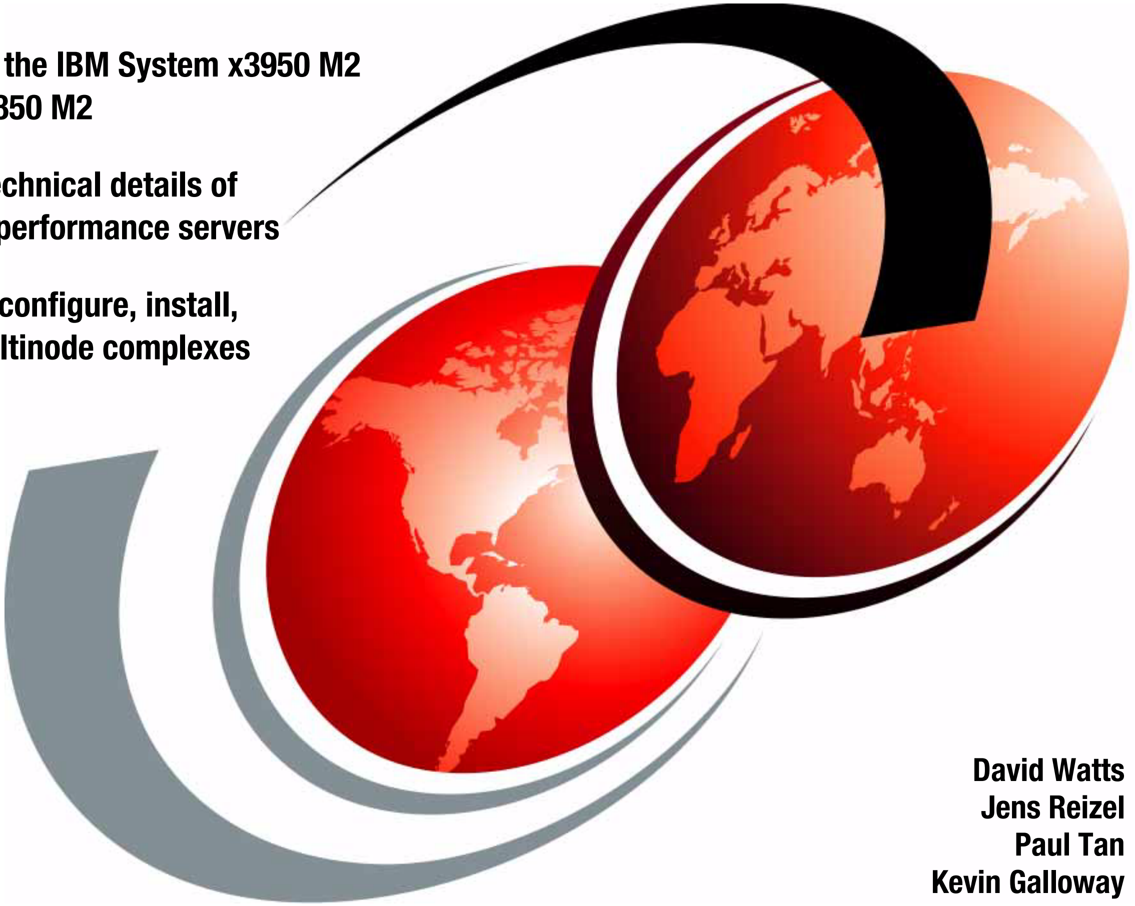
Planning, Installing, and Managing the IBM System x3950 M2

Understand the IBM System x3950 M2 and IBM x3850 M2

Learn the technical details of these high-performance servers

See how to configure, install, manage multinode complexes



David Watts
Jens Reizel
Paul Tan
Kevin Galloway

Redbooks

ibm.com/redbooks

International Technical Support Organization

Planning, Installing, and Managing the IBM System x3950 M2

November 2008

SG24-7630-00

Note: Before using this information and the product it supports, read the information in "Notices" on page ix.

First Edition (November 2008)

This edition applies to the following systems:

- IBM System x3950 M2, machine types 7141 and 7233
- IBM System x3850 M2, machine types 7141 and 7233

Contents

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

- Active Memory™
- BladeCenter®
- Chipkill™
- Cool Blue™
- DB2®
- DPI®
- IBM Systems Director Active Energy Manager™
- IBM®
- iDataPlex™
- Lotus®
- PowerExecutive™
- PowerPC®
- Predictive Failure Analysis®
- Redbooks®
- Redbooks (logo) ®
- RETAIN®
- ServeRAID™
- ServerGuide™
- ServerProven®
- ServicePac®
- System x™
- Tivoli®
- Wake on LAN®
- WebSphere®
- X-Architecture®
- Xcelerated Memory Technology™
- xSeries®

The following terms are trademarks of other companies:

Advanced Micro Devices, AMD, AMD-V, ATI, ES1000, Radeon, the AMD Arrow logo, and combinations thereof, are trademarks of Advanced Micro Devices, Inc.

Cognos, and the Cognos logo are trademarks or registered trademarks of Cognos Incorporated, an IBM Company, in the United States and/or other countries.

Snapshot, and the NetApp logo are trademarks or registered trademarks of NetApp, Inc. in the U.S. and other countries.

Novell, SUSE, the Novell logo, and the N logo are registered trademarks of Novell, Inc. in the United States and other countries.

Oracle, JD Edwards, PeopleSoft, Siebel, and TopLink are registered trademarks of Oracle Corporation and/or its affiliates.

SAP, and SAP logos are trademarks or registered trademarks of SAP AG in Germany and in several other countries.

Virtual SMP, VMotion, VMware, the VMware "boxes" logo and design are registered trademarks or trademarks of VMware, Inc. in the United States and/or other jurisdictions.

Java, OpenSolaris, Solaris, Ultra, and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

BitLocker, Hyper-V, Microsoft, SQL Server, Windows NT, Windows Server, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Intel Core, Intel SpeedStep, Intel Xeon, Intel, Itanium-based, Itanium, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States, other countries, or both.

Preface

The x3950 M2 server and x3850 M2 are the System x™ flagship servers and implement the fourth generation of the IBM® X-Architecture®. They delivers innovation with enhanced reliability and availability features to enable optimal performance for databases, enterprise applications, and virtualized environments.

The x3950 M2 four-socket system is designed for extremely complex, compute-intensive applications that require four sockets, plus processing power and large memory support.

The x3950 M2 and x3850 M2 features make the servers ideal for handling complex, business-critical On Demand Business applications such as database serving, business intelligence, transaction processing, enterprise resource planning, collaboration applications, and server consolidation.

Up to four x3950 M2 servers can be connected to form a single-system image comprising of up to 16 six-core processors, up to 1 TB of high speed memory, and support for up to 28 PCI Express adapters. The capacity gives you the ultimate in processing power, ideally suited for very large relational databases. The x3850 M2 is the equivalent of the x3950 M2 however it can only be used as a single four-processor node

This IBM Redbooks® publication describes the technical details of the x3950 M2 scalable server and the x3850 M2 server. We explain what the configuration options are, how 2-node, 3-node, and 4-node complexes are cabled and implemented, how to install key server operating systems, and what management tools are available to systems administrators.

The team that wrote this book

This book was produced by a team of specialists from around the world working at the International Technical Support Organization, Raleigh Center.

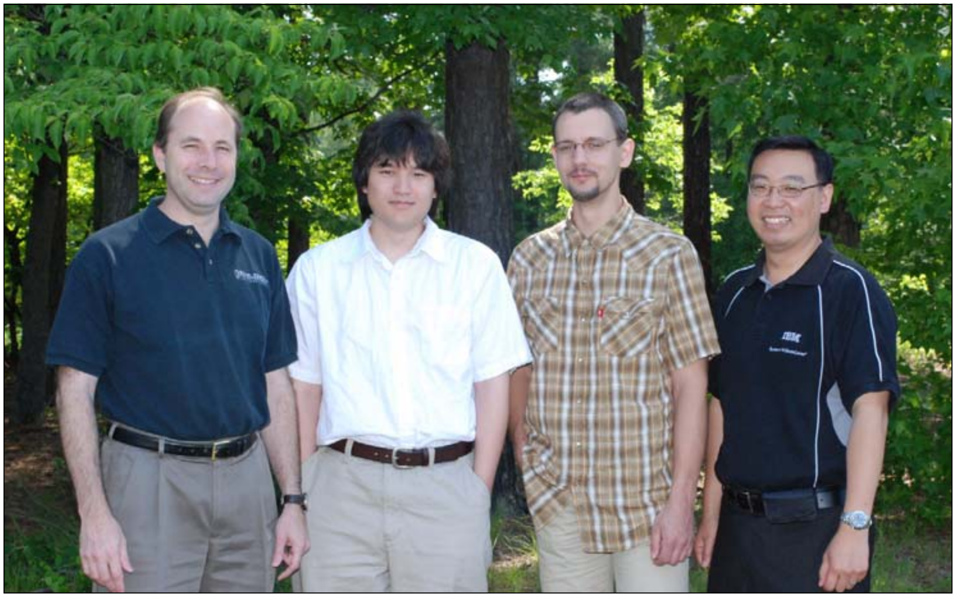
David Watts is a Consulting IT Specialist at the IBM ITSO Center in Raleigh. He manages residencies and produces IBM Redbooks publications on hardware and software topics related to IBM System x and BladeCenter® servers, and associated client platforms. He has authored over 80 books, papers, and technotes. He holds a Bachelor of Engineering degree from the University of

Queensland (Australia) and has worked for IBM both in the United States and Australia since 1989. He is an IBM Certified IT Specialist.

Jens Reizel is a Support Specialist at IBM Germany and is responsible for the post-sales technical support teams in the EMEA region. He has been working in this function and with IBM for nine years. His areas of expertise include IBM System x high end systems, management hardware, and Windows®, Linux®, and VMware® operating systems.

Paul Tan works as a presales System x, BladeCenter and Storage Technical Specialist at IBM Systems and Technology Group in Melbourne, Australia. He regularly leads customer presentations and solution workshops based around key leading IBM technologies with a particular focus on x86-based virtualization products such as VMware. He has been working in this role for more than two years and prior to that for five years as an IBM Infrastructure Consultant, specializing in Microsoft® and Linux systems. He holds a Bachelor of Science (Computer Science) and Bachelor of Engineering (Computer Engineering) from the University of Melbourne, Australia. He also holds industry certifications such as Microsoft Certified Systems Engineer and Red Hat Certified Technician.

Kevin Galloway is a graduate student at the University of Alaska, Fairbanks. He is currently working toward a Master of Science degree in Computer Science, with a focus on computer security and software development. He joined the ITSO as an IBM Redbooks intern.



The team (left to right): David, Kevin, Jens, and Paul

Thanks to the following people for their contributions to this project:

From the International Technical Support Organization:

- Jeanne Alderson
- Tamikia Barrow
- Emma Jacobs
- Linda Robinson
- Diane Sherman
- Erica Wazewski

From IBM Marketing

- Beth McElroy
- Heather Richardson
- Kevin Powell
- Don Roy
- Scott Tease
- Bob Zuber

From IBM Development

- Paul Anderson
- Chia-Yu Chu
- Richard French
- Joe Jakubowski
- Mark Kapoor
- Don Keener
- Dan Kelaher
- Randy Kolvick
- Josh Miller
- Thanh Ngo
- Chuck Stephan

From IBM Service and Support

- Khalid Ansari
- Brandon Church

Become a published author

Join us for a two- to six-week residency program! Help write a book dealing with specific products or solutions, while getting hands-on experience with leading-edge technologies. You will have the opportunity to team with IBM technical professionals, Business Partners, and Clients.

Your efforts will help increase product acceptance and customer satisfaction. As a bonus, you will develop a network of contacts in IBM development labs, and increase your productivity and marketability.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an e-mail to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

1

Technical overview

The IBM System x3950 M2 and IBM System x3850 M2 are the IBM System x flagship systems. They are based on eX4 technology, which is the fourth generation of IBM X-Architecture. This technology leverages the extensive research and development by IBM in XA-64e chipset based on the scalable Intel® Xeon MP system.

This chapters discusses the following topics:

- 1.1, “IBM eX4-based servers” on page 2
- 1.2, “Model numbers and scalable upgrade options” on page 9
- 1.3, “Multinode capabilities” on page 14
- 1.4, “x3950 M2 Windows Datacenter models” on page 15
- 1.5, “Integrated virtualization: VMware ESXi” on page 19
- 1.6, “IBM fourth generation XA-64e chipset” on page 27
- 1.7, “Processors” on page 33
- 1.8, “Memory subsystem” on page 39
- 1.9, “SAS controller and ports” on page 42
- 1.10, “PCI Express subsystem” on page 43
- 1.11, “Networking” on page 44
- 1.12, “Systems management features” on page 47
- 1.13, “Trusted Platform Module and where used” on page 51

1.1 IBM eX4-based servers

IBM eX4 technology offers a balanced system design with unique scalability, reliability, availability, and performance capabilities to take full advantage of Intel's latest multi-core processors. By connecting four servers together, the single-system image can have up to 16 processor sockets (96 cores), up to 128 DIMM sockets and 1 TB of RAM, 28 PCI Express slots, and 34.1 GBps of memory bandwidth for each 256 GB RAM server. This results in a high-capacity system with significant processing and I/O performance, and greater power efficiency.

The two servers based on IBM eX4 technology are:

- IBM System x3850 M2
- IBM System x3950 M2

Although they have the same technical specifications and features, the x3850 M2 cannot be used to form a multinode unless you upgrade it to an IBM System x3950 M2 by adding the ScaleXpander Option Kit, as described in section 1.2, "Model numbers and scalable upgrade options" on page 9.

1.1.1 Features of the x3950 M2 and x3850 M2 servers

The x3950 M2 and x3850 M2 look very similar, as shown in Figure 1-1.

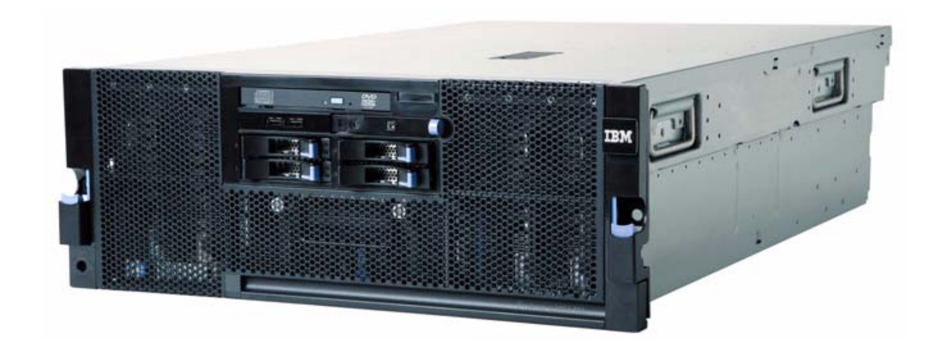


Figure 1-1 IBM System x3950 M2 and IBM System x3850 M2

Front and rear panels

The components and connectors on the front and rear of the system are shown in Figure 1-2 on page 3 and Figure 1-3 on page 4.

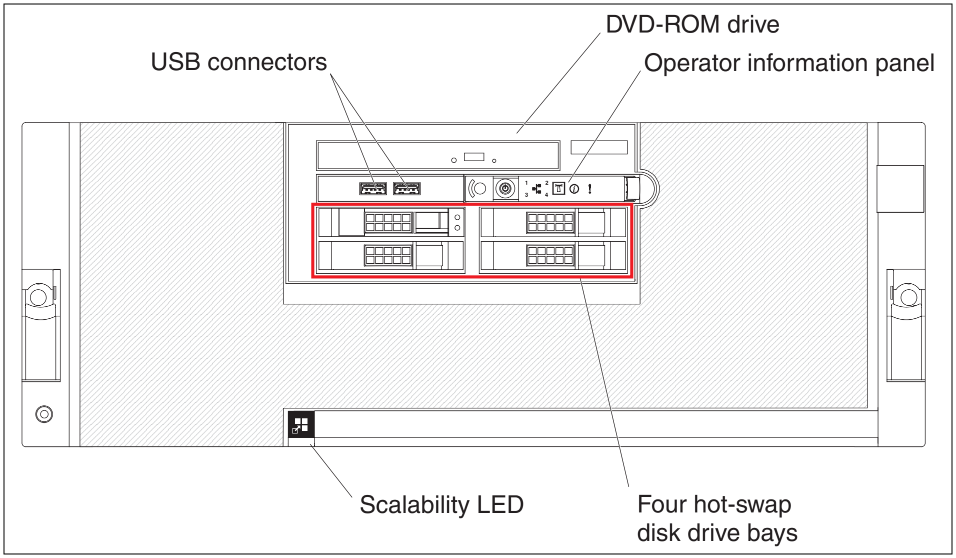


Figure 1-2 Front panel of x3850 M2 and x3950 M2

The front panel of the x3850 M2 and the x3950 M2, as shown in Figure 1-2, provides easy access to a maximum of four hot-swap 2.5-inch SAS drives, DVD-ROM, two USB 2.0 ports, an operator information panel with power on/off button, and LEDs indicating information such as scalability, network activity, and system errors and warnings.

The scalability LED on an x3950 M2 indicates whether the node (building block in a scalable system) is participating in a multinode x3950 M2 complex. After each node has successfully merged with the primary node in a partition, the scalability LED is lit on all nodes in a partition of a multinode complex.

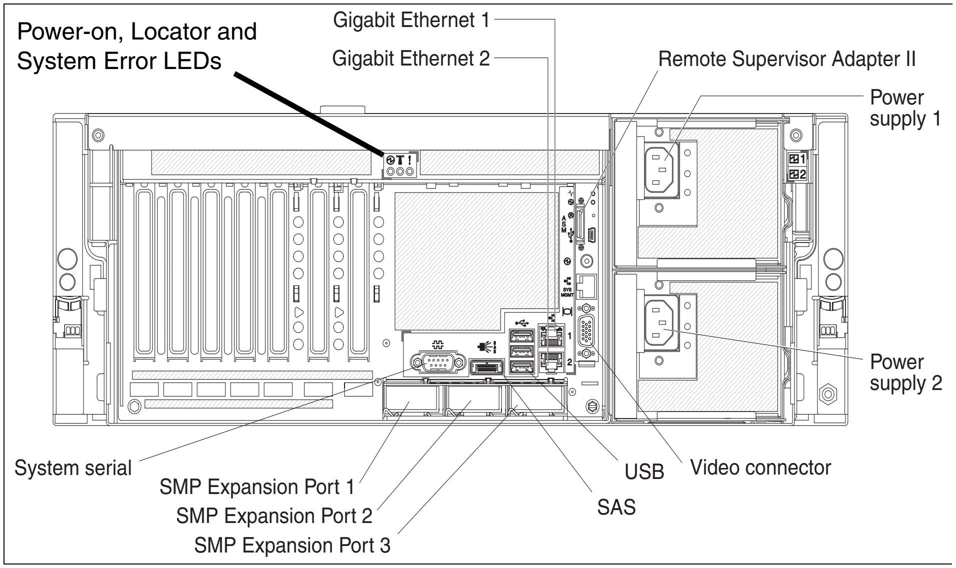


Figure 1-3 Rear panel of x3850 M2 and x3950 M2

The rear panel of the x3850 M2 and x3950 M2, as shown in Figure 1-3, has:

- PCI Express (PCIe) slots 1 to 7 (from left to right on the panel)
- System serial port
- Three scalability SMP expansion ports used for multinode x3950 M2 complexes
- External SAS port
- Three USB 2.0 ports
- Integrated dual-port Broadcom Gigabit Ethernet RJ45 ports
- Remote Supervisor Adapter II panel, which contains the servers video connector port, 10/100 Mbps RJ45 out-of-band remote management port (there is also a mini-USB port and a power adapter socket that is not used for the x3850 M2/x3950 M2)
- Two hot-swap redundant power supplies

Hypervisor models of the x3850 M2

Inside the server is an additional USB socket used exclusively for the embedded virtualization feature. This device, shown in Figure 1-4 on page 5, is standard on hypervisor models of the x3850 M2.

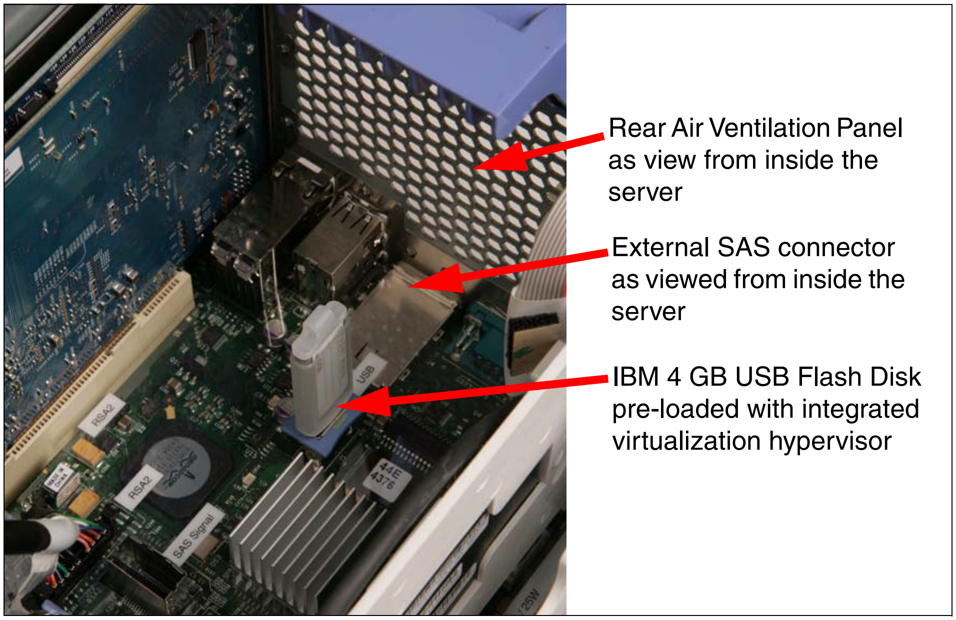

Rear Air Ventilation Panel as view from inside the server
External SAS connector as viewed from inside the server
IBM 4 GB USB Flash Disk pre-loaded with integrated virtualization hypervisor
RSA2
RSA2
SAS Signal
USB

Figure 1-4 On the hypervisor models of x3850 M2, a USB flash drive is pre-installed in the internal USB socket and contains VMware ESXi 3.5 pre-loaded

Standard features for both systems

The x3950 M2 and x3850 M2 have the following standard features. We discuss these in greater detail in sections later in this chapter.

Processors

Processor features include:

- One 4U Rack-optimized sever with one of the following Intel processors:
  - Xeon 7200 series (Tigerton) dual-core processors
  - Xeon 7300 series (Tigerton) quad-core processors
  - Xeon 7400 series (Dunnington) quad-core processors
  - Xeon 7400 series (Dunnington) 6-core processors
- Two processors standard, with support for up to four processors
- One IBM eX4 “Hurricane 4” chipset with four 1066 MHz front-side buses
- Support for Intel Virtualization Technology (Intel VT), Intel 64 technology (EM64T), and Execute Disable Bit feature

Memory subsystem

Memory subsystem features include:

- 4 GB or 8 GB memory standard expandable to 256 GB
- Support for 1, 2, 4, and 8 GB DDR2 registered DIMMs
- Maximum of 32 DIMM slots by installing four memory card (each card has eight DIMMs sockets)
- Active Memory™ with Memory ProteXion, hot-swap memory with memory mirroring, hot-add memory with supported operating systems, and Chipkill™.

I/O slots and integrated NICs

I/O subsystem features include:

- Seven 64-bit PCIe x8 full height (half-length) slots; two of these seven slots are hot-swap
- Integrated dual-port Broadcom NeXtreme II 5709C PCI Express Gigabit Ethernet controller with Jumbo Frame support

> **Note:** TCP Offload Engine (TOE) support is planned.

SAS RAID controller and HDD slots

Disk subsystem features include:

- Integrated LSI 1078 SAS controller with support for RAID-0 and RAID-1
- External JBOD SAS storage through external SAS x4 port (if IBM ServeRAID™ MR10k SAS/SATA Controller is installed)

  The SAS SFF-8088 connector is located above SMP Expansion Port 2 in Figure 1-3 on page 4.
- Up to four hot-swap 2.5-inch SAS hard drives (up to a maximum of 584 GB of internal storage)

Systems management and security

Management and security features include:

- Onboard BMC shares integrated Broadcom Gigabit Ethernet 1 interface
- Remote Supervisor Adapter II with dedicated 10/100 Mbps Ethernet management interface

  The RSA Adapter II's 10/100 Mbps Ethernet port is located above the video connector in Figure 1-3 on page 4.
- Operator information panel (see Figure 1-2 on page 3), which provides light path diagnostics information

- Windows Hardware Error Architecture (WHEA) support in the BIOS
- Trusted Platform Module support. The module is a highly secure start-up process from power-on through to the startup of the operating system boot loader. Advanced Configuration and Power Interface (ACPI) support is provided to allow ACPI-enabled operating systems to access the security features of this module.

System ports and media access

Ports and media access features include:

- Six USB 2.0 ports, two on the front panel, three on the rear panel, and one internal for USB Flash Disk

  The Hypervisor model of x3850 M2 includes an integrated hypervisor for virtualization on a 4 GB USB Flash Disk with VMware ESXi pre-loaded. See Figure 1-4 on page 5.
- An ATI™ Radeon™ ES1000™ SVGA video controller (DB-15 video connector on RSA II card as shown in Figure 1-3 on page 4) on the Remote Supervisor Adapter II
- Optical drive:
  - On machine type 7141: One standard 24x/8x IDE CD-RW/DVD-ROM combo drive
  - One machine type 7233: SATA CD-RW/DVD-ROM combo drive
- USB keyboard and mouse
- System serial port
- Three SMP expansion ports for use in scalable multinode complex.

Power

Two hot-swap redundant 1440 W power supplies are standard. At 220 V, one power supply is redundant. At 110 V, the power supplies are non-redundant.

1.1.2 x3950 M2: scalable hardware components

The x3950 M2 includes the following additional scalable hardware components as standard compared to the x3850 M2. The additional components enable the x3950 M2 to scale up to a multinode complex comprising of up to a maximum four x3950 M2s.

- ScaleXpander chip (see Figure 1-5 on page 8)
- One 3.08 m scalability cable (see Figure 1-6 on page 9)

- Larger cable management arm to accommodate use of scalability cables connecting to SMP expansion ports (see Figure 1-7 on page 12 and Figure 1-8 on page 13)

All necessary hardware components are provided for forming a three-node x3950 M2 complex with the order of three x3950 M2 servers. However, to form a four-node x3950 M2 complex, you must have four x3950 M2 *and* a Scalability Upgrade Option 2, which contains one 3.08m and one 3.26m Scalability cable (see Table 4-1 on page 202 for details of part numbers). Refer to Chapter 4, “Multinode hardware configurations” on page 195 for more details about scaling the x3950 M2 to complexes of two, three, and four nodes.

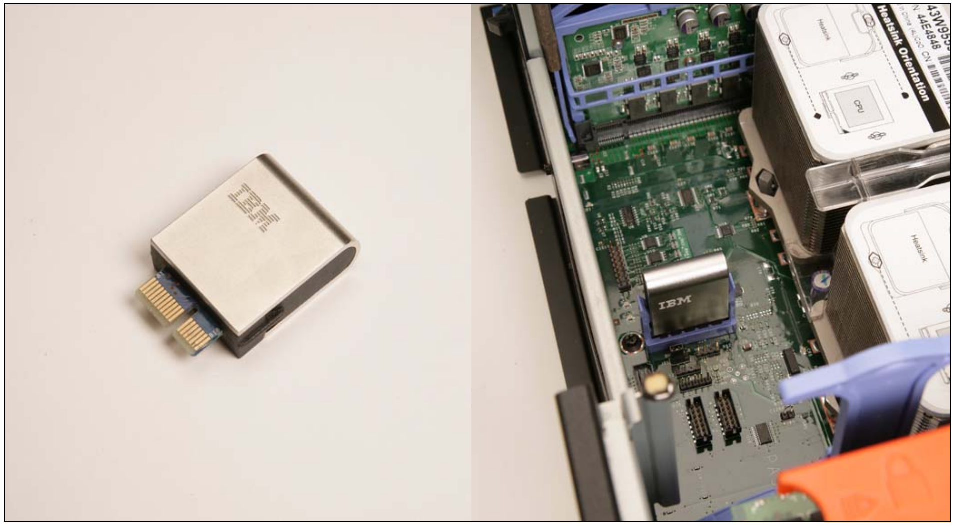


Figure 1-5 ScaleXpander chip (left); ScaleXpander chip installed on processor board near the front panel of the x3950 M2 (right)

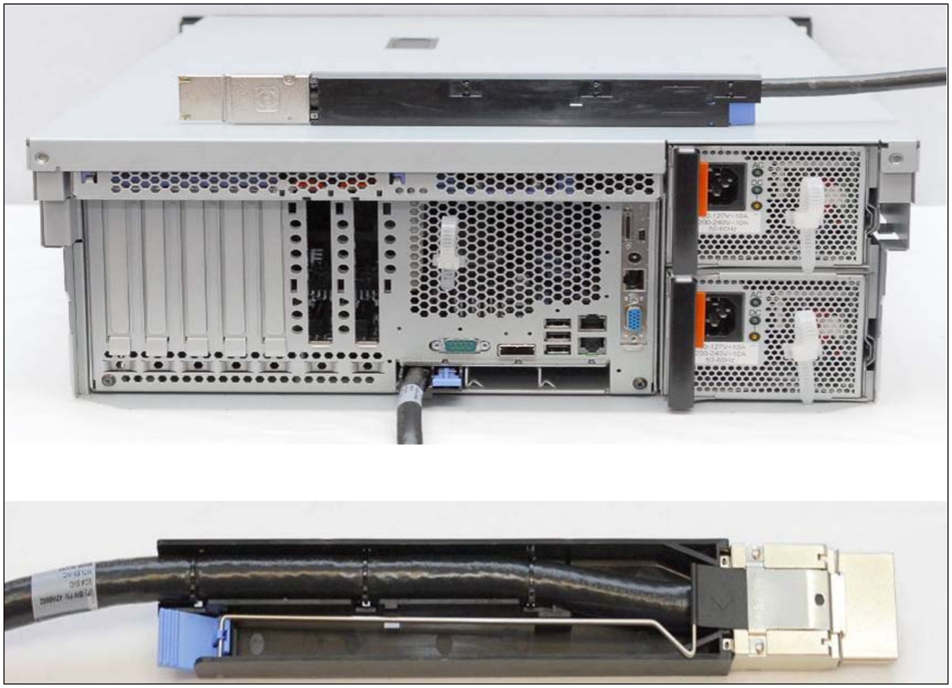


Figure 1-6 Scalability cable (top); cable installed in SMP Expansion Port 1 (bottom)

1.2 Model numbers and scalable upgrade options

As discussed previously, the x3850 M2 and x3950 M2 servers are based on IBM eX4 technology. This section lists the available models for each server and where to find more information about models available in your country.

The tables in this section use the following nomenclature:

n Indicates variations between server models relating to the processor type and the number of memory cards and memory DIMMs installed.

c Indicates the country in which the model is available: **U** is for countries in North America and South America. **G** is for EMEA (for example, 1RG). For Asia-Pacific countries, the letter varies from country to country.

1.2.1 Finding country-specific model information

For the specific models available in your country, consult one of the following sources of information:

- Announcement letters; search for the machine type (such as 7141):

  http://www.ibm.com/common/ssi/

- Configuration and Options Guide (COG) for System x:

  http://www.ibm.com/support/docview.wss?uid=psg1SCOD-3ZVQ5W

  Direct link to COG page for the System x3850/3950 M2 servers:

  http://www.ibm.com/systems/xbc/cog/x3850m2/x3850m2aag.html

- IBM BladeCenter and System x Reference Sheets (xREF):

  http://www.redbooks.ibm.com/xref

1.2.2 x3850 M2 model information

The model numbers of the x3850 M2 are listed in Table 1-1.

Table 1-1 Models of x3850 M2

Models	Description
7141-***n*****R*****c***	Standard models of x3850 M2 with dual-core or quad-core Xeon 7200 and Xeon 7300 (Tigerton) processors
7141-3**H*****c***	Integrated hypervisor models of x3850 M2 with Xeon E7330 (Tigerton) processors. See 1.5, “Integrated virtualization: VMware ESXi” on page 19.
7233-***n*****R*****c***	Standard models of x3850 M2 with quad-core and six-core Xeon 7400 (Dunnington) processors
7233-4**H*****c***	Integrated hypervisor models of x3850 M2 with quad-core Xeon E7440 (Dunnington) processors. See 1.5, “Integrated virtualization: VMware ESXi” on page 19.

1.2.3 x3950 M2 model information

The model numbers of the x3950 M2 are listed in Table 1-2.

Table 1-2 Models of x3950 M2

Models	Description
7141-**_n_S_c_**	Standard models of the x3950 M2 with dual-core or quad-core Xeon 7200 or Xeon 7300 (Tigerton) processors
7233-**_n_S_c_**	Standard Models of x3950 M2 with quad-core and six-core Xeon 7400 (Dunnington) processors[a]
7141-**_n_A_c_**	Datacenter Unlimited Virtualization with High Availability models certified for 32-bit Windows 2003 Datacenter Edition. See 1.4.2, “IBM Datacenter Unlimited Virtualization with High Availability” on page 16.
7141-**_n_B_c_**	Datacenter Unlimited Virtualization with High Availability models certified for 64-bit Windows 2003 Datacenter Edition. See 1.4.2, “IBM Datacenter Unlimited Virtualization with High Availability” on page 16.
7141-**_n_D_c_**	Datacenter Unlimited Virtualization models certified for 32-bit Windows 2003 Datacenter Edition. See 1.4.1, “IBM Datacenter Unlimited Virtualization offering” on page 16.
7141-***n***E***c***	Datacenter Unlimited Virtualization models certified for 64-bit Windows 2003 Datacenter Edition. See 1.4.1, “IBM Datacenter Unlimited Virtualization offering” on page 16.

[a] Dunnington quad-core and six-core processors include L2 and L3 shared cache unlike Tigerton processors with only L2 shared cache. See 1.7, “Processors” on page 33 for more details.

1.2.4 Scalable upgrade option for x3850 M2

Unlike the x3850 server (based on X3 technology), the x3850 M2 can be converted to an x3950 M2 through the use of the IBM ScaleXpander Option Kit, part number 44E4249. After this kit is installed, the x3850 M2 functionally becomes an x3950 M2, and is therefore able to form part of a multinode complex comprising of up to four x3950 M2s.

The IBM ScaleXpander Option Kit contains the following items:

- Scalability cable 3.08m (See Figure 1-6 on page 9.)
- Larger cable management arm, which replaces the existing arm to allow the easy installation of the scalability cables. See Figure 1-7 on page 12 and Figure 1-8 on page 13.

- ScaleXpander chip required to convert the x3850 M2 to an x3950 M2. See Figure 1-5 on page 8.
- x3950 M2 bezel, which replaces the existing bezel and shows the x3850 M2 has the kit installed and is now functionally equal to an x3950 M2. See Figure 1-9 on page 13.

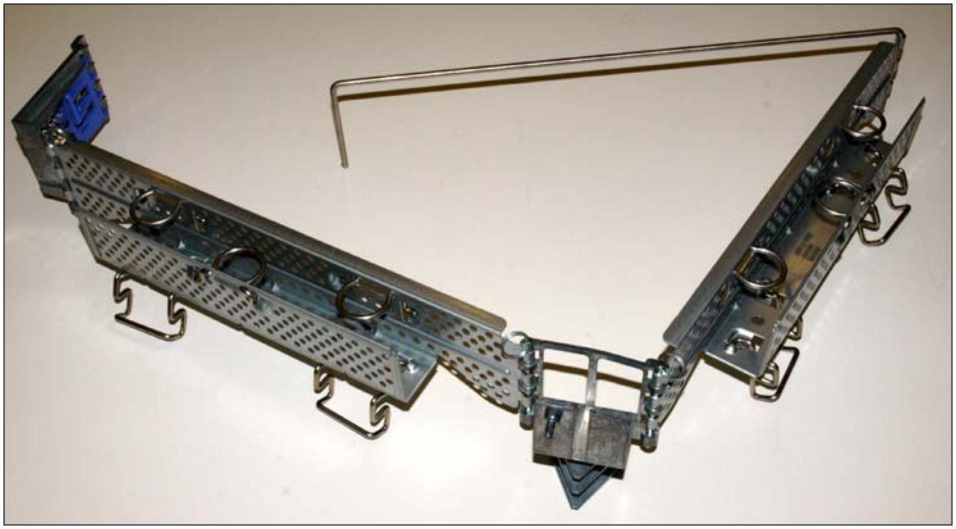


Figure 1-7 x3950 M2 enterprise cable management arm

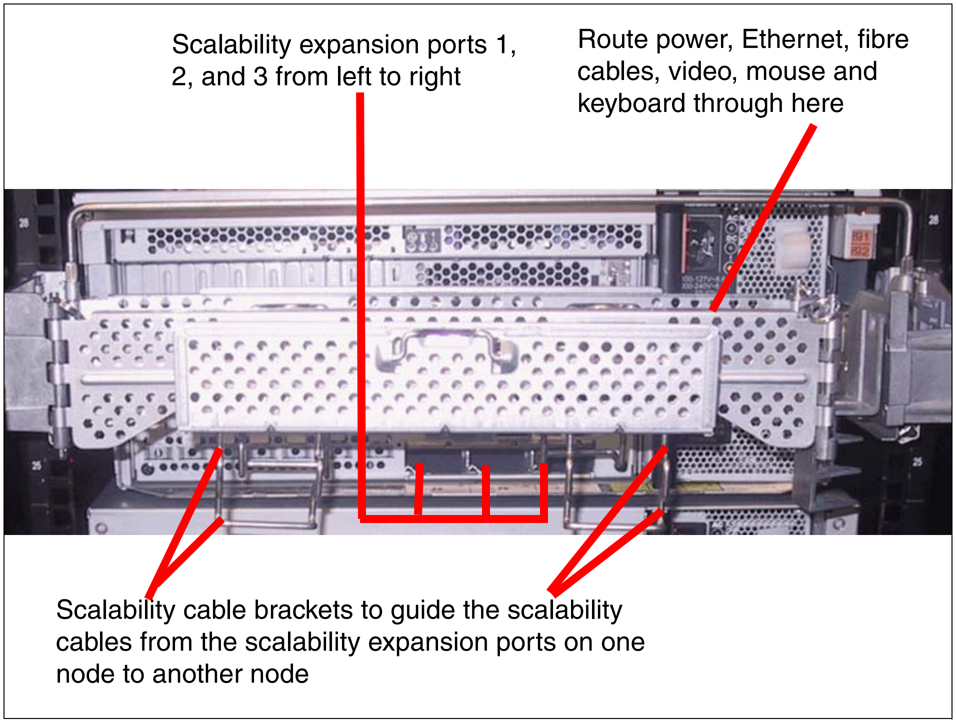

Scalability expansion ports 1, 2, and 3 from left to right
Route power, Ethernet, fibre cables, video, mouse and keyboard through here
Scalability cable brackets to guide the scalability cables from the scalability expansion ports on one node to another node

Figure 1-8 x3950 M2 cable management arm mounted on server rails

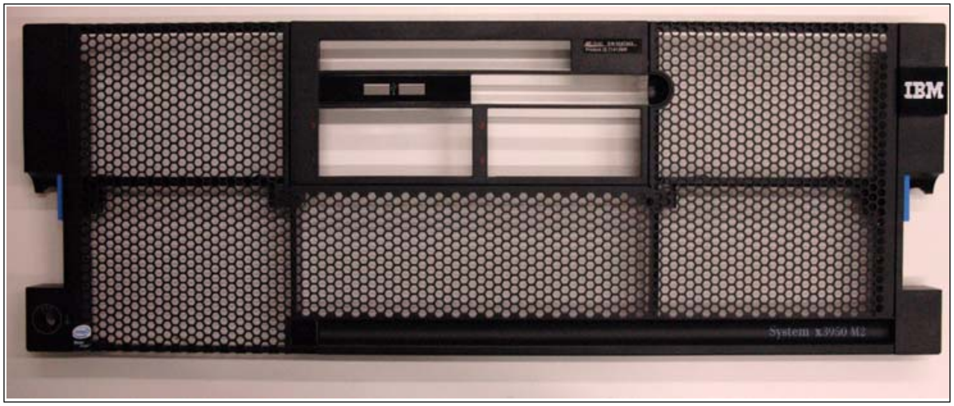


Figure 1-9 x3950 M2 bezel

1.3 Multinode capabilities

The x3950 M2 is the base building block, or *node*, for a scalable system. At their most basic, these nodes are comprised of four-way SMP-capable systems with processors, memory, and I/O devices. The x3950 M2 is the building block that allows supported 8-way, 12-way, and 16-way configurations by adding more x3950 M2s as required.

Unlike with the System x3950 and xSeries® 460, the x3950 M2 does not require a special modular expansion enclosure. The multinode configuration is simply formed by using another x3950 M2 or an x3850 M2 that has the ScaleXpander Option Kit installed as described previously in 1.2.4, "Scalable upgrade option for x3850 M2" on page 11

Note: When we refer to an x3950 M2, we mean either an x3950 M2 or an x3850 M2 that has the ScaleXpander Option Kit installed.

Multinode configurations

The x3950 M2 can form a multinode configuration by adding one or more x3950 M2 servers. A number of configurations are possible as shown in Figure 1-10.

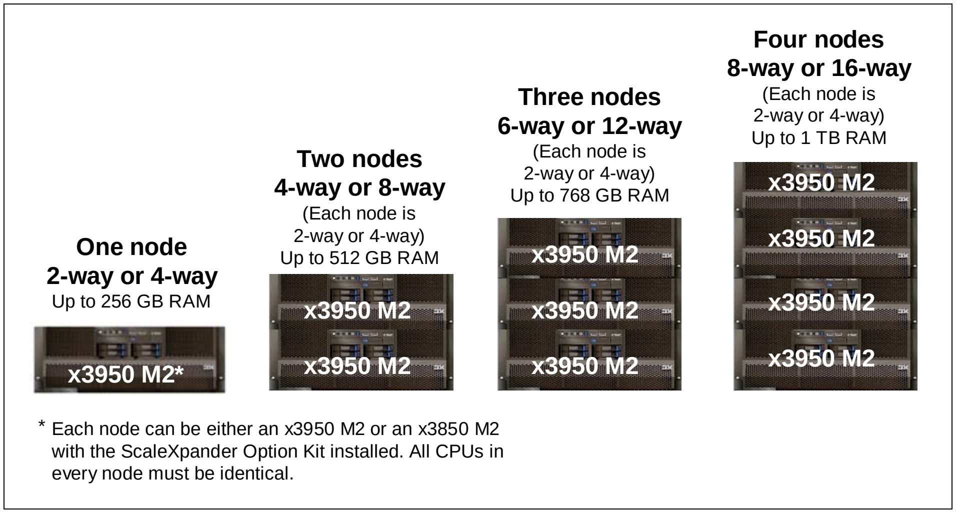


Figure 1-10 Possible multinode configurations

The possible configurations are:

- A one-node system is a one x3950 M2 server or one x3850 M2 server, with one, two, three, or four processors and up to 256 GB of RAM.
- A two-node complex is comprised of two x3950 M2 servers, with up to eight processors, and up to 512 GB RAM installed.
- A three-node complex is comprised of three x3950 M2 servers, up to 12 processors, and up to 768 GB RAM installed.
- A four-node complex is comprised of four x3950 M2 servers, up to 16 processors, and up to 1 TB RAM installed.

Note: At the time of writing, only Windows 2003 Enterprise and Datacenter 64-bit editions, RHEL 5 64-bit, and SLES 10 64-bit support this amount of memory. See 2.6, “Operating system scalability” on page 66 for details.

Partitioning

Partitioning is the concept of logically splitting a multinode complex into separate systems. You can then install an operating system on a partition and have it run independently from all other partitions. The advantage of partitioning is that you can create and delete partitions without having to recable the complex. The only requirement is that partitions be formed on node boundaries.

The interface where you set up and maintain partitions is an extension of the Remote Supervisor Adapter II Web interface. It is used to create, delete, control, and view scalable partitions.

Multinode complexes support partitioning on node boundaries. This means, for example, you can logically partition your 2-node 8-way system as two 4-way systems, while still leaving the complex cabled as 2 nodes. This increases flexibility. You can reconfigure the complex by using the Web interface without changing the systems or cabling.

For more information about multinode complexes and partitioning, see Chapter 4, “Multinode hardware configurations” on page 195.

1.4 x3950 M2 Windows Datacenter models

IBM offers Windows 2003 Datacenter Edition as part of the following two IBM offerings, which are described in this section:

- IBM Datacenter Unlimited Virtualization
- IBM Datacenter Unlimited Virtualization with High Availability

1.4.1 IBM Datacenter Unlimited Virtualization offering

The IBM Datacenter Unlimited Virtualization offering is ideal for customers who already have a well-managed IT infrastructure and want only a Windows operating system that scales from 4-way to 32-way and offers maximum performance and scalability in a nonclustered environment.

IBM Datacenter Unlimited Virtualization solution is tested and certified on specific System x servers and with standard ServerProven® options:

- Datacenter-specific, certified configurations are no longer required
- Supported on all ServerProven configurations for designated System x Datacenter servers

Installation can be performed by IBM, the Business Partner, or the customer. Optional System x Lab Services onsite and IBM Global Services - Remote Support Services can be provided.

Table 1-2 on page 11 shows the system models for both 32-bit and 64-bit versions of the operating system.

With the IBM Datacenter Unlimited Virtualization option, the x3950 M2 models come with two processors, 8 GB of memory (eight 1 GB DIMMs), four memory cards, and no disks. The system is shipped with the Datacenter installation CD, OS documentation, recovery CD, and a 4-socket Certificate of Authenticity (COA) to license the system. Windows Server® 2003 R2 Datacenter Edition is not preloaded. This offering is available in both English and Japanese languages.

> **Note:** IBM no longer offers a Software Update Subscription for this offering. Customers should purchase a Microsoft Software Assurance contract for operating system maintenance and upgrades. For information see:
>
> `http://www.microsoft.com/licensing/sa`

1.4.2 IBM Datacenter Unlimited Virtualization with High Availability

The IBM Datacenter Unlimited Virtualization with High Availability (UVHA) Program offering delivers a fully certified solution on 4-way through 32-way server configurations that support up to 8-node Microsoft cluster certified solutions for a tightly controlled, end-to-end supported environment for maximum availability.

This end-to-end offering provides a fully configured and certified solution for customers who want to maintain a tightly controlled environment for maximum availability. To maintain this high availability, the solution must be maintained as a certified configuration.

IBM Datacenter UVHA solution offerings are tested and certified on specific System x servers and with standard ServerProven options, storage systems, and applications. All components must be both ServerProven and Microsoft cluster-logo certified.

The operating system is not preloaded. Installation must be performed by IBM System x Lab Services or IBM Partners certified under the EXAct program. Standard IBM Stage 2 manufacturing integration services can be used. IBM Global Services - Remote Support Services are mandatory.

Table 1-2 on page 11 shows the models for both 32-bit and 64-bit versions of the operating system.

With this option, the x3950 M2 models come with two processors, 8 GB memory (eight 1 GB DIMMs), four memory cards, and no disks. Unlike previous high-availability offerings, Windows Server 2003 R2 Datacenter Edition is not preloaded. Also shipped with the system are a recovery CD, OS documentation, and a 4-socket Certificate of Authenticity (COA) to license the system. This offering is available in both English and Japanese languages.

Note: IBM no longer offers a Software Update Subscription for this offering. Customers should purchase a Microsoft Software Assurance contract for operating system maintenance and upgrades. For information see:

http://www.microsoft.com/licensing/sa

1.4.3 Upgrading to Datacenter Edition

If you are using another Windows operating system on your x3950 M2, such as Windows Server 2003 Enterprise Edition, and want to upgrade to Datacenter Edition, you can order the appropriate upgrade as described in this section.

IBM Datacenter preload upgrades can be ordered only after receiving approval from the IBM world-wide System x marketing team. IBM Sales Representatives should notify their geography Marketing Product Manager and Sales Managers of these opportunities, and the Product and Sales Managers should, in turn, notify the World Wide Marketing Product Manager of the sales opportunity. Business Partners should notify their IBM Sales Representative, who should engage with geography Product Marketing and Sales Managers.

IBM validates the customer's current x3950 M2 hardware configuration as a certified Datacenter configuration. Orders are allowed only after a Solutions

Assurance Review is completed. For the IBM Datacenter Unlimited Virtualization with High Availability offering, the appropriate service and support contracts must also be in place.

Upgrading to IBM Datacenter Unlimited Virtualization

To upgrade to the IBM Datacenter Unlimited Virtualization offering, order one or more of the part numbers listed in Table 1-3. You must have one 4-CPU license for each x3950 M2 in your configuration. Licenses are cumulative.

Table 1-3 Upgrade options for the IBM Datacenter Unlimited Virtualization offering

Upgrade kits	Order number
Windows Server 2003 Datacenter Edition R2, 32-bit, 1-4 CPUs	4818-NCU
Windows Server 2003 Datacenter Edition R2 x64, 1-4 CPUs	4818-PCU
Windows Server 2003 Datacenter Edition R2, 32-bit, 1-4 CPUs (Japanese)	4818-NCJ
Windows Server 2003 Datacenter Edition R2 x64, 1-4 CPUs (Japanese)	4818-PCJ

Note: These upgrade order numbers can only be ordered from the IBM World Wide System x Brand and might not appear in IBM standard configuration tools.

Upgrading to IBM Datacenter Unlimited Virtualization with High Availability

To upgrade to the IBM Datacenter Unlimited Virtualization with High Availability (UVHA) offering, order one or more of the part numbers listed in Table 1-4. You must have one 4-CPU license for each x3950 M2 in your configuration. Licenses are cumulative.

Table 1-4 Upgrade options for IBM Datacenter UVHA

Upgrade kits	Order number
Windows Server 2003 Datacenter Edition R2, 32-bit, 1-4 CPUs	4816-NCU
Windows Server 2003 Datacenter Edition R2 x64, 1-4 CPUs	4816-PCU

Upgrade kits	Order number
Windows Server 2003 Datacenter Edition R2, 32-bit, 1-4 CPUs (Japanese)	4816-NCJ
Windows Server 2003 Datacenter Edition R2 x64, 1-4 CPUs (Japanese)	4816-PCJ

Note: These upgrade order numbers can only be ordered from the IBM World Wide System x Brand and might not appear in IBM standard configuration tools.

1.4.4 Datacenter multinode configurations

Configurations greater than 4-way on the x3950 M2 are comprised of an x3950 M2 primary node with a number of x3950 M2 systems, up to a maximum of four nodes to make a 16-way Datacenter system. Forming these scalable systems requires additional scalability cables, as explained in 4.4, “Prerequisites to create a multinode complex” on page 201.

1.4.5 Datacenter cluster configurations

Microsoft Cluster Server (MSCS) is supported only under the UVHA offering. Check for updates to the Microsoft Hardware Compatibility List (HCL) at:

`http://www.microsoft.com/whdc/hcl/default.mspx`

1.5 Integrated virtualization: VMware ESXi

VMware ESXi is the next-generation hypervisor that is integrated into IBM servers such as the x3850 M2. It provides a cost-effective, high-capacity virtual machine platform with advanced resource management capabilities. This innovative architecture operates independently from any general-purpose operating system, offering improved security, increased reliability, and simplified management. The compact architecture is designed for integration directly into virtualization-optimized server hardware like the IBM x3850 M2, enabling rapid installation, configuration, and deployment.

1.5.1 Key features of VMware ESXi

As discussed in the white paper *The Architecture of VMware ESX Server 3i*[1], VMware ESXi has equivalent functions to ESX Server 3.5. However, the ESXi hypervisor footprint is less than 32 MB of memory because the Linux-based service console has been removed. The function of the service console is replaced by new remote command line interfaces in conjunction with adherence to system management standards.

Like the ESX Server, VMware ESXi supports the entire VMware Infrastructure 3 suite of products, including VMFS, Virtual SMP®, VirtualCenter, VMotion®, VMware Distributed Resource Scheduler, VMware High Availability, VMware Update Manager, and VMware Consolidated Backup.

The VMware ESXi architecture comprises the underlying operating system, called VMkernel, and processes that run on it. VMkernel provides the means for running all processes on the system, including management applications and agents as well as virtual machines. VMkernel also manages all hardware devices on the server, and manages resources for the applications.

The main processes that run on top of VMkernel are:

- Direct Console User Interface (DCUI), which is the low-level configuration and management interface, accessible through the console of the server, and used primarily for initial basic configuration.
- The virtual machine monitor, which is the process that provides the execution environment for a virtual machine, as well as a helper process known as VMX. Each running virtual machine has its own VMM and VMX process.
- Various agents are used to enable high-level VMware Infrastructure management from remote applications.
- The Common Information Model (CIM) system, which is the interface that enables hardware-level management from remote applications through a set of standard APIs.

Figure 1-11 on page 21 shows a components diagram of the overall ESXi 3.5 architecture.

For detailed examination of each of these components, refer to the previously mentioned white paper, *The Architecture of VMware ESX Server 3i*, at:

http://www.vmware.com/files/pdf/ESXServer3i_architecture.pdf

[1] Available from http://www.vmware.com/files/pdf/ESXServer3i_architecture.pdf. This section contains material from VMware. Used with permission.

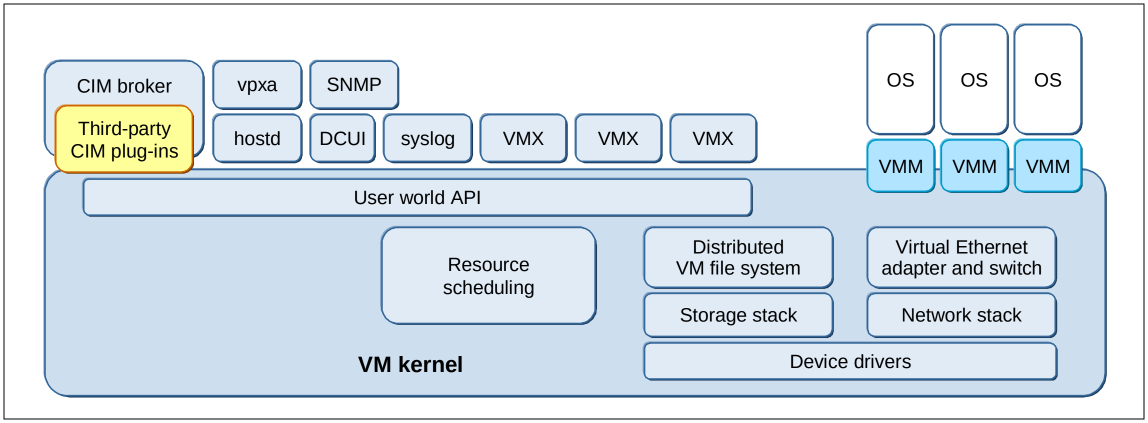


Figure 1-11 The architecture of VMware ESXi eliminates the need for a service console

1.5.2 VMware ESXi on x3850 M2

Although VMware ESXi can be booted from flash memory or installed on a hard disk, ESXi is currently available from IBM only on specific systems, including specific models of the x3850 M2. These models have an integrated bootable USB flash drive that is securely installed to an internal USB port.

With an IBM eX4 server running VMware ESXi (or VMware ESX), applications and services can be deployed in highly reliable and secure virtual machines. Virtual machines can be provisioned, consolidated, and managed centrally without having to install an operating system, thus simplifying the IT infrastructure and driving down total cost of ownership for businesses with constrained IT budgets and resources.

One important recommended consideration when selecting a server to run VMware ESX is to ensure you have sufficient headroom in capacity. The IBM x3850 M2 is optimized for VMware ESXi because of its vertical scalability in the key areas such as processor, memory, and I/O subsystems. VMware discusses the benefits of CPU dense ESX server hosts by saying[2]:

> The chance that the scheduler can find room for a particular workload without much reshuffling of virtual machines will always be better when the scheduler has more CPUs across which it can search for idle time. For this reason, it will generally be better to purchase two four-way ESX Server licenses than to purchase four two-way machines.

[2] See *Tips and Techniques for Implementing Infrastructure Services on ESX Server*, available at: http://www.vmware.com/vmtn/resources/409. Reproduced by permission.

Similarly, two eight-way servers will provide more scheduling flexibility than four 4-way servers. Refer to the white paper, *Tips and Techniques for Implementing Infrastructure Services on ESX Server* available from:

http://www.vmware.com/vmtn/resources/409

Table 1-5 shows that scheduling opportunities scale exponentially rather than linearly when more cores are available.

Table 1-5 Scheduling opportunities scale exponentially when there are more cores

ESX Host Server	Number of Cores	Scheduling opportunities (VM = 2 vCPUs)
4-way Dual Core	8	28
8-way Dual Core	16	120
8-way Quad Core	32	496

1.5.3 Comparing ESXi to other VI3 editions

VMware ESXi is one of the new VMware VI Editions being offered from VMware and provides the same functionality as ESX 3.5. VMware ESXi can be upgraded to VI Foundation, Standard, and Enterprise Editions to provide additional management features as detailed in Table 1-6.

Table 1-6 Feature comparison

Feature	VMware ESXi	VI Foundation	VI Standard	VI Enterprise
VMFS Virtual SMP	Yes	Yes	Yes	Yes
VC Agent - Central management	No	Yes	Yes	Yes
Update manager	No	Yes	Yes	Yes
Consolidated backup	No	Yes	Yes	Yes
High availability	No	No	Yes	Yes
DRS - Resource management	No	No	No	Yes
DPM - Power management	No	No	No	Yes
VMotion - Live VM migration	No	No	No	Yes
Storage VMotion - Live VM disk file migration	No	No	No	Yes

VMware is available in several editions, including:

- VMware Infrastructure Enterprise Edition

  This edition contains the entire array of virtual infrastructure capabilities for resource management, workload mobility, and high availability. It includes:

  - VMware ESX Server
  - VMware ESXi
  - VMware Consolidated Backup
  - VMware Update Manager
  - VMware VMotion
  - VMware Storage VMotion
  - VMware DRS with Distributed Power Management (DPM)
  - VMware HA

- VMware Infrastructure Standard Edition

  This edition is designed to bring higher levels of resiliency to IT environments at greater value. It includes:

  - VMware HA
  - VMware ESX Server
  - VMware ESXi
  - VMware Consolidated Backup
  - VMware Update Manager

- VMware Infrastructure Foundation Edition

  Unlike the previous VMware Infrastructure 3 Starter Edition, VMware Infrastructure Foundation Edition has no restrictions on shared storage connectivity, memory utilization, or number of CPUs of the physical server. It includes:

  - VMware ESX Server
  - VMware ESXi
  - VMware Consolidated Backup
  - VMware Update Manager

  New features such as VMware High Availability (VMware HA), Distributed Resource Scheduler (DRS), and Consolidated Backup provide higher availability, guaranteed service level agreements, and quicker recovery from failures than was previously possible, and comes close to the availability you get from more expensive and complicated alternatives such as physically clustered servers.

Note: For all VMware VI3 Infrastructure editions (Enterprise, Standard, and Foundation), two Socket licenses must be purchased with a corresponding subscription and support for the VI3 Edition purchased. The licenses are also valid for use with ESXi Installable Edition. VMware ESXi is now available free of cost, with no subscription required, however additional VI3 features are licensed separately.

- VMware Infrastructure ESXi Edition

  This edition has no restrictions on shared storage connectivity, memory utilization, or number of CPUs of the physical server. However, if you purchase IBM x3850 M2 with VMware ESXi integrated hypervisor and subsequently require additional functionality, you can upgrade ESXi to the VI Enterprise, Standard, or Foundation Editions. See "License upgrades from ESXi to VI3 Editions" on page 25 for details about upgrade options.

The System x3850 M2 and x3950 M2 servers are designed for balanced system performance, and are therefore uniquely positioned to take advantage of the larger workloads now available to be virtualized.

Table 1-7 shows the limitations of each VMware distribution that is supported on the x3850 M2 and x3950 M2 (single node).

Table 1-7 Features of the VMware ESX family

Feature	ESX Server 3.0.2 update 1	VMware ESXi VMware ESX V3.5
Maximum logical CPUs[a]	32	32 (64 logical CPUs are supported experimentally by VMware)
Maximum memory	64 GB	256 GB
Size of RAM per virtual machine	16,384 MB	65,532 MB

[a] Each core is equal to a logical CPU.

Note: The values in the table are correct at the time of writing and may change as testing completes. The values do not reflect the theoretical values but set the upper limit of support for either distribution.

For more information about the configuration maximums of ESX Server, see:

- ESX Server 3.0.2 configuration maximums:

  http://www.vmware.com/pdf/vi3_301_201_config_max.pdf
- VMware ESX V3.5 and VMware ESXi V3.5 configuration maximums:

  http://www.vmware.com/pdf/vi3_35/esx_3/r35/vi3_35_25_config_max.pdf

1.5.4 VMware ESXi V3.5 licensing

As described in 1.2.2, “x3850 M2 model information” on page 10, specific hypervisor models of the x3850 M2 includes VMware ESXi V3.5, the embedded virtualization engine on an IBM customized USB Flash Disk. These models include a license for VMware ESXi V3.5 for up to four processor sockets.

> **Note:** VMware ESXi is available only in a dedicated model of the x3850 M2 (as described in 1.2.2, “x3850 M2 model information” on page 10) or in configure-to-order (CTO) models pre-loaded as a Factory Install (product number 5773-VMW).

Subscriptions for updates

In addition, subscriptions for updates to VMware ESXi V3.5 are recommended, but not mandatory, to be purchased for each ESXi V3.5 (four-socket) license using product number 5773-VMW:

- Subscription for two processor sockets: Feature code 0997
- Subscription for four processor sockets: Feature code 0998

For more details, see the IBM Announcement Letter 208-071:

http://www.ibm.com/isource/cgi-bin/goto?it=usa_annred&on=208-071

License upgrades from ESXi to VI3 Editions

VMware ESXi can be upgrade to provide the additional features available in the VMware Infrastructure Enterprise, Standard or Foundation Editions, with a purchase of licenses from IBM as shown in Table 1-8 on page 26.

Table 1-8 VMware license, subscription, support options for ESX 3.5 and ESXi

Description	Quantity for [x3850 M2 / x3950 M2] (single node) with 2 sockets	Quantity for [x3850 M2 / x3950 M2] (single node) with 4 sockets
VMware ESX 3i to [Enterprise, Standard, or Foundation] upgrade, 2-socket license only	1	2
Subscription only VMware ESX 3i to [Enterprise, Standard, or Foundation] upgrade, 2-socket,1 or 3 year support	1	2
Virtual Infrastructure 3, [Enterprise, Standard, or Foundation], 2-socket, 1 or 3 year support	1	2

For details of part numbers, refer to VMware Offerings in the IBM System x Configuration and Options Guide:

http://www.ibm.com/systems/xbc/cog/vmwareesx.html

For example, to upgrade an ESXi 4-socket license for x3850 M2 hypervisor model (with 4 x processor sockets populated), purchase the following items:

- Two VMware ESX Server 3i to Enterprise Upgrade, 2-socket license only
- Two Subscription Only VMware ESX Server 3i to Enterprise Upgrade, 2-socket, 3-year support
- Two VMware Infrastructure 3, Enterprise, 2-socket, 3-year support

The exact description of the parts above might differ slightly from country to country, or by the length (in years) of subscription and support. License upgrades, subscription upgrades, and support must be purchased as a complete set to upgrade ESXi Edition to Virtual Infrastructure Foundation, Standard, and Enterprise Editions.

1.5.5 Support for applications running on VMware ESX and ESXi

Ensure that the applications you plan to use on the x3850 M2 and x3950 M2 running VMware ESX Server are supported by the application vendor:

- Microsoft

  See the following Microsoft support Web site for details about its support of applications and operating systems running on ESX Server:

  http://support.microsoft.com/kb/897615/

- IBM software

  If you are running IBM software such as WebSphere®, Lotus®, and Tivoli® products on VMware ESX Server, you must have an IBM Remote Technical Support ServicePac® or IBM VMware Support Line agreement through the IBM Support Line or the IBM equivalent. You must have a current Software Maintenance Agreement to receive support for the IBM software products in this environment. Individual IBM software products can have a level of customer support beyond that described in the product announcement. If applicable, information about the added support will be included in the specific product announcement letter.

  VMware ESX software is designed to be transparent to the middleware and applications that operate above the VMware guest operating system. If an IBM software problem occurs only within a VMware ESX environment, it will be considered a transparency problem and will be directed to VMware for resolution. The IBM VMware Support Line is available to provide assistance in working with the customer and the VMware Business Partner to resolve this type of problem.

  Customer implementations that are not covered by an IBM VMware Support Line agreement are required to re-create the problem in a native environment without VMware in order to use Software Maintenance support services for the IBM software that is experiencing the problem.

1.6 IBM fourth generation XA-64e chipset

The x3850 M2 and x3950 M2 use the fourth generation of the IBM XA-64e or eX4 chipset.

This chipset is designed for the Xeon MP processor family from Intel. The IBM eX4 chipset provides enhanced functionality and capability with significant improvements in scalability, decreased memory latency, increased memory bandwidth and increased I/O bandwidth. The architecture consists of the following components:

- One to four Xeon dual-core, quad-core, or 6-core processors
- Hurricane 4 Memory and I/O Controller (MIOC)
- Eight high-speed memory buffers
- Two PCI Express bridges
- One Enterprise Southbridge Interface

Figure 1-12 on page 28 shows a block diagram of the x3850 M2 and x3950 M2.

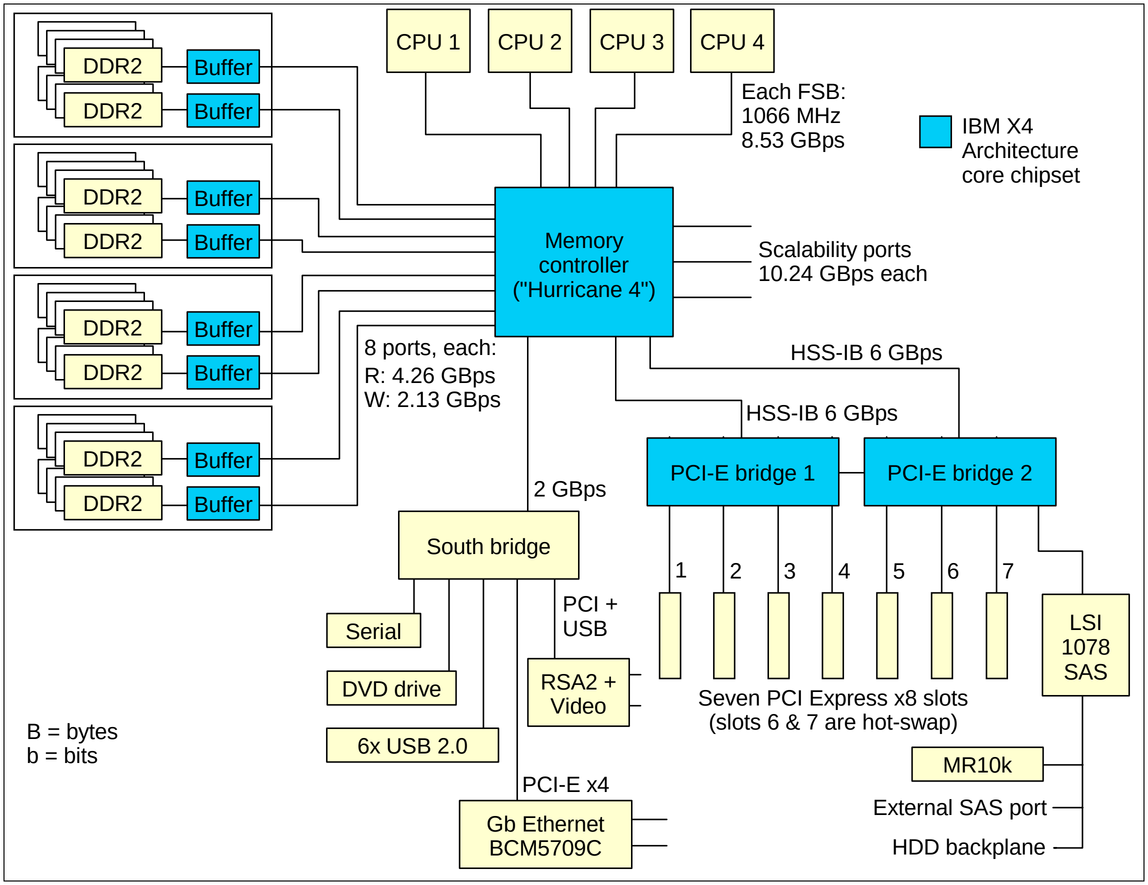


Figure 1-12 x3850 M2 and x3950 M2 system block diagram

Each memory port out of the memory controller has a peak read throughput of 4.26 GBps and a peak write throughput of 2.13 GBps. DIMMs are installed in matched pairs, two-way interleaving, to ensure the memory port is fully utilized. Peak throughput for each PC2-5300 DDR2 DIMM is 4.26 GBps.

There are eight memory ports; spreading the installed DIMMs across all ports can improve performance. The eight independent memory ports provide simultaneous access to memory. With four memory cards installed, and eight DIMMs in each card, peak read memory bandwidth is 34.1 GBps and peak write bandwidth is 17.1 GBps. The memory controller routes all traffic from the eight memory ports, four microprocessor ports, and the three PCIe bridge ports.

The memory controller also has an embedded DRAM that, in the x3850 M2 and x3950 M2, holds a snoop filter lookup table. This filter ensures that snoop requests for cache lines go to the appropriate microprocessor bus and not all four of them, thereby improving performance.

The three scalability ports are each connected to the memory controller through individual scalability links with a maximum theoretical bidirectional data rate of 10.24 GBps per port.

IBM eX4 has two PCIe bridges and each are connected to a HSS-IB port of the memory controller with a maximum theoretical bidirectional data rate of 6 GBps. As shown in Figure 1-12 on page 28, PCIe bridge 1 supplies four of the seven PCI Express x8 slots on four independent PCI Express buses. PCIe bridge 2 supplies the other three PCI Express x8 slots plus the onboard SAS devices, including the optional ServeRAID-MR10k and a 4x external onboard SAS port.

A separate Southbridge is connected to the Enterprise Southbridge Interface (ESI) port of the memory controller through a PCIe x4 link with a maximum theoretical bidirectional data rate of 2 GBps. The Southbridge supplies all the other onboard PCI devices, such as the USB ports, onboard Ethernet and the standard RSA II.

1.6.1 Hurricane 4

Hurricane 4 is the north bridge component of the IBM eX4 chipset designed for latest Intel Core™ Architecture-based processors which feature a new architecture for the processor front-side bus. Hurricane 4 supports the processors in the Xeon 7000 family of processors, including those with code names of *Tigerton* and *Dunnington*.

Hurricane 4 is an enhanced memory controller with Level 4 (L4) scalability cache. Hurricane 4 contains processor scalability support for up to 16 sockets across four NUMA nodes. Hurricane 4 provides the following features:

- Reduced local and remote latency compared to the X3 chipset in the x3950
- Integrated memory controller, NUMA controller, two I/O channels and three scalability ports
- Local memory used for scalability L4 cache for a multinode environment
- Connectivity to high speed memory hub module, two PCIe bridges, and a Southbridge
- Scalability to 16 socket SMP system providing industry leading performance
- Support for four front-side buses, one for each of the four Intel Xeon® MP processors

1.6.2 XceL4v dynamic server cache

The XceL4v dynamic server cache is a technology developed as part of the IBM XA-64e fourth-generation chipset. It is used in two ways:

- As a single four-way server, the XceL4v and its embedded DRAM (eDRAM) is used as a snoop filter to reduce traffic on the front-side bus. It stores a directory of all processor cache lines to minimize snoop traffic on the four front-side buses and minimize cache misses.
- When the x3950 M2 is configured as a multinode server, this technology dynamically allocates 256 MB of main memory in each node for use as an L4 cache directory and scalability directory. This means an eight-way configuration has 512 MB of XceL4v cache.

Used in conjunction with the XceL4v Dynamic Server Cache is an embedded DRAM (eDRAM), which in single-node configurations contains the snoop filter lookup tables. In a multinode configuration, this eDRAM contains the L4 cache directory and the scalability directory.

Note: The amount of memory that BIOS reports is minus the portion used for the XceL4v cache.

1.6.3 PCI Express I/O bridge chip

Two single-chip PCIe host bridges are designed to support PCIe adapters for IBM x3850 M2 and x3950 M2 servers. The PCIe bridges each have one HSS-IB port that provide connectivity to Hurricane 4 memory controller chip and also another HSS-IB link between the PCIe bridges for redundancy in the event one of the links from the Hurricane 4 chipset to the two PCIe bridges are not working. The HSS-IB links are capable of up to 3.0 GBps bandwidth in each direction per port or up to 6.0 GBps bidirectional bandwidth (see Figure 1-12 on page 28).

Each PCIe chip provides four separate PCIe x8 buses to support four PCIe x8 slots. PCIe Bridge 1 supports slots 1-4 of the PCIe x8 slots and PCIe Bridge 2 supports slots 5-7 of the PCIe x8 slots and a dedicated PCIe x8 slot for ServeRAID MR10k SAS/SATA RAID controller.

1.6.4 High-speed memory buffer chips

The two high-speed memory buffer chips on each memory card are used to extended memory capacity. They provide the necessary functions to connect up to 32 8-byte ranks of DDR-II memory (see 1.6.5, “Ranks” on page 31 for an explanation of ranks of memory). Each buffer supports multiple data flow modes that allow configurable combinations of controller data bus widths. These modes allow its bandwidth and capacity to be optimized for many system configurations.

1.6.5 Ranks

A *rank* is a set of DRAM chips on a DIMM that provides eight bytes (64 bits) of data.

Memory in servers is implemented in the form of DIMMs, which contain a number of SDRAM (or just DRAM) chips.

The capacity of each DRAM is a number of *words* where each word can be 4 bits (x4), 8 bits (x8) and, starting to become prevalent, 16 bits in length (x16).

The word length is usually written as x4 for 4 bits, and so on. The number of words in the DRAM is sometimes written on the label of the DIMM, such as a DRAM chip on a DIMM.

DIMMs are typically configured as either single-rank or double-rank devices but four-rank devices are becoming more prevalent.

The DRAM devices that make up a rank are often, but not always, mounted on one side of the DIMM, so a single-rank DIMMs can also be referred to as a single-sided DIMM. Likewise a double-ranked DIMM can be referred to as a double-sided DIMM.

Refer to “Chapter 8, Memory subsystem” of the IBM Redbooks publication *Tuning IBM System x Servers for Performance*, SG24-5287 for more details.

1.6.6 Comparing IBM eX4 to X3 technologies

This section discusses the improvements in the design of IBM eX4 technology as compared to the design of previous X3 technology. A block diagram of the X3 technology is shown in Figure 1-13 on page 32.

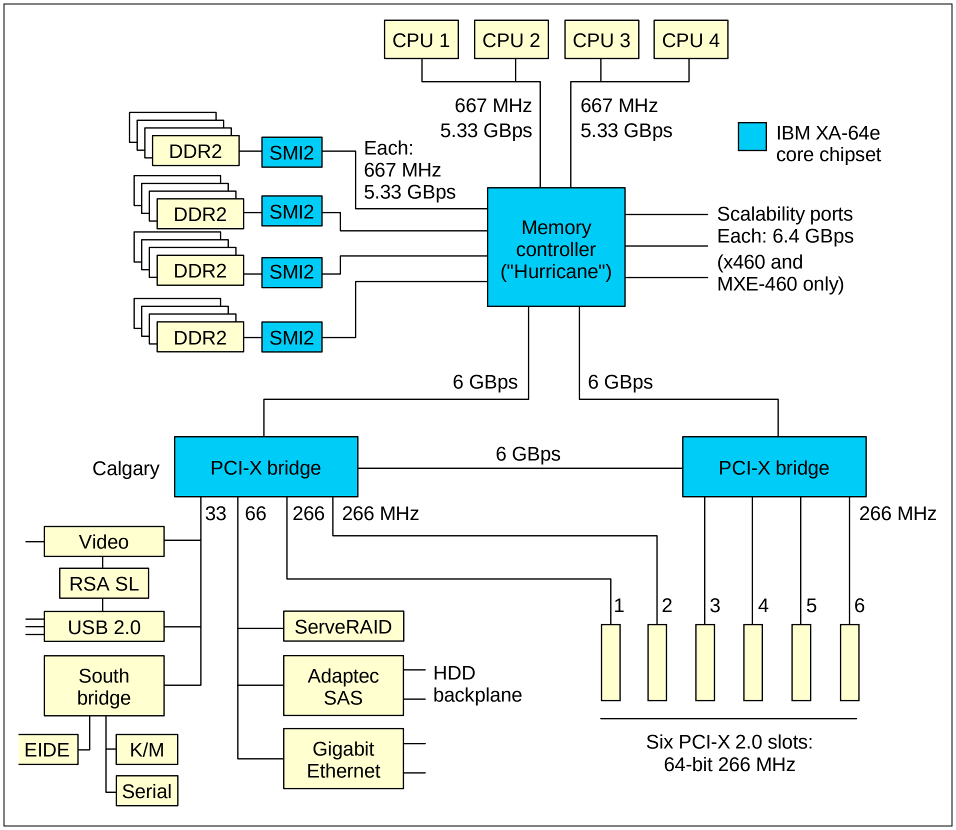


Figure 1-13 Block diagram of IBM X3 x3850/x366 and x3950/x460

IBM eX4 technology builds and improves upon its previous generation X3 technology. The key enhancements are:

- Processor interface
  - Quad 1066 MHz front-side bus (FSB), which has a total bandwidth of up to 34.1 GBps. In X3, the maximum bandwidth was 10.66 GBps.
  - The front-side bus is increased to 1066 MHz from 667 MHz for 3.2x bandwidth improvement.
  - Snoop filter is for quad FSB coherency tracking compared to X3 with only dual FSB coherency tracking.
- Memory
  - Increased (four-fold) memory capacity (2X from chipset, 2X from DRAM technology) compared to X3.

  - Eight memory channels are available in eX4, compared to four memory channels in X3.
  - Memory bandwidth improved 1.6x is eX4. The eX4 has 34.1 GBps read and 17.1 GBps write aggregate peak memory bandwidth versus 21.33 GBps aggregate peak memory bandwidth in X3.
  - Lower memory latency in the eX4 because the eX4 uses DDR2 533 MHz memory compared to DDR2 333 MHz memory in X3.
- Direct connect I/O
  - Earlier X3 models used a PCI-X chipset instead of the PCI Express chipset in eX4.
  - The ServeRAID MR10K SAS/SATA RAID controller in the x3950 M2 no longer shares bandwidth on a shared bus such as the ServeRAID 8i SAS RAID controller in the x3950 did with devices like the Gigabit Ethernet controllers in X3. With its own dedicated PCIe slot, the ServeRAID MR10k has improved throughput and ability to support external SAS devices through the integrated external SAS port.
  - Dedicated Southbridge ESI port to support Southbridge devices such as RSA II, dual Gigabit Ethernet controllers, IDE DVD-ROM, USB port, Serial port and Video interface.
- Scalability
  - Almost twice the increase in scalability port bandwidth for improved scaling. The eX4 has three scalability ports with increased bandwidth of 30.72 GBps compared to 19.2 GBps in X3.

1.7 Processors

As mentioned previously, the x3850 M2 and x3950 M2 models use one of the following Intel Xeon Processor models:

Tigerton (code name) processors:

- Xeon 7200 series dual-core processors
- Xeon 7300 series quad-core processors

Dunnington (code name) processors:

- Xeon 7400 series quad-core processors
- Xeon 7400 series 6-core processors

Refer to 1.2, “Model numbers and scalable upgrade options” on page 9 for details about the current models.

All standard models of the x3850 M2 and x3950 M2 have two processors installed. One, two, three, or four processors are supported. Installed processors must be identical in model, speed, and cache size.

Tip: For the purposes of VMware VMotion, the Dunnington processors are compatible with the Tigerton processors.

As described in 1.3, “Multinode capabilities” on page 14, you can also connect multiple x3950 M2s to form larger single-image configurations.

The processors are accessible from the top of the server after opening the media hood. The media hood is hinged at the middle of the system and contains the SAS drives, optical media, USB ports and light path diagnostic panel.
Figure 1-14 shows the media hood half-way open.

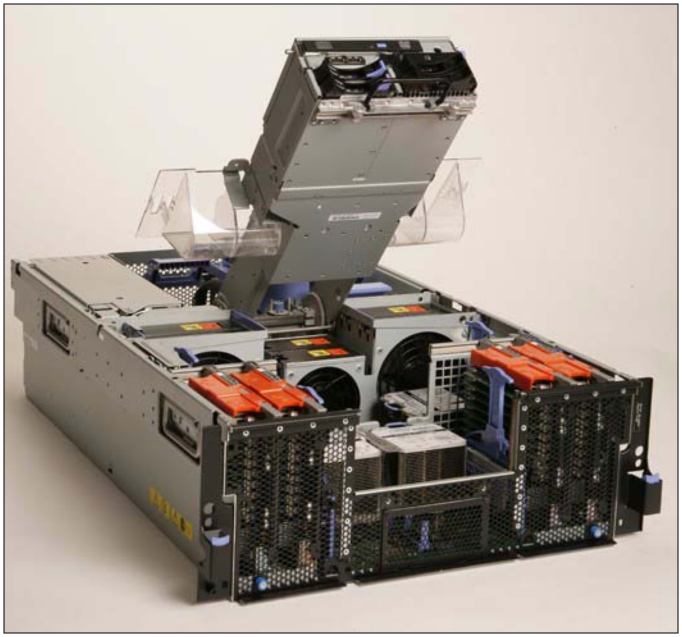


Figure 1-14 The x3950 M2 with the media hood partially open

The processors are each packaged in the 604-pin Flip Chip Micro Pin Grid Array (FC-mPGA) package. It is inserted into surface-mount mPGA604 socket. The processors use a large heat-sink to meet thermal specifications.

The Xeon E7210 and E7300 Tigerton processors have two levels of cache on the processor die:

- Each pair of cores in the processor has either 2, 3, or 4 MB shared L2 cache for a total of 4, 6, or 8 MB of L2 cache. The L2 cache implements the Advanced Transfer Cache technology.
- L1 execution trace cache in each core is used to store micro-operations and decoded executable machine instructions. It serves those to the processor at rated speed. This additional level of cache saves decode-time on cache-hits.

The Tigerton processors do not have L3 cache.

Figure 1-15 compares the layout of the Tigerton dual-core and quad-core processors.

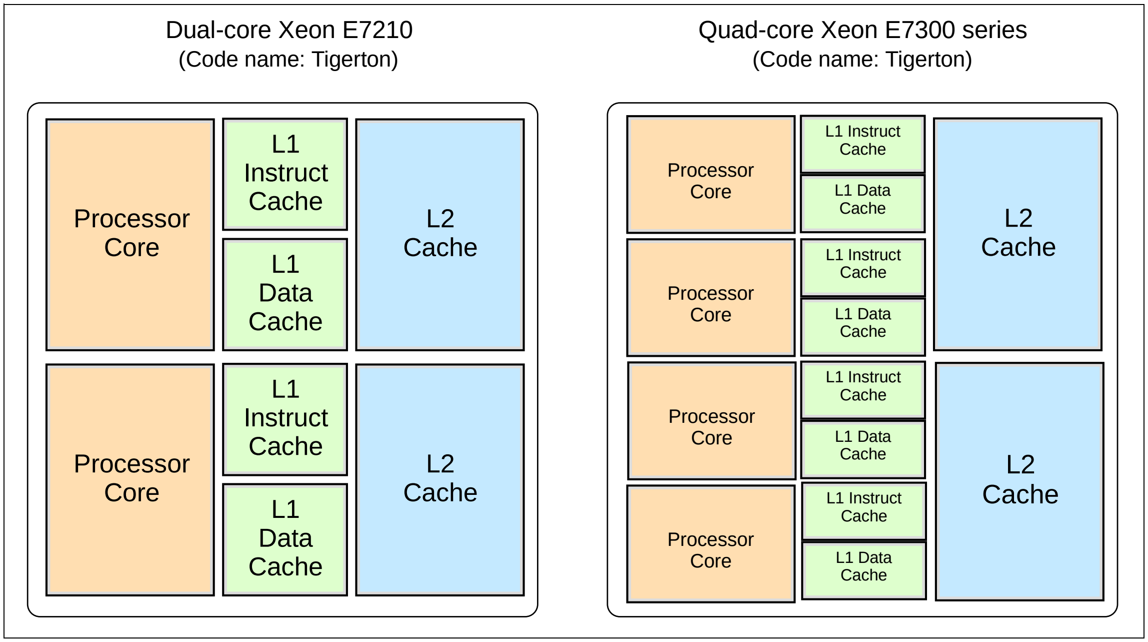


Figure 1-15 Comparing the dual-core and quad-core Tigerton

The Xeon E7400 Series Dunnington processors, both 4-core and 6-core models, have shared L2 cache between each pair of cores but also have a shared L3 cache across all cores of the processor. While technically all Dunnington Xeon processors have 16MB of L3 cache, 4-core models only have 12MB of L3 cache enabled and available. See Figure 1-16 on page 36.

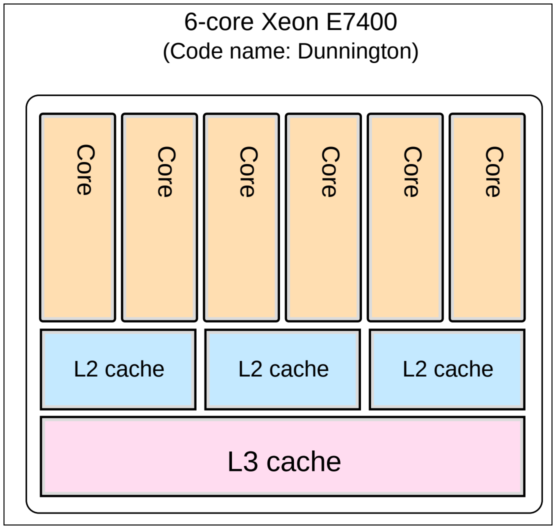


Figure 1-16 Block diagram of Dunnington 6-core processor package

Key features of the processors used in the x3850 M2 and x3950 M2 include:

- Multi-core processors

  The Tigerton dual-core processors are a concept similar to a two-way SMP system except that the two processors, or *cores*, are integrated into one silicon die. This brings the benefits of two-way SMP with lower software licensing costs for application software that licensed per CPU socket plus the additional benefit of less processor power consumption and faster data throughput between the two cores. To keep power consumption down, the resulting core frequency is lower, but the additional processing capacity means an overall gain in performance.

  The Tigerton quad-core processors add two more cores onto the same die, and some Dunnington processors also add two more. Hyper-Threading Technology is not supported.

  Each core has separate L1 instruction and data caches, and separate execution units (integer, floating point, and so on), registers, issue ports, and pipelines for each core. A multi-core processor achieves more parallelism than Hyper-Threading Technology because these resources are not shared between the two cores.

  With two times, four times, and even six times the number of cores for the same number of sockets, it is even more important that the memory subsystem is able to meet the demand for data throughput. The 34.1 GBps peak throughput of the x3850 M2 and x3950 M2's eX4 technology with four memory cards is well suited to dual-core and quad-core processors.

- 1066 MHz front-side bus (FSB)

  The Tigerton and Dunnington Xeon MPs use two 266 MHz clocks, out of phase with each other by 90°, and using both edges of each clock to transmit data. This is shown in Figure 1-17.

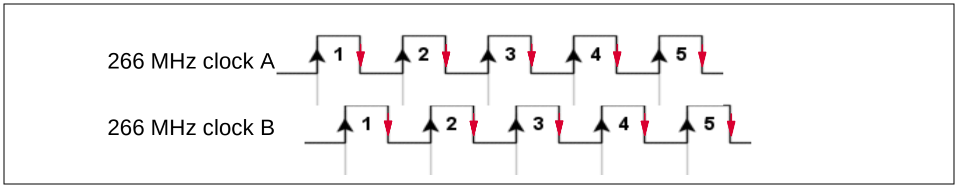


Figure 1-17 Quad-pumped front-side bus

  A quad-pumped 266 MHz bus therefore results in a 1066 MHz front-side bus.

  The bus is 8 bytes wide, which means it has an effective burst throughput of 8.53 GBps. This can have a substantial impact, especially on TCP/IP-based LAN traffic.

- Intel 64 Technology (formerly known as EM64T)

  Intel 64 Technology is a 64-bit extension to the industry-standard IA32 32-bit architecture. Intel 64 Technology adds:

  - A set of new 64-bit general purpose registers (GPRs)
  - 64-bit instruction pointers
  - The ability to process data in 64-bit chunks

  Although the names of these extensions suggest that the improvements are simply in memory addressability, Intel 64 Technology is, in fact, a fully functional 64-bit processor.

  The processors in the x3850 M2 and x3950 M2 include the Intel 64 Technology extensions from Intel. This technology is compatible with IA-32 software while enabling new software to access a larger memory address space.

  To realize the full benefit of this technology, you must have a 64-bit operating system and 64-bit applications that have been recompiled to take full advantage of this architecture. Existing 32-bit applications running on a 64-bit operating system can also benefit from EM64T.

  The Tigerton processors limit memory addressability to 40 bits of addressing.

  Intel 64 Technology provides three distinct operation modes:

  - 32-bit legacy mode

    The first and, in the near future, probably most widely used mode will be the *32-bit legacy mode*. In this mode, processors with Intel 64 Technology act just like any other IA32-compatible processor. You can install your 32-bit operating system on such a system and run 32-bit applications, but you cannot use the new features such as the flat memory addressing above 4 GB or the additional GPRs. Thirty-two-bit applications run as fast as they would on any current 32-bit processor.

Most of the time, IA32 applications run even faster because numerous other improvements boost performance regardless of the maximum address size.

– Compatibility mode

  The *compatibility mode* is an intermediate mode of the full 64-bit mode, which we describe next. To run in compatibility mode, you have to install a 64-bit operating system and 64-bit drivers. When a 64-bit OS and drivers are installed, the processor can support both 32-bit applications and 64-bit applications.

  The compatibility mode provides the ability to run a 64-bit operating system while still being able to run unmodified 32-bit applications. Each 32-bit application still is limited to a maximum of 4 GB of physical memory. However, the 4 GB limit is now imposed on a per-process level, not on a system-wide level. This means that every 32-bit process on this system gets its very own 4 GB of physical memory space, provided sufficient physical memory is installed. This is already a huge improvement compared to IA32, where the operating system kernel and the application had to share 4 GB of physical memory.

  Because the compatibility mode does not support the virtual 8086 mode, real-mode applications are not supported. However, sixteen-bit protected mode applications are supported.

– Full 64-bit mode

  The *full 64-bit mode* is referred to by Intel as the *IA-32e mode*. (For AMD™, it is the *long mode*.) This mode is applied when a 64-bit OS and 64-bit application are used. In the full 64-bit operating mode, an application can have a virtual address space of up to 40 bits, equating to one terabyte (TB) of addressable memory. The amount of physical memory is determined by how many DIMM slots the server has, and the maximum DIMM capacity supported and available at the time.

  Applications that run in full 64-bit mode have access to the full physical memory range, depending on the operating system, and also have access to the new GPRs as well as to the expanded GPRs. However, it is important to understand that this mode of operation requires not only a 64-bit operating system (and, of course, 64-bit drivers) but also a 64-bit application that has been recompiled to take full advantage of the various enhancements of the 64-bit addressing architecture.

For more information about the features of the Xeon processors, go to:

- Intel server processors:

  http://www.intel.com/products/server/processors/index.htm?iid=process+server

- Intel Xeon processor 7000 sequence:

  http://www.intel.com/products/processor/xeon7000/index.htm?iid=servproc+body_xeon7000subtitle

For more information about Intel 64 architecture, see:

http://www.intel.com/technology/intel64/index.htm

1.8 Memory subsystem

The standard x3850 M2 and x3950 M2 models have either 4 GB or 8 GB of RAM standard, implemented as four or eight 1 GB DIMMs.

Memory DIMMs are installed in the x3850 M2 and x3950 M2 using memory cards, each card has eight DIMM sockets. The server supports up to four memory cards, giving a total of up to 32 DIMM sockets.

Some models have two memory cards, others have all four cards as standard.

Using 8 GB DIMMs in every socket, the server can hold 256 GB of RAM. With four nodes, the combined complex can hold up to 1 TB RAM.

With a multinode configuration, the memory in all nodes is combined to form a single coherent physical address space. For any given region of physical memory in the resulting system, certain processors are *closer* to the memory than other processors. Conversely, for any processor, some memory is considered *local* and other memory is *remote*. The system's partition descriptor table ensures that memory is used in the most optimal way.

The memory is two-way interleaved, meaning that memory DIMMs are installed in pairs. Figure 1-12 on page 28 shows eight ports from the Hurricane 4 memory controller to memory, with each supporting up to 4.26 GBps read data transfers and 2.13 GBps write data transfers.

The DIMMs operate at 533 MHz, to be in sync with a front-side bus. However the DIMMs are 667 MHz PC2-5300 spec parts because they have better timing parameters than the 533 MHz equivalent. The memory throughput is 4.26 GBps, or 533 MHz x 8 bytes per memory port for a total of 34.1 GBps with four memory cards.

See 3.2, “Memory subsystem” on page 111 for further discussion of how memory is implemented in the x3850 M2 and x3950 M2 and what you should consider before installation.

A number of advanced features are implemented in the x3850 M2 and x3950 M2 memory subsystem, collectively known as *Active Memory*:

- Memory ProteXion

  The Memory ProteXion feature (also known as *redundant bit steering*) provides the equivalent of a hot-spare drive in a RAID array. It is based in the memory controller, and it enables the server to sense when a chip on a DIMM has failed and to route the data around the failed chip.

  Normally, 128 bits of every 144 are used for data and the remaining 16 bits are used for error checking and correcting (ECC) functions. However, the x3850 M2 and x3950 M2 require only 12 bits to perform the same ECC functions, thus leaving 4 bits free. In the event that a chip failure on the DIMM is detected by memory scrubbing, the memory controller can reroute data around that failed chip through these spare bits.

  It reroutes the data automatically without issuing a Predictive Failure Analysis® (PFA) or light path diagnostics alerts to the administrator, although an event is recorded to the service processor log. After the second DIMM failure, PFA and light path diagnostics alerts would occur on that DIMM as normal.

- Memory scrubbing

  Memory scrubbing is an automatic daily test of all the system memory that detects and reports memory errors that might be developing before they cause a server outage.

  Memory scrubbing and Memory ProteXion work in conjunction and do not require memory mirroring to be enabled to work properly.

  When a bit error is detected, memory scrubbing determines whether the error is recoverable:

  - If the error is recoverable, Memory ProteXion is enabled, and the data that was stored in the damaged locations is rewritten to a new location. The error is then reported so that preventative maintenance can be performed. If the number of good locations is sufficient to allow the proper operation of the server, no further action is taken other than recording the error in the error logs.
  - If the error is not recoverable, memory scrubbing sends an error message to the light path diagnostics, which then turns on the proper lights and LEDs to guide you to the damaged DIMM. If memory mirroring is enabled,

the mirrored copy of the data from the damaged DIMM is used until the system is powered down and the DIMM replaced.

As x3850 M2 and x3950 M2 is now capable of supporting large amounts of memory, IBM has added the **Initialization Scrub Control** setting to the BIOS, to let customers choose when this scrubbing is done and therefore potentially speed up the boot process. See 3.2.3, “Memory mirroring” on page 118for more details on these settings.

- Memory mirroring

  Memory mirroring is roughly equivalent to RAID-1 in disk arrays, in that usable memory is halved and a second copy of data is written to the other half. If 8 GB is installed, the operating system sees 4 GB once memory mirroring is enabled. It is disabled in the BIOS by default. Because all mirroring activities are handled by the hardware, memory mirroring is operating-system independent.

  When memory mirroring is enabled, certain restrictions exist with respect to placement and size of memory DIMMs and the placement and removal of memory cards. See 3.2, “Memory subsystem” on page 111 and 3.2.3, “Memory mirroring” on page 118for details.

- Chipkill memory

  Chipkill is integrated into the XA-64e chipset, so it does not require special Chipkill DIMMs and is transparent to the operating system. When combining Chipkill with Memory ProteXion and Active Memory, the x3850 M2 and x3950 M2 provides very high reliability in the memory subsystem.

  When a memory chip failure occurs, Memory ProteXion transparently handles the rerouting of data around the failed component as previously described. However, if a further failure occurs, the Chipkill component in the memory controller reroutes data. The memory controller provides memory protection similar in concept to disk array striping with parity, writing the memory bits across multiple memory chips on the DIMM. The controller is able to reconstruct the missing bit from the failed chip and continue working as usual. One of these additional failures can be handled for each memory port for a total of eight Chipkill recoveries.

- Hot-add and hot-swap memory

  The x3850 M2 and x3950 M2 support the replacing of failed DIMMs while the server is still running. This hot-swap support works in conjunction with memory mirroring. The server also supports adding additional memory while the server is running. Adding memory requires operating system support.

  These two features are mutually exclusive. Hot-add requires that memory mirroring be disabled, and hot-swap requires that memory mirroring be enabled. For more information, see 3.2, “Memory subsystem” on page 111.

In addition, to maintain the highest levels of system availability, when a memory error is detected during POST or memory configuration, the server can automatically disable the failing memory bank and continue operating with reduced memory capacity. You can manually re-enable the memory bank after the problem is corrected by using the Setup menu in the BIOS.

Memory ProteXion, memory mirroring, and Chipkill provide the memory subsystem with multiple levels of redundancy. Combining Chipkill with Memory ProteXion allows up to two memory chip failures for each memory port on the x3850 M2 and x3950 M2, for a total of eight failures sustained.

The system takes the following sequence of steps regarding memory failure detection and recovery:

1. The first failure detected by the Chipkill algorithm on each port does not generate a light path diagnostics error because Memory ProteXion recovers from the problem automatically.
2. Each memory port can sustain a second chip failure without shutting down.
3. Provided that memory mirroring is enabled, the third chip failure on that port sends the alert and takes the DIMM offline, but keeps the system running out of the redundant memory bank.

1.9 SAS controller and ports

The x3850 M2 and x3950 M2 have a disk subsystem comprised of an LSI Logic 1078 serial-attached SCSI (SAS) controller and four internal 2.5-inch SAS hot-swap drive bays. The x3850 M2 and x3950 M2 support internal RAID-0 and RAID-1. The optional ServeRAID MR10k provides additional RAID levels and a 256 MB battery-backed cache.

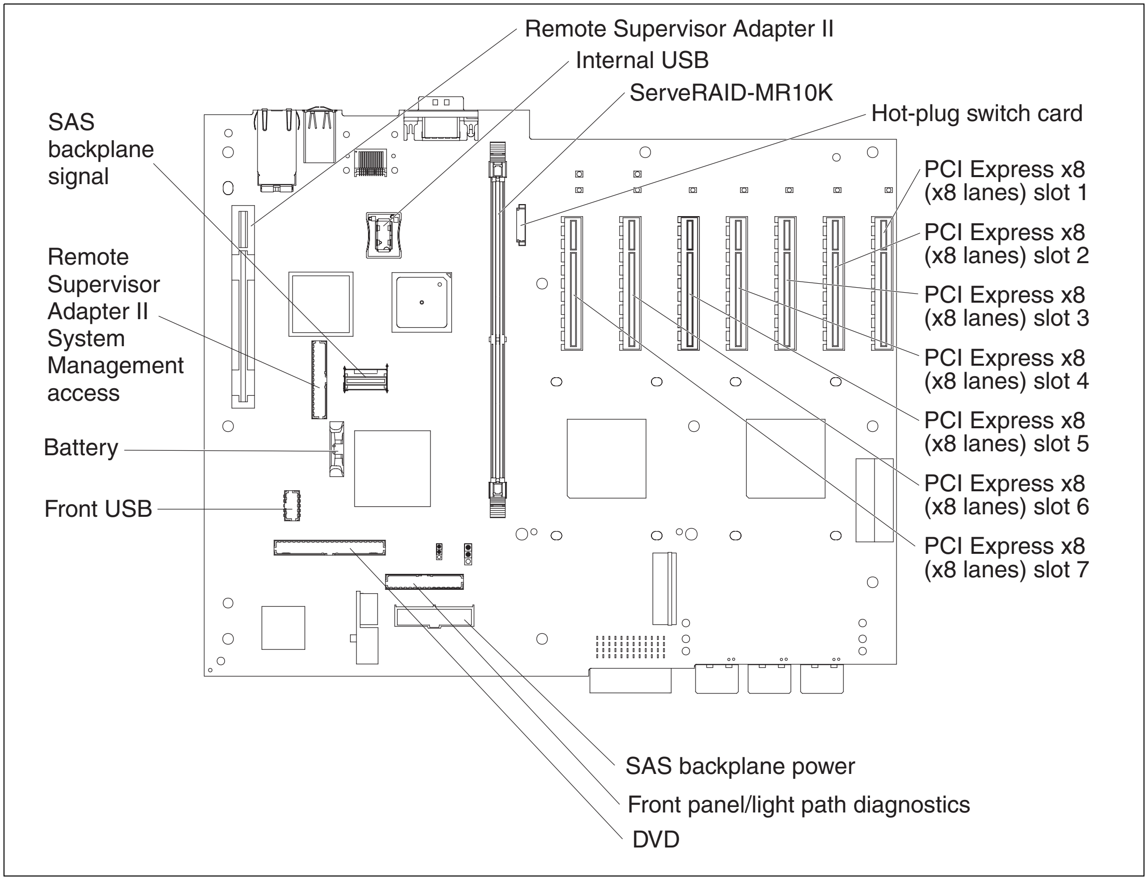
SAS is the logical evolution of SCSI. SAS uses much smaller interconnects than SCSI, while offering SCSI compatibility, reliability, performance, and manageability. In addition, SAS offers longer cabling distances, smaller form factors, and greater addressability.

The x3850 M2 and x3950 M2 has an external SAS x4 port used in conjunction with the optional ServeRAID MR10k. This external port supports SAS non-RAID disk enclosures such as the EXP3000. This port has an SFF-8088 connector.

For more information about the onboard SAS controller and the ServeRAID MR10k daughter card, see Figure 3-22 on page 131 in section 3.3.3, “ServeRAID-MR10k RAID controller” on page 128.

1.10 PCI Express subsystem

As shown in Figure 1-18, five half-length full-height PCI Express x8 slots and two half-length full-height active PCI Express x8 slots are internal to the x3850 M2 and x3950 M2, and all are vacant in the standard models.



Remote Supervisor Adapter II
Internal USB
ServeRAID-MR10K
Hot-plug switch card
SAS backplane signal
PCI Express x8 (x8 lanes) slot 1
PCI Express x8 (x8 lanes) slot 2
Remote Supervisor Adapter II System Management access
PCI Express x8 (x8 lanes) slot 3
PCI Express x8 (x8 lanes) slot 4
PCI Express x8 (x8 lanes) slot 5
Battery
PCI Express x8 (x8 lanes) slot 6
Front USB
PCI Express x8 (x8 lanes) slot 7
SAS backplane power
Front panel/light path diagnostics
DVD

Figure 1-18 System planar layout showing the seven PCI Express slots

All seven slots have the following characteristics:

- Separation of the bus from the other slots and devices
- PCI Express x8
- 40 Gbps full duplex (assumes PCIe (v1.1) x1 capable of maximum 2.5 Gbps unidirectional or half duplex)
- Slots 6 and 7 also support Active PCI hot-swap adapters

Note: Microsoft Windows Server 2003 is required so you can use Active PCI on the x3850 M2 and x3950 M2. Support in Linux distributions is planned for later in 2008.

The optional ServeRAID MR10k adapter does not use a PCIe x8 slot because it has a dedicated PCIe x8 customized 240-pin slot on the I/O board.

The PCI subsystem also supplies these I/O devices:

- LSI 1078 serial-attached SCSI (SAS) controller
- Broadcom dual-port 5709C 10/100/1000 Ethernet
- Six USB ports, two on the front panel, three on the rear, one onboard
- Remote Supervisor Adapter II adapter in a dedicated socket on the I/O board. This adapter also provides the ATI ES1000 16 MB video controller.
- EIDE interface for the DVD-ROM drive (some models)
- Serial port

1.11 Networking

The IBM x3950 M2 and x3850 M2 servers have an integrated dual 10/100/1000 Ethernet controller that uses the Broadcom NetXtreme II BCM5709C controller. The controller contains two standard IEEE 802.3 Ethernet MACs which can operate in either full-duplex or half-duplex mode.

1.11.1 Main features

The Broadcom NetXtreme II dual port Gigabit capable Ethernet ports have the following main features:

- Shared PCIe interface across two internal PCI functions with separate configuration space
- Integrated dual 10/100/1000 MAC and PHY devices able to share the bus through bridge-less arbitration
- IPMI enabled
- TOE acceleration (support for in the x3950 M2 and x3850 M2 is planned but was not available at the time of writing).

Note: These onboard Ethernet controllers do not support iSCSI nor RDMA.

On the back of the server, the top port is Ethernet 1 and the bottom is Ethernet 2. The LEDs for these ports are shown in Figure 1-19.

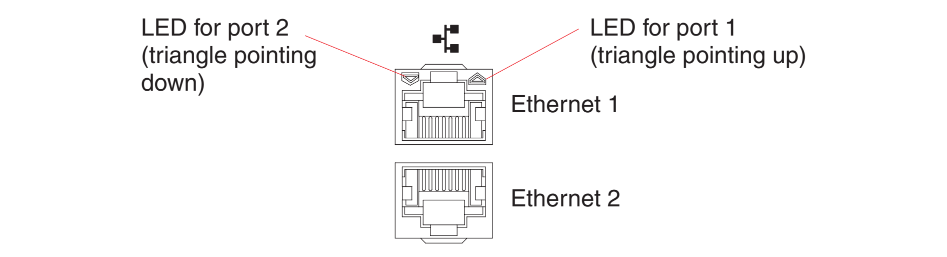

LED for port 2 (triangle pointing down)
LED for port 1 (triangle pointing up)
Ethernet 1
Ethernet 2

Figure 1-19 Onboard dual-port Gigabit Ethernet controller

The LEDs indicate status as follows:

- LED Blink = port has activity
- LED On = port is linked
- LED Off = port is not linked

1.11.2 Redundancy features

The x3850 M2 and x3950 M2 have the following redundancy features to maintain high availability:

- Six hot-swap, multispeed fans provide cooling redundancy and enable individual fan replacement without powering down the server. Each of the three groups of two fans is redundant. In the event of a fan failure, the other fans speed up to continue to provide adequate cooling until the fan can be hot-swapped by the IT administrator. In general, failed fans should be replaced within 48 hours following failure.
- The two Gigabit Ethernet ports can be configured as a team to form a redundant pair.
- The memory subsystem has a number of redundancy features, including memory mirroring and Memory ProteXion, as described in 1.8, “Memory subsystem” on page 39.
- RAID disk arrays are supported for both servers, each with the onboard LSI 1078 for RAID-0 and RAID-1. The optional ServeRAID MR10k provides additional RAID features and a 256 MB battery-backed cache. The x3850 M2 and x3950 M2 has four internal, hot-swap disk drive bays.
- The two, standard 1440 W hot-swap power supplies are redundant in all configurations at 220 V. At 110 V, each power supply draws approximately 720 W and the second power supply is not redundant.

- The 1440 W power supplies have a power factor correction of 0.98 so the apparent power (kVA) is approximately equal to the effective power (W).

The layout of the x3850 M2 and x3950 M2, showing the location of the memory cards, power supplies, and fans is displayed in Figure 1-20.

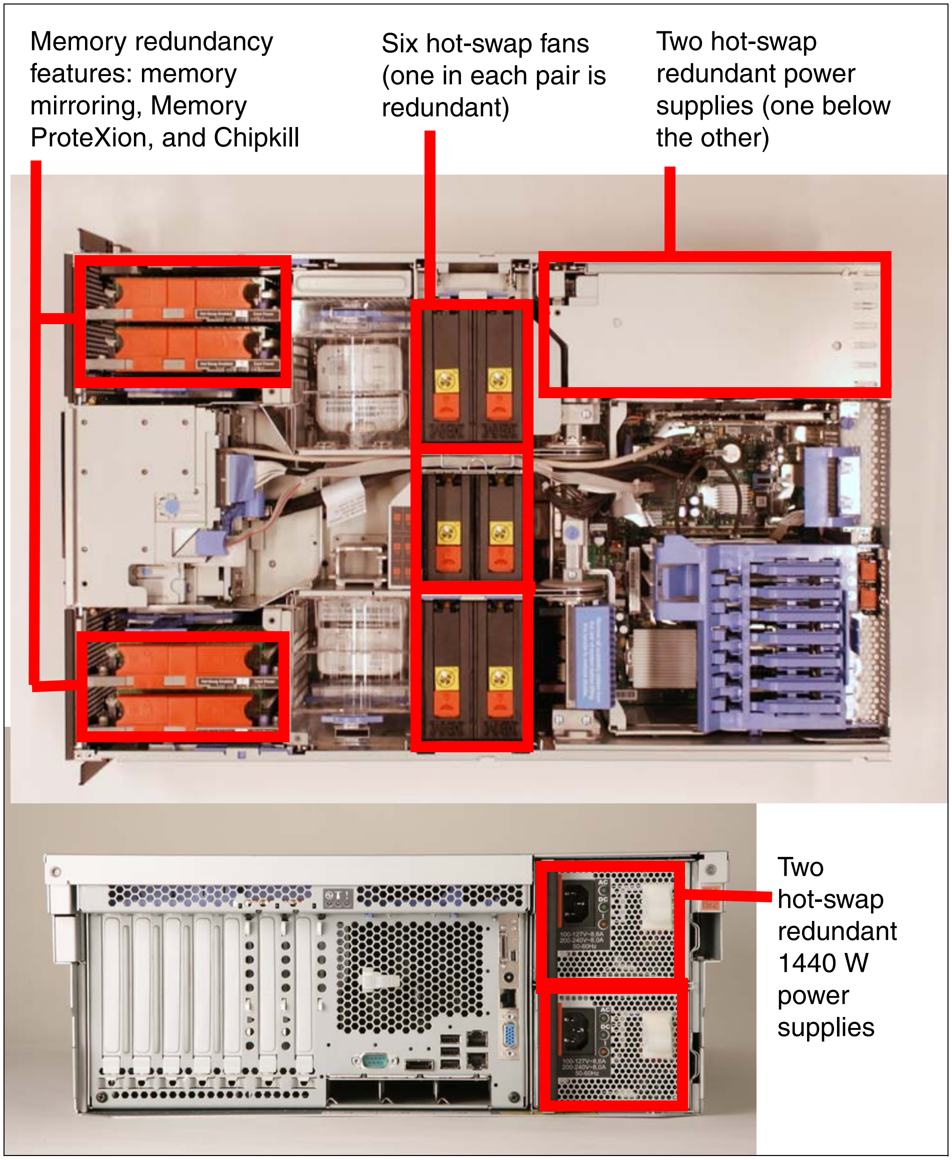

Memory redundancy features: memory mirroring, Memory ProteXion, and Chipkill
Six hot-swap fans (one in each pair is redundant)
Two hot-swap redundant power supplies (one below the other)
Two hot-swap redundant 1440 W power supplies

Figure 1-20 Redundant memory, fans, and power supplies features

1.12 Systems management features

This section provides an overview of the system management features such as Light Path Diagnostics, Baseboard Management Controller, Remote Supervisor Adapter II, and Active Energy Manager for the IBM x3850 M2 and x3950 M2.

1.12.1 Light path diagnostics

To eliminate having to slide the server out of the rack to diagnose problems, a Light Path Diagnostics panel is located at the front of x3850 M2 and x3950 M2, as shown in Figure 1-21. This panel slides out from the front of the server so you can view all light path diagnostics-monitored server subsystems. In the event that maintenance is required, the customer can slide the server out of the rack and, using the LEDs, find the failed or failing component.

Light path diagnostics can monitor and report on the health of microprocessors, main memory, hard disk drives, PCI adapters, fans, power supplies, VRMs, and the internal system temperature. See 6.1, “BMC configuration options” on page 300 for more information.

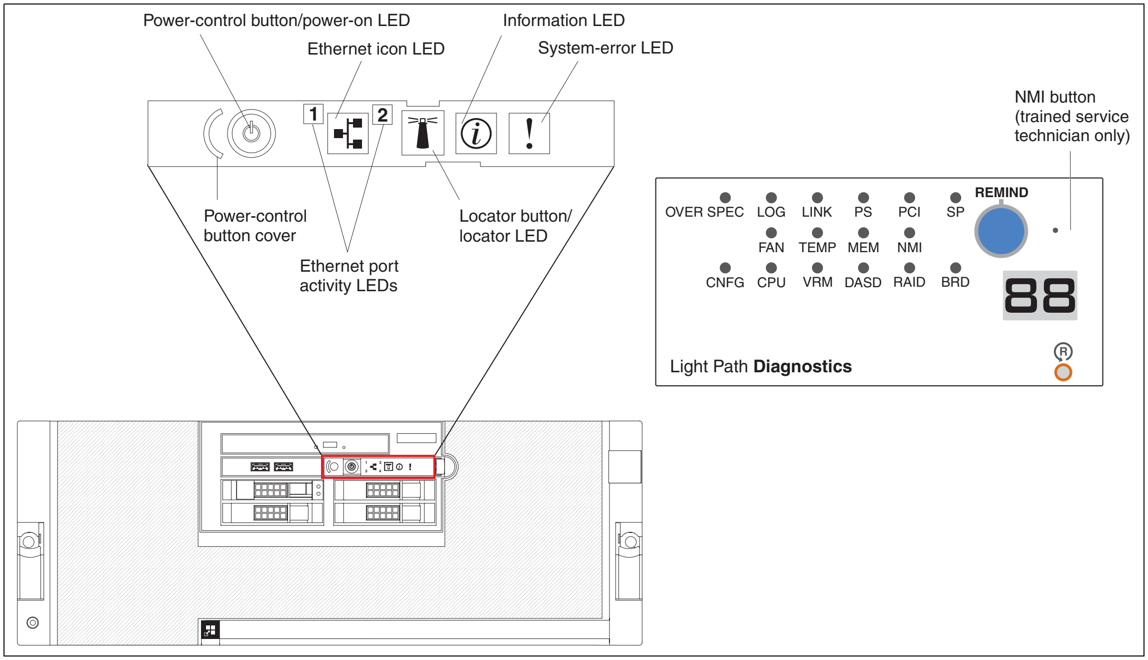

Power-control button/power-on LED
Ethernet icon LED
Information LED
System-error LED
NMI button (trained service technician only)
1
2
Power-control button cover
Ethernet port activity LEDs
Locator button/ locator LED
OVER SPEC
LOG
LINK
PS
PCI
SP
REMIND
FAN
TEMP
MEM
NMI
CNFG
CPU
VRM
DASD
RAID
BRD
88
Light Path Diagnostics

Figure 1-21 Light Path Diagnostics panel

1.12.2 BMC service processor

The baseboard management controller (BMC) is a small, independent controller that performs low-level system monitoring and control functions, and interface functions with the remote Intelligent Platform Management Interface (IPMI). The BMC uses multiple I2C bus connections to communicate out-of-band with other onboard devices. The BMC provides environmental monitoring for the server. If environmental conditions exceed thresholds or if system components fail, the BMC lights the LEDs on the Light Path Diagnostics panel to help you diagnose the problem. It also records the error in the BMC system event log.

The BMC functions are as follows:

- Initial system check when AC power is applied

  The BMC monitors critical I2C devices in standby power mode to determine if the system configuration is safe for power on.

- BMC event log maintenance

  The BMC maintains and updates an IPMI-specified event log in nonvolatile storage. Critical system information is recorded and made available for external viewing.

- System power state tracking

  The BMC monitors the system power state and logs transitions into the system event log.

- System initialization

  The BMC has I2C access to certain system components that might require initialization before power-up.

- System software state tracking

  The BMC monitors the system and reports when the BIOS and POST phases are complete and the operating system has booted.

- System event monitoring

  During run time, the BMC continually monitors critical system items such as fans, power supplies, temperatures, and voltages. The system status is logged and reported to the service processor, if present.

- System fan speed control

  The BMC monitors system temperatures and adjusts fan speed accordingly.

We describe more about the BMC in 6.1, “BMC configuration options” on page 300.

The BMC also provides the following remote server management capabilities through the OSA SMBridge management utility program:

- Command-line interface (the IPMI shell)
- Serial over LAN (SOL)

For more information about enabling and configuring these management utilities, see the *x3850 M2 and x3950 M2 User's Guide*:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073029

1.12.3 Remote Supervisor Adapter II

The x3850 M2 and x3950 M2 have the Remote Supervisor Adapter II service processor as a standard component. This adapter is installed in a dedicated PCI 33 MHz slot and provides functionality similar to the Remote Supervisor Adapter II PCI option available for other System x servers. However, only the Ethernet and video connectors are used on the x3850 M2 and x3950 M2. The other external ports (including remote power and the ASM interconnect) are not supported on these servers.

The video adapter on this RSA II card is an ATI Radeon RN50 (ES1000) SVGA video controller. A DB-15 video connector (shown in Figure 1-3 on page 4) is provided on the Remote Supervisor Adapter II. The RSA II provides up to 1024x768 resolution, with a color depth of maximum of 32 bits at 85 Hz maximum refresh rate, with 16 MB of video memory.

Figure 1-22 on page 50 shows the Remote Supervisor Adapter II.

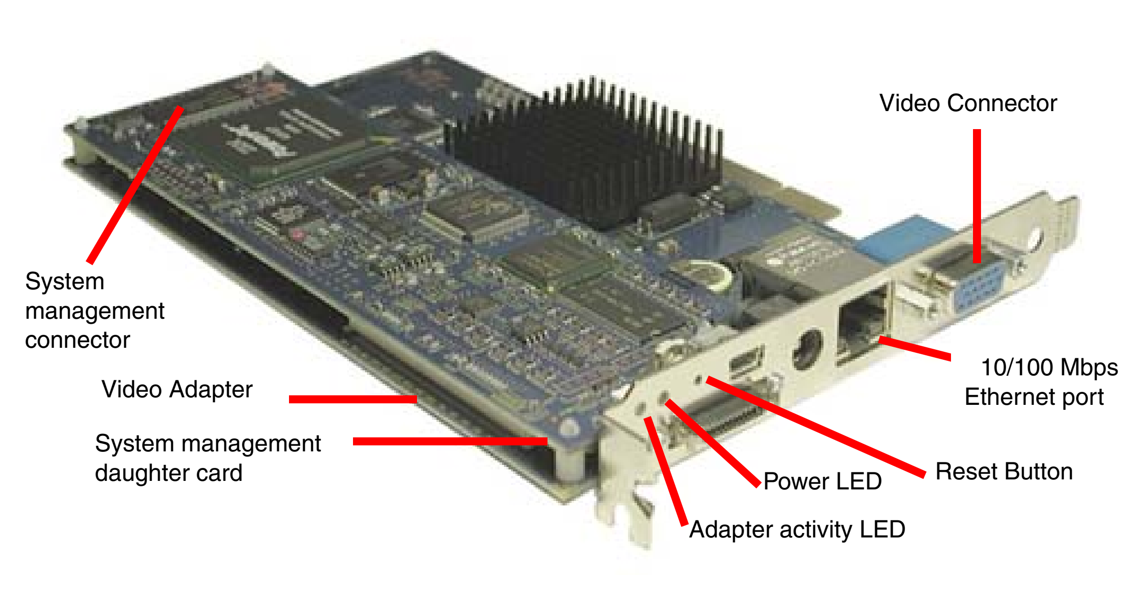

Video Connector
System management connector
Video Adapter
System management daughter card
10/100 Mbps Ethernet port
Power LED
Reset Button
Adapter activity LED

Figure 1-22 Remote Supervisor Adapter II

The most useful functions and features of the Remote Supervisor Adapter II include:

- IBM ASIC with integrated PowerPC® 405 core executing at 200 MHz
- Automatic notification and alerts

  The RSA II automatically sends different types of alerts and notifications to another server such as IBM Director and SNMP destination, or it sends e-mail directly to a user by using SMTP.
- Continuous health monitoring and control

  The RSA II continuously monitors all important system parameters such as temperature, voltage, and so on. If a fan fails, for example, the RSA II forces the remaining fans to increase speed to compensate for the failing fan.
- Event log

  You can access the server event logs and the power-on-self-test (POST) log, and export them while the server is running.
- Operating system failure window capture

  When the operating system hangs, for example, with a blue screen, you might want to capture the window for support purposes. Additionally, the RSA II stores the last failure window in memory so you can refer to it later.

- Remote media

  As a part of the remote control feature, the remote media capability lets you use diskette drives, diskette images, optical drives such as DVD or CD-ROM drives, or optical drive images of the system where the Web interface of RSA II is running on the remote PC. With this feature, the drives can be made to appear as local drives.

See 6.2, “Remote Supervisor Adapter II” on page 316 for more information about the service processor.

1.12.4 Active Energy Manager

IBM Systems Director Active Energy Manager™ (formerly known as IBM PowerExecutive™) is a combination of hardware and software that enables direct power monitoring through IBM Director. By using an OS that supports this feature, you can monitor the power consumption of the x3850 M2 and x3950 M2 and then modify or cap the consumption if required.

The application software enables you to track actual power consumption trends and corresponding thermal loading of servers running in your environment with your applications.

Active Energy Manager enables customers to monitor actual power draw and thermal loading information. This helps you with:

- More efficient planning of new datacenter construction or modification
- Proper power input sizing based on physical systems
- Justification of incremental hardware purchases based on available input power capacity
- Better utilization of existing resources

For more information see:

- Section 6.4, “Active Energy Manager” on page 334
- Extensions: Active Energy Manager at:

  http://www.ibm.com/systems/management/director/extensions/actengmrg.html

1.13 Trusted Platform Module and where used

The x3850 M2 and x3950 M2 implement the Trusted Platform Module (TPM), which ensures that the process from power-on to hand-off to the operating system boot loader is secure. The Core Root of Trusted Measurements (CRTM) code is embedded in the BIOS for logging and signing of the BIOS. In addition, you can enable the advanced control and power interface (ACPI) setting in the BIOS. The setting, which is disabled by default, assists any OS that has support written into its code to use the security features of this module.

The TPM is TCG V1.2-compliant and is ready for use with software purchased from the third-party list of the TPM Ecosystem partners who are also in compliance with the TPM V1.2 specification.

TPM can be used for the following purposes:

- Disk encryption (For example BitLocker™ Drive Encryption in Windows Server 2008)
- Digital Rights Management
- Software license protection and enforcement
- Password protection

2

Product positioning

The new IBM System x3850 M2 and x3950 M2 servers (collectively called the eX4 servers) enhance the IBM System x product family by providing new levels of performance and price-for-performance. These servers feature a high-density, 4U mechanical platform that supports quad-core and 6-core Xeon MP processors, PCIe architecture, and high-capacity high-speed DDR2 memory.

The IBM System x3850 M2 and x3950 M2 servers deliver additional processing, expandability, and high-availability features over those of their predecessors, the IBM System x3850 and x3950 servers. They are ideal for handling complex, business-critical On Demand Business applications that must be supported by space-saving, rack-optimized servers.

This chapters discusses the following topics:

2.1 Focus market segments and target applications

The eX4 servers from IBM are designed for the demands of the application and database serving tiers offering leadership performance and the proven reliability of the Intel Xeon MP processor architecture to power mission-critical stateful workloads, such as:

- Server consolidation

  Server consolidation is a process of centralizing business computing workloads to reduce cost, complexity, network traffic, management overhead and, in general, to simplify the existing IT infrastructure and provide a foundation for new solution investment and implementation.

  Key server consolidation software vendors are VMware (VMware ESX 3.5 and VMware ESXi 3.5), Xen, Virtual Iron and Microsoft (Hyper-V™)

- Database

  The eX4 servers are ideal as database servers or application servers, with their fast multi-core processors and their large and very fast memory subsystems. The x3950 M2 in particular provides an extremely scalable platform with room to scale to additional nodes. These configurations use an external storage enclosure or SAN, depending on the size of the database, which is driven by the number of users.

  The 16-way four-node configuration can deliver a highly reliable and capable platform for clients who have to run multiple instances of databases that can scale beyond eight processors.

  Key database software vendors are Microsoft SQL Server® 2005 and 2008, IBM (DB2®), and Oracle®.

- Enterprise Resource Planning (ERP)

  ERP is an industry term for the broad set of activities supported by multi-module application software that helps a manufacturer or other companies to manage the important parts of its business, including product planning, parts purchasing, maintaining inventories, interacting with suppliers, providing customer service, and tracking orders. ERP can also include application modules for the finance and human resources aspects of a business. Typically, an ERP system uses or is integrated with a relational database system.

  These applications today use a Web-based infrastructure with interfaces to suppliers, clients, and internal company employees. The three general architectures used by enterprise solutions are:

  - Four-tier architecture (often referred to as an Internet architecture) with client systems, Web servers, application servers, and database servers

  - Three-tier architecture, which includes client systems, Web, application servers, and database servers
  - Two-tier architecture, which includes client systems and database servers

  Key ERP software vendors are SAP® (SAP Business Suite and SAP Business All-in-One), Oracle (PeopleSoft® and JD Edwards®), Microsoft (Axapta), and Infor ERP Baan.

- Customer Relationship Management (CRM)

  CRM is an IT-industry term for methodologies, software, and usually Internet capabilities that help an enterprise manage client relationships in an organized way. The application can use a four-tier, three-tier, or two-tier architecture similar to ERP applications.

  Key CRM software vendors are Siebel®, Oracle (PeopleSoft and JD Edwards), SAP (SAP Business Suite and SAP Business All-in-One), Infor CRM Baan, and Onyx.

- Supply Chain Management (SCM)

  SCM is the oversight of materials, information, and finances as they move, through a process, from supplier to manufacturer to wholesaler to retailer to consumer. SCM involves coordinating and integrating these flows both within and among companies. The application also can use a four-tier, three-tier, or two-tier architecture.

  Key SCM software vendors are I2, SAP (SAP Business Suite and SAP Business All-in-One), Oracle (JD Edwards and PeopleSoft) and International Business System (IBS).

- Business Intelligence (BI)

  BI is a broad category of applications and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions. BI applications include the activities of decision-support systems, query and reporting, online analytical processing (OLAP), statistical analysis, forecasting, and data mining.

  Key BI software vendors are SAS, Oracle, Cognos®, and Business Objects.

- eCommerce

  eCommerce is the use of Internet technologies to improve and transform key business processes. This includes Web-enabling core processes to strengthen customer service operations, streamlining supply chains, and reaching existing and new clients. To achieve these goals, e-business requires a highly scalable, reliable, and secure server platform.

  Key software vendors are IBM (WebSphere) and BEA.

2.2 Positioning the IBM x3950 M2 and x3850 M2

The IBM eX4 servers are part of the broader IBM System x portfolio, which encompasses both scale-out and scale-up servers, storage, and tape products.

2.2.1 Overview of scale-up, scale-out

The System x scale-out servers start from the tower range of two-way servers with limited memory and I/O expansion, limited redundancy and system management features to the rack-optimized two-way servers with increased memory and I/O expansion, higher levels of redundancy, and increased system management features.

The IBM eX4 servers are part of the IBM System x scale-up server offering, which is designed to provide: the highest level of processor scalability with support for up to 16 multi-core processors; up to 1 TB of addressable memory with higher levels of memory availability; and flexible I/O expansion with support for up to 28 PCIe adapters.

Figure 2-1 on page 57 provides an overview of the scale-up and scale-out IBM System x product portfolio including x3850 M2 and x3950 M2.

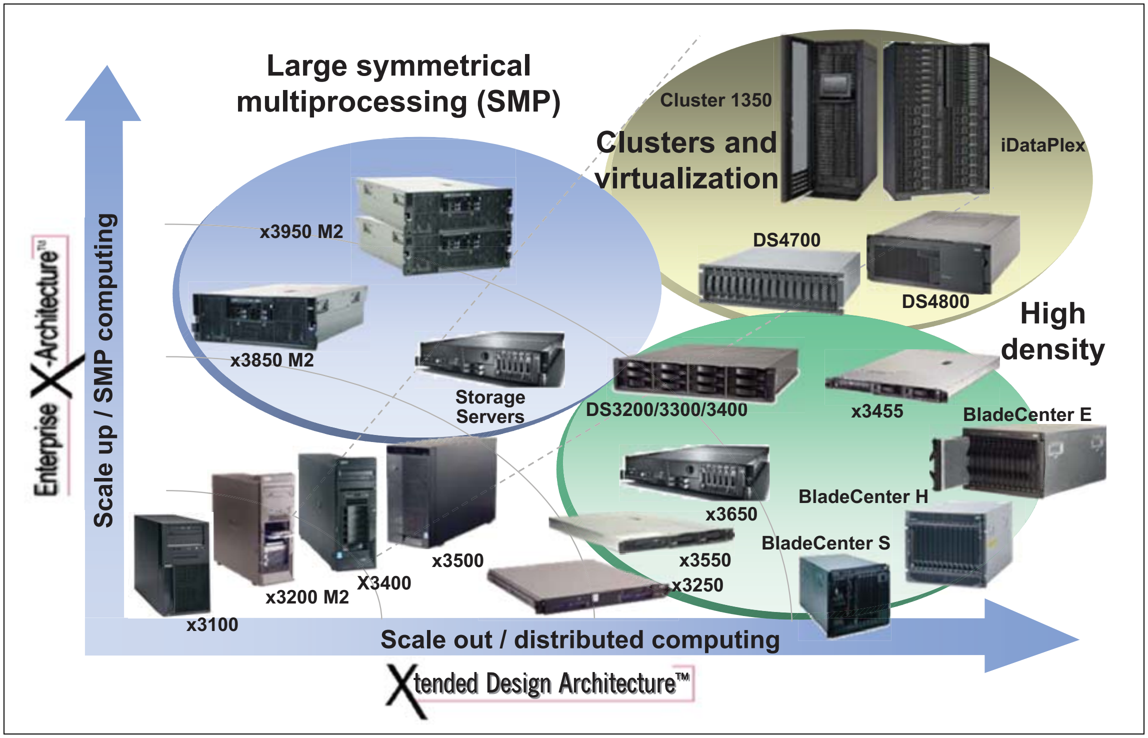


Figure 2-1 The IBM x3950 M2 and x3850 M2 are part of the high-end scale-up portfolio

2.2.2 IBM BladeCenter and iDataPlex

IBM BladeCenter and iDataPlex™ are part of the IBM System x scale-out portfolio of products; IBM eX4 is part of the IBM System x scale-up portfolio of products.

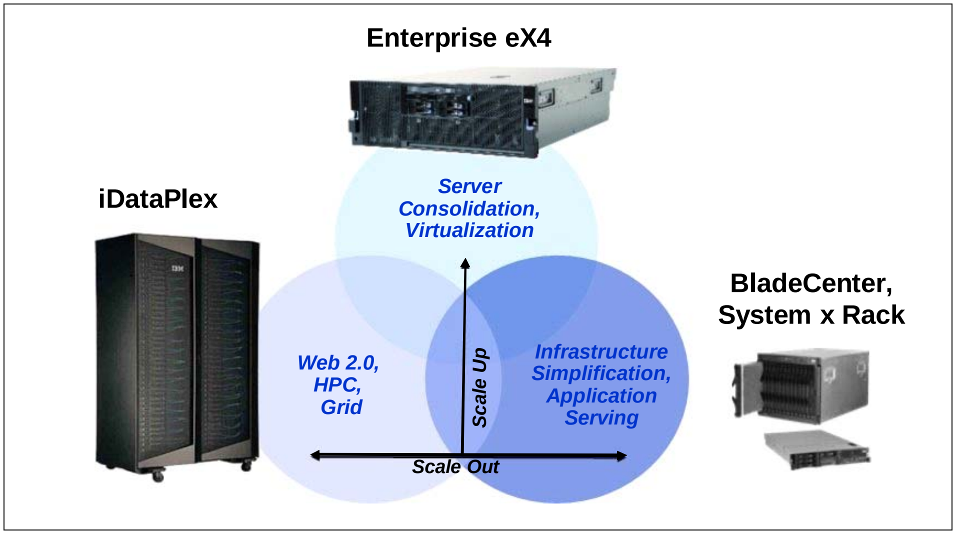

Enterprise eX4
iDataPlex
Server Consolidation, Virtualization
BladeCenter, System x Rack
Web 2.0, HPC, Grid
Scale Up
Infrastructure Simplification, Application Serving
Scale Out

Figure 2-2 IBM eX4 compared to IBM BladeCenter and System x Rack and iDataPlex

IBM iDataPlex

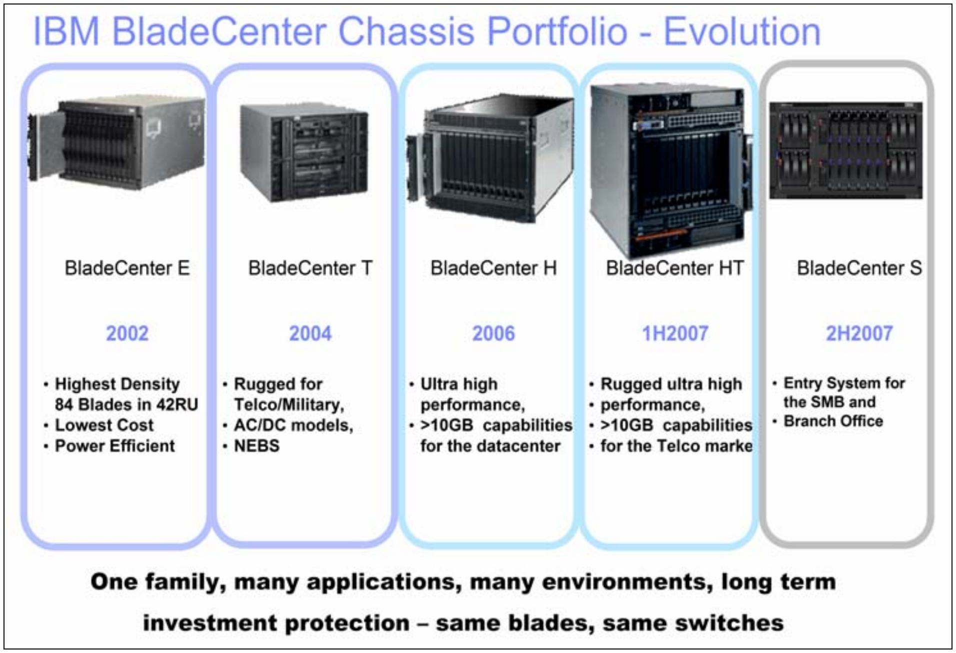
iDataPlex is massive scale-out solution that is deployed in customized rack units. It is designed for applications where workloads can be divided and spread across a very large pool of servers that are configured identically from the application workload perspective. Web 2.0, High Performance Clusters, and Grid Computing are some of the targeted applications for IBM iDataPlex in which the applications are stateless and use software for workload allocation across all nodes.

For more information about iDataPlex, refer to the paper *Building an Efficient Data Center with IBM iDataPlex*, REDP-4418 available from:

http://www.redbooks.ibm.com/abstracts/redp4418.html

IBM BladeCenter

IBM BladeCenter products are designed for complete infrastructure integration, ease of management, energy efficient servers, hardware and software Reliability, Availability, and Serviceability (RAS), and network virtualization through Open Fabric Manager. Figure 2-3 on page 59 shows the evolution of BladeCenter.


IBM BladeCenter Chassis Portfolio - Evolution

BladeCenter E

2002

- Highest Density 84 Blades in 42RU
- Lowest Cost
- Power Efficient

BladeCenter T

2004

- Rugged for Telco/Military,
- AC/DC models,
- NEBS

BladeCenter H

2006

- Ultra high performance,
- >10GB capabilities for the datacenter

BladeCenter HT

1H2007

- Rugged ultra high
- performance,
- >10GB capabilities
- for the Telco marke

BladeCenter S

2H2007

- Entry System for the SMB and
- Branch Office

One family, many applications, many environments, long term investment protection – same blades, same switches

Figure 2-3 IBM BladeCenter Chassis Portfolio

IBM BladeCenter can provide an infrastructure simplification solution; it delivers the ultimate in infrastructure integration. It demonstrates leadership in power use, high speed I/O, and server density. It provides maximum availability with industry standard components and reduces the number of single points of failure.

IBM BladeCenter offers industry's best flexibility and choice in creating customized infrastructures and solutions. BladeCenter Open Fabric Manager can virtualize the Ethernet and Fibre Channel I/O on BladeCenter. BladeCenter has a long life cycle and preserves system investment with compatible, proven, field-tested platforms and chassis.

For more information about IBM BladeCenter, refer to the IBM Redbooks publication, *IBM BladeCenter Products and Technology*, SG24-7523:

`http://www.redbooks.ibm.com/abstracts/sg247523.html`

2.3 Comparing x3850 M2 to x3850

The x3850 M2 can scale to more than four processor sockets unlike its predecessors, the System x3850 and the xSeries 366. It has twice the number of memory slots (from 16 to 32), and benefits from the increased number of processor cores (from 8 to 16 cores) and increased front-side bus bandwidth (from 667 MHz to 1066 MHz). It also derives benefits from a dedicated front side bus for each multi-core processor, as compared to the x3850 and x366, which used a shared front-side bus for each pair of processor sockets.

The Hurricane 4 chipset also adds improved bandwidth for its three scalability ports and has increased memory throughput with eight high speed memory buffer chips. Furthermore, the x3850 M2 supports more I/O slots from previously having six PCI-X (Tulsa based x3850 has four PCIe and two PCI-X slots) to seven PCIexpress slots.

The onboard LSI 1078 RAID controller and the optional ServeRAID MR10k installed in a dedicated PCIe x8 slot have significantly improved storage subsystem bandwidth compared to the x3850's Adaptec ServeRAID 8i RAID controller which shared a slower common PCI-X 66 MHz bus to the Southbridge with the onboard Broadcom Gigabit Ethernet controllers. The Hurricane 4 has a dedicated Enterprise Southbridge Interface (ESI) for the dual port onboard PCIe x4 Broadcom 5709C controllers, RSA II, Video, USB 2.0 and Serial interfaces.

The x3850 M2 also has an onboard 4x SAS port which can be used in conjunction with the ServeRAID MR10k for additional disk drive expansion (for example, using one or more EXP3000 storage enclosures) not previously possible with the x3850, without the use of one PCIe slot for a MegaRAID 8480 SAS RAID controller.

Table 2-1 compares major differences between x3850 M2 and the x3850.

Table 2-1 Comparing the x3850 M2 to x3850 servers

Feature	System x3850 server	System x3850 M2 server
Processors	Dual-core Intel Xeon 7100 series	Dual-core Intel Xeon E7210 and quad-core Intel Xeon 7300 series processors and quad-core and 6-core Intel Xeon 7400 series processors
Front-side bus	Two 667 MHz (two processors on each bus)	Four 1066 MHz (one processor on each bus)
Memory controller	Hurricane 3.0	Hurricane 4.0

Feature	System x3850 server	System x3850 M2 server
Memory	Maximum of four memory cards, each with four DDR2 DIMM slots running at 333 MHz supporting a total of 16 DDR2 DIMMs	Maximum of four memory cards, each with eight DDR2 DIMM slots running at 533 MHz supporting a total of 32 DDR2 DIMMs
Scalability	None	Upgradeable to support multinode scaling with the ScaleXpander Option Kit, 44E4249
Disk subsystem	Adaptec AIC9410 SAS	LSI 1078 SAS
External disk port	None	Yes (SAS x4) with the addition of the ServeRAID MR10k
RAID support	Standard not supported only through optional ServeRAID-8i	Standard RAID-0 and RAID-1; additional RAID features through optional ServeRAID-MR10k
PCI-X slots	Two or six depending on model	None
PCI Express slots	Some models have four PCI Express x8 full-length slots	Seven PCI Express x8 half-length slots
Active PCI slots (hot-swap)	Six	Two
Video controller	ATI Radeon 7000M 16 MB onboard	ATI ES1000 16 MB memory on RSA II
USB ports	Three (front: one, rear: two)	Six (front: two, rear: three, internal: one)
Keyboard and mouse connectors	PS/2	USB
Service processor	RSA II SlimLine adapter (optional on some models)	RSA II PCI-X adapter
Embedded virtualization	None	VMware ESXi integrated hypervisor (specific model only; for models, see Table 1-1 on page 10)
Mechanical	3U height	4U height

Feature	System x3850 server	System x3850 M2 server
Trusted Platform Module (TPM)	None	TPM with TCG V1.2 compliance
Power supplies	One or two 1300 W power supplies, depending on model	Two 1440 W power supplies

2.4 Comparing x3950 M2 to x3950

The x3950 M2 has the ability to scale to more than four processor sockets similar to its predecessors, the System x3950 and the xSeries 460. It has twice the number of memory slots (from 16 to 32), and benefits from the increased number of supported processor cores (from 8 to 24 cores), and increased front-side bus bandwidth (from 667 MHz to 1066 MHz). It also derives benefits from a dedicated front-side bus for each multi-core processors as compared to the x3950 and x460 which used a shared front-side bus for each pair of processor sockets.

For multinode configurations, the x3950 M2 scales to a four-node configuration with potentially more cores (up to 96 cores) than the maximum eight nodes possible with the x3950 (at most 64 cores).

The Hurricane 4 chipset also adds improved bandwidth for its three scalability ports and has increased memory throughput with eight high speed memory buffer chips. Furthermore, the x3950 M2 has support for more I/O slots from previously having six PCI-X slots to seven PCIe slots.

The onboard LSI 1078 RAID controller and the optional ServeRAID MR10k installed in a dedicated PCIe x8 slot have significantly improved storage subsystem bandwidth compared to the x3950’s Adaptec ServeRAID 8i RAID controller which shared a slower common PCI-X 66 MHz bus to the Southbridge with the onboard Broadcom Gigabit Ethernet controllers. The Hurricane 4 has a dedicated Enterprise Southbridge Interface (ESI) for the dual port onboard PCIe x4 Broadcom 5709C controllers, RSA II, Video, USB 2.0 and Serial interfaces.

The x3950 M2 also has an onboard 4x SAS port which can be used in conjunction with the ServeRAID MR10k for additional disk drive expansion (for example, using one or more EXP3000 storage enclosures) not previously possible with the x3950.

Table 2-2 on page 63 compares the major differences between the x3950 M2 and the x3950.

Table 2-2 Comparing the x3950 to x3950 M2 servers

Feature	x3950 server	x3950 M2 server
X-Architecture	Third-generation XA-64e chipset	Fourth generation XA-64e chipset
Processors	Dual-core Intel Xeon 7100 series	Dual-core Intel Xeon E7210 and quad-core Intel Xeon 7300 series processors and quad-core and 6-core Intel Xeon 7400 series processors
Front-side bus	Two 667 MHz (two processors on each bus)	Four 1066 MHz (one processor on each bus)
Memory controller	Hurricane 3.0	Hurricane 4.0
Maximum SMP	32 sockets using eight chassis; with dual-core processors, maximum of 64 cores	16 sockets using four chassis; with quad-core processors, maximum of 64 cores; with 6-core Dunnington processors, the maximum core count is 96
Memory	16 DDR2 DIMM sockets per node. Maximum of four memory cards, each with four DDR2 DIMM slots running at 333 MHz; 64 GB maximum per node; 512 GB maximum with eight nodes	32 DDR2 DIMM sockets per node. Maximum of four memory cards, each with eight DDR2 DIMM slots running at 533 MHz; 256 GB maximum per node; 1 TB maximum with four nodes
Internal disks	Six hot-swap bays	Four hot-swap bays
Disk subsystem	Adaptec AIC9410 SAS	LSI 1078 SAS
RAID support	No support standard. RAID support optional with the addition of a ServeRAID 8i adapter	Standard RAID-0 and RAID-1, additional RAID features through optional ServeRAID-MR10k
PCI-X slots per node	Two or six depending on model	None
PCI Express slots per node	Some models have four PCI Express x8 full-length slots	Seven PCI Express x8 half-length slots

Feature	x3950 server	x3950 M2 server
Active PCI slots (Hot Swap)	Six	Two
Ethernet controller	Broadcom 5704 dual Gigabit Ethernet	Broadcom 5709C dual Gigabit Ethernet
Video controller	ATI Radeon 7000M 16 MB onboard	ATI ES1000 16 MB memory on RSA II
Keyboard and mouse connectors	PS/2	USB
Service processor	RSA II SlimLine standard	RSA II standard
Trusted Platform Module (TPM)	None	TPM with TCG V1.2 compliance
Power supply	Two 1300W supplies	Two 1440W supplies
Mechanical	3U height	4U height

2.5 System scalability

If you plan to increase performance of your system, consider the following issues:

- Application scalability
- Operating system scalability
- Server scalability
- Storage scalability

This section discusses application and operating system scalability.

Adding processors can improve server performance under certain circumstances because software instruction execution can be shared among the additional processors. However, both the operating system and, more important, the applications must be designed to take advantage of the extra processors. Merely adding processors does not guarantee a performance benefit.

For example, not all applications can use the full power of four processors in one server. File and print servers often only take advantage of one or two processors and popular mail systems typically only scale well to four processors. Table 2-3 on page 65 shows the suitability of multi-processor systems to application types.

Table 2-3 Processor scalability by application type

Processors	File and print	Web server	E-mail collaboration	Business logic	Database	Server consolidation
1 way	Suitable	Suitable	Suitable	Suitable	—	—
2 way	Suitable	Suitable	Suitable	Suitable	Suitable	Suitable
4 way	—	—	Suitable	Suitable	Suitable	Suitable
8 way	—	—	—	—	Suitable	Suitable
16 way	—	—	—	—	Suitable	Suitable

Processors are only one part of the scalability story. Typically, an important task is to examine the following items for scalability: memory, disk, and networking subsystems. Normally, the performance gains from adding processors are only realized when memory is added in parallel. For disk-intensive applications, such as OLTP-type applications, it is essential to have a large disk array to stream data to the CPU and memory subsystems so that any disk-related delays are kept to a minimum.

Also important is to plan your system in advance according to your business requirements. Plan so that you will not have to replace your server, operating system, or storage subsystem because your server no longer meets your processing requirements, for example, if the operating system does not support more than four processors, or your server is not able to hold more than seven PCIe adapters.

Table 2-4 shows how application types scale and what is required to achieve peak performance. This table lists the server configurations used to produce several recent benchmark results. As you can see, the amount of memory and disks varies widely depending on the application.

Table 2-4 Differences in benchmark resource configurations

Benchmark	Cores	Threads	Processors	Memory	Disk Drives
TPC-C (Databases)	32	32	8	512 GB	1361
	16	16	4	256 GB	775
TPC-H (Decision Support)	32	32	8	32 GB	304
TPC-E (Databases)	32	32	8	256 GB	544
	16	16	4	128 GB	384
SPEC CPU2006	16	16	4	64 GB	1

Benchmark	Cores	Threads	Processors	Memory	Disk Drives
SAP SD (2 Tier) - ERP	16	16	4	64	28

The different server configurations reflect the different workloads of the these benchmarks. The workload that the benchmark generates causes the server to bottleneck in a particular subsystem.

As the table indicates, the SPEC CPU2006 Benchmark also highlights the component-focused nature of the SPEC benchmarks and the CPU-intensive applications they serve. This 8-way dual-core server required only 64 GB of memory and one disk. Clearly, the workload isolates the CPU with very little dependency on other subsystems. This means that the benchmark might be very good for comparing raw CPU performance, but it provides limited information regarding the performance of the entire system. The CPUs in a system can be very fast, but performance remains poor if the memory or I/O subsystems cannot supply data to them quickly enough.

2.6 Operating system scalability

This section discusses scaling of the following operating systems:

- VMware ESX
- Microsoft Windows Server 2003
- Microsoft Windows Server 2008 and Hyper-V
- Linux server operating systems

In the single chassis 4-way configuration, the IBM eX4 server acts as an industry standard symmetric multiprocessor (SMP) system. Each processor has equal access to all system resources. In most industry standard SMP systems, scaling beyond 4-way configurations has inherent processor, memory, and I/O subsystem contention issues. These issues can limit the ability of the system to scale linearly with the increased number of processors, memory, and I/O resources greater than 4-way SMP systems.

The Non-Uniform Memory Access (NUMA) architecture on the x3950 M2 multinode complex helps to address the resource contention issues faced by SMP systems by providing the ability for NUMA-aware operating systems to limit applications competing for access to processor, memory, and I/O resources to the local node’s resource pools. This minimizes the chance of scheduling applications that inefficiently access resource pools on multiple x3950 M2 NUMA nodes.

To realize the full benefit of NUMA systems, such as the x3950 M2, it is very important that operating systems have NUMA support. A NUMA-aware operating system must have the ability to schedule the use of system resource pools on each NUMA node. This must be done so that any request for processor, memory, and I/O resources for any given application process (which can spawn multiple threads) be serviced from one NUMA node ideally, or from as few NUMA nodes as possible to minimize inefficient multinode resource allocations.

The x3950 M2 multinode complex implements NUMA by connecting the scalability ports of each node together. These ports are directly connected to the Hurricane memory controller and allow high speed communication between processors located in different nodes. The ports act like hardware extensions to the CPU local buses. They direct read and write cycles to the appropriate memory or I/O resources, and maintain cache coherency between the processors.

In such multinode configurations, the physical memory in each node is combined to form a single coherent physical address space. For any given region of physical memory in the resulting system, some processors are closer to physical memory than other processors. Conversely, for any processor, some memory is considered local and other memory is remote.

The term NUMA is not completely correct because memory and I/O resources can be accessed in a non-uniform manner. PCIe and USB devices may be associated with nodes. The exceptions to this situation are existing I/O devices, such as DVD-ROM drives, which are disabled because the classic PC architecture precludes multiple copies of these existing items.

The key to this type of memory configuration is to limit the number of processors that directly access a piece of memory, thereby improving performance because of the much shorter queue of requests. The objective of the operating system is to ensure that memory requests be fulfilled by local memory when possible.

However, an application running on CPUs in node 1 might still have to access memory physically located in node 2 (a remote access). This access incurs longer latency because the travel time to access remote memory on another expansion module is clearly greater. Many people think this is the problem with NUMA. But this focus on latency misses the actual problem NUMA is attempting to solve: shorten memory request queues.

The performance implications of such a configuration are significant. It is essential that the operating system recognize which processors and ranges of memory are local and which are remote.

So, to reduce unnecessary remote access, the x3950 M2 maintains a table of data in the firmware called the Static Resource Allocation Table (SRAT). The data in this table is accessible by operating systems such as VMware ESX, Windows Server 2003 and 2008 (Windows 2000 Server does not support it) and current Linux kernels.

These modern operating systems attempt to allocate resources that are local to the processors being used by each process. So, when a process and its threads start on node 1, all execution and memory access will be local to node 1. As more processes are added to the system, the operating system balances them across the nodes. In this case, most memory accesses are evenly distributed across the multiple memory controllers, reducing remote access, greatly reducing queuing delays, and improving performance.

2.6.1 Scaling VMware ESX

This section describes the NUMA features of VMware ESX 3.0.x and 3.5 as discussed in the IBM Redbooks publication, *Virtualization on the IBM System x3950 Server*, SG24-7190, available from:

http://www.redbooks.ibm.com/abstracts/sg247190.html

VMware ESX implements NUMA scheduling and memory placement policies to manage all VMs transparently, without requiring administrators to manual oversee the complex task of balancing VMs across multiple NUMA nodes. VMware ESX does provide manual override controls for administrators with advanced skills to optimize their systems to the specific requirements of their environments.

These optimizations work seamlessly regardless of the types of guest operating systems running. VMware ESX provides transparent NUMA support even to guests that do not support NUMA hardware. This unique feature of VMware ESX allows you to take advantage of cutting-edge new hardware, even when tied to earlier operating systems.

Home nodes

VMware ESX assigns each VM a home node when the VM begins running. A VM only runs on processors within its home node. Newly-allocated memory comes from the home node also. Thus, if a VM’s home node does not change, the VM uses only local memory, avoiding the performance penalties associated with remote memory accesses to other NUMA nodes. New VMs are assigned to home nodes in a round-robin fashion. The first VM goes to the first node, the second VM to the second node, and so on. This policy ensures that memory is evenly used throughout all nodes of the system.

Several commodity operating systems, such as Windows 2003 Server, provide this level of NUMA support, which is known as initial placement. It might be sufficient for systems that only run a single workload, such as a benchmarking configuration, which does not change over the course of the system's uptime. However, initial placement is not sophisticated enough to guarantee good performance and fairness for a datacenter-class system that is expected to support changing workloads with an uptime measured in months or years.

To understand the weaknesses of an initial-placement-only system, consider the following example:

> An administrator starts four VMs. The system places two of them on the first node and two on the second node. Now, consider what happens if both VMs on the second node are stopped, or if they simply become idle. The system is then completely imbalanced, with the entire load placed on the first node. Even if the system allows one of the remaining VMs to run remotely on the second node, it will suffer a serious performance penalty because all of its memory will remain on its original node.

Dynamic load balancing and page migration

To overcome the weaknesses of initial-placement-only systems, as described in the previous section, VMware ESX combines the traditional initial placement approach with a dynamic rebalancing algorithm. Periodically (every two seconds by default), the system examines the loads of the various nodes and determines whether it should rebalance the load by moving a virtual machine from one node to another. This calculation takes into account the relative priority of each virtual machine to guarantee that performance is not compromised for the sake of fairness.

The rebalancer selects an appropriate VM and changes its home node to the least-loaded node. When possible, the rebalancer attempts to move a VM that already has some memory located on the destination node. From that point on, the VM allocates memory on its new home node, unless it is moved again. It only runs on processors within the new home node.

Rebalancing is an effective solution to maintain fairness and ensure that all nodes are fully utilized. However, the rebalancer might have to move a VM to a node on which it has allocated little or no memory. In this case, the VM can incur a performance penalty associated with a large number of remote memory accesses. VMware ESX can eliminate this penalty by transparently migrating memory from the virtual machine's original node to its new home node. The system selects a page, 4 KB of contiguous memory, on the original node and copies its data to a page in the destination node. The system uses the VM monitor layer and the processor's memory management hardware to seamlessly remap the VM's view of memory, so that it uses the page on the destination node for all further references, eliminating the penalty of remote memory access.

When a VM moves to a new node, VMware ESX immediately begins to migrate its memory in this fashion. It adaptively manages the migration rate to avoid overtaxing the system, particularly when the VM has very little remote memory remaining or when the destination node has little free memory available. The memory migration algorithm also ensures that it will not move memory needlessly if a VM is moved to a new node for only a short period of time.

When all these techniques of initial placement, dynamic rebalancing, and intelligent memory migration work in tandem, they ensure good memory performance on NUMA systems, even in the presence of changing workloads. When a major workload change occurs, for instance when new VMs are started, the system takes time to readjust, migrating VMs and memory to new, optimal locations. After a short period of time, the system completes its readjustments and reaches a steady state.

Transparent page sharing optimized for NUMA

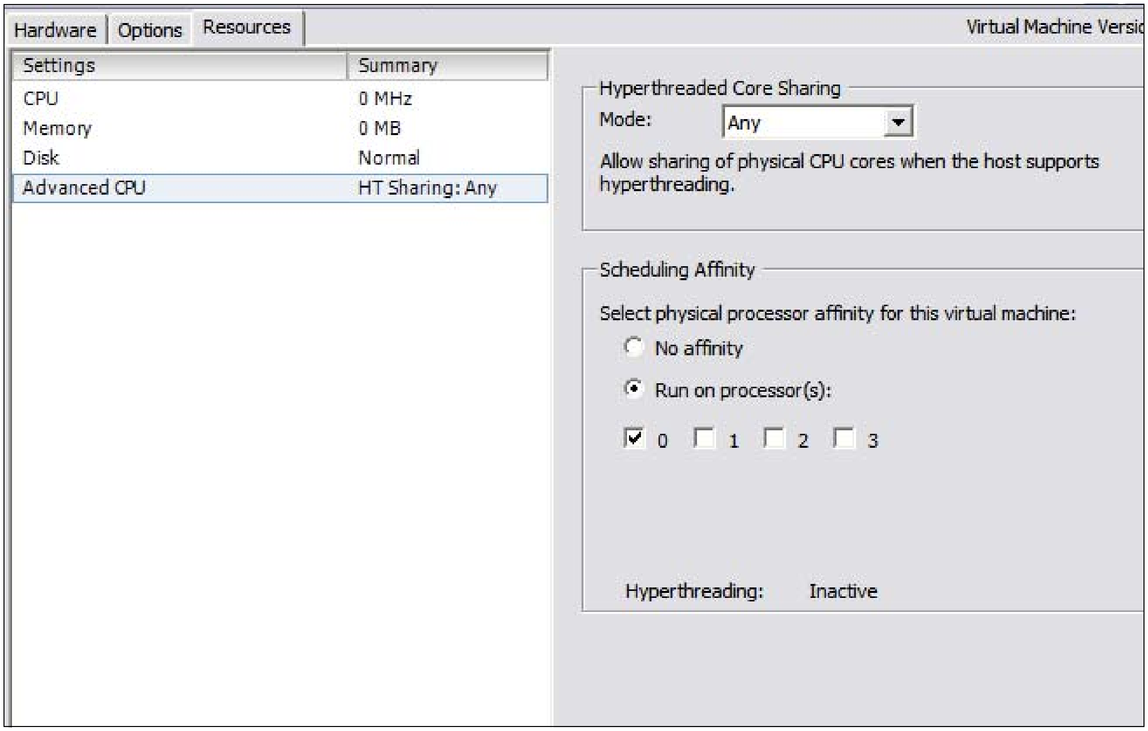
Many VMware ESX workloads present opportunities for sharing memory across virtual machines. For example, several virtual machines might be running instances of the same guest operating system, have the same applications or components loaded, or contain common data. In such cases, VMware ESX systems use a proprietary transparent page-sharing technique to securely eliminate redundant copies of memory pages. With memory sharing, a workload running in virtual machines often consumes less memory than it would when running on physical machines. As a result, higher levels of overcommitment can be supported efficiently.

Transparent page sharing for VMware ESX systems has also been optimized for use on NUMA systems like the IBM x3950 M2. With VMware ESX running on a multinode IBM x3950 M2 partition, pages are shared per node, so each NUMA node has its own local copy of heavily shared pages. When virtual machines use shared pages, they do not have to access remote memory.

Manual NUMA controls

Some administrators with advanced skills might prefer to control the memory placement and processor use manually. See Figure 2-4 on page 71. This can be helpful, for example, if a VM runs a memory-intensive workload, such as an in-memory database or a scientific computing application with a large dataset. Such an application can have performance improvements if 100% of its memory is allocated locally, whereas VMs managed by the automatic NUMA optimizations often have a small percentage (5-15%) of their memory located remotely. An administrator might also want to optimize NUMA placements manually if the system workload is known to be simple and unchanging. For example, an eight-processor system running eight VMs with similar workloads would be easy to optimize by hand.

ESX Server provides two sets of controls for NUMA placement, so that administrators can control memory and processor placement of a virtual machine.



Settings	Summary
CPU	0 MHz
Memory	0 MB
Disk	Normal
Advanced CPU	HT Sharing: Any

Hyperthreaded Core Sharing

Mode: Any

Allow sharing of physical CPU cores when the host supports hyperthreading.

Scheduling Affinity

Select physical processor affinity for this virtual machine:

○ No affinity

● Run on processor(s):

☑ 0 □ 1 □ 2 □ 3

Hyperthreading: Inactive

Figure 2-4 Setting CPU or memory affinity on VMware ESX 3.0.x / 3.5.x in VI Client or Virtualcenter

The VI Client allows you to specify:

- CPU affinity: A virtual machine should use only the processors on a given node.
- Memory affinity: The server should allocate memory only on the specified node.

If both options are set before a virtual machine starts, the virtual machine runs only on the selected node and all of its memory is allocated locally.

An administrator can also manually move a virtual machine to another node after the virtual machine has started running. In this case, the page migration rate of the virtual machine should also be set manually, so that memory from the virtual machine's previous node can be moved to its new node.

See the VMware ESX *Resource Management Guide* for a full description of how to set these options:

http://www.vmware.com/pdf/vi3_35/esx_3/r35/vi3_35_25_resource_mgmt.pdf

> **Note:** In most situations, an ESX host's automatic NUMA optimizations result in good performance. Manual NUMA placement can interfere with the ESX Server resource management algorithms, which attempt to give each VM a fair share of the system's processor resources. For example, if ten VMs with processor-intensive workloads are manually placed on one node, and only two VMs are manually placed on another node, then the system cannot possibly give all twelve VMs equal shares of the system's resources. You should consider these issues when using manual placement.

VMware ESX 3.5 scalability

At the time of the writing, VMware ESX 3.5 was the latest major release of VMware's hypervisor. ESX 3.5 provides many other enhanced features compared to ESX 3.0.x but the main features that relate to scalability on the x3850 M2 and x3950 M2 are:

- Large memory support for both ESX hosts and virtual machines

  VMware ESX 3.5 supports 256 GB of physical memory and virtual machines with 64 GB of RAM. Upon booting, ESX Server 3.5 uses all memory available in the physical server.

  > **Note:** VMware ESX 3.5 currently supports no more than 256 GB of RAM installed.

- ESX Server hosts support for up to 32 logical processors

  ESX Server 3.5 fully supports systems with up to 32 logical processors. Systems with up to 64 logical processors are supported experimentally by VMware. To enable experimental support for systems with up to 64 logical processors in ESX Server 3.5, run the following commands in the service console and then reboot the system:

  ```
  # esxcfg-advcfg -k 64 maxPCPUS
  # esxcfg-boot -b
  ```

- SATA support

  ESX Server 3.5 supports selected SATA devices connected to dual SAS/SATA controllers. For a list of supported dual SAS/SATA controllers see the ESX Server 3.x *I/O Compatibility Guide*:

  http://www.vmware.com/pdf/vi35_io_guide.pdf

SATA drives typically come in larger capacities than SAS. Because of the lower 7.2K RPM speeds for SATA versus 15K RPM for SAS, and because SATA drives are designed for lower duty cycles, SAS is still the preferred drive for production-level virtualization workloads. SATA, however, can be appropriate for a multi-tiered archiving solution for less frequently used virtual machines.

- VMware HA

  At the time of this writing, although ESX 3.5 Update 1 added support for VMware HA feature, it had restrictions-—swap space must be enabled on individual ESXi hosts, and only homogeneous (no mixing of ESX 3.5 and ESXi hosts) HA clusters are supported. See VMware release notes for more details:

  http://www.vmware.com/support/vi3/doc/vi3_esx3i_e_35u1_vc25u1_rel_notes.html

Note: At the time of the writing, VMware ESX 3.5 Update 1 was not supported by IBM for multinode x3950 M2 with up to 64 processor cores. Although not currently supported, VMware ESX 3.5 Update 1 support from IBM is planned for 2-node x3950 M2 (32-cores) in 2H/2008. VMware ESXi is not supported on multinode x3950 M2 complexes.

2.6.2 Scaling Microsoft Windows Server 2003

Both Enterprise and Datacenter Editions of Windows Server 2003 x64 scale well with support for 8 and 64 multi-core processors respectively[1]. These operating systems also support up to 2 TB of RAM and support the NUMA capabilities of the IBM x3950 M2 multinode complex.

Windows Server 2003 Enterprise and Datacenter Editions are NUMA-aware and are able to assign application threads to use processor and memory resource pools on local NUMA nodes. Scheduling application threads to run on local resource pools can improve application performance because it minimizes internode traffic on the x3950 M2 scalability ports and also reduces contention for resources with other applications and potential resources bottlenecks by assigning the different applications to run on different NUMA nodes.

For a detailed discussion of the features of Windows Server 2003 pertaining to NUMA systems, see the Microsoft document *Application Software Considerations for NUMA-Based Systems*, available from:

http://www.microsoft.com/whdc/archive/numa_isv.mspx

Table 2-5 on page 74 lists features of the various Microsoft Server 2003 editions.

[1] At the time of the writing, Windows Server 2003 was able to detect and use only up to 64 cores.

Table 2-5 Features of the Windows Server 2003 family

Features	Standard Edition	Enterprise Edition	Datacenter Edition	Web Edition
Edition availability				
32-bit release	Yes	Yes	Yes	Yes
x64 release	Yes	Yes	Yes	No
64-bit release (Itanium®)	Yes	Yes	Yes	No
Scalability				
Maximum Processors Sockets	4	8	32-bit: 32 x64: 64[a]	2
Number of x3950 M2 nodes	One	x64: Two	x64: Two[b] 32-bit: One	None
Memory: 32-bit	4 GB	64 GB	128 GB	2 GB
Memory: x64	32 GB	2 TB[c]	2 TB[c]	N/A
Hyper-threading	Yes	Yes	Yes	Yes
Hot-add memory	No	Yes	Yes	No
NUMA support	No	Yes	Yes	No
Cluster Service	No	1-8 nodes	1-8 nodes	No

[a] At the time of writing, Windows 2003 Datacenter Edition x64 is licensed for up to 64 sockets but can detect a maximum of only 64 cores; Windows 2003 Datacenter Edition (32-bit) is licensed for up to 32 sockets but can detect a maximum of only 64 cores.

[b] For Datacenter Edition x64, four-node support is planned for 2H2008.

[c] The x3950 M2 is limited to 256 GB per node, which means 512 GB in a two-node configuration and 1 TB in a four-node configuration.

The following documents have more details about other features available on each edition of Windows Server 2003:

- Comparison of Windows Server 2003 editions

  http://technet2.microsoft.com/windowsserver/en/library/81999f39-41e9-4388-8d7d-7430ec4cc4221033.mspx?mfr=true

- Virtual memory address space limits for Windows editions

  http://msdn.microsoft.com/en-us/library/aa366778(VS.85).aspx#memory_limits

- Physical memory limits

  http://msdn.microsoft.com/en-us/library/aa366778(VS.85).aspx#physical_memory_limits_windows_server_2003

- Microsoft White Paper: *Application Software Considerations for NUMA-Based Systems*

  http://www.microsoft.com/whdc/archive/numa_isv.mspx

2.6.3 Scaling Microsoft Windows Server 2008 and Hyper-V

For a detailed discussion of the features of Microsoft Windows Server 2008 and Hyper-V for use on the IBM x3850 M2 and x3950 M2, see the Microsoft white paper, *Inside Windows Server 2008 Kernel Changes*, available from:

http://technet.microsoft.com/en-us/magazine/cc194386.aspx

Table 2-6 lists features supported by the various Windows Server 2008 editions.

Table 2-6 Features of the Windows Server 2008 family

Features	Standard Edition	Enterprise Edition	Datacenter Edition	Web Edition
Edition availability				
32-bit release[a]	Yes	Yes	Yes	Yes
x64 release	Yes	Yes	Yes	Yes
64-bit release (Itanium)[b]	No	No	No	No
Scalability				
Maximum Processors Sockets	4	8	x64: 64 32-bit: 32[c]	4
Number of x3950 M2 nodes	x64: Two[d]	x64: Two	Two[e]	None
Memory — 32-bit	4 GB	64 GB	64 GB	4 GB
Memory — x64 (64-bit)	32 GB	2 TB[f]	2 TB[f]	32 GB
Hyper-Threading	Yes	Yes	Yes	Yes
Hot-add memory	No	Yes	Yes	No
Hot-replace memory	No	No	Yes	No
Hot-add processors	No	No	Yes[g]	No
NUMA support	No	Yes	Yes	No

Features	Standard Edition	Enterprise Edition	Datacenter Edition	Web Edition
Cluster Service	No	1-8 nodes	1-8 nodes	No

a. Windows Server 2008 (32-bit) Standard, Enterprise, and Datacenter Editions are not yet supported by IBM on x3850 M2, x3950 M2 (single or multinode). Windows 2008 Web Edition (32-bit) and (64-bit) are not supported on IBM x3850 M2 or x3950 M2.

b. Microsoft has released a separate Windows Server 2008 Itanium Edition for IA64 to be used solely on Itanium-based™ hardware platforms. Previously, Windows Server 2003 family had IA64 versions of Windows Server 2003 Standard, Enterprise, and Datacenter Editions that supported running these Windows Server 2003 Editions on the Itanium hardware platform.

c. At the time of writing, Windows 2008 Datacenter Edition x64 was licensed for up to 64 sockets but could only detect a maximum of 64 cores and Windows 2008 Datacenter Edition (32-bit) is licensed for up to 32 sockets but could only detect a maximum of 64 cores.

d. Two nodes are supported with two processors (sockets) in each node.

e. Four-node support is planned.

f. The x3950 is limited to 256 GB per node. This means 512 GB in a two-node configuration and 1 TB in an four-node configuration.

g. Windows Server 2008 hot-add processors features is not supported by the x3850 M2 or x3950 M2.

The following Web pages and documents have more details about other features available on each edition of Windows Server 2008:

- Compare Technical Features and Specifications

  http://www.microsoft.com/windowsserver2008/en/us/compare-specs.aspx

- Virtual memory address space limits for Windows limits

  http://msdn.microsoft.com/en-us/library/aa366778(VS.85).aspx#memory_limits

- Physical memory address space limits

  http://msdn.microsoft.com/en-us/library/aa366778(VS.85).aspx#physical_memory_limits_windows_server_2008

- *Inside Windows Server 2008 Kernel Changes*, by Mark Russinovich

  http://technet.microsoft.com/en-us/magazine/cc194386.aspx

2.6.4 Scaling Linux server operating systems

Review the following documents for information about how to scale Linux (and particular SLES and RHEL):

- *Inside the Linux scheduler*

  http://www.ibm.com/developerworks/linux/library/l-scheduler/

- *Linux Scalability in a NUMA World*

  http://oss.intel.com/pdf/linux_scalability_in_a_numa_world.pdf

- *What Every Programmer Should Know About Memory*, by Ulrich Drepper

  http://people.redhat.com/drepper/cpumemory.pdf

- *A NUMA API for Linux*

  http://www.novell.com/collateral/4621437/4621437.pdf

- *Anatomy of the Linux slab allocator*

  http://www.ibm.com/developerworks/linux/library/l-linux-slab-allocator/

The documents describe features of the Linux 2.6 kernel and components such as the Linux task scheduler and memory allocator, which affect the scaling of the Linux operating system on the IBM x3850 M2 and x3950 M2.

Factors affecting Linux performance on a multinode x3950 M2

The overall performance of a NUMA system depends on:

- The local and remote CPU cores on which tasks are scheduled to execute

  Ensure threads from the same process or task are scheduled to execute on CPU cores in the same node. This can be beneficial for achieving the best NUMA performance, because of the opportunity for reuse of CPU core's cache data, and also for reducing the likelihood of a remote CPU core having to access data in the local node's memory.

- The ratio of local node to remote node memory accesses made by all CPU cores

  Remote memory accesses should be kept to a minimum because it increases latency and reduces the performance of that task. It can also reduce the performance of other tasks because of the contention on the scalability links for remote memory resources.

The Linux operating system determines where processor cores and memory are located in the multinode complex from the ACPI System Resource Affinity Table (SRAT) and System Locality Information Table (SLIT) provided by firmware. The SRAT table associates each core and each contiguous memory block with the node they are installed in. The connections between the nodes and the number of hops between them is described by the SLIT table.

In general, memory is allocated from the memory pool closest to the core on which the process is running. Some system-wide data structures are allocated evenly from all nodes in the complex to spread the load across the entire complex and to ensure that node 0 does not run out of resources, because most boot-time code is run from that node.

Features of Red Hat Enterprise Linux 5

Table 2-7 describes several features supported by the various Red Hat Enterprise Linux 5 Editions.

Table 2-7 Features of Red Hat Enterprise Linux 5

Features	Base server product subscription	Advanced platform product subscription
Server support limits as defined by Red Hat Enterprise Linux Product Subscription		
Maximum physical CPUs (sockets)	2	Unlimited
Maximum memory	Unlimited	Unlimited
Maximum virtualized guests, instances	4	Unlimited
Technology capabilities and limits		
Maximum Logical Processors (x86)[a]	8 (assuming maximum of quad-core processors limited by subscription to 2 sockets)	32
Maximum Logical Processors (EM64T and AMD64)[a]	8 (limited by subscription to 2 sockets)	64 (certified)[b] 255 (theoretical)
Number of x3950 M2 nodes	One	Two
Memory: x86 (32-bit)	16 GB[c]	16 GB[c]
Memory: x64 (64-bit)	256 (certified)[b] 1 TB (theoretical)	256 (certified)[b] 1 TB (theoretical)
NUMA support	Yes	Yes

a. Red Hat defines a logical CPU as any schedulable entity. So every core/thread in a multi-core/thread processor is a logical CPU.

b. Certified limits reflect the current state of system testing by Red Hat and its partners, and set the upper limit of support provided by any Red Hat Enterprise Linux subscription if not otherwise limited explicitly by the subscription terms. Certified limits are subject to change as on-going testing completes.

c. The x86 *Hugemem* kernel is not provided in Red Hat Enterprise Linux 5.

For more information, review these Web pages:

- Red Hat Enterprise Linux Server Version comparison chart

  `http://www.redhat.com/rhel/compare/`

- IBM ServerProven NOS Support for RedHat Enterprise Linux chart

  `http://www.ibm.com/servers/eserver/serverproven/compat/us/nos/redchat.html`

Features of SUSE Linux Enterprise Server

Table 2-8 describes the features supported by the various Novell® SUSE® Linux Enterprise Server (SLES):

Table 2-8 Features of the SUSE Enterprise Linux 10

Features	SLES 10 (2.6.16.60) x86	SLES 10 (2.6.16.60) x86_64
Maximum Logical Processors	32 (up to 128 with bigsmp kernel on certified systems)	32 (up to 128 on certified systems)
Number of x3950 M2 nodes	One	Two
Maximum Memory	16 GB (certified) 64 GB (theoretical)[1]	512 GB (certified) 64 TB (theoretical)
NUMA support	Yes	Yes

For more information, review the SUSE Linux Enterprise Server 10 Tech Specs & System Requirements at:

http://www.novell.com/products/server/techspecs.html

2.7 Application scalability

Enterprise applications enable you to run your business more effectively and are often referred to as back-office applications. As discussed briefly in 2.1, “Focus market segments and target applications” on page 54, they bring together four major application groups to create integrated end-to-end solutions:

- Business Intelligence (BI)
- Customer Relationship Management (CRM)
- Enterprise Resource Planning (ERP)
- Supply Chain Management (SCM)

Enterprise applications work with your most critical business data so it is important that these applications are highly available and secure. As as shown in Figure 2-5 on page 80, the following three general architectures are used by these applications:

- A three-tier architecture (often referred to as an Internet architecture) includes client systems, Web servers, application servers, and database servers.
- A two-tier architecture includes client systems, Web servers, and database/application servers.
- A three-in-one tier architecture includes client systems and database servers.

While three-tier architecture has far greater complexity, it also allows for greater scalability. The architecture selected for a solution depend on your business requirements, the type of application deployed, and the number of planned users. In most cases, if you have to scale your applications, use a two-tier or three-tier architecture. Smaller clients might prefer to implement a *three-in-one* implementation, simply because it is easier to manage and the number of users supported can be handled by the three-in-one solution.

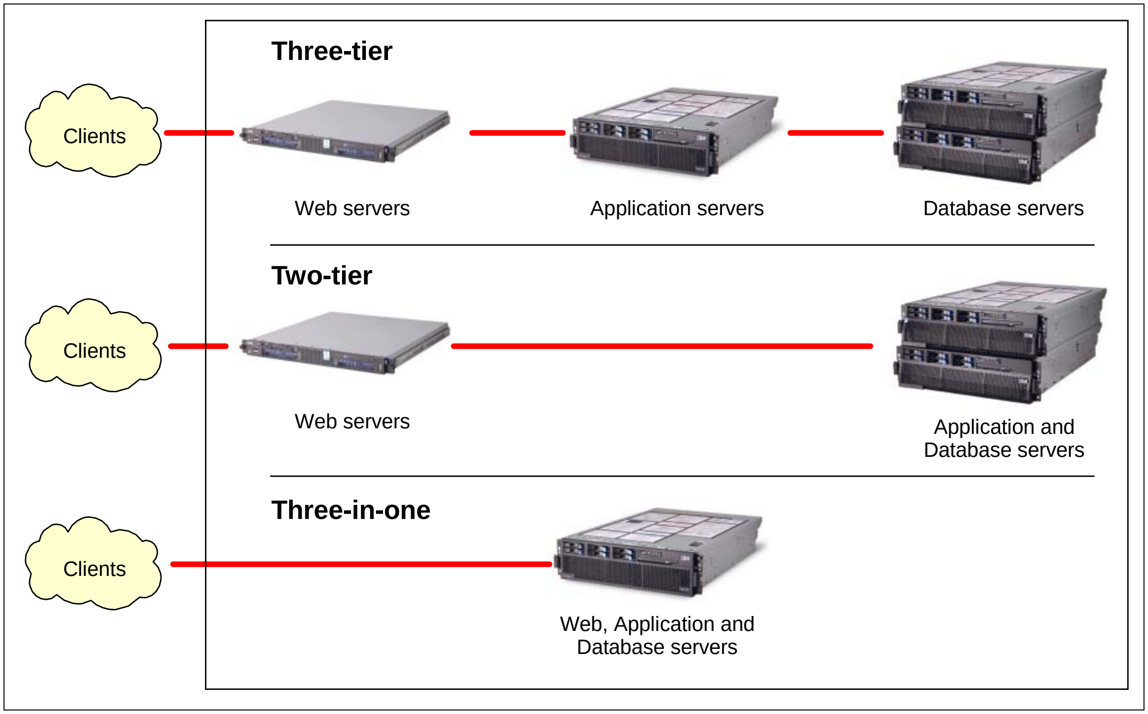


Figure 2-5 Enterprise solution architectures

2.7.1 Microsoft SQL Server 2005

Microsoft SQL Server 2005 has many features that allow it to scale from a small single-user database to a huge enterprise-wide, mission-critical, multi-user database. This section highlights these features and discusses how they are relevant to server consolidation.

Support for 64-bit computing (x64)

The combination of Windows Server 2003 for x64 and SQL Server 2005 for x64 offers directly addressable physical memory up to the memory limit of the operating system: 32 GB for Windows Server 2003 Standard Edition, and

1024 GB (1 TB) for Windows Server 2003 Enterprise Edition. This effectively resolves the memory constraint that exists with the 32-bit versions of Windows Server and SQL Server.

Several editions of SQL Server 2005 have varying support for x64; however, only these versions are suitable for creating a consolidated SQL Server environment:

- SQL Server 2005 Enterprise Edition (32-bit and 64-bit)
- SQL Server 2005 Standard Edition (32-bit and 64-bit)

For medium and large-scale SQL Server consolidation projects, the Standard Edition and Enterprise Edition versions both have native x64 versions; however, many of the advanced scalability features are only found in the Enterprise Edition. Developer Edition has all the features of Enterprise Edition, but is licensed only for development and testing, not for production use.

It is important to note that a database created using SQL Server 2005 Express Edition can be moved to an installation of SQL Server 2005 Enterprise Edition without any modifications. This provides a clear growth strategy for all new databases created with SQL Server 2005 and demonstrates the ease with which databases can be scaled-up on this platform.

Hot-add memory

Additional physical memory can be installed in a running server, and SQL Server 2005 will recognize and use the additional memory immediately. This could prove useful if you must increase available memory to service new business requirements without affecting database availability. This feature also requires hot-add memory support as provided in servers such as the IBM System x3850 M2 and x3950 M2.

Feature comparisons

Most of the features that are mentioned in the following sections are found only in the Enterprise Edition. For a detailed analysis of what is supported by Standard Edition and Enterprise Edition, see the following documents:

- Comparison Between SQL Server 2005 Standard and Enterprise Editions

  http://www.microsoft.com/sql/editions/enterprise/comparison.mspx

- SQL Server 2005 Features Comparison

  http://www.microsoft.com/sql/prodinfo/features/compare-features.mspx

Server Resource Management

Given the availability of server hardware that has large memory capacity, up to 32 processors and multiple network cards, having control over how those considerable resources are allocated is necessary. This section introduces hardware and software features that can provide that control and ensure the most appropriate use of the available resources.

Non-uniform memory addressing (NUMA)

NUMA is a scalability technology for splitting servers with numerous processors (CPUs) and large amounts of memory into resource groups, or NUMA nodes. The processors in a NUMA node work primarily with the local memory in that NUMA node while still having access to memory in other NUMA nodes (remote memory). Using local memory is quicker than remote memory because of the configuration of the NUMA node.

Because SQL Server 2005 is NUMA-aware, it tries to write data to physical memory that is associated with the requesting CPU to benefit from the better local memory access performance. If the requesting CPU does not have enough memory available, it is allocated from another NUMA node.

Soft-NUMA, CPU affinity, and I/O affinity

Soft-NUMA is a SQL Server 2005 feature that you can use to group CPUs and network interfaces into soft-NUMA nodes. However, you cannot allocate memory to a soft-NUMA node and all memory requests are served from all memory available to SQL Server.

To group CPUs, you must edit the registry directly using a node configuration affinity mask. After the soft-NUMA nodes have been created, you can assign individual SQL Server instances to one or more soft-NUMA nodes.

You might create soft-NUMA nodes if your server hardware does not have hardware NUMA capabilities or to sub-divide a NUMA node further. Each soft-NUMA node gets its own I/O thread and lazy writer thread. If the SQL instance has a high I/O requirement, it could be assigned two soft-NUMA nodes. The SQL instance then has two I/O threads that can help it process I/O requests better. Soft-NUMA provides the ability to fine-tune the use of the server resources to ensure that critical databases get the resources that they require within a consolidated environment.

CPU affinity and I/O affinity are SQL Server 2005 features for configuring each database instance to use specific CPUs for database processing and I/O requests. Assigning a set of CPUs only to handle I/O processing might provide performance benefits with a database that relies heavily on I/O operations.

Designating a certain number of CPUs to a critical database ensures that performance is not affected by other processes running on the same server because those processes are run on other CPUs in the server. CPU and I/O affinity are used for fine-tuning the allocation of server resources to where they are most required.

Windows Server Resource Manager (WSRM)

WSRM comes with Windows Server 2003 Enterprise and Datacenter Editions and can be applied to any application running on the server. Using WSRM policies, it is possible to manage CPU and memory use by process or user. WSRM helps ensure that a process or user does not use more than its allotted quantity of CPU and memory resources meaning that multiple applications can run safely together.

SQL Server 2005 is also able to allocate individual CPUs to a SQL database instance using soft-NUMA and CPU affinity, so be sure no contention issues arise during the configuration of WSRM.

2.7.2 Microsoft SQL Server 2008

This section discusses new features slated for inclusion in SQL Server 2008.

Organizations across the board clearly are experiencing exponential growth in the volume and variety of data that they must process, analyze, protect, and store. The growing importance of regulatory compliance and increasing globalization dictates that data must be stored securely and be available at all times. Because the costs of disk storage have dropped to record lows, organizations can store more data per dollar. Users must be able to examine and analyze this data quickly and easily on any device using their regular office productivity programs. The management of this information explosion and hike in user expectations creates severe challenges for the enterprise.

Microsoft has positioned its new database platform as the answer to these challenges and we have highlighted some of the new key features here. For a comprehensive review of all the new features in SQL Server 2008, visit:

http://www.microsoft.com/sql/2008/default.mspx

Enhanced SQL Server resource management

With the Resource Governor in SQL Server 2008, administrators can control how resources are consumed in SQL Server. Windows Server Resource Manager (WSRM) provided some ability to control processes within Windows by permitting restrictions on what resources the process sqlservr.exe could consume. However, this affected every activity in the SQL Server instance. With Resource Governor, administrators can configure how various workloads use the available SQL Server-specific resources (only CPU and memory in CTP V5). This is an important feature for SQL Server consolidation because administrators can ensure the best use of available SQL Server resources based on business requirements.

Hot-add CPU

Building on the existing support for hot-add memory, SQL Server 2008 introduces the installation of additional CPUs in supported server hardware so that you can use the new CPU resources immediately without downtime. This feature extends the ability of SQL Server to scale up the available hardware resources without disrupting the environment. Currently, support is not available for SQL 2008 hot-add CPU feature with IBM eX4 servers.

2.8 Scale-up or scale-out

The goal of system scalability is to increase performance at a rate that is proportional to increases in system resources for a given workload. The two methods to achieving system scalability are:

- Scale-up: Increasing the capacity of the single system image by adding (in particular) processors, memory, and disk.
- Scale-out: Adding systems that can be managed and run together.

2.8.1 Scale-up

Scaling-up is achieved by adding resources, such as memory, processors, and storage, to an existing system that runs on a single server. It is also referred to as vertical scaling. The benefit to scaling up is that it is relatively easy, because in general it requires only hardware or software that is designed to take advantage of additional memory, processor, and storage resources.

With the mass adoption of virtualization as a means for server consolidation and driving up resource utilization on under utilized mulit-core processor based systems, virtualization infrastructures are increasingly required to support larger numbers of virtual and more diverse workloads with higher levels of software and hardware redundancy. This has translated to virtualization trends seeking to deploy virtualization platforms that deliver performance and availability, and also the agility and flexibility to grow and shrink in line with business demands. Scale-up systems, such as the IBM eX4 servers, are increasingly being exploited by NUMA-aware virtualization hypervisors, such as VMware ESX, and the demands of virtualized workloads.

For example, your database server might start out on a 2-way SMP system with 4 GB of memory and six hard drives. As the database grows in size or the number of users increases, you can easily scale-up by adding more processors, memory, and disk resources to maintain the same level of performance. You might eventually have to replace the server with one that is capable of supporting more resources. However, today on x3950 M2 servers, you can scale-up to systems that support 16 processors (96 cores) and 64 GB of memory on 32-bit versions of operating systems and 1 TB of memory on operating systems with 64-bit extension EM64T.

NUMA architecture, when compared to other architectures, provides near linear scalability and minimum overhead in resource management that limits the scalability of a single large systems when you add processors, memory, and storage.

The x3950 M2 server is a good example of an SMP system based on eX4 and NUMA technologies. The server starts with a base 2-way configuration and, as your requirements grow, you can add incremental capacity to a maximum of 16 processor sockets (64 cores). Likewise, memory can be expanded from 4 GB to 1 TB. This modular server architecture delivers investment protection without the up front costs of expensive switch-based alternatives.

Advantages of scale-up include:

- Easier to configure and administer
- Good when most queries access small blocks of data
- Best for applications that maintain state (OLTP)
- Add CPU and memory as required (scheduled downtime especially for CPUs)
- All tools and queries work as expected
- Can be maintained by lesser skilled DBAs

Disadvantages of scale-up include:

- Requires higher cost hardware
- The database has finite capacity limits tied to hardware
- Must balance CPU, memory, and I/O to achieve peak performance
- Fail-over cluster server usually configured equal to primary

2.8.2 Scale-out

Scale-out means adding discrete servers to your server *farm* to gain more processing power. Although many options exist for implementing a farm comprised of small low-end servers, we consider the use of the IBM BladeCenter, 1U rack servers or iDataPlex for large scale-out implementations such as the System x3550 as the most viable alternative when discussing this requirement.

Scale-out is sometimes called horizontal scaling, and in general referred to as clustering. However, clustering can sometimes be ambiguous because there are distinct types of clusters, which include high availability, load balancing, and high-performance computing. Load balancing is the goal of scaling out. That is to say, we scale-out by adding one or more servers to an existing system to balance the system load as we add additional demands on the system.

For example, your database server might start out on a 2-way system with 4 GB of memory and six hard drives. As the database grows in size or the number of users increase, you scale-out by adding another server with two processors, 4 GB of memory, and six disk drives to maintain the same level of performance. Although you do not necessarily have to add another server with the exact specifications, adding one does reduce the complexity of scaling out.

The benefit to scaling-out is that you can achieve near linear scalability. That is, as you add each additional server to the system, you effectively increase your system capacity proportionally. Thus, scaling-out provides much better returns in terms of the additional costs associated with adding more servers to the system. Another benefit inherent with scaling-out is that a cluster of smaller servers generally costs less than a single large system.

The drawback to scaling-out is that it requires system and database administrators who understand the technology well enough so that it can be implemented effectively. Another drawback is that clustering requires software specifically designed for the task.

Advantages of scale-out include:

- It uses lower cost hardware.
- Scaling is near linear.
- The database size is not gated by hardware.
- It is preferred when queries access large blocks of data.
- It is best for serving stateless applications (Web).

Disadvantages of scale-out include:

- It requires more skilled DBA to maintain clusters.
- Management and scheduling are more complex.
- It depends on intelligent data partitioning.
- It introduces query overhead.
- Maintenance activities require downtime.
- Cluster applications can be much more expensive than stand-alone versions.

Architectures for scaling out

The two distinct approaches to scaling out database management systems are are generally referred to as a shared architecture and a shared-nothing architecture. Both architectures attempt to achieve the same goal, which is to implement a database management system that consists of a cluster of servers, provides linear scalability, and appears as single database to the end users.

A *shared architecture* attempts to accomplish this goal while sharing the database. As more servers are added to the system, they all share or attempt to share the same database, which resides on shared storage, hence the name shared architecture. Oracle is an example of a database application that implements a shared-disk approach.

A *shared-nothing architecture* accomplishes the same goal by dividing a large database into smaller and more manageable parts, called partitions. The term shared-nothing simply refers to the fact that as more servers are added to the system, each server manages a clearly defined portion of the database. The fact that the database is partitioned should not imply that the system cannot be implemented on shared storage. IBM DB2 and Microsoft SQL Server both implement a shared-nothing approach.

Choosing scale-up or scale-out

Microsoft SQL Server 2005 and SQL Server 2008 are well-suited for scale-up configurations, such as a multinode x3950 M2 configuration. It follows a single server, shared-nothing approach and it is a high performance solution for Windows environments.

Oracle uses a shared-disk approach and is suited to scale-up or scale-out. It is a leading solution for middle market UNIX®, Windows, and Linux environments. Scale-out capabilities can be extended with Oracle 9i or 10g RAC.

DB2 is suited to scale-up or scale-out. It is developed following a multi-server, shared nothing approach, and is the highest performing database environment for mainframe, UNIX, and Linux environments.

Scale-up is preferred for smaller databases (150-200 GB). For larger databases, large block I/O, data warehousing and decision support applications, use a scale-out deployment.

3

Hardware configuration

In this chapter, we highlight the different subsystems that should understand and configure before the hardware configuration at your x3850 M2 or x3950 M2 is completed.

This chapters discusses the following topics:

- 3.1, “Processor subsystem” on page 90
- 3.2, “Memory subsystem” on page 111
- 3.3, “Internal drive options and RAID controllers” on page 124
- 3.4, “Configuring RAID volumes” on page 154
- 3.5, “PCI Express options” on page 188

3.1 Processor subsystem

The eX4 architecture is designed to support the following Intel Xeon processors:

- Xeon 7200 series (Tigerton) dual-core processors
- Xeon 7300 series (Tigerton) quad-core processors
- Xeon 7400 series (Dunnington) quad-core processors
- Xeon 7400 series (Dunnington) 6-core processors

Each server can be upgraded to a maximum of four processors. One, two, three, or four processors are supported. Installed processors must be identical in model, speed, and cache size.

Figure 3-1 on page 90 shows the locations of the four processors (CPUs), locations of the required voltage regulator modules (VRMs), which you can identify by the blue handles, and the memory cards that we describe later in 3.2, “Memory subsystem” on page 111.

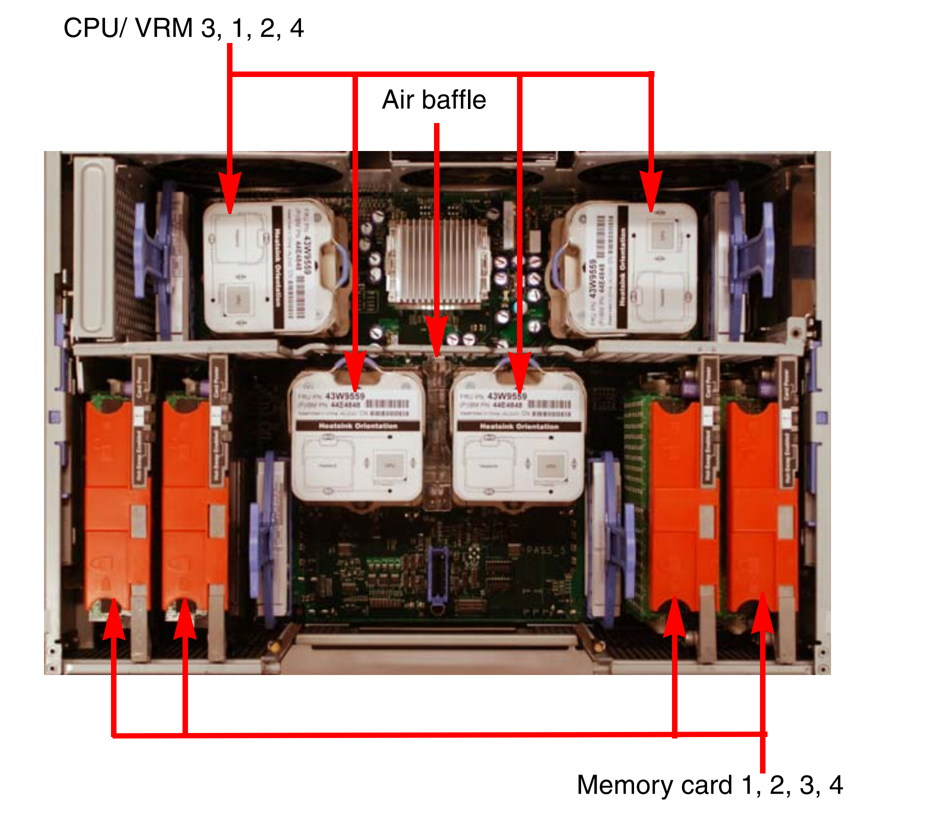


Figure 3-1 Top view of the x3850 M2 and x3950 M2 processor board; order of installation

If you require more than four processors, you can create a single scalable system by connecting up to three additional x3950 M2 systems. You may also upgrade any x3850 M2 system to a scalable x3950 M2 system as described in 1.2.4, "Scalable upgrade option for x3850 M2" on page 11.

Every processor in such a multinode configuration must be identical — same processor type, speed, and cache size. The number of processors in each node in the configuration may vary, so you can have a three-node configuration where the nodes have two, three and two processors respectively, for a total of seven processors.

Note: If you have a multinode complex and one or more nodes has only one processor installed, and that node fails, then the complex will automatically reboot without the memory resources and I/O resources that were in that node. Therefore from a system availability perspective, we recommend you have at least two processors in every node.

3.1.1 Processor options

Table 3-1 on page 91 lists the processors supported in the x3950 M2 and x3850 M2 (machine type 7141/7144) with Xeon 7200 and 7300 *Tigerton* processors. Each processor option includes a CPU, heat-sink, and VRM.

Table 3-1 Tigerton processor options for x3950 M2 and x3850 M2: machine type 7141/7144

Processor	Cores	Speed GHz	L2 / L3 cache	TDP[a]	Part number	Feature code[b]
Xeon E7210	2-core	2.40	2x4 /0 MB	80 W	44E4244	3053 / 3476
Xeon E7310	4-core	2.16	2x2 /0 MB	80 W	44W2784	3613 / 3397
Xeon E7320	4-core	2.13	2x2 /0 MB	80 W	44E4241	3052 / 3473
Xeon E7330	4-core	2.40	2x3 /0 MB	80 W	44E4242	3051 / 3474
Xeon X7350	4-core	2.93	2x4 /0 MB	130 W	44E4243	3050 / 3475
Xeon L7345	4-code	1.86	2x4 /0 MB	50 W	None	6969 / 4432

a. Thermal Design Power (TDP): The thermal specification shown is the maximum case temperature at the maximum TDP value for that processor. It is measured at the geometric center on the topside of the processor integrated heat spreader. For processors without integrated heat spreaders, such as mobile processors, the thermal specification is referred to as the junction temperature (Tj). The maximum junction temperature is defined by an activation of the processor Intel Thermal Monitor, which has an automatic mode to indicate that the maximum Tj has been reached.

b. The first feature code is the base configuration. Additional processor orders should use the second feature code.

Table 3-2 contains the processor options supported in the x3850 M2 and x3950 M2 (machine type 7233) with Xeon 7400 *Dunnington* processors. Each processor option includes a CPU, heat-sink, and VRM.

Table 3-2 Dunnington processor options for x3950 M2 and x3850 M2: machine type 7233

Processor	Cores	Speed	L2 / L3 cache	TDP[a]	Part number	Feature code[b]
Xeon L7445	4-core	2.13 GHz	2x3 /12 MB	50 W	44E4517	6976 / 4435
Xeon E7420	4-core	2.13 GHz	2x3 /12 MB	90 W	44E4469	3647 / 4419
Xeon E7430	4-core	2.13 GHz	2x3 /12 MB	90 W	44E4470	3648 / 4420
Xeon E7440	4-core	2.4 GHz	2x3 /12 MB	90 W	44E4471	3646 / 4418
Xeon L7455	6-core	2.13 GHz	3x3 /16 MB	65 W	44E4468	3645 / 4417
Xeon E7450	6-core	2.4 GHz	3x3 /16 MB	90 W	44E4472	3649 / 4421
Xeon X7460	6-core	2.67 GHz	3x3 /16 MB	130 W	44E4473	3644 / 4416

a. For a description of TDP, see *footnote a* in Table 3-1.

b. The first feature code is the base configuration. Additional processor orders should use the second feature code.

Note: The x3950 M2 and x3850 M2 models with Xeon 7200 and 7300 *Tigerton* processors (machine types 7141/7144) do not support the installation of Xeon 7400 *Dunnington* processors.

Processor cache: Tigerton

The Intel Xeon 7200 and 7300 series processors (Tigerton) have two levels of cache on the processor die:

L1 cache The L1 execution, 32 KB instruction and 32 KB data, for data trace cache in each core is used to store micro-operations, which are decoded executable machine instructions. It serves those to the processor at rated speed. This additional level of cache saves decoding time on cache hits.

L2 cache Each pair of cores in the processor has either 2 MB, 3 MB, or 4 MB of shared L2 cache, for a total of 4 MB, 6 MB, or 8 MB of L2 cache. The L2 cache implements the Advanced Transfer Cache technology.

L3 cache Tigerton processors do not have L3 cache.

Processor cache: Dunnington

The Intel Xeon 7400 series processors (Dunnington) have three levels of cache on the processor die:

L1 cache The L1 execution, 32 KB instruction and 32 KB data, for data trace cache in each core is used to store micro-operations, which are decoded executable machine instructions. It serves those to the processor at rated speed. This additional level of cache saves decoding time on cache hits.

L2 cache Each pair of cores in the processor has 3 MB of shared L2 cache, for a total of 6 MB, or 9 MB of L2 cache. The L2 cache implements the Advanced Transfer Cache technology.

L3 cache Dunnington processors have 12 MB (4-core), or 16 MB (6-core) shared L3 cache.

3.1.2 Installation of processor options

The processors are accessible from the top of the server after opening the media hood, as shown in Figure 3-2 on page 94. The media hood is hinged at the middle of the system and contains the SAS drives, optical media and the light path diagnostics panel.

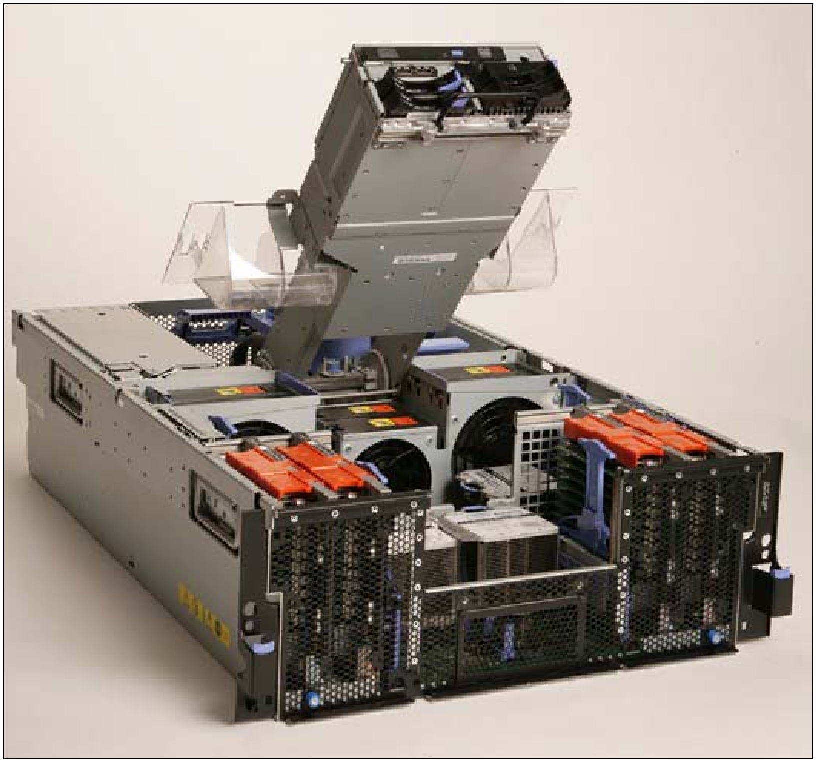


Figure 3-2 Media hood opened to access the processors

All processors must be installed with the VRMs and heat-sink, included in the option packaging. If the VRM is missing, the following error message is displayed and the system does not power up:

```
Processor configuration missmatch error
```

To install the processor options:

1. Turn off the server and peripheral devices.
2. Disconnect the power cords.
3. Wait approximately 20 seconds, then *make sure the blue LED light is off*.

Important: Any installation and removal of processors or VRM can result in damage if the system is not removed from the AC power source. Check that the system is without power after removal of the power cables.

The blue locator LED is located on the rear side of the server. This LED indicates whether AC power is connected to the system but the system is powered off. The LED remains lit for up to 20 seconds after you remove the power cords. Use this as a guide as to when you can start working in a system. After the blue LED is off you can be sure that all components are without power.

4. Remove the server bezel and the cover.
5. Loosen the captive screws and rotate the media hood to the fully open position.

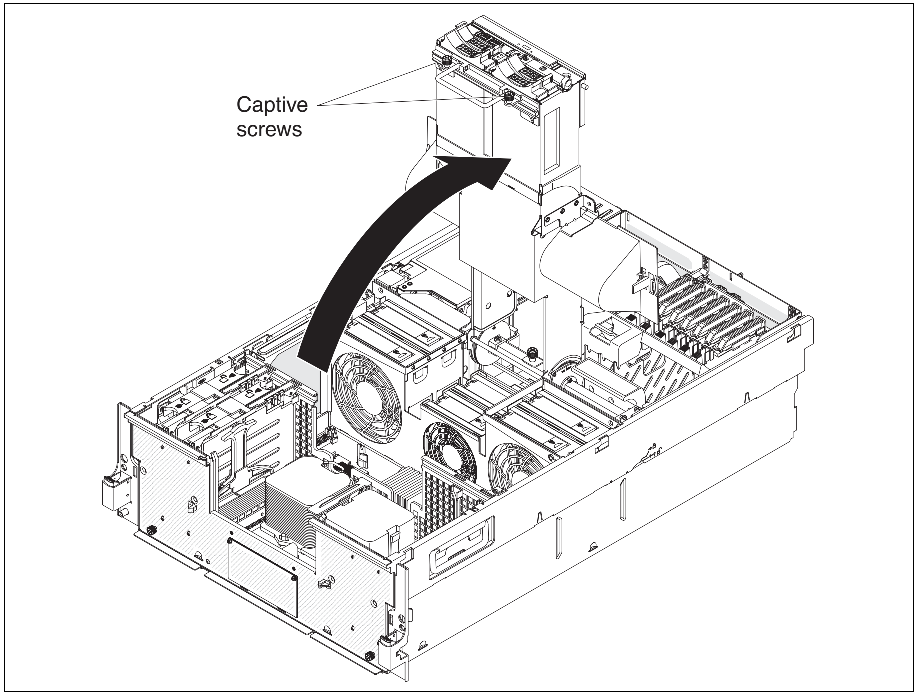


Figure 3-3 x3850 M2 and x3950 M2 with fully opened media hood

6. Review the installation instructions included in the processor option kit.
7. Install the processors in the order shown in Figure 3-1 on page 90.

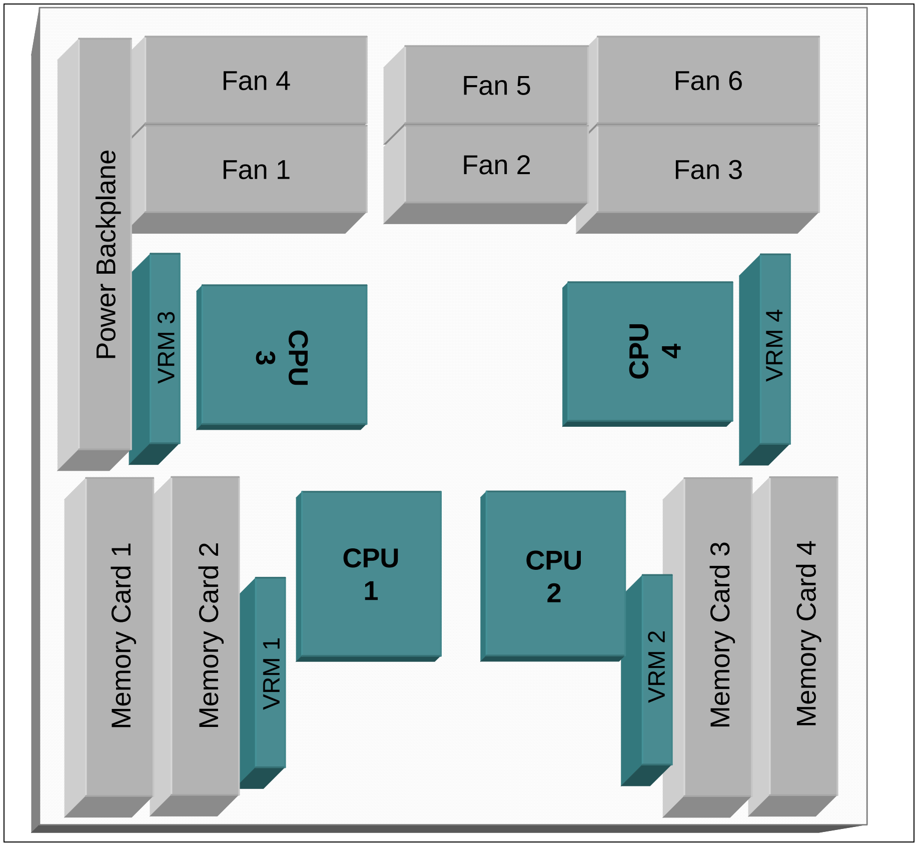


Figure 3-4 Processor board with CPU socket orientation

8. If you are installing a processor in socket 2, remove the heat-sink blank and store it for future use. Remove the protective cover, tape, or label from the surface of the microprocessor socket, if any is present.
9. Note that sockets 3 and 4 are mounted on the processor board with the processor's release levers on opposite sides. These sockets are oriented 180° from each other on the processor board.

   Verify the orientation of the socket *before you install* the processor in either of these sockets. Figure 3-5 on page 97 shows the orientation of the sockets.
10. Note that the processor air baffle is always located between socket 1 and socket 2, as shown in Figure 3-1 on page 90.
11. Lift the processor-release lever to the fully opened position, which is approximately a 135° angle. See in Figure 3-5 on page 97.

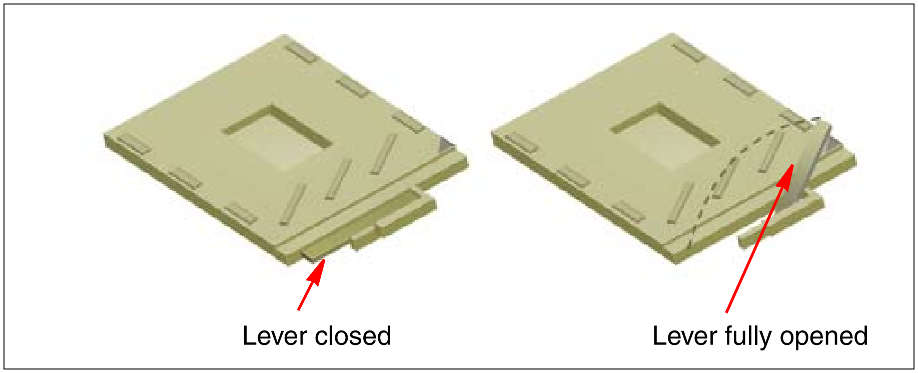


Figure 3-5 Processor board showing processor release lever

12. Position the processor over the socket, as shown in Figure 3-6, and then carefully press the processor into the socket. Close the processor release lever.

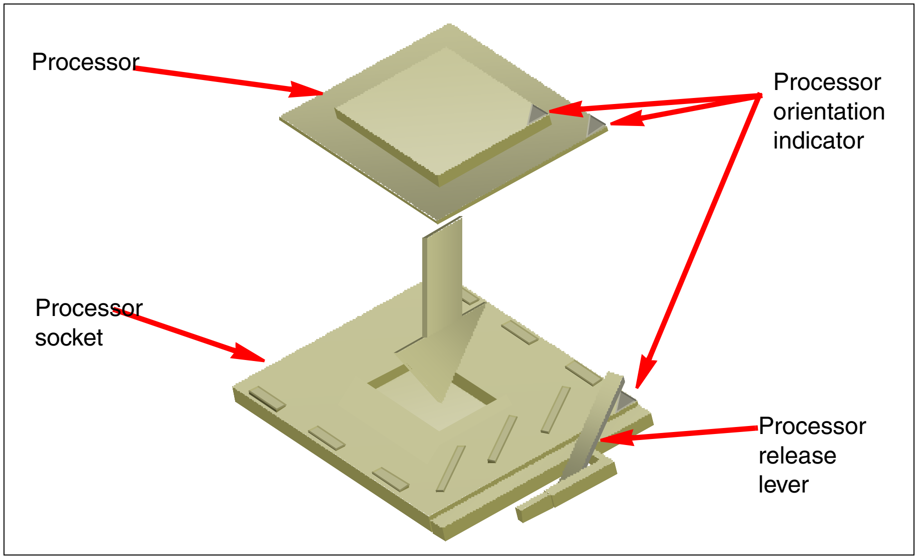


Figure 3-6 Processor orientation

13. Remove the heat-sink from its package and remove the cover from the bottom of it.

> **Hint:** The heat-sink in the processor option kit is already covered with the correct amount of thermal grease.

14. Release and rotate heat-sink retention clip to its fully opened state. See Figure 3-7. Position the heat-sink above the processor and align it with the alignment posts. Put the heat-sink on the processor, press on the top of the heat-sink and rotate the heat-sink retention lever up until it is locked.

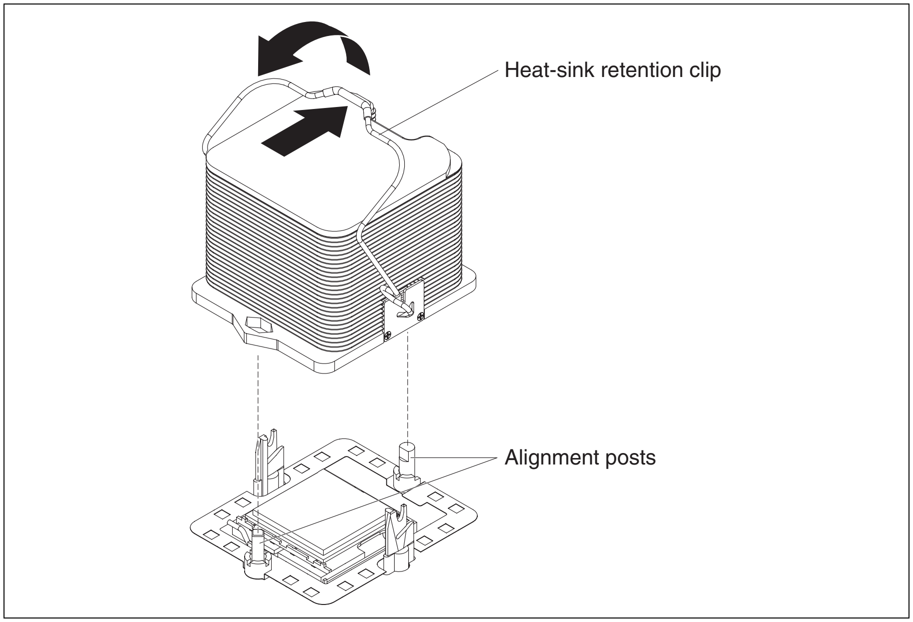


Figure 3-7 Heat-sink installation

15. Ensure that the air baffle between processors is correctly installed, as shown in Figure 3-1 on page 90.

16. Install a VRM in the connector next to the processor socket. The VRM and handle are shown in Figure 3-8 on page 99.

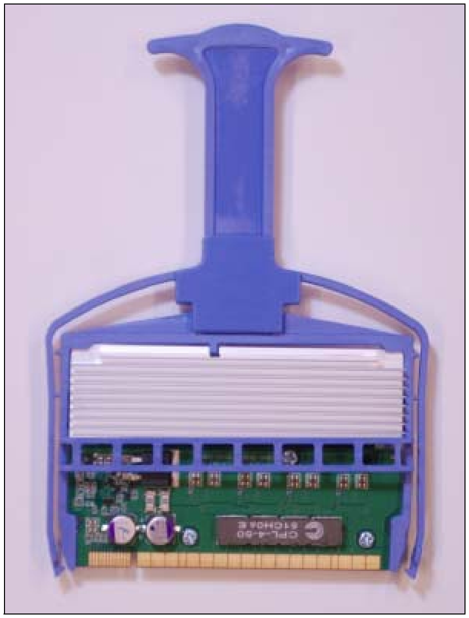


Figure 3-8 x3850 M2 and x3950 M2: voltage regulator module and handle

> **Note:** Make sure that the `Front` label on the VRM is facing the front of the server.

17. Close the media hood, replace the cover and the bezel if you are not installing other options.

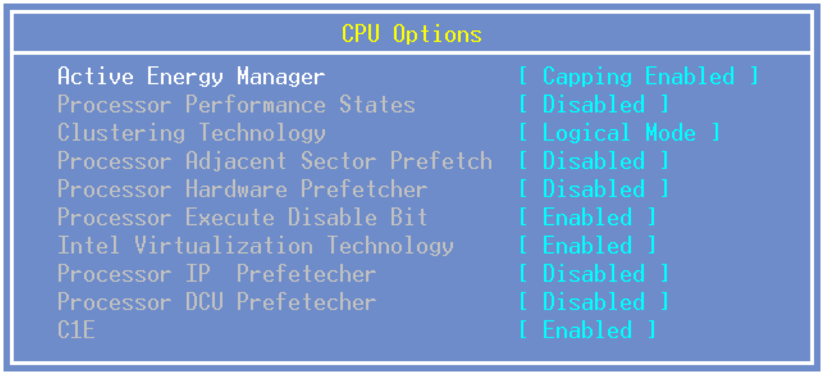
The installation of the processor option is now finished.

3.1.3 Processor (CPU) configuration options

The BIOS includes various processor settings, which you can adjust to the installed operating system and performance purposes.

To access the processor settings:

1. Power on or reboot the system.
2. When prompted during system startup, press `F1`.
3. In the Main Menu window, select **Advanced Setup**.
4. Select **CPU Options**. The selection menu in Figure 3-9 on page 100 opens.



```
                         CPU Options
  Active Energy Manager                     [ Capping Enabled ]
  Processor Performance States              [ Disabled ]
  Clustering Technology                     [ Logical Mode ]
  Processor Adjacent Sector Prefetch        [ Disabled ]
  Processor Hardware Prefetcher             [ Disabled ]
  Processor Execute Disable Bit             [ Enabled ]
  Intel Virtualization Technology           [ Enabled ]
  Processor IP  Prefetecher                 [ Disabled ]
  Processor DCU Prefetecher                 [ Disabled ]
  C1E                                       [ Enabled ]
```

Figure 3-9 BIOS CPU Options menu

The settings listed in the CPU Options menu are described in the following sections:

- "Active Energy Manager (power capping)" on page 100
- "Processor Performance States" on page 101
- "Clustering Technology" on page 105
- "Processor Adjacent Sector Prefetch" on page 107
- "Processor Hardware Prefetcher" on page 108
- "Processor Execute Disable Bit" on page 108
- "Intel Virtualization Technology" on page 109
- "Processor IP Prefetcher" on page 109
- "Processor DCU Prefetcher" on page 110
- "C1E" on page 110

Active Energy Manager (power capping)

Select this option to enable or disable the Active Energy Manager Power Capping. When power capping is enabled, the Active Energy Manager application can limit the maximum power that this system consumes.

Active Energy Manager is part of a larger power-management implementation that includes hardware and firmware components. Use Active Energy Manager to manage the power and thermal requirements of IBM servers and BladeCenter systems. We cover AEM in more detail in 6.4, "Active Energy Manager" on page 334.

Active Energy Manager 3.1 is an extension to IBM Director software. A stand-alone version of Active Energy Manager, which runs on top of Embedded Director, is also available.

Only the onboard Base Management Controller (BMC) has Active Energy Manager support. The BMC is on the system board and it shares the Ethernet port with the system. For Active Energy Manager to detect the system, you must configure the BMC IP address through BIOS. This is described in 6.1, “BMC configuration options” on page 300.

> **Note:** The IBM performance team observed a drop in performance when power capping is used in a x3950 M2 multinode configuration. The setting is therefore disabled and hidden if a multinode configuration is started. Power capping is still an available option for single-node configurations.

Processor Performance States

The Advanced Configuration and Power Interface (ACPI) specification defines three major controls (states) over the processor:

- Processor power states (C-states)
- Processor clock throttling (T-states)
- Processor performance states (P-states)

These controls are used to achieved the desired balance of:

- Performance
- Power consumption
- Thermal requirements
- Noise-Level requirements

Processor power states (C-states) are low-power idle states. This means the operating system puts the processor into different quality low-power states (which vary in power and latency), depending on the idle time estimate, if the operating system is idle. C0 is the state where the system is used. C1 to Cn are states that are then set to reduce power consumption. After the idle loop is finished the system goes back to the C0 state.

Processor clock throttling (T-states) is a passive cooling mechanism, which allows the platform to control and indicate the temperature at which clock throttling, for example, will be applied to the processor residing in a given thermal zone. Unlike other cooling policies, during passive cooling of processors, operating system power management (OSPM) may take the initiative to actively monitor the temperature in order to control the platform.

Processor performance states (P-states) are power consumption and capability states within the active or executing states. P-states allow the OSPM to make trade-offs between performance and energy conservation. It has the greatest effect when the states invoke different processor efficiency levels, as opposed to a linear scaling of performance and energy consumption, so they are more efficient than the power management features. Those P-states are placed in Intel’s SpeedStep technology, which knows one low and one maximum power state only. More states are defined in the Enhanced Intel SpeedStep® (EIST) technology, which is a major feature of the processors in the x3850 M2 and x3950 M2 servers.

Lowering the processor performance state when processor demand is low can significantly reduce CPU dynamic power consumption. These processor performance states can be changed very quickly in response to processor demand while software continues to execute. This technique, sometimes referred to as demand-based switching (DBS), allows the operating system to provide automatic scaling of the processor's power consumption in response to varying workloads, with no required user intervention and no perceivable effect to system performance.

The number of P-states depends on the type of processor. This information is defined in the ACPI table of the BIOS. The operating system reads the ACPI information and adjusts the core voltage followed by the core frequency, while the processor continues to run. The Processor Performance States option (shown in Figure 3-9 on page 100) is disabled by default; you enable it in BIOS.

The following sections show examples in Linux, Windows Server 2003, Windows 2008, and how you can identify the operating system power management

Linux

Various distributions of Linux operating systems, such as SUSE SLES10/SP2 or RHEL5U2 based on kernel 2.6, integrate power management features by default. The kernel processor frequency scaling subsystem can adjust the core frequency as it goes. After the P-states are enabled in the BIOS, the subdevice is found in the operating system at the following location:

`/sys/devices/system/cpu/cpu0/cpufreq`

To show the available core frequencies, which are adjustable, use the **`cat`** command, shown in Example 3-1.

Example 3-1 Command to show available core frequencies

```
# cat /sys/devices/system/cpu/cpu0/cpufreq/scaling_available_frequencies
2128000 1862000 1862000 1596000
#
```

The command can also show the current value of the core frequency, as indicated in Example 3-2.

Example 3-2 Command to show current core frequency

```
# cat /sys/devices/system/cpu/cpu0/cpufreq/cpuinfo_cur_freq
1596000
#
```

Various governors can be set to affect the power management policy: *ondemand*, *userspace*, *powersave*, and *performance*. Depending on the configured governor, the balance between power saving and performance can be adjusted:

- To display the current governor, use the **cat** command as shown in Example 3-3.
- To scale the frequency, and change the governor from the default `ondemand` to `userspace` (not case-sensitive), use the command shown in Example 3-4.
- To change to another frequency, use the command shown in Example 3-5.

Example 3-3 Command to display the current scaling governor

```
# cat /sys/devices/system/cpu/cpu0/cpufreq/scaling_governor
ondemand
#
```

Example 3-4 Change the governor to USERSPACE

```
cpufreq-set -g USERSPACE
```

Example 3-5 Command to change the frequency

```
# cpufreq-info -f -c 5
1862000
# cpufreq-set -c 5 -f 2128000
# cpufreq-info -f -c 5
2128000
```

Windows Server 2003

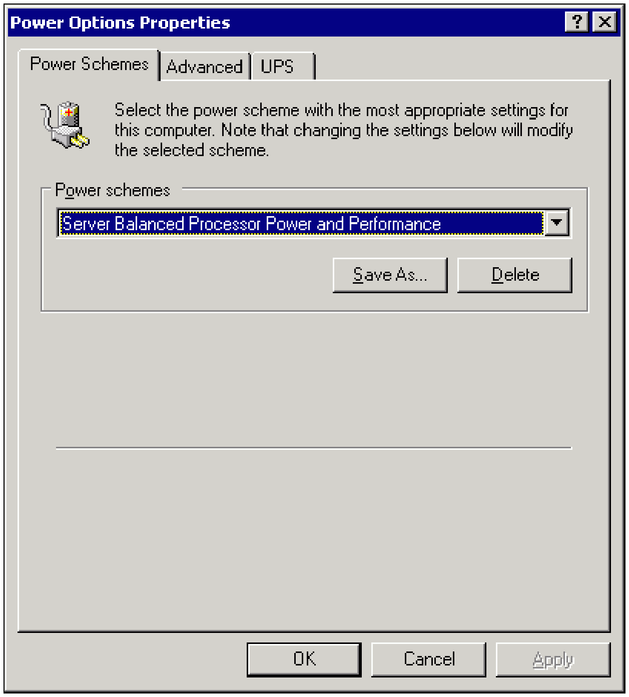
To utilize processor performance states, Windows Server 2003 must be using a power policy that enables DBS. By default, the power scheme `Always On` is set. This has the effect of running the system at full power regardless of workload demands. Windows includes a power policy named Server Balanced Processor Power and Performance that implements DBS by using the entire range of performance states that are available on the system.

Select the power scheme through either of the following methods:

- Command line, by using the **powercfg.exe** executable command:

  ```
  powercfg.exe -s "Server Balanced Processor Power and Performance"
  ```

- Power Options Properties windows, shown in Figure 3-10 on page 104.



Select the power scheme with the most appropriate settings for this computer. Note that changing the settings below will modify the selected scheme.

Power schemes: Server Balanced Processor Power and Performance

Figure 3-10 Windows Power Options

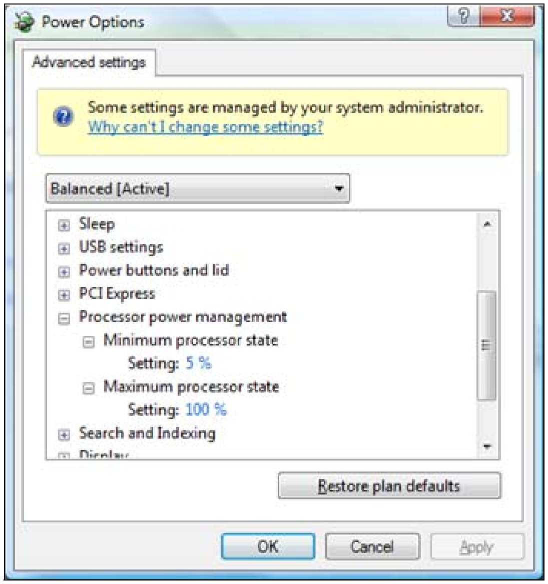
Windows Server 2008

Windows Server 2008 includes updated support for ACPI processor power management (PPM) features, including support for processor performance states and processor idle sleep states on multiprocessor systems. Windows Server 2008 implements PPM features by processor drivers that contain processor-specific routines to determine the presence of PPM capabilities. The correct loaded driver can be found in the following folder:

%SYSTEMDRIVE%\Windows\Inf %SYSTEMDRIVE%\Windows\Inf

For Intel processors, the Intelppm.sys file is used. Because the Intel Xeon is not listed in the support matrix, the Microsoft generic driver Processr.sys is used instead, until the support is available.

The power policies can be adjusted in the Power Options window. Each processor power policy includes an upper and lower limit, referred to as the *Maximum processor state* and the *Minimum processor state*. They determine the range of currently available P-states that Windows may use. These values are exposed in the Advanced settings panel of the Power Options window, shown in Figure 3-11 on page 105.



Some settings are managed by your system administrator.
Why can't I change some settings?

Balanced [Active]

- ⊞ Sleep
- ⊞ USB settings
- ⊞ Power buttons and lid
- ⊞ PCI Express
- ⊟ Processor power management
  - ⊟ Minimum processor state
    - Setting: 5 %
  - ⊟ Maximum processor state
    - Setting: 100 %
- ⊞ Search and Indexing
- ⊞ Display

Figure 3-11 Windows Server 2008: Power Options window, Advanced settings panel

You may set these values independently to define the bounds for any contiguous range of performance states, or they may be set to the same value to force the system to remain at a specific state. When you select a new target performance state, Windows Server 2008 chooses the closest match between the current power policy setting and the states available on the system, rounding up if necessary.

For details about adjusting and optimizing your system to balance performance to power consumption efficiency, see the following Microsoft Web page:

http://www.microsoft.com/whdc/system/pnppwr/powermgmt/ProcPowerMgmt.mspx

Clustering Technology

For certain operating systems, you must configure how the routing of processor interrupts in a multi-processor system is handled. A low-level value sets the multi-processor interrupt communication protocol (XAPIC). The settings are functional only, and do not affect performance.

In the Clustering Technology menu, choose the appropriate mode for your operating system, as advised in the operating system requirements, described in 5.4, “Installing the operating system” on page 264.

Although the IBM scalability chip can scale up to a total of eight nodes, IBM supports only up to four nodes. The Hurricane controller supports four front-side buses (FSBs), each of which connects one multi-core processor package. By default, the FSBs represent two processor agent clusters. Beyond this, the Hurricane controller can subdivide each of the four FSBs into its own cluster, which has a logical hierarchal cluster mode. A single node can represent four or eight processor packages.

> **Note:** Remember a single node is considered to have a maximum of four physical populated processor packages. However, each processor package has one, two, or three dies with two cores each.

The 7300-series processors that we support with the x3850 M2 and x3950 M2 are dual-core or quad-core processors. Therefore, they have four processor agents per processor package.

The IBM scalability chip can handle up to 128 agents in a maximum of eight x3950 M2 scaled system complexes.

The following sections discuss the three available cluster technology modes:

- Special mode
- Logical mode
- Physical mode

The sections also discuss the Linux and Windows operating systems regarding clustering.

Special mode

This mode was created temporarily to allow 64-bit Linux operating systems to run on the eX4 technology servers and the older X3 technology servers, if they do not support clustering. Although flat mode is not really a clustering mode, it is a mode by which most non-clustering operating systems can work simply, because it abstracts them from the real physical clustering of the processors as a result of the architecture. We do not recommend using this mode on X3 systems with other operating systems, because it could cause situations resulting in failure to boot. By definition, a maximum of eight agents are allowed, all are logical.

Because the special mode was developed for single-threaded processors only, this mode is not used on x3850 M2 and x3950 M2.

Logical mode

This mode is applicable to XAPIC-based systems. A maximum of 60 logical agents are supported, which means that 15 cluster IDs, each with four logical agents, can be used.

Physical mode

This mode is applicable to XAPIC based systems too. It allows up to 255 physical agent IDs. Although there is no definition of what has to be considered as a cluster ID in this mode, the chipset design has imposed a limitation because of clustering controllers for 16 possible cluster IDs. The physical mode allows up to 15 physical mode agents per cluster.

Linux operating systems

Linux distributions, such as Red Hat and SUSE, require special capabilities to support clustering beyond the shrink-wrapped flat mode. Future Linux distributions in general will boot scaled systems in physical mode and therefore are not affected when booting and POST has set up the topology within the restrictions of logical mode; logical mode is essentially a subset of physical mode.

Additionally, Linux distributions in physical mode can possibly support processor packages with four agents, each beyond the four-chassis limit of logical mode operating systems, and attain a maximum partition topology of eight chassis incorporating 128 processor agents. The actual number of processor agents supported by a particular Linux distribution, depends on the vendor. The assumption is that the newer IA32E(64T) distributions will support greater than 32 agents.

Windows operating systems

Windows Server does not require a custom hardware abstraction layer (HAL) because it has been incorporated into 2003 SP1 in order to support scaled systems. The 32-bit Windows operating systems are typically limited to 32 processor agents, by design. Windows Server 2003 x64 supports a maximum of 60 agents purely in logical mode. A special kind of implementation within the operating system is required to allow 64 agents, which is implemented in the 64-bit Windows Datacenter editions.

> **Note:** Changing the mode in clustering technology from logical to physical mode is not necessary unless you want to upgrade the system to work with a Linux distribution in a multinode configuration.

Processor Adjacent Sector Prefetch

When this setting is disabled (the default), the processor fetches only the sector of the cache line that contains the data currently required by the processor.

When it is enabled the processor fetches both sectors of a cache line when it requires data that is not currently in its cache.

This change here affects all CPUs.

Note: Intel recommends testing both enabled and disabled settings and then setting the adjacent sector prefetch accordingly after performance evaluation.

For instance, only one 64-byte line from the 128-byte sector will be prefetched with this setting disabled. This setting can affect performance, depending on the application running on the server and memory bandwidth utilization. Typically, it affects certain benchmarks by a few percent, although in most real applications it will be negligible. This control is provided for benchmark users who want to fine-tune configurations and settings.

Processor Hardware Prefetcher

When this setting is enabled, the processors are able to prefetch extra cache lines for every memory request. Recent tests in the performance lab have shown that you can get the best performance for most commercial application types if you disable this feature. The performance gain can be as much as 20% depending on the application.

For high-performance computing (HPC) applications, we recommend that you enable the Processor Hardware Prefetch option; for database workloads, we recommend you disable the option.

Note: Intel recommends that the Processor Hardware Prefetcher be enabled for server workloads similar to the Streams Benchmark, but that the actual setting should be determined by performance testing in your intended workload environment.

Processor Execute Disable Bit

Processor Execute Disable Bit (EDB or XD) is a function of new Intel processors which lets you prevent the execution of data that is in memory as though it was code. When this setting is enabled (the default), viruses or other malicious code are prevented from gaining unauthorized access to applications by exploiting buffer overruns in those applications.

If this option is enabled, and the operating system has marked the memory segment as containing data, then the processor will not execute any code in the segment. This parameter can be disabled in the BIOS, if the applications to run on the server have problems with Execution Prevention. For added protection, you might want to enable it, but you should first test your applications to ensure they can continue to run as expected before you enable the option in a production environment.

Note: This function is only used for 32-bit operating environments where the processor is in one of the following modes:

- Legacy protected mode, if Physical Address Extension (PAE) is enabled on a 32-bit operating system.
- IA-32e mode, when EM64T is enabled on a 64-bit operating system.

The operating system must also implement this function.

The XD feature is implemented in Linux OS as Data Execution Prevention (DEP) and Windows OS as Execute Disable Bit (EDB).

For more details of the Execute Disable Bit function, see:

`http://www.intel.com/technology/xdbit/index.htm`

Intel Virtualization Technology

To reduce the complexity of the hypervisor, which can reduce overhead and improve performance significantly, enable (default) this setting, if it is not already.

Processor IP Prefetcher

The purpose of the IP prefetcher, as with any prefetcher, is to predict what memory addresses will be used by the program and deliver that data just in time.

To improve the accuracy of the prediction, the IP prefetcher tags the history of each load using the Instruction Pointer (IP) of the load. For each load with an IP, the IP prefetcher builds a history and keeps it in the IP history array. Based on load history, the IP prefetcher tries to predict the address of the next load according to a constant stride calculation (a fixed distance or *stride* between subsequent accesses to the same memory area).

The IP prefetcher then generates a prefetch request with the predicted address and brings the resulting data to the Level 1 data cache.

This setting is disabled by IBM (the opposite of the Intel default). The observations in the IBM performance lab have shown that this prefetcher is negligible in most real applications. Although Intel recommends that the Processor Hardware Prefetcher option be enabled for some server workloads similar to the Streams benchmark, the actual setting should be determined by performance testing in your intended workload environment.

Processor DCU Prefetcher

The Data Cache Unit (DCU) prefetcher detects multiple readings from a single cache line for a determined period of time. It then loads the following line in the L1 cache; and one for each core too.

The hardware prefetch mechanisms are efficient, and in practice can increase the success rate of the cache subsystem. However, the prefetch can also have the opposite result. Frequent errors tend to pollute cache with useless data, reducing its success rate. This is why you may deactivate most of the hardware prefetch mechanisms. Intel recommends deactivating the DCU prefetch in processors that are intended for servers, because it can reduce performances in some applications.

> **Note:** In the BIOS, all four prefetchers are disabled by default. Most benchmarks in the performance lab indicate that this combination offers the best performance because many benchmarks typically have very high CPU and FSB utilization rates. Prefetching adds extra overhead, often slowly in a very busy system.
>
> However, try all combinations of these prefetchers on their specific workload; in many cases, other combinations help. Systems not running at high CPU and FSB utilizations might find prefetching beneficial.

C1E

This setting allows you to enable (by default) or disable the Enhanced Halt State (C1E).

If the Enhanced Halt State is enabled, the operating system allows the processor to alter the core frequency after sending an idle command such as HALT or MWAIT. The processor core speed slows down and then transitions to the lower voltage.

Enhanced Halt State is a low power state entered when all processor cores have executed the HALT or MWAIT instructions and Extended HALT state has been enabled. When one of the processor cores executes the HALT instruction, that processor core is halted; however, the other processor cores continue normal operation. The processor automatically transitions to a lower core frequency and voltage operating point before entering the Extended HALT state.

Note: The processor FSB frequency is not altered; only the internal core frequency is changed. When entering the low power state, the processor first switches to the lower bus to core frequency ratio and then transitions to the lower voltage.

While in the Extended HALT state, the processor will process bus snoops.

For more information about power saving modes, refer to the Intel Xeon 7200 and 7300 Processor Series Datasheet:

http://download.intel.com/design/xeon/datashts/318080.pdf

Important: All changes you make within the BIOS affects the particular local node only. Any changes you make must also be made on all nodes within a multinode configuration. Although you may have different settings on different nodes, it is not recommended.

3.2 Memory subsystem

The x3850 M2 and x3950 M2 systems allow you to add memory DIMMs to a maximum of four memory cards.

3.2.1 Memory options

All memory modules you want to install in your x3850 M2 or x3950 M2 server must be 1.8 V, 240-pin, PC2-5300 DDR II, registered SDRAM with ECC DIMMs.

The supported DIMM options are listed in Table 3-3:

Table 3-3 Supported DIMM options for x3850 M2 and x3950 M2

Option Part number[a]	Feature code[b]	Description
41Y2762	3935	2 GB kit (2x 1 GB DIMM) PC2-5300 CL5 ECC DDR2 667 MHz SDRAM LP RDIMM
41Y2771	3939	4 GB kit (2x 2 GB DIMM) PC2-5300 CL5 ECC DDR2 667 MHz SDRAM LP RDIMM
41Y2768	3937	8 GB kit (2x 4 GB DIMM) PC2-5300 CL3 ECC DDR2 SDRAM RDIMM

Option Part number[a]	Feature code[b]	Description
43V7356[c]	3938	16 GB kit (2x 8 GB DIMM) PC2-5300 CL5 ECC DDR2 667 MHz SDRAM LP RDIMM

a. The option part number contains two DIMMs of the same size.
b. The feature code contains one DIMM.
c. This option is supported only at x3950 M2 one-node and two-node.

The number of installed DIMMs and memory cards as shipped is specified in the server model description. You can obtain details in the latest version of BladeCenter and System x Reference Sheet (xRef) at:

http://www.redbooks.ibm.com/xref

The DIMMs operate at 533 MHz, to be in sync with the front-side bus. However, the DIMMs are 677 MHz PC2-5300 spec parts because these have better timing parameters than the 533 MHz equivalent. The memory throughput is 4.26 GBps, or 533 MHz x 8 bytes per memory port, for a total of 34.1 GBps with four memory cards.

The server supports up to four memory cards. See Figure 3-12 on page 113. Each memory card holds up to eight DIMMs to allow thirty-two DIMMs per chassis.

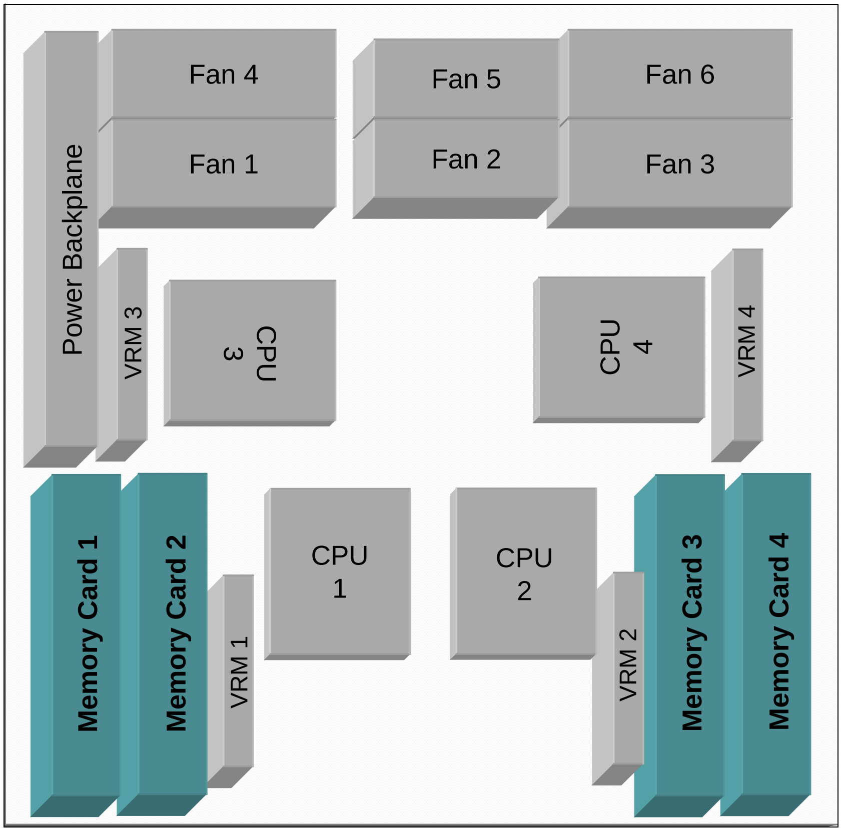


Figure 3-12 Memory card location

3.2.2 Memory card

The memory card embeds two memory controllers Nova x4, as components of the designated IBM eX4 chipset, to address up to four DIMMs on each memory controller. The memory controller embeds Xcelerated Memory Technology™.

You may order additional memory cards; use the part number in Table 3-4.

Table 3-4 Memory card part number

Option part number[a]	Feature code	Description
44E4252	4894	8-port Memory card

[a] You may install four memory cards in a chassis.

A memory card is fully populated with eight of the specified options DIMMs. Figure 3-13 shows a full populated memory card with eight DIMMs.

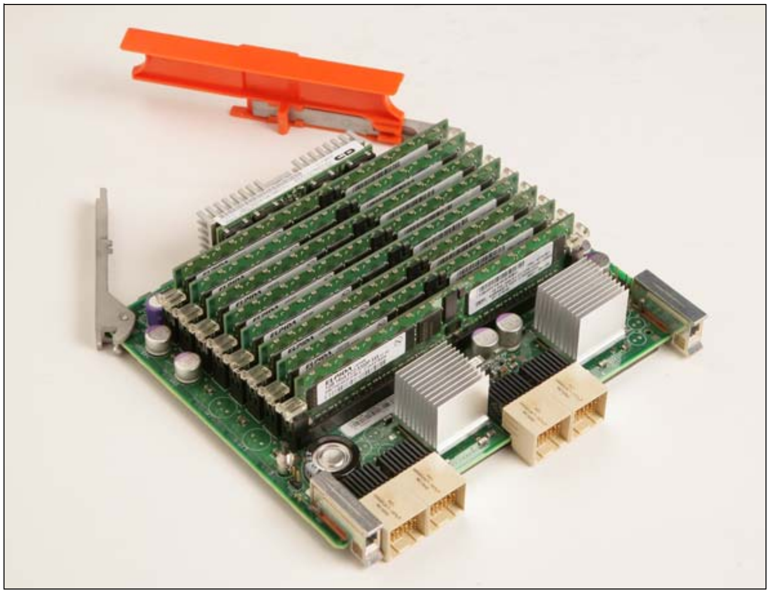


Figure 3-13 Memory Card, fully populated

You must install at least one memory card with one pair of DIMMs because it is two-way interleaved. These pairs of DIMMs must be the same size and type. The server must have this to operate.

> **Note:** When you install additional DIMMs on a memory card or a new memory card, make sure they are installed in pairs.
>
> You do not have to save new configuration information to the BIOS when you install or remove DIMMs. The only exception is if you replace a DIMM that was designated as `Disabled` in the Memory Settings menu. In this case, you must re-enable the row in the Configuration/Setup Utility program or reload the default memory settings.

In a multinode configuration, the memory in all nodes is combined to form a single, coherent physical address space.

If you replace the standard pair of DIMMs and install 32x 8 GB DIMMs and four memory cards, both the x3850 M2 and x3950 M2 can be expanded to 256 GB.

The XceL4v Dynamic Server Cache consumes 256 MB in each chassis of the main memory for use as L4 cache if a multinode system is formed, therefore giving a reduction in overall memory that is available to the operating system of 256 MB per node. The XceL4v architecture is discussed in 1.6.2, “XceL4v dynamic server cache” on page 30.

A minimum of 4 GB of memory must be installed in one node if you want to form a multinode complex. This is discussed in 4.4, “Prerequisites to create a multinode complex” on page 201.

From a performance view of point, all nodes should have the same amount of memory and the population within and over the memory cards should be balanced. This reduction in memory is reflected in the power-on self-test (POST) with the addition of a new line of text specifying the amount of available system main memory after the L4 scalability cache for each node has been subtracted. Example 3-6 shows what you see in a two-node complex.

Example 3-6 Two-node system

```
64GB Memory: Installed
512MB Memory: Consumed by Scalability
```

To replace or add any DIMMs, remove one or more of the installed memory cards, or add a new one.For an explanation of how this can even be done while the system and the operating system are up and running, refer to sections:

- 3.2.3, “Memory mirroring” on page 118
- 3.2.4, “Hot-swap memory” on page 119
- 3.2.5, “Hot-add memory” on page 120

Figure 3-14 on page 116 shows the layout of a memory card with its available status and failure LED indicators.

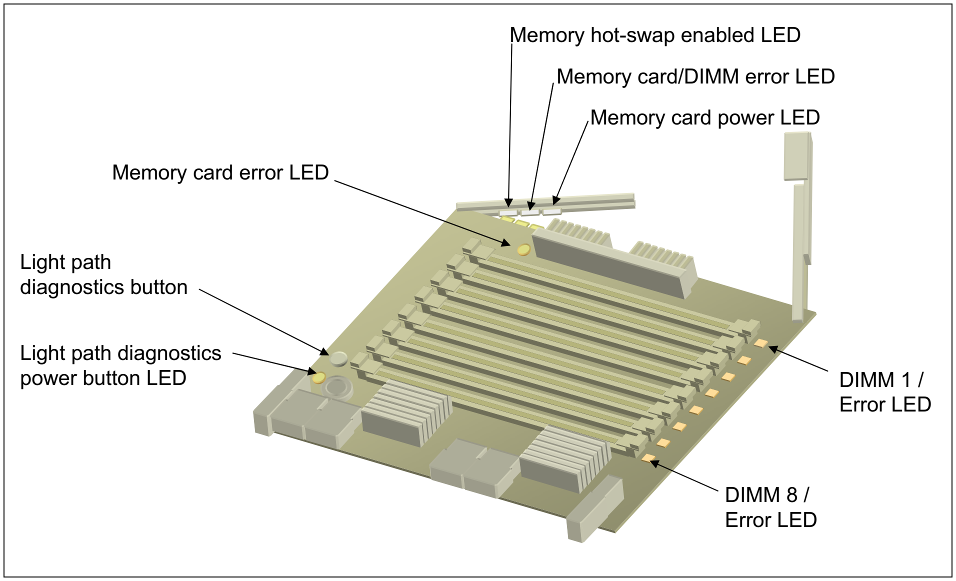

Memory hot-swap enabled LED
Memory card/DIMM error LED
Memory card power LED
Memory card error LED
Light path diagnostics button
Light path diagnostics power button LED
DIMM 1 / Error LED
DIMM 8 / Error LED

Figure 3-14 Memory card layout

Note the following key configuration rules:

- Because the x3850 M2 and x3950 M2 use two-way interleaving memory, DIMMs must be installed in matched pairs.
- Every x3850 M2 and single-node x3950 M2 must have at least one memory card installed and at least 2 GB of RAM installed
- For multi-node complexes, the primary node must have at least one memory card with 4 GB memory (4x 1 GB or 2x 2 GB) installed. The extra memory is required for node management. The additional nodes must have at least one memory card and at least 2 GB of RAM installed.
- Memory cards have the part number 44E4252. Two or four are standard in the x3850 M2 and x3950 M2, depending on the model. A maximum of four memory cards can be installed in a node. See the Reference Sheets (xREF) for specific information about models currently available:

  `http://www.redbooks.ibm.com/Redbooks.nsf/pages/xref`
- Each memory card has eight DIMM sockets.
- The three ways to fill the DIMMs sockets, depending on whether cost, performance, or reliability is the more important consideration, are:
  - Cost-effective configuration

    To minimize cost, install the memory DIMMs by filling each memory card before adding DIMMs to the next memory card. See Table 3-5.

Table 3-5 Low-cost installation sequence

DIMM pair	Memory card	Connector
1-4	1	1/5, 2/6, 3/7, 4/8
5-8	2	1/5, 2/6, 3/7, 4/8
9-12	3	1/5, 2/6, 3/7, 4/8
13-16	4	1/5, 2/6, 3/7, 4/8

– Performance-optimized configuration

Section 1.6, “IBM fourth generation XA-64e chipset” on page 27, describes eight independent memory ports.

Therefore, to optimize performance, install four memory cards and then spread the DIMMs, still installed in matched pairs, across all four memory cards before filling each card with two more DIMMs, see Table 3-6.

Table 3-6 High-performance installation sequence

DIMM pair	Memory card	Connector
1-4	1, 2, 3, 4	1/5
5-8	1, 2, 3, 4	2/6
9-12	1, 2, 3, 4	3/7
13-16	1, 2, 3, 4	4/8

– Reliability-increased configuration

To improve the reliability of your system fill your memory cards, as listed in Table 3-7. Depending on the memory population, if a DIMM fails, it can be removed and replaced by a new DIMM. The use of mirroring halves the memory that is available for the operating system.

Table 3-7 Memory-mirroring configuration

DIMM pair	Memory card	Connector
1, 2	1, 2, 3, 4	1/5
3, 4	1, 2, 3, 4	2/6
5, 6	1, 2, 3, 4	3/7
7, 8	1, 2, 3, 4	4/8

A more detailed description and the exact sequence for installation is provided in the *IBM: System x3850 M2 and x3950 M2 Installation Guide*, available from:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073028

If you want to install the full 256 GB, remove the existing DIMMs and fully populate the x3850 M2 and x3950 M2 with four memory cards, each with 8 GB DIMMs.

3.2.3 Memory mirroring

Memory mirroring is available on the x3850 M2 and x3950 M2 for increased fault tolerance. Memory mirroring is operating system-independent, because all mirroring activities are handled by the hardware. This setting can be changed as described in section 3.2.6, "Memory configuration in BIOS" on page 121.

The x3850 M2 and x3950 M2 have four separate memory power buses that each power one of the four memory cards. Figure 3-12 on page 113 shows the location of the memory cards, which are numbered 1 to 4, from left to right. The DIMM sockets and Memory card LEDs are shown in Figure 3-14 on page 116.

Mirroring takes place across two memory cards, as follows:

- The memory DIMMs in card 1 are mirrored to the memory DIMMs in card 2.
- The memory DIMMs in card 3 are mirrored to the memory DIMMs in card 4.

Therefore, with memory mirroring enabled in the BIOS, you can hot-swap any memory card if the hot-swap enabled LED is lit. For instructions to hot-swap a memory card, see *IBM: System x3850 M2 and System x3950 M2 User's Guide*:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073029

After memory-mirroring is enabled, the data that is written to memory is stored in two locations. For read operations, data is read from the DIMMs with the least amount of reported memory errors reported through memory scrubbing.

Table 3-8 shows the possible BIOS settings for the initialization of scrub control. The setting is accessed by going to (from the system startup Main Menu) **Advanced Setup** → **Memory Settings** → **Initialization Scrub Control**.

Table 3-8 Initialization scrub control

Setting	Function
Scrub on Every Boot	Performs full memory test on every boot
Scrub only after AC Cycle	Performs scrub only after AC has been removed or applied

Setting	Function
Disabled	Relies on standard memory test and run-time scrub engine to ensure memory is *good*

Note: A standard test is still performed across all memory and a run-time scrub engine is always enabled regardless of these settings.

If memory mirroring is enabled, then the mirrored copy of the data from the damaged DIMM is used until the DIMM replaced. After the damaged DIMM is replaced, memory mirroring copies the mirrored data back to the new DIMM.

Key configuration rules of memory mirroring are as follows:

- Memory mirroring must be enabled in the BIOS (it is disabled by default).
- Both memory cards must have the same total amount of memory, and must have identical DIMMs. In other words, DIMMs must be installed in matched quads to support memory mirroring. Partial mirroring is not supported. Refer to Table 3-7 on page 117 for information about the exact installation order required.

Note: Because of memory mirroring, you have only half of the total amount of memory available. If 8 GB is installed, for example, then the operating system sees 4 GB minus half the total XceL4v Dynamic Server Cache, if this is a multinode system, after memory mirroring is enabled.

3.2.4 Hot-swap memory

The x3850 M2 and x3950 M2 support hot-swap memory. If a DIMM fails, it can be replaced with a new DIMM without powering down the server. This advanced feature allows for maximum system availability. Hot-swap memory requires that memory mirroring be enabled. This setting can be changed as described in section 3.2.6, “Memory configuration in BIOS” on page 121.

To easily identify whether hot-swap is enabled and the status of power to the memory card, each memory card has a green memory hot-swap enabled LED, and a green memory card power LED on the top panel of the memory card, as shown in Figure 3-14 on page 116. The memory card has eject levers with sensors, so that the system can recognize when a memory card is being removed and power down that card’s slot accordingly.

To hot-swap a failed DIMM:

1. Verify that memory mirroring and hot-swap are enabled by checking the memory hot-swap enabled LED on the memory cards.

   When a DIMM fails, you are alerted with the memory LED on the light path diagnostics panel and by other means with the service processor, if this has been configured.

2. Locate the memory card that has the failed DIMM by identifying which memory card has the lit memory error LED.
3. Remove the memory card containing the failed DIMM.
4. Press the button on the memory card to identify which DIMM has failed. The LED next to the failed DIMMs light up.
5. Replace the failed DIMM and reinsert the memory card.

For details about hot-swapping memory correctly and which sequence to follow, see *IBM: System x3850 M2 and System x3950 M2 User's Guide*:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073029

3.2.5 Hot-add memory

The hot-add memory feature enables you to add DIMMs without turning off the server. This setting can be changed as described in 3.2.6, "Memory configuration in BIOS" on page 121.

Requirements for enabling the hot-add memory feature on the server are:

- The operating system must support the adding of usable system memory to a running operating system. This is done with an ACPI sequence.

  Currently, the only operating systems that have this capability and support on the x3850 M2 and x3950 M2 are Windows Server 2003 and Windows Server 2008, both Enterprise Edition and Datacenter Edition.

- Memory hot-add must be specifically enabled in the BIOS setup. When this is done, the system allocates blank windows of memory space for future memory additions. By enabling hot-add, memory mirroring will automatically be disabled.
- Memory cards 2 and 4 must not be installed because these are the *only* cards that can be hot-added.
- If only one memory card, memory card 1, is installed prior to the hot-add operation, then *only* one more memory card may be added in slot 2.
- If two memory cards are installed in slots 1 and 3, then two additional memory cards can be added in slots 2 and 4.

- The DIMMs must be added in matched pairs, that is, two at a time, and they must also match the equivalent pair of DIMMs on the matching memory card on the other power bus.
- A minimum of 4 GB of memory must be installed in the server for hot-add memory to be available. Additionally, for 32-bit operating systems, the Physical Address Extension (PAE) mode has to be enabled to take advantage of the additional memory.

For details about performing a hot-add operation, and restrictions, see *IBM: System x3850 M2 and System x3950 M2 User's Guide*:

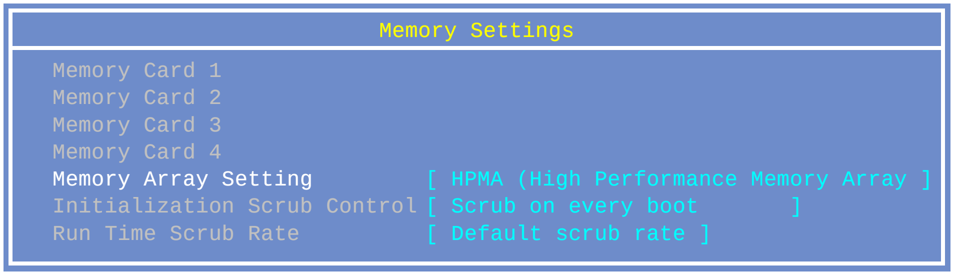
http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073029

3.2.6 Memory configuration in BIOS

The BIOS includes settings related to processor and operating system configuration and performance purposes. You can adjust the settings.

To access the settings and configure the memory subsystem in the server's BIOS setup:

1. Power on or reboot the system.
2. Press F1 during system startup, when prompted.
3. From the Main Menu window, choose **Advanced Setup**.
4. Select **Memory Options**. The Memory Settings selection menu opens, as shown in Figure 3-15.



```
                         Memory Settings
  Memory Card 1
  Memory Card 2
  Memory Card 3
  Memory Card 4
  Memory Array Setting          [ HPMA (High Performance Memory Array ]
  Initialization Scrub Control [ Scrub on every boot        ]
  Run Time Scrub Rate           [ Default scrub rate ]
```

Figure 3-15 Memory BIOS settings

Notes:

- As previously mentioned, hot-add and hot-swap are mutually exclusive. You can enable only one of these features.
- If you plan to enable hot-add memory and you have a x3850 M2 or x3950 M2 system that comes standard with two memory cards, move memory card 2 to slot 3 to be able to hot-add memory cards in slots 2 and 4.
- After you add a memory card that has two DIMMs, you cannot add more memory to that same memory card without powering off the server.
- Enabling hot-add reserves a portion of the memory map for the memory that can be hot-added in the future. If you do not plan to use hot-add, we recommend that you do not enable this feature in BIOS.

Memory Array Setting (memory mode)

Available memory modes and the features they enable are listed in Table 3-9.

Table 3-9 Memory configuration modes in BIOS

Mode	Memory ProteXion	Memory Mirroring	Hot-swap memory	Hot-add memory
HPMA (high performance memory array)	Yes	Disabled	Disabled	Disabled
FAMM (full array memory mirroring)	Yes	Yes	Yes	Disabled
HAM (hot-add memory)	Yes	Disabled	Disabled	Yes

The memory configuration mode you select depends on the memory features you want to use. Select one of the following modes:

- **HPMA** if you are *not* using mirroring, hot-swap, or hot-add. This is now the default or standard setting.
- **FAMM** enables memory mirroring and hot-swap.
- **HAM** enables hot-add in the future.

Note: The memory setting must be the same for all nodes in a multinode complex before merging the scalable partition. This requires a KVM connection to each node before the scalable partition is created.

Unlike with the x3850, the x3850 M2 and x3950 M2 support Memory ProteXion with the HPMA setting, providing maximum performance, and they continue to provide the reliability of Redundant Bit Steering (RBS).

Initialization Scrub Control

This setting allows you to configure the frequency of the memory initialization scrub which occurs at the beginning of POST/BIOS execution. Memory correction technologies are described in 1.8, “Memory subsystem” on page 39.

In very large memory arrays, this particular memory scrub can take up to 10 minutes so you may choose to either disable this feature or only perform the scrub when power has first been applied to the system.

The purpose of the memory initialization scrub is to detect catastrophic or *hard* memory errors, and disable memory devices that generate these errors.

Note that disabling the scrub or performing only the scrub after an AC cycle does not eliminate the normal memory error detection capabilities of the system. Any run-time correctable or uncorrectable memory error is still detected and the failing memory device or devices are logged in the system event logs.

Run Time Scrub Rate

Use this setting to configure the rate of the run-time hardware memory scrub engine. The hardware memory scrub engine is responsible for checking system memory looking for correctable or uncorrectable memory errors.

If you set the scrub rate to default, the chipset is configured to scrub the entire memory array every 24 hours. The setting can help to ensure maximum system performance by allowing the scrub engine to run only in this low speed mode.

You can use the fastest scrub setting if maximum system performance is not as essential as maximum memory reliability. With this setting, the chipset scrubs the entire memory array every 10 minutes.

Important: All changes you do within the BIOS affects the particular local node only. Any changes you make must also be made on all nodes within a multinode configuration. Although you can have different settings on several nodes, it is not recommended.

3.3 Internal drive options and RAID controllers

This section describes the disk storage subsystem of your x3850 M2 and x3950 M2 server system and upgrade options.

3.3.1 LSI 1078 SAS onboard controller

The x3850 M2 and the x3950 M2 contain an onboard LSI 1078 SAS controller also defined as RAID on motherboard (RoMB) that you can use to setup RAID-0 and RAID-1 with a fixed stripe size of 64 KB. As shown in Figure 3-16, the LSI 1078 SAS controller is wired with two x4 3.0 Gbps PCI Express ports, one for internal connectivity and one for external connectivity.

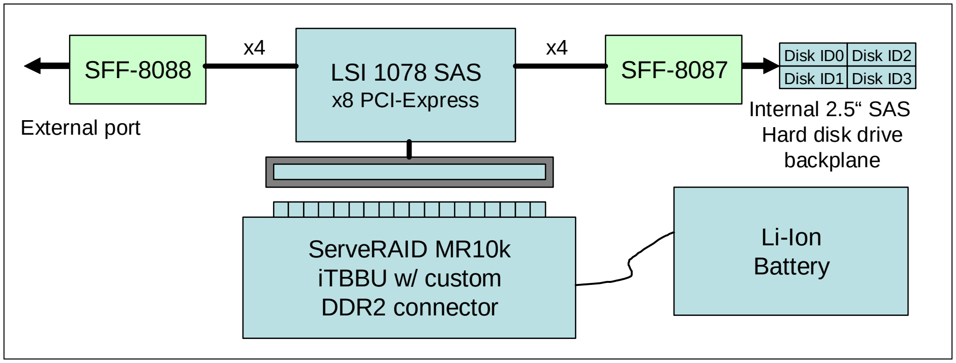


Figure 3-16 x3850 M2 and x3950 M2 SAS storage subsystem

The LSI 1078 SAS controller operates in the Integrated RAID (IR) mode that supports the internal port by communicating through a x4 SAS cable connection to the SAS 4-port hot- swap backplane.

The SAS LSI1078 IR controller has the following features:

- RAID level 0: Integrated Mirroring (IM)
- RAID level 1: Integrated Striping (IS)
- Two-disk IM mirrored volumes
- Two volumes maximum in a mixture of RAID-0 and RAID-1 arrays
- 10 disks total in one volume of IS
- 14 disks total in two volumes
- 20 TB virtual disk size limit (limitation by 64-bit addressing)
- All physical disks are visible in the operating system

- RAID level migration from IR to MR mode:
  - RAID 1 + RAID 1 → RAID 1 + RAID 1

    Two created RAID 1 volumes in the IR (integrated) mode can be imported to the MR (MegaRAID) mode.
  - RAID 0 + RAID 1 → RAID 0 + RAID 1

    A striped and mirrored mode created in IR mode can be imported to a RAID 0 and RAID 1 in MR mode.
  - Two hot spare drives can be imported

    Working in the IR mode allows the implementing of up to two hot spare drives, which will also be imported to the MR mode.

Figure 3-17 shows the internal SAS cable with the hot-swap backplane. The cable is an industry standard MiniSAS 4i cable with SFF-8087 cable and board connectors from the I/O board to the hot-swap backplane.

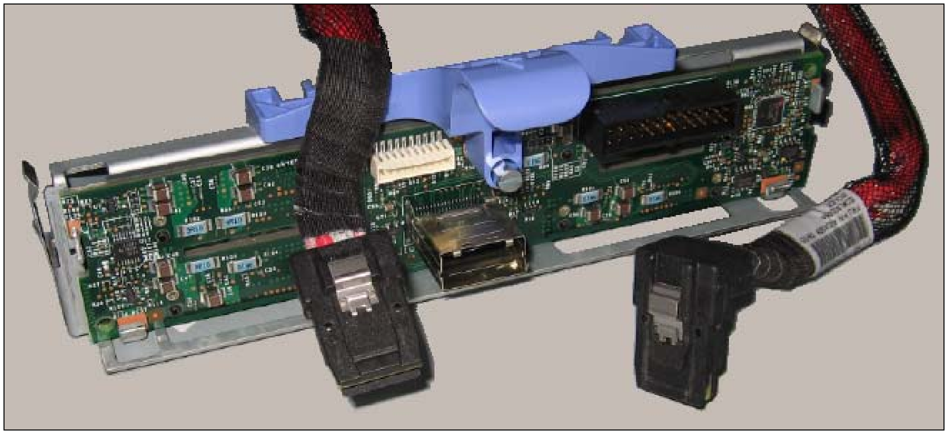


Figure 3-17 x3850 M2 and x3950 M2 showing internal SAS cabling

The hot-swap backplane is mounted in the media hood assembly and allows it to attach up to four internal 2.5-inch SAS hot-swap disks drives. The internal SAS connector on the I/O board located near the RSA II cable connector, as shown in Figure 3-18 on page 126.

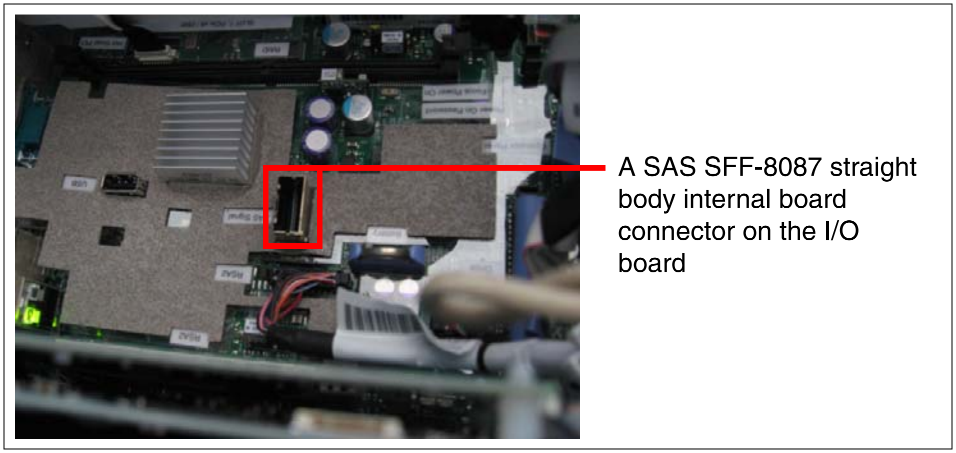

A SAS SFF-8087 straight body internal board connector on the I/O board

Figure 3-18 x3850 M2 and x3950 M2: internal SAS SFF-8087 board connector

Attaching an external storage expansion EXP3000 to the external port is possible. See Figure 3-19. The LSI1078 SAS onboard controller can use this port in the IR mode in configuration of up to 10 disk in RAID-0 in one volume.

> **Note:** For significant performance reasons, IBM recommends you attach an EXP3000 to the external SAS port only when you have the ServeRAID-MR10k installed. Connecting an EXP3000 to the external SAS port with just the onboard LSI 1078 controller is not recommended and is not supported.

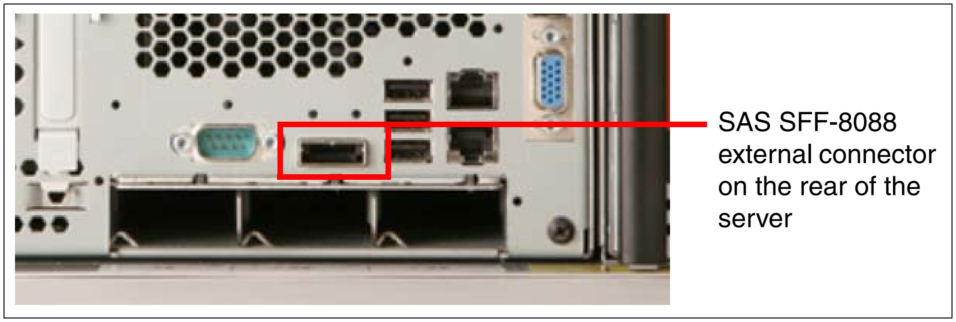


Figure 3-19 x3850 M2 and x3950 M2: external SAS SFF-8088 connector

By upgrading your server with the ServeRAID-MR10k adapter, you enable further RAID features. This card works in the MegaRAID (MR) mode. The adapter is described in 3.3.3, “ServeRAID-MR10k RAID controller” on page 128.

You may also add the ServeRAID-MR10M adapter to increase disk space with additional expansion units. Read more about this adapter, and the connection with expansion units in:

- Section 3.3.4, “ServeRAID-MR10M SAS/SATA II controller” on page 135
- Section 3.3.5, “SAS expansion enclosure (unit)” on page 142

3.3.2 SAS disk drive options

The x3850 M2 and x3950 M2 servers have four internal 2.5-inch hot-swap SAS disk drive bays. The hard disk drive tray is located in the media hood assembly in the middle of the server as shown in Figure 3-20.

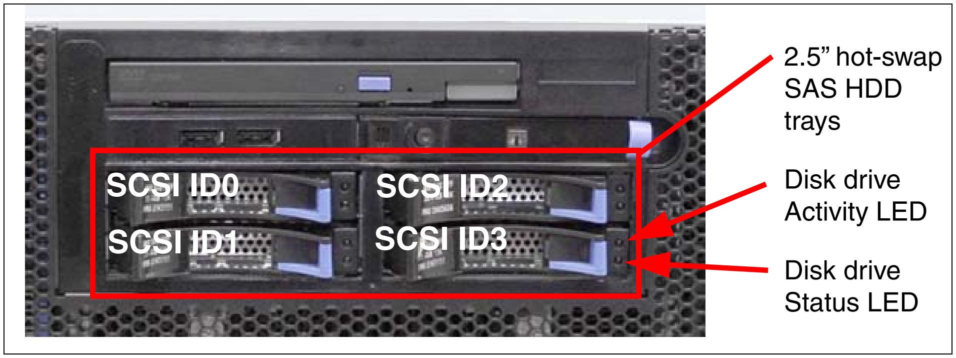


Figure 3-20 Media hood: hard disk drive bays

Important: To prevent damage to the disk, we recommend you firmly secure the media hood before installing or removing a SAS disk.

Depending on the level of RAID you want to configure, up to 584 GB of disk space can be used internally by using four 146 GB drives. Table 3-10 shows the supported disks.

Table 3-10 Supported internal disk options for x3850 M2 and x3950 M2

Part number	Description
40K1052	2.5 inch SAS 73 GB 10 K SAS
43X0824	2.5 inch SAS 146 GB 10 K SAS
43X0837	2.5 inch SAS 73 GB 15 K SAS

3.3.3 ServeRAID-MR10k RAID controller

To extend the basic RAID-0 (IS) and RAID-1 (IM) provided by the internal LSI1078 SAS controller and support the external SAS port, install the optional ServeRAID-MR10k controller. The ServeRAID-MR10k controller is a PCIe x8 RAID controller implemented as a DIMM card with 240 pins and has 256 MB ECC DDR2 cache.

The ServeRAID-MR10k is installed into a dedicated socket on the I/O board near the PCI Express slots as shown in Figure 3-21.

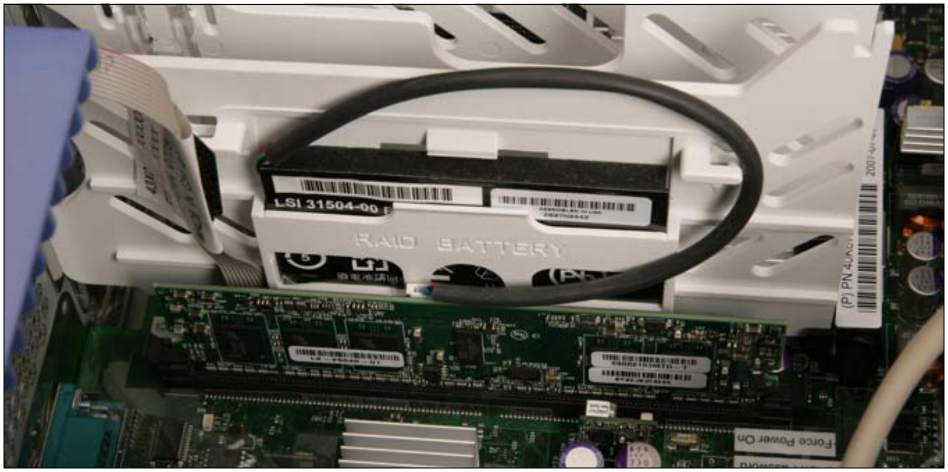


Figure 3-21 Optional installed ServeRAID-MR10k

The ServeRAID-MR10k supports stripe sizes from 8 KB to 1024 KB. The default stripe size is 128 KB. The SAS drive can be driven with up to 3 GBps for each port in full-duplex mode.

The controller supports up to 64 virtual disks, and up to 64 TB logical unit numbers (LUNs). It supports up to 120 devices on the external x4 port. IBM supports the use of up to nine cascaded and fully populated EXP3000 enclosures on the external port with up to a total of 108 disk drives.

The ServeRAID-MR10k controller has an LSI 1078 ASIC on the designated DIMM card. The cache data is secured by the iTBBU package as described in “Intelligent transportable battery backup unit (iTBBU)” on page 130.

RAID levels

After installing the ServeRAID-MR10k RAID controller, the following RAID levels may be used:

- RAID-0

  Uses striping to provide high data throughput, especially for large files in an environment that does not require fault tolerance.

- RAID-1

  Uses mirroring so that data written to one disk drive is simultaneously written to another disk drive. This is good for small databases or other applications that require small capacity but complete data redundancy.

- RAID-5

  Uses disk striping and parity data across all drives (distributed parity) to provide high data throughput, especially for small random access.

- RAID-6

  Uses distributed parity, with two independent parity blocks per stripe, and disk striping. A RAID 6 virtual disk can survive the loss of two disks without losing data.

  Note: The MR10k implements a variation of RAID-6 to allow usage of three hard disk drives.

- RAID-10

  A combination of RAID-0 and RAID-1, consists of striped data across mirrored spans. It provides high data throughput and complete data redundancy but uses a larger number of spans.

- RAID-50

  A combination of RAID-0 and RAID-5, uses distributed parity and disk striping, and works best with data that requires high reliability, high request rates, high data transfers, and medium-to-large capacity.

  Note: Having RAID-0 and RAID-5 virtual disks in the same physical array is not recommended. If a drive in the physical RAID-5 array has to be rebuilt, the RAID-0 virtual disk can result in a rebuild failure.

- RAID-60

  A combination of RAID-0 and RAID-6, uses distributed parity, with two independent parity blocks per stripe in each RAID set, and disk striping. A RAID-60 virtual disk can survive the loss of two disks in each of the RAID-6 sets without losing data. It works best with data that requires high reliability, high request rates, high data transfers, and medium-to-large capacity.

Note: RAID-50 and RAID-60, which are supported only externally, require at least six hard disk drives.

Key features

The ServeRAID-MR10k controller has the following features:

- Logical drive migration

  You can increase the size of a virtual disk while the disk is online by using RAID-level migration and by installing additional disks.

- Global hot spare, dedicated hot spare

  LSI defines any additional disks as global hot spares. These disk drives can reconstruct your virtual disk in case of disk failures. Global hot spares are defined to manage failing disks over all virtual drives.

  You can assign a hot spare to a specific volume instead of to all available volumes. Dedicated hot spares are used to recover from a failed drive in the assigned virtual drive only.

- Rebuild and rapid restore features

  These reliability features help to secure your data.

- Check consistency

  This feature fixes media errors and inconsistencies.

- Patrol read

  This background operation, also known as data scrubbing, checks for media errors in configured drives.

- Selectable boot virtual disk

  The first eight virtual disks can be chosen as a boot device.

Intelligent transportable battery backup unit (iTBBU)

An intelligent transportable battery backup unit (iTBBU) is attached to the ServeRAID-MR10k by a short cable as shown in Figure 3-22 on page 131. The iTBBU has the following characteristics:

- Intelligent: The iTBBU can automatically charge the battery pack and communicate to the server the battery status information such as voltage, temperature, and current.
- Transportable: The iTBBU can be used to move cached data to another system while the battery package is connected to the DIMM, if that data has

not been written to a disk. For example, this could be necessary if the server fails after an unexpected power failure. After you install the iTBBU and the RAID DIMM to another system, it flushes the unwritten data preserved in the cache to the disk.

- Ability to monitor: An integrated chip monitors capacity and other critical battery parameters. It uses a voltage-to-frequency converter with automatic offset error correction for charge and discharge counting. This chip communicates data by using the system management bus (SMBus) 2-wire protocol. The data available includes the battery's remaining capacity, temperature, voltage, current, and remaining run-time predictions.

The ServeRAID-MR10k controller and the iTBBU battery package shown in Figure 3-22 can be ordered with the part number in Table 3-11.

Table 3-11 Part number for ServeRAID MR10k

Part number	Description
43W4280	ServeRAID-MR10k controller and iTBBU battery package

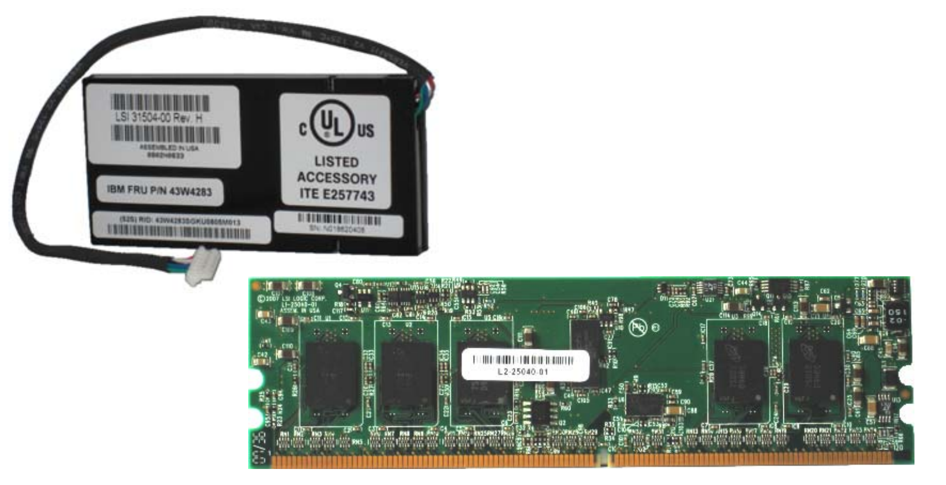


Figure 3-22 ServeRAID MR10k and iTBBU battery package

The battery protects data in the cache for up to 72 hours, depending on operating environment. IBM recommends that the battery be replaced annually. Replacement part numbers are listed in Table 3-12 on page 132.

Important: Before the battery in the iTBBU can effectively protect from a failure, it must be charged for at least six hours under normal operating conditions. To protect your data, the firmware changes the Write Policy to *write-through* (effectively bypassing the cache and battery) until the battery unit is sufficiently charged. When the battery unit is charged, the RAID controller firmware changes the Write Policy to *write-back* to take advantage of the performance benefits of data caching.

Table 3-12 Customer replacement part numbers

Customer replaceable unit (CRU) part numbers	Description
43W4283	iTBBU battery package
43W4282	ServeRAID MR10k adapter

Note: After the iTBBU battery is attached to the ServeRAID-MR10k controller, the battery must be charged for 24 hours at the first-time use to become fully charged. This ensures the maximum usefulness of the Li-Ion battery.

While the battery is charging for the first time, the controller cache is disabled and set to the *write-through* (cache disabled) mode and is changed back to the
write-back (cache enabled) mode automatically.

Installation guidelines

Consider the following guidelines when using RAID or installing the ServeRAID MR10k:

- No rewiring of the existing internal cabling is required when the ServeRAID-MR10k is installed in an x3850 M2 or x3950 M2.
- A RAID array created with the SAS LSI 1078 can be migrated for use with the ServeRAID-MR10k, but the reverse is not possible.

  This means that if you create RAID-0 (IS) and RAID-1 (IM) arrays using the onboard LSI 1078 Integrated RAID controller, and later install a ServeRAID-MR10k, you are given the option to convert those arrays to the format used by the MR10k. However, if you want to later remove the MR10k, you must first *save* all your data because the data in those arrays will be *inaccessible* by the LSI 1078 Integrated RAID controller.

For more details, see Chapter 3 of the *IBM: ServeRAID-MR10k User's Guide* available from:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5074104

- The onboard LSI 1078 and the ServeRAID-MR10k are not supported with the ServeRAID Manager tool. Use the MegaRAID Storage Manager (MSM) instead.
- One or more arrays can be formed using both the four internal disks and disks in an external disk enclosure such as the EXP3000 attached to the external SAS port.

Warnings:

- Prior to inserting the ServeRAID MR10k and converting your arrays you *must* install the ServeRAID MR10 driver. Failure to do so prior to the conversation will render all data on those drives inaccessible, permanently.
- Existing arrays (created using the onboard RAID controller) will be imported into MegaRAID arrays and they cannot be converted back again. This is a permanent migration.

Installation instructions

To install the ServeRAID-MR10k controller and the battery package:

1. Turn off the server and peripheral devices, and disconnect the power cords and all external cables as necessary to replace the device.
2. Remove the server cover.
3. From the server, remove the divider that contains the battery holder.
4. Open the retaining clip on each end of the connector.
5. Touch the static-protective package that contains the DIMM to any unpainted metal surface on the outside of the server; then, remove the DIMM from the package.
6. Turn the DIMM so that the keys align correctly with the slot.
7. Insert the DIMM into the connector by aligning the edges of the DIMM with the slots at the ends of the connector.
8. Firmly press the DIMM straight down into the connector by applying pressure on both ends simultaneously. The retaining clips snap into the locked position when the DIMM is seated in the connector.
9. Install the iTBBU in the divider that contains the battery holder.
10. Install the divider that contains the iTBBU holder in the server.

11. Route the iTBBU cable through the cable routing guides on the divider to the DIMM.

12. Insert the battery pack harness at the end of the cable into the J1 connector on the backside of the DIMM. See Figure 3-23.

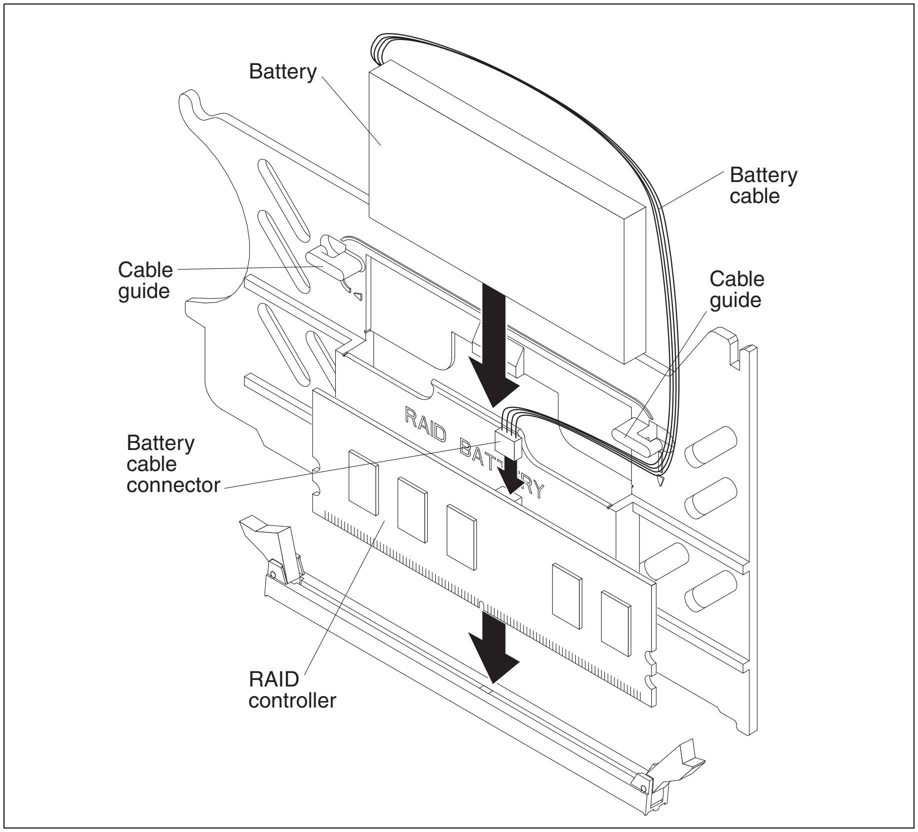


Figure 3-23 Installing the ServeRAID-MR10k

13. Reinstall the server cover and reconnect the power cords.

14. Turn on the power: first to the enclosure (if one is connected), then to the system.

15. Check that the ServeRAID MR10k controller is initialized correctly during POST. The text in Example 3-7 appears.

Example 3-7 ServeRAID-MR10k initialization in POST

```
LSI MegaRAID SAS-MFI BIOS
Version NT16 (Build Nov 20, 2007)
Copyright(c) 2007 LSI Corporation
```

```
HA -0 (Bus 4 Dev 0) IBM ServeRAID-MR10k SAS/SATA Controller
FW package: 8.0.1-0029
```

You have completed the installation of the ServeRAID MR10k controller.

For guidance with installing an SAS expansion enclosure, see 3.3.5, "SAS expansion enclosure (unit)" on page 142. To configure your RAID controller, see 3.4, "Configuring RAID volumes" on page 154.

3.3.4 ServeRAID-MR10M SAS/SATA II controller

The PCI Express ServeRAID-MR10M SAS/SATA II controller is also based on the LSI 1078 ASIC chip RAID-on-card (ROC) design. It is a PCIe x8 card and is a small form factor MD2 adapter with 2U or 3U bracket capability. Use the 3U bracket for the x3950 M2 and x3850 M2 servers. Figure 3-24 is a block diagram of the MR10M.

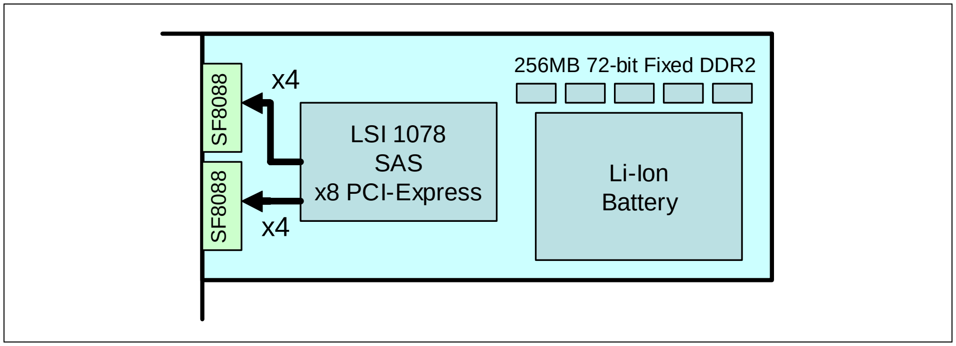


Figure 3-24 ServeRAID MR10M layout

Each channel is attached to an external x4 port SAS SFF-8088 connector. You can use these ports to connect more local disks to your x3850 M2 and x3950 M2 using external expansion boxes. Read more about the external expansion capability in section 3.3.5, "SAS expansion enclosure (unit)" on page 142.

This adapter has 256 MB 667 MHz ECC SDRAM memory for cache on the card as shown in Figure 3-25 on page 136. The cache has battery backup; the battery is standard with the card.

The controller can have up to:

- 64 virtual disks
- 64 TB LUN
- 120 devices on each of both external x4 SAS ports.

IBM supports using up to 9 cascaded, fully-populated EXP3000 per SAS/SATA connector on each of the external ports, with up to 108 SAS/SATA II hard disk drives per channel to maximum of 216 SAS/SATA II hard disk drives on both channels.

The ServeRAID-MR10M has the same key features as the MR10k as described in “Key features” on page 130.

Figure 3-25 shows the components of the ServeRAID-MR10M.

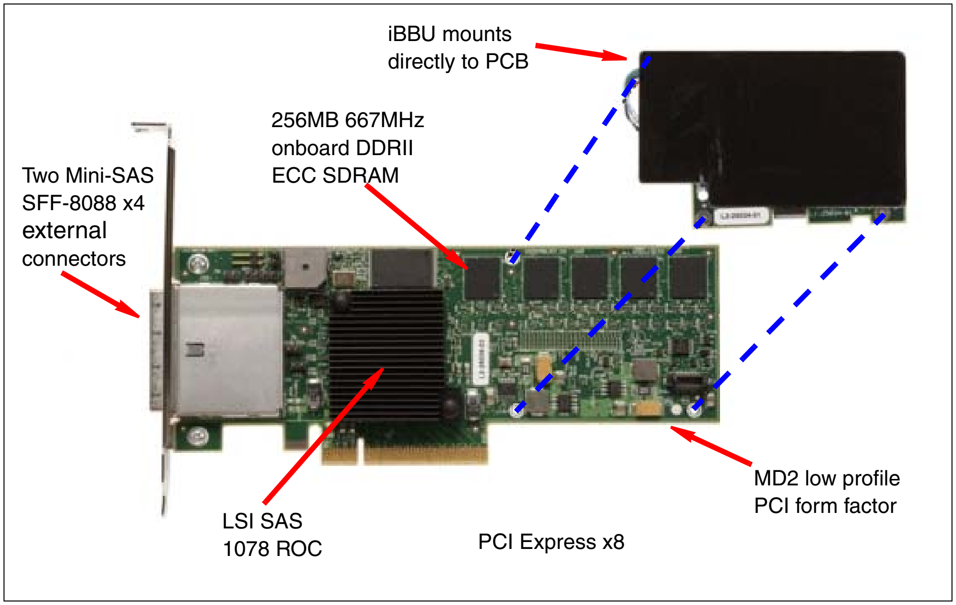

iBBU mounts directly to PCB
256MB 667MHz onboard DDRII ECC SDRAM
Two Mini-SAS SFF-8088 x4 external connectors
MD2 low profile PCI form factor
LSI SAS 1078 ROC
PCI Express x8

Figure 3-25 ServeRAID-MR10M

This controller can be ordered with the part number listed in Table 3-13:

Table 3-13 ServeRAID-MR10M part number

Part number	Description
43W4339[a]	IBM ServeRAID-MR10M SAS/SATA Controller

[a] Kit includes SeveRAID-MR10M, iBBU battery, brackets, and installation guide.

Intelligent backup battery unit (iBBU)

The option kit contains the battery package, which is the intelligent battery backup unit (iBBU). This backup battery must be mounted on the circuit.

An iBBU can charge the battery pack automatically and communicate battery status information such as voltage, temperature, and current, to the host computer system.

The backup battery protects cache up to 72 hours, depending on operating environment. IBM recommends that the battery be replaced annually. A damaged or no-recoverable battery can be replaced if you submit a service claim. Table 3-14 lists the replacement part numbers.

> **Important:** The battery in the iBBU must charge for at least six hours under normal operating conditions. To protect your data, the firmware changes the Write Policy to *write-through* until the battery unit is sufficiently charged. When the battery unit is charged, the RAID controller firmware changes the Write Policy to *write-back* to take advantage of the performance benefits of data caching.

Table 3-14 Customer replacement part numbers

Customer replaceable unit (CRU) part numbers	**Description**
43W4342	iBBU battery package
43W4341	ServeRAID MR10M adapter
43W4343	Carrier

Installation guidelines

Consider the following guidelines when using RAID or installing the ServeRAID-MR10M:

- The ServeRAID-MR10M can be used in any slot of a x3850 M2 and x3950 M2 complex.
- The ServeRAID-MR10M is not supported with the ServeRAID manager. Use the MegaRAID Storage Manager (MSM) instead.
- One or more arrays can be formed using up to nine cascaded EXP3000. An x4 SAS SFF8088 cable is required to attach the external storage subsystem to the ServeRAID-MR10M connectors.

You may have 32 ServeRAID adapters installed in a 4-node multinode complex, one MR10k and seven MR10M controllers in each node. However, at the time of writing, the MegaRAID Storage Manager and the WebBIOS in current ServeRAID firmware are limited to address 16 adapters only.

To configure all 32 adapters you may remove the first sixteen adapters, configure the remaining installed adapters, and then reinsert the adapters you removed and configure them. Or, you may use the LSI MegaCli command line utility, which does not have this limitation. This limitation should be fixed in future controller firmware and MSM.

See the following RETAIN® tips for further information:

- RETAIN tip H193590: MegaRAID Storage Manager cannot recognize greater than 16 controllers

  `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5076491`

- RETAIN tip H193591: WebBIOS only recognizes sixteen ServeRAID controllers

  `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5076492`

Installation instructions

To install the ServeRAID-MR10M SAS/SATA controller and the battery package:

1. Mount the iBBU package on the controller board.
2. Prepare the system: Unplug the power cords from the power supplies, disconnect the computer from the network.

   > **Important:** Another blue locator LED is on the rear side of the server. This LED indicates power in the system. It remains on (lit) for 20 seconds after you remove the power cables.

3. Remove the computer cover.
4. Open the blue adapter retention bracket as shown in Figure 3-26 on page 139.

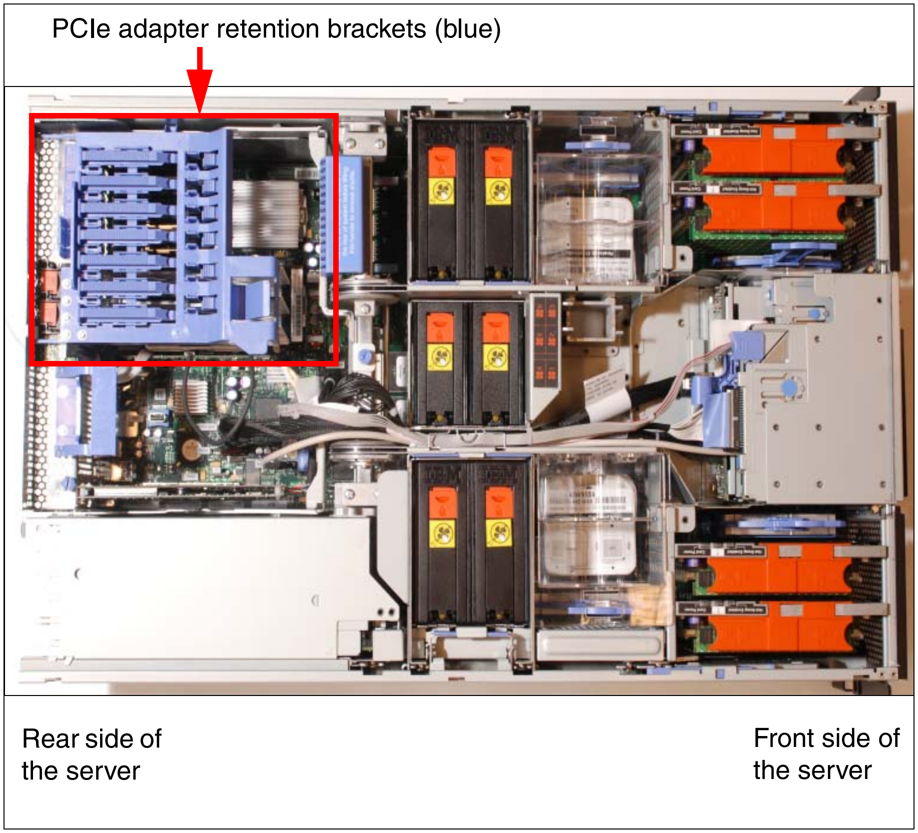


Figure 3-26 x3850 M2 and x3950 M2 - PCIe slot location

The PCIe slots are beyond the adapter-retention bracket on the right at the rear of the server shown in Figure 3-26 and Figure 3-27 on page 140.

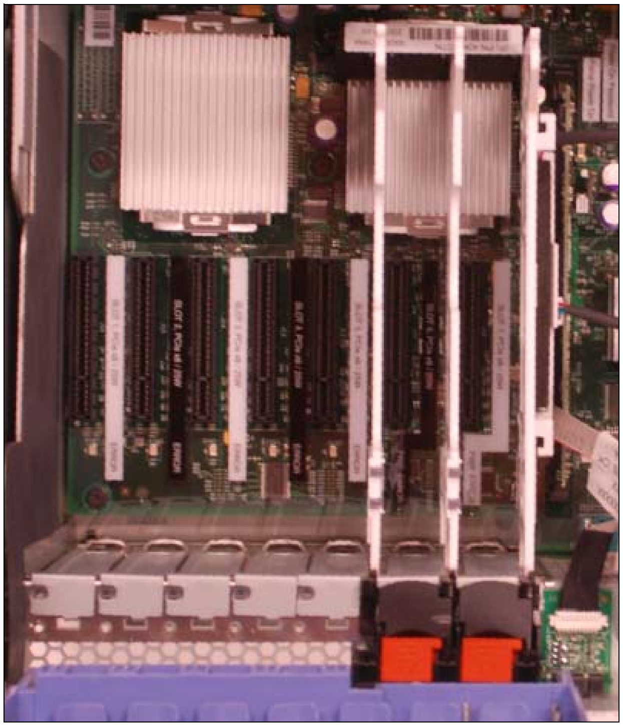


Figure 3-27 x3850 M2 and x3950 M2: PCIe slots

5. Remove the screw for the non hot-plugable slot, or push the orange adapter retention latch toward the rear of the server at the hot-plug, and open the tab. That is where you add a PCIe adapter. The power LED for slot 6 and slot 7 turn off.
6. Insert the controller in a PCIe slot as shown in Figure 3-28 on page 141

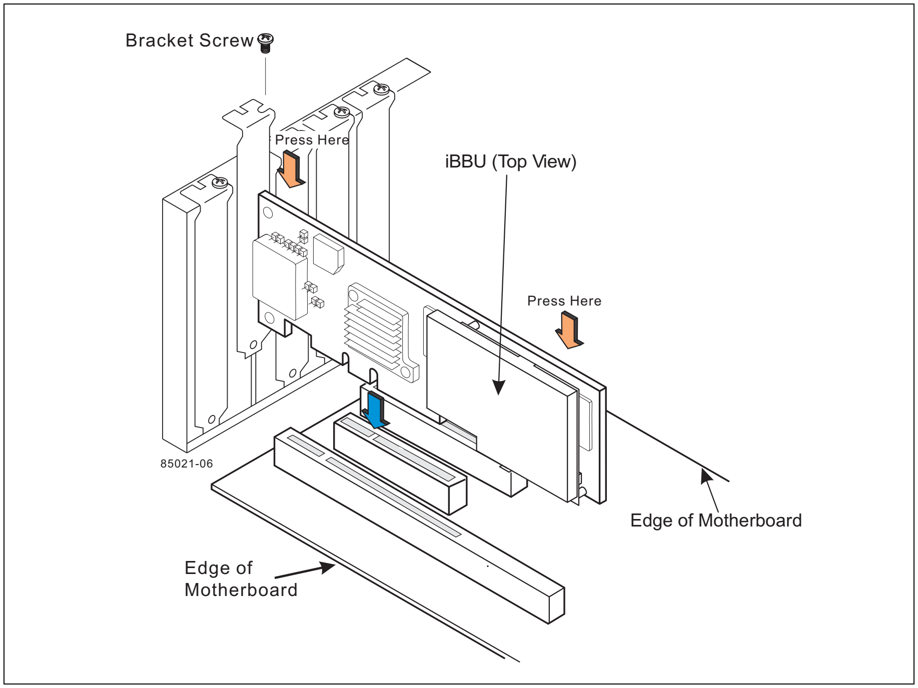


Figure 3-28 ServeRAID MR10M installation

7. Secure the controller to the computer chassis with the bracket screw. Close the adapter-retention bracket, and use the retention pin as shown in Figure 3-29 on page 142

> **Note:** Secure all adapters with the retention pin. Adapters can loosen during shipment or pushing the server out of the racks.

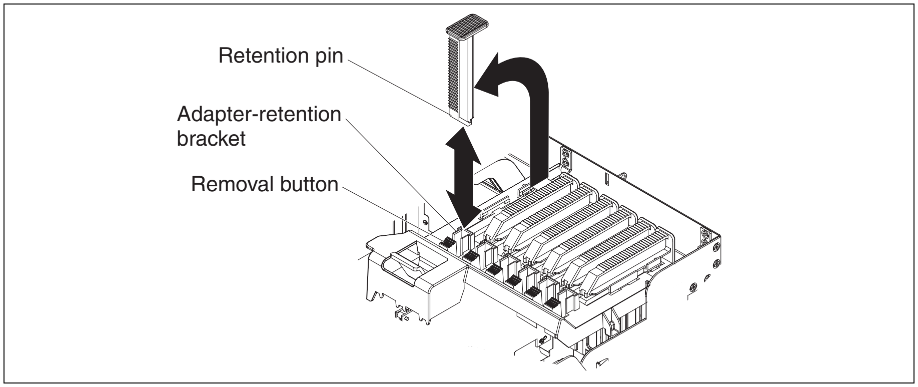


Figure 3-29 Mounting adapters by the retention pin

8. Close the cover and replug the power cables.
9. Power on the system and check whether the controller is initialized, as shown in the message in Example 3-8.

Example 3-8 ServeRAID-MR10M initialization in POST

```
LSI MegaRAID SAS-MFI BIOS Version NT16 (Build Nov 20, 2007)
Copyright(c) 2007 LSI Corporation
HA -0 (Bus 84 Dev 0) IBM ServeRAID-MR10M SAS/SATA Controller
FW package: 8.0.1-0029
```

10. Power off the system again. You can now install the disks in the external enclosures and cable with the adapter.

3.3.5 SAS expansion enclosure (unit)

The x3850 M2 and x3950 M2 SAS storage controllers can be extended to expand your disk space by the attachment of the EXP3000 external storage expansion unit. This expansion unit is a 2U device, which can hold up to twelve 3.5-inch SAS or SATA II hot-swap disk drives. The expansion unit is shown in Figure 3-30 on page 143.

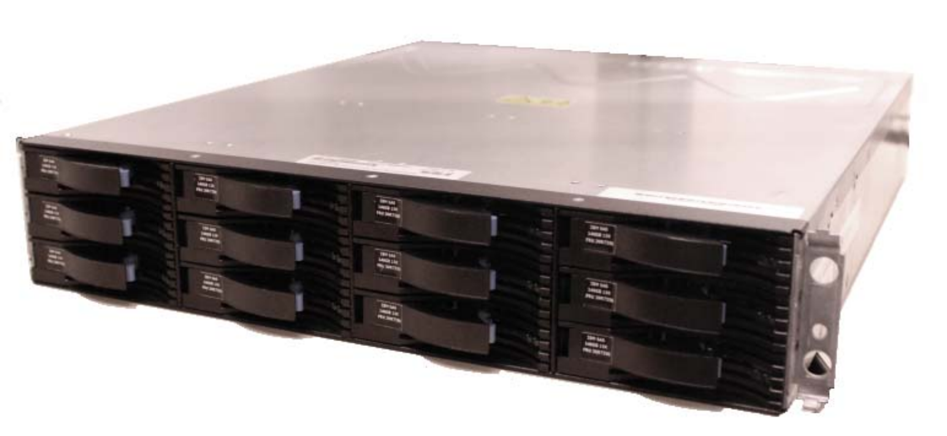


Figure 3-30 EXP3000 storage expansion unit

You can attach the EXP3000 storage expansion unit to the SAS controllers described in the following sections:

- 3.3.3, “ServeRAID-MR10k RAID controller” on page 128,
- 3.3.4, “ServeRAID-MR10M SAS/SATA II controller” on page 135

Table 3-15 lists the hard disk drive options.

Table 3-15 EXP3000: hard disk drive options

Part number	Description
39M4558	500 GB 3.5-inch hot-swap SATA II
43W7580	750 GB 3.5-inch hot-swap SATA dual port
43W7630	1 TB 3.5-inch hot-swap SATA dual port, 7200 RPM
40K1043	73 GB 3.5-inch hot-swap SAS, 15000 RPM
40K1044	146 GB 3.5-inch hot-swap SAS, 15000 RPM
43X0802	300 GB 3.5-inch hot-swap SAS, 15000 RPM

The EXP3000 machine type 1727-01X is available and shipped in the following configurations:

- Dual hot-swap redundant power supplies and two rack power cables
- No disk; add up to 12 SAS, SATA or SATA II 3.5-inch hard disk drives
- One ESM board (2 SAS ports); a second ESM is required for attachment to dual controller models of DS3200/DS3400

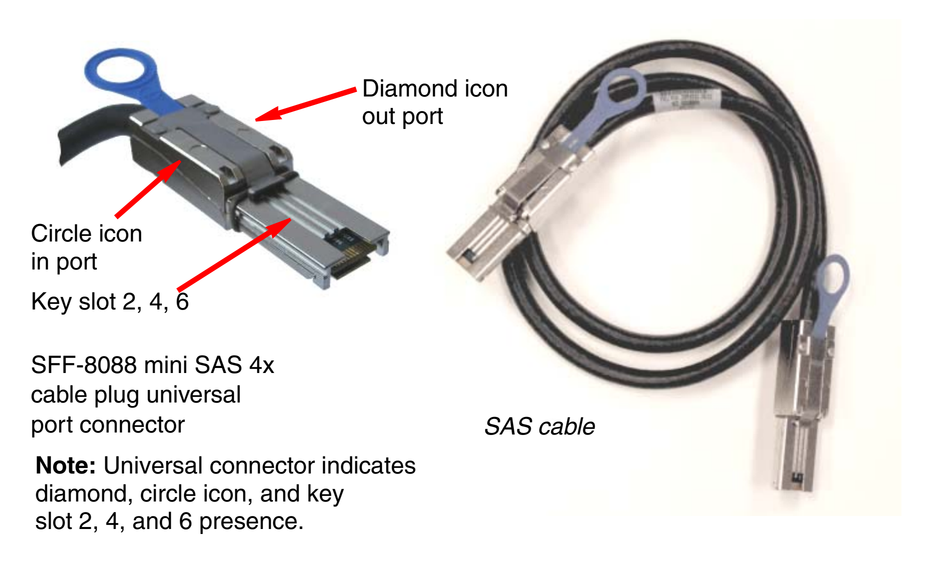
Table 3-16 lists part numbers for SAS cable, shown in Figure 3-31, and the EXP3000 ESM.

Table 3-16 EXP3000 Part numbers

Part number	Description
39R6531	IBM 3m SAS cable
39R6529	IBM 1m SAS cable
39R6515[a]	EXP3000 Environmental Service Module (ESM)

[a] Second ESM required for attachment to dual controller models of DS3200/DS3400

The cables connect to the SFF-8088 sockets on the ServeRAID-MR10M, or the external SAS port on the rear of the server to the ESM of the EXP3000. Chain multiple EXP3000 units through an external cable connection from this EXP3000 to up to nine EXP3000 enclosures. See Figure 3-31.



Note: Universal connector indicates diamond, circle icon, and key slot 2, 4, and 6 presence.

Figure 3-31 EXP3000: 1m SAS cable, part number 39R6529

Performance

Depending on your disk space requirement, type of application, and required performance, the amount of disks in a number of chained EXP3000 can result in higher performance, but does not increase linearly up to the maximum of nine EXP3000 enclosures with up to 108 disk drives.

It might be more sensible to chain, for example, two EXP3000 units to one controller and add another controller, which is linked to the second PCIe bridge chip. Figure 1-12 on page 28 shows the slot assignment to the PCIe bridge chips.

Installation guidelines

The EXP3000 is ready to use after minimal installation and configuration steps. You must have the required rack space of 2U for each expansion unit.

To install and configure the EXP3000:

1. Insert the EXP3000 in your rack.

   > **Tip:** You can reduce the weight for easier installation in your rack, while you remove the power supplies. Press the orange release tab to the right, just enough to release the handle as you rotate the handle downward.

2. Install the hard disk drives.

   > **Note:** Install a minimum of four hard disk drives for each power supply to operate in a redundant mode.

3. Start with the cabling of your expansion unit by connecting the power cables:
   a. Attach one end of your 3 m SAS cable to the host port, shown in Figure 3-32 on page 146.
   b. Connect the other end to the closest EXP3000 IN-port.
   c. Attach one end of the 1 m SAS cable to the EXP3000 OUT-port and the other end to the IN-port of the next expansion. Repeat this step if you want to attach multiple expansion units to this chain.

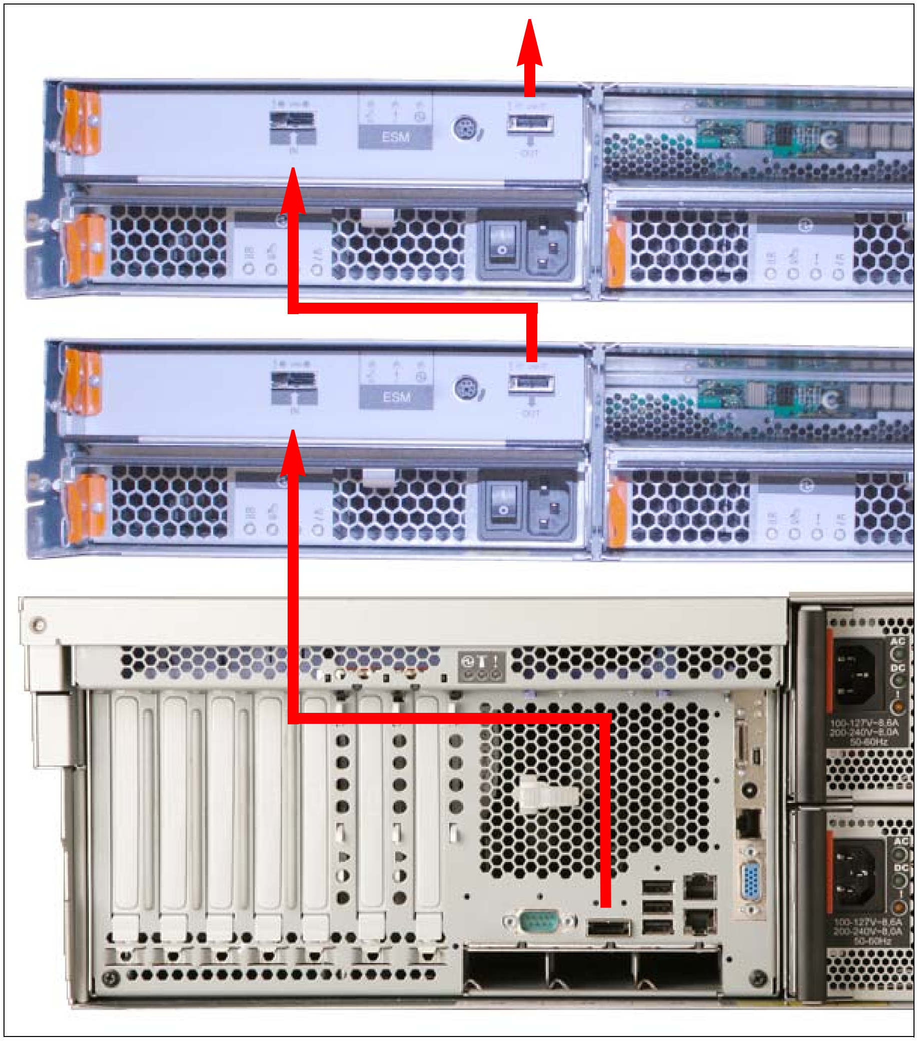


Figure 3-32   EXP3000: rear view cabling

4. Turn on the EXP3000 expansion units before or at the same time as you turn on the device that contains the RAID controller.

   Figure 3-33 on page 147 shows the components of the ESM.

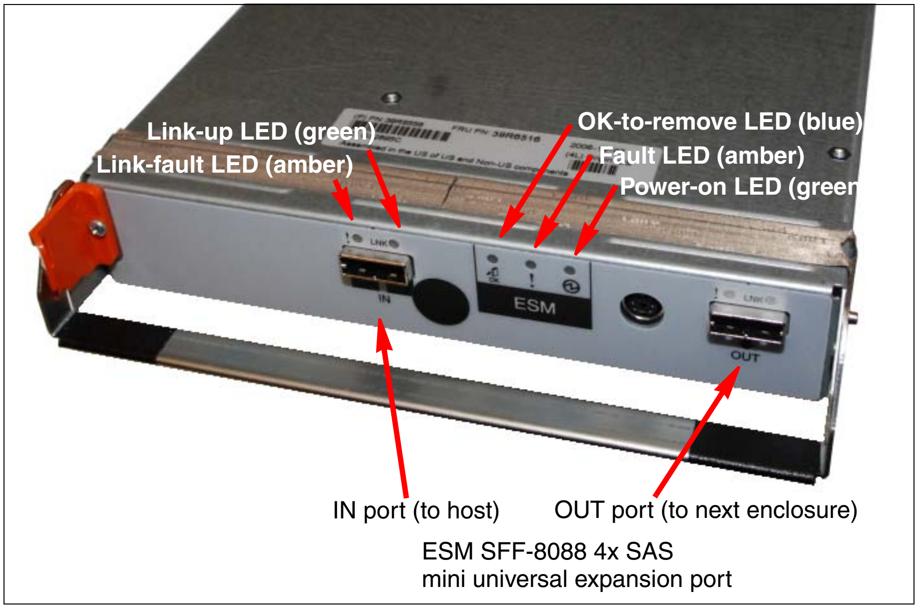


Figure 3-33 EXP3000: ESM

5. Check the states of the controller LEDs.

   – After you power on the expansion unit by using the power switch on the rear side of each power supply, the Power-on LED (green) lights up.

   – Check that none of the fault LEDs (amber) are on.

   – Check the Link-up LEDs (green) are on.

6. If you observe any abnormal behavior, review the EXP3000 user guide that is in your package.

> **Note:** The IBM Support for System x Web page contains useful technical documentation, user guides, and so on. It is located at:
>
> `http://www.ibm.com/systems/support/x`
>
> Enter the system **Type** and **Model** 172701X in the **Quick path** field to link to the information for this enclosure. We recommend you also check regularly for new codes and tips. Use the **Download** and **Troubleshoot** links.

Your enclosures are now prepared for configuration RAID volumes, which is described in the next section.

3.3.6 Updating the SAS storage controllers

Before you create RAID volumes, we recommend that you update the SAS controller BIOS and firmware, and the disk subsystems. The following Web addresses are for the latest updates:

- LSI1078 BIOS and firmware:

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073134

  See also "Updating the onboard SAS LSI1078" on page 148.
- IBM ServeRAID-MR10k SAS controller firmware (Windows)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073139

  See also "Updating the onboard ServeRAID-MR10k/MR10M controller" on page 150.
- IBM ServeRAID-MR10M SAS controller firmware (Windows)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073389

  See also "Updating the onboard ServeRAID-MR10k/MR10M controller" on page 150.
- IBM SAS hard drive update program

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-62832

  See also "Updating the SAS hard disk drives" on page 152.
- EXP3000 ESM update

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073877

  See also "Updating the EXP3000" on page 153.

Updating the onboard SAS LSI1078

The program updates all SAS LSI1078 controllers found in the system. This allows you to update all referred controllers in a multinode configuration, without the requirement to boot each particular node in stand-alone mode.

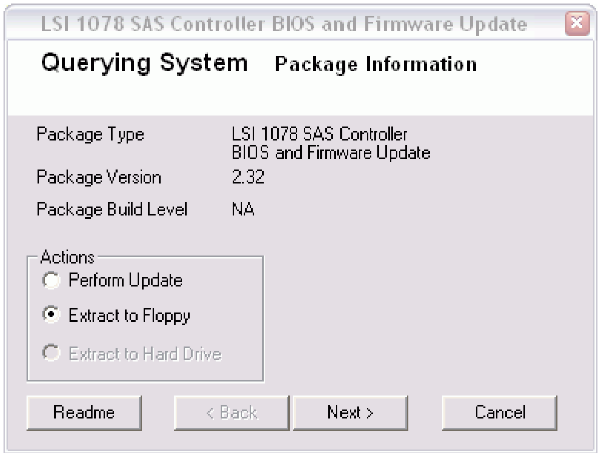
We used the following steps to update the controller:

1. Download the code release from the following Web address.

   http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073134

   We used the Microsoft Windows package and ran it on a Windows workstation to create bootable diskettes. With these diskettes, you can boot the server by an attached USB floppy drive or mount it in the Remote Supervisor Adapter II remote drive in the Web interface as described in 6.2.4, "Remote console and media" on page 324.

2. Start the executable package and create two diskettes (the first is bootable) by selecting of **Extract to Floppy** as shown in Figure 3-34.



Package Type LSI 1078 SAS Controller BIOS and Firmware Update

Package Version 2.32

Package Build Level NA

Actions

○ Perform Update

● Extract to Floppy

○ Extract to Hard Drive

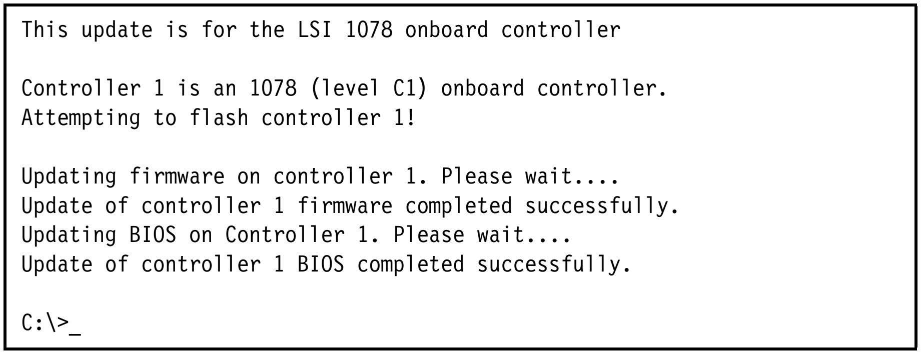
Figure 3-34 LSI1078 code package extraction

3. Boot the server with disk 1 and insert disk 2 when prompted. The utility attempts to flash the onboard LSI 1078 controllers in you system. You might see warning messages stating that this flash is not compatible with all controllers, as shown in Figure 3-35. These are not error messages.

```
****************************************
*                                      *
*    SAS Firmware & BIOS Flash Disk    *
*                                      *
****************************************
NOTE: This utility will scan all LSI 1064, 1068, and 1078 based
      controllers in the system. Only the controllers which match
the update type will be flashed.
.
Do you wan to continue[Y,N]?_
```

Figure 3-35 LSI1078: update process

4. Press **Y** to start the flash process. See Figure 3-35.



```
This update is for the LSI 1078 onboard controller

Controller 1 is an 1078 (level C1) onboard controller.
Attempting to flash controller 1!

Updating firmware on controller 1. Please wait....
Update of controller 1 firmware completed successfully.
Updating BIOS on Controller 1. Please wait....
Update of controller 1 BIOS completed successfully.

C:\>_
```

Figure 3-36 LSI1078: update process

5. After the update is completed, reboot your server.
6. If all updates are performed, go to section 3.4, “Configuring RAID volumes” on page 154.

Updating the onboard ServeRAID-MR10k/MR10M controller

We used the steps in this section o update the controllers.

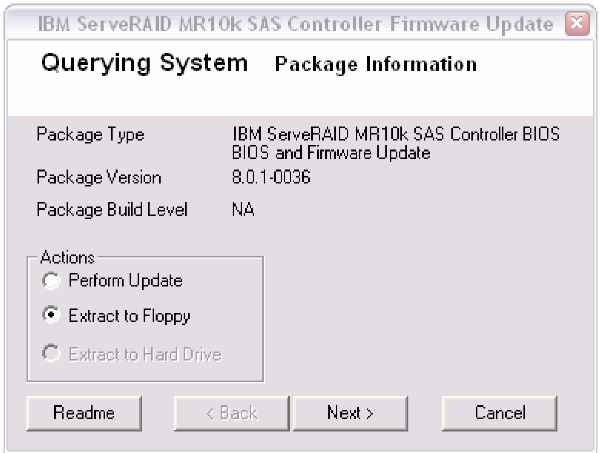
> **Note:** The program will update all ServeRAID-MR10k/MR10M LSI1078 controllers found in the system, depending of the package you downloaded. This allows you to update all referred controllers in a multinode configuration, without the requirement to boot each particular node in stand-alone mode.

1. Extract the contents of the package to floppy disks and then flash the ServeRAID controller.
2. Download the most recent firmware:
   - IBM ServeRAID-MR10k firmware update (Windows)

     http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073139
   - IBM ServeRAID-MR10M firmware update (Windows)

     http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073389

   We used the Microsoft Windows package and ran it on a Windows workstation to create bootable diskettes. With these diskettes, you can boot the server by an attached USB floppy drive or mount it in the Remote Supervisor Adapter II remote drive in the Web interface as described in 6.2.4, “Remote console and media” on page 324.
3. Start the executable package and create three diskettes (the first is bootable) by selecting **Extract to Floppy** (in Figure 3-37 on page 151).



IBM ServeRAID MR10k SAS Controller Firmware Update

Querying System   Package Information

Package Type: IBM ServeRAID MR10k SAS Controller BIOS BIOS and Firmware Update

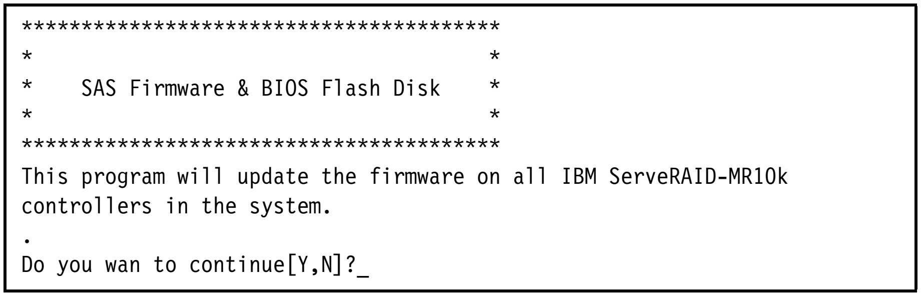
Package Version: 8.0.1-0036

Package Build Level: NA

Actions
- ○ Perform Update
- ● Extract to Floppy
- ○ Extract to Hard Drive

Figure 3-37 ServeRAID-MR10k code package extraction

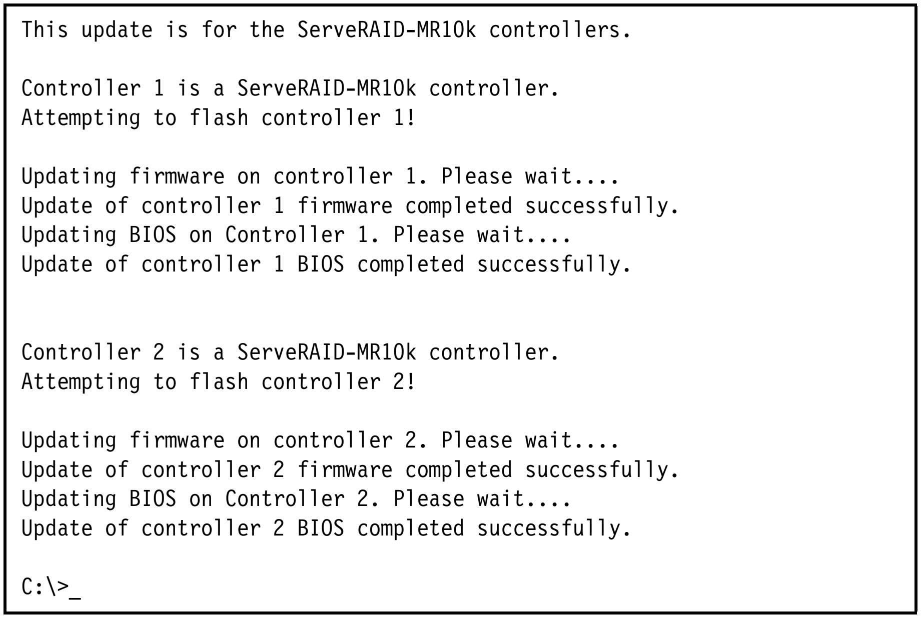
4. Boot the server with disks 1, 2, and 3 as prompted. It attempts to flash all ServeRAID-MR10k or ServeRAID-MR10M controllers in the system. You might see warning messages stating that this flash is not compatible with all controllers, as shown in Figure 3-38. These are not error messages.



```
****************************************
*                                      *
*    SAS Firmware & BIOS Flash Disk    *
*                                      *
****************************************
This program will update the firmware on all IBM ServeRAID-MR10k
controllers in the system.
.
Do you wan to continue[Y,N]?_
```

Figure 3-38 ServeRAID-MR10k: update process

5. Press **Y** to start the flash process. The update starts, as shown in Figure 3-39 on page 152.



```
This update is for the ServeRAID-MR10k controllers.

Controller 1 is a ServeRAID-MR10k controller.
Attempting to flash controller 1!

Updating firmware on controller 1. Please wait....
Update of controller 1 firmware completed successfully.
Updating BIOS on Controller 1. Please wait....
Update of controller 1 BIOS completed successfully.


Controller 2 is a ServeRAID-MR10k controller.
Attempting to flash controller 2!

Updating firmware on controller 2. Please wait....
Update of controller 2 firmware completed successfully.
Updating BIOS on Controller 2. Please wait....
Update of controller 2 BIOS completed successfully.

C:\>_
```

Figure 3-39 ServeRAID-MR10M: update process

6. After the update completes, reboot your server.
7. If all updates are performed, go to 3.4, “Configuring RAID volumes” on page 154.

Updating the SAS hard disk drives

We used the steps in this section to update the controllers.

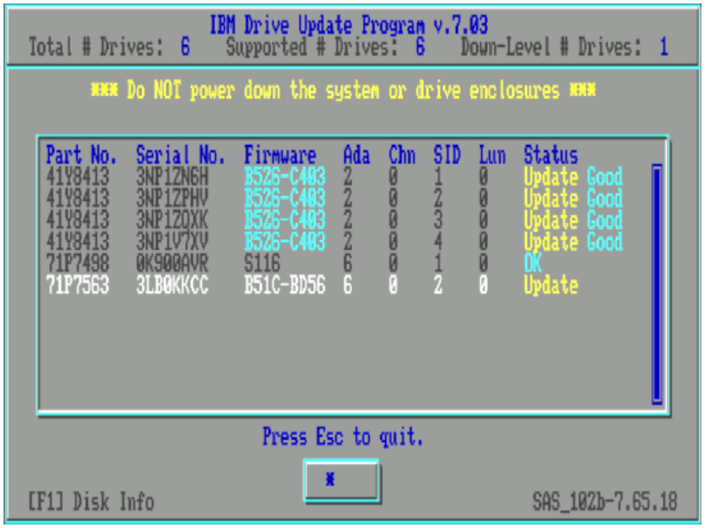
To update the SAS hard disk drives internal in the media hood assembly and external in the EXP3000 installed:

1. Download the bootable CD from the following Web address:

   http://www.ibm.com/support/docview.wss?uid=psg1MIGR-62832

2. Burn a CD or mount the ISO image in the Remote Supervisor Adapter II remote drive of the Web interface as described in 6.2.4, “Remote console and media” on page 324.
3. Boot your server from this CD image.
4. Update the hard drive disks by clicking the **Update** button.

   The update program window is shown in Figure 3-40 on page 153.



```
                    IBM Drive Update Program v.7.03
 Total # Drives:  6    Supported # Drives:  6    Down-Level # Drives:  1

         *** Do NOT power down the system or drive enclosures ***

   Part No.  Serial No.  Firmware   Ada  Chn  SID  Lun  Status
   41Y8413   3NP1ZN6H    B526-C403  2    0    1    0    Update Good
   41Y8413   3NP1ZPHV    B526-C403  2    0    2    0    Update Good
   41Y8413   3NP1ZQXK    B526-C403  2    0    3    0    Update Good
   41Y8413   3NP1V7XV    B526-C403  2    0    4    0    Update Good
   71P7498   0K900AVR    S116       6    0    1    0    OK
   71P7563   3LB0KKCC    B51C-BD56  6    0    2    0    Update

                           Press Esc to quit.
                               [ * ]
 [F1] Disk Info                                          SAS_102b-7.65.18
```

Figure 3-40 SAS HDD update

> **Note:** Do *not* power down the system or drive enclosures.

5. After the update is completed, reboot your server.
6. If all updates are complete, go to section 3.4, "Configuring RAID volumes" on page 154.

Updating the EXP3000

We used the steps in this section to update the ESM in the attached EXP3000.

We recommend also updating the EXP3000. At the time of writing, the latest update contained a critical fix for the ESM, which fixed the problem of an unexpected reboot of the EXP3000.

To update the EXP3000:

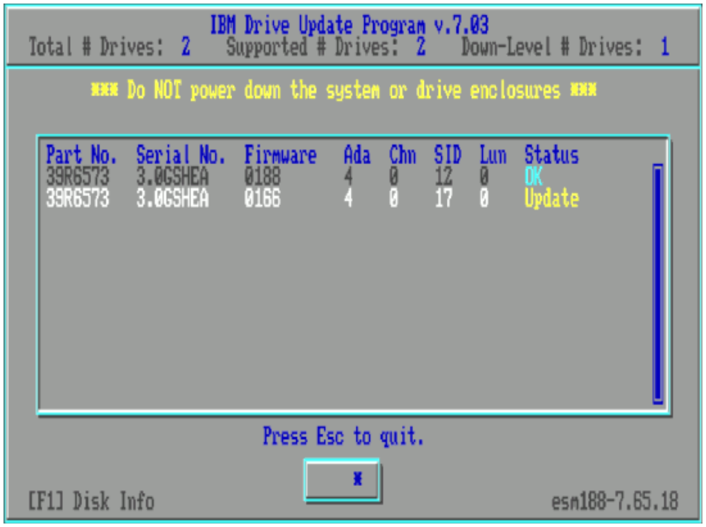
1. Download the bootable CD image from the following Web address:

   `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073877`

2. Burn a CD or mount the ISO image in the Remote Supervisor Adapter II remote drive of the Web interface as described in 6.2.4, "Remote console and media" on page 324.
3. Boot your server from this CD image.

4. Update the ESM modules by clicking the **Update** button.

   The update program window is shown in Figure 3-41.



```
                IBM Drive Update Program v.7.03
 Total # Drives:  2   Supported # Drives:  2   Down-Level # Drives:  1

        *** Do NOT power down the system or drive enclosures ***

   Part No.  Serial No.  Firmware  Ada  Chn  SID  Lun  Status
   39R6573   3.0GSHEA    0188      4    0    12   0    OK
   39R6573   3.0GSHEA    0166      4    0    17   0    Update

                      Press Esc to quit.

 [F1] Disk Info                                   esm188-7.65.18
```

Figure 3-41   EXP3000 ESM firmware update

Note: Do *not* power down the system or drive enclosures.

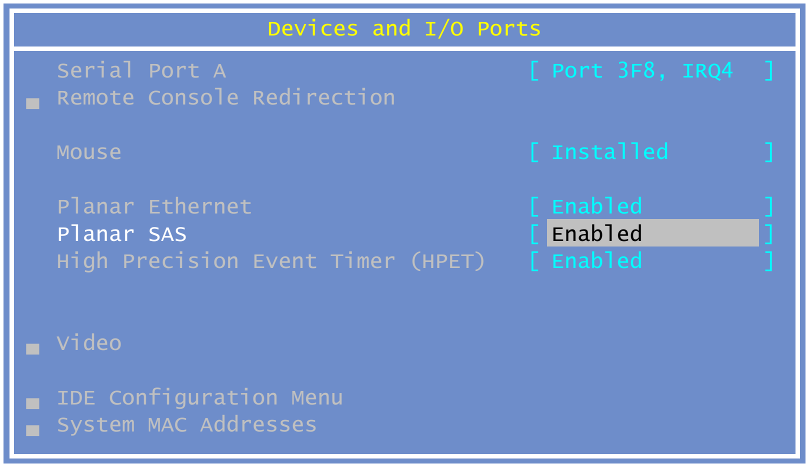
5. After the update is completed, reboot your server.
6. If all updates are performed, go to section 3.4, “Configuring RAID volumes” on page 154.

3.4 Configuring RAID volumes

In this section, we describe the various ways to configure your RAID volumes. We provide the steps and rules for the controllers and describe the differences.

3.4.1 Starting the LSI1078 controller BIOS

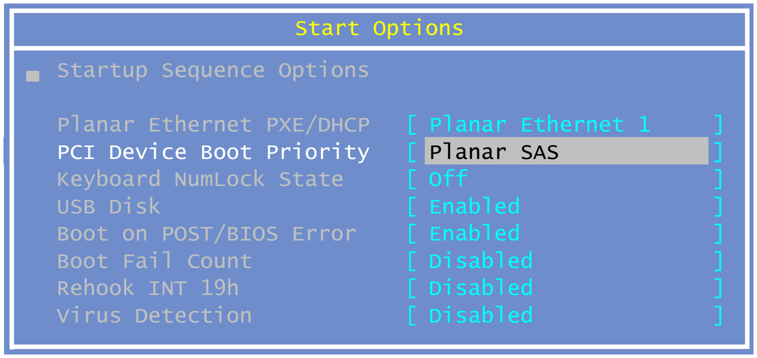
The onboard LSI controller is initialized in POST, if it is enabled in the BIOS of the server. Figure 3-42 on page 155 and Figure 3-43 on page 155 shows the BIOS settings you should know about to gain access to the SAS storage subsystem.



```
                       Devices and I/O Ports

   Serial Port A                              [ Port 3F8, IRQ4   ]
■  Remote Console Redirection

   Mouse                                      [ Installed        ]

   Planar Ethernet                            [ Enabled          ]
   Planar SAS                                 [ Enabled          ]
   High Precision Event Timer (HPET)          [ Enabled          ]


■  Video

■  IDE Configuration Menu
■  System MAC Addresses
```

Figure 3-42 BIOS: system devices and ports

Decide which controller you want to boot the server from and select it in the PCI Device Boot Priority field in **Start Options** menu, Figure 3-43.

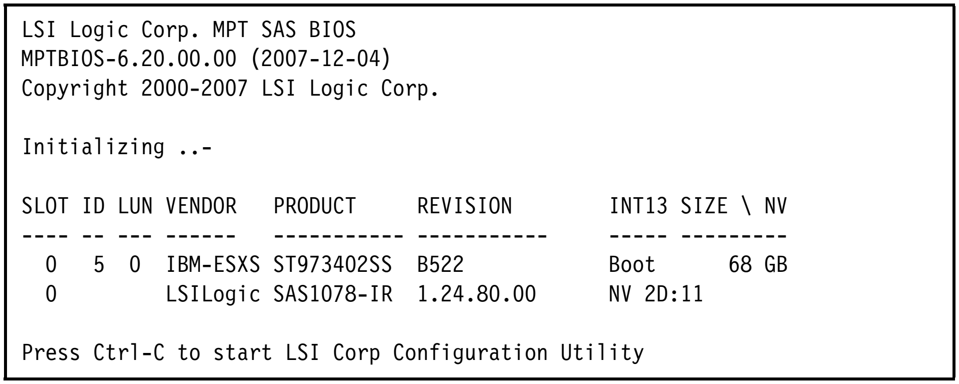


```
                      Start Options
■  Startup Sequence Options

   Planar Ethernet PXE/DHCP   [ Planar Ethernet 1    ]
   PCI Device Boot Priority   [ Planar SAS           ]
   Keyboard NumLock State     [ Off                  ]
   USB Disk                   [ Enabled              ]
   Boot on POST/BIOS Error    [ Enabled              ]
   Boot Fail Count            [ Disabled             ]
   Rehook INT 19h             [ Disabled             ]
   Virus Detection            [ Disabled             ]
```

Figure 3-43 BIOS: Start Options

Note: If you have to change BIOS settings on secondary nodes in a multinode complex, you must first break the partition and boot the nodes into standalone mode. You can then change the BIOS settings as required.

Power on your server. Watch the window in POST. Press `Ctrl+C` while the messages in Figure 3-44 on page 156 are shown.

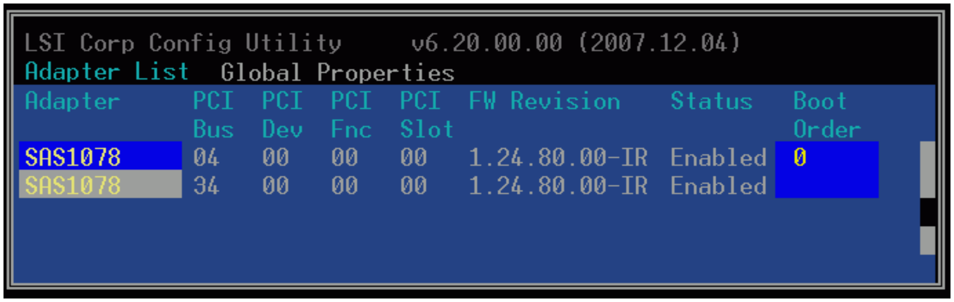


```
LSI Logic Corp. MPT SAS BIOS
MPTBIOS-6.20.00.00 (2007-12-04)
Copyright 2000-2007 LSI Logic Corp.

Initializing ..-

SLOT ID LUN VENDOR   PRODUCT     REVISION        INT13 SIZE \ NV
---- -- --- ------   ----------- -----------     ----- ---------
  0   5  0  IBM-ESXS ST973402SS  B522            Boot      68 GB
  0         LSILogic SAS1078-IR  1.24.80.00      NV 2D:11

Press Ctrl-C to start LSI Corp Configuration Utility
```

Figure 3-44   LSI 1078 SAS controller initialization in POST

The LSI 1078 onboard SAS controller BIOS starts after the POST finishes.

Select the controller to start the LSI controller configuration utility. Figure 3-45 shows a two-node multinode system; the first controller is the primary node, the second controller is the onboard SAS controller of the secondary node.



```
LSI Corp Config Utility     v6.20.00.00 (2007.12.04)
Adapter List   Global Properties
Adapter        PCI  PCI  PCI  PCI  FW Revision    Status    Boot
               Bus  Dev  Fnc  Slot                          Order
SAS1078        04   00   00   00   1.24.80.00-IR  Enabled   0
SAS1078        34   00   00   00   1.24.80.00-IR  Enabled
```

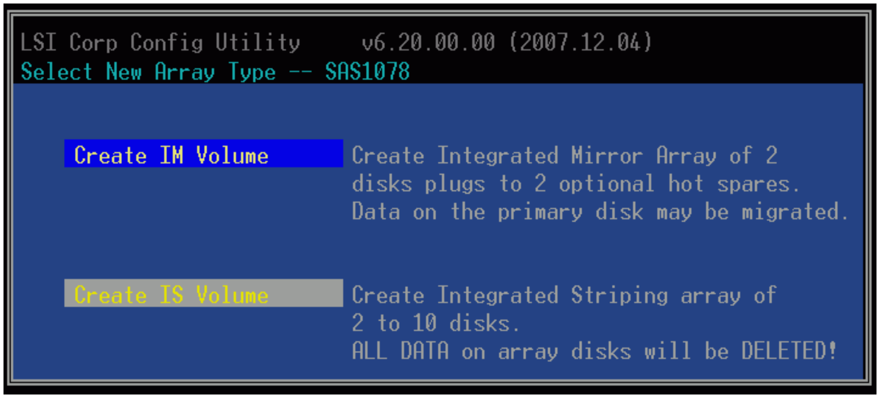
Figure 3-45   LSI1078: adapter selection menu

The LSI controller configuration utility has three submenus:

- RAID Properties: Use to create and manage your RAID volumes.
- SAS Topology: Shows a tree of all available devices.
- Advanced Adapter Properties: Use advanced settings for the controller and devices.

Creating volumes

To create a volume, select the LSI controller you want to configure by highlighting it, as shown in Figure 3-45, and then pressing **Enter**. From the submenu that appears, select the **RAID Properties**. The submenu, shown in Figure 3-46, lists the RAID types.

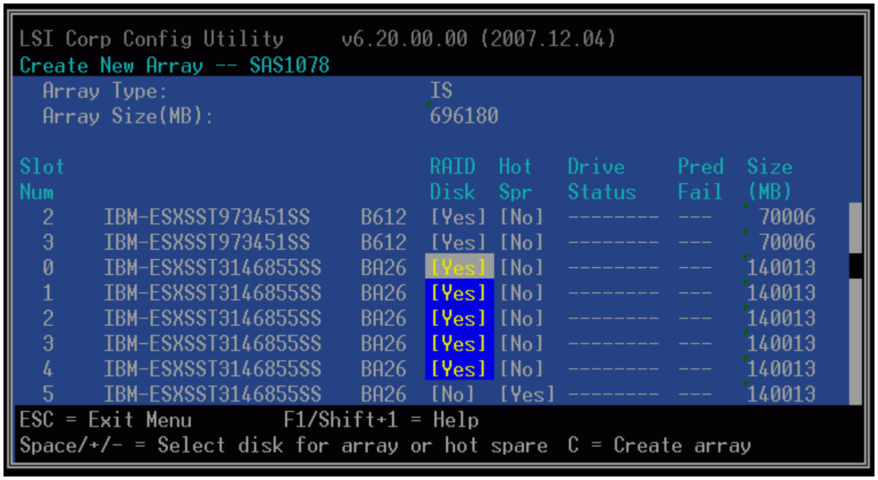


```
LSI Corp Config Utility     v6.20.00.00 (2007.12.04)
Select New Array Type -- SAS1078


     Create IM Volume           Create Integrated Mirror Array of 2
                                disks plugs to 2 optional hot spares.
                                Data on the primary disk may be migrated.


     Create IS Volume           Create Integrated Striping array of
                                2 to 10 disks.
                                ALL DATA on array disks will be DELETED!
```

Figure 3-46 LSI1078: Array type selection

You can create two volumes of the same type or a mixture of Integrated Mirroring (IM) or Integrated Striping (IS). Select the RAID level IM or IS.

In the RAID creation menu, Figure 3-47, select the hard disk drives you want to add, for example, to a RAID-0 (IS). You can add up to 10 hard disk drives to one volume. Press **C** to create this array.



```
LSI Corp Config Utility     v6.20.00.00 (2007.12.04)
Create New Array -- SAS1078
  Array Type:                          IS
  Array Size(MB):                      696180

Slot                                   RAID  Hot   Drive    Pred  Size
Num                                    Disk  Spr   Status   Fail  (MB)
  2     IBM-ESXSST973451SS     B612   [Yes] [No]  -------- ---    70006
  3     IBM-ESXSST973451SS     B612   [Yes] [No]  -------- ---    70006
  0     IBM-ESXSST3146855SS    BA26   [Yes] [No]  -------- ---   140013
  1     IBM-ESXSST3146855SS    BA26   [Yes] [No]  -------- ---   140013
  2     IBM-ESXSST3146855SS    BA26   [Yes] [No]  -------- ---   140013
  3     IBM-ESXSST3146855SS    BA26   [Yes] [No]  -------- ---   140013
  4     IBM-ESXSST3146855SS    BA26   [Yes] [No]  -------- ---   140013
  5     IBM-ESXSST3146855SS    BA26   [No]  [Yes] -------- ---   140013
ESC = Exit Menu          F1/Shift+1 = Help
Space/+/- = Select disk for array or hot spare  C = Create array
```

Figure 3-47 LSI1078: array creation

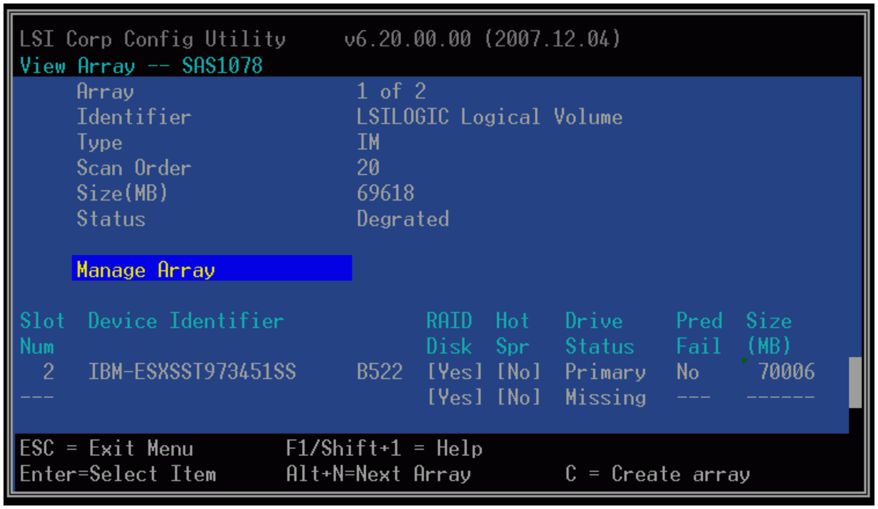
Figure 3-47 shows a previously created volume in RAID-1 (IM) and one assigned hard disk drive as a hot-spare disk. see in Figure 3-48 on page 158.

```
Slot                                  RAID  Hot    Drive     Pred Size
Num                                   Disk  Spr    Status    Fail (MB)
  2    IBM-ESXST973451SS     B612  [YES] [No]   -------- ---    70006
  3    IBM-ESXST973451SS     B612  [YES] [No]   -------- ---    70006
...
  2    IBM-ESXST3146855SS    BA26  [No]  [Yes] -------- ---    70006
```

Figure 3-48 RAID-0 (Integrated Mirroring), hot spare

You can assign a maximum of 12 hard disk drives to two volumes plus two hard disk drives to hot spares.

After you create the volumes, select the RAID Properties menu again. You can check the state of the created volumes. Figure 3-49 shows a missing disk in the mirror, which will be synchronized after a new disk is inserted.

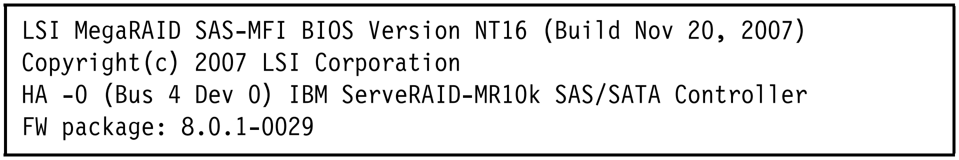


```
LSI Corp Config Utility      v6.20.00.00 (2007.12.04)
View Array -- SAS1078
     Array                        1 of 2
     Identifier                   LSILOGIC Logical Volume
     Type                         IM
     Scan Order                   20
     Size(MB)                     69618
     Status                       Degrated

     Manage Array

Slot  Device Identifier           RAID  Hot   Drive     Pred  Size
Num                               Disk  Spr   Status    Fail  (MB)
  2   IBM-ESXSST973451SS    B522  [Yes] [No]  Primary   No     70006
---                               [Yes] [No]  Missing   ---   ------

ESC = Exit Menu          F1/Shift+1 = Help
Enter=Select Item        Alt+N=Next Array        C = Create array
```

Figure 3-49 LSI1078: Manage Array

3.4.2 Starting the ServeRAID-MR10k controller WebBIOS

After correct installation of the ServeRAID card, the card is initialized in POST, as shown in Figure 3-50 on page 159.



```
LSI MegaRAID SAS-MFI BIOS Version NT16 (Build Nov 20, 2007)
Copyright(c) 2007 LSI Corporation
HA -0 (Bus 4 Dev 0) IBM ServeRAID-MR10k SAS/SATA Controller
FW package: 8.0.1-0029
```

Figure 3-50 ServeRAID-MR10k initialization in POST

The following message is prompting you to start the MR10k WebBIOS:

```
Press <CTRL><H> for WebBIOS or press <CTRL><Y> for Preboot CLI
```

After you press `Ctrl+H` keys the WebBIOS starts after the POST finishes:

```
WebBIOS will be executed after POST completes
```

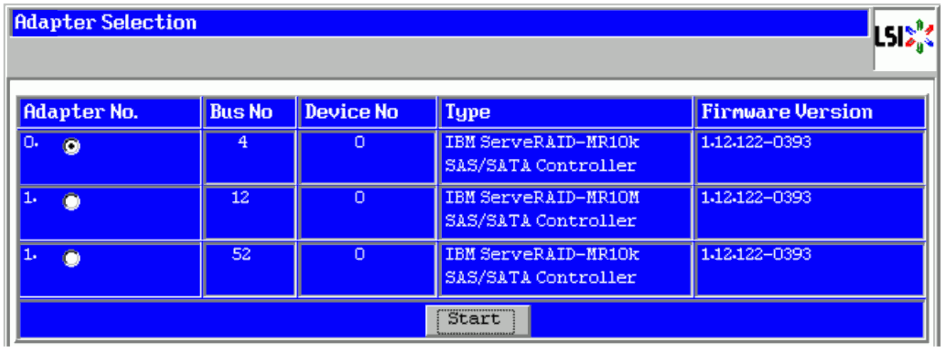
You can migrate RAID-0 and RAID-1 volumes that you created using the onboard LSI1078 SAS controller to MegaRAID (MR) mode by installing the ServeRAID-MR10k. The created volumes are automatically imported after the first installation of this controller.

> **Note:** You cannot reverse this step to go back to using the onboard SAS controller if you remove the ServeRAID-MR10k again.

3.4.3 Working with LSI MegaRAID controller WebBIOS

The WebBIOS starts up with the initial window, which indicates all available LSI SAS controllers that are installed in either the x3850 M2, x3950 M2, or x3950 M2 multinode complex, and initialized in the POST correctly.

Figure 3-51 shows both installed ServeRAID MR10k controllers, within a two-node x3950 M2 system.



Adapter Selection

Adapter No.	Bus No	Device No	Type	Firmware Version
0. ●	4	0	IBM ServeRAID-MR10k SAS/SATA Controller	1.12.122-0393
1. ○	12	0	IBM ServeRAID-MR10M SAS/SATA Controller	1.12.122-0393
1. ○	52	0	IBM ServeRAID-MR10k SAS/SATA Controller	1.12.122-0393

Figure 3-51 Adapter Selection menu: initialized controllers in POST

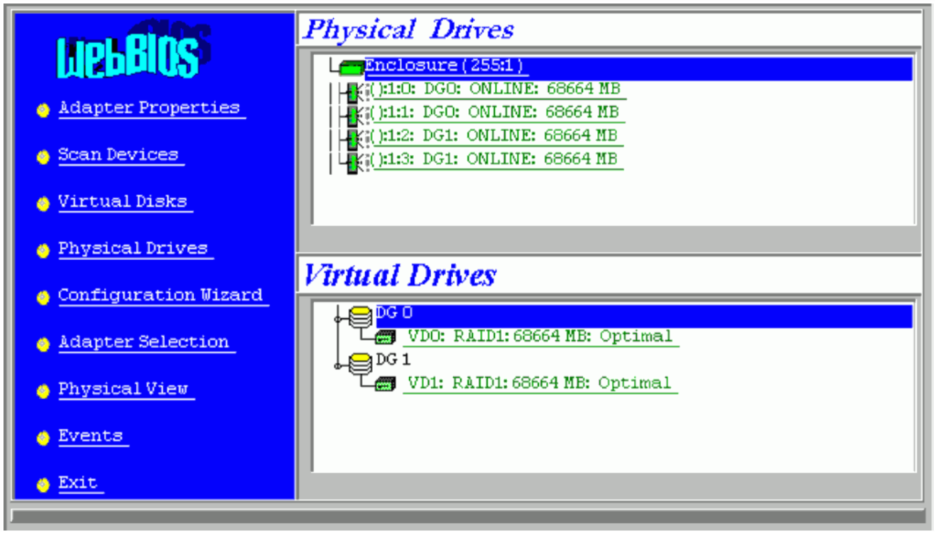
According the PCIe bus scan order (see 3.5, “PCI Express options” on page 188 for details), adapter number 0 is installed in the primary node, and adapter number 1 in the secondary node.

To choose the adapter you want to configure, use either the mouse or keyboard (`Tab` key) to select the adapter, and then click **Start** or press `Enter` to begin the process. The WebBIOS main window opens.

Note: If a mouse is not available, you can always operate with the `Tab` key to switch between options and buttons and spacebar or `Enter` to select any.

WebBIOS main window

The WebBIOS has been started. As shown in Figure 3-52, the window is separated into three areas: the Controller Function menu, Physical Drives panel, and Virtual Drives panel.

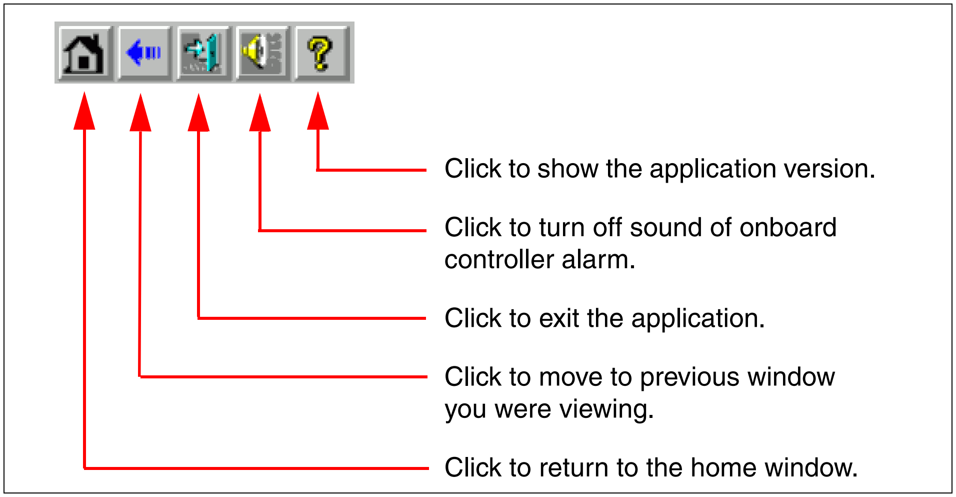


```
Enclosure(255:1)
( ):1:0: DG0: ONLINE: 68664 MB
( ):1:1: DG0: ONLINE: 68664 MB
( ):1:2: DG1: ONLINE: 68664 MB
( ):1:3: DG1: ONLINE: 68664 MB
```

```
DG 0
  VD0: RAID1: 68664 MB: Optimal
DG 1
  VD1: RAID1: 68664 MB: Optimal
```

Figure 3-52 WebBIOS main window

The window also includes icons, as shown in Figure 3-53 on page 161.


Click to show the application version.

Click to turn off sound of onboard controller alarm.

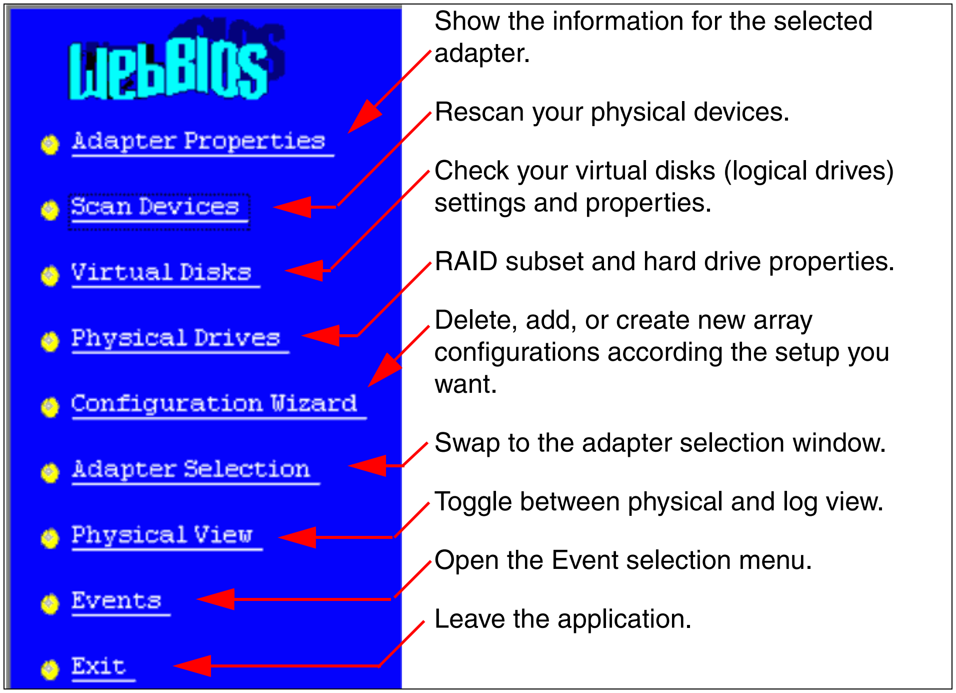
Click to exit the application.

Click to move to previous window you were viewing.

Click to return to the home window.

Figure 3-53   WebBIOS control icons

Options in the Controller Function menu are explained in Figure 3-54. Details about several of these options are provided in sections that follow.



Show the information for the selected adapter.

Rescan your physical devices.

Check your virtual disks (logical drives) settings and properties.

RAID subset and hard drive properties.

Delete, add, or create new array configurations according the setup you want.

Swap to the adapter selection window.

Toggle between physical and log view.

Open the Event selection menu.

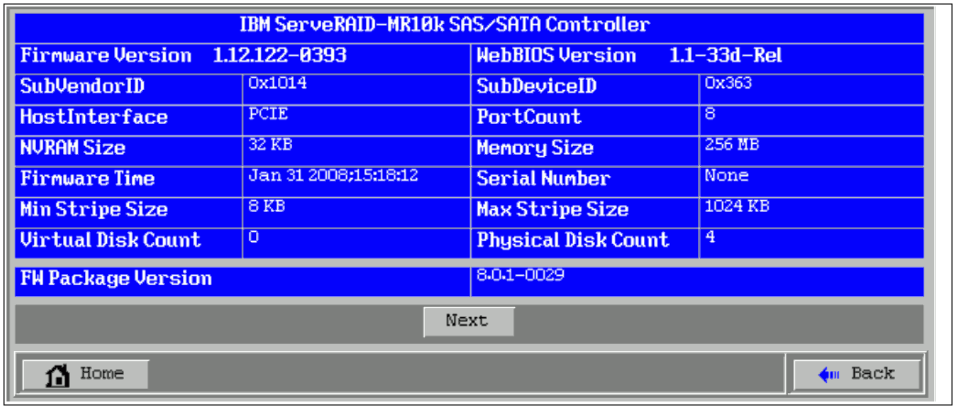
Leave the application.

Figure 3-54   WebBIOS Controller Function menu

Adapter Properties option

To get the details for this adapter, select the **Adapter Properties** option, shown in Figure 3-54 on page 161. This provides information about the adapter you selected in the Adapter Selection menu (Figure 3-51 on page 159).

The first page, in Figure 3-55, shows the physical specification of the adapter and version for BIOS, firmware package, and summarized information for physical and defined virtual drives.



IBM ServeRAID-MR10k SAS/SATA Controller

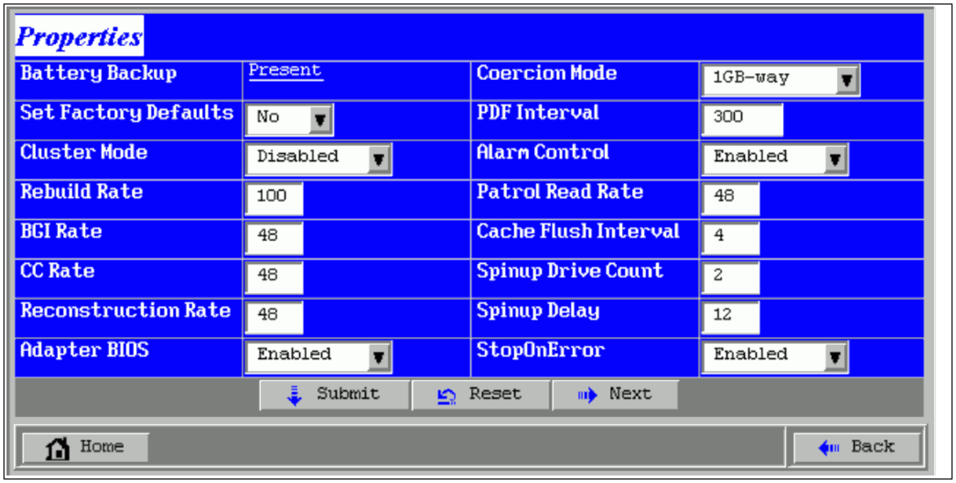
Firmware Version	1.12.122-0393	WebBIOS Version	1.1-33d-Rel
SubVendorID	0x1014	SubDeviceID	0x363
HostInterface	PCIE	PortCount	8
NVRAM Size	32 KB	Memory Size	256 MB
Firmware Time	Jan 31 2008;15:18:12	Serial Number	None
Min Stripe Size	8 KB	Max Stripe Size	1024 KB
Virtual Disk Count	0	Physical Disk Count	4
FW Package Version		8.0.1-0029	

Figure 3-55 Adapter properties (page 1): physical specifications

> **Note:** If a background initialization is in progress, you may click **Background Init Progress** to determine its state of completion.

Click **Next** to view the second page of the Adapter Properties window, shown in Figure 3-56 on page 163. On the second page, click **Next** again to view the third page, shown in Figure 3-57 on page 163. You may change any of the default settings of the ServeRAID adapter properties.

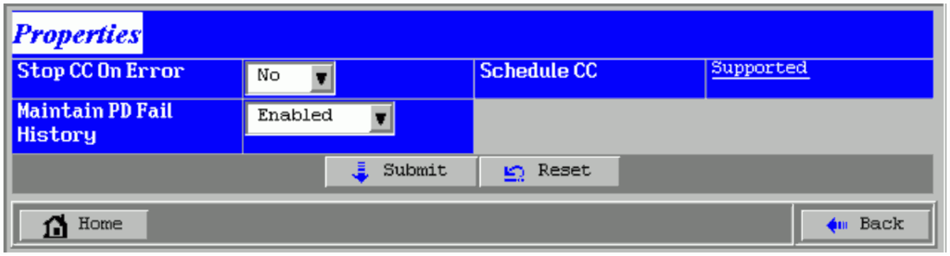
> **Note:** The second page allows you to change the Rebuild Rate, which can negatively affect the performance of the SAS subsystem such as by performance degradation within the operating system and its applications.



Properties

Battery Backup	Present	Coercion Mode	1GB-way
Set Factory Defaults	No	PDF Interval	300
Cluster Mode	Disabled	Alarm Control	Enabled
Rebuild Rate	100	Patrol Read Rate	48
BGI Rate	48	Cache Flush Interval	4
CC Rate	48	Spinup Drive Count	2
Reconstruction Rate	48	Spinup Delay	12
Adapter BIOS	Enabled	StopOnError	Enabled

Figure 3-56 Adapter properties (page 2): settings



Properties

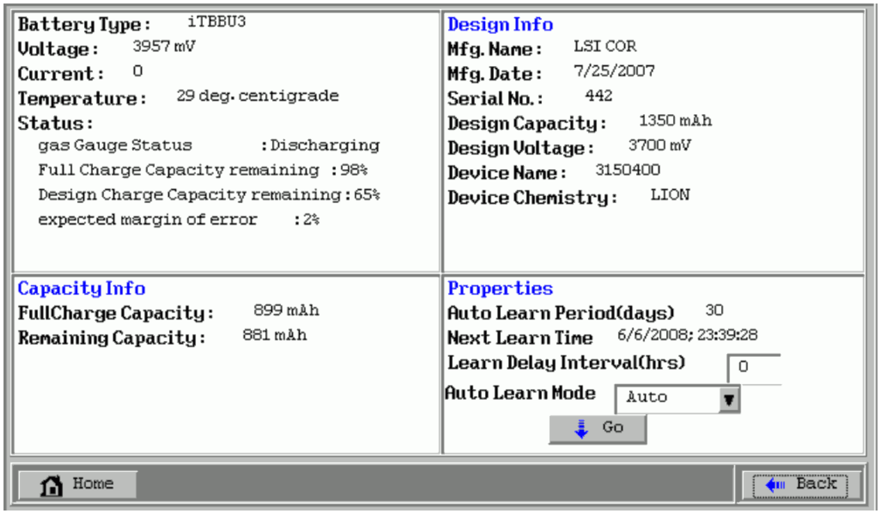
Stop CC On Error	No	Schedule CC	Supported
Maintain PD Fail History	Enabled		

Figure 3-57 Adapter properties (page 3): settings

In the second and third pages, change any of the following default settings:

- Battery Backup

  This indicates whether you installed a battery backup unit (BBU) on the selected controller. If a BBU is present, click **Present** to view details about it. See details in Figure 3-58 on page 164.


Battery Type: iTBBU3
Voltage: 3957 mV
Current: 0
Temperature: 29 deg. centigrade
Status:
gas Gauge Status :Discharging
Full Charge Capacity remaining :98%
Design Charge Capacity remaining :65%
expected margin of error :2%

Design Info
Mfg. Name: LSI COR
Mfg. Date: 7/25/2007
Serial No.: 442
Design Capacity: 1350 mAh
Design Voltage: 3700 mV
Device Name: 3150400
Device Chemistry: LION

Capacity Info
FullCharge Capacity: 899 mAh
Remaining Capacity: 881 mAh

Properties
Auto Learn Period(days) 30
Next Learn Time 6/6/2008; 23:39:28
Learn Delay Interval(hrs) 0
Auto Learn Mode Auto

Figure 3-58 iTBBU properties

- Set Factory Defaults: `No` is the default

  This loads the default configurations.

- Cluster Mode: `Disabled` is the default

  This provides additional cluster services while accessing the same data storage. The default setting cannot be changed because it is not supported.

- Rebuild Rate: `48` is the default

  This is the percentage of system resources dedicated to rebuilding a failed drive. The higher the value is, the more system resources are devoted to a rebuild. The value can be 1-100%.

> **Tip:** We do not recommend setting the lowest nor the highest value.
>
> A very high percentage can effect performance negatively because this is a background process that additionally uses the SAS subsystem.
>
> A very low percentage can result in a loss of data, in case further hard drives within the same RAID subset (VD) are going offline while the Rebuild process is running.

- BGI Rate: `48` is the default

  This affects the amount of system resources dedicated to background initialization (BGI) of virtual disks connected to the selected adapter. The value can be 1-100%.

- CC Rate: `48` is the default

  This indicates the amount of system resources dedicated to consistency checking (CC) of virtual disks connected to the selected adapter. The value can be 1-100%.

- Reconstruction Rate: `48` is the default

  This is where you define the reconstruction rate of physical drives to the selected adapter dedicated to the amount of system resources. The value can be 1-100%.

- Adapter BIOS: `Enabled` is the default

  This enables or disables the Adapter BIOS. If the selected controller is connected to the boot drive, then Adapter BIOS must be enabled on that controller.

> **Tip:** If more than one controller is installed (for example, in a x3950 M2 complex), one of them typically owns the bootable device.
>
> We recommend that you enable Adapter BIOS only on the controller that contains the boot device, and disable it on all others. This maximizes the available PCI ROM space and reduces the chance of getting PCI ROM Allocation errors.

- Coercion Mode: `1GB-way` is the default

  This forces physical disks of varying capacities to the same size so that they can be used in the same array.

  The Coercion Mode can be set to None, 128 MB-way, and 1 GB-way. This number depends on how much the drives from various vendors vary in their actual size.

> **Note:** LSI recommends that you use the 1 GB-way setting.

- PDF Interval: `300` (seconds) is the default (this equals 5 minutes)

  This specifies how frequently the controller polls for physical drives report a Predictive Drive Failure (PDF) Self-Monitoring, Analysis and Reporting Technology (S.M.A.R.T.) error.

- Alarm Control: Disabled is the default

  This enables, disables, or silences the onboard alarm tone generator on the controller.

- Patrol Read Rate: `48` is the default

Definition: A patrol read scans the system for possible physical disk drive errors that could lead to drive failure, then takes action to correct the errors. The goal is to protect data integrity by detecting physical drive failure before the failure can damage data. IBM calls this *data scrubbing*.

This option indicates the rate for patrol reads for physical drives connected to the selected adapter. The patrol *rate* is the percentage of system resources dedicated to running a patrol *read*.

Tip: The corrective action depend on the virtual configuration and the type of errors. It affects performance, the more iterations there are, the greater the impact.

- Cache Flush Interval: `4` (seconds) is the default

  This controls the interval, in seconds, at which the contents of the onboard data cache are flushed.

- Spinup Drive Count: `2` is the default

  This is the number of disks that are started (spun up) simultaneously. Other disks are started after these disks have started (after waiting the number of seconds indicated in Spinup Delay).

- Spinup Delay: `12` (seconds) is the default

  This controls the interval in seconds between spinup of physical disks connected to the selected controller.

  These two Spinup settings (drive count and delay) are important for preventing a drain on the system's power supply that would occur if all disks spin up at the same time.

- StopOnError: `Enabled` is the default

  When the controller encounters an error during boot-up, it stops if this setting is enabled.

- Stop CC On Error: `No` is the default

  An error found in checking the consistency causes further checking to stop if you enable this setting.

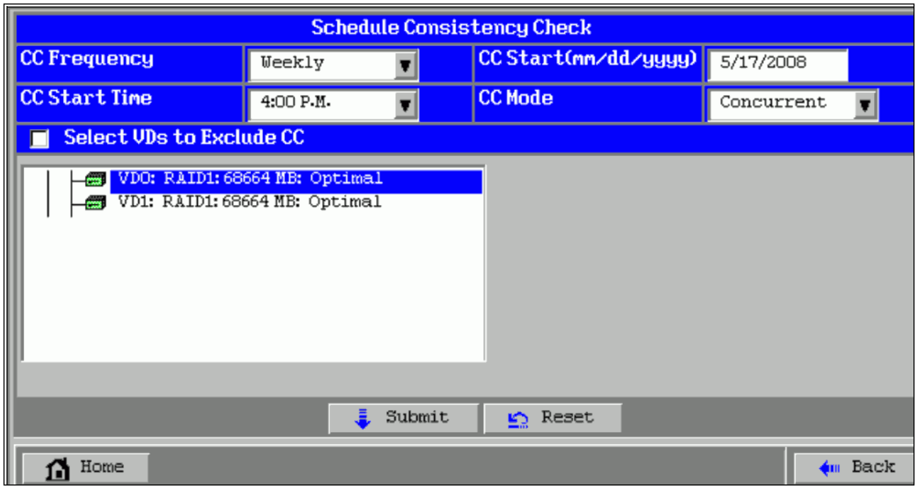
- Maintain PD Fail History: `Enabled` is the default

  If enabled, this option maintains the problem determination failure history.

Note: Disabling this setting is not recommended. Enabling the history is important for recovering multiple disk failures.

- Schedule CC: `Supported`

  This schedules a consistency check. Click **Supported**. The Schedule Consistency Check window opens, shown in Figure 3-59.



Schedule Consistency Check

CC Frequency	Weekly	CC Start(mm/dd/yyyy)	5/17/2008
CC Start Time	4:00 P.M.	CC Mode	Concurrent

□ Select VDs to Exclude CC

VD0: RAID1: 68664 MB: Optimal
VD1: RAID1: 68664 MB: Optimal

Figure 3-59 Adapter properties: settings to schedule a consistency check

The options in this dialog are:

– CC Frequency: `Weekly` is the default

  This controls the frequency of the consistency check. Values can be: continuous, hourly, daily, weekly, monthly, or disable.

  > **Important:** You should not *disable* the consistency check to prevent any impact of data lost in case the amount of disk errors may reach a rate of inconsistency.
  >
  > Consistency checking is a background operation that can affect performance. If you use performance-sensitive applications that are related to high storage I/O traffic, use this setting cautiously.

– CC Start Time: No default

  This is the time of day the consistency check should start.

– CC Start (mm/dd/yyyy): No default.

  This is the date the first consistency check should start.

– CC Mode: `Concurrent` is the default

  This specifies whether to concurrently or sequentially check consistency of the virtual drives.

Virtual Disks option

In the WebBIOS main window, select a particular virtual drive from the list of Virtual Drives and then click **Virtual Disks** from the menu. The window view changes, listing the features for managing virtual disks.

This view is available after creation of virtual disks only. Virtual disks, also known as logical drives, are arrays or spanned arrays that are available to the operating system. The storage space in a virtual disk is spread across all the physical drives in the array.

A virtual drive can have the following states:

- Optimal

  The virtual disk operating condition is good. All configured physical drives are online.

- Degraded

  The virtual disk operating condition is not optimal. One of the configured physical drives has failed or is offline.

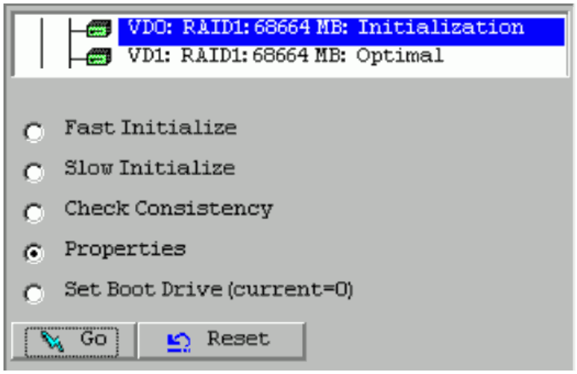
- Failed

  The virtual disk has failed.

- Offline

  The virtual disk is not available to the RAID controller.

This section discusses the features shown in Figure 3-60.



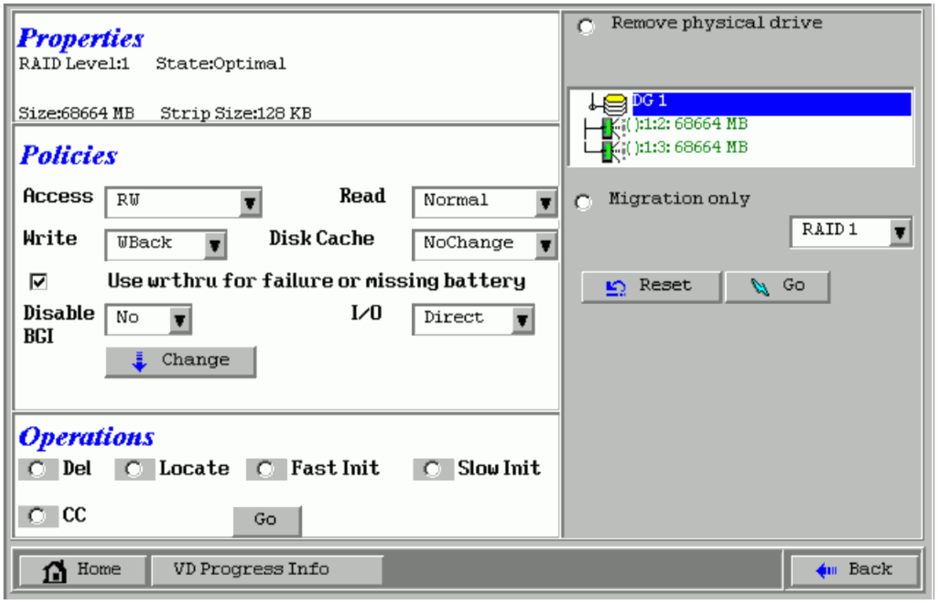
VD0: RAID1: 68664 MB: Initialization
VD1: RAID1: 68664 MB: Optimal

○ Fast Initialize
○ Slow Initialize
○ Check Consistency
● Properties
○ Set Boot Drive (current=0)

Figure 3-60 Virtual disk features

First, however, access all of the features and properties:

1. Select the appropriate virtual disk (VD) from the list (shown in Figure 3-60 on page 168).
2. Select **Properties**.
3. Click **Go**. The properties, policies, and operations are displayed, as shown in Figure 3-61.



Properties

RAID Level:1 State:Optimal

Size:68664 MB Strip Size:128 KB

Policies

Access: RW
Read: Normal
Write: WBack
Disk Cache: NoChange
☑ Use wrthru for failure or missing battery
Disable BGI: No
I/O: Direct

○ Remove physical drive

DG 1
( ):1:2: 68664 MB
( ):1:3: 68664 MB

○ Migration only
RAID 1

Operations

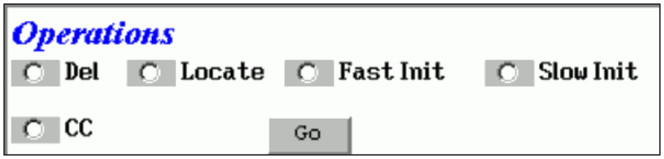
○ Del ○ Locate ○ Fast Init ○ Slow Init
○ CC

Figure 3-61 Virtual disk Properties, Policies, and Operations panels

Note: You can watch the status of any ongoing background process by clicking the **VD Progress Info** button.

Operations panel

Many of the same operations listed in Figure 3-60 on page 168 can be performed from the Operations panel shown in Figure 3-62.



Operations

○ Del ○ Locate ○ Fast Init ○ Slow Init
○ CC

Figure 3-62 WebBIOS virtual disk Operations panel

The Operations options are:

- Delete (Del)

  You can delete any virtual disk on the controller if you want to reuse that space for a new virtual disk. The WebBIOS utility provides a list of configurable arrays where there is a space to configure. If multiple virtual disks are defined on a single array, you can delete a virtual disk without deleting the whole array.

  > **Important:** Be aware that any kind of *initialization* or *deletion* erases all data on your disks. Ensure you have a valid backup of any data you want to keep.
  >
  > The following message appears; confirm by answering `YES`:
  >
  > ```
  > All data on selected Virtual Disks will be lost.
  > Want to Proceed with Initialization? <YES> / <NO>
  > ```

- Locate (Loc)

  Causes the LED on the drives in the virtual disk to flash

- Fast Initialize (Fast Init)

  This initializes the selected virtual disk by quickly writing zeroes to the first and last 10 MB regions of the new virtual disk. It then completes the initialization in the background.

- Slow Initialize (Slow Init)

  This also initializes the selected virtual disk but it is not complete until the entire virtual disk has been initialized with zeroes.

  > **Note:** Slow initialization is rarely used because your virtual drive is completely initialized after creation.

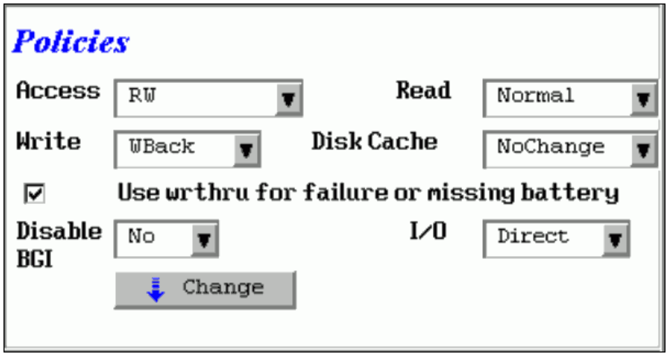
- Check Consistency (CC)

  If the controller finds a difference between the data and the parity value on the redundant array, it assumes that the data is accurate and automatically corrects the parity value.

  > **Note:** Be sure to back up the data before running a consistency check if you think the consistency data might be corrupted.

Policies panel

Figure 3-63 shows the panel where you change the policies of the selected virtual disk.



Policies

Access: RW

Read: Normal

Write: WBack

Disk Cache: NoChange

☑ Use wrthru for failure or missing battery

Disable BGI: No

I/O: Direct

Figure 3-63 The WebBIOS virtual disk Policies panel

Use this window to specify the following policies for the virtual disk:

- Access: `RW` is the default

  Specify whether a virtual drive can be accessed by read/write (RW, default), read-only (R), or no access (Blocked).

- Read: `Normal` is the default

  This is the read-ahead capability. The settings are:

  - Normal: This is the default. It disables the read ahead capability.
  - Ahead: This enables read-ahead capability, which allows the controller to read sequentially ahead of requested data and to store the additional data in cache memory, anticipating that the data will be needed soon. This speeds up reads for sequential data, but there is little improvement when accessing random data.
  - Adaptive: This enables the controller to begin using read-ahead if the two most recent disk accesses occurred in sequential sectors. If the read requests are random, the controller reverts to Normal (no read-ahead).

- Write: `WBack` is the default

  This is the write policy. The settings are:

  - WBack: This is the default. In Writeback mode, the controller sends a data transfer completion signal to the host when the controller cache has received all the data in a transaction. This setting is recommended in Standard mode.

  - WThru: In Writethrough mode, the controller sends a data transfer completion signal to the host when the disk subsystem has received all the data in a transaction.

    Set a check mark for the following setting if you want the controller to use Writeback mode but the controller has no BBU or the BBU is bad:

    `Use wthru for failure or missing battery`

    If you do not set this option, the controller firmware automatically switches to Writethrough mode if it detects a bad or missing BBU.

- Disk Cache: `NoChange` is the default

  The settings are:

  - Enable: Enable the disk cache.
  - Disable: Disable the disk cache.
  - NoChange: Do not change the current disk cache policy.

- I/O: `Direct` is the default

  The I/O policy applies to reads on a specific virtual disk. It does not affect the read ahead cache. The settings are:

  - Direct: In direct I/O mode, reads are not buffered in cache memory. Data is transferred to the cache and the host concurrently. If the same data block is read again, it comes from cache memory. This is the default.
  - Cached: In cached I/O mode, all reads are buffered in cache memory.

- Disable BGI: `No` is the default

  Specify the background initialization (BGI) status. A setting of `No` means that a new configuration can be initialized in the background.

  > **Tip:** A setting of `Yes`, disables BGI. Changing this setting is not recommended because the controller is blocked while any outstanding background process have to be completed, before further actions are allowed.

Physical drive and Migration panel

Options are shown in Figure 3-64 on page 173.

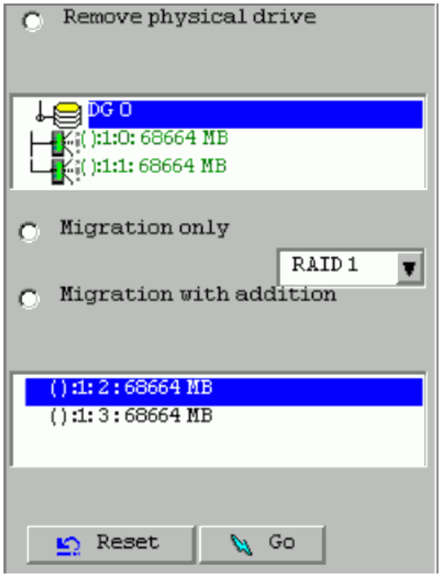


Figure 3-64 WebBIOS virtual disk Properties migration options panel

The following list explains the properties on this panel:

- To set the RAID-level migration:

  a. Select either **Migration only** to change the RAID level, or **Migration with addition** to expand the virtual drive with further available disk space.

  b. Choose the RAID level you want to migrate.

  c. Click **Go.**

  As the amount of data and the number of disk drives in your system increases, you can use RAID-level migration to change a virtual disk from one RAID level to another. You do not have to power down or reboot the system. When you migrate a virtual disk, you can keep the same number of drives, or you can add drives. You can use the WebBIOS CU to migrate the RAID level of an existing virtual disk.

  > **Important:** It is strongly recommended that you have a valid backup of any data you want to keep; there is no reversible way to recover data after a failing migration. The following message reminds you:
  >
  > ```
  > Migration is not a reversible operation. Do you want to proceed?
  > <YES> / <NO>
  > ```

Migrations are allowed for the following RAID levels:

- RAID 0 to RAID 1
- RAID 0 to RAID 5
- RAID 1 to RAID 5
- RAID 1 to RAID 6
- RAID 1 to RAID 0[1]

Note: Although you can apply RAID-level migration at any time, we recommend you do so when no reboots are occurring. Many operating systems issue I/O operations serially during a reboot. With a RAID-level migration running, a boot can often take more than 15 minutes.

After a migration has been started, it runs in the background, in case you did not change the default for the background initialization (BGI) setting. The status of the virtual disk changes from *Optimal* to *Reconstruction* and changes back after the task has completed successfully.

- To remove physical drives:

  a. Select **Remove physical drive** to remove a physical drive.
  b. Select the physical drive you want to remove.
  c. Click **Go**.

  Note: When removing a drive, follow the steps, otherwise the drive might not be stopped properly, which could damage the drive. The hard disk must be allowed to spin down before you remove it from the slot.

- To add physical drives:

  a. Install any new disk and select **Add physical drive** to add a physical drive.
  b. Click **Go**.

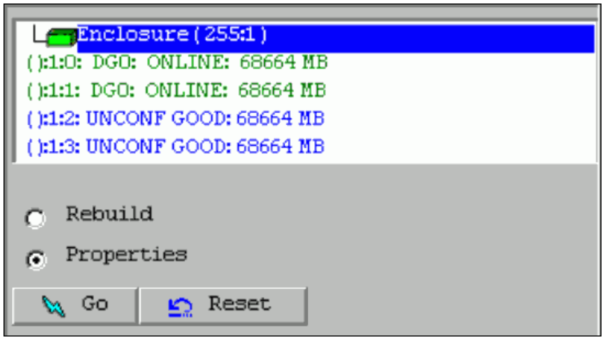
Note: To watch states of any ongoing background process, click the **VD Progress Info** button shown in Figure 3-61 on page 169.

Physical Drives option

In the WebBIOS main window (Figure 3-54 on page 161), select **Physical Drives** from the Controller Function menu (see Figure 3-54 on page 161).

On the panel that is displayed (shown in Figure 3-65 on page 175), you can rebuild a drive either by selecting **Rebuild**, or by selecting **Properties** and then working with the extended properties to do the rebuilding.

[1] At the time of writing, the software User's Guide does not list this migration option. However we verified that the option is available in our tests in our lab.



```
Enclosure(255:1)
( ):1:0: DG0: ONLINE: 68664 MB
( ):1:1: DG0: ONLINE: 68664 MB
( ):1:2: UNCONF GOOD: 68664 MB
( ):1:3: UNCONF GOOD: 68664 MB
```

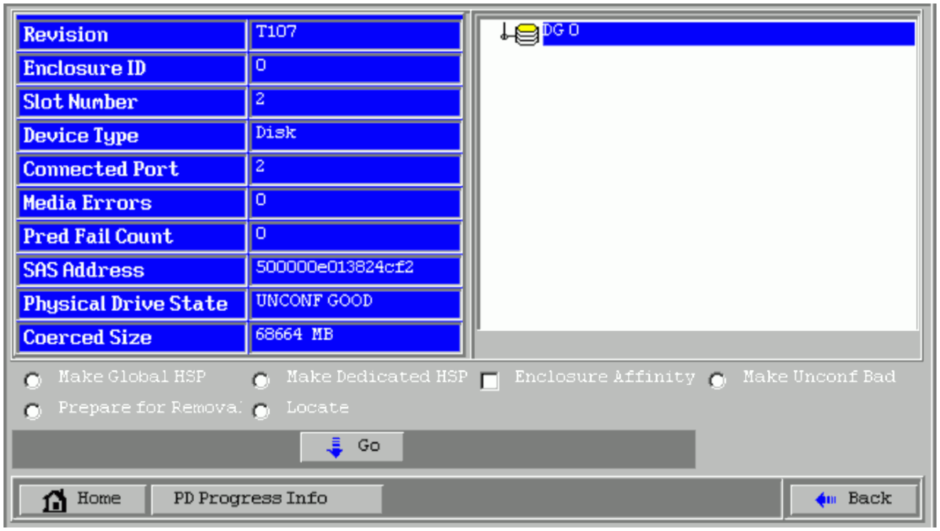
○ Rebuild

● Properties

Figure 3-65 WebBIOS: physical drives

To rebuild by using the extended properties:

1. Select the appropriate physical disk.
2. Select **Properties**.
3. Click **Go**. The extended properties about the selected disk are displayed, as shown in Figure 3-65.



Revision	T107
Enclosure ID	0
Slot Number	2
Device Type	Disk
Connected Port	2
Media Errors	0
Pred Fail Count	0
SAS Address	500000e013824cf2
Physical Drive State	UNCONF GOOD
Coerced Size	68664 MB

Figure 3-66 WebBIOS: physical properties

The properties are:

- Revision information about the firmware
- Enclosure ID, slot number, device type, connected port, SAS World Wide Name (WWN)

- Media errors and predicted failure counters
- SAS address
- Physical drive state, which can be:
  - Unconfigured Good
  - Unconfigured Bad
  - Online
  - Offline
  - Failed
  - Missing
  - Global or Dedicated Hotspare
  - Rebuild
- Coerced size

Physical drive states introduction

The bottom area has settings you can use depending on the state of the physical disk. You can also manage the disk through the drive states:

Unconfigured Good

This is a disk that is accessible to the RAID controller but not configured as a part of a virtual disk or as a hot spare. See Figure 3-67.

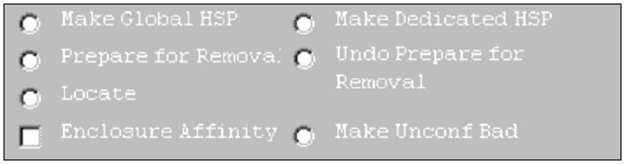


Figure 3-67 Physical drives: unconfigured good drive state

Select one of the following settings:

- Make Global HSP: Set the selected disk to a global hotspare which is assigned to all defined virtual disks. A reconstruction process starts immediately, if any degraded virtual drive has been found.
- Make Dedicated HSP: Set the selected disk to a dedicated hotspare for a specific selected virtual disk only. A reconstruction process starts immediately, if any other physical disk is failed or missed in this virtual disk.
- Prepare for Removal: The firmware spins down this disk drive and the disk drive state is set to unaffiliated, which marks it as offline even though it is not a part of configuration.
- Undo Prepare for Removal: This undoes this operation. If you select undo, the firmware marks this physical disk as unconfigured good.

- Locate: Use this command to flash the LED on the selected disk.

Unconfigured Bad

This state is a physical disk on which the firmware detects an unrecoverable error, the physical disk was *Unconfigured Good*, or the physical disk could not be initialized.

Online

You can access at a physical disk if it is in this state because it is member of a virtual disk configured by you. Figure 3-68 shows the choices.

Figure 3-68 Physical drives: online drive state

Select one of the following options:

- Make Drive Offline: This forces the selected physical disk to an offline state.

> **Important:** If you set a drive offline, this can result in loss of data or redundancy. We recommend to backup all important data, because the selected physical disk is member of the virtual disks.
>
> The following reminder continues to be displayed:
>
> ```
> Making a Drive Offline may result in data and/or redundancy loss.
> It is advisable to backup data before this operation. Do you want
> to continue? <YES> / <NO>
> ```

  A reconstruction process starts immediately if any other physical disk is assigned as global hotspare or as dedicated hotspare for this virtual disk.

> **Note:** We recommend using this option before you remove any disk from the enclosure or the internal disk bays, to prevent damage to a disk that did not spin down.

- Locate: This flashes the LED on the selected disk.

Offline

A drive you forced from Online to Offline can result in loss of data, redundancy, or both. This depends of the RAID level you have configured. Figure 3-69 on page 178 shows the choices.

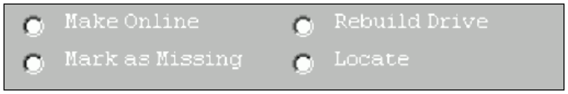


Figure 3-69 Physical drive: offline drive state

Select one of the following options:

- Mark Online: This sets the offline drive back to online.
- Rebuild Drive: This starts the rebuilding with the selected disk at the virtual disks that are in a *degraded* state. The drive state changes to *rebuild*. A status window opens, so you can follow the progress, as shown in Figure 3-70.

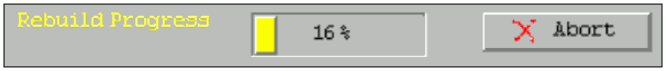


Figure 3-70 Physical drive: offline drive showing Rebuild Progress

- Mark as Missing: This sets this command to mark this drive as missing.
- Locate: This flashes the LED on the selected disk.

Failed

This shows you a physical disk that was configured as online or hotspare but failed, but the controller firmware detected an unrecoverable error and marked this disk as failed.

> **Tip:** After a physical disk failed you should check the events at this disk. We recommend to replace a failed disk. However after you have a valid backup for any data you want to keep, you can try to rebuild this disk. If it failed or a critical event is logged against this disk, replace it.

Missing

A missing drive is a physical drive that is removed or marked missing from an enclosure and is a member of a configured virtual disk.

> **Tip:** If the drive state is Missing but the drive is physically present, try reseating it and seeing if that changes the state back to Unconfigured Good.

A physical disk previously forced to offline can be marked as missing also. See “Offline” on page 177.

Tip: A disk you marked as missing, will show as *Unconfigured Good* typically. To redefine this disk you have to set this disk to *Global* or *Dedicated Hotspare*. The reconstruction process starts immediately.

Global or Dedicated Hotspare

This shows you that the disk is assigned as Global or Dedicated Hotspare.

Note: When a disk is missing that brings a defined virtual disk to the degraded state, a rebuild process (reconstruction) of the array starts immediately, if a Global or Dedicated Hotspare is enabled.

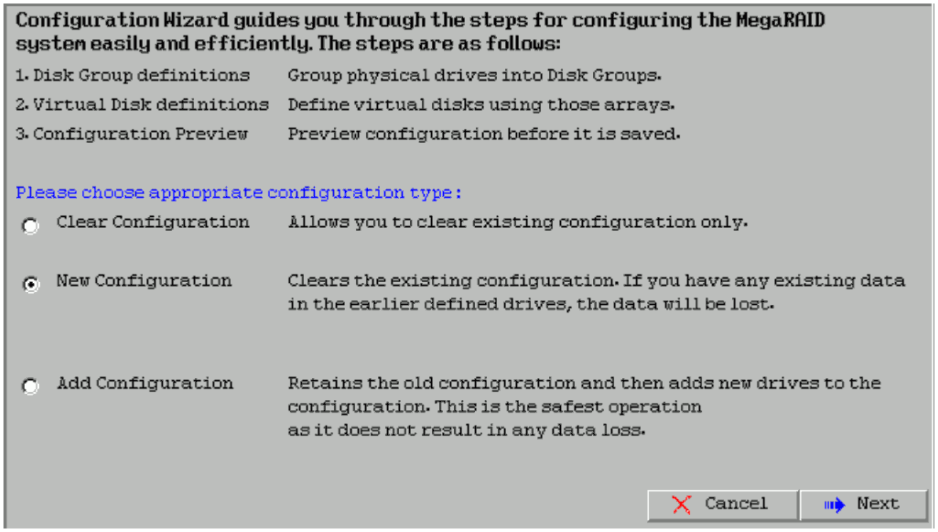
Rebuild

This state indicates that the selected disk is in the progress of reconstruction of a virtual disk (array). A rebuilding starts if you have a missing or failed disk and another *Unconfigured Good* disk was enabled as *Global* or *Dedicated Hotspare,* or if a physical disk that is in state *Unconfigured Good* is forced by you to *Rebuild*. In both situations, reconstruction of the array starts immediately.

Configuration Wizard option

The Configuration Wizard advises you to clear, create new, or add a virtual disk to any type of array that is supported at the type of installed adapter you selected in the Adapter Selection menu (in Figure 3-51 on page 159).

To start the Configuration Wizard, select it from the Function Controller panel shown in Figure 3-54 on page 161. The Configuration Wizard window opens, as shown in Figure 3-71 on page 180.



Configuration Wizard guides you through the steps for configuring the MegaRAID system easily and efficiently. The steps are as follows:

1. Disk Group definitions Group physical drives into Disk Groups.

2. Virtual Disk definitions Define virtual disks using those arrays.

3. Configuration Preview Preview configuration before it is saved.

Please choose appropriate configuration type:

○ Clear Configuration Allows you to clear existing configuration only.

● New Configuration Clears the existing configuration. If you have any existing data in the earlier defined drives, the data will be lost.

○ Add Configuration Retains the old configuration and then adds new drives to the configuration. This is the safest operation as it does not result in any data loss.

Figure 3-71 WebBIOS: Configuration Wizard selection menu

The wizard helps you to perform the following steps:

1. (Disk Group definitions): Group physical drives into Disk Groups.

   In this section you define the disks that will be assigned to create a virtual disk

2. (Virtual Disk definitions): Define virtual disks using those arrays.

   Within the virtual disk definition window, you create and adapt the attributes of any virtual disk.

3. (Configuration Preview): Preview a configuration before it is saved.

   This shows you, your configured virtual disk before you save it.

The three configuration types, shown in the figure, are discussed in the following sections:

- “Clear Configuration” on page 180
- “New Configuration” on page 181
- “Add Configuration” on page 187

Clear Configuration

Use this to clear existing configurations.

Important: If you clear the configuration, all data will be lost. The following reminder appears in the window:

```
This is a Destructive Operation! Original configuration and data
will be lost. Select YES, if desired so. <YES> / <NO>
```

New Configuration

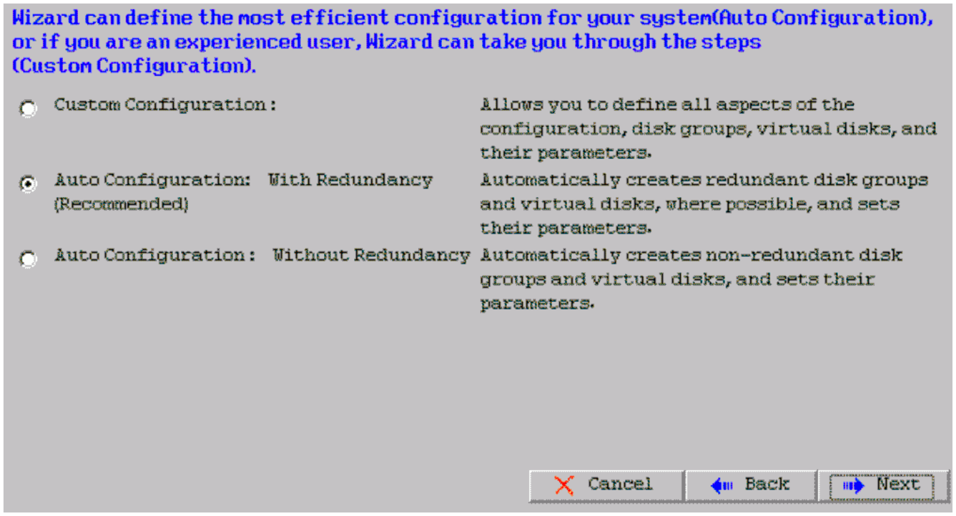
This option clears the existing configuration and guides you to add new virtual disks.

Important: If you use this operation, all existing data will be lost. Be sure that you do not require any existing virtual disks and their data. The following reminder appears on the window:

```
This is a Destructive Operation! Original configuration and data
will be lost. Select YES, if desired so. <YES> / <NO>
```

After you select **Yes,** the following options are available for creating new disks, as shown in Figure 3-72:

- Custom configuration
- Auto configuration with redundancy
- Auto configuration without redundancy



Wizard can define the most efficient configuration for your system(Auto Configuration), or if you are an experienced user, Wizard can take you through the steps (Custom Configuration).

○ Custom Configuration : Allows you to define all aspects of the configuration, disk groups, virtual disks, and their parameters.

● Auto Configuration: With Redundancy (Recommended) Automatically creates redundant disk groups and virtual disks, where possible, and sets their parameters.

○ Auto Configuration : Without Redundancy Automatically creates non-redundant disk groups and virtual disks, and sets their parameters.

Figure 3-72 WebBIOS: Configuration Wizard starting a new configuration

Note: Use the **Back** and **Next** buttons to display the previous or next window in the wizard. You may also cancel the operation by clicking **Cancel** at any time before the new created configuration is saved.

- Auto Configuration: *with* and *without* redundancy

  Although Custom Configuration is first in the list on the wizard window, we describe it second. We guide you though the automatic configuration here, but we recommend you use the Custom Configuration if you have strong experience of working with LSI controllers.

  To create a configuration with automatic configuration, with or without redundancy:

  a. When WebBIOS displays the proposed new configuration, review the information about the window. If you accept the agreement, click **Accept** (or click **Back** to go back and change the configuration.) The RAID level selected depends on whether you selected with or without redundancy and the number of drives you have installed, as shown in Table 3-17.

  b. Click **Yes** when you are prompted to save the configuration.

  c. Click **Yes** when you are prompted to initialize the new virtual disks.

     WebBIOS CU begins a background initialization of the virtual disks.

Table 3-17 Algorithm to select the RAID level for automatic configuration

If you select this option...	And you have this many drives installed...	...then this RAID level is automatically selected
Auto configuration without redundancy	Any	RAID-0
Auto configuration with redundancy	Two	RAID-1
Auto configuration with redundancy	3	RAID-6[a]
Auto configuration with redundancy	4 or 5	RAID-10
Auto configuration with redundancy	6 or more	RAID-60[b]

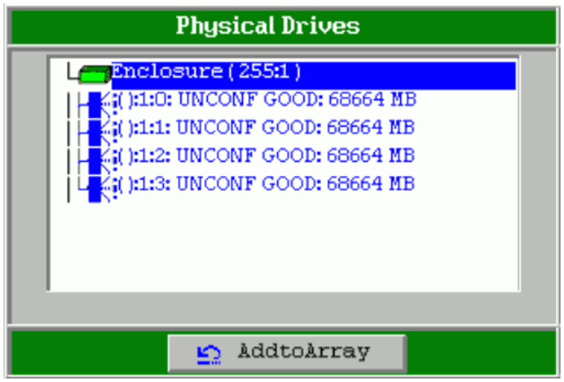
[a] The LSI controller implements RAID-6 that can be comprised of three drives (instead of a minimum of four).

[b] The LSI controller implements RAID-60 that can be comprised of six drives (instead of a minimum of eight).

- Custom Configuration

  When you want to set up different or more virtual disks, use the custom setup.

Click Custom Configuration. In the Physical Drives panel, shown in Figure 3-73, define the number of disks that you want to add to a virtual disk to assign to a disk group.



Enclosure(255:1)
( ):1:0: UNCONF GOOD: 68664 MB
( ):1:1: UNCONF GOOD: 68664 MB
( ):1:2: UNCONF GOOD: 68664 MB
( ):1:3: UNCONF GOOD: 68664 MB

Figure 3-73 WebBIOS: Configuration Wizard custom DG definition (on left)

You can assign all available disks that have the status of *Unconfigured good* to different disk groups as follow:

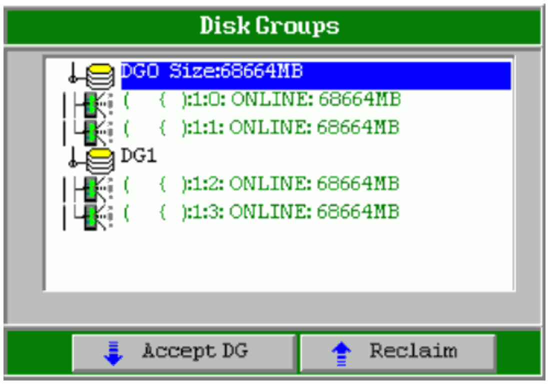
a. Select the disk in the panel view.

b. Click **AddtoArray** to assign to a disk group. Repeat this step to add more disks.

> **TIp:** You can assign multiple disks at the same time if you press the `Shift` key and use the `Up` and `Down arrow` keys to select the disks, can then click **AddtoArray**.

c. Confirm this disk group by clicking on the **Accept DG** button, shown in Figure 3-74 on page 184. The WebBIOS increases the disk groups and allows you to assign more disks. Repeat this step if you want another disk group.

d. If you want to undo a drive addition, click the **Reclaim** button.



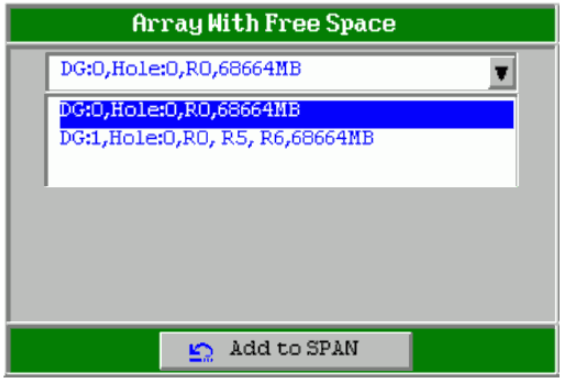
```
DG0 Size:68664MB
(    (   ):1:0: ONLINE: 68664MB
(    (   ):1:1: ONLINE: 68664MB
DG1
(    (   ):1:2: ONLINE: 68664MB
(    (   ):1:3: ONLINE: 68664MB
```

Figure 3-74   WebBIOS: Configuration Wizard custom DG definition

> **Tips:** All disks that you define within a disk group will be assigned to the virtual drives that you create later.
>
> At the time of writing, the Reclaim button does not have any affect to undo a disk addition.Instead, go back one window to the configuration selection menu and start the process of creating a disk group again.

e. Click **Next**.

f. In the Array With Free Space panel, shown in Figure 3-75, select the disk group for which you want to create a virtual disk.



```
DG:0,Hole:0,R0,68664MB
DG:0,Hole:0,R0,68664MB
DG:1,Hole:0,R0, R5, R6,68664MB
```

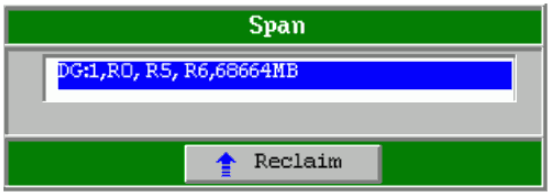
Figure 3-75   WebBIOS: Configuration Wizard custom Span definition

> **Note:** The WebBIOS suggests all available RAID levels for disk groups.

g. Click the **Add to SPAN** button.

To span a disk group to create RAID 10, 50 or 60 you must have two disk groups with the same number of disks and capacity.

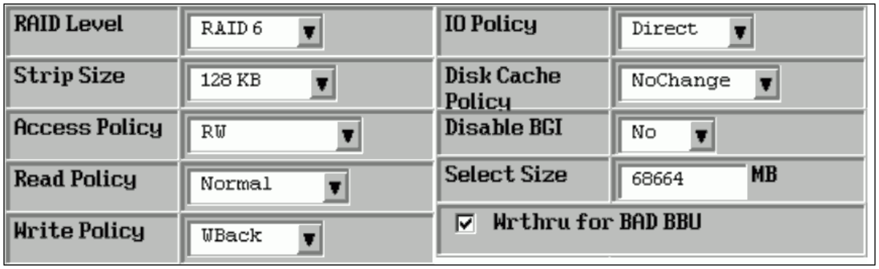
The selected disk group is added to the **Span** panel, shown in Figure 3-76.



DG:1,R0, R5, R6,68664MB

Figure 3-76 WebBIOS: Configuration Wizard custom Span definition on right

h. Click **Next** to change the properties of the virtual disk, as shown in Figure 3-77.



RAID Level	RAID 6	IO Policy	Direct
Strip Size	128 KB	Disk Cache Policy	NoChange
Access Policy	RW	Disable BGI	No
Read Policy	Normal	Select Size	68664 MB
Write Policy	WBack	☑ Wrthru for BAD BBU	

Figure 3-77 WebBIOS: Configuration Wizard custom with RAID properties

You can change the following settings:

- RAID Level: Select the RAID level to use with the disks you selected.
- Stripe Size: Select a stripe size from 8 KB to 1024 KB blocks. The recommended size is 128 KB.

> **Note:** The value of the stripe size affects system performance and depends on the application you want to use. Learn what the best value for your application is and fine tune later, if necessary.

- Access Policy: Select a policy for the virtual drive. Values are read/write, read-only or not blocked.
- Read Policy: Specify the read policy for this virtual drive:

  Normal: This disables the read ahead capability. This is the default.

Ahead: This enables read ahead capability, which allows the controller to read sequentially ahead of requested data and to store the additional data in cache memory, anticipating that the data will be needed soon. This speeds up reads for sequential data, but there is little improvement when accessing random data.

Adaptive: When Adaptive read ahead is selected, the controller begins using read ahead if the two most recent disk accesses occurred in sequential sectors. If the read requests are random, the controller reverts to Normal (no read ahead).

- Write Policy: Specify the write policy for this virtual drive:

  WBack*:* In Writeback mode the controller sends a data transfer completion signal to the host when the controller cache has received all the data in a transaction. This setting is recommended in Standard mode.

  WThru: In Writethrough mode the controller sends a data transfer completion signal to the host when the disk subsystem has received all the data in a transaction. This is the default.

  Bad BBU: Select this mode if you want the controller to use Writeback mode but the controller has no BBU or the BBU is bad. If you do not choose this option, the controller firmware automatically switches to Writethrough mode if it detects a bad or missing BBU.

- IO Policy: The IO Policy applies to reads on a specific virtual disk. It does not affect the read-ahead cache.

  Direct: In direct I/O mode, reads are not buffered in cache memory. Data is transferred to the cache and the host concurrently. If the same data block is read again, it comes from cache memory. This is the default.

  Cached: In cached I/O mode, all reads are buffered in cache memory.

- Disk Cache Policy: Specify the disk cache policy:

  Enable: Enable the disk cache.

  Disable: Disable the disk cache. This is the default.

  Unchanged: Leave the current disk cache policy unchanged.

- Disable BGI: Specify the background initialization status:

  Leave background initialization enabled with set to `No`.

- Select Size: Specify the size of the virtual disk in megabytes.

  By default the size is set to the maximum allowed size defined by the RAID level at this virtual disk. You can specify a smaller size if you want to create further virtual disks on the same disk group. The space that is available is called a *hole*.

i. After reviewing and changing all options, click **Accept** to confirm the changes or **Reclaim** to return to previous settings.

j. Click **Next** when you are finished defining virtual disks.

k. If necessary, click **Back** to create a further virtual drive,

l. Click **Next** to jump to the wizard's step **3 - Configuration Preview.**

m. Press **Accept** to save the configuration. At the following message, select **Yes** to save and initialize:

```
Save this Configuration? <No> / <Yes>
All Data on the new Virtual Disks will be lost. Want to
Initialize? <No> / <Yes>
```

Note: After you confirm to save the new configuration, the new created virtual disks are shown in the Virtual Drives view at the Home Screen View.

Add Configuration

This option retains the old configuration and then adds new drives to the configuration. You can add new virtual disks by using **Auto Configuration** after inserting new disks, or by using **Custom Configuration** if you have unallocated space (a *hole*) in the assigned disk groups.

Note: This is the safest operation because it does not result in any data loss.

Events option

The WebBIOS Event Information panel keeps any actions and errors that are reported by the selected controller. The events can be filtered at the different RAID components for:

- Virtual Disk
- Physical Device
- Enclosure
- BBU
- SAS
- Boot/Shutdown
- Configuration
- Cluster

The events can be filtered to the following criteria:

- Informal
- Warning

- Critical
- Fatal
- Dead

Define the start sequence and the number of events you are looking for.

3.5 PCI Express options

This section describes the PCI Express (PCIe) subsystem and the supported PCI Express adapters of the x3850 M2 and x3950 M2. The x3850 M2 and x3950 M2 both have seven half-length full-height PCIe x8 slots, two of which support the hot removal (hot-swap) of adapters.

The slots are connected to one of two 4-port PCIe x8 controllers as follows:

- PCIe controller 1: Slot 1-4
- PCIe controller 2: Slot 5-7, onboard SAS LSI1078

The onboard SAS LSI1078 uses a dedicated PCIe port to the second PCIe controller. The onboard SAS controller and the optional ServeRAID-MR10k are discussed in 3.3, "Internal drive options and RAID controllers" on page 124.

The Remote Supervisor Adapter II and the new Broadcom NIC interface are linked to the Intel ICH7 Southbridge, so that it is no longer shared with the onboard SAS and RAID controller because it is at X3 chipset based systems.

3.5.1 PCI and I/O devices

The following I/O slots/devices are available at Intel ICH7 Southbridge:

- RSA II adapter: dedicated PCI 33MHz/32bit slot to RSA II video
- Onboard Broadcom NetXtreme II 10/100/1000 NIC 5709C: PCIe x4 connection to the single-chip high performance multi-speed dual port Ethernet LAN controller
- ServeRAID-MR10k: dedicated 128-pin DIMM socket

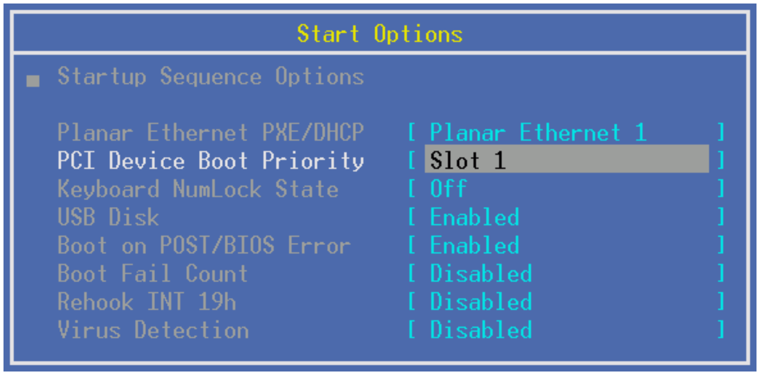
3.5.2 PCI device scan order

The system scans the slots in the following sequence:

1. Slot internal SAS devices.
2. PCIe slot 1, 2, 3, 4, 5, 6, 7

3. Integrated Ethernet controller
4. PCIe slot 8, 9, 10, 11, 12, 13, 14; then 15, 16, 17, 18, 19, 20, 21, and so on, in a multinode configuration

You can use the Configuration Setup utility program to change the sequence and have the server scan one of the first six PCIe slots before it scans the integrated devices. Press `F1` when prompted during system POST. In the Start Options panel, set the **PCI Device Boot Priority** to the slot you want to boot from as shown in Figure 3-78. The default is **Planar SAS**.



```
                    Start Options
■ Startup Sequence Options

  Planar Ethernet PXE/DHCP    [ Planar Ethernet 1      ]
  PCI Device Boot Priority    [ Slot 1                 ]
  Keyboard NumLock State      [ Off                    ]
  USB Disk                    [ Enabled                ]
  Boot on POST/BIOS Error     [ Enabled                ]
  Boot Fail Count             [ Disabled               ]
  Rehook INT 19h              [ Disabled               ]
  Virus Detection             [ Disabled               ]
```

Figure 3-78 BIOS: PCI Device Boot Priority

> **Note:** If you want to change this setting on secondary nodes in a multinode configuration, you should first boot to stand-alone (non-merged) and then make the necessary changes. Once you have complete the changes, you can then reemerge the multinode configuration.

3.5.3 PCI adapter installation order

We recommend that you install the adapters in a specific add-in order to balance bandwidth between the two PCIe controllers.

The recommended order for PCIe adapter installation is:

1. Slot 1
2. Slot 5
3. Slot 2
4. Slot 6
5. Slot 3
6. Slot 7
7. Slot 4

The server BIOS controls the PCI adapter boot sequence by executing the slot's onboard ROM in this sequence. If multiple adapters from the same vendor are installed (for example, multiple ServeRAID MR10M adapters), the ROM from one of them will be used for all of them. These adapters have their own policy regarding which one will be the boot adapter. ServeRAID or HBA adapter typically fall into this category. Because of this, the boot order might appear to be incorrect. System BIOS does not and cannot control this behavior.

A re-arrangement might be required in case the adapters, installed in your system, do not map their ROM space dynamically, which can result in overlapping to the mapped ROM space of other adapters in the same system. A PCI ROM space allocation event can occur in POST after you power on the server. This means that not all PCI Express adapters are initialized during POST, because the PCI ROM space limit was reached or any overlapping of PCI resources remained. This behavior is seen typically if you add-in, for example, Ethernet adapters in slots with a higher slot-scan priority before the Fibre channel HBA or ServeRAID adapters.

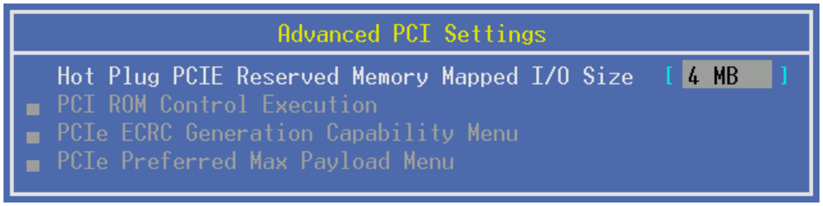
Disabling of specific features such as PXE boot at Ethernet adapters or the BIOS at HBA adapters, that you do not have to boot from SAN devices, and disabling of onboard devices are also options for solving any PCI ROM allocation errors. Your in country technical support team may advice here.

3.5.4 PCI Express device-related information in the BIOS

Information about a PCI Express device can be found in the BIOS of your x3850 M2 and x3950 M2 servers.

BIOS PCI Express options

BIOS PCI Express options are displayed after you press `F1` during system POST and then select **Advanced Setup** → **Advanced PCI Settings**. See Figure 3-79.



```
                    Advanced PCI Settings
  Hot Plug PCIE Reserved Memory Mapped I/O Size  [ 4 MB    ]
■ PCI ROM Control Execution
■ PCIe ECRC Generation Capability Menu
■ PCIe Preferred Max Payload Menu
```

Figure 3-79 BIOS: Advanced PCI Settings

The options in this windows are:

- Hot plug PCI Express Reserved Memory Mapped I/O Size

  By default, 4 MB of memory mapped I/O (MMIO) space is reserved for hot-swap PCIe adapters. If an adapter exceeds the reserved resources during a hot-plug operation, the device driver does not load. The Windows operating system shows a message similar to:

  ```
  This device cannot find enough free resources that it can use.
  ```

  You can reserve more MMIO space with this entry, up to a maximum of 32 MB.

- PCI ROM Control Execution

  This allows you to enable or disable the ROM execution on a per slot-basis.

- PCI Express ECRC Generation Menu

  ECRC (end-to-end cyclic redundancy check) is the data transmission error detection feature. You can override the generation capability in each slot.

  An adapter that is not capable of checking PCI Express ECRC must be run with the PCI Express ECRC generation turned off below the PCI Express root port for proper operation.

- PCI Express Preferred Max Payload Menu

  This sets the Preferred Max Payload Size for each PCI Express slot. At POST, BIOS sets the working Max Payload Size to the lower of the Preferred Max Payload Size and the Max Payload Size capability reported by the adapter.

  The default Preferred Max Payload Size is 512 bytes.

  For multinode configurations, this setting must be changed independently on each node.

  The default Max Payload Size optimizes performance for high speed adapters. Though performance can be affected, a lower size might be necessary to avoid issues after changing device power states in certain operating systems without an intervening reboot.

  This can happen, for example, if an adapter device power state changes when removing and reloading the adapter driver within the operating system. An adapter originally set to 512 bytes can then stop working if the adapter Max Payload Size is inadvertently reset to 128 bytes by the device power state transition while the root port (chipset) May Payload Size remains at the 512 value set at POST. In this example, the adapter Preferred Max Payload Size and functionality are restored at the next reboot.

Hot plug support

Hot plugging is not natively supported by Microsoft Windows 2003 and prior versions of the operating systems. The system BIOS must initialize and configure any PCI Express native features that buses, bridges, and adapters require before they boot these operating systems.

IBM provides the relevant Active PCI slot software for Windows 2000 and Windows Server 2003 from the following Web address:

`http://www.ibm.com/support/docview.wss?uid=psg1MIGR-62127`

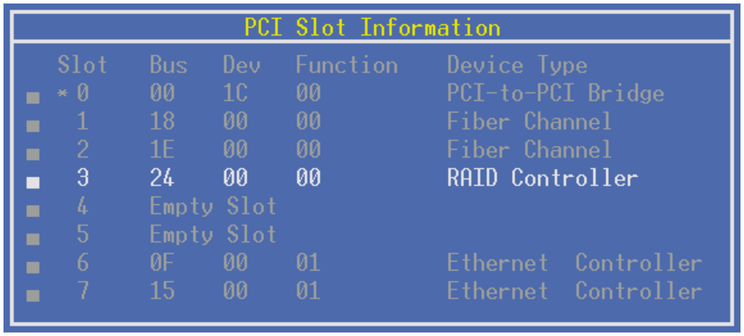
Microsoft Windows 2008 supports several PCI Express features natively, such as Active State Power Management (ASPM), Advanced Error Reporting as part of Windows Hardware Error Handling Architecture (WHEA), Extended configuration space access through the Memory-Mapped PCI Configuration Space (MCFG) table, Hot-plug, Message Signaled Interrupt (MSI), and Power Management Event (PME).

IBM provides Active PCI software for Windows Server 2008 from the following Web address:

`http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5074966`

PCI device information

You can review detailed information at all onboard devices and optional installed PCIe adapters. Boot to the server configuration menu by pressing `F1` while the message is shown in the POST. Then select the **PCI Slot/Device Information** in the **Advanced Setup** and select any line that is not marked empty to get the detailed information, see Figure 3-80.

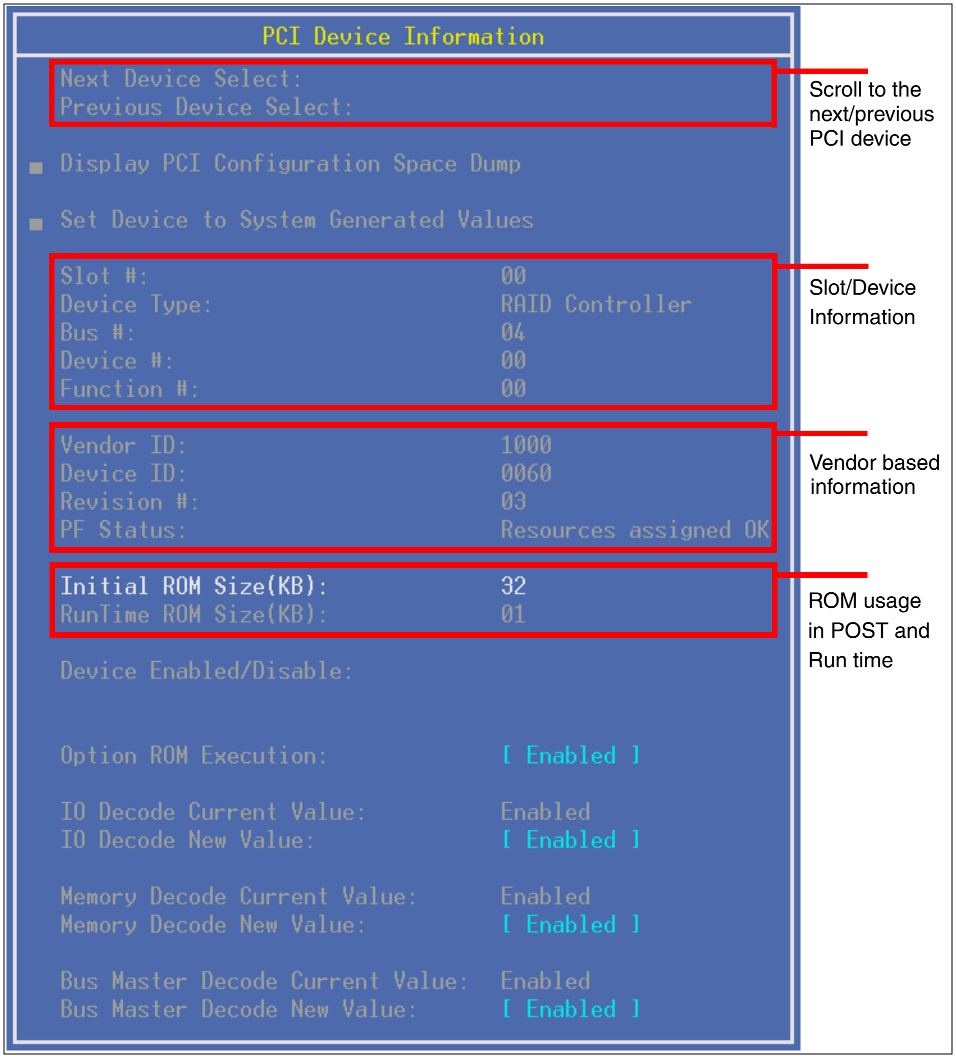


```
                     PCI Slot Information
    Slot     Bus    Dev    Function     Device Type
■  * 0       00     1C     00           PCI-to-PCI Bridge
■    1       18     00     00           Fiber Channel
■    2       1E     00     00           Fiber Channel
■    3       24     00     00           RAID Controller
■    4       Empty Slot
■    5       Empty Slot
■    6       0F     00     01           Ethernet  Controller
■    7       15     00     01           Ethernet  Controller
```

Figure 3-80   BIOS: PCI Slot Information

An asterisk (*) next to the slot number indicates that more than one device is in this slot. Slot 0 is the planar system board, which contains the devices found at all chassis in a multinode configuration in the POST after its merged.

In Figure 3-81 an example is shown that details the ServeRAID-MR10k device information.



```
                    PCI Device Information
  Next Device Select:
  Previous Device Select:

■ Display PCI Configuration Space Dump

■ Set Device to System Generated Values

  Slot #:                            00
  Device Type:                       RAID Controller
  Bus #:                             04
  Device #:                          00
  Function #:                        00

  Vendor ID:                         1000
  Device ID:                         0060
  Revision #:                        03
  PF Status:                         Resources assigned OK

  Initial ROM Size(KB):              32
  RunTime ROM Size(KB):              01

  Device Enabled/Disable:


  Option ROM Execution:              [ Enabled ]

  IO Decode Current Value:           Enabled
  IO Decode New Value:               [ Enabled ]

  Memory Decode Current Value:       Enabled
  Memory Decode New Value:           [ Enabled ]

  Bus Master Decode Current Value:   Enabled
  Bus Master Decode New Value:       [ Enabled ]
```

Scroll to the next/previous PCI device

Slot/Device Information

Vendor based information

ROM usage in POST and Run time

Figure 3-81 BIOS: ServeRAID-MR10k PCI Device Information

The most important information that can be checked here is the initial ROM size and whether all resources can be assigned correctly. If a system reports an initialization error, by showing a message in POST or in the server error logs. The initial ROM size of each installed and enabled PCI device must be determined.

3.5.5 Supported PCI Express adapter options

After completing intensive tests with different operating systems and different external environments, IBM provides all the options that are supported with IBM System x systems, in table form, on the IBM ServerProven Web page:

http://www.ibm.com/servers/eserver/serverproven/compat/us/

This list is updated regularly and might contain limitations for using a specific PCI Express option.

Choose your system and scroll to the required subcategory, based on the type of adapter you are looking for. Select the link at the relevant option to get detailed information.

4

Multinode hardware configurations

This chapter describes what you have to consider in working with multinode hardware configurations. We provide prerequisites and setup information for creating a multinode complex, and how to work with multinode systems.

This chapters discusses the following topics:

4.1 Introduction and terminology

A multinode configuration is multiple individual servers connected together by a high speed bus to form a single system image. This system is seen in an installed operating system as one system and the operating system has access to all resources in each server.

We often refer to this multinode configuration as a *complex* or an *n-node scalable system*. Each individual server within this complex is called a *node*.

The following terms are used when describing a multinode configuration:

- Node

  An x3950 M2 server that contains at least one processor, one memory card with a pair of DIMMs installed, and has the capability to scale.

- 1-node system or stand-alone system

  This is a server that has been temporarily removed from a complex, or has been booted separately for maintenance.

- 2-node, 3-node, 4-node systems

  A scaled system of two nodes, three nodes, or four nodes.

- Complex

  This is at least two systems that have the ScaleXpander chip installed as described in 4.5, “Upgrading an x3850 M2 to an x3950 M2” on page 204, and are cabled with the scalability cables as described in 4.6, “Cabling of multinode configurations” on page 209.

- Complex Descriptor

  This is a data record that is generated by a BMC in an x3950 M2 and that is capable in scaling of partitions. A ScaleXpander chip is required and installed.

  The data record contains local system-specific information, and the system state and partition information of all other systems in this complex.

  See details in 4.3.2, “Complex Descriptor contents” on page 200.

- Merging

  This is when the nodes are in the process of forming a multinode complex to bind their resources to one system.

  Progress messages appear in the POST, followed by a `Merge completed` message. During this process you have the option to cancel the merge process and boot each node individually in an *unmerged* state (such as for maintenance). See “Start a multinode system” on page 234.

- Merging timeout

  This is the time that the system waits for the merge process to complete. If the merge process does not complete within this time, the nodes in the partition are booted in standalone mode instead. The default value is six seconds.

- Multinode mode

  A created partition is started, merges successfully, and is running as a single image.

- Stand-alone mode

  A node is booted separately and not merged in a partition.

- Partition

  A partition is the merging process of one or more nodes within a multinode complex. It is seen in any type of a supported operating system as one system. Partitioning can be performed at any time to add or remove a node in it to scale resources, but does require a reboot to activate.

  This is a more flexible alternative to recabling your systems in a complex.

- ScaleXpander

  This is the scalability enablement chip, which is a hardware component that is inserted into a dedicated socket to enable a system to join a multinode complex. See Figure 4-5 on page 205.

- Primary node

  This is the lead node in a configured partition.

- Secondary nodes

  These are all other nodes in a configured partition.

4.2 Multinode capabilities

The x3950 M2 is the initial base building block, or *node*, for a scalable system. At their most basic, these nodes are comprised of four-way SMP-capable systems with processors, memory, and I/O devices. The x3950 M2 is the building block that allows supported 8-way, 12-way, and 16-way configurations by adding other x3950 M2s as required.

> **Note:** When we refer to an x3950 M2, we mean both an x3950 M2 or an x3850 M2 that has the ScaleXpander Option Kit installed.

The x3950 M2 can form a multinode configuration by adding one or more x3950 M2 servers. Various configurations are possible as shown in Figure 4-1.

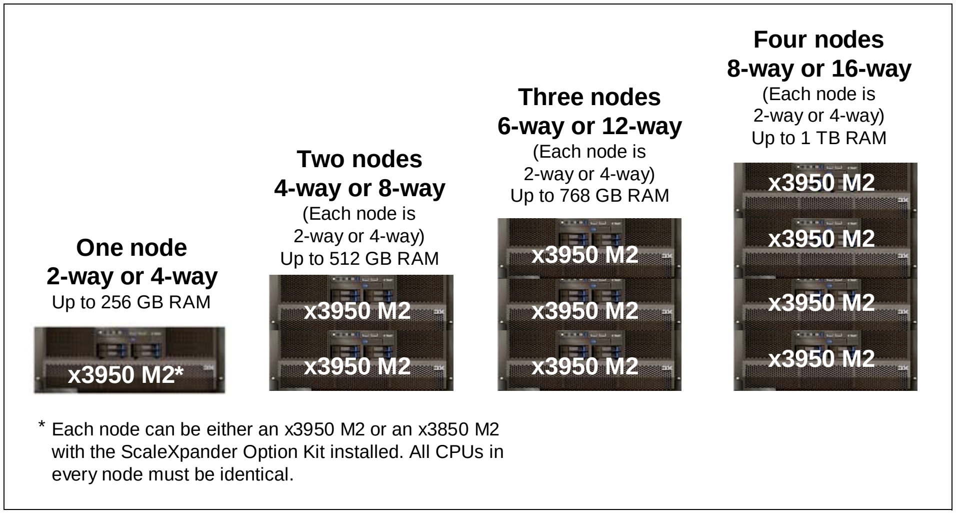


Figure 4-1 Supported multinode configurations

The possible configurations are:

- A two-node complex comprised of two x3950 M2 servers, with four or eight processors, and up to 512 GB RAM installed
- A three-node complex comprised of three x3950 M2 servers, with six or 12 processors, and up to 768 GB RAM installed
- A four-node complex comprised of four x3950 M2 servers, with eight or 16 processors, and up to 1 TB RAM installed

4.3 Understanding scalability

The science behind the eX4 technology allows multiple x3950 M2 systems to form a single scalable multinode system. Building such a complex is simply a matter of connecting the nodes using scalability cables as described in 4.6, “Cabling of multinode configurations” on page 209, we use the term *scalability*. This is the requirement to prepare all hardware resources in all of these nodes and assign then to one single system.

After multiple nodes are cabled together to form a multinode complex, the next step is to define the single system image, a procedure called *partitioning*. As a result, all hardware resources of each node in a configured partition are bound to the one system and are shown to the installed operating system as a single system.

For example, consider a complex of four nodes. Your current production environment might dictate that you configure the complex as a single two-node partition plus two nodes as separately booting systems. As your business needs change, you might have to reconfigure the complex to be a single three-node complex with an extra single-node system, and then later, form one partition out of all four nodes. A recabling in these circumstances is not required.

The new partitioning technology is managed by the Baseboard Management Controller (BMC) firmware in the x3950 M2 systems and a ScaleXpander key is required to enable the scalability features. This is different from the previous X3 generation of x3950 systems, which manages the partitioning by the Remote Supervisor Adapter (RSA) II firmware.

Any RSA Adapter II in any x3950 M2 in this complex can be used to control the partitions. The RSA II is necessary for configuration and management purposes, and offers a very clear Web interface layout. Learn more about working with the RSA2 Web interface by reading sections 4.7, “Configuring partitions” on page 220 and 4.8, “Working with partitions” on page 228.

Because of the new scalability features that are included in the BMC firmware, a disconnection of any RSA2 LAN interface does not affect the stability of your production running multinode system. The BMC handles all instructions and discovers the complex information through the scalability cables.

4.3.1 Complex Descriptor

Each BMC generates a *unique* Complex Descriptor after the ScaleXpander chip is detected by the local BMC in all systems in this complex. See Figure 4-2 on page 200. All Complex Descriptor information is read by the RSA II firmware from the BMC in this system. Therefore, you can use each RSA II Web interface in this complex to manage your partitions.

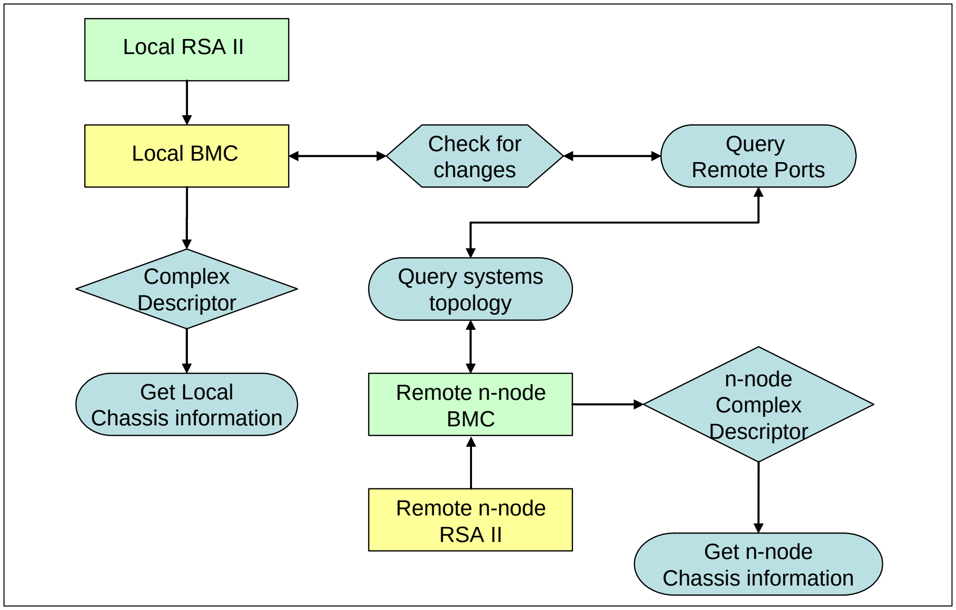


Figure 4-2 Scalability protocol communication path in a complex

The BMC gets the partition descriptor information of all other nodes in this complex by exchanging information through the connected scalability cables on the scalability ports at the rear side of your server. Figure 4-10 on page 211 shows the scalability ports.

Note: The Complex Descriptor is unique; it contains all system and partition information.

4.3.2 Complex Descriptor contents

The contents of the Complex Descriptor include:

- A unique Complex Descriptor signature
- Number of chassis
- UUID and structure of each chassis
- Logical component ID (to specific which one is the primary)
- Partition ID (a unique partition identifier)
- Node power state
- Number of partitions

- Partition-related information (such as merge timeout definition)
- Chassis information for serial and UUID checksum
- Scalability port information

The BMC firmware discovers scalable systems automatically, to check for changes at the scalability port connections, adding new systems and any changes at any existing system in this complex. This is all done by the attached scalability cables. The BMC reads all the Complex Descriptors and keeps the information consistent in all the BMC.

> **Note:** The BMC in each system discovers and keeps consistent complex and partition information of all nodes in a complex. This allows the RSA2 in any system in a complex to manage the all of the partitioning.

The BMC monitors the Complex Descriptors information in the complex. A change in the Complex Descriptor information, which always holds information about the state of all systems and partitions, presents a red alert in the RSA II Scalability Management menu Web interface if an error occurs. It also indicates static errors, which is a problem of a particular system, or the cabling at a scalability port, shown in Figure 4-24 on page 224.

> **Note:** In the case of any unexpected behavior, which can result in an unexpected reboot of the whole partition, but can be recovered after a reboot, it is not reported in the RSA II Scalability Management menu. There you can see static errors only, that affect any system in a partition or the complex.
>
> To isolate the failure, we recommend that you check the system event information in the RSA II Web interface on all connected nodes.
> See Figure 4-31 on page 238.

4.4 Prerequisites to create a multinode complex

This section describes what to consider before you can create a multinode complex. Before you configure a scalability partition, be aware of the following prerequisites:

- A scalable system must have a scalability hardware enabler key, also known as the ScaleXpander chip, installed into the x3850 M2 and x3950 M2. This provides the capability to scale the system, or add to an existing multinode configuration to create new or add to existing hardware partitions. See 4.5, "Upgrading an x3850 M2 to an x3950 M2" on page 204

The ScaleXpander key is standard in x3950 M2 model numbers nSn. See 1.2, "Model numbers and scalable upgrade options" on page 9.

The ScaleXpander key can be installed in an x3850 M2 with the ScaleXpander Option Kit, part number 44E4249. An x3850 M2 with the installed ScaleXpander Option Kit is considered equivalent to an x3950 M2.

- All nodes within a multinode configuration must have the same processor family, speed, and cache sizes.
- Two or four processors are supported in each node.
- A multinode complex can be formed using two, three, or four nodes. A complex of eight nodes is not supported.
- For two-node and three-node complexes, the necessary cables are included with the x3950 M2 or with the ScaleXpander Option Kit.

  For a four-node complex, additional cables are required; they are included in the IBM XpandOnDemand Scalability Kit 2, part number 44E4250.

  Table 4-1 lists part numbers for the components required when you set up a multinode configuration.

Table 4-1 Supported and required scalability option part numbers

Option part number	Description
44E4249	IBM ScaleXpander Option Kit, to upgrade an x3850 M2 to an x3950 M2 to enable scalability features. Kit contains: ► One ScaleXpander key ► One 3.08m (9.8ft) scalability cable ► One x3950 M2 front bezel ► One Enterprise Cable Management Arm
44E4250	IBM Scalability Upgrade Option 2 provides additional cables to form a 4-node complex. It contains: ► One 3.08m (9.8ft) scalability cable ► One 3.26m (10.8ft) scalability cable

  Refer to 4.6, "Cabling of multinode configurations" on page 209.

- The primary node must have at least 4 GB of RAM to form a multinode complex.

- The Hurricane 4 controller in each node in a complex consumes 256 MB of the main memory by the embedded XceL4v dynamic server cache technology, described in 1.6.2, “XceL4v dynamic server cache” on page 30.

  After initialization of main memory, you see the amount of memory that is consumed for XceL4 cache after a successful merging is completed.

- All systems you integrate in a complex *must* have the same firmware levels for RSA II, FPGA, BMC and BIOS.

> **Note:** Our recommendation is to flash all the system components after you install the ScaleXpander chip. We recommend you reflash these components, even if the latest versions are already installed.
>
> The UpdateXpress-based update packages do not allow the reflashing, if the same level of code is installed already. In our lab, we used the DOS-based update utilities to do this.

- The RSA II network interface should be configured and attached to a separate management LAN.
- You must have access to at least one RSA II network connection to control or configure your scalability configurations.

  You can manage all partitions by the RSA Adapter II’s scalability management menu in any system in this complex.

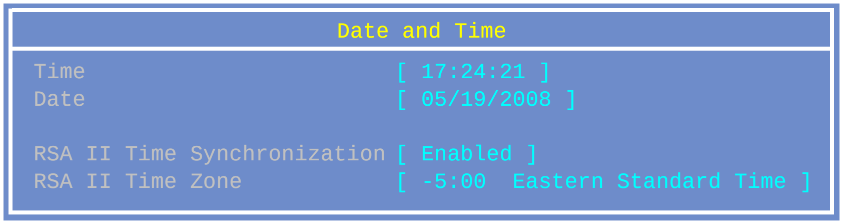
  We recommend that you connect the RSA II in each node to your management LAN. This ensures that you can connect to the RSA II Web interface of at least one node to manage the complex. This can be important if you require assistance by your local IBM technical support if the system experiences any hardware-related error.

  Remember, the remote video and remote drive features are available from the RSA II interface of the primary system only after a merge is completed.

- The BIOS date and time settings should be adjusted to your local time settings.

> **Note:** We recommend synchronizing the system time and the RSA II in the BIOS to be sure they have the same time stamp information for events that can occur. See Figure 4-3 on page 204. You may also set up and use a time server for the operating system and Remote Supervisor Adapter II.

Boot to System Setup (press `F1` when prompted) and select Date and Time. Figure 4-3 on page 204 appears.



```
                       Date and Time
Time                               [ 17:24:21 ]
Date                               [ 05/19/2008 ]

RSA II Time Synchronization [ Enabled ]
RSA II Time Zone                   [ -5:00  Eastern Standard Time ]
```

Figure 4-3   BIOS: Date and Time settings

- After you install the ScaleXpander chip, the Scalable Partitioning submenu appears in the RSA II Web interface, shown in Figure 4-4.

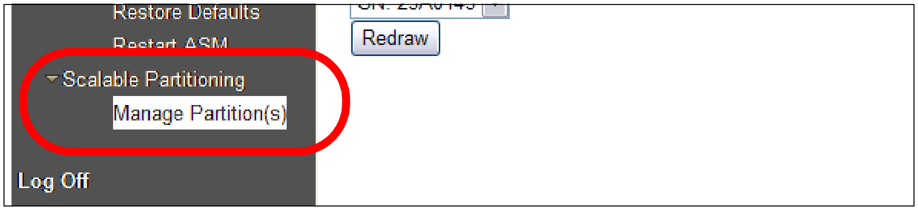


Figure 4-4   RSA II Web interface with enabled scalability features

4.5 Upgrading an x3850 M2 to an x3950 M2

You may upgrade your x3850 M2 system to a x3950 M2 system. For each x3850 M2 that you want to upgrade, add a scalability partition by ordering the IBM ScaleXpander option kit, part number 44E4249.

Review section 4.4, “Prerequisites to create a multinode complex” on page 201.

4.5.1 Installing the ScaleXpander key (chip)

Figure 4-5 on page 205 shows the ScaleXpander key (chip), which must be installed in a vertical position on the processor board.

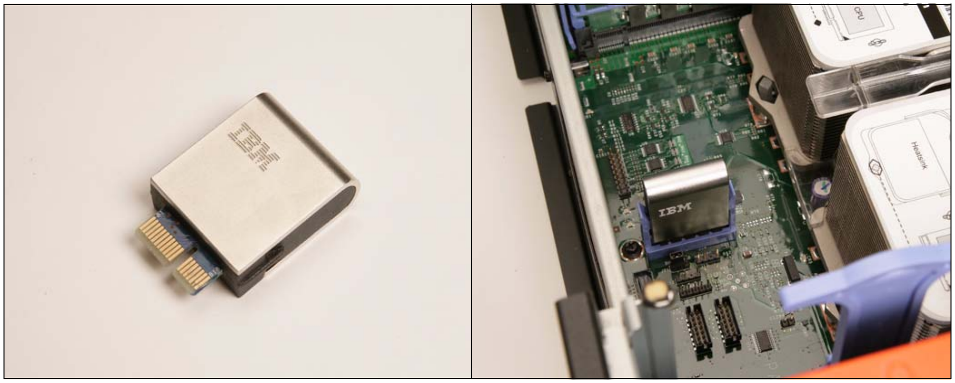
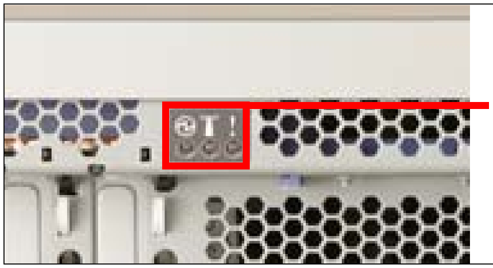


Figure 4-5 ScaleXpander key (left); ScaleXpander key installed on processor board near the front of the x3950 M2 (right)

To install the ScaleXpander key:

1. Unpack your ScaleXpander option kit.
2. Remove the power cables and ensure that the blue locator LED at the rear side of the server is off, indicating that the system is without power. See Figure 4-6.



Rear side LED from left to right

- Green: System power LED
- **Blue: Locator LED**
- Amber: System error LED

Figure 4-6 x3850 M2 and x3950 M2: rear side LED

3. Remove the cover and the front bezel of x3850 M2 server. See instructions in the *System x3850 M2 and x3950 M2 User's Guide*.
4. Loosen the captive screws and rotate the media hood (Figure 3-3 on page 95) to the fully opened position. You now an open view to the processor board.
5. Locate the ScaleXpander key connector at the front of the processor board. See Figure 4-5 and Figure 4-7 on page 206.
6. Check that the ScaleXpander key is oriented correctly direction and push it into the blue slides, which hold the ScaleXpander key in place, until it is firmly seated. You hear a clicking sound. See Figure 4-7 on page 206.

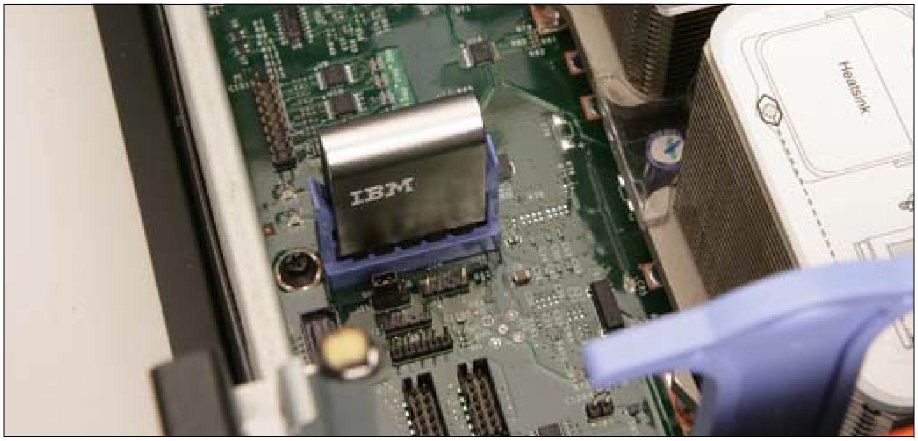


Figure 4-7   x3950 M2: installation of the ScaleXpander key

7. Close the media hood assembly and tighten the captive screws.

The x3850 M2 is now enabled for scalability. To indicate that the server is now an x3950 M2, you might want to replace the front bezel of the x3850 M2 with the one provided in the ScaleXpander Option Kit. Follow the replacement instructions included in the ScaleXpander Option Kit.

Note: The new bezel does not have a sticker where you can enter the original serial number of your server. You should record this data separately.

4.5.2 Configuring for a LAN connection

At least one system in a complex *must* have a valid LAN connection to the Remote Supervisor Adapter (RSA) II. The RSA II has a 10/100 Mbps network interface.

Note: We recommend configuring the RSA II in each node with unique network settings, not just at the primary node. Configuring each RSA II with an IP address means you can more easily manage the complex from any node.

To configure the LAN interface:

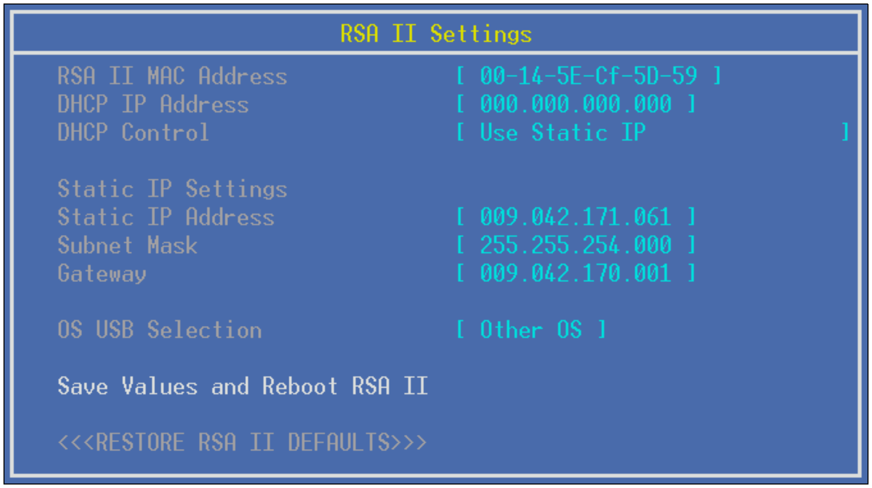
1. Plug in the power cables and power on your system.
2. When prompted, press `F1` to enter the System Setup in BIOS.
3. Select **Advanced Setup** → **RSA II Settings** → **RSA LAN interface**.

Tip: Another way to access the RSA II Web interface is by connecting your workstation directly to the RSA II using a crossover Ethernet cable, and then connecting to the RSA II with the following address and subnet:

- IP address: 192.168.70.125
- Subnet: 255.255.255.0

You could also set up the RSA IP for each DHCP request. The following IP settings are offered in the BIOS and RSA II Settings:

– Try DHCP then use Static IP (default)
– DHCP Enabled
– Use Static IP



```
                           RSA II Settings

   RSA II MAC Address                  [ 00-14-5E-Cf-5D-59 ]
   DHCP IP Address                     [ 000.000.000.000 ]
   DHCP Control                        [ Use Static IP           ]

   Static IP Settings
   Static IP Address                   [ 009.042.171.061 ]
   Subnet Mask                         [ 255.255.254.000 ]
   Gateway                             [ 009.042.170.001 ]

   OS USB Selection                    [ Other OS ]

   Save Values and Reboot RSA II

   <<<RESTORE RSA II DEFAULTS>>>
```

Figure 4-8 BIOS: RSA II Settings

4. Configure the RSA II LAN interface to fit in your LAN segment.
5. Save the settings and select **Restart the RSA II.** The RSA II resets, enabling the new settings.

Note: The RSA Adapter II should provide its LAN settings, at the latest after 20 seconds, to the switch. The RSA II is pingable.

6. Use your workstation to ping and access the Remote Supervisor adapter's Web interface as described in 6.2.3, "Web interface" on page 321.

4.5.3 Updating the code levels, firmware

You can now consider the x3850 M2 to be an x3950 M2 because it has been upgraded with the ScaleXpander key.

Ensure that all systems are at the same code levels. Failure to do so could cause unpredictable results. As we described earlier, we recommend you force a reflash of all firmware to prevent unexpected behavior.

The system firmware that is relevant to scalability terms includes:

- RSA II, FPGA, BMC and BIOS
- The minimum required build level for a two-node system is shown in Table 4-2.

Table 4-2 Two-node x3950 M2: system code levels

Description	Image release	Build level
RSA2	v1.01	A3EP20A
FPGA	v1.18	A3UD18B
BMC	v2.32	A3BT31B
BIOS	v1.03	A3E129A

- The minimum required code levels for three-node and four-node system are shown in Table 4-3.

Table 4-3 Three -node and four-node x3950 M2: system code level

Description	Image release	Build level
RSA2	v1.02	A3EP26A
FPGA	v1.22	A3UD22A
BMC	v3.40	A3BT40A
BIOS	v1.05	A3E145A

Download the following firmware:

- Remote Supervisor Adapter II firmware:

  `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073124`

- Field Programmable Gate Array (FPGA) firmware:

  `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5074499`

- Baseboard Management Controller (BMC) firmware:

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073127

- System BIOS update:

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073120

  > **Important:** After updating the BIOS, reset the BIOS to the default settings.

Our observation has shown that using system flash utilities that are based on the following items, do not reflash any component if the version in the update source is the same as that on the system:

- UpdateXpress CD or UpdateXpress system packs
- Operating system firmware update packages for Windows (wflash) or Linux (lflash) systems

See section 5.1, "Updating firmware and BIOS" on page 246.

> **Recommendation**: Use DOS-based packages, which provide an option to re-flash (overwrite) the same code level. This should be done at least for FPGA and BMC on all systems.

After applying the new codes, power off all systems, and then remove the power source for 30 seconds.

4.6 Cabling of multinode configurations

This section describes how the x3950 M2 systems are cabled.

This section assumes you have already installed the systems into your rack using the instructions in *Rack Installation Instructions* for System x3850 M2 and x3950 M2, available from:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073030

> **Note:** The scalability cables used to join the nodes together are a specific length; no empty U-spaces should exist between the nodes.

Before you begin the multinode cabling, install the Enterprise Cable Management Arm (CMA) that is shipped with your x3950 M2.

If your x3950 M2 has been upgraded from an x3850 M2, remove the standard cable management arm that came with the x3850 M2 and replace it with the

Enterprise Cable Management Arm that was included with the ScaleXpander Option Kit.

Figure 4-9 shows the CMA installed.

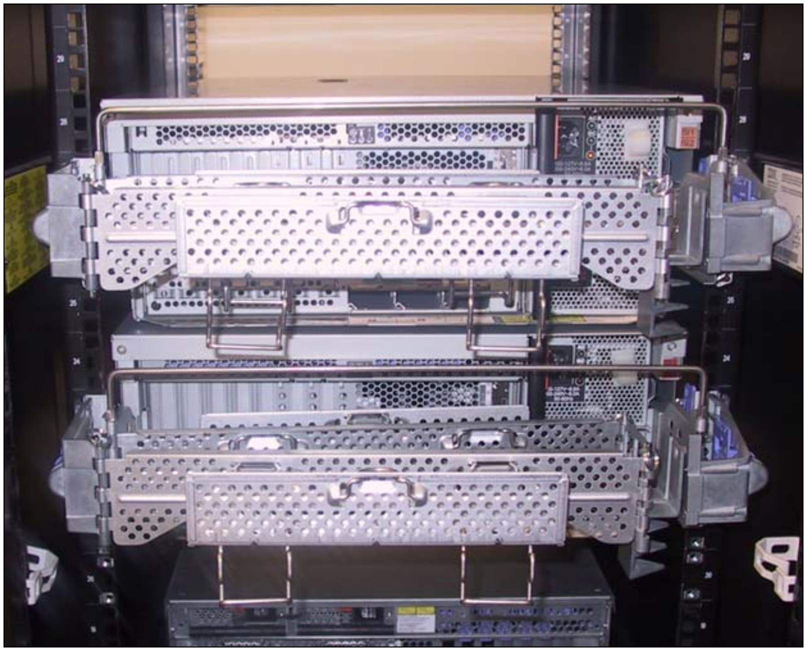


Figure 4-9 x3950 M2: Enterprise Cable Management Arm

Figure 4-10 on page 211 shows the ports of the x3950 M2 where the scalability cables are installed. Depending on the number of nodes you plan to connect together, you use either two or three of these ports in each node.

The ports are named port 1, port 2, and port 3, starting from the left. Each ports has an indicator LED that shows an active connection on this port.

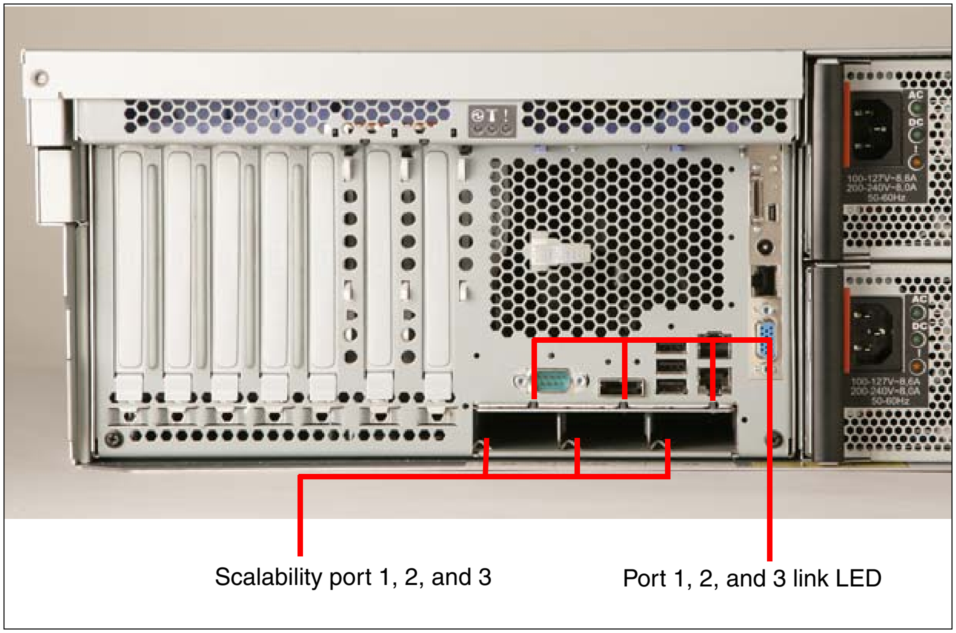


Figure 4-10 x3950 M2: Rear view with scalability connectors

Figure 4-11 shows the scalability cable connector. The connectors are the same on both ends of the cable.

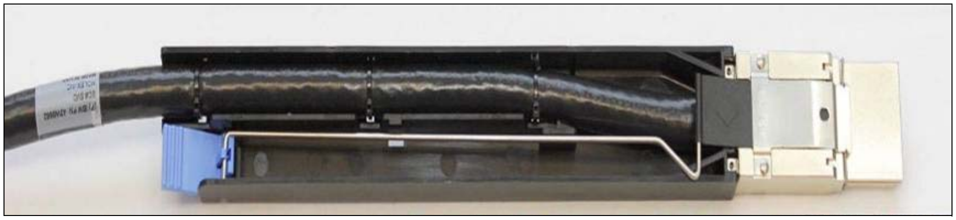


Figure 4-11 x3950 M2: Scalability cable connector

Insert the cables on one server first, follow the installation guidance in the next sections to route the scalability cables through the cable management arm. Figure 4-12 on page 212 shows cables inserted.

> **Tip:** After you insert a cable into the scalability port, ensure that the cable is fully inserted and securely in place.

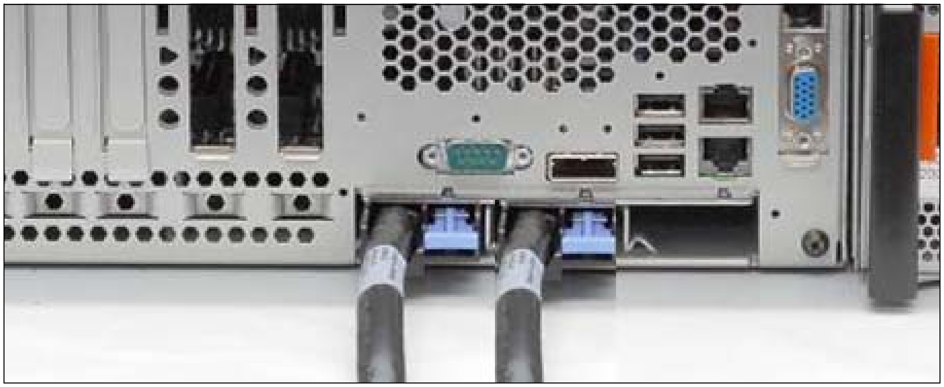


Figure 4-12 x3950 M2 - connected scalability cables

4.6.1 Two-node configuration

A two-node configuration requires two 3.0 m (9.8 feet) ScaleXpander cables. The cable diagram is shown in Figure 4-13 on page 213. One 3.0 m cable is included with every x3950 M2 system, or in the ScaleXpander Option Kit (for the upgraded x3850 M2 systems).

The scalability cables required for a two-node configuration are listed in Table 4-4.

Table 4-4 Scalability cables for two-node configuration

Scalability cable	**Connection (→)**
3.0 m cable	Port 1 of node 1 → port 1 of node 2
3.0 m cable	Port 2 of node 1 → port 2 of node 2

To cable a two-node configuration:

1. Label each end of each ScaleXpander cable according to where it will be connected to each server.
2. Connect the ScaleXpander cables to node1:
   a. Connect one end of a ScaleXpander cable to port 1 on node 1; then, connect the opposite end of the cable to port 1 of node 2.
   b. Connect one end of a ScaleXpander cable to port 2 on node 2; then, connect the opposite end of the cable to port 2 of node 2.

   See Figure 4-13 on page 213.

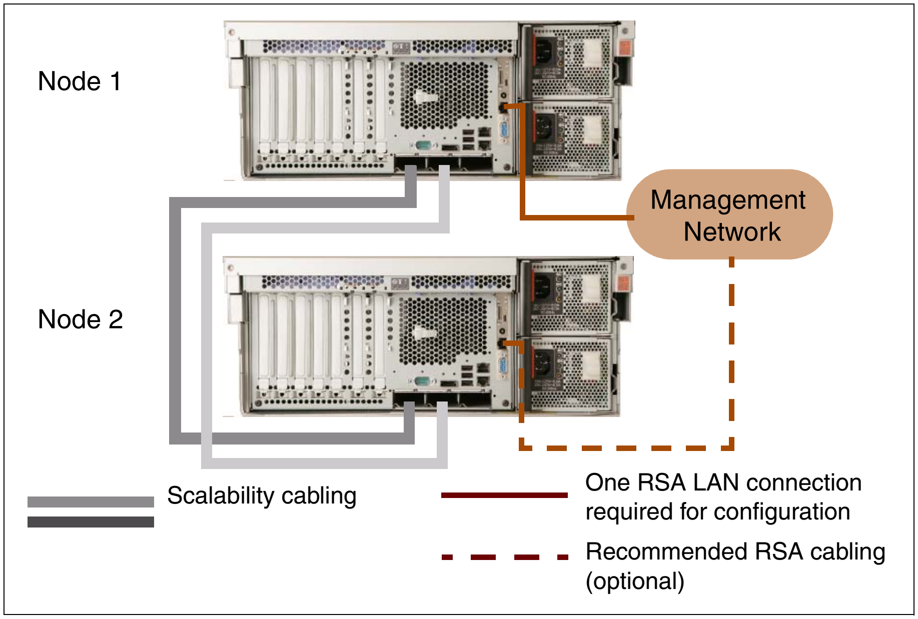
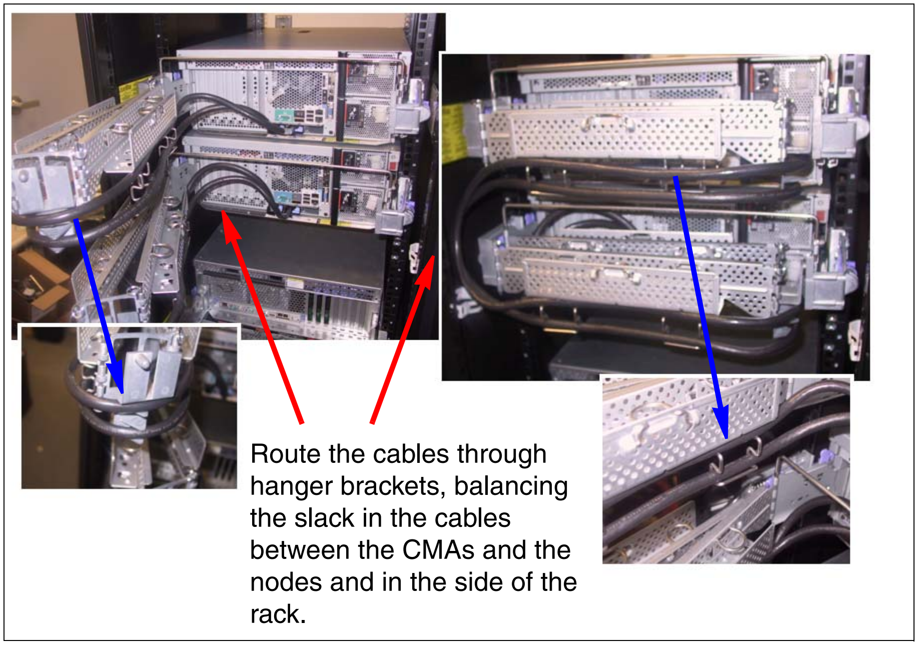

Node 1
Node 2
Management Network
Scalability cabling
One RSA LAN connection required for configuration
Recommended RSA cabling (optional)

Figure 4-13 Two-node x3950 M2: ScaleXpander cable port assignment

3. Route the cables through the Enterprise Cable Management Arms (CMAs) as follows (refer to Figure 4-14 on page 214), ensuring that in each case, you have enough spare length of the cable between the rear side of the server and the first hanger bracket, and on the front side of the cable management arm:

   a. Route the cable connected on port 1 of node 1 through the hanger brackets on the cable management arms.

   b. Route the cable connected on port 2 of node 1 through the hanger brackets on the cable management arms.



Route the cables through hanger brackets, balancing the slack in the cables between the CMAs and the nodes and in the side of the rack.

Figure 4-14 x3950 M2: ScaleXpander cable routing

4. Ensure each server can be pulled out fully at the front of the rack and the scalability cables are clearly routed.

4.6.2 Three-node configuration

A three-node configuration requires three 3.0 m (9.8-foot) ScaleXpander cables. The cable diagram is shown in Figure 4-15 on page 215. One 3.0 m cable is included with every x3950 M2 system, or with the ScaleXpander Option Kit (for upgraded x3850 M2 systems).

The scalability cables required for a three-node configuration are listed in Table 4-5.

Table 4-5 Scalability cables for three-node configuration

Scalability cable	Connection
3.0 m cable	Node 1: scalability port 1 → node 2: scalability port 1
3.0 m cable	Node 1: scalability port 2 → node 3: scalability port 1
3.0 m cable	Node 2: scalability port 2 →node 3: scalability port 2

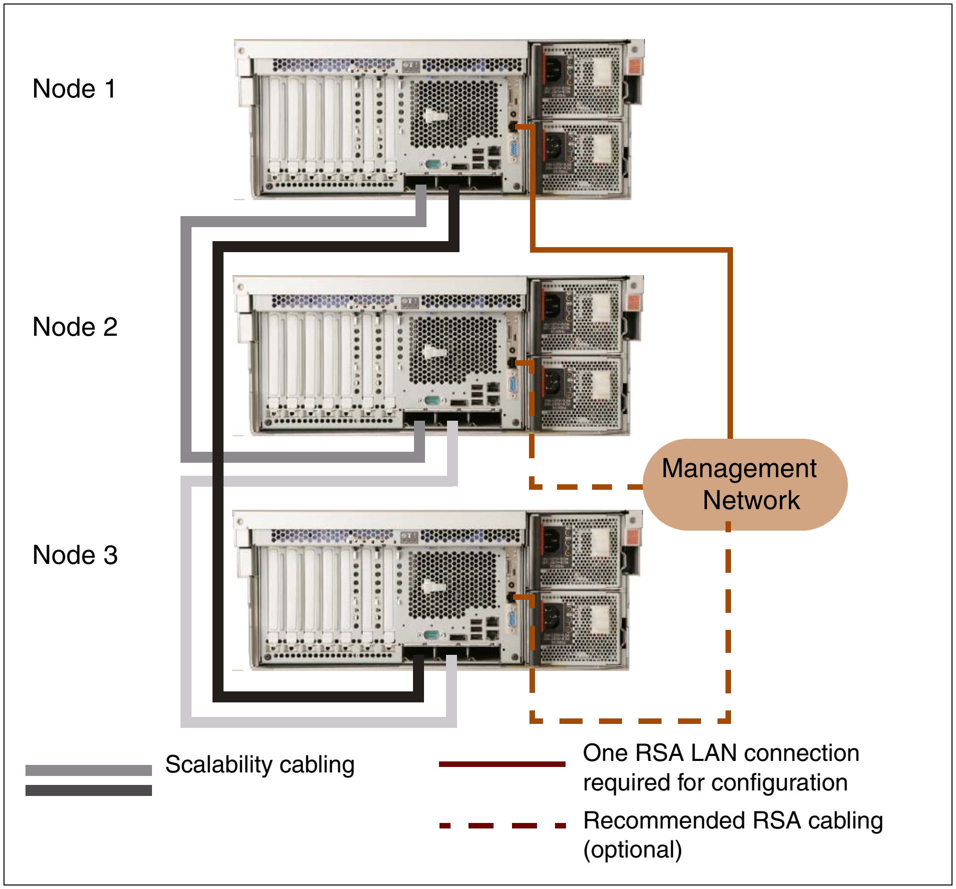

Node 1
Node 2
Node 3
Management Network
Scalability cabling
One RSA LAN connection required for configuration
Recommended RSA cabling (optional)

Figure 4-15 Three-node x3950 M2: ScaleXpander cable port assignment

To cable a three-node configuration:

1. Label each end of each ScaleXpander cable according to where it will be connected to each server.
2. Connect the ScaleXpander cables to node 1:
   a. Connect one end of a ScaleXpander cable to port 1 on node 1; then, connect the opposite end of the cable to port 1 of node 2.
   b. Connect one end of a ScaleXpander cable to port 2 on node 1; then, connect the opposite end of the cable to port 1 of node 3.
   c. Connect one end of a ScaleXpander cable to port 2 on node 2; then, connect the opposite end of the cable to port 2 of node 3.
3. Route the cables through the enterprise cable management arms as follows (refer to Figure 4-14 on page 214, ensuring that in each case, you have

enough spare length of the cable between the rear side of the server and the first hanger bracket, and on the front side of the cable management arm:

a. Route the cable connected on port 1 of node 1 through the hanger brackets on the cable management arms.

b. Route the cable connected on port 2 of node 1 through the hanger brackets on the cable management arms.

c. Route the cable connected on port 2 of node 2 through the hanger brackets on the cable management arms.

4. Ensure each server can be pulled out fully at the front of the rack and the scalability cables are clearly routed.

4.6.3 Four-node configuration

A four-node configuration requires five 3.08 m (9.8 feet) and one 3.26m (10.7 feet) ScaleXpander cables. The cable diagram is shown in Figure 4-16 on page 217.

One 3.0 m cable is included with every the x3950 M2 or the ScaleXpander Option Kit (for upgraded x3850 M2s). The fifth longer (3.26 m) cable is included in the IBM Scalability Upgrade Option 2 Kit. These kits are described in Table 4-1 on page 202.

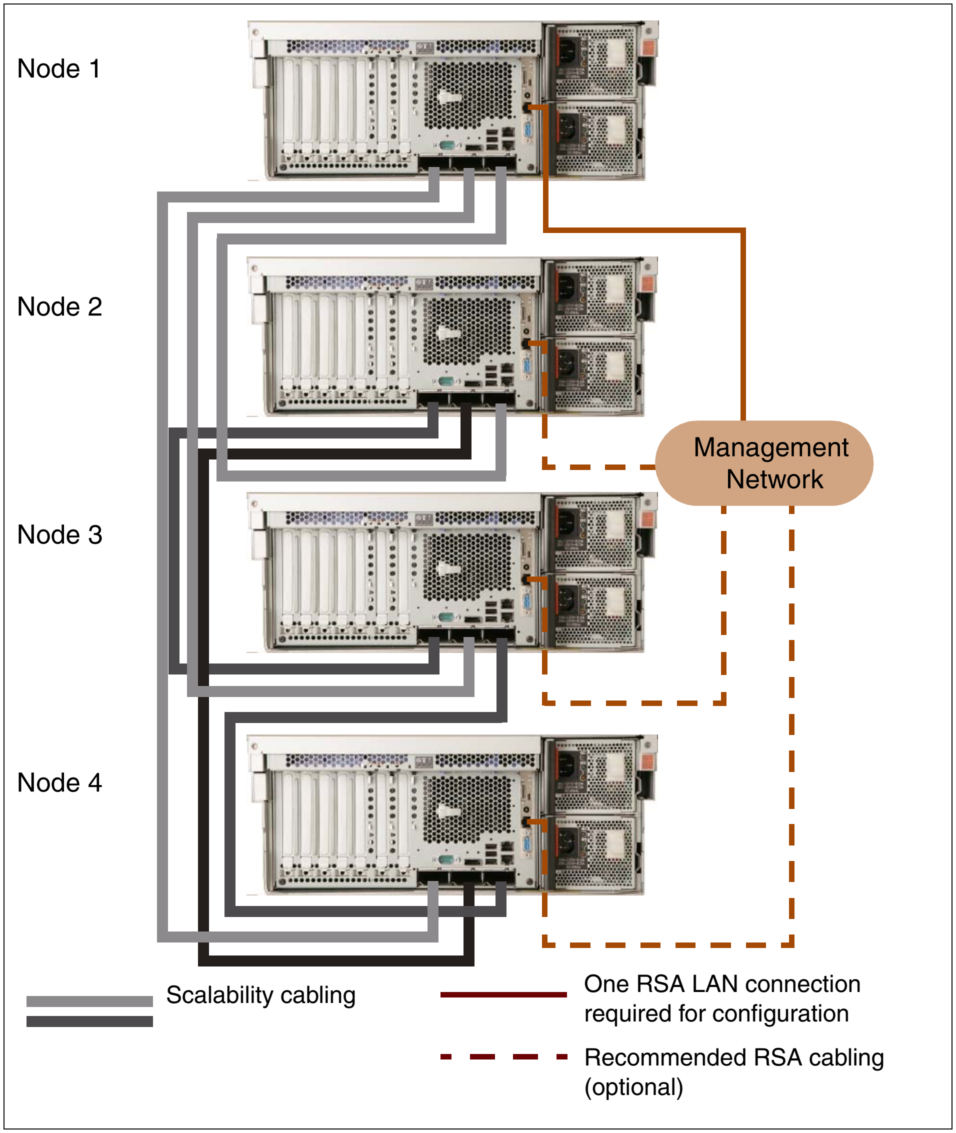

Node 1
Node 2
Node 3
Node 4
Management Network
Scalability cabling
One RSA LAN connection required for configuration
Recommended RSA cabling (optional)

Figure 4-16 Four-node x3950 M2: ScaleXpander cable port assignment

The scalability cables required for a four-node configuration are listed in Table 4-6 on page 218.

Table 4-6 Scalability cables for four-node configuration

Scalability cable	Connection
3.26m cable (longer)	Node 1: scalability port 1 → node4: scalability port 1
3.0 m cable	Node 1: scalability port 2 → node3: scalability port 2
3.0 m cable	Node 1: scalability port 3 → node2: scalability port 3
3.0 m cable	Node 2: scalability port 1 → node3: scalability port 1
3.0 m cable	Node 2: scalability port 2 → node4: scalability port 2
3.0 m cable	Node 3: scalability port 3 → node4: scalability port 3

To cable a four-node configuration for up to 16-socket operation:

1. Label each end of each ScaleXpander cable according to where it will be connected to each server.
2. Connect the ScaleXpander cables to node1:
   a. Connect one end of the long ScaleXpander cable to port 1 on node 1; then, connect the opposite end of the cable to port 1 of node 4.
   b. Connect one end of a short ScaleXpander cable to port 2 on node 1; then, connect the opposite end of the cable to port 2 of node 3.
   c. Connect one end of a short ScaleXpander cable to port3 on node1; then, connect the opposite end of the cable to port 2 of node 3.
3. Route the cables through the enterprise cable management arms as follows (refer to Figure 4-14 on page 214), ensuring that in each case, you have enough spare length of the cable between the rear side of the server and the first hanger bracket, and on the front side of the cable management arm:
   a. Route the cable connected on port 1 of node 1 through the hanger brackets on the cable management arms.
   b. Route the cable connected on port 2 of node 1 through the hanger brackets on the cable management arms.
   c. Route the cable connected on port 2 of node 2 through the hanger brackets on the cable management arms.
4. Connect the ScaleXpander cables to node 2:
   a. Connect one end of the long ScaleXpander cable to port 1 on node 2; then, connect the opposite end of the cable to port 1 of node 3.
   b. Connect one end of a short ScaleXpander cable to port 2 on node 2; then, connect the opposite end of the cable to port 2 of node 4.

5. Connect the ScaleXpander cables to node 3:
   a. Connect one end of the long ScaleXpander cable to port 3 on node 3; then, connect the opposite end of the cable to port 3 of node 4.
6. Repeat step 3 on page 218.
7. Ensure each server can be pulled out fully at the front of the rack and the scalability cables are clear routed.

Figure 4-17 shows the completed cabling.

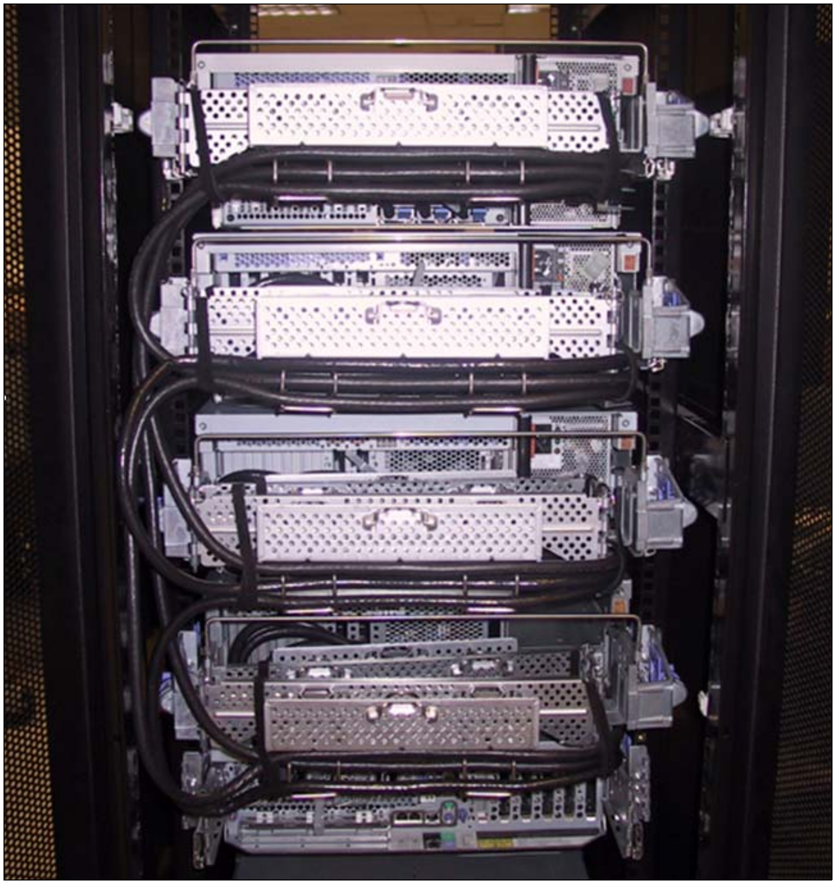


Figure 4-17 x3950 M2: four-node scalability cabling

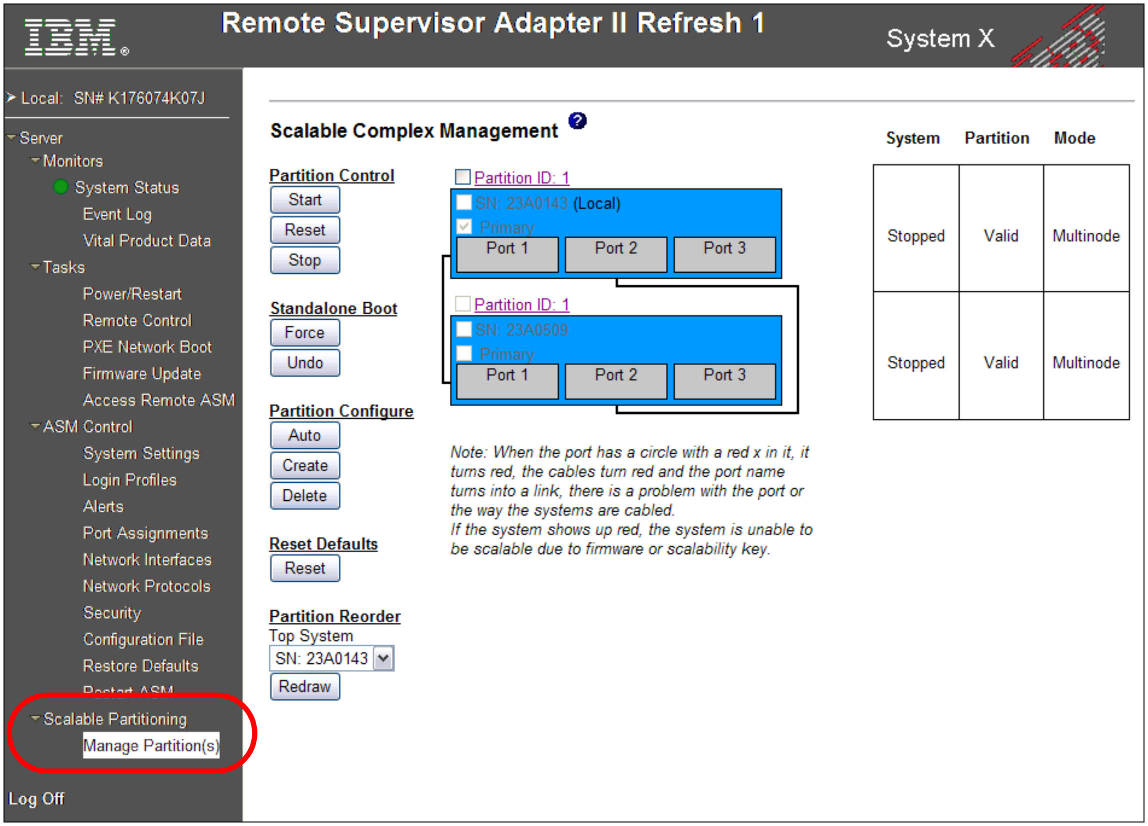
The cabling of the scalability cables is now finished. Read the next section to create a partition.

4.7 Configuring partitions

In this section, we describe how to create scalable partitions by first discussing the options in the scalability management menu.

4.7.1 Understanding the Scalable Partitioning menu

The Remote Supervisor Adapter II (RSA II) Web interface includes panels to configure scalable partitions, as shown in Figure 4-18.



System	Partition	Mode
Stopped	Valid	Multinode
Stopped	Valid	Multinode

Note: When the port has a circle with a red x in it, it turns red, the cables turn red and the port name turns into a link, there is a problem with the port or the way the systems are cabled.
If the system shows up red, the system is unable to be scalable due to firmware or scalability key.

Figure 4-18 RSA II Web interface with enabled scalability features

The **Manage Partition(s)** option lists a variety of functions in the Scalable Complex Management panel, shown in Figure 4-18 on page 220. The following buttons enable you to configure, monitor, and manage scalability partitions:

- Partition Control buttons are shown in Figure 4-19.

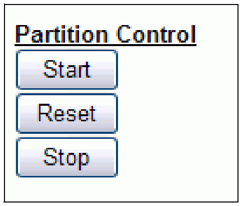


Figure 4-19 RSA II Web interface: Partition Control buttons

The buttons provide the following functions:

- Start partition

  The **Start** button powers on a system or partition. To start a partition, select the serial number (SN) or a partition ID (PID) check box, and then click **Start**.

  Notes:

  - Single systems behave as a partition of one.
  - Multiple systems and partitions can be controlled at the same time.
  - If the systems are in a partition, the partitions must be shut down first to make sure that these will merge.

- Reset partition

  The **Reset** button reboots a system or partition that is currently powered on. If the system is off, the command is ignored. To reset the partition, select the SN or a PID check box, then click **Reset**.

  Notes:

  - Single systems behave as a partition of one.
  - Multiple systems and partitions can be controlled at the same time.

- Stop partition

  The **Stop** button shuts down a system or partition that is currently powered on. If the system is off, the command is ignored. To stop a partition, select the SN or a PID check box, and click **Stop**.

  Notes:

  - Single systems behave as a partition of one.
  - Multiple systems and partitions can be controlled at the same time.

- Standalone Boot buttons are shown in Figure 4-20.

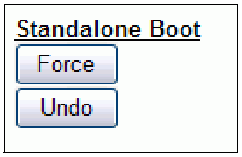


Figure 4-20 RSA II Web interface: Standalone Boot buttons

  The Standalone Boot buttons can either force a partition out of stand-alone mode, or put back in stand-alone mode, as follows:

  – Stand-alone mode

    The **Force** button resets a multinode system and boots a selected system in stand-alone mode. First select the PID of the partition you want to control, and then click **Force**.

    Notes:

    - The stand-alone mode allows the systems to boot separately for debugging or diagnostic purposes.
    - If the system is in this mode already, the command is ignored.

  – Multinode mode

    A system in a partition that is started in stand-alone mode can be reassigned to boot in the multinode mode again. Select the SN (serial number), then click **Undo**.

    Notes:

    - One or more systems that are members in a partition and booted in stand-alone boot, swap back to the boot mode multinode.
    - If the system is in multinode mode already, this command is ignored.

- Partitions Configure buttons are shown in Figure 4-21.

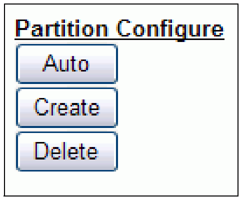


Figure 4-21 RSA II Web interface: Partition Configure buttons (create partitions)

The buttons provide the following functions:

- Auto partition

  The **Auto** button puts all systems in the complex into a single partition. First, power off all systems. Any pre-existing partitions are deleted during power-off. Select a single primary system, then click **Auto** to create one partition with all systems.

- Create partition

  The **Create** button forces systems in the complex to behave as a single entity. First, power off all systems. Select the SN check boxes of each system you would like in this partition, select the primary system, then click **Create**. There can only be one primary in a partition.

- Delete partition

  The **Delete** button removes systems in the complex from a partition. To delete a partition, first select the PID of the partition that you want to delete and then click **Delete**.

  - If the systems are powered on, they need to be powered off first.

- Reset Defaults:

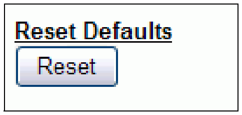
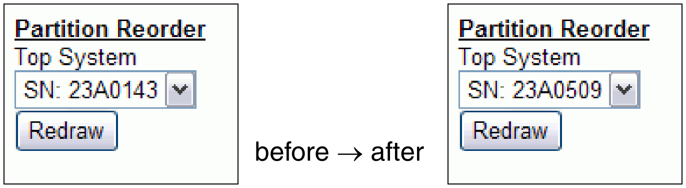


Figure 4-22 RSA II Web interface: Reset Defaults (reset the partition information)

The Reset button clears all the partition information in the complex. Use this function when data is corrupted and a system becomes stuck in an invalid partition where the firmware cannot detect it. All partition information from all systems is cleared. Click the **Reset** button to delete the partition information of all systems.

- Partition Reorder functions redraws the boxes and serials numbers, as shown in the before and after reordering in Figure 4-23.



Top System
SN: 23A0143

Top System
SN: 23A0509

Figure 4-23 RSA II Web interface: Partition Reorder (redrawing)

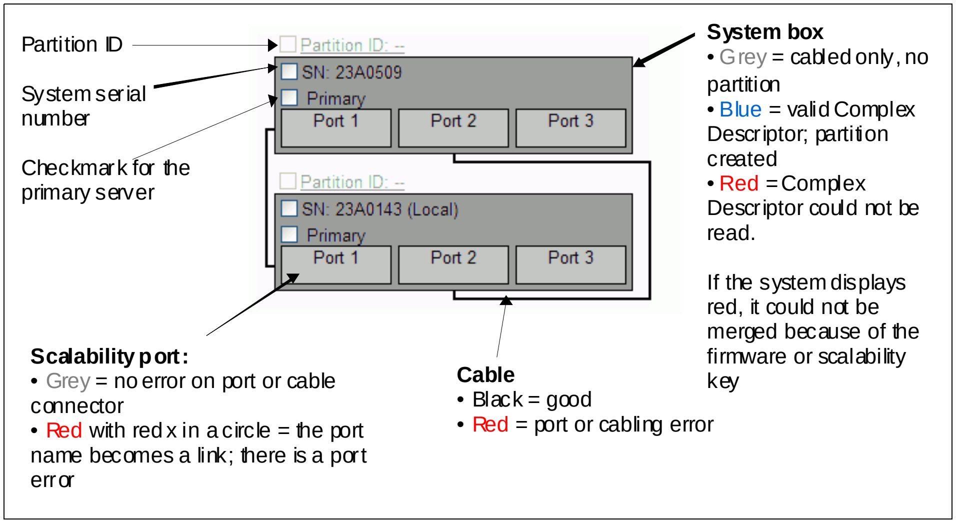
The **Redraw** button allows you to select the system that should be displayed first to match the cabled configuration. To redraw the boxes and the serial numbers, first select the serial number you want to be at the top of the list and then click **Redraw**.

Notes:

- The box on which you are connected is always described as (Local), the Primary system is always marked by a check mark.
- The sequence of the boxes in the middle section of the panel can change.

This middle section of the Scalable Complex Management section, in Figure 4-18 on page 220, shows you the health of your partitioned systems. It provides an overview of the following information, as indicated in Figure 4-24:

- Scalability ports
- Scalability cable connection
- Complex Descriptors
- Firmware problems
- Complex and Partition overview



System box
- Grey = cabled only, no partition
- Blue = valid Complex Descriptor; partition created
- Red = Complex Descriptor could not be read.

If the system displays red, it could not be merged because of the firmware or scalability key

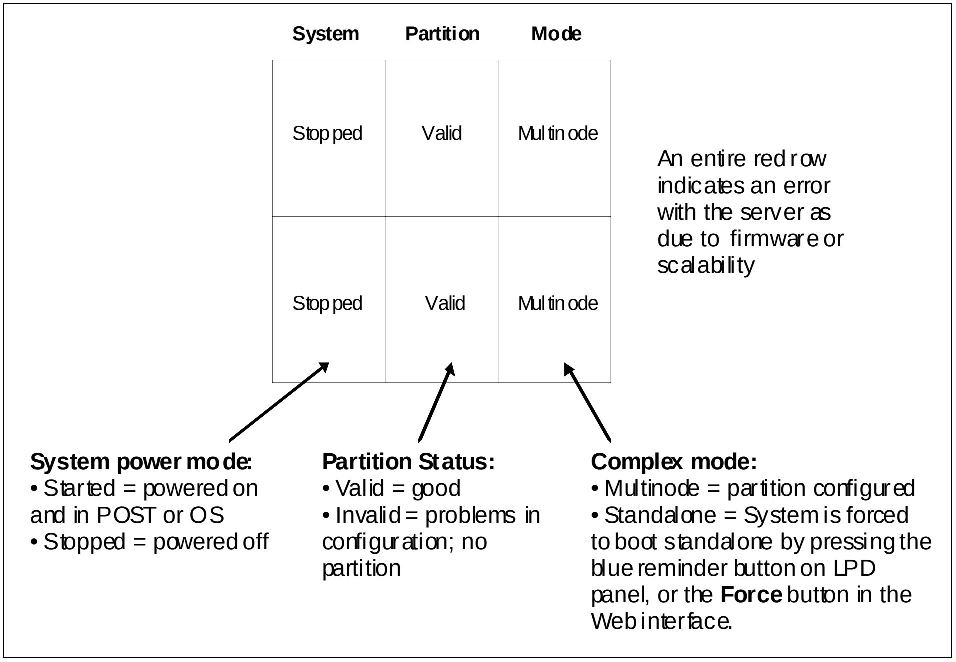
Scalability port:
- Grey = no error on port or cable connector
- Red with red x in a circle = the port name becomes a link; there is a port error

Cable
- Black = good
- Red = port or cabling error

Figure 4-24 RSA II Web interface: complex health

The section on the right side in Figure 4-18 on page 220 (and also shown in Figure 4-25 on page 225) indicates:

- System power state
- Partition state
- Scalability mode



System	Partition	Mode
Stopped	Valid	Multinode
Stopped	Valid	Multinode

An entire red row indicates an error with the server as due to firmware or scalability

System power mode:
• Started = powered on and in POST or OS
• Stopped = powered off

Partition Status:
• Valid = good
• Invalid = problems in configuration; no partition

Complex mode:
• Multinode = partition configured
• Standalone = System is forced to boot standalone by pressing the blue reminder button on LPD panel, or the **Force** button in the Web interface.

Figure 4-25 RSA II Web interface: complex status

4.7.2 First steps in configuring the partition

Before configuring the partition:

1. Open the RSA II Web interface of each node in a browser.
2. Go to:

   Scalable Partitions → **Manage Partitions** → **Scalability Management**.

3. Confirm that all nodes are cabled correctly and that no errors exist, as follows:

   a. In the middle section, which shows the cabling (see Figure 4-24 on page 224), if you see any red-marked cable, check that you followed the cabling order in 4.6, “Cabling of multinode configurations” on page 209.

   b. Confirm the boxes representing each node are colored gray.

      If any box representing a scalability port is red or has a red circle, this indicates an error. You must resolve the error before you can configure.

> **Tips:** If the Web page indicates a cable error, ensure that all cables are seated correctly. After you connect the cables, pull on the cable ends to confirm that the connectors are fixed in the scalability cable slots.
>
> If a behavior is unexpected, such as a red indicator on the scalability ports, an unreadable partition descriptor, or other error condition that is unexpected, our tests showed that an AC power cycle of all nodes in the complex may help to return the complex to a healthy state.

4. Check the system and partition states on the right side of the Scalable Complex Management panel (shown in Figure 4-18 on page 220 and Figure 4-25 on page 225). The states should indicate:

   – Each system is `Stopped`.

   – The partition state at each node is `Invalid` and is colored *black*.

   > **Note:** The state `Invalid` also indicates a non-partitioned system.

   If any node is in the state `Invalid` or is colored *red*, an error occurred.

   > **Note:** An `Invalid` system that is colored *red* shows you *red* colored symbolic boxes in the middle section too.

5. Identify the server that is the local system.

   > **Tip:** The serial number that is highlighted as `Local` system, as shown in Figure 4-24 on page 224, is the physical node that you are currently accessing by using the RSA II Web interface.

4.7.3 Creating partitions

This section describes how to create the first partition.

> **Tip:** We recommend that you reorder the systems in the configuration window to match the physical installation in your rack. See Partition Reorder in 4.7.1, “Understanding the Scalable Partitioning menu” on page 220 for details.

To create a partition, use one of the following methods:

- Automatic partitioning (autopartitioning)

  Autopartitioning generates a partition, without your intervention, by using all nodes in the complex. You assign a designated server as the *primary* system and then click **Auto**.

  The default settings are:

  – Partition merge timeout: 6 minutes
  – On merge failure, attempt partial merge?: Yes
  – Partial merge allowed: Yes

- Manual partitioning

  This allows you to assign the servers you want to have in one partition and to set up your complex to at least:

  – One 2-node partition
  – One 3-node partition
  – One 4-node partition
  – Two 2-node partitions (a second Partition ID is generated)

  Select the nodes you want to put in one partition and assign the box that is the primary system. Then, click **Create**.

  The default settings are the same as in autopartitioning:

  – Partition merge timeout: 6 minutes
  – On merge failure, attempt partial merge?: Yes
  – Partial merge allowed: Yes

  To review and change the settings for the partition, click the **Partition ID: n** link (where **n** is the number of the partition) as shown in Figure 4-26 on page 228. This can only be done at a valid partition.

  If no error occurs, the view in the middle and right sections change as follows:

  – The color for the server boxes change from *grey* to *light blue*.

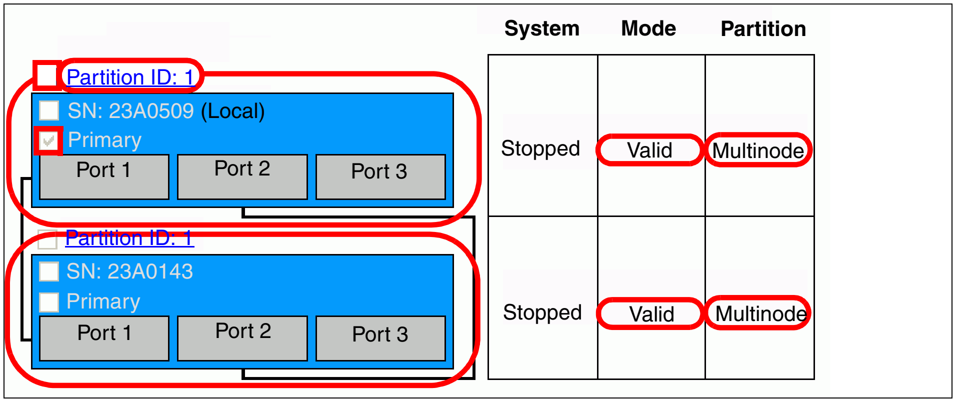
  – The scalability mode at all assigned nodes changes from `Standalone` to `Multinode`.

  – The partition state changes from `Invalid` to `Valid`.

  – The partition ID above the boxes becomes a selectable *soft link.* which takes you to the partition settings for you to review.

  – The check box on the top of the primary node is selectable,

  – You see a check mark in the primary's symbolic box



System	Mode	Partition
Stopped	Valid	Multinode
Stopped	Valid	Multinode

Figure 4-26 RSA II Web interface: two-node partition configuration

You have completed the step to create partitions.

If you want to create a second partition in this complex, repeat the steps. You may create partitions that are comprised of two, three, or four nodes.

4.8 Working with partitions

This section describes how to work with and manage your partitions, and the various management and control features.

4.8.1 Managing partitions

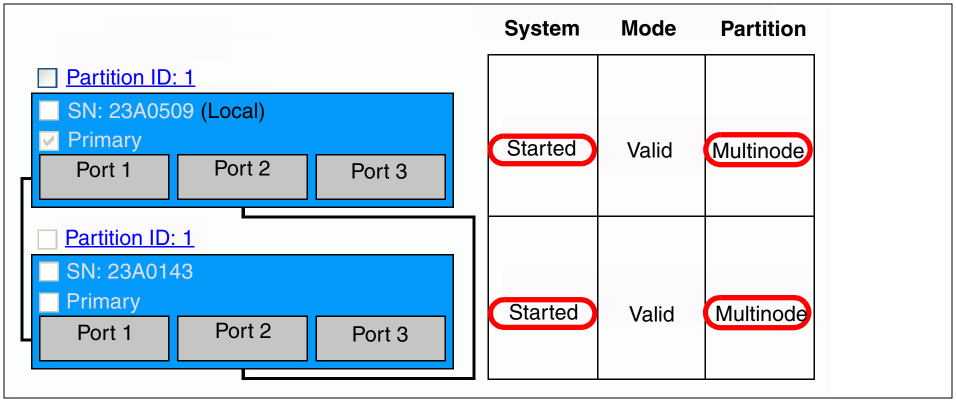
You can start, reset, power off, and delete partitions. You can also swap between Standalone and Multinode mode, and clear partition information in a complex.

Start a partition

All nodes in the partition power on immediately, after a partition is started.

To start a partition, select the Partition ID (PID) at the top of the blue symbolic box at the primary node. Click **Start** (under Partition Control). The following events occur, as shown in Figure 4-27 on page 229:

- The selected partition starts immediately.
- The system’s state changes from *Stopped* to *Started*.



□ Partition ID: 1

□ SN: 23A0509 (Local)

☑ Primary

Port 1 | Port 2 | Port 3

□ Partition ID: 1

□ SN: 23A0143

□ Primary

Port 1 | Port 2 | Port 3

System	Mode	Partition
Started	Valid	Multinode
Started	Valid	Multinode

Figure 4-27 RSA II Web interface: two-node partition started

The system starts the POST. In the POST you can watch the merge process on the screens of all nodes in this partition. The process indicates:

- The primary node searches for the secondary node, shown in Example 4-1.

Example 4-1 x3950 M2: primary server merge process

```
IBM BIOS - (c) Copyright IBM Cooperation 2008
Symmetric Multiprocessing System
Quad-Core Intel Xeon MP ~2.93GHz

Searching for secondary server
Press BLUE Remind button to bypass partition merge and boot standalone
```

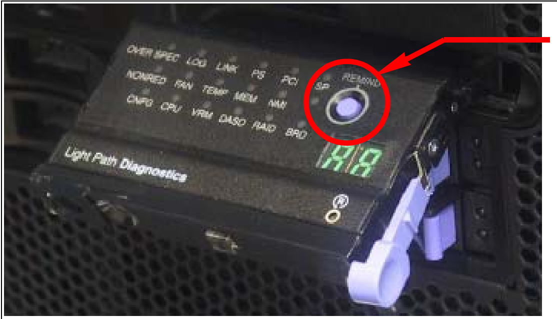
- The secondary node searches for the primary node.
- The primary node searches for the secondary node, shown in Example 4-2.

Example 4-2 x3950 M2: secondary server merge process

```
IBM BIOS - (c) Copyright IBM Cooperation 2008
Symmetric Multiprocessing System
Quad-Core Intel Xeon MP ~2.93GHz

Searching for primary server
Press BLUE Remind button to bypass partition merge and boot standalone
```

The merging process can bypassed by pressing the blue **REMIND** button located on the Light Path Diagnostics panel, shown in Figure 4-28 on page 230.

Pressing the button breaks this node from the merging process. The remaining nodes merge as normal, if partial merge is enabled.



Blue REMIND button

Figure 4-28 x3950 M2: Light Path Diagnostics panel with blue REMIND button

- After the boxes merged successful you can see the following information about the window on the different boxes:
  - Primary server display, shown in Example 4-3.

Example 4-3 x3950 M2: merging process successful, Primary server

```
Chassis Number       Partition Merge Status       Installed Memory
       1             Primary                       32GB
       2             Merged                        32GB



Partition merge successful

64 GB Memory:  Installed
512 MB Memory:  Consumed by Scalability



Press ESC to reboot and bypass merge attempt on next boot
Press F1 for Setup
Press F2 for Preboot Diagnostics (DSA)
Press F12 to select boot devices
```

  - Secondary server display, shown in Example 4-4.

Example 4-4 x3950 M2: merging process successful, Secondary server

```
Merge complete - see primary server display
```

- The *white* scalability LED on the front of all nodes, which merged successfully to this partition, becomes solid *on*. This LED goes *off* again after a system in this partition is swapped to the Standalone mode, followed by a reboot.

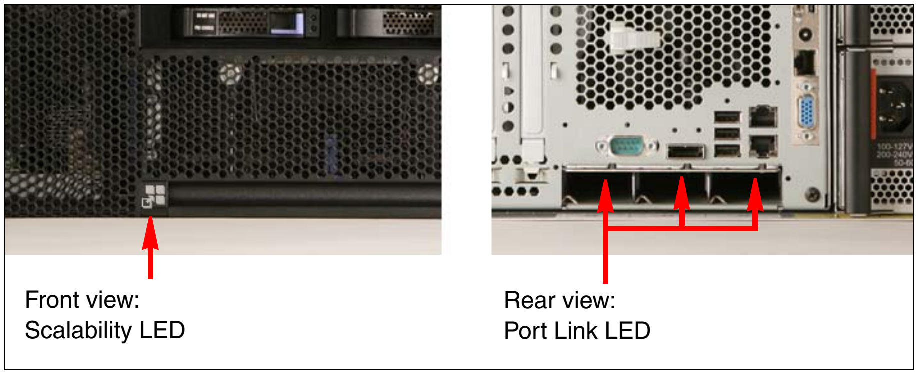


Figure 4-29 x3950 M2: Scalability LED and port link LED

- The scalability port LED becomes solid *on* in *green* at the ports where a scalability connection has been established.

Reset a partition:

To reset a partition, select the Partition ID you want to reset. Then, click **Reset** (under Partition Control). The following events occur:

- The Multinode resets.
- A running operating system resets also.

Power off a partition

To power off a partition, select the Partition ID you want to power off. Then click **Stop** (under Partition Control). The following events occur:

- All nodes that are configured in this partition are powered off.
- The System state changes back to *Stopped*.

Swap between stand-alone and multinode mode

Depending on the partition, you can change the scalability modes:

- To switch all nodes in the partition to stand-alone mode, select the Partition ID (PID) and then click **Force** (under Partition Control). The systems reset immediately and boot in stand-alone mode.
- To switch the system back to multinode mode, select the PID and click **Undo**.

You can also interrupt nodes *after* they are merged by pressing the `Esc` key on your keyboard in POST. The nodes reset and bypass to merge on next boot. See Example 4-5 on page 232.

Example 4-5   POST: special keystrokes

```
Press ESC to reboot and bypass merge attempt on next boot
Press F1 for Setup
Press F2 for Preboot Diagnostics (DSA)
Press F12 to select boot device
```

Delete a partition

To change your partitions or add another system, clear this partition first and then configure your new partition. Select the Partition ID, power off the systems in this partition, and then click the **Delete** button.

Clear partition information in a complex

To delete all partition information in a complex, power off all systems in the complex, then click **Reset** (under Reset Defaults) to get the defaults.

Note: A partition or all partition settings cannot be deleted if any of the system is powered on.

4.8.2 Behaviors of scalability configurations

The typical behaviors you might observe when working with systems in a multinode complex are described in this section.

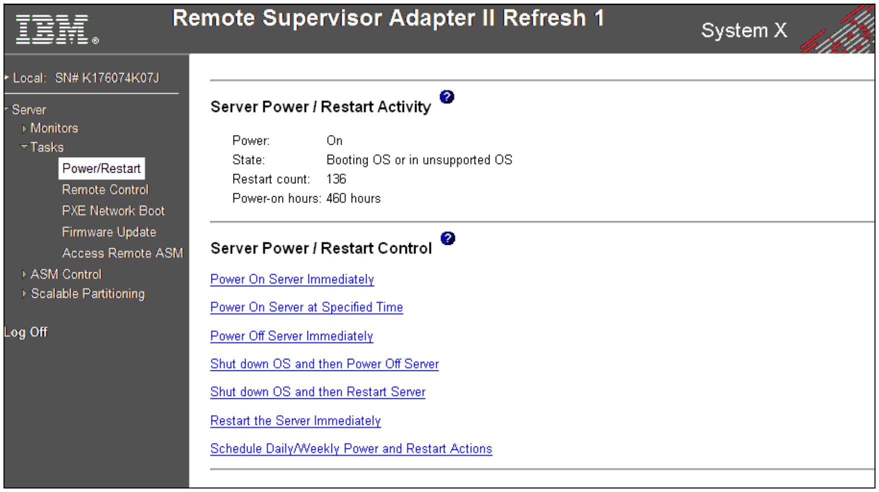
First, however, review the following checklist again if you experience difficulties creating partitions:

- Did you upgrade each x3850 M2 system to a x3950 M2 system to be capable for scalability configuration as described in 4.5, “Upgrading an x3850 M2 to an x3950 M2” on page 204?
- Does each system match the prerequisites as described in 4.4, “Prerequisites to create a multinode complex” on page 201?
- Did you provide the correct cabling at all nodes depending on the number of nodes in your complex as described in 4.6, “Cabling of multinode configurations” on page 209?
- Are you familiar with the Scalability Manager in the Remote Supervisor Adapter II as described in 4.7, “Configuring partitions” on page 220?

Add a x3950 M2 server system to a complex

After the server is installed in the rack and cabled with the scalability cables of each system, you can observe the following behavior:

- The system can be powered on by one of the following methods:
  - Use the power (on/off) button on the front of the server on the Light Path Diagnostics panel, see Figure 4-28 on page 230.
  - Use the power (on/off) features in the Remote Supervisor Adapter II Web interface, as shown in Figure 4-30.


IBM
Remote Supervisor Adapter II Refresh 1
System X
Local: SN# K176074K07J
Server
Monitors
Tasks
Power/Restart
Remote Control
PXE Network Boot
Firmware Update
Access Remote ASM
ASM Control
Scalable Partitioning
Log Off
Power On Server Immediately
Power On Server at Specified Time
Power Off Server Immediately
Shut down OS and then Power Off Server
Shut down OS and then Restart Server
Restart the Server Immediately
Schedule Daily/Weekly Power and Restart Actions

Server Power / Restart Activity

Power: On
State: Booting OS or in unsupported OS
Restart count: 136
Power-on hours: 460 hours

Server Power / Restart Control

Figure 4-30 RSA II Web interface: Power/Restart options

- This system behaves as a single-server system as follows:
  - When it is not configured as a member in a multinode partition.
  - All changes of BIOS settings in BIOS affect this particular node only.
  - If it is part of a partition, but is bypassed during the merging process when you press the blue **REMIND** button, shown in Figure 4-28 on page 230.
  - If its part of a partition, but is forced to boot a stand-alone node.
- The scalability management menu in the Remote Supervisor Adapter II Web interface (see Figure 4-4 on page 204), which is installed in each server in this complex, can see each other system by reading the partition descriptor information stored in the BMC. It also contains a scalability cabling diagram as described in 4.7.1, “Understanding the Scalable Partitioning menu” on page 220.

Access the RSA II Web interface

The Remote Supervisor Adapter II (RSA II) is configured and can be accessed through a browser. It has the following behaviors:

- Each RSA II in any of the servers in a complex, has the scalability management menu enabled in the Web interface.
- The scalability management drawing in the RSA II in each node in this complex contains the same complex and partition information.
- You can use the RSA II that is installed in any system in this complex to configure and manage partitions, provided all RSA IIs are configured and available in the network.
- To determine which system you are accessing, check the Scalability Manager panel in the Remote Supervisor Adapter II Web interface and look for the node that is flagged with `Local`.

Create a multinode configuration

After you create a scalability partition to get a multinode system, you can observe the following behaviors:

- All nodes of a partition in a complex can be managed as one system.
- All nodes of a partition in a complex can be managed as single systems. If you work remotely, you must be able to access each RSA II.

Start a multinode system

If you start a multinode system (after creating the required first partition), after all nodes in a partition are powered on, you can observe the behaviors described in this section.

Use any of the following methods to power on the multinode:

- Use the power on/off button on the front of the server on the Light Path Diagnostic panel, shown in Figure 4-28 on page 230.
- Use the power on/off option in the RSA II Web interface, as shown in Figure 4-30 on page 233.
- Use the power on option in the scalability management menu of the RSA II Web interface. See “Start a partition” on page 228.

Observe the following behaviors:

- You can watch the merging process after a partition is started in multinode mode on the window of each member in this partition, see “Start a partition” on page 228.

- In the merging process:
  - The primary system scans for resources of all merged nodes in the installed processor packages and memory, and then indicates how much of it is available.
  - The main memory is decreased by 256 MB for each merged system.
- After a merge is successful:
  - The video window on the primary system is active and contains all information in POST or in the operating system.
  - The video window at all other nodes is active and tells you that you can watch the window on the primary node.
  - The white scalability LED on the front of all systems that merged to a partition as part of a multinode system becomes solid *on*. See Figure 4-29 on page 231.
  - The port link LEDs on the rear of all systems that merged to a partition becomes solid *on* green. See Figure 4-29 on page 231.
  - The scalability LED at the front and the port link LEDs at the rear of the systems go off again, if the scalability mode is changed after a partition is started and rebooted to start in stand-alone mode.
- If a merge is unsuccessful:
  - The node or nodes that could not be merged boot in stand-alone mode.
  - All nodes that could be merged boot as multinode system if the *partial merge* flag is enabled.
  - If the scalability LED at the front and the port link LEDs of a system are off, then this system is not part of a scalability partition or has not merged to a partition.
  - The system event log in the RSA II and Baseboard Management Controller report the timeout and additional error events, at least on the particular affected node.
- The POST process and the loading of the operating system are shown on the window only at the primary system, as follows:
  - PCIe devices that are integrated or installed in the other systems at this partition are visible on the window on the primary server.
  - The started operating system provides all resources at this partition as though it is one system.

Reset a multinode system

If a reset is performed on all nodes in the partition (a partition is created) reboot, you can various behaviors.

First, to reset a multinode system, use one of the following methods:

- Use the Reset button on the top of the Light Path Diagnostic panel at any node in this partition.
- Use the reset options in the Power/Restart control of the RSA II Web interface, which is installed in any of the x3950 M2 systems. See Figure 4-30 on page 233.
- Use the reset option in the scalability management menu of the RSA II Web interface. See "Reset a partition:" on page 231.
- Perform a reset in the operating system.

The following behaviors occur if an automatic unexpected reboot (a software or hardware error occurred) occurred without your intervention:

- All nodes in a partition are affected.
- The event log of the operating system, RSA II and the Baseboard Management Controller might contain logged events.
- The x3950 M2 embedded error correction features try to recover a hardware error and boot up the partition
- The operating system attempts to reboot.

Power off a Multinode system

Powering off a started multinode system affects all nodes in a partition.

To power off a running Multinode system, use any of the following methods:

- Use the power button on the front of the Light Path Diagnostic panel at any of the merged nodes in this partition. See Figure 4-28 on page 230
- Use the reset options in the Power/Restart control of the RSA II Web interface, which is installed in any of the x3950 M2 systems. See Figure 4-30 on page 233.
- Use the reset option in the scalability management menu of the RSA II Web interface. See "Power off a partition" on page 231.
- Perform a shutdown in the operating system.

A shutdown is generated automatically in any form of abnormal fatal conditions, such running at temperatures that are too high, specification overages, hardware malfunctions, or defects, and not recoverable errors.

4.9 Observations with scalability configurations

This section highlights the experiences we gained when we tested and worked with the servers in our lab. You should regularly check the IBM Support Web page for tips, newly released documentation, and system firmware updates at:

http://www.ibm.com/systems/support/x

4.9.1 Problem with merging if prerequisites were met

If you have problems forming a multinode complex but you have met all prerequisites as described in 4.4, “Prerequisites to create a multinode complex” on page 201, check the error indicators in the Scalable Complex Management health page as shown in Figure 4-24 on page 224. This section provides additional guidance.

Problem: Members of partition not found

During the process of merging, a successful merge is not completed. The members in this partition do not find each other and cannot find a member to merge.

The systems cannot be bypassed for merging to boot in stand-alone mode. This cannot be done remotely if the merge process is not completed.

If the RSA II scalability management menu does not enable you to manage any partition, restart or power off the systems; the system can be recovered only by an AC power cycle.

> **Note:** You can bypass the boot to stand-alone remote if the scalability management interface is unresponsive only by a support person who is on-site, who must push the blue **REMIND** button.
>
> No function in the remote control interface simulates the **REMIND** button.

How to identify the problem

The RSA Log can contain the following events:

- Primary node:

```
E Merge Failure-No secondary servers found to merge.
E Merge Failure-Timed out waiting for secondary server.
I System Complex Powered Up
```

- Secondary nodes:

```
E Merge Failure-Timeout occurred waiting for primary server.
I System Complex Powered Up
```

Ensure that no hardware error occurred. See Figure 4-31. In the RSA II Web interface, select **Monitors** → **Event Log** at each node in this partition.

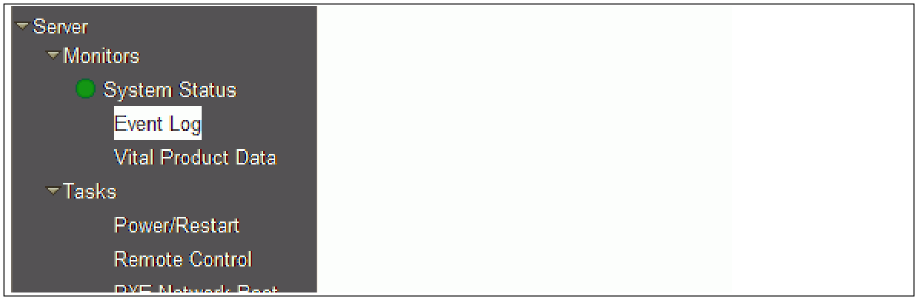
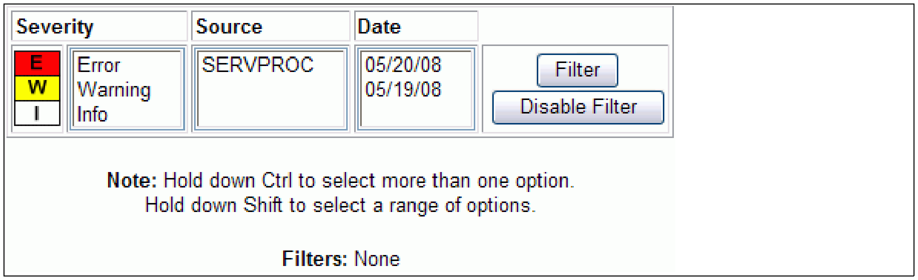


Figure 4-31 RSA II Web interface - Event log selection

Check the Severity column to determine if any events are colored with red or yellow at this specific time frame (see Figure 4-32). You can filter the events at Severity, Source, and Date.



Note: Hold down Ctrl to select more than one option.
Hold down Shift to select a range of options.

Filters: None

Figure 4-32 RSA II Web interface: event log filtering

How to resolve the problem

To resolve the problem, try the following steps:

1. Power off both systems and start the multinode again. If the scalability management menu is unresponsive, perform a power off at all systems in this partition in the power options of the RSA II Web interface. See Figure 4-30 on page 233.
2. Perform an AC power cycle and wait 30 seconds before you replug the power cables. Try a merge.

Note: Bear in mind this expects a local response if you do not have access to an uninterruptible power supply that can manage the power connections to shut off/ or turn on the power to the outlets.

3. After powering off all boxes:
   - Delete the partition, as described in 4.8, “Working with partitions” on page 228. Create a new partition and try to merge.
   - Clear all Complex information as described in 4.8, “Working with partitions” on page 228. Create a new partition and try to merge.

4.9.2 Problems with merging if prerequisites were not met

This section describes situations where the merge process fails, particularly if the prerequisites listed in 4.4, “Prerequisites to create a multinode complex” on page 201 are not met.

Problem 1: CPU mismatch

The merging of nodes fails with information about the window, as shown in Example 4-6.

Example 4-6 x3950 M2: merging process unsuccessful, Primary server

```
Chassis Number      Partition Merge Status      Installed Memory
       1            Primary                      32GB
       2            Failed: CPU Mismatch


Partition merge failed: No secondary chasiss to merge successful

64 GB Memory:  Installed
512 MB Memory:  Consumed by Scalability
```

How to identify the problem

To identify the issue:

- Check that all systems in this complex have two or four CPUs installed.
- Check that all servers you added to this complex have the same type of CPU regarding speed, cache size, and number of cores. This is especially important if you ordered your servers on different dates or added a new one to an existing complex.

- Check the RSA II event logs at each node (as described in “Problem with merging if prerequisites were met” on page 237) if, for example, a CPU error occurred that is not linked to having the incorrect type of CPU.

How to resolve the problem

Install the same CPU (two or four of them) in all systems in this complex.

Problem 2: Insufficient memory in the primary node

The merging of nodes fails with information about the window, as shown in Example 4-7.

Example 4-7 x3950 M2: Merging process unsuccessful; primary server

```
IBM BIOS - (c) Copyright IBM Corporation 2008
Symmetric Multiprocessing System
Dual-Core Intel Xeion MP ~2.4 GHz

Primarynode has less than 4GB memory installed
...merge not supported


2GB Memory:  Installed
4 Processor Packages Installed


>> BIOS Version 1.03 <<
```

How to identify the problem

Verify that the primary system in this complex is added with at least one pair of DIMMs according the DIMM installation rules. This is important if a system is shipped with 2 GB DIMM.

How to resolve the problem

Install at least 4 GB of memory in the primary node to meet the prerequisites.

4.9.3 Known problems

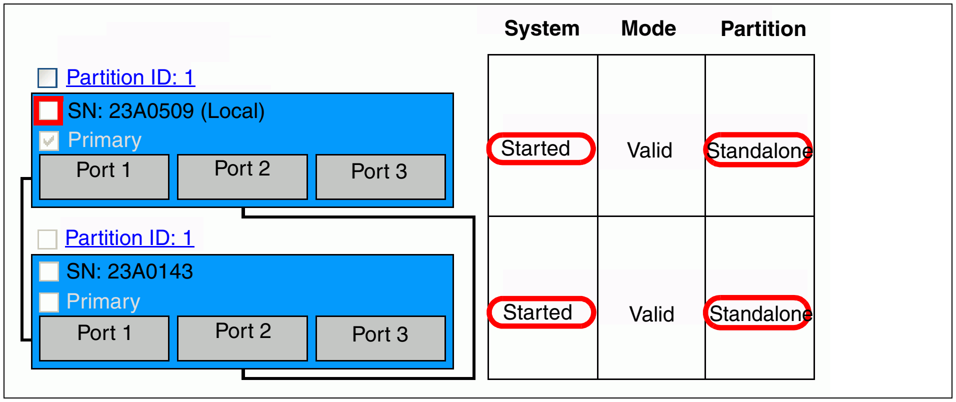
This section describes the known problems with multinode configurations and the workarounds currently available.

Problem 1: Resetting a node in a complex, unexpected reboot

A started partition is in the multinode mode and can be changed to the stand-alone mode in the scalability management menu of the RSA II (see “Swap between stand-alone and multinode mode” on page 231) to boot the systems in stand-alone mode after a reboot.

After the scalability mode is changed to stand-alone mode, the serial numbers of each system is selectable, as shown in Figure 4-33.

.



□ Partition ID: 1

□ SN: 23A0509 (Local)

☑ Primary

□ Partition ID: 1

□ SN: 23A0143

□ Primary

System	Mode	Partition
Started	Valid	Standalone
Started	Valid	Standalone

Figure 4-33 RSA II Web interface: two-node Partition configuration

If you add a check mark in one of the serial number (SN) boxes, then click **Reset**, the server immediately resets. The secondary server logs an error and reboots unexpectedly several seconds later.

How to identify the problem

Check the RSA II event logs at each node as described in 4.9.1, “Problem with merging if prerequisites were met” on page 237 if, for example, a CPU error occurred that is not linked to having the wrong type of CPU. See Example 4-8.

Example 4-8 RSA II event log

```
1 Err SERVPROC 04/16/08, 22:02:03 Resetting system due to an unrecoverable error
2 Err SERVPROC 04/16/08, 22:02:03 Machine check asserted - SPINT, North Bridge
```

Solution and workaround

This problem is reported to IBM development. Newer RSA II firmware prevents the serial number of any node from being selectable.

Do not reboot one system in stand-alone mode after the scalability mode was changed to stand-alone mode and all other systems are still running in the operating system.

Boot any of the nodes in stand-alone mode only if they are powered off, or select all of the nodes in a partition to reboot at the same time. They then reboot at the same time, without reporting the error listed in Example 4-8 on page 241.

Problem 2: Power state incorrect

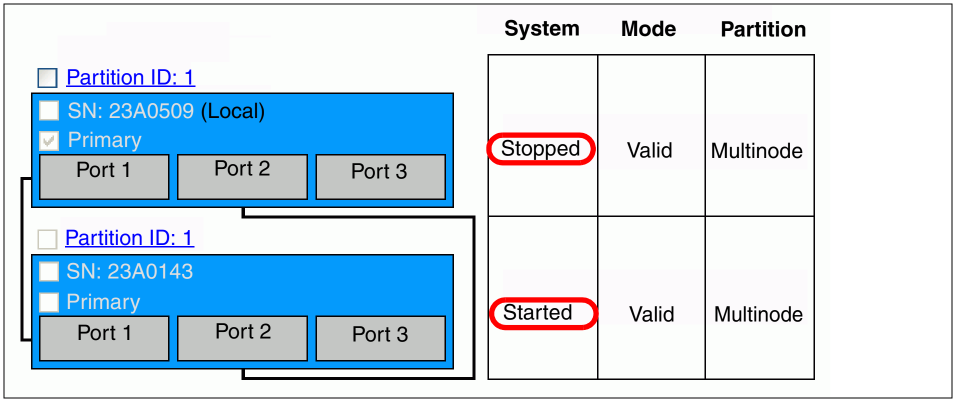
A node in a configured partition results in becoming of unresponsive to requests to control the partition by the scalability management menu.

The system power state is incorrectly shown as powered off.

How to identify the problem

In the scalability management menu, check the System power state of each system in the partition. See Figure 4-34.

.



☐ Partition ID: 1

☐ SN: 23A0509 (Local)

☑ Primary

☐ Partition ID: 1

☐ SN: 23A0143

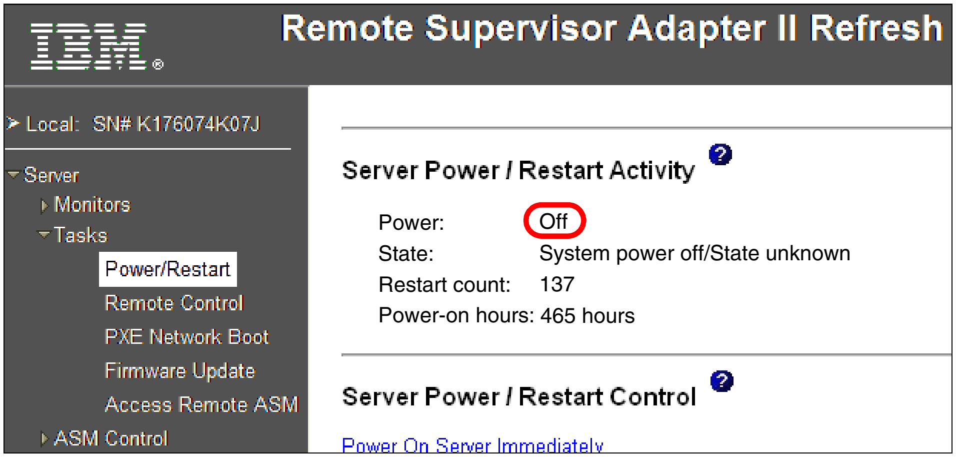
☐ Primary

System	Mode	Partition
Stopped	Valid	Multinode
Started	Valid	Multinode

Figure 4-34 RSA II Web interface: two-node Partition configuration

As Figure 4-34 shows, the first system in the list is `Stopped`, however the second system in the list is `Started`.

The power state is shown as `Off` in the Server Power/Restart Activity panel, shown in Figure 4-35 on page 243.



Server Power / Restart Activity

Power: Off
State: System power off/State unknown
Restart count: 137
Power-on hours: 465 hours

Server Power / Restart Control

Figure 4-35 RSA II Web interface: power-off state

Solution and workaround

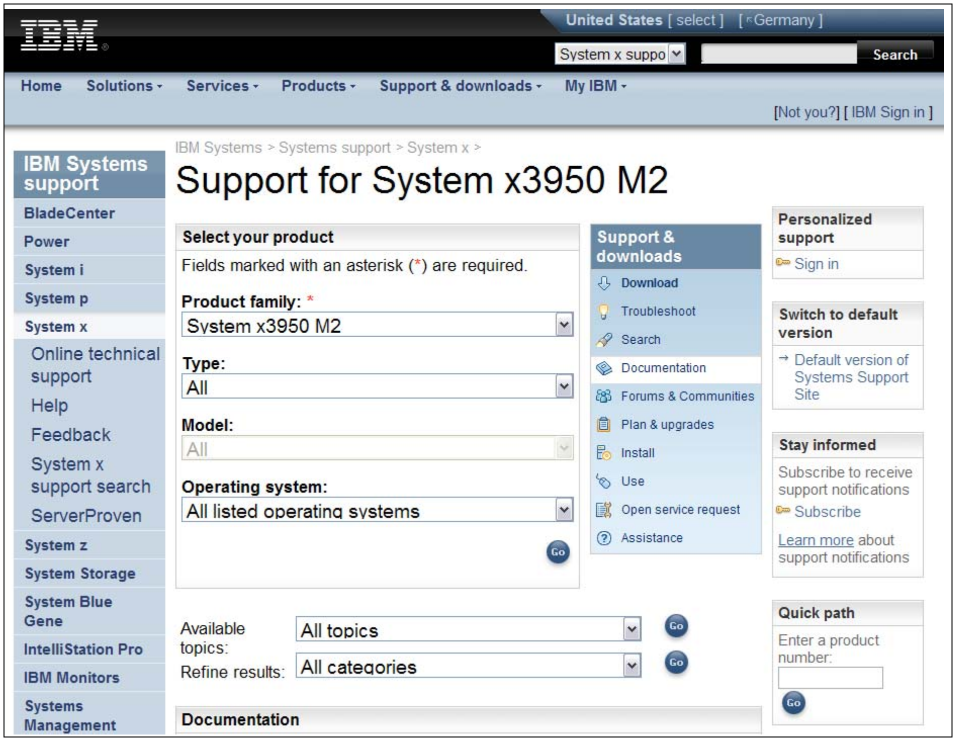
A newer BMC firmware solves the incorrect reading of the power state that is reported by the field programmable gate array (FPGA) firmware code.

You can work around this error by removing all systems in the partitions from the AC power source. Wait 30 seconds and then replug the AC power.

We recommend that you regularly check for newer system firmware code and updated product publications such as the Problem Determination and Support Guide and new tips on the Support for IBM System Support Web pages at:

`http://www.ibm.com/systems/support/x`

As shown in Figure 4-36 on page 244, select your product family x3850 M2 or x3950 M2, or the machine type to find the required information.



IBM Systems > Systems support > System x >

Support for System x3950 M2

Select your product

Fields marked with an asterisk (*) are required.

Product family: *
System x3950 M2

Type:
All

Model:
All

Operating system:
All listed operating systems

Available topics: All topics

Refine results: All categories

Documentation

Personalized support
Sign in

Switch to default version
Default version of Systems Support Site

Stay informed
Subscribe to receive support notifications
Subscribe
Learn more about support notifications

Quick path
Enter a product number:

Figure 4-36 IBM System x: technical support on the Web

5

Installation

This chapter describes the steps involved in installing and configuring supported operating systems on the IBM x3850 M2 and x3950 M2. Firmware and BIOS updates, BIOS settings and operating system support are also discussed as important preliminary tasks prior to beginning the installation of operating systems on the server.

This chapters discusses the following topics:

- 5.1, “Updating firmware and BIOS” on page 246
- 5.2, “Confirming BIOS settings” on page 254
- 5.3, “Supported operating systems” on page 257
- 5.4, “Installing the operating system” on page 264

5.1 Updating firmware and BIOS

Before updating firmware, ensure that all hardware components have been installed in the x3850 M2 or x3950 M2. Ensure that all hardware components required to upgrade the x3850 M2 to the x3950 M2 are installed including the ScaleXpander key (chip).

Information about installing all hardware components for the x3850 M2 or the x3950 M2, both as a single server and as a multinode complex, is in Chapter 3, “Hardware configuration” on page 89 and Chapter 4, “Multinode hardware configurations” on page 195.

In this section, we discuss updating the following components:

- System BIOS
- DSA Preboot diagnostics
- BMC firmware
- FPGA firmware
- Remote Supervisor Adapter II firmware
- Additional devices such as RAID controller firmware and NIC firmware

5.1.1 Prerequisite checklist

Make sure to perform the following tasks before updating the firmware and installing the operating systems:

- You have installed servers and additional components according to instructions in Chapter 3, “Hardware configuration” on page 89 and Chapter 4, “Multinode hardware configurations” on page 195.
- Check the BIOS and firmware levels on the items listed before you start installing the operating systems. Refer to 6.9, “DSA Preboot” on page 366 for details on how to check the BIOS and firmware levels of your server.
- If you plan to install multiple x3950 M2 systems in a multinode complex, update all the nodes in the complex to the same firmware levels.
- Even if the firmware is reported to be at the latest level by the update utility, we still recommend that you use the firmware update utility to re-apply the most recent IBM supported firmware level to your node. This ensures that all x3950 M2 nodes are optimally prepared for any stand-alone or multinode configuration operating system deployment.

Note: When you further update the systems in a multinode complex at a later date, you will be able to update certain BIOS and firmware on the primary node; that BIOS and firmware will then be automatically propagated to the other nodes. Be aware, however, that you will have to update FPGA, BMC and BIOS firmware levels on all nodes separately.

5.1.2 Downloading the firmware

To download the latest firmware for the servers you will be installing:

1. Navigate to the "Support for IBM System x" Web page:

   http://www.ibm.com/systems/support/x

2. In the **Product family** list, select the model of your server (for example x3850 M2 or x3950 M2).
3. In the **Type** list, select the 4-digit server machine type of your server (7141 or 7233).
4. Click **Go** to retrieve latest versions of firmware for your server.

If you have not already installed an operating system on your server, you can download the non-OS specific firmware update utilities for the components listed previously. Otherwise, there are also OS specific firmware update utilities which you can use to update your server firmware from within the operating system.

Note: If firmware for a non-specific OS is *not listed*, download one of the OS-specific utilities. It should provide a method to extract the non-OS specific firmware update utilities to a local directory. This directory should provide instructions to update the firmware of the server through bootable media such as a floppy disk image file or a bootable ISO image file.

5.1.3 Performing the updates

This section lists the sequence we followed, and the server components that require firmware updates.

To prepare the x3950 M2 servers in our labs for operating system installation, we used the following firmware update sequence:

1. Remote Supervisor Adapter II firmware (RSA II requires a firmware update specific to system x3850 M2 or x3950 M2.)
2. FPGA firmware
3. BMC firmware

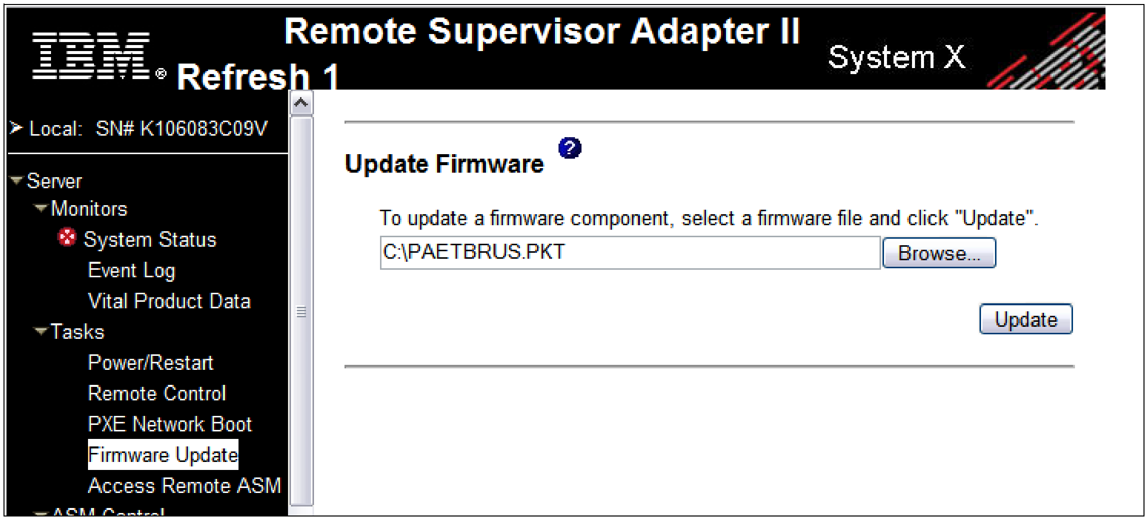
4. System BIOS firmware
5. All other firmware such as the DSA Preboot, LSI1078, ServeRAID MR10k, ServeRAID MR10M, Fibre Channel HBAs, and others.

The next sections list the server components that require firmware updates.

Remote Supervisor Adapter (RSA) II firmware

You can update the RSA firmware from its Web interface, shown in Figure 5-1. The update involves uploading two separate packet (PKT) files before you restart the adapter.

In x3950 M2 multinode configurations, you should update the RSA II firmware separately on each node preferably before forming a multinode partition.



Update Firmware

To update a firmware component, select a firmware file and click "Update".

C:\PAETBRUS.PKT

Figure 5-1 Remote Supervisor Adapter II firmware. separately on each x3950 M2 node

FPGA firmware

This update is in the form of a bootable diskette or ISO image. If you have the RSA II configured and connected to your network, you can use the Remote Console feature to mount a remote diskette IMG file or ISO file, and boot the x3850 M2 or x3950 M2 from the mounted IMG or ISO file.

On x3950 M2 multinode configurations, you can update the FPGA firmware version on all chassis from the primary node by booting the IMG or ISO file from the primary node.

At the completion of the FPGA update on the primary node, you will be prompted to press `Enter` to power off all nodes in the same partition. See Example 5-1 on page 249. The FPGA is then reloaded on all nodes in the partition, then each node is powered off and automatically powered on again (this can be 30-second delays in this sequence). Removing any power cable to activate new updated FPGA firmware is not required.

Example 5-1   Updating the FPGA code on a two-node complex

```
MNFPGA.EXE v2.60
------------------------------------
|  Node 0          0x004E / 0x004F  |
------------------------------------
Erasing the CPU Card SPI ROM
Programming CPU Card SPI ROM (with Verification)
Sector 7 <=> Page 54
CPU Card SPI ROM Programming is complete

Erasing the PCI Card SPI ROM
Programming PCI Card SPI ROM (with Verification)
Sector 4 <=> Page 54
PCI Card SPI ROM Programming is complete

MNFPGA.EXE v2.60
------------------------------------
|  Node 1          0x604E / 0x604F  |
------------------------------------

Erasing the CPU Card SPI ROM
Programming CPU Card SPI ROM (with Verification)
Sector 7 <=> Page 54
CPU Card SPI ROM Programming is complete

Erasing the PCI Card SPI ROM
Programming PCI Card SPI ROM (with Verification)
Sector 4 <=> Page 54
PCI Card SPI ROM Programming is complete

*************************************************************************
*                                                                       *
*        DC Power Cycle IS required to Reload the FPGAs                 *
*                                                                       *
*                 >>>>>  Remove the diskette  <<<<<                     *
*                                                                       *
*            Press the [Enter] key to automatically Power off          *
*   the Athena System and power back ON within 38 seconds               *
*    >>>  FPGAs will reload during the Power off phase   <<<            *
*************************************************************************
```

BMC firmware

This update is in the form of a bootable diskette. If you have the RSA II configured and connected to your network, you can use the Remote Console feature with a remote diskette and boot each node from the diagnostics diskette.

On x3950 M2 multinode complex, if you perform the update from the primary node, then all other nodes are automatically updated as shown in Example 5-2.

Example 5-2 Updating the BMC code on a two-node complex

```
Detecting RAMDRIVE
Copying files to RAMDRIVE D:
Decompressing BMC firmware
reading fullfw.cmt…wrote 16360 SRECs to fullfw.mot
Acquiring BMC attributes
Updating BMC Firmware
Do you want to clear the SEL (y or n)?y

> SEL Cleared.
>
Flash Loader v1.30.0.54, OSA Technologies. Inc. (c)2007

                     firmware            image
IPMI Version=          2.0                2.0
major Revision=         2                  2
minor Revision =       32                 32
manufacturer ID =       2                  2
product ID=            77                 77
build Name=           A3BT31B            A3BT31B

Firmware and the image have the same version and build name.

Start to program flash? (Y/N) Y_

Start programming…
Writing to Address: 0x0007FF80......OK.
Download to Flash OK.

BMC initialization…OK.
BMC Firmware and SDRs updated successfully!!!
Do you want to clear the SEL (y or n)?y

> SEL Cleared.
>

Flash Loader v1.30.0.54, OSR Technologies. Inc. (c)2007

                     firmware            image
IPMI Version=          2.0                2.0
```

```
major Revision=          2                  2
minor Revision =        32                 32
manufacturer ID =        2                  2
product ID=             77                 77
build Name=           A3BT31B         A3BT31B

Firmware and the image have the same version and build name.

Start to program flash? (Y/N) Y_

Start programming…
Writing to Address: 0x0007FF80……OK.
Download to Flash OK.

BMC initialization…OK.
BMC Firmware and SDRs updated successfully!!!

Please remove the disk from the drive and restart the system
D :\>
```

System BIOS

In a x3950 M2 multinode partition, you can update the BIOS of all nodes from the primary node. You are presented with a menu from which you can select:

- Update the BIOS on the current system.
- Update all nodes in a multinode x3950 M2 partition from the primary node.

The BIOS update process involves a sequence of erasing the current flash image, updating the new flash image, requesting confirmation of the serial number (SN) and model type for each node in consecutive order (from first node to last node).

Preboot Diagnostics (DSA Preboot)

This update is in the form of a bootable diskette. If you have the RSA II configured and connected to your network, you can use the Remote Console feature with a remote diskette and boot each node from the diagnostics diskette.

On x3950 M2 multinode complex, if you perform the update from the primary node, then all other nodes are automatically updated as shown in Example 5-3.

Example 5-3 Updating the DSA Preboot code on a two-node complex

```
Commands:
    update - Update/Recover your embedded usb chip
    help - Display this help message.
    exit - Quit program.
             Note: This will reboot the system.
```

```
Please enter a command.  (Type 'help' for commands)
>update_
Node count:  2
Flashing node:  0
Waiting for device to become available:
.usb 9-6: new high speed USB device using ehci_hcd and address 4
usb 9-6: new device found, idVendor=0483, idProduct=fada
usb 9-6: new device strings: Mfr=2, Product=1, SerialNumber=0
usb 9-6: Product: ST72682  High Speed Mode
usb 9-6: Manufacturer: STMicroelectronics
usb 9-6: configuration #1 chosen from 1 choice
scsi1 : SCSI emulation for USB Mass Storage devices
.....  Vendor:  ST         Model:  ST72682                Rev: 2.10
   Type:      Direct-Access                       ANSI SCSI revision:  02
SCSI device sda: 1024000 512-byte hdwr sectors (524 MB)
sda: Write Protect is off
sda: assuming drive cache:  write through
SCSI device sda: 1024000 512-byte hdwr sectors (524 MB)
.sda: Write Protect is off
sda: assuming drive cache:  write through
sda: sda1
sd 1:0:0:0: Attached scsi removable disk sda
sd 1:0:0:0: Attached scsi generic sg1 type 0

Device flashing in progress:
.........................................................................
.........................................................................
.........................................................................
Flashing node:  1
Waiting for device to become available:
.sda : READ CAPACITY failed.
sda : status=0, message=00, host=7, driver=00
sda : sense not available.
sda : Write Protect is off
sda : assuming drive cache: write through
sda : READ CAPACITY failed.
sda : status=0, message=00, host=7, driver=00
sda : sense not available.
usb 10-6: new high speed USB device using ehci_hcd and address 4
sda : Write Protect is off
sda : assuming drive cache: write through
INQUIRY host_status=0x7
sda : READ CAPACITY failed.
sda : status=0, message=00, host=7, driver=00
sda : sense not available.
sda : Write Protect is off
sda : assuming drive cache: write through
sda : READ CAPACITY failed.
sda : status=0, message=00, host=7, driver=00
sda : sense not available.
sda : Write Protect is off
sda : assuming drive cache: write through
INQUIRY host_status=0x7
usb 10-6: new device found, idVendor=0483, idProduct=fada
usb 10-6: new device strings: Mfr=2, Product=1, SerialNumber=0
usb 10-6: Product: ST72682  High Speed Mode
usb 10-6: Manufacturer: STMicroelectronics
usb 10-6: configuration #1 chosen from 1 choice
scsi2 : SCSI emulation for USB Mass Storage devices
usb 9-6: USB disconnect, address 4
......  Vendor:  ST             Model:  ST72682                Rev: 2.10
   Type:      Direct-Access                       ANSI SCSI revision:  02
SCSI device sda: 1024000 512-byte hdwr sectors (524 MB)
sda: Write Protect is off
sda: assuming drive cache:  write through
SCSI device sda: 1024000 512-byte hdwr sectors (524 MB)
.sda: Write Protect is off
sda: assuming drive cache:  write through
sda: sda1
sd 2:0:0:0: Attached scsi removable disk sda
sd 2:0:0:0: Attached scsi generic sg1 type 0

Device flashing in progress:
.........................................................................
.........................................................................
.........................................................................
DSA key has been flashed successfully
usb 10-6: USB disconnect, address 4

Please enter a command.  (Type 'help' for commands)
>
```

Additional devices

For additional components with upgradable firmware, we recommend you apply the latest versions of firmware on each. This includes:

- LSI1078 Integrated SAS/SATA RAID controller
- ServeRAID MR10k SAS/SATA RAID controller
- ServeRAID MR10M SAS/SATA RAID controller
- Network adapters
- Fibre Channel HBAs
- iSCSI HBAs
- SAS HBAs

5.2 Confirming BIOS settings

Before installing an operating system, ensure that all system components are correctly detected in the BIOS and that the settings are correct for those devices.

To check the settings:

1. At System Boot up press F1 to enter the BIOS **Configuration/Setup Utility** window, shown in Figure 5-2.

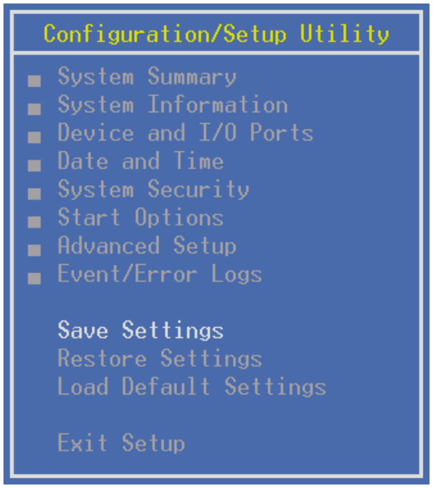
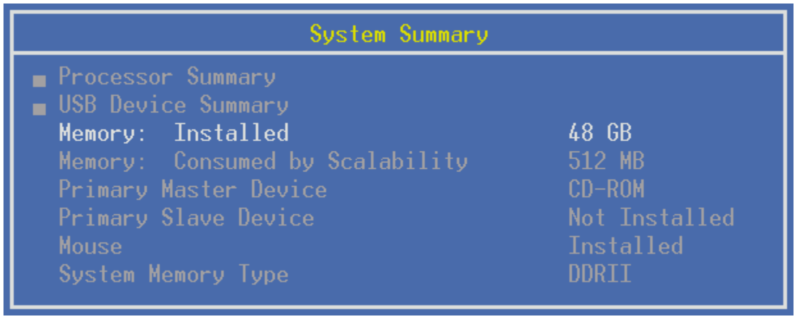


Figure 5-2 BIOS Configuration/Setup Utility window

2. Select **System Summary** to open the System Summary window shown in Figure 5-3.

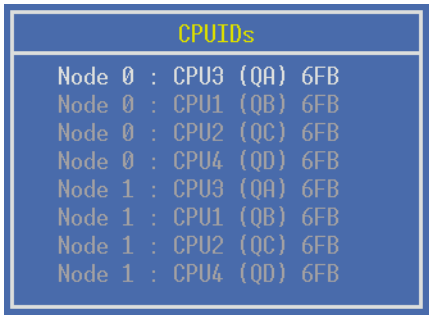


```
                         System Summary
■ Processor Summary
■ USB Device Summary
  Memory:  Installed                       48 GB
  Memory:  Consumed by Scalability         512 MB
  Primary Master Device                    CD-ROM
  Primary Slave Device                     Not Installed
  Mouse                                    Installed
  System Memory Type                       DDRII
```

Figure 5-3 System Summary window

In the next steps, you check the system BIOS to ensure all processors, memory cards and capacity, and PCIe adapters installed on the x3850 M2, x3950 M2, or both nodes are correctly detected.

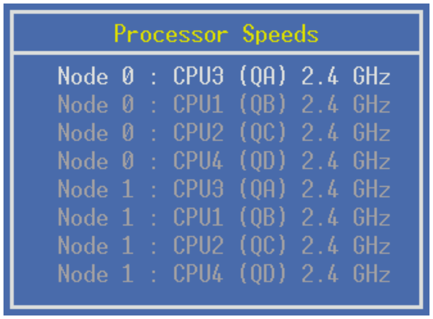
3. From the System Summary window, select **Processor Summary** → **CPUIDs** (Figure 5-4) and check that all processors on stand-alone or merged nodes are detected as expected.



```
CPUIDs

Node 0 : CPU3 (QA) 6FB
Node 0 : CPU1 (QB) 6FB
Node 0 : CPU2 (QC) 6FB
Node 0 : CPU4 (QD) 6FB
Node 1 : CPU3 (QA) 6FB
Node 1 : CPU1 (QB) 6FB
Node 1 : CPU2 (QC) 6FB
Node 1 : CPU4 (QD) 6FB
```

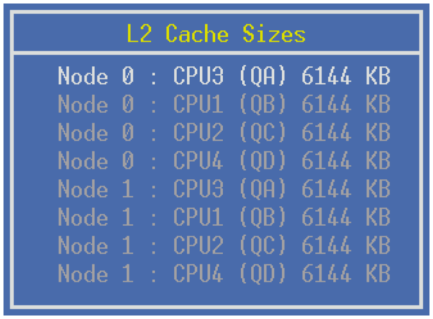
Figure 5-4 CPUIDs on a two-node x3950 M2

4. From the System Summary window, select **Processor Summary** → **Processor Speeds** (Figure 5-5) to ensure that all processors are matched, and check **L2 Cache Sizes** (Figure 5-6 on page 256).



```
Processor Speeds

Node 0 : CPU3 (QA) 2.4 GHz
Node 0 : CPU1 (QB) 2.4 GHz
Node 0 : CPU2 (QC) 2.4 GHz
Node 0 : CPU4 (QD) 2.4 GHz
Node 1 : CPU3 (QA) 2.4 GHz
Node 1 : CPU1 (QB) 2.4 GHz
Node 1 : CPU2 (QC) 2.4 GHz
Node 1 : CPU4 (QD) 2.4 GHz
```

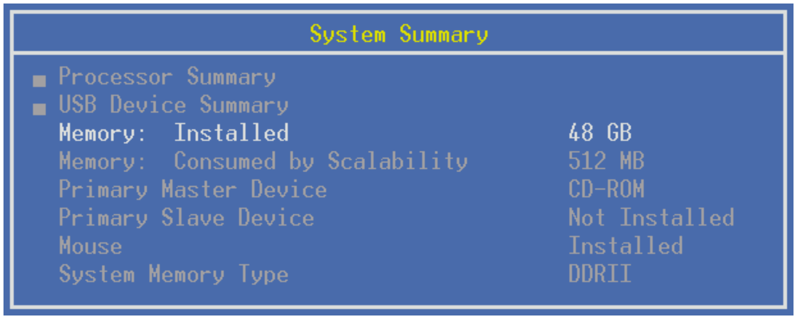
Figure 5-5 Processor Speeds on a two-node x3950 M2



```
            L2 Cache Sizes

  Node 0 : CPU3 (QA) 6144 KB
  Node 0 : CPU1 (QB) 6144 KB
  Node 0 : CPU2 (QC) 6144 KB
  Node 0 : CPU4 (QD) 6144 KB
  Node 1 : CPU3 (QA) 6144 KB
  Node 1 : CPU1 (QB) 6144 KB
  Node 1 : CPU2 (QC) 6144 KB
  Node 1 : CPU4 (QD) 6144 KB
```

Figure 5-6 Processor L2 Cache Sizes

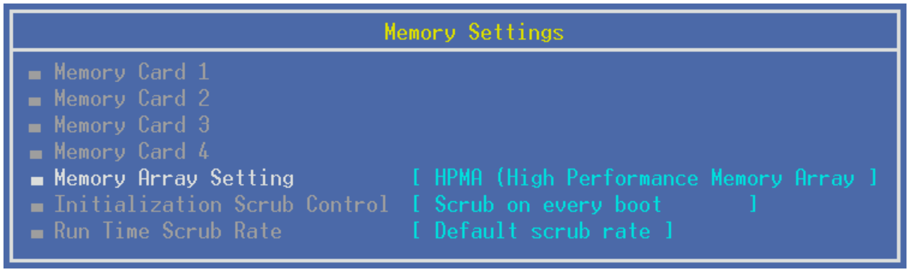
5. In the **System Summary** window (Figure 5-7), check that all memory installed is visible to the server, as expected.



```
                        System Summary

 ■ Processor Summary
 ■ USB Device Summary
   Memory:  Installed                      48 GB
   Memory:  Consumed by Scalability        512 MB
   Primary Master Device                   CD-ROM
   Primary Slave Device                    Not Installed
   Mouse                                   Installed
   System Memory Type                      DDRII
```

Figure 5-7 Memory Installed, detailed in System Summary

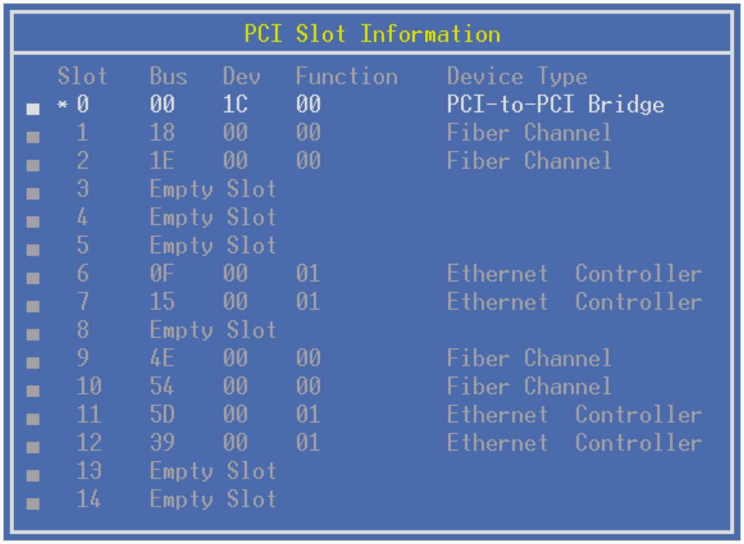
6. Check that the memory cards installed in the server are detected as shown in the Memory Settings window (Figure 5-8), which also shows memory array, scrub on every boot, and the scrub rate at run time. (To open the Memory Settings window, select Advanced Setup in the Configuration/Setup Utility window, shown in Figure 5-2 on page 254.)



```
                          Memory Settings

 ■ Memory Card 1
 ■ Memory Card 2
 ■ Memory Card 3
 ■ Memory Card 4
 ■ Memory Array Setting          [ HPMA (High Performance Memory Array ]
 ■ Initialization Scrub Control  [ Scrub on every boot        ]
 ■ Run Time Scrub Rate           [ Default scrub rate ]
```

Figure 5-8 Detected memory cards in the primary node

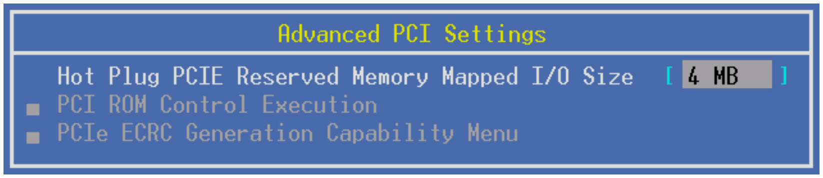
7. Check that all PCIe adapters are detected in the respective slots in which they were installed, as shown in Figure 5-9. Slots that have an asterisk (*) next to the slot number indicate that the adapter has more than one interface (for example, dual-port or quad-port NICs, dual-port FC HBAs, and others).



```
                    PCI Slot Information

    Slot    Bus    Dev    Function      Device Type
■   * 0     00     1C     00            PCI-to-PCI Bridge
■     1     18     00     00            Fiber Channel
■     2     1E     00     00            Fiber Channel
■     3     Empty Slot
■     4     Empty Slot
■     5     Empty Slot
■     6     0F     00     01            Ethernet  Controller
■     7     15     00     01            Ethernet  Controller
■     8     Empty Slot
■     9     4E     00     00            Fiber Channel
■     10    54     00     00            Fiber Channel
■     11    5D     00     01            Ethernet  Controller
■     12    39     00     01            Ethernet  Controller
■     13    Empty Slot
■     14    Empty Slot
```

Figure 5-9   PCIe slot information and the installed adapters

8. Depending on the numbers and types of PCIe adapters in a multinode configuration, you might have to configure the Reserved Memory Mapped I/O Size, shown in Figure 5-10. For details see 3.5.4, “PCI Express device-related information in the BIOS” on page 190.



```
                    Advanced PCI Settings

    Hot Plug PCIE Reserved Memory Mapped I/O Size  [ 4 MB    ]
■   PCI ROM Control Execution
■   PCIe ECRC Generation Capability Menu
```

Figure 5-10   Advanced PCI Settings

5.3 Supported operating systems

Before you install your intended operating system, check the IBM ServerProven Web pages, indicated in the following list, to verify that the particular version of an operating system has been tested for compatibility with the x3850 M2 or x3950 M2:

- IBM ServerProven home

  http://www.ibm.com/servers/eserver/serverproven/compat/us/index.html
- IBM ServerProven operating system support matrix

  http://www.ibm.com/servers/eserver/serverproven/compat/us/nos/matrix.shtml
- IBM ServerProven compatibility for System x

  http://www.ibm.com/servers/eserver/serverproven/compat/us/indexsp.html

IBM ServerProven program also extends to testing and publishing the IBM System x options (for example, network adapters, storage adapters, storage expansions, storage subsystems and other common options.

In addition, we also recommend that you check the particular operating system vendor's hardware compatibility list (HCL). Operating system vendors typically publish and update their HCL as new models of servers are introduced by the server hardware vendors. The commonly referenced Web pages of the operating system vendors are:

- VMware

  http://www.vmware.com/pdf/vi35_systems_guide.pdf
  http://www.vmware.com/pdf/vi3_systems_guide.pdf
  http://www.vmware.com/pdf/vi35_io_guide.pdf
  http://www.vmware.com/pdf/vi3_io_guide.pdf
- Microsoft

  http://www.windowsservercatalog.com/
  http://www.windowsservercatalog.com/item.aspx?idItem=dbf1ed79-c158-c428-e19d-5b4144c9d5cd
  http://www.microsoft.com/whdc/hcl
- Red Hat

  https://hardware.redhat.com/
- Novell SUSE Linux

  http://developer.novell.com/yessearch/Search.jsp
- Solaris™

  http://www.sun.com/bigadmin/hcl/
  http://www.sun.com/bigadmin/hcl/data/systems/details/3406.html

5.3.1 VMware ESX operating systems

The ServerProven NOS support matrix for VMware is located at:

http://www.ibm.com/servers/eserver/serverproven/compat/us/nos/vmware.html

At the time of writing this book, the VMware operating systems are supported on x3850 M2 and x3950 M2 as indicated in Table 5-1. The shaded table cells indicate *Yes*, they are supported.

Table 5-1 VMware ESX OS support (current and planned)

Operating System	**x3850 M2**	**x3950 M2 1-node**	**x3950 M2 2-node**	**x3950 M2 3-node**	**x3950 M2 4-node**
VMware ESX 3.5	Yes	Yes with patch ESX350-200802301-BG[a]	Yes with Update 1	No	Yes with Update 2
VMware ESXi 3.5	Yes (on selected x3850 M2 hypervisor models)	Yes with Update 1	Yes with Update 2	No	No
VMware ESXi 3.5 Installable	Yes	Yes with Update 1	No	No	No
VMware ESX 3.0.2	Yes with Update 1 or later	Yes with Update 1 or later	Check Server Proven	No	No

a. This supersedes patch ESX350-200712401-BG.

5.3.2 Windows Server 2003 and 2008 operating systems

The ServerProven NOS Support Matrix for Windows is located at:

http://www.ibm.com/servers/eserver/serverproven/compat/us/nos/microsoft.html

At the time of writing this book, the Windows operating systems are supported on x3850 M2 and x3950 M2 as indicated in Table 5-2 on page 260. The shaded table cells indicate *Yes*, they are supported.

Table 5-2 Microsoft Windows Server 2003/2008 OS support (current and planned)

Operating System	**x3850 M2**	**x3950 M2 1-node**	**x3950 M2 2-node**	**x3950 M2 3-node**	**x3950 M2 4-node**
Windows Server 2008					
Microsoft Windows Server 2008, Datacenter x64 Edition	Yes	Yes	Yes	Yes	Yes
Microsoft Windows Server 2008, Datacenter x64 Edition with Hyper-V	Yes	Yes	Planned	Planned	Planned
Microsoft Windows Server 2008, Enterprise x64 Edition	Yes	Yes	Yes	No	No
Microsoft Windows Server 2008, Enterprise x64 Edition with Hyper-V	Yes	Yes	Planned	No	No
Microsoft Windows Server 2008, Standard x64 Edition	Yes	Yes	Yes (limited to 2 sockets in each node)	No	No
Microsoft Windows Server 2008, Standard x64 Edition with Hyper-V	Yes	Yes	Planned	No	No
Microsoft Windows Server 2008, Web x64 Edition	No	No	No	No	No
Microsoft Windows Server 2008, Datacenter x86 Edition	No	No	No	No	No
Microsoft Windows Server 2008, Enterprise x86 Edition	No	No	No	No	No
Microsoft Windows Server 2008, Standard x86 Edition	No	No	No	No	No
Microsoft Windows Server 2008, Web x86 Edition	No	No	No	No	No
Windows Server 2003					
Microsoft Windows Server 2003 R2 x64 Datacenter Edition Unlimited Virtualization	Yes	Yes, with SP2 or later	Yes, with SP2 or later	Planned	Planned

Operating System	x3850 M2	x3950 M2 1-node	x3950 M2 2-node	x3950 M2 3-node	x3950 M2 4-node
Microsoft Windows Server 2003 R2 Datacenter Edition Unlimited Virtualization	Yes	Yes, with SP2 or later	No	No	No
Microsoft Windows Server 2003 R2 x64 Datacenter Edition Unlimited Virtualization with High Availability Program	Yes	Yes, with SP2 or later	Yes, with SP2 or later	Planned	Planned
Microsoft Windows Server 2003 R2 Datacenter Edition Unlimited Virtualization with High Availability Program	Yes	Yes, with SP2 or later	No	No	No
Microsoft Windows Server 2003/2003 R2 Enterprise x64 Edition	Yes	Yes, with SP2 or later	Yes, with SP2 or later	No	No
Microsoft Windows Server 2003/2003 R2 Standard x64 Edition	Yes	Yes, with SP2 or later	Yes, with SP2 or later (limited to 2 sockets in each node)	No	No
Microsoft Windows Server 2003/2003 R2 Web x64 Edition	No	No	No	No	No
Microsoft Windows Server 2003/2003 R2 Enterprise x86 Edition	Yes	Yes, with SP2 or later	No	No	No
Microsoft Windows Server 2003/2003 R2 Standard x86 Edition	Yes	Yes, with SP2 or later	No	No	No
Microsoft Windows Server 2003/2003 R2 Web x86 Edition	No	No	No	No	No

5.3.3 Red Hat Enterprise Linux operating systems

See the ServerProven NOS support matrix for Red Hat Enterprise Linux at:

http://www.ibm.com/servers/eserver/serverproven/compat/us/nos/redchat.html

At the time of writing this book, the Red Hat operating systems are supported on x3850 M2 and x3950 M2 as indicated in Table 5-3. The shaded table cells indicate *Yes*, they are supported.

Table 5-3 Red Hat Enterprise Linux OS support (current and planned)

Operating System	x3850 M2	x3950 M2 1-node	x3950 M2 2-node	x3950 M2 3-node	x3950 M2 4-node
Red Hat Enterprise Linux 5					
Red Hat Enterprise Linux 5 Server x64 Edition	Yes	Yes	Yes with Update 1 or later	No	No
Red Hat Enterprise Linux 5 Server with Xen x64 Edition	Yes	Yes	Yes with Update 1 or later	No	No
Red Hat Enterprise Linux 5 Server Edition	Yes	Yes	No	No	No
Red Hat Enterprise Linux 5 Server with Xen Edition	Yes	Yes	No	No	No
Red Hat Enterprise Linux 4					
Red Hat Enterprise Linux 4 AS for AMD64/EM64T	Yes with Update 5 or later	Yes with Update 5 or later	Planned	No	No
Red Hat Enterprise Linux 4 AS for x86	Yes with Update 5 or later	Yes with Update 5 or later	No	No	No
Red Hat Enterprise Linux 4 ES for AMD64/EM64T	Yes, with Update 5	Yes, with Update 5	Planned	No	No
Red Hat Enterprise Linux 4 ES for x86	Yes, with Update 5	Yes, with Update 5	No	No	No

5.3.4 SUSE Linux Enterprise Server operating systems

The ServerProven NOS support matrix for SUSE Linux Enterprise Server is located at:

http://www.ibm.com/servers/eserver/serverproven/compat/us/nos/suseclinux.html

At the time of writing this book, the SUSE Linux operating systems are supported on x3850 M2 and x3950 M2 as indicated in Table 5-4 on page 263. The shaded table cells indicate *Yes*, they are supported.

Table 5-4 SUSE Linux Enterprise Server OS support (current and planned)

Operating System	x3850 M2	x3950 M2 1-node	x3950 M2 2-node	x3950 M2 3-node	x3950 M2 4-node
SUSE Linux Enterprise Server 10					
SUSE Linux Enterprise Server 10 for AMD64/EM64T	Yes, with SP1 or later	Yes, with SP1 or later	Yes, with SP1 or later	No	No
SUSE Linux Enterprise Server 10 with Xen for AMD64/EM64T	Yes, with SP1 or later	Yes, with SP1 or later	Yes, with SP1 or later	No	No
SUSE Linux Enterprise Server 10 with Xen for x86	Yes, with SP1 or later	Yes, with SP1 or later	No	No	No
SUSE Linux Enterprise Server 10 for x86	Yes, with SP1 or later	Yes, with SP1 or later	No	No	No
SUSE Linux Enterprise Server 9					
SUSE Linux Enterprise Server 9 for AMD64/EM64T	Yes, with SP3 U2 or later	Yes, with SP3 U2 or later	Yes, with SP4 or later	No	No
SUSE Linux Enterprise Server 9 for x86	Yes, with SP3 U2 or later	No	No	No	No

5.3.5 Solaris operating systems

The ServerProven NOS support matrix for Solaris is located at:

http://www.ibm.com/servers/eserver/serverproven/compat/us/nos/suseclinux.html

At the time of writing this book, only Solaris 10 was supported on x3850 M2 and x3950 M2 as indicated in Table 5-5. The shaded table cells indicate *Yes*, they are supported.

Table 5-5 Solaris 10 OS support (current and planned)

Operating System	x3850 M2	x3950 M2 1-node	x3950 M2 2-node	x3950 M2 3-Node	x3950 M2 4-node
Solaris 10	Yes	Yes, with Solaris 10 08/07 to Solaris 10 05/08 (including Solaris Express, Developer Edition 01/08) OpenSolaris™ 2008.05	No	No	No

5.4 Installing the operating system

This section describes the tasks required to install the following supported operating systems for the IBM x3850 M2 and x3950 M2 servers:

- 5.4.1, “Installing (configuring) VMware ESXi 3.5 embedded” on page 264
- 5.4.2, “Installing VMware ESXi 3.5 Installable” on page 279
- 5.4.3, “Installing VMware ESX 3.5 Update 1” on page 281
- 5.4.4, “Installing Windows Server 2003” on page 285
- 5.4.5, “Installing Windows Server 2008” on page 289
- 5.4.6, “Installing Red Hat Enterprise Linux 5 Update 1” on page 293
- 5.4.7, “Installing SUSE Linux Enterprise Server 10 SP1” on page 295

Installation of Solaris operating systems is not described.

5.4.1 Installing (configuring) VMware ESXi 3.5 embedded

VMware ESXi 3.5 boots from an IBM customized USB Flash Drive. It does not have to be installed onto a local server storage. IBM does provide ESXi 3.5 Installable Edition, which can be installed onto local storage but cannot be installed onto remote storage such as a SAN LUN or iSCSI target.

The following topics are described in this section:

- “Prerequisites” on page 264
- “Configure server BIOS for Embedded Hypervisor Boot” on page 265
- “Boot ESXi hypervisor and customize settings” on page 266

Prerequisites

Before you configure ESXi:

- Ensure you have local or remote (through RSA II Remote Control) video and keyboard console access to the server with the IBM ESXi flash drive installed.
- Download the following three VMware ESXi guides:
  - Getting Started with ESX Server 3i Installable

    http://www.vmware.com/pdf/vi3_35/esx_3i_i/r35/vi3_35_25_3i_i_get_start.pdf
  - ESX Server 3i Embedded Setup Guide

    http://www.vmware.com/pdf/vi3_35/esx_3i_e/r35/vi3_35_25_3i_setup.pdf

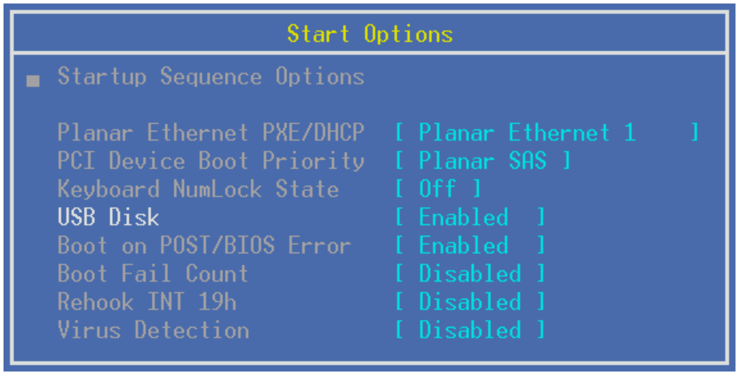
- ESX Server 3i Configuration Guide

  http://www.vmware.com/pdf/vi3_35/esx_3i_e/r35/vi3_35_25_3i_server_config.pdf

Configure server BIOS for Embedded Hypervisor Boot

To configure the server's BIOS:

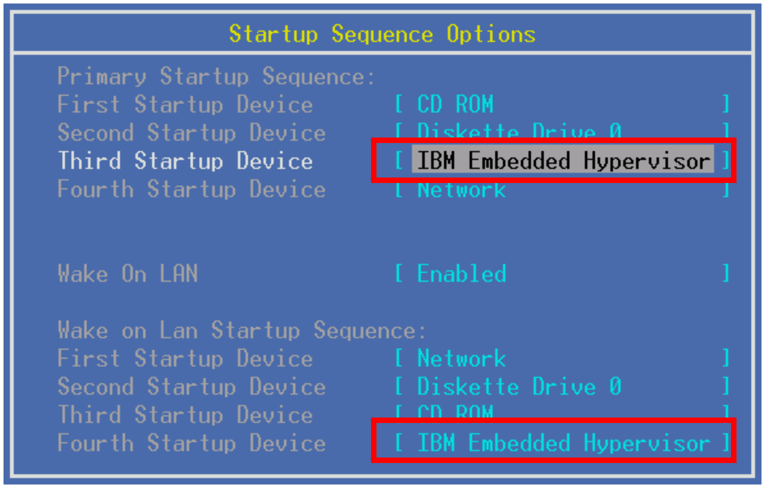
1. Press `F1` to boot the server in the BIOS.
2. Go to **Start Options** → **USB Disk** and set it to **Enabled**, as shown in Figure 5-11.



```
                         Start Options
■ Startup Sequence Options

  Planar Ethernet PXE/DHCP   [ Planar Ethernet 1     ]
  PCI Device Boot Priority   [ Planar SAS ]
  Keyboard NumLock State     [ Off ]
  USB Disk                   [ Enabled  ]
  Boot on POST/BIOS Error    [ Enabled  ]
  Boot Fail Count            [ Disabled ]
  Rehook INT 19h             [ Disabled ]
  Virus Detection            [ Disabled ]
```

Figure 5-11 Example of x3850 M2 BIOS Start Options with USB Disk enabled

3. In **Start Options** → **Startup Sequence Options**, set the following startup sequences to **IBM Embedded Hypervisor**, also shown in Figure 5-12 on page 266:
   - Primary Startup Sequence: Third Startup Device
   - Wake on LAN® Startup Sequence: Fourth Startup Device



```
                 Startup Sequence Options

   Primary Startup Sequence:
   First Startup Device        [ CD ROM                 ]
   Second Startup Device       [ Diskette Drive 0       ]
   Third Startup Device        [ IBM Embedded Hypervisor ]
   Fourth Startup Device       [ Network                ]


   Wake On LAN                 [ Enabled                ]

   Wake on Lan Startup Sequence:
   First Startup Device        [ Network                ]
   Second Startup Device       [ Diskette Drive 0       ]
   Third Startup Device        [ CD ROM                 ]
   Fourth Startup Device       [ IBM Embedded Hypervisor ]
```

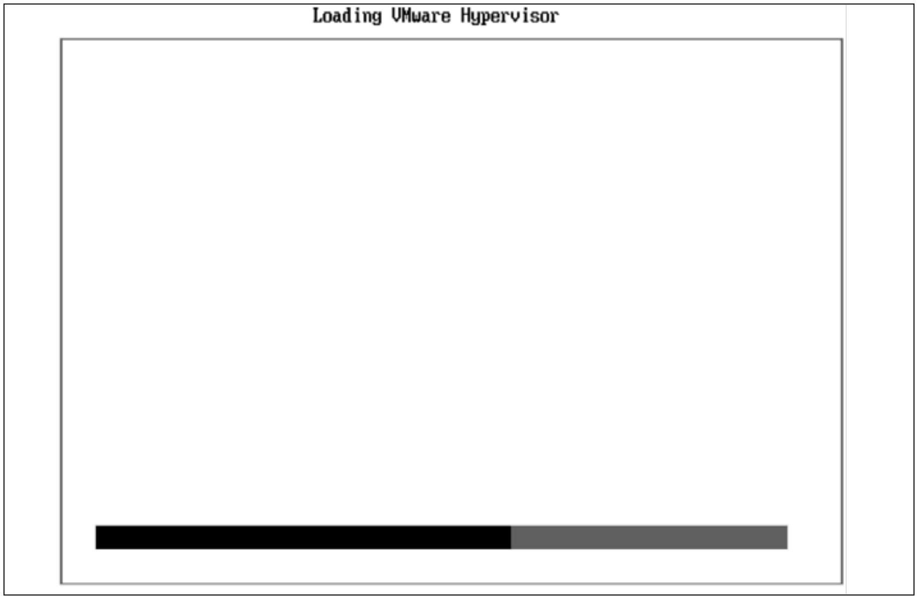
Figure 5-12 IBM Embedded Hypervisor configured as a startup device after CD-ROM and diskette drive 0 in system BIOS

4. Exit the BIOS utility and save you settings when prompted.

Boot ESXi hypervisor and customize settings

After setting the IBM Embedded Hypervisor to boot as a startup device in the system BIOS, and after POST completes, the server starts to boot VMware ESXi hypervisor from the USB Flash Disk. The following steps are the sequence of events that take place and how you interact with the process:

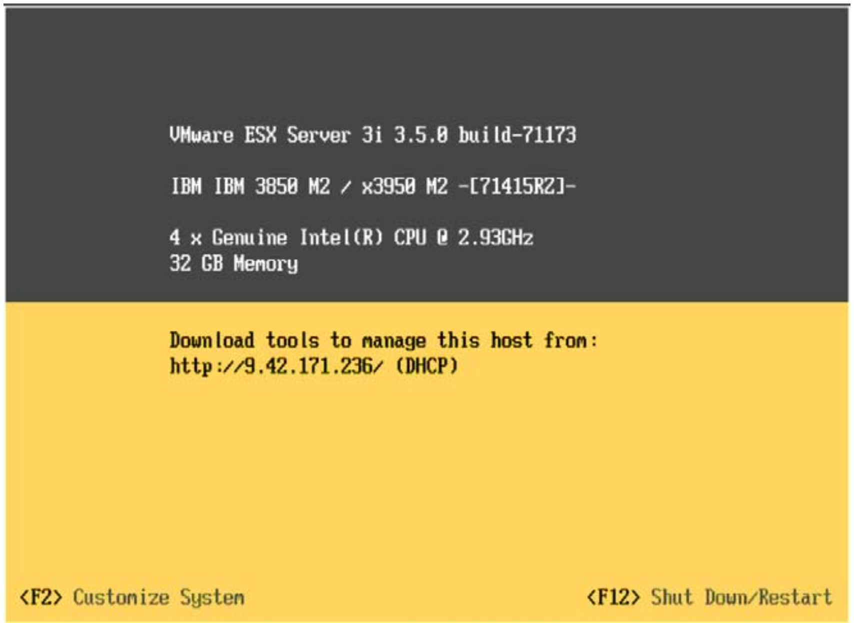
1. After the server POST has completed, the server boots from the IBM USB Flash Disk and loads ESXi hypervisor, shown in Figure 5-13.



Loading VMware Hypervisor

Figure 5-13 ESXi starting to boot from IBM customized USB Flash Disk with IBM Customized ESXi image

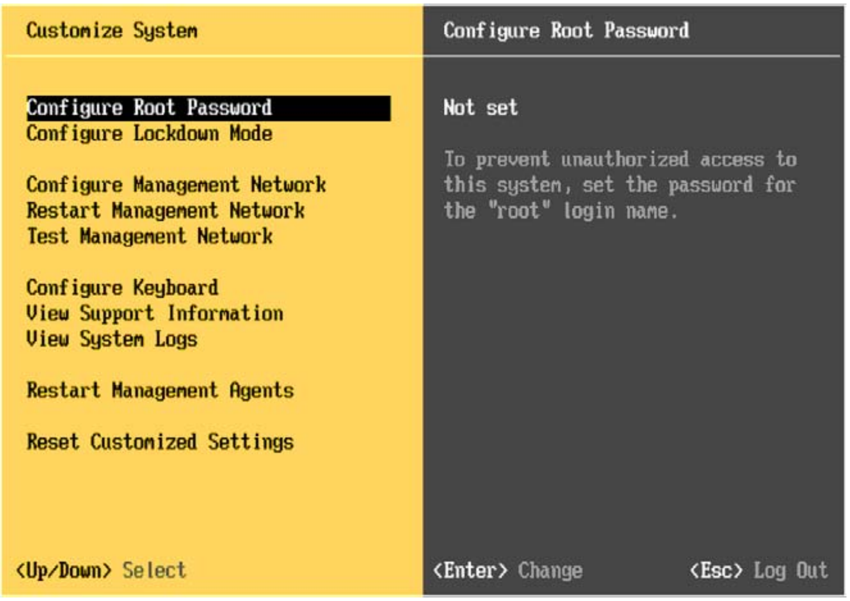
2. After ESXi finishes loading, a Direct Console user interface (DCUI) opens, shown in Figure 5-14, which is the primary interface for basic configuration of ESXi. The default network setting and behavior for ESXi is to request a DHCP IP Address. In the event that it cannot find a DHCP server to obtain a DHCP lease, it defaults to the 169.254.0.1/255.255.0.0 IP address and Class B netmask.



```
VMware ESX Server 3i 3.5.0 build-71173

IBM IBM 3850 M2 / x3950 M2 -[71415RZ]-

4 x Genuine Intel(R) CPU @ 2.93GHz
32 GB Memory

Download tools to manage this host from:
http://9.42.171.236/ (DHCP)

<F2> Customize System                          <F12> Shut Down/Restart
```

Figure 5-14 The x3850 M2 is successfully booting the IBM Customized ESXi on the USB Flash Disk, and detecting processor sockets and total memory installed

3. In the Customize System window, shown in Figure 5-15, change the default root password to a more secure password in accordance with your root/administrator password policies; select **Configure Root Password**.



Customize System

Configure Root Password
Configure Lockdown Mode

Configure Management Network
Restart Management Network
Test Management Network

Configure Keyboard
View Support Information
View System Logs

Restart Management Agents

Reset Customized Settings

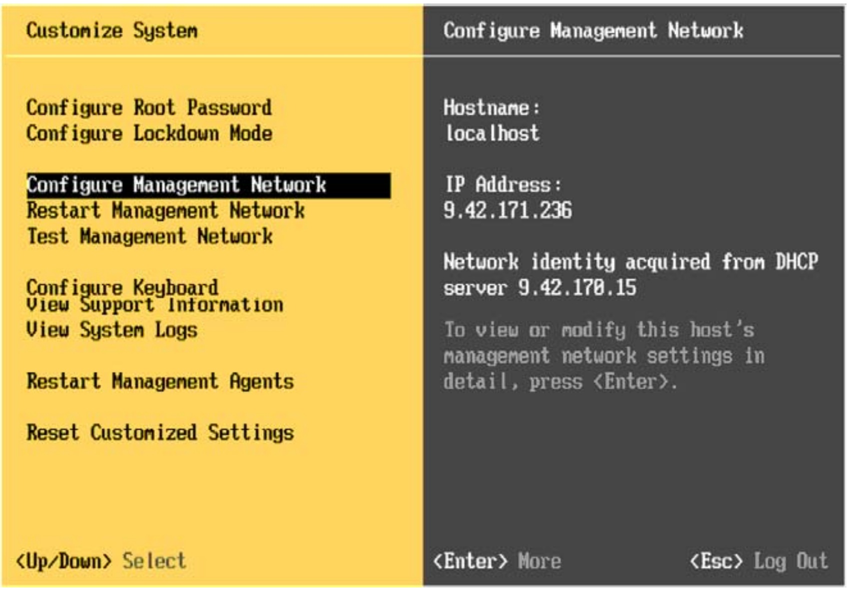
Configure Root Password

Not set

To prevent unauthorized access to this system, set the password for the "root" login name.

Figure 5-15 The Customize System menu when you press F2 at the DCUI ESXi interface

4. Set the host name (for example DNS FQDN) and DNS servers IP Addresses by selecting **Configure Management Network**. See Figure 5-16 on page 270. You might want to set the ESXi host management IP Address (this interface should be separate to the interfaces you would set in VI Client for Virtual Machine traffic) to static rather than DHCP, depending on your management network policies.



Configure Management Network

Hostname:
localhost

IP Address:
9.42.171.236

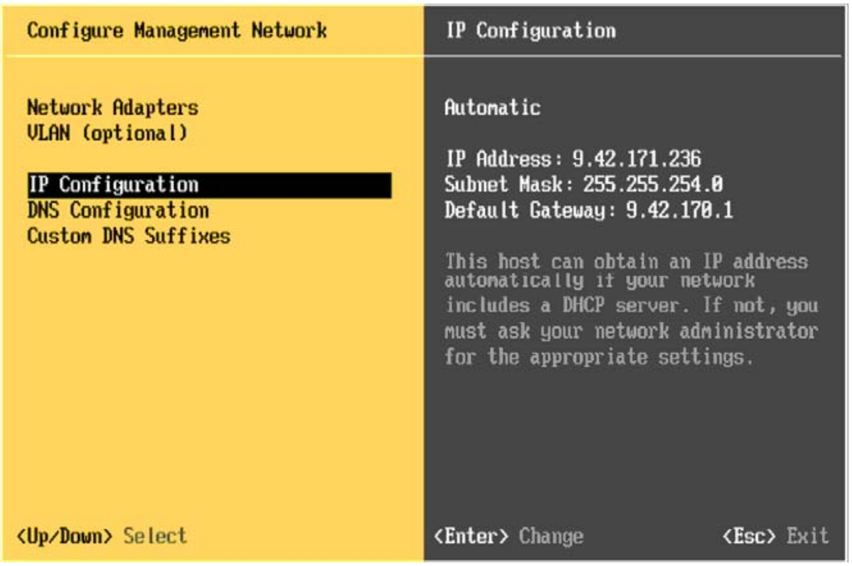
Network identity acquired from DHCP server 9.42.170.15

To view or modify this host's management network settings in detail, press <Enter>.

Figure 5-16 Configuring the management network IP address settings

5. Under Network Adapters, Figure 5-17, you can select which vmnic to assign to the ESXi host management interface for fault-tolerance and management network traffic load-balancing.

   Selecting **VLAN (optional)** allows you to set a VLAN ID (if your network policy uses VLANs for network isolation or segregation) for this ESXi host management interface.



Automatic

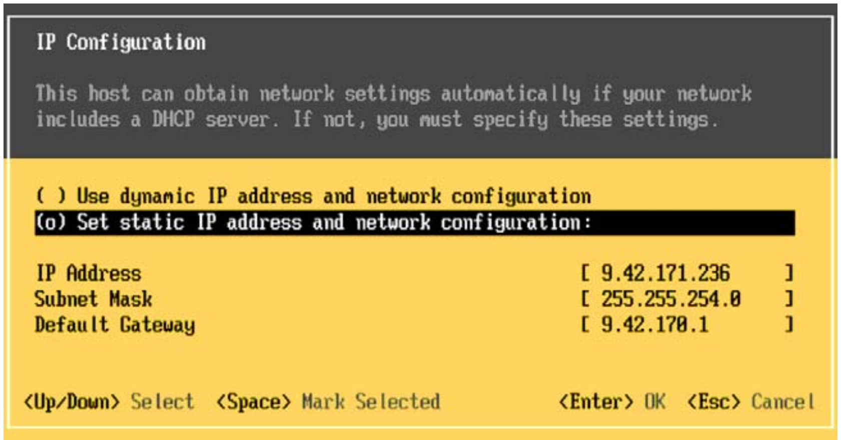
IP Address: 9.42.171.236
Subnet Mask: 255.255.254.0
Default Gateway: 9.42.170.1

This host can obtain an IP address automatically if your network includes a DHCP server. If not, you must ask your network administrator for the appropriate settings.

Figure 5-17 Options available under Configure Management Network

6. Select **IP Configuration** and then set the IP Address to match your management network policies (static IP or DHCP). See Figure 5-18 on page 272. Make sure to register the host name in your DNS server (or servers), and set DNS server IP addresses and DNS suffixes for the ESXi host, especially if you plan to use management tools like VMware VirtualCenter which relies heavily on DNS name resolution.

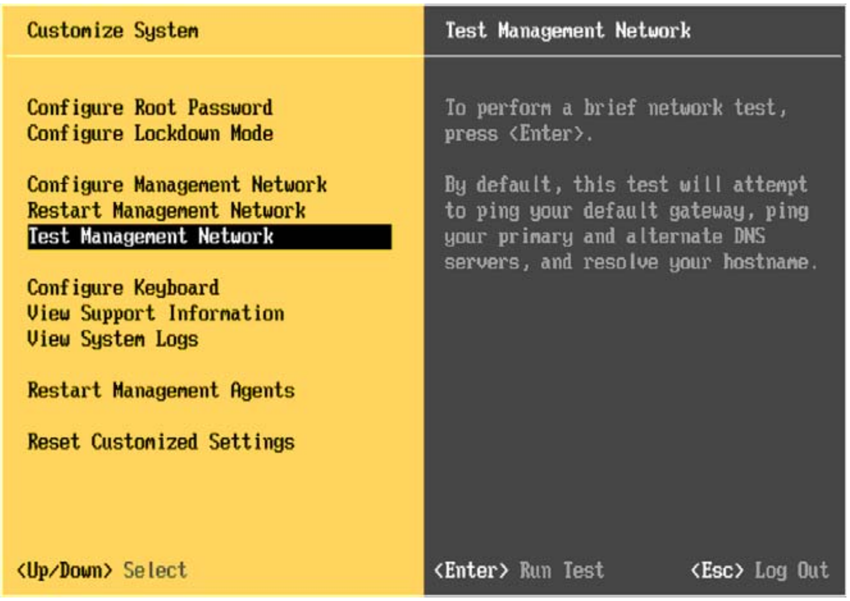
   IP Configuration, Figure 5-18 on page 272, allows you to set **Use dynamic IP address and network configuration** (ESXi requests a DHCP lease) or **Set static IP addresses and network configuration**. Note that if ESXi cannot find a DHCP server, it defaults to the 169.254.0.1/255.255.0.0 IP address and Class B netmask.



```
IP Configuration

This host can obtain network settings automatically if your network
includes a DHCP server. If not, you must specify these settings.

( ) Use dynamic IP address and network configuration
(o) Set static IP address and network configuration:

IP Address                                          [ 9.42.171.236    ]
Subnet Mask                                         [ 255.255.254.0   ]
Default Gateway                                     [ 9.42.170.1      ]

<Up/Down> Select  <Space> Mark Selected          <Enter> OK  <Esc> Cancel
```

Figure 5-18 Setting a static IP Address for ESXi host management interface

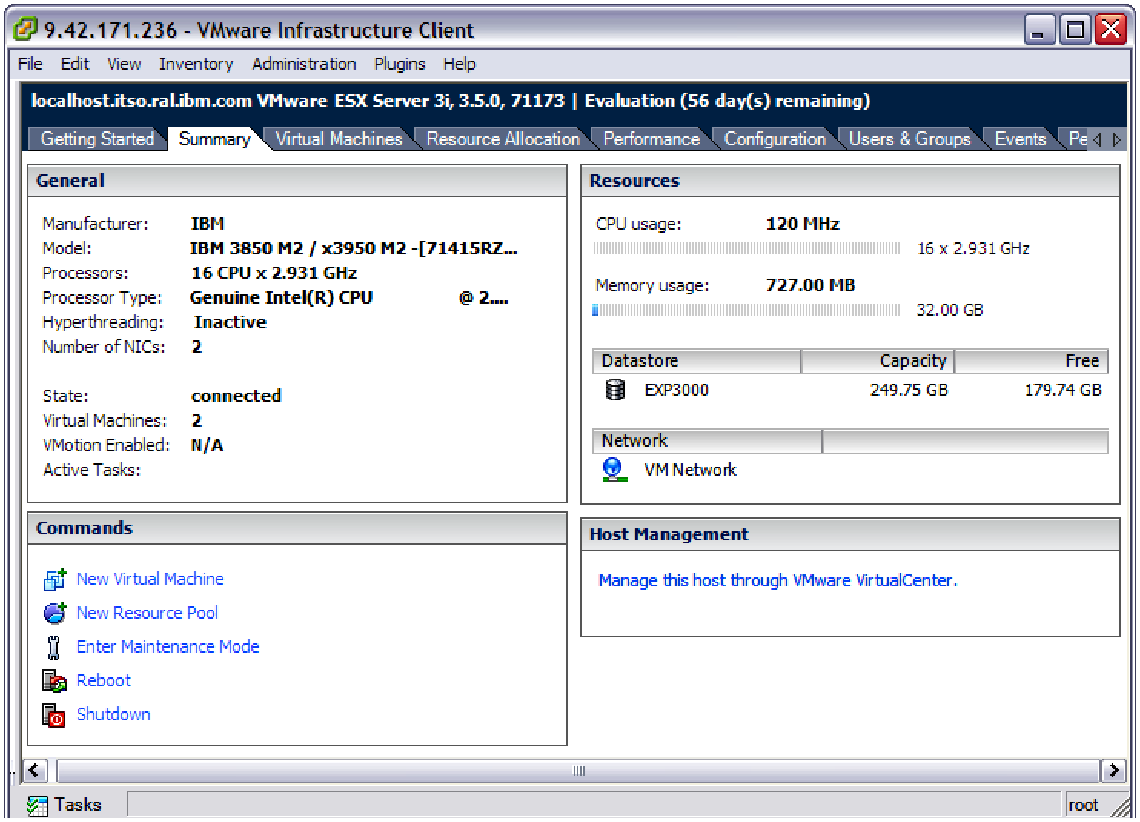
7. Test that the network interface settings are correctly configured to ensure that you can connect to the ESXi host for further configuration using VMware VI Client or VirtualCenter. By default, this test tries to ping your Default Gateway, DNS server IP addresses and performs DNS resolution of your ESXi host name. See Figure 5-19.



```
Customize System                     Test Management Network

Configure Root Password              To perform a brief network test,
Configure Lockdown Mode              press <Enter>.

Configure Management Network         By default, this test will attempt
Restart Management Network           to ping your default gateway, ping
Test Management Network              your primary and alternate DNS
                                     servers, and resolve your hostname.
Configure Keyboard
View Support Information
View System Logs

Restart Management Agents

Reset Customized Settings


<Up/Down> Select                     <Enter> Run Test        <Esc> Log Out
```

Figure 5-19 Testing of the management network settings

8. Most of the other settings, typically configured for ESX 3.5, can be also configured similarly by using VI Client, as shown in Figure 5-20.



localhost.itso.ral.ibm.com VMware ESX Server 3i, 3.5.0, 71173 | Evaluation (56 day(s) remaining)

General

Manufacturer:	IBM
Model:	IBM 3850 M2 / x3950 M2 -[71415RZ...
Processors:	16 CPU x 2.931 GHz
Processor Type:	Genuine Intel(R) CPU @ 2....
Hyperthreading:	Inactive
Number of NICs:	2
State:	connected
Virtual Machines:	2
VMotion Enabled:	N/A
Active Tasks:	

Resources

CPU usage: 120 MHz
16 x 2.931 GHz

Memory usage: 727.00 MB
32.00 GB

Datastore	Capacity	Free
EXP3000	249.75 GB	179.74 GB

Network	
VM Network	

Commands

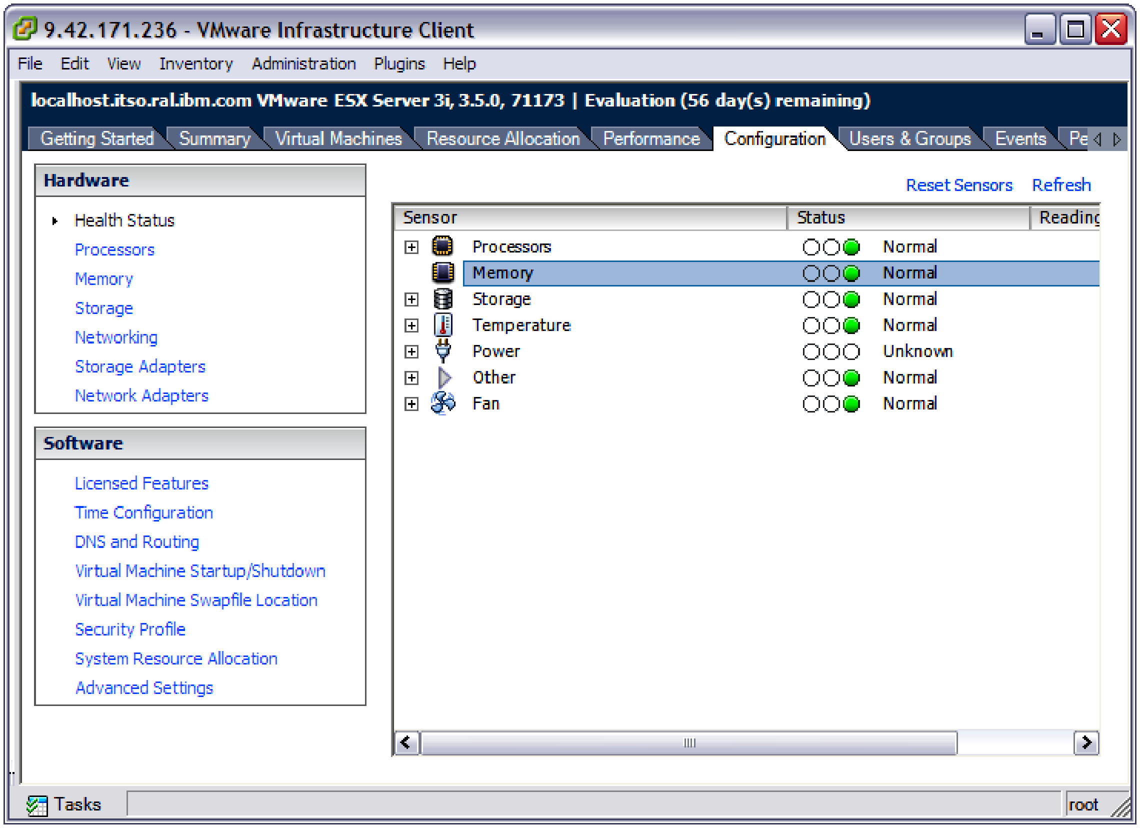
- New Virtual Machine
- New Resource Pool
- Enter Maintenance Mode
- Reboot
- Shutdown

Host Management

Manage this host through VMware VirtualCenter.

Figure 5-20 Summary tab after connecting to ESXi using VI Client

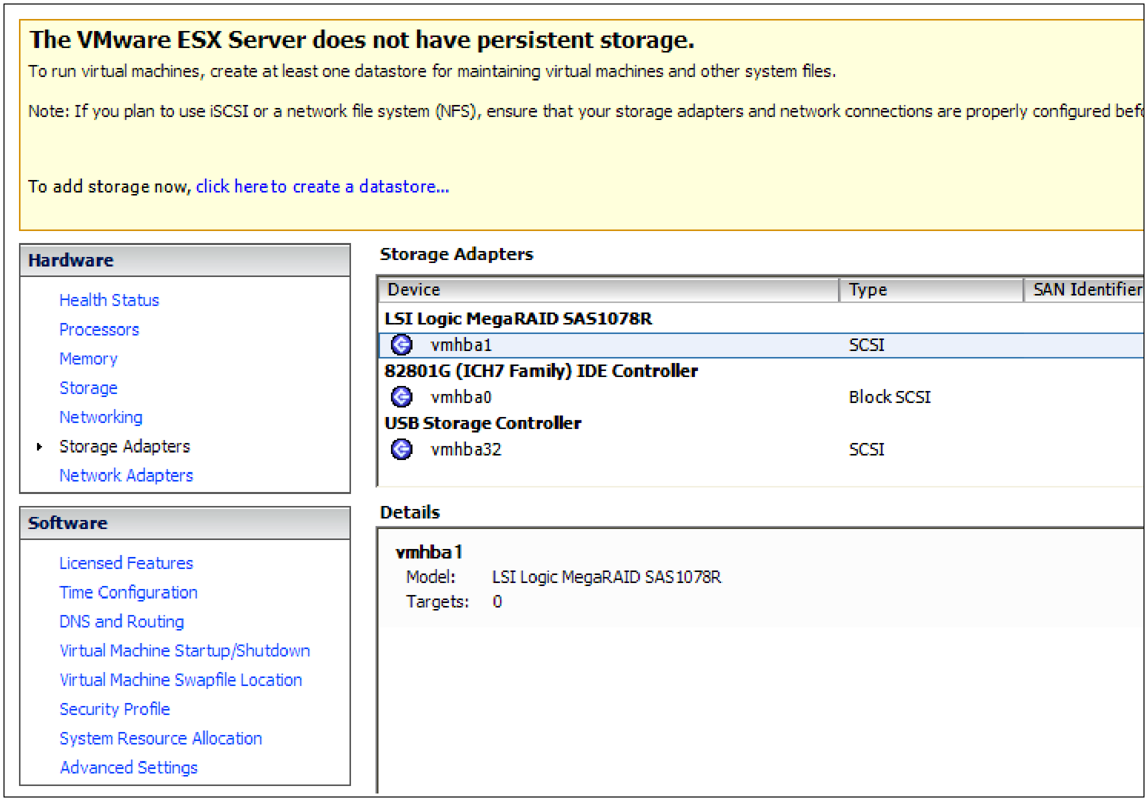
9. ESXi can interface with Common Information Model (CIM) through a set of standard APIs to provide basic processor, memory, storage, temperature, power, and fan hardware-level alerts. See Figure 5-21.


9.42.171.236 - VMware Infrastructure Client
File Edit View Inventory Administration Plugins Help
localhost.itso.ral.ibm.com VMware ESX Server 3i, 3.5.0, 71173 | Evaluation (56 day(s) remaining)
Getting Started Summary Virtual Machines Resource Allocation Performance Configuration Users & Groups Events Pe
Hardware
Health Status
Processors
Memory
Storage
Networking
Storage Adapters
Network Adapters
Software
Licensed Features
Time Configuration
DNS and Routing
Virtual Machine Startup/Shutdown
Virtual Machine Swapfile Location
Security Profile
System Resource Allocation
Advanced Settings
Reset Sensors
Refresh
Tasks
root

Sensor	Status	Reading
⊞ Processors	○○● Normal	
Memory	○○● Normal	
⊞ Storage	○○● Normal	
⊞ Temperature	○○● Normal	
⊞ Power	○○○ Unknown	
⊞ Other	○○● Normal	
⊞ Fan	○○● Normal	

Figure 5-21 The VI Client Configuration tab for ESXi host displaying Health Status basic alerts for the ESXi's server major subsystems

10. In the Hardware panel on the left of the Configurator tab in VI Client, check that ESXi can successfully detect each processor's sockets and cores, total memory installed, network adapters installed, and storage adapters installed. Figure 5-22 on page 275 shows the storage adapters detected.


The VMware ESX Server does not have persistent storage.

To run virtual machines, create at least one datastore for maintaining virtual machines and other system files.

Note: If you plan to use iSCSI or a network file system (NFS), ensure that your storage adapters and network connections are properly configured bef

To add storage now, click here to create a datastore...

Storage Adapters

Device	Type	SAN Identifier
LSI Logic MegaRAID SAS1078R		
vmhba1	SCSI	
82801G (ICH7 Family) IDE Controller		
vmhba0	Block SCSI	
USB Storage Controller		
vmhba32	SCSI	

Details

vmhba1

Model:	LSI Logic MegaRAID SAS1078R
Targets:	0

Figure 5-22 Configuration tab: Detecting Storage Adapters in a x3580 M2 with ESXi

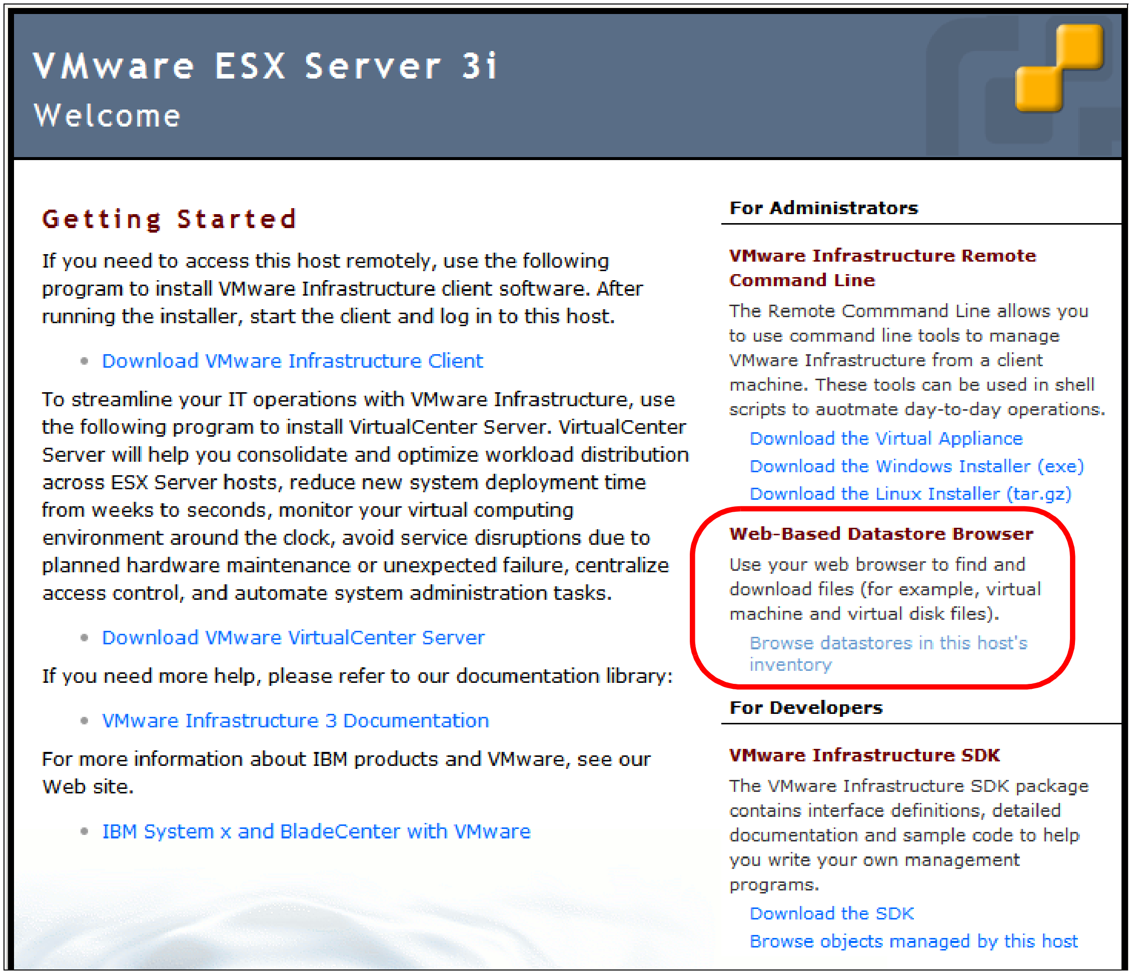
11. ESXi should be able to detect all internal hard drives in the x3850 M2 and any external hard drives connected through the x3850 M2's external SAS connector to EXP3000 enclosures. To confirm this:

    – Check Storage sensors in the Hardware panel of the Configurator tab, Hardware pane, under **Health Status**.

    – Check the ESXi Web Interface accessible by using the address format:

    ```
    https://<IP Address of ESXi host management interface>
    ```

    Click **Browse datastores in this host's inventory** link as shown in Figure 5-24 on page 277.



VMware ESX Server 3i

Welcome

Getting Started

If you need to access this host remotely, use the following program to install VMware Infrastructure client software. After running the installer, start the client and log in to this host.

- Download VMware Infrastructure Client

To streamline your IT operations with VMware Infrastructure, use the following program to install VirtualCenter Server. VirtualCenter Server will help you consolidate and optimize workload distribution across ESX Server hosts, reduce new system deployment time from weeks to seconds, monitor your virtual computing environment around the clock, avoid service disruptions due to planned hardware maintenance or unexpected failure, centralize access control, and automate system administration tasks.

- Download VMware VirtualCenter Server

If you need more help, please refer to our documentation library:

- VMware Infrastructure 3 Documentation

For more information about IBM products and VMware, see our Web site.

- IBM System x and BladeCenter with VMware

For Administrators

VMware Infrastructure Remote Command Line

The Remote Commmand Line allows you to use command line tools to manage VMware Infrastructure from a client machine. These tools can be used in shell scripts to auotmate day-to-day operations.

Download the Virtual Appliance
Download the Windows Installer (exe)
Download the Linux Installer (tar.gz)

Web-Based Datastore Browser

Use your web browser to find and download files (for example, virtual machine and virtual disk files).

Browse datastores in this host's inventory

For Developers

VMware Infrastructure SDK

The VMware Infrastructure SDK package contains interface definitions, detailed documentation and sample code to help you write your own management programs.

Download the SDK
Browse objects managed by this host

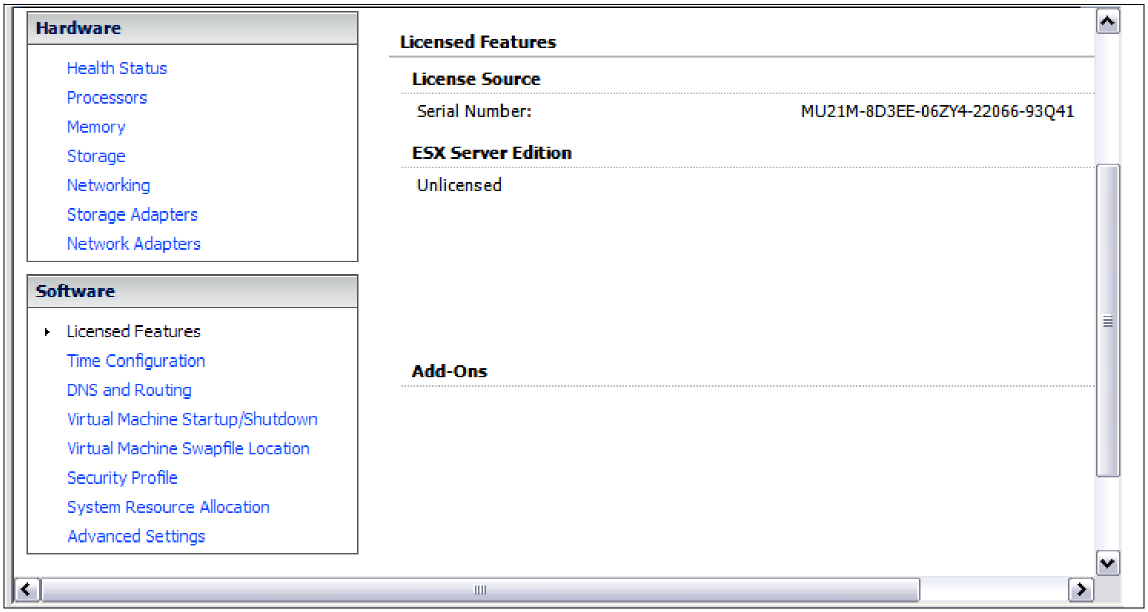
Figure 5-23 Web-Based Datastore Browser

12. ESXi has an ESXi Edition license (serial number) with limited hypervisor management features. See Figure 5-24 on page 277. Perform one of the following actions:

   – If you intend to use VirtualCenter to manage all your ESXi hosts under **Licensed Features** → **License** → **Source Edit**, configure the ESXi host's license source to point to the IP address of your designated License Server (this is typically configured before or during the installation of VMware VirtualCenter Server).

   – If you use host-based license files, configure the ESXi host's license source to point to a license file.

– If you have VI3 Enterprise, Standard, or Foundation license files, under Licensed Features, you can also enable the appropriate ESX Server Edition (such as Standard) and add-on features (HA, DRS, VCB) permitted by your license.



Licensed Features

License Source

Serial Number: MU21M-8D3EE-06ZY4-22066-93Q41

ESX Server Edition

Unlicensed

Add-Ons

Figure 5-24 Example of Serial Number license for ESXi

13. If you intend to use VirtualCenter to manage all your ESXi hosts, we recommend that you exit the VI Client sessions that are directly connected to the individual ESXi hosts and connect to your VirtualCenter server using VI Client (by using your VirtualCenter IP address, add the appropriate ESXi hosts (using DNS registered host names and ESXi host root password)) to VirtualCenter and make all configuration changes to ESXi hosts, HA/DRS Clusters through VirtualCenter.

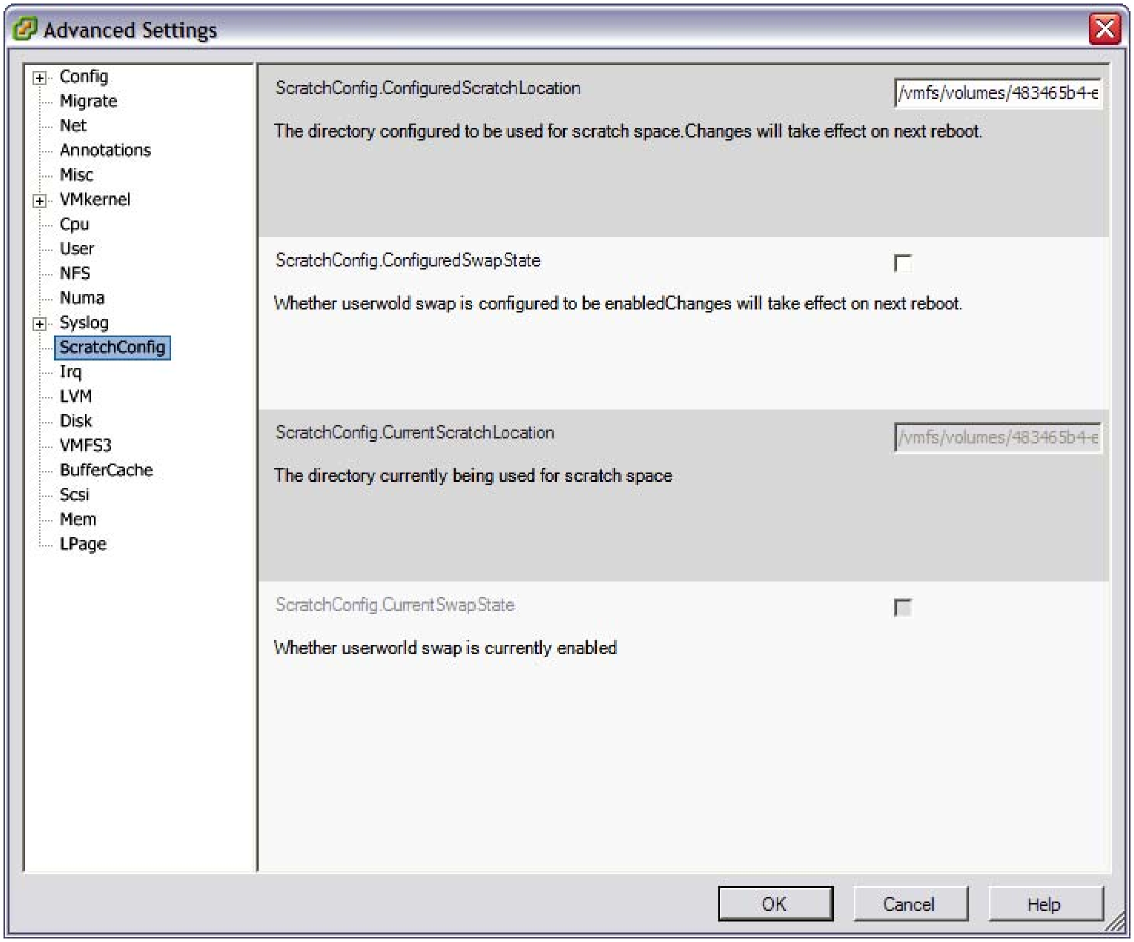
14. If you intend to use VMware VirtualCenter 2.5 Update 1 to manage ESXi hosts in a High Availability (HA) Cluster, you have to enable swap space on the ESXi hosts before adding them to an HA Cluster.

    To enable Swap:

    a. In the VirtualCenter Server, select the ESXi host.

    b. In the Configuration tab page, click **Advanced Settings**.

    c. Choose **ScratchConfig**. See Figure 5-25 on page 278.



ScratchConfig.ConfiguredScratchLocation /vmfs/volumes/483465b4-e

The directory configured to be used for scratch space.Changes will take effect on next reboot.

ScratchConfig.ConfiguredSwapState □

Whether userwold swap is configured to be enabledChanges will take effect on next reboot.

ScratchConfig.CurrentScratchLocation /vmfs/volumes/483465b4-e

The directory currently being used for scratch space

ScratchConfig.CurrentSwapState □

Whether userworld swap is currently enabled

Figure 5-25 Example of ScratchConfig settings to enable Swap space for ESXi host

d. Set the data store for **ScratchConfig.CurrentScratchLocation** to a valid directory with sufficient space (for example, 1 GB) to hold the userworld swap file. The userworld swap can be configured on local storage or shared storage.

   For example, the VMFS volume /vmfs/volumes/DatastoreName after reboot ESXi might not show DatastoreName but instead a unique volume ID.

e. Check the **ScratchConfig.ConfiguredSwapState** check box.

f. Click **OK**.

g. Reboot the ESXi host.

Note: At the time of writing, only VMware Virtual Center 2.5 Update 1 has support for ESXi in HA Clusters. The following restrictions apply when using VMware HA in conjunction with VMware ESX Server 3i hosts:

- Swap space must be enabled on individual ESX Server 3i hosts,
- Only homogeneous (non-mixed) clusters are supported at this time.

See the following knowledge base articles from VMware regarding ESXi:

- "Limited configurations are supported for VMware HA and ESX Server 3i hosts" (KB 1004656)

  http://kb.vmware.com/kb/1004656

- "ESXi 3 Hosts without swap enabled cannot be added to a VMware High Availability Cluster" (KB 1004177)

  http://kb.vmware.com/kb/1004177

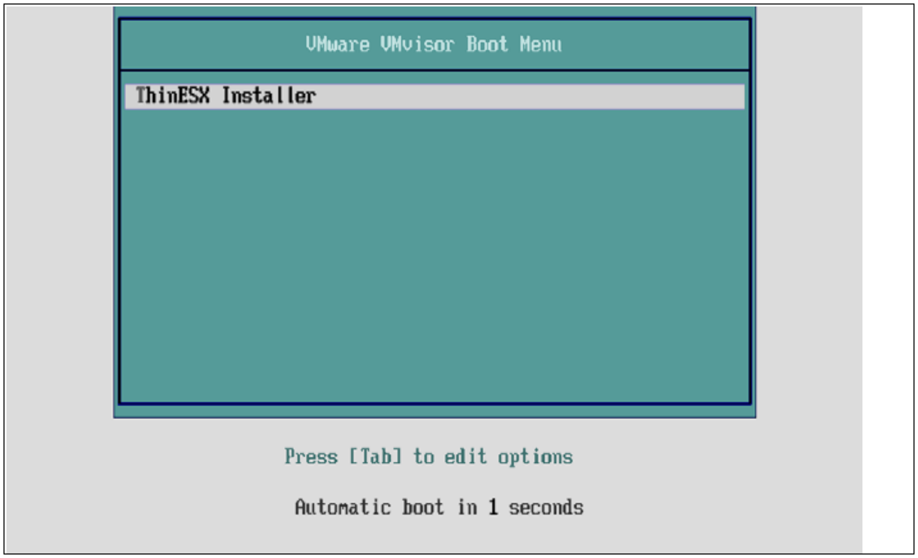
Detailed VMware ESX configuration guidance is beyond the scope of this book. See the VMware ESXi setup and configuration guidelines listed in "Prerequisites" on page 264.

5.4.2 Installing VMware ESXi 3.5 Installable

Use the ESXi Installable CD-ROM to install ESXi Installable onto a hard drive (SAS or SATA). This procedure presumes you are using a keyboard and monitor attached to the server to perform the installation.

To install VMware ESXi:

1. Configure any RAID arrays using the MegaRAID Manager tool.
2. Insert the ESX Server 3i Installable CD into the CD drive and boot the system.
3. When prompted during the boot process, press `F12` to boot from the CD. You see the boot menu, as shown in Figure 5-26 on page 280.

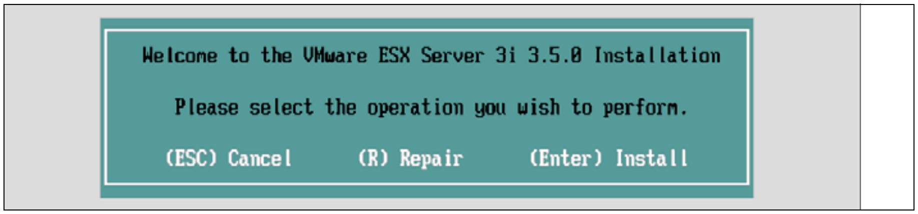


Press [Tab] to edit options

Automatic boot in 1 seconds

Figure 5-26 VMware VMvisor ESXi Installable Boot Menu

4. In the VMware VMvisor Boot Menu, select the **ThinESX Installer** and press `Enter` to boot the installer.
5. The ESX Server runs through its boot process until the Welcome window opens, as shown in Figure 5-27.

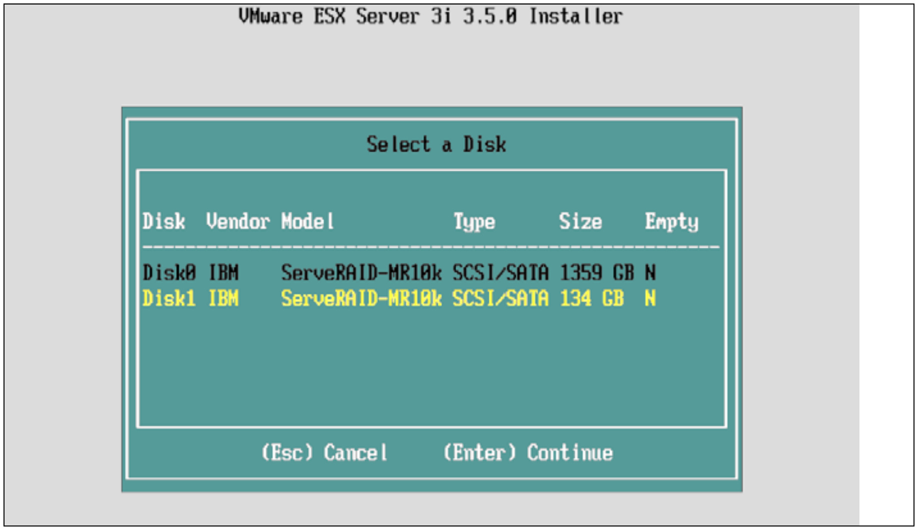

Welcome to the VMware ESX Server 3i 3.5.0 Installation

Please select the operation you wish to perform.

Figure 5-27 Welcome window for ESXi Installer

6. Press `Enter` to continue with the Installation.
7. If you accept the End User License Agreement, press `F11`.

8. Select the appropriate disk on which to install VMware ESXi Installable (see Figure 5-28) and press `Enter` to continue.



```
                VMware ESX Server 3i 3.5.0 Installer

                           Select a Disk

       Disk  Vendor Model            Type      Size    Empty
       -----------------------------------------------------
       Disk0 IBM    ServeRAID-MR10k SCSI/SATA 1359 GB N
       Disk1 IBM    ServeRAID-MR10k SCSI/SATA 134 GB  N

                  (Esc) Cancel      (Enter) Continue
```

Figure 5-28 Selecting the disk on which to install ESXi

9. When the installer completes the operation, the Installation Complete window opens.
10. Remove the CD and reboot the host.
11. After the host reboots, the direct console window opens and allows you to set up your ESX Server host configuration, as described in “Boot ESXi hypervisor and customize settings” on page 266.

For detailed information about installing ESX Server 3i, see the *ESX Server 3i Embedded Setup Guide*:

`http://www.vmware.com/pdf/vi3_35/esx_3i_e/r35/vi3_35_25_3i_setup.pdf`

5.4.3 Installing VMware ESX 3.5 Update 1

VMware ESX 3.5 is can be installed on local server storage or remote storage like a SAN LUN using a Fibre Channel Host Bus Adapter or iSCSI target using a iSCSI Host Bus Adapter.

Prerequisites

Before you install and configure ESX 3.5 Update 1:

- Configure any RAID arrays using the MegaRAID Manager tool.

- Ensure you have local or remote (through RSA II remote control) video and keyboard console access to the server onto which you will be installing.
- Ensure only intended local boot disk or remote bootable storage LUN is visible to VMware ESX 3.5 Update 1 installer; avoid installing on the wrong disk or remote storage disk (SAN LUNs to be used to create VMFS datastores) that is not intended for booting.
- If you are planning to boot from SAN/iSCSI disks, set the system BIOS boot start sequence to boot from the particular device. Consult your storage vendor and host bus adaptor setup and configuration guides to establish the appropriate settings for booting from the SAN/iSCSI remote storage.
- Check for storage system compatibility in *Storage / SAN Compatibility Guide For ESX Server 3.5 and ESX Server 3i* (including support for boot from SAN):

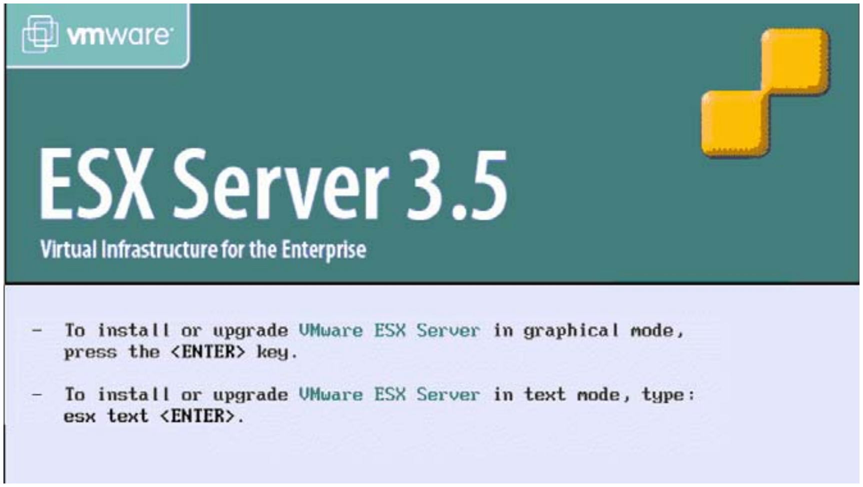
  http://www.vmware.com/pdf/vi35_san_guide.pdf
- Download the VMware ESX 3.5 installation and configuration guides:

  http://www.vmware.com/pdf/vi3_35/esx_3/r35/vi3_35_25_installation_guide.pdf
  http://www.vmware.com/pdf/vi3_35/esx_3/r35/vi3_35_25_3_server_config.pdf

Install ESX 3.5 Update 1

To install ESX 3.5 Update 1:

1. Boot the VMware ESX 3.5 Update 1 CD or ISO file (mounted using Remote Media using the RSA II **Remote Control** → **Mount Drive** feature). See Figure 5-29.



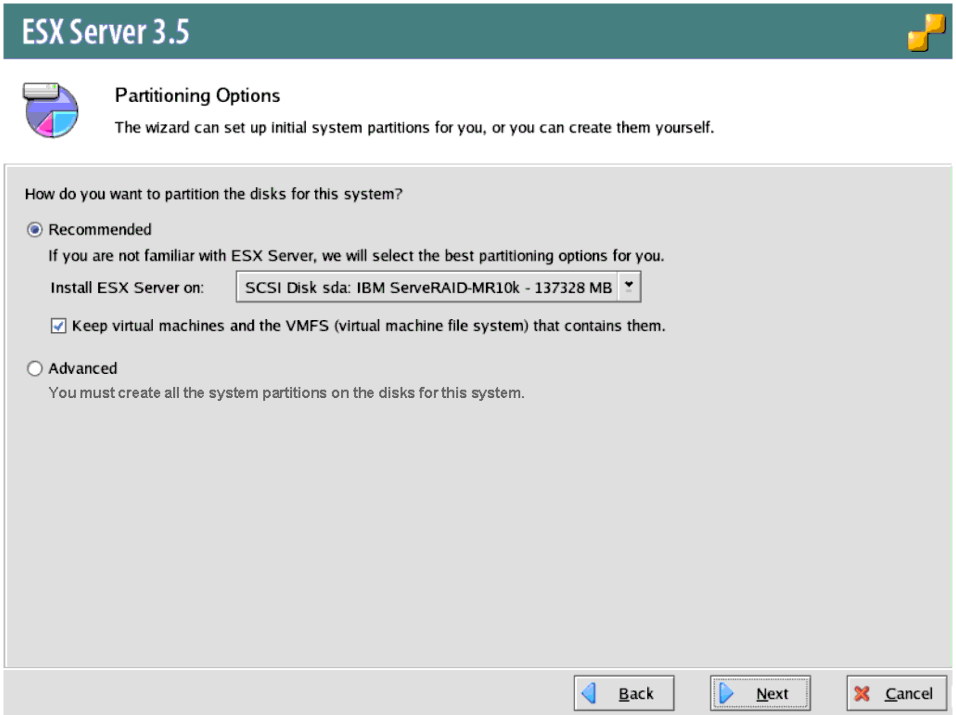
- To install or upgrade VMware ESX Server in graphical mode, press the <ENTER> key.

- To install or upgrade VMware ESX Server in text mode, type: esx text <ENTER>.

Figure 5-29 Booting from VMware ESX 3.5 Update 1 from CD or ISO file

2. Select the keyboard and mouse as indicated and click **Next**.

3. If you accept the End User License Agreement, click **Next**.
4. For Partitioning Options, Figure 5-30, select the available local or remote disk (bootable SAN LUN) on which to install ESX 3.5 Update 1. If multiple disks are available, be careful to select the correct disk to install ESX 3.5 Update 1, otherwise the server might not boot correctly into ESX 3.5 Update 1 hypervisor.



Partitioning Options

The wizard can set up initial system partitions for you, or you can create them yourself.

How do you want to partition the disks for this system?

● Recommended

If you are not familiar with ESX Server, we will select the best partitioning options for you.

Install ESX Server on: SCSI Disk sda: IBM ServeRAID-MR10k - 137328 MB

☑ Keep virtual machines and the VMFS (virtual machine file system) that contains them.

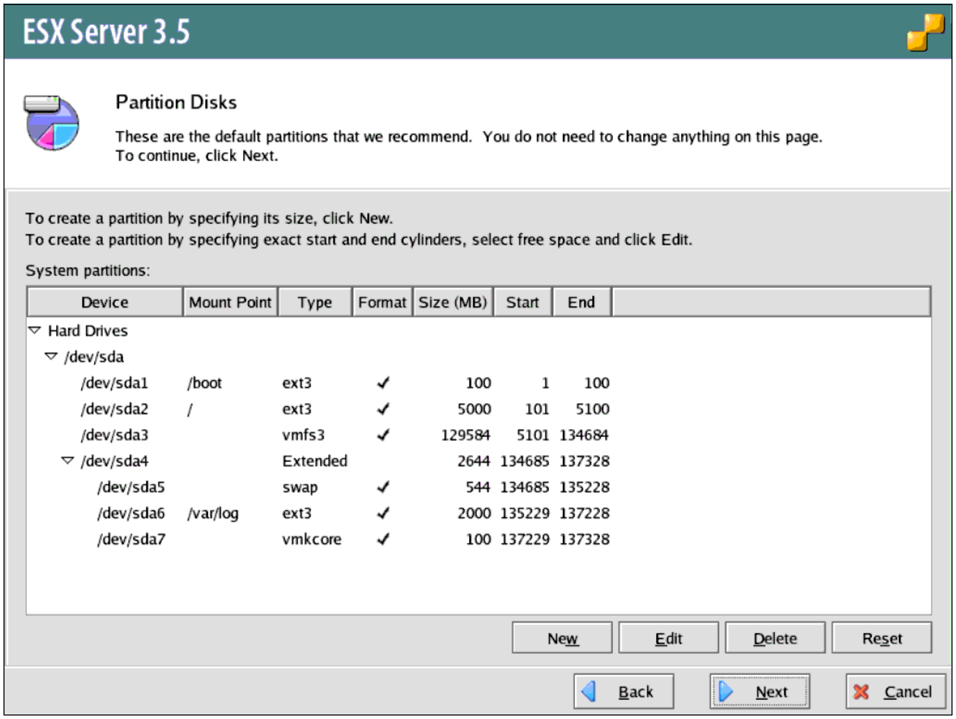
○ Advanced

You must create all the system partitions on the disks for this system.

Figure 5-30 ESX 3.5 Update 1 Partitioning Options

5. Click **Next** to continue with the installation.

6. Default partitions are provided by the ESX 3.5 Update 1 installer. You can change the default partition layout if required, as shown in Figure 5-31. Otherwise, accept the defaults and click **Next** to continue.



Partition Disks

These are the default partitions that we recommend. You do not need to change anything on this page. To continue, click Next.

To create a partition by specifying its size, click New.
To create a partition by specifying exact start and end cylinders, select free space and click Edit.

System partitions:

Device	Mount Point	Type	Format	Size (MB)	Start	End
▽ Hard Drives						
▽ /dev/sda						
/dev/sda1	/boot	ext3	✓	100	1	100
/dev/sda2	/	ext3	✓	5000	101	5100
/dev/sda3		vmfs3	✓	129584	5101	134684
▽ /dev/sda4		Extended		2644	134685	137328
/dev/sda5		swap	✓	544	134685	135228
/dev/sda6	/var/log	ext3	✓	2000	135229	137228
/dev/sda7		vmkcore	✓	100	137229	137328

Figure 5-31 Choosing the Partition Layout for ESX 3.5 Update 1.

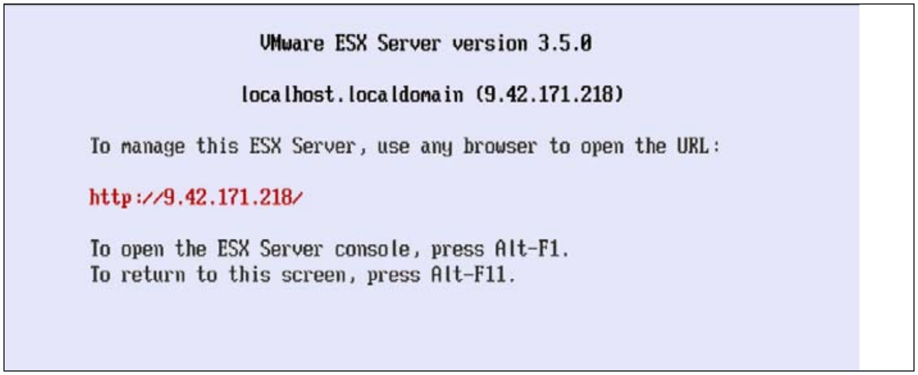
The default boot loader setting from VMware ESX 3.5 Update 1 installer configures the boot loader to be installed on the MBR of the disk, which you selected earlier, on which to install VMware ESX 3.5 Update 1.

7. Network Configuration enables you to select the available Network Adapters to use for the default vSwitch and vSwif0 interface for shared Virtual Machine and Service Console access. You may modify the physical Network Adapter that is configured for the vSwitch and the vSwif interface.
8. Select the time zone as appropriate to your location.
9. Under Account Configuration, you may configure the Root Password and add new users to the ESX 3.5 Udpate 1 system.
10. In the summary window of all the selections you have made during the installation process, confirm the selections and click **Next** to begin the installation.

11. The installer formats the selected local or remote storage you selected earlier and transfer or load the ESX 3.5 Update 1 hypervisor image to the selected disks.

12. Upon completion of the installation, click **Finish** and reboot the server.

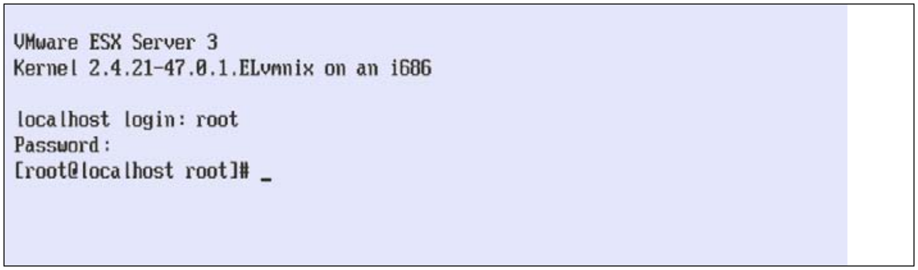
    The ESX 3.5 Update 1 Hypervisor will boot into the ESX 3.5 startup window with the Service Console IP Address as shown in Figure 5-32.



```
                    VMware ESX Server version 3.5.0

                  localhost.localdomain (9.42.171.218)

     To manage this ESX Server, use any browser to open the URL:

     http://9.42.171.218/

     To open the ESX Server console, press Alt-F1.
     To return to this screen, press Alt-F11.
```

Figure 5-32 ESX 3.5 Update 1 Hypervisor startup window

13. Press `Alt+F1` to get the service console logon prompt (Figure 5-33); to return to the startup window press `Alt+F11`.



```
VMware ESX Server 3
Kernel 2.4.21-47.0.1.ELvmnix on an i686

localhost login: root
Password:
[root@localhost root]# _
```

Figure 5-33 Logging on to ESX 3.5 Update 1 service console

14. To logon to the ESX 3.5 Service Console, type root and the password you set earlier.

5.4.4 Installing Windows Server 2003

This section describes the key aspects of installing Windows Server 2003 on the x3950 M2.

Prerequisites

Before you install and configure Windows Server 2003:

- Configure any RAID arrays using the MegaRAID Manager tool.
- Ensure you have local or remote (through RSA II remote control) video and keyboard console access to the server onto which you will be installing.
- Download the appropriate Windows Server 2003 installation instructions:
  - Installing Windows Server 2003 (32-bit) on x3950 M2 and x3850 M2

    `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5075238`

  - Installing Windows Server 2003 x64 on x3950 M2 and x3850 M2

    `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5075237`

- If you are using a ServeRAID MR10k SAS/SATA RAID controller or the onboard LSI 1078 integrated RAID controller, download the latest drivers from:

  `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073138`

- Download the Broadcom NetXtreme II 5709 drivers from:

  `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5070012`

- Download the Intel-based Gigabit Ethernet drivers from:

  `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5070807`

- In BIOS, ensure the following parameters are set:
  - In CPU Options, ensure that the Clustering Technology parameter is set to **Logical Mode**, as shown in Figure 3-9 on page 100.
  - In **Advanced Settings** → **RSA II Settings**, ensure that the OS USB Selection setting is set to **Other OS**, as shown in Figure 6-14 on page 320.
- If you plan to install Windows using a regular Windows installation CD, ensure you have a USB diskette drive to supply the necessary disk controller device drivers.
- If you plan to boot from the internal SAS drives and will be using the Windows installation CD-ROM, press `F6` when you see:

  ```
  Setup is inspecting your computer's hardware configuration
  ```

  Insert the disk controller driver diskette when prompted. Create the diskette from:

  `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073138`

  > **Tip:** If you are using ServerGuide™ to install Windows, you do not have to obtain these drivers separately.

If you do not have a USB diskette drive, you can mount a remote diskette drive using the RSA II remote media function.

a. Select the diskette drive **A**.

b. Select **No** when asked if you want to upload the diskette image to the RSA II adapter.

c. Browse to the driver diskette.

d. Click **Mount drive**.

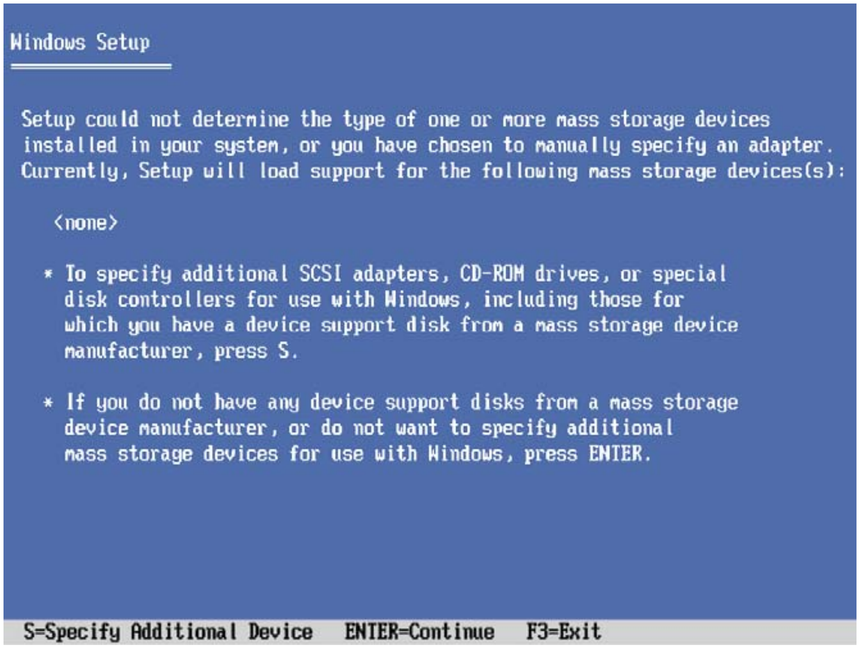
Install Windows Server 2003

To install Windows Server 2003:

1. Boot the server from Windows Server 2003 Installation CD, DVD, or ISO file.
2. Press `F6` to load additional storage device drivers when you see the following text:

   ```
   Setup is inspecting your computer's hardware configuration
   ```

   The Windows Setup window opens (see Figure 5-34).



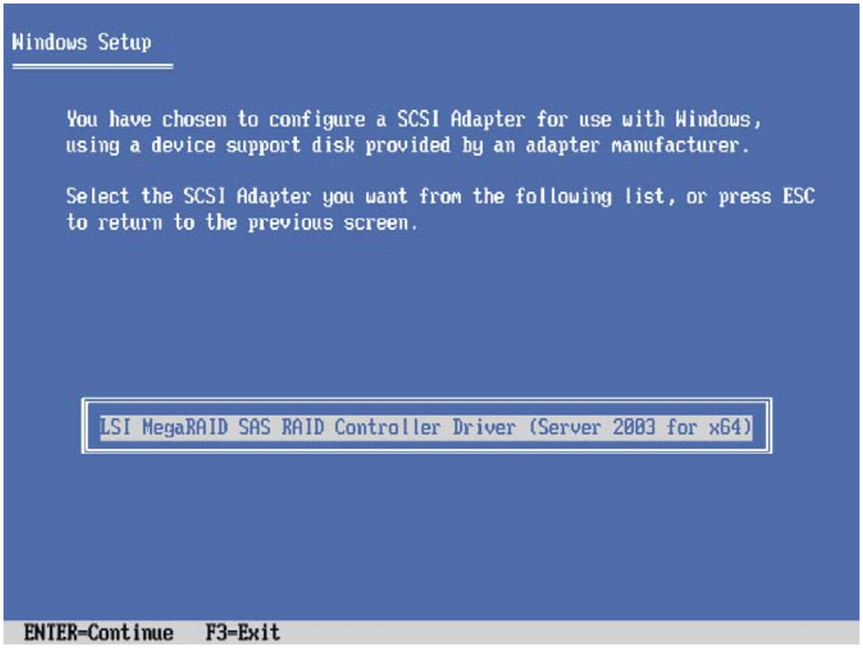
```
Windows Setup

 Setup could not determine the type of one or more mass storage devices
 installed in your system, or you have chosen to manually specify an adapter.
 Currently, Setup will load support for the following mass storage devices(s):

    <none>

   * To specify additional SCSI adapters, CD-ROM drives, or special
     disk controllers for use with Windows, including those for
     which you have a device support disk from a mass storage device
     manufacturer, press S.

   * If you do not have any device support disks from a mass storage
     device manufacturer, or do not want to specify additional
     mass storage devices for use with Windows, press ENTER.

 S=Specify Additional Device   ENTER=Continue   F3=Exit
```

Figure 5-34 Specify additional storage device drivers

3. Press **S** (Specify Additional Device) to select appropriate storage device drivers.

4. Insert the driver diskette for the disk controller. See Figure 5-35. The LSI MegaRAID SAS RAID Controller driver files for the ServeRAID MR10k SAS/SATA Controller should include the following files:

   – msas2k3.cat
   – msas2k3.sys
   – nodev.inf
   – oemsetup.inf
   – txtsetup.oem



```
Windows Setup
══════════════

     You have chosen to configure a SCSI Adapter for use with Windows,
     using a device support disk provided by an adapter manufacturer.

     Select the SCSI Adapter you want from the following list, or press ESC
     to return to the previous screen.



        LSI MegaRAID SAS RAID Controller Driver (Server 2003 for x64)



  ENTER=Continue   F3=Exit
```

Figure 5-35 Windows Setup detects the LSI MegaRAID SAS RAID Controller Driver

5. Proceed with the normal installation steps to complete the installation of Windows Server 2003.

6. After he installation finishes and the server boots into the partition on which Windows Server 2003 has successfully installed, log on as Administrator. Proceed to install additional drivers for any other adapters you might have installed in the x3850 M2 or x3950 M2 (single and multinode), as detailed in the next section.

Post-installation

The key points to the installation are as follows:

- After installation, install additional drivers. Consult the post-install steps in the installation instructions (guides listed in “Prerequisites” on page 286). In addition, install:
  - The OSA IPMI driver (for the BMC)
  - The RSA II driver
  - Onboard Broadcom NetXtreme II 5709 drivers
  - Drivers and firmware updates for additional adapters you might have installed, such as Fibre Channel HBAs, Network Adapters, and others.
- If you are installing the 32-bit version of Windows Server 2008 and you have more than 4 GB of RAM installed, you should add the /PAE switch to the boot.ini file after installation is complete, so that the operating system can access the memory about the 4 GB line (see the last line in Example 5-4).

Example 5-4 boot.ini for accessing more than 4 GB memory

```
[boot loader]
timeout=3
default=multi(0)disk(0)rdisk(1)partition(1)\WINDOWS
[operating systems]
multi(0)disk(0)rdisk(1)partition(1)\WINDOWS="Windows Server 2003, Enterprise" /fastdetect /PAE
```

5.4.5 Installing Windows Server 2008

This section describes the key aspects of installing Windows Server 2008 on the x3950 M2.

Prerequisites

Before you install and configure Windows Server 2008:

- Configure any RAID arrays using the MegaRAID Manager tool.
- Ensure you have local or remote (through RSA II remote control) video and keyboard console access to the server you will be installing onto.
- Download the appropriate Windows Server 2008 installation instructions:
  - Installing Windows Server 2008 (32-bit) on x3950 M2 and x3850 M2:

    http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5074895
  - Installing Windows Server 2008 x64 on x3950 M2 and x3850 M2:

    http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5074896

- No specific storage device drivers are required for installing Windows Server 2008 on the x3850 M2 and x3950 M2. Windows Server 2008 already has the necessary drivers to detect any logical disk created using the ServeRAID MR10k or LSI 1078 Integrated SAS/SATA RAID Controllers.

  However, it is good practice to download the latest ServeRAID MR10k SAS/SATA RAID controller or the onboard LSI 1078 integrated RAID controller from:

  `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073138`

- Download the Broadcom NetXtreme II 5709 Gigabit Ethernet drivers from:

  `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5070012`

- Download the Intel-based Gigabit Ethernet drivers from:

  `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5070807`

- In BIOS, ensure the following parameters are set:

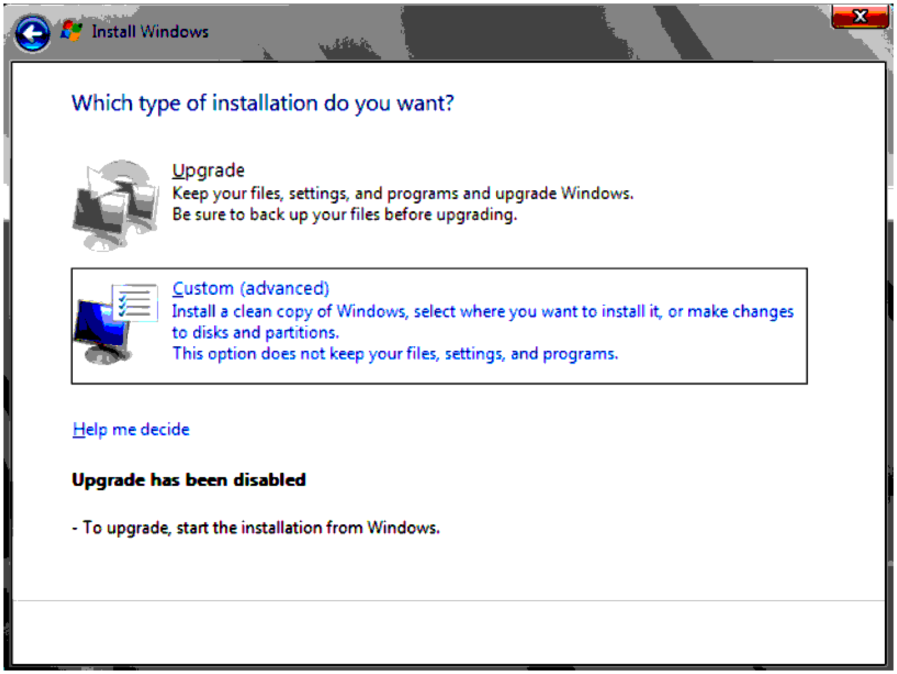
  - In CPU Options, ensure that the Clustering Technology parameter is set to **Logical Mode**, as shown in Figure 3-9 on page 100.
  - In **Advanced Settings** → **RSA II Settings**, ensure that the OS USB Selection setting is set to **Other OS**, as shown in Figure 6-14 on page 320

- A USB diskette drive is not required because Windows Server 2008 already includes boot drives for the IBM ServeRAID MR10k and LSI1078 disk controllers.

Install Windows Server 2008

To install Windows Server 2008:

1. Boot the server from Windows Server 2003 Installation CD/DVD or ISO file.
2. Select **Install Now**.
3. Enter the Product Key for the version of Windows Server 2008 you purchased; the installer automatically detects which version of Windows Server 2008that the key is valid for and begins installing that version of Windows Server 2008.
4. If you accept the Microsoft Software License terms, click **Next**.
5. Select **Custom (advanced)** to install a clean copy of Windows as shown in Figure 5-36 on page 291.



Which type of installation do you want?

Upgrade
Keep your files, settings, and programs and upgrade Windows.
Be sure to back up your files before upgrading.

Custom (advanced)
Install a clean copy of Windows, select where you want to install it, or make changes to disks and partitions.
This option does not keep your files, settings, and programs.

Help me decide

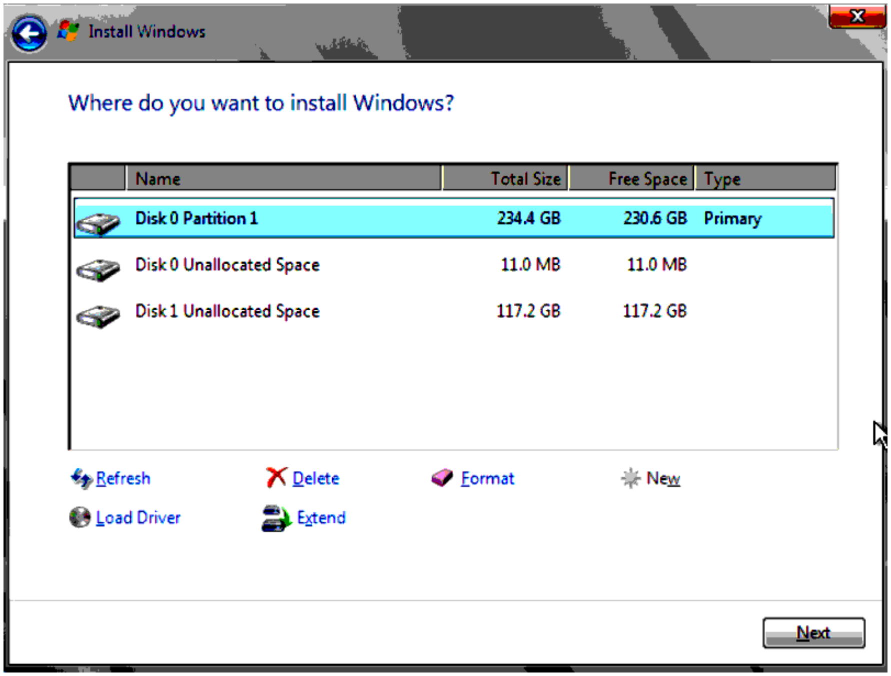
Upgrade has been disabled

- To upgrade, start the installation from Windows.

Figure 5-36 Select type of installation: Upgrade or Custom (clean installation)

6. The Windows Server 2008 installer should detect the logical disks configured through the ServeRAID MR10k or onboard LSI 1078 integrated RAID controllers.

   If you are installing on remote storage (for example SAN LUN), you might have to click **Load Driver** and select the appropriate storage host bus adapter driver from diskette drive A:



Where do you want to install Windows?

Name	Total Size	Free Space	Type
Disk 0 Partition 1	234.4 GB	230.6 GB	Primary
Disk 0 Unallocated Space	11.0 MB	11.0 MB	
Disk 1 Unallocated Space	117.2 GB	117.2 GB	

Figure 5-37 Selecting the disk on which to install Windows Server 2008

7. Continue with the installation process until completed.
8. After Windows Server 2008 boots from the disk partition you selected during the installation process, log on as Administrator and follow the instructions listed in the next section to install additional updated devices drivers for any other devices you might have installed in the x3850 M2 or x3950 M2 (single and multinode).

Post-installation information

The key points to the installation are as follows:

- After installation, install additional drivers. Consult the post-install steps in the installation instructions (see list in “Prerequisites” on page 289). In addition, install the following drivers:
  - OSA IPMI driver (for the BMC)
  - RSA II driver
  - Onboard Broadcom NetXtreme II 5709 drivers
  - Drivers and the latest firmware for additional adapters you might have installed, such as Fibre Channel HBAs, network adapters, and others.
- If you are installing a 32-bit version of Windows and you have more than 4 GB of RAM installed, add the /PAE switch to the boot.ini file once installation is complete, so that the operating system can access the memory about the 4 GB line (see the last line in Example 5-5).

Example 5-5 boot.ini for accessing more than 4 GB memory

```
[boot loader]
timeout=3
default=multi(0)disk(0)rdisk(1)partition(1)\WINDOWS
[operating systems]
multi(0)disk(0)rdisk(1)partition(1)\WINDOWS="Windows Server 2008" /fastdetect /PAE
```

5.4.6 Installing Red Hat Enterprise Linux 5 Update 1

This section describes the key aspects of installing RHEL 5 Update 1.

Prerequisites

Before you install and configure Red Hat Enterprise Linux 5 Update 1:

- Configure any RAID arrays using the MegaRAID Manager tool.
- Ensure you have local or remote (through RSA II Remote Control) video and keyboard console access to the server you will be installing onto.
- Download the RHEL 5 Update 1 installation guide:

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5074155
- If you are planning to boot from the internal SAS drives or remote storage (for example SAN LUN) and will be installing using the RHEL5 Update 1 installation CD/DVD-ROM, you do not have to insert a driver diskette for SAS controller drivers for local SAS drives, or host bus adapter drivers for remote

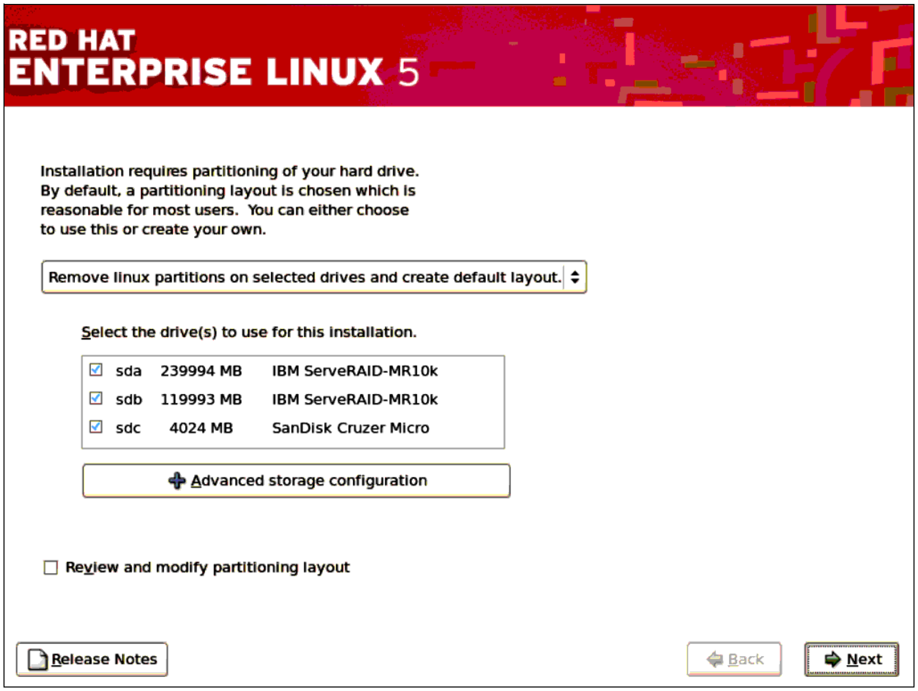
storage. The required drivers should already be included in your RHEL5 Update 1 installation CD/DVD, or ISO file.

If you have the RSA II adapter installed in the server, you can also install RHEL5 Update 1 remotely using the remote console and remote media functions of the RSA II. Select the RHEL5 Update 1 CD/DVD (for example `D:` in your windows management workstation or a mount point if you are managing the RSA from a linux management workstation) or choose **Select File** and browse to the CD/DVD, ISO file second and click the **>>** button followed by the **Mount Drive** button.

Install Red Hat Enterprise Linux 5 Update 1

To install Red Hat Enterprise Linux 5 Update 1:

1. Boot the server from the Red Hat Enterprise Linux 5 Update 1 DVD, CD, or ISO file.
2. When prompted, press `Enter` to install in graphical mode.
3. Enter the serial number.
4. Select the partitions (drives) to use for the installation.



Installation requires partitioning of your hard drive. By default, a partitioning layout is chosen which is reasonable for most users. You can either choose to use this or create your own.

Remove linux partitions on selected drives and create default layout.

Select the drive(s) to use for this installation.

☑	sda	239994 MB	IBM ServeRAID-MR10k
☑	sdb	119993 MB	IBM ServeRAID-MR10k
☑	sdc	4024 MB	SanDisk Cruzer Micro

□ Review and modify partitioning layout

Figure 5-38 Red Hat Enterprise Linux 5 Update 1 select partition window

5. Select the network device you want to use, click **Next**, and then select your time zone information and root password.

Post-Installation Information

If you are using a 32-bit versions of Red Hat Enterprise Linux 5, you must install the kernel-PAE (Physical Address Extension) kernel to detect more than 4 GB of memory.

Support for the Trusted Platform Module

The System x3950 M2 and x3805 M2 include a Trusted Platform Module (TPM) for added data security. TrouSerS and tpm-tools are planned to be included as a Technology Preview in RHEL 5 Update 2 to enable use of TPM hardware.

TPM hardware features include:

- Creation, storage, and use of RSA keys securely (without being exposed in memory)
- Verification of a platform's software state using cryptographic hashes

TrouSerS is an implementation of the Trusted Computing Group's Software Stack (TSS) specification. You can use TrouSerS to write applications that make use of TPM hardware. The tpm-tools is a suite of tools used to manage and utilize TPM hardware.

For more information about TrouSerS, refer to:

http://trousers.sourceforge.net/

5.4.7 Installing SUSE Linux Enterprise Server 10 SP1

This section describes the key aspects of installing SLES 0 SP1.

Prerequisites

Before you install and configure SUSE Linux Enterprise Server 10 SP1:

- Configure any RAID arrays using the MegaRAID Manager tool.
- Ensure you have local or remote (through RSA II Remote Control) video and keyboard console access to the server onto which you will be installing.
- Download the following SLES 10 SP1 installation guide:

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073088
- If you plan to boot from the internal SAS drives or remote storage (for example SAN LUN) and will be installing using the SLES10 Service Pack 1 installation CD/DVD-ROM, you do not have to insert a driver diskette for SAS

controller drivers for local SAS drives or host bus adapter drivers for remote storage. The required drivers should already be included in your SLES10 Service Pack 1 installation CD/DVD or ISO file.

If you have the RSA II adapter installed in the server, you can also install SLES10 Service Pack 2 remotely using the remote console and remote media functions of the RSA II. Select the SLES10 Service Pack 1 CD/DVD (for example `D:` in your windows management workstation or a mount point if you manage the RSA from a linux management workstation) or choose **`Select File...`** and browse to the CD/DVD ISO file second and click the **>>** button, then click **`Mount Drive`** button.

Install SUSE Enterprise Linux 10 SP1

To install SUSE Enterprise Linux 10 SP1:

1. Boot the server from the SUSE Enterprise LInux 10 SP1 Installation CD/DVD, or ISO file. See Figure 5-39.

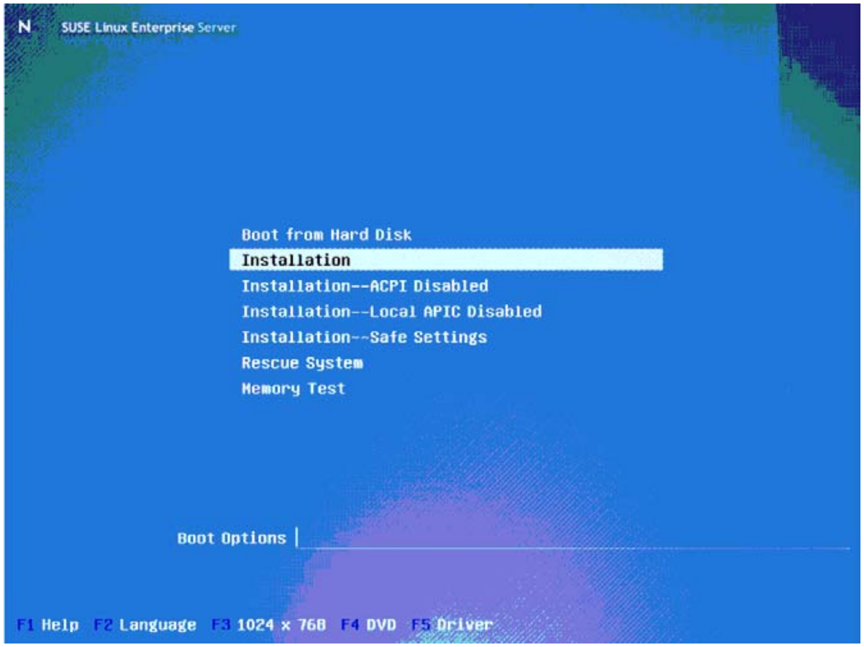
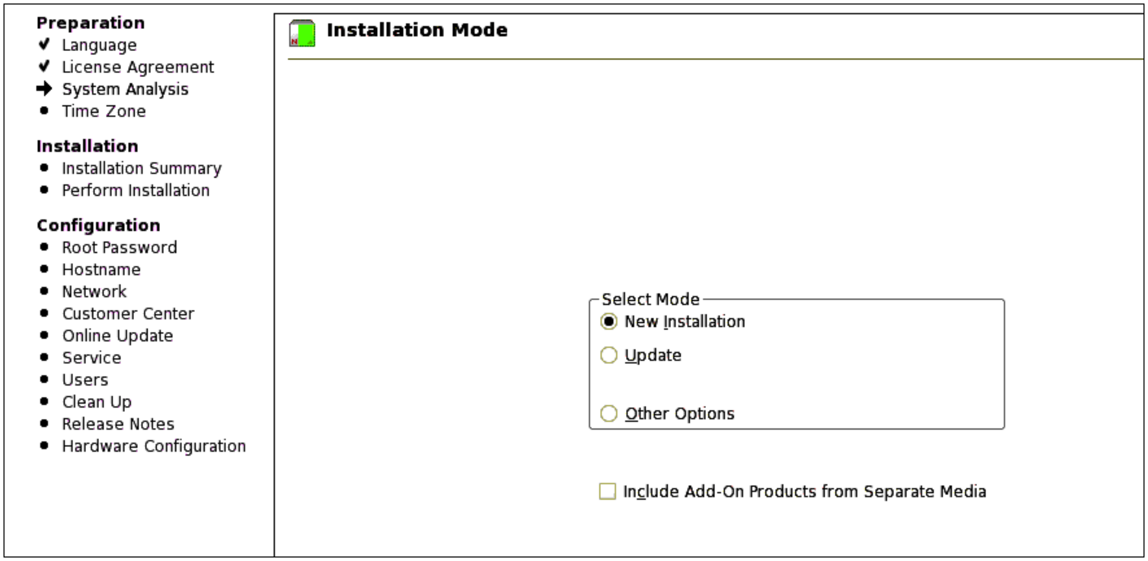

N
SUSE Linux Enterprise Server
Boot from Hard Disk
Installation
Installation--ACPI Disabled
Installation--Local APIC Disabled
Installation--Safe Settings
Rescue System
Memory Test
Boot Options
F1 Help
F2 Language
F3 1024 x 768
F4 DVD
F5 Driver

Figure 5-39 SUSE Enterprise Linux Server 10 installer start-up window

2. Select your language and accept the license agreement.
3. Select the type of installation you want to use: new or update. See Figure 5-40 on page 297.



Installation Mode

Select Mode

● New Installation

○ Update

○ Other Options

□ Include Add-On Products from Separate Media

Figure 5-40 SUSE Enterprise Linux 10 Installation Mode window

4. Continue with the installation process. Refer to the installation instructions for additional information.

Post-Installation

If you are using a 32-bit version of SUSE Enterprise Linux 10, then install the kernel-PAE kernel to access more than 4 GB of memory.

6

Management

This section explains system management capabilities of the x3850 M2 and x3950 M2 servers. It addresses the various subsystems, which are implemented to help you manage and service your servers, and also describes embedded features to help you reduce energy costs.

This chapters discusses the following topics:

- 6.1, “BMC configuration options” on page 300
- 6.2, “Remote Supervisor Adapter II” on page 316
- 6.3, “Use of IBM Director with VMware ESX” on page 334
- 6.4, “Active Energy Manager” on page 334
- 6.5, “IBM Director: Implementation of servers” on page 346
- 6.6, “System management with VMware ESXi 3.5” on page 355
- 6.7, “Power management” on page 356
- 6.8, “Power Distribution Units (PDU)” on page 357
- 6.9, “DSA Preboot” on page 366

6.1 BMC configuration options

The Baseboard Management Controller (BMC) in the x3850 M2 and x3950 M2 systems is a service processor based on the Hitachi 2166 chip that complies with the Intelligent Platform Management Interface, version 2.0 (IPMI v2.0) specification. The Intel specification document is available at:

http://www.intel.com/design/servers/ipmi/ipmiv2_0_rev1_0_markup_2.pdf

The BMC stores information related to up to 512 events. After 512 events have been stored, the log must be cleared before further events are recorded. A first-in-first-out algorithm is not used here. The time it takes to fill the BMC log area depends of the kind of events that have occurred.

Tip: We recommend you save the BMC log before clearing it, in case the service history is required. The logged events in the BMC can be transferred in-band to the Remote Supervisor Adapter II (RSA II) or to the operating system through a driver.

The BMC communicates with the RSA II through the Intelligent Platform Management Bus (IPMB) and controls the components about a whole set of standardized and system-specific sensors. The components communicate with the BMC through an embedded Inter-Integrated Circuit (I^2C) bus interface.

Logged events that are defined as errors are reported, as follows:

- Into the RSA II event data store
- Through the BMC driver interface into the operation system and management applications
- Through error indicator LEDs on different components in the server
- By illuminating of the respectively error LED on the Light Path Diagnostic (LPD) panel, which you can pull out at the front of the x3850 M2 and x3950 M2 servers
- By IPMI messaging over LAN

The BMC protects the server and powers the server off or prevents DC power-on in any of the following circumstances:

- At temperature of out-of-specification
- Invalid or wrong component installation
- Power voltage faults

The server is reset if any of the following events occur:

- Software non-maskable interrupt (NMI)
- Service processor interrupt (SPINT) routine was started
- Internal error (IERR)
- Automatic boot recovery (ABR) request is received

The BMC and the x3850 M2 and x3950 M2 baseboard management firmware enable the following features:

- Environmental monitoring for:
  - Fans
  - Temperature
  - Power supply
  - Disk drives
  - Processor status
  - NMI detection
  - Cable, card, component presence
  - Voltages and Power Good settings for battery and system components
- System LED Control (power, HDD activity, error, and others)
- Fan speed control
- Power, and reset control
- Interface to subassemblies to provide the vital product data (VPD), which contains information about components and the system, such as:
  - Field replacement unit (FRU) information
  - System firmware code levels like BIOS or BMC
- Non-maskable interrupt (NMI), system management interrupt (SMI) generation
- Button handling (Power, Reset, Locate, Remind, and others)
- Serial redirect
- Serial over LAN
- Machine check error capture.
- Scalability partition management for multinode support
- Platform event filtering (PEF) and alert policies
- Local and remote update of BMC, PFGA firmware, and BIOS

6.1.1 BMC connectivity

The BMC is accessible in the following ways:

- An in-band communication through the IPMB to the RSA2 interface to enable a simple Web interface to control the server without having a management server, which watches the x3850 M2 and x3950 M2 server systems.
- An in-band communication through the BMC driver interface. Any system management software or utility, which you installed in the operation system or is part of any operating system vendor, can communicate with the BMC.
- The BMC is capable of network access through Ethernet port 1. The BMC supports 10/100Mb/s full duplex negotiation. NIC port 1 allows access also to an internal 1 Gbps Broadcom interface.

> **Tip:** The switch configuration in your network environment should allow two MAC addresses on the switch port, which is connected to NIC port 1 on each x3850 M2 and x3950 M2, to gain access to the BMC and the NIC interface.
>
> The switch port should allow different port speeds and negotiation modes to guarantee the Broadcom network interface at 1 Gbps and the BMC at the 10/100Mb/s rate is still working after the operating system is loaded.

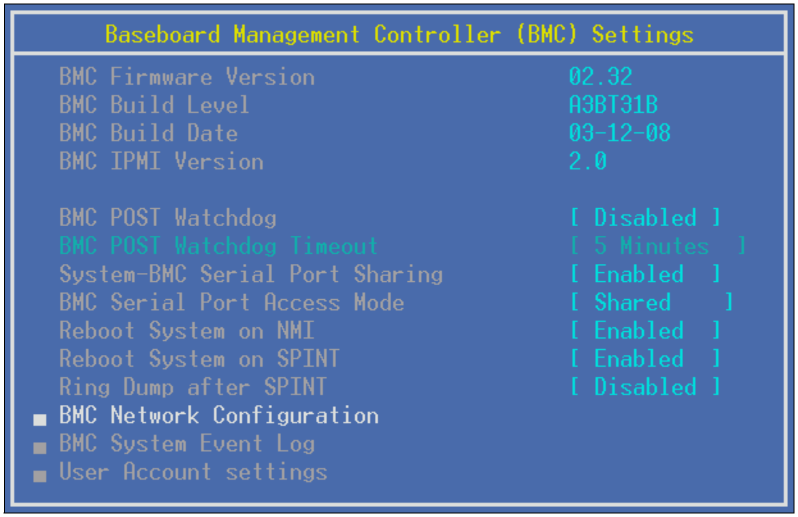
6.1.2 BMC LAN configuration in BIOS

By default, the BMC is configured with the following defaults:

- IP address: 10.1.1.97
- Subnet 255.0.0.0.

You can change the IP settings in the server BIOS, as follows:

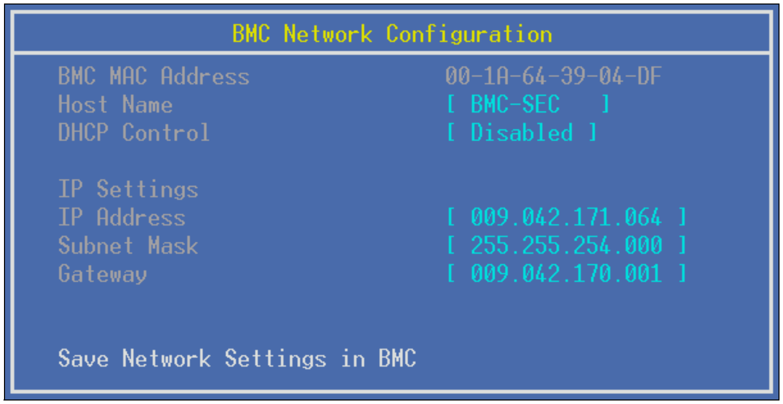
1. Boot or reset the server and press `F1` when prompted to enter Setup.
2. Select the **Advanced Setup** → **Baseboard Management Controller (BMC) Settings.** The BMC menu is shown in Figure 6-1 on page 303.



```
              Baseboard Management Controller (BMC) Settings
   BMC Firmware Version                      02.32
   BMC Build Level                           A3BT31B
   BMC Build Date                            03-12-08
   BMC IPMI Version                          2.0

   BMC POST Watchdog                         [ Disabled ]
   BMC POST Watchdog Timeout                 [ 5 Minutes  ]
   System-BMC Serial Port Sharing            [ Enabled  ]
   BMC Serial Port Access Mode               [ Shared     ]
   Reboot System on NMI                      [ Enabled  ]
   Reboot System on SPINT                    [ Enabled  ]
   Ring Dump after SPINT                     [ Disabled ]
 ■ BMC Network Configuration
 ■ BMC System Event Log
 ■ User Account settings
```

Figure 6-1 BIOS: BMC configuration settings menu

3. Select the **BMC Network Configuration** to open the BMC Network Configuration menu, shown in Figure 6-2.



```
                     BMC Network Configuration
   BMC MAC Address                  00-1A-64-39-04-DF
   Host Name                        [ BMC-SEC    ]
   DHCP Control                     [ Disabled ]

   IP Settings
   IP Address                       [ 009.042.171.064 ]
   Subnet Mask                      [ 255.255.254.000 ]
   Gateway                          [ 009.042.170.001 ]


   Save Network Settings in BMC
```

Figure 6-2 BIOS: BMC Network Configuration settings menu

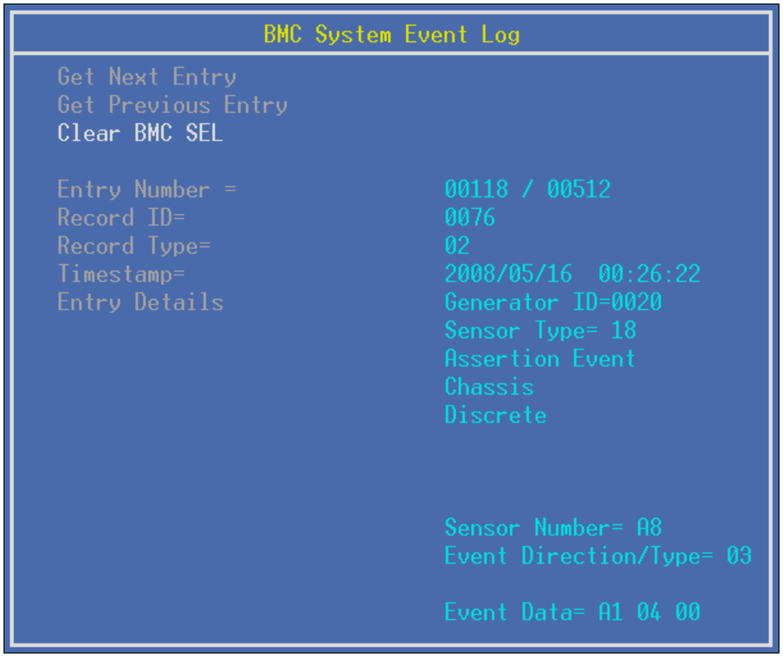
4. Enter a host name.
5. Enter the appropriate IP address, subnet mask, and gateway address, or enable DHCP control.
6. Select **Save Network Settings in BMC** and press `Enter`.

6.1.3 Event Log

The BMC maintains a separate event log that contains entries such as power events, environmental events, or chipset specific entries. This event log records all the hardware events or alerts for the server. The logs are defined in different priorities: informal, warning or error event entries, critical or non-critical events.

You can access the **BMC System Event Log** (SEL) through the menu shown in Figure 6-1 on page 303, or by using tools such as OSA SMBridge, SvcCon, Dynamic System Analysis (DSA), or IBM Director.

An example event log is shown in Figure 6-3. You can select **Get Next Entry** and **Get Previous Entry** to page through the events.



```
                         BMC System Event Log
     Get Next Entry
     Get Previous Entry
     Clear BMC SEL

     Entry Number =                 00118 / 00512
     Record ID=                     0076
     Record Type=                   02
     Timestamp=                     2008/05/16  00:26:22
     Entry Details                  Generator ID=0020
                                    Sensor Type= 18
                                    Assertion Event
                                    Chassis
                                    Discrete



                                    Sensor Number= A8
                                    Event Direction/Type= 03

                                    Event Data= A1 04 00
```

Figure 6-3 BMC System Event Log entry in BIOS

The total number of events recorded is listed in the Entry Number field.

Tip: You might notice that several events have a time stamp of 1970. These events are defined in the IPMI specification at Intel, and do not include time stamp information. It fills the respective bytes in the event raw data with empty values, which are interpreted by the SEL as a date in 1970.

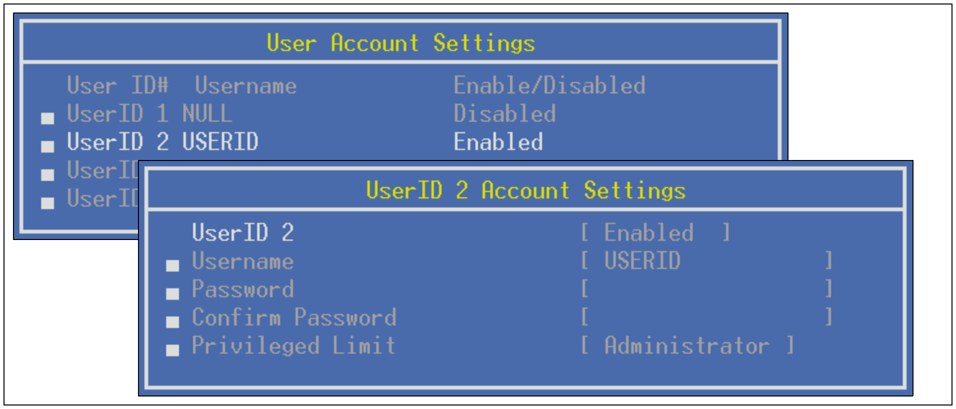
As shown in Figure 6-3 on page 304, the number of events reached the maximum level of 512 entries. You must clear this, freeing it for new events. To clear it, select **Clear BMC SEL** or use tools we mentioned previously. The BMC alerts you if the log reaches 75, 90, or 100 percent full.

The BMC makes the events available to the installed RSA II. The RSA II also maintains a separate log, which you can view in the RSA II. Both the RSA II and BMC-based events are listed. However, the reverse is not true: you cannot view the events recorded by the RSA II in the BMC event log.

6.1.4 User Account Settings menus

You can add, change, or set user privileges in the BMC. User access is through account UserID 2 as shown in Figure 6-4. (UserID 1 is `NULL` and cannot be changed.) The default credentials for UserID 2 are:

- Username: USERID (all uppercase)
- Password: PASSW**0**RD (all uppercase, where 0 is the number zero)



```
                 User Account Settings

  User ID#  Username                Enable/Disabled
■ UserID 1 NULL                     Disabled
■ UserID 2 USERID                   Enabled
■ UserII
■ UserII
                      UserID 2 Account Settings

             UserID 2                        [ Enabled  ]
           ■ Username                        [ USERID           ]
           ■ Password                        [                  ]
           ■ Confirm Password                [                  ]
           ■ Privileged Limit                [ Administrator ]
```

Figure 6-4 BIOS: BMC User Account Settings menu and individual user settings

Select the UserID you want to enable or disable and make the required settings for username, password, and the privileges level.

6.1.5 Remote control using SMBridge

The BMC supports remote control using the OSA SMBridge utility and Serial over LAN. This provides a text-only console interface that lets you control BIOS screens and specific operating system consoles. Both Linux and Windows provide text-only consoles.

The SMBridge utility and documentation can be found at:

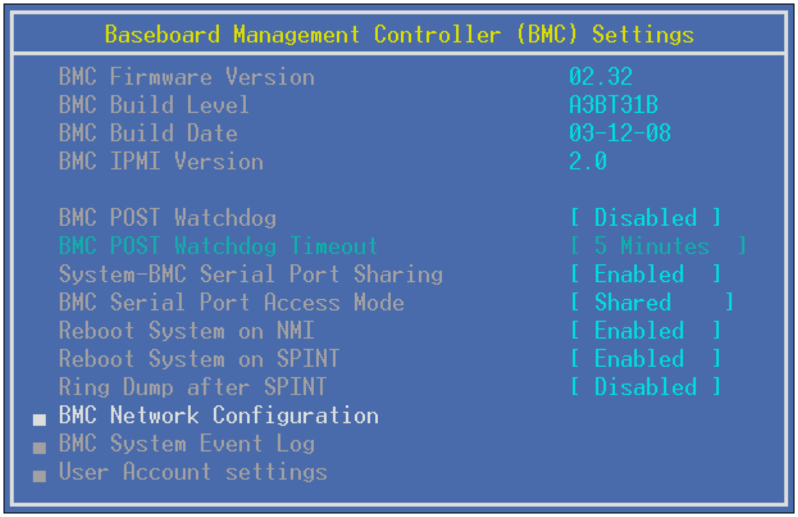
http://www.ibm.com/support/docview.wss?uid=psg1MIGR-62198

The IBM Redbooks technote, *Enabling Serial Over LAN for a Remote Windows Text Console using OSA SMBridge*, TIPS0551 also provides helpful information:

http://www.redbooks.ibm.com/abstracts/tips0551.html

6.1.6 BMC monitoring features

The BMC settings menu (Figure 6-5) in the server BIOS allows to change the behavior of the server and how to proceed if an error occurs. Various system management routines are controlled by the BMC.



```
            Baseboard Management Controller (BMC) Settings

   BMC Firmware Version                   02.32
   BMC Build Level                        A3BT31B
   BMC Build Date                         03-12-08
   BMC IPMI Version                       2.0

   BMC POST Watchdog                      [ Disabled ]
   BMC POST Watchdog Timeout              [ 5 Minutes  ]
   System-BMC Serial Port Sharing         [ Enabled  ]
   BMC Serial Port Access Mode            [ Shared    ]
   Reboot System on NMI                   [ Enabled  ]
   Reboot System on SPINT                 [ Enabled  ]
   Ring Dump after SPINT                  [ Disabled ]
 ■ BMC Network Configuration
 ■ BMC System Event Log
 ■ User Account settings
```

Figure 6-5 BIOS: BMC configuration settings

The settings as shown in Figure 6-5 are:

- BMC POST Watchdog

  This is Disabled by default. It monitors the system initialization in the POST phase. If you enable this watchdog, the system remains at reboot or shutdown after its defined time-out. An event is recorded to the BMC log.

- Reboot System on NMI

  When a non-maskable interrupt (NMI) signal is received, the processor immediately drops what it was doing and attends to it. The NMI signal is normally used only for critical problem situations, such as serious hardware errors.

This setting is Enabled by default. When an NMI is issued by a critical event the BMC performs the system to reset for recovering the system. The BMC logs the reboot and additional error events in the SEL.

- Reboot on SPINT

  When any unrecoverable error condition occurs, the service processor interrupt (SPINT) routine catches chipset register information of the machine check (MCK) error registers and stores it in the BMC NVRAM. Each main component of the IBM eX4 is embedded by a different machine check error register that monitors the system operability and holds the state of that condition.

  This information is cleared after the system is recovered by a reboot. However the BMC starts the SPINT routine to store the information in its NVRAM area. This information is available, until the next MCK occurs.

  An MCK is reported in the BMC and is caused by a fatal situation that cannot be handled by the chipset. Typically, an MCK is issued by an essential component of the IBM eXA4 chipset, particular components such as a DIMM, processor, or I/O linked devices.

  > **Note:** We recommend reporting a machine check error to IBM Technical Support, if the information is not sufficient in the *System x3850 M2 and x3950 M2 Problem Determination and Service Guide*.

- Ring Dump after SPINT

  This setting is Disabled by default. By enabling the setting, the BMC catches Ring Dump information from the chipset, after an MCK occurred, to log unexpected error conditions. IBM Technical Support might request you to enable this setting.

6.1.7 BMC firmware update

The BMC can be flashed by a bootable image or by Linux or a Windows-based update package. You may download updates from the IBM System x support site:

- BMC flash update (DOS bootable ISO image)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073127

- BMC flash update (Windows executable)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073128

- BMC flash update (Linux executable)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073129

Review each *readme* file for considerations and limitations. To update the BMC, follow the instructions in 5.1, “Updating firmware and BIOS” on page 246.

6.1.8 Installing the BMC device drivers

The device drivers are required to provide operating system support and in-band communication with IBM Director. This section describes how to install the IPMI device drivers on Windows and Linux platforms. The required device drivers are listed in Table 6-1.

Table 6-1 IPMI required device drivers

Device driver	Additional commands
OSA[a] IPMI device driver	Required for in-band communication with IBM Director and BMC system utilities
OSA IPMI library	This is the OSA IPMI mapping layer. Includes the BMC Mapping Layer, which maps the dot.commands to IPMI commands. Required for in-band communication with IBM Director.
IBM ASR service	Required for Automatic Server Restart functionality
Microsoft IPMI	Microsoft IPMI driver interface, support for new IBM update tools in Windows OS
OpenIPMI	Open source IPMI driver for Linux OS

[a] OSA Technologies, Inc. is part of the Avocent corporation.

The device drivers must be installed in a specific order or the installation can fail:

1. Either of the following drivers can be installed first:
   - OSA IPMI device driver, then OSA IPMI library (mapping layer)
   - Microsoft IPMI driver (any third-party driver must be uninstalled first)
2. IBM IPMI ASR service

Download the BMC device drivers

To download the drivers appropriate for your server:

1. Go to the Support for IBM System x Web page:

   `http://www.ibm.com/systems/support/x`
2. Under Popular links, click **Software and device drivers**.
3. Select your product, for example, **System x3850 M2 (7141,7144)**
4. Use the category **OSA IPMI**.

5. Select the links to download each component.

> **Note:** The IBM System x Web page, does not have OSA IPMI driver and layer software available for Linux, Novell, and Windows operating systems to download. For instructions, see the following sections.

Install the device drivers on Windows

Although the OSA IPMI driver limits the support for single node systems only, the Microsoft IPMI driver is required for multinode x3950 M2 configurations.

You may use both drivers, but the OSA IPMI and the Microsoft IPMI drivers cannot coexist. If the OSA driver is already installed, you must uninstall it before installing the Microsoft IPMI driver.

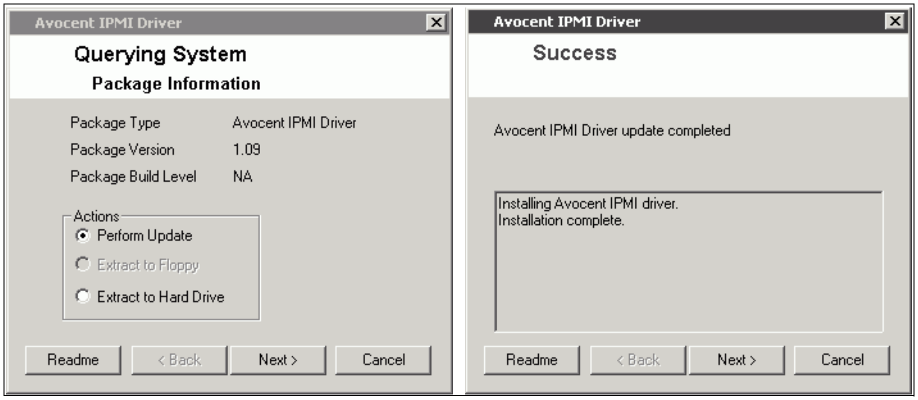
> **Notes:**
>
> - On multinode x3950 M2 configurations, the Microsoft IPMI device driver is the only supported IPMI driver. If you have the OSA IPMI device driver installed you must remove that and replace it with the Microsoft IPMI device driver. The Microsoft IPMI driver is not installed in Windows Server 2003 by default. See the readme file in the driver package for more details.
> - The Microsoft IPMI device driver is required for UXSP, wFlash (embedded in any IBM update packages for Windows), and online Dynamic System Analysis (DSA) software.

This section describes how to install the drivers under Windows.

OSA IPMI device driver in Windows Server 2003

To install the OSA IPMI device driver:

1. Run Setup.exe. Click **Next** to proceed through the usual initial windows.
2. When prompted select **Perform Update** and click the **Next** button (Figure 6-6 on page 310).



Package Type Avocent IPMI Driver
Package Version 1.09
Package Build Level NA

Avocent IPMI Driver update completed

Installing Avocent IPMI driver.
Installation complete.

Figure 6-6 OSA IPMI driver in a Windows installation

3. Click **Next** to continue the installation. When it completes, you are prompted to reboot the server, the installer does not do this automatically.
4. After the reboot open Device Manager again. Enable **Show hidden devices** in the submenu **View**. The ISMP device driver is listed in the section System devices as shown in Figure 6-7.

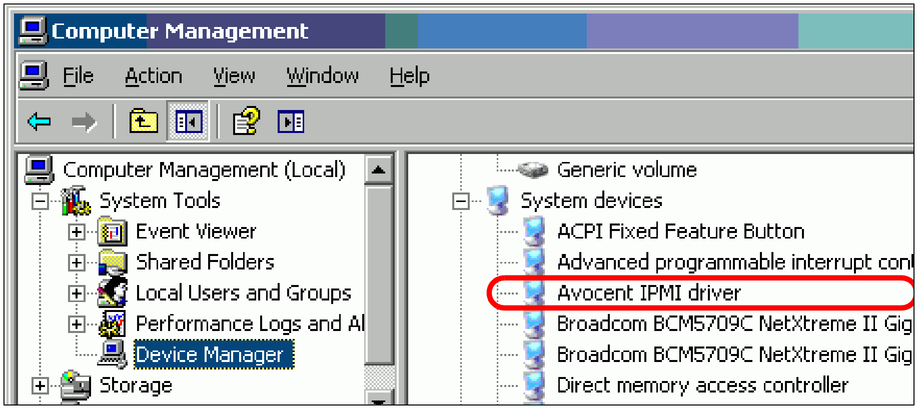


Figure 6-7 Device Manager showing list of system devices

OSA IPMI mapping layer (library) files in Windows Server 2003

To install the IPMI mapping layer (library) files:

1. Ensure that the OSA IPMI device driver is installed before installing the library software.
2. Download the executable file from the link listed in “Download the BMC device drivers” on page 308, and run it.

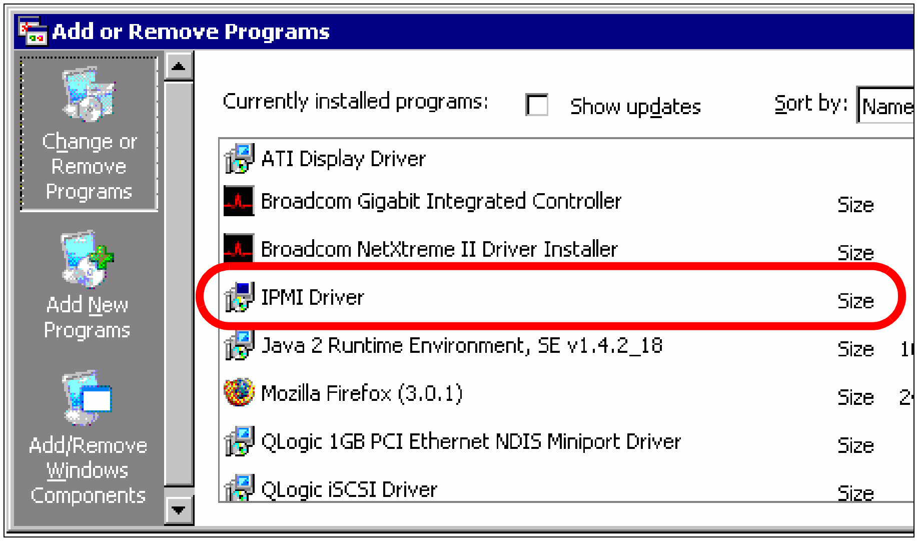
3. Follow the program's instructions, shown in the window.
4. Reboot the server if the installation procedure prompts you to do so.

> **Note:** If the Microsoft IPMI driver is installed, you do not have to install the OSA IPMI mapping layer.

OSA IPMI driver uninstall in Windows Server 2003

We added this section in case you want to uninstall the Microsoft IPMI driver. To remove the IPMI driver:

1. Click **Start** → **Settings** → **Control Panel** → **Add or Remove Programs**. The window shown in Figure 6-8 opens.


Add or Remove Programs
Change or Remove Programs
Add New Programs
Add/Remove Windows Components
Currently installed programs:
□ Show updates
Sort by: Name

ATI Display Driver	
Broadcom Gigabit Integrated Controller	Size
Broadcom NetXtreme II Driver Installer	Size
IPMI Driver	Size
Java 2 Runtime Environment, SE v1.4.2_18	Size 1
Mozilla Firefox (3.0.1)	Size 2
QLogic 1GB PCI Ethernet NDIS Miniport Driver	Size
QLogic iSCSI Driver	Size

Figure 6-8 Windows 2003 utility for adding or removing programs

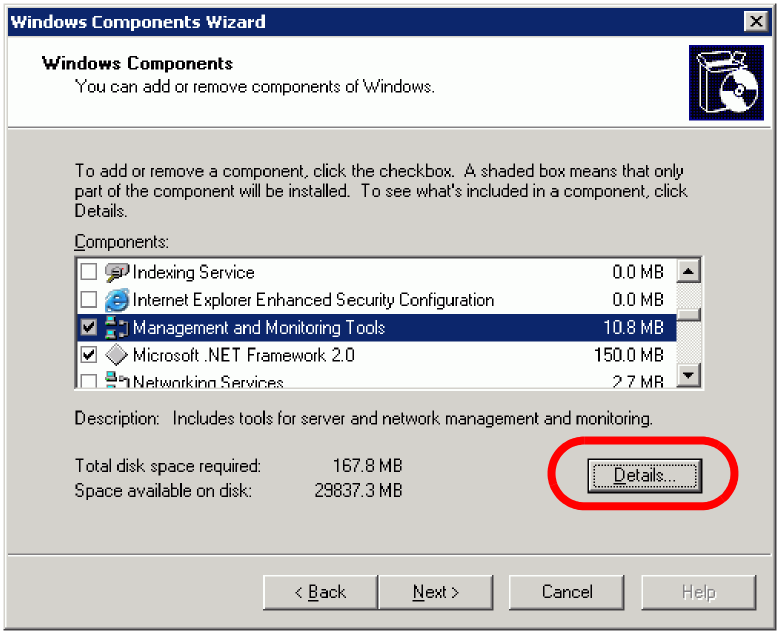
2. Select the entry **IBM_msi_server** or **IPMI Driver** (depending on the version you have installed) and click the **Remove** button.
3. The Add or Remove Programs utility prompts you to restart the system to complete the driver removal process.

> **Note:** The OSA IPMI driver should not be removed using Device Manager. Doing so only partially removes the driver, which prevents the installation of a new driver version or re-installation of the previously installed version. If an attempt has already been made to remove the driver using Device Manager, then follow the driver removal instructions listed in this section and reboot the system.

Microsoft IPMI driver in Windows Server 2003

The Microsoft IPMI driver is not installed by default, however, it is available for manual installation if you have applied Release 2 Service Pack 2, as follows:

1. Click **Start** → **Settings** → **Control Panel** → **Add or Remove Programs** → **Add/Remove Windows Components**. The window shown in Figure 6-9 opens.



Windows Components Wizard

Windows Components
You can add or remove components of Windows.

To add or remove a component, click the checkbox. A shaded box means that only part of the component will be installed. To see what's included in a component, click Details.

Components:

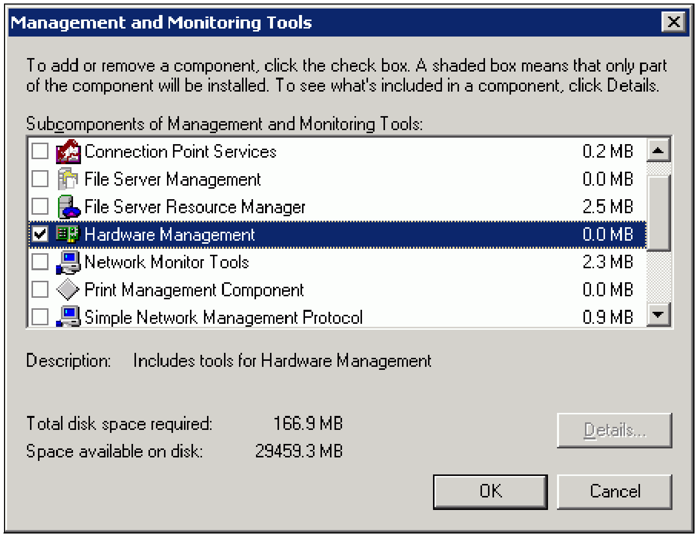
□	Indexing Service	0.0 MB
□	Internet Explorer Enhanced Security Configuration	0.0 MB
☑	Management and Monitoring Tools	10.8 MB
☑	Microsoft .NET Framework 2.0	150.0 MB
□	Networking Services	2.7 MB

Description: Includes tools for server and network management and monitoring.

Total disk space required: 167.8 MB
Space available on disk: 29837.3 MB

Figure 6-9 Windows Server 2003: adding or removing Windows components

2. Select **Management and Monitoring Tools** and click **Details**. The window shown in Figure 6-10 on page 313 opens.
3. Add a check mark to **Hardware Management**.


Management and Monitoring Tools

To add or remove a component, click the check box. A shaded box means that only part of the component will be installed. To see what's included in a component, click Details.

Subcomponents of Management and Monitoring Tools:

☐	Connection Point Services	0.2 MB
☐	File Server Management	0.0 MB
☐	File Server Resource Manager	2.5 MB
☑	Hardware Management	0.0 MB
☐	Network Monitor Tools	2.3 MB
☐	Print Management Component	0.0 MB
☐	Simple Network Management Protocol	0.9 MB

Description: Includes tools for Hardware Management

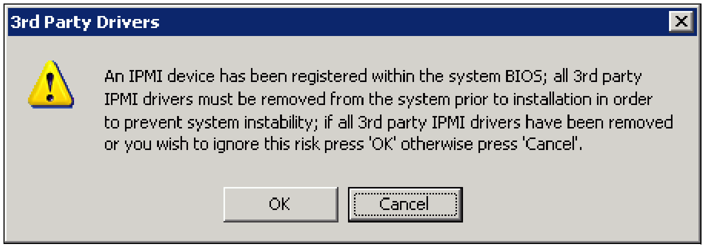
Total disk space required: 166.9 MB

Space available on disk: 29459.3 MB

Figure 6-10 Windows Server 2003: selecting a subcomponent

4. Click **OK**.

   Windows reminds you that any third-party IPMI driver must be uninstalled. See Figure 6-11. Ensure that the OSA IBM driver is uninstalled, as described in “OSA IPMI driver uninstall in Windows Server 2003” on page 311.


3rd Party Drivers

An IPMI device has been registered within the system BIOS; all 3rd party IPMI drivers must be removed from the system prior to installation in order to prevent system instability; if all 3rd party IPMI drivers have been removed or you wish to ignore this risk press 'OK' otherwise press 'Cancel'.

Figure 6-11 Microsoft hardware management notice of third-party drivers

5. Click **OK**.
6. Click **Next**.
7. Click **Finish**.

When the installation is completed, the Microsoft-compliant IPMI device is listed as a hidden device under **Device Manager** → **System Devices** as shown in Figure 6-12 on page 314.

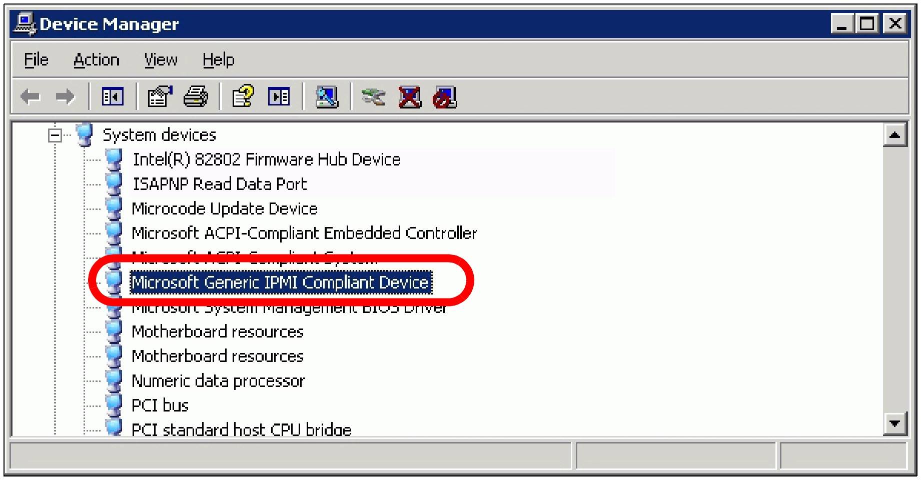

Device Manager
File
Action
View
Help
System devices
Intel(R) 82802 Firmware Hub Device
ISAPNP Read Data Port
Microcode Update Device
Microsoft ACPI-Compliant Embedded Controller
Microsoft ACPI-Compliant System
Microsoft Generic IPMI Compliant Device
Microsoft System Management BIOS Driver
Motherboard resources
Motherboard resources
Numeric data processor
PCI bus
PCI standard host CPU bridge

Figure 6-12 Device Manager: Microsoft IPMI driver in Windows 2003

IPMI ASR service

To install the Automatic Server Restart (ASR) service, do the following:

1. Ensure that the IPMI device driver (ipmi.sys or ipmidrv.sys) and IPMI library files are installed before installing this software.
2. Download the executable, listed in “Download the BMC device drivers” on page 308, and run it.
3. Follow the program’s instructions, shown in the window.
4. Reboot the server if the installation procedure prompts you to do so.

Microsoft IPMI driver in Windows 2008

The Microsoft IPMI driver is installed in Windows 2008 by default.

Install the device drivers on Linux

The support for OSA IPMI driver and layer software for RHEL4 and SLES9 are deprecated. In May 2008, IBM transitioned to the Open Source IPMI (OpenIPMI) software instead. See the following support document for details:

- IBM now supporting Linux open source IPMI driver and utility

  `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5069569`

- Linux Open Source watchdog daemon support replaces IPMI ASR

  `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5069505`

The OpenIPMI library supports the Intelligent Platform Management Interface (IPMI) versions 1.5 and 2.0.

The libraries for OpenIPMI are part of the most recent versions of Linux operating systems. As a result, IBM does not supply these drivers.

> **Note:** SUSE Linux Enterprise Server 10 (and later) and Red Hat Enterprise Linux 5 (and later) are shipped with the OpenIPMI driver natively.
>
> For older operating systems SUSE Linux Enterprise Server 9 (and earlier) and Red Hat Enterprise Linux 4 (and earlier), see the OpenIPMI Web site:
>
> `http://openipmi.sourceforge.net/`

Operating systems supporting OpenIPMI driver

The operating systems that support the OpenIPMI driver include:

- Red Hat Enterprise Linux 4 Update 3, Update 4, Update 5, Update 6
- Red Hat Enterprise Linux 5, any update
- Red Hat Enterprise Linux 5.1, any update
- SUSE Linux SLED 10, any Service Pack
- SUSE Linux SLED 10 for AMD64/EM64T, any Service Pack
- SUSE Linux SLES 10, any Service Pack
- SUSE Linux SLES 10 for AMD64/EM64T, any Service Pack
- SUSE Linux SLES 9 Service Pack 2, Service Pack 3
- SUSE Linux SLES 9 for AMD64/EM64T Service Pack 2
- VMware ESX Server 3.0.2, any update
- VMware ESX 3.5, any update
- VMware ESX 3i installable

Working with Open Source IPMITool

Download the latest version of ipmitool from the following location:

`http://ipmitool.sourceforge.net`

> **Note:** Problems have been reported with versions of ipmitool prior to v1.8.9. See the following location for details:
>
> `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5069538`

To build ipmitool:

> **Note:** The installation must be run with root permissions to overlay the existing ipmitool utility in /usr/local/bin.

1. Unzip and untar the downloaded ipmitool package:

   ```
   tar xvzf ipmitool*.tar.gz
   ```

2. Configure ipmitool for your system:

   ```
   cd ipmitool*
   ./configure
   ```

3. Change to the created ipmitool directory to build ipmitool and install it:

   ```
   make
   make install
   ```

6.1.9 Ports used by the BMC

The BMC uses several TCP/UDP ports for communication, as shown in Table 6-2. If the communication with the BMC passes firewalls, it is important to know which ports you must enable on the firewalls to communicate properly.

Table 6-2 TCP/IP ports used by the BMC

Port number	Description
623	IPMI communications to SMBridge and IBM Director
664	IPMI communications (secondary)
161	SNMP get and set commands
162	SNMP traps and PET alerts to IBM Director

6.2 Remote Supervisor Adapter II

The Remote Supervisor Adapter II, shown in Figure 6-13 on page 317, is a system management card that ships with the x3850 M2 and x3950 M2 server systems.

The adapter includes two components: the main board has an embedded video chip ATI RN50 and the networking interface; and the daughter card is the RSA II adapter that is connected by a separate internal cable to the Intelligent Platform Management Bus (IPMB) on the Serial IO/PCI-X board.

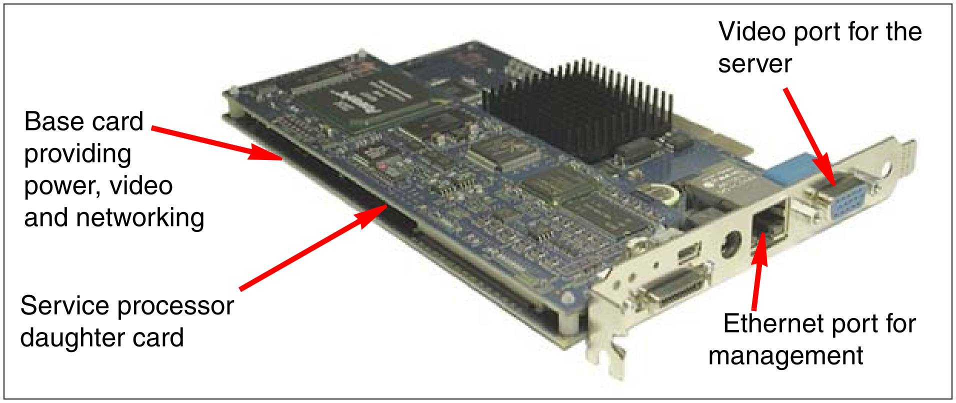

Base card providing power, video and networking
Service processor daughter card
Video port for the server
Ethernet port for management

Figure 6-13 Remote Supervisor Adapter (RSA) II

The RSA II communicates with the BMC and periodically polls for new events. The RSA II catches events of the BMC and translates them to more user-friendly event information. The event information is then stored in an assigned space in the NVRAM of the RSA II.

The card contains a real-time clock (RTC) timer chip, however the RSA II timer can be synchronized with the system BIOS time or by a Network Time Protocol (NTP) server.

The primary user management interface of the RSA II is Web-based and provides a complete set of functions. The most useful functions and features of the RSA II include:

- Web and Telnet interface

  The RSA II is managed through its built-in Web interface. From the interface, you can manage the local server plus other nodes if the local server is part of a multinode complex.

- Vital Product Data

  This provides an overview of the most essential system firmware codes for BIOS, RSA II, BMC without rebooting the server.

- Continuous health monitoring and control

  The RSA II continuously monitors all important system parameters such as temperature, voltage, and more. If a fan fails, for example, the RSA II forces the remaining fans to increase speed to compensate for the failing fan.

- Automatic notification and alerts

  The RSA II automatically sends various types of alerts and notifications to another server, such as IBM Director, to an SNMP destination, or as e-mail directly to a user by using SMTP.

- Event log

  You can access the event logs of the server and the power-on-self-test (POST) log and export them while the server is running.

- Remote control

  The RSA II card offers full remote control, including mouse, keyboard and video from power-up and setup and diagnostics panels, all the way through to the operating system running as normal. In fact, combined with the remote media function, you are able to boot the server and remotely install an operating system from a remote media such as CD-ROM, ISO files, or floppy images.

  A connection through Windows Terminal Server can also be established if the remote desktop software is installed and enabled.

- Remote media

  As a part of the remote control feature, the remote media capability lets you use diskette drives, diskette images, optical drives (such as DVD or CD-ROM), ISO files on your hard disk, mounted in a virtual drive or anywhere in your network. You only have to map this network drive or file system to the local workstation to which you have established the connection to the server. Either way makes them appear to be local drives on the remotely managed server.

  This feature and additionally the remote control features allow you to remotely install any operating system on your x3850 M2 and x3950 M2 without requiring access to any resources in your network environment, or to manage your operating system or any application. You only have to access the RSA II Web interface.

- Remote power control

  The RSA II supports remote power control to power on, power off, or restart the server with or without operating system shutdown over LAN or even a WAN connection.

- Event log

  You can access the event logs of the server and the POST log and export them while the server is up and running.

- Scalability Management

  The Remote Supervisor Adapter II offers an easy-to-use Scalability Management Interface to configure, control, and manage Scalability partitions in a complex of x3950 M2 server systems. Read more about scalability, how it works, and how to manage in Chapter 4, “Multinode hardware configurations” on page 195.

6.2.1 RSA II connectivity

The Remote Supervisor Adapter II (RSA II) provides two ways to communicate with system management software:

- In-band communication through the RSA II driver interface, allowing any system management software or utility that you installed in the operation system to communicate with the RSA II
- Out-of-band communication through the RSA II network interface, allowing various features such as alerting or Web interface access

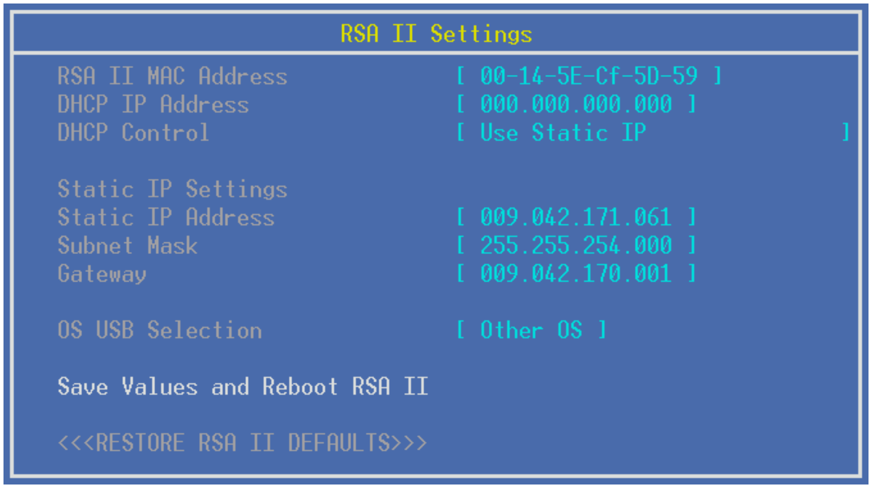
6.2.2 RSA LAN configuration in BIOS

The default network address for the RSA II is:

- IP address: 192.168.70.125
- Subnet: 255.255.255.0

The RSA II LAN settings are available in the server BIOS. To change the IP settings in the server BIOS:

1. Boot or reset the server and press `F1` when prompted to enter Setup.
2. Select the **Advanced Setup** → **RSA II Settings.** The RSA II menu opens, as shown in Figure 6-14 on page 320*.*



```
                         RSA II Settings
   RSA II MAC Address                 [ 00-14-5E-Cf-5D-59 ]
   DHCP IP Address                    [ 000.000.000.000 ]
   DHCP Control                       [ Use Static IP                ]

   Static IP Settings
   Static IP Address                  [ 009.042.171.061 ]
   Subnet Mask                        [ 255.255.254.000 ]
   Gateway                            [ 009.042.170.001 ]

   OS USB Selection                   [ Other OS ]

   Save Values and Reboot RSA II

   <<<RESTORE RSA II DEFAULTS>>>
```

Figure 6-14 BIOS: RSA II configuration settings

> **Tip:** You may also modify these settings by connecting to the service processor port using a cross-over cable and opening a Web browser to the card's default IP address: 192.168.70.125
>
> See details in 6.2.3, "Web interface" on page 321.

3. Set your static IP settings. You may also set up the RSA IP for each DHCP server. You can change host name in the Web interface. Enter the following IP settings in the DHCP Control field of the RSA II Settings menu in BIOS:

   – Try DHCP then Use Static IP (default)
   – DHCP Enabled
   – Use Static IP

   Configure the RSA II LAN interface to fit in your LAN segment. Use the right and left arrow keys for selection.

4. In the OS USB Selection option, select either **Other OS** (for Windows) or select **Linux OS**. Use the right and left arrow keys to make the selection.

   The purpose of this selection is to prevent a known problem with Linux and its generic human interface device (HID) driver. Linux cannot establish USB communication with the RSA II using the generic HID (which Windows uses). By selecting **Linux OS** here, it makes the RSA II appear as an OEM HID instead of generic HID, which then functions properly.

5. Select **Save the Values and Reboot RSA II**, and then press `Enter`.

6. Exit the utility.

7. Save the settings and restart the ASM. The RSA2 adapter is restarted and the new settings are enabled.

   > **Tip:** This process can take up to 20 seconds before you can **`ping`** the new IP address of the RSA II adapter.

6.2.3 Web interface

Through port 80, the RSA II enables the embedded Web server daemon, which you use to access the Web interface after you assigned the correct LAN interface settings as described in 6.2.2, “RSA LAN configuration in BIOS” on page 319.

As described in that section, the adapter is configured by default to look for a DHCP server to obtain an IP address, and if none is available, to use the default IP address of:

```
192.168.70.125
```

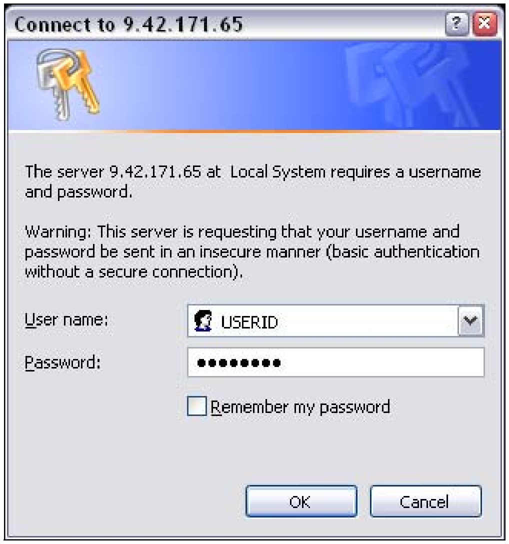
You can view the assigned address from the BIOS in the **Advanced Settings** menu.

To access the RSA II Web interface:

1. Open a supported browser. Enable the use of a Java™ Plugin, if you want to use remote console features, as described in 6.2.4, “Remote console and media” on page 324.

   Use the default user ID and password (see Figure 6-15 on page 322):

   – User ID: `USERID` (all uppercase)
   – Password: `PASSW0RD` (all uppercase, where `0` is the number zero)


Connect to 9.42.171.65

The server 9.42.171.65 at Local System requires a username and password.

Warning: This server is requesting that your username and password be sent in an insecure manner (basic authentication without a secure connection).

User name: USERID

Password: ●●●●●●●●●

□ Remember my password

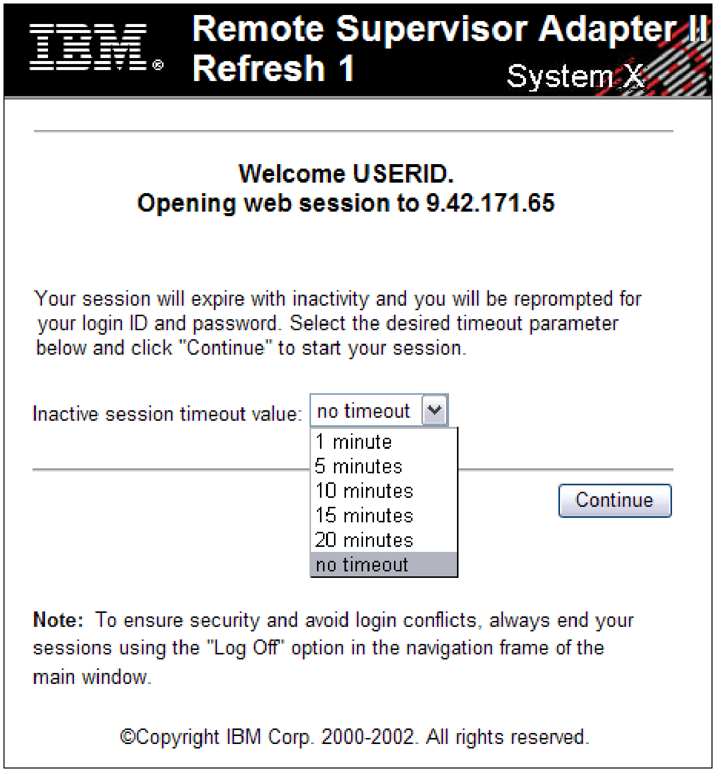
Figure 6-15   RSA II Web interface login window

Note: In the Web interface, you will have to change the login credentials.

Safeguard your credentials. Losing them requires rebooting the server BIOS and restoring the RSA settings to the factory defaults.

After you log on, the welcome window opens.

2. Set the session time-out values as shown in Figure 6-16 on page 323.



Welcome USERID.
Opening web session to 9.42.171.65

Your session will expire with inactivity and you will be reprompted for your login ID and password. Select the desired timeout parameter below and click "Continue" to start your session.

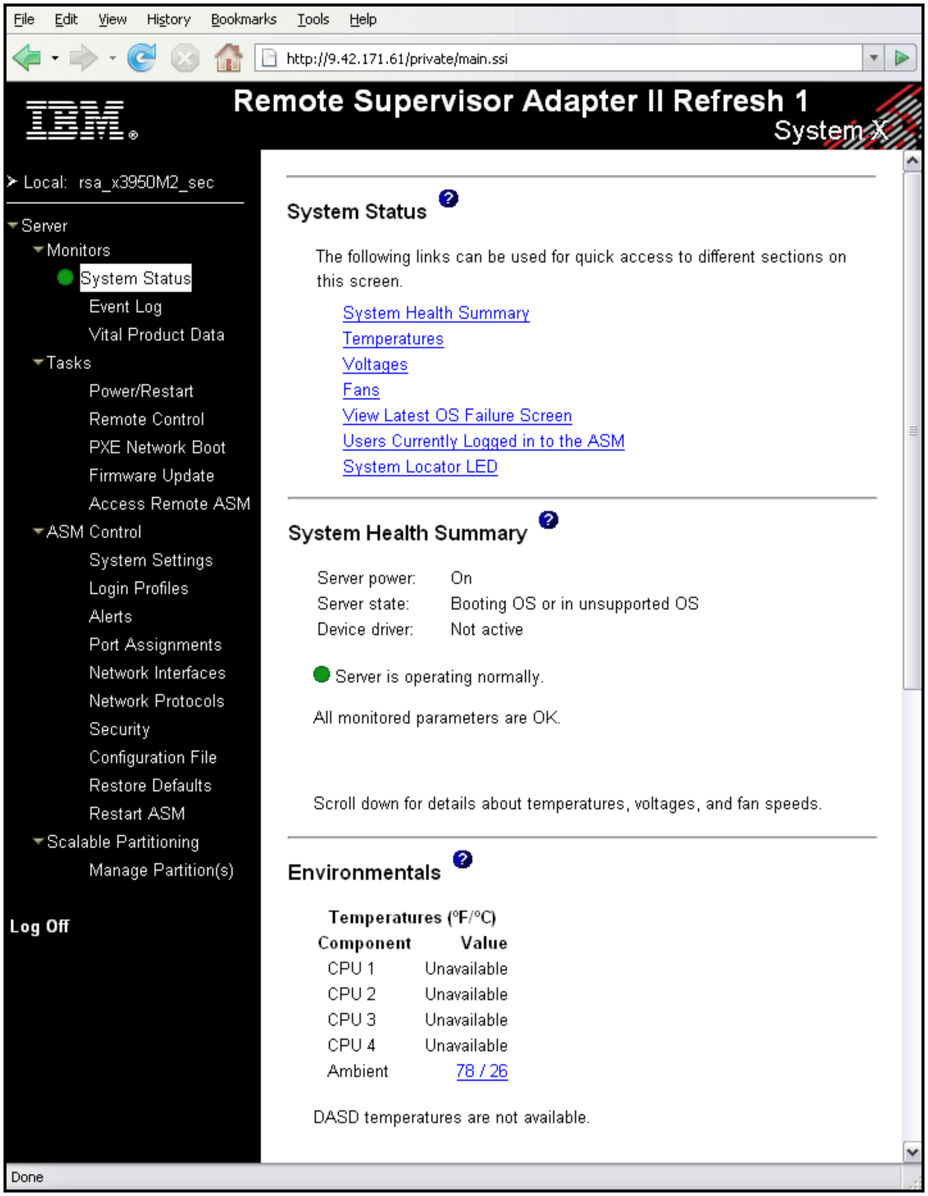
Inactive session timeout value: no timeout

Note: To ensure security and avoid login conflicts, always end your sessions using the "Log Off" option in the navigation frame of the main window.

©Copyright IBM Corp. 2000-2002. All rights reserved.

Figure 6-16 RSA II Web interface Welcome window.

The home window of the Web interface opens. See Figure 6-17 on page 324.



System Status

The following links can be used for quick access to different sections on this screen.

- System Health Summary
- Temperatures
- Voltages
- Fans
- View Latest OS Failure Screen
- Users Currently Logged in to the ASM
- System Locator LED

System Health Summary

Server power:	On
Server state:	Booting OS or in unsupported OS
Device driver:	Not active

● Server is operating normally.

All monitored parameters are OK.

Scroll down for details about temperatures, voltages, and fan speeds.

Environmentals

Temperatures (°F/°C)

Component	Value
CPU 1	Unavailable
CPU 2	Unavailable
CPU 3	Unavailable
CPU 4	Unavailable
Ambient	78 / 26

DASD temperatures are not available.

Figure 6-17 RSA II Web interface

6.2.4 Remote console and media

To manage servers from a remote location, you often use more than just keyboard-video-mouse (KVM) redirection. For example, for the remote installation of an operating system or patches, you may require remote media to connect a CD-ROM or diskette to the server. The RSA II offers the ability to make available a local diskette, CD-ROM, or image to a remote server and have that server treat the devices as though they were a local USB-attached device. See Figure 6-16 on page 323.

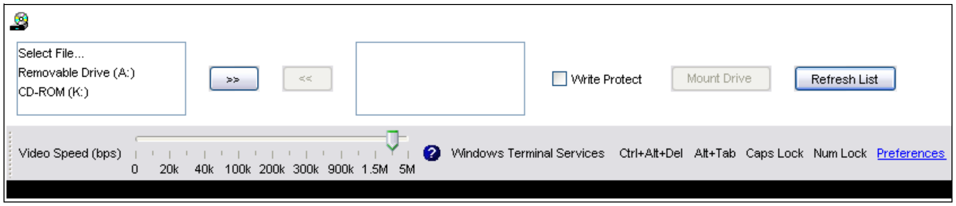


Figure 6-18 RSA II Web interface: remote control

Using remote media enables USB support after the server is powered up and the RSA II is initialized. During installation of the operating system or in the operating system support, USB is required. The correct setting for earlier USB support has can be done in the RSA II Settings menu, shown in Figure 6-14 on page 320.

Tip: You can mount more than one remote drive concurrently. This means you may add a CD-ROM and a disk drive to your remote managed server, or use ISO and diskette image files.

Remote media works with the following operating systems:

- Windows Server 2003
- Windows Server 2008 64bit
- Red Hat Enterprise Linux AS 4, but not for OS installation
- Red Hat Enterprise Linux AS 5, but not for OS installation
- SUSE LINUX Enterprise Server 9, but not for OS installation
- SUSE LINUX Enterprise Server 10, but not for OS installation
- VMware ESX 3i v3.5
- VMware ESX 3.5

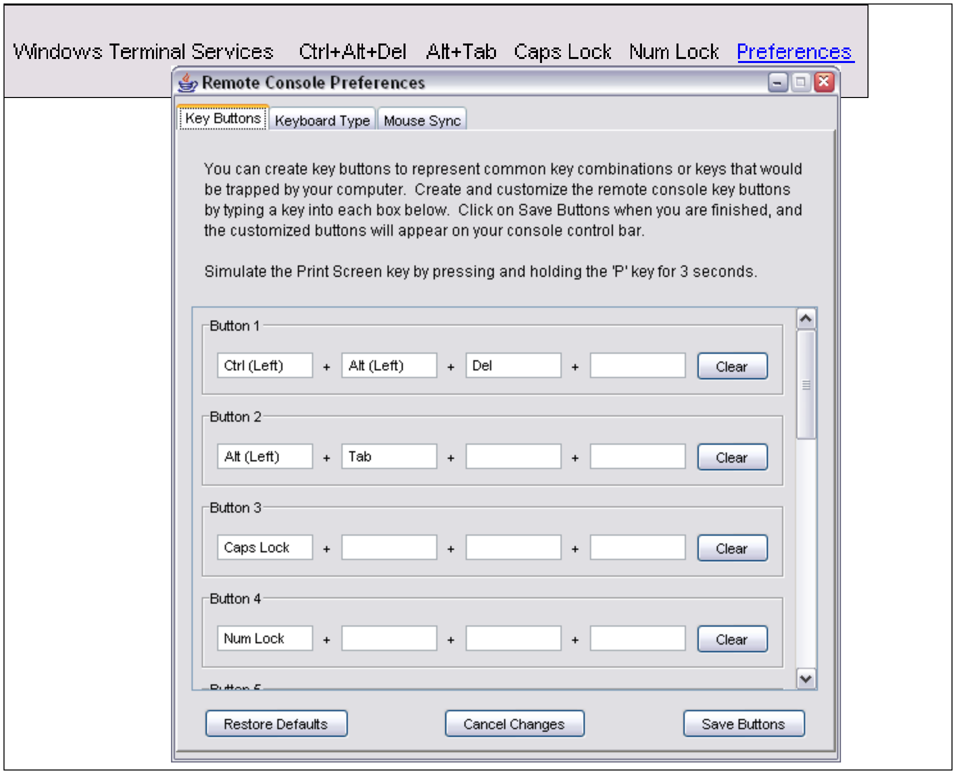
A Java run-time is required, which can be downloaded from:

`http://www.java.com/en/download/manual.jsp`

Tip: We recommend to use at least SDK 1.4.2_06, or 1.5.0_10. We have concerns with the use of remote control features from builds v1.6.0, which were available at the time of writing this book.

The remote control window enables a button bar with predefined softkeys that simulate specific key stroke characters and the video speed selector. In the button bar you can access to the Windows Terminal Server if the Remote Desktop Protocol (RDP) is enabled.

Each of the buttons represents a key or a combination of keys. If you click a button, the corresponding key-stroke sequence is sent to the server. If you require additional buttons, click **Preferences**, and then modify or create new key buttons. To detach the button bar, click anywhere in the grey background, and then dragging and dropping the bar, a process that creates a separate window.



You can create key buttons to represent common key combinations or keys that would be trapped by your computer. Create and customize the remote console key buttons by typing a key into each box below. Click on Save Buttons when you are finished, and the customized buttons will appear on your console control bar.

Simulate the Print Screen key by pressing and holding the 'P' key for 3 seconds.

Button 1: Ctrl (Left) + Alt (Left) + Del +

Button 2: Alt (Left) + Tab + +

Button 3: Caps Lock + + +

Button 4: Num Lock + + +

Button 5

Figure 6-19 RSA II Web interface: softkeys and preferences

The preferences link allows you to specify up to twelve key-stroke sequences and enable mouse synchronization (that is, ensure the mouse pointer on the remote system precisely follows the local mouse pointer).

The following keyboard types are supported:

- US 104-key keyboard
- Belgian 105-key keyboard

- French 105-key keyboard
- German 105-key keyboard
- Italian 105-key keyboard
- Japanese 109-key keyboard
- Spanish 105-key keyboard
- UK 105-key keyboard

The Video Speed selector, shown in Figure 6-20 is used to limit the bandwidth that is devoted to the Remote Console display on your computer. Reducing the Video Speed can improve the rate at which the Remote Console display is refreshed by limiting the video data that must be displayed.

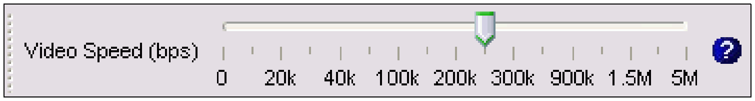


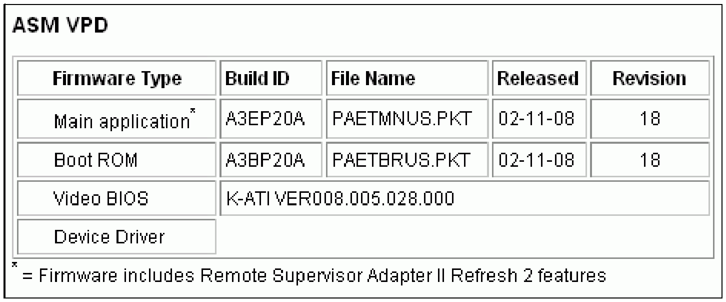
Figure 6-20 RSA II Remote Control: Video Speed selector

You may reduce, or even stop, video data to allow more bandwidth for Remote Disk. Move the slider left or right until you find the bandwidth that achieves the best results. The changes do not require a restart of RSA II or the server.

6.2.5 Updating firmware

Flash the RSA II card if it is not at the latest level of firmware to ensure that all known fixes are implemented. As described in the 4.4, “Prerequisites to create a multinode complex” on page 201, having a specific firmware level on the card is important.

Check the installed firmware in the RSA II Web interface. Select the task **Monitors** → **Vital Product Data** and scroll down to ASM VPD, as shown in Figure 6-21.



ASM VPD

Firmware Type	Build ID	File Name	Released	Revision
Main application*	A3EP20A	PAETMNUS.PKT	02-11-08	18
Boot ROM	A3BP20A	PAETBRUS.PKT	02-11-08	18
Video BIOS	K-ATI VER008.005.028.000			
Device Driver				

* = Firmware includes Remote Supervisor Adapter II Refresh 2 features

Figure 6-21 RSA II firmware level

To update the firmware:

1. Download the latest firmware for the RSA II from one of the following pages:

   – RSA II flash update (PKT files for remote use)

     `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073124`

   – RSA II flash update (Windows executable for local use)

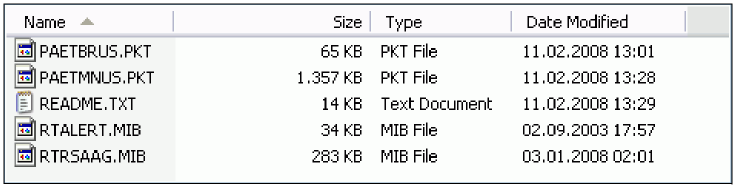
     `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073125`

   – RSA II flash update (Linux executable for local use)

     `http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073123`

   The easiest method is probably to use the PKT files because they can be deployed through the Web browser interface. The remaining steps describe the use of this method.

2. Unpack the ZIP file and you will find the two PKT firmware files as shown in Figure 6-22.



Name	Size	Type	Date Modified
PAETBRUS.PKT	65 KB	PKT File	11.02.2008 13:01
PAETMNUS.PKT	1.357 KB	PKT File	11.02.2008 13:28
README.TXT	14 KB	Text Document	11.02.2008 13:29
RTALERT.MIB	34 KB	MIB File	02.09.2003 17:57
RTRSAAG.MIB	283 KB	MIB File	03.01.2008 02:01

Figure 6-22 RSA II firmware files

> **Tip:** Read the README.TXT file and review all dependencies.

3. Connect to the RSA II by using a Web browser by simply entering the IP address in the address bar. Log in using your credentials (the default user and password are: USERID and PASSW0RD)
4. Select **Tasks** → **Firmware Update**.
5. Click **Browse** to select the first of two files for firmware update. Select and apply the files in the correct order, as follows:

   a. BRUS file (for example, RAETBRUS.PKT), which is the RSA Boot ROM
   b. MNUS file (for example, RAETMNUS.PKT), which is the RSA Main Application

   Restart the RSA only after applying both files.

6. Click **Update** to begin the process. The PKT file is transferred to the RSA II.
7. Click **Continue** to begin the flash writing process.

8. When prompted, do not restart the RSA adapter. You will do this after loading the second firmware PKT file.
9. Repeat the steps for the second PKT file.
10. Restart the adapter by selecting **ASM Control** → **ASM Restart**.
11. After the RSA II is restarted, select **Monitors** → **Vital Product Data** to verify that the new code is applied. This is reported in the RSA II event log.

6.2.6 Implementing the RSA II in the operating system

The device drivers provide operating support and in-band communication with IBM Director. This section describes how to install the RSA II device driver on Windows and Linux platforms. The required device drivers are listed in Table 6-1 on page 308

After you install the operating system, also install the driver or service for the RSA II SlimLine adapter.

Download the RSA II device driver

Download one of the following drivers:

- RSA II service for 32-bit Windows Server 2003 and Windows Server 2008

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5071025
- RSA II service for x64 Windows Server 2003 and Windows Server 2008

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5071027
- RSA II daemon for Linux (RHEL 3 and 4, SLES 9 and 10, ESX 3)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5071676

Windows service installation

The installation of the RSA II server software package is unlike the driver installations of older systems management adapters. It is done by executing the downloaded executable file.

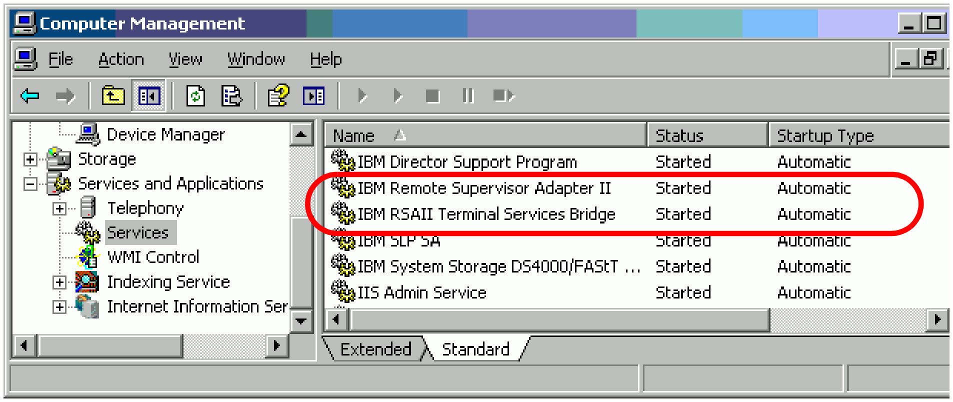
To install the RSA II device driver on the Windows service:

1. Execute the downloaded EXE file on the server with the RSA II.
2. Optionally, click **Change** to specify an alternate temporary folder for the installation files.

   The installation process starts automatically after the files are copied.
3. Follow the instructions.

4. When the installation finishes, you may delete the files in the temporary folder.

To determine if the installation was successful, check the services for the IBM Remote Supervisor Adapter II by selecting **Start** → **All Programs** → **Administrative Tools** → **Services**. Scroll to the service IBM RSAII and verify that the Status indicates Started (Figure 6-23).



Name	Status	Startup Type
IBM Director Support Program	Started	Automatic
IBM Remote Supervisor Adapter II	Started	Automatic
IBM RSAII Terminal Services Bridge	Started	Automatic
IBM SLP SA	Started	Automatic
IBM System Storage DS4000/FAStT ...	Started	Automatic
IIS Admin Service	Started	Automatic

Figure 6-23 RSA II service in Windows operating system

Note: If you have not already done so, change the setting OS USB Selection to **Other OS** in the system BIOS, as shown in Figure 6-14 on page 320.

Linux daemon installation

To install the RSA II device driver (daemon) on Linux, ensure that you downloaded the correct RPM for your Linux distribution. Available packages at the time of writing this book are:

- Red Hat Enterprise Linux 3
- Red Hat Enterprise Linux 4
- Red Hat Enterprise Linux 5
- SUSE Linux Enterprise Server 9
- SUSE Linux Enterprise Server 10
- VMware ESX Server 3
- VMware ESX 3i version 3.5 (embedded or installed)
- VMware ESX 3.5

RHEL and SLES daemon installation

Make sure both the run-time and development *libusb* libraries are installed on your Linux system. Execute the following **rpm query** command to check that the libraries are installed and the version numbers:

```
rpm -qa | grep libusb
```

The command returns the following two libusb entries (if your version numbers are different, that is okay):

- libusb-0.1.6-3
- libusb-devel-0.1.6-3

Review the appropriate readme file of the RPM Package for prerequisites and installation steps.

To install RSA II on RHEL and SLES:

1. Copy the downloaded file to a folder on the Linux server, for example to /tmp/inst. The following RPMs are contained in the installation package:
   - ibmusbasm-1.xx-2.src.rpm for VMware ESX 3.x and 32-bit Linux except RHEL5
   - ibmusbasm64-1.xx-2.src.rpm for 64-bit Linux except RHEL5
   - ibmusbasm-rhel5-1.xx-2.src.rpm for 32-bit RHEL5
   - ibmusbasm64-rhel5-1.xx-2.src.rpm for 64-bit RHEL5
2. Install the daemon (for example, SUSE, where xx is the version) by running:

   ```
   rpm -ivh ibmusbasm-1.xx.i386.rpm
   ```
3. Check that the daemon is running by using the **ps** command, as shown in Example 6-1.

   Example 6-1 RSA II daemon status in Linux OS

   ```
   nux:~ # ps -ef | grep ibmasm
   root 11056 1 0 10:47 pts/1 00:00:00 /sbin/ibmasm
   root 11060 11056 0 10:47 pts/1 00:00:00 /sbin/ibmasm
   root 11062 10996 0 10:48 pts/1 00:00:00 grep ibmasm
   linux:~ #
   ```

   If `/sbin/ibmasm` appears in the list, the daemon is running. The ibmusbasm daemon is started automatically during the boot process of the operating system.

To start the daemon manually, use the command **ibmspup**. To stop the daemon, enter **ibmspdown**.

VMware daemon installation

Make sure both the run-time and development *libusb* libraries are installed on your VMware system. Execute the following **rpm query** command to check that the libraries are installed and the version numbers:

```
rpm -qa | grep libusb
```

The command returns the following two libusb entries (if your version numbers are different, that is okay):

- libusb-0.1.6-3
- libusb-devel-0.1.6-3

If the command does not return these two libusb entries, install them. Both libusb RPMs are in the /VMware/RPMS/ subdirectory on your VMware CD.

To install RSA II on VMware:

1. Download or copy the following RSA II daemon package to the VMware server (xx is the version number of the RSA II package):

   ibm_svc_rsa2_hlp2xxa_linux_32-64.tgz

2. Expand the .tgz file by using the following command:

   ```
   tar xzvf ibm_svc_rsa2_hlp2xxa_linux_32-64.tgz
   ```

   (Expanding the file creates an SRPMS directory.)

3. Change directory to SRPMS:

   ```
   cd SRPMS
   ```

4. Install the 32-bit source RPM Package by issuing an **rpm** command, changing directories, and issuing a second **rpm** command (xx is the version number of the RSA II package):

   ```
   rpmbuild --rebuild ibmusbasm-1.xx-2.src.rpm
   cd /usr/src/redhat/RPMS/i386
   rpm -ivh ibmusbasm-1.xx-2.i386.rpm
   ```

The RSA II daemon is ready to use.

6.2.7 TCP/UDP ports used by the RSA II

The RSA II uses several TCP/UDP ports for communication. If the communication with the RSA II passes through firewalls, it is important to know which ports you have, so you can enable the firewalls and be able to communicate with the RSA. Table 6-3 on page 333 lists the default ports. Remember when you change the ports in the RSA you have to change them in the firewalls too.

Table 6-3 User configurable TCP/IP ports used by the RSA II

Port name	Port number	Description
http	80 (default)	Web server HTTP connection (TCP)
https	443 (default)	SSL connection (TCP)
telnet	23 (default)	Telnet command-line interface connection (TCP)
SSH	22 (default)	Secure Shell (SSH) command-line interface (TCP)
SNMP Agent	161 (default)	SNMP get/set commands (UDP)
SNMP Traps	162 (default)	SNMP traps (UDP)
—	2000 (default)	Remote Console video direct (TCP)
—	3389 (default)	Windows Terminal Service (TCP)
—	6090 (default)	TCP Command Mode (TCP)

Other ports are fixed and cannot be changed, as listed in Table 6-4.

Table 6-4 Fixed TCP/IP ports used by the RSA II

Port number	Description
427	SLP connection (UDP)
1044	Remote disk function (TCP)
1045	Persistent remote disk (disk on card) (TCP)
7070-7077	Partition Management

6.2.8 MIB files

The RSA II supports SNMP from many management tools, including IBM Director. If you require management information base (MIB) files, they are on the RSA II firmware ZIP file that also includes the PKT files:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073124

6.2.9 Error logs

The event log data is available in the NVRAM of the RSA II and kept there until deleted or the battery is exhausted. To access the event log data, select the task

Monitors → **Event Log**. The Web interface provides all information, sorted with the newest first.

You may save this list in ASCII text format by clicking the **Save Log as Text File** button on the bottom of the event log table. After saving it, you may clear the event log by clicking the button **Clear Log**.

You may filter the events at severity (Error, Warning, Info), source, and date. The events are displayed in different colors.

6.3 Use of IBM Director with VMware ESX

With the Service Update1 of version 5.20.2, the IBM Director supports systems in level 0 in ESX Server 3i, and levels 1 and 2 in ESX 3.0.2U1, 3.5.

To manage a Level-0 system, you must use an out-of-band network connection to the RSA II. To manage the server as a Level-1 system, you must have IBM Director Core Services installed. The system management driver is required to manage Level-2 systems by the IBM Director agent.

For more information about implementing and managing objects of levels 0, 1, and 2, see section 6.5, “IBM Director: Implementation of servers” on page 346.

The IBM Systems Software Information center has additional support and guidance for using the IBM Director:

http://publib.boulder.ibm.com/infocenter/eserver/v1r2/index.jsp?topic=/diricinfo_all/diricinfoparent.html

6.4 Active Energy Manager

IBM Systems Director Active Energy Manager (AEM) is an extension of IBM Director. The Active Energy Manager you can use to monitor and manage the energy and thermal consumption of IBM servers and BladeCenter systems.

Systems that are not IBM systems can also be monitored with metering devices, such as PDU+ and sensors. Active Energy Manager is part of a larger energy-management implementation that includes hardware and firmware components.

The Active Energy Manager is also available in a stand-alone version, which runs on top of Embedded Director.

For information and guidance about Active Energy Manager implementation, see the IBM Systems Software Information center at:

http://publib.boulder.ibm.com/infocenter/eserver/v1r2/index.jsp?topic=/aem_310/frb0_main.html

To download the extension to IBM Director and the stand-alone version, go to:

http://www.ibm.com/systems/management/director

The requirements for using the AEM are as follows:

- IBM Director 5.20.2 requirements, see IBM Director information center:

  http://publib.boulder.ibm.com/infocenter/eserver/v1r2/index.jsp?topic=/diricinfo_all/diricinfoparent.html

- Supported operating systems:
  - Red Hat Enterprise Linux (AS and ES, 4.0, 5.0, and 5.1)
  - SUSE Linux Enterprise Server (9 and10)
  - Windows Server 2003 Enterprise Edition
- Supported managed hardware:
  - Rack-mounted server
  - BladeCenter Director-managed objects
  - iPDUs

6.4.1 Active Energy Manager terminology

The following terms apply to the hardware in an Active Energy Manager environment:

management server A server on which both IBM Director Server and Active Energy Manager Server are installed.

management console A system on which both IBM Director Console and Active Energy Manager Console are installed.

BladeCenter system A chassis and a number of modules. The chassis is the physical enclosure that houses the modules. The modules are the individual components, such as blade servers and blowers, that are inserted into bays in the chassis.

rack-mounted server A stand-alone system with a Baseboard Management Controller (BMC).

managed system Either a BladeCenter system or a rack-mounted server in an IBM Director environment.

module A BladeCenter component that is inserted in a bay in a BladeCenter chassis. See BladeCenter system. The *management module* and *blade server* are two types of modules you can insert in a BladeCenter chassis.

management module The BladeCenter component that handles system-management functions. It configures the BladeCenter chassis and switch modules, communicates with the blade servers and all I/O modules, multiplexes the keyboard/video/mouse (KVM), and monitors critical information about the chassis and blade servers.

blade server A complete server with a combination of processors, memory, and network interfaces, and with supporting hardware and firmware. The blade server may occupy one or more slots in a BladeCenter chassis.

6.4.2 Active Energy Manager components

The three major Active Energy Manager (AEM) components are:

- Active Energy Manager Server
- Active Energy Manager Console
- Active Energy Manager Database

This section also discusses AEM’s monitoring capabilities.

Active Energy Manager Server

Active Energy Manager Server maintains the AEM environment and manages all AEM operations. AEM Server communicates out-of-band with each managed object to collect power information. In the case of a BladeCenter chassis, it communicates with the management module in the chassis. In the case of a rack-mounted server, it communicates with the BMC.

AEM Server also communicates with the Director Server to provide event filtering and event actions that support IBM Director event action plans, and communicates with Director Console to display status and to allow the user to perform operations. You must install AEM Server on the IBM Director management server. Power data is collected only while the Active Energy Manager Server is running. When you install AEM, the server starts running. It runs when the IBM Director Server is running. By default, it collects data on the BladeCenter chassis and rack-mounted servers every minute, but the collection interval can be configured on the Manage Trend Data window in the a graphical user interface (GUI).

Active Energy Manager Console

Active Energy Manager Console provides the GUI to AEM in IBM Director Console. AEM Console displays rack-mounted servers and BladeCenter chassis at the same level in the navigation tree. Rack-mounted servers are represented by leaf nodes (lowest point in the tree), while BladeCenter chassis have sub-nodes representing the power domains and modules within the chassis. When you install AEM Server, AEM Console is installed automatically. You can also install AEM Console on all management consoles from which a system administrator remotely accesses the management server and performs AEM tasks.

Active Energy Manager Database

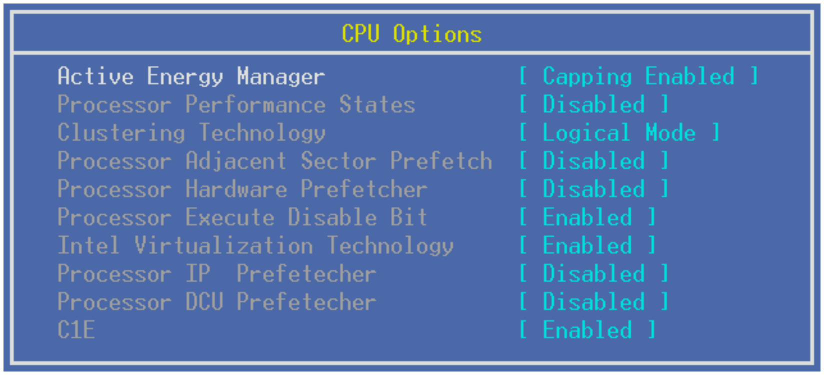
The Active Energy Manager Database stores information collected by AEM Server. AEM saves configuration information and historical power and temperature data for managed systems in a Derby database. The AEM Database is created in the Data subdirectory of the IBM Director installation directory. When you uninstall IBM Director Server, it does not remove the AEM Database unless you request that customizations be deleted during the uninstallation of IBM Director Server. The size of the AEM Database directly correlates to the short-term data-collection interval, the long-term data-collection interval, and the number of days that short-term and long-term trend data is kept. All of these values are configurable. You can control them and the size of the AEM Database by using the Manage Trend Data window.

Managed systems

Active Energy Manager can monitor power consumption for selected rack servers, BladeCenter chassis and blade servers, and intelligent PDUs.

You can implement each rack server, or modular system, which enables power capping features to provide the power measurement information to the AEM. Older server or expansion units, that do not have this feature can be implemented by using iPDUs. See 6.8, “Power Distribution Units (PDU)” on page 357. The iPDUs monitor the power usage on their outlets and provide the captured information by the integrated SNMP out-of-band support.

The power capping feature in the BIOS of your x3850 M2 and x3950 M2, as shown in Figure 6-24 on page 338, is enabled by default. This feature is available for single node x3950 M2 systems. For details see section 3.1.3, “Processor (CPU) configuration options” on page 99.



```
                         CPU Options
  Active Energy Manager                  [ Capping Enabled ]
  Processor Performance States           [ Disabled ]
  Clustering Technology                  [ Logical Mode ]
  Processor Adjacent Sector Prefetch     [ Disabled ]
  Processor Hardware Prefetcher          [ Disabled ]
  Processor Execute Disable Bit          [ Enabled ]
  Intel Virtualization Technology        [ Enabled ]
  Processor IP  Prefetecher              [ Disabled ]
  Processor DCU Prefetecher              [ Disabled ]
  C1E                                    [ Enabled ]
```

Figure 6-24 x3850 M2 and x3950 M2: power capping feature in BIOS is enabled

6.4.3 Active Energy Manager tasks

After the AEM is launched, it displays the AEM console in the IBM Director Console. Its functions enable you to monitor and collect power-consumption data from power devices, create trend data, export data and manage energy on certain hardware, for example by using capping features.

A new Active Energy Manager task icon is added in the IBM Director Management console after installation of the AEM extension (see Figure 6-25).

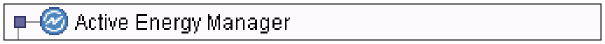
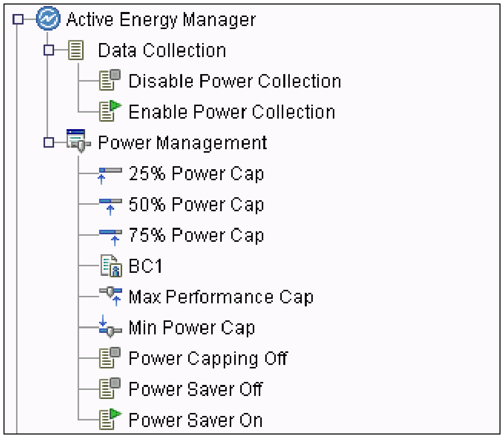


Figure 6-25 Active Energy Manager task icon

You use the Active Energy Manager task on the IBM Director Console menu to launch the Active Energy Manager GUI. Use the GUI to view and monitor power consumption on various rack-mounted servers, BladeCenter chassis and iPDUs in the IBM Director environment (see Figure 6-26 on page 339).



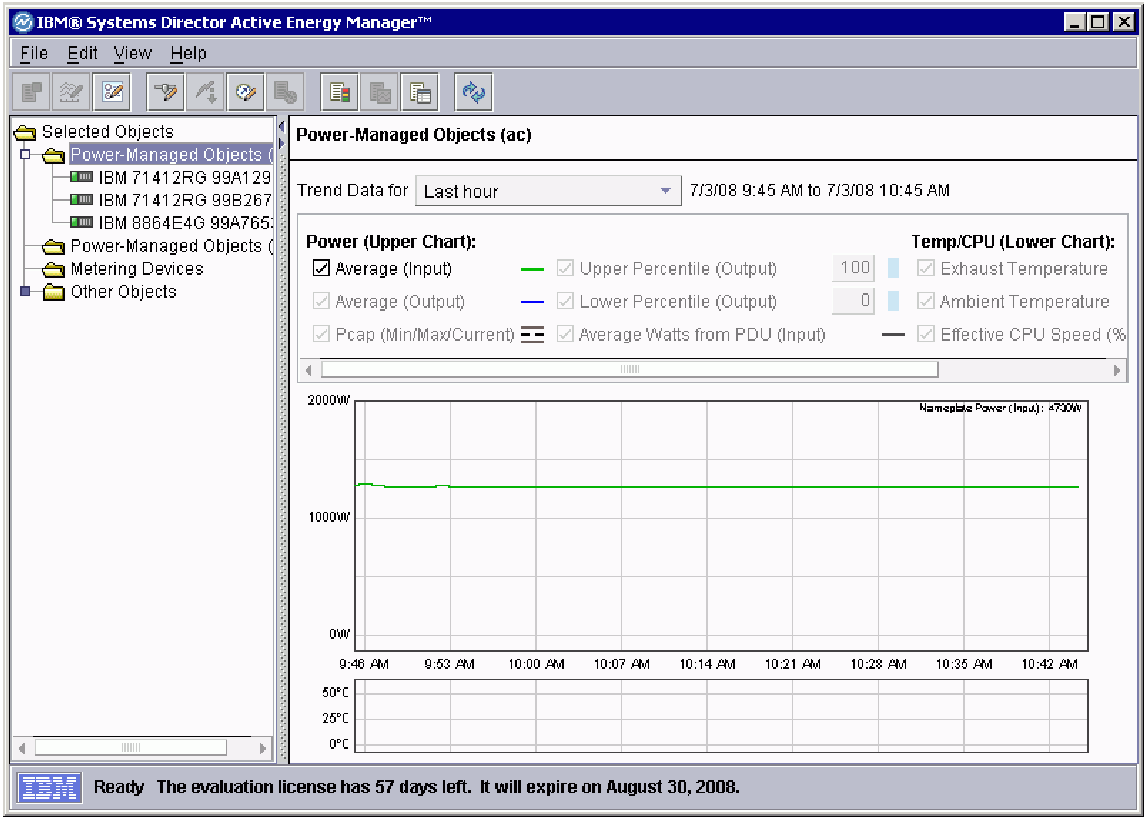
Active Energy Manager
- Data Collection
  - Disable Power Collection
  - Enable Power Collection
- Power Management
  - 25% Power Cap
  - 50% Power Cap
  - 75% Power Cap
  - BC1
  - Max Performance Cap
  - Min Power Cap
  - Power Capping Off
  - Power Saver Off
  - Power Saver On

Figure 6-26 Active Energy Manager options

When you install the AEM extension in your IBM Director environment, its task is added to IBM Director Console. Start the Active Energy Manager task by dragging and dropping the task icon onto any of the following targets:

- If the target is a blade server, the display consists of the BladeCenter chassis containing the blade with the targeted blade preselected.
- If the target is a BladeCenter chassis, the display consists of the chassis with the chassis itself preselected.
- If the target is a rack-mounted server like your x3850 M2 or x3950 M2, the display consists of the server with the server itself preselected.
- If the target is a group of rack-mounted servers or BladeCenter chassis, or both, the display consists of all chassis or rack-mounted servers in the group with the first chassis or Start Active Energy Manager for all managed systems.
- If the target is an iPDU, the current date for the PDU and the capacity and usage for the load groups is displayed with the iPDU itself preselected.

When you start AEM for only a single managed system or for a group of managed systems, the Active Energy Manager window opens and displays a tree of only the selected objects. See Figure 6-27 on page 340. The left panel in Active Energy Manager contains only the selected managed systems, those chassis that contain the systems that have also been selected, and the managed systems in a selected group.



Trend Data for Last hour 7/3/08 9:45 AM to 7/3/08 10:45 AM

☑ Average (Input)
☑ Average (Output)
☑ Pcap (Min/Max/Current)
☑ Upper Percentile (Output) 100
☑ Lower Percentile (Output) 0
☑ Average Watts from PDU (Input)
☑ Exhaust Temperature
☑ Ambient Temperature
☑ Effective CPU Speed (%

Ready The evaluation license has 57 days left. It will expire on August 30, 2008.

Figure 6-27 Active Energy Manager: managed systems

When you select a system, Active Energy Manager displays the entire tree for all managed systems. The advantage of starting Active Energy Manager in this manner is that you see only the subset of managed systems that you are interested in. The disadvantage is that you can manage only those managed systems that were selected when you started Active Energy Manager. If you have to manage other managed systems at a later time, you must start another Active Energy Manager task for those additional managed systems.

6.4.4 Active Energy Manager 3.1 functions

This section provides an overview of the functions supported through the Active Energy Manger. The two classes of functions are monitoring functions, which are always available, and management functions, which require a license fee to be paid to allow them to work beyond a 90-day trial period:

- Monitoring functions (no charge)
  - Power Trending
  - Thermal Trending
  - iPDU support
  - Display details of current power and temperature
  - Monitor and generate events
- Management functions (fee-based license)
  - Power Capping
  - Power Savings Mode

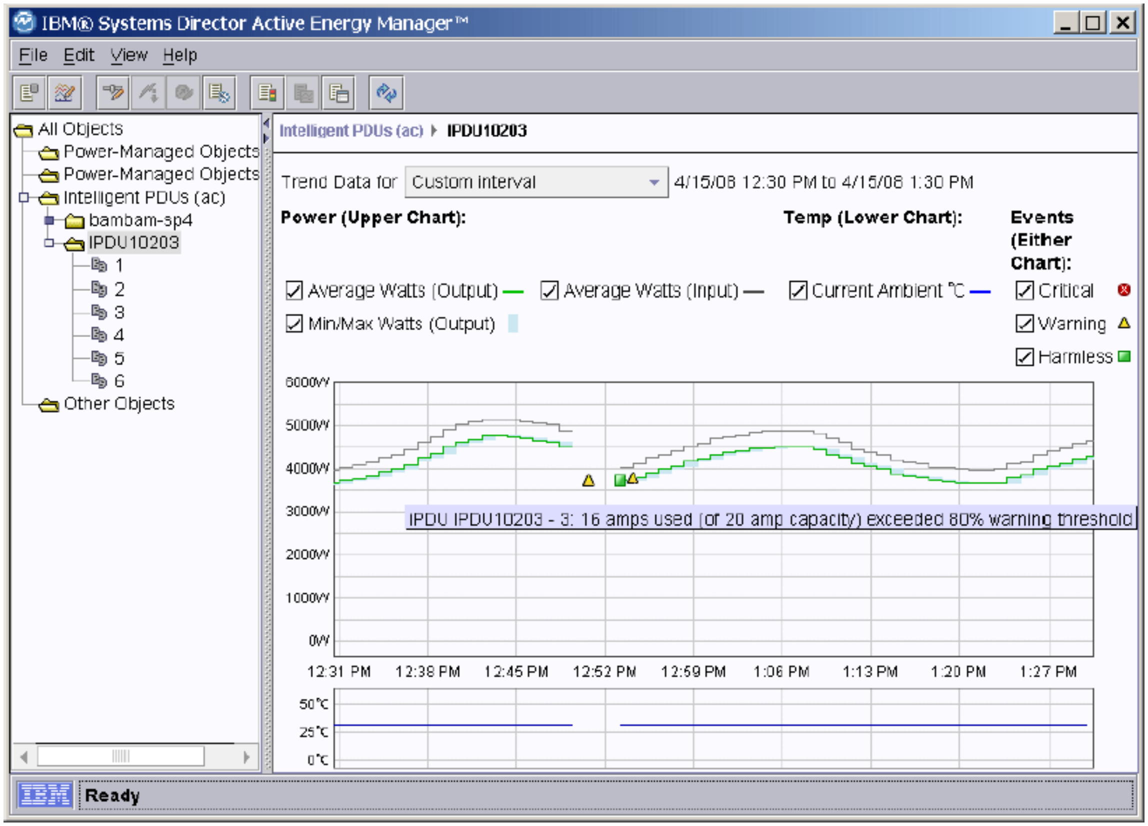
Monitoring functions

These functions provide the general information about the power consumption of the data center and possible anomalies. You control what information to collect about power and temperatures, how often to collect it, and how long to store the data. You can view graphs or tables of data trends, and also export the tabular data as a spreadsheet, XML, or HTML formats.

Power trending

The Energy Scale architecture provides continuous power usage data collection. The power usage data can be displayed from Active Energy Manager. Data administrators can use the information to project the power consumption of the data center at various times of the day, week, or month and use that information to identify anomalies, manage loads when electrical demands or costs are high and, if the system supports power capping, determine appropriate times and levels for power caps (see “Power capping” on page 346).

In Figure 6-28 on page 342, the top chart shows an example of the power trending information, set at custom intervals. Figure 6-27 on page 340 shows an example of power trending in the last hour. Various controls on the panel help define the interval and the data to display.



Trend Data for Custom interval 4/15/08 12:30 PM to 4/15/08 1:30 PM

☑ Average Watts (Output) ☑ Average Watts (Input) ☑ Current Ambient °C ☑ Critical ☑ Warning ☑ Harmless
☑ Min/Max Watts (Output)

IPDU IPDU10203 - 3: 16 amps used (of 20 amp capacity) exceeded 80% warning threshold

Figure 6-28 Active Energy Manager: graphical trending view for an iPDU

The trending graph can also display events that are monitored by the Active Energy Manager. Details about each event recorded can be displayed by dragging the mouse pointer over the event symbol.

The types of events monitored by the Active Energy Manager include power component failures, offline and online actions, power management actions, and power and thermal critical warnings.

Thermal trending

The thermal sensors are primarily based on the digital thermal sensors available on the systems. All of the logical sensors are the result of firmware running on the thermal and power management device (TPMD) using the raw data provided by the hardware, and converting it into values that are then fed into the control loops.

In Figure 6-27 on page 340, the bottom chart shows an example of thermal trending information. The device in this example shows the ambient and exhaust temperature. Other devices might not available to show an exhaust temperature.

iPDU support

Section 6.8, “Power Distribution Units (PDU)” on page 357 briefly describes on how iPDUs work with the AEM. An iPDU can record similar information to other power monitored hardware. The power trending data in Figure 6-28 on page 342 is from an iPDU.

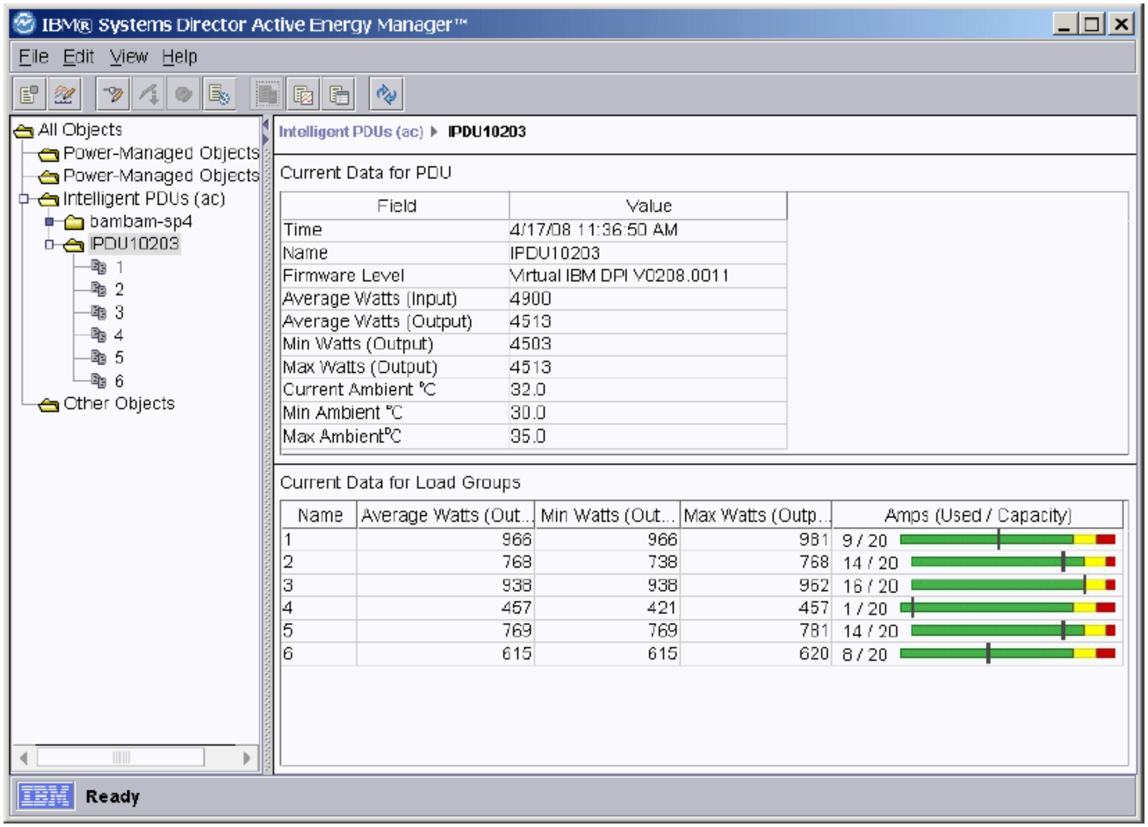
Other detailed information can also be reported for each iPDU:

- Time at which the current data was collected
- Name assigned to the iPDU
- Firmware level of the iPDU
- Average input and output watts for the iPDU over the last polling interval
- Minimum and maximum output watts for the iPDU over the last polling interval
- Current, minimum, maximum ambient temperature of the iPDU at the last polling interval

You can also display detailed information about each load group in the iPDU:

- Name assigned to the load group
- Average output watts which is the DC power currently being consumed as reported by the power meter, or shown as two dashes (`--`) if power data is not available
- Minimum and maximum output watts for the load group over the last polling interval
- Amps used in relation to capacity

Figure 6-29 on page 344 shows how information for an iPDU and its load groups is presented.



Current Data for PDU

Field	Value
Time	4/17/08 11:36:50 AM
Name	IPDU10203
Firmware Level	Virtual IBM DPI V0208.0011
Average Watts (Input)	4900
Average Watts (Output)	4513
Min Watts (Output)	4503
Max Watts (Output)	4513
Current Ambient °C	32.0
Min Ambient °C	30.0
Max Ambient°C	35.0

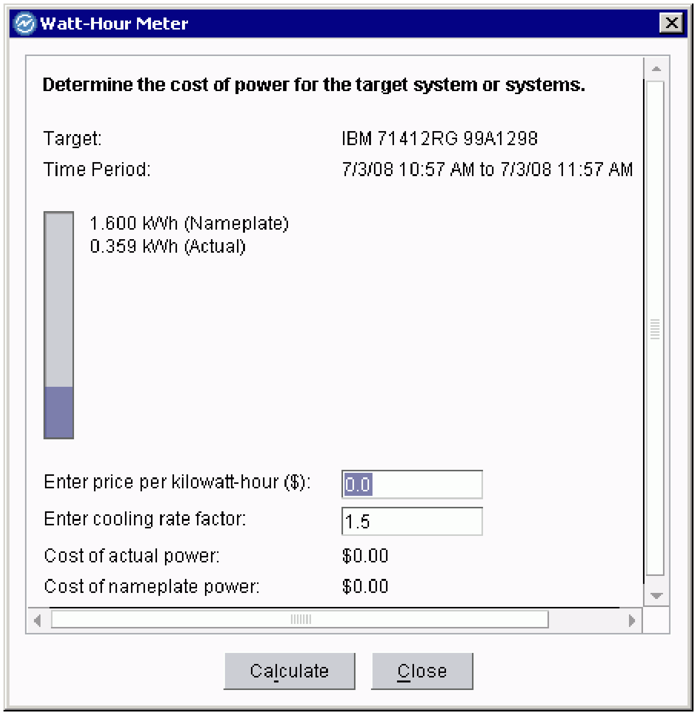
Current Data for Load Groups

Name	Average Watts (Out...	Min Watts (Out...	Max Watts (Outp...	Amps (Used / Capacity)
1	966	966	981	9 / 20
2	768	738	768	14 / 20
3	938	938	962	16 / 20
4	457	421	457	1 / 20
5	769	769	781	14 / 20
6	615	615	620	8 / 20

Figure 6-29 Active Energy Manager: iPDU details

Watt-Hour Meter

The Watt-Hour Meter displays the amount of energy used for a given system or group of systems over a specified period of time, and calculates the corresponding cost of that energy, shown in Figure 6-30 on page 345. It gives a visual indication of the number of watt-hours consumed by the target object or objects over the specified time period and provides a comparison to what would have been consumed had nameplate power been drawn over that entire period.



Watt-Hour Meter

Determine the cost of power for the target system or systems.

Target: IBM 71412RG 99A1298

Time Period: 7/3/08 10:57 AM to 7/3/08 11:57 AM

1.600 kWh (Nameplate)
0.359 kWh (Actual)

Enter price per kilowatt-hour ($): 0.0

Enter cooling rate factor: 1.5

Cost of actual power: $0.00

Cost of nameplate power: $0.00

Figure 6-30 Active Energy Manager: Watt-Hour Meter

Management functions

The management functions are Power Saver and Power Capping. They provide methods to reduce energy consumption by dropping the processor voltage and frequency. The functions are available for a 90-day trial use. You must purchase a license and install a key to use these functions beyond the trial period.

Power Saver mode

Power Saver mode provides a way to save power by dropping the voltage and frequency a fixed percentage. This percentage is predetermined to be within a safe operating limit and is not user configurable. Under current implementation this is a 14% frequency drop. In the Active Energy Manager user interface the Power Saver mode is set to enable or disable.

One possible use for Power Saver would be to enable it when workloads are minimal, such as at night, and then disable it in the morning. When Power Saver is used to reduce the peak energy consumption, it can lower the cost of all power used. At low processor utilization, the use of Power Saver increases processor utilization such that the workload notices no performance effect. Depending on workload, this can reduce the processor power usage by 20-30%.

The IBM Director scheduler can be used to automate the enabling and disabling of Power Saver mode based on projected workloads. Scripts can also be run to enable and disable it based on processor utilization.

Power Saver mode is available on the x3850 M2 and x3950 M2 systems to.

- Set a power cap:
  - Guarantees server does not exceed a specified number of watts
  - If cap is reached, processor is throttled and voltage is reduced
  - Available on P6 Blades and selected System x servers and blades
- Drag a function onto a system:
  - For example, drag *Power Saver* on to a server in middle pane
  - Perform a task now, or schedule it for later

Power capping

Power Capping enforces a user-specified limit on power usage. You set and enable a power cap from the Active Energy Manager interface. In most data centers and other installations, when a machine is installed, a certain amount of power is allocated to it.

Generally, the amount is what is considered to be a *safe* value, which is often the label power for the system. This means that a large amount of reserved, extra power is never used. This is called the *margined power*. The main purpose of the power cap is not to save power but rather to allow a data center operator the ability to reallocate power from current systems to new systems by reducing the margin assumed for the existing machines. Thus the basic assumption of power capping allows an operator to add extra machines to a data center, which previously had all the data center power allotted to its current systems.

Power capping provides the guarantee that a system does not use more power than assigned to it by the operator.

6.5 IBM Director: Implementation of servers

This section does not discuss the general use of the IBM Director. Rather, it discusses how to implement your x3850 M2 and x3950 M2 into an existing IBM Director environment.

The IBM Systems Software Information Center provides technical information about working with the IBM Director:

http://publib.boulder.ibm.com/infocenter/eserver/v1r2/topic/diricinfo_all/diricinfoparent.html

Requirements for implementation include:

- IBM Director version 5.20.2 and above:
  - x3850 M2, machine type 7141 without installed ScaleXpander chip
  - x3850 M2, LSI 1078 IR onboard SAS controller
  - ServeRAID-MR10k SAS/SATA Controller, part number 43W4280[1]
  - ServeRAID-MR10M SAS/SATA Controller, part number 43W4339
- IBM Director version 5.20.2 Service Update 1
  - x3850 M2, machine type 7141 with installed ScaleXpander chip
  - x3950 M2, machine type 7141

The managing of objects is implemented by three levels.

- Level-0 manageable objects:

  The requirement to manage a Level-0 object is the capability to access a service processor out-of-band (through LAN connection)
- Level-1 manageable objects

  A system must have installed the IBM Director Core Services
- Level-2 manageable objects

  The system management driver is required for managing Level-2 systems by the IBM Director agent.

6.5.1 Integrating x3850 M2 and x3950 M2 into IBM Director

Understand the requirements and detection of manageable objects.

Requirements

Integrating your x3850 M2 or x3950 M2 server as a Level-0 managed system does not require that the drivers or the IBM Director agent be installed. Level-0 systems are managed by the embedded service processor or optional installed RSA II adapter out-of-band through the service processor LAN interface communication.

Use of Level-1 or Level-2 managed systems is based on the Common Information Model (CIM) standard. After installation of the IBM Director Core Services (Level 1) and the IBM Director agent (Level 2), a service processor device driver, and possibly a shared library driver to access the device driver are required.

[1] Level-1 systems with a LSI1078-based MegaRAID controller require an installed LSI-MegaRAID-Provider, after the installation of the IBM Director agent.

The IBM Director differs from the embedded service processor and Remote Supervisor Adapter device drivers. Because the x3850 M2 and x3950 M2 are shipped with an RSA II, the installation of the embedded BMC driver is not required. The BMC device is superposed if the RSA II is found in the system; the complete management is done through the RSA interface.

Detection of manageable objects

The IBM Director detects your systems automatically or you can detect it manually. After a system is detected, an icon is added in the IBM Director console. The basic management server saves the addresses of the detected systems in the IBM Director database and supports, based on the x3850 M2 and x3950 M2 server.

6.5.2 Level 0: Implementation by service processors

Understand configuration of LAN interface and implementation of Level 0.

LAN interface

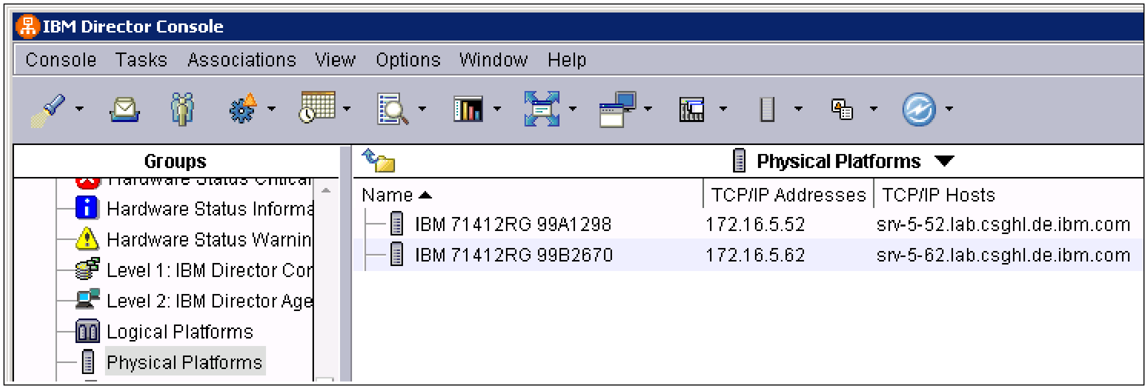
Systems that have an embedded service processor or installed RSA II can be managed and monitored if the communication to the LAN interface is guaranteed.

The RSA II adapter in the x3850 M2 and x3950 M2 replaces the embedded BMC, so the LAN interface for the RSA II must be configured only to manage it by the IBM Director. Section 6.2.2, “RSA LAN configuration in BIOS” on page 319 discusses how to set a valid network configuration.

Implementation of Level-0 systems in the IBM Director

The IBM Director polls the network to scan for new manageable objects automatically, but can be discovered manually also. If a valid service processor can be found, it adds an icon in the Physical Platforms group, as shown in Figure 6-31 on page 349.

To access the system, right-click the system name and select **Request access**. Enter the service processor login credentials to be able to manage the RSA II adapter from IBM Director.



Name	TCP/IP Addresses	TCP/IP Hosts
IBM 71412RG 99A1298	172.16.5.52	srv-5-52.lab.csghl.de.ibm.com
IBM 71412RG 99B2670	172.16.5.62	srv-5-62.lab.csghl.de.ibm.com

Figure 6-31 IBM Director Console: Physical Platforms

6.5.3 Level 1: Implementation by the IBM Director Core Services

The Core Services component of IBM Director provides hardware-specific functions for intercommunication between managed systems. Systems with installed IBM Director Core Services are called *Level-1 managed systems*.

The services use standard agent functions, such as detection, authentication, and management. Core Services also installs a Service Location Protocol (SLP) agent, and a Common Information Model Object Manager (CIMOM) with SSL support (at Linux), CIM-classification for WMI (at Windows), and SSH server and platform-specific instrumentation.

Supported tasks

Core services allow the following tasks to be performed:

- Inventory information
- Upgrading to a Level-2 by deployment of the IBM Director agent
- Management of events by using of event action plans
- Monitoring hardware status
- Deployment of system update package
- Remote management (if SSH service started)
- Running of command-line scripts

Installation of the IBM Director Core Services

The IBM Director Core Services is on the IBM Director installation CD or can be downloaded as a single installation package from:

http://www.ibm.com/systems/management/director/downloads.html

Linux installation

Install the software as follows:

1. Extract the RPM Package:

   ```
   tar -xvf dir5.20.2_coreservices_linux.tar
   ```

2. Change to the extraction folder:

   ```
   cd /install_files_directory/FILES
   ```

3. Run the installation script by using default settings:

   ```
   ./dir5.20.2_coreservices_linux.sh
   ```

4. Start the core services by using the response file:

   ```
   ./dir5.20.2_coreservice_linux.sh -r /directory/response.rsp
   ```

   The directory is the folder, which contains the response file.

Windows installation

Install the software as follows:

1. Extract the package:

   ```
   unzip dir5.20.2_coreservices_windows.zip
   ```

2. Change to the extraction folder:

   ```
   cd /install_files_directory/FILES
   ```

3. Copy the response file (\directory\FILES\coresvcs.rsp) to another location.
4. Open the coresvc.rsp with an ASCII editor and change the contents as commented in this file. Save it with a new filename, such as:

   ```
   responsefile.rsp
   ```

5. Select **Start** → **Run**, and then type:

   ```
   \directory\FILES\dir5.20.2_coreservices_windows.exe /s /a
   installationtype rsp=responsefile.rsp option
   ```

   – *directory* is the installation files folder or cd-rom path CD:\coresvc\agent\windows\i386\

   – /s is optional to hide file extraction dialog

   – *installationtype* is either:

     - `unattended`: installation progress, no user intervention
     - `silent`: hide progress

   – *responsefile.rsp* is path and name of the created response file

   – *option* indicates:

     - `waitforme` (ensure complete installation)

- `debug` (logging of the Microsoft Windows installation program messages)
- `log=logfile` (creates logfile)
- `verbose` (detailed logging)

6. Restart the operating system after the installation is completed.

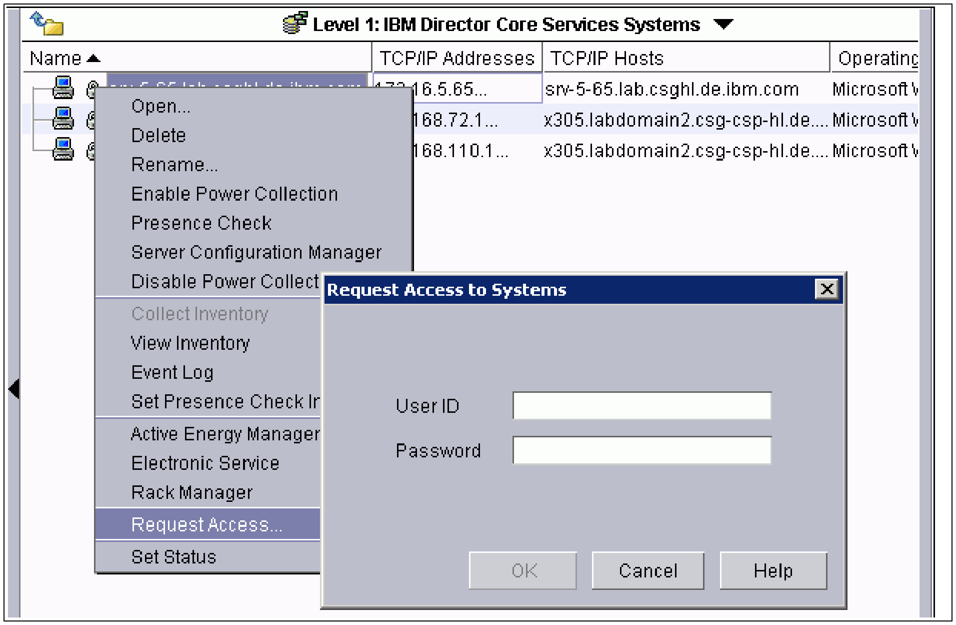
Implementation of Level-1 systems in IBM Director

After the IBM Director Core Services are installed and an automatic or manual Discovery is performed, the IBM Director detects the systems.

After IBM Director detects the system, that system is listed under Groups as:

> Level1: IBM Director Core Services Systems

To unlock the system, right-click the system name, select **Request Access**, and provide your operating system logon credentials. See Figure 6-32. The lock disappears and the system can now be managed from IBM Director Console.



Name ▲	TCP/IP Addresses	TCP/IP Hosts	Operating
	16.5.65...	srv-5-65.lab.csghl.de.ibm.com	Microsoft W
	168.72.1...	x305.labdomain2.csg-csp-hl.de....	Microsoft W
	168.110.1...	x305.labdomain2.csg-csp-hl.de....	Microsoft W

Figure 6-32 IBM Director: requesting access to the discovered system

6.5.4 LSI MegaRAID Provider

The LSI MegaRAID Provider is required to receive events of the following options in your server:

- LSI 1064e (SAS HBA)
- LSI 1078 IR (Integrated RAID)
- ServeRAID-MR10k-SAS/SATA-Controller, part number 43W4280
- ServeRAID-MR10M-SAS/SATA-Controller, part number 43W4339

The LSI MegaRAID Provider is supported on Level-1 systems with the following operating systems:

- Red Hat Enterprise Linux, version 4.0
- Red Hat Enterprise Linux, version 5.0
- SUSE Linux Enterprise Server 9 for x86
- SUSE Linux Enterprise Server 10 for x86
- Microsoft Windows

The IBM Director Core Services must be installed before you can install the LSI MegaRAID Provider. Download this extension from:

http://www.ibm.com/systems/management/director/downloads.html

LSI MegaRAID Provider is not supported with:

- VMware operating systems
- Operating systems with enabled Xen virtualization layer

Linux installation of LSI MegaRAID Provider

Install this extension by using the following command:

```
rpm -ivh lsi_mr_hhr_xx.xx.xx.xx-x.os.kernel.rpm
```

The installation is completed. Restart your operating system.

Windows installation of LSI MegaRAID Provider

Install this extension by using the batch file:

IndicationSubscription.bat

This file is located in either of the following folders:

- C:\Program Files\Common Files\IBM\ICC\cimom\bin
- C:\Program Files (x86)\Common Files\IBM\ICC\cimom\bin

The installation is completed. Restart your operating system.

6.5.5 Level 2: Implementation by the IBM Director agent

The IBM Director agent enables all the agent options, which are used for communication and management of the system. Functions vary, depending on the operating system and type of hardware.

The use of the IBM Director agent requires installation of the service processor driver. The x3850 M2 and x3950 M2 is shipped with the RSA II so installing the BMC driver is not necessary.

Installation of the RSA II device driver

Section 6.2.6, "Implementing the RSA II in the operating system" on page 329 describes how to install the device driver in the Linux and Windows operating systems.

Installation of the IBM Director agent

IBM Director agent is on IBM Director installation CD or you can download it as single installation package from:

http://www.ibm.com/systems/management/director/downloads.html

Linux installation of IBM Director agent

To install IBM Director agent:

1. Extract the RPM Package, by using the following command:

   ```
   tar -xvf dir5.20.2_agent_linux.tar
   ```

2. Change to the extraction folder:

   ```
   cd /install_files_directory/FILES
   ```

3. Run the installation script by using default settings:

   ```
   ./dir5.20.2_agent_linux.sh
   ```

4. Optional: By default, AES-algorithm (Advanced Encryption Standard) is enabled. If you want to disable it or change the security settings, run the following security command (where *install_root* is the root directory of your IBM Director installation):

   ```
   install_root/bin/cfgsecurity
   ```

5. If you want to start or stop the agent, use the response file:

   – Start: `install_root/bin/twgstart`
   – Stop: `install_root/bin/twgstop`

Windows installation of IBM Director agent

To install IBM Director agent:

1. Extract the package by using the command:

   ```
   unzip dir5.20.2_agent_windows.zip
   ```

2. Change to the extraction folder:

   ```
   cd /install_files_directory/FILES
   ```

3. Copy the response file (\directory\FILES\coresvcs.rsp) to another location.
4. Open the coresvc.rsp with an ASCII editor and change the contents as commented in this file. Save it with a new filename, such as:

   ```
   responsefile.rsp
   ```

5. Select **Start** → **Run**, then type:

   ```
   \directory\FILES\dir5.20.2_agent_windows.exe /s /a installationtype
   rsp=”responsefile.rsp option
   ```

   – *directory* is installation files folder or cd-rom path CD:\coresvc\agent\windows\i386\

   – /s is optional to hide file extraction dialog

   – *installationtype* can be:

     - `unattended`: installation progress, no user intervention
     - `silent`: hide progress

   – *responsefile.rsp* is path and name of the created response file

   – *option* can be:

     - `waitforme` (ensure complete installation),
     - `debug` (logging of the Microsoft Windows installation program messages),
     - `log=logfile` (creates logfile),
     - `verbose` (detailed logging)

6. If you enabled the following option in the response file, you must restart the operating system:

   ```
   RebootIfRequired=Y (Yes)
   ```

Implementation of Level-2 systems in IBM Director

After IBM Director detects the system, that system is listed under Groups as:

Level2: IBM Director Agents

To unlock the system, right-click the system name, select **Request Access**, and provide your operating system logon credentials. See Figure 6-32 on page 351.

The lock disappears and the system can now be managed from IBM Director Console.

6.6 System management with VMware ESXi 3.5

Understand hypervisors and implementation of them.

6.6.1 Hypervisor systems

The VMware ESXi 3.5 Embedded and Installable hypervisors offer several management features through integrated CIM and SNMP protocols.

The Embedded hypervisor ESX 3i does not have a service console, unlike the Installable version 3.5 has. This means an IBM Director agent cannot be installed to manage this system as a Level-2 Agent system in your system with the embedded hypervisor running. Instead, the Embedded hypervisor relies on CIM and SNMP for remote management.

6.6.2 Implementation of x3850 M2 Hypervisor systems

IBM Director manages an embedded hypervisor system as a Level-0 system to discover, inventory, and report color-coded hardware failure and RAID events.

You do not have to possess special operating system knowledge, do not have to maintain user accounts and passwords, and do not have to install additional security tools, antivirus tools, or the need to backup the operating system.

A remote command-line interface is enabled to run scripts in the same syntax as with previous ESX versions. It includes features for:

- Host configuration (esxcfg-advcfg)
- Storage configuration (esxcfg-nas, esxcfg-swiscsi, esxcfg-mpath, vmkfstools)
- Network configuration (esxcfg-vswitch, esxcfg-vnic)
- Maintenance and patch (esxcfg-dumpart)
- Backup (VCBMounter, fast copy)
- Monitoring (esxtop, vmkuptime)

6.7 Power management

An aspect that is becoming more significant in the maintenance of a data center is reducing costs for power consumption by ensuring the server infrastructure is more efficient and utilized more effectively. In the past, systems were typically designed for maximum performance without the requirement of keeping power consumption to a minimum.

Newer systems such as the x3850 M2 and x3950 M2 servers have power supplies that regulate the power usage by active electronic parts and by the system-adjusted firmware. As a result, the efficiency can now reach a level above 90%. New power supplies have a power factor to reduce the ratio between the amount of dissipated (or consumed) power and the amount of absorbed (or returned) power.

However this is not the only approach to optimize your yearly IT environment costs in your data center. More components such as processor cores, memory, or I/O components for each system and rack can result in less space necessary in the rack for your IT solution.

6.7.1 Processor features

The Intel Xeon processors have a number of power-saving features. The C1E Enhanced Halt State as discussed in “C1E” on page 110, reduces the internal core frequency, and is followed by reducing the internal core voltages if the operating system is in a low state.

Enhanced Intel SpeedStep Technology (EIST), developed by Intel as the advancement of the C1E Enhanced Halt State, typically allows only one low (idle) or high power state. EIST allows a gradual reduction in the core frequency and core voltages.

Another EIST feature allows you to control the power usage on a system. This feature is known as *power capping* and can be enabled in the BIOS of the x3850 M2 and x3950 M2 server systems. This feature must be enabled so it works in IBM Active Energy Manager as described in “Active Energy Manager (power capping)” on page 100. IBM added this feature as a key component of the IBM Cool Blue™ portfolio within Project Big Green.

Project Big Green is an IBM initiative that targets corporate data centers where energy constraints and costs can limit their ability to grow.

Getting a server controlled and managed to a defined level of power usage is important. It can be important for you, if you must limit the power usage at your data center IT environment to reduce costs for energy. Section 6.4, “Active Energy Manager” on page 334 explains the use of this application as an IBM Director extension.

Power capping is an effective way of keeping control of energy costs. However, it might not be appropriate for all systems, especially those running applications that are processor-intensive, because the method of capping power often involves slowing down the processors. As a result, IBM decided not to make power capping available on multinode x3950 M2 systems because this would be contrary to the design and purpose of these systems. Power capping is still an option on single-node systems, however.

6.7.2 Power consumption measurement and capping

Monitoring the power usage of components that do not have built-in power monitoring (such as I/O subsystems and older servers) can be achieved by using an intelligent Power Distribution Unit (iPDU). An iPDU has power usage measurement capabilities that report through SNMP to applications such as IBM Active Energy Manager.

We discuss these in detail in 6.8, “Power Distribution Units (PDU)” on page 357.

6.7.3 Virtualization

Intelligent applications are developed to help a system be more efficient and fully loaded. Virtualization technologies in the various NUMA-aware operating systems are enabled by Intel Virtualization Technology architecture in Xeon processors, as follows:

- VMware: ESX Server now in ESX 3.5 and ESXi 3.5
- Red Hat Enterprise Linux 5 with Xen
- SUSE Linux Enterprise Server 10 with Xen
- Microsoft: Windows 2003 with Microsoft Virtualization Server
- Microsoft: Windows 2008 and the Hyper V

This allows the usage of efficient multicore systems by running multiple virtual machines in parallel on one system.

6.8 Power Distribution Units (PDU)

IBM completes the offering of rack power solutions with Ultra™ Density Enterprise Power Distribution Units (PDU). This adds further four-rack or six-rack power components to the IBM rack power distribution solution. IBM Ultra Density

Enterprise PDUs share a common space-efficient, power-dense design across the line.

These help to quickly and simply deploy, protect, and manage your high-availability IBM System x3850 M2 and x3950 M2 and non-server equipment such as expansions for storage, tape backup, or non-IBM hardware.

All enterprise PDUs are designed in 1U full-rack size and can be mounted vertically inside the rack.

6.8.1 Key features

The enterprise PDU has the following key features

- Designed with input cable connection, C19 outlets, communication connections, and breakers on one face to improve usability and cable management
- C19 (16A) or C13 (10A) outlets on front or rear panel
- Includes hardware to mount in either a standard rack space or side pocket of rack
- Easily accessible individual breakers per load group for high availability environments
- Single phase 30A, 32A, 60A, and 63A offerings (dependent on the line cord selected)
- Three-phase 32A (removable line cord) and 60A fixed line cord offerings. 60A/208V/3 phase model includes fixed IEC60309 3P+G, 60A line cord

The intelligent PDU+ models have the following power management features:

- Monitored power draw at the breaker level
- Monitoring / measurement of power data
- Advanced remote monitoring capability
- Detailed data-logging for statistical analysis and diagnostics
- Includes an Environmental Monitoring Probe to provide both temperature and humidity values
- Integrates with IBM Active Energy Manager for consolidated rack power monitoring
- Comprehensive power management and flexible configuration through a Web browser, NMS, Telnet, SNMP, or HyperTerminal
- SNMP v1 support

- SNMP v2 & v3 support
- IPv6 (Ultra Density PDU+ only)
- SLP support

6.8.2 Availability and flexibility of Enterprise PDUs

Table 6-5 lists available Enterprise PDUs. Check availability of the Enterprise PDU and power input cables in your geographical region. Refer to the Rack Power configurator:

http://www.ibm.com/systems/xbc/cog/rackpwr/rackpwrconfig.html

Table 6-5 IBM Enterprise PDU part numbers

Part number[a]	Feature code	Description	Outlets		Input power
			C19	C13	
71762MX	6090	Ultra Density Enterprise PDU+	9	3	Varies by line cord, see Table 6-6
71763MU	6091	Ultra Density Enterprise PDU+	9	3	60A/208V/3 phased fixed cable
71762NX	6050	Ultra Density Enterprise PDU	9	3	Varies by line cord, see Table 6-6
71763NU	6051	Ultra Density Enterprise PDU	9	3	60A/208V/3 phase fixed cable
39M2816	6030	DPI® C13 Enterprise PDU+	0	12	Varies by line cord, see Table 6-6
39M2818	6080	DPI C19 Enterprise PDU+	6	3	Varies by line cord, see Table 6-6
39M2819	6081	DPI C19 Enterprise PDU+	6	3	3 phase fixed cable
39Y8923	6061	DPI C19 Enterprise PDU	6	0	60A/208V/3P+G fixed cable
39Y8941	6010	DPI C13 Enterprise PDU	0	12	Varies by line cord, see Table 6-6
39Y8948	6060	DPI C19 Enterprise PDU	6	0	Varies by line cord, see Table 6-6

[a] Last letter: U = available in U.S. and Canada; X = available worldwide
Second last letter: N = non-monitored (PDU);, M = monitored (PDU+)

Table 6-6 lists the available detachable line cords and the power source they support. Only those PDUs in Table 6-5 without fixed line cords support these.

Table 6-6 IBM Ultra Density Enterprise PDU power cords

Geography	Part number	FC	Description	Connector	Power source
EMEA, AP, LA	40K9611	N/A	IBM DPI 32A cord	IEC 309 3P+N+G	250V 32A
EMEA, AP, LA	40K9612	N/A	IBM DPI 32A cord	IEC 309 P+N+G	250 V 32A

Geography	Part number	FC	Description	Connector	Power source
EMEA, AP, LA	40K9613	N/A	IBM DPI 63A cord	IEC 309 P+N+G	250 V 63A
US, Canada	40K9614	6500	4.3m, 30A/208V,	NEMA L6-30P	208 V 30A
US, Canada	40K9615	6501	4.3m, 60A/208V,	IEC 309 2P+G	208 V 60A
Australia, NZ	40K9617	N/A	IBM DPI 32A cord	IEC 309 P+N+G	250 V 32A
Korea	40K9618	N/A	IBM DPI 30A cord	IEC 309 P+N+G	250 V 30A

6.8.3 Comparing PDU and intelligent PDU

This section compares the PDU with the intelligent PDU (iPDU, or PDU+).

PDU

PDUs are engineered to split a power source to different power outlets so that the power can be distributed to several systems. PDUs are either single-phase or three-phase, and which one you use depends on the power sources you have to attach it to.

Three-phase PDUs are more independent of power overload compared to single-phase PDUs because the power is supplied by the three phases. The outlets are combined to different load groups that are ideal for power redundancy.

Tip: We recommend you distribute the power cables to load groups that are joined to different phases, to ensure even greater power redundancy.

Intelligent PDU

Intelligent PDUs (iPDU, with the IBM designation “PDU+”) can be remotely managed and have an Ethernet or serial port for management. The PDU can be accessed from a browser interface or through Telnet. Section 6.8.5, “Intelligent PDU power management Web interface” on page 364 provides a description of the Web interface.

The iPDU collects power and temperature data at the power outlets and can send this information through SNMP. This data is collected once per second. You can also view this data in table form from the iPDU Web interface or by using tools such as IBM Active Energy Manager. They report power and thermal trending also for devices that is plugged into their individual load groups.

Collecting of this data is independent of the devices that are connected, so iPDUs are ideal for connecting older servers and systems that do not have their own power and thermal monitoring capabilities.

The IBM Enterprise C13+ and IBM C19 PDU+ DPI and ultra-density units are intelligent PDUs that are supported for use with Active Energy Manager. They report power and thermal trending.

In addition, Active Energy Manager allows a user to associate an IBM Director managed object with an iPDU outlet, allowing the power consumption for the server to be displayed in the Active Energy Manager Console. In cases where there are multiple power supplies for a server and each is plugged into a different iPDU in the rack, Active Energy Manager adds the different values together to display a graph of the total power being consumed by the server. Through this iPDU support, Active Energy Manager can monitor power usage on earlier servers that do not have power metering built in.

Additionally, Active Energy Manager monitors how many amps are being consumed by an iPDU overall, and how this compares to the maximum current that the iPDU can support.

Users are alerted when an iPDU approaches its capacity.

Note: IBM DPI C13+ and IBM DPI C19 PDU+ PDUs were available before the announcement of Active Energy Manager. Any existing versions of these iPDUs in the field must be upgraded to the November 7th, 2007 version of the iPDU firmware so that Active Energy Manager can support them.

6.8.4 Assembling of intelligent PDU

The IBM enterprise intelligent PDUs are available in three outlet versions for power distribution. The figures in this section show the devices.

IBM Enterprise DPI C13 PDU+

Figure 6-33 shows the back and front of the Enterprise DPI C13 PDU+ (part number 39M2816).

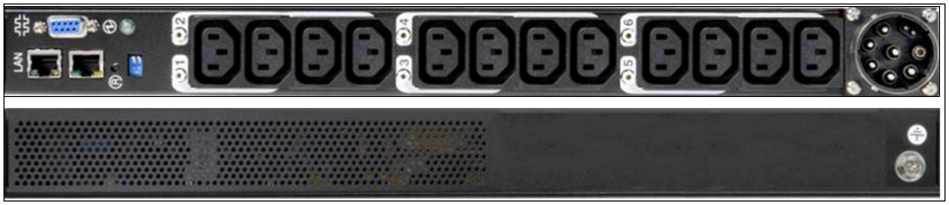


Figure 6-33 IBM Enterprise DPI C13 PDU+ UTG connector, part number: 39M2816

The back of the PDU+ shows (from left to right):

- Left: is the management interface with RJ45 LAN and RJ45 serial console port, RS232 serial port (DB-9) interface, reset button, input voltage status LED, and switches.
- Middle:12x IEC-320-C13 outlets. The outlets on the back are protected by six branch type circuit breakers 20A. Each pair of C13 outlets is linked to a load group, six load groups overall.
- Right: UTG0247 inlet connector

The front of this PDU+ has no outlets.

IBM Enterprise DPI C19 PDU+

Figure 6-34 shows the back and front of the C19 DPI Enterprise PDU+, part number 39M2818.

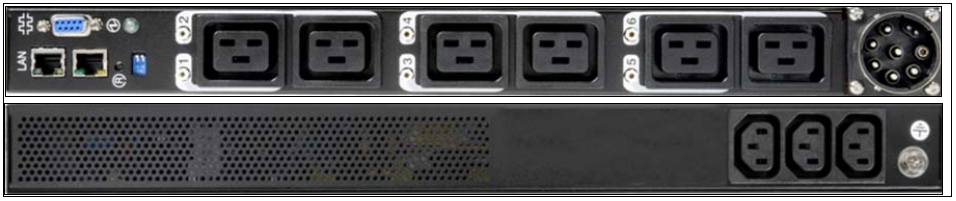


Figure 6-34 IBM Enterprise DPI C19 PDU+ UTG connector, part number: 39M2818

The back of the PDU+ shows (from left to right):

- Left: Management interface with RJ45 LAN, RJ45 serial console port, RS232 serial port (DB-9) interface, reset button, input voltage status LED, and operating DIP switch.
- Middle: 6x IEC-320-C19 outlets
- Right: UTG0247 inlet connector

The front of the PDU+ has 3x IEC-320-C13 outlets on the right.

The outlets are protected by six branch type circuit breakers 20A. Each of the C19 outlets on the front represents one load group, to six load groups overall. Every one of the C13 outlets at the back of the PDU+ is shared with load group 1, 3, or 5 at the C19 outlets.

IBM Ultra Density Enterprise PDU+

Figure 6-35 on page 363 shows the front of the C19 DPI Enterprise PDU+, part number 71762MX.

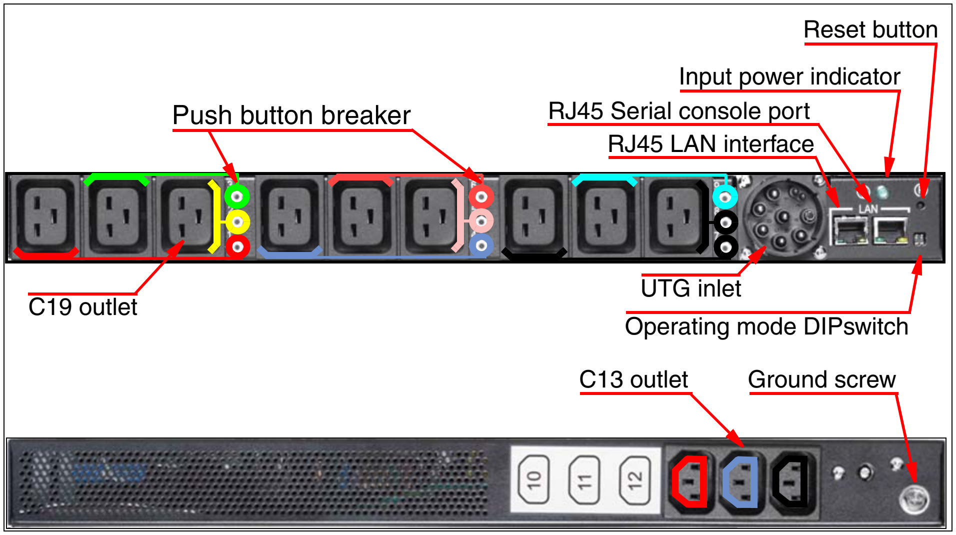


Figure 6-35 IBM Ultra Density Enterprise PDU+, part number: 71762MX

The front of the PDU+ shows (from left to right):

- Left: 9x IEC-320-C19 outlets
- Right: UTG0247 inlet connector
- Far right: Management interface with RJ45 LAN, RJ45 serial console port, reset button, input voltage status LED, and operating DIP switch.

The bottom view in Figure 6-35 shows the back of PDU+, which has 3x IEC-320-C13 outlets.

The outlets are protected by six branch type circuit breakers 20A. Each of the C19 outlets on the front side represents one load group, to nine load groups overall. Load group 1,4,7 each are shared with one C13 outlets at the rear side of the PDU+.

Description of the PDU+ components

The following components are on the iPDUs in the previous figures:

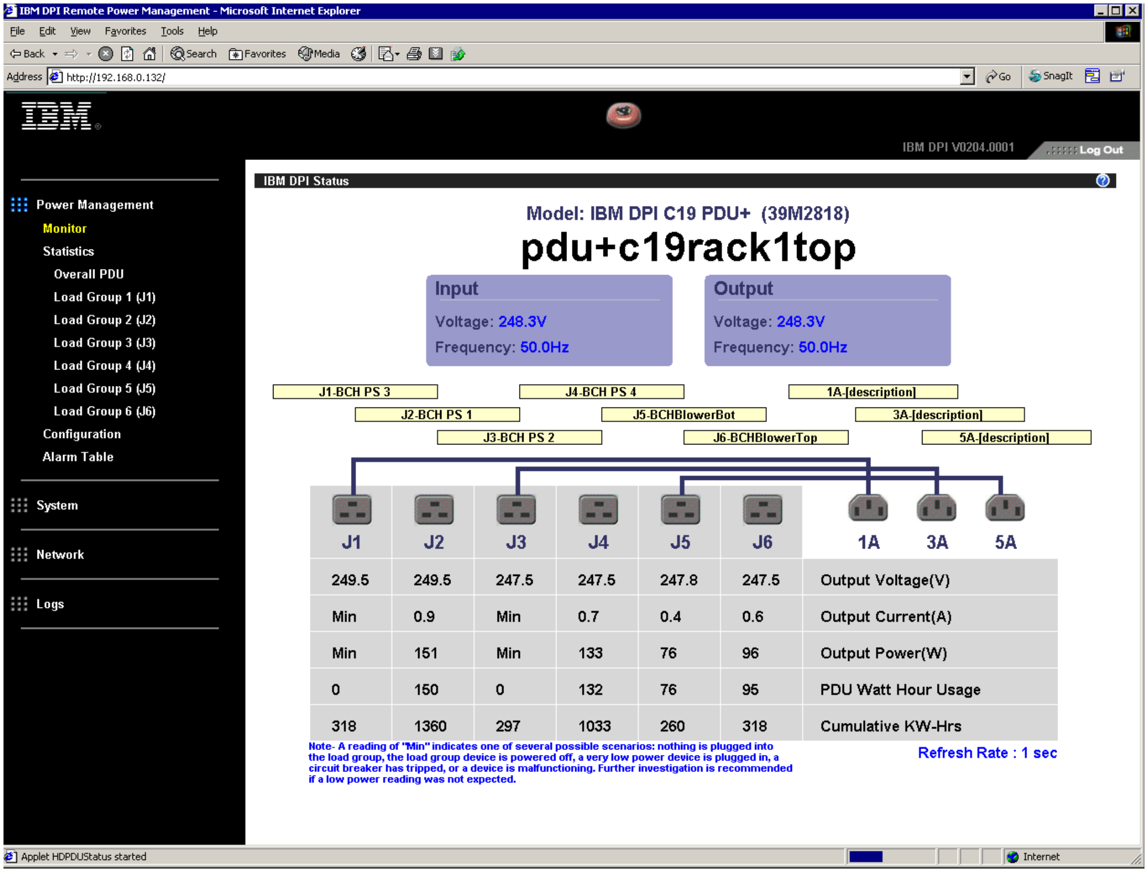
- Input power connector: Connect a power cord to this connector. Some PDU models have an attached power cord.
- Power outlets: You can connect a device to each power outlet. There are either nine or 12 power outlets, depending on the PDU model.
- RS-232 Serial connector: Use the RS-232 serial connector to update the iPDU firmware. Ultra Density PDU+ models doesn't provide this port.

- Green LED: This shows the iPDU input voltage status. When this LED is lit, the iPDU is receiving voltage. If the input voltage is too low, this LED is flashing.
- Operation model DIP switch: Sets the mode to operation for this iPDU. The default mode is S1 off, S2 off for normal operation. Other settings:
  - 1=Off, 2=Off: The card can run normal operational firmware
  - 1=On, 2=On: The card can start in diagnostics mode.
  - 1=On, 2=Off: Serial upgrade mode. You can upgrade the iPDU firmware from the serial connection if the network upgrade is not available.
  - 1=Off, 2=On: Read only mode. The device can run normal operational firmware, but all parameters of the device cannot be changed by a user.
- Reset button: Reset only the iPDU communication functions. This reset does not affect the loads.
- RJ-45 console connector: Connect the DB9 to RJ45 cable that comes with the iPDU to this connector and to the serial (COM) connector on a workstation or notebook computer and use the workstation or notebook as a configuration console. You can also connect an environment-monitored probe to this connector. The environment-monitored probe monitors humidity and temperature. The connection of an environment-monitored probe is automatically detected.
- RJ-45 Ethernet (LAN) connector: Use this connector to configure the iPDU through a LAN. The Ethernet connector supports 10/100 auto sense network connection.

6.8.5 Intelligent PDU power management Web interface

Each intelligent PDU can be accessed by a Web browser. The section describes the IBM Enterprise DPI C19 PDU+ (part number 39M2818) in particular.

The Web interface, in Figure 6-36 on page 365, shows the actual power data measurement at the load groups, which are refreshed once per second.



Model: IBM DPI C19 PDU+ (39M2818)

pdu+c19rack1top

Input
Voltage: 248.3V
Frequency: 50.0Hz

Output
Voltage: 248.3V
Frequency: 50.0Hz

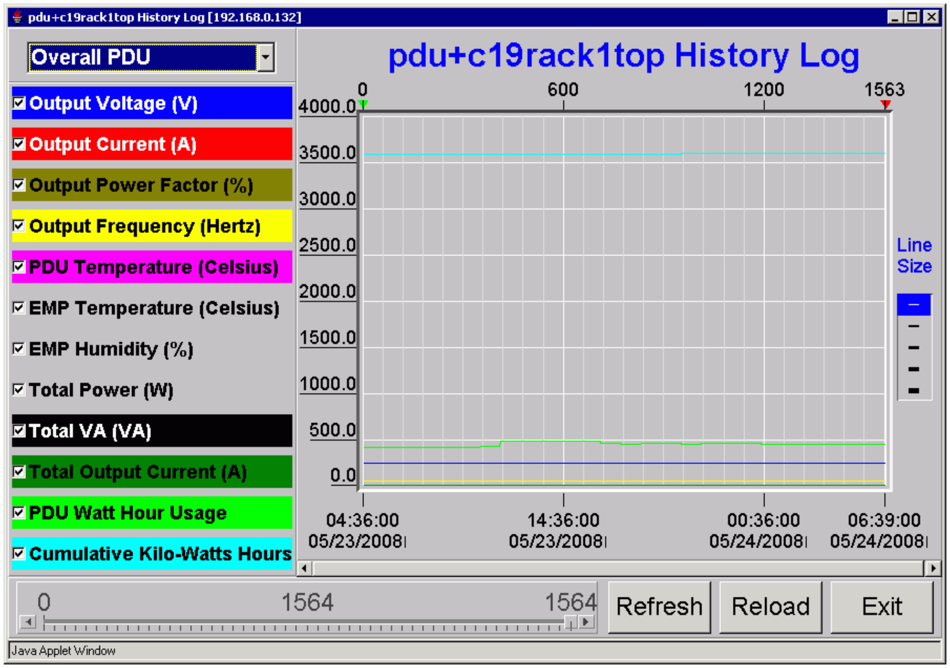
J1	J2	J3	J4	J5	J6	
249.5	249.5	247.5	247.5	247.8	247.5	Output Voltage(V)
Min	0.9	Min	0.7	0.4	0.6	Output Current(A)
Min	151	Min	133	76	96	Output Power(W)
0	150	0	132	76	95	PDU Watt Hour Usage
318	1360	297	1033	260	318	Cumulative KW-Hrs

Note- A reading of "Min" indicates one of several possible scenarios: nothing is plugged into the load group, the load group device is powered off, a very low power device is plugged in, a circuit breaker has tripped, or a device is malfunctioning. Further investigation is recommended if a low power reading was not expected.

Refresh Rate : 1 sec

Figure 6-36 Home page of the IBM Enterprise DPI C19 PDU+ Web interface

The iPDU stores the measurement of power data and reports it to a graph, shown in Figure 6-37 on page 366, and presents a chart of power consumption tendencies.


pdu+c19rack1top History Log [192.168.0.132]

Overall PDU

☑ Output Voltage (V)
☑ Output Current (A)
☑ Output Power Factor (%)
☑ Output Frequency (Hertz)
☑ PDU Temperature (Celsius)
☑ EMP Temperature (Celsius)
☑ EMP Humidity (%)
☑ Total Power (W)
☑ Total VA (VA)
☑ Total Output Current (A)
☑ PDU Watt Hour Usage
☑ Cumulative Kilo-Watts Hours

Figure 6-37 The iPDU power measurement graph

The measured results are reported by the built-in SNMP support to the IBM Active Energy Manager.

6.9 DSA Preboot

IBM developed Dynamic System Analysis (DSA) Preboot for the x3850 M2 and x3950 M2 because the previous tool (PC Doctor) did not meet the requirements for newer systems. DSA Preboot will become the standard embedded diagnostic tool for System x and BladeCenter servers.

DSA Preboot is an NVRAM-based version of the of the DSA tool, which is used by the Technical Support teams to collect the following information when determining the cause of errors:

- System and component level information,
- Operating system driver information,
- Hardware event logs of various components,
- Operating system event logs

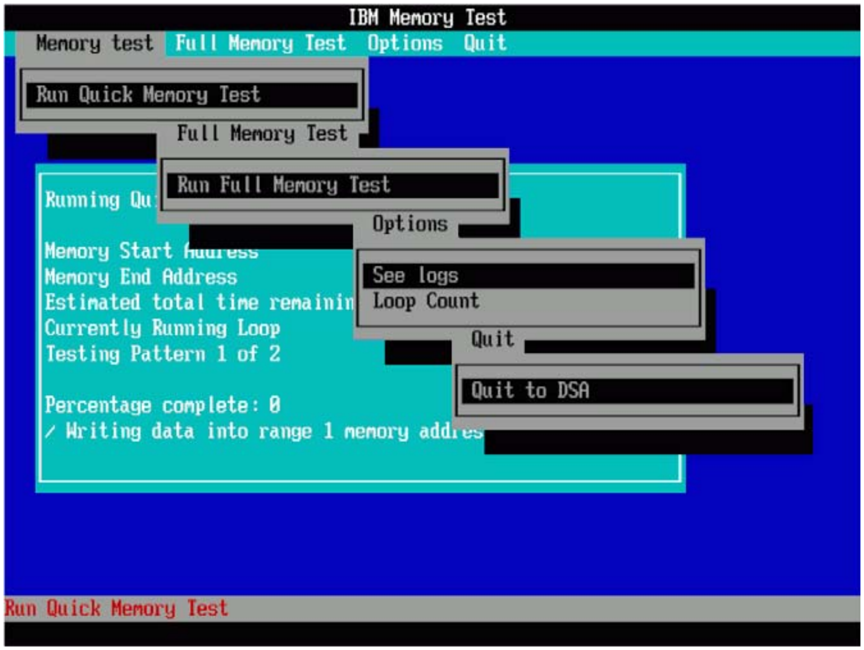
DSA and DSA Preboot collect the information that can be viewed locally or uploaded to an IBM internal FTP server for the Technical Support teams to have remote access from different locations around the world if further analysis of system state information or error logs is required.

DSA Preboot performs tests on all subsystems to check:

- System configuration
- Memory
- Processor
- Hard drives
- Network interfaces and settings
- Hardware inventory, including PCI and USB information
- Optical devices
- LSI 1064/1068/1078 SAS RAID controllers
- Remote Supervisor Adapter
- BMC
- Check-point panel test
- CPU and Memory Stress tests
- IBM light path diagnostics status
- Service Processor status and configuration
- Vital product data, firmware, and BIOS information
- Drive Health Information
- ServeRAID configuration
- LSI RAID & Controller configuration
- Event logs for Service Processors
- Merged Devices information
- Memory Diagnostics log
- DSA Preboot Error log

DSA Preboot works with single-node and multinode configurations.

To run DSA Preboot, press `F2` during POST when prompted. The system first starts the memory test menu as shown in Figure 6-38 on page 368.



```
Running Qu
Memory Start Address
Memory End Address
Estimated total time remainin
Currently Running Loop
Testing Pattern 1 of 2

Percentage complete: 0
/ Writing data into range 1 memory addres
```

Figure 6-38 DSA Preboot memory tests (showing all menus)

To leave the memory test window use the keyboard's right arrow key to move to the menu **Quit** and select **Quit to DSA**. The DSA Preboot is then started.

A list of commands is displayed is shown in Figure 6-39.

```
Starting IBM DSA Preboot v1.00
Extracting...

Commands:
  gui - Enter GUI Environment.
  cmd - Enter Command Line Environment.
  copy - Copy DSA results to removable media.
  exit - Quit the program.
         Note: This will reboot the system.
  help - Display this help message.

Please enter a command. (Type 'help' for commands)
>
```

Figure 6-39 DSA Preboot: Commands on the main menu

6.9.1 Updating DSA Preboot

This section describes how easily DSA Preboot can be updated. Download DSA Preboot from:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5072294

The DSA can be updated by:

- Windows operating system
- Linux operation system (not VMware ESX 3i embedded)
- Bootable CD or mounted ISO image in the RSA II Web interface
- USB memory key, if it is attached locally at the server or mounted in the remote control window

> **Note:** The DSA Preboot updates stand-alone systems and all nodes that are partitioned to a multinode system. See Chapter 4, “Multinode hardware configurations” on page 195.

Update by using the ISO image

After you download the ISO file, you may burn it to a CD-ROM or mount it at the remote control feature in the RSA II Web interface, described in 6.2.4, “Remote console and media” on page 324.

To update the DSA Preboot:

1. Mount the ISO image or created CD-ROM in the RSA II Web interface remote control window, or insert the CD-ROM in the CD-ROM drive of your server.
2. Boot the server to the CD-ROM drive.

   The initial sequence detects the internal 4 GB USB flash device, and addresses it by SCSI emulation for USB Mass Storage support, shown in Example 6-2.

   Example 6-2 DSA Preboot boot initialization sequence

   ```
   ...
   USB Mass Storage support registered.
     Vendor: ST        Model: ST72682         Rev. 2.10
     Type:   Direct Access                    ANSI SCSI revision: 02
   ...
   ```

3. The DSA Preboot flash update menu is displayed, shown in Example 6-3 on page 370.

Example 6-3 DSA Preboot: Commands on the update menu

```
Starting DSA Preboot v1.00

Commands:
  update - Update/Recover your embedded usb chip
  help - Display this help message
  exit - Quit program.
         Note: This will reboot the syste,-

Please enter a command. (Type ’help’ for commands)
> update
```

4. Use the command **update**.

   The flashing of the device starts after a re-init of the USB device, shown in Example 6-4.

Example 6-4 DSA Preboot: flash progress

```
> update
Node count: 1
Flashing node: 0
Waiting for device to become available:
.usb 5-6: new high speed USB device using ehci_hcd and address 3
usb 5-6: new device found. idVendor=0483, idProduct=fada
...
usb 5-6: Product: ST72682  High Speed Mode
usb 5-6: Manufacturer STMicroelectronics
scsi1 : SCSI emulation for USB Mass Storage devices
.....  Vendor: ST        Model: ST72682         Rev: 2.10
  Type:   Direct-Access                    ANSI-SCSI revision:02
SCSI device sda_: 1024000 512-byte hdwr sectors (524 MB)
sda: Write Protect is off
sda: assuming drive cache: write through
 sda
sd 1:0:0:0: Attached scsi removable disk sda
sd 1:0:0:0: Attached scsi generic sg1 type 0

Device flashing in progress
....................................................................
....................................................................
....................................................................
....................................................................
DSA key has been flashed successfully

Please enter a command. (Type ’help’ for commands)
usb 5-6: USB disconnect, address 3
> _
```

5. Verify you receive the following message:

   `DSA key has been flashed successfully`

6. Type **exit** and quit the program. The server reboots.
7. Unmount the image or remove your CD-ROM from the local drive at your server.

Update in the Linux operating system

You may update the DSA Preboot in 32-bit and 64-bit Linux operating systems.

Read the instructions for and then download the `*.sh` for Linux operating system.

To update the DSA Preboot within the Linux operating system:

1. Save the package in a user directory: /usr/
2. Unpack the package, by using the command in Example 6-5.

Example 6-5 DSA Preboot: extraction in Linux

```
linux-x3850m2:/usr/ # ./ibm_fw_dsa_a3yt71a_linux_32_64.sh -x /usr/dsa
```

3. Change to this directory.
4. Perform the update as shown in Example 6-6.

Example 6-6 DSA Preboot: flash progress of the Update process in Linux

```
srv-253-217:/usr/dsa # ./lflash64 -z
Node count: 1
Flashing node: 0
Waiting for device to become available:
.......
Device flashing in progress:
...........................................................................
...........................................................................
...........................................................................
..............................................................
DSA key has been flashed successfully
srv-253-217:/tmp/dsa #
```

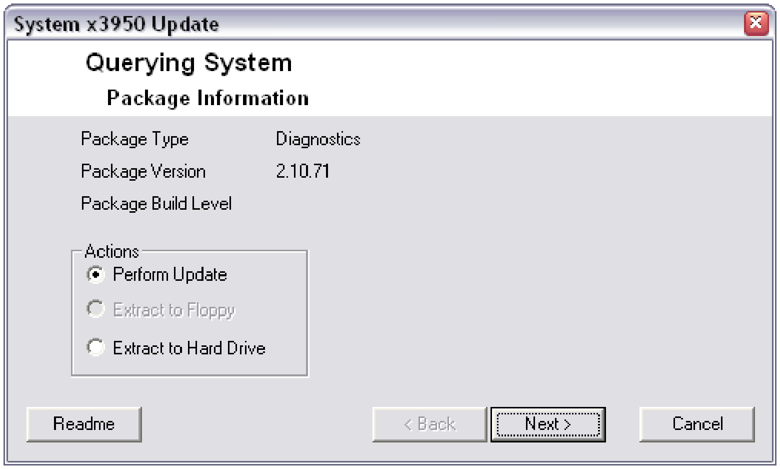
Update in the Windows operating system

To update the DSA Preboot in 32-bit and x64 Windows:

1. Download the following (or similarly named) EXE file to a local disk on your server:

   `ibm_fw_dsa_a3yt71a_windows_32_64.exe`

2. Double-click the file name. The Update window opens, as shown in Figure 6-40.



System x3950 Update

Querying System

Package Information

Package Type Diagnostics

Package Version 2.10.71

Package Build Level

Actions

● Perform Update

○ Extract to Floppy

○ Extract to Hard Drive

Figure 6-40 DSA Preboot: the Update process in Windows

3. Select the **Perform Update** action, and then click **Next**.

The server diagnostics are updated.

6.9.2 Working with the command line interface

After the memory text finishes, described in 6.9.1, “Updating DSA Preboot” on page 369, the main menu is started as shown in Figure 6-39 on page 368. Select any of the commands listed in Table 6-7.

Table 6-7 DSA Preboot commands

Command	Description
`gui`	This starts the graphical user interface.
`cmd`	The command line environment interface you can use to perform the set of diagnostics, collect the system information, and show the date in the local text viewer. The data and results can be sent to the IBM FTP server to diagnose the information. See “Use of the interactive menu” on page 373.
`copy`	Use this command to save the captured archive to any USB device.

Use of the interactive menu

You can access the interactive menu by entering the `cmd` command at the prompt. The interactive menu has the following options:

Table 6-8 DSA Preboot `cmd` commands

Command	Description
`collect`	Collects system information. The collected information can be captured, without creating XML output, by adding option **`-x`**. The data collector creates additional HTML output by using **`-v`**.
`view`	Displays the collected data on the local console in text viewer. To exit viewer, type **`:x`** then press `Enter`.
`enumtests`	Lists available tests ► 1: BMC I2C Bus ► 2-17: Extended onboard network port tests ► 18: CPU stress test ► 19: Memory stress test ► 20-22: Optical media tests ► 23: Checkpoint panel test ► 24: RSA restart test
`exectest`	Presents a menu in which you can select a test to execute. Use this command to execute the specific test.
`getextendedresults`	Retrieves and displays additional diagnostic results for the last diagnostic test that was executed.

`transfer`	Transfers collected data and results to IBM. To transfer data and results to IBM, a network connection is required. You see:
	`*******************`
	`Attempted data upload to IBM by means of an unencrypted channel will proceed in 10 seconds. Press any key to cancel the attempted upload.`
	`*******************`
	`..........`
	`Configuring Network Controllers.`
	`Transferring collected data to IBM Service.`

`quit`	Exits the DSA Preboot menu, reboots the system.

Save the collected information

Data that you can save for reviewing or sending to IBM must be previously collected in the interactive menu (by using the commands **cmd** and then **collect**).

To save the collected information:

1. Attach a USB key at the front of the server. If you are working remotely from the server, attach the USB key locally at your workstation and mount it as an accessible remote device at the Remote Supervisor remote media window.
2. Use the command **copy** to save the data to a local or remote USB storage device. See Example 6-7.

Example 6-7 DSA Preboot command line: mounted USB storage devices

```
>copy
1: SanDisk Cruzer Micr
2: USB 2.0 Flash Disk
Enter Number (type x to exit)
```

> **Note:** The example shows two mounted USB devices. Number 1 in this example is an attached Hypervisor key. You cannot use this device to save any captured data.

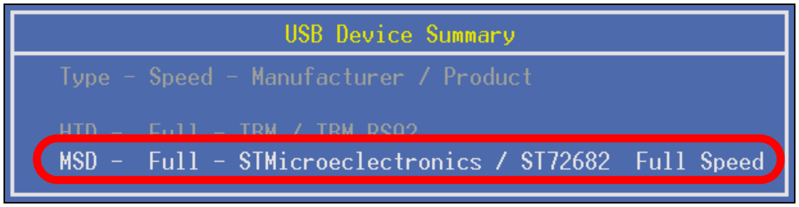
3. Unmount the USB storage device. The data is now available for reviewing by you and by IBM Service.

Save the collected information in the operating system

The internal 4 GB flash device is hidden in the operating system by default.

To retrieve the DSA logs that the DSA Preboot collected, enable this device temporarily and make it show up as a drive in your operating system. Do this by pressing `Ctrl+End` on your keyboard during POST when the IBM logo appears. Although no message indicates that it is enabled, the device shows up in BIOS and the operating system. See Figure 6-41, Figure 6-42 on page 375, and Figure 6-43 on page 375.

If you boot the system, you can view the added USB devices in the BIOS. See Figure 6-41.



```
USB Device Summary

Type - Speed - Manufacturer / Product

HID - Full - IBM / IBM RSA2
MSD - Full - STMicroeclectronics / ST72682 Full Speed
```

Figure 6-41 Unhidden embedded USB device in BIOS

Windows adds it as a storage device in the Device Manager (Figure 6-42 on page 375).

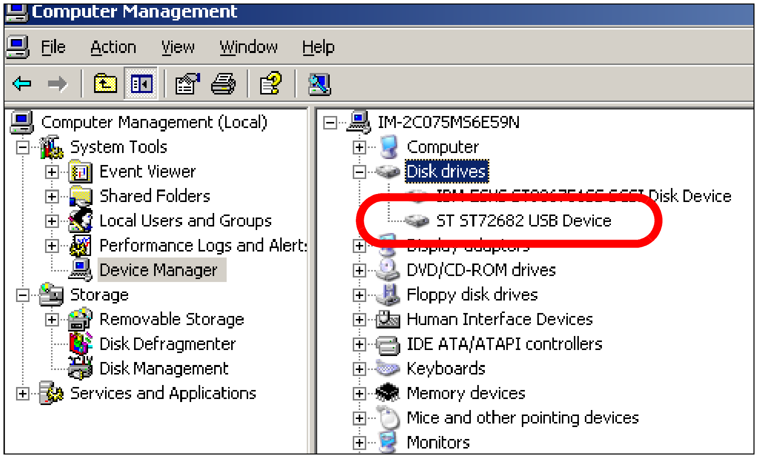
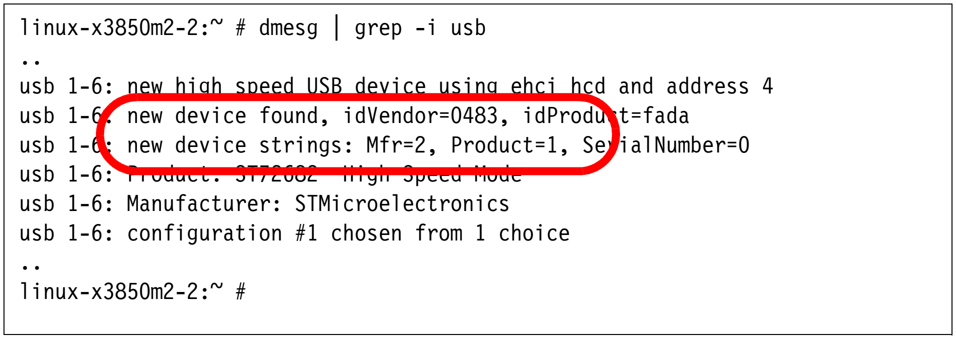


Figure 6-42 Embedded USB device unhidden in Windows OS

In the Linux operating system, shown in Figure 6-43, you see the USB device listed, which can be mounted as storage device.



```
linux-x3850m2-2:~ # dmesg | grep -i usb
..
usb 1-6: new high speed USB device using ehci_hcd and address 4
usb 1-6: new device found, idVendor=0483, idProduct=fada
usb 1-6: new device strings: Mfr=2, Product=1, SerialNumber=0
usb 1-6: Product: ST72682  High Speed Mode
usb 1-6: Manufacturer: STMicroelectronics
usb 1-6: configuration #1 chosen from 1 choice
..
linux-x3850m2-2:~ #
```

Figure 6-43 DSA Preboot flash device in Linux

Check that a partition is available so you can then use the following command:

`mount /dev/sdb1 /mnt/dsa`

The results are shown in Example 6-8.

Example 6-8 Flash storage device in Linux

```
linux-x3850m2-2:~ # fdisk -l /dev/sdb1

Disk /dev/sdb1: 1036 MB, 1036353024 bytes
```

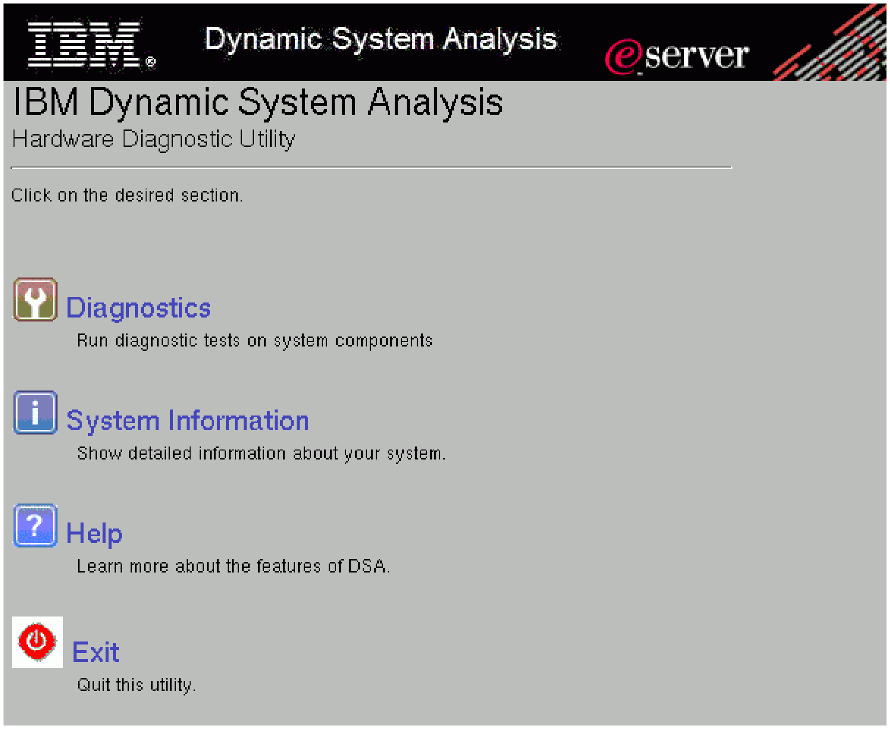
```
255 heads, 63 sectors/track, 125 cylinders
Units = cylinders of 16065 * 512 = 8225280 bytes

linux-x3850m2-2:~ #
```

Copy the files with the extension `*.xml.gz` to a local disk. Provide the log files to the IBM technical support team if requested. After you reboot or power cycle your server this device is hidden again.

6.9.3 Working with the graphical user interface (GUI)

The new embedded DSA Preboot in your x3850 M2 and x3950 M2 enables a graphical user interface (GUI) that you control by keyboard and mouse.

You start the Dynamic System Analysis GUI in the main menu by using the command **`gui`**. The GUI window opens, shown in Figure 6-44.



IBM Dynamic System Analysis
Hardware Diagnostic Utility

Click on the desired section.

Diagnostics
Run diagnostic tests on system components

System Information
Show detailed information about your system.

Help
Learn more about the features of DSA.

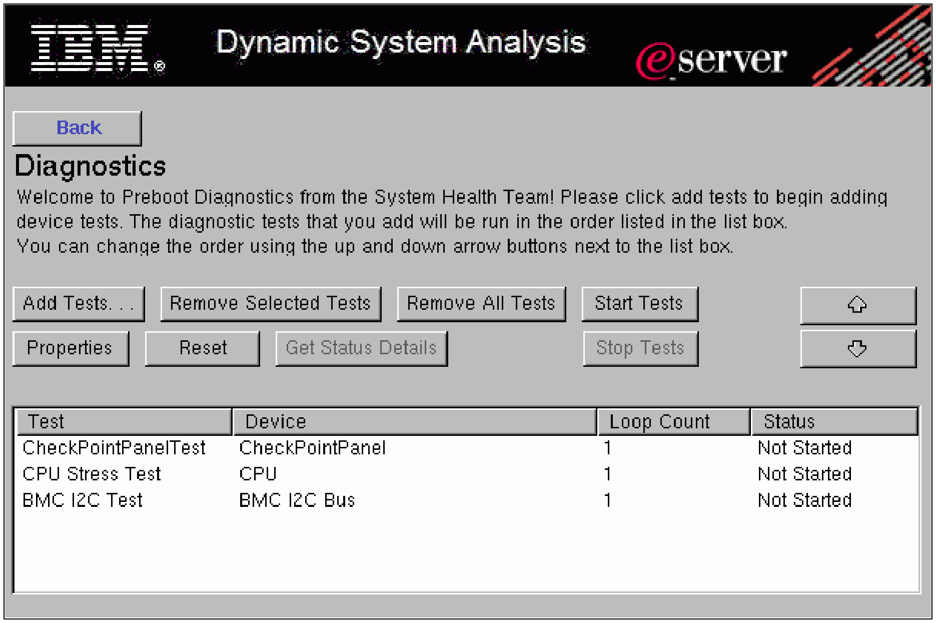
Exit
Quit this utility.

Figure 6-44 DSA Preboot home window

The following sections provide an overview of the Dynamic System Analysis options in the GUI window.

Diagnostics

Click **Diagnostics** to open the Diagnostics window, shown in Figure 6-45. Use this to perform various tests, such as CPU or Memory stress test. The complete list of implemented tests at the time of writing this book are in “Use of the interactive menu” on page 373.



Diagnostics

Welcome to Preboot Diagnostics from the System Health Team! Please click add tests to begin adding device tests. The diagnostic tests that you add will be run in the order listed in the list box. You can change the order using the up and down arrow buttons next to the list box.

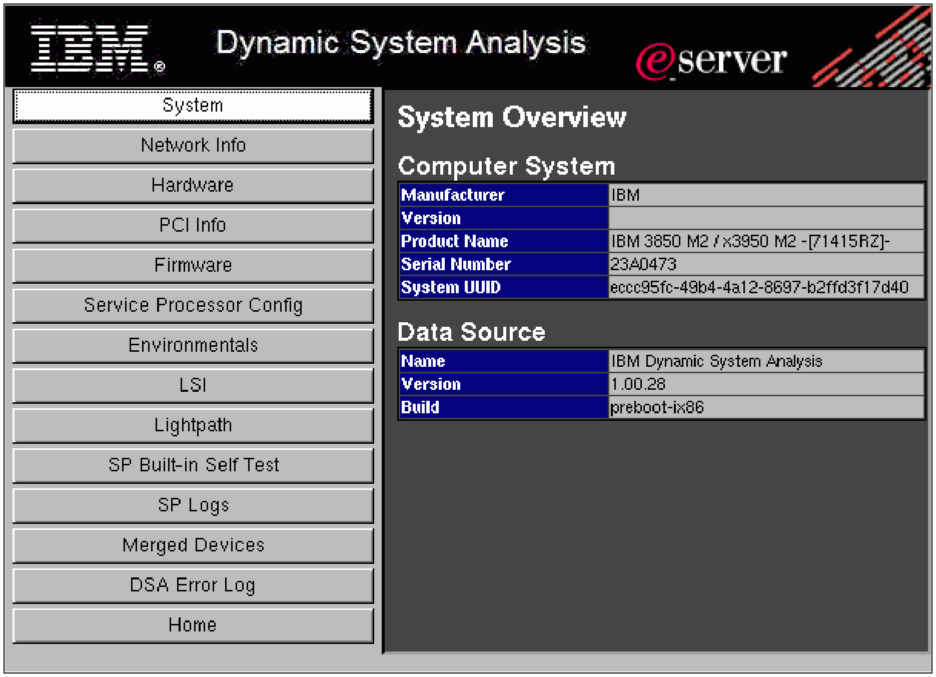
Test	Device	Loop Count	Status
CheckPointPanelTest	CheckPointPanel	1	Not Started
CPU Stress Test	CPU	1	Not Started
BMC I2C Test	BMC I2C Bus	1	Not Started

Figure 6-45 DSA Preboot: Diagnostics window in the GUI

After adding tests you select, start the tests by clicking the button **Start Tests**. The Status column shows the result of this test. For details about the results, click **Get Status Details**.

System Information

Use the system information window to get an overview about your system or your multinode partition. See Figure 6-46 on page 378.



System Overview

Computer System

Manufacturer	IBM
Version	
Product Name	IBM 3850 M2 / x3950 M2 -[71415RZ]-
Serial Number	23A0473
System UUID	eccc95fc-49b4-4a12-8697-b2ffd3f17d40

Data Source

Name	IBM Dynamic System Analysis
Version	1.00.28
Build	preboot-ix86

Figure 6-46 DSA Preboot: System Overview information window in the GUI

The selections to get information include:

System Information about your system: type, model, serial number, and UUID.

Network info NIC port information, such as speed, negotiation, or MAC addresses

Hardware System component information, such as CPU, memory, scalability ports.

PCI Info PCI slot device description with information about the installed PCIe adapters.

Firmware Details about the installed firmware of your system.

Environmentals Snapshot of fan speed and environmental monitor

LSI SAS subsystem information, such as type and firmware of the controller, physical and virtual drive information.

Lightpath Snapshot information about error LED states and information about the Light Path Diagnostic (LPD) panel on the x3850 M2 and x3950 M2, error LED status at the failing component, such as a single DIMM, processor or VRM.

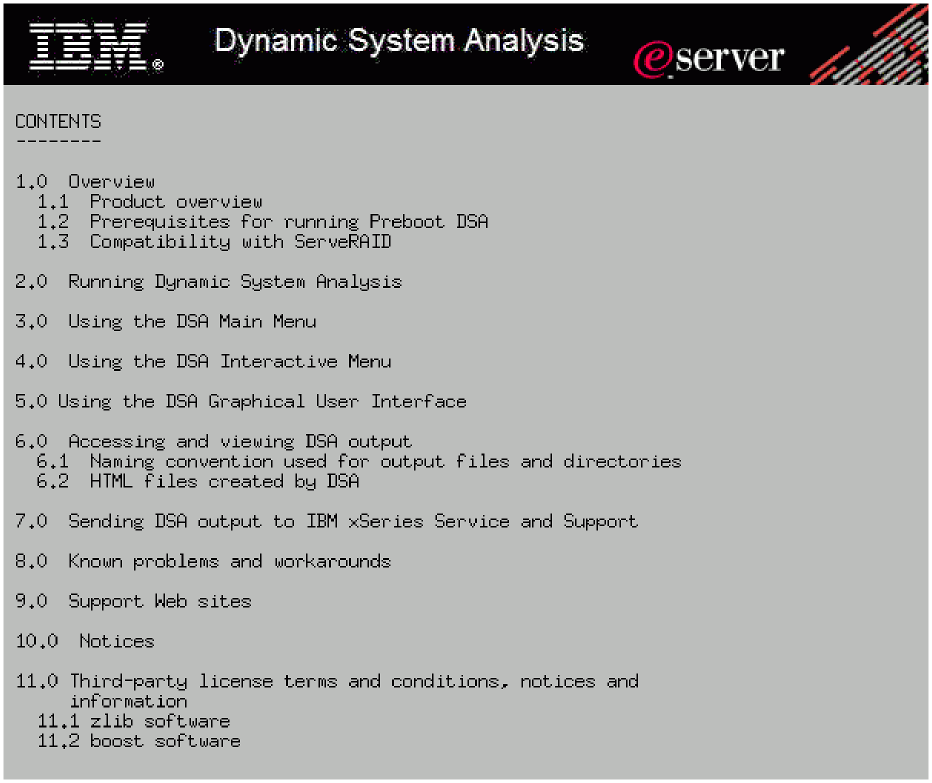
SP Built-In Self TestRSA and BMC self test

SP Logs Service Processor (SP) logs are split in a table for RSA and BMC (IPMI) event logs.

> **Note:** If your system indicates an error and you are able to boot to the DSA Preboot diagnostics, we recommend you review each of these entries.

Help system

The embedded Help (Figure 6-47) describes prerequisites, handling instructions, known problems, and provides tips for using the DSA Preboot.



```
CONTENTS
--------

1.0  Overview
  1.1  Product overview
  1.2  Prerequisites for running Preboot DSA
  1.3  Compatibility with ServeRAID

2.0  Running Dynamic System Analysis

3.0  Using the DSA Main Menu

4.0  Using the DSA Interactive Menu

5.0 Using the DSA Graphical User Interface

6.0  Accessing and viewing DSA output
  6.1  Naming convention used for output files and directories
  6.2  HTML files created by DSA

7.0  Sending DSA output to IBM xSeries Service and Support

8.0  Known problems and workarounds

9.0  Support Web sites

10.0  Notices

11.0 Third-party license terms and conditions, notices and
     information
  11.1 zlib software
  11.2 boost software
```

Figure 6-47 DSA Preboot: list of help topics in the GUI

6.9.4 Scalability partition management

You manage scalable partitions by using the RSA II Web interface. This is discussed in Chapter 4, “Multinode hardware configurations” on page 195.

Abbreviations and acronyms

ABR	Automatic BIOS recovery
AC	alternating current
ACPI	Advanced Configuration and Power Interface
AEM	Active Energy Manager
AMD	Advanced Micro Devices™
ANSI	American National Standards Institute
API	application programming interface
AS	Australian Standards
ASCII	American Standard Code for Information Interchange
ASIC	application-specific integrated circuit
ASM	Advanced System Management
ASPM	Active State Power Management
ASR	automatic server restart
AWE	Address Windowing Extensions
BBU	battery backup unit
BCD	Boot Configuration Database
BGI	background initialization
BI	Business Intelligence
BIOS	basic input output system
BMC	Baseboard Management Controller
CC	consistency check
CD	compact disc
CD-ROM	compact disc read only memory
CIM	Common Information Model
CIMOM	Common Information Model Object Manager
CLI	command-line interface
COA	Certificate of Authenticity
COG	configuration and option guide
COM	Component Object Model
CPU	central processing unit
CRM	Customer Relationship Management
CRTM	Core Root of Trusted Measurements
CTO	configure-to-order
CTP	composite theoretical performance
DASD	direct access storage device
DBA	database administrator
DBS	demand-based switching
DC	domain controller
DCU	data cache unit
DCUI	Direct Console User Interface
DDR	Double Data Rate
DEP	Data Execution Prevention
DG	disk group
DHCP	Dynamic Host Configuration Protocol
DIMM	dual inline memory module
DMA	direct memory access
DNS	Domain Name System
DOS	disk operating system
DPC	deferred procedure call
DPM	Distributed Power Management
DRAM	dynamic random access memory

DRS	Distributed Resource Scheduler
DSA	Dynamic System Analysis
ECC	error checking and correcting
ECRC	end-to-end cyclic redundancy check
EDB	Execute Disable Bit
EIDE	enhanced IDE
EIST	Enhanced Intel SpeedStep
EMEA	Europe, Middle East, Africa
ERP	enterprise resource planning
ESI	Enterprise Southbridge Interface
ESM	Ethernet switch modules
ETW	Event Tracing for Windows
FAMM	Full Array Memory Mirroring
FC	Fibre Channel
FPGA	Field Programmable Gate Array
FQDN	fully qualified domain name
FRU	field replaceable unit
FSB	front-side bus
FTP	File Transfer Protocol
GB	gigabyte
GPR	general purpose register
GUI	graphical user interface
HA	high availability
HAL	hardware abstraction layer
HAM	hot-add memory
HBA	host bus adapter
HCL	Hardware Compatibility List
HDD	hard disk drive
HID	human interface device
HPC	high performance computing
HPMA	High Performance Memory Array
HSP	hotspare
HTML	Hypertext Markup Language
HW	hardware
I/O	input/output
IBM	International Business Machines
IBS	International Business System
ID	identifier
IDE	integrated drive electronics
IEEE	Institute of Electrical and Electronics Engineers
IERR	internal error
IIS	Internet Information Server
IM	instant messaging
IP	Internet Protocol
IPMB	Intelligent Platform Management Bus
IPMI	Intelligent Platform Management Interface
IR	Integrated RAID
IS	information store
ISMP	Integrated System Management Processor
ISO	International Organization for Standards
IT	information technology
ITSO	International Technical Support Organization
JBOD	just a bunch of disks
KB	kilobyte
KVM	keyboard video mouse
LAN	local area network
LED	light emitting diode
LLHEL	Low-Level Hardware Error Handlers
LPD	light path diagnostic
LUN	logical unit number
MAC	media access control

MB	megabyte
MBR	Master Boot Record
MCFG	Memory-Mapped PCI Configuration Space
MCK	machine check
MIB	management information base
MIOC	Memory and I/O Controller
MMIO	memory mapped I/O
MP	multiprocessor
MR	MegaRAID
MSCS	Microsoft Cluster Server
MSI	Message Signaled Interrupt
MSM	MegaRAID Storage Manager
NIC	network interface card
NMI	non-maskable interrupt
NMS	Network Management System
NOS	network operating system
NTP	Network Time Protocol
NUMA	Non-Uniform Memory Access
NVRAM	non-volatile random access memory
OCA	Online Crash Analysis
OEM	other equipment manufacturer
OLAP	online analytical processing
OLTP	online transaction processing
OS	operating system
OSPM	operating system power management
PAE	Physical Address Extension
PC	personal computer
PCI	Peripheral Component Interconnect
PCIe	PCI Express
PD	problem determination
PDF	Predictive Drive Failure
PDU	power distribution unit
PEF	platform event filtering
PFA	Predictive Failure Analysis
PID	partition ID
PKT	packet
PME	power management event
POST	power-on self test
PPM	processor power management
PSHED	Platform Specific Hardware Error Driver
PXE	Preboot eXecution Environment
RAC	Real Application Clusters
RAID	redundant array of independent disks
RAM	random access memory
RAS	remote access services; row address strobe
RBS	redundant bit steering
RDP	Remote Desktop Protocol
RETAIN	Remote Electronic Technical Assistance Information Network
RHEL	Red Hat Enterprise Linux
ROC	RAID-on-card
ROM	read-only memory
RPM	Red Hat Package Manager
RSA	Remote Supervisor Adapter
RTC	real-time clock
SAN	storage area network
SAS	Serial Attached SCSI
SATA	Serial ATA
SCM	Supply Chain Management
SCSI	Small Computer System Interface
SDR	Single Data Rate
SDRAM	static dynamic RAM

SEL	System Event Log
SLED	SUSE Linux Enterprise Desktop
SLES	SUSE Linux Enterprise Server
SLIT	System Locality Information Table
SLP	Service Location Protocol
SMP	symmetric multiprocessing
SMTP	simple mail transfer protocol
SN	serial number
SNMP	Simple Network Management Protocol
SOL	Serial over LAN
SP	service processor
SPEC	Standard Performance Evaluation Corporation
SPI	SCSI-3 parallel interface
SPINT	service processor interrupt
SQL	Structured Query Language
SRAT	Static Resource Allocation Table
SSH	Secure Shell
SSL	Secure Sockets Layer
TB	terabyte
TCG	Trusted Computing Group
TCP/IP	Transmission Control Protocol/Internet Protocol
TOE	TCP offload engine
TPM	Trusted Platform Module
TPMD	thermal and power management device
TSS	Trusted Computing Group Software Stack
URL	Uniform Resource Locator
USB	universal serial bus
UUID	Universally Unique Identifier
UVHA	Unlimited Virtualization with High Availability
VCB	VMware Consolidated Backup
VI	VMware Infrastructure
VLAN	virtual LAN
VM	virtual machine
VMFS	virtual machine file system
VMM	virtual machine manager
VMX	virtual machine extensions
VPD	vital product data
VRM	voltage regulator module
VSC	Virtual Service Clients
VSP	Virtual Service Providers
VT	Virtualization Technology
WAN	wide area network
WHEA	Windows Hardware Error Architecture
WMI	Windows Management Instrumentation
WSRM	Windows System Resource Manager
WWN	World Wide Name
XAPIC	multi-processor interrupt communication protocol
XD	execute disable
XML	Extensible Markup Language

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

For information about ordering these publications, see “How to get Redbooks” on page 394. Note that some of the documents referenced here might be available in softcopy only.

- *Tuning IBM System x Servers for Performance*, SG24-5287
- *IBM eServer xSeries and BladeCenter Server Management*, SG24-6495
- *Virtualization on the IBM System x3950 Server*, SG24-7190
- *IBM BladeCenter Products and Technology*, SG24-7523
- *Building an Efficient Data Center with IBM iDataPlex*, REDP-4418
- *Enabling Serial Over LAN for a Remote Windows Text Console using OSA SMBridge*, TIPS0551

Product publications

These product publications are also relevant as further information sources:

- *System x3850 M2 and x3950 M2 Installation Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073028
- *System x3850 M2 and x3950 M2 User’s Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073029
- *System x3850 M2 and x3950 M2 Problem Determination and Service Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073027
- *System x3850 M2 and x3950 M2 Rack Installation Instructions*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073030
- *IBM ScaleXpander Option Kit Installation Instructions*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5075330

- IBM ServeRAID-MR10k documentation (installation and user guides)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5074104
- IBM ServeRAID-MR10M documentation (installation and user guides)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5074105
- IBM Information Center: IBM Systems Director Active Energy Manager V3.1.1

  http://publib.boulder.ibm.com/infocenter/eserver/v1r2/index.jsp?topic=/aem_310/frb0_main.html
- IBM Information Center: IBM Director

  http://publib.boulder.ibm.com/infocenter/eserver/v1r2/index.jsp?topic=/diricinfo_all/diricinfoparent.html

Online resources

These Web sites are also relevant as further information sources:

IBM production information

See the following Web addresses for more information:

- IBM Systems home page

  http://www.ibm.com/systems
- IBM System x3850 M2 product page

  http://www.ibm.com/systems/x/hardware/enterprise/x3850m2
- IBM System x3950 M2 product page

  http://www.ibm.com/systems/x/hardware/enterprise/x3950m2
- IBM Announcement letter search

  http://www.ibm.com/common/ssi/
- VMware ESX Server 3i on IBM hardware product announcement

  http://www.ibm.com/isource/cgi-bin/goto?it=usa_annred&on=208-071
- System x3850 M2 product announcement

  http://www.ibm.com/isource/cgi-bin/goto?it=usa_annred&on=107-606
- System x3950 M2 2-node product announcement

  http://www.ibm.com/isource/cgi-bin/goto?it=usa_annred&on=108-081
- x3950 M2 3-node and 4-node product announcement

  http://www.ibm.com/isource/cgi-bin/goto?it=usa_annred&on=108-345

- xREF Reference Sheets

  http://www.redbooks.ibm.com/xref
- IBM System x Configuration and Options Guide

  http://www.ibm.com/systems/xbc/cog
- IBM Director download page

  http://www.ibm.com/systems/management/director/downloads.html
- Active Energy Manager extensions page

  http://www.ibm.com/systems/management/director/extensions/actengmrg.html
- Rack power configurator

  http://www.ibm.com/systems/xbc/cog/rackpwr/rackpwrconfig.html
- Anatomy of the Linux slab allocator

  http://www.ibm.com/developerworks/linux/library/l-linux-slab-allocator/
- Inside the Linux scheduler

  http://www.ibm.com/developerworks/linux/library/l-scheduler/

IBM support documents

See the following Web addresses for more information:

- IBM System x support home page

  http://www.ibm.com/systems/support/x
- ServerProven compatibility home page

  http://www.ibm.com/servers/eserver/serverproven/compat/us
- ServerProven operating system compatibility home page

  http://www.ibm.com/servers/eserver/serverproven/compat/us/nos/matrix.shtml
- Linux Open Source watchdog daemon support replaces IPMI ASR

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5069505
- Download latest Linux Open Source ipmitool utility

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5069538
- IBM now supporting Linux open source IPMI driver and utility

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5069569

- Broadcom NetXtreme II device driver for Microsoft Windows Server 2008 and Windows Server 2003

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5070012

- Intel-based Gigabit Ethernet drivers for Microsoft Windows 2003 and 2008

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5070807

- IBM Remote Supervisor Adapter II Daemon for IA32 Windows

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5071025

- IBM Remote Supervisor Adapter II Daemon for Microsoft Windows Server 2003/2008 x64

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5071027

- IBM Remote Supervisor Adapter II Daemon for Linux

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5071676

- Dynamic System Analysis (DSA)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5072294

- Installing SUSE Linux Enterprise Server 10 SP 1

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073088

- Flash BIOS update v1.06 (DOS package)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073120

- Remote Supervisor Adapter II update (Linux package)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073123

- Remote Supervisor Adapter II update (DOS package)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073124

- Remote Supervisor Adapter II update (Microsoft Windows package)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073125

- Baseboard Management Controller (BMC) flash update (ISO)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073127

- Baseboard Management Controller (BMC) flash update (Windows)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073128

- Baseboard Management Controller (BMC) flash update (Linux)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073129

- LSI 1078 SAS controller BIOS and firmware update (Windows)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073134

- IBM and LSI Basic or Integrated RAID SAS driver (Windows)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073138

- ServeRAID MR10k SAS controller firmware update (Windows)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073139

- ServeRAID MR10M SAS controller firmware update (Windows)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073389

- IBM HBA EXP3000 ESM 1.88 update program

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073877

- Installing Red Hat Enterprise Linux Version 5 Update 1

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5074155

- Installing Microsoft Windows 2008 32-bit

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5074895

- Installing Microsoft Windows 2008 x64 Edition

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5074896

- Installing Microsoft Windows Server 2003 x64 Edition

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5075237

- Installing Microsoft Windows Server 2003 Edition (32-bit)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5075238

- IBM Active PCI Software for Windows Server 2008

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5074966

- IBM Active PCI Software for Microsoft Windows 2000 and 2003

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-62127

- RETAIN tip H193590: MegaRAID Storage Manager cannot recognize greater than 16 controllers

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5076491

- RETAIN tip: H193591: WebBIOS only recognizes sixteen ServeRAID controllers

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5076492

- SMBridge

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-62198

- IBM SAS hard drive update program

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-62832

Microsoft

See the following Web addresses for more information:

- Support policy for Microsoft software running in non-Microsoft hardware virtualization software

  http://support.microsoft.com/kb/897615/
- Inside Windows Server 2008 Kernel Changes

  http://technet.microsoft.com/en-us/magazine/cc194386.aspx
- Comparison of Windows Server 2003 Editions

  http://technet2.microsoft.com/windowsserver/en/library/81999f39-41e9-4388-8d7d-7430ec4cc4221033.mspx?mfr=true
- Microsoft Software Assurance

  http://www.microsoft.com/licensing/sa
- SQL Server 2008

  http://www.microsoft.com/sql/2008
- Comparison Between SQL Server 2005 Standard and Enterprise Editions

  http://www.microsoft.com/sql/editions/enterprise/comparison.mspx
- SQL Server 2005 Features Comparison

  http://www.microsoft.com/sql/prodinfo/features/compare-features.mspx
- Application Software Considerations for NUMA-Based Systems

  http://www.microsoft.com/whdc/archive/numa_isv.mspx
- Products Designed for Microsoft Windows – Windows Catalog and HCL

  http://www.microsoft.com/whdc/hcl
- Processor Power Management in Windows Vista and Windows Server 2008

  http://www.microsoft.com/whdc/system/pnppwr/powermgmt/ProcPowerMgmt.mspx
- Windows Server 2008, Compare Technical Features and Specifications

  http://www.microsoft.com/windowsserver2008/en/us/compare-specs.aspx
- WindowsServer catalog

  http://www.windowsservercatalog.com/
- WindowsServer catalog, IBM System x3950 M2

  http://www.windowsservercatalog.com/item.aspx?idItem=dbf1ed79-c158-c428-e19d-5b4144c9d5cd

VMware

See the following Web addresses for more information:

- ESXi 3 Hosts without swap enabled cannot be added to a VMware High Availability Cluster

  http://kb.vmware.com/kb/1004177

- Limited configurations are supported for VMware HA and ESX Server 3i hosts

  http://kb.vmware.com/kb/1004656

- The Architecture of VMware ESX Server 3i

  http://www.vmware.com/files/pdf/ESXServer3i_architecture.pdf

- I/O Compatibility Guide For ESX Server 3.5 and ESX Server 3i

  http://www.vmware.com/pdf/vi35_io_guide.pdf

- Storage / SAN Compatibility Guide For ESX Server 3.5 and ESX Server 3i

  http://www.vmware.com/pdf/vi35_san_guide.pdf

- Systems Compatibility Guide For ESX Server 3.5 and ESX Server 3i

  http://www.vmware.com/pdf/vi35_systems_guide.pdf

- VMware Infrastructure 3, Configuration Maximums

  http://www.vmware.com/pdf/vi3_301_201_config_max.pdf
  http://www.vmware.com/pdf/vi3_35/esx_3/r35/vi3_35_25_config_max.pdf

- ESX Server 3 Configuration Guide

  http://www.vmware.com/pdf/vi3_35/esx_3/r35/vi3_35_25_3_server_config.pdf

- ESX Server 3 Installation Guide

  http://www.vmware.com/pdf/vi3_35/esx_3/r35/vi3_35_25_installation_guide.pdf

- Resource Management Guide

  http://www.vmware.com/pdf/vi3_35/esx_3/r35/vi3_35_25_resource_mgmt.pdf

- ESX Server 3i Configuration Guide

  http://www.vmware.com/pdf/vi3_35/esx_3i_e/r35/vi3_35_25_3i_server_config.pdf

- ESX Server 3i Embedded Setup Guide

  http://www.vmware.com/pdf/vi3_35/esx_3i_e/r35/vi3_35_25_3i_setup.pdf

- Getting Started with ESX Server 3i Installable

  http://www.vmware.com/pdf/vi3_35/esx_3i_i/r35/vi3_35_25_3i_i_get_start.pdf

- I/O Compatibility Guide For ESX Server 3.0.x

  http://www.vmware.com/pdf/vi3_io_guide.pdf
- Systems Compatibility Guide For ESX Server 3.0.x

  http://www.vmware.com/pdf/vi3_systems_guide.pdf
- VMware Infrastructure 3 Release Notes

  http://www.vmware.com/support/vi3/doc/vi3_esx3i_e_35u1_vc25u1_rel_notes.html
- Tips and Tricks for Implementing Infrastructure Services on ESX Server

  http://www.vmware.com/vmtn/resources/409

Novell SUSE Linux

See the following Web addresses for more information:

- YES CERTIFIED Bulletin Search

  http://developer.novell.com/yessearch/Search.jsp
- A NUMA API for LINUX

  http://www.novell.com/collateral/4621437/4621437.pdf
- SUSE Linux Enterprise Server 10 Tech Specs & System Requirements

  http://www.novell.com/products/server/techspecs.html

Intel

See the following Web addresses for more information:

- Execute Disable Bit and Enterprise Security

  http://www.intel.com/technology/xdbit/index.htm
- Intelligent Platform Management (IPMI) Interface Specification

  http://www.intel.com/design/servers/ipmi/ipmiv2_0_rev1_0_markup_2.pdf
- Intel Xeon Processor 7000 Sequence

  http://www.intel.com/products/processor/xeon7000/index.htm?iid=servproc+body_xeon7000subtitle
- Server Processors

  http://www.intel.com/products/server/processors/index.htm?iid=process+server
- Intel 64 Architecture

  http://www.intel.com/technology/intel64

- Dual-Core Intel Xeon Processor 7200 Series and Quad-Core Intel Xeon Processor 7300 Series

  http://download.intel.com/design/xeon/datashts/318080.pdf

- Linux Scalability in a NUMA world

  http://oss.intel.com/pdf/linux_scalability_in_a_numa_world.pdf

Red Hat

See the following Web addresses for more information:

- What Every Programmer Should Know About Memory

  http://people.redhat.com/drepper/cpumemory.pdf

- Red Hat Enterprise Linux Server Version comparison chart

  http://www.redhat.com/rhel/compare/

- Red Hat Hardware Catalog

  https://hardware.redhat.com/

Other

See the following Web addresses for more information:

- IPMItool

  http://ipmitool.sourceforge.net

- OpenIPMI

  http://openipmi.sourceforge.net/

- TrouSerS

  http://trousers.sourceforge.net/

- Java Downloads for all operating systems

  http://www.java.com/en/download/manual.jsp

- Solaris OS: Hardware Compatibility Lists

  http://www.sun.com/bigadmin/hcl/

- HCL for OpenSolaris, Solaris OS

  http://www.sun.com/bigadmin/hcl/data/systems/details/3406.html

How to get Redbooks

You can search for, view, or download Redbooks, Redpapers, Technotes, draft publications and Additional materials, as well as order hardcopy Redbooks, at this Web site:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Index

Planning, Installing, and Managing the IBM System x3950 M2

(0.5” spine)
0.475”<->0.875”
250 <-> 459 pages

Planning, Installing, and Managing the IBM System x3950 M2

Understand the IBM System x3950 M2 and IBM x3850 M2

Learn the technical details of these high-performance servers

See how to configure, install, manage multinode complexes

The x3950 M2 server is the System x flagship server and implements the fourth generation of the IBM X-Architecture. It delivers innovation with enhanced reliability and availability features to enable optimal performance for databases, enterprise applications, and virtualized environments.

The x3950 M2 features make the server ideal for handling complex, business-critical On Demand Business applications such as database serving, business intelligence, transaction processing, enterprise resource planning, collaboration applications, and server consolidation.

Up to four x3950 M2 servers can be connected to form a single-system image comprising of up to 16 six-core processors, up to 1 TB of high speed memory and support for up to 28 PCI Express adapters. The capacity gives you the ultimate in processing power, ideally suited for very large relational databases.

This IBM Redbooks publication describes the technical details of the x3950 M2 scalable server as well as the x3850 M2 server. We explain the configuration options, how 2-node, 3-node and 4-node complexes are cabled and implemented, how to install key server operating systems, and the management tools available to systems administrators.

SG24-7630-00 ISBN 0738431788