C122-B001-10EN

PRIMEQUEST
580A/540A/580/540/480/440

SYSTEM DESIGN GUIDE

FUJITSU

FOR SAFE OPERATION

This manual contains important information regarding the use and handling of this product. Read this manual thoroughly. Pay special attention to the section "NOTE ON SAFETY" Use the product according to the instructions and information available in this manual. Keep this manual handy for further reference.

Fujitsu makes every effort to prevent users and bystanders from being injured or from suffering damage to their property. Use the product according to this manual.

ABOUT THIS PRODUCT

This product is designed and manufactured for use in standard applications such as office work, personal device, household appliance, and general industrial applications. This product is not intended for use in nuclear-reactor control systems, aeronautical and space systems, air traffic control systems, mass transportation control systems, medical devices for life support, missile launch control systems or other specialized uses in which extremely high levels of reliability are required, the required levels of safety cannot be guaranteed, or a failure or operational error could be life-threatening or could cause physical injury (referred to hereafter as "high-risk" use). You shall not use this product without securing the sufficient safety required for high-risk use. If you wish to use this product for high-risk use, please consult with sales representatives in charge before such use.

RADIO FREQUENCY INTERFERENCE STATEMENT

The following notice is for EU users only.

WARNING: This is a product which meets Class A of EN55022. In a domestic environment this product may cause radio interference in which case the user may be required to take adequate measures.

The following notice is for USA users only.

This equipment has been tested and found to comply with the limits for a Class A digital device, pursuant to Part 15 of the FCC Rules. These limits are designed to provide reasonable protection against harmful interference when the equipment is operated in a commercial environment.
This equipment generates, uses, and can radiate radio frequency energy and, if not installed and used in accordance with the instruction manual, may cause harmful interference to radio communications. Operation of this equipment in a residential area is likely to cause harmful interference in which case the user will be required to correct the interference at his own expense.

Laser standards

This equipment includes Class 1 laser products and complies with FDA Radiation Performance Standards, 21 CFR 1040.10 and 1040.11, and the International Laser Safety Standards IEC60825-1: 2001.

TRADEMARK ACKNOWLEDGEMENTS

- Microsoft, Windows, MS, Windows NT, and Windows Server are trademarks or registered trademarks of Microsoft Corporation in the United States and/or other countries.
- Linux is a trademark or registered trademark of Linus Torvalds in the United States and other countries.
- Red Hat, RPM, and all Red-Hat-based marks and logos are trademarks or registered trademarks of Red Hat, Inc. in the United States and other countries.
- SUSE is a registered trademark of Novell Inc. in the United States and other countries.
- Intel, Xeon, and Itanium are trademarks or registered trademarks of Intel Corporation.
- Ethernet is a registered trademark of Fuji Xerox Co., Ltd. and Xerox Corporation in the United States and other countries.
- All other product names mentioned herein are the trademarks or registered trademarks of their respective owners.
- System and product names in this manual are not always noted with trademark or registered trademark symbols (™), (®).

TERMS AND CONDITONS

The product includes software provided by third parties in addition to that provided by Fujitsu Ltd. You are granted permission to use the third parties' software subject to the terms and conditions below. If you acquire the source code of the software to which the following terms and conditions apply, refer to LICENSE1_EN.pdf and LICENSE2_EN.pdf, which are provided with the 'PRIMEQUEST Manuals' (C122-E013-C2).

THIS SOFTWARE IS PROVIDED "AS IS" AND FUJITSU LIMITED MAKES NO WARRANTIES, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER WHATSOEVER REGARDING TO THIS SOFTWARE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE.

IN NO EVENT SHALL FUJITSU LIMITED BE LIABLE FOR ANY CLAIM FROM A THIRD PARTY, OR SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE, COPYING, MODIFICATION OR DISTRIBUTION OF THIS SOFTWARE.

Fujitsu Ltd, April 2006

The contents of this manual shall not be disclosed in any way or reproduced in any media without the express written permission of Fujitsu Limited.

All Rights Reserved, Copyright © FUJITSU LIMITED 2005-2008

C122-B001-10EN

Revision History

(1/1)

Edition	Date	Revised section (Added/ Deleted/ Altered) (Note)	Details
01	2005-07-11	—	—
02	2005-09-16	All chapters (correction) Section 9.3 (addition)	Technical brushup Addition of description for cluster of Windows
03	2005-10-31	Chapter 2, 3, 5, 9, 10, 11 (correction) Appendix A, B (correction)	Error corrections
04	2006-2-20	All chapters (correction)	Technical brushup Modification of description of Mirror Mode
05	2006-4-17	Section 4.1, 4.3 (addition)	Addition of notes
06	2006-8-11	All chapters (correction) Chapter 1, 2, and 7	Addition of descriptions for PRIMEQUEST 500 series and Linux SUSE Correction of description for partition configuration with XPAR installed
07	2007-03-05	Chapter 1, 2, 4, 7 (correction) Chapter 9 (correction)	Addition of descriptions for GTHB Addition of description for supported backup software
08	2007-08-31	Section 2.14 (correction) Chapter 10 (correction)	Addition of description for PEXU Addition of Linux dump function kdump
09	2008-03-10	Entire manual (addition)	Addition of description for PRIMEQUEST 580A/540A
10	2008-04-10	Section 4.2.8 (correction)	Correction of description for NTP operation

Note: In this table, the revised section is indicated by its section number in the current edition.
An asterisk (*) indicates a section in the previous edition.

Preface

This manual describes conditions and points for consideration for PRIMEQUEST system operation design and provides related essential notes.

The manual is intended for system administrators. Read the manual together with the reference manuals cited in it.

This section explains

- Structure and Contents of This Manual
- Other Reference Manuals
- Abbreviations
- Text Conventions
- Syntax of the Command Line Interface (CLI)
- Notes Regarding Notations Used in This Manual
- Conventions for Alert Messages
- Environmental Requirements for Using This Product
- Reader Feedback

Structure and Contents of This Manual

This manual is organized as described below.

CHAPTER 1 PRIMEQUEST Overview

Describes features of the main unit and software.

CHAPTER 2 General Hardware Configuration

Describes the hardware mounted in the main unit.

CHAPTER 3 General Software Configuration

Describes the OS, firmware, supplied software, and software that can be linked.

CHAPTER 4 Hardware System Management

Describes the basic mechanism related to hardware system management, functions of the Management Board (MMB) that manages the main unit as a whole, and functions of the PRIMEQUEST Server Agent (PSA) managed by the MMB.

CHAPTER 5 Redundant Configuration

Describes the redundant configuration of main unit components, including the System Mirror Mode and redundant LANs and HDDs.

CHAPTER 6 Hot Swapping

Describes hot swapping of hardware components of the main unit.

CHAPTER 7 Partition

Describes how to set up and operate partitions of the main unit and describes the necessary items for determining the partition configuration.

CHAPTER 8 Clustering

Describes clustering and provides essential notes on configuring a cluster system.

CHAPTER 9 Backup and Restoration

Describes system maintenance by explaining the necessity of backups, how to back up and restore the PRIMEQUEST system, and other such topics.

CHAPTER 10 Considerations for Maintenance

Describes the system maintenance considerations that must be made during system design.

Appendix A Software Supplied with PRIMEQUEST Hardware

Describes the software supplied with the PRIMEQUEST.

Glossary

Explains the terms used in this manual.

Index

Provides keywords and corresponding reference page numbers so that the reader can easily search for items in this manual as necessary.

Other Reference Manuals

The following manuals are provided for reference:

a) PDF manuals included on the *PRIMEQUEST Manuals* CD-ROM disk (C122-E013-C2)

Title	Description	Manual code
PRIMEQUEST 580A/540A/580/540/480/440 System Design Guide	Explains requirements, considerations, and notes on the system operation design of the PRIMEQUEST 580A/540A/580/540/480/440.	C122-B001EN
PRIMEQUEST 580A/540A/580/540/480/440 Installation Planning Manual	Explains specifications and requirements for installation sites that are applicable to the installation of the PRIMEQUEST 580A/540A/580/540/480/440.	C122-H001EN
PRIMEQUEST 500A/500/400 Series Installation Manual	Explains the setup of the PRIMEQUEST, including the preparation for the installation, initial settings, and software installation.	C122-E001EN
PRIMEQUEST 580A/540A/520A/500/400 Series Reference Manual: Basic Operation/GUI/Commands	Explains operations, setup methods, and the system management method that are required for the system operation of the PRIMEQUEST. The explanation covers basic operations and functions of the MMB, PSA, and EFI.	C122-E003EN
PRIMEQUEST 500A/500/400 Series Reference Manual: Tools/Operation Information	Explains system maintenance, Hot Plug, REMCS, and LEDs and other information required for system operation. Also, the manual provides supplementary information such as information on the physical locations of components.	C122-E074EN
PRIMEQUEST 500A/500/400 Series Reference Manual: Messages/Logs	Explains measures to be taken against problems that occur during operation and describes various types of messages.	C122-E004EN
PRIMEQUEST GSWB User's Manual	Explains the requirements, points to consider, and notes concerning installing and operating GSWB, an optional product.	C122-E028EN
SPARC Enterprise/PRIMEQUEST Common Installation Planning Manual	Explains basic information and policy on installation planning and facilities planning that are required for the installation of the SPARC Enterprise series and PRIMEQUEST series.	C120-H007EN
PRIMEQUEST 580A/580 Dynamic Partitioning (DP) Manual	Explains the Dynamic Partitioning (DP) function which the PRIMEQUEST 580A/540A supports.	C122-E085EN

b) Printed manual

For the printed manual (sold separately), contact your certified service engineer.

- *PRIMEQUEST 500A/500/400 Series Installation Manual* (C122-E001EN)

Abbreviations

In this manual, the product names are abbreviated as follows:

Long title	Abbreviations
Red Hat® Enterprise Linux® AS (v.4 for Itanium)	Red Hat (*1)
Red Hat® Enterprise Linux® 5 (for Intel Itanium)	
Red Hat® Enterprise Linux® AS (v.4 for Itanium)	RHEL-AS4 (IPF)
Red Hat® Enterprise Linux® 5 (for Intel Itanium)	RHEL5 (IPF) (*2)
SUSE™ Linux Enterprise Server 9 for Itanium Processor Family	SUSE
SUSE™ Linux Enterprise Server 10 for Itanium Processor Family	
SUSE™ Linux Enterprise Server 9 for Itanium Processor Family	SUSE9
SUSE™ Linux Enterprise Server 10 for Itanium Processor Family	SUSE10
Microsoft® Windows Server® 2003, Enterprise Edition for Itanium-based Systems	Windows Windows Server 2003
Microsoft® Windows Server® 2003, Datacenter Edition for Itanium-based Systems	
Microsoft® Windows Server® 2008 for Itanium-Based Systems	Windows Windows Server 2008

*1: Version-independent abbreviation

*2: A description in the form of "RHEL5.x (IPF)" indicates an updated version.

Text Conventions

This manual uses the following fonts and symbols to express specific types of information.

Fonts/symbols	Meaning	Example
Italic	Indicates names of manuals.	See the *PRIMEQUEST 580A/540A/580/540/480/440 System Design Guide*.
" "	Indicates names of chapters, sections, items, buttons, or menus.	See CHAPTER 5, "Business LAN Management by GSWB."
[]	Indicates window names, window button names, tab names, and dropdown menu selections.	Click the [OK] button.

Syntax of the Command Line Interface (CLI)

The command syntax is described below.

Command syntax

The command syntax is as follows:

- A variable that requires input of a value must be enclosed in < >.
- An optional element must be enclosed in [ ].
- A group of options for an optional keyword must be enclosed in [ ] and delimited by |.
- A group of options for a mandatory keyword must be enclosed in { } and delimited by |.

The command syntax is shown in a frame such as this one.

Notes Regarding Notations Used in This Manual

- Items marked with "Linux" apply to both Red Hat® Enterprise Linux® AS (v.4 for Itanium), Red Hat® Enterprise Linux® 5 (for Intel Itanium), SUSE™ Linux Enterprise Server 9 for Itanium Processor Family, and SUSE™ Linux Enterprise Server 10 for Itanium Processor Family (*).
  * For details, contact your Fujitsu certified service engineer.
- The IO Unit is indicated as "IOU" in the MMB Web-UI and in the figures shown in this manual.

Conventions for Alert Messages

This manual uses the following conventions to show alert messages. An alert message consists of an alert signal and alert statements.

WARNING	This indicates a hazardous situation that *could* result *in serious personal injury* if the user does not perform the procedure correctly.
CAUTION	This indicates a hazardous situation that *could* result *in minor or moderate personal injury* if the user does not perform the procedure correctly. This signal also indicates that damage to the product or other property *may* occur if the user does not perform the procedure correctly.
IMPORTANT	This indicates information that could help the user to use the product more effectively.

Alert messages in the text

In the text, alert messages are indented to distinguish then from regular text. A wider space precedes and follows the message to show where the message begins and ends.

WARNING

Certain tasks in this manual should only be performed by a certified service engineer. Users must not perform these tasks. Incorrect operation of these tasks may cause electric shock, injury, or fire.

- Installation and reinstallation of all components, and initial settings
- Removal of front, rear, or side covers
- Mounting/de-mounting of optional internal devices

Environmental Requirements for Using This Product

This product is a computer which is intended to be used in a computer room. For details on the operational environment, see the *PRIMEQUEST 580A/540A/580/540/480/440 Installation Planning Manual* (C122-H001EN).

Reader Feedback

- In this manual, it is assumed that two BMMs (optional products) can be connected to a single IO Unit; this is reflected both in the explanations and in the figures included in this manual. At present, however, the PRIMEQUEST 480/400 supports only connection to one BMM (BMM#0) per IO Unit.
- In this manual, the term BP (BackPlane) used in descriptions for the PRIMEQUEST 480/440 series actually stands for MP (MidPlane).
- The screen images in this manual may be different from the actual screen images.
- If you find any errors or unclear statements in this manual, please fill in the "Reader's Comment Form" sheet at the back of this manual and forward it to the address indicated at the bottom of the sheet.
- This manual is subject to revision without prior notice.
- The PDF version of this manual is best viewed in Adobe® Reader® with a magnification of 100% and Single Page for the page layout.

NOTE ON SAFETY

Alert Labels

The following labels are attached to this product:

These labels provide information to the users of this product.

CAUTION

Do not peel off the labels.

- Main unit

Front view

Rear view

Unit rear

- Extended Power Cabinet

- Extended I/O cabinet

- PCI_Box

Left side view

Front view (without cover)

Product Handling

Maintenance

WARNING

Certain tasks in this manual should only be performed by a certified service engineer. Users must not perform these tasks. Incorrect operation of these tasks may cause electric shock, injury, or fire.

- Installation and reinstallation of all components, and initial settings
- Removal of front, rear, or side covers
- Mounting/de-mounting of optional internal devices
- Plugging or unplugging of external interface cards
- Maintenance and inspections (repairing, and regular diagnosis and maintenance)

CAUTION

The following tasks regarding this product and the optional products provided from Fujitsu should only be performed by a certified service engineer. Users must not perform these tasks. Incorrect operation of these tasks may cause malfunction.

- Unpacking optional adapters and such packages delivered to the users

Remodeling/Rebuilding

CAUTION

Do not make mechanical or electrical modifications to the equipment.
Using this product after modifying or overhauling may cause unexpected injury or damage to the property, the user, or bystanders.

Contents

Preface i
Structure and Contents of This Manual i
Other Reference Manuals iii
Abbreviations iv
Text Conventions iv
Syntax of the Command Line Interface (CLI) iv
Notes Regarding Notations Used in This Manual v
Conventions for Alert Messages v
Environmental Requirements for Using This Product vi

NOTE ON SAFETY vii
Alert Labels vii

Product Handling xi
Maintenance xi
Remodeling/Rebuilding xi

CHAPTER 1 PRIMEQUEST Overview 1-1

1.1 Basics of PRIMEQUEST 1-1
1.1.1 PRIMEQUEST conceptual diagram 1-3
1.1.1.1 Hardware components 1-4
1.1.2 Product lineup 1-6
1.2 PRIMEQUEST Features 1-8
1.2.1 Hardware technologies 1-9
1.2.1.1 PRIMEQUEST-dedicated chipset 1-9
1.2.1.2 IPF 1-10
1.2.1.3 Symmetric multi-processing (SMP) technology 1-10
1.2.1.4 System Mirror Mode 1-11
1.2.1.5 Partition function 1-14
1.2.1.6 Dynamic Partitioning (DP) function (PRIMEQUEST 580A/540A only) 1-16
1.2.1.7 Flexible I/O mechanism 1-17
1.2.1.8 Hardware redundancy and hot maintenance function 1-21
1.2.1.9 Management board (MMB) 1-23
1.2.1.10 GSWB (internal LAN switching hub) 1-24
1.2.2 Software technologies 1-25
1.2.2.1 Firmware 1-25
1.2.2.2 Linux 1-26
1.2.2.3 Windows 1-26
1.2.2.4 Clustering software 1-27

1.2.2.5 Fujitsu middleware products 1-28

CHAPTER 2 General Hardware Configuration. 2-1

2.1 Cabinet 2-6
2.2 System Board (SB) 2-7
2.2.1 CPU 2-8
2.2.2 Memory 2-8
2.3 IO Unit 2-10
2.3.1 BMM 2-12
2.3.2 HDD 2-12
2.3.3 PCI slot 2-13
2.3.4 Compatible PCI cards 2-14
2.3.5 Mounting conditions of PCI cards and usable built-in I/Os 2-15
2.3.5.1 Mounting conditions of PCI cards 2-15
2.3.5.2 Usable built-in I/Os 2-17
2.3.6 PCI segment mode and PCI bus mode 2-23
2.4 Management Board (MMB) 2-24
2.5 Gigabit Through Board (GTHB) 2-27
2.6 Crossbar 2-29
2.6.1 Xbar address interconnect board (XAI board) 2-30
2.6.2 Xbar data interconnect board (XDI board) 2-31
2.7 KVM Interface Unit (KVM) 2-32
2.8 Clock and PCI_Box Control Board (CPCB) 2-33
2.9 Operator Panel (Board) (OPL) 2-35
2.10 AC Section (ACS) 2-36
2.11 Power Supply Unit (PSU) 2-37
2.12 Fan (Cooling Unit) 2-39
2.13 Dual Power Feed Option 2-40
2.14 PCI_Box 2-40
2.14.1 PCIU 2-44
2.14.1.1 PCI slot in the PCIU 2-44
2.14.1.2 PCIU card cassette 2-44
2.14.1.3 Compatible PCI cards 2-45
2.14.2 PEXU 2-45
2.14.2.1 PCI slots in the PEXU 2-46
2.14.2.2 Card cassettes in the PEXU 2-46
2.14.2.3 Compatible PCI cards 2-47
2.14.3 Requirements on PCIU and PEXU connections and combinations 2-47
2.14.3.1 Requirements on connecting PCIUs and PEXUs to one IO Unit 2-47
2.14.3.2 IOU and PEXU connection examples 2-48

2.14.4 Connection requirements when XPAR partitioning is used 2-49
2.14.5 Requirements on a combination of PCIUs and PEXUs in one PCI_Box 2-49
2.14.6 Requirements on a combination of PCIUs and PEXUs in one partition 2-50

CHAPTER 3 General Software Configuration. 3-1

3.1 Firmware 3-2
3.1.1 Processor Abstraction Layer (PAL) 3-2
3.1.2 System Abstraction Layer (SAL) 3-2
3.1.3 Extensible Firmware Interface (EFI) 3-2
3.1.4 BMC firmware 3-2
3.1.5 MMB firmware 3-3
3.1.6 GSWB firmware 3-3
3.2 OS (Linux or Windows) 3-3
3.2.1 Linux 3-4
3.2.2 Windows 3-5
3.3 Supplied Software 3-5
3.3.1 System management software (PSA) 3-5
3.3.2 Drivers 3-5
3.3.3 SystemcastWizard Lite 3-5
3.3.4 System data output tool (fjsnap) 3-6
3.4 Linked Software 3-6
3.4.1 PRIMECLUSTER 3-6
3.4.2 PRIMECLUSTER GDS 3-6
3.4.3 PRIMECLUSTER GLS 3-7
3.4.4 PRIMEQUEST System Disk Mirror for Windows (PSDM) 3-7
3.4.5 Systemwalker 3-7
3.4.6 Backup and restoration software 3-8

CHAPTER 4 Hardware System Management 4-1

4.1 Basic Mechanism 4-1
4.1.1 LAN configuration (management LAN/private LAN/business LAN) 4-3
4.1.1.1 Management LAN 4-4
4.1.1.2 Private LAN 4-9
4.1.1.3 Business LAN 4-10
4.1.2 IP addresses of PRIMEQUEST management LAN 4-11
4.2 MMB Functions 4-13
4.2.1 User privilege management 4-14
4.2.2 Hardware configuration display 4-15
4.2.3 Hardware monitor 4-16

4.2.4 Partition settings and configuration display 4-18
4.2.5 KVM/USB connection switching 4-19
4.2.6 Console redirection 4-21
4.2.7 MMB use environment setting 4-23
4.2.8 Time synchronization 4-25
4.2.9 Power control/scheduled operation 4-26
4.2.10 Firmware maintenance 4-26
4.2.11 Setting information save and restore 4-27
4.3 PSA Functions 4-29
4.3.1 Operation management GUI 4-30
4.3.2 Hardware configuration management 4-32
4.3.3 OS information display 4-32
4.3.4 Hardware error monitoring 4-33
4.3.5 Log collection, analysis, and display 4-34
4.3.6 Maintenance operation 4-35
4.3.7 REMCS linkage 4-35
4.3.8 PRIMECLUSTER linkage 4-36
4.3.9 Operation management software linkage 4-38
4.3.10 Control targets of MMB user interface 4-39
4.3.11 Information managed on partition 4-41

CHAPTER 5 Redundant Configuration 5-1

5.1 Redundancy of Components 5-2
5.2 Redundancy of Private LAN and Management LAN 5-5
5.3 System Mirror Mode 5-8
5.4 Redundancy of an HDD (Use of PRIMECLUSTER GDS) 5-11
5.4.1 Redundancy provided to system disk drives 5-12
5.4.2 Redundancy of external storage 5-15
5.5 Redundancy of the Business LAN (Use of PRIMECLUSTER GLS) 5-17
5.5.1 Duplication of a transmission line between servers (high-speed switch method) 5-18
5.5.2 Duplication between a server and a hub and between switches on the same network (NIC switch method) 5-19
5.6 Degradation Function 5-21

CHAPTER 6 Hot Swapping 6-1

6.1 Hot Swappable Components 6-2

CHAPTER 7 Partition 7-1

7.1 Overview of Partition Functions 7-1
7.2 PRIMEQUEST Partition Functions 7-3

7.2.1 Mounting requirements 7-3
7.2.1.1 Mounting requirements of SBs 7-3
7.2.1.2 Mounting requirements of IO Units 7-3
7.2.2 PCI_Box allocation 7-4
7.3 Partitioning Function of the PRIMEQUEST 580A/540A/580/540 with XPAR Installed 7-6
7.3.1 Dividing SBs and IO Units when XPAR has been installed 7-6
7.3.2 Partition granularity 7-7
7.3.2.1 Partition granularity of SB 7-7
7.3.2.2 Partition granularity of IO Unit 7-8
7.3.3 Mounting requirements (when XPAR has been installed) 7-10
7.3.3.1 Mounting requirements of SBs 7-10
7.3.3.2 Mounting requirements of IO Units 7-12
7.3.4 PCI_Box allocation 7-15
7.4 Partition Configuration Definition 7-18
7.5 Points to Consider when Determining Configurations 7-19
7.5.1 SB and IO Unit maintenance policy 7-19
7.5.2 Flexible I/O component and reserved SB 7-19
7.5.3 Notes on System Mirror Mode operation 7-21
7.5.4 Notes on the memory configuration 7-21
7.5.5 Operating environment 7-21
7.6 Function for Changing Partition Configurations 7-22
7.7 Notes to Observe when Configuring Partitions 7-22
7.8 Notes on Use of Mixed Mirror Modes 7-23

CHAPTER 8 Clustering 8-1

8.1 The basics of clustering 8-1
8.2 Cluster Functions Provided in Linux Environments 8-5
8.2.1 Cluster Configurations 8-5
8.2.1.1 Inter-cabinet cluster configuration 8-5
8.2.1.2 Intra-cabinet cluster configuration 8-7
8.2.2 PRIMECLUSTER 8-8
8.3 Cluster Functions Provided in Windows Environments 8-9
8.3.1 Basic cluster configuration 8-9
8.3.2 Virtual server 8-10
8.3.3 Scope of remedies 8-11
8.3.4 Notes on construction 8-11
8.4 Points to be Considered at Cluster System Installation 8-11

CHAPTER 9 Backup and Restoration 9-1

9.1 Importance of Backing Up Data 9-1
9.2 General Methods of Backup and Restoration 9-1

9.2.1 OS-standard utility (Linux (Red Hat): dump/restore, Windows: NTBackup) 9-2
9.2.2 Cloning software (SystemcastWizard Lite) 9-5
9.2.3 Snapshot software (PRIMECLUSTER GDS Snapshot) 9-7
9.2.4 Backup software (ETERNUS SF ACM, VERITAS NetBackup, NetVault, NetWorker) 9-9
9.3 Comparison of Backup/Restoration Methods 9-14

CHAPTER 10 Considerations for Maintenance 10-1

10.1 Estimating the Dump-use Area 10-1
10.1.1 RHEL-AS4 (IPF) 10-1
10.1.2 RHEL5 (IPF) 10-3
10.2 Estimating the Dump-use Memory 10-4
10.2.1 RHEL5 (IPF) 10-4

Appendix A Software Supplied with PRIMEQUEST Hardware A-1

Glossary GL-1

Index IN-1

Figures

Figure 1.1 PRIMEQUEST 580A/540A/580/540/480/440 1-2
Figure 1.2 Concept of PRIMEQUEST configuration 1-3
Figure 1.3 Configuration of all hardware (PRIMEQUEST 580A/580/480) . . . 1-5
Figure 1.4 PRIMEQUEST technologies 1-8
Figure 1.5 System Mirror Mode 1-12
Figure 1.6 Outline of System Mirror Mode 1-12
Figure 1.7 Improved business continuity with System Mirror Mode 1-13
Figure 1.8 Partitions 1-14
Figure 1.9 Independent resource expansion in each partition 1-15
Figure 1.10 Operation on different platforms 1-16
Figure 1.11 Connection between SBs and IO Units in a conventional model 1-17
Figure 1.12 Example of the flexible I/O mechanism and partition configuration 1-18
Figure 1.13 Reduced MTTR 1-20
Figure 1.14 Duplication of the main components 1-22
Figure 1.15 MMB 1-23
Figure 1.16 GSWB 1-24
Figure 1.17 Clustering 1-27
Figure 1.18 Fujitsu middleware products 1-28
Figure 2.1 PRIMEQUEST 580A/580/480 physical configuration (full package) 2-2
Figure 2.2 PRIMEQUEST 540A/540/440 physical configuration (full package) 2-3
Figure 2.3 Main unit 2-6
Figure 2.4 SB 2-7
Figure 2.5 Memory (DIMM) slot numbers and locations 2-9
Figure 2.6 IO Unit (Full package) 2-11
Figure 2.7 BMMs 2-12
Figure 2.8 PCI card cassette 2-13
Figure 2.9 Mounting conditions of SCSI cards and usable built-in I/Os (within the same partition) 2-18
Figure 2.10 Display example of Device Manager (SCSI cards) 2-19
Figure 2.11 Display example of Device Manager (LAN cards) 2-21
Figure 2.12 MMB 2-25
Figure 2.13 GTHB external view 2-27
Figure 2.14 XAI board 2-30
Figure 2.15 XDI board 2-31
Figure 2.16 KVM 2-32

Figure 2.17 CPCB 2-33
Figure 2.18 OPL 2-35
Figure 2.19 PSU 2-37
Figure 2.20 Fan types 2-39
Figure 2.21 PCI_Box 2-41
Figure 2.22 PCI_Box connections (only with PCIU) 2-42
Figure 2.23 PCI_Box connections (with a combination of PCIUs and PEXUs) 2-43
Figure 2.24 PCIU (with PCI card cassettes inserted) 2-44
Figure 2.25 PCI card cassette (PCI_Box) 2-45
Figure 2.26 PEXU (with PEXU card cassettes inserted) 2-46
Figure 2.27 PCI Express card cassette 2-46
Figure 2.28 IO Unit and PEXU connections 2-48
Figure 2.29 IO Unit and PEXU connections (prohibited patterns) 2-49
Figure 2.30 Interface assignment 2-51
Figure 3.1 PRIMEQUEST software configuration 3-1
Figure 4.1 Basic mechanism 4-1
Figure 4.2 Management LAN 4-4
Figure 4.3 Private LAN 4-9
Figure 4.4 Business LAN 4-10
Figure 4.5 Management LAN network configuration and IP addresses 4-11
Figure 4.6 Main MMB functions 4-13
Figure 4.7 Example of the configuration display window 4-15
Figure 4.8 Example of the system event log display window 4-16
Figure 4.9 Example of the filter condition setting window 4-17
Figure 4.10 Example in which different OSs and OSs of different versions are mounted in respective partitions 4-18
Figure 4.11 Example of display of the partition configuration window 4-19
Figure 4.12 KVM/USB connection switching 4-20
Figure 4.13 Example of the KVM/USB individual connection switching setting window 4-20
Figure 4.14 Flow of redirected information 4-21
Figure 4.15 [Console Redirection Switch] window 4-22
Figure 4.16 System account management method 4-23
Figure 4.17 Example of setting the usage evironment 4-24
Figure 4.18 Example of the access control settings screen 4-24
Figure 4.19 Time synchronization image 4-25
Figure 4.20 Example of the scheduled operation display window 4-26
Figure 4.21 Schematic of configuration information save and restore operation 4-28
Figure 4.22 PSA functions 4-30

Figure 4.23 Web-UI function 4-31
Figure 4.24 REMCS linkage 4-36
Figure 4.25 PRIMECLUSTER linkage 4-37
Figure 4.26 Operation management software linkage 4-38
Figure 5.1 General redundant configuration 5-2
Figure 5.2 Redundant configuration of management LANs and private LANs 5-5
Figure 5.3 Example of automatic switching between MMB and management LAN at the occurrence of an MMB failure 5-7
Figure 5.4 Extended Mirror Mode 5-9
Figure 5.5 Logical numbers of memory (DIMM) for which a redundant configuration has been adopted 5-10
Figure 5.6 I/O Configuration 5-11
Figure 5.7 Minimum redundancy configuration 5-12
Figure 5.8 Redundancy including SCSI controllers 5-13
Figure 5.9 HDD redundant configuration where all I/O paths are mirrored 5-14
Figure 5.10 Disk device mirroring between cabinets 5-15
Figure 5.11 ETERNUS multipath connection 5-16
Figure 5.12 Example of implementing redundancy using the high-speed switching method 5-18
Figure 5.13 High-speed switching method 5-19
Figure 5.14 Example of implementing redundancy using the NIC switch method 5-19
Figure 5.15 NIC switching method 5-20
Figure 6.1 Layout of included components 6-2
Figure 7.1 Example of partition configuration 7-2
Figure 7.2 Allocation to a PCI_Box partition (only if a PCIU is mounted) . . . 7-4
Figure 7.3 Allocation to a PCI_Box partition (both a PCIU and a PEXU are mounted) 7-5
Figure 7.4 Example of partition configuration (after logical division) 7-6
Figure 7.5 Partition granularity of an SB 7-8
Figure 7.6 SB partition granularities and allocation rule 7-11
Figure 7.7 Partition granularities and allocation rule of IO Units 7-13
Figure 7.8 Example of allocation of partition granularity to IO Units (1 of 3) 7-14
Figure 7.9 Example of allocation of partition granularity to IO Units (2 of 3) 7-14
Figure 7.10 Example of allocation of partition granularity to IO Units (3 of 3) 7-15

Figure 7.11 Allocation of a PCI_Box to partitions (only if a PCIU is mounted) 7-16
Figure 7.12 Allocation of a PCI_Box to partitions (both a PCIU and a PEXU are mounted) 7-17
Figure 7.13 [Partition Configuration] window 7-18
Figure 7.14 Flexible I/O component 7-20
Figure 7.15 Data communication among SBs, IO Units, and Other Ports 7-22
Figure 8.1 Example for 1:1 standby 8-2
Figure 8.2 Example for N:1 standby 8-3
Figure 8.3 Example for mutual standby 8-3
Figure 8.4 Example for a cascade 8-4
Figure 8.5 Inter-cabinet cluster system configuration 8-6
Figure 8.6 Intra-cabinet cluster system configuration 8-7
Figure 8.7 Basic cluster configuration (Windows) 8-9
Figure 8.8 Virtual server 8-10
Figure 8.9 Position of the interconnect 8-12
Figure 8.10 Extended Mirror Mode and cluster configuration 8-13
Figure 8.11 Example of using flexible operation in an intra-cabinet cluster configuration 8-14
Figure 9.1 Local backup using a standard utility (when the KVM console is used) 9-3
Figure 9.2 Backup to the remote host by using a standard utility (when using the KVM console) 9-4
Figure 9.3 Backup using SystemcastWizard Lite 9-5
Figure 9.4 System volume backup using PRIMECLUSTER GDS Snapshot . 9-7
Figure 9.5 Data area backup using PRIMECLUSTER GDS Snapshot 9-8
Figure 9.6 Backup of data on the ETERNUS by using ACM 9-9
Figure 9.7 Data backup using VERITAS NetBackup 9-10
Figure 9.8 Online backup of a DB using VERITAS NetBackup 9-11
Figure 9.9 Data backup using NetWorker 9-12

Tables

Table 1.1 Hardware specifications 1-6
Table 1.2 Relationship of the above partitions, SBs, and IO Units 1-19
Table 1.3 Linux OS features 1-26
Table 2.1 Types and maximum numbers of boards and components 2-4
Table 2.2 SB specifications 2-7
Table 2.3 IO Unit external interfaces 2-11
Table 2.4 PCI slot specifications (IO Unit). 2-13
Table 2.5 Compatible PCI cards 2-14
Table 2.6 PCI card mounting conditions 2-15
Table 2.7 Priority ranks of IO Units (within the same partition) 2-16
Table 2.8 Usable built-in I/Os (within the same partition) 2-17
Table 2.9 Number of interrupt vectors required for each type of PCI card . . 2-22
Table 2.10 PCI bus mapping modes 2-23
Table 2.11 MMB external interfaces 2-25
Table 2.12 GTHB functions 2-28
Table 2.13 GTHB external interface 2-28
Table 2.14 Crossbar specifications 2-29
Table 2.15 KVM external interfaces 2-32
Table 2.16 CPCB external interfaces 2-34
Table 2.17 LED display on the OPL 2-35
Table 2.18 ACS specifications 2-36
Table 2.19 PSU specifications 2-37
Table 2.20 Number of mounted SBs and IO Units, and PSU configuration (PRIMEQUEST 580A/580/480). 2-38
Table 2.21 Number of mounted SBs and IO Units, and PSU configuration (PRIMEQUEST 540A/540/440). 2-38
Table 2.22 PCI_Box configuration for PRIMEQUEST. 2-42
Table 2.23 PCI slot specifications (PCI_Box) 2-44
Table 2.24 Compatible PCI Express cards 2-47
Table 2.25 Combination of PCIUs and PEXUs in one partition (pattern 1) . . . 2-50
Table 2.26 Combination of PCIUs and PEXUs in one partition (pattern 2) . . . 2-50
Table 2.27 Combination of PCIUs and PEXUs in one partition (pattern 3) . . . 2-51
Table 2.28 Combination of PCIUs and PEXUs in one partition (pattern 4) . . . 2-51
Table 4.1 Inter-partition communication connection mode (No VLAN mode) 4-7
Table 4.2 Inter-partition communication disconnection mode (VLAN mode). 4-8
Table 4.3 All-partition communication disconnection mode (Port disable mode) 4-8

Table 4.4 Log file information 4-34
Table 4.5 Operations available from GUI on a partition. 4-40
Table 4.6 Information managed on partition 4-41
Table 5.1 System Mirror Mode Specifications 5-8
Table 6.1 Component hot swapping and/or redundant configuration availability. 6-3
Table 7.1 Effects of LSB failure 7-8
Table 7.2 SB partition granularities and allocation rule 7-11
Table 7.3 IO Unit allocation rule 7-13
Table 7.4 Data communication among SBs, IO Units, and Other Ports 7-23
Table 9.1 Comparison of backup methods 9-14
Table 10.1 Areas used to collect dump data 10-2
Table 10.2 Devices that can be used as a dump device 10-2
Table A.1 Software programs supplied with PRIMEQUEST hardware A-1

CHAPTER 1 PRIMEQUEST Overview

This chapter provides an overview of PRIMEQUEST as follows:

- Basics of PRIMEQUEST
- PRIMEQUEST Features

1.1 Basics of PRIMEQUEST

PRIMEQUEST is a server platform based on the latest IA architecture.

The PRIMEQUEST series includes a number of high-reliability and high-performance technologies that Fujitsu has long cultivated over the years. It can thus serve large-scale mission-critical systems, including those supporting social infrastructures or enterprise business operations.

PRIMEQUEST is a server platform that includes, as its central processing units (CPUs), Itanium Processor Family processors (hereafter referred to as IPF processors or simply as CPUs), which are 64-bit processors manufactured by Intel Corporation.

This manual describes the high-end server model (PRIMEQUEST 580A/580/480) that houses up to 32 CPUs and the mid-range server model (PRIMEQUEST 540A/540/440) that houses up to 16 CPUs.

The PRIMEQUEST 580A/580 is equipped with up to 2 TB of memory, internal HDDs providing a total storage capacity of up to 4.70 TB, and PCI slots that can be expanded to up to 128 slots. The unit inside one cabinet can be flexibly divided into up to 8 partitions or, when XPAR (option) has been installed, up to 16 partitions.

The PRIMEQUEST 540A/540 is equipped with up to 1 TB of memory, internal HDDs providing a total storage capacity of up to 2.35 TB, and PCI slots that can be expanded to up to 64 slots. The unit inside one cabinet can be flexibly divided into up to 4 partitions or, when XPAR (option) has been installed, up to 8 partitions.

The PRIMEQUEST 480 is equipped with up to 1 TB of memory, internal HDDs providing a total storage capacity of up to 4.70 TB, and up to 128 PCI slots. The unit inside one cabinet can be flexibly divided into up to 8 partitions.

The PRIMEQUEST 440 is equipped with up to 512 GB of memory, internal HDDs providing a total storage capacity of up to 2.35 TB, and up to 64 PCI slots. The unit inside one cabinet can be flexibly divided into up to 4 partitions.

These models support Linux and Windows, operating systems (hereafter sometimes abbreviated as OSs) that are available in 64-bit processor configurations.

Figure 1.1 shows the appearance of the PRIMEQUEST 580A/540A/580/540/480/440.

Figure 1.1 PRIMEQUEST 580A/540A/580/540/480/440

1.1.1 PRIMEQUEST conceptual diagram

The PRIMEQUEST cabinet contains hardware elements. These hardware elements are interconnected within the cabinet and work independently of other elements and, at the same time, cooperate with each other to serve as an integral system.

Figure 1.2 Concept of PRIMEQUEST configuration

A partition is a system that is equipped with the functions of a processing unit. Each partition contains software resources, such as an operating system and application programs, as well as hardware resources, such as system boards (SBs) and IO Units. Because partitions are independent of each other, they can be used as single servers that are assigned different tasks. Each partition can be connected to an external storage device via a PCI card (within the cabinet or through a PCI_Box). It can also interact with the operating system via a remotely connected or serially connected console.

The Management Board (MMB) is a feature that manages the PRIMEQUEST hardware as a whole. The MMB can be connected to the MMB console, the Field Support Tool (FST), and the Internet via a LAN. The MMB console is a console used to configure or display the hardware resource settings. The FST is a console used by maintenance personnel to maintain the PRIMEQUEST hardware. It is connected to the Remote Customer Support System (REMCS) via the Internet or a P-to-P connection to remotely monitor the unit status.

The Gigabit Switch Board (GSWB) is a Gigabit-Ethernet-compatible switching hub. The Gigabit Through Board (GTHB) without a switching function is also available. The GSWB and GTHB are both optional products used to configure a business LAN.

The clock and PCI_Box control board (CPCB) is a component with two functions: one is to generate system clock signals and send them to PRIMEQUEST hardware, and the other is to control a PCI_Box and uninterruptible power supply (UPS).

The KVM interface unit (KVM) is a feature used to connect devices including the keyboard, display, and mouse.

1.1.1.1 Hardware components

The main PRIMEQUEST components include the SB, IO Unit, crossbar, MMB, and GSWB (GTHB). Figure 1.3 shows the connection configuration of these main components. In Figure 1.3, a rectangle represents one SB, IO Unit, MMB, or GSWB (GTHB) board.

System boards (SBs) are boards on which CPUs, memory modules, PRIMEQUEST-dedicated chipsets, and other components are mounted. The PRIMEQUEST 580A/580/480 can contain up to eight SBs and the PRIMEQUEST 540A/540/440 can contain up to four SBs within the cabinet.

One SB can accommodate one to four CPUs. Therefore, the PRIMEQUEST 580A/540A/580/480 can contain up to 32 CPUs and the PRIMEQUEST 540A/540/440 can contain up to 16 CPUs in total.

One PRIMEQUEST 580A/540A/580/540 SB can carry up to 256 GB of memory and one PRIMEQUEST 480/440 SB can carry up to 128 GB of memory. This means that PRIMEQUEST 580A/580, PRIMEQUEST 540A/540, PRIMEQUEST 480, and PRIMEQUEST 440 servers can have up to 2 TB, 1 TB, 1 TB, and 512 GB of memory, respectively.

IO Units are boards on which hard disk drives (HDDs), PCI card slots, and other components are mounted. The PRIMEQUEST 580A/580/480 can contain up to eight IO Units and the PRIMEQUEST 540A/540/440 can contain up to four IO Units within the cabinet. One IO Unit can contain four HDDs. This means that the PRIMEQUEST 580A/580/480 can contain up to 32 HDDs (4.7 TB) and the PRIMEQUEST 540A/540/440 can contain up to 16 HDDs (2.35 TB). To accommodate PCI cards, the PRIMEQUEST 580A/580/480 has 32 PCI card slots and the PRIMEQUEST 540A/540/440 has 16 PCI card slots. By using PCI_BOXes (optional), you can add additional PCI card slots.

The system interconnect is a bus with the crossbar function, which sends and receives addresses and data between SBs or between an SB and IO Unit. Two types of boards compose the system interconnect: the Xbar address interconnect (XAI) for the address bus, and the Xbar data interconnect (XDI) for the data bus. The XAI and XDI are duplicated to increase the reliability of each bus.

The MMB is connected to each component in the cabinet and manages hardware as a whole. The MMB is also duplicated to increase its reliability.

The GSWB and GTHB, which are optional products, are boards used to construct a business LAN. The GSWB has a switching function, and carries 8 1000Base-T ports. The GTHB does not have a switching function and outputs signals directly to external connectors. It carries 16 1000Base-T ports.
The GSWB and GTHB also can be duplicated to increase their reliability.

Figure 1.3 Configuration of all hardware (PRIMEQUEST 580A/580/480)

1.1.2 Product lineup

The PRIMEQUEST lineup consists of two models: a high-end server that accommodates up to 32 CPUs, and a mid-range server that accommodates up to 16 CPUs.

Table 1.1 Hardware specifications

Item		Specification	
		PRIMEQUEST 580A/580/480	PRIMEQUEST 540A/540/440
CPU	Type	Itanium	Itanium
	Freq./L3cache	• PRIMEQUEST 580A 1.66 GHz/24 MB, 1.66 GHz/8 MB • PRIMEQUEST 580 1.60 GHz/24 MB, 1.42 GHz/12 MB • PRIMEQUEST 480 1.60 GHz/9 MB, 1.50 GHz/4 MB	• PRIMEQUEST 540A 1.66 GHz/24 MB, 1.66 GHz/12 MB • PRIMEQUEST 540 1.60 GHz/24 MB, 1.42 GHz/12 MB • PRIMEQUEST 440 1.60 GHz/9 MB, 1.50 GHz/4 MB
	Number of CPUs	Up to 32 CPUs	Up to 16 CPUs
System board (SB)		Up to 8 boards	Up to 4 boards
IO Unit		Up to 8 boards	Up to 4 boards
Chipset		PRIMEQUEST-dedicated chipset	PRIMEQUEST-dedicated chipset
Memory (maximum capacity)(*1) (*2)		PRIMEQUEST 580A/580: 2 TB PRIMEQUEST 480: 1 TB	PRIMEQUEST 540A/540: 1 TB PRIMEQUEST 440: 512 GB
Interconnect		Point-to-point crossbar	Point-to-point crossbar
System Mirror Mode (option)		Standard Mirror Mode Extended Mirror Mode	Standard Mirror Mode Extended Mirror Mode
Flexible I/O		Supported	Supported
Partition		PRIMEQUEST 580A/580: Up to 8 partitions * Up to 16 partitions if XPAR (option) is installed	PRIMEQUEST 540A/5540: Up to 4 partitions * Up to 8 partitions if XPAR (option) is installed
		PRIMEQUEST 480: Up to 8 partitions	PRIMEQUEST 440: Up to 4 partitions
Built-in I/O	HDD	Up to 32 drives (4.7 TB)	Up to 16 drives (2.35 TB)
	SCSI port	Up to 16 ports	Up to 8 ports
	External port (*3)	Video, keyboard (USB), mouse (USB)	Video, keyboard (USB), mouse (USB)
	DVD-ROM drive (*3)	1	1

Item	Specification	
	PRIMEQUEST 580A/580/480	PRIMEQUEST 540A/540/440
Number of PCI slots	Up to 128 slots • Built-in: 32 slots • PCI_Box: 96 slots (12 × 8 boxes)	Up to 64 slots • Built-in: 16 slots • PCI_Box: 48 slots (12 × 4 boxes)
PCI_Box (option)	Up to 8 boxes 12 PCI slots per PCI_Box	Up to 4 boxes 12 PCI slots per PCI_Box
Server management	MMB	MMB
Redundant configuration	SB, GSWB, GTHB (*4), MMB, XAI, XDI, PSU, HDD (Linux only), memory, fan, PCI card (*5), IO_PSU	SB, GSWB, GTHB (*4), MMB, XAI, XDI, PSU, HDD (Linux only), memory, fan, PCI card (*5), IO_PSU
Hot swapping	GSWB, GTHB (PRIMEQUEST 580A/540A/580/540 only), MMB, PSU, fan, HDD (Linux only), KVM, OPL, PCI card	GSWB, GTHB (PRIMEQUEST 580A/540A/580/540 only), MMB, PSU, fan, HDD (Linux only), KVM, OPL, PCI card
External dimensions (WxDxH) (mm)	738 × 1,100 × 1,800	738 × 1,100 × 1,800
Weight	PRIMEQUEST 580A/580: 730 kg PRIMEQUEST 480: 720 kg	PRIMEQUEST 540A/540: 605 kg PRIMEQUEST 440: 600 kg
Power consumption	PRIMEQUEST 580A/580: 12.5 KVA (*6) PRIMEQUEST 480: 10.7 KVA (*6)	PRIMEQUEST 540A/540: 6.79 KVA PRIMEQUEST 440: 6.4 KVA

MMB: Management board
GSWB: Gigabit Switch Board
GTHB: Gigabit Through Board
PSU: Power supply unit

*1 This is the maximum memory capacity when the PRIMEQUEST 580A/540A/580/540 contains the maximum number of 8-GB memory modules or the PRIMEQUEST 480/440 contains the maximum number of 4-GB memory modules. The effective memory size in Extended Mirror Mode is one half of that capacity.

*2 The 8GB DIMM (32-GB memory) is DDR2-400 memory. It works at 266 MHz.

*3 The component is commonly used by each partition.

*4 PRIMEQUEST 580A/540A/580/540 only.

*5 A redundant configuration using mirroring software such as a multipath driver can be used.

*6 This indicates a value for the main unit containing at least five SBs or IO Units.

1.2 PRIMEQUEST Features

PRIMEQUEST has the following three features:

- High performance
  The high-speed multiprocessor and high-speed crossbar provide high throughput.
- High reliability and high availability
  PRIMEQUEST-series machines detect errors and failures in their own systems, and they correct minor errors and isolate faulty components to provide continuous system operation. Faulty components can be easily located and replaced, enabling quick system recovery.
- High operability
  All server management operations such as those for partition-related settings and status monitoring can be centrally managed from the MMB Web-UI.

Figure 1.4 shows the technologies underlying the three features. The following sections describe the overview of these technologies.

Figure 1.4 PRIMEQUEST technologies

1.2.1 Hardware technologies

This section describes the following essential hardware technologies shown in Figure 1.4.

- Technologies underlying high performance
  - PRIMEQUEST-dedicated chipset
  - IPF
  - Symmetric multi-processing (SMP) technology
- Technologies underlying high reliability and high availability
  - System Mirror Mode
  - Partition function
  - Flexible I/O mechanism
  - Hardware redundancy and hot swapping function
- Technologies underlying high operability
  - MMB (management board)
  - GSWB (built-in switching board)
  - GTHB (built-in hub)

1.2.1.1 PRIMEQUEST-dedicated chipset

The chipset is an LSI device that controls data transfer among CPUs, memory, and IO Units.

The PRIMEQUEST-series machines contain this new chipset that has been developed by Fujitsu for improved performance and increased reliability.

The following three achievements are made by the PRIMEQUEST-dedicated chipset:

- Large-scale SMP

  The high-speed memory access technology in the chipset suppresses deteriorations in performance in a large-scale SMP configuration, resulting in a high-performance system with high scalability.

- System Mirror Mode

  The chipset supports dual reading/writing of memory and synchronizes duplication of crossbar. System Mirror Mode, one of the new technologies of the PRIMEQUEST-series machines, is thus implemented.

- High reliability of the internal chipset

  By using highly reliable technologies (e.g., duplicated internal chipset) originally developed for mainframes, the chipset has greatly increased reliability.

1.2.1.2 IPF

The PRIMEQUEST-series machines contain IPF processors, which are Intel 64-bit processors.

The IPF features reliability and performance levels, as described below, suitable for large-scale mission-critical systems:

- High reliability

  The reliability, availability and serviceability (RAS) function is enhanced by the error detection and correction function incorporated for cache memory and other components. This improves business continuity.

- High performance

  The IPF improves performance because its 64-bit processors use an expanded memory address space and have higher computing speeds.

- Openness

  Since the IPF processors are open-standard IA processors, a wide range of IPF independent software vendor (ISV) and independent hardware vendor (IHV) products can be used, thereby expanding the selection of products for system construction.

1.2.1.3 Symmetric multi-processing (SMP) technology

Each processor is the same as any other processor in a multiprocessor system using SMP technology, so processing can be assigned to any processor. High scalability can thus be obtained with applications unchanged.

In an SMP system configuration with multiple SBs linked, the PRIMEQUEST 580A/580/480 supports up to 32 CPUs, and the PRIMEQUEST 540A/540/440 supports up to 16 CPUs.

By using large-scale SMP technology originally developed for mainframes and UNIX servers, high performance suitable for a high-reliability large-scale mission-critical system is implemented.

1.2.1.4 System Mirror Mode

System Mirror Mode duplicates the memory, crossbar and the inside of the chip set. Thus, even if one of the duplicated components fails, the other component, which is its mirror, can continue the system processing.

The use of System Mirror Mode can thus increase reliability and improve continuity of business.

PRIMEQUEST has two types of Mirror Mode, Standard Mirror Mode and Extended Mirror Mode. These modes can be switched using MMB Web-UI. A cabinet having more than one partition allows the combined use of Standard Mirror Mode and Extended Mirror Mode ; the appropriate mode can be used selectively according to the operation type.

Remarks: Partitions running in Extended Mirror Mode cannot contain any divided SB or IO Unit. For more information, see CHAPTER 7, "Partition."

The features of each mode are described below.

- Standard Mirror Mode

  Reliability can be improved by duplicating the crossbar of the address path. The memory is not duplicated and up to the maximum amount of memory can be used.

- Extended Mirror Mode

  Substantial improvements in reliability can be achieved by duplicating the crossbar and the data part of the address path, the inside of the chip set, and the memory. For example, if a memory failure occurs, Extended Mirror Mode stops only the system that has the failure, and is able to continue the operation using the other system which is running normally. Therefore, the operation can be continued without stopping the entire system.

  For details of System Mirror Mode, see Section 5.3, "System Mirror Mode."

Figure 1.5 System Mirror Mode

Figure 1.6 Outline of System Mirror Mode

Improved business continuity with System Mirror Mode

The PRIMEQUEST System Mirror Mode greatly improves business continuity.

Figure 1.7 shows a comparison of operation with and without System Mirror Mode. In operation without System Mirror Mode, if a hardware failure occurs while the system is operating, the system (task) stops. Then, the cause of the failure is investigated, which is followed by repair work and other processing. Finally, the system is recovered and restarted. The task remains stopped throughout this period.

In Extended Mirror Mode, one of the System Mirror Modes, if a hardware failure such as a memory failure occurs while the system is running, the failure location (or system) is isolated and the current task is, without interruption, taken over by the other system and run on that system only. The disconnection stops neither the system nor the task. After the disconnection, the task is performed only by one system, while a search for the cause of the failure is made so that the faulty component can be replaced quickly and operation can return to system mirrored operation quickly. Thus, Extended Mirror Mode can improve business continuity in the event of a hardware failure.

Figure 1.7 Improved business continuity with System Mirror Mode

Remarks: Also in Standard Mirror Mode, if an address bus system fails, the other system takes over the current task without stopping it.

1.2.1.5 Partition function

The partition function divides hardware resources in one cabinet into multiple systems, and each of the systems operates independently. The maximum number of partitions depends on the model.

- PRIMEQUEST 580A/580: Up to 8 partitions or, when XPAR (option) has been installed, up to 16 partitions
- PRIMEQUEST 540A/540: Up to 4 partitions or, when XPAR (option) has been installed, up to 8 partitions
- PRIMEQUEST 480: up to 8 partitions
- PRIMEQUEST 440: up to 4 partitions

Remarks: In Extended Mirror Mode, a PRIMEQUEST 580A/580 system can contain up to 8 partitions and a PRIMEQUEST 540A/540 system can contain up to 4 partitions.

Figure 1.8 Partitions

The partition function provides the following advantages:

- XPAR (eXtended PARtitioning)

  XPAR (eXtended PARtitioning) enables each SB and IO Unit as well as each divided SB and IO Unit to be configured in individual partitions. Divided and undivided SBs and IO Units can belong to the same partition.

  Divided SBs and IO Units cannot be incorporated in the partition for which Extended Mirror Mode is set. XPAR is an option of the PRIMEQUEST 580A/540A/500 series. An XPAR option is required for implementing XPAR.

- Flexible system construction

The partition function can be used to construct multiple business systems in the same cabinet, thereby enabling flexible system construction.

- Robustness

  Hardware protects each partition from being affected by another partition. If a failure occurs in a partition, the other partitions are not affected by it.

Independent resource expansion in each partition

The PRIMEQUEST flexible I/O mechanism and partition function can be used to allocate resources effectively. For example, a system with the minimum configuration can be constructed first, and required hardware resources can then be added to the system at any time to improve its capabilities.

For details on the flexible I/O mechanism, see Section 1.2.1.7, "Flexible I/O mechanism."

Figure 1.9 Independent resource expansion in each partition

Operation on different platforms

The partition function enables a mixture of systems in the same cabinet. Each partition in the cabinet can run a different OS or a different OS version.

This makes possible flexible development and capability improvements, so hardware resources can be used effectively.

Figure 1.10 Operation on different platforms

1.2.1.6 Dynamic Partitioning (DP) function (PRIMEQUEST 580A/540A only)

Dynamic Partitioning (hereinafter, DP) is a function for adding or replacing SBs in a partition for which the OS is busy. For details, see the *PRIMEQUEST Dynamic Partitioning (DP) Manual*.

1.2.1.7 Flexible I/O mechanism

The flexible I/O mechanism provides greater flexibility in connecting SBs and IO Units.

In operation without the flexible I/O mechanism, some resources remain unused because CPUs, memory, and I/O resources have fixed connections, as shown in Figure 1.11.

Figure 1.11 Connection between SBs and IO Units in a conventional model

In the above figure, Partition #0 operation requires more I/O resources than CPU resources, and the CPU resources are underused. This state is typical in large-scale database processing and other processing that handles a large volume of data.

In the above figure, Partition #1 operation requires more CPU resources than I/O resources, and the I/O resources are underused. This state is typical in processing that requires a high CPU processing capacity, such as scientific and technological computational processing.

As explained above, resources such as CPUs and memory cannot be freely allocated when SBs and IO Units have fixed connections, and this sometimes results in wasted resources.

The system configuration can be flexibly set and changed by using this mechanism and partitioning function.

The flexible I/O mechanism provides the following advantages:

- Free resource allocation

  If CPUs, memory, and I/O resources have fixed connections, resources may sometimes be used ineffectively. The flexible I/O mechanism enables CPUs, memory, and I/O resources to be allocated as required according to the operation mode to reduce costs.

- Reduced mean time to repair (MTTR)

  If an SB, CPU, or memory fails, use of the flexible I/O mechanism and a reserved SB enables system recovery simply with a reboot, thereby significantly shortening the MTTR.

  The MTTR is the average time necessary for recovery of a faulty computer. It is the period in which the faulty computer does not operate before it is repaired. The shorter the MTTR, the easier the maintenance and repair.

Free resource allocation with the flexible I/O mechanism

The flexible I/O mechanism enables flexible use of hardware resources. Specifically, SBs and IO Units are freely connected by the flexible I/O mechanism in the crossbar, and the flexible I/O mechanism provides diversity and flexibility in such connections. Figure 1.12 shows an example of the flexible I/O mechanism and partition configuration.

Note: The lines in the crossbar represents connections between SBs and IO Units.

Figure 1.12 Example of the flexible I/O mechanism and partition configuration

Table 1.2 Relationship of the above partitions, SBs, and IO Units

Partition	SB	IO Unit
Partition #0	SB #0	IO Unit #0
Partition #1	SBs #1, #2, #3, and #4	IO Units #1, #2, and #3
Partition #2	SBs #5, #6, and #7	IO Units #4, #5, #6, and #7

In Partition #0, SB #0 and IO Unit #0 are connected in a one-to-one basis. In Partition #1, SBs #1, #2, #3, and #4 are connected to IO Units #1, #2, and #3 in a four-to-three basis by the flexible I/O mechanism. In Partition #2, SBs #5, #6, and #7 are connected to IO Units #4, #5, #6, and #7 in a three-to-four basis. In Partition #0 to Partition #2, the flexible I/O mechanism can establish connections in which the number of SBs does not match the number of IO Units.

As shown in this example, CPUs, memory, and I/O resources can be allocated independently because of the flexible I/O mechanism. Therefore, system resources can be used more effectively. That is, resources can be allocated as required according to the operation mode.

If an SB, CPU, or memory fails in a partition that has a reserved SB prepared, the current task can be continued with the reserved SB, which replaces the faulty SB, CPU, or memory after the partition is restarted. The reserved SB can be incorporated simply with a reboot of the partition. Therefore, the task can be restarted without waiting for the faulty SB to be repaired.

Reduced MTTR

Figure 1.13 shows an example where the MTTR is shortened by a reserved SB. Operation is switched to reserved SB #4 when a hardware failure occurs on SB #3. SB #3 is disconnected from Partition #1. The active IO Unit #2 can be connected to the reserved SB #4 as is by the flexible I/O mechanism. This shortens the MTTR to the time necessary for rebooting only, resulting in a great improvement in system availability.

Figure 1.13 Reduced MTTR

Remarks:

- Fujitsu recommends specifying a reserved SB that has the same memory and CPU configurations as those of the SBs configured in the target partition. If the number of CPUs used increases during switching to the specified reserved SB, additional licenses are required.
- CPUs of the same type as those on the SBs composing the partition must be mounted on the reserved SB. If the SB has a different type of CPU mounted on it, it is not used as a reserved SB.

1.2.1.8 Hardware redundancy and hot maintenance function

The main components in the PRIMEQUEST-series machines are made redundant to improve business continuity.

Also, PRIMEQUEST supports the hot maintenance function under certain conditions, so that maintenance components can be disconnected and installed during system operation. However, replacing components is a task handled by Fujitsu maintenance personnel.

Hardware redundant configuration

The redundant configurations of the main hardware components are described below:

- Internal chipset — The important control part of the PRIMEQUEST-dedicated internal chipset is duplicated.
- Board — For most board types, two or more boards can be mounted for a redundant configuration. Therefore, if one board fails, operation can continue with another board.
- Power supply unit — Redundant power supply units are provided, so if one power supply unit fails, operation can continue with another power supply unit. A dual power feed has been prepared as an option.
- Cooling fan unit — Redundant cooling fan units are also provided, so if one fan unit becomes faulty, operation can continue with another fan unit cooling the system.
- Other main components — The MMB, different types of LANs, the crossbar, and other main components are duplicated (standard), and if one system fails, operation can continue with the other system.

Remarks: The optional GSWB can also be duplicated, depending on the operating method.

Figure 1.14 Duplication of the main components

Hot maintenance function

The hot maintenance function disconnects a unit from the system for maintenance work on the unit, without stopping the partition operation.

Hot maintenance is possible because the main components and important units in a PRIMEQUEST-series machine are in redundant configurations, as described below:

- Since the power supply units (PSUs) and fans are in redundant configurations as standard, hot maintenance of PSUs and fans is possible.
- Since the MMB, which provides the server management function, is in a redundant configuration, hot maintenance of the MMB is possible.
- HDD, PCI card, GSWB, and other devices is possible under specific conditions.
- Hot maintenance of a non-duplicated KVM unit and OPL is also possible.

Remarks: Maintenance on an SB or IO Unit when the partition is stopped is possible.

1.2.1.9 Management board (MMB)

No dedicated management server need be installed because the PRIMEQUEST cabinet contains the MMB, which has the server management function.

The MMB functions include hardware status monitoring, displaying configuration and error information, partition management, network environment management, and power control. Because the PRIMEQUEST-series machines are duplicated by the two MMBs, if one of the MMBs fails, the duplicate MMB can continue processing. Hot maintenance of the MMB is possible.

Figure 1.15 MMB

The built-in MMB provides the following advantages:

- Reduced management costs

  Management costs can be greatly reduced because the server management function does not require a dedicated management server or dedicated software.

- Centralized management

  The MMB can centrally monitor, control, and set up hardware resources of the server.

1.2.1.10 GSWB (internal LAN switching hub)

Since PRIMEQUEST cabinet has a built-in switching hub (GSWB) compatible with Gigabit Ethernet, effective network infrastructures can be constructed.

The GSWB can be operated from the MMB Web-UI for centralized and simple operation.

Figure 1.16 GSWB

The built-in LAN switching hub in the cabinet provides the following advantages:

- Cabling on board

  Since no cable is required for connecting an external LAN switch hub (wireless), the time necessary for cable connection and management can be reduced. This also can prevent mistaken cable connections.

- Centralized management

  Centralized management of GSWB status monitoring (for failure occurrences) and GSWB control can be implemented from the MMB Web-UI.

- Virtual LAN (VLAN)

  A single GSWB port can be shared by multiple groups, and each group can function as an independent logical network.

  Remarks: If VLAN is used, the spanning tree protocol (STP) function must be set to off.

For details of GSWB functions, see the *PRIMEQUEST GSWB User's Manual* (C122-E028EN).

1.2.2 Software technologies

The following software is adopted by the PRIMEQUEST-series machines to increase reliability, availability, and operability:

- Firmware
- Linux
- Windows
- Clustering software
- Fujitsu middleware products

1.2.2.1 Firmware

The firmware listed below is the main hardware management and operation firmware installed on PRIMEQUEST-series machines.

For details, see CHAPTER 3, "General Software Configuration."

- MMB firmware
  Monitoring of all hardware and centralized operation
- PSA (PRIMEQUEST Server Agent: System management software)
  Monitoring of hardware in individual partitions
- GSWB firmware
  Centralized management of LAN environments

1.2.2.2 Linux

PRIMEQUEST supports Red Hat Enterprise Linux, which are platforms adopted as global standards. It also supports SUSE Linux Enterprise Server, mainly for overseas markets.

Table 1.3 Linux OS features

Growth potential	Linux's strong growth potential is underscored by the following characteristics: • Not dependent on the principles and business of any particular enterprise (OS vendor) • Continuous technological improvement and development of new technologies by many developers and researchers
High reliability	Objective of continuous increases in reliability, such as with the following high-reliability functions: • Quick fixes to security holes because engineers around the world tackle defects • Logging of detailed hardware error information • Enhanced maintenance functions (trace function and dump function) • High-speed and high-reliability cluster switching synchronized with hardware and cluster software
Long-time support	Support can be provided for a long time because Linux is open-source software and its source is open to the public.

1.2.2.3 Windows

PRIMEQUEST supports Windows Server 2003, which supports 64-bit processors, and will support the next-generation of Windows when it is released.

1.2.2.4 Clustering software

The Linux OS supports clustering using PRIMECLUSTER.

Windows Server 2003 supports clustering using the OS standard function (Microsoft Cluster Service).

The availability of the whole system can be increased in a cluster configuration. If a failure occurs in the active partition or the cabinet, the current task can be handed to a standby partition or cabinet, thereby providing high availability. This function also enables the automatic detection of failures in important system resources ranging from hardware to middleware and applications, implements autonomous control such as failover, and thereby enables operation to continue safely. Even the work done in periodic maintenance and scheduled stoppages such as a system configuration change can be performed without affecting the resources in use. As a result, the service availability time is greater.

Figure 1.17 Clustering

1.2.2.5 Fujitsu middleware products

A high-reliability and high-availability system can be constructed by linking the PRIMEQUEST-series machines with the following middleware products that use Fujitsu's high-reliability and high-performance technologies and accumulated know-how:

- Interstage .....Business integration software
- Symfoware .....Database software
- Systemwalker .....Integrated operation management software

Figure 1.18 Fujitsu middleware products

*1 This is supported only for Linux.

CHAPTER 2 General Hardware Configuration

This chapter describes the following hardware devices that compose a PRIMEQUEST-series machine:

- Main unit
  - Cabinet
  - System Board (SB)
  - IO Unit
  - Management Board (MMB)
  - Gigabit Through Board (GTHB)
  - Crossbar
  - KVM Interface Unit (KVM)
  - Clock and PCI_Box Control Board (CPCB)
  - Operator Panel (Board) (OPL)
  - AC Section (ACS)
  - Power Supply Unit (PSU)
  - Fan (Cooling Unit)
  - Dual Power Feed Option
  - PCI_Box

The configurations of PRIMEQUEST components are shown below.

Physical configuration (PRIMEQUEST 580A/580/480)

Front

Rear

PDB:	Power Distribution Board
PSU:	Power Supply Unit
OPL:	OP-Panel
BP:	Back Plane
SB:	System Board
IOU:	IO Unit
BMM:	BMC Module
PCI:	PCI slot
HDD:	Hard Disk Drive

ACS:	AC Section
XDI:	Xbar Data Interconnect board
XAI:	Xbar Address Interconnect board
FAN:	Fan
KVM:	KVM interface unit
CPCB:	Clock and PCI_Box Control board
MMB:	Management Board
GSWB:	Gigabit Switch Board (option)
GTHB:	Gigabit Through Board (option)

Figure 2.1 PRIMEQUEST 580A/580/480 physical configuration (full package)

Remarks:

- BMM#1 is supported in the PRIMEQUEST 580A/540A/500 series only.
- GTHB is supported in the PRIMEQUEST 580A/580/540 only.

Physical configuration (PRIMEQUEST 540A/540/440)

Front

Rear

PDB:	Power Distribution Board
PSU:	Power Supply Unit
OPL:	OP-Panel
BP:	Back Plane
SB:	System Board
IOU:	IO Unit
BMM:	BMC Module
PCI:	PCI slot
HDD:	Hard Disk Drive

ACS:	AC Section
XDI:	Xbar Data Interconnect board
XAI:	Xbar Address Interconnect board
FAN:	Fan
KVM:	KVM interface unit
CPCB:	Clock and PCI_Box Control board
MMB:	Management Board
GSWB:	Gigabit Switch Board (option)
GTHB:	Gigabit Through Board (option)

Figure 2.2 PRIMEQUEST 540A/540/440 physical configuration (full package)

Remarks:

- BMM#1 is supported in the PRIMEQUEST 580A/540A/500 series only.
- GTHB is supported in the PRIMEQUEST 580A/540A/580/540 only.
- The Dual Power Feed Option supports PSU#A06 to PSU#A11, ACS#1, PDB#2, and PDB#3.

Table 2.1 Types and maximum numbers of boards and components

Board/component		Maximum number		Remarks
		PRIMEQUEST 580A/580/480	PRIMEQUEST 540A/540/440	
SB		8	4	The minimum is one.
	CPU	32 (4 × 8 SB)	16 (4 × 4 SB)	Each LSB requires at least one CPU.
	Maximum memory capacity (DIMM) (*1)(*2)	PRIMEQUEST 580A/580: 2 TB (32 × 8 SB) PRIMEQUEST 480: 1 TB (32 × 8 SB)	PRIMEQUEST 540A/540: 1 TB (32 × 8 SB) PRIMEQUEST 440: 512 GB (32 × 8 SB)	Each LSB requires at least four memory modules.
IO Unit		8	4	The minimum is one.
	BMM	PRIMEQUEST 580A/580: 8 * Up to 16 partitions if XPAR (option) is installed	PRIMEQUEST 540A/540: 4 * Up to 8 partitions if XPAR (option) is installed	One BMM is required per partition.
		PRIMEQUEST 480: 8	PRIMEQUEST 440: 4	
	PCI	32 (4 × 8IOU)	16 (4 × 4IOU)	-
	HDD	32	16	-
BP		1	1	-
XDI		4	4	-
XAI		2	2	-
MMB		2	2	-
KVM		1	1	-
	DVD drive	1	1	-
CPCB		1	1	-
GTHB		2	2	Optional product Supported only in the PRIMEQUEST 580A/540A/580/540
GSWB		2	2	Optional product
PDB		4	2 (4)	The number in parentheses is for the Dual Power Feed Option.

Board/component		Maximum number		Remarks
		PRIMEQUEST 580A/580/480	PRIMEQUEST 540A/540/440	
ACS		2	1 (2)	The number in parentheses is for the Dual Power Feed Option.
PSU		12	12	The minimum is six.
FAN	FAN_A	6	4	-
	FAN_B	2	2	-
OPL		1	1	-
PCI_Box		8	4	Optional product
	PCIU	4/PCI_Box	4/PCI_Box	per PCI_Box
	PEXU	4/PCI_Box	4/PCI_Box	per PCI_Box

*1 The maximum capacity of a PRIMEQUEST 580A/540A/580/540 server is reached when the maximum number of 8-GB DIMMS is mounted and the maximum capacity of PRIMEQUEST 480/440 is reached when the maximum number of 4-GB DIMMS is mounted

*2 The 8GB DIMM (32-GB memory) is DDR2-400 memory. It works at 266 MHz.

2.1 Cabinet

The main unit of the cabinet contains system boards (SB), IO Units, and other components. The cabinet also contains the management board (MMB) (dedicated unit for server management) for controlling and monitoring components in the cabinet. The MMB can be operated through a Web interface from a general-purpose personal computer.

Two types of Gigabit LAN boards can be mounted as optional products: the gigabit switch board (GSWB), which is a Gigabit Ethernet switching hub, and the gigabit through board (GTHB), which does not have a switching function.

Figure 2.3 shows the appearance of the cabinet with its front and rear covers removed (example of the full package of the PRIMEQUEST 580A/540A/580/480).

Figure 2.3 Main unit

2.2 System Board (SB)

CPUs, memory and various chipsets are mounted on the SB:

- The PRIMEQUEST 580A/580/480 main unit contains one to eight SBs, and the PRIMEQUEST 540A/540/440 main unit contains one to four SBs.
- The SB contains one to four CPUs and one set (four modules) to eight sets (32 modules) of memory.
- At least one CPU and one set of memory must be mounted on an SB that composes a partition.
- In the PRIMEQUEST 580A/540A/580/540 series, one SB can be divided into two for use.

Remarks: For details on the partition function, see CHAPTER 7, "Partition."

Figure 2.4 SB

Table 2.2 SB specifications

Item	Specifications
Number of CPUs	1 to 4
Number of memory modules	DDR2-DIMM × 4 to 32 (added in sets of four modules)

2.2.1 CPU

This section describes the CPU:

- The CPU is an IPF. The reliability, availability, and serviceability (RAS) function is enhanced by the error detection and correction function incorporated for cache memory, etc. This improves business continuity.
- The IPF improves performance because its 64-bit processors use an expanded memory address space and have higher computing speeds.
- One to four CPUs can be mounted on one SB.

2.2.2 Memory

This section describes the memory:

- The memory module (DIMM) uses double data rate synchronous dynamic random access memory (DDR2 SDRAM).
- The memory module can be added in sets of four modules of the same type.
- Single 4-bit block error correction - double 4-bit block error detection (S4EC-D4ED) is used as the error correction function for high system reliability. S4EC-D4ED is the DIMM error check function for 16 bytes (128 bits).
- The memory modules can be added in sets of four modules that are mounted in the same slots of the same channels in individual LDXs.
- One (four modules) to eight sets (32 modules) can be mounted on one SB.
- Different types of DIMMs can be mixed. For details, contact your Fujitsu certified service engineer.

Figure 2.5 Memory (DIMM) slot numbers and locations

Notes on mounting DIMMs

Note the following points on mounting DIMMs:

- DIMMs must be added in sets of four DIMMs. One set is the minimum configuration unit.
- The four DIMMs in each set must be the same type.
- If the initial diagnosis detects a fixed one-bit error in a DIMM, isolation of the DIMM is by the unit of addition (four modules). For this reason, it is recommended to mount two sets of DIMMs or more (eight modules or more).

2.3 IO Unit

This section provides an overview of the IO Unit. An I/O board, BMM (BMC module), HDD, PCI cassette, and PCI-X slot are mounted in the IO Unit, which provides instructions to devices and inputs and outputs data:

- The PRIMEQUEST 580A/580/480 main unit contains one to eight IO Units, and the PRIMEQUEST 540A/540/440 main unit contains one to four IO Units.
- SCSI, USB (USB 2.0), and COM ports are provided as the external interfaces.
- One to four HDDs, which are hot-pluggable, can be mounted.
- Up to four PCI cards, which are hot-pluggable, can be mounted.
- Up to two BMMs, which would be IO Unit control boards, can be mounted as options. However, only one BMM is valid at a time in a partition.
- Four PCI_Box interface ports of a serial bus can be mounted.
- In the PRIMEQUEST 580A/540A/580/540 series, one SB can be divided into two for use.

Remarks: Hot swapping of PCI cards is available depending on the OS installed on the partition, as follows:

- Linux: A PCI card can be hot-swapped with the PCI card inserted in a card cassette.
- Windows Server 2008: See the *PRIMEQUEST 500A/500 Series Microsoft Windows Server 2008 User's Guide* (C122-E087EN).
- Windows Server 2003: Hot swapping of PCI cards is not supported.

Note: BMM#1 is supported in the PRIMEQUEST 580A/540A/580/540 only.
For details of the BMM, see Section 2.3.1, "BMM ."

Figure 2.6 IO Unit (Full package)

Table 2.3 IO Unit external interfaces

External interface	Count	Remarks
USB	2/BMM	Mounted on optional BMM product
RS-232C	1/BMM	
SCSI	2	Ultra320-SCSI
PCI_Box interface	4	Connected to optional PCI_Box
PCI_Box interface side band	4	Connected to optional PCI_Box

2.3.1 BMM

The BMM is a board that manages input and output of a partition. Each partition requires one BMM. Two BMMs can be mounted in an IO Unit:

- Each partition requires one BMM.
- A Video port, two USB ports, and a COM port are also mounted on the BMM.

Remarks:

- As shown in Figure 2.7, the locations of the USB ports are different between BMM A and BMM B, but the ports have the same functions.
- BMM A is mounted in the PRIMEQUEST 580/540/480/440. BMM A or B is mounted in the PRIMEQUEST 580A/540A.

Figure 2.7 BMMs

2.3.2 HDD

The HDD is connected to the SCSI-BP (SCBP). Four U320 SCSI 3.5-inch HDDs, which are hot-pluggable, can be mounted.

Remarks: Hot plugging of HDDs is supported only under Linux.

2.3.3 PCI slot

The PCI slot is a slot in which a PCI-X card is installed:

- One IO Unit contains four slots.
- A PCI card is inserted into a PCI card cassette (adapter), and the cassette is inserted into the PCI slot.
- PCI cassettes are individually hot-swappable.

Remarks: Inserting or removing a PCI card in a hot swapping operation is supported under Linux and Windows Server 2008.

Figure 2.8 PCI card cassette

Table 2.4 PCI slot specifications (IO Unit)

PCI slot No.	Bus width	Operating clock	PCI bus operating voltage	Size
PCI Slot#0	64 bit	100 MHz	3.3 V	Short
PCI Slot#1	64 bit	133 MHz	3.3 V	Short
PCI Slot#2	64 bit	100 MHz	3.3 V	Short
PCI Slot#3	64 bit	133 MHz	3.3 V	Short

Remarks:

1. The PCI slot is compatible with only the standard-type (full height) PCI card bracket.
2. Only short PCI slots are mounted in an IO Unit. If a long PCI slot is required, use an optional PCI_Box.

2.3.4 Compatible PCI cards

PRIMEQUEST supports the following PCI cards.

Remarks: For the latest information on the supported PCI cards, make an inquiry of a certified service engineer.

Table 2.5 Compatible PCI cards

Type	Function
SCSI card	Ultra320 SCSI (dual channel)
Fibre Channel card	Fibre Channel (2 Gbps × 1 port)
	Fibre Channel (4 Gbps × 1 port)
	Fibre Channel (4 Gbps × 2 ports)
Network-related PCI card	Gigabit Ethernet 1000Base-T (dual channel)
	Gigabit Ethernet 1000Base-SX (single channel)
	Gigabit Ethernet 1000Base-SX (dual channel)
	10Gigabit Ethernet 10GBase-SR (single channel) (*)

*: Supported only in the PRUMEQUEST 580A/540A/500 series.

2.3.5 Mounting conditions of PCI cards and usable built-in I/Os

2.3.5.1 Mounting conditions of PCI cards

Table 2.6 contains the mounting conditions of PCI cards for the PRIMEQUEST series.

- A gigabit LAN card can be mounted only in a 100-MHz PCI slot within an IO Unit or PCI_Box. Up to eight 10-gigabit LAN cards per partition can be mounted in arbitrary PCI slots in the PRIMEQUEST 580A/540A/500 series.
- A SCSI card can be mounted only in PCI Slot #1 or PCI Slot #3 of an IO Unit with a priority rank of 1 or 2 (see Table 2.7).
- A Fibre Channel (FC) Card can be mounted in any PCI slot.

Table 2.6 PCI card mounting conditions

	PCI slots in an IO Unit				PCI slots in a PCI unit of a PCI_Box			PCI slots in a PCI Express unit of a PCI_Box (*4)	
Slot#	0	1	2	3	1(*3)	2	0(*3)	0	1
PCI Clock[MHz]	100	133	100	133	100	100	133		
Ultra320 SCSI Card (dual channel)		Y(*1)		Y(*1)					
Fibre Channel Card (2-Gbps × 1 port)	Y(*2)	Y(*2)	Y(*2)	Y(*2)	Y	Y	Y		
[PCI-X card-FC] Fibre Channel Card (4-Gbps × 1 port)	Y(*2)	Y(*2)	Y(*2)	Y(*2)	Y	Y	Y		
[PCI Express card-FC] Fibre Channel Card (4-Gbps × 1 port)								Y	Y
[PCI-X card-FC] Fibre Channel Card (4-Gbps × 2 port)	Y(*2)	Y(*2)	Y(*2)	Y(*2)	Y	Y	Y		
[PCI Express card-FC] Fibre Channel Card (4-Gbps × 2 port)								Y	Y
1000Base-T LAN card (dual channel)	Y		Y		Y	Y			

	PCI slots in an IO Unit				PCI slots in a PCI unit of a PCI_Box			PCI slots in a PCI Express unit of a PCI_Box (*4)	
1000Base-SX LAN card (single channel)	Y		Y		Y	Y			
1000Base-SX LAN card (dual channel)	Y		Y		Y	Y			
10GBase-SR LAN card (single channel)	Y	Y	Y	Y	Y	Y	Y		

Y: Slot in which the card can be mounted

*1: A SCSI card can be mounted only in PCI Slot #1 or PCI Slot #3 of an IO Unit with a priority rank of 1 or 2 (see Table 2.7). If two or more partitions are defined and operated, these priority ranks apply to each partition.

*2: For the PRIMEQUEST 580A/540A/580/540

In a SAN boot configuration, the OS can be installed via PCI Slot#1 or PCI Slot#3 in an IO Unit. After OS installation, making multipath settings that include the slot used for OS installation enables SAN boot from any slot of the IO Unit.

For the PRIMEQUEST 480/440

An SAN-bootable FC card can be mounted in only PCI Slot#1 and PCI Slot#3 of an IO Unit with a priority rank 1 or 2 (see Table 2.7). When two or more partitions are defined for operation, these priority ranks apply to each partition (same conditions as the SCSI card slots in *1).

*3: The PCI slot numbers of the PCI units in the PCI_Box are PCI Slot#1, PCI Slot#2, and PCI Slot#0 from the left when viewed from the front.

*4: Available for the PRIMEQUEST 580A/540A/500 series

Table 2.7 Priority ranks of IO Units (within the same partition)

Priority rank	IO Unit
1	Home IO Unit
2	The IO Unit with the smallest IO Unit number among all IO Units except the Home IO Unit
3	The IO Unit with the IO Unit number next smallest after that of the IO Unit with a priority rank of 2 except the Home IO Unit
4	The IO Unit with the IO Unit number next smallest after that of the IO Unit with a priority rank of 3 except the Home IO Unit

Priority rank	IO Unit
5	The IO Unit with the IO Unit number next smallest after that of the IO Unit with a priority rank of 4 except the Home IO Unit
6	The IO Unit with the IO Unit number next smallest after that of the IO Unit with a priority rank of 5 except the Home IO Unit
7	The IO Unit with the IO Unit number next smallest after that of the IO Unit with a priority rank of 6 except the Home IO Unit
8	The IO Unit with the IO Unit number next smallest after that of the IO Unit with a priority rank of 7 except the Home IO Unit

2.3.5.2 Usable built-in I/Os

The IO Unit includes the I/Os listed below. These I/Os are referred to as built-in I/Os. Table 2.8 contains usable built-in I/Os. If two or more partitions are defined and operated, the contents in this table apply to each partition.

- USB port on the BMM (connected to the USB port connected to the USB connector provided on the BMM front panel, and to the KVM)
- Video port on the BMM (connected to the KVM video port)
- COM port on the BMM
- 100MbE port on the BMM
- Built-in HDD
- External SCSI port
- GbE port

Table 2.8 Usable built-in I/Os (within the same partition)

Built-in I/O		IO Unit with a priority rank of 1 (Home IO Unit)	IO Unit with a priority rank of 2	IO Unit with a priority rank of 3	IO Unit with a priority rank of 4	IO Unit with a priority rank of 5 or higher
On the BMM (*1)	USB port	Y	N	N	N	N
	Video port	Y	N	N	N	N
	COM port	Y	N	N	N	N
	100MbE port	Y	N	N	N	N
Built-in HDD		Y	Y	Y	Y	N
External SCSI port		Y	Y	Y	Y	N
GbE port		Y	Y	Y	Y	Y

Y: Usable built-in I/O

N: Unusable built-in I/O

Mounting conditions of SCSI cards and usable built-in I/Os

Figure 2.9 shows an example that explains the priority ranks of IO Units, mounting conditions of SCSI cards, and usable built-in IOs within the same partition. In the partition shown in Figure 2.9, eight IO Units are allocated (the SB is omitted) and IOU #1 is specified as the Home IO Unit. If a SCSI card is mounted in this partition, it needs to be mounted in IOU #1, which has a priority rank of 1, or in IOU #0, which has a priority rank of 2. If a SCSI port on an IO Unit in this partition is used, IOU #1, IOU #0, IOU #2, and IOU #3, which have priority ranks of 1, 2, 3, and 4, respectively, can be used. If a built-in HDD is mounted in an IO Unit in this partition, it needs to be mounted in IOU #1, IOU #0, IOU #2, or IOU #3, which have priority ranks of 1, 2, 3, and 4, respectively.

Figure 2.9 Mounting conditions of SCSI cards and usable built-in I/Os (within the same partition)

Display of unusable devices in a Windows system

The display of unusable devices as described below involves a partition in which Windows Server 2003 Datacenter Edition or Windows 2003 Enterprise Edition is installed.

The [ ] mark that is displayed in the [Device Manager] window in the Windows system indicates a device that cannot be used due to a mounting restriction.

This indication means that devices are physically mounted but cannot actually be used due to the mounting restriction.

Remarks: Unusable devices are not affected.

Figure 2.10 shows an example of the [Device Manager] window in which the [ ] mark is displayed for SCSI controllers.

Figure 2.10 Display example of Device Manager (SCSI cards)

Mounting conditions of LAN cards

The mounting conditions of LAN cards for a partition in which Windows Server 2003 Datacenter Edition or Windows 2003 Enterprise Edition is installed are described below.

The mounting conditions of the 1000BASE-T LAN card and 1000BASE-SX LAN card are also described below.

Note: The mounting conditions described below do not apply to the GbE port of a built-in I/O, and 10Gigabit LAN card.

If LAN cards are mounted in a partition in which Windows is installed, the total number of ports of the LAN cards (number of connectors) must be 32 or less.

If the mounting conditions described in 2.3.5.1, "Mounting conditions of PCI cards" are satisfied, a LAN card can be mounted in any slot. However, the total number of ports used must be 32 or less.

If the mounted LAN cards were to have a total of more than 32 ports, up to 32 ports could be used and the remaining ports would be unusable. The ports that would be unusable in such cases cannot be specified in advance. Therefore, make sure that the total number of ports is 32 or less.

In a single partition, up to 32 single-port GbE cards can be mounted, or up to 16 dual-port GbE cards can be mounted.

In a multiple-partition configuration, the mounting conditions must be satisfied for each partition.

The total number of Giga LAN ports in a single partition is up to 64 ports, which include 32 additional ports in built-in I/Os when the partition is configured with eight IO Units.

If the number of mounted LAN cards is greater than the limit defined by the above conditions, the display of Device Manager is as shown in Figure 2.11.

Figure 2.11 Display example of Device Manager (LAN cards)

Number of PCI cards that can be mounted under control of Linux

For partitions in which Linux is installed, the number PCI cards that can be mounted is restricted. When PCI cards are connected to the system, "interrupt vectors" used to control interrupts are required. The Linux specifications have a restriction limiting the number of interrupt vectors used in the system to a maximum of 184. If a number of PCI cards exceeding this number is mounted, the system cannot be started.

For this reason, in a partition in which Linux is installed, the number of interrupt vectors used must not exceed 184. If the number of interrupt vectors exceeds 184, reduce the number of mounted PCI cards so that the total number of interrupt vectors remains within 184.

Interrupt vectors are also used for internal IO in IO Units, and PCI units and PCI Express units in the PCI_Box. The number of interrupt vectors that can be used in PCI cards can thus be determined according to the following formula:

Number if interrupt vectors that can be used = 184 - (6 + 12 x number of IO Units + 2 x number of PCI units in PCI_Box + 2 x number of PCI Express units in PCI_Box)

Remarks: The greater the number of units mounted in the partition, the smaller the number of interrupt vectors that can be used. Assume "Number of IO Unit x 4 + number of PCI_Box units x 4" as the maximum configuration.

- Interrupt vectors required for each type of PCI card
  The required number of interrupt vectors depends on the type of PCI card to be added.
  Adjust the number of PCI cards to be mounted such that the total number of interrupt vectors used does not exceed the maximum number of vectors that is determined from the above formula.
  The table below lists the numbers of interrupt vectors required for individual types of PCI cards.

Table 2.9 Number of interrupt vectors required for each type of PCI card

PCI card type	Required number of interrupt vectors
SCSI card (dual channel)	2
Fibre Channel card (1 port)	1
Fibre Channel card (2 ports)	2
LAN card (single channel)	1
LAN card (dual channel)	2

2.3.6 PCI segment mode and PCI bus mode

PRIMEQUEST supports two PCI bus mapping modes: PCI segment mode and PCI bus mode.

- PCI segment mode
  All PCI_Box interfaces of each IO Unit can be used.
- PCI bus mode
  Of the PCI_Box interfaces of each IO Unit, only PCI Slot#0 and PCI Slot#1 can be used.

The mapping mode can be defined for each partition.

Table 2.10 lists the PCI bus mapping modes supported by each operating system.

Table 2.10 PCI bus mapping modes

OS	PCI segment mode	PCI bus mode
Red Hat Enterprise Linux	Supported (*1)	Supported (*1)
SUSE Linux Enterprise Server	Supported (*1)	Supported (*1)
Windows Server 2003, Enterprise Edition	Unsupported	Supported
Windows Server 2003, Datacenter Edition	Unsupported	Supported

*1 Fujitsu recommends the PCI segment mode when Linux is used.

2.4 Management Board (MMB)

The MMB is integrated in the main unit, and it manages the unit as a whole.

The MMB has a specific CPU and memory separate from the SBs. MMB firmware runs on the MMB.

The MMB firmware provides a Web interface. As a result, a server can be managed and operated from a connected PC and other connected devices via a Web browser. Since the MMB centrally manages the whole system, unit management does not require any specific console or software.
The MMB has the following management functions:

- Cabinet power control
- Remote operations including configuring settings, displaying the status, and various other operations from the MMB Web-UI
- Remote control of unit power and partition resets
- Monitoring sensors and other devices
- Partition configuration setup
- Setting the mirror mode
- Reset processing
- Scheduling operations
- KVM operation
- GSWB operation
- REMCS operation
- Setting the alarm e-mail

Figure 2.12 MMB

Table 2.11 MMB external interfaces

External interface		Count	Remarks
RS-232C		1	Used when the Fujitsu certified service engineer performs the setup of the equipment. It is not used during usual operation.
LAN	100Base-TX	3	• Used for console connection: user port • Used for maintenance - [LOCAL]: port of the Fujitsu certified service engineer - [REMOTE]: REMCS port

Two MMBs are mounted to set up a redundant configuration. If one MMB fails, the system can switch to the other MMB to continue processing uninterrupted. There are two types of MMB failures, as follows:

(1) Failure of one MMB: Failure of one MMB owing to an MMB hardware failure
(2) Failure of one MMB Management LAN: Failure of management LAN that includes MMB internal hub, and LAN cable/external switch

Of the above two cases, MMB switching automatically occurs for (1).

For (2), because the MMB itself is operating normally, MMB switching is not executed.

In this case, the MMB is carrying out important functions such as hardware monitoring, but is no longer able to access the MMB Web-UI. To implement recovery, the failure location must be identified, and the faulty unit must be replaced. Moreover, in the case of an MMB internal hub failure, switching from the active MMB must be done manually.

Hot maintenance is supported for the MMB.

The MMB includes a built-in hub that has the VLAN function. The VLAN function controls communication between partitions and communication between the REMCS port and a partition, thereby achieving a high security level. For details, see Section 4.1.1.1, "Management LAN"

Remarks:

When Speed/Duplex for the LAN port on the MMB is set to an item other than "AUTO," use a cross cable to connect a switching hub to the MMB LAN port.

However, when using an REMCS port of a PRIMEQUEST 580A/540A MMB, or an MMB with a part number (*) that ends in 004AB or lower, the following notes apply.

* Confirm the part number of the MMB with the Part Number item in the [MMB#x] window (Click [System] → [MMB] → [MMB#x].).

- If the Speed/Duplex setting of the REMCS port is set to other than "AUTO," use a LAN cable.
- If the Speed/Duplex setting of the device to connect to the REMCS port is set to other than "AUTO," set the Speed/Duplex setting of the REMCS port of the MMB to the same setting as that of the device.

2.5 Gigabit Through Board (GTHB)

Remarks: GTHB is supported in the PRIMEQUEST 580A/540A/580/540 only.

The GTHB without a switching function is a board used for output to external connectors by receiving the GbE interfaces of individual IO Units via the backplane (BP).

Each interface on the IO Unit side (internal) is fixed to a bandwidth of 1 Gbps, and only 1000BASE-T is supported for the external interfaces connecting LAN cables. The GTHB has 16 1 GbE ports each on the IO Unit side (internal) and external connector side.

Up to two GTHB boards can be mounted. Note, however, that the GTHB cannot be mounted together with a Gigabit Switch Board (GSWB) that has a switching function.

For more information about the GSWB, see the *GSWB User's Manual* (C122-E028EN).

Figure 2.13 GTHB external view

Table 2.12 GTHB functions

Item	Description		Remarks
Interface (external)	Fixed to 1000BASE-T	16 ports	RJ-45 connector
Interface (internal)	Fixed to 1000 Mbps (SERDES)	16 ports	Connected to a GbE port of an IO Unit
Support layer	The SERDES interface from an IO Unit is converted into 1000BASE-T via the Media Converter chip.		
Auto negotiation	Supported (external port only)		Linked through only 1000BASE-T. Not linked through 10BASE-T or 100BASE-TX. Full duplex and half duplex are automatically switched.
Auto MDI/MDI-X	Supported (external port only)		
Flow control	PAUSE frame		In full duplex mode
	Back pressure		In half duplex mode
Jumbo frame	Up to 4508 bytes (4.5 Kbytes)		
Hot Plug	Supported		

Table 2.13 GTHB external interface

Interface	Number of ports	Connector	Remarks
1000BASE-T	16	RJ-45	[Interface standard] • Compliant with IEEE802.3ab (1000BASE-T) • 1000BASE-T: Twisted paired cable (category 5 enhanced cable) • Transmission distance: 100 m • Transfer rate: 1000 Mbps (1000BASE-T) Note: The GTHB supports 1000BASE-T only.

2.6 Crossbar

The crossbar provides sharing of memory in up to eight SBs and the symmetric multi-processing (SMP) system that supports 32 CPUs. It also provides the flexible I/O mechanism, which enables up to eight SBs and up to eight IO Units to be flexibly connected.

The crossbar consists of the following two type of boards:

- Xbar address interconnect board (XAI board)
- Xbar data interconnect board (XDI board)

Table 2.14 Crossbar specifications

Item	Specifications
Transmission mode	Stored and forward
Number of ports	SB port: 8 IO Unit port: 8
Bandwidth/port	SB port: 18.96 GB/s × 2 dir (1.333 GHz) IO Unit port: 4.7 GB/s × 2 dir (667 MHz)
Transmission frequency	PRIMEQUEST 580A/540A: 1333 MHz (FSB 667 MHz) PRIMEQUEST 580/540: 1066 MHz (FSB 533 MHz) PRIMEQUEST 480/440: 800 MHz (FSB 400 MHz)
Data protection	ECC

2.6.1 Xbar address interconnect board (XAI board)

A chip for controlling the address bus system of the crossbar is mounted on the Xbar address interconnect board (XAI board). The XAI board connect eight SBs and eight IO Units, and the XAI board transmits address information between the boards.

- Two XAI boards are mounted in the main unit.

Figure 2.14 XAI board

2.6.2 Xbar data interconnect board (XDI board)

A chip for controlling the data bus system of the crossbar is mounted on the Xbar data interconnect board (XDI board). The XDI board connect eight SBs and eight IO Units, and the XDI board transmits data between the boards.

- Four XDI boards are mounted in the main unit.

Figure 2.15 XDI board

2.7 KVM Interface Unit (KVM)

The KVM has the function for switching between the keyboard, video, and mouse.

The KVM collects the VGA interface and the three USB interfaces from IO Units and selects one of them to provide external output. The interface that is output from the KVM is specified with the MMB Web-UI. For information on KVM operation with the MMB Web-UI, see Section 4.2.5, "KVM/USB connection switching"

Figure 2.16 KVM

Table 2.15 KVM external interfaces

Interface	Count	Remarks
USB	4	Uses USB 2.0
VGA	1	D-sub 15 pin
DVD-ROM drive	1	For software installation

2.8 Clock and PCI_Box Control Board (CPCB)

The CPCB generates system clock signals, it has a function that sends the signals to SBs, boards in IO Units, and other boards, and it has an interface connector for controlling a PCI_Box:

- Two sets of clock generator circuits are mounted on the CPCB. If one clock generator circuit fails, MMB firmware control automatically switches operation to the other one.
  However, the whole system must be rebooted for this switching operation.
- The CPCB contains eight interface connectors (D-SUB 9-pin small connector) for PCI_Box control.
- The CPCB contains two UPC interface connectors (D-SUB 9-pin) for un-interruptable power supply (UPS) control.
- If only one UPS is connected, connect it to UPC #0.

Figure 2.17 CPCB

Table 2.16 CPCB external interfaces

Interface	Count	Remarks
UPC interface	2	Connected to an external uninterruptible power supply (UPS). It can be used for the Dual power feed option.
PCI_Box interface	8	Connected to a PCI_Box

2.9 Operator Panel (Board) (OPL)

The OPL contained in this unit has a four-LED display function.
Each LED status indicates the unit status and whether a failure has occurred.

The OPL has the following functions:

- Displaying the power status of the main unit (Power-LED)
- Indicating whether a unit is faulty (Alarm-LED)
- Identifying the unit under operation (Location-LED)
- Displaying the MMB status, for MMB status checks from outside the main unit (MMB-Ready-LED)

Figure 2.18 OPL

Table 2.17 LED display on the OPL

LED	Color	LED status	Unit status
Power-LED	Green	Off	Power to unit is off
		On	Power to unit is on
Alarm-LED	Orange	Off	Normal status
		On	Unit failure
Location-LED	Blue	Off	Unit cannot be recognized
		On/Blinking	Unit is recognized
MMB-Ready-LED	Green	Off	Power to unit is off
		On	MMB is active
		Blinking	MMB is being initialized

2.10 AC Section (ACS)

This section explains the AC section (ACS).

The ACS has a redundant configuration only for the PRIMEQUEST 540A/540/440 using dual power feed.

- Two power distribution board (PDB) units can be connected to an ACS unit. Three power supply units (PSUs) can be connected to a PDB.
- Two PSUs can be used with one AC cable.
- The single power feed option consists of six PSUs. Up to 12 PSUs can be connected to the option.
- The AC section supports dual power feed as an option.

Table 2.18 lists ACS specifications.

Table 2.18 ACS specifications

Item			Specifications
Input	Voltage	Rated voltage	200 to 240 VAC
		Number of phases	Single phase
		Fluctuation range	+10% to -10%
	Frequency	Rated frequency	50/60 Hz
		Fluctuation range	2%, -4%

2.11 Power Supply Unit (PSU)

The PSU has n + 1 redundant configuration of the single power feed as standard:

- The PSU supports dual power feed as an option.
- For PRIMEQUEST 580A/580/480, the number of mounted PSUs depends on the number of mounted SBs and IO Units.

Figure 2.19 PSU

Table 2.19 PSU specifications

Item	Specifications	
Output	Rated voltage	48 VDC
	Rated power	1450 W
Size		69.5 mm × 123.5 mm × 344.3 mm
Weight		3.3 kg

Table 2.20 Number of mounted SBs and IO Units, and PSU configuration (PRIMEQUEST 580A/580/480)

Number of SBs and IO Units	Receiving power mode	Number of mounted PSUs	PSU configuration	Remarks
Four SBs and IO Units or fewer	Single power feed	6	5 + 1 redundancy	
Five SBs or IO Units or more		10	9 + 1 redundancy	
Four SBs and IO Units or fewer	Dual power feed	12	(5 + 1) + (5 + 1) redundancy (six per power feed)	An expansion cabinet is required.
Five SBs or IO Units or more		20	(9 + 1) + (9 + 1) redundancy (10 per power feed)	An expansion cabinet is required.

Table 2.21 Number of mounted SBs and IO Units, and PSU configuration (PRIMEQUEST 540A/540/440)

Power feed	Number of mounted PSUs	PSU configuration	Remarks
Single power feed	6	5 + 1 redundancy	The configuration on the left is used regardless of the number of SBs or IO Units.
Dual power feed	12	(5 + 1) + (5 + 1) redundancy (6 per power feed)	

2.12 Fan (Cooling Unit)

- The fan is used to cool the unit as a whole. This unit has the following two types of fans:
  - Fan A (front): FANA#0 to FANA#5
  - Fan B (rear): FANB#0 to FANB #1

Figure 2.20 Fan types

- The fans have redundant configuration as standard. Periodic replacement is not required.
- Hot maintenance of fans is possible in units of trays.

2.13 Dual Power Feed Option

The Dual power feed option can be mounted in the PRIMEQUEST-series machines. The Dual power feed option is a system that allows feeding of AC power to the main unit from two different sources.

Remarks: The PRIMEQUEST 580A/580/480 requires the expansion power supply cabinet.

If power outage occurs in either source, processing can continue in the main unit because its AC power is connected to power from the dual sources.

Remarks:

- Also, connect the AC power source of a PCI_Box with two separate sources.
- If a UPS is set up, be sure to set it up for each power source.

2.14 PCI_Box

The PCI_Box is an optional unit for PCI slot expansion, and it is mounted in an Extended I/O cabinet:

- Up to eight PCI_Boxes can be mounted in one PRIMEQUEST 580A/580/480 main unit, and up to four PCI_Boxes can be mounted in one PRIMEQUEST 540A/540/440 main unit
- Up to four PCI slot units (PCIUs) or PCI Express unit (PEXUs) can be mounted in one PCI_Box.
- The PCI_Box is connected to the CPCB by a CPCB control cable and controlled by the CPCB.
- A PCI-X card is inserted into a PCI card cassette, and the cassette is inserted into the PCI slot of a PCIU.
- A PCI Expres card is inserted into a PCI card cassette, and the cassette is inserted into the PCI slot of a PEXU.
- The PCI_Box is used while connected to a PCI_Box interface of IO Unit and PCI_Box interface side band.
- Power supply units (IO_PSU) and fans use redundant configurations, increasing their reliability.
- If dual power feed is applied to the main unit, also connect the AC power source of a PCI_Box using two separate sources.

Remarks: PEXUs and PCI Express cards are supported in the PRIMEQUEST 580A/540A/500 series.

Figure 2.21 PCI_Box

Table 2.22 PCI_Box configuration for PRIMEQUEST

Unit	Addition unit	Maximum number		Count at maximum configuration	
		580A/ 580/480	540A/ 540/440	580A/ 580/480	540A/ 540/440
PCI_Box	1	8	4	8	4
PCI slot unit (PCIU)	1	4/PCI_Box	4/PCI_Box	32	16
PCI slot (PCIU)	1	3/PCIU	3/PCIU	96	48
PCI Express Unit (PEXU) (*1)	1	4/PCI_Box	4/PCI_Box	16	8
PCI slot (PEXU) (*1)	1	2/PEXU	2/PEXU	32	16

*1 PEXUs are supported in the PRIMEQUEST 580A/540A/500 series.

The following figure is a conceptual diagram of PCI_Box connections.

Figure 2.22 PCI_Box connections (only with PCIU)

Figure 2.23 PCI_Box connections (with a combination of PCIUs and PEXUs)

Remarks: PEXUs are supported in the PRIMEQUEST 580A/540A/500 series

2.14.1 PCIU

Three PCI cards can be inserted in the PCIU. The PCIU contains three PCI card cassettes as standard.
The PCIU is connected to an IO Unit with a PCI_Box interface cable and a PCI_Box side band interface cable.

Figure 2.24 PCIU (with PCI card cassettes inserted)

2.14.1.1 PCI slot in the PCIU

The following table lists the types of PCI slots in the PCIU.

Only the standard-type (full height) PCI card bracket is supported.

Table 2.23 PCI slot specifications (PCI_Box)

PCI slot No.	Operating clock	Operating voltage	Bus width	Size
PCI slot#1	100 MHz	3.3 V	64 bit	Long-length
PCI slot#2	100 MHz	3.3 V	64 bit	Long-length
PCI slot#0	133 MHz	3.3 V	64 bit	Long-length

2.14.1.2 PCIU card cassette

To use a PCI card in the PCIU, a PCI card cassette (adapter) is required. A PCI card is mounted in a PCI card cassette, which is then mounted in the PCIU.

Remarks: This PCI cassette is different from the PCI cassette used in the IO Unit.

Figure 2.25 PCI card cassette (PCI_Box)

2.14.1.3 Compatible PCI cards

The PCI cards compatible with the PRIMEQUEST PCI_Box are the same as the PCI cards compatible with IO Units.

For details, see Section 2.3.4, "Compatible PCI cards"

2.14.2 PEXU

Up to two PCI cards can be mounted in one PEXU, in which two PCI card cassettes are inserted as standard equipment.

The PEXU is connected to an IO Unit in the PRIMEQUEST 580A/540A/580/540 through two PCI Express cables and two PEXU control cables.

Notes:

- When mounting a PEXU, use PSA version 1.10 or later.
- The Linux OS versions that can be used in a partition containing a PEXU are as follows:
  - Red hat® Enterprise Linux® AS (4.5 for Itanium)
  - Red Hat® Enterprise Linux® 5 (for Intel Itanium)
  - SUSE™ Linux Enterprise Server 10 for Itanium Processor Family Service Pack 1
  - Microsoft® Windows Server® 2003, Enterprise Edition for Itanium-based Systems Service Pack 1 or later
  - Microsoft® Windows Server® 2003, Datacenter Edition for Itanium-based Systems Service Pack 1 or later

Figure 2.26 PEXU (with PEXU card cassettes inserted)

2.14.2.1 PCI slots in the PEXU

PCI Express cards are inserted in PCI slots in the PEXU.

- The PEXU has two PCI Express slots.
- A PCI Express card is mounted in a PCI card cassette (adapter), which is then inserted in a PCI Express slot.
- The PEXU can be used in the PRIMEQUEST 580A/540A/500 series.

2.14.2.2 Card cassettes in the PEXU

To use a PCI card in the PEXU, a PCI Express card cassette (adapter) is required.

A PCI Express card is mounted in a PCI Express card cassette, which is then inserted in one of the slots.

Figure 2.27 PCI Express card cassette

2.14.2.3 Compatible PCI cards

The following PCI Express cards are compatible with the PRIMEQUEST PCI_Box.

Table 2.24 Compatible PCI Express cards

Type	Function
Fibre Channel card	PCI Express Fibre channel (4Gbps x 1 port)
Fibre Channel card	PCI Express Fibre channel (4Gbps x 2 port)

2.14.3 Requirements on PCIU and PEXU connections and combinations

2.14.3.1 Requirements on connecting PCIUs and PEXUs to one IO Unit

One PCI-Box interface port and one PCI-Box interface side band port compose a set of ports for connecting a PCIU. An IO Unit is equipped with a total of four ports (Ports #0 to #3) for PCIU connections.

PCIU

A PCIU has one set of ports for connecting an IO Unit, and the set consists of one PCI Express port and one PCIU Control port. These ports that connect an IO Unit are connected with PCI_Box interface ports on the IO Unit.

PEXU

A PEXU has two sets of ports for connecting IO Units, and each set consists of one PCI Express port and one PCIU Control port.

One PEXU must be connected to two PCI_Box interface ports on an IO Unit.

Make sure that PCI_Box interface ports #0 and #1 or PCI_Box interface ports #2 and #3 on one IO Unit are paired for a connection to one PEXU. PCI_Box interface ports #1 and #2, PCI_Box interface ports #0 and #3, PCI_Box interface ports #0 and #2, or PCI_Box interface ports #1 and #3 cannot be paired for a connection to one PEXU.

2.14.3.2 IOU and PEXU connection examples

Examples of permitted and prohibited IOU and PEXU connection patterns are shown below.

- One PEXU is connected (pattern 1)
  Port #0 on the IO Unit is connected to Port #0 on the PEXU, and Port #1 on the IO Unit is connected to Port #1 on the PEXU.
- One PEXU is connected (pattern 2)
  Port #2 on the IO Unit is connected to Port #0 on the PEXU, and Port #3 on the IO Unit is connected to Port #1 on the PEXU.
- Two PEXUs are connected
  The PEXUs is connected as follows:
  Port #0 on the IO Unit to Port #0 on the first PEXU
  Port #1 on the IO Unit to Port #1 on the first PEXU
  Port #2 on the IO Unit to Port #0 on the second PEXU
  Port #3 on the IO Unit to Port #1 on the second PEXU

Figure 2.28 IO Unit and PEXU connections

Connections such as those shown below are prohibited.

Figure 2.29 IO Unit and PEXU connections (prohibited patterns)

2.14.4 Connection requirements when XPAR partitioning is used

On a partitioned IO Unit, the ports for connecting PCIUs (Ports #0 to #3) can be freely assigned to LIOU_A and LIOU_B.

To connect a PEXU, however, the connection must be made in units of groups (group A consisting of Ports #0 and #1, or group B consisting of Ports #2 and #3), and the ports in one group (Ports #0 and #1, or Ports #2 and #3) must be assigned to one LIOU.

Also, the ports (two ports) connected with a PEXU cannot be assigned to a different partition.

Note: To assign ports connected with a PEXU to a different partition, the partition configuration must be changed after their cables are disconnected from the PEXU or after the PEXU is disconnected.

2.14.5 Requirements on a combination of PCIUs and PEXUs in one PCI_Box

PCIUs and PEXUs can be mounted together in one PCI_Box.

2.14.6 Requirements on a combination of PCIUs and PEXUs in one partition

There are some restrictions on mounting PCIUs and PEXUs together in one partition.

The MMB automatically recognizes the connection patterns of IO Units, PCIUs, and PEXUs. If a recognized pattern does not correspond to any of the patterns in Table 2.25 to Table 2.28, the MMB turns off the power to the relevant partition.

If the connection state of a partition corresponds to multiple connection patterns, the MMB chooses a pattern according to the following order of priority: pattern 1, pattern 2, pattern 3, and pattern 4.

For example, if PCI_Box interface ports #0 and #1 on IOU #0 are the only ports in a partition and these ports are connected to a PCIU, the partition corresponds to connection patterns 1, 3, and 4. In this case, the MMB chooses pattern 1, which has the highest priority.

Remarks: If there were no PCIU or PEXU connected to the partition, the MMB would choose pattern 1.

Table 2.25 Combination of PCIUs and PEXUs in one partition (pattern 1)

PCI_Box interface port on IO Unit	IOU#0	IOU#1	IOU#2	IOU#3	IOU#4	IOU#5	IOU#6	IOU#7
#0	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU
#1	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU
#2	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU
#3	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU

Table 2.26 Combination of PCIUs and PEXUs in one partition (pattern 2)

PCI_Box interface port on IO Unit	IOU#0	IOU#1	IOU#2	IOU#3	IOU#4	IOU#5	IOU#6	IOU#7
#0	PEXU	PEXU	PEXU	PEXU	PEXU	PEXU	PEXU	PEXU
#1								
#2	PEXU	PEXU	PEXU	PEXU	PEXU	PEXU	PEXU	PEXU
#3								

Table 2.27 Combination of PCIUs and PEXUs in one partition (pattern 3)

PCI_Box interface port on IO Unit	IOU#0	IOU#1	IOU#2	IOU#3	IOU#4	IOU#5	IOU#6	IOU#7
#0	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU
#1	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU	PCIU
#2	PEXU	PEXU	PEXU	PEXU	PEXU	PEXU	PEXU	PEXU
#3								

Table 2.28 Combination of PCIUs and PEXUs in one partition (pattern 4)

PCI_Box interface port on IO Unit	IOU#0	IOU#1	IOU#2	IOU#3	IOU#4	IOU#5	IOU#6	IOU#7
Port#0	PCIU	PEXU	PCIU	PEXU	PCIU	PEXU	PCIU	PEXU
Port#1	PCIU		PCIU		PCIU		PCIU	
Port#2	PCIU	PEXU	PCIU	PEXU	PCIU	PEXU	PCIU	PEXU
Port#3	PCIU		PCIU		PCIU		PCIU	

If the LIOUs partitioned with XPAR in an IO Unit do not belong to the same partition, some of the connection patterns in the entire system do not correspond to any of the connection patterns such as IOU #3, IOU #4, and IOU#5 in the table below. In this case, the MMB determines the appropriate patterns in units of partitions. Therefore, in the example below, the MMB turns off the power to only Partition #4.

The partitions in the figure below correspond to the following patterns:

- Partition #0 and Partition #2: Connection pattern 1
- Partition #3: Connection pattern 2
- Partition #1: Connection pattern 4
- Partition #4: Connection violation (The pattern does not belong to the connection patterns.)

Figure 2.30 Interface assignment

CHAPTER 3 General Software Configuration

This chapter describes the PRIMEQUEST software configuration.

As shown in Figure 3.1, the software configuration consists of MMB firmware installed on the management board (MMB), GSWB firmware installed on the gigabit switch board (GSWB), and the software installed in each partition, such as the OS (Linux or Windows), PRIMEQUEST Server Agent (PSA), and related software.

System management operations and configuration are performed from the MMB Web-UI. All system management operations, including operations related to PSA installed in a partition, can be performed from a Web browser running on a general-purpose PC through linkage with a Web server using the MMB firmware.

PRIMEQUEST supports large-scale enterprise environments that require high reliability, by ensuring close linkage between hardware and software.

Figure 3.1 PRIMEQUEST software configuration

3.1 Firmware

This section describes the different types of PRIMEQUEST firmware.

3.1.1 Processor Abstraction Layer (PAL)

PAL is part of the Itanium architecture, and it abstracts the functions unique to IPF CPU hardware to higher layers.

PAL encapsulates processor functions that may vary from one IPF CPU to another. It gives a consistent view of the hardware to SAL and the OS, which are located on a higher layer.

The PAL functions include diagnosis, initialization, and error processing of the CPU hardware.

3.1.2 System Abstraction Layer (SAL)

SAL is firmware that abstracts the platform and provides the OS with the specified interface.

The SAL functions include diagnosis and initialization of the PRIMEQUEST platform including ASIC and memory configuration control.

3.1.3 Extensible Firmware Interface (EFI)

EFI is PRIMEQUEST boot firmware.

EFI provides a boot environment that is not dependent on any platform for an OS such as Linux and Windows.

The EFI functions include loading EFI drivers, diagnosis and initialization of the OS memory area, write/read diagnosis of I/O registers, boot device selection, and setting the order of boot devices.

3.1.4 BMC firmware

BMC firmware is incorporated in each IO Unit. BMC firmware manages the IO Unit configuration, and monitors and controls hardware in mounted units.

3.1.5 MMB firmware

MMB firmware is incorporated in the MMB. MMB firmware is linked with different types of firmware (e.g., PAL/SAL/EFI, BMC firmware, and GSWB firmware) for operations such as managing the entire PRIMEQUEST hardware configuration, managing partition configurations, hardware monitoring, and power control.

For details on the MMB, see CHAPTER 4, "Hardware System Management."

3.1.6 GSWB firmware

GSWB firmware is incorporated in the GSWB, and it controls a LAN switching hub (GSWB). GSWB firmware is operated via MMB firmware. GSWB firmware is used to construct and manage network systems using user LANs (business LANs).

For details of GSWB functions, see the *PRIMEQUEST GSWB User's Manual* (C122-E028EN).

3.2 OS (Linux or Windows)

PRIMEQUEST supports the following OSs. The OSs are installed and used in individual partitions.

- Linux
  - Red Hat® Enterprise Linux® AS (v.4 for Itanium)
  - Red Hat® Enterprise Linux® 5 (for Intel Itanium)
  - SUSE™ Linux Enterprise Server 9 for Itanium Processor Family (*1)
  - SUSE™ Linux Enterprise Server 10 for Itanium Processor Family (*1)

    *1: Mainly supported for markets outside Japan
- Microsoft Windows Server
  - Microsoft® Windows Server® 2003, Enterprise Edition for Itanium-based Systems
  - Microsoft® Windows Server® 2003, Datacenter Edition for Itanium-based Systems
  - Microsoft® Windows Server® 2008 for Itanium-Based Systems

3.2.1 Linux

Linux supports large-scale enterprise environments that require high reliability:

(1) Support of large-scale systems

Linux supports construction of large-scale systems by increasing scalability with the following functions:

- 64-bit virtual space

  Large-scale business applications can be developed without consideration of the virtual space size.

- Support of large-scale SMP

  A high-performance system with a large-scale SMP configuration can be constructed using high-speed interconnect technology.

- Increased number of connected devices

  Large-scale storage systems can be managed.

- Support of large-capacity storage

  Systems that handle large-volume data can be managed.

(2) Higher reliability and availability

Linux works together with hardware to increase reliability and availability as follows:

- Enhancing drivers

  Error handling is enhanced, thereby increasing reliability.

- Machine Check Abort (MCA) information logging function

  Information on recoverable errors is always collected in a log and used for protection against the errors and maintenance.

- PCI hot plugging

  PCI cards that support PCI hot plugging are hot-swappable.

- Guaranteed uniqueness of device names

  Each device name is guaranteed to be unique even if a device is removed or replaced because of a disk failure.

3.2.2 Windows

PRIMEQUEST supports Microsoft Windows Server 2003 and its subsequent versions.

3.3 Supplied Software

The software programs supplied as standard with the PRIMEQUEST-series machines are described below. The software is installed and used in individual partitions.

For details on the supplied software, see Appendix A, "Software Supplied with PRIMEQUEST Hardware."

3.3.1 System management software (PSA)

PSA handles hardware failure monitoring, configuration management, and other tasks in a partition.

For details on the PSA, see CHAPTER 4, "Hardware System Management."

SIRMS

When REMCS is used for operation, SIRMS collects software configuration information on each partition and troubleshooting materials in case of a software failure.

3.3.2 Drivers

SCSI-related drivers, LAN-related drivers, and other types of drivers are supplied.

3.3.3 SystemcastWizard Lite

SystemcastWizard is used to install OSs, back up systems, and restore systems.

3.3.4 System data output tool (fjsnap)

fjsnap collects the system information (e.g., configuration information, operation information, definition information, and logs) required for a failure investigation in the event of a PRIMEQUEST system failure. This software can run in Linux.

3.4 Linked Software

This section describes optional software that can be used in linkage with the PRIMEQUEST-series machines.

3.4.1 PRIMECLUSTER

PRIMECLUSTER is software for cluster operations. Install these programs as necessary. In the PRIMEQUEST system, when Linux is installed in a partition as its operating system, cluster operations can be performed between cabinets and between partitions in a cabinet.

For details on cluster operations, see CHAPTER 8, "Clustering."

For details on PRIMECLUSTER, see the PRIMECLUSTER manuals.

3.4.2 PRIMECLUSTER GDS

PRIMECLUSTER GDS is volume management software that helps increase the availability of data stored in disk drive units and facilitate operation management of data when Linux is installed in a partition as its operating system. PRIMECLUSTER GDS protects disk data against hardware failures and operating errors by users, and it provides assistance in operation management of disk drive units.

For details on the redundant configuration, see CHAPTER 5, "Redundant Configuration."

For details on PRIMECLUSTER GDS, see the PRIMECLUSTER GDS manuals.

3.4.3 PRIMECLUSTER GLS

PRIMECLUSTER GLS is software that uses multiple network interface cards (NICs) to ensure redundancy of the network transmission paths connected to its system in order to provide overall high reliability of communications when Linux is installed in a partition as its operating system.

For details on the redundant configuration, see CHAPTER 5, "Redundant Configuration."

For details on PRIMECLUSTER GLS, see the PRIMECLUSTER GLS manuals.

3.4.4 PRIMEQUEST System Disk Mirror for Windows (PSDM)

PRIMEQUEST System Disk Mirror for Windows (referred to below as PSDM) is software that provides a redundant configuration for a PRIMEQUEST system by mirroring the system disk on which the Windows OS is installed. It thereby enhances system availability.

Mirroring is a function for data redundancy, duplicating disk data to multiple disks. This ensures continuous accessibility to a valid copy of the data even if failures occur on some of the disks.

The advantages provided by PSDM include mirroring in units of individual disks and dynamic setting of EFI boot entries.

For details on PSDM, see the *PRIMEQUEST System Disk Mirror for Windows User's Guide* (B1FN-5771).

3.4.5 Systemwalker

Systemwalker is integrated operation management software that manages systems, networks, resources, etc., to support Internet-based businesses.

For details on Systemwalker, see the Systemwalker manuals.

3.4.6 Backup and restoration software

The following backup and restoration software can be used in linkage with the PRIMEQUEST-series machines:

- PRIMECLUSTER GDS Snapshot
- ETERNUS SF AdvancedCopy Manager
- VERITAS NetBackup™
- NetVault
- NetWorker

For details on backup and restoration, see CHAPTER 9, "Backup and Restoration."

Remarks:

Access the Fujitsu Web site to confirm the latest information on linkage software.

CHAPTER 4 Hardware System Management

This chapter describes the basic mechanism of hardware system management, the functions of the management board (MMB) that is the core component that performs total hardware management, and the functions of the PRIMEQUEST Server Agent (PSA) that monitors each partition to be managed.

- Basic Mechanism
- MMB Functions
- PSA Functions

4.1 Basic Mechanism

Figure 4.1 shows the basic mechanism for hardware system management in PRIMEQUEST series machines.

Figure 4.1 Basic mechanism

The MMB performs total management of the PRIMEQUEST hardware system.

The MMB manages the mounted components (SBs, IO Unit, chip sets, power supply units, fans, LANs, etc.) and partitions, and controls the system clock, management LAN and private LAN, and PCI_Box. The MMB sets various types of operations such as console redirection, KVM switching, and mirror mode. These functions can be operated from the MMB Web-UI.

As shown in Figure 4.1, the MMB is connected to each hardware component through the LANs. The MMB is also connected with various types of sensors through the internal bus. The MMB firmware always monitors and manages the entire system by linking with the BMC firmware mounted in the IO Unit for each partition, the GSWB firmware operating in the GSWB, and the PSA installed in each partition, by using the networks in the cabinet.

The PSA, software that runs on Linux or Windows, manages the partitions.

The PSA monitors the state of the SB and IO Unit forming the partition, and the configuration and state of the connected hardware units.

Monitoring targets include all of the hardware units incorporated in the partition such as the CPUs and memory of the SB, the HDDs and PCI slots of the IO Unit and, if the PCI_Box is connected to the IO Unit, the PCI card to which the PCI_Box is connected.

The PSA of each partition, operating independently of the PSAs in the other partitions, links with the MMB. The PSA is operated from the MMB Web-UI.

The GSWB is a switching hub unit that manages the business LAN. The GSWB links with the MMB and is operated from the MMB Web-UI.

To monitor the hardware, information is exchanged by using an information path separated from the information path of the business system, that is, the private LAN and management LAN. Not depending on the business system, a particular information system with extensive multiple connections encompassing the hardware unit is used. More specifically, various types of firmware around the MMB firmware acquire information detected by the OS and various types of drivers and information detected by various types of sensors and prompts users to take corrective action if necessary.

Monitoring items are set and changed and details such as management values are set from the MMB Web-UI.

Each partition has console information. This console information is input from and output to the COM port of the IO Unit incorporated in the partition.

The PRIMEQUEST series has a function that redirects the console information to the management LAN having the external interface through the firmware and private LAN (console redirection). This function enables the console of each partition to be operated from a general-purpose PC connected to the management LAN.

For details of the console redirection function, see Section 4.2.6, "Console redirection."

The PRIMEQUEST series has the KVM to which the keyboard, mouse, and DVD unit can be connected. The KVM switching function switches the partitions to which the KVM is to be connected. This function enables sharing the keyboard, mouse, and DVD unit among the various partitions.

For details of the KVM switching function, see Section 4.2.5, "KVM/USB connection switching."

Processing for the MMB, PSA, and the GSWB can be executed from the MMB Web-UI. Rather than from a dedicated console, this can be done from a general-purpose PC connected to the management LAN.

Accordingly, functions used by the system administrator or user during operation through the MMB and MMB Web-UI are operated through the Web-UI.

The main functions of the MMB are described in Section 4.2, "MMB Functions."

The main functions of the PSA are described in Section 4.3, "PSA Functions."

The specifications and main functions of the GSWB are described in the *PRIMEQUEST GSWB User's Manual* (C122-E028EN).

4.1.1 LAN configuration (management LAN/private LAN/business LAN)

The PRIMEQUEST cabinet contains three different LANs in addition to the internal bus connection. Each of these LANs has a different use.

These three LANs are the management LAN, private LAN, and business LAN. They are separated for security and load distribution purposes.

The management LAN is provided for system management purposes and has external ports for the user (system administrator), Fujitsu certified service engineer, and remote customer support system (REMCS) on the MMB.

The private LAN is provided for internal control of the system.

Each partition is connected to the GSWB through the intra-cabinet LAN, called the business LAN, and then connected to a LAN outside of the cabinet through the GSWB.

The business LAN is provided for the business system. The GSWB manages the network of the entire business LAN. Various types of GSWB settings are operated through the MMB from a general-purpose PC connected to the management LAN.

Note: The management LAN and the business LAN must be configured on different subnets.

4.1.1.1 Management LAN

The configuration of the management LAN is shown in Figure 4.2.

Note: BMM#1 is supported in the PRIMEQUEST 580A/540A/580/540 only.

Figure 4.2 Management LAN

To manage the PRIMEQUEST system, the management LAN connects the MMB to each partition, and the MMB to a LAN outside the cabinet. The MMB operates and controls each partition through the management LAN. Each partition sends its log information and hardware configuration information to the MMB through the management LAN.

Applications of the management LAN are described below.

- Communication between the MMB and each partition
  Sends and receives information about Web linkage, SNMP linkage, and REMCS linkage between the MMB and the PSA managing each partition.
- Operator console (Web-UI, SSH, etc.)
  Connected as the console of the MMB and each partition
- Management console
  Used for linkage with the operation management system of an external server such as the Server System Manager
- Integrated monitoring software linkage
  Used for linkage with integrated monitoring software such as Systemwalker
- Firmware update
  Used for updating various types of firmware from the MMB
- Time synchronization
  Used for synchronizing the time of each OS while the MMB operates as the NTP server
- External connection
  The following three ports are provided for external connection of the management LAN.
  - User port
    For management purposes such as connecting an external management server and using the operation console
  - Fujitsu certified services engineer’s port
    For connection to a maintenance terminal
  - REMCS port
    For connection to the REMCS Center

The ports for REMCS connections can be selected based on the customer's requirements as follows:

- P-P connection
  Connect a dial-up router directly to the REMCS port.
  - Si-R30/130 dial-up router (10BASE-T4 port on the LAN side)
  - Si-R 170 broadband router (10BASE-T4 port on the LAN side)

  * For the connection to either of the above routers, the settings of the MMB port must be 10 Mbps and half-duplex fixed.

- Internet connection
  For an REMCS connection established through the network that is connected to a user port through a firewall, the REMCS port need not be used.
  For an REMCS connection established through a network that is not connected to a user port because of security requirements, use the REMCS port.

Speed/Duplex setting for the management LAN port

Set "AUTO" for Speed/Duplex of the USER port.

Remarks: If Speed/Duplex of the USER port is set to a value other than "AUTO," the Auto MDI/MDI-X function is disabled. To prevent this problem, use the following cable:
To connect a switching hub device: Crossover cable
To directly connect a PC: Straight cable

Also, note the following about the REMCS port:

- For the PRIMEQUEST 580/540
  The REMCS port also has the Auto MDI-X function. If Speed/Duplex is set to a value other than "AUTO," the appropriate cable for use is the same as that for the USER port.
- For the PRIMEQUEST 580A/540A
  - The REMCS port does not have the Auto MDI-X function.
    Use the following cable regardless of the Speed/Duplex setting:
    To connect a switching hub device: Straight cable
    To directly connect a PC: Crossover cable
    Note: A different cable is used compared with that of the USER port.
    Be careful.
  - If Speed/Duplex of an external device connected to the REMCS port is set to a value other than "AUTO," also set "AUTO" for Speed/Duplex of the REMCS port. (Be sure to set the same value as that of the external device.)

VLAN function of the management LAN

The hub for the management LAN can be found on the MMB. On this hub, cables for each partition network, user port, Fujitsu certified service engineer's port, and REMCS port are concentrated and the Virtual LAN (VLAN) function is mounted.

In a VLAN, the respective ports of one switching hub are logically divided into groups, each functioning as an independent LAN.

The VLAN function of the PRIMEQUEST management LAN provides the following three types of mode. You can select the mode appropriate for the operation method you are using.

- VLAN mode of the management LAN

(1) Inter-partition communication connection mode (No VLAN mode)
This mode enables communication between ports of all partitions. However, communication between the REMCS/Fujitsu certified service engineer's port and the user port, and communication between the REMCS/Fujitsu certified service engineer's port and partition ports are not supported. Y in Table 4.1 indicates that communication between the corresponding ports is available.

Table 4.1 Inter-partition communication connection mode (No VLAN mode)

No VLAN mode	User port	CE maintenance port	REMCS port	MMB	Partition ports
User port	Y	N	N	Y	Y
Fujitsu certified service engineer's port	N	Y	N	Y	N
REMCS port	N	N	Y	Y	N
MMB	Y	Y	Y	Y	Y
Partition ports	Y	N	N	Y	Y

Y: Communication possible, N: Communication impossible

(2) Inter-partition communication disconnect mode (VLAN mode)

This mode disconnects communication with partitions. However, communication between the user port and partition ports and between the MMB and partition ports are enabled in the cabinet. Y in Table 4.2 indicates that communication between the corresponding ports is available.

Table 4.2 Inter-partition communication disconnection mode (VLAN mode)

VLAN mode	User port	Fujitsu certified service engineer's port	REMCS port	MMB	Partition ports
User port	Y	N	N	Y	Y
Fujitsu certified service engineer's port	N	Y	N	Y	N
REMCS port	N	N	Y	Y	N
MMB	Y	Y	Y	Y	Y
Partition ports	Y	N	N	Y	N

Y: Communication possible, N: Communication impossible

(3) All-partition communication disconnection mode (Port disable mode)

This mode is provided for enhancing security.

This mode disconnects communication of the management LAN with ports of all partitions. Intra-cabinet communication between an MMB and port in the same partition is not possible. Y in Table 4.3 indicates that communication between the corresponding ports is available.

Table 4.3 All-partition communication disconnection mode (Port disable mode)

Port disable mode	User port	Fujitsu certified service engineer's port	REMCS port	MMB	Partition ports
User port	Y	N	N	Y	N
Fujitsu certified service engineer's port	N	Y	N	Y	N
REMCS port	N	N	Y	Y	N
MMB	Y	Y	Y	Y	N
Partition ports	N	N	N	N	N

Y: Communication possible, N: Communication impossible

4.1.1.2 Private LAN

The configuration of the private LAN is shown in Figure 4.3.

Note: BMM#1 is supported in the PRIMEQUEST 580A/540A/580/540 only.

Figure 4.3 Private LAN

The private LAN is an internal control LAN for communication among firmware items installed in hardware components. The private LAN enables communication among the BMC firmware mounted on an IO Unit, the MMB firmware, and the GSWB firmware. The private LAN cannot be used from the OS or an applications.

4.1.1.3 Business LAN

The business LAN is used for configuring the user's business system.

The following two methods are available for building a business LAN in an external device:

- Mount a GSWB to use the switching hub function. Manage and operate the GSWB by using the MMB Web-UI.
- Mount a GTHB to use hub functions other than the switching function. Manage the GTHB through the Web-UI of the MMB.
- Mount a 1000BASE-SX card or another card in a PCI slot of an IO Unit, and manage and operate the LAN independently.

The configuration of the business LAN is shown in Figure 4.4.

Note: BMM#1 is supported in the PRIMEQUEST 580A/540A/580/540 only.

Figure 4.4 Business LAN

Figure 4.4 shows two GSWBs that are mounted to configure a business LAN. The GSWB supports layer 2 switching hub functions including the tag VLAN function and port VLAN function. In Figure 4.4, four 1-Gigabit LAN ports are provided for each IO Unit and two ports are connected to each GSWB.

The two GSWBs (GTHBs) are connected to all IO Units in a dual configuration.

One GSWB has eight 1-Gigabit LAN ports for LAN connection with external devices outside the cabinet.

The GSWB specifications and general functions are described in the *PRIMEQUEST GSWB User's Manual* (C122-E028EN).

The GTHB specifications and general functions are described in Section 2.5, "Gigabit Through Board (GTHB)."

4.1.2 IP addresses of PRIMEQUEST management LAN

The PRIMEQUEST management LAN uses the IP addresses shown in Figure 4.5.

Figure 4.5 Management LAN network configuration and IP addresses

The main applications of IP addresses are described below.

- MMB IP addresses

  Each of MMB #0 and #1 has one physical IP address. In addition to these IP addresses, a virtual IP address and a REMOTE MAINTENANCE connection use virtual IP address can be used for both MMBs. The physical IP address is used to control linkages, such as PRIMECLUSTER linkage and Systemwalker linkage.

The virtual IP address is used as the SNMP access destinations from the management server or the MMB Web-UI operation. GSWB and PSA Web-UI operations are also performed via the MMB Web-UI.

- IP address set by the OS on a partition

  An IP address needs to be specified for the operating system on a partition. Physically, this corresponds to an NIC of IOU#0 to IOU#7.

  These IP addresses are used for control of Web-UI linkage, SNMP linkage, and REMCS linkage between the PSA and the MMB or for linkage between the PRIMECLUSTER and the MMB.

  The BMMs in each IO Unit have two network interfaces. These network interfaces are made redundant by Linux Bonding Driver under Linux, and by such methods as INTEL PROset teaming in PRIMECLUSTER GLS and under Windows.

  Note: If the switching hub that supports a loop prevention function, such as the spanning tree protocol or domain separation, is used for an external switch, suspend the loop prevention function by setting disable for the spanning tree protocol of the connection port between the switching hub and the main unit or by setting on to the domain separation.

- The MMB transmits packets in communication of the following types: NTP, alarm e-mail, REMCS, and SNMP trap. If both a physical IP address and a virtual address are set for the MMB, the IP address of the packet sender becomes the physical IP address of the MMB.

  Under the following conditions, firewall, mail, and other servers must be configured such that the packets with the physical IP addresses of MMB#0 and MMB#1 can pass through:

  - The destination server is an external server outside the firewall.
  - IP addresses are restricted by the mail server used.

  Also, if the REMOTE MAINTENANCE port is used with the PRIMEQUEST 580A/540A under the above conditions, the servers must be configured such that the packets with the IP addresses for REMOTE MAINTENANCE can pass through.

4.2 MMB Functions

The MMB functions are indicated by ◆ MMB in Figure 4.6 in which all of the PRIMEQUEST functions are shown.

This section describes the MMB functions.

The MMB functions include the switching function for using the PSA and GSWB functions.

Figure 4.6 Main MMB functions

4.2.1 User privilege management

The user privileges necessary for accessing the MMB are classified as follows. System account elements such as a user name and a user privilege need to be registered. Users can use functions corresponding to their access privileges.

- User..... A user having this privilege is allowed only to reference the system state. The user cannot set system configuration information and cannot turn on and off power supplies of partitions.
- CE ..... A user having this privilege is allowed to reference the system state and to perform maintenance. The user cannot manage users and change network settings.
- Operator ..... A user having this privilege is allowed to reference and set a system state. The user cannot manage users and change the configuration of a management LAN.
- Administrator..... A user having this privilege is allowed to execute all operations.

4.2.2 Hardware configuration display

This function displays the configuration and state of hardware.

This function displays the configuration of all components such as SBs and IO Units including fans and power supply units mounted in the PRIMEQUEST system and units in the components such as CPUs, memory, and HDDs.

Figure 4.7 Example of the configuration display window

Figure 4.7 shows an example of displaying the configuration of the entire system. The states in which the configuration elements of the system are placed are also displayed in the configuration information. For example, whether the configuration elements are in the normal or abnormal state is displayed. In some cases where these elements are in the abnormal state, the system administrator or person in charge of maintenance needs take corrective action. Detailed information can be displayed by tracing the link in the display item field.

Also, the configuration can be displayed in units of partitions. Detailed information on the configuration and states of the SBs and IO Units allocated to each partition are displayed in this case.

4.2.3 Hardware monitor

This function monitors for hardware failures and errors.

Hardware failures and errors are detected by various types of check mechanisms and sensors. If a failure or an error is detected, the required corrective action must be taken according to the location where the failure or error was detected, and the elements of the system configuration in place, such as mirror mode, redundant configuration, and the like. All logs including logs about processing that was able to proceed following a failure or error occurrence (such as parts replacement and parts disconnection) are collected and error messages are issued.

Failures and errors are displayed by LEDs according to their severity. In addition, different types of reports are supported as follows: reports to applications for operation management following SNMP trap configuration, mail reports to the system administrator following alarm e-mail configuration, and reports to the REMCS Center following REMCS (remote customer support system) registration.

To reduce the amount of displayed information, all logs and messages can be filtered prior to their display. The conditions for notifying information on abnormal states that have been detected can also be set.

An example of the display of system event logs is shown in Figure 4.8, and an example of the filter selection screen is shown in Figure 4.9.

Severity	Date/Time	Unit	Source	Event ID	Description	Detail
Info	2005-12-14 14:36:52	Partition#0	P#0 Sys Status	C06F02FF	Reset	Detail
Info	2005-12-14 14:32:45	Partition#0	P#0 Sys Status	C06F02FF	Reset	Detail
Info	2005-12-14 14:28:35	Partition#0	P#0 Sys Status	C06F02FF	Reset	Detail
Info	2005-12-14 14:24:28	Partition#0	P#0 Sys Status	C06F02FF	Reset	Detail
Info	2005-12-14 14:20:07	Partition#0	P#0 Sys Status	C06F02FF	Reset	Detail
Info	2005-12-14 14:19:22	Partition#0	P#0 Sys Status	C06F01FF	Power On In Progress	Detail
Info	2005-12-14 11:43:09	Partition#0	P#0 Sys Status	C06F00FF	Power Off	Detail
Info	2005-12-14 11:42:21	Partition#0	P#0 Sys Status	C06F08FF	Power Off In Progress	Detail
Info	2005-12-14 11:40:27	Partition#0	P#0 Sys Status	C06F02FF	Reset	Detail
Info	2005-12-14 11:36:20	Partition#0	P#0 Sys Status	C06F02FF	Reset	Detail

Figure 4.8 Example of the system event log display window

System | Partition | Switch | User Administration | Network Configuration | Maintenance Logout

>System >System Event Log

- System Status
- System Event Log
- System Information
- Firmware Information
- System Setup
- System Power Control
- LEDs
- Power Supply
- Fans
- Temperature
- SB
- IOU
- System Interconnect
- Other Boards
- MMB
- GSWB
- PCI_Box

System Event Log Filtering Condition Help

Select the filtering conditions and click the Apply Button to take effect.
Note : The followings are AND conditions.

1)Severity: ☑ Error ☑ Warning ☑ Info

2)Partition: ◉ All
○ Specified ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10 ☐ 11 ☐ 12 ☐ 13 ☐ 14 ☐ 15

3)Unit: ◉ All
○ Specified ☐ PSUs ☐ Fans
☐ All SB(s) ☐ SB#0 ☐ SB#1 ☐ SB#2 ☐ SB#3
☐ SB#4 ☐ SB#5 ☐ SB#6 ☐ SB#7
☐ All IOU(s) ☐ IOU#0 ☐ IOU#1 ☐ IOU#2 ☐ IOU#3
☐ IOU#4 ☐ IOU#5 ☐ IOU#6 ☐ IOU#7
☐ XAI#0 ☐ XAI#1
☐ XDI#0 ☐ XDI#1 ☐ XDI#2 ☐ XDI#3
☐ CPCB ☐ KVM ☐ Panel
☐ MMB#0 ☐ MMB#1
☐ GSWB#0 ☐ GSWB#1 ☐ GTHB#0 ☐ GTHB#1
☐ All PCI_Box ☐ PCI_Box#0 ☐ PCI_Box#1 ☐ PCI_Box#2 ☐ PCI_Box#3
☐ PCI_Box#4 ☐ PCI_Box#5 ☐ PCI_Box#6 ☐ PCI_Box#7

4)Source: ◉ All
○ Specified ☐ CPU ☐ DIMM ☐ Chipset
☐ Voltage ☐ Temperature ☐ Other

5)Sort by Date/Time: ◉ New event first ○ Old event first

6)Start Date/Time: ◉ First event ○ Specified Time 2005 - 1 - 1 0 : 0 : 0

7)End Date/Time: ◉ End event ○ Specified Time 2005 - 1 - 1 0 : 0 : 0

8)Number of events to display: 100 /995

Apply Cancel Default Setting

Figure 4.9 Example of the filter condition setting window

4.2.4 Partition settings and configuration display

This function is used to set a partition and display its configuration. Partition settings are used to divide mounted hardware resources and to determine the units of control and management.Figure 4.10 shows an example in which three partitions are set.

Figure 4.10 Example in which different OSs and OSs of different versions are mounted in respective partitions

A partition is set by specifying a combination of SBs and IO Units. The partition configuration information can be displayed in detail, showing incorporated SBs and IO Units and their configuration elements. As with the settings, the partition can be changed by changing the combination of the SBs and IO Units.

An example of the display of the partition configuration screen is shown in Figure 4.11.

Figure 4.11 Example of display of the partition configuration window

4.2.5 KVM/USB connection switching

This function switches the connection destinations of the KVM, a component for connecting the keyboard, mouse, display, and DVD drive.

As shown in Figure 4.12, arbitrary partitions comprise the connection destinations. The connection destination can be set from the MMB Web-UI. An example of the switching setting screen of the KVM switch of the MMB Web-UI is shown in Figure 4.13. This function enables the keyboard or mouse USB, video output, and DVD-ROM drive USB to be connected to the required partitions individually.

Figure 4.12 KVM/USB connection switching

Figure 4.13 Example of the KVM/USB individual connection switching setting window

4.2.6 Console redirection

This function sets I/O destinations of each partition's console information through a management LAN.

The two methods below are available for OS input and output in each partition. These methods can be set independently in each partition:

- Using the COM port of the IO Unit incorporated in each partition
- Using the management LAN through the MMB

Figure 4.14 shows the flow of console information sent when the information is redirected to the management LAN.

Figure 4.14 Flow of redirected information

When the information is redirected to the management LAN, a terminal connected to the management LAN can be used as a console.

#	Partition Name	Status	Power Status	Connect to	Telnet/SSH	Port#
0	MasterDBS	OK	Standby	COM MMB	None Telnet SSH	7000
1	ManDBS	OK	Standby	COM MMB	None Telnet SSH	7001
2		None		COM MMB	None Telnet SSH	7002
3		None		COM MMB	None Telnet SSH	7003
4		None		COM MMB	None Telnet SSH	7004
5		None		COM MMB	None Telnet SSH	7005
6		None		COM MMB	None Telnet SSH	7006
7		None		COM MMB	None Telnet SSH	7007
8		None		COM MMB	None Telnet SSH	7008
9		None		COM MMB	None Telnet SSH	7009
10		None		COM MMB	None Telnet SSH	7010
11		None		COM MMB	None Telnet SSH	7011
12		None		COM MMB	None Telnet SSH	7012
13		None		COM MMB	None Telnet SSH	7013
14		None		COM MMB	None Telnet SSH	7014
15		None		COM MMB	None Telnet SSH	7015

Figure 4.15 [Console Redirection Switch] window

4.2.7 MMB use environment setting

This function sets and changes the MMB use environments.

- Centralized system account

The MMB and GSWB support interfaces such as Telnet and Web-UI and their system accounts are managed centrally by the MMB. The MMB also performs GSWB user authentication. For centralized management, the Lightweight Directory Access Protocol (LDAP) is used. Because the MMB adopts a redundant configuration, MMB #0 is used as the master LDAP server and MMB #1 is used as the backup server.

Figure 4.16 shows how system accounts are managed.

Figure 4.16 System account management method

- User management
  For information on the user privilege for accessing the MMB, see Section 4.2.1, "User privilege management."
- Making and changing the network environment settings
  In this kind of operation, the MMB usage environments for the HTTP and Telnet is set. The network environments of a business LAN is set up and changed from the Switch item in the Navigation Bar on the setup and modification screen.

Figure 4.17 Example of setting the usage evironment

- Access control

  To ensure security, an IP filter for permitting access to the MMB is set up. (Usable IP addresses are specified for each protocol.). Access to the MMB is allowed only from the set IP addresses.

Figure 4.18 Example of the access control settings screen

- Secure Sockets Layer (SSL) support
  Web and Telnet access is ciphered in the SSL. Secret keys and electronic certificates are generated in this layer.

4.2.8 Time synchronization

To assure synchronization among partitions, this function synchronizes the partition times with that of the MMB.

The NTP server function and the NTP client function are used to synchronize partition times with the MMB time. As shown in Figure 4.19, synchronization with the NTP server time is possible by accessing the other NTP server. However, the MMB need not be specified for the NTP servers in any of the partitions. For stable NTP operation, specify multiple NTP servers from the NTP client (Fujitsu recommends at least three).

Note: NTP server 1 to NTP server 5 are NTP servers using very high-precision clocks for connections to the Internet or intranets.

Figure 4.19 Time synchronization image

4.2.9 Power control/scheduled operation

This function sets a schedule for turning on and off the power supply of the partitions and automatically controls the operation of the partitions in accordance with the set schedule. Figure 4.20 shows an example of the display for scheduled operation.

Figure 4.20 Example of the scheduled operation display window

4.2.10 Firmware maintenance

This function updates various types of firmware. The update is performed by Fujitsu certified service engineer.

The following firmware components are subject to update:

- MMB firmware (installed in the MMB)
- GSWB firmware (installed in the GSWB)
- BMC firmware (installed in each partition)
- PAL/SAL/EFI firmware (installed in each partition)

4.2.11 Setting information save and restore

This function saves and restores PAL/SAL/EFI setting information for each partition and MMB setting information.

The MMB provides a save and restore function for PAL/SAL/EFI setting information of each partition as shown in Figure 4.21.

This function enables the following operations:

- After PAL/SAL/EFI settings are made from the PAL/SAL/EFI SETUP screen in one partition, the PAL/SAL/EFI setting information for that partition can be saved and applied to another partition from the MMB.
- If a failure occurs in an SB and the SB is swapped, the saved PAL/SAL/EFI information enables restoration of the original state.

Saved information can be stored on a remote terminal. Data saved on a remote terminal can be restored.

In addition to the function for the PAL/SAL/EFI configuration information, a save and restore function for saving and restoring MMB and GSWB configuration information is provided as well.

Figure 4.21 Schematic of configuration information save and restore operation

4.3 PSA Functions

The PSA functions are indicated by ◆ PSA in Figure 4.22 which shows all of the PRIMEQUEST functions.

This section describes each PSA function.

Remarks: If a PRIMEQUEST series machine is operated, be sure to install the PSA. If the PSA is not installed, the following restrictions apply:

- Notification of errors of I/O devices (PCI cards, HDDs, etc.) and trap notification to the administrator cannot be performed.
- Partition monitoring by using the Watchdog function cannot be performed.
- The following error notification through failure prediction and trap notification to the administrator cannot be performed:
  - Correctable error count of a CPU, DIMM, or chip set, that exceeds the threshold
  - Exceeded threshold in the S.M.A.R.T. (Self-Monitoring Analysis and Reporting Technology system) monitoring of an HDD
- The operation management software cannot collect information from the partition side.
- When a REMCS contract is entered, software errors are not reported.
- HDD cannot be hot-maintained. The partition must be stopped during maintenance.
- PRIMECLUSTER linkage cannot be used.

Note: When using this function under Windows, do not stop the Windows print spooler service.

For the management of the hardware configuration, Windows Management Instrumentation (WMI) is used in the processing for collecting information from the operating system. However, when the print spooler service is stopped, an error is reported to WMI and the information is not collected correctly.

Figure 4.22 PSA functions

4.3.1 Operation management GUI

This function is a Web-UI function that can be used to manage partition operation.

By linking the PSA in each partition to MMB firmware, the partition can be displayed and operated via the MMB Web-UI without a Web server function at the partition end.

As shown in Figure 4.23, the MMB firmware has the Common Gateway Interface (CGI-WebGateCGI) which is used for communication with the Web server function. The PSA in each partition has the operation management GUI functions comprising the WebGate and HTML template group.

When receiving a request from a user, the WebGateCGI communicates with the WebGate by TCP/IP and delivers a corresponding HTML. The WebGate acquires information from the data source (configuration information, etc.) in accordance with the request and embeds the information in an HTML template.

In this manner, the Web-UI function for managing partition operation is provided.

Figure 4.23 Web-UI function

The following functions, which enable script operation and operation in general by using the command line from an OS console, are provided by the Command Line Interface (CLI). For details, see the *PRIMEQUEST 580A/540A/520A/500/400 Series Reference Manual: Basic Operation/GUI/Commands* (C122-E003EN). For the DP operation commands, see *PRIMEQUEST 580A/580 Dynamic Partitioning (DP) Manual* (C122-E085EN).

- SAF_TE operation command (used by a Fujitsu SE when hot swapping disks)
- PSA start/stop command
- PSA survey data collection command
- Filter definition update command
- Local partition number acquisition command
- Serial number acquisition command
- SNMP security setting command
- Firmware information acquisition command
- fjsvdrc (DP operation command)

4.3.2 Hardware configuration management

This function displays the hardware resources that form a partition.

This function displays the following configuration information:

- SB configuration, IO Unit configuration
- CPU configuration (maximum number of CPUs which can be mounted, CPU mounting locations, identification information such as the CPU type, error status information)
- Memory configuration (mounting locations, detailed information such as the memory type, error status information (PSA 1.11 or earlier))
- PCI configuration (PCI card mounting, PCI device mounting, detailed information such as the PCI device type, error status information)
- SCSI/FC connection unit configuration (HDD, tape drive, etc.)
- Network configuration (network interfaces, error status information (PSA 1.11 or earlier))

4.3.3 OS information display

This function displays information on the OS installed in a partition.

This function displays the following information:

- OS information (OS type, OS version, and package installation information)
- Storage configuration information (device and capacity information)
- Network configuration information (interface, connection state, speed, routing information)
- OS status (operating hours and login count)
- Process list

4.3.4 Hardware error monitoring

PSA monitors the partition for errors output by PCI cards and SCSI device drivers, and periodically monitors predictive signs detected by the S.M.A.R.T. (Self-Monitoring Analysis and Reporting Technology) function of the hard disk. When an error is detected, error analysis is performed to identify the unit. The analysis result is recorded as logging information and reported to the MMB and the management software at an upper layer.

PSA records the counts of various errors detected in the CPU, memory, chip sets, and interchip buses by the OS and firmware.

The PSA performs an error analysis to identify a component or unit that had more correctable errors than defined in the threshold for a given period. The PSA then records this information as log information.

- PSA versions 1.11 and earlier support the functions of error count analysis and degradation reservation.
- The hardware can detect errors of the following two types:
  - Uncorrectable error (UE)
  - Correctable error (CE)

  If an UE occurs, the hardware stops all partitions affected by the error, disconnects the faulty component, and attempts restart. Or the hardware keeps the partition stopped and waits for maintenance.

  CEs are corrected by the hardware function. Therefore, the partition need not be stopped or the faulty component need not be disconnected immediately. However, if a CE occurs frequently, the component may be degraded, making it likely that a fatal failure will occur in the future.

4.3.5 Log collection, analysis, and display

This function collects, analyzes, and displays logs related to the hardware error monitoring function.

When various types of events and messages are posted from firmware, various types of drivers, and the OS, this function records them in log files and takes a predefined action (notification by e-mail, notification to the REMCS, and log output). The recorded logs can be filtered based on conditions such as the time period messages are posted and the target message type, so as to display an appropriate number of log records on the screen.

The following table lists log files and information the PSA collects in the files.

Table 4.4 Log file information

Log file type	Description
Agent message log	Information related to the events (excluding event IDs 00000 to 09999 detected in the PSA) for which the PSA took action (recording to an OS log, SNMP trap, etc.) is stored in the agent log. This log can be displayed in a format supported by the GUI and can be downloaded in the form of a CSV-format file.
Error record log	Information on firmware-detected errors is recorded by the OS machine check handler. The PSA monitors the error information recorded by the OS machine check handler and stores it in an OS log file.
System event log	System event logs are logged by the MMB. From a partition, information related to the partition can be collected. The PSA periodically performs polling of the system event logs and stores them in OS log files. The stored SEL information can be displayed in a format supported by the GUI and can be downloaded in the form of a binary file.

4.3.6 Maintenance operation

This function supports hot swapping of hard disk in partitions.

The SCSI controller of the hard disk used in the PRIMEQUEST series enables setting the disk power to on or off and the disk to be inserted or extracted with the SCSI Accessed Fault-Tolerant Enclosures (SAF-TE) function. The PSA provides a function by an SAF-TE operator command for safe maintenance performed by the SAF-TE function when the disk is swapped at detection of a hardware failure or at disk expansion.

Remarks: The SAF-TE operator command is supported only under Linux.

4.3.7 REMCS linkage

This function works with the MMB to report resource information and errors of the partition to the REMCS Center (remote maintenance center).

REMCS Agent reports PRIMEQUEST system errors and log information to the REMCS Center via the Internet or a P-to-P connection.

PRIMEQUEST REMCS Agent consists of the MMB firmware and the PSA and SIRMS installed in each partition. As shown in Figure 4.24, the MMB firmware monitors for errors in the entire system and, when it detects an error, reports the error to the REMCS Center. The PSA reports hardware error information and hardware configuration information detected by the OS on the partition to the REMCS Center via the MMB firmware. The PSA also reports software configuration information and software error information detected by the SIRMS to the REMCS Center via the MMB firmware.

For information on REMCS, see Chapter 7 in the *PRIMEQUEST 500A/500/400 Series Reference Manual: Tools/Operation Information* (C122-E074EN).

Figure 4.24 REMCS linkage

4.3.8 PRIMECLUSTER linkage

These are the functions for linkage with the PRIMECLUSTER (clustering system).

The function allows for a clustering configuration that works with PRIMECLUSTER machines. A cluster consisting of multiple nodes (a node can be considered one server, and for PRIMEQUEST-series machines, a node can be considered one partition) can compose a redundant configuration that has active and standby systems.

If the active system becomes unusable such as because of a fault, the operation on the active system can be taken over by the standby system.

The PRIMECLUSTER linkage function is installed in MMB firmware and PSA to support the following functions:

- System status monitor/display..... Monitors and displays the status of the specified node (partition).
- System status notification..... Notifies the associated nodes, which together with the local node form the cluster, of a status change of the local node.

- Event reception from another system.....Receives a status change of a remote node.
- Instruction to another system .....Issues an instruction to the specified remote node.

As shown in Figure 4.25, control information is taken over from the active system to the standby system to continue operation.

Remarks: The PRIMECLUSTER linkage is supported only under Linux.

Figure 4.25 PRIMECLUSTER linkage

4.3.9 Operation management software linkage

This function enables a linkage with operation management software.

Figure 4.26 shows software that is installed on a partition and that can work with the PSA.

The function enables the PSA to work with operation management software such as Systemwalker. It uses SNMP (Simple Network Management Protocol) for the linkage.

Figure 4.26 Operation management software linkage

4.3.10 Control targets of MMB user interface

The following is an overview of the MMB management and PSA management targets.

MMB management targets

- Hardware mounting information and status in cabinet
  (SB, IO Unit, PCI_Box, XAI, XDI, FAN TRAY, PSU, KVM, MMB, GSWB, etc.)
- System information display and settings (including cabinet setting, MMB, and GSWB)
- Partition configuration management and settings
- Maintenance operation (hot swapping of PCI cards, collection and display of logs for partitions)

PSA management targets

- Information management and operation within a partition
  (PCI card, connection I/O, OS information display, and OS resource performance)
- Maintenance operation (hot swapping of disks, collection and display of logs for partitions)

Available functions

The following operations are available from GUI on a partition.

Table 4.5 Operations available from GUI on a partition

Available operation	Overview
Displaying partition configuration information	SB configuration and IO Unit configuration
	CPU configuration (maximum number of CPUs which can be mounted, CPU mounting location, identification information such as the CPU type)
	Memory configuration (mounting location, detailed information such as the memory type, error status)
	PCI configuration (PCI card mounting, PCI device mounting, detailed information such as the PCI device type, error status information)
	SCSI/FC connection unit configuration (HDD, tape drive, etc)
	Network configuration (network interfaces and error status information)
Displaying and manipulating OS information	OS information (OS type, OS version, and package installation information)
	Storage configuration information (device and capacity)
	Network configuration information (interface, connection state, speed, routine information)
	OS status (operating hours and login count)
	Process list
Maintenance operation	Displaying and saving log information (agent log)
Exporting	Exporting the current configuration and status in a partition

4.3.11 Information managed on partition

Table 4.6 Information managed on partition

Information type	Description	
Hardware information	Model information	• CPU information: mounting information, status, type, version, and frequency
		• Memory information: mounting information, status and type (size)
		• SB/IO Unit/PCI_Box information: mounting information
		• PCI card information: mounting information, adaptor name, and detailed information
		• Connection I/O information: mounting information, type, and detailed information
System information	• Operating system: OS type and version number (revision number)	
	• Disk-related information: file system configuration	
	• Network-related information: Interface Name, Network Type, MAC Address, Interface Speed Current Status (Up/Down/Link Down), and PacketSize	
	• Other I/O information	

Note: MMB Web-UI supports the browsers listed below. If another browser is used, the Web-UI window may not be displayed correctly.

- Microsoft(R) IE (Internet Explorer) v5.5 (SP2) or later
- Netscape v7.02 or later

CHAPTER 5 Redundant Configuration

Redundant configuration refers to a configuration where even if some components making up the hardware configuration fail, the remaining components maintain adequate throughput.

PRIMEQUEST provides redundant configuration for most elements of the hardware configuration, such as memory, CPUs, IO Units, chip sets, transmission lines, and power units, to achieve high reliability and high availability.

This chapter describes the following types of redundant configuration adopted by PRIMEQUEST to improve reliability and availability.

For details on the PRIMEQUEST System Disk Mirror for Windows (PSDM), see the *PRIMEQUEST System Disk Mirror for Windows User's Guide* (B1FN-5771).

- Redundancy of Components
- Redundancy of Private LAN and Management LAN
- System Mirror Mode
- Redundancy of an HDD (Use of PRIMECLUSTER GDS)
- Redundancy of the Business LAN (Use of PRIMECLUSTER GLS)

5.1 Redundancy of Components

Figure 5.1 shows components for which redundant configuration can be adopted.

Figure 5.1 General redundant configuration

(1) PRIMEQUEST 580A/580/480

This part describes redundant configuration of the following components for the PRIMEQUEST 580A/580/480 family.

- System Board (SB)

  From one to up to eight SBs can be mounted. Redundant configuration can be adopted depending on the operation methods. For details, see CHAPTER 7, "Partition," and CHAPTER 8, "Clustering." An SB can have one to four IPF CPUs and up to 8 set (32) memory boards (DIMMs).

- Xbar Data Interconnect board (XDI)

  An XDI is a board on which ASIC (GDX) that controls data paths of the crossbar is mounted. An XDI has ports to connect to SBs and IO Units, and transfers data between ports. As basic configuration, four XDIs are mounted. It is possible to adopt redundant configuration (duplication) handling two XDIs as a set. The redundant configuration is referred to as a mirror mode. For details, see Section 5.3, "System Mirror Mode."

- Xbar Address Interconnect board (XAI)

  An XAI is a board on which ASIC (GAC) that controls address paths of the crossbar is mounted. An XAI has ports to connect SBs and IO Units, and transfers address information between ports. As basic configuration, two XAIs are mounted. It is possible to adopt redundant configuration (duplication). The redundant configuration is referred to as a mirror mode. For details, see Section 5.3, "System Mirror Mode."

- Management Board (MMB)

  Two MMBs are mounted in the basic configuration. One of them is in the active state while the other is in the standby state.

- IO Unit

  One to up to eight IO Units can be mounted. Redundant configuration can be adopted depending on the operation methods. For details, see 5.4, "Redundancy of an HDD (Use of PRIMECLUSTER GDS)," CHAPTER 7, "Partition," and CHAPTER 8, "Clustering." An IO Unit has four HDDs and four PCI cassettes. An IO Unit has four PCI_Box interface ports to connect to a PCI_Box that expands the number of PCI slots. To increase the number of PCI slots, connect a PCI_Box interface port to a PCI_Box with cables.

- PCI_Box

  Up to 12 PCI cards can be inserted in a PCI_Box that is connected as an option to the PCI_Box interface output from an IO Unit. Redundant configuration can be adopted depending on the operation method.

- IO_PSU (in the PCI box)

  This unit converts AC input to 48 VDC. This unit adopts the redundant configuration as the basic configuration.

- Gigabit Through Board (GTHB) (PRIMEQUEST 580A/540A/580/540 only)

  In the basic configuration, no GTHB is mounted. Two boards can be mounted as an option. If two boards are mounted, redundant configuration can be adopted.

- Gigabit Switch Board (GSWB)

  In the basic configuration, no GSWB is mounted. Two boards can be mounted as an option. If two boards are mounted, redundant configuration can be adopted.

- Power Supply Unit (PSU)

  A PSU converts AC input to 48 VDC. This unit adopts redundant configuration as basic configuration.

- FAN Tray (FAN)

  Redundant configuration can be adopted as basic configuration for each cooling system in units of individual fans, with the fans mounted on a FAN tray.

(2) PRIMEQUEST 540A/540/440

Redundant configuration is the same as for the PRIMEQUEST 580A/580/480 family.

However, only a maximum of four SBs, IOUs, and PCI_Box units can be mounted.

5.2 Redundancy of Private LAN and Management LAN

The management board (MMB) manages the PRIMEQUEST system as a whole. Figure 5.2 shows connections to management LANs and private LANs from the MMBs in the PRIMEQUEST 580A/580/480 cabinet.

Note: BMM#1 is supported in the PRIMEQUEST 580A/580/540 only.

Figure 5.2 Redundant configuration of management LANs and private LANs

A CPU in an MMB and a Baseboard Management Controller (BMC) mounted on a BMM in an IO Unit manage partitions in a cabinet. A BMC performs management in partitions. An MMB manages functions common across the cabinet and partitions.

Remarks: At least one optional BMM is required per partition. If no BMM is mounted in any of IO Units that are used to configure a partition, the partition does not work. If multiple BMMs exist in a partition, the specified BMM (specified from the MMB Web-UI as the Home IO Unit) operates.

Redundancy of the private LAN

An intra-cabinet LAN referred to as a private LAN connects MMBs and BMCs. A BMC has a LAN interface for a 100 Mbps private LAN for each redundant MMB. The MMB includes a hub for terminating a 100 Mbps LAN signal from a BMC. The private LAN, used to link MMBs to BMCs, is duplicated according to the duplication of MMBs. The system is configured such that it can operate without any problem in the event of a failure on either side of a connection.

Redundancy of the management LAN

A management LAN is used to connect an external device of PRIMEQUEST series machines to partitions in the PRIMEQUEST, and to connect partitions in the PRIMEQUEST to one another. A management LAN is intended for critical system management tasks, such as connection to a console unit (such as a general-purpose PC) and node-to-node connection in a cluster configuration. In the same way as for a private LAN, a hub incorporated in an MMB connects the components in a cabinet. CPUs of an MMB are also connected to a management LAN. A management LAN has three external interface ports. Two ports are allocated for maintenance tasks performed by Fujitsu certified service engineers and REMCS. The remaining port can be used to connect to a console unit, or for connection between nodes across cabinets in a cluster configuration. A management LAN, as in the case of a private LAN, is duplicated according to the duplication of MMBs. The system is configured such that it can operate without any problem in the event of a failure on either side of a connection.

Example of automatic switching at an MMB failure

Redundant configuration is adopted for MMBs. Therefore, there are two management LAN interfaces for MMBs in the entire system. A virtual IP address to be shared is assigned to each of the management LAN interfaces. If MMBs are switched due to an MMB failure, the virtual IP address is inherited and the communication destination unit can continue operation without having to take the MMB status into consideration.

Redundant configuration is also adopted for management LAN interfaces for each partition. A port is connected to MMB #0 and a port is connected to MMB #1. A virtual IP address to be shared is assigned to these ports as well, and a communication destination unit can continue operation without having to give consideration to the redundant status.

Figure 5.3 shows the connection setup of management LAN and an example of where current MMB #0 fails and the system switches to standby MMB #1 to continue operation.

As shown in this example, a management LAN is switched by switching BMM NIC on the partition at the same time that an MMB is switched from MMB #0 to MMB #1.

Figure 5.3 Example of automatic switching between MMB and management LAN at the occurrence of an MMB failure

5.3 System Mirror Mode

In the PRIMEQUEST series, Standard Mirror Mode and Extended Mirror Mode are available as system mirror modes. In Standard Mirror Mode, the address bus crossbar is duplicated. In Extended Mirror Mode, main server components, such as memory, the address bus crossbar, the data bus crossbar, and chip set are duplicated, and data is read from and written to both components at the same time. Since hardware components and read/write are duplicated, even if a hardware component fails, operation can be continued without stopping the system.

Table 5.1 contains the specifications of Mirror Mode.

Table 5.1 System Mirror Mode Specifications

Mode / Item	Mode	
	Standard Mirror Mode	Extended Mirror Mode
Address bus crossbar	Duplicated	Duplicated
Data bus crossbar	Not duplicated	Duplicated
Memory	Not duplicated	Duplicated
Inside of a chip set	Not duplicated	Duplicated

Remarks: Any divided SB or IO Unit cannot be allocated to a partition for which Extended Mirror Mode is set.

Figure 5.4 outlines the system configuration of Extended Mirror Mode.

Extended Mirror Mode duplicates important hardware in the system, i.e., the Xbar address interconnect board (XAI) on which an address crossbar is mounted, and Xbar data interconnect board (XDI) on which a data crossbar is mounted, memory, and the inside of the chip set. Data is also simultaneously processed by using both of duplicated components. Thus, even if one of two duplicated components fails, the other component can continue operation without stopping the system.

Note: BMM#1 is supported in the PRIMEQUEST 580A/580/540 only.

Figure 5.4 Extended Mirror Mode

Mirroring of memory

In Extended Mirror Mode, memory is divided into two groups for mirroring. Therefore, the logical memory capacity becomes half of the physical memory capacity installed. However, data read and write operation is duplicated and performed simultaneously. Thus, even in the case where a part of memory fails, the operation continues using the rest of the memory. Reliability increases with the mirroring of memory. Figure 5.5 shows the logical numbers of memory that are mirrored.

Figure 5.5 Logical numbers of memory (DIMM) for which a redundant configuration has been adopted

5.4 Redundancy of an HDD (Use of PRIMECLUSTER GDS)

This section describes the redundant configuration for a HDD.

In the PRIMEQUEST series, an IO Unit includes SCSI controllers, each of which has its own system. Each SCSI controller connects to two HDDs. An IO Unit has four PCI cassette slots. The unit also has a PCI_Box interface to allow for PCI slot expansion.

The I/O section does not have a redundancy function that would work independent from without software. Therefore, to configure redundancy of the I/O function, combine multiple I/O devices for the intended purpose and install the software required to implement the redundancy function. Optional software is required for redundancy.

Only Linux supports redundancy using optional software. Windows supports mirroring using external RAID.

Note: BMM#1 is supported in the PRIMEQUEST 580A/580/540 only.

Figure 5.6 I/O Configuration

5.4.1 Redundancy provided to system disk drives

PRIMEQUEST supports software mirroring of internal HDDs and system disks as an option. To support this function, optional software is required. Linux supports mirroring using PRIMECLUSTER Global Disk Services (GDS), which is provided as an optional software product. For details, see the PRIMECLUSTER GDS manual. Windows supports mirroring using external RAID product.

Remarks: Mirroring by PRIMECLUSTER GDS does not cover the dump areas (diskdump, kdump).

Minimum configuration for HDD redundancy

Figure 5.7 shows the minimum configuration for HDD redundancy. In this configuration, a redundant configuration of HDDs is implemented by connecting two HDDs to one SCSI bus. Redundancy, however, is not applied to SCSI controllers, and other I/O devices other than HDDs.

Note: BMM#1 is supported in the PRIMEQUEST 580A/540A/580/540 only.

Figure 5.7 Minimum redundancy configuration

Configuration providing redundancy of an HDD and a SCSI controller

Figure 5.8 shows a configuration providing redundancy of an HDD and a SCSI controller. In this configuration, HDD-SCSI controller PXHs are mirrored. I/O paths including FLP and FLI are not mirrored.

Note: BMM#1 is supported in the PRIMEQUEST 580A/540A/580/540 only.

Figure 5.8 Redundancy including SCSI controllers

HDD redundant configuration where all I/O paths are mirrored

Figure 5.9 shows a HDD redundant configuration where all I/O paths are mirrored.

Note: BMM#1 is supported in the PRIMEQUEST 580A/540A/580/540 only.

Figure 5.9 HDD redundant configuration where all I/O paths are mirrored

5.4.2 Redundancy of external storage

Mirroring between cabinets

PRIMECLUSTER GDS supports mirroring between cabinets. This, however, does not include the case where a disk is used as a system disk.

Figure 5.10 shows an example where mirroring between cabinets is enabled using external SCSI ports incorporated in an IO Unit. In this example, external SCSI ports of the same IO Unit are used. To improve system reliability, use external SCSI ports of different IO Units.

Note: Use together with the multipath connection (grmpd).

Figure 5.10 Disk device mirroring between cabinets

Path redundancy

To connect PRIMEQUEST series machines to ETERNUS machines over multiple paths (multipaths), use the ETERNUS multipath driver. If one of the paths in use becomes unavailable, operation is switched to another path working normally to continue operation. When this happens, an application can continue operation, since access is not interrupted. To improve access performance, use all usable paths to control the load balance (distribution of load). For details, see the manual for the ETERNUS multipath driver.

Figure 5.11 ETERNUS multipath connection

5.5 Redundancy of the Business LAN (Use of PRIMECLUSTER GLS)

If MMBs, GSWBs, and GTHBs are duplicated, networks in a PRIMEQUEST cabinet, such as the management LAN, the private LAN, and the business LAN are also duplicated. This section describes the implementation of redundancy for a business LAN as related to the construction of a business system.

To implement redundancy for of a network and the network interface, at least two network interfaces and optional software to switch the interfaces are needed.

Linux supports PRIMECLUSTER Global Link Services (GLS) as optional software. Windows supports duplication using a teaming driver. This section shows an example of network redundancy that can be implemented by combining PRIMEQUEST and PRIMECLUSTER GLS. For details, see the manual for PRIMECLUSTER GLS.

5.5.1 Duplication of a transmission line between servers (high-speed switch method)

The high-speed switch method controls transmission lines by using the PRIMECLUSTER GLS-specific method. The method uses multiplexed transmission lines at the same time. When a failure occurs, a failed line is isolated and the system switches to degraded operation. Since PRIMECLUSTER GLS itself controls multiplexed transmission lines, it is possible to detect a failure at an early stage, but destination units that can be communicated with are restricted to PRIMEQUEST, PRIMERGY, PRIMEPOWER, GP7000F, FUJITSU S series, and GP-S units. When this method is used, the unit cannot communicate with a host on a different network across a router.

Figure 5.12 Example of implementing redundancy using the high-speed switching method

With this method, as shown in Figure 5.13,each of Network Interface Cards (NICs) is connected to a different network, and these NICs are activated and used at the same time. Transmission packets are sent to appropriate paths according to the operational status (whether operation is normal) of transmission paths.

A virtual interface is generated to logically handle multiple NICs as a single NIC. A TCP/IP application can communicate with a destination system without being aware that a network adopts redundant configuration physically by using the IP address (virtual IP address) set in the virtual interface as the own system IP address.

As the connection pattern, destination systems should be connected over the same network. It is impossible to communicate with a system on another network.

Figure 5.13 High-speed switching method

5.5.2 Duplication between a server and a hub and between switches on the same network (NIC switch method)

In the NIC switch method, as shown in Figure 5.14, duplicated Network Interface Card (NICs) are connected on the same network, and switch of transmission lines is controlled by exclusive use. With this method, communication destinations are not restricted. It is possible to communicate with hosts on the same network and on different networks. Since the range of duplication is restricted to the directly connected switches and hubs, to duplicate the entire communication path, it is necessary to duplicate routers and other network devices located halfway as well as the transmission line of the communication destination.

Figure 5.14 Example of implementing redundancy using the NIC switch method

With this method, as shown in Figure 5.15, duplicated NICs are connected on the same network, and switching of transmission lines is controlled by exclusive use (during normal operation, put one NIC in the "up" status for communication). A TCP/IP application can communicate with a destination system without having to take into consideration switching of NICs, by using the IP address set in the physical interface in the "up" status as the own system IP address.

As the connection pattern, duplicated NICs should be connected on the same network. It is possible to connect to a communication destination system either on the same network or on a different network via a router.

If each of the network devices in a multi-vendor environment (hub, router, and other devices) has a duplication function, the reliability of the whole system can be improved by combining devices. In this case, the range of duplication is defined separately for each vendor.

Figure 5.15 NIC switching method

5.6 Degradation Function

If a hardware abnormality is detected in the hardware power-on diagnosis or during system operation, the degradation function stops the faulty hardware and has the system inherit the operation with the remaining hardware.

Degradation reservation is the act of notifying a MMB about hardware that caused more correctable errors than the threshold in a given period, and that is likely to fail. The MMB automatically isolates such hardware at the next reboot.

Whether to use degradation varies depending on the location where the failure occurred, the nature of the failure, the redundant configuration of a component where a failure has occurred, and the mirror mode status. The following discussion includes a case where the location of a failure can be identified and a case where it cannot be identified.

Case where the location of a failure can be identified

If the location of a failure can be identified, degradation is performed as appropriate for the failed location.

If an SB, CPU, or memory fails, and reserved SB set to be switched to in case of an SB failure exists, reserve degradation and stop the partition. Incorporate the reserved SB into the partition at a reboot.

Details on partitions are described in CHAPTER 7, "Partition."

Case where the location of a failure cannot be identified

If the location of a failure cannot be identified, the partition is stopped.

If, after the partition is stopped, the user specifies reboot, reboot or the retry of reboot is performed. If the system cannot be recovered after the specified times of retries has been reached, the partition is stopped.

Remarks: It is very rare that the location of a failure cannot be identified. In most cases, the location of a failure is identified and degradation is performed.

CHAPTER 6 Hot Swapping

This chapter describes hot swapping of hardware components.

Hot swapping is implemented for the purpose of replacing a component that was operating but stopped due to a failure during system operation, with a normal component, without stopping the partition operation. The component that had been swapped can be incorporated into the running system again. In this chapter, hot swapping also refers to deleting or adding a component without stopping the partition, such as deliberately deleting a component from system operation, adding a component for expansion purposes, as well as replacing a component that was operating but stopped due to a failure.

In principle, hot swapping should be implemented by a Fujitsu certified service engineer.

6.1 Hot Swappable Components

Figure 6.1 shows the layout of components included in PRIMEQUEST series machines. Table 6.1 lists whether hot swapping and/or redundant configuration is available for components.

PDB:	Power Distribution Board
PSU:	Power Supply Unit
OPL:	OP-Panel
BP:	Back Plane
SB:	System Board
IOU:	IO Unit
BMM:	BMC Module
PCI:	PCI slot
HDD:	Hard Disk Drive

ACS:	AC Section
XDI:	Xbar Data Interconnect board
XAI:	Xbar Address Interconnect board
FAN:	Fan
KVM:	KVM interface unit
CPCB:	Clock and PCI_Box Control board
MMB:	Management Board
GSWB:	Gigabit Switch Board (option)
GTHB:	Gigabit Through Board (option)

Figure 6.1 Layout of included components

Table 6.1 Component hot swapping and/or redundant configuration availability

Item	Component	Hot swapping	Redundant configuration	Notes on redundant configuration
1	SB	N	Y (optional)	
2	CPU	N	N	
3	Memory (DIMM)	N	Y	Redundancy is ensured when Extended Mirror Mode is set
4	IO Unit	N	N	A redundancy function that is ensured by hardware alone is not supported.
5	BMM	N	N	
6	HDD	Y	Y	Linux: Redundancy is ensured by PRIMECLUSTER GDS (software mirror) See Section 5.4, "Redundancy of an HDD (Use of PRIMECLUSTER GDS)." Windows: System disks only. Redundancy is ensured by a PSDN.
7	PCI Card	Y	Y	Windows Server 2008: See the *PRIMEQUEST 500A/500 Series Microsoft Windows Server 2008 User's Guide* (C122-E087EN). Windows Server 2003: Hot swapping of PCI cards is not supported.
8	BP	N	N	
9	XAI	N	Y	Redundancy is ensured when Standard Mirror Mode or Extended Mirror Mode is set.
10	XDI	N	Y	Redundancy is ensured when Extended Mirror Mode is set.
11	MMB	Y	Y	
12	KVM	Y	N	
13	DVD drive	Y	N	
14	CPCB	N	N	
15	GTHB	Y	Y (optional)	Redundancy is ensured by PRIMECLUSTER GLS (software mirror) See Section 5.5, "Redundancy of the Business LAN (Use of PRIMECLUSTER GLS)."
16	GSWB	Y	Y (optional)	Redundancy is ensured by PRIMECLUSTER GLS (software mirror) See Section 5.5, "Redundancy of the Business LAN (Use of PRIMECLUSTER GLS)."
17	OP Panel	Y	N	

Item	Component	Hot swapping	Redundant configuration	Notes on redundant configuration
18	PSU	Y	Y	
19	FAN TRAY	Y	N	
20	FAN	Y	Y	
21	PCI_Box	N	-	Redundancy is ensured depending on the operation.
22	PCIU	N	-	Redundancy is ensured depending on the operation.
23	PCI Card	Y	Y	Windows Server 2008: See the *PRIMEQUEST 500A/500 Series Microsoft Windows Server 2008 User's Guide* (C122-E087EN). Windows Server 2003: Hot swapping of PCI cards is not supported.
24	PEXU	N	-	Redundancy is ensured depending on the operation.
25	PCI Express Card	Y	Y	Windows Server 2008: See the *PRIMEQUEST 500A/500 Series Microsoft Windows Server 2008 User's Guide* (C122-E087EN). Windows Server 2003: Hot swapping of PCI cards is not supported.
26	IO_PSU	Y	Y	

Y = Available, N = Unavailable

*1 This is supported only for Linux.

Redundant configuration and hot maintenance of each unit are described below.

Remarks: In a state where System Mirror Mode is not set, memory or crossbars are not mirrored. In this state, Mirror Mode in the [System Setup] window is set to [Disable].

(1) SB

A reserved SB can be used for a redundant configuration, but hot swapping is not available.

If an SB, CPU, or memory fails in a partition for which a reserved SB is set, the reserved SB is incorporated into the partition to replace the failed SB. The partition must be rebooted to incorporate the reserved SB into the partition.

(2) CPUs

Hot swapping is not available.

To swap the failed CPU, stop the partition that includes the SB containing the failed CPU, extract the SB containing the failed CPU, and maintain (swap) the CPU. Insert the SB containing a new CPU into the partition, and then restart the partition.

If a CPU fails, the partition stops, isolates the failed CPU, and restarts automatically.

Isolating the failed CPU may cause no normal CPU to be available in the SB. In this case, the entire SB is subjected to isolation even if a normal DIMM exists. If no normal SB is available in the partition, the partition stops.

(3) Memory (DIMMs)

Memory (DIMM) cannot be subjected to hot swapping. Stop the partition that includes the SB containing the failed DIMM, extract the SB containing the failed DIMM, and replace the failed DIMM.

- When the system is operated in Extended Mirror Mode:
  Only redundant configuration is available. If a DIMM fails, the system containing the failed DIMM stops, and operation continues with the other normal system alone.
- When the system is operated in Standard Mirror Mode or without Mirror Mode set:
  If a DIMM fails, the partition stops. The failed DIMM is automatically isolated from the partition when the partition restarts. The unit of isolation is a set of four DIMMs including the failed DIMM (same as the unit of expansion). The partition automatically restarts after the DIMM failure.
  Isolating the failed DIMM may cause no normal DIMM to be available in the SB. In this case, the entire SB is subjected to isolation even if a normal CPU exists. If no normal SB is available in the partition, the partition stops.

(4) IO Units

A redundant configuration and hot swapping are not available.

Partition may stop due to a failure.

After the partition stops, the failed component can be swapped. To add a new IO Unit to the partition, the relevant partition must be restarted.

(5) BMMs

A redundant configuration and hot swapping are not available.

The partition stops if a BMM fails. Turn off the power to the IO Unit containing the failed BMM, and replace the failed unit.

(6) HDDs

- In a partition with Linux installed:
  A redundant configuration can be created with software for mirroring (such as PRIMECLUSTER GDS).
  If an HDD adopts the mirror configuration using mirroring software (such as PRIMECLUSTER GDS) or if deleting an HDD does not affect system operation, hot swapping is available.
- In a partition with Windows installed:
  Redundancy is ensured for system disks only by a PSDN. Hot swapping is not supported.

(7) PCI cards (in IO Units)

A redundant configuration using mirroring software such as a multipath driver can be used.

Also, in a partition with Linux installed as its OS, a PCI card can be hot-swapped with the PCI card inserted in a card cassette.

For a partition with Windows Server 2008 installed as its OS, see the *PRIMEQUEST 500A/500 Series Microsoft Windows Server 2008 User's Guide* (C122-E087EN).

In a partition with Windows Server 2003 installed as its OS, hot swapping of PCI cards is not supported.

Remarks: For PCI card replacement using PCI Hot Plug, consult with a Fujitsu certified service engineer.

(8) BPs

Redundant configuration is not adopted, and hot swapping is not available.

(9) XAIs

A redundant configuration are available while System Mirror Mode is enabled. Hot swapping is not available.

- When the system is running in Standard Mirror Mode or Extended Mirror Mode:
  Two XAI systems form a redundant configuration. If an XAI fails, the failed XAI is automatically isolated from the system operation configuration and operation continues with the other normal XAI in degradation mode without a reboot. The failed XAI thus isolated can be swapped. To incorporate a new XAI into the system, all the partitions must be restarted.
- When Standard Mirror Mode or Extended Mirror Mode is not set:
  An XAI does not form a redundant configuration. If the XAI fails, all partitions stop. The failed XAI can be maintained (swapped) while all the partitions are stopped. A new XAI after swapping is automatically incorporated into the system when any partition is started.

(10) XDIs

A redundant configuration may become available depending on the operation type. Hot swapping is not available.

- When the operation is performed in Extended Mirror Mode:
  Two XDI systems form a redundant configuration. If an XDI fails, the failed XDI is automatically isolated from the system operation configuration, and the operation continues with the other normal XDI in degradation mode without a reboot. The failed XDI thus isolated can be swapped. To incorporate a new XDI into the system, the entire server must be restarted.
- When the operation is performed in Standard Mirror Mode or without Mirror Mode:
  An XDI does not form a redundant configuration. If the XDI fails, all partitions stop. The failed XDI can be maintained (swapped) while all partitions are stopped. A new XDI after swapping is automatically incorporated into the system when any partition is started.

(11) MMBs

The MMB is usually duplicated and hot swapping is enabled.

If the active MMB fails, it automatically switches to the standby MMB. The failed MMB can be hot-swapped.

(12) KVM

Redundant configuration is not adopted. Hot swapping can be performed for the KVM only if there is no USB, Video, or DVD connection to any partition.

To perform hot swapping, specify [Free] for all the USB, Video, and DVD items in the [USB/Video/DVD Switch] window of the MMB Web-UI that is displayed by selecting [Partition].

(13) DVD drive

Redundant configuration is not adopted, but hot swapping is available. However, the DVD drive cannot solely be swapped since it is the built-in drive of the KVM.

(14) CPCBs

Redundant configuration is not adopted, and hot swapping is not available.

Two Clock Generators (CLGs) are mounted on a CPCB. When a CLG fails, an MMB controls operation to switch from the failed CLG to the normal CLG. When a CLG fails, the entire server is restarted automatically.

(15) GTHBs

Redundant configuration and hot swapping are available. If two GTHBs are mounted and redundant operation is performed by LAN redundancy software (PRIMECLUSTER GLS), hot swapping is available.

(16) GSWBs

Redundant configuration and hot swapping are available.

If two GSWBs are mounted and redundant operation is performed by LAN redundancy software (PRIMECLUSTER GLS), hot swapping is available.

(17) OP Panels

Redundant configuration is not adopted, but hot swapping is available.

(18) PSUs

Redundant configuration is adopted, and hot swapping is available.

(19) FAN TRAYs

Redundant configuration is not adopted, but hot swapping is available.

(20) FANs

Redundant configuration is adopted, and hot swapping is available. However, since FANs are contained in a FAN TRAY, replacement of only a single FAN is not possible.

(21) PCI_Boxes

Hot swapping is not available.

Redundant configurations are available for some operation patterns.

(22) PCIUs (in PCI_Boxes)

Hot swapping is not available.

Redundant configurations are available for some operation patterns.

(23) PCI cards (in PCIU)

A redundant configuration using mirroring software such as a multipath driver can be used.

Also, in a partition with Linux installed as its OS, a PCI card can be hot-swapped with the PCI card inserted in a card cassette.

For a partition with Windows Server 2008 installed as its OS, see the *PRIMEQUEST 500A/500 Series Microsoft Windows Server 2008 User's Guide* (C122-E087EN).

In a partition with Windows Server 2003 installed as its OS, hot swapping of PCI cards is not supported.

Remarks: For PCI card replacement using PCI Hot Plug, consult with a Fujitsu certified service engineer.

(24) PEXU (in PCI_Boxes)

Hot swapping is not available.

Redundant configurations are available for some operation patterns.

(25) PCI Express cards (in PCIU)

A redundant configuration using mirroring software such as a multipath driver can be used.

Also, in a partition with Linux installed as its OS, a PCI card can be hot-swapped with the PCI Express card inserted in a card cassette.

For a partition with Windows Server 2008 installed as its OS, see the *PRIMEQUEST 500A/500 Series Microsoft Windows Server 2008 User's Guid*e (C122-E087EN).

In a partition with Windows Server 2003 installed as its OS, hot swapping of PCI cards is not supported.

Remarks: For PCI card replacement using PCI Hot Plug, consult with a Fujitsu certified service engineer.

(26) IO_PSUs (in PCI_Boxes)

Redundant configuration is adopted, and hot swapping is available.

CHAPTER 7 Partition

This chapter describes the partition function the PRIMEQUEST provides and the following items that are required to determine a system configuration:

- Overview of Partition Functions
- PRIMEQUEST Partition Functions
- Partitioning Function of the PRIMEQUEST 580A/540A/580/540 with XPAR Installed
- Partition Configuration Definition
- Points to Consider when Determining Configurations
- Function for Changing Partition Configurations
- Notes to Observe when Configuring Partitions
- Notes on Use of Mixed Mirror Modes

7.1 Overview of Partition Functions

The PRIMEQUEST series allows one system to be divided into multiple systems, with each system operating as an independent system. Each of the individual systems resulting from such division is called a partition. An operating system (OS) can run independently on each partition. This enables running multiple servers, which operate independently from one another, as a single server.

Each partition is protected from the effects of other partitions. Therefore, a failure (such as an OS panic) occurring in one partition does not affect job processing in other partitions directly. The OS can be rebooted and shut down independently in each partition. These operations are performed by the system administrator of each partition.

In each partition, system boards (SB) and IO Units can be combined as required. The configuration of a partition can be defined by using the Web-UI of the management board (MMB).

Figure 7.1 Example of partition configuration

Without XPAR installed, the PRIMEQUEST 480 and PRIMEQUEST 580A/580 can each be configured into up to eight partitions, and the PRIMEQUEST 440 and PRIMEQUEST 540A/540 can each be configured into up to four partitions. Each SB and IO Unit in the PRIMEQUEST 580A/540A/580/540 can be logically divided.

Logically divided SBs and IO Units in the PRIMEQUEST 580A/580 can be configured in up to 16 partitions, as compared with up to 8 partitions in the PRIMEQUEST 540A/540. The process of logically dividing an SB and IO Unit to configure a partition is called "XPAR" (eXtended PARtitioning). An XPAR option is required for implementing XPAR. The XPAR feature is an option of the PRIMEQUEST 580A/540A/500 series.

Remarks: Up to 8 partitions are available in Extended Mirror Mode.

7.2 PRIMEQUEST Partition Functions

This section describes the mounting requirements of SBs and IO Units and their allocation to the PCI Box.

7.2.1 Mounting requirements

When a partition is defined as a combination of SBs and IO Units, the individual mounting requirements of these units must be met.

7.2.1.1 Mounting requirements of SBs

When defining a partition, ensure that the following mounting requirements are fulfilled:

- An SB having memory mapped in the area from 0 GB to 4 GB is called the Home SB. It is set up automatically.
- On each SB, at least one CPU and one set of memory modules (four modules) must be mounted.
- DIMMs on a bus must have the same capacity.
- CPUs on the same SB mounted on the same partition must be of the same type.
- An SB with 32 GB of memory mounted (8-GB DIMM x 4) cannot be mixed in the same partition with an SB having another memory configuration.
- An SB with 16 GB of memory mounted (4-GB DIMM x 4) cannot be mixed in the same partition with an SB having another memory configuration.

7.2.1.2 Mounting requirements of IO Units

When defining partitions, ensure that the following mounting requirements are fulfilled:

- The partition must include one IO Unit having a BMC module (BMM) mounted.
- When a partition contains two or more IO Units with an MBB mounted on each , only one BMM actually operates. The IO Unit that contains the operating MBB is called the Home IO Unit.

Remarks: The Home IOU can be set on the [Home of Partition] screen, which is displayed by clicking the [Home] button on the [Partition Configuration] screen of the MMB Web-UI that is displayed by selecting [Partition].

7.2.2 PCI_Box allocation

Each PCIU/PEXU of a PCI_Box can be allocated to a partition. One PCI_Box can be divided and allocated to up to four partitions. Each PCIU/PEXU is allocated to the partition to which the PCI_Box interface connector of the connected IO Unit belongs.

Figure 7.2 Allocation to a PCI_Box partition (only if a PCIU is mounted)

Figure 7.3 Allocation to a PCI_Box partition (both a PCIU and a PEXU are mounted)

7.3 Partitioning Function of the PRIMEQUEST 580A/540A/580/540 with XPAR Installed

This section describes mounting requirements and allocation rules for SBs and IO Units in the PRIMEQUEST 580A/540A/580/540 with XPAR installed.

7.3.1 Dividing SBs and IO Units when XPAR has been installed

The hardware resources of each SB and IO Unit in the PRIMEQUEST 580A/540A/580/540 can be logically divided into halves internally. By dividing the hardware resources, the PRIMEQUEST 580A/580 with XPAR installed can be configured into up to 16 partitions, as compared with up to 8 partitions in the PRIMEQUEST 540A/540, and this ensures flexibility in the system architecture.

Notes:

- Extended Mirror Mode cannot be set in a partition containing a divided SB or IO Unit. Divided SBs or IO Units cannot be allocated to a partition for which Extended Mirror Mode is set.
- A divided SB cannot be allocated as a reserved SB to a partition for which Extended Mirror Mode is set. Similarly, Extended Mirror Mode cannot be set in a partition to which a divided SB is allocated as a reserved SB.

Figure 7.4 Example of partition configuration (after logical division)

7.3.2 Partition granularity

The minimum unit of partition constituents such as SBs and IO Units is referred to as partition granularity.

The partition granularity of an SB is referred to as an LSB (Logical System Board) and the partition granularity of an IO Unit is referred to as an LIOU (Logical IO Unit).

7.3.2.1 Partition granularity of SB

There are two kinds of SB partition granularity (LSB): SB and logically halved SB. The 1LSB mode uses one SB for partition granularity (LSB) and the 2LSB mode uses logically halved SB for partition granularity (LSB).

Remarks

1: When a partition is operated in Extended Mirror Mode, the SB belonging to this partition can be operated in 1LSB mode. (Divided SBs cannot be set for a partition for which Extended Mirror Mode is set)

2: When a partition is operated in Standard Mirror Mode, the SB belonging to this partition can be operated in either 1LSB mode or 2LSB mode.

3: The LSB must always contain CPUs and memory. Without a CPU and memory, the LSB cannot run.

A logically divided SB together with a logically divided IO Unit (details below) can configure a partition. The maximum number of partitions with such divided components depends on the model and operation mode and XPAR (option) settings as follows:

- PRIMEQUEST 580A/580
  - Operating in Standard Mirror Mode and System Mirror Mode not set
    XPAR (option) not installed: Up to 8 partitions
    XPAR (option) installed: Up to 16 partitions
  - Operating in Extended Mirror Mode: Up to 8 partitions
- PRIMEQUEST 540A/540
  - Operating in Standard Mirror Mode and System Mirror Mode not set
    XPAR (option) not installed: Up to 4 partitions
    XPAR (option) installed: Up to 8 partitions
  - Operating in Extended Mirror Mode: Up to 4 partitions

Figure 7.5 Partition granularity of an SB

The table below shows how the partitions are affected when partition granularity (LSB) fails.

Table 7.1 Effects of LSB failure

Partition granularity	Constituent of partition granularity	Effect of an SB failure on other partitions
LSB (Logical System Board)	1LSB mode	None
	2LSB mode	May have an effect on other partitions belonging to the same LSB when the SB is logically divided.

7.3.2.2 Partition granularity of IO Unit

The IO Unit partition granularity (LIOU) is an IO Unit logically divided into two halves.

A logically divided SB together with a logically divided IO Unit can configure a partition . The maximum number of partitions with such divided components depends on the model and operation mode and XPAR (option) settings as follows:

- PRIMEQUEST 580A/580
  - Operating in Standard Mirror Mode and System Mirror Mode not set
    XPAR (option) not installed: Up to 8 partitions
    XPAR (option) installed: Up to 16 partitions
  - Operating in Extended Mirror Mode: Up to 8 partitions

- PRIMEQUEST 540A/540
  - Operating in Standard Mirror Mode and System Mirror Mode not set
    XPAR (option) not installed: Up to 4 partitions
    XPAR (option) installed: Up to 8 partitions
  - Operating in Extended Mirror Mode: Up to 4 partitions

When an IO Unit is logically divided into two, some components are always allocated to one of the granularities while other components can be arbitrarily allocated to either of the granularities. For details, see 7.3.3.2, "Mounting requirements of IO Units."

Remarks

1: When a partition is operated in Extended Mirror Mode, the IO Unit belonging to this partition can be operated in 1LIOU mode only.
(The IO Unit division cannot be set for a partition for which Extended Mirror Mode is set)

2: When a partition is operated in Standard Mirror Mode, the IO Unit belonging to this partition can be operated in either 1LIOU or 2LIOU mode.

7.3.3 Mounting requirements (when XPAR has been installed)

When a partition is defined as a combination of SBs (LSB) and IO Units (LIOU), the individual mounting requirements of these units must be met.

7.3.3.1 Mounting requirements of SBs

When defining a partition, ensure that the following mounting requirements are fulfilled:

- The SB that has memory mapped in the lowest address location is called the "Home SB," and it is automatically configured.
- On each SB, at least one CPU and one set of memory modules (four modules) must be mounted.
- DIMMs on a bus must have the same capacity.
- The CPUs in a SB and a partition must be of the same type.

Allocation of CPUs and memory in 2LSB mode

When an SB is operated in 2LSB mode, CPUs and memory are always allocated to partitions A and B as shown below.

- An SB with 32 GB of memory mounted (8-GB DIMM x 4) cannot be mixed in the same partition with an SB having another memory configuration.
- An SB with 16 GB of memory mounted (4-GB DIMM x 4) cannot be mixed in the same partition with an SB having another memory configuration.
- Partition granularity LSB#A; Memory A (Channel #A and Channel #B) and CPUA
- Partition granularity LSB#B; Memory B (Channel #C and Channel #D) and CPUB

Figure 7.6 SB partition granularities and allocation rule

Table 7.2 SB partition granularities and allocation rule

Partition granularity	CPU memory		Partition granularity	CPU memory	
LSB#A	CPU#A	0	LSB#B	CPU#B	0
		1			1
	Channel#A	0a0		Channel#C	0c0
		0a1			0c1
		1a0			1c0
		1a1			1c1
		2a0			2c0
		2a1			2c1
		3a0			3c0
		3a1			3c1
	Channel#B	0b0		Channel#D	0d0
		0b1			0d1
		1b0			1d0
		1b1			1d1
		2b0			2d0
		2b1			2d1
		3b0			3d0
		3b1			3d1

7.3.3.2 Mounting requirements of IO Units

When defining partitions, ensure that the following mounting requirements are fulfilled:

- The partition must include one IO Unit having a BMC module (BMM) mounted.

  Remarks: The BMM products are optional.

- When a partition contains more than one IO Unit with BMM, only one of the BMMs can actually run. The LIOU containing the active BMM is called the Home IOU.

  Remarks: The Home IOU can be set on the [Home of Partition] screen, which is displayed by clicking the [Home] button on the [Partition Configuration] screen of the MMB Web-UI that is displayed by selecting [Partition].

- Four mountable HDDs can permanently allocated to an IO Unit at a ratio of two HDDs per LIOU.
- Two PCI slots comprising one set are allocated to an arbitrary LIOU.
- PCI Express connectors are allocated to an arbitrary LIOU in units of one port.

Granularity allocation rule

The allocation of IO Unit granularities is explained below.

When one IO Unit is divided into two granularities, some components are always allocated to one of the granularities and other components are arbitrarily allocated to either of the granularities.

IO Units are allocated to partition granularities (A and B) in accordance with the following rule:

- Blocks a0 and a1 that contain HDDs and blocks e0 and e1 that contain BMMs are always allocated as follows: Blocks a0 and e0 are allocated to partition granularity A. Blocks a1 and e1 are allocated to partition granularity B.
- Blocks b0, b1, c0, c1, d0, and d1 that contain PCI slots and PCI Express connector ports are allocated arbitrarily to either partition granularities A or B.

Figure 7.7 Partition granularities and allocation rule of IO Units

Table 7.3 IO Unit allocation rule

	a0	a1	b0	b1	c0	c1	d0	d1	e0	e1
			#0, #1	#2, #3	#4	#5	#6	#7		
When IO Unit is not divided	A	A	A	A	A	A	A	A	A	A
When IO Unit is divided	A	B	AorB	AorB	AorB	AorB	AorB	AorB	A	B

The example shows an IO Unit allocation for the case that IO Units are divided.

- Blocks b0, b1, c0, c1, d0, and d1, that can be allocated arbitrarily, are all allocated to partition granularity A.

Figure 7.8 Example of allocation of partition granularity to IO Units (1 of 3)

- One half of blocks b0, b1, c0, c1, d0, and d1, which can be allocated arbitrarily, is allocated to partition granularity A and the other half is allocated to partition granularity B.

Figure 7.9 Example of allocation of partition granularity to IO Units (2 of 3)

- Example of allocation of ports to PCI_Box at a ratio of 3:1

Figure 7.10 Example of allocation of partition granularity to IO Units (3 of 3)

7.3.4 PCI_Box allocation

Each PCIU/PEXU of a PCI_Box can be allocated to a partition. One PCI_Box can be divided and allocated to up to four partitions. Each PCIU/PEXU is allocated to the partition to which the PCI_Box interface connector of the connected IO Unit belongs.

Figure 7.11 Allocation of a PCI_Box to partitions (only if a PCIU is mounted)

Figure 7.12 Allocation of a PCI_Box to partitions
(both a PCIU and a PEXU are mounted)

7.4 Partition Configuration Definition

To change the configuration of the partition, define the partition by using the MMB Web-UI.

To do this, first, create a partition having a specific partition ID (PartitionName) , and then register SBs, IO Units, and Home IOU belonging to the partition on the [Partition Configuration] screen of the MMB Web-UI that is displayed by selecting [Partition]. Further, set Reserved SB as required.

For information on the setting procedure, see the *PRIMEQUEST 580A/540A/520A/500/400 Series Reference Manual: Basic Operation/GUI/Commands* (C122-E003EN).

Notes:

- Extended Mirror Mode cannot be set in a partition containing a divided SB or IO Unit. Divided SBs or IO Units cannot be allocated to a partition for which Extended Mirror Mode is set.
- A divided SB cannot be allocated as a reserved SB to a partition for which Extended Mirror Mode is set. Similarly, Extended Mirror Mode cannot be set in a partition to which a divided SB is allocated as a reserved SB.

Figure 7.13 [Partition Configuration] window

7.5 Points to Consider when Determining Configurations

This section describes the points to consider when determining the configuration and operation of partitions. The partition configuration must be determined by considering the requirements of each system.

7.5.1 SB and IO Unit maintenance policy

- Replacement of an SB or IO Unit involves work with partitions stopped.
- When a SB or IO unit is logically divided into two and its components are housed in different partitions, SB or IO unit maintenance may affect both partitions.

7.5.2 Flexible I/O component and reserved SB

A flexible I/O component is mounted in PRIMEQUEST series machines. The flexible I/O component provides a function which separates SBs and IO Units and allows the SBs and IO Units to be combined freely. This component can connect SBs and IO Units freely in comparison with a method that allocates SBs and IO Units permanently.

Although a reserved SB does not belong to any partition, it can be incorporated in a partition to replace a faulty SB. To replace a faulty SB with a reserved SB, the reserved SB can be set as the reserved SB of the partition on the [Reserved SB Configuration] screen of the MMB Web-UI that is displayed by selecting [Partition].

If an SB in a partition becomes faulty, the system switches from the faulty SB to the reserved SB. At the same time, the connections between the faulty SB and IO Units are switched to those of the reserved SB to continue job processing. The flexible I/O component enables the system to restore only by reboot.

Mounting the flexible I/O component and reserved SBs can widely reduce the restore time necessary if an SB becomes faulty.

Notes:

- If an SB failure causes the reboot of a partition, a switch is made to the reserved SB indicated as the Reserved SB in the partition.

If a failure occurs in one of the components (such as memory) in a duplicated component configuration in Extended Mirror Mode and the other component continues operating without interruption, switching to a reserved SB is not made.

- A divided SB cannot be allocated as a reserved SB to a partition for which Extended Mirror Mode is set. Similarly, Extended Mirror Mode cannot be set in a partition to which a divided SB is allocated as a reserved SB.
- An SB with 32 GB of memory mounted (8-GB DIMM x 4) cannot be mixed in the same partition with an SB having another memory configuration. Therefore, an SB with 32 GB of memory mounted (8-GB DIMM x 4) cannot be set as a reserved SB in a partition consisting of SBs having another memory configuration.
- An SB with 16 GB of memory mounted (4-GB DIMM x 4) cannot be mixed in the same partition with an SB having another memory configuration. Therefore, an SB with 16 GB of memory mounted (4-GB DIMM x 4) cannot be set as a reserved SB in a partition consisting of SBs having another memory configuration.

Remarks:

- Fujitsu recommends specifying a reserved SB that has the same memory and CPU configurations as those of the SBs configured in the target partition. If the number of CPUs used increases during switching to the specified reserved SB, additional licenses are required.
- CPUs of the same type as those on the SBs composing the partition must be mounted on the reserved SB. If the SB has a different type of CPU mounted on it, it is not used as a reserved SB.

Figure 7.14 Flexible I/O component

7.5.3 Notes on System Mirror Mode operation

PRIMEQUEST series machines support Extended Mirror Mode, in which memory, crossbar, and chip set contents are duplicated and perform the same operations simultaneously. They also support Standard Mirror Mode, in which address bus crossbars are duplicated and perform the same operations simultaneously.

This enables continuation of the current job without interruption even if one of the duplicated systems fails. These system mirror modes can therefore assure high system reliability. For details of the system mirror modes (Extended Mirror Mode and Standard Mirror Mode), see Section 5.3, "System Mirror Mode."

Note:

- Extended Mirror Mode cannot be set in a partition containing a divided SB or IO Unit. Divided SBs or IO Units cannot be allocated to a partition for which Extended Mirror Mode is set.

7.5.4 Notes on the memory configuration

The SB requires at least one CPU operating normally and one set of memory (DIMM × 4) operating normally. Consequently, if the memory degrades due to detection of an error in the memory and no set of memory operating normally is available, the entire SB is degraded. To prevent the CPU from being degraded concurrently with the SB degradation, two or more sets of memory should be mounted in each SB.

Remarks: The term "degradation" means that if a hardware error is detected by the initial diagnosis hardware executed when the power is turned on or during system operation, the system runs after disconnecting the respective hardware elements.

7.5.5 Operating environment

Two types of OSs (Linux and Windows) can be installed on PRIMEQUEST series machines. Because the OS can be selected for each partition, determine the partition configuration by considering the operation method.

7.6 Function for Changing Partition Configurations

The partition configuration of PRIMEQUEST series machines is changed using the [Partition Configuration] screen of the MMB Web-UI that is displayed by selecting [Partition]. For details of the operation, see the *PRIMEQUEST 580A/540A/520A/500/400 Series Reference Manual: Basic Operation/GUI/Commands* (C122-E003EN).

Remarks: A change to the partition configuration is applied by power-off and subsequent power-on in units of partitions. The change is not applied by a partition reboot. Be sure to power-off the partition and power it on again.

7.7 Notes to Observe when Configuring Partitions

The XDI is composed of two chip sets (GDX #0 and GDX #1), and the chip sets are connected by four internal ports. These internal ports are referred to as Other Ports. When a partition configuration contains Other Ports, the partition may stop due to a port failure, because if an Other Port is shared with another partition, a failure in one partition may affect the other partition.

However, the possibility of an Other Port failure is very low.

Figure 7.15 Data communication among SBs, IO Units, and Other Ports

The send and receive relationships among SBs, IO Units, and Other Ports are described in the table below.

Figure 7.5 shows which Other Port is used for data communication between IO Units. The table indicates that if, for example, the partition is configured by combining one of SB #0 to SB #3 and IO Unit #7, Other Port #0 is used.

Table 7.4 Data communication among SBs, IO Units, and Other Ports

		Receive							
		SB #3 IO Unit #5	SB #2 IO Unit #1	SB #1 IO Unit #4	SB #0 IO Unit #0	SB #7 IO Unit #7	SB #6 IO Unit #3	SB #5 IO Unit #6	SB #4 IO Unit #2
Send	SB #0-SB #3 IO Units #0,1,4,5					Other Port #0	Other Port #1	Other Port #2	Other Port #3
	SB #4-SB #7 IO Units #2,3,6,7	Other Port #0	Other Port #1	Other Port #2	Other Port #3				

7.8 Notes on Use of Mixed Mirror Modes

If a cabinet contains a Standard Mirror Mode partition that spans other ports and an Extended Mirror Mode partition that spans other ports and if the Extended Mirror Mode partition runs in a single data bus system, the Standard Mirror Mode partition cannot be restarted (reset and power-on).

To restart the Standard Mirror Mode partition in such a case, it is necessary to turn off all power supplies to the Standard Mirror Mode partition and other partitions.

Also, the Extended Mirror Mode partition can keep on running in a single data bus system. If an instruction is issued to turn on or reset a partition that cannot be restarted, a warning message is output and the partition cannot be powered on or reset.

CHAPTER 8 Clustering

This chapter provides an overview of clustering.

To perform the cluster operation in Linux, the clustering software PRIMECLUSTER is required. For details of PRIMECLUSTER, see the relevant PRIMECLUSTER manuals.

In Windows Server 2003, clustering is supported by Microsoft Cluster Service (MSCS), which is a standard function of the operating system. For details on Microsoft Cluster Service (MSCS) provided in Windows Server 2003, see the MSCS manuals.

- The basics of clustering
- Cluster Functions Provided in Linux Environments
- Cluster Functions Provided in Windows Environments
- Points to be Considered at Cluster System Installation

8.1 The basics of clustering

Clustering is the grouping of multiple systems, each operating independently, to have them function as one system. Each of the systems forming a cluster is called a node. Even if one node fails, another node takes over the processing, thereby enabling job processing to continue.

Clustering can increase the availability of the entire system.

Availability is an index for the reliability of a system. A system in which few failures occur, which continues operation even if a partial error occurs, and which can be restored to its original state quickly following the occurrence of a system failure is called a system having high availability.

By implementing auto detection of hardware failures and errors in middleware and applications, and autonomous control such as taking over of job processing, job processing can be continued securely.

During a scheduled stop for periodical maintenance or to change the system configuration, the work can be done without affecting operating resources.

In PRIMEQUEST series machines, one cabinet can be divided into multiple logical servers by the partition function so as to implement a cluster system, where each partition acts as one node.

For internode communication of management information in a PRIMEQUEST cluster system, the GSWB and LAN are duplicated and used for error monitoring in each node. Because an error during communication to each node and an error in the MMB may otherwise cause the entire system to stop, the management LAN and MMB are duplicated to increase the reliability of the system. TCP/IP is used as the communication protocol to prevent data delay and loss when the amount of information on the communication path between an active node and a standby node increases temporarily.

Cluster operation modes are described below.

1:1 standby

A standby node is prepared for the active node. This configuration features a high redundancy where operation can continue even if the active node fails.

Figure 8.1 Example for 1:1 standby

N:1 standby

One standby node is prepared for multiple active nodes. Since only one standby node is required for multiple active nodes, a cluster system with good cost-performance can be constructed.

Figure 8.2 Example for N:1 standby

Mutual standby

Each active node also operates as a standby node for another node. If an error occurs in one node, the other node takes over job processing. During normal operation, all node can be used, but when a malfunction occurs, the job processing for two nodes needs to be performed on a single node. Therefore, the node needs to be constructed in such a way that a single node can perform the processing for the jobs on both nodes.

Figure 8.3 Example for mutual standby

Cascade

Multiple standby nodes are prepared for one active node. When one node stops, the remaining nodes provide a redundant configuration, thereby maintaining availability during periodical maintenance tasks, etc.

Figure 8.4 Example for a cascade

8.2 Cluster Functions Provided in Linux Environments

8.2.1 Cluster Configurations

Two types of PRIMEQUEST cluster configuration are available: Intra-cabinet cluster and inter-cabinet cluster.

8.2.1.1 Inter-cabinet cluster configuration

For the inter-cabinet cluster configuration, the cabinet is treated as one node or multiple nodes to construct a cluster system.

Because the same operation is performed in another cabinet, the degree of redundancy is high compared with the intra-cluster configuration, thereby increasing reliability.

The following example shows a connection of shared disk units by fibre channel. The system disk is connected to two IO Units and mirroring is applied by the PRIMECLUSTER Global Disk Services (GDS). Two separately mounted IO Units increase the resilience against failures.

Figure 8.5 Inter-cabinet cluster system configuration

8.2.1.2 Intra-cabinet cluster configuration

In the intra-cabinet cluster system configuration, one partition is treated as one node and a cluster system is constructed among partitions.

By treating one partition as one node, a cluster system can be configured in one cabinet (unit).

The following example shows a connection of shared disk units by fibre channel. The system disk is connected to two IO Units and mirroring is applied by the PRIMECLUSTER Global Disk Services (GDS). Two separately mounted IO Units increase the resilience against failures.

Inter-system paths are connected via the GSWB. In this configuration, Fujitsu recommends disabling the spanning tree protocol (STP).

Figure 8.6 Intra-cabinet cluster system configuration

8.2.2 PRIMECLUSTER

PRIMECLUSTER is a software product for constructing a cluster system.

PRIMECLUSTER virtualizes system-configuring elements such as servers, storage, networks, and middleware, and enables job processing to be continued autonomously by error detection, failover (job takeover), and degradation functions to implement high reliability of the entire system.

Even if a failure occurs in an operating partition or in the cabinet, high availability can be implemented by having a standby partition and –cabinet take over job processing.

Machines of the PRIMEQUEST series perform communication between PRIMECLUSTER and the MMB to provide the following functions:

- System state monitoring
  This function displays the state of a specified node and reports it to the MMB.
- System state notification
  This function reports changes in the operation state of the local node to another node.
  1. OS-detected events such as OS running and shutdown are reported to the MMB.
  2. When a change in the partition state is detected, the MMB sends a trap to an associated node registered in advance.
  3. A trap is sent to both of the duplicated management LANs.
- Event reception from the other system (in the pair active server ↔ standby server)
  This function receives information about changes in the state of the other node.
  This function receives information from the other MMB and filters duplicated trap data.
- Instruction to the other system (in the pair active server ↔ standby server)
  This function sends a Panic instruction and a Reset instruction to a specified other node.
  1. A control instruction from PRIMECLUSTER is reported to the MMB of the other system.
  2. The MMB controls the partition through a private LAN.

8.3 Cluster Functions Provided in Windows Environments

Windows Server 2003 supports cluster functions as standard features of the operating system. This section explains Microsoft Cluster Service (MSCS), which is provided by Windows Server 2003.

A cluster system consisting of up to eight partitions (*1) can be configured with Windows Server 2003. Each server (partition) in a cluster is called a node. When Windows is used with PRIMEQUEST, the most basic cluster system is a two-node inter-partition cluster using a shared disk. This section explains the inter-partition cluster as the basic configuration.

*1: The partitions referred to in this section may be contained in a single or multiple PRIMEQUEST cabinets.

Note:

For the configuration other than a two-node inter-partition cluster using a shared disk, contact your Fujitsu certified service engineer.

8.3.1 Basic cluster configuration

The following figure shows a typical two-node cluster configuration.

Figure 8.7 Basic cluster configuration (Windows)

- Business application software
  Business application software must be installed on each node used for business operations and on each standby node.

- Shared disk
  The data used by the business application software is stored on a shared disk unit so that it can be accessed and used from any node. However, note that the data cannot be accessed simultaneously from multiple nodes.
- Network configuration
  The nodes that belong to a cluster are connected via a dedicated LAN to ensure that they can monitor errors encountered by any of the other nodes and that data consistency can be maintained among the nodes. This network is called a private network. In contrast, the LAN used for business operations is called a public network.

8.3.2 Virtual server

A virtual IP address instead of the IP address of a node is used as the IP address for access to business application software, because it is not affected by business application operations on any node. The server whose virtual network name is associated with the virtual IP address is called a virtual server. The virtual server can be referenced from clients as a Windows network computer.

As a rule, clients access the application software via a virtual server. During switching from a node on which business application software is running, the corresponding session is disconnected once. However, the session can be resumed by reestablishing the connection to the same virtual server after the business application software restarts.

Because of this mechanism, clients in a cluster system need not consider the particular node on which business application software is running.

Figure 8.8 Virtual server

8.3.3 Scope of remedies

Remedies can be found for hardware (e.g., CPUs, motherboards, and LAN cards) failures and software failures (e.g., application errors). Note that a server cluster is not a fault tolerant (FT) system. If a business application failover occurs, operation stops for a short time.

8.3.4 Notes on construction

MSCS construction requires a separate environment for the Windows Server 2003 directory service (Active Directory). You can use the cluster construction service for the Active Directory and MSCS construction.

8.4 Points to be Considered at Cluster System Installation

This section discusses points to be considered before determining a cluster configuration. The cluster configuration should be determined after the requirements of the individual systems are considered.

Interconnect (network connecting nodes) disposition

In the PRIMEQUEST cluster configuration, all nodes must be connected by a dedicated LAN. The network connecting nodes is called interconnect. Multiplexing the interconnect increases the availability of the system.
The PRIMEQUEST interconnect should be connected to either of the following types of NIC card:

- NIC card added to the PCI slot in an IO Unit
- NIC card added to the PCI slot in the PCI_Box

Remarks

1: Never use the management LAN for interconnect communication because it is used only for controlling the PRIMEQUEST system.
2: Never use a LAN connected to the GSWB through the NIC card built in an IO Unit for interconnect communication because it is used for business communication.

The following figure shows a configuration in which NIC cards are added to the PCI slot in the IO Unit.

Figure 8.9 Position of the interconnect

Remarks: Do not use the spanning tree protocol (STP) for a switch to be used as interconnect. Otherwise, the system will enter the non-communication state at reconfiguration of a link, thereby disabling interconnect communication.

System Mirror Mode and clustering

PRIMEQUEST series machines support two system mirror modes: Extended Mirror Mode and Standard Mirror Mode. By way of Extended Mirror Mode, in which memory, crossbars, and chip set contents are divided into two systems which always perform the same operation simultaneously, a PRIMEQUEST server can continue the current job without interruption even when one of the systems fails. In Standard Mirror Mode, address bus crossbars are duplicated and always perform the same operation simultaneously. Operation in System Mirror Mode provides higher availability.

Thus the PRIMEQUEST series increases the availability of hardware by operation in System Mirror Mode. However, System Mirror Mode does not support software redundancy.

Installation of a cluster system can increase the availability of the entire system also as far as software is concerned.

By constructing a system using both System Mirror Mode and clustering functions, the availability of the system can be improved even further.

Figure 8.10 Extended Mirror Mode and cluster configuration

Cluster operation by using a flexible I/O component and reserved SB

Using a flexible I/O component and reserved SBs in a cluster configuration can shorten the time required for restoration to normal cluster operation even if a failure occurs in an SB, CPU, or memory.

Figure 8.11 shows an example for using a flexible I/O component in an intra-cabinet cluster configuration. A cluster switch occurs due to a failure in SB #3 of partition #1 (the active node) and partition #1 becomes the standby node. Partition #1 is stopped; it can be placed in a state in which partition #1 can operate as the standby node by disconnection of SB #3 and incorporation of SB #4 after a restart.

Figure 8.11 Example of using flexible operation in an intra-cabinet cluster configuration

Processing at failure occurrence

- Clustering
  If a failure occurs in the active node, the cluster system recognizes the failure and the standby node takes over job processing. The standby node must check the consistency of the taken over data. During this period of time, job processing is stopped temporarily.
  If a failure is detected in the active node, PRIMEQUEST system starts switching from the active node before the kernel stops to shorten the switching time.
  If a failure occurs in the active node, collection of active node dump and system switching to the standby node are executed at the same time to shorten the time necessary for taking over job processing.
- System Mirror Mode
  If an error occurs in hardware during operatzion in System Mirror Mode, switching time is not required. This is because the faulty location is simply disconnected. However, if an error is detected in software, this approach is not supported.
- System Mirror Mode and clustering
  If an error occurs in hardware in System Mirror Mode during cluster operation, System Mirror Mode disconnects the faulty location to continue the operation. If another failure occurs during operation with the faulty location disconnected, the cluster system recognizes the failure and the standby node takes over job processing.

CHAPTER 9 Backup and Restoration

This chapter describes methods of backing up and restoring system volumes in a partition with Linux (Red Hat) or Windows installed. For details of SUSE backup/ restore methods, see the SUSE manual.

9.1 Importance of Backing Up Data

PRIMEQUEST uses components and hardware with high reliability and duplicates many units to maintain high reliability. To be prepared for failure occurrences, however, be sure to back up data periodically.

If a system is damaged due to some problem, or if data on a server is deleted by an operator error and the like, the data on the PRIMEQUEST system needs to have been backed up to recover the data in the server to its previous status.

As long as data is backed up, if data in a hard disk drive is damaged due to a hardware failure, an operator error, and the like, the user can restore the backup data and recover normal system operation. If the data is not backed up, it will be lost.

Be sure to back up data periodically to operate a system with confidence.

9.2 General Methods of Backup and Restoration

The following four methods of backing up and restoring a system are available:

- Using the standard utilities included in an OS
  Linux (Red Hat): dump/restore
  Windows: NTBackup
- Using cloning software
  Common to Linux (Red Hat) and Windows: SystemcastWizard Lite
- Using snapshot software
  Linux (RedHat): PRIMECLUSTER GDS Snapshot
- Using backup software
  Linux (Red Hat): ETERNUS SF ACM, VERITAS NetBackup, NetVault, NetWorker

Ask the sales representative in charge for details including the versions of the supported operating systems.

The methods above include a method that uses a backup server with a backup unit. Consult with the sales representative in charge, and prepare backup software and a backup server.

An overview of each backup and restoration method is provided below.

9.2.1 OS-standard utility (Linux (Red Hat): dump/restore, Windows: NTBackup)

There are two types of file backups and restoration. One is backup on a relatively small scale, such as backing up or restoring a specific file and directory, and the other is backup on a large scale, such as backing up the entire system volume.

This section explains the backup and restoration of the entire system volume by using the dump and restore commands, which are Linux (Red Hat) standard utilities, and the backup and restoration of files by using NTBackup, which is a Windows standard utility.

Configuration example when using an OS-standard utility

- Local backup

  For Linux (Red Hat):

  Prepare an external unit to which backup data is to be saved for each partition. Enter the dump or restore command from the OS console of the partition (Partition#0 in Figure 9.1) to be backed up or restored. A serial console or KVM console can be used as the OS console.

  To perform restoration, the system needs to be started in rescue mode by using the installation CD.

  For Windows:

  For the backup or restore of a data area of a Windows system, NTBackup is used. The backup can be performed either from the GUI screen when the KVM console has been connected or with the backup command from the command prompt screen.

  However, the restoration can be performed only from the GUI screen. Files cannot be restored with a backup command from the command prompt screen.

Figure 9.1 Local backup using a standard utility (when the KVM console is used)

- Backup to a remote host

  For Linux (Red Hat):

  Prepare a server with a backup unit,and back up or restore data using ssh.

  Using the LAN, back up or restore the backup/restoration target partition (Partition#0 in Figure 9.2) to a tape unit in a server (remote host) equipped with a backup unit.

  Perform the operation from the OS console of the partition to be backed up or restored (Partition #0). A serial console or KVM console can be used as the OS console.

  When this method is used, the network needs to have been activated. For the data transmission line, Fujitsu recommends using a LAN exclusively prepared for backup purposes. It is also possible to use the management LAN.

  Backup is performed by outputting data using the standard output as the output destination of the dump command, and copying (dd) the result to a tape unit of a remote host after logging in remotely using the ssh command.

  Restoration is performed by logging in remotely by using the ssh command, copying (dd) the tape unit data to standard output, and then expanding the contents of standard output by using the restore command. At this time, the system needs to be started in rescue mode by using the installation CD.

  A server outside the cabinet can also be used as the remote host.

  For Windows:

  If data in Partition #0 is backed up, it is necessary to use NTBackup on the remote host side (Partition #n) for the backup and restoration work.

First, perform remote-login to Partition #n from Partition #0 that is connected to the KVM console by using remote desktop connection. Backup and restoration can be performed after assigning the Partition #0 as a network drive.

However, the server to be backed up and the remote host must be placed in the same network.

Figure 9.2 Backup to the remote host by using a standard utility (when using the KVM console)

9.2.2 Cloning software (SystemcastWizard Lite)

A cloning software program (SystemcastWizard Lite) is a tool for easily installing a system on multiple computers and performing maintenance in a short time.

SystemWizard Lite is a software program that performs OS installation, backs up hard disk data, and restores hard disk data via a network.

This software can be used for backup at the system construction or verification stage and also for recovery if a problem occurs. The PRIMEQUEST network activation function can be used to facilitate recovery of entire system volumes, even from a state in which the OS cannot start.

Configuration example when using SystemcastWizard Lite (common to Linux: Red Hat and Windows)

Figure 9.3 Backup using SystemcastWizard Lite

- Terminal for an MMB console
  A Web browser must be installed. It can be used on unit (b) deployment server console.
- Deployment server console
  This is a server or PC with Windows OS installed. SystemcastWizard Lite must be installed on the server, too. SystemcastWizard Lite is included in the supplied software. For details, see the *PRIMEQUEST SystemcastWizard Lite User's Guide* (C122-E010EN).

Notes:

- Since deployment server console has an interface for direct exchanging information with an MMB, it needs to connect to the management LAN. Place deployment server console in the same segment as the PRIMEQUEST system. ((a) terminal for an MMB console can be used also when placed in another segment.)
- SystemcastWizard Lite uses MMB#0 for communication concerning remote startup, backup, or restoration of a target partition. Therefore, make sure that MMB#0 is connected to the management LAN. However, MMB#0 need not be active.
- For installing the OS using SystemcastWizard Lite, specify [PCI Bus Mode] for the PCI address mode of the partition (PCI Address Mode (setting)) in the [Mode] window of the MMB Web-UI that is displayed by selecting [Partition].
- If SystemcastWizard Lite is used for backup and restoration, up to only four partitions can be used in parallel.

9.2.3 Snapshot software (PRIMECLUSTER GDS Snapshot)

By using PRIMECLUSTER GDS Snapshot, a snapshot (copy of data at a certain point in time) of a logical volume managed by PRIMECLUSTER GDS can be collected and stored in a logical backup volume.

By using the snapshot function, the amount of time that business is suspended for a backup or restore operation can be reduced.

Configuration example when PRIMECLUSTER GDS Snapshot is used (Linux: Red Hat)

- Backup of system volume (internal disk)
  If data on the system volume is destroyed, the backup destination disk can be used as a new system volume. It is also possible to restore data to the original system disk by starting the system from the backup destination disk.

Note: A backup copy can be output also to a disk on the ETERNUS system. However, since booting from a backup volume on the ETERNUS system is not supported, restoration cannot be performed from a backup volume on the ETERNUS system.

Figure 9.4 System volume backup using PRIMECLUSTER GDS Snapshot

- Backup of data area
  - Whether online backup is enabled or not depends on the data management software such as database management software or the file system. Generally, business operation needs to be stopped temporarily when a snapshot is created.
  - Backup from the backup destination disk to tape is enabled during business operation. For this purpose, however, backup software that supports backup from the GDS volume to tape is required separately.
  - The restoration destination volume can be used immediately after the start of restoration.

Figure 9.5 Data area backup using PRIMECLUSTER GDS Snapshot

9.2.4 Backup software (ETERNUS SF ACM, VERITAS NetBackup, NetVault, NetWorker)

To design a schedule for backup in conformity with basic business operations, and to back up data in a database operation in cooperation with database software, use a backup-dedicated software program.

Backup software programs for the PRIMEQUEST series include ETERNUS SF ACM, VERITAS NetBackup, NetVault, NetWorker, and others. The characteristics and configuration examples of each software programs are described below.

Configuration example when using ETERNUS SF AdvancedCopy Manager (Linux: Red Hat)

ETERNUS SF AdvancedCopy Manager (ACM), software for backup purposes, is used in combination with the ETERNUS8000/ETERNUS4000/ETERNUS2000/ETERNUS6000/ETERNUS3000 disk array unit. Data of a disk array unit can be backed up or restored by using this software without passing through the business server or LAN.

Figure 9.6 Backup of data on the ETERNUS by using ACM

- ACM Manager
  Comprehensively manages the device information and policy of all the agents.
- ACM Agent
  Controls the AdvancedCopy function of the ETERNUS. ACM Agent is installed in the business server that is a backup restoration target.
- ACM Tape Server
  Controls the tape library and the backup to and restoration from tape media. ACM Tape Server is installed on the Solaris server and the tape library is connected to the server.

Configuration example when VERITAS NetBackup is used (Linux: Red Hat)

- Normal backup of data
  Use the LAN or SAN, back up client data to a tape unit connected to a backup server. As a backup destination, the user can specify a disk area connected to the backup server.

Figure 9.7 Data backup using VERITAS NetBackup

- Online backup of a DB (Oracle)
  For Oracle, Recovery Manager (RMAN) provided by Oracle becomes the interface to the DB. VERITAS NetBackup Oracle Agent is linked to RMAN, and data is passed from Oracle Agent to VERITAS NetBackup.

Figure 9.8 Online backup of a DB using VERITAS NetBackup

- Backup of a system volume
  This operation is not supported.

Configuration example when NetVault is used (Linux: Red Hat)

- Normal backup of data
  The configuration is the same as the case where VERITAS NetBackup is used. The user can use the GUI on a NetVault server terminal or a NetVault remote management terminal via a LAN to back up or restore data. To perform backup of multiple terminals at the same time, a NetVault server disk, as a backup destination, should be a virtual tape library.
- Online backup of a DB (Oracle)
  The configuration is the same as the case where VERITAS NetBackup is used. To perform online backup of a database by NetVault GUI, use the NetVault application plug-in module (APM) to seamlessly integrate the online backup API provided by Oracle and NetVault.
- Backup of a system volume
  Not supported.

Configuration example when NetWorker is used (Linux: Red Hat)

- Normal backup of data

  The configuration is the same as when VERITAS NetBackup is used.
  Using the LAN or SAN, back up client data to a tape unit connected to a backup server. As a backup destination, the user can also specify a disk area connected to the backup server. Data backed up on a disk can be restored at high speed, enabling quick completion of staging, cloning, and other operations.
  If the user purchases an optional product (Dedicated Storage Node), the user can back up or restore data to tape without the intervention of a server in the SAN environment.

Figure 9.9 Data backup using NetWorker

- Online backup of a DB (Oracle)

  The configuration is the same as when VERITAS NetBackup is used.
  For Oracle, the API for online backup is linked to NetWorker Module for Oracle, and data is passed from Oracle Agent to the NetWorker server.
  Oracle databases can be backed up with the dedicated storage node and NetWorker Module for Oracle to enable linkage between the online backup APIs and NetWorker Module for Oracle.

- Backup of system volume

  This operation is not supported.

9.3 Comparison of Backup/Restoration Methods

Table 9.1 lists the usage method and characteristics of various backup and restoration methods. The requirements for the backup methods vary with the operation. The methods appropriate to each requirement need to be combined to back up data.

Table 9.1 Comparison of backup methods

Method		Main usage	Characteristics
OS-standard	dump/restore (Linux: Red Hat) NTBackup (Windows)	Backup of a system volume (single)	• It is possible to back up and restore a system volume. • These are OS-standard functions. The user is normally accustomed to their operation.
Cloning software	Systemcast Wizard Lite (Linux: Red Hat/ Windows)	Backup of a system volume (batch)	• It is possible to back up and restore a system volume. • Automatic processing is possible in cooperation with a MMB via a network. • It is possible to perform concurrent parallel backup and restoration of multiple partitions.
Snapshot software	PRIMECLUSTER GDS Snapshot (Linux: Red Hat)	Backup of GDS logical volume (system volume and data area)	• It is possible to perform online backup and restoration of a system volume and data area. • For a data area on an ETERNUS disk array, it is possible to perform backup and restoration without applying a load to the PRIMEQUEST CPU and LAN. Remarks: Setting and operation for GDS mirror configuration are not necessary • No backup server needs to be prepared. • This software has high affinity with the PRIMECLUSTER and PRIMECLUSTER GDS.

Method		Main usage	Characteristics
Backup software Backup software	ETERNUS SF ACM (Linux: Red Hat)	General data stored in the ETERNUS disk array unit (backup of a system volume by booting SAN and backup of data areas)	• It is possible to perform backup without applying load to a LAN. • It is possible to perform backup during DB (Oracle/Symfoware) operation. • It is possible to back up and restore a system volume while booting from a SAN. • PRIMECLUSTER is supported.
	VERITAS NetBackup (Linux: Red Hat)	Backup of data areas (large scale DB)	• It is possible to perform backup or restoration during DB (Oracle) operation. • NetBackup has a track record for DB backup of large-scale systems. • It is possible to manage a mixed platform environment by using one server. • PRIMECLUSTER is supported.
	NetVault (Linux: Red Hat)	Backup of data areas	• It is possible to perform backup during DB (Oracle) operation. • Setup for a small-scale system is possible at low cost. • Operation with a Japanese GUI is possible. • PRIMECLUSTER is supported.
	NetWorker (Linux: Red Hat)	Backup of data areas	• It is possible to perform backup during DB (Oracle) operation. • It is possible to perform backup or restoration of a wide range of environments from a small-scale system to a large-scale system. • It is possible to perform concurrent parallel backup/restoration of multiple partitions. • It is possible to manage a mixed platform environment by using one server.

CHAPTER 10 Considerations for Maintenance

This chapter describes the following items that are related to the system maintenance of a partition with Linux (Red Hat) installed, and that the user must take into consideration when designing a system.

For details of a partition with SUSE installed, see the SUSE manual.

- Estimating the Dump-use Area
- Estimating the Dump-use Memory

10.1 Estimating the Dump-use Area

In a PRIMEQUEST system using Linux (Red Hat), the estimate of the dump-use area varies depending on the version of the operating system. This section explains the following estimation methods:

- RHEL-AS4 (IPF)
- RHEL5 (IPF)

10.1.1 RHEL-AS4 (IPF)

If using RHEL-AS4 (IPF) in PRIMEQUEST systems, the dump function provided by Linux distribution (diskdump) is available.

This section explains how to estimate the disk spaces necessary for the use of this dump function. The explanation is provided with respect to the following items in the order given:

- Area used to collect dump data
- Area used to save dump data

Area used to collect dump data

When a system becomes faulty, a disk partition (dump device) area exclusively used for dump collection is required.

A dump device is a device on which the dump function temporarily stores information on problem occurrences. Table 10.1, "Areas used to collect dump data" lists the sizes of the areas required for a dump device.

Table 10.1 Areas used to collect dump data

No.	Function name	Size	Remarks
1	diskdump	Amount of installed memory *1 + 512 MB	

*1 When Extended Mirror Mode is used, the Linux system can only recognize one half of the amount of the actually installed memory. The amount of memory above indicates the amount of memory that Linux can recognize.

The capacity of the dump device needs to be expanded according to the amount of memory installed. Therefore, if the user plans to expand the amount of installed memory, estimate the size of the dump device based on the amount of installed memory after expansion.

When the amount of installed memory is large, the partial dump function or compression function can be used to reduce the disk space requirements for the dump function. The partial dump function can collect data from only the areas used for kernel control. Although the function can reduce the disk space requirements, it may take a little more time for troubleshooting upon the occurrence of a kernel error or system failure.

A dump device can be allocated either on an internal disk or on an external disk. Table 10.2, "Devices that can be used as a dump device" lists the devices that can be used as a dump device. Allocate a dump device on a non-duplicated disk.

Table 10.2 Devices that can be used as a dump device

No.	Type	Driver	Remarks
1	Internal disk	LSILogic SCSI driver	Disk not managed by PRIMECLUSTER GDS
2	External disk	Emulex FC driver	

As the capacity of internal disks is limited, the user may be unable to construct a system using only internal disks. In such a case, prepare a dump device on an external disk. Note that a dump device prepared on an external disk cannot be shared by multiple clusters.

Area used to save dump data

When the system is restarted, the dump information that has been collected by diskdump on a dump device is saved after it is automatically converted to a format that a dump tool can reference. Therefore, an area with the size indicated below must be created for each type of dump information to be saved. The dump saving areas must be in /var/crash.

Area size required per saved dump = Amount of memory installed + 1 GB

When creating a dump saving area, such as /var/crash, as a dedicated disk partition, estimate the area size and increase the estimated value by 10% in consideration of the file system management area.

10.1.2 RHEL5 (IPF)

If RHEL5 (IPF) is used in the PRIMEQUEST system, the dump function provided by the Linux distribution (kdump) is available.

This section explains how to estimate the amounts of disk space necessary for the use of these dump functions. The explanation is provided with respect to the following items in the order given:

- Area used to save dump data

Area used to save dump data

At the time the system becomes faulty, dump information is directly saved in this area by kdump. Therefore, an area with the size indicated below must be created for each type of dump information to be saved.

Area size required per saved dump = Amount of memory installed + 1 GB

The areas must be created on a disk not managed by PRIMECLUSTER GDS.

When creating a dump saving area as a dedicated disk partition, estimate the area size and increase the estimated value by 10% in consideration of the file system management area.

10.2 Estimating the Dump-use Memory

In a PRIMEQUEST system using Linux (Red Hat), RHEL5 (IPF), if used, affects the design of memory used for dump functions.

10.2.1 RHEL5 (IPF)

kdump starts a kernel (OS) for kdump separately from the kernel that is run to collect a dump. For this reason, the memory required for running the kernel for kdump is reserved in advance. This reserved memory cannot be used for an ordinary operation. The size of this memory is fixed at 512 MB in PRIMEQUEST. Take this value into consideration in memory design.

Appendix A Software Supplied with PRIMEQUEST Hardware

The software programs supplied with PRIMEQUEST hardware are listed below.

Table A.1 Software programs supplied with PRIMEQUEST hardware

No.	Name	Function	EFI	Target OS		
				Linux (RedHat)	Linux (SUSE)	Windows
1	EFI/BIOS	Firmware	Y	Y	Y	Y
2	PSA	Hardware system management	-	Y	Y	Y
3	SIRMS	Collecting software configuration information	-	Y (*1)	-	Y (*1)
SCSI Related						
4	Driver	Driver	Y	Y	Y	Y
5	Fusion MPT EFI appl.	F/W update tool	Y	-	-	-
6	Fusion MPT management tool	CIMS Browser	-	Y	Y	Y
Broadcom LAN Related						
7	Driver	Driver	-	Y	-	Y
8	BCM EFI appl.		Y	-	-	-
9	BACS(Win)	VLAN, aggregation, etc.	-	-	-	Y
Intel LAN Related						
10	Driver	Driver	-	Y	Y	Y
11	Intel PROSet (Win)	VLAN, AFT/ALB, etc.	-	-	-	Y
Neterion LAN Related						
12	Driver	Driver	-	Y	-	Y
13	Xframe Control Panel (Win)	VLAN, Team, etc.	-	-	-	Y
Emulex FC Related						
14	Driver	Driver	Y	Y	Y	Y
15	lputil		-	Y	Y	Y
16	HBAnyware		-	Y	Y	Y
Others						
17	HBA block driver for PCL		-	Y	-	-
18	Log trace capture	Collecting logs (compatible with extended drivers)	-	Y (*1)	-	-

No.	Name	Function	EFI	Target OS		
				Linux (RedHat)	Linux (SUSE)	Windows
19	Installation Support Tool	Executed on an external PC (Windows) and creates a Config file.	-	Y	Y	-
20	Installation Support Tool for Windows	Tool for creating an installation floppy disk for automatically installing Windows	-	-	-	Y
21	Bundled-Software Package Installer	Installation tool (bundled software)	-	Y	Y	-
22	DSNAP	Collects the basic Windows OS information (which is the basis of OS operation) through command operation.	-	-	-	Y
23	High-reliability tool package installer	Installation tool (bundled software)	-	-	-	Y
24	System Parameter Check Tool	Checking the environment variable (argument necessary for OS operation) that the user is going to set	-	Y	-	-
25	System Data Output Tool (fjsnap)	Executing the command for collecting the information required for support under OS operation.	-	Y	Y	-
26	SystemcastWizard Lite	• Remote installation • Backup or restoration of a disk	-	Y	Y	Y
27	SNMP MIB	MIB	-	Y	Y	Y
28	Software Support Guide	Collects the basic OS information required for support (which is the basis of OS operation) through command operation.	-	-	-	Y
29	HRM/server	Maintenance support tool	-	Y	-	Y

Y: Supported

*1 This is supported only in Japan.

*2 This is supported only in Japan and the Asian region.

Glossary

ACS (AC Section)

AC power input section

ASIC (Application Specific Integrated Circuit)

Integrated circuit (IC) designed and manufactured for specific applications

API (Application Program Interface)

A set of instructions and functions used for developing operating systems and middleware

BIOS (Basic Input Output System)

Part of the operating system (OS) function. The BIOS is the system that controls input/output to devices. For the PRIMEQUEST-series machine, BIOS is a general term for PAL, SAL, and EFI.

BMC (Baseboard Management Controller)

The BMC is a system management controller that continuously monitors the system for serious hardware errors and notifies the OS of such errors.

BMM (BMC Module)

Board on which legacy I/O ports such as BMC, VGA, USB, and COM ports are mounted

BP (Backplane)

The backplane is connected to SBs, IO Units, and other devices. Together with the XAI and XDI, it constitutes the memory and I/O interconnect (crossbar).

Business LAN

LAN used to configure a user business system

CLI (Command Line Interface)

This interface with UNIX or DOS allows the user to enter commands and optional arguments to communicate with the OS.

CoA

Abbreviation for Certificate of Authenticity. This is a visual identifier that helps identify genuine Microsoft software and components.

COM Port (Communication Port)

RS-232C serial port for PC/AT compatible machines. A COM port is also called an "RS-232C port." Most PC/AT compatible machines each have two COM ports, which are often used to connect a modem, terminal adapter, or scanner. Most of these ports use D-Sub 25-pin or D-Sub 9-pin connectors. A COM port enables an interface with the OS.

CPCB (Clock and PCI_Box Control Board)

Board equipped with a system clock oscillator and a PCI_Box control interface

Crossbar

This concept covers the address crossbar and data crossbar that transfer data and control the data transfer between SBs and IO Units. Memory and I/O interconnect has the same meaning as crossbar.

DDR2 (Double Data Rate 2)

Standards for the next generation of memory that operates at higher speeds and consumes less power than conventional DDR memory

DIMM (Dual Inline Memory Module)

This compact memory module has pins on both sides and is mainly used in notebook PCs.

DP

Abbreviation for Dynamic Partitioning. This is a function for dynamically adding, deleting, or replacing CPU or memory resources as well as PCI cards and onboard I/O units in a partition while the operating system is running in the partition.

DVD-ROM (Digital Versatile Disc-Read-Only Memory)

Digital format for high-volume storage of data on optical disks

ECC (Error Checking Correction)

Error correction code or a method of using the error correction code to check for and correct errors

EFI (Extensible Firmware Interface)

Specifications for an interface between an OS and firmware. Instead of the BIOS, EFI is used for hardware control.

FC (Fibre Channel)

One of the serial interface standards. The Fibre Channel standard uses fiber cables as the transmission medium.

Firmware

Built-in software for basic hardware control

FWH (Firmware Hub)

LSI device from Intel Corporation. FWH is flash memory that stores SAL (BIOS). The PRIMEQUEST-series machine uses two types of FWH: one type is mounted on an SB and the other is mounted in an IO Unit.

GAC (Global Address Controller)

One of the ASICs developed by Fujitsu for the PRIMEQUEST-series machine

GbE (Gigabit Ethernet)

Ethernet standards for high-speed communication of up to 1 Gbps

GDS

Abbreviation for PRIMECLUSTER GDS

GDX (Global Data Xbar)

One of the ASICs developed by Fujitsu for the PRIMEQUEST-series machine

GLS

Abbreviation for PRIMECLUSTER GLS

GSWB (Gigabit Switch Board)

Board with a switching hub function and a connector that receives Gigabit Ethernet interface output from an IO Unit via the BP and outputs it to a destination outside the cabinet

GTHB (Gigabit Through Board)

Board with connectors that receive Gigabit Ethernet interface output from an IO Unit via the BP and output it without change to a destination outside the cabinet

HBA

Abbreviation for a host bus adapter

HDD (Hard Disk Drive)

Device that reads a hard disk. HDD may also be an abbreviation for the hard disk itself.

Hot Plug

Method of replacing components while power is on

HTTP (Hypertext Transfer Protocol)

Protocol used by Web servers and clients for data transmission

I2C (Inter Integrated Circuit)

Protocol used for high-speed communication between integrated circuits (ICs)

IA (Intel Architecture)

Generic term for the basic design (architecture) of Intel's microprocessors

IFT (Instruction Fetch)

Mechanism for reading instructions stored in memory

IHV (Independent Hardware Vendor)

This hardware provider has no special relationship with a particular hardware or OS maker.

IO Unit

Input/Output control unit that contains PCI card slot, HDD, SCSI controller, GbE controller, and other I/O interfaces

IP (Internet Protocol Address)

Identification number assigned to each computer connected to an IP network, such as the Internet and intranets

IPMI (Intelligent Platform Management Interface)

Standardized interface specifications established so that SNMP and server management software can monitor server hardware independently of specific hardware systems and OSs

ISV (Independent Software Vendor)

This application software provider has no special relationship with a particular hardware or OS maker.

KVM (KVM interface unit)

Unit used to select the USB interface for keyboard and mouse, or the VGA interface, for external output from a partition

LAN (Local Area Network)

Using optical fiber, for example, this network allows data to be transferred among computers and printers connected in a facility.

LDAP (Lightweight Directory Access Protocol)

Protocol used to access directory databases in a TCP/IP network, such as the Internet and intranets

LDX (Local Data Xbar)

One of the ASICs developed by Fujitsu for the PRIMEQUEST-series machine

LED

Abbreviation for a light emitting diode

MAC address (Media Access Control Address)

Unique address assigned to each network interface device, switch, or router mounted on a network interface card (NIC) or motherboard

Management LAN

This LAN connects the MMB to partitions and to LANs outside the cabinet so that the PRIMEQUEST system can be managed.

MIB (Management Information Base)

Information released by a network device managed by SNMP in order to post the device status to an external destination

Middleware

Software that runs under an OS and provides application software with more advanced and detailed functions than the OS. It is positioned between the OS and application software in terms of its characteristics.

MMB (Management Board)

This server management board is a system control unit whose tasks include control and monitoring of cabinet hardware, partition management, and system initialization.

NIC (Network Interface Controller)

Hardware that supports network functions

NTP (Network Time Protocol)

Standard time information protocol used on the Internet. Highly precise time information with consideration of line speeds and load changes in paths can be obtained with this protocol.

PAL (Physical Abstract Layer)

Firmware that provides platform initialization and operating system boot functions

Partition

System equipped with the functions of a processing unit. Each partition contains software resources such as an OS and applications as well as hardware resources such as SBs and IO Units.

PCI_Box

Device used for PCI slot expansion

PCI Hot Plug

Technology that enables PCI cards to be mounted and removed while the system is operating

PCI (Peripheral Component Interconnect)

Bus architecture established by PCI SIG for connecting PC components

PCIU (PCI Unit)

PCI-X card slot expansion unit that is mounted in a PCI_Box

PEXU

PCI Express card slot expansion unit that is mounted in a PCI_Box

Platform

OS type or environment that is the basis for operation of application software

POST (Power-On Self Test)

Hardware test that is automatically run when the computer is powered on

Private LAN

LAN used for internal control, under which firmware programs installed on hardware components communicate with one another. MMB firmware, GSWB firmware, and BMC firmware installed on IO Units can use a private LAN for communication with one another. OSs and applications cannot use a private LAN.

PSA (PRIMEQUEST Server Agent)

Software that performs hardware error monitoring and configuration management over PRIMEQUEST partitions

PSU (Power Supply Unit)

Component that converts AC voltage to DC voltage as a DC power supply

PXE

PXE (Preboot eXecution Environment)

Network boot standard based on BIOS technology that enables remote operation of management tasks such as system start and OS installation/update

RAID (Redundant Array of Independent Disks)

Technology that increases reliability and processing speeds by using multiple hard disks as a single disk

REMCS (Remote Customer Support System)

Fujitsu's remote customer support service

Reserved SB

Standby SB reserved for possible failures

RHEL (Red Hat Enterprise Linux)

Linux distribution released by Red Hat, Inc.

SAF-TE

Abbreviation for a SCSI accessed fault-tolerant enclosure

SAL (System Abstraction Layer)

Firmware that supports processor initialization and error recovery functions

SAN (Storage Area Network)

Dedicated network for connections between a server and storage devices

SB (System Board)

Board on which a CPU and memory are mounted

SCSI (Small Computer System Interface)

Standards for connections between PCs and peripherals. SCSI was established by the American Standards Association.

SDRAM (Synchronous DRAM)

Memory standard for access speeds that are higher than those of DRAM

SEL (System Event Log)

Information on the processing parameters, processing, and processing results logged during hardware and software operations

SERDES (Serializer Deserializer)

Parallel-to-serial converter (from low speeds to high speeds)

SIRMS (Software Product Information Collection for Remote Maintenance Support)

Software that collects configuration information on software installed in PRIMEQUEST partitions

S.M.A.R.T. (Self-Monitoring Analysis Reporting Technology)

Function that enables a hard disk to monitor its own conditions and notify the BIOS of any error detected

SMP (Symmetric Multiple Processor)

Parallel processing system in which all processors work together through common memory resources

SNMP (Simple Network Management Protocol)

TCP/IP-compliant protocol for managing devices in a network

SSL (Secure Sockets Layer)

Protocol under which information is encrypted for transmission. SSL was developed by Netscape Communications Corp.

Systemwalker

One of Fujitsu's middleware products. Systemwalker is integrated operation management software.

Telnet

Protocol or standard method for remote control of computers connected to a TCP/IP network, such as the Internet and intranets

UPS (Uninterruptible Power Supply)

Power supply unit that stores power and protects against possible damage and loss of computer data from a momentary voltage drop or unexpected power failure

USB (Universal Serial Bus)

One of the standards on connecting peripheral devices such as keyboards and mice

VLAN (Virtual LAN)

Function that logically groups the ports of one switching hub so each group works as an independent LAN

Web UI (Web User Interface)

Interface that uses a Web browser for displaying information to users and for user operations

XAI (Xbar Address Interconnect Board)

Board that transfers address information and controls the information transfer between SBs and IO Unit boards

XDI (Xbar Data Interconnect Board)

Board that transfers data and controls the data transfer between SBs and IO Unit boards

Index

A

access control 4-24
ACS 2-36
agent message log. 4-34
Alarm-LED 2-35
alert message, convention for v
allocation PCI_Box to partition 7-16, 7-17
all-partition communication
disconnection mode 4-8

B

backup 3-8
general method of backup and
restoration 9-1
importance of backing up data 9-1
backup software
NetBackup. 9-9
NetVault. 9-9
Softek ACM 9-9
Softek TSM 9-9
backup/restoration method,
comparison of. 9-14
BMC firmware 3-2
BMM. 2-12
business LAN. 4-10
redundancy of business LAN 5-17

C

cabinet 2-6
card cassettes in PEXU 2-46
chipset 1-9
cloning software
SystemcastWizard 9-5
cluster configuration. 8-5
cluster system installation, point to be
considered at 8-11
clustering
processing at failure occurrence 8-15
clustering, basic of 8-1
COM port 2-12
compatible PCI card. 2-14, 2-47
Compatible PCI Express card 2-47
conceptual diagram
PRIMEQUEST conceptual diagram . . 1-3
configuration all hardware
(PRIMEQUEST 580A/580/480). 1-5
connection requirement. 2-49
console redirection 4-3, 4-21
cooling unit 2-39
CPCB 2-33
CPCB: Clock and PCI_Box
Control Board 1-4
CPU. 2-8
crossbar 2-29
custering 1-27

D

data area backup using PRIMECLUSTER
GDS Snapshot. 9-8
data backup using NetWorker 9-12
DDR2 SDRAM 2-8
degradation function 5-21
degradation reservation. 5-21
dividing SBs and IO Units when
XPAR installed. 7-6
driver 3-5
dual power feed option 2-40
dump 9-2
area used to collect dump data. 10-1
area used to save dump data 10-3
dump-use area, estimating 10-1
dump-use memory, estimating 10-4
DVD-ROM drive 2-32
dynamic partitioning function. 1-16

E

EFI: Extensible Firmware Interface 3-2
environmental requirement for
using product. vi
error
error correction 2-8
record log 4-34
ETERNUS multipath driver 5-15
ETERNUS SF AdvancedCopy
Manager 9-9
example of allocation partition
granularity to IO Unit 7-14, 7-15

example partition configuration
(after logical division) 7-6
Extended Mirror Mode 5-9

F

fan 2-39
firmware 1-25, 3-2
BMC firmware 3-2
GSWB firmware. 3-3
MMB firmware 3-3
fjsnap 3-6
flexible I/O component 8-14
flexible I/O component and reserved
SB. 7-19
flexible I/O mechanism 1-17
function
MMB function. 4-13
PSA function 4-29

G

GSWB 1-24
GSWB firmware 3-3
GSWB: Gigabit Switch Board. 1-4
GTHB 2-27
GTHB external interface 2-28
GTHB external view 2-27
GTHB function 2-28
GUI
operation management 4-30

H

hard disk drive 2-12
hardware
configuration display 4-15
configuration management 4-32
hardware error monitoring 4-33
hardware technology 1-9
monitor. 4-16
hardware component 1-4
ACS 2-36
CPCB. 2-33
crossbar. 2-29
dual power feed option 2-40
fan 2-39
IO Unit 2-10
KVM. 2-32
MMB 2-24
OPL. 2-35
PCI_Box 2-40
PSU. 2-37
SB 2-7
XAI 2-30
XDI 2-31
HDD. 2-12
redundancy of HDD 5-11
high-speed switch method
duplication of transmission line
between servers 5-18
hot maintenance function. 1-21
hot swappable component. 6-2

I

independent resource expansion in
each partition 1-15
inter-cabinet cluster configuration 8-5
interface
CPCB external interface 2-34
IO Unit external interface 2-11
KVM external interface 2-32
MMB external interface 2-25
interface assignment 2-51
internal LAN switching hub 1-24
inter-partition communication
connection mode 4-7
inter-partition communication
disconnect mode 4-8
intra-cabinet cluster configuration 8-7
IO Unit 1-5, 2-10
IP address set by OS on partition 4-12
usable built-in I/O 2-17
IO Unit allocation rule. 7-13
IO Unit and PEXU connection 2-48
IO Unit and PEXU connection
(prohibited). 2-49
IOU and PEXU connection example 2-48
IP address 4-11
IP address set by OS on partition 4-12
IPF 1-10

K

KVM. 2-32
KVM/USB connection switching 4-19
KVM: Keyboard/Video/Mouse-USB 1-4

L

LAN
- business LAN 4-10
- configuration 4-3
- management LAN 4-4
- private LAN 4-9

large-scale SMP. 1-9
LED display 2-35
linked software
- backup. 3-8
- PRIMECLUSTER 3-6
- PRIMECLUSTER GDS 3-6
- PRIMECLUSTER GLS 3-7
- restoration 3-8
- Systemwalker 3-7

Linux 1-26, 3-4
load balance. 5-15
load distribution 5-15
Location-LED 2-35
log
- analysis 4-34
- collection 4-34
- display 4-34

M

maintenance
- firmware maintenance. 4-26

maintenance operation 4-35
management board 1-4, 2-24
management LAN 4-4
- redundancy of management LAN 5-5

memory 2-8
memory configuration, note on. 7-21
middleware product 1-28
mirror mode operation, note on 7-21
Mixed Mirror Mode, note on use 7-23
MMB. 1-4, 1-23, 2-24
- function 4-13
- IP address 4-11

MMB external interface 2-25
MMB firmware 3-3
MMB management targets. 4-39
MMB use environment setting 4-23
MMB-Ready-LED. 2-35
mounting condition
- SCSI card and usable built-in I/O . . . 2-18

mounting DIMMs, note on 2-9
mounting requirement 7-3
- IO Unit 7-3, 7-12
- SB. 7-3, 7-10

mounting requirements
(when XPAR installed). 7-10
MTTR 1-18
- reduced MTTR 1-20

N

NetVault. 9-11
network environment setting 4-23
NetWorker 9-12
NIC switch method
- duplication between server and hub and between switches on same network. 5-19

node 4-36, 8-1
NTBackup 9-2
NTP
- NTP server 4-25

O

operating environment. 7-21
operation management GUI 4-30
operation management software
linkage 4-38
OPL. 2-35
OS. 3-3
- Linux. 3-4
- Windows. 3-5

OS console 4-31
OS information display 4-32
OS input and output 4-21
OS-standard utility. 9-2
Other Port 7-22

P

PAL: Processor Abstraction Layer. 3-2
partition 1-3
- configuration definition 7-18
- configuration display. 4-18
- function for changing configuration. . . 7-22
- note to observe when configuring. . . . 7-22
- point to consider when determining configuration. 7-19
- setting. 4-18

partition function 1-14

partition function overview 7-1
partition granularity 7-7
SB 7-7, 7-8
partition granularity and allocation rule of IO Unit 7-13
partitioning function of the PRIMEQUEST 580/540 with XPAR installed 7-6
PCI card
compatible PCI card. 2-14, 2-45
mounting condition of PCI card 2-15
PCI card mounting condition 2-15
PCI card cassette 2-40
PCI Express card cassette 2-46
PCI slot. 2-13
PCI slot in PCIU 2-44
PCI slot in PEXU 2-46
PCI_Box 2-40
PCI_Box allocation 7-4, 7-15
PCI_Box connection with PCIU and PEXU 2-43
PCI_Box partition, allocation. 7-4, 7-5
PCIU. 2-44
PCIU and PEXU connection/combination 2-47
PCIU and PEXU in one partition. . .2-50, 2-51
PCIU and PEXU to IO Unit. 2-47
PCIU card cassette. 2-44
PEXU 2-45
PEXU (with PEXU card cassette inserted) 2-46
physical configuration
PRIMEQUEST 540A/540/440 2-3
PRIMEQUEST 580A/580/480 2-2
platform, operation on different. 1-16
power control 4-26
power supply unit 2-37
Power-LED 2-35
PRIMECLUSTER 3-6, 8-8
PRIMECLUSTER GDS 3-6, 5-12
redundancy of HDD (use of PRIMECLUSTER GDS) 5-11
PRIMECLUSTER GDS Snapshot 9-7
PRIMECLUSTER GLS. 3-7, 5-17
redundancy of business LAN (use of PRIMECLUSTER GLS). 5-17
PRIMEQUEST
feature 1-8
PRIMEQUEST 580A/540A/580/540/480/440 1-2
PRIMEQUEST conceptual diagram. 1-3
PRIMEQUEST management LAN, IP address of 4-11
PRIMEQUEST partition function 7-3
PRIMEQUEST System Disk Mirror for Windows 3-7
PRIMEQUEST-dedicated chipset 1-9
private LAN 4-9
redundancy of private LAN 5-5
product handling xi
product lineup 1-6
PSA
function 4-29
management targets 4-39
system management software 3-5
PSDM 3-7
PSU 2-37

R

reduced mean time to repair 1-18
redundancy
hardware redundancy 1-21
redundancy of component. 5-2
REMCS linkage 4-35
requirement on combination of PCIU and PEXU. 2-49, 2-50
reserved SB, cluster operation using 8-14
restoration 3-8
restore 9-2
setting information restore. 4-27
RHEL5 (IPF) 10-3, 10-4
RHEL-AS4 (IPF) 10-1

S

safety, note on vii
SAF-TE function 4-35
SAL: System Abstraction Layer 3-2
save
setting information save 4-27
SB 2-7
SB and IO Unit maintenance policy 7-19
SB partition granularitie and allocation rule 7-11
SBs: system boards. 1-4
scheduled operation 4-26

SCSI controller 4-35
secure socket layer 4-25
setting
MMB use environment setting 4-23
setting information restore 4-27
setting information save 4-27
simple network management
protocol 4-38
SMP 1-9, 1-10
snapshot software 9-7
SNMP 4-38
software
linked software 3-6
software supplied with
PRIMEQUEST A-1
supplied software 3-5
software configuration 3-1
software mirroring 5-12
software technology 1-25
specification
ACS specification 2-36
PRIMEQUEST 540A/540/440 1-6
PRIMEQUEST 580A/580/480 1-6
PSU specification 2-37
SSL 4-25
support 4-25
Standard Mirror Mode 1-11
storage redundancy 5-12
external storage redundancy 5-15
STP 8-12
structure and contents of manual i
supplied software 3-5
symmetric multi-processing
technology 1-10
system account 4-23
system board 2-7
system clock 2-33
system event log 4-34
system interconnect 1-5
system management
basic mechanism 4-1
System Mirror Mode 1-9, 1-11, 5-8
improved business continuity with . . . 1-13
System Mirror Mode and clustering 8-12
System Mirror Mode operation, note 7-21
system volume backup using
PRIMECLUSTER GDS Snapshot 9-7
SystemcastWizard Lite 3-5, 9-5
Systemwalker 3-7

T

text convention iv
time synchronization 4-25

U

UPC interface connector 2-33
update
firmware updating 4-26
USB port 2-12
user authentication 4-23
user privilege management 4-14

V

VERITAS NetBackup 9-10
Video port 2-12
VLAN
VLAN function of management LAN . . . 4-7

W

Windows 1-26, 3-5

X

XAI board 2-30
XDI board 2-31

Reader's Comment Form

We would appreciate your comments and suggestions for improving this publication.

Date:_____
Your Name:_____
Company: _____
Address: _____
City/State/Zip:_____
Phone/Email address: _____

Publication No.: C122-B001-10EN
Publication Name: PRIMEQUEST 580A/540A/580/540/480/440 SYSTEM DESIGN GUIDE

Your Comments:

Page	Line	Comments
Reply requested: o Yes o No		

Please evaluate the overall quality of this manual by checking (✓)the appropriate boxes

	Good	Fair	Poor		Good	Fair	Poor		Good	Fair	Poor
Organization:	o	o	o	Use of examples:	o	o	o	Legibility:	o	o	o
Accuracy:	o	o	o	Index coverage:	o	o	o	Binding:	o	o	o
Clarity:	o	o	o	Cross referencing:	o	o	o	Figures and tables:	o	o	o
Overall rating of this publication:	o	o	o					General appearance:	o	o	o

Technical level: o Too detailed o Appropriate o Not enough detail

All comments and suggestions become the property of Fujitsu Limited.

For Users in U.S.A., Canada, and Mexico

Fold and fasten as shown on back
No postage necessary if mailed in U.S.A.

Fujitsu Computer Systems
Attention: Engineering Ops M/S 249
1250 East Arques Avenue
Sunnyvale, CA 94085-5401
FAX: (408) 746-6676

For Users in Other Countries

Fax this form to the number below or send this form to the address below.

Fujitsu Learning Media Limited
FAX: 81-3-3730-3702
37-10 Nishi-Kamata 7-chome
Oota-Ku
Tokyo 144-0051
JAPAN

FUJITSU LIMITED

FOLD AND TAPE

NO POSTAGE
NECESSARY
IF MAILED
IN THE
UNITED STATES

BUSINESS REPLY MAIL

FIRST-CLASS MAIL PERMIT NO 741 SUNNYVALE CA

POSTAGE WILL BE PAID BY ADDRESSEE

FUJITSU

FUJITSU COMPUTER SYSTEMS
ATTENTION ENGINEERING OPS M/S 249
1250 EAST ARQUES AVENUE
SUNNYVALE CA 94085-5401

FOLD AND TAPE

FUJITSU