transfer the required driver and packages to the system management node.

The NVIDIA driver and the executable files and packages of the NVIDIA Container Toolkit are listed below:

- NVIDIA driver
- NVIDIA CUDA Toolkit
- NVIDIA Container Toolkit packages
  - nvidia-container-runtime
  - nvidia-container-toolkit
  - libnvidia-container1
  - libnvidia-container-tools
- NVIDIA Fabric Manager (compute nodes equipped with NVLink)

See

- For details on the NVIDIA driver, see https://www.nvidia.com/Download/index.aspx.
- For details on the NVIDIA CUDA Toolkit, see https://developer.nvidia.com/cuda-downloads.
- For details on how to obtain the NVIDIA Container Toolkit packages, see https://nvidia.github.io/nvidia-container-runtime/.
- For details on how to obtain and apply NVIDIA Fabric Manager, see https://docs.nvidia.com/datacenter/tesla/fabric-manager-user-guide/index.html.

Note

The applicable nvidia-container-runtime package of the NVIDIA Container Toolkit varies depending on the docker-ce version. Check the version of docker-ce installed on each target compute node.
The following example checks the docker-ce version. In the example, the docker-ce version is 18.06.03.

```
[Compute node]
# rpm -qa | grep docker-ce
docker-ce-18.06.3.ce.k3-3.el7.x86_64
```

The applicable package for docker-ce version 18.06.03 is nvidia-container-runtime-*version*.docker18.06.3.

version: Version of nvidia-container-runtime
Use version 2.0.0 of nvidia-container-runtime.

A.2 Transitioning to Software Maintenance Mode

Place compute nodes equipped with GPUs in software maintenance mode.

1. Isolating the compute nodes from operation

   Isolate the target compute nodes from operation by using the --disable option of the paclstmgr command.

   ```
   # paclstmgr -c clstname --rscunit rscuname --disable
   ```

   clstname: Cluster name
   rscuname: Resource unit name

2. Confirming the state of isolation from operation

   Check the status of nodes by using the -v option of the pashowclst command.
   Confirm transitions to "Disable" in the STATUS field and "Manual" in the REASON field for each target compute node.

   ```
   # pashowclst -v -c clstname
   [ CLST: clstname ]
   NODE        NODETYPE STATUS   REASON           PWR_STATUS   ARCH_STATUS   SRV_STATUS
   ...
   0xFFFF0004  CN       Disable  Manual           os-running   -             PLE(o),NRD(o),PWRD(o)
   ...
   ```

   clstname: Cluster name

3. Transitioning to software maintenance mode

   Use the --soft-mainte option of the paclstmgr command to get the target compute nodes to transition to software maintenance mode.
   At this time, specify the --service-stop option, which means the Job Operation Software services will stop.

   ```
   # paclstmgr -c clstname --rscunit rscuname --soft-mainte --service-stop
   ```

   clstname: Cluster name
   rscuname: Resource unit name

4. Confirming that the state is software maintenance mode

   Check the status of nodes by using the -v option of the pashowclst command.
   Confirm transitions to "SoftMaintenance" in the REASON field and a hyphen (-) in the SRV_STATUS field for each target compute node.

   ```
   # pashowclst -v -c clstname
   [ CLST: clstname ]
   NODE        NODETYPE STATUS   REASON           PWR_STATUS   ARCH_STATUS   SRV_STATUS
   ...
   0xFFFF0004  CN       Disable  SoftMaintenance  os-running   -             -
   ...
   ```

   clstname: Cluster name

A.3 Applying the NVIDIA Driver

Apply the NVIDIA driver to compute nodes equipped with GPUs. For details on how to apply the NVIDIA driver, see the NVIDIA installation guide.

A.4 Applying the NVIDIA Fabric Manager

Apply the NVIDIA Fabric Manager to compute nodes equipped with NVLink. For details on how to apply the NVIDIA Fabric Manager, see the NVIDIA installation guide.

A.5 Applying the NVIDIA CUDA Toolkit

Apply the NVIDIA CUDA Toolkit to compute nodes equipped with GPUs. For details on how to apply the NVIDIA CUDA Toolkit, see the NVIDIA installation guide.

A.6 Applying the NVIDIA Container Toolkit

Apply each NVIDIA Container Toolkit package to compute nodes equipped with GPUs.

1. Confirming the path to the package storage directory

   The package will be registered in a repository. Check the repository by using the repo subcommand of the pxinst command.

```
# pxinst repo list
REPONAME      PATH                          PKGDIR     OS      VERSION    ARCH
--------      ----                          ------     --      -------    ----
reponame      repopath                      dirname    os      version    arch
...
```

   reponame: Repository name
   repopath: Repository path
   dirname: Directory for placing the package within the repository
   os: OS type
   version: OS version
   arch: Architecture

   Check the line showing the repository used by nodes to apply the package.
   "*repopath*/*dirname*" is the package storage path.

2. Adding the package

   Place the package in the confirmed directory.

```
# cp pkgname repopath/dirname
```

   pkgname: Package name
   repopath: Repository path
   dirname: Directory for placing the package within the repository

3. Updating the repository

   Update the repository.

```
# pxinst repo update -R reponame
[INFO] INST 0800 pxinst repo completed successfully.(update repository)
```

   reponame: Repository name

4. Synchronizing the repository

   Apply the repository changes to the standby system management node and the compute cluster sub management node.
   If the system contains neither a standby system management node nor a compute cluster sub management node, this operation is unnecessary.

```
# pxinst sync repo
...
[INFO] INST 1805 pxinst File replication completed.
```

5. Deleting cache information

   Before applying the package to a node, delete cache information from the node by using the pmexe command.
   The following example uses the pmexe command to execute the yum command on all nodes in the target resource unit.

```
# pmexe -c clstname --rscunit rscuname "yum clean all"
```

clstname: Cluster name
rscuname: Resource unit name

6. Applying the package

   Apply the package to the target nodes by using the pmexe command.
   The following example uses the yum command to apply the package to all nodes in the target resource unit.

```
# pmexe -c clstname --rscunit rscuname "yum -y install pkgname"
```

   clstname: Cluster name
   rscuname: Resource unit name
   pkgname: Package name

A.7 Configuring Docker for the NVIDIA Container Toolkit

Configure Docker to enable the NVIDIA Container Toolkit.

1. Creating daemon.json

   Create a Docker configuration file (/etc/docker/daemon.json).
   Distribute the sample Docker configuration file (/etc/opt/FJSVtcs/sample/docker.daemon.nvidia.json) to the compute nodes where the NVIDIA Container Toolkit has been applied.
   The following example distributes the file to all nodes in the target resource unit.

```
[System management node]
# pmscatter -c clstname --rscunit rscuname etc/opt/FJSVtcs/sample/docker.daemon.nvidia.json \
/etc/docker/daemon.json
```

   clstname: Cluster name
   rscuname: Resource unit name

A.8 Configuring the Number of GPUs on a Compute Node with NVLink

If the compute node has NVLink, create a configuration file for the number of GPUs.
This step is not required if the compute node does not have NVLink.

The following example creates a GPU count configuration file on a system management node and distribute it to the nodes belonging to the target resource unit.

1. Create a GPU count configuration file (/var/opt/FJSVtcs/psm/gpunumsettings.conf)

```
[System management node]
# echo GPUNUM > gpunumsettings.conf
```

   GPUNUM: Number of GPUs

2. Set permissions on gpunumsettigs.conf

```
[System management node]
# chmod 400 gpunumsettings.conf
```

3. Distribute to the nodes belonging to the target resource unit

```
[System management node]
# pmscatter -c clstname --rscunit rscuname gpunumsettings.conf \
/var/opt/FJSVtcs/psm/gpunumsettings.conf
```

   clstname: Cluster name
   rscuname: Resource unit name

A.9 Recovering From Software Maintenance Mode

Recover compute nodes equipped with GPUs from software maintenance mode.

1. Recovering from software maintenance mode

   Recover the target compute nodes from software maintenance mode by using the --recover option of the paclstmgr command. Also, specify --cold-reboot to restart the nodes.

   ```
   # paclstmgr -c clstname --rscunit rscuname --recover --cold-reboot
   ```

   clstname: Cluster name
   rscuname: Resource unit name

2. Confirming the cancellation of software maintenance mode

   Check the status of the target compute nodes by using the -v option of the pashowclst command.
   Confirm that the nodes are in the same state as that confirmed in step 2 in "2.3.11 Reflecting Cluster Configuration Information [SMM]."

A.10 Configuring Job Operations to Use GPUs

Configure settings for using GPUs in job operations.

For details on how to configure the settings, see "Settings for Job Execution Using GPUs" in the "Job Operation Software Administrator's Guide for Job Management."

Appendix B Configuring Singularity

Singularity is container virtualization software for HPC.
This appendix describes the procedures for installing and configuring Singularity.

See

For details on Singularity, see https://sylabs.io/singularity/ or other sites.

B.1 Preparing the Package

Obtain and transfer the Singularity package (package name: singularity) to the system management node.

See

See https://github.com/sylabs/singularity for the Singularity package.
You can obtain some versions of the Singularity package from EPEL (https://fedoraproject.org/wiki/EPEL).

Note

Use version 3.2.1 of Singularity.

B.2 Transitioning to Software Maintenance Mode

To apply Singularity package to compute nodes, place the nodes in software maintenance mode.

1. Isolating the compute nodes from operation

   Isolate the target compute nodes from operation by using the --disable option of the paclstmgr command.

   ```
   # paclstmgr -c clstname --rscunit rscuname --disable
   ```

   clstname: Cluster name
   rscuname: Resource unit name

2. Confirming the state of isolation from operation

   Check the status of nodes by using the -v option of the pashowclst command.
   Confirm transitions to "Disable" in the STATUS field and "Manual" in the REASON field for each target compute node.

   ```
   # pashowclst -v -c clstname
   [ CLST: clstname ]
   NODE        NODETYPE STATUS   REASON          PWR_STATUS   ARCH_STATUS   SRV_STATUS
   ...
   0xFFFF0004  CN       Disable  Manual          os-running   -             PLE(o),NRD(o),PWRD(o)
   ...
   ```

   clstname: Cluster name

3. Transitioning to software maintenance mode

   Use the --soft-mainte option of the paclstmgr command to get the target compute nodes to transition to software maintenance mode.
   At this time, specify the --service-stop option, which means the Job Operation Software services will stop.

   ```
   # paclstmgr -c clstname --rscunit rscuname --soft-mainte --service-stop
   ```

   clstname: Cluster name
   rscuname: Resource unit name

4. Confirming that the state is software maintenance mode

   Check the status of nodes by using the -v option of the pashowclst command.
   Confirm transitions to "SoftMaintenance" in the REASON field and a hyphen (-) in the SRV_STATUS field for each target compute node.

```
# pashowclst -v -c clstname
[ CLST: clstname ]
NODE        NODETYPE STATUS   REASON           PWR_STATUS   ARCH_STATUS   SRV_STATUS
...
0xFFFF0004  CN       Disable  SoftMaintenance  os-running   -             -
...
```

   clstname: Cluster name

B.3 Applying the Package

Apply the Singularity package to the compute nodes.

1. Confirming the path to the package storage directory

   The package will be registered in a repository. Check the repository by using the repo subcommand of the pxinst command.

```
# pxinst repo list
REPONAME      PATH                          PKGDIR     OS      VERSION    ARCH
--------      ----                          ------     --      -------    ----
reponame      repopath                      dirname    os      version    arch
...
```

   reponame: Repository name
   repopath: Repository path
   dirname: Directory for placing the package within the repository
   os: OS type
   version: OS version
   arch: Architecture

   Check the line showing the repository used by nodes to apply the package.
   "*repopath*/*dirname*" is the package storage path.

2. Adding the package

   Place the package in the confirmed directory.

```
# cp pkgname repopath/dirname
```

   pkgname: Package name
   repopath: Repository path
   dirname: Directory for placing the package within the repository

3. Updating the repository

   Update the repository.

```
# pxinst repo update -R reponame
[INFO] INST 0800 pxinst repo completed successfully.(update repository)
```

   reponame: Repository name

4. Synchronizing the repository

   Apply the repository changes to the standby system management node and the compute cluster sub management node.
   If the system contains neither a standby system management node nor a compute cluster sub management node, this operation is unnecessary.

```
# pxinst sync repo
...
[INFO] INST 1805 pxinst File replication completed.
```

5. Deleting cache information

   Before applying the package to a node, delete cache information from the node by using the pmexe command.
   The following example uses the pmexe command to execute the yum command on all nodes in the target resource unit.

```
# pmexe -c clstname --rscunit rscuname "yum clean all"
```

   clstname: Cluster name
   rscuname: Resource unit name

6. Applying the package

   Apply the package to the target nodes by using the pmexe command.
   The following example uses the yum command to apply the package to all nodes in the target resource unit.

```
# pmexe -c clstname --rscunit rscuname "yum -y install pkgname"
```

   clstname: Cluster name
   rscuname: Resource unit name
   pkgname: Package name

B.4 Recovering From Software Maintenance Mode

After applying the Singularity package to compute nodes, recover the nodes from software maintenance mode.

1. Recovering from software maintenance mode

   Recover the target compute nodes from software maintenance mode by using the --recover option of the paclstmgr command.
   Also, specify --service-restart to restart services.

```
# paclstmgr -c clstname --rscunit rscuname --recover --service-restart
```

2. Confirming the cancellation of software maintenance mode

   Check the status of the target compute nodes by using the -v option of the pashowclst command.
   Confirm that the nodes are in the same state as that confirmed in step 2 in "2.3.11 Reflecting Cluster Configuration Information [SMM]."

B.5 Configuring Job Operations to Use Singularity

Configure settings for using Singularity in job operations.

For details on how to configure the settings, see "Settings for Using Singularity" in the "Job Operation Software Administrator's Guide for Job Management."

Appendix C Preparing Subscripted Servers

If you are installing Red Hat Enterprise Linux 8 on a node in the cluster, you must have a subscripted server to obtain the Red Hat Enterprise Linux 8 update packages and repository metadata.

At least one subscripted server that can connect to the internet must be available as a repository download server for each architecture (x86_64, aarch64) used in the cluster. A subscripted server can also serve as a node in the cluster.

Information

Systems that use the Red Hat Enterprise Linux 8 DVD for OS installation and do not plan to apply the package fixes managed by the AppStream repository do not require the provision of a subscripted server.