

The Lenovo logo is displayed in white text on a black rectangular background.

Lenovo NeXtScale System Water Cool Technology Planning and Implementation Guide

The ultimate in performance and energy efficiency

Describes the warm water cooled offerings from Lenovo

Covers the n1200 WCT Enclosure and nx360 M5 WCT Compute Node

Addresses power, cooling, water flow, and racking

David Watts

Matt Archibald

Jerrold Buterbaugh

Duncan Furniss

David Latino





Lenovo NeXtScale System Water Cool Technology Planning and Implementation Guide

August 2015

Note: Before using this information and the product it supports, read the information in “Notices” on page vii.

Last update on August 2015

This edition applies to:

NeXtScale n1200 WCT Enclosure, type 5468

NeXtScale WCT Water Manifold, type 5469

NeXtScale nx360 M5 WCT Compute Node, type 5467

42U 1100mm Enterprise V2 Dynamic Rack, 93634PX

Rear Door Heat eXchanger for 42U 1100 mm Enterprise V2 Dynamic Racks, 1756-42X

© Copyright Lenovo 2015. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract

Contents

Notices	vii
Trademarks	viii
Preface	ix
The team who wrote this book	ix
Comments welcome	xi
Chapter 1. Introduction	1
1.1 Evolution of the data center	2
1.1.1 Density	2
1.1.2 Scale out applications	2
1.2 Summary of the key components	3
1.2.1 Lenovo NeXtScale n1200 WCT Enclosure	4
1.2.2 NeXtScale nx360 M5 WCT	5
1.3 Design points of the system	6
1.4 NeXtScale System cooling choices	7
1.5 This book	7
Chapter 2. Positioning	9
2.1 Market positioning	10
2.1.1 Three key messages with NeXtScale	11
2.1.2 Optimized for workloads	13
2.2 System x and ThinkServer overview	14
2.3 NeXtScale System versus iDataPlex	15
2.4 Ordering and fulfillment	16
2.5 Cooling with water	16
Chapter 3. NeXtScale n1200 WCT Enclosure	19
3.1 Overview	20
3.1.1 Front components	20
3.1.2 Rear components	21
3.1.3 Fault tolerance features	22
3.2 Standard chassis models	22
3.3 Supported compute nodes	22
3.3.1 nx360 M5 WCT tray support	23
3.4 Power supplies	25
3.5 Fan modules	28
3.6 n1200 WCT Manifold	28
3.7 Midplane	29
3.8 Fan and Power Controller	30
3.8.1 Ports and connectors	30
3.8.2 Internal USB memory key	32
3.8.3 Overview of functions	32
3.8.4 Web GUI interface	33
3.9 Power management	34
3.9.1 Power Restore policy	34
3.9.2 Power capping	35
3.9.3 Power supply redundancy modes	35
3.9.4 Power supply oversubscription	35

3.9.5 Acoustic mode	37
3.9.6 Smart Redundancy mode	37
3.10 Specifications	37
3.10.1 Physical specifications	37
3.10.2 Supported environment.	38
Chapter 4. NeXtScale nx360 M5 WCT Compute node	39
4.1 Overview	40
4.1.1 Scalability and performance	41
4.1.2 Manageability and security	42
4.1.3 Energy efficiency.	42
4.1.4 Availability and serviceability.	43
4.1.5 Locations of key components and connectors	43
4.1.6 System Architecture	44
4.2 Specifications	46
4.3 Processor options	48
4.4 Memory options.	49
4.5 I/O expansion options	53
4.6 Network adapters	53
4.7 Local server management.	56
4.8 Remote server management.	57
4.9 Supported operating systems	59
4.10 Physical and electrical specifications	60
4.11 Regulatory compliance	61
Chapter 5. Rack planning	63
5.1 Power planning	64
5.1.1 NeXtScale WCT Rack Power Reference Examples	65
5.1.2 Examples	66
5.1.3 PDUs.	68
5.1.4 UPS units	69
5.2 Cooling	69
5.2.1 Planning for air cooling	69
5.2.2 Planning for water cooling.	71
5.3 Density	75
5.4 Racks	75
5.4.1 Rack Weight	75
5.4.2 The 42U 1100mm Enterprise V2 Dynamic Rack	76
5.4.3 Installing NeXtScale WCT System in other racks	81
5.4.4 Shipping the chassis.	82
5.4.5 Rack options	83
5.5 Cable management.	87
5.6 Rear Door Heat eXchanger.	89
5.7 Top-of-rack switches	93
5.7.1 Ethernet switches	93
5.7.2 InfiniBand switches	94
5.7.3 Fibre Channel switches.	94
5.8 Rack-level networking: Sample configurations	95
5.8.1 Non-blocking InfiniBand	96
5.8.2 A 50% blocking InfiniBand	97
5.8.3 10 Gb Ethernet, one port per node	98
5.8.4 10 Gb Ethernet, two ports per node	99
5.8.5 Management network	100

Chapter 6. Factory integration and testing	101
6.1 Lenovo Intelligent Cluster	102
6.2 Lenovo factory integration standards	102
6.3 Factory testing.	103
6.4 Documentation provided	105
6.4.1 HPLinpack testing results: Supplied on request	106
Chapter 7. Managing a NeXtScale environment	109
7.1 Managing compute nodes	110
7.1.1 Integrated Management Module II	110
7.1.2 Unified Extendible Firmware Interface	111
7.1.3 ASU.	121
7.1.4 Firmware upgrade.	122
7.2 Managing the chassis	123
7.2.1 FPC web browser interface.	123
7.2.2 FPC IPMI interface	140
7.3 ServeRAID C100 drivers: nx360 M4	148
7.4 Integrated SATA controller: nx360 M5	148
7.5 VMware vSphere Hypervisor	148
7.6 eXtreme Cloud Administration Toolkit.	149
Abbreviations and acronyms	153
Related publications	157
Lenovo Press publications	157
Other publications and online resources	157

Notices

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

Lenovo 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:

Lenovo (United States), Inc.
1009 Think Place - Building One
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO 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 jurisdictions 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. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary.

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

Any references in this publication to non-Lenovo 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 Lenovo product, and use of those Web sites is at your own risk.

Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Trademarks

Lenovo, the Lenovo logo, and For Those Who Do are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. These and other Lenovo 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 Lenovo at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of Lenovo trademarks is available on the Web at <http://www.lenovo.com/legal/copytrade.html>.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

Advanced Settings Utility™	Lenovo®	ServerProven®
BladeCenter®	NeXtScale™	System x®
Dynamic System Analysis™	NeXtScale System®	ThinkServer®
Flex System™	RackSwitch™	TruDDR4™
iDataPlex®	Lenovo (logo)®	UpdateXpress System Packs™
Intelligent Cluster™	ServerGuide™	

The following terms are trademarks of other companies:

Intel, Intel Xeon, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.

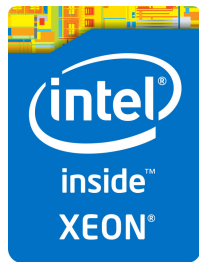
Preface

NeXtScale™ System is a dense computing offering based on Lenovo®'s experience with iDataPlex® and Flex System™ and with a tight focus on emerging and future client requirements. The NeXtScale n1200 Enclosure and NeXtScale nx360 M5 Compute Node are designed to optimize density and performance within typical data center infrastructure limits.

The 6U NeXtScale n1200 Enclosure fits in a standard 19-inch rack and up to 12 compute nodes can be installed into the enclosure. The WCT warm water-cooled versions of the NeXtScale n1200 and nx360 M5 offer the ultimate in performance and energy efficiency. With more computing power per watt and the latest Intel Xeon processors, you can reduce costs while maintaining speed and availability.

This Lenovo Press publication is for customers who want to understand and implement a water-cooled NeXtScale System® solution. It introduces the offering and the innovations in its design, outlines its benefits, and positions it with other x86 servers. The book provides details about NeXtScale System components and supported options, and provides rack, power and water planning considerations.

Lenovo NeXtScale System servers are based on Intel Xeon processors.



The team who wrote this book

This document is produced by the following subject matter experts working in the Lenovo offices in Morrisville, NC, USA.

David Watts is a Senior IT Consultant with Lenovo Press in the Lenovo Enterprise Business Group in Morrisville, North Carolina in the USA. He manages residencies and produces pre-sale and post-sale technical publications for hardware and software topics that are related to System x®, Flex System, and BladeCenter® servers. He has authored over 300 books and papers. Prior to working for Lenovo starting in 2014, David had worked for IBM both in the U.S. and Australia since 1989. David holds a Bachelor of Engineering degree from the University of Queensland (Australia).

Matthew Archibald is the Global Installation & Planning Architect for Lenovo Professional Services. He has been with System x for 12 years and has been working for the last nine years in the data center space. Previous to this, Matt worked as a development engineer in the System x and BladeCenter power development lab doing power subsystem design for BladeCenter and System x servers. Matt is also responsible for the development, maintenance, and support of the System x Power Configurator program and holds 29 patents for various data center and hypervisor technologies. Matt has four degrees from Clarkson University in Computer Engineering, Electrical Engineering, Software Engineering, and

Computer Science and a Bachelor of Engineering in Electronics Engineering from Auckland University of Technology.

Jerrod Buterbaugh is the Worldwide Data Center Services Principal for Lenovo Professional Services. He has been a Data Center practice lead for the last 7 years focused on facility and IT installations and optimization. Previous to this, Jerrod worked as a power development engineer in the System x and IBM POWER Systems power development labs, designing server power subsystems. Jerrod is also the key focal point supporting HPC installations and holds 18 patents for various data center and server technologies.

Duncan Furniss is a Consulting Client Technical Specialist for Lenovo in Canada. He currently provides technical sales support for iDataPlex, NeXtScale, BladeCenter, Flex and System x products. He has co-authored several Lenovo Press and IBM Redbooks publications, including NeXtScale System M5 with Water Cool Technology and NeXtScale System Planning and Implementation Guide. Duncan has designed and provided oversight for the implementation of many large-scale solutions for HPC, distributed databases, and rendering of computer generated images.

David Latino is the lead HPC Architect and HPC Center of Competency leader for Lenovo Middle East & Africa. Prior to working for Lenovo starting in 2015, he was a Consulting IT Specialist, performing the same role for IBM Middle East, Turkey & Africa. David has 12 years of experience in the HPC field. He led a wide spectrum of consulting projects, working with HPC users in academic research and industry sectors. His work covered many aspects of the HPC arena and he was technical leader for the design and implementation of multiple large HPC systems that appeared in the top500 list. David worked extensively on HPC application development, optimization, scaling, and performance benchmark evaluation, which resulted in several highly optimized application software packages. He also spent several years based at customer sites to train system administrators, users, and developers to manage and efficiently use IBM Blue Gene systems.

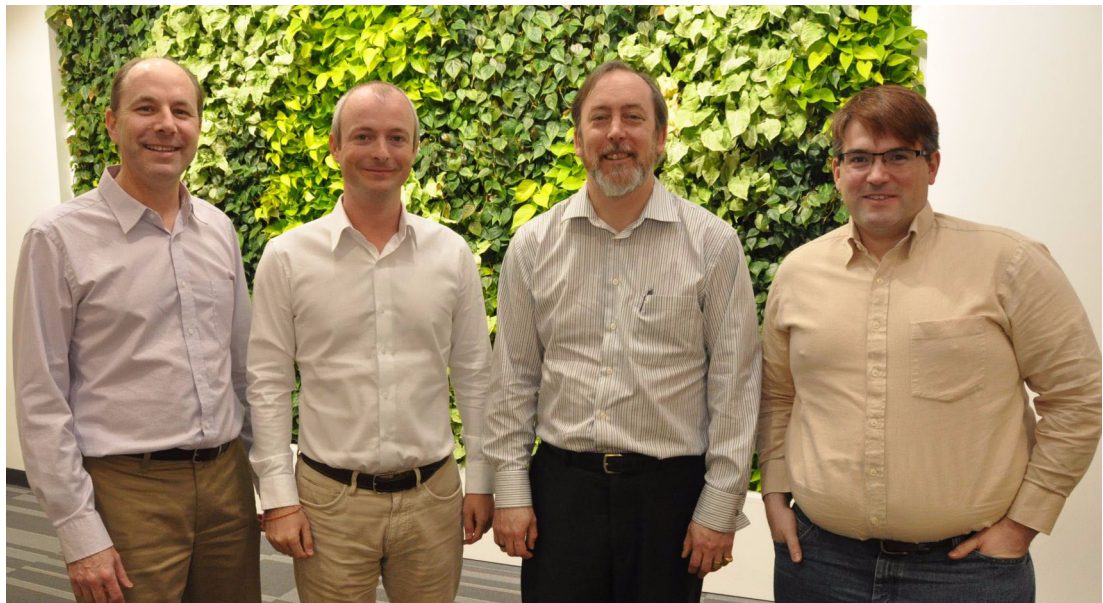


Figure 1 Four of the team (l-r): David Watts, David Latino, Duncan Furniss, and Matt Archibald

Thanks to the following people who contributed to the project:

Lenovo Press:

- ▶ Ilya Krutov

System x Marketing:

- ▶ Jill Caugherty
- ▶ Keith Taylor
- ▶ Scott Tease

NeXtScale Development:

- ▶ Vinod Kamath
- ▶ Edward Kung
- ▶ Mike Miller
- ▶ Wilson Soo
- ▶ Mark Steinke

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 in one of the following ways:

- ▶ Use the online **Contact us** review Redbooks form found at:
ibm.com/redbooks
- ▶ Send your comments in an email to:
redbooks@us.ibm.com

Introduction

NeXtScale System is the next generation of dense computing. It is an open, flexible, and simple data center solution for users of technical computing, grid deployments, analytics workloads, and large-scale cloud and virtualization infrastructures.

NeXtScale System is built with industry-standard components to create flexible configurations of servers, chassis, and networking switches that integrate easily in a standard 19-inch rack. It is a general-purpose platform that provides flexibility to clients for creating unique and differentiated solutions by using off-the-shelf components. Front-access cabling enables you to quickly and easily make changes in networking and power connections.

NeXtScale System Water Cool Technology (WCT) is warm-water cooling technology that offers the highest performance offerings and maximized energy efficiency in the data center. The use of water cooling means that the highest spec processors from Intel can be used, something you cannot do with traditional air cooling.

The NeXtScale n1200 WCT Enclosure and NeXtScale nx360 M5 WCT server are the major components of the offering. These components optimize density and performance within typical data center infrastructure limits. The 6U n1200 enclosure fits in a standard 19-inch rack and up to 12 nx360 M5 servers can be installed into the enclosure.

This chapter includes the following topics:

- ▶ 1.1, “Evolution of the data center” on page 2
- ▶ 1.2, “Summary of the key components” on page 3
- ▶ 1.3, “Design points of the system” on page 6
- ▶ 1.5, “This book” on page 7

1.1 Evolution of the data center

There is an increasing number of computational workloads that can be run on groups of servers, which are often referred to by such names as clusters, farms, or pools. This type of computing can be described as scale-out; however, as a convention, we refer to these groups as *clusters*. As the computing community's proficiency with implementing and managing clusters improved, there is a trend to create large clusters, which are becoming known as *hyper-scale environments*.

In the past, when the number of servers in a computing environment was lower, considerable hardware engineering effort and server cost was expended to create servers that were highly reliable to reduce application downtime. With clusters of servers, we strive to create a balance between the high availability technologies that are built in to every server and reduce the cost and complexity of the servers, which allows more of them to be provisioned.

The mainstream adoption of virtualization and cloud software technologies in the data center caused a paradigm shift at the server level that further removes the reliance on high availability hardware. The focus is now on providing high availability applications to users on commodity hardware with workloads that are managed and brokered with highly recoverable virtualization platforms. With this shift away from hardware reliability, new server deployment methods emerged that allow previous data center barriers to grow without the need for large capital investments.

1.1.1 Density

As the number of servers in clusters grows and the cost of data center space increases, the number of servers in a unit of space (also known as the *compute density*) becomes an increasingly important consideration. NeXtScale System optimizes density while addressing other objectives, such as, providing the best performing processors, minimizing the amount of energy that is used to cool the servers, and providing a broad range of configuration options.

Increased density brings new challenges for facility managers to cool high-performance, highly dense rack-level solutions. To support this increased heat flux, data center facilities teams are investigating the use of liquid cooling at the rack. NeXtScale System was designed with this idea in mind, in that it can use traditional air cooling or can be cooled by using the latest Water Cool Technology.

1.1.2 Scale out applications

The following applications are among the applications that lend to clusters of servers:

- ▶ High performance computing (HPC)

HPC is a general category of applications that are computationally complex, can deal with large data sets, or consist of vast numbers of programs that must be run. Examples of computationally complex workloads include weather modeling or simulating chemical reactions. Comparing gene sequences is an example of a workload that involves large data sets. Image rendering for animated movies and Monte Carlo analysis for particle physics are examples of workloads where there are vast numbers of programs that must be run. The use of several HPC clusters in a Grid architecture is an approach that gained popularity.

- ▶ Cloud services

Cloud services that are privately owned and those services that are publicly available from managed service providers provide standardized computing resources from pools of homogeneous servers. If a consumer requires more or less server capacity, the servers are provisioned from or returned to the pools. This paradigm often also includes consumer self-service and usage metering with some form of show back, charge back, or billing.

- ▶ Analytics

Distributed databases and extensive use of data mining, or analytics, is another use case that is increasing in prevalence and is applied to a greater range of business and technical challenges.

1.2 Summary of the key components

The NeXtScale n1200 WCT Enclosure and NeXtScale nx360 M5 WCT server optimize density and performance within typical data center infrastructure limits. The 6U n1200 enclosure fits in a standard 19-inch rack and up to 12 nx360 M5 servers can be installed into the enclosure.

The NeXtScale WCTs system is the next generation of dense computing. It is an open, flexible, and simple data center solution for users of technical computing, grid deployments, analytics workloads, and large-scale cloud and virtualization infrastructures.

The NeXtScale n1200 WCT enclosure and new NeXtScale nx360 M5 WCT server optimize density and performance within typical data center infrastructure limits. The 6U NeXtScale n1200 WCT enclosure fits in a standard 19-inch rack. Up to 12 nx360 M5 WCT servers on 6 WCT compute trays can be installed into the enclosure.

The NeXtScale WCT solution operates by using warm water, up to 45°C (113°F). Chillers are not needed for most customers, which results in even greater savings and a lower total cost of ownership (TCO).

NeXtScale System M5 WCT is the new generation dense water-cooled platform from System x, following on from the iDataPlex water-cooled system. The NeXtScale WCT system includes a dense chassis, two half-wide compute nodes on the WCT Compute Tray, all fitting in a standard rack. With WCT M5, Lenovo drives increase compute density, performance, and cooling efficiency for High Performance Computing and other workloads that require dense compute performance, such as Cloud, Grid, and Analytics.

One of the most notable features of WCT products is direct water cooling. Direct water cooling is achieved by circulating the cooling water directly through cold plates that contact the CPU thermal case, DIMMs, and other high-heat-producing components in the server.

One of the main advantages of direct water cooling is the water can be relatively warm and still be effective because water conducts heat much more effectively than air. Depending on the server configuration, 85% - 90% of the heat is removed by water cooling; the rest can be easily managed by a standard computer room air conditioner. With allowable inlet temperatures for the water being as high as 45°C (113°F), in many cases the water can be cooled by using ambient air and chilled water and a heat exchanger is not required.

1.2.1 Lenovo NeXtScale n1200 WCT Enclosure

The NeXtScale WCT System is based on a six-rack unit (6U) high chassis with six full-width bays, as shown in Figure 1-1.

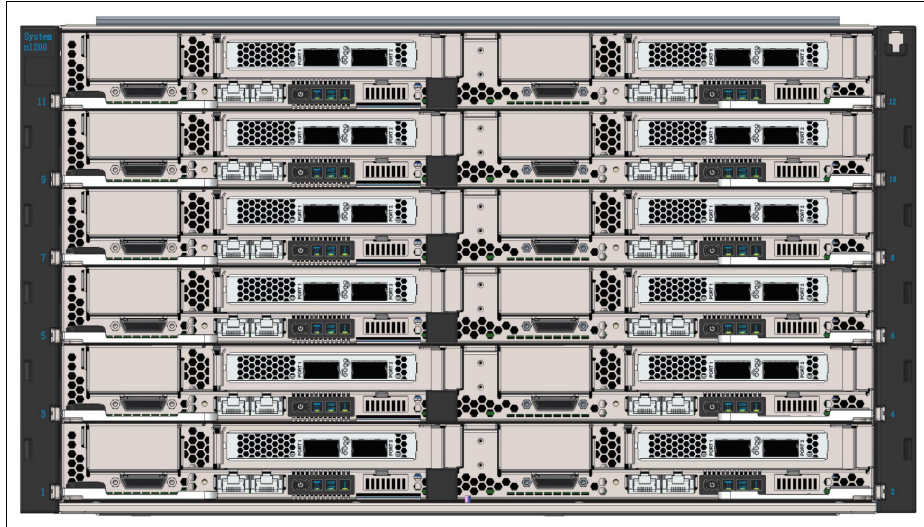


Figure 1-1 Front of NeXtScale n1200 WCT enclosure with six nx360 M5 WCT compute trays

Chassis power and cooling

The NeXtScale n1200 WCT Enclosure includes the n1200 WCT manifold and six hot swappable power supplies, which are installed in the rear of the chassis, as shown in Figure 1-2.



Figure 1-2 Rear of NeXtScale n1200 Enclosure with 1300 W power supplies and WCT manifold shown

Six single-phase power supplies enable power feeds from one or two sources of three-phase power.

Also, in the rear of the chassis is the Fan and Power Controller, which controls power and cooling aspects of the chassis.

1.2.2 NeXtScale nx360 M5 WCT

The first water-cooled server that is available for the NeXtScale System is the NeXtScale nx360 M5 WCT, as shown in Figure 1-3. The NeXtScale nx360 M5 WCT is implemented as two independent compute nodes that are housed on one full-wide tray. Rather than using fans, water is circulated through cooling tubes within the server to cool the processor, memory DIMMs, I/O, and other heat producing components. This feature supports water inlet temperatures of up to 45° C (113 °F), which makes expensive water chillers unnecessary. On the front of the server are the power buttons, status LEDs, and connectors for each compute node.

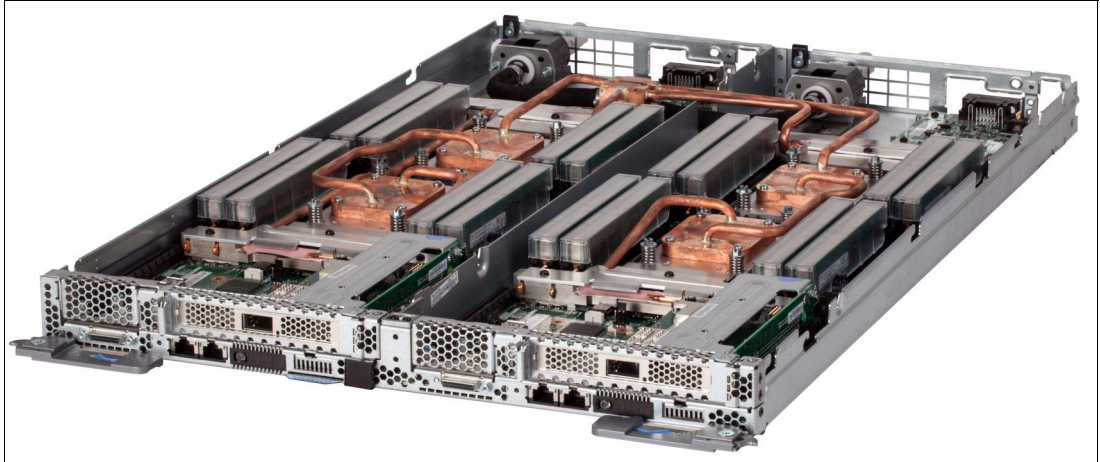


Figure 1-3 NeXtScale nx360 M5 WCT compute tray

Inside, the nx360 M5 WCT compute tray contains two independent nx360 M5 compute nodes. Each planar supports two Intel Xeon E5-2600 v3 series processors and 16 DDR4 DIMMs. There is a full-height, half-length PCI Express card slot and a PCI Express mezzanine card slot that uses the same mezzanine card type as our rack mount servers. The server is shown in Figure 1-4 on page 6.

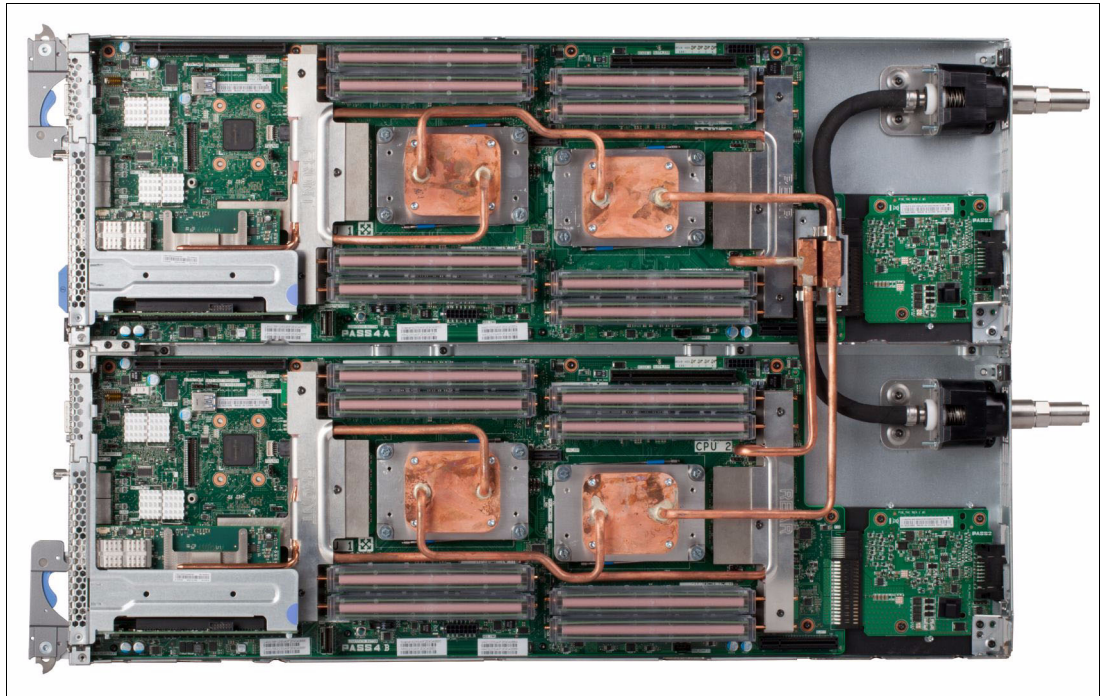


Figure 1-4 NeXtScale nx360 M5 WCT compute tray

1.3 Design points of the system

This section describes some of the following design points that are included in the Lenovo NeXtScale System:

- ▶ System is designed for flexibility

The power supplies in the back of the chassis are modular and hot swappable. The servers slide in and out of the front and have their cable connections at the front.

- ▶ Fits in a standard rack

The NeXtScale n1200 Enclosure can be installed into many standard 19-inch racks (which might require more cable routing brackets) and the rack can have a mixture of NeXtScale and other components. Alternatively, you can have Lenovo install the servers into racks with switches and power distribution units and all of the cables connected.

- ▶ Factory integration available

More configuration and testing are done when the systems are factory-integrated. For more information about Lenovo factory integration, see Chapter 6, “Factory integration and testing” on page 101.

- ▶ NeXtScale System is focused on computational density

Compared to an iDataPlex system with 84 servers in each iDataPlex rack, with six NeXtScale chassis in each 42U standard rack (which leaves 6U per rack for more components), 28% more servers can be fit in the same floor tile configuration. With standard racks, clients can design compact data center floor layouts for all their equipment.

- ▶ System is designed for simplicity

Cable access to the server is from the front. The servers are directly accessed for their local console, management, and data networks, which eliminates contention.

- ▶ Uses standard components

The servers support standard PCI Express (Generation 3) adapters and have RJ-45 copper Ethernet interfaces on board. The chipsets that are used in the server were selected with broad industry acceptance in mind.

1.4 NeXtScale System cooling choices

NeXtScale System is a dense IT solution that is deployable in various data center cooling environments from traditional forced air, raised floor data centers to highly efficient data centers that distribute water to the IT racks and use free cooling.

The following cooling solutions are supported by NeXtScale System:

- ▶ Air-cooled

A traditional cooling topology in which the cooled air is delivered to the IT systems at the front of the rack and the hot exhaust air exits the rear of the rack and is returned to the forced air cooling system.

For more information about the air-cooled NeXtScale options, see *Lenovo NeXtScale System Planning and Implementation Guide*, SG24-8152, which is available at this website:

<http://lenovopress.com/sg248152>

- ▶ Air-cooled and water-cooled with Rear Door Heat Exchanger (RDHX)

Similar to the traditional cooling topology in which cool air is delivered to the front of the rack, but the heat that exits the IT systems is immediately captured by the water-cooled RDHX and returned to the facility chilled water cooling system.

For more information about the benefits of the use of an RDHX with air-cooled NeXtScale options, see *Lenovo NeXtScale System Planning and Implementation Guide*, SG24-8152, which is available at this website:

<http://lenovopress.com/sg248152>

- ▶ Direct water-cooled (WCT)

A non-traditional cooling approach in which cool or warm water is distributed directly to the system CPUs, memory, and other subsystem heat sinks via a rack water manifold assembly.

- ▶ Hybrid - direct water-cooled (WCT) coupled with RDHX

A non-traditional cooling approach that combines WCT with an external RDHX to remove 100% of the rack heat load and provides a room neutral solution.

1.5 This book

In this book, we compare NeXtScale System to other systems and raise points to help you select the right systems for your applications. We then take an in-depth look at the chassis, servers, and fan and power controller (FPC). Next, we take a broader view and cover implementations at scale and review racks and cooling. We then describe Lenovo's process for assembling and testing complete systems in to Intelligent Cluster™ solutions.

Positioning

NeXtScale is ideal for fastest-growing workloads, such as, social media, analytics, technical computing, and cloud delivery, which are putting increased demands on data centers.

This chapter describes how NeXtScale System is positioned in the marketplace compared with other systems that are equipped with Intel processors. The information helps you to understand the NeXtScale target audience and the types of workloads for which it is intended.

This chapter includes the following topics:

- ▶ 2.1, “Market positioning” on page 10
- ▶ 2.2, “System x and ThinkServer overview” on page 14
- ▶ 2.3, “NeXtScale System versus iDataPlex” on page 15
- ▶ 2.4, “Ordering and fulfillment” on page 16
- ▶ 2.5, “Cooling with water” on page 16

2.1 Market positioning

NeXtScale WCT System is a new x86 offering that introduces a new category of dense computing into the marketplace. NeXtScale WCT System includes the following key characteristics:

- ▶ Direct warm-water cooled compute trays to enable the best energy efficiency and compute power.
- ▶ Strategically, this system is the next generation dense system from Lenovo that includes the following features:
 - A building block design that is based on a low function and low-cost chassis.
 - Flexible compute node configurations that are based around a 1U half-wide compute node supports various application workloads.
 - A standard rack platform.
- ▶ Built for workloads that require density
- ▶ NeXtScale performs well in scale-out applications, such as, cloud, HPC, grid, and analytics
- ▶ Is central in OpenStack initiatives for public clouds

NeXtScale WCT System includes the following key features:

- ▶ Supports up to six chassis in a 42U rack, which means up to a total of 72 systems and 2,592 processor cores in a standard 19-inch rack.
- ▶ Industry-standard components for flexibility, ease of maintenance, and adoption.
- ▶ Approved for data centers with up to 40°C ambient air temperature, which lowers cooling costs.
- ▶ Can be configured as part of the Intelligent Cluster processor for complete pre-testing, configuration, and arrival ready to plug in.
- ▶ The use of direct water cooling allows the use of “high bin” processors, such as the 165 W Intel Xeon E5-2698A v3 processor with new 2133 MHz memory.
- ▶ Direct water cooling can enable the Intel Turbo Boost performance feature more to further increase processor performance.
- ▶ Supports 100 - 127 V and 200 - 240 V power.
- ▶ Standard form factor and components make it ideal for Business Partners.

Direct water cooling includes the following benefits:

- ▶ High heat producing components in the server are directly cooled by circulating water across cold plates that are attached to CPUs, DIMMs, and platform control hubs (chipsets).
- ▶ Because water conducts heat 4000% more efficiently than air, even warm water can be effective at removing heat.
- ▶ Depending on the server configuration, 85% - 90% of heat is removed by water cooling.
- ▶ Designed to operate by using warm water with allowable water supply temperature up to 45°C (113°F).

Chillers are not needed for most customers, which means greater savings and a lower total cost of ownership (TCO).

The customer that benefits the most from NeXtScale is an enterprise that is looking for a low-cost, high-performance computing system to start or optimize cloud, big data, Internet, and technical computing applications, which include the following uses:

- ▶ Large data centers that require efficiency, density, scale, and scalability.
- ▶ Public, private, and hybrid cloud infrastructures.
- ▶ Data analytics applications, such as, customer relationship management, operational optimization, risk and financial management, and enabling new business models.
- ▶ Internet media applications, such as, online gaming and video streaming.
- ▶ High-resolution imaging for applications ranging from medicine to oil and gas exploration.
- ▶ “Departmental” uses in which a small solution can increase the speed of outcome prediction, engineering analysis, and design and modeling.

2.1.1 Three key messages with NeXtScale

The three key messages about NeXtScale WCT System is that it is flexible, simple, and scalable, as shown in Figure 2-1.

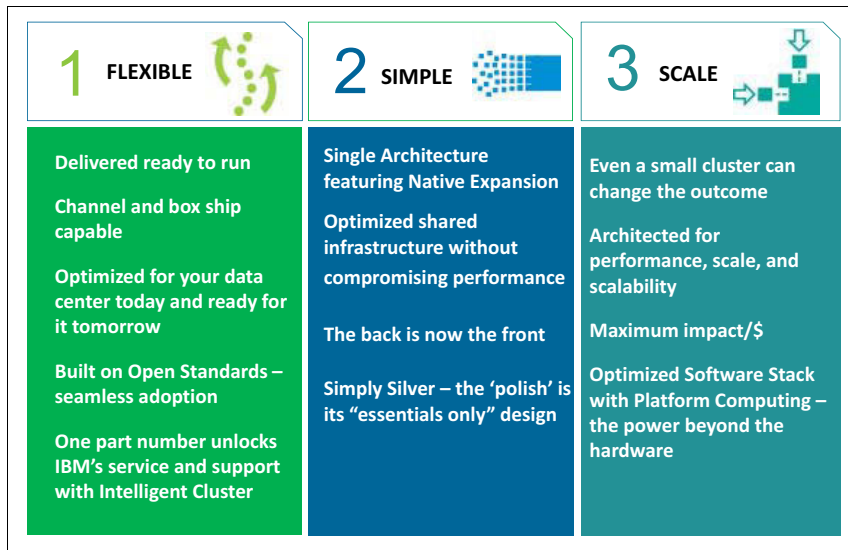


Figure 2-1 NeXtScale System key messages

NeXtScale WCT System is flexible in the following ways:

- ▶ Ordering and delivery

The question of how you want your NeXtScale System configuration to be ordered and delivered is complex because there are many choices. Some clients want everything in parts so that they can mix and match to build what they want. Others want systems that they tested and approved to show that it is configured to their liking. Still others want complete solutions of racks to arrive ready to plug in. With NeXtScale, the choice is yours.
- ▶ Fit it into your data center seamlessly in a Lenovo rack or most 19-inch standard racks.

The NeXtScale n1200 WCT Enclosure is installed in the 42U 1100mm Enterprise V2 Dynamic Rack because it provides the best cabling features. However, the chassis can also be installed in many third-party, 4-post, 19-inch racks. This flexibility ensures maximum flexibility regarding deploying NeXtScale WCT System into your data center.

- ▶ NeXtScale WCT System is backed by leading Lenovo service and support no matter how you buy it, where you use it, or what task you have it running.
- ▶ Support for open standards.

A client needs more than hardware to use IT. We designed NeXtScale to support an open stack of industry standard tools to allow clients that have protocols and tools to migrate easily by using NeXtScale System.

The nx360 M5 compute node offers the Integrated Management Module II service processor and the n1200 Enclosure has the Fan and Power Controller. Both support the IPMI protocol for flexible and standards-based systems managements.

NeXtScale System is simple in the following ways:

- ▶ A design that is based on the half-wide compute node.

The architecture of NeXtScale System revolves around a low-function chassis that hosts compute nodes.

- ▶ A chassis that includes shared cooling.

The n1200 Enclosure supplies the cooling and power to maximize energy efficiency, but leaves management and connectivity to the compute nodes, which minimizes cost.

- ▶ Cables, connectors, and controls at the front.

Except for the power cords and water connections, all cabling is at the front of the chassis. All controls and indicators also are at the front. This configuration makes access, server swap-outs, and overall systems management easier.

Because the cables do not clog up the back of the rack, air flow is improved and thus energy efficiency also is improved. The harder the fans work to move air through server, the more power they use.

The NeXtScale WCT Enclosure uses the power supply fans to pull air through the chassis to cool non-water cooled devices, such as PCI and ML2 adapters. Minimizing cabling at the front and the back of the rack ensures that the minimal amount of power is used to pull air through the chassis, which makes the NeXtScale WCT System one of the most efficient x86 platforms on the market.

Your support staff who work in the data center can tell you the front of the rack is a much more enjoyable environment to spend time in because it might easily be 30 °F (16 °C) cooler at the front than at the back. People cannot stay in the rear of the rack for long before it is no longer comfortable. Also, the front of rack is less noisy than the rear of the rack because of fan noise.

It is also difficult to locate a single dense server in a row of dense racks and then go to the back to service the cabling. Having all of the cabling on the front simplifies and reduces the chances of mis-cabling or pulling the wrong server.

- ▶ Installation in a three-phase power data center.

The design of six power supplies per chassis allows seamless installation into data centers with three-phase power. With six supplies and two, three-phase feeds, power delivery is optimized and balanced; there is no waste, no inefficiency.

- ▶ The compute nodes are unpainted metal.

- ▶ Unlike every other x86 server Lenovo offers, these servers do not have a black front to them, which indicates simplicity and efficiency.

NeXtScale System is scalable in the following ways:

► **Scaling is for everyone**

As we describe scale, it is important to understand that scale is not for massive deployments only; even a small, one-chassis solution can change what users believe can be done.

Whether you start small and grow or start huge and grow enormously, NeXtScale can be run and managed at scale as a single solution.

► **NeXtScale System is built on what was learned about the financial aspects of scale-out.**

Every decision about the product was aimed at improving our clients impact per dollar, whether that meant removing components that are not required or by selecting more energy efficient parts to reduce power usage and, therefore, power costs.

► **Scalable to the container level.**

NeXtScale System can meet the needs of clients who want to add IT at the rack level or even at the container level. Racks can be fully configured, cabled, labeled, and programmed before they are shipped. Lenovo also can take configured racks and assemble them into complete, containerized solutions with power, cooling, and infrastructure delivered ready to go.

2.1.2 Optimized for workloads

NeXtScale System is best-suited for the following primary workloads:

- Public and private cloud
- HPC and technical computing

Although these areas do have much in common, they also have unique needs that are served with NeXtScale System, as shown in Figure 2-2.

Workload	Fine Tuned Server Characteristics
<div style="display: flex; justify-content: space-around; align-items: center;"> <div style="border: 1px solid #0056b3; border-radius: 10px; padding: 5px 15px; background-color: #0056b3; color: white; margin: 5px;">Private Cloud</div> <div style="border: 1px solid #0056b3; border-radius: 10px; padding: 5px 15px; background-color: #0056b3; color: white; margin: 5px;">Public Cloud</div> </div>	<ul style="list-style-type: none"> • Processor and Memory performance and choice Full Intel stack support with memory for performance and/or cost optimization • Standard Rack optimized Fits into client data centers seamlessly • Right sized IO Choice of networking options – 1Gb, 10Gb, or InfiniBand, all SDN ready • Infinitely Scalable from small to enormous grid deployments all built on open standards • High energy efficiency means more impact/watt
<div style="display: flex; justify-content: center; align-items: center;"> <div style="border: 1px solid #008080; border-radius: 10px; padding: 5px 15px; background-color: #008080; color: white; margin: 5px;">High Performance Computing</div> </div>	<ul style="list-style-type: none"> • Top bin Intel Xeon processors, large memory bandwidth, and high IOPS for rapid transaction processing and analytics • Workload optimized software stack with Platform Computing and IBM xCAT • Architected for low latency with choice of high speed fabric support • Supported as one part number no matter the size of the solution and content with Intelligent Cluster

Figure 2-2 NeXtScale System: Optimized for cloud computing and HPC solutions

For cloud, the important factors are that the entire processor stack is supported; therefore, no matter what the client's goal is, it can be supported by the right processor performance and cost point. The same is true of memory; cost and power-optimized choices and performance-optimized alternatives are available. The nodes feature the networking on board with 1 Gb NICs embedded, with options for up to four other high-speed fabric ports. With these features, the entire solution can scale to any size.

HPC and technical computing have many of the same attributes as cloud; a key factor is the need for the top-bin 145 W processors. NeXtScale System can support top bin 145 W processors, which means more cores and higher frequencies than others.

2.2 System x and ThinkServer overview

The world is evolving, and the way that our clients do business is evolving with it. That is why Lenovo has the broadest x86 portfolio in our history and is expanding even further to meet the needs of our clients, whatever those needs might be.

The x86 server market is segmented on the following key product areas:

- ▶ High-end systems

Lenovo dominates this space with enterprise-class X6 four-socket and eight-socket server that offer unprecedented x86 performance, resiliency, and security.

- ▶ Blades and integrated systems

Integrated systems is a fast growing market where Lenovo adds value by packaging Lenovo software and hardware assets in a way that helps our clients optimize value.

- ▶ Dense systems

As with NeXtScale or iDataPlex, dense systems is a fast growing segment that is pushed by data center limitations and new workloads that require scale out architecture. These systems transformed how clients optimize space-constrained data centers with extreme performance and energy efficiency.

- ▶ High volume systems

The volume space is over half the total x86 server market and Lenovo has a broad portfolio of rack and tower servers to meet a wide range of client needs, from infrastructure to technical computing. This portfolio includes offerings from the System x and ThinkServer® brands.

- ▶ System Networking

System Networking solutions are designed for complex workloads that are tuned for performance, virtualization, and massive scale. Lenovo takes an interoperable, standards-based approach to implement the latest advances in today's high-speed, converged data center network designs with optimized applications and integrated systems.

- ▶ Enterprise Storage

Lenovo delivers simplified, centralized storage solutions for small businesses to enterprises with excellent performance and reliability, advanced data protection, and virtualization capabilities for your business-critical data.

2.3 NeXtScale System versus iDataPlex

Although iDataPlex and NeXtScale look different, many of the ideas and innovations we pioneered with iDataPlex remain in the new NeXtScale System.

When we introduced iDataPlex in 2008, we introduced a chassis that was dedicated to power and cool independent nodes. With NeXtScale system, we reuse the same principle, but we are extending it to bring more flexibility to the users.

The NeXtScale n1200 WCT Enclosure supports up to 6 1U full-wide compute trays, while the iDataPlex chassis can house only two 1U half-deep compute nodes.

The NeXtScale n1200 WCT Enclosure fits in the 42U 1100mm Enterprise V2 Dynamic Rack, but it also fits in many standard 19-inch racks. Although iDataPlex also can be installed in a standard 19-inch rack, the use of iDataPlex racks and different floor layouts was required to use its high-density capability. NeXtScale System brings more flexibility by allowing users to use standard 19-inch racks and does not require a special data center layout, which does not affect customer best practices and policies. This flexibility also allows the Lenovo NeXtScale System to achieve greater data center density when it is used among other standard 19-inch racks.

As with iDataPlex servers, NeXtScale servers support S3 mode. S3 allows systems to come back into full production from low-power state much quicker than a traditional power-on. In fact, cold start normally takes about 270 seconds; with S3, it takes only 45 seconds. When you know that a system is not to be used because of time of day or state of job flow, you can send it into a low-power state to save power and, when needed, bring it back online quickly.

Table 2-1 compares the features of NeXtScale System to those features of iDataPlex.

Table 2-1 Comparing NeXtScale System to iDataPlex

Feature	iDataPlex	NeXtScale	Comments
Form factor	Unique rack 1200 mm x 600 mm	Standard rack 600 mm x 1100 mm	NeXtScale System allows for lower-cost racks and customer's racks.
Density in a standard 42U rack	Up to 42 servers	Up to 84 servers (72 with space for switches)	NeXtScale System can provide up to twice the server density when both types of servers are installed in a standard 42U rack.
Density in two consecutive floor tiles ^a	84 servers 8 ToR switches (iDataPlex rack)	144 servers 12 ToR switches (two standard racks next to each other)	NeXtScale can provide up to 71% more servers per row when top-of-rack (ToR) switches are used.
Density/400 sq. ft. (with Rear Door Heat Exchanger)	1,680 servers 84 servers/iDataPlex rack; four rows of five racks	2,160 servers 72 servers/standard rack; three rows of 10 racks	10x10 floor tiles A 28% density increase because of standard rack layout.
Power/tile (front)	22 kW maximum 15 kW typical 42 servers + switches	37 kW maximum 25 kW typical 72 servers + switches	Similar power/server
GPU support	Two GPUs per server in 2U	Two GPUs per server in 1U effective space	GPU tray + base node = 1U. NeXtScale System has twice the density of iDataPlex.

Feature	iDataPlex	NeXtScale	Comments
Direct attached storage	None ^b	Other storage-rich offerings include eight drives in 1U effective space	More flexibility with NeXtScale storage plan.
Direct water cooling	Available Now	NeXtScale System design supports water cooling	Opportunity to optimize cost with NeXtScale.

a. Here we compare the density of servers that can be fitted in a single row of racks while top-of-rack switches are used. We use a single iDataPlex rack for iDataPlex servers that is 1200 mm wide, and we compare it with two standard racks for NeXtScale servers that also are 1200 mm wide.

b. None are available with Intel E5-2600 v2 series processor.

2.4 Ordering and fulfillment

The Lenovo NeXtScale WCT System is ordered through the configure-to-order (CTO) process, as there are no standard models. Because of this fact, it is not possible to order extra n1200 WCT Enclosures, n1200 WCT Manifolds, and nx360 M5 WCT Compute Trays separately. Therefore, the NeXtScale WCT System is not field upgradeable; more chassis cannot be added to partially full racks, more nodes cannot be added to partially full chassis, and the n1200 WCT Manifolds are not serviceable in the field.

2.5 Cooling with water

The advantages of the use of water cooling over air cooling result from water's higher specific heat capacity, density, and thermal conductivity. These features allow water to transmit heat over greater distances with much less volumetric flow and reduced temperature difference as compared to air.

For cooling IT equipment, this heat transfer capability is its primary advantage. Water has a tremendously increased ability to transport heat away from its source to a secondary cooling surface, which allows for large, more optimally designed radiators or heat exchangers rather than small, inefficient fins that are mounted on or near a heat source, such as a CPU.

NeXtScale WCT uses the benefits of water by distributing it directly to the highest heat generating server subsystem components. By doing so, NeXtScale WCT, realizes 7% - 10% direct energy savings when compared to its air-cooled counterpart. That energy savings results from the removal of the system fans and the lower operating temp of the direct water-cooled system components.

Because NeXtScale WCT supports up to 45 °C (113 °F) inlet water for all available processors (with only one exception: the E5-2698A v3 supports inlet water temperatures up to 35 °C), data centers can realize significant operational energy savings by using water-side economizers to cool the NeXtScale WCT chassis. In most climates, water-side economizers can supply water at temperatures below 45°C (113 °F) for most of the year. This ability allows the data center chilled water system to be bypassed thus saving energy because the chiller is the most significant energy consumer in the data center. Typical economizer systems, such as dry-coolers, use only a fraction of the energy that is required by chillers, which produce 6 °C - 10 °C (43 °F - 50 °F) water. The facility energy savings are the largest component of the total energy savings that are realized when NeXtScale WCT is deployed.

The direct energy savings at the NeXtScale WCT chassis level, combined with the potential for significant facility energy savings, makes NeXtScale WCT an excellent choice for customers that are burdened by high energy costs or with a Green mandate.

NeXtScale n1200 WCT Enclosure

The foundation on which NeXtScale System is built is the NeXtScale n1200 WCT Enclosure.

Providing shared, high-efficiency power and cooling for up to 12 compute nodes, this chassis scales with your business needs. There is no built-in networking or switching capabilities, which requires no chassis-level management beyond power and cooling.

This chapter includes the following topics:

- ▶ 3.1, “Overview” on page 20
- ▶ 3.2, “Standard chassis models” on page 22
- ▶ 3.3, “Supported compute nodes” on page 22
- ▶ 3.4, “Power supplies” on page 25
- ▶ 3.5, “Fan modules” on page 28
- ▶ 3.6, “n1200 WCT Manifold” on page 28
- ▶ 3.7, “Midplane” on page 29
- ▶ 3.8, “Fan and Power Controller” on page 30
- ▶ 3.9, “Power management” on page 34
- ▶ 3.10, “Specifications” on page 37

3.1 Overview

The NeXtScale n1200 WCT Enclosure is a 6U next-generation dense server platform with integrated Fan and Power Controller. The n1200 WCT enclosure efficiently powers and cools up to 12 1U half-wide compute nodes, with which clients can install in a standard 42U 19-inch rack that is twice the number of servers per rack-U space that is compared to traditional 1U rack servers.

The founding principle behind NeXtScale System is to allow clients to adopt this new hardware with minimal or no changes to their data center infrastructure, management tools, protocols, and best practices.

The enclosure looks similar to an BladeCenter or Flex System chassis, but it is different as there is no consolidated management interface or integrated switching.

The NeXtScale n1200 WCT Enclosure includes the following components:

- ▶ Up to 12 compute nodes
- ▶ Six power supplies, each separately powered
- ▶ One Fan and Power Controller

3.1.1 Front components

The NeXtScale n1200 WCT Enclosure supports up to 6 1U full-wide compute trays, as shown in Figure 3-1. Each compute tray contains 2 1U half-wide compute nodes.

All compute nodes are front accessible with front cabling as shown in Figure 3-1. From this angle, the chassis looks to be simple because it was designed to be simple, low-cost, and efficient.

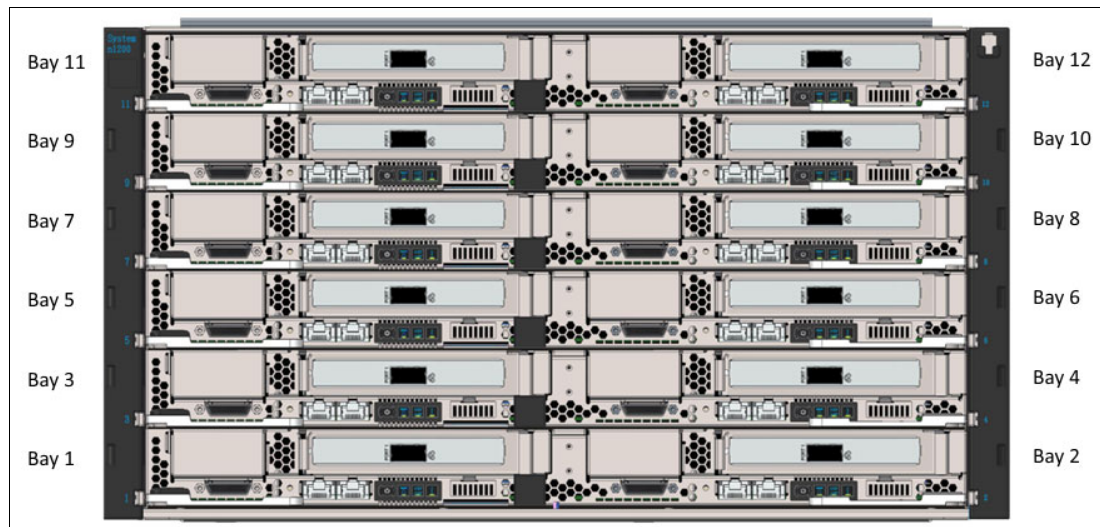


Figure 3-1 NeXtScale n1200 WCT Enclosure front view with 12 compute nodes

With this simple design, clients can access some powerful IT inside a simple and cost effective base compute node that is highly efficient.

3.1.2 Rear components

As with BladeCenter and Flex System, the NeXtScale System compute nodes connect to a midplane, but this connection is for power and control only; the midplane does not provide any I/O connectivity.

Figure 3-2 shows the major components that are accessible from the rear of the chassis.

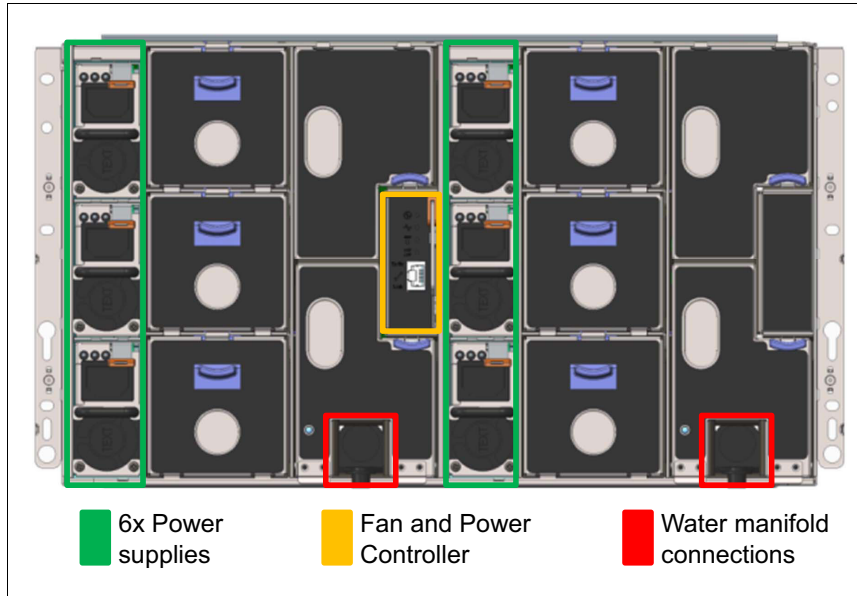


Figure 3-2 NeXtScale n1200 WCT Enclosure rear view

At the rear of the chassis, the following types of components are accessible:

► Power supplies

The NeXtScale n1200 WCT Enclosure has a six-power supply design, all in one power domain. This configuration allows clients with 100 V - 240 V utility power (in single or three-phase) to power up the chassis by using their available infrastructure. For three-phase power, the phases are split in the PDU for single phase input to the chassis power supplies.

For more information about the power supplies, see 3.4, “Power supplies” on page 25.

► Fan modules

The NeXtScale n1200 WCT Enclosure does not have any chassis fans for active cooling.

The power supplies provide cooling for components that are not cooled by the WCT.

► WCT manifold

Each WCT Enclosure contains two manifolds: one for water supply and one for water return.

The manifolds for each chassis are connected in series with other WCT Enclosures installed in a rack.

Note: The WCT Manifold is not expandable after the initial system installation, meaning that the number of WCT Enclosures that are installed in a rack is fixed based on the rack configuration that is ordered.

- ▶ Fan and Power Controller

The Fan and Power Controller (FPC) module is the management device for the chassis and as its name implies, controls the power and cooling features of the enclosure.

For more information about the FPC, see 3.8, “Fan and Power Controller” on page 30.

You might notice that the enclosure does not contain space for network switches. All I/O is routed directly out of the servers to top-of-rack switches. This configuration provides choice and flexibility and keeps the NeXtScale n1200 WCT Enclosure flexible and low-cost.

3.1.3 Fault tolerance features

The chassis implements a fault-tolerant design. The following components in the chassis enable continued operation if one of the components fails:

- ▶ Power supplies

The power supplies support a single power domain that provides DC power to all of the chassis components. If a power supply fails, the other power supplies can continue to provide power.

Power policies: The power management policy that you implemented for the chassis determines the affect on chassis operation if there is a power supply failure. Power policies can be N+N, N+1, or no redundancy. Power policies are managed by the FPC.

- ▶ FPC

The FPC enables the Integrated Management Module to monitor the fans and control fan speed. If the FPC fails, the enclosure fans ramp up to maximum, but all systems continue to operate by using the power management policy.

3.2 Standard chassis models

There are no standard models; all NeXtScale n1200 WCT chassis must be configured by using the CTO process. The machine type is 5468.

The water manifold a separate machine type 5469.

The NeXtScale n1200 WCT Enclosure ships with the following items:

- ▶ Rail kit
- ▶ Four detachable chassis lift handles
- ▶ One Console breakout cable (also known as a KVM Dongle)

3.3 Supported compute nodes

The NeXtScale n1200 WCT Enclosure supports the NeXtScale nx360 M5 WCT Compute Tray only. The number of compute nodes that can be powered on depends on the following factors:

- ▶ The power supply and power policy that is selected (N+N, N+1, or no redundancy)
- ▶ The AC input voltage

- ▶ The components that are installed in each compute node (such as processor, memory, drives, and PCIe adapters)
- ▶ UEFI settings

To size for a specific configuration, Lenovo Data Center Services can assist by providing complete environmental sizing information by contacting power@lenovo.com.

The following number of nodes can be operated with no performance compromise within the chassis depending on the power policy required. The tables use the following conventions:

- ▶ A green cell means that the chassis can be filled with trays up to the maximum number that is supported in the chassis (that is, 6 compute trays).
- ▶ A yellow cell means that the maximum number of trays that the chassis can hold is fewer than the total available bays. Other bays in the chassis must remain empty.

Consider the following points regarding the tables:

- ▶ Oversubscription (OVS) of the power system allows for more efficient use of the available system power. By using oversubscription, users can make the most of the extra power from the redundant power supplies when the power supplies are in healthy condition.
- ▶ OVS and Power supply redundancy options are set via one of the available user interfaces to the Fan and Power Controller in the chassis.

Power Configurator: Use the Power Configurator to determine an accurate power model for your configuration. The configurator is available at this website:

<http://ibm.com/support/entry/portal/docdisplay?lnocid=LNVO-PWRCONF>

3.3.1 nx360 M5 WCT tray support

The nx360 M5 WCT Compute Tray contains two independent nx360 M5 planars. Both planars must be installed in a tray and both processors must be installed on each planar to complete the water loop. Therefore, each nx360 M5 WCT Compute Tray contains four CPUs.

This section describes the maximum number of nodes that can be supported under various configurations.

The nx360 M5 includes the following support tables:

- ▶ 1300 W power supply: 200 - 240 V AC input, no GPU Trays; see Table 3-1 on page 24
- ▶ 900 W power supply:
 - 200 - 240 V AC input, no GPU Trays; see Table 3-2 on page 24
 - 100 - 127 V AC input, no GPU Trays; see Table 3-3 on page 24

Table 3-1 shows the supported quantity of compute nodes with six 1300 W power supplies installed in the chassis.

Table 3-1 Number of supported compute nodes (200 - 240 V AC Input, with 6 x 1300 W PSUs)

Compute nodes		Power policy: 6 x 1300 W power supplies			
CPU TDP	Number of CPUs	Non-redundant or N+1 with OVS ^a	N+1	N+N	N+N with OVS ^a
120 W	4	6	6	3	4
135 W	4	6	6	3	4
145 W	4	6	5	3	4
165 W	4	6	5	3	3

a. Oversubscription (OVS) of the power system allows for more efficient use of the available system power.

Table 3-2 shows the supported quantity of compute nodes with six 900 W power supplies installed in the chassis with those power supplies connected to a 200 - 240 V (high-line) AC input.

Table 3-2 Number of supported compute nodes (200 - 240 V AC Input, with 6 x 900 W PSUs)

Compute nodes		Power policy: 6 x 900 W power supplies			
CPU TDP	Number of CPUs	Non-redundant or N+1 with OVS ^a	N+1	N+N	N+N with OVS ^a
120 W	4	5	4	2	3
135 W	4	4	4	2	2
145 W	4	4	3	2	2
165 W	4	4	3	2	2

a. Oversubscription (OVS) of the power system allows for more efficient use of the available system power.

Table 3-3 shows the supported quantity of compute nodes with six 900 W power supplies installed in the chassis with those power supplies connected to a 100 - 127 V (low-line) AC input.

Table 3-3 Number of supported compute nodes (100 - 127 V AC Input, with 6 x 900 W PSUs)

Compute nodes		Power policy: 6 x 900 W power supplies			
CPU TDP	Number of CPUs	Non-redundant or N+1 with OVS ^a	N+1	N+N	N+N with OVS ^a
120 W	4	3	2	1	2
135 W	4	3	2	1	1
145 W	4	3	2	1	1
165 W	4	2	2	1	1

a. Oversubscription (OVS) of the power system allows for more efficient use of the available system power.

3.4 Power supplies

The NeXtScale n1200 Enclosure supports up to six high-efficiency autoranging power supplies. The standard model includes all six power supplies. Available power supplies are shown in Figure 3-3.



Figure 3-3 Available power supplies for NeXtScale n1200 WCT Enclosure

Table 3-4 lists the ordering information for the supported power supplies.

Table 3-4 Power supplies

Part number	Feature code	Description	Min / Max supported	Chassis model where used
00Y8569	A41T	CFF 900 W Power Supply (80 PLUS Platinum)	6 / 6	A2x, B2x
00Y8652	A4MM	CFF 1300 W Power Supply (80 PLUS Platinum)	2 / 6	A3x, A4x, B3x, B4x
00MU774	ASYH	NeXtScale n1200 1300W Titanium Power Supply	2 / 6	-
00MU775	ASYJ	NeXtScale n1200 1500W Platinum Power Supply	2 / 6	-

The power supply options include the following features:

- ▶ Supports N+N or N+1 Power Redundancy, or Non-redundant power configurations to support higher density
- ▶ Power management controller and configured through the Fan and Power Controller
- ▶ Integrated 2500 RPM fan
- ▶ 80 PLUS Platinum or Titanium certified
- ▶ Built-in overload and surge protection

The 900 W AC power supply features the following specifications:

- ▶ Supports dual-range voltage: 100 - 240 V
- ▶ 100 - 127 (nominal) V AC; 50 or 60 Hz; 6.8 A (maximum)
- ▶ 200 - 240 (nominal) V AC; 50 or 60 Hz; 5.0 A (maximum)

The 1300 W AC power supply features the following specifications:

- ▶ Supports high-range voltage only: 200 - 240 V
- ▶ 200 - 240 (nominal) V AC; 50 or 60 Hz; 6.9 A (maximum)

1500 W AC power supply specifications:

- ▶ Supports high-range voltage only: 200 - 240 V
- ▶ 200 - 240 (nominal) V AC; 50 or 60 Hz; 8.2 A (maximum)

200-240V only: The 1500 W AC and 1300 W AC power supplies do not support low-range voltages (100 - 127 V).

The location and numbering of the power supplies are shown in Figure 3-4.



Figure 3-4 NeXtScale n1200 WCT Enclosure rear view with power supply numbering

The power supplies that are used in NeXtScale System are hot-swap, high-efficiency 80 PLUS Platinum power supplies that are operating at 94% peak efficiency. The efficiency varies by load, as shown in Table 3-5. The 80 PLUS report is available at the following websites:

- ▶ 900 W AC Power Supply 80 PLUS report:
http://www.plugloadsolutions.com/psu_reports/IBM_7001700-XXXX_900W_S0-571_Report.pdf
- ▶ 1300 W Power Supply 80 PLUS report:
http://www.plugloadsolutions.com/psu_reports/IBM_700-013496-XXXX_1300W_S0-628_Report.pdf

Table 3-5 Power efficiencies at different load levels

	20% load	50% load	100% load
80 PLUS Platinum standard	90.00%	94.00%	91.00%
NeXtScale n1200 900 W power supply	92.00%	94.00%	91.00%
NeXtScale n1200 1300 W power supply	92.00%	94.00%	91.00%

The 80 PLUS performance specification is for power supplies that are used within servers and computers. To meet the 80 PLUS standard, the power supply must have an efficiency of 80% or greater, at 20%, 50%, and 100% of rated load with PF of 0.9 or greater. The standard includes several grades, such as, Bronze, Silver, Gold, Platinum, and Titanium. For more information about the 80 PLUS standard, see this website:

<http://www.80PLUS.org>

The power supplies receive electrical power from a 100 V - 127 V AC or 200 V- 240 V AC power source and convert the AC input into DC outputs. The power supplies can autorange within the input voltage range.

Use with 110 V - 127 V AC: When low input voltage (100 V - 127 V AC) is used, the power supply is limited to 600 W.

There is one common power domain for the chassis that distributes DC power to each of the nodes and modules through the system midplane.

DC redundancy is achieved when there is one more power supply available than is needed to provide full power to all chassis components. AC redundancy is achieved by distributing the AC power cord connections between independent AC circuits. For more information, see 5.1, “Power planning” on page 64.

Each power supply includes presence circuitry, which is powered by the midplane. This circuitry allows the FPC to recognize when power supplies are installed in the enclosure but are not powered by the AC supply.

The power supplies support oversubscription. By using oversubscription, users can make the most of the extra power from the redundant power supplies when the power supplies are in a healthy condition. You can use the power capacity of all installed power supplies while still preserving power supply redundancy if there is a power supply failure. For more information, see 3.9.4, “Power supply oversubscription” on page 35.

As shown in Figure 3-3 on page 25, the following LEDs are on each power supply:

- ▶ AC power LED
When this LED is lit (green), it indicates that AC power is supplied to the power supply.
- ▶ DC power LED
When this LED is lit (green), it indicates that DC power is supplied from the power supply to the chassis midplane.
- ▶ Fault LED
When this LED is lit (yellow), it indicates that there is a fault with the power supply.

Removing a power supply: To maintain proper system cooling, do not operate the NeXtScale n1200 Enclosure without a power supply (or power supply filler) in every power supply bay. Install a power supply within 1 minute of the removal of a power supply.

3.5 Fan modules

The NeXtScale n1200 WCT Enclosure does not use cooling fans. Cooling to the nodes is provided through the Water Cool Technology. Other component, such as I/O cards and hard disk drives, are cooled with air that is drawn through the chassis by the power supplies. Water Cool Technology can remove up to approximately 85% of the total chassis heat with the remaining heat removed with the power supply fans.

3.6 n1200 WCT Manifold

The WCT Manifold delivers water to each of the nodes from the CDU. Each manifold section attaches to a chassis and connects directly to the water inlet and outlet connectors for each compute node to safely and reliably deliver water to and from each WCT Compute Tray.

The WCT Manifold is modular and is available in multiple configurations that are based on the number of chassis drops that are required in a rack. The WCT Manifold scales to support up to six WCT Enclosures in a single rack, as shown in Figure 3-5.

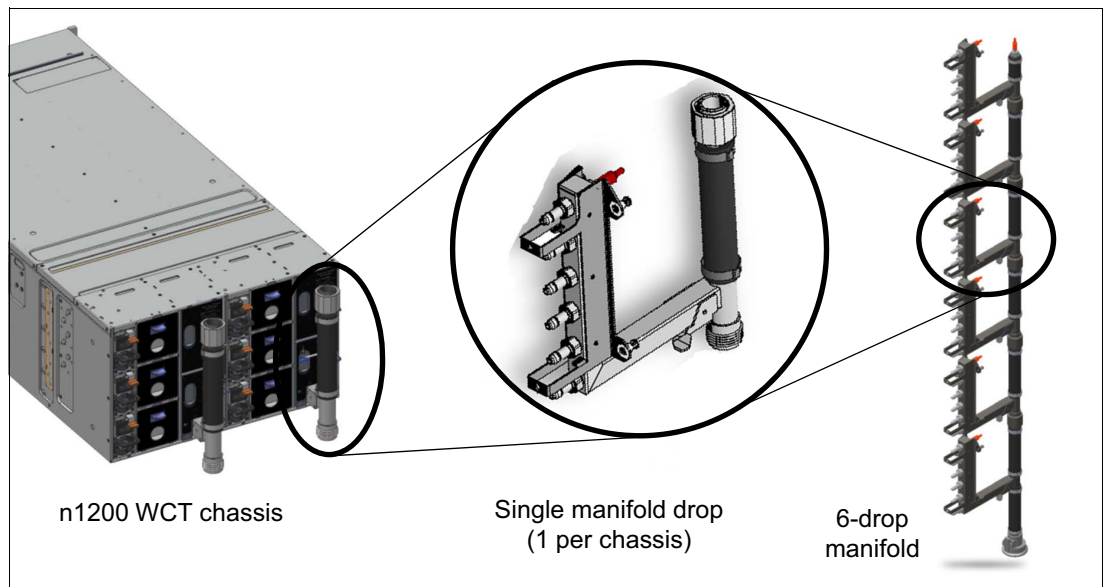


Figure 3-5 n1200 WCT Enclosure and Manifold assembly showing scaled up WCT Manifold

3.7 Midplane

The enclosure midplane is the bridge to connect the compute nodes with the power supplies, fan modules, and the FPC. Figure 3-6 shows the front and rear of the midplane.

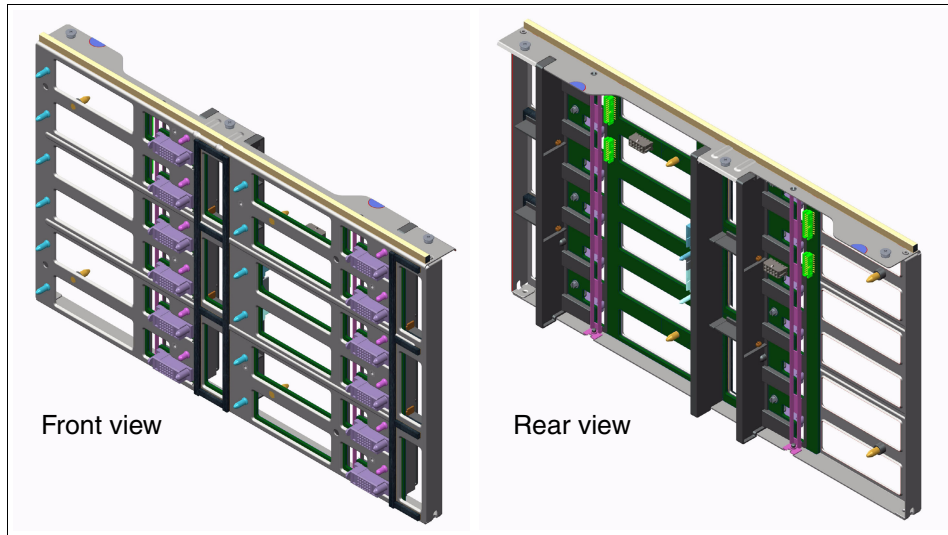


Figure 3-6 Front and rear view of the n1200 Enclosure Midplane Assembly

The midplane is used to provide power to all elements in the chassis. It also provides signals to control fan speed, power consumption, and node throttling.

The midplane was designed with no active components to improve reliability and minimize serviceability. Unlike BladeCenter, the midplane is removed by removing a cover from the top of the chassis.

Figure 3-7 shows the connectivity of the chassis components through the midplane.

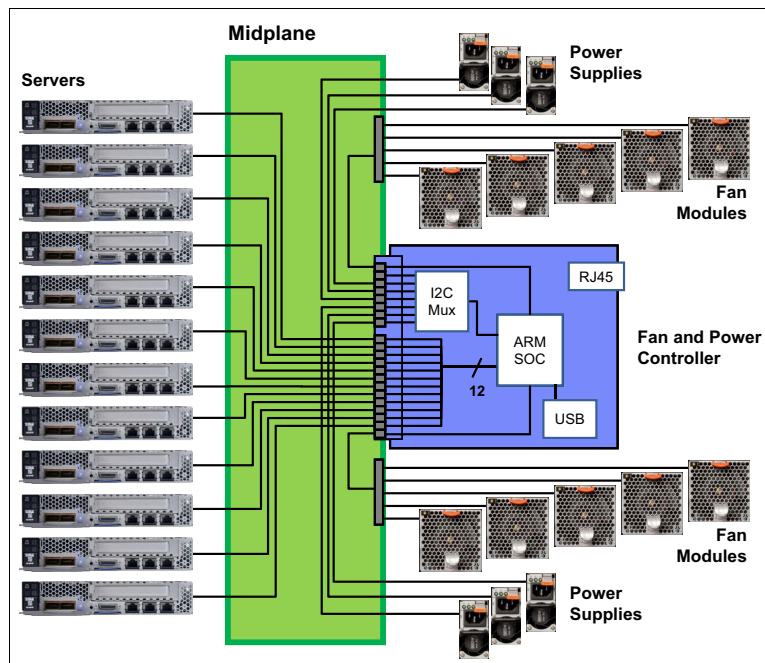


Figure 3-7 Midplane connectivity

3.8 Fan and Power Controller

The Fan and Power Controller (FPC) controls the power budget, provides the power permission to each node, and controls the speed of the fans. The FPC is installed inside the chassis and is accessible from the rear of the chassis, as shown in Figure 3-8. The FPC is a hot-swap component, as indicated by the orange handle.



Figure 3-8 Rear view of the chassis that shows the location of the FPC

3.8.1 Ports and connectors

The FPC provides integrated systems management functions. The user interfaces (browser and CLI) are accessible remotely via the 10/100 Mbps Ethernet port.

Figure 3-9 shows the FPC and its LEDs.

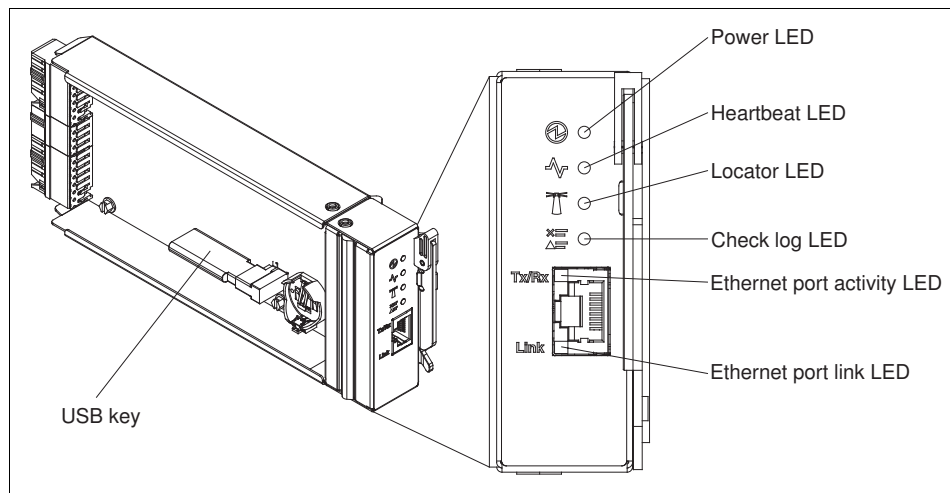


Figure 3-9 FPC

The FPC has the following LEDs and connector that you can use to obtain status information and restart the FPC:

- ▶ Power LED

When this LED is lit (green), it indicates that the FPC has power.

- ▶ Heartbeat LED

When this LED is lit (green), it indicates that the FPC is actively controlling the chassis.

- ▶ Locator LED

When this LED is lit or flashing (blue), it indicates the chassis location in a rack. The locator LED is lights or flashes in response to a request for activation via the FPC web interface or a systems management application.

- ▶ Check log LED

When this LED is lit (yellow), it indicates that a system error occurred.

- ▶ Ethernet port activity LED

When this LED is flashing (green), it indicates that there is activity through the remote management and console (Ethernet) port over the management network.

- ▶ Ethernet port link LED

When this LED is lit (green), it indicates that there is an active connection through the remote management and console (Ethernet) port to the management network.

- ▶ Remote management and console (Ethernet) connector

The remote management and console RJ45 connector is the management network connector for all chassis components. This 10/100 Mbps Ethernet connector is connected to the management network through a top-of-rack switch.

Note: The FPC is not a point of failure for the chassis. If the FPC fails, the compute nodes, power supplies, and fans remain functional to keep the systems running. The power capping policies that are set for the chassis and compute nodes remain in place. The fans speed up to maximum to provide cooling for the compute nodes until the FPC is replaced.

3.8.2 Internal USB memory key

The FPC also includes a USB key that is housed inside the unit, as shown in Figure 3-10.

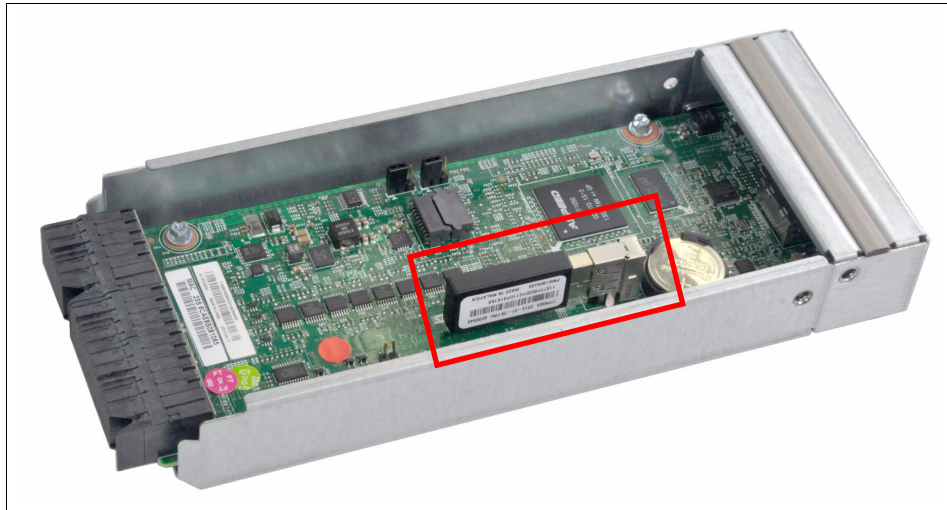


Figure 3-10 Internal view of the FPC

The USB key saves the following information:

- ▶ Event log
- ▶ Enclosure configuration data:
 - PSU redundancy setting
 - Oversubscription mode setting
 - Chassis/node-level power capping value and settings
 - Power restore policy
 - Acoustic mode setting
- ▶ Midplane vital product data (VPD)

If the FPC fails and must be replaced, users can restore the configuration data to the new FPC by transferring the USB from the old unit to the new unit.

3.8.3 Overview of functions

The FPC performs the following functions:

- ▶ Controls the power and power management features of the chassis, compute nodes, and power supplies. The FPC prevents a compute node from powering on if there is insufficient power available from the power supplies.
- ▶ Controls the cooling of the chassis. The FPC ramps up power supply fan speeds if conditions require more cooling or slows down the power supply fans to conserve energy if less cooling is possible.
- ▶ Provides the following user interfaces:
 - Web interface
 - IPMI command line (for external management tools, such as ipmitool or xCAT)
- ▶ Allows you to select one of the following power supply redundancy policies:
 - N+1, where one power supply is redundant and allows for a single power supply to fail without any loss of function.

- N+N, where half the power supplies are redundant backups of the other half. This interface is useful if you have two power utility sources and want the chassis to survive the failure of one of the utility sources.
 - No redundancy, which maximizes the power that is available to the compute nodes at the expense of power supply redundancy.
 - Oversubscription, which can be enabled with N+1 and N+N policies.
 - Smart Redundancy mode, which can disable power conversion on some power supplies to increase the efficiency of the other power supplies during times with low-power requirements.
- ▶ Supports updating the FPC firmware.
 - ▶ Monitors and reports power supply, and chassis status and other failures in the event log and with corresponding LEDs.

3.8.4 Web GUI interface

Through the FPC web interface, the user or system administrator can perform the following tasks. For more information about the FPC web interface, see 7.2.1, “FPC web browser interface” on page 123:

- ▶ View summary of elements status:
 - Front and rear view of the chassis
 - Compute nodes location and status
 - FPC, power supplies, and status
- ▶ View current power usage:
 - Voltage overview of the chassis
 - Total chassis power consumption (AC-in)
 - Total PSU power consumption (DC-out)
 - Per-node power consumption
 - Power supply fan speeds
- ▶ View and set power supply redundancy, oversubscription, and smart redundancy:
 - Select No Redundancy, N+1, or N+N Redundant mode
 - Enable or disable oversubscription mode
 - Select Disabled or 10-, 30-, or 60-minute scanning periods for Smart Redundancy mode
- ▶ View and set power capping:
 - Node level: Set value within a defined range for each node separately, or choose between one of the three predefined modes.
 - Chassis level: Set value within a defined range for the enclosure, or choose between one of the three predefined modes.
- ▶ View and set power restore policy: Enable or disable (for each node or chassis)
- ▶ View current fan speeds
- ▶ View and set Acoustic mode (three modes)
- ▶ View Chassis, Midplane, and FPC vital product data (VPD) details
- ▶ View, save, and clear the system event log

- ▶ View and set network configuration:
 - SMTP configuration
 - SNMP traps and email alert configuration
 - Host name, DNS, Domain, IP, and IP version configuration
 - SNMP traps email alert configuration
 - Web server (http or https) ports configuration
- ▶ Perform a Virtual Reset or Virtual Reseat of each compute node
- ▶ Set Locator (Identify) LED to on, off, or flash
- ▶ Turn off Check Log LED
- ▶ Back up and restore FPC configuration to USB key; reset to default
- ▶ Perform firmware update
- ▶ Set date and time
- ▶ Perform user account management

3.9 Power management

The FPC controls the power on the NeXtScale n1200 Enclosure. If there is sufficient power available, the FPC allows a compute node to be powered on.

The power permission includes the following two-step process:

1. Pre-boot (BMC stage) inventory power (standby power) is pre-determined based on the node type.
2. Post-boot (UEFI stage) inventory power is a more accurate estimation of the node's maximum power usage that is based on power maximizer test. The following values are generated:
 - Maximum power usage value under stressed condition.
 - Maximum power usage value under stressed condition when P-state is capped at the lowest level.

The FPC uses these values to compare the total node power and total available power to determine power-on and boot permissions.

3.9.1 Power Restore policy

By using the Power Restore policy, you can specify whether you want the compute nodes to restart when a chassis AC power is removed and restored. This policy is similar to the Automatic Server Restart (ASR) feature of many System x servers. For more information about the Power Restore Policy, see “Power Restore Policy tab” on page 132.

When Power Restore policy is enabled, the FPC turns the compute node back on automatically when power is restored if the compute node was powered on before the AC power cycle.

However, a compute node is restarted only if the FPC determines there is sufficient power available to power on the server, which is based on the following factors:

- ▶ Number of working power supplies
- ▶ Power policy that is enabled
- ▶ The oversubscription policy

If there is insufficient power, some compute nodes are not powered on. The nodes are powered on based on their power usage, lowest first. The objective is to maximize the number of nodes that can be powered on in a chassis. The power-on sequence nominally takes approximately 2 minutes with most of that time spent by the nodes running a power maximizer. After that process is complete, the FPC can quickly release the nodes to continue to boot.

3.9.2 Power capping

Users can choose chassis-level capping or saving, or node-level capping or saving, through the power capping configuration options. Power capping allows users to set a wattage limit on power usage. When it is applied to an individual node, the node power consumption is capped at an assigned level. When it is applied to a chassis, the whole chassis power usage is capped. When power saving is enabled, an individual node or all nodes (chassis level) run in modes of different throttling level, depending on the modes that are chosen.

3.9.3 Power supply redundancy modes

The FPC offers the following power supply redundancy modes:

- ▶ No redundancy mode

The loss of any power supply can affect the system's operation or performance. If the chassis is evaluated to be vulnerable, because of the failure of one or multiple power supplies, throttle signals are sent to all nodes in the chassis to be throttled down to the lowest power level possible (CPU or Memory lowest P-state). If the power usage remains too high, the chassis is shut down.

This mode is the default mode and does not support the oversubscription mode (see 3.9.4, "Power supply oversubscription" on page 35).

- ▶ N+1 Mode

One installed power supply is used as redundant. The failure of one power supply is allowed without affecting the system's operation or performance (performance can be affected if oversubscription mode is enabled).

This mode can be enabled with oversubscription mode.

- ▶ N+N Mode

Half of the power supplies that are installed are used as redundant. The failure of up to half the number of the power supplies is allowed without affecting the system's operation or performance (performance can be affected if oversubscription mode is enabled). This mode is useful if you have two power sources from two separate PDU for example.

This mode can be enabled with oversubscription mode.

3.9.4 Power supply oversubscription

By using oversubscription, users can make the most of the extra power from the redundant power supplies when the power supplies are in healthy condition.

For example, when oversubscription is enabled with N+1 redundancy mode, the total power that is available is equivalent to No Redundancy mode with six power supplies in the chassis. This configuration means that the power of six power supplies can be counted on instead of five for normal operation.

When oversubscription mode is enabled with redundant power (N+1 or N+N redundancy), the total available power is 120% of the label ratings of the power supplies. Therefore, for a 1300 W power supply, it can be oversubscribed to $1300\text{ W} \times 120\% = 1,560\text{ W}$.

For example, with an N+1 power policy and six power supplies, instead of $5 \times 1300\text{ W}$ (6500 W) of power, there are $5 \times 1300\text{ W} \times 120\%$ (7800 W) of power that is available to the compute nodes.

Table 3-6 lists the power budget that is available, depending on the redundancy and oversubscription mode that is selected.

Table 3-6 Power budget for 6 x 1300 W power supplies

Redundancy Mode	Oversubscription mode	Power budget ^a
Non-redundant	Not available	7800 W (= 6 x 1300 W)
N+1	Disabled	6500 W (= 5 x 1300 W)
	Enabled	7800 W (= 5 x 1300 W x 120%)
N+N	Disabled	3900 W (= 3 x 1300 W)
	Enabled	4680 W (= 3 x 1300 x 120%)

a. The power budget that is listed in this table is based on power supply ratings. Actual power budget can vary.

When oversubscription mode is enabled with redundant power (N+1 or N+N redundancy), the chassis' total available power can be stretched beyond the label rating (up to 120%). However, the power supplies can sustain this oversubscription for a limited time (approximately 1 second).

In healthy condition (all power supplies are in normal-operational mode), the redundant power supplies provide the extra 20% power oversubscription load for the rest of the normal-operational power supplies (none of the power supplies are oversubscribed).

When redundant power supplies fail (that is, one power supply failure in N+1 mode, or up to N power supplies fail in N+N mode), the remaining normal-operational power supplies provide the extra 20% power oversubscription load. This extra power is provided for a limited time only to allow the compute nodes to throttle to the lowest P-state to reduce their power usage back to a supported range. By design, the compute nodes perform this action quickly enough and operation continues.

Non-redundant mode: It is not possible to enable the oversubscription mode without any power redundancy.

The Table 3-7 on page 37 lists the consequences of redundancy failure in the chassis with and without oversubscription mode.

Table 3-7 Consequences of power supply failure, depending on the oversubscription

Redundancy mode	Oversubscription mode	Consequences of redundancy failure ^a	
		Compute nodes might be throttled ^b	Chassis might power off
Non-redundant	Not available	Yes	Yes ^c
N+1	Disabled	No	No
	Enabled	Yes	No
N+N	Disabled	No	No
	Enabled	Yes	No

- a. Considering one power supply failure in non-redundant and N+1 mode and three power supplies failures in N+N mode.
- b. Compute nodes are throttled only if they require more power than what is available on the remaining power supplies.
- c. The chassis is powered off only if after throttling the compute nodes the enclosure power requirement still exceeds the power that is available on the remaining power supplies.

3.9.5 Acoustic mode

The use of Acoustic Mode 1, 2, or 3 on NeXtScale n1200 WCT is not supported. Acoustic Mode must be set to disabled.

3.9.6 Smart Redundancy mode

The use of Smart Redundancy mode on the NeXtScale n1200 WCT enclosure is not supported. Smart Redundancy mode must be set to Disabled. In data centers with three-phase power, this mode can unbalance the load on the phases, which can lead to larger neutral currents that have the potential to negate the power savings at the chassis level.

3.10 Specifications

This section describes the specifications of the NeXtScale n1200 Enclosure.

3.10.1 Physical specifications

The enclosure features the following physical specifications:

- ▶ Dimensions:
 - Height: 263 mm (10.37 in.)
 - Depth: 915 mm (36 in.)
 - Width: 447 mm (17.6 in.)
- ▶ Weight:
 - Fully configured (stand-alone): 122.4 kg (270 lb)
 - Empty chassis: approximately 38 kg (84 lb) (including water manifold section)
- ▶ Approximate heat output:
 - Minimum configuration: 341 Btu/hr (100 watts)
 - Maximum configuration: 20,471 Btu/hr (6,000 watts)

- ▶ Declared sound power level: 7.0 bels

3.10.2 Supported environment

The NeXtScale n1200 Enclosure complies with the following ASHRAE class A3 specifications.

- ▶ Power on¹:
 - Temperature: 5 °C - 40 °C (41 °F - 104 °F)²
 - Humidity, non-condensing: -12 °C dew point (10.4 °F) and 8% - 85% relative humidity
 - Maximum dew point: 24 °C (75 °F)
 - Maximum altitude: 3048 m (10,000 ft.)
 - Maximum rate of temperature change: 5 °C/hr. (41 °F/hr.)³
- ▶ Power off⁴:
 - Temperature: 5 °C - 45 °C (41 °F - 113 °F)
 - Relative humidity: 8% - 85%
 - Maximum dew point: 27 °C (80.6 °F)
- ▶ Storage (non-operating):
 - Temperature: 1 °C to 60 °C (33.8 °F - 140 °F)
 - Altitude: 3050 m (10,006 ft.)
 - Relative humidity: 5% - 80%
 - Maximum dew point: 29 °C (84.2 °F)
- ▶ Shipment (non-operating)⁵:
 - Temperature: -40°C - 60°C (-40°F - 140°F)
 - Altitude: 10700 m (35,105 ft.)
 - Relative humidity: 5% - 100%
 - Maximum dew point: 29 °C (84.2 °F)⁶
- ▶ The following specific component restrictions apply to the Intel Xeon Processor E5-2697 v3 and E5-2698A v3:
 - Temperature: 5 °C - 35 °C (41 °F - 95 °F)
 - Altitude: 0 - 950 m (3,117 ft.)

Cooling water requirements

The water that is required to initially fill the system side cooling loop must be reasonably clean and bacteria-free water (less than 100 colony forming units [CFU]/ml), such as demineralized water, reverse osmosis water, deionized water, or distilled water. The water must be filtered with an inline 50-micron filter. The water must be treated with antibiological and anticorrosion measures. The following requirements must be met:

- ▶ Minimum flow rate: 6 liters per minute
- ▶ Inlet water temperature: 18 °C - 45 °C (18 °C - 35 °C for Intel Xeon E5-2698A v3)

¹ Chassis is powered on.

² A3: Derate maximum allowable temperature 1 °C/175 m above 950 m.

³ 5 °C per hour for data centers that use tape drives and 20 °C per hour for data centers that use disk drives.

⁴ Chassis is removed from original shipping container and is installed but not in use; for example, during repair, maintenance, or upgrade.

⁵ The equipment acclimation period is 1 hour per 20 °C of temperature change from the shipping environment to the operating environment.

⁶ Condensation is acceptable, but not rain.

NeXtScale nx360 M5 WCT Compute node

NeXtScale System M5 WCT is the new generation dense water-cooled platform from System x, which follows on from the iDataPlex water-cooled system. The NeXtScale WCT system includes a dense chassis and two half-wide compute nodes on the WCT Compute Tray, all fitting in a standard rack. With WCT M5, Lenovo drives increased compute density, performance, and cooling efficiency for High Performance Computing and other workloads that require dense compute performance, such as Cloud, Grid, and Analytics.

Perhaps the most notable feature of WCT products is direct water cooling. Direct water cooling is achieved by circulating the cooling water directly through cold plates that contact the CPU thermal case, DIMMs, and other high-heat-producing components in the server.

The use of direct water cooling allows the use of “high-bin” processors, such as the 165 W Intel Xeon E5-2698A v3 processor. The nx360 M5 WCT is the only server that is available in the market that supports this processor.

This chapter includes the following topics:

- ▶ 4.1, “Overview” on page 40
- ▶ 4.2, “Specifications” on page 46
- ▶ 4.3, “Processor options” on page 48
- ▶ 4.4, “Memory options” on page 49
- ▶ 4.5, “I/O expansion options” on page 53
- ▶ 4.6, “Network adapters” on page 53
- ▶ 4.7, “Local server management” on page 56
- ▶ 4.8, “Remote server management” on page 57
- ▶ 4.9, “Supported operating systems” on page 59
- ▶ 4.10, “Physical and electrical specifications” on page 60
- ▶ 4.11, “Regulatory compliance” on page 61

4.1 Overview

The NeXtScale nx360 M5 WCT server is shown in Figure 4-1.

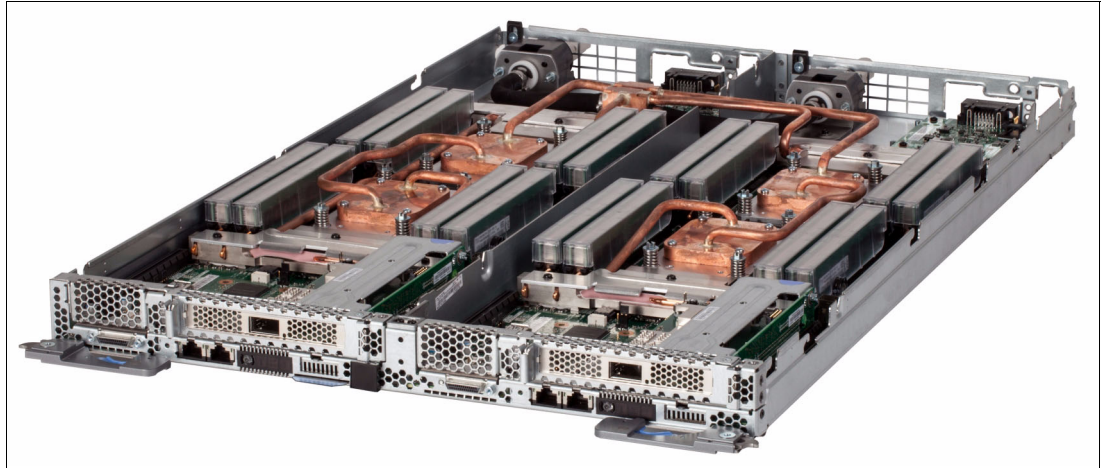


Figure 4-1 Two NeXtScale nx360 M5 WCT servers on the Compute Tray that provides water cooling

One of the main advantages of direct water cooling is the water can be relatively warm and still be effective because water conducts heat much more effectively than air. Depending on the server configuration, 85 - 90% of the heat is removed by water cooling; the rest can easily be managed by a standard computer room air conditioner. With allowable inlet temperatures for the water being as high as 45 °C (113 °F), in many cases the water can be cooled by using ambient air and chilled water and a heat exchanger is not required.

The rear view of the nx360 M5 WCT is shown in Figure 4-2.

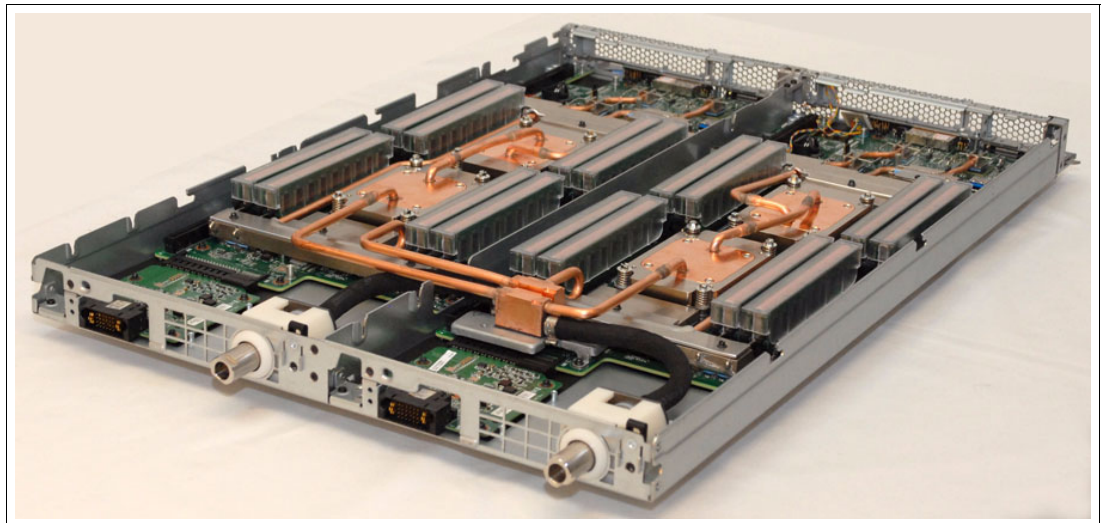


Figure 4-2 Rear view of the nx360 M5 WCT server showing the water inlet and outlet

Designed to industry-standards, NeXtScale Systems are general-purpose platforms that give customers a flexible IT infrastructure. Customized solutions can be configured to provide an application-appropriate platform with a choice of servers, networking switches, adapters, and racks.

This modular system scales and grows with data center needs to protect and maximize IT investments. Because it is optimized for standard racks, users can easily mix high-density NeXtScale server offerings and non-NeXtScale components within the same rack. The NeXtScale WCT System can be pre-configured by Lenovo, which enables users to start using it more quickly.

NeXtScale nx360 M5 WCT servers provide a dense, flexible solution with a low total cost of ownership (TCO). The half-wide, dual-socket NeXtScale nx360 M5 WCT server is designed for data centers that require high performance but are constrained by floor space. By using less physical space in the data center, the NeXtScale server enhances density and supports the Intel Xeon processor E5-2600 v3 series up to 165 W and 18-core processors, which provides more performance per server. The nx360 M5 WCT compute node contains only essential components in the base architecture to provide a cost-optimized platform.

In addition, direct water cooling means that the processors operate at a lower temperature, which enable the Intel Turbo Boost performance feature to increase processor performance.

The NeXtScale n1200 WCT Enclosure is an efficient, 6U, 12-node chassis with no built-in networking or switching capabilities; therefore, it requires no chassis-level management. Sensibly designed to provide shared, high-efficiency power and cooling for housed servers, the n1200 WCT enclosure scales with your business needs.

4.1.1 Scalability and performance

The NeXtScale n1200 WCT chassis and the NeXtScale nx360 M5 WCT server offer the following features to boost performance, improve scalability, and reduce costs:

- ▶ Up to 12 compute nodes, each with two of the latest Xeon processors, 16 DIMMs, and two PCIe slots in 6U of rack space. It is a highly dense, scalable, and price-optimized offering.
- ▶ The Intel Xeon processor E5-2600 v3 product family improves productivity by offering superior system performance processors with up to 18 cores, core speeds up to 3.2 GHz, L3 cache sizes up to 45 MB, four channels of DDR4 memory that are running at speeds up to 2133 MHz, and QPI interconnect links of up to 9.6 GTps.
- ▶ The direct water cooling capability of NeXtScale WCT enables the use of top-bin processors, including the Intel Xeon E5-2698A processor. The E5-2698A has a thermal design power (TDP) of 165 W, which is higher than most servers can support. This processor has 16 cores, 40 MB of L3 cache, and a base core frequency of 2.8 GHz. This is the fastest Intel Xeon E5-2600 V3 processor and it is available with NeXtScale WCT only.
- ▶ Two processors, up to 36 cores, and 72 threads maximize the concurrent execution of multi-threaded applications.
- ▶ Intelligent and adaptive system performance with Intel Turbo Boost Technology 2.0 allows CPU cores to run at maximum speeds during peak workloads by temporarily going beyond processor TDP.
- ▶ By providing better cooling than possible with air cooling, water-cooled servers optimize Turbo Boost 2.0 enablement.
- ▶ Intel Hyper-Threading Technology boosts performance for multi-threaded applications by enabling simultaneous multi-threading within each processor core, up to two threads per core.
- ▶ Intel Advanced Vector Extensions 2 (AVX2) doubles the number of floating-point operations per second (FLOPS) per clock cycle, enables 256 bit integer operation, and provides more instructions to improve performance for compute-intensive technical and scientific applications.

- ▶ A total of 16 DIMMs of registered 2133 MHz DDR4 ECC memory provide speed, high availability, and a memory capacity of up to 256 GB.
- ▶ Two usable PCIe slots internal to the nx360 M5 WCT, a full-height half-length x16 PCIe Gen 3 slot and a mezzanine LOM Generation 2 (ML2) slot, which is also x16 PCIe Gen 3.
- ▶ Supports new mezzanine LOM Generation 2 (ML2) cards for 40 Gb Ethernet and FDR InfiniBand that offer network performance in the smallest footprint.
- ▶ PCI Express 3.0 I/O expansion capabilities almost double (1.97x) the usable lane bandwidth compared with PCI Express 2.0.
- ▶ With Intel Integrated I/O Technology, the PCI Express 3.0 controller is integrated into the Intel Xeon processor E5 family, which reduces I/O latency and increases overall system performance.

4.1.2 Manageability and security

The following powerful systems management features simplify local and remote management of the nx360 M5 WCT:

- ▶ The server includes an Integrated Management Module II (IMM2) to monitor server availability and perform remote management.
- ▶ The first standard 1 Gbps Ethernet port can be shared between the operating system and IMM2 for remote management or can be dedicated to the IMM2. The second standard Ethernet port provides 1 Gbps Ethernet connectivity.
- ▶ IMM2 functionality can be enhanced with optional Features on Demand upgrades. The first upgrade enables a browser based interface; the second upgrade adds remote console and media functionality.
- ▶ An integrated industry-standard Unified Extensible Firmware Interface (UEFI) enables improved setup, configuration, and updates and simplifies error handling.
- ▶ Integrated Trusted Platform Module (TPM) 1.2 support enables advanced cryptographic functions, such as digital signatures and remote attestation.
- ▶ Intel Trusted Execution Technology provides enhanced security through hardware-based resistance to malicious software attacks, which allows the application to run in its own isolated space that is protected from all other software that is running on a system.
- ▶ The Intel Execute Disable Bit function can prevent certain classes of malicious buffer overflow attacks when combined with a supporting operating system.
- ▶ The n1200 WCT chassis includes drip sensors that monitor the inlet and outlet manifold quick connect couplers; leaks are reported via the Fan and Power Controller (FPC).

4.1.3 Energy efficiency

NeXtScale System offers the following energy efficiency features to save energy, reduce operational costs, increase energy availability, and contribute to a green environment:

- ▶ Water cooling eliminates power that is drawn by cooling fans in the chassis and dramatically reduces the required air movement in the server room, which also saves power.
- ▶ The processors and other microelectronics are run at lower temperatures because they are water cooled, which uses less power.
- ▶ Support is available for S3 standby power states in the processor.
- ▶ Shared 80 Plus Platinum power supplies ensure energy efficiency.

- ▶ The Intel Xeon processor E5-2600 v3 product family offers better performance over the previous generation while fitting into the similar TDP limits. These processors have their voltage regulators in the processor die (as opposed to externally); therefore, although the TDP numbers increased as compared to the v2 products, overall system power consumption is reduced.
- ▶ Intel Intelligent Power Capability can power on and off individual processor elements as needed to reduce power draw.
- ▶ Low-voltage 1.2 V DDR4 memory DIMMs use up to 20% less energy, compared to 1.35 V DDR3 DIMMs.
- ▶ There are power monitoring and power capping capabilities through the FPC in the chassis.

4.1.4 Availability and serviceability

NeXtScale n1200 WCT chassis and the nx360 M5 WCT server provide the following features to simplify serviceability and increase system uptime:

- ▶ The NeXtScale n1200 WCT chassis supports N+N and N+1 power policies for its six power supplies, which means greater system uptime.
- ▶ The power supplies are hot-swappable.
- ▶ Toolless cover removal provides easy access to upgrades and serviceable parts, such as adapters and memory.
- ▶ Predictive Failure Analysis (PFA) detects when system components (processors, memory, and PCI devices) operate outside of standard thresholds and generates proactive alerts in advance of possible failure, which increases uptime.
- ▶ The built-in IMM2 continuously monitors system parameters, triggers alerts, and performs recovering actions if there are failures to minimize downtime.
- ▶ The IMM2 offers optional remote management capability and can enable remote keyboard, video, and mouse (KVM) control and remote media for the server.
- ▶ There is a three-year customer replaceable unit and onsite limited warranty, with next business day 9x5 coverage. Optional warranty upgrades and extensions are available.

4.1.5 Locations of key components and connectors

The front of the nx360 M5 WCT server tray is shown in Figure 4-3.

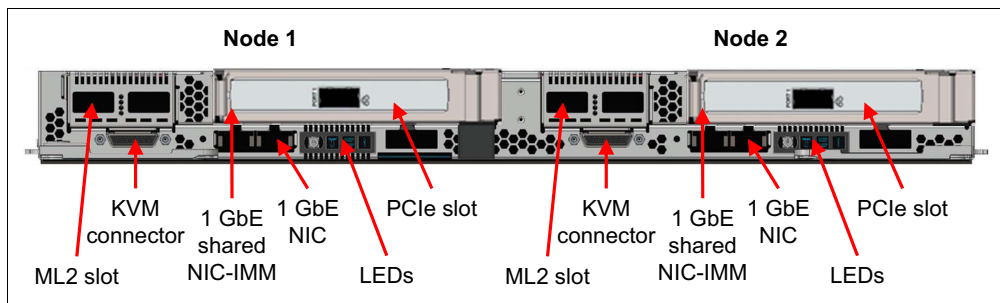


Figure 4-3 Front view of NeXtScale nx360 M5 WCT server Tray.

Figure 4-4 on page 44 shows the locations of key components inside the servers tray.

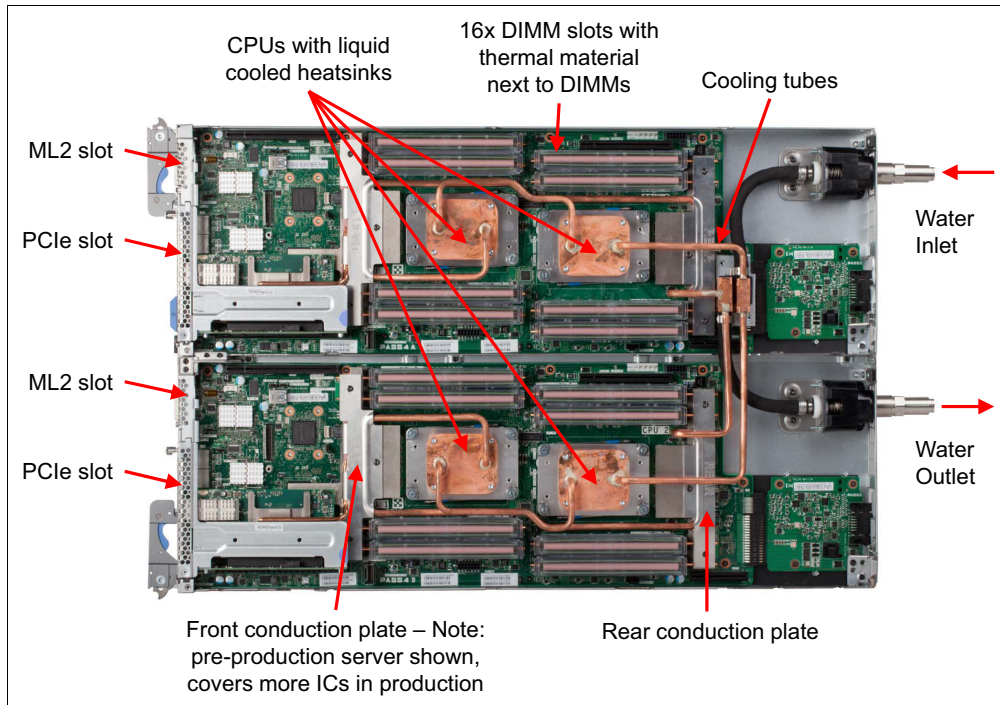


Figure 4-4 Inside view of the NeXtScale nx360 M5 WCT Tray

4.1.6 System Architecture

The NeXtScale nx360 M5 WCT compute node features the Intel E5-2600 v3 series processors. The Xeon E5-2600 v3 series processor has models with 4, 6, 8, 10, 12, 14, 16, or 18 cores per processor with up to 36 threads per socket.

The Intel Xeon E5-2600 v3 series processor (formerly known by the Intel code name *Haswell-EP*) is the third implementation of Intel's micro architecture that is based on tri-gate transistors. It uses a 22nm manufacturing process.

The processor architecture allows data on-chip to be shared through a high-speed ring that is interconnected between all processor cores, the last level cache (LLC), and the system agent. The system agent houses the memory controller and a PCI Express root complex that provides 40 PCIe 3.0 lanes.

The Integrated memory controller in each CPU supports four memory channels with three¹ DDR4 DIMMs per channel that are running at a speed that is up to 2133 MHz. Two QPI links connect the two CPU in a dual-socket installation.

The Xeon E5-2600 v3 series is available with up to 18 cores and 45 MB of last-level cache. It features an enhanced instruction set that is called Intel Advanced Vector Extensions 2 (AVX2). Intel AVX2 extends the Intel AVX with 256-bit integer instructions, floating-point fused multiply add (FMA) instructions and gather operations. Intel AVX2 doubles the number of flops per clock, doubling the core's theoretical peak floating point throughput. However, when executing some Intel AVX instructions, the processor can run at a less than rated frequency to remain within the TDP limit.

Table 4-1 on page 45 lists the improvements of the instruction set of the Intel Xeon E5-2600 v3 over previous generations.

¹ Only two DIMMs per channel implemented on the nx360 M5 WCT compute node.

Table 4-1 Instructions sets and floating point operations per cycle of Intel processors

Processor Family	Instruction Set	Single Precision Flops Per Clock	Double Precision Flops Per Clock
Intel Xeon 5500 Series (Nehalem)	SSE 4.2	8	4
Intel Xeon E5-2600 and v2 (Sandy Bridge / Ivy Bridge)	AVX	16	8
Intel Xeon E5-2600 v3 (Haswell)	AVX2	32	16

The implementation architecture includes Intel Turbo Boost Technology 2.0 and improved power management capabilities. Intel Turbo Boost Technology dynamically increases the processor's frequency as needed by using thermal and power headroom to give a burst of speed when the workload needs it, and increased energy efficiency when it does not.

As with iDataPlex servers, NeXtScale servers support S3 mode. S3 allows systems to come back into full production from low-power state much quicker than a traditional power-on. In fact, cold boot normally takes about 270 seconds; with S3, cold boot occurs in only about 45 seconds. When you know that a system is not to be used because of time of day or state of job flow, you can send it into a low power state to save power and bring it back online quickly when needed.

Table 4-2 lists the differences between the current and the previous generation of Intel's micro architecture implementations. Improvements are highlighted in gray.

Table 4-2 Comparison between Xeon E5-2600 v2 and Xeon E5-2600 v3

	Xeon E5-2600 v2 (Ivy Bridge-EP)	Xeon E5-2600 v3 (Haswell-EP)
QPI Speed (GT/s)	8.0, 7.2 and 6.4 GT/s	9.6, 8.0 and 6.4 GTps
Addressability	46 bits physical, 48 bits virtual	
Cores	Up to 12	Up to 18
Threads per socket	Up to 24 threads	Up to 36 threads
Last-level Cache (LLC)	Up to 30 MB	Up to 45MB
Intel Turbo Boost Technology	Yes	
Memory population	4 channels of up to 3 RDIMMs, 3 LRDIMMs, or 2 UDIMMs	4 channels of up to 3 DIMMs per channel and 24 DIMM slots
Maximum memory speed	Up to 1866 MHz DDR3	Up to 2133 MHz DDR4
Memory RAS features	ECC, Patrol Scrubbing, Sparring, Mirroring, Lockstep Mode, x4/x8 SDDC	
PCIe lanes	40 PCIe 3.0 lanes at 8 GTps	40 PCIe 3.0 lanes at 10GTps
TDP values (W)	130, 115, 96, 80, 70, 60, 50 W	165, 145, 135, 120, 105, 90, 85, 65, 55 W
Idle power targets (W)	10.5 W or higher 7.5 W for low-voltage SKUs	9 W or higher 7.5 W for low-voltage SKUs
Instruction Set	AVX	AVX2

Figure 4-5 shows the NeXtScale nx360 M5 WCT system board block diagram.

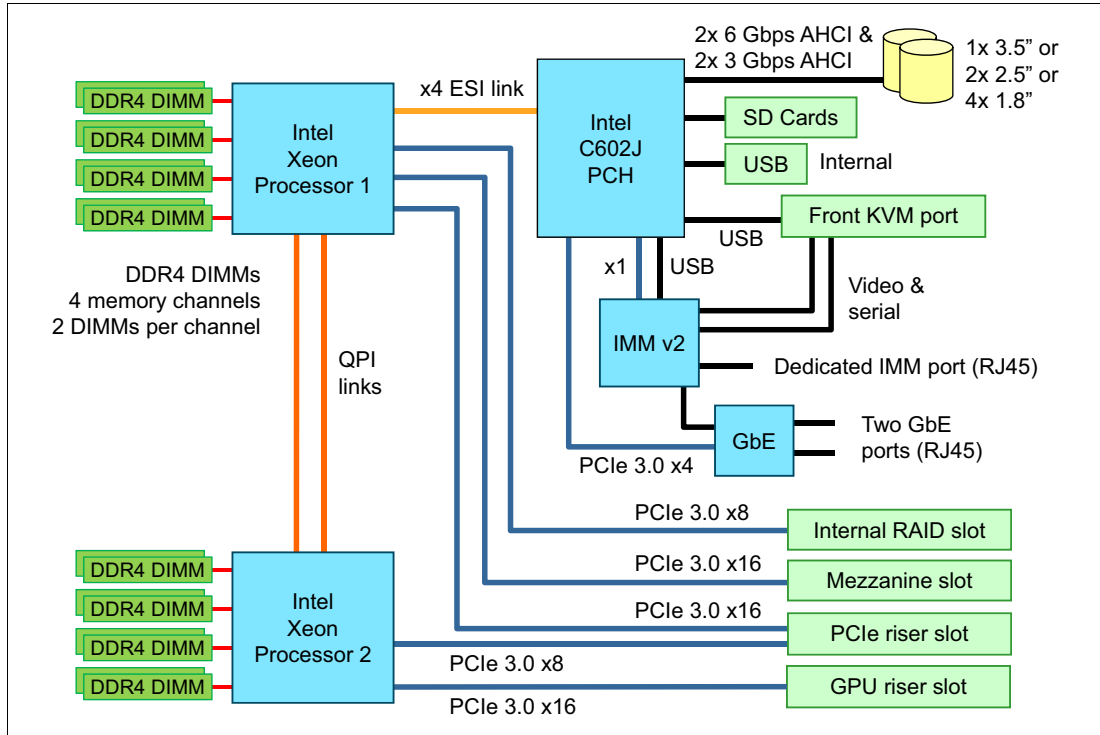


Figure 4-5 NeXtScale nx360 M5 WCT system board block diagram

4.2 Specifications

Table 4-3 lists the standard specifications of the NeXtScale nx360 M5 WCT compute.

Table 4-3 Standard specifications: nx360 M5 WCT

Components	Specification
Machine type	5467
Firmware	Lenovo signed firmware
Form factor	Standard server: Half-wide, 1U compute node; 2 per full wide 1U compute tray.
Supported chassis	NeXtScale n1200 WCT enclosure, 6U high; up to 12 compute nodes per chassis.
Processor	Two Intel Xeon Processor E5-2600 v3 series processors; QuickPath Interconnect (QPI) links speed up to 9.6 GTps. Hyper-Threading Technology and Turbo Boost Technology. Intel C612 chipset: <ul style="list-style-type: none"> ▶ 8-core processors at 3.2 GHz with 20 MB L3 cache ▶ 12-core processors up to 2.6 GHz with 30 MB L3 cache ▶ 14-core processors up to 2.6 GHz with 35 MB L3 cache ▶ 16-core processors at 2.8 GHz with 40 MB L3 cache ▶ 18-core processors at 2.3 GHz with 45 MB L3 cache
Memory	16 DIMM sockets (8 DIMMs per processor) supporting 8 GB or 16 GB DDR4 RDIMMs at 2133 MHz. Four memory channels per processor (two DIMMs per channel).

Components	Specification
Memory maximum	Up to 256 GB with 16x 16 GB RDIMMs and two processors.
Memory protection	<ul style="list-style-type: none"> ▶ 8 GB DIMM: Error Checking and Correcting (ECC), memory mirroring^a, memory rank sparing^a. ▶ 16 GB DIMM: Chipkill and ECC, memory mirroring^a and memory rank sparing^a.
Disk drive bays	Disk drives not currently supported on nx360 M5 WCT; boot from LAN (PXE boot) only.
Optical drive bays	No internal bays. Use an external USB drive.
Tape drive bays	No internal bays. Use an external USB drive.
Network interfaces	Integrated two-port Gigabit Ethernet (Broadcom BCM5717) with RJ45 connectors. One port dedicated for use by the operating system and one configurable as shared by the operating system and IMM or as dedicated to the IMM. Optionally, PCIe and mezzanine LOM Gen 2 (ML2) adapters can be added to provide more network interfaces. ML2 Ethernet adapters support shared access to the IMM.
PCI Expansion slots	One PCIe 3.0 x16 Mezzanine LOM Gen 2 (ML2) adapter slot. One PCIe 3.0 x16 full-height half-length slot.
Ports	Front of the server: KVM connector; with the addition of a console breakout cable (1 cable standard with the chassis) supplies one RS232 serial port, one VGA port, and two USB 1.1 ports for local console connectivity. Two 1 Gbps Ethernet ports with RJ45 connectors.
Cooling	Supplied by the NeXtScale n1200 WCT enclosure via water loop and power supply fans.
Power supply	Supplied by the NeXtScale n1200 WCT enclosure. Up to six hot-swap power supplies 900 W or 1300 W, depending on the chassis model. Support power policies N+N or N+1 power redundancy and non-redundant. 80 PLUS Platinum certified.
Systems Management	UEFI, Integrated Management Module II (IMM2.1) with Renesas SH7758 controller, Predictive Failure Analysis, Light Path Diagnostics, Automatic Server Restart, and ServerGuide™. Browser-based chassis management through an Ethernet port on the Fan and Power Controller at the rear of the n1200 WCT enclosure. IMM2 upgrades are available to IMM2 Standard and IMM2 Advanced for web GUI and remote presence features.
Video	Matrox G200eR2 video core with 16 MB DDR3 video memory that is integrated into the IMM2. Maximum resolution is 1600 x 1200 with 16M colors (32 bpp) at 75 Hz, or 1680 x 1050 with 16M colors at 60 Hz.
Security features	Power-on password, administrator's password, and Trusted Platform Module 1.2.
Operating systems supported	Microsoft Windows Server 2012 and 2012 R2, SUSE Linux Enterprise Server 11 SP3 and 12, Red Hat Enterprise Linux 6 U5 and 7, VMware vSphere 5.1 U2, 5.5 U2 and 6.0.
Limited warranty	Three-year customer-replaceable unit and onsite limited warranty with 9x5/NBD.
Service and support	Optional service upgrades are available through Lenovo Services: 4-hour or 2-hour response time, 8-hour fix time, 1 year or 2 year warranty extension, remote technical support for hardware and some Lenovo, and OEM software.

Components	Specification
Dimensions	Compute tray width: 432 mm (17 in.), height: 41.0 mm (1.6 in.), depth: 658.8 mm (25.9 in.).
Weight	Compute tray (2 servers): 13.3 kg (29.3 lb)

a. planned for 3Q/2015

The nx360 M5 WCT servers are shipped with the following items:

- ▶ Statement of Limited Warranty
- ▶ Important Notices
- ▶ Documentation CD that contains the Installation and Service Guide

4.3 Processor options

The nx360 M5 WCT supports the processor options that are listed in Table 4-4. Two processors per node need to be selected. There is no support for single processor per node.

Table 4-4 Processor support

Feature code	Intel Xeon processors ^a
AS4T	Intel Xeon Processor E5-2667 v3 8C 3.2GHz 20MB 2133MHz 135W
A5L9	Intel Xeon Processor E5-2670 v3 12C 2.3GHz 30MB 2133MHz 120W
A5L8	Intel Xeon Processor E5-2680 v3 12C 2.5GHz 30MB 2133MHz 120W
A5L7	Intel Xeon Processor E5-2690 v3 12C 2.6GHz 30MB 2133MHz 135W
A5V9	Intel Xeon Processor E5-2683 v3 14C 2.0GHz 35MB 2133MHz 120W
A5L6	Intel Xeon Processor E5-2695 v3 14C 2.3GHz 35MB 2133MHz 120W
A5L5	Intel Xeon Processor E5-2697 v3 14C 2.6GHz 35MB 2133MHz 145W
AS4S	Intel Xeon Processor E5-2698 v3 16C 2.3GHz 40MB 2133MHz 135W
A5VA	Intel Xeon Processor E5-2698A v3 16C 2.8GHz 40MB 2133MHz 165W
AS4U	Intel Xeon Processor E5-2699 v3 18C 2.3GHz 45MB 2133MHz 145W

a. Processor detail: Model, core count, core speed, L3 cache, memory speed, and TDP power.

Floating point performance: The number of sockets and the processor option that are selected determine the theoretical double precision floating point peak performance, as shown in the following example:

$$\#sockets \times \#cores \text{ per processor} \times freq \times 16 \text{ flops per cycle} = \#Gflops$$

An nx360 M5 compute node with dual socket E5-2698 v3 series 16-core that operates at 2.3 GHz has the following peak performance:

$$2 \times 16 \times 2.3 \times 16 = 1177.6 \text{ Gflops}$$

However, because of the higher power consumption that is generated by AVX instructions, the processor frequency typically^a is reduced by 0.4 GHz to stay within the TDP limit. As a result, a more realistic peak performance formula is used:

$$\#sockets \times \#cores \text{ per processor} \times (freq - 0.4) \times 16 \text{ flops per cycle} = \#Gflops$$

Therefore, an nx360 M5 compute node with dual socket E5-2698 v3 series 16-core that operates at 2.3 GHz has the following peak performance:

$$2 \times 16 \times (2.3 - 0.4) \times 16 = 972.8 \text{ Gflops}$$

By using LINPACK benchmark, single node sustained performance is approximately 92% of the peak, which correspond to approximately 895 Gflops with the node that is used in our example.

a. The processor frequency reduction is not identical for each SKU.

Memory performance: The processors with eight cores and less have a single memory controller. The processors with 10 cores and more have two memory controllers. As a result, it is recommended that a 10+ core processor is selected to get maximum performance on memory-intensive application.

4.4 Memory options

TruDDR4™ Memory from Lenovo uses the highest quality components that are sourced from Tier 1 DRAM suppliers and only memory that meets the strict requirements of Lenovo is selected. It is compatibility tested and tuned on every System x server to maximize performance and reliability. TruDDR4 Memory has a unique signature that is programmed into the DIMM that enables System x servers to verify whether the memory installed is qualified or supported by Lenovo. Because TruDDR4 Memory is authenticated, certain extended memory performance features can be enabled to extend performance over industry standards.

The NeXtScale nx360 M5 WCT supports up to 16 TruDDR4 Memory DIMMs. Each processor has four memory channels, and there are two DIMMs per memory channel (2 DPC). RDIMMs are supported.

The supported memory that is available for the nx360 M5 WCT server is listed in Table 4-5 on page 50.

Table 4-5 Memory options

Feature code	Description	Maximum supported
A5B8	8GB TruDDR4 Memory (2Rx8, 1.2V) PC4-17000 CL15 2133MHz LP RDIMM	16
A5B7	16GB TruDDR4 Memory (2Rx4, 1.2V) PC4-17000 CL15 2133MHz LP RDIMM	16

Memory operates at 2133 MHz, with 1 or 2 DIMMs per channel (DPC), in the NeXtScale nx360 M5 WCT.

For optimal performance, use DIMMs in multiple of eight (four per processor) to make use of the four memory channels of the processor.

The following memory protection technologies are supported:

- ▶ ECC
- ▶ Chipkill (only available with x4 memory; for example, 1Rx4, 2Rx4, and 4Rx4); therefore, available on the 16GB DIMMs only
- ▶ Memory mirroring (planned for 3Q/2015)
- ▶ Memory sparing (planned for 3Q/2015)

If memory mirroring is used, DIMMs must be installed in pairs (minimum of one pair per CPU), and both DIMMs in a pair must be identical in type and size.

If memory rank sparing is used, a minimum of two DIMMs must be installed per populated channel (the DIMMs do not need to be identical). In rank sparing mode, one rank of a DIMM in each populated channel is reserved as spare memory.

DIMM installation order

The NeXtScale nx360 M5 WCT boots with only one memory DIMM installed per processor. However, the suggested memory configuration is to balance the memory across all the memory channels on each processor to use the available memory bandwidth.

The locations of the DIMM sockets relative to the processors are shown in Figure 4-6 on page 51.

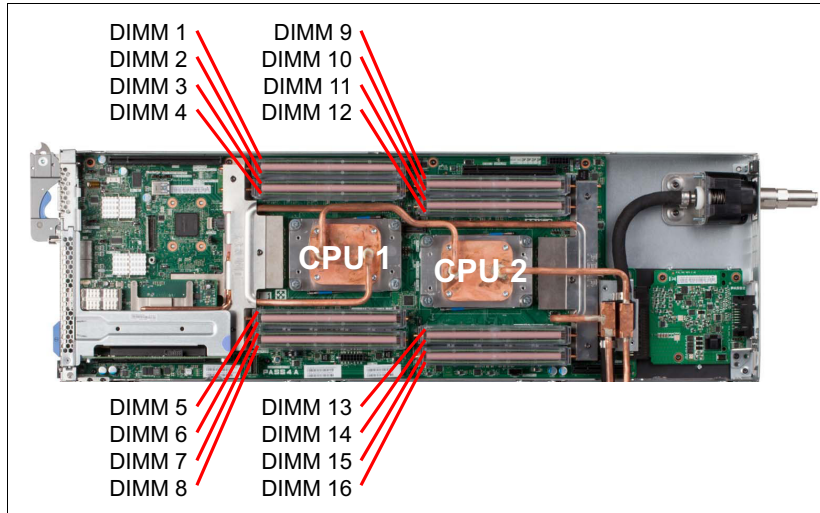


Figure 4-6 DIMM locations

Memory DIMM installation: Independent channel mode

Table 4-6 shows DIMM installation order. A minimum of two memory DIMMs (one for each processor) are required.

Table 4-6 Memory population table (independent channel mode)

Number of DIMMs	Processor 1								Processor 2							
	DIMM 1	DIMM 2	DIMM 3	DIMM 4	DIMM 5	DIMM 6	DIMM 7	DIMM 8	DIMM 9	DIMM 10	DIMM 11	DIMM 12	DIMM 13	DIMM 14	DIMM 15	DIMM 16
2								x	x							
3	x								x	x						
4	x							x	x							x
5	x					x		x	x							x
6	x					x		x	x		x					x
7	x		x			x		x	x		x					x
8	x		x			x		x	x		x			x		x
9	x		x			x	x	x	x		x			x		x
10	x		x			x	x	x	x	x	x			x		x
11	x	x	x			x	x	x	x	x	x			x		x
12	x	x	x			x	x	x	x	x	x			x	x	x
13	x	x	x		x	x	x	x	x	x	x			x	x	x
14	x	x	x		x	x	x	x	x	x	x	x		x	x	x
15	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x
16	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Memory DIMM installation: Mirrored-channel mode

In mirrored channel mode, the channels are paired and both channels in a pair store the same data. Because of the redundancy, the effective memory capacity of the compute node is half the installed memory capacity.

The pair of DIMMs that are installed in each channel must be identical in capacity, type, and rank count.

Table 4-7 shows DIMM installation order. A minimum of four memory DIMMs (two for each processor) are required.

Table 4-7 Memory population table (mirrored channel mode)

	Processor 1								Processor 2							
Number of DIMMs	DIMM 1	DIMM 2	DIMM 3	DIMM 4	DIMM 5	DIMM 6	DIMM 7	DIMM 8	DIMM 9	DIMM 10	DIMM 11	DIMM 12	DIMM 13	DIMM 14	DIMM 15	DIMM 16
4						x		x	x		x					
6	x		x			x		x	x		x					
8	x		x			x		x	x		x			x		x
10	x		x		x	x	x	x	x		x			x		x
12	x		x		x	x	x	x	x	x	x	x		x		x
14	x	x	x	x	x	x	x	x	x	x	x	x		x		x
16	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Memory DIMM installation: Rank-sparing mode

In rank sparing mode, a minimum of two DIMMs must be installed per populated channel (the DIMMs do not need to be identical). In rank sparing mode, one rank of a DIMM in each populated channel is reserved as spare memory.

Table 4-8 shows DIMM installation order. A minimum of four DIMMs (two for each processor) are required.

Table 4-8 Memory population table (rank sparing mode)

	Processor 1								Processor 2							
Number of DIMMs	DIMM 1	DIMM 2	DIMM 3	DIMM 4	DIMM 5	DIMM 6	DIMM 7	DIMM 8	DIMM 9	DIMM 10	DIMM 11	DIMM 12	DIMM 13	DIMM 14	DIMM 15	DIMM 16
4							x	x	x	x						
6	x	x					x	x	x	x						
8	x	x					x	x	x	x					x	x
10	x	x			x	x	x	x	x	x					x	x
12	x	x			x	x	x	x	x	x	x	x			x	x
14	x	x	x	x	x	x	x	x	x	x	x	x			x	x
16	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

4.5 I/O expansion options

The nx360 M5 WCT offers the following I/O expansion options:

- ▶ One PCIe 3.0 x16 ML2 adapter slot (optional, front accessible)
- ▶ One PCIe 3.0 x16 full-height half-length slot (optional, front accessible)

Note: Each slot requires a riser card, as listed in Table 4-9 on page 53.

The front accessible slots are shown in Figure 4-7.

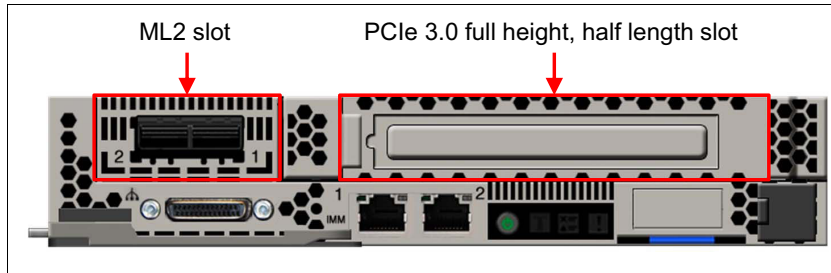


Figure 4-7 Optional front accessible PCIe slots

The ordering information for optional riser cards for the two slots is listed in Table 4-9.

Table 4-9 Riser card options

Feature code	Description	Maximum supported
A5JV	nx360 M5 ML2 Riser	1
AS9R	nx360 M5 WCT Compute Node Front Riser	1

4.6 Network adapters

The nx360 M5 WCT provides two Gigabit Ethernet ports standard, with the following features:

- ▶ Broadcom BCM5717 Gigabit Ethernet controller
- ▶ TCP/IP Offload Engine (TOE) support
- ▶ Wake on LAN support
- ▶ Receive side Scaling (RSS) and Transmit side Scaling (TSS) support
- ▶ MSI and MSI-X capability (up to five MSI-X vectors)
- ▶ VLAN tag support (IEEE 802.1Q)
- ▶ Layer 2 priority encoding (IEEE 802.1p)
- ▶ Link aggregation (IEEE 802.3ad)
- ▶ Full-duplex flow control (IEEE 802.3x)
- ▶ IP, TCP, and UDP checksum offload (hardware based) on Tx/Rx over IPv4/IPv6
- ▶ Hardware TCP segmentation offload over IPv4/IPv6
- ▶ Jumbo frame support
- ▶ NIC Teaming (Load Balancing and Failover)

- ▶ One port that is shared with IMM2 by using the Network Controller-Sideband Interface (NC-SI)

The nx360 M5 WCT server supports a Mezzanine LOM Generation 2 (ML2) adapter with a dedicated slot at the front of the server, as shown in Figure 4-7 on page 53. The use of an ML2 adapter also requires the installation of the ML2 riser card. The riser card and supported adapter are listed in Table 4-10.

Table 4-10 Mezzanine LOM Gen 2 (ML2) adapters

Feature code	Description
A5JV	nx360 M5 ML2 Riser
A5KL	Mellanox ConnectX-3 Pro 40GbE / FDR IB VPI ML2 for nx360 M5 WCT

The Mellanox ConnectX-3 Pro 40GbE / FDR IB VPI ML2 adapter has the following features:

- ▶ Two QSFP ports that support FDR-14 InfiniBand or 40 Gb Ethernet
- ▶ Mezzanine LOM Generation 2 (ML2) form factor
- ▶ Support for InfiniBand FDR speeds of up to 56 Gbps (auto-negotiation FDR-10, QDR, DDR, and SDR)
- ▶ Support for Virtual Protocol Interconnect (VPI), which enables one adapter for InfiniBand and 10/40 Gb Ethernet. Supports the following configurations:
 - 2 ports InfiniBand
 - 2 ports Ethernet
 - 1 port InfiniBand and 1 port Ethernet
- ▶ SR-IOV support; 16 virtual functions that are supported by KVM and Hyper-V (OS-dependent) up to a maximum of 127 virtual functions that are supported by the adapter
- ▶ Enables Low Latency RDMA over 40 Gb Ethernet (supported by non-virtualized and SR-IOV enabled virtualized servers); latency as low as 1 μ s
- ▶ Microsoft VMQ/VMware NetQueue support
- ▶ Sub 1 μ s InfiniBand MPI ping latency
- ▶ Support for QSFP to SFP+ for 10 Gb Ethernet support
- ▶ Traffic steering across multiple cores
- ▶ Legacy and UEFI PXE network boot support (Ethernet mode only)
- ▶ Offers NVGRE hardware offloads
- ▶ Offers VXLAN hardware offloads
- ▶ Offers access to IMM2 by using the Network Controller-Sideband Interface (NC-SI)

The supported network adapters for use in the standard full-height half-length PCIe slot are listed in Table 4-11. The use of an adapter in this slot also requires the installation of the PCIe riser card.

Table 4-11 PCIe Network adapters

Feature code	Description
AS9R	nx360 M5 WCT Compute Node Front Riser
AS9Q	Intel QLE7340 for WCT Single-port QDR IB PCIe 2.0 x8 HCA

Feature code	Description
AS4V	Mellanox Single-Port Connect-IB PCIe x16 Adapter for nx360 M5 WCT
AS96	Mellanox DWC Connect-IB FDR IB Single-port PCIe 3.0 x16 HCA
ASWQ	Mellanox ConnectX-4 EDR IB VPI Single-port x16 PCIe 3.0 HCA
ASWR	Mellanox ConnectX-4 EDR IB VPI Dual-port x16 PCIe 3.0 HCA

Although there are two different feature codes, AS4V and AS96 are references to the same Connect-IB adapter. It is recommended to select AS96 in the x-config configurator.

The Mellanox Connect-IB PCIe adapter includes the following features:

- ▶ One QSFP port
- ▶ PCI Express (PCIe) 3.0 x8
- ▶ FDR 56 Gbps
- ▶ Greater than 130M messages/sec
- ▶ 16 million I/O channels
- ▶ 256 to 4Kbyte MTU, 1 GB messages
- ▶ Protocol support: OpenMPI, IBM PE, Intel MPI, OSU MPI (MVAPICH/2), Platforms MPI, UPC, Mellanox SHMEM, TCP/UDP, IPoIB, RDS, SRP, iSER, NFS RDMA, SMB Direct, uDAPL
- ▶ Flexboot technology; remote boot over InfiniBand
- ▶ Multiple queues per virtual machine
- ▶ 16 physical functions, 256 virtual functions
- ▶ VMware NetQueue support
- ▶ Enhanced QoS for vNICs and vHCAs

The QLogic QLE7340 HCA has the following summary of features and specifications:

- ▶ One QDR 4X InfiniBand external port (QSFP)
- ▶ PCI Express Gen 2 x8 host bus interface
- ▶ 40 Gbps InfiniBand interface (40/20/10 Gbps auto-negotiation)
- ▶ 3400 MBps unidirectional throughput
- ▶ Approximately 30 million messages processed per second (non-coalesced)
- ▶ Approximately 1.0 microsecond latency that remains low as the fabric is scaled
- ▶ Multiple virtual lanes (VLs) for unique quality of service (QoS) levels per lane over the same physical port
- ▶ TrueScale architecture, with MSI-X interrupt handling, optimized for multi-core compute nodes
- ▶ Operates without external memory
- ▶ Optional data scrambling in InfiniBand link
- ▶ Complies with InfiniBand Trade Association 1.2 standard
- ▶ Supports OpenFabrics Alliance (OFED) software distributions

4.7 Local server management

The nx360 M5 WCT provides local console access through the KVM connector at the front of the server. A console breakout cable is used with this connector, which provides a VGA port, two USB ports, and a DB9 serial port. The cable is shown in Figure 4-8.



Figure 4-8 Console breakout cable

One console breakout cable is shipped with the NeXtScale n1200 WCT enclosure. More cables can be ordered per Table 4-12.

Table 4-12 Console breakout cable

Feature code	Description	Maximums supported
A4AK	Console breakout cable (KVM Dongle cable)	1

To aid with problem determination, the server includes light path diagnostics, which is a set of LEDs on the front of the server and inside the server that show you which component is failing. The LEDs are shown in Figure 4-9.

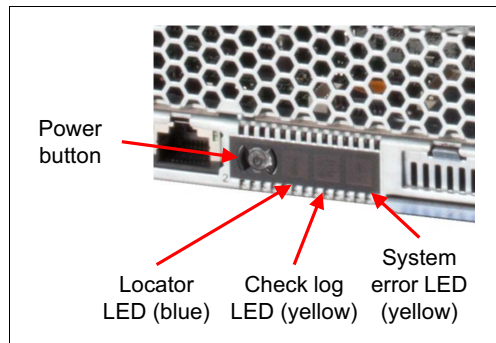


Figure 4-9 Power button and system LEDs

When an error occurs, the system error LED lights up. Review the logs through the interface of the IMMv2 (see 4.8, “Remote server management” on page 57). If needed, power off the server and remove it from the enclosure. Then, press and hold the light path button on the system board (see location on Figure 4-4 on page 44) to activate the system board LEDs. The LED next to the failed component lights up.

4.8 Remote server management

Each NeXtScale nx360 M5 WCT compute node has an IMM 2.1 onboard and uses the UEFI.

The IMM provides advanced service-processor control, monitoring, and an alerting function. If an environmental condition exceeds a threshold or if a system component fails, the IMM lights LEDs to help you diagnose the problem, records the error in the event log, and alerts you about the problem. Optionally, the IMM also provides a virtual presence capability for remote server management capabilities. The IMM provides remote server management through the following industry-standard interfaces:

- ▶ Intelligent Platform Management Interface (IPMI) version 2.0
- ▶ Simple Network Management Protocol (SNMP) version 3.0
- ▶ Common Information Model (CIM)
- ▶ Web browser

The IMM2.1 also provides the following remote server management capabilities through the **ipmitool** management utility program:

- ▶ Command-line interface (IPMI Shell)

The command-line interface provides direct access to server management functions through the IPMI 2.0 protocol. Use the command-line interface to issue commands to control the server power, view system information, and identify the server. You can also save one or more commands as a text file and run the file as a script.

- ▶ Serial over LAN

Establish a Serial over LAN (SOL) connection to manage servers from a remote location. You can remotely view and change the UEFI settings, restart the server, identify the server, and perform other management functions. Any standard Telnet client application can access the SOL connection.

The NeXtScale nx360 M5 WCT server includes IMM Basic and can be upgraded to IMM Standard and IMM Advanced with FoD licenses.

IMM Basic has the following features:

- ▶ Industry-standard interfaces and protocols
- ▶ Intelligent Platform Management Interface (IPMI) Version 2.0
- ▶ Common Information Model (CIM)
- ▶ Advanced Predictive Failure Analysis (PFA) support
- ▶ Continuous health monitoring
- ▶ Shared Ethernet connection
- ▶ Domain Name System (DNS) server support
- ▶ Dynamic Host Configuration Protocol (DHCP) support
- ▶ Embedded Dynamic System Analysis™ (DSA)
- ▶ LAN over USB for in-band communications to the IMM
- ▶ Serial over LAN (SOL)
- ▶ Remote power control
- ▶ Server console serial redirection

IMM Standard (as enabled by using the FoD software license key that uses FC A1MK) has the following features to the IMM Basic features:

- ▶ Remote access through a secure web console
- ▶ Access to server vital product data (VPD)
- ▶ Automatic notification and alerts

- ▶ Continuous health monitoring and control
- ▶ Email alerts
- ▶ Syslog logging support
- ▶ Enhanced user authority levels
- ▶ Event logs that are time stamped, saved on the IMM, and that can be attached to email alerts
- ▶ Operating system watchdogs
- ▶ Remote configuration through Advanced Settings Utility™ (ASU)
- ▶ Remote firmware updating
- ▶ User authentication by using a secure connection to a Lightweight Directory Access Protocol (LDAP) server

IMM Advanced (as enabled by using the FoD software license key that uses FC A1ML) adds the following features to IMM Standard:

- ▶ Remotely viewing video with graphics resolutions up to 1600 x 1200 at 75 Hz with up to 23 bits per pixel color depths, regardless of the system state
- ▶ Remotely accessing the server by using the keyboard and mouse from a remote client
- ▶ Mapping the CD or DVD drive, diskette drive, and USB flash drive on a remote client, and mapping ISO and diskette image files as virtual drives that are available for use by the server
- ▶ Uploading a diskette image to the IMM memory and mapping it to the server as a virtual drive

The blue-screen capture feature captures the video display contents before the IMM restarts the server when the IMM detects an operating system hang condition. A system administrator can use the blue-screen capture to assist in determining the cause of the hang condition.

Table 4-13 lists the remote management options.

Note: The IMM Advanced upgrade requires the IMM2 Standard upgrade.

Table 4-13 Remote management options

Feature codes	Description	Maximum supported
A1MK	Integrated Management Module Standard Upgrade	1
A1ML	Integrated Management Module Advanced Upgrade (requires Standard Upgrade, A1MK)	1

The nx360 M5 WCT provides two Ethernet ports standard, one of which (port 1) is configured in UEFI by default to be shared between the operating system and the IMM2. In shared mode, this port enables you to connect remotely to the IMM2 to perform systems management functions. Figure 4-10 on page 59 shows the location of the ports.

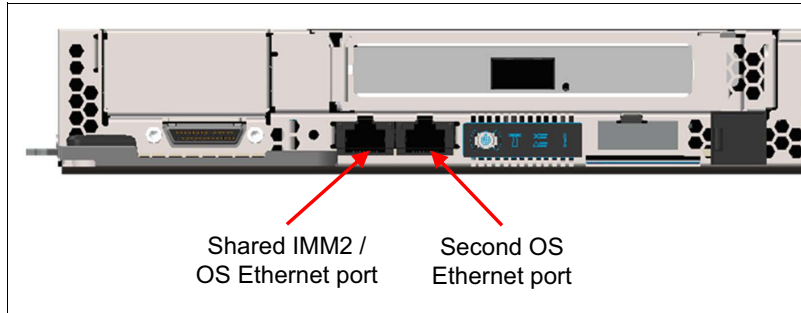


Figure 4-10 IMM port (some front panel perforations are not present on the production model)

UEFI-compliant server firmware

System x Server Firmware (server firmware) offers several features, including UEFI 2.1 compliance; Active Energy Manager technology; enhanced reliability, availability, and serviceability (RAS) capabilities; and basic input/output system (BIOS) compatibility support. UEFI replaces the BIOS and defines a standard interface between the operating system, platform firmware, and external devices. UEFI-compliant System x servers can boot UEFI-compliant operating systems, BIOS-based operating systems, and BIOS-based adapters and UEFI-compliant adapters.

For more information about the IMM, see the following User's Guide website:

<http://ibm.com/support/entry/portal/docdisplay?lnocid=migr-5086346>

4.9 Supported operating systems

The nx360 M5 WCT server supports the following operating systems:

- ▶ Microsoft Windows Server 2012 R2
- ▶ Microsoft Windows Server 2012
- ▶ Red Hat Enterprise Linux 7
- ▶ Red Hat Enterprise Linux 6 Server x64 Edition, U5
- ▶ SUSE Enterprise Linux Server (SLES) 12
- ▶ SUSE Linux Enterprise Server 12 with XEN
- ▶ SUSE Linux Enterprise Server 11 for AMD64/EM64T, SP3
- ▶ SUSE Linux Enterprise Server 11 with Xen for AMD64/EM64T, SP3
- ▶ VMware vSphere 6.0 (ESXi)
- ▶ VMware vSphere 5.5 (ESXi), U2
- ▶ VMware vSphere 5.1 (ESXi), U2

For more information about the specific versions and service levels that are supported and any other prerequisites, see the following ServerProven® website:

<http://www.ibm.com/systems/info/x86servers/serverproven/compat/us/nos/somatrix.shtml>

4.10 Physical and electrical specifications

The NeXtScale nx360 M5 WCT compute tray features the following dimensions:

- ▶ Width: 432 mm (17.0 in.)
- ▶ Height: 41.0 mm (1.6 in.)
- ▶ Depth: 658.8 mm (25.9 in.)
- ▶ Weight (two servers on each compute tray): 13.3 kg (29.3 lb)

Supported environment

The NeXtScale nx360 M5 WCT compute node complies with ASHRAE class A3 specifications. The node supports the following environments when it is powered on:

- ▶ Temperature: 5 °C - 40 °C (41 °F - 104 °F) up to 950 m (3,117 ft.)
- ▶ Above 950m, de-rated maximum air temperature 1 °C/175 m
- ▶ Humidity, non-condensing: -12 °C dew point (10.4 °F) and 8% - 85% relative humidity
- ▶ Maximum dew point: 24 °C (75 °F)
- ▶ Maximum altitude: 3050 m (10,000 ft.) and 5 °C - 28 °C (41 °F - 82 °F)

The minimum humidity level for class A3 is the higher (more moisture) of the -12 °C (10.4 °F) dew point and the 8% relative humidity. These intersect at approximately 25 °C (77 °F). Below this intersection (approximately 25°C or 77°F), the dew point (-12 °C or 10.4 °F) represents the minimum moisture level, while above it relative humidity (8%) is the minimum.

Moisture levels lower than 0.5 °C (32 °F) dew point, but not lower -10 °C (14 °F) dew point or 8% relative humidity, can be accepted if appropriate control measures are implemented to limit the generation of static electricity on personnel and equipment in the data center. All personnel and mobile furnishings and equipment must be connected to ground through an appropriate static control system.

The following items are considered the minimum requirements:

- ▶ All conductive flooring, conductive footwear on all personnel that go into the data center, and all mobile furnishings and equipment must be made of conductive or static dissipative materials.
- ▶ During maintenance on any hardware, a properly functioning wrist strap must be used by any personnel who come into contact with IT equipment.

To adhere to ASHRAE Class A3, Temperature: 36 °C - 40 °C (96.8 °F - 104 °F) with relaxed support, consider the following points:

- ▶ A support cloud-like workload with no performance degradation is acceptable (Turbo-Off).
- ▶ Under no circumstance can any combination of worst case workload and configuration result in system shutdown or design exposure at 40 °C (104 °F).
- ▶ The worst case workload (such as Linpack and Turbo-On) might have performance degradation.

The Intel Xeon Processor E5-2697 v3 and E5-2698A v3 feature the following specific component restrictions:

- ▶ Temperature: 5 °C - 35 °C (41 °F - 95 °F)
- ▶ Altitude: 0 - 950 m (3,117 ft.).

4.11 Regulatory compliance

The server conforms to the following international standards:

- ▶ FCC - Verified to comply with Part 15 of the FCC Rules, Class A
- ▶ Canada ICES-003, issue 5, Class A
- ▶ UL/IEC 60950-1
- ▶ CSA C22.2 No. 60950-1
- ▶ NOM-019
- ▶ Argentina IEC60950-1
- ▶ Japan VCCI, Class A
- ▶ IEC 60950-1 (CB Certificate and CB Test Report)
- ▶ China CCC GB4943.1, GB9254, Class A, and GB17625.1
- ▶ Taiwan BSMI CNS13438, Class A; CNS14336-1
- ▶ Australia/New Zealand AS/NZS CISPR 22, Class A; AS/NZS 60950.1
- ▶ Korea KN22, Class A, KN24
- ▶ Russia/GOST ME01, IEC-60950-1, GOST R 51318.22, and GOST R 51318.24,
- ▶ GOST R 51317.3.2, GOST R 51317.3.3
- ▶ IEC 60950-1 (CB Certificate and CB Test Report)
- ▶ CE Mark (EN55022 Class A, EN60950-1, EN55024, and EN61000-3-2,
- ▶ EN61000-3-3)
- ▶ CISPR 22, Class A
- ▶ TUV-GS (EN60950-1/IEC 60950-1, and EK1-ITB2000)

Rack planning

A NeXtScale Water Cool Technology (WCT) System configuration can consist of many chassis, nodes, switches, cables, and racks. In many cases, it is relevant for planning purposes to think of a system in terms of racks or multiple racks.

In this chapter, we describe best practices for configuring the individual racks. After the rack level design is established, we provide some guidance for designing multiple rack solutions.

This chapter includes the following topics:

- ▶ 5.1, “Power planning” on page 64
- ▶ 5.2, “Cooling” on page 69
- ▶ 5.3, “Density” on page 75
- ▶ 5.4, “Racks” on page 75
- ▶ 5.5, “Cable management” on page 87
- ▶ 5.6, “Rear Door Heat eXchanger” on page 89
- ▶ 5.7, “Top-of-rack switches” on page 93
- ▶ 5.8, “Rack-level networking: Sample configurations” on page 95

5.1 Power planning

In this section, we provide example best practices for configuring power connections and power distribution to meet electrical safety standards while providing a wanted level of redundancy and cooling for racks that contain NeXtScale WCT System chassis and servers.

For more information, see *NeXtScale System Power Requirements Guide*, which is available at this website:

<http://ibm.com/support/entry/portal/docdisplay?lndocid=LNVO-POWINF>

NeXtScale WCT System offers N+1 and N+N power supply redundancy policies at the chassis level. To minimize system cost, it is expected that N+1 or non-redundant power configurations are used; therefore, this use is how the power system was optimized.

When you are planning your power sources for a NeXtScale WCT System rack configuration, consider the following important points:

- ▶ Input voltage

Power supply input is single phase 100 - 120 V or 200 - 240 V alternating current (VAC). For NeXtScale WCT System servers in production environments, 200 - 240 VAC is preferred because it reduces the electrical current requirement. Another consideration is that the power supplies produce less DC output at the lower range.

Restrictions at 110 V: The 900 W CFF power supply is limited to 600 W capacity when operated at low-range voltage (100 - 127 V).

The 1300 W power supply does not support low-range voltages (100 - 127 V).

- ▶ Power distribution unit (PDU) input: single-phase or three-phase power

PDUs can be fed with single-phase or three-phase power. Three-phase power provides more usable power to each PDU and to the equipment. The Lenovo three-phase PDUs separate the phases, which provide single-phase power to the power supplies. The NeXtScale n1200 WCT Enclosure's six power supplies evenly balance the load on three-phase power systems.

- ▶ Single or dual power feed (N+N) to the rack

With a dual-power feed, half of the rack PDUs are powered by one feed and the other half is powered by the second feed, which provides redundant power to the rack.

N+N designs can provide resilience if a power feed must be powered down for maintenance, or if there is a power disruption. However, careful power planning must be done to assure there is adequate power for the NeXtScale systems to keep running on only one power feed.

- ▶ PDU control

PDUs can be switched and monitored, monitored only, or non-monitored. It might be of interest to monitor the power usage at the outlet level of the PDU. More power savings and data center control can be gained with PDUs on which the outlets can be turned on and off.

Single-phase power: In some countries, single-phase power can also be used in such configurations. For more information, see *NeXtScale System Power Requirements Guide*, which is available at this website:

<http://ibm.com/support/entry/portal/docdisplay?lnocid=LNVO-POWINF>

5.1.1 NeXtScale WCT Rack Power Reference Examples

Table 5-1 lists the typical steady-state rack power for common NeXtScale WCT configurations. The power information represents a full rack with 36x dual compute trays and six common switches.

Table 5-1 Typical steady-state power for NeXtScale WCT System rack configurations

Compute Tray Configuration (dual compute nodes)	Single Rack Steady-State Power (w/Linpack, Turbo ON/OFF)	Single Rack Steady-State kVA
4x E5-2698 v3 (165 W), 16x 16 GB DDR4 DIMMs, 2x NICs	33.5 kW	34.2 kVA
4x E5-2697 v3 (145 W), 16x 16 GB DDR4 DIMMs, 2x NICs	30.7 kW	31.3 kVA
4x E5-2690 v3 (135 W), 16x 16 GB DDR4 DIMMs, 2x NICs	29.4 kW	30 kVA
4 x E5-2680 v3 (120 W), 16x 16 GB DDR4 DIMMs, 2x NICs	27.3 kW	27.9 kW

5.1.2 Examples

Data center power can be supplied from a single power feed, which can be protected by a facility UPS. The power cabling that is shown in Figure 5-1 uses three PDUs. The PDUs can be supplied by using a 60 A, 200 - 240 V three-phase source or a 32 A, 380 - 415 V three-phase source. The colors (red, blue, green) show the phases as separated by the PDUs.

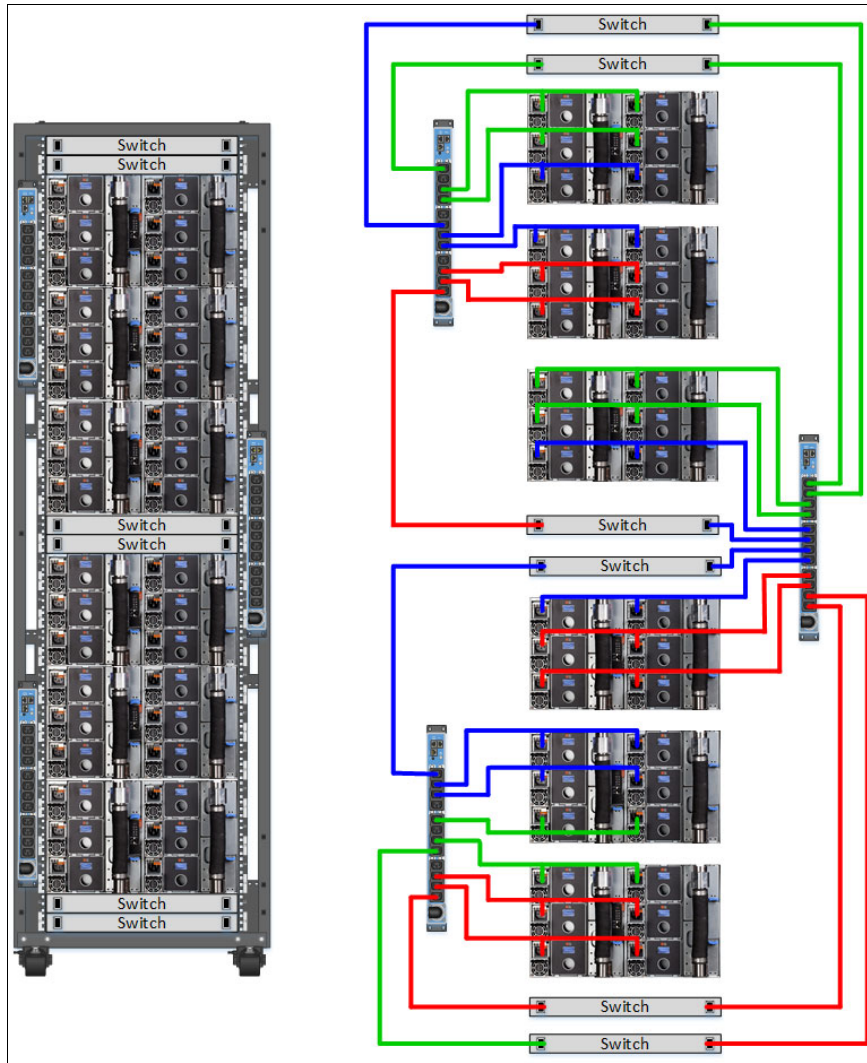


Figure 5-1 Six chassis and six switches that are connected to three PDUs

The PDUs that are shown are in vertical pockets in the rear of an 42U 1100mm Enterprise V2 Dynamic Rack. For more information about the rack, see 5.4.1, “Rack Weight” on page 75.

This configuration requires “Y” power cables for the chassis power supplies. Part numbers are listed in Table 5-2.

Table 5-2 Y cable part numbers

Feature code	Description
A3SW	1.2 m, 16A/100-250V, 2 Short C13s to Short C20 Rack Power Cable
A3SX	2.5 m, 16A/100-250V, 2 Long C13s to Short C20 Rack Power Cable

Figure 5-2 shows the connections from four 1U PDUs. Each PDU has 12 outlets; therefore, 48 outlets are available. There are six NeXtScale n1200 WCT Enclosures, each with six power supplies, so there are 36 power supplies to be connected. In all, there are 12 outlets that are not connected to chassis.

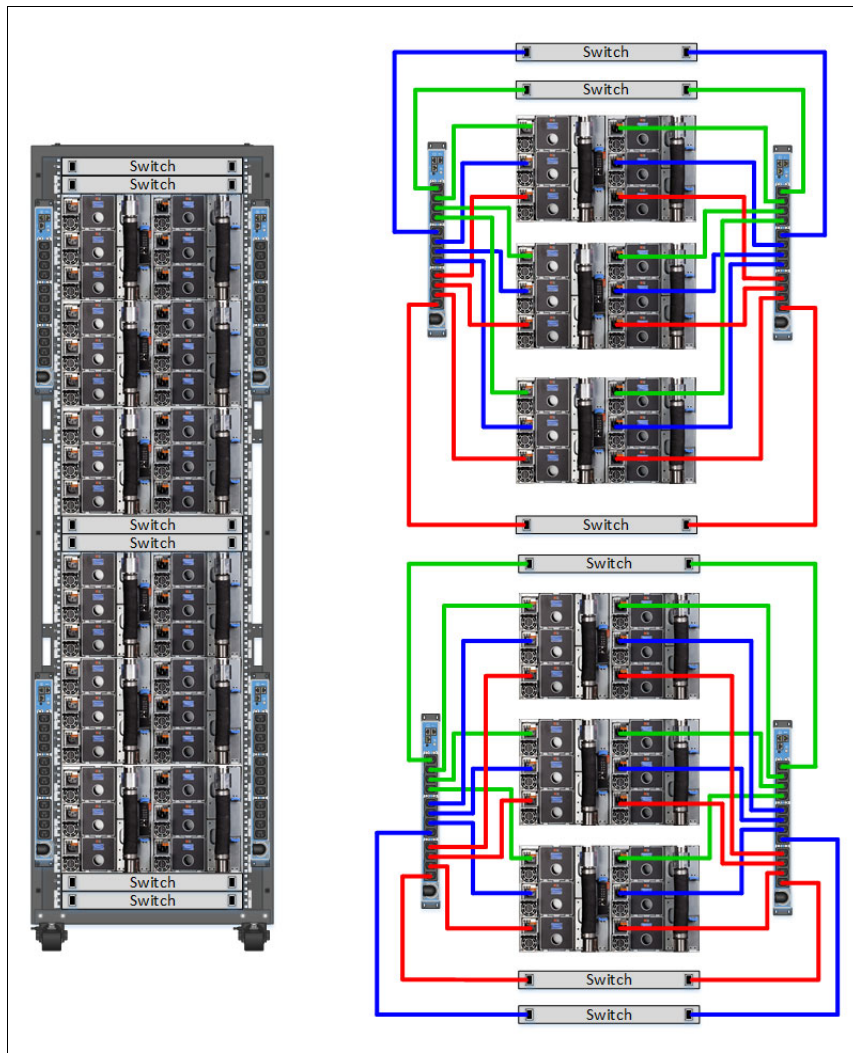


Figure 5-2 Sample power cabling: Six chassis and six switches

With 6U of rack space left open, it is possible to put six 1U devices in the rack and provide two independent power connections to each device. This configuration can provide power

redundancy to the chassis, servers, and optional devices that are installed in the rack, depending on the specifications of the equipment.

5.1.3 PDUs

There are several power distribution units that can be used with NeXtScale WCT System. Listed in Table 5-3 are 1U rack units, which have 12 C13 outlets and are supplied with 200 - 240 V 60 A, three-phase power or 380 - 415 V 32 A, three-phase power.

0U PDUs: The Lenovo 0U PDU should not be used in the 42U 1100mm Enterprise V2 Dynamic Rack with the NeXtScale n1200 Enclosure because there is inadequate clearance between the rear of the chassis and the PDU.

Table 5-3 PDUs for use with NeXtScale System

Part number	Description
Switched and Monitored PDU	
46M4005	1U 12 C13 Switched and Monitored 60 A 3-Phase PDU
46M4004	1U 12 C13 Switched and Monitored PDU without line cord
Monitored PDU	
39M2816	DPI C13 Enterprise PDU without line cord
Basic PDU	
39Y8941	Enterprise C13 PDU

The switched and monitored PDU (part number 46M4005) includes an attached line cord with IEC 609 3P+G plug. The other PDUs require a line cord, which is listed in Table 5-4.

Table 5-4 Line cord part number

Part number	Description
40K9611	32 A 380-415V IEC 309 3P+N+G (non-US) Line Cord

For more information about selecting appropriate PDUs to configure, see the Lenovo Power Guide, which is available at this website:

<http://ibm.com/support/entry/portal/docdisplay?lnocid=LNVO-POWINF>

For more information about PDUs, see the Lenovo Press Product Guides that are available at this website:

<http://lenovopress.com/systemx/power>

5.1.4 UPS units

There are several rack-mounted UPS units that can be used with the NeXtScale systems, which are listed in Table 5-5. In larger configurations, UPS service is often supplied at the data center level.

Table 5-5 UPS units for use with NeXtScale System

Part number	Description
55945KX	RT5kVA 3U Rack UPS (200-240Vac)
55946KX	RT6kVA 3U Rack UPS (200-240Vac)
55948KX	RT8kVA 6U Rack UPS (200-240Vac)
55949KX	RT11kVA 6U Rack UPS (200-240Vac)
55948PX	RT8kVA 6U 3:1 Phase Rack UPS (380-415Vac)
55949PX	RT11kVA 6U 3:1 Phase Rack UPS (380-415Vac)

For more information about UPS units, see the Lenovo Press Product Guides that are available at this website:

<http://lenovopress.com/systemx/power>

5.2 Cooling

Cooling is an important consideration in designing any server solution. It can require careful planning for large-scale environments. After the power planning is complete, calculating the amount of heat to be dissipated is relatively straightforward. For each W of power that is used, 3.414 British Thermal Units per hour (BTU/hr) of cooling is required.

5.2.1 Planning for air cooling

It is often more difficult to determine the volume of air, which is commonly measured in cubic feet per minute (CFM), that is required to cool the servers. The following factors influence the air volume requirement of the servers:

- ▶ The intake air temperature, which affects how much air is required to cool the components to an acceptable temperature.
- ▶ Humidity, which affects the air's thermal conductivity.
- ▶ Altitude and barometric pressure, which affect air density.
- ▶ Airflow impedance in the environment.

Table 5-6 shows typical and maximum airflow values for the NeXtScale n1200 WCT Enclosure at 25C inlet ambient temperature.

Table 5-6 Typical and maximum airflow for the NeXtScale n1200 WCT Enclosure

Typical Airflow (CFM)	Maximum Airflow (CFM)
35 CFM	60 CFM

In data centers that contain several power dense racks, extracting the used warm air and providing enough chilled air to the equipment intakes can be challenging.

In these environments, one of the primary considerations is preventing warm air from recirculating directly from the equipment exhaust into the cold air intake. It is important to use filler panels to block any unused rack space. Part numbers for kits of five filler panels, which install quickly and without tools, are listed in Table 5-7.

Table 5-7 Blank filler panel kits: Five panels per kit

Part number	Description
25R5559	1U Quick Install Filler Panel Kit (quantity five)
25R5560	3U Quick Install Filler Panel Kit (quantity five)

The blank filler panel kits that are listed in Table 5-7 can be used next to the NeXtScale n1200 Enclosure. However, next to switches that re mounted in the front of the rack requires the use of the 1U Pass Through Bracket (part number 00Y3011), as described in 5.4.5, “Rack options” on page 83. This bracket is required if there is an empty 1U space above or below a front-mounted switch to prevent air recirculation from the rear to the front of the switch. Front-mounted 1U switches that are used with NeXtScale are recessed 75 mm behind the front rack mounting brackets to provide sufficient room for cables. A standard filler panel does not contact a switch that is recessed 75 mm; therefore, hot air recirculation can occur.

The next areas to watch for air recirculation are above, below, and around the sides of the racks. Avoiding recirculation above the racks can be difficult to address because this area is typically the return path to the air conditioner and often there is overhead lighting, plumbing, cable trays, or other things with which to contend.

Examples of the use of recirculation prevention plates or stabilizer brackets to prevent under rack air return, and joining racks to restrict airflow around the sides is described in 5.4.1, “Rack Weight” on page 75. We also suggest covering any unused space in rack cable openings.

There are various approaches to addressing the cooling challenge. The use of traditional computer room air conditioners with a raised floor and perforated floor tiles requires numerous floor space that is dedicated to perforated tiles and generous space for air return paths.

Smaller rooms are possible with air curtains, cold air and warm air separation ducting, fans or blowers, or other airflow modification schemes. These approaches typically create restricted spaces for airflow, which are windy, noisy, inflexible in their layout, and difficult to move around in.

Another approach is to contain a single rack or a row of racks in purpose built enclosure with its own air movement and cooling equipment. When multiplied by several racks or rows of racks, this approach is expensive and less space efficient. It also often requires the rack or row to be powered off if the enclosure must be opened for maintenance of the equipment or the enclosure.

Lenovo suggests the use of Rear Door Heat Exchanger, as described in 5.6, “Rear Door Heat eXchanger” on page 89. This option is a relatively low-cost, low-complexity, space, and power efficient solution to the cooling challenge.

The NeXtScale WCT System still does require some treated and conditioned air supply for operation. The power supplies are not cooled with the water cooling and thus require fans to remove heat. This approach also gives the NeXtScale WCT System the added flexibility of

supporting some pluggable adapters in the compute nodes that are not directly cooled with water.

The use of water cooling, as in the NeXtScale WCT System, significantly reduces the effect of warm air recirculation because the airflow requirements of the solution are so low compared to traditional air-cooled solutions. The amount of heat that is added to the room is also less because most of the heat that is extracted through the water loops that run throughout the system.

5.2.2 Planning for water cooling

Planning for the deployment of a NeXtScale WCT System in a data center environment is similar to planning for air-cooled installations in that the facility must operate within a specific set of parameters. In addition to making sure that the facility can supply conditioned and treated air within specified ranges for temperature and humidity, the water that is providing the cooling to the NeXtScale WCT System must be supplied within specified ranges for temperature, flow rate, and water quality.

NeXtScale WCT Manifold connection to facility water supply

The NeXtScale n1200 WCT Manifold connects to the facility side secondary water loop through two 1-inch EPDM hose connections. Each hose connection is terminated with 1-inch Eaton ball valves, as listed in Table 5-8 and shown in Figure 5-3 on page 72. The facility side hoses and termination valves often are supplied by the firm that is installing the NeXtScale WCT System.

Table 5-8 Connection hose termination part number

Description	Eaton Part number
Eaton 1-inch ball valve	45D6934

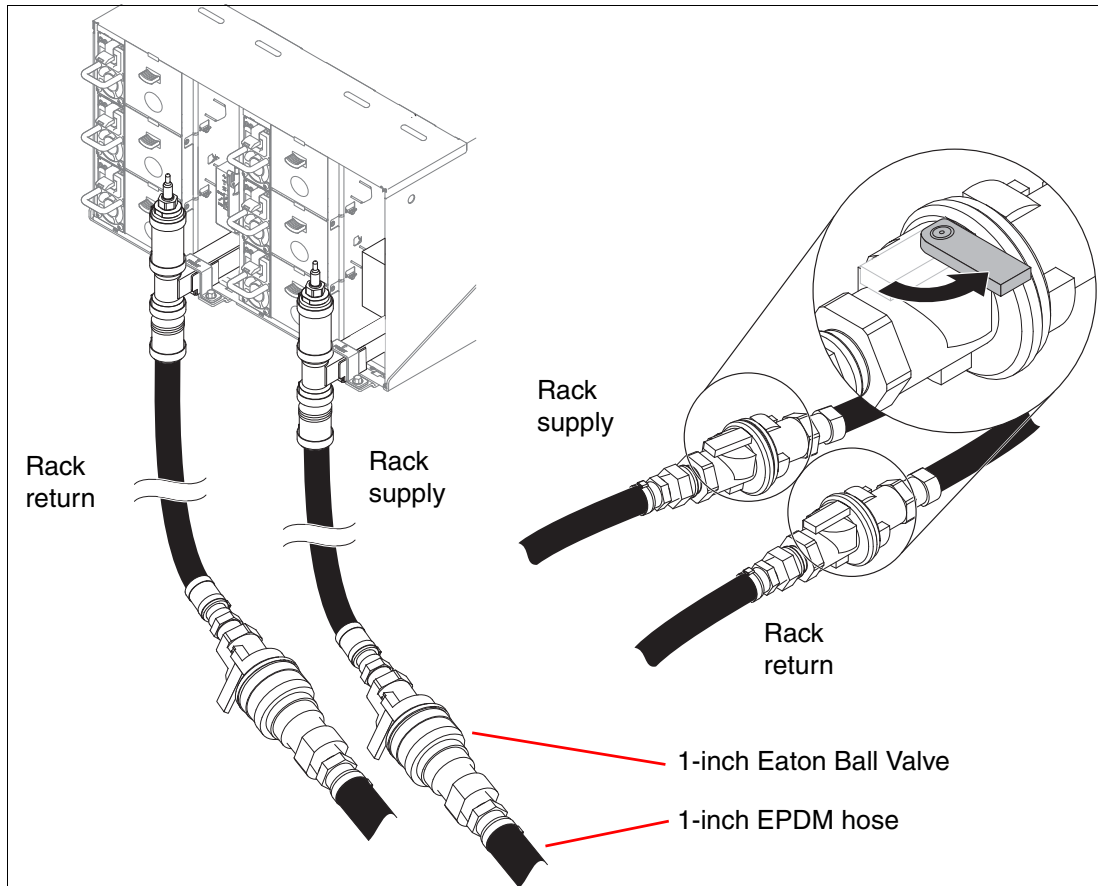


Figure 5-3 WCT Manifold connection with EPDM hose and Eaton ball valves

Table 5-9 lists the water requirements for a rack with six NeXtScale n1200 WCT Enclosures that support 36 nx360 M5 WCT Compute Trays.

Table 5-9 Water requirements for NeXtScale WCT Systems

Parameter	Minimum	Typical	Maximum
Flow Rate	36 liters/min.	N/A	N/A
Water Pressure	N/A	N/A	50 psi
Water Temperature	18 C ¹	N/A	45 C ²
Water Quality ³	N/A	50 µM	100 CFU/ml
Rack Pressure Drop	1.5 psi	2.4 psi	4 psi

1. The WCT manifold and supply hose are not insulated. Avoid any condition that might cause condensation. The water temperature inside the supply hose, return hose, and WCT Manifold must be kept above the dew point of the environment where the NeXtScale WCT System is operating.
2. The maximum is 35° C for processor option E5-2698 v3 (165 W).
3. The water must be treated with anti-biological and anti-corrosion measures.

Data center topology

The most common implementation of a NeXtScale WCT System or a Rear Door Heat Exchanger solution uses a chiller distribution unit (CDU) to isolate the facility primary side chilled water loop from the secondary side water loop. Figure 5-4 on page 73 shows a typical CDU connection that supports multiple Rear Door Heat Exchangers. Most chilled-water

primary loops operate 6° C - 12° C. The CDU regulates the water temperature, flow rate, and monitors dew point of the local environment to ensure the water that is supplied to the NeXtScale WCT System or Rear Door Heat Exchanger is within the correct operating parameters. Multiple CDUs can be connected in parallel on a common external secondary side water manifold to provide for redundancy.

Although CDUs are commonly used, they are not a fixed requirement if the appropriate controls are in place. For example, a stand-alone dry-cooler can cool a NeXtScale WCT System solution if deployed in certain geographies.

The high operating water temperature of the NeXtScale WCT System allows for increased use of free cooling with economizers. This use allows the facility chillers to be bypassed for longer periods of time throughout the year, which can yield significant energy savings at the facility level. In many configurations, the return water temperature is high enough to be used for pre-heating building water for use in heating facility common areas.

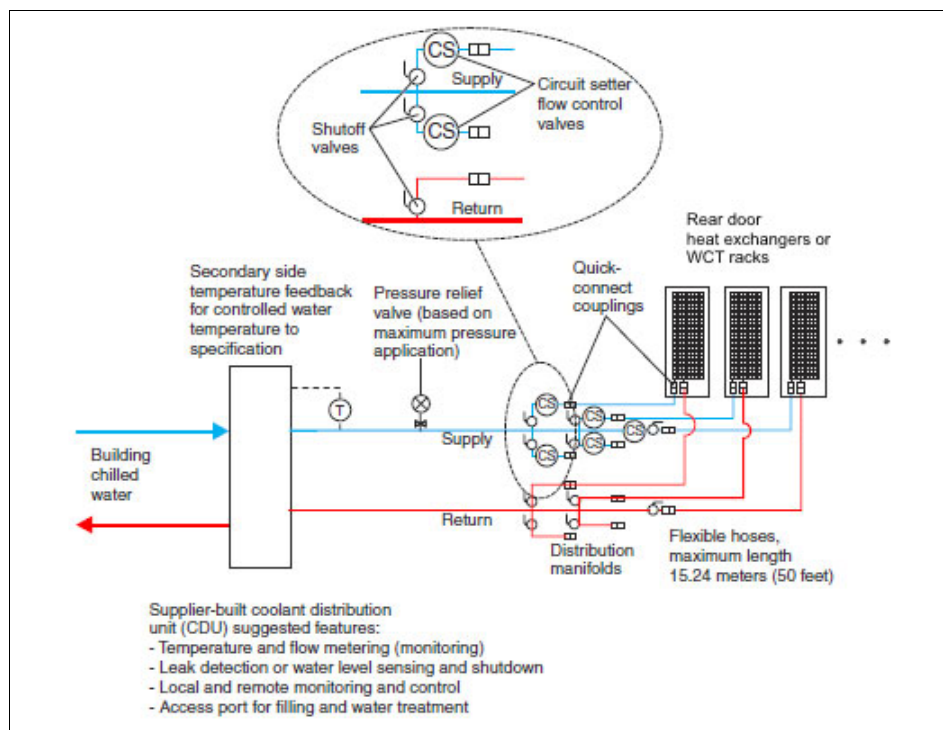


Figure 5-4 Cooling distribution unit connection between facility primary and secondary water loops

Figure 5-5 on page 74 shows a typical under-floor manifold for connecting to multiple racks from a CDU. It does not show the shut-off valves, pressure regulator, or inline filter that are shown in Figure 5-4.

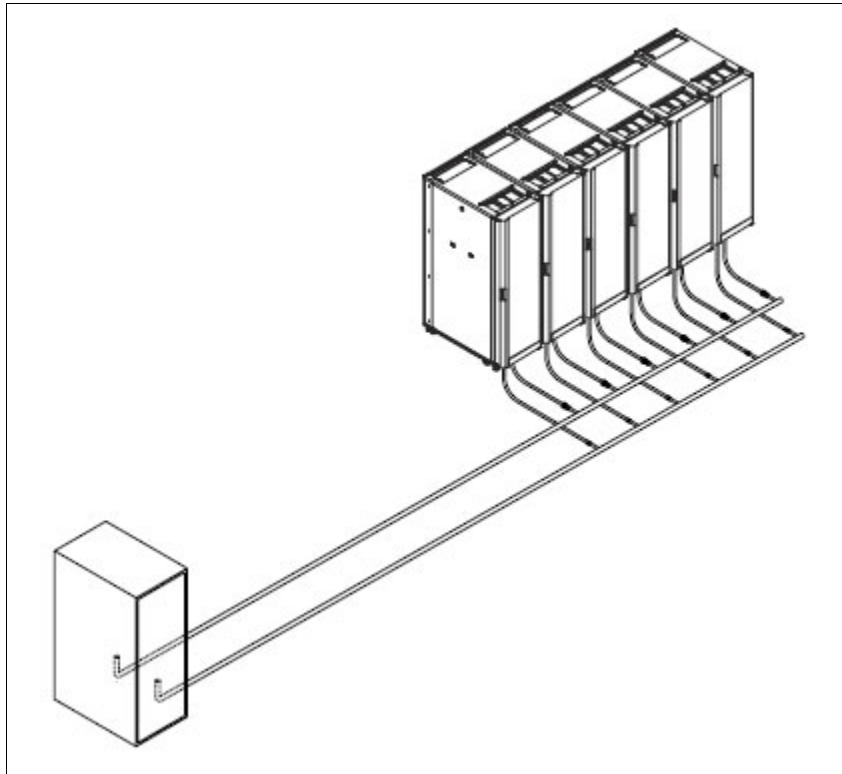


Figure 5-5 Typical extended CDU manifold connection WCT racks or Rear Door Heat Exchangers

The manifold pipe diameter is dependent on several variables, such as length of run, required flow rate, pumps capacity, and back-pressure because of routing. It also depends on whether a conventional CDU is used. Pipe dimensions must be independently verified by the plumbing contractor or engineering firm that is installing the facility side cooling infrastructure.

Manifolds that accept large diameter piping feed pipes from a pump unit are the preferred method for splitting the flow of water to smaller diameter pipes or hoses that are routed to individual WCT racks. Manifolds must be constructed of materials that are compatible with the pump unit and related piping. The manifolds must provide enough connection points to allow a matching number of supply and return lines to be attached, and the manifolds must match the capacity rating of the pumps and the loop heat exchanger (between the secondary cooling loop and the building chilled-water source). All manifolds must be anchored or restrained to provide support against movement when couplings are connected and disconnected to and from the manifolds.

To allow stopping the flow of water in individual legs of multiple circuit loops, install shutoff valves for each supply line that exits the manifold. This configuration provides a way of servicing or replacing WCT elements in a rack without affecting the operation of other WCT racks in the same loop.

A best practice for managing a water environment to ensure that water specifications are met and that the optimum heat removal is occurring and temperature and flow metering (monitoring) are used in secondary loops.

5.3 Density

With the NeXtScale WCT System, most data centers should sustain a density increase of approximately 28% as compared to iDataPlex (Lenovo's previous high-density server design) even with 6U in each 42U rack that is reserved for switches or cabling. These server types are compared in Figure 5-6 and use an area 10 x 10 floor tile-wide deep-rack. NeXtScale servers can easily provide double the density of 1U rack server configurations.

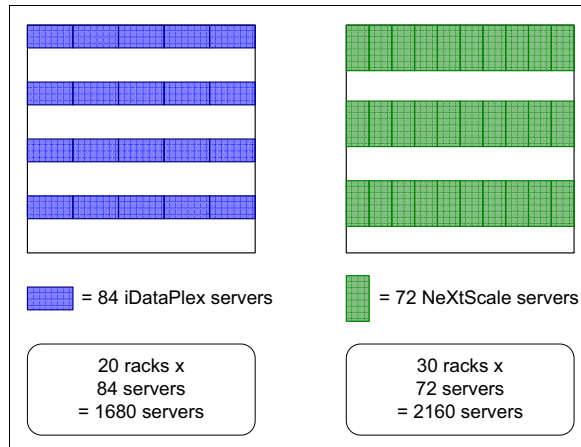


Figure 5-6 Increase in server density with NeXtScale compared to iDataPlex

The examples that are shown in Figure 5-6 are at the higher end of the density spectrum. They are meant to show what is possible with high-density designs.

5.4 Racks

In this section, we describe installing NeXtScale WCT System in the 42U 1100mm Enterprise V2 Dynamic Rack because this rack is the rack that is used in our Intelligent Cluster solution. It is also the rack we recommend for NeXtScale System implementations. Examples of several best practices also are described. Other considerations for installing servers in other racks and other rack options also are described.

5.4.1 Rack Weight

The NeXtScale WCT System can pose some weight challenges when deployed in raised floor environments. Table 5-10 shows the building block weights for the typical building block components that make up NeXtScale configurations.

Table 5-10 Building block weights for NeXtScale WCT System components

Component	Weight
Chassis + Manifold (Loaded)	127.62 kg
Network Switch	6.4 kg
Power distribution unit + power cord	11.7 kg
Cable weight per chassis (Ethernet + Power)	12 kg
Rack (Empty)	187 kg

Table 5-11 shows the total rack weights and floor loading for NeXtScale WCT Systems. Typical configurations consist of each chassis having 6 dual planar compute trays, and six 1300 W power supplies. Each rack consists of six NeXtScale n1200 WCT enclosures, six network switches, six drop WCT manifold assembly, and four PDUs.

Table 5-11 Total rack weights for NeXtScale WCT System

Component	Weight
Shipping weight (Crated)	1150 kg
Weight (Uncrated)	1110 kg
Point load (four points per rack)	277 kg
Floor loading - Rack only (kg/m ²)	1682 kg/m ²
Floor loading - Rack + service area (kg/m ²)	661 kg/m ²

5.4.2 The 42U 1100mm Enterprise V2 Dynamic Rack

The 42U 1100mm Enterprise V2 Dynamic Rack is Lenovo’s leading open systems server rack. The word *dynamic* in the name means that it is shipped with equipment installed. As shown in Figure 5-7, it features removable outriggers to prevent it from tipping when it is moved with equipment that is installed in it.

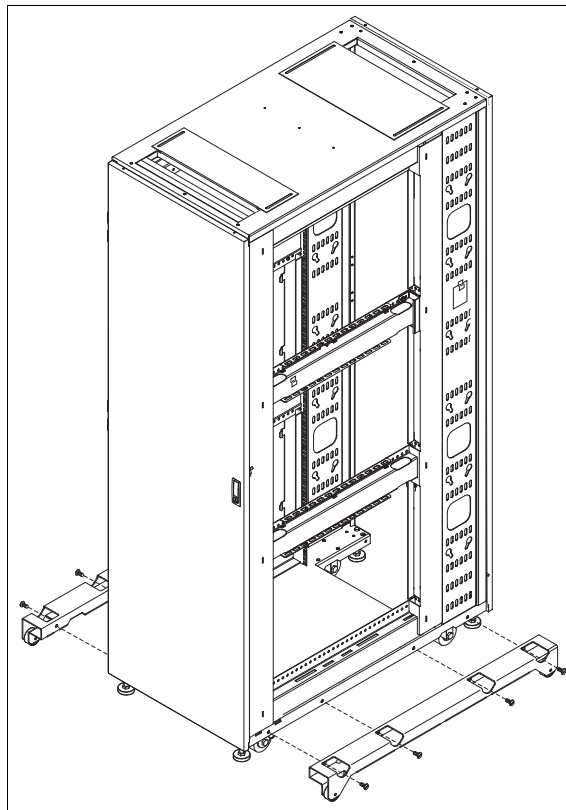


Figure 5-7 Outrigger removal or attachment

The rack features fully welded construction and is rated for a load of 953 kilograms (2100 pounds). The rack is one standard data center floor tile (600 mm) wide, and 1100 mm deep. When the optional Rear Door Heat Exchanger (which is described in 5.6, “Rear Door

Heat eXchanger” on page 89) is added, the rack is two standard data center floor tiles (1200 mm) deep.

The base model of the rack includes side panels. The expansion model of the rack includes the hardware that is used to join it to another rack, and no side panels. A row of racks (sometimes called a suite) can be created with one base model rack and one or more expansion racks; the side panels from the base rack are installed on each end of the row. Figure 5-8 shows joining two racks. Racks that are connected in this way match up with standard floor tile widths, which is not possible by moving the base model racks next to each other.

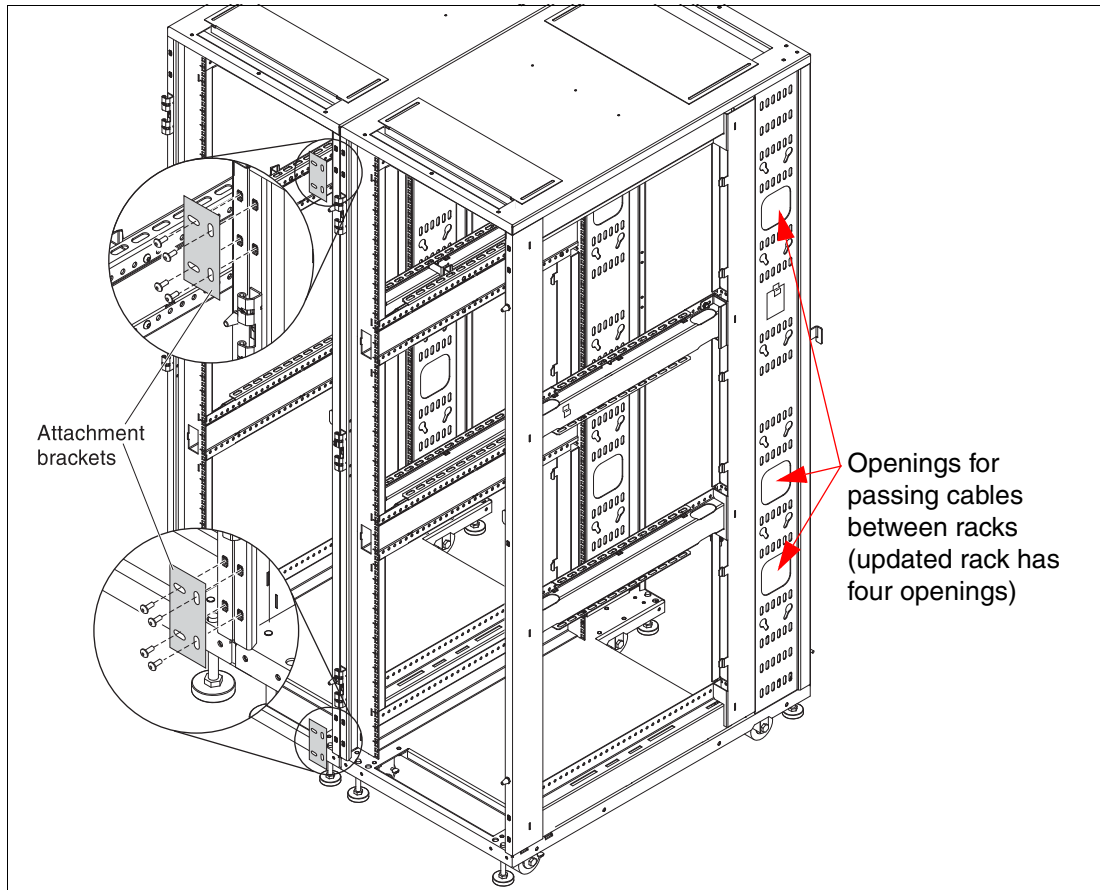


Figure 5-8 Joining two racks to make a row

The following features also are included:

- ▶ A front stabilizer bracket, which can be used to secure the rack to the floor, as shown in Figure 5-9.

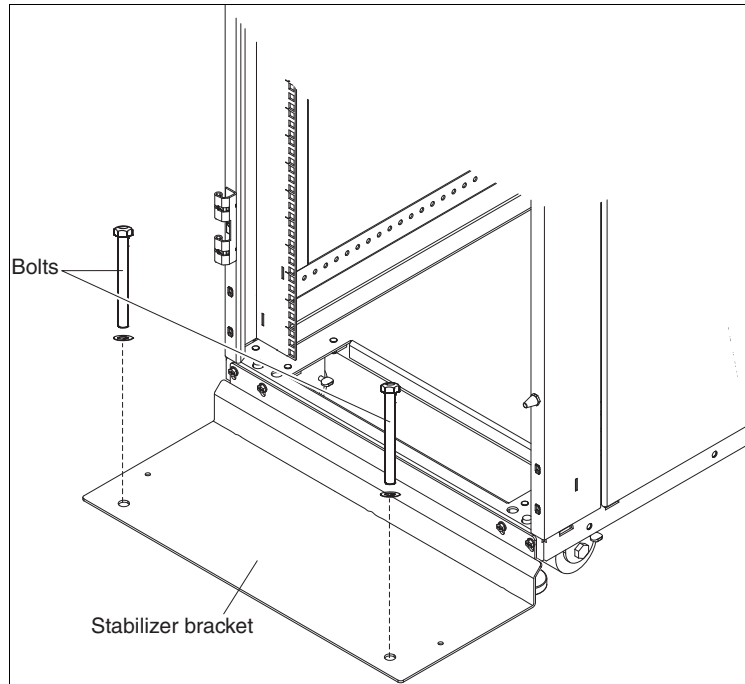


Figure 5-9 Rack with stabilizer bracket attached that is bolted to the floor

- ▶ As shown in Figure 5-10, a recirculation plate is used to prevent warm air that is entering from the rear of the rack from passing under the rack and into the front of the servers. This plate is not required if the stabilizer bracket is installed.

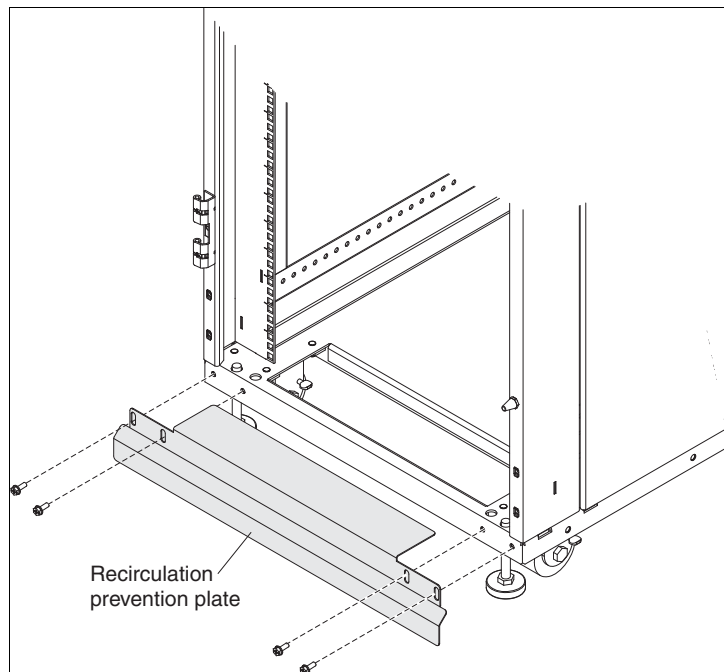


Figure 5-10 Attaching the recirculation prevention plate

Note: In addition to the recirculation plate or stabilizer bracket, another seal kit (part number 00Y3001) is required to prevent air recirculation through the opening at the front, bottom of the rack if a recessed switch is in the U space 1. For more information, see 5.4.5, “Rack options” on page 83.

- ▶ One-piece perforated front and rear doors with reversible hinges, so the doors can be made to open to the right or the left.
- ▶ Lockable front and rear doors, and side panels.
- ▶ Height of less than 80 inches, which enables it to fit through the doors of most elevators and doorways.
- ▶ Reusable, ship-loadable packaging. For more information about the transportation system, see this website:

<http://ibm.com/support/entry/portal/docdisplay?lnocid=migr-5091922>

- ▶ Six 1U vertical mounting brackets in the rear post flanges, which can be used for power distribution units, switches, or other 1U devices, as shown in Figure 5-11.

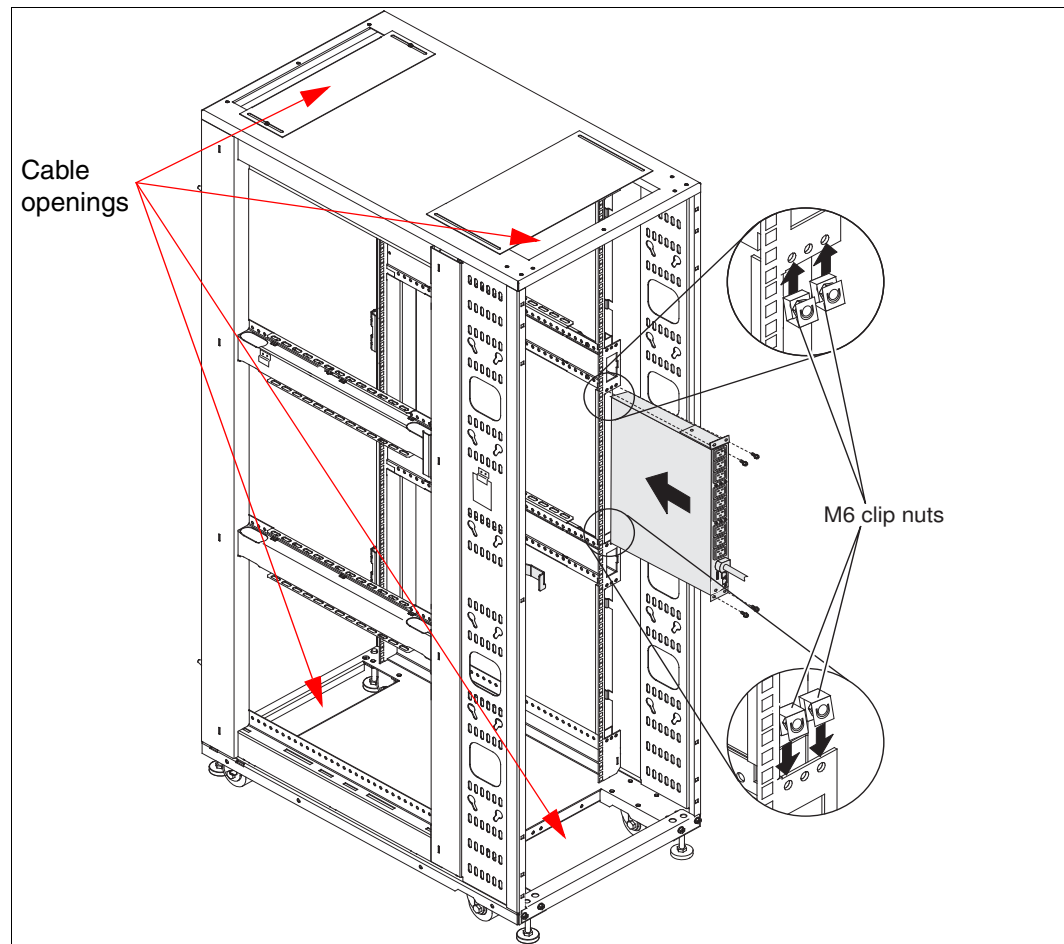


Figure 5-11 1U Power distribution unit mounted in 1U pocket on flange of rear post

- ▶ Two front-to-rear cable channels on each side of the rack, as shown in Figure 5-12.

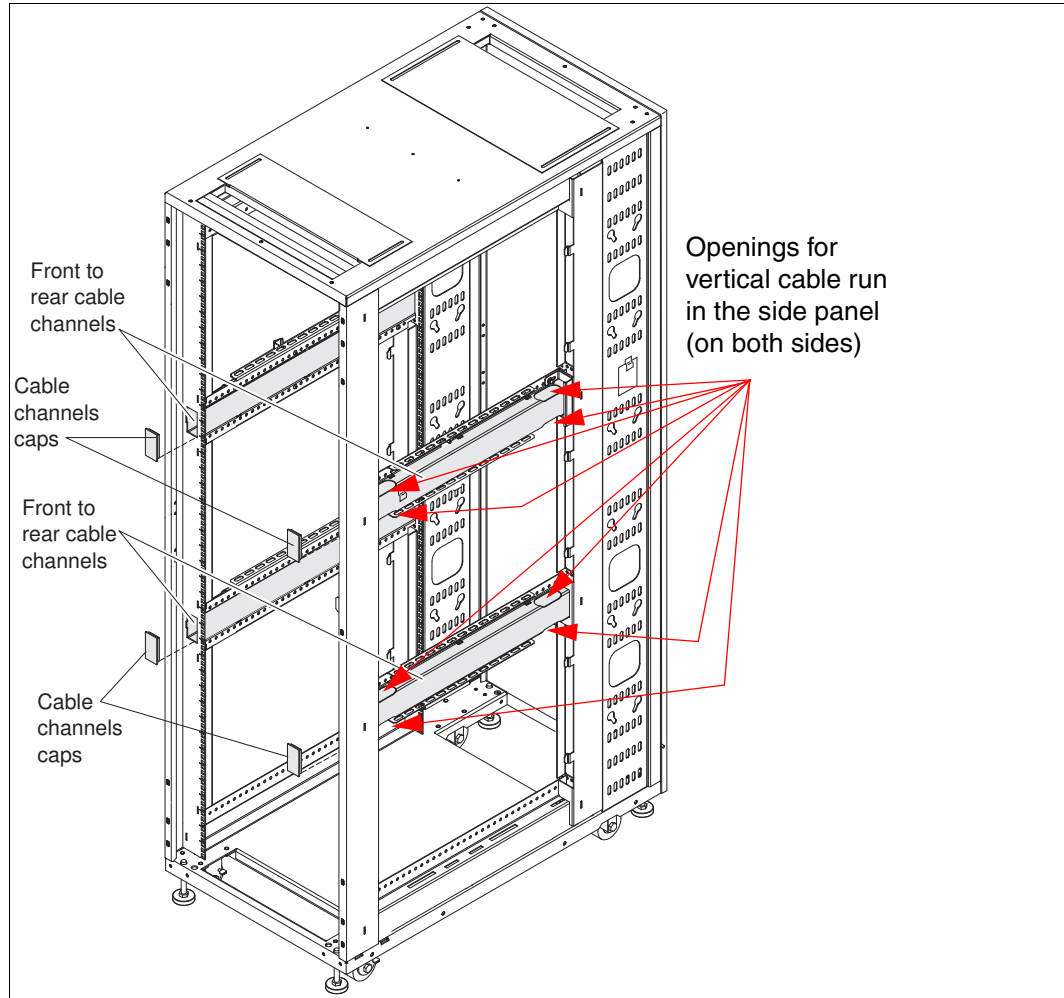


Figure 5-12 Front to rear cable channels

- ▶ Openings in the cable channels on each side of the rack to enable vertical cabling within the side panels, as indicated by arrows in Figure 5-12.
- ▶ Openings in the side walls behind the rear posts through which cables can be routed between racks in a row, as indicated by the arrows that are shown in Figure 5-8 on page 77. Also included are attachment points above and below these openings to hold cables out of the way and reduce cable clutter. New versions of this rack have four openings in each side wall.
- ▶ Front and rear cable access openings at the top and bottom of the rack, as indicated by the arrows that are shown in Figure 5-11 on page 79. The top cable access opening doors slide to restrict the size of the opening. For most effective use, do not tightly bundle cable that is passing through these openings. Instead, use the doors to flatten them in to a narrow row.
- ▶ Pre-drilled holes in the top of the rack for mounting third-party cable trays to the top of the rack.

The part numbers for the racks are listed in Table 5-12.

Table 5-12 Enterprise V2 Dynamic Rack part numbers

Part Number	Description
93634PX	42U 1100 mm Enterprise V2 Dynamic Rack
93634EX	42U 1100 mm Enterprise V2 Dynamic Expansions Rack

For more information about this rack, see the *Installation Guide*, which available at this website:

<http://www.ibm.com/support/entry/portal/docdisplay?lnocid=migr-5089535>

5.4.3 Installing NeXtScale WCT System in other racks

The NeXtScale WCT System chassis can be installed in other racks. The NeXtScale n1200 WCT Enclosure can be installed in most industry-standard, four-post server racks. Figure 5-13 shows the dimensions of the NeXtScale n1200 WCT Enclosure and included rail kit.

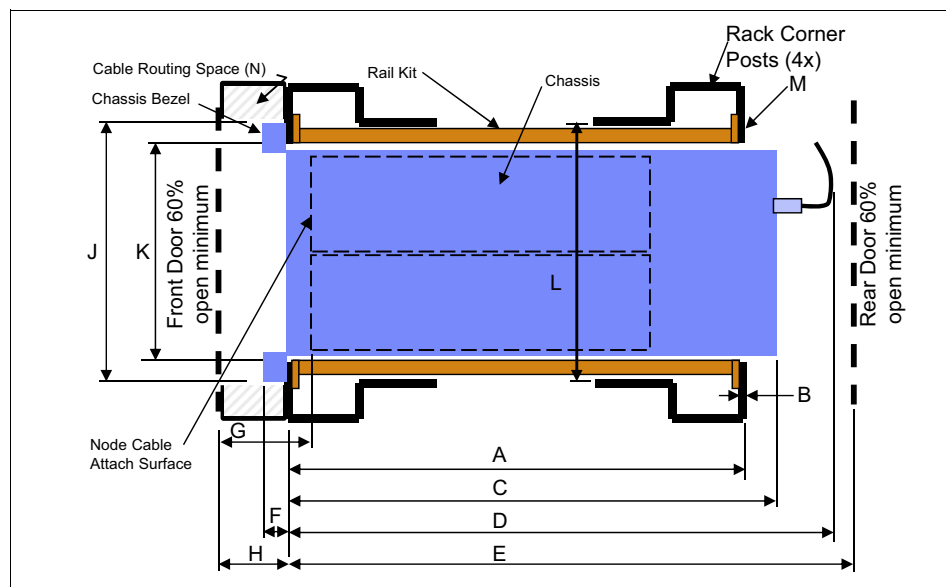


Figure 5-13 Dimensions for mounting the n1200 WCT Enclosure rails and brackets (top view)

The following features are highlighted in Figure 5-13:

- The distance between the outside of front and outside of the rear rack EIA flanges can range 609 - 780 mm. This distance must be 719 mm to support the supplied rear shipping brackets when shipping is configured in a rack.
- The thickness of rack EIA flanges should be 2 mm - 4.65 mm. The supplied cage and clip nuts do not fit on material thicker than 3 mm.
- The distance from front EIA flange to the rear of the system is 915 mm (including handles and latches, the distance is approximately 935 mm).
- The distance from front EIA flange to the bend radius of the rear cables is 980 mm.
- The minimum distance from the front EIA flange to the closest features on the inside of the rear door is 985 mm.

- F. The distance from front EIA flange to the front bezel of the system is 35 mm.
- G. A recommended minimum distance from front of node cable plug surface to inside of front door is 120 mm.
- H. A recommended minimum distance from front EIA flange to the closest features on the inside of the front door is approximately 80 mm.
- J. The width of the front of the system bezel is 482 mm.
- K. The minimum horizontal opening of the inside of the rack at the front and rear EIA flanges is 450 mm.
- L. The minimum width between the required internal structure to the rack to mount rail kits is 475 mm (481 mm within 50 mm of the front and rear EIA flanges).
- M. Mounts use 7.1 mm round or 9.5 mm square hole racks. Tapped hole racks with standard rail kit are not supported.
- N. NeXtScale WCT System data and management cables attach to the front of the node (chassis). Cable routing space must be provided at the front of the rack, inside or outside the rack. Cable space (N) should from the bottom to the top of the rack and be approximately 25 x 75 mm in size. Some cable configurations might require more cable routing space.

We suggest installing the chassis 1 or 2 rack units from the bottom to allow space for cables and the power line cords to exit the rack without blocking service access to the chassis.

5.4.4 Shipping the chassis

Lenovo supports shipping the NeXtScale n1200 WCT Enclosure in an 42U 1100mm Enterprise V2 Dynamic Rack if the shipping brackets that are supplied with the chassis are installed. Shipping the chassis in any other rack is at your discretion.

When it is installed in a rack and moved or shipped from one location to another, the NeXtScale WCT System shipping brackets must be reattached. For more information, see the documentation that is included with the chassis for the installation of the shipping brackets.

Cooling considerations

It is important that the NeXtScale n1200 WCT Enclosure be installed in an environment that results in proper cooling. The following rack installation considerations are important:

- ▶ The rack must have front (server node side) to back airflow.
- ▶ The rack must have front and rear doors with at least 60% open area for airflow.
- ▶ To allow adequate airflow into and out of the chassis, adhere to the recommended minimum distances to the front and rear doors, as shown in Figure 5-13 on page 81.
- ▶ It is important to prevent hot air recirculation from the back of the rack to the front. The following points should be considered:
 - All U spaces must be occupied at the front by a device or a blank filler panel.
 - All other openings should be covered, including air openings around the EIA flanges (rack posts) and cable passage ways.
 - If multiple racks are arranged in a row, gaps between the rack front EIA flanges must be blocked to prevent hot air recirculation.
- ▶ Seal openings under the front of the racks.

5.4.5 Rack options

The rack options that are listed in Table 5-13 are available for the Enterprise V2 Dynamic Racks and other racks.

Table 5-13 Rack option part numbers

Part number	Description
Monitor kits and keyboard trays	
17238BX	1U 18.5-inch Standard Console
17238EX	1U 18.5-inch Enhanced Media Console
172317X	1U 17-inch Flat Panel Console Kit
172319X	1U 19-inch Flat Panel Console Kit
Console switches	
1754D2X	Global 4x2x32 Console Manager (GCM32)
1754D1X	Global 2x2x16 Console Manager (GCM16)
1754A2X	Local 2x16 Console Manager (LCM16)
1754A1X	Local 1x8 Console Manager (LCM8)
Console cables	
43V6147	Single Cable USB Conversion Option (UCO)
39M2895	USB Conversion Option (four Pack UCO)
39M2897	Long KVM Conversion Option (four Pack Long KCO)
46M5383	Virtual Media Conversion Option Gen2 (VCO2)
46M5382	Serial Conversion Option (SCO)

The options that are listed in Table 5-14 are unique to configuring NeXtScale environments.

Table 5-14 Racking parts that are specific to NeXtScale

Part number	Description
00Y3011	1U Pass Through Bracket
00Y3016	Front cable management bracket kit
00Y3026	Cable Routing Tool
00Y3001	Rack and Switch Seal Kit

The 1U Pass Through Bracket is shown in Figure 5-14. The front for the component has brushes to block the airflow around any cables that are passed through it. It can also serve to block air flow around a switch that is recessed in the rack (for cable routing reasons), that pass around a blank filler panel.

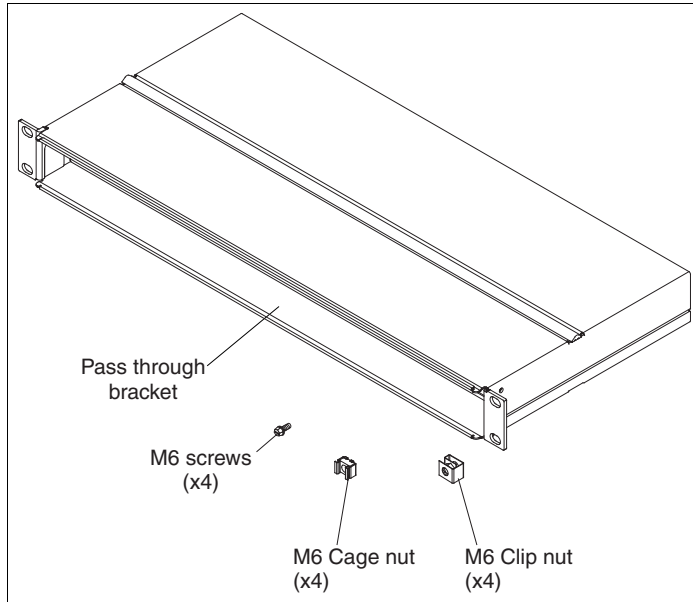


Figure 5-14 1U filler with brushes, allows cables through but blocks air flow

The Front Cable Management Bracket kit (part number 00Y3016) that attaches to the front of the rack is shown in Figure 5-15. This kit includes four brackets, which are enough for one rack (two brackets are installed on each side of the rack).

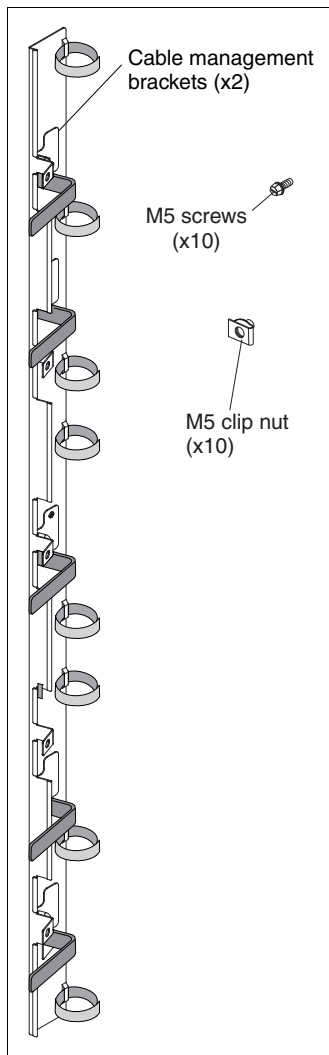


Figure 5-15 Cable management bracket

The Cable Routing Tool (part number 00Y3026) that is shown in Figure 5-15 is a plastic rod to which a cable or cables can be attached by using a hook-and-loop fastener. After it is assembled, the tool is used to pull the cables through the cable channels.

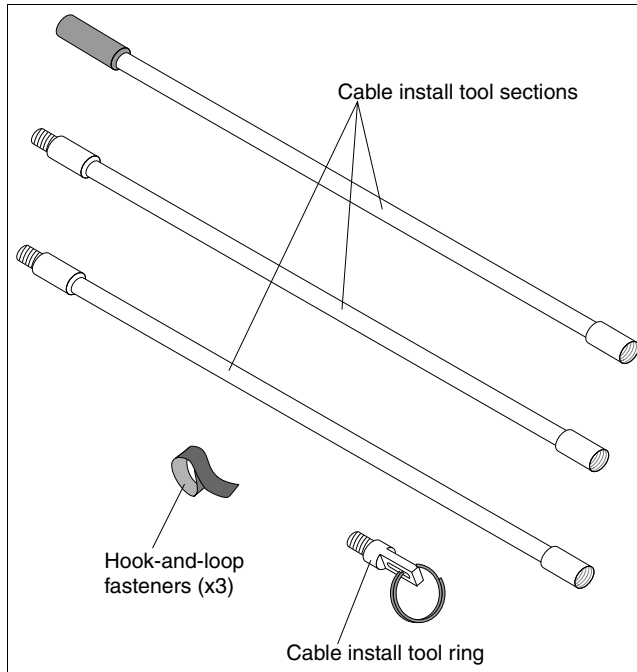


Figure 5-16 Cable routing tool

The Rack and Switch Seal Kit (part number 00Y3001) has the following purposes:

- ▶ Provides a means to seal the opening in the bottom front of the 42U 1100 mm Enterprise V2 Dynamic Rack. The opening at the bottom front of the rack must be sealed if a switch is at the front of the rack in U space one.
- ▶ Provides air sealing of switch mounting rails of switches that are mounted at the front of a rack and are recessed behind the rack mounting rails. The seal kit includes enough switch seals for six switches.

Figure 5-17 shows the components of the Rack and Switch Seal Kit.

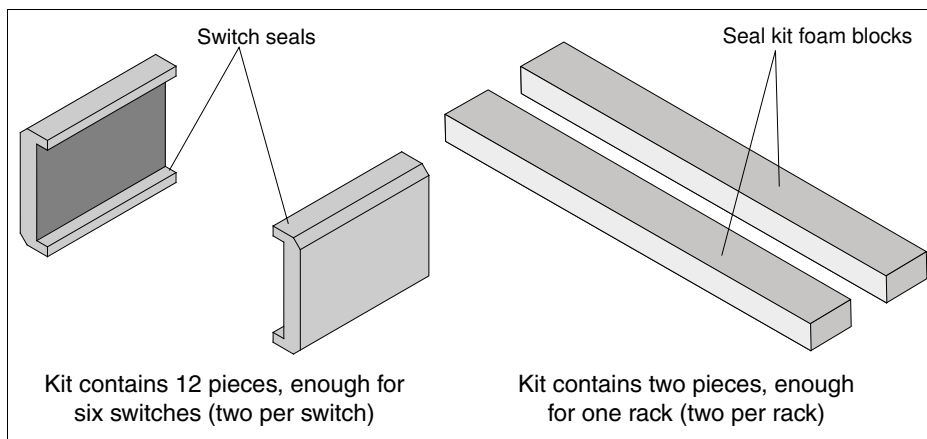


Figure 5-17 Rack and Switch Seal Kit contents

Figure 5-18 shows where these pieces are used.

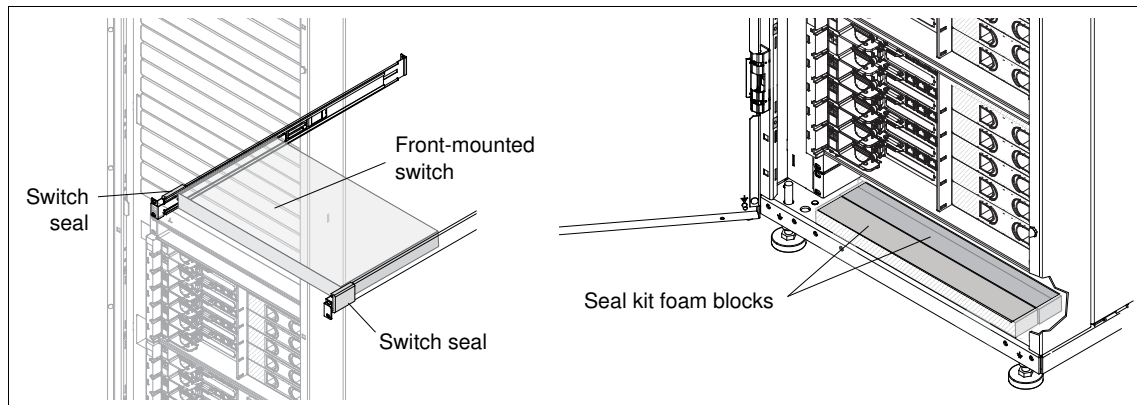


Figure 5-18 Placement of the components of the Rack and Switch Seal Kit

5.5 Cable management

NeXtScale WCT System was designed with serviceability in mind. All of the compute tray connectivity is at the front so trays can be serviced without any guesswork as to which cables belong to which trays. Managing and routing all of this cabling can cause problems; however, Lenovo developed NeXtScale WCT System with options to make cable management simple.

Cable routing

As seen in Figure 5-12 on page 80, the 42U 100 mm Enterprise V2 Dynamic Rack contains two front to rear cable channels on each side of the rack so routing cables to the back of the rack does not waste any U space in the rack. Having these multiple openings on each side of the rack at different heights reduces the overall size of the cable bundles that are supporting the compute trays.

Cables can be routed between racks in a row through openings in the side walls behind the rear posts, as indicated by arrows in Figure 5-8 on page 77. Also included are attachment points above and below these openings to hold cables out of the way and reduce cable clutter. New versions of this rack have four openings in each side wall.

Front and rear cable access openings are available at the top and bottom of the rack, as indicated by arrows in Figure 5-11 on page 79. The top cable access opening doors slide to restrict the size of the opening. For most effective use, do not tightly bundle cable that is passing through these openings. Instead, use the doors to flatten them in to a narrow row.

Pre-drilled holes are available in the top of the rack for mounting third-party cable trays to the top of the rack.

Cable management brackets

An optional cable management bracket kit is available to help with cable management of the cable bundles at the front of the rack. This kit consists of six brackets (enough for three chassis) that attach to the front left and right of the chassis. To use the bracket, the front door of the rack often must be removed because the bracket can extend beyond the front of the rack. Figure 5-19 shows the cable bracket.

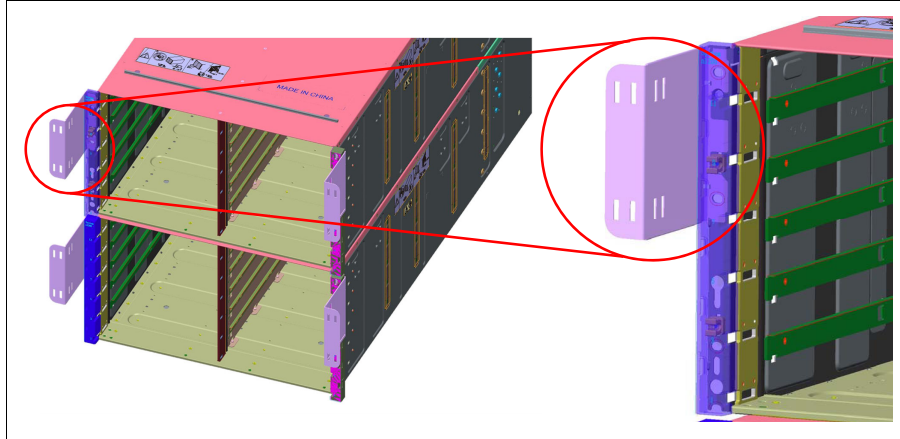


Figure 5-19 Cable management bracket kit for third-party racks, part 00Y3040

The part number of the Cable Management Bracket kit is listed in Table 5-15.

Table 5-15 Cable Management Bracket Kit part number

Part number	Description
00Y3040	Cable management bracket kit (contains 6 brackets and 20 hook-and-loop fasteners)

Figure 5-20 shows a top view of the bracket and possible location for cable bundles.

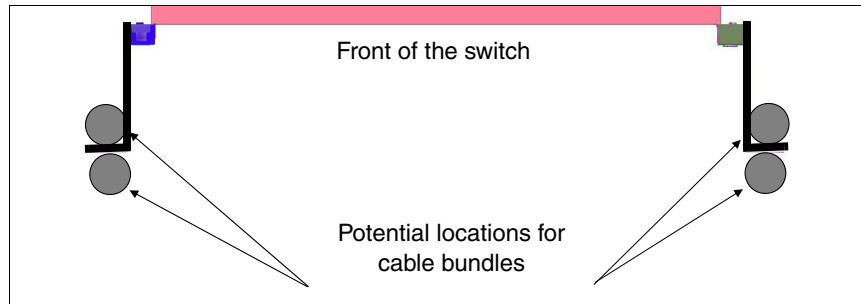


Figure 5-20 Top view of cable management bracket kit, showing cable bundles

Another option for managing cable bundles at the front of the rack is the Front Cable Management Bracket kit (part number 00Y3016). This bracket assembly attaches to the front of the rack, as shown in Figure 5-15 on page 85. This kit includes four brackets, which are enough for one rack (two brackets are installed on each side of the rack).

Cable routing support

The Cable Routing Tool (part number 00Y3026) that is shown in Figure 5-16 on page 86 is a plastic rod to which a cable or cables can be attached by using a hook-and-loop fastener. After it is assembled, the tool is used to pull the cables through the cable channels. This process helps make changes to the rack cabling much easier and reduces the potential for damage to the cabling by not forcing cables through confined environments.

5.6 Rear Door Heat eXchanger

The heat exchanger is a water-cooled door that is mounted on the rear of an 42U 1100 mm Deep Dynamic Rack Type 9363 to cool the air that is heated and exhausted by devices inside the rack. A supply hose delivers chilled, conditioned water to the heat exchanger. A return hose delivers warmed water back to the water pump or chiller. In this document, this configuration is referred to as a *secondary cooling loop*. The primary cooling loop supplies the building chilled water to secondary cooling loops and air conditioning units. The rack on which you install the heat exchanger can be on a raised floor or a non-raised floor. Each heat exchanger can remove 100,000 BTU per hour (or approximately 30,000 watts) of heat from your data center.

The part number for the Rear Door Heat eXchanger for 42U 1100 mm rack is shown in Table 5-16.

Table 5-16 Part number for Rear Door Heat eXchanger for 42U 1100 mm rack

Part number	Description
175642X	Rear Door Heat eXchanger for 42U 1100 mm Enterprise V2 Dynamic Racks

For more information, see *Rear Door Heat eXchanger V2 Type 1756 Installation and Maintenance Guide*, which is available at this website:

<http://www.ibm.com/support/entry/portal/docdisplay?ln docid=migr-5089575>

Rear Door Heat eXchanger V2 overview

Table 5-17 lists the specifications of the Rear Door Heat eXchanger V2 type 1756.

Table 5-17 Rear Door Heat eXchanger V2 specifications

Parameter	Specification
Door dimensions	
Depth	129 mm (5.0 in.)
Height	1950 mm (76.8 in.)
Width	600 mm (23.6 in.)
Door weight	
Empty	39 kg (85 lb)
Filled	48 kg (105 lb)

Parameter	Specification
Pressure	
Normal operation	<137.93 kPa (20 psi)
Maximum	689.66 kPa (100 psi)
Volume	
Water volume	9 liters (2.4 gallons)
Flow rate	
Nominal flow	22.7 lpm (6 gpm)
Maximum flow	56.8 lpm (15 gpm)
Water Temperature	
Minimum (non-condensing)	Above dew point
ASHRAE Class 1	18 °C +/- 1 °C (64.4 °F +/- 1.8 °F)
ASHRAE Class 2	22 °C +/- 1 °C (71.6 °F +/- 1.8 °F)

The Rear Door Heat eXchanger (RDHX) also includes the following features:

- ▶ Attaches in place of the perforated rear door and adds 100 mm, which makes the overall package 1200 mm (the depth of two standard data center floor tiles).
- ▶ The doors use 3/4-inch quick connect couplers, which include automatic valves that restrict water leakage (often a few drops at most) when the doors are connected or disconnected.
- ▶ Each door has a capacity of 9 liters (2.4 US gallons), and supports flow rates of 22.7 liters (6 US gallons) to 56.8 liters (15 US gallons) per minute.
- ▶ The doors have no moving parts; the fans in the equipment move air through the heat exchanger as easily as a standard rack door.
- ▶ If the water flow is disrupted, the rack reverts to standard air cooling.

Rear Door Heat eXchanger performance

Each door can remove 100% of the heat that is generated by servers that use 30 kW of power and 90% of the heat that is generated by servers that use 40 kW. It removes the heat by using 18 °C (64 °F) water at a 27 °C (81 °F) server inlet air temperature.

Figure 5-21 shows more information about the capability of this rear door heat exchanger.

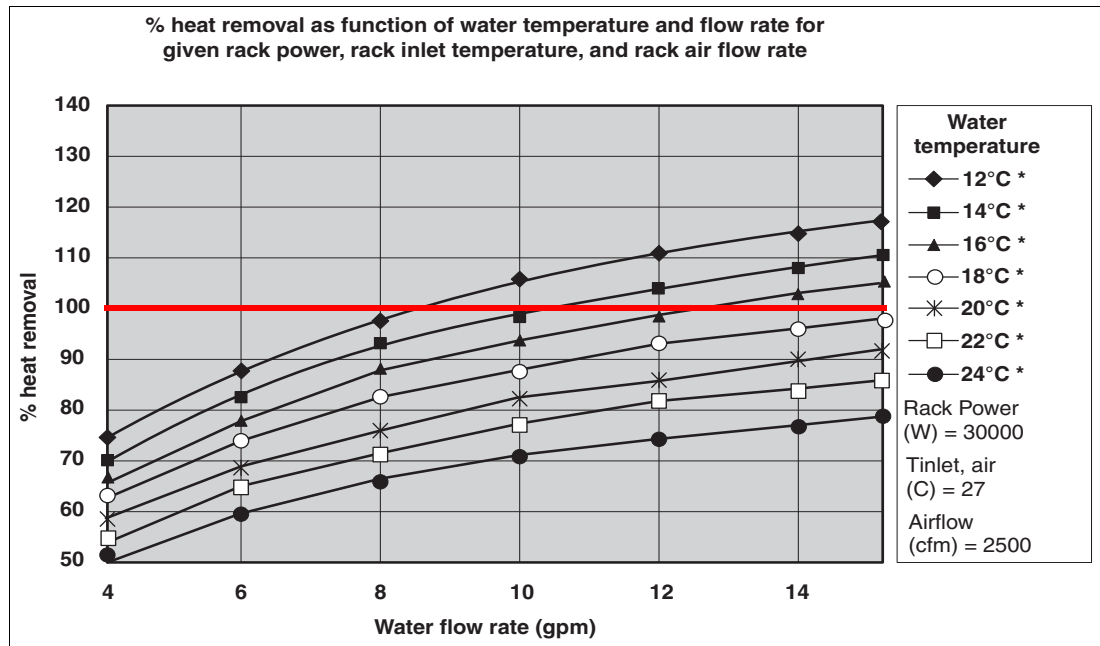


Figure 5-21 Heat removal performance with a 30 kW load

Although even more heat can be extracted if the water is cooler, the water temperature cannot be below the dew point in the server room or condensation forms on the rear door.

Some standard computer room air conditioning often is provisioned to control humidity and enable doors to be disconnected for maintenance or other requirements

The reduced air conditioning requirement typically saves about 1 KW per rack that is used to compress refrigerant and move air.

The reduction in air conditioner noise, which is coupled with the acoustic dampening effect of the heat exchangers and the decrease in high velocity cold air, makes the data center environment less hostile.

NeXtScale WCT System with Rear Door Heat eXchanger

Water cooling with NeXtScale WCT System often removes only approximately 85% of the heat that is generated by the system. The remaining heat is removed with air that is drawn through the system by the power supply fans. In 30 kW racks, this rate can equate to 4.5 kW of heat that is rejected to the facility and must be removed through other methods of cooling, such as traditional Computer Room Air Handling (CRAH) units.

The NeXtScale WCT System rack solution can be coupled with a Rear Door Heat eXchanger (RDHX) to remove 100% of the rack heat. This configuration can be accomplished by having two separate secondary water loops (one for the WCT rack and one for the RDHX). Another option consists of a single secondary water loop with a water temperature of approximately 25 °C. Operating the RDHX at approximately 25 °C allows for up to 10 kW of heat removal from a rack.

Most management and storage racks that are typical to HPC solutions use less than 10 kW. Therefore, on a single secondary loop water temp of 25 °C, Lenovo can support a rack of NeXtScale WCT Systems and RDHX equipped racks.

Figure 5-22 shows a typical connection of an RDHX to the secondary water loop. This configuration simplifies the data center infrastructure that is required to support water-cooled solutions and is unique in the industry. In the current form, the facility must provide only one supply and one return for each rack. The connection type is the same across the two different rack setups, by using 1-inch (25.4 mm) Eaton ball valves that are currently specified for connecting a NeXtScale WCT Manifold assembly to the secondary loop.

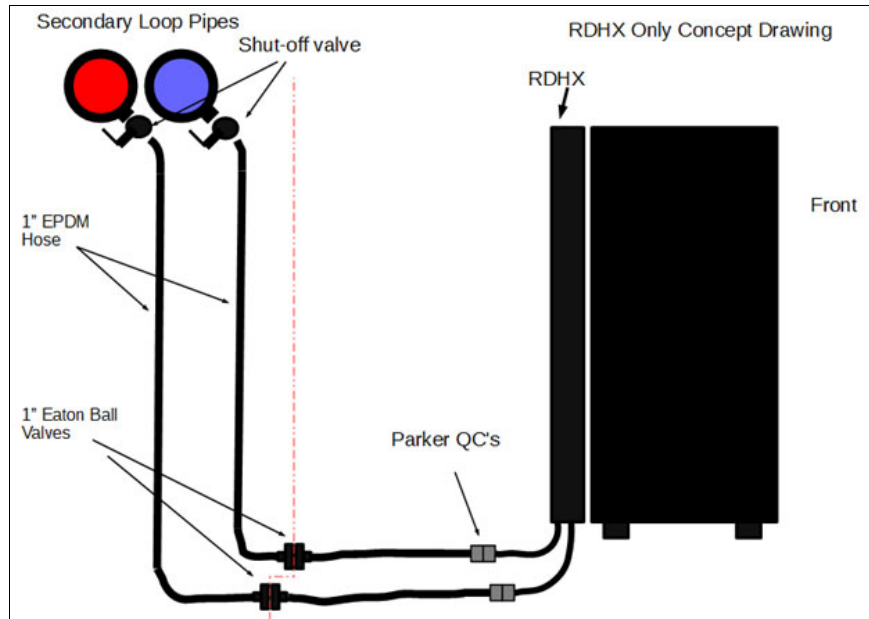


Figure 5-22 Storage or management rack with the RDHX

Extending this concept further, a single loop can support NeXtScale WCT Systems and RDHX that are installed on one rack. The RDHX removes the excess heat that is rejected from the WCT chassis, which makes the rack system cooled with water. Figure 5-23 on page 93 shows a WCT Manifold assembly and an RDHX that is connected to a single secondary water loop.

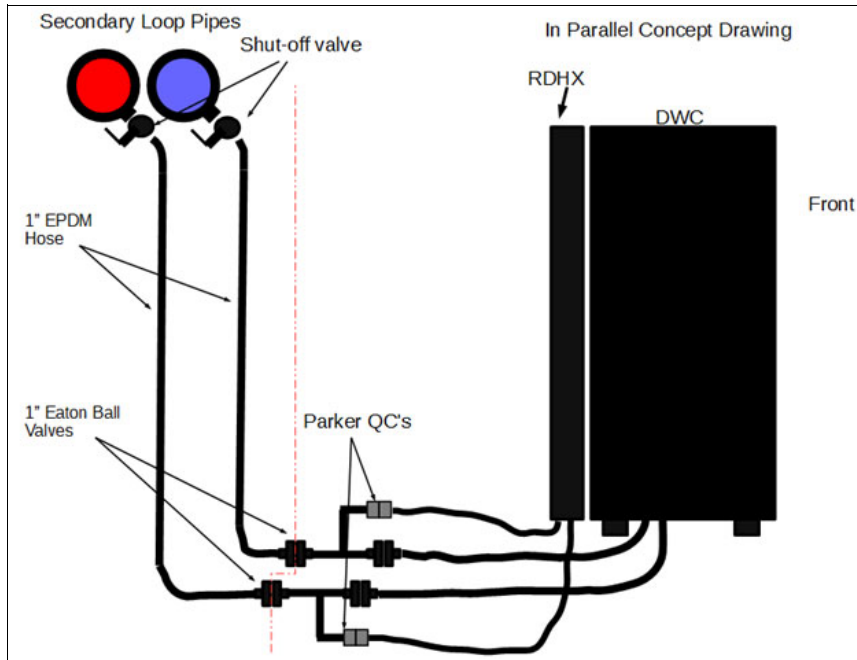


Figure 5-23 NeXtScale WCT System rack with RDHX, 100% heat-to-water

5.7 Top-of-rack switches

NeXtScale System does not include integrated switching in the chassis, unlike BladeCenter or Flex System. Integrated switching was not included to maximize the modularity and flexibility of the system and prevent any chassis-level networking contention. The Intelligent Cluster offering includes various switches from Lenovo Networking and others that can be configured and ordered as an integrated solution. Users who are building solutions (including NeXtScale System) can combine them with switches of their own choosing. In the following sections, we describe the switches that are available from Lenovo.

Note: There is no rule that smaller rack mountable switches (typically 1U or 2U) must be mounted in the top of a rack, but “top-of-rack” is the common name they were given. This idea is in contrast to larger switches that are deployed for rows of rack-mounted servers, which are referred to as “end-of-row” switches; or the often large, modular switches, which also often contain routing functionality, which are commonly called “core switches”.

5.7.1 Ethernet switches

NeXtScale System features forward-facing cabling. To connect cables from the NeXtScale servers to Ethernet switches, it is easiest to mount the switches in the racks with the switch ports facing the front of the racks as well.

It is important that the cooling air for the switches flow from the front (port side) of the switch to the rear (non-port side). The switches that are listed in Table 5-18 on page 94 meet this criteria. Switches that are cooled from rear to front can be used, but these switches must be mounted facing the rear of the rack. Also, all the cables from the servers are routed from the front of the rack to the back to connect.

Table 5-18 Top-of-rack switches

Part number	Description
1 Gb top-of-rack switches	
715952F	Lenovo RackSwitch™ G8052 (Front to Rear)
10 Gb top-of-rack switches	
7159BF7	Lenovo RackSwitch G8124E (Front to Rear)
715964F	Lenovo RackSwitch G8264 (Front to Rear)
7159DFX	Lenovo RackSwitch G8264CS (Front to Rear)
7159CFV	Lenovo RackSwitch G8272 (Front to Rear)
7159GR5	Lenovo RackSwitch G8296 (Front to Rear)
40 Gb top-of-rack switches	
7159BFX	Lenovo RackSwitch G8332 (Front to Rear)
Rail kit	
00CG089	Recessed 19-inch 4-Post Rail Kit

The Recessed 19-inch, 4-Post Rail Kit (00CG089) locates the front of the switch 75 mm behind the front posts of the rack to provide more cable bend radius. The rail kit is recommended for mounting 1U switches in racks with NeXtScale chassis.

5.7.2 InfiniBand switches

The InfiniBand switches (as listed in Table 5-19) are available as part of the Intelligent Cluster offering. As with the Ethernet switches that require cooling from the ports to the back of the switch, the InfiniBand switches need the same air flow path. However, on the InfiniBand switches, it is referred to as opposite port side exhaust (oPSE). It is recommended that the System Networking Recessed 19-inch 4-Post Rail Kit is installed for cable bend radius reasons.

Table 5-19 InfiniBand switch feature codes

Feature code	Description
A2EZ	Mellanox SX6036 FDR10 InfiniBand Switch (oPSE)
A2Y7	Mellanox SX6036 FDR14 InfiniBand Switch (oPSE)
6676	Intel 12200 QDR IB Redundant Power Switch (oPSE)
6925	Intel 12200 QDR InfiniBand Switch (oPSE)

5.7.3 Fibre Channel switches

For I/O intensive applications, 8 Gb and 16 Gb Fibre Channel networks (which are compatible with 8 Gb and 4 Gb storage devices) are popular because of their high I/Os per second (IOPS) capability and high reliability. In larger clusters or systems with high I/O demands, there are several nodes that are connected to SAN storage, which then share storage via a parallel file system (such as IBM GPFS) to the rest of the servers. Table 5-20 on page 95 lists the current 16 Gb Fibre Channel switch offerings. These switches should be mounted facing

the rear of the rack because they all have rear to front (port side) cooling air flow. The fiber optic connections must pass from the adapters that are installed in the front of the NeXtScale nodes to the ports at the rear of the rack.

Table 5-20 Fibre Channel switch part numbers

Part number	Description
16 Gb Fibre Channel switches	
2498F24	IBM System Storage SAN24B-5
2498F48	IBM System Storage SAN48B-5
2498F96	IBM System Storage SAN96B-5
8 Gb Fibre Channel switches	
249824E	IBM System Storage SAN24B-4 Express
241724C	Cisco MDS 9124 Express
2417C48	Cisco MDS 9148 for IBM System Storage
2498B80	IBM System Storage SAN80B-4

For more information about Fibre Channel switches, see this website:

<http://www.ibm.com/systems/networking/switches/san/>

5.8 Rack-level networking: Sample configurations

In this section, we describe the following sample configurations that use NeXtScale systems:

- ▶ InfiniBand non-blocking
- ▶ InfiniBand 50% blocking
- ▶ 10 Gb Ethernet configuration with one port per node
- ▶ 10 Gb Ethernet configuration with two ports per node

The location of the chassis and the switches within the rack are shown in a way that optimizes the cabling of the solution. The chassis and switches are color-coded to indicate which InfiniBand or Ethernet switches support which chassis.

Management network: Management networking is the same for all configurations. For more information, see 5.8.5, “Management network” on page 100.

For more information about networking with NeXtScale System, see *NeXtScale System Network and Management Cable Guide*, which is available at this website:

<http://ibm.com/support/entry/portal/docdisplay?lnocid=LNVO-POWINF>

5.8.1 Non-blocking InfiniBand

Figure 5-24 shows a non-blocking InfiniBand configuration. On the left side is a rack with six chassis for a total of 72 compute nodes. Four 36-port InfiniBand switches and two 48-port 1 Gb Ethernet switches are used.

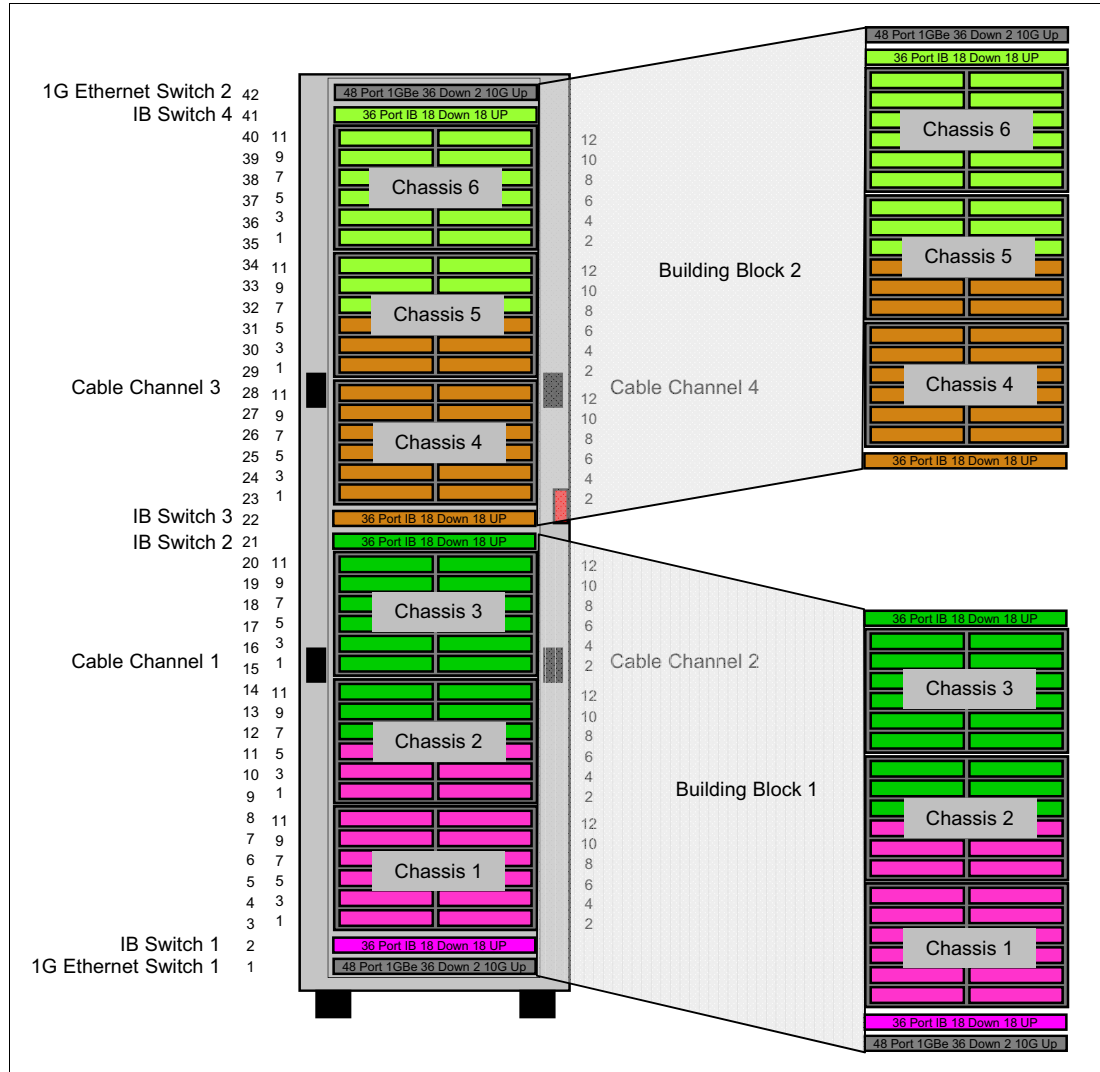


Figure 5-24 Non-blocking InfiniBand

5.8.2 A 50% blocking InfiniBand

Figure 5-25 shows a 50% blocking (two node ports for every uplink port) InfiniBand configuration. On the left side is a rack with six chassis for a total of 72 CPU nodes. Three 36-port InfiniBand switches and two 48-port 1 Gbps Ethernet switches provide the network connectivity.

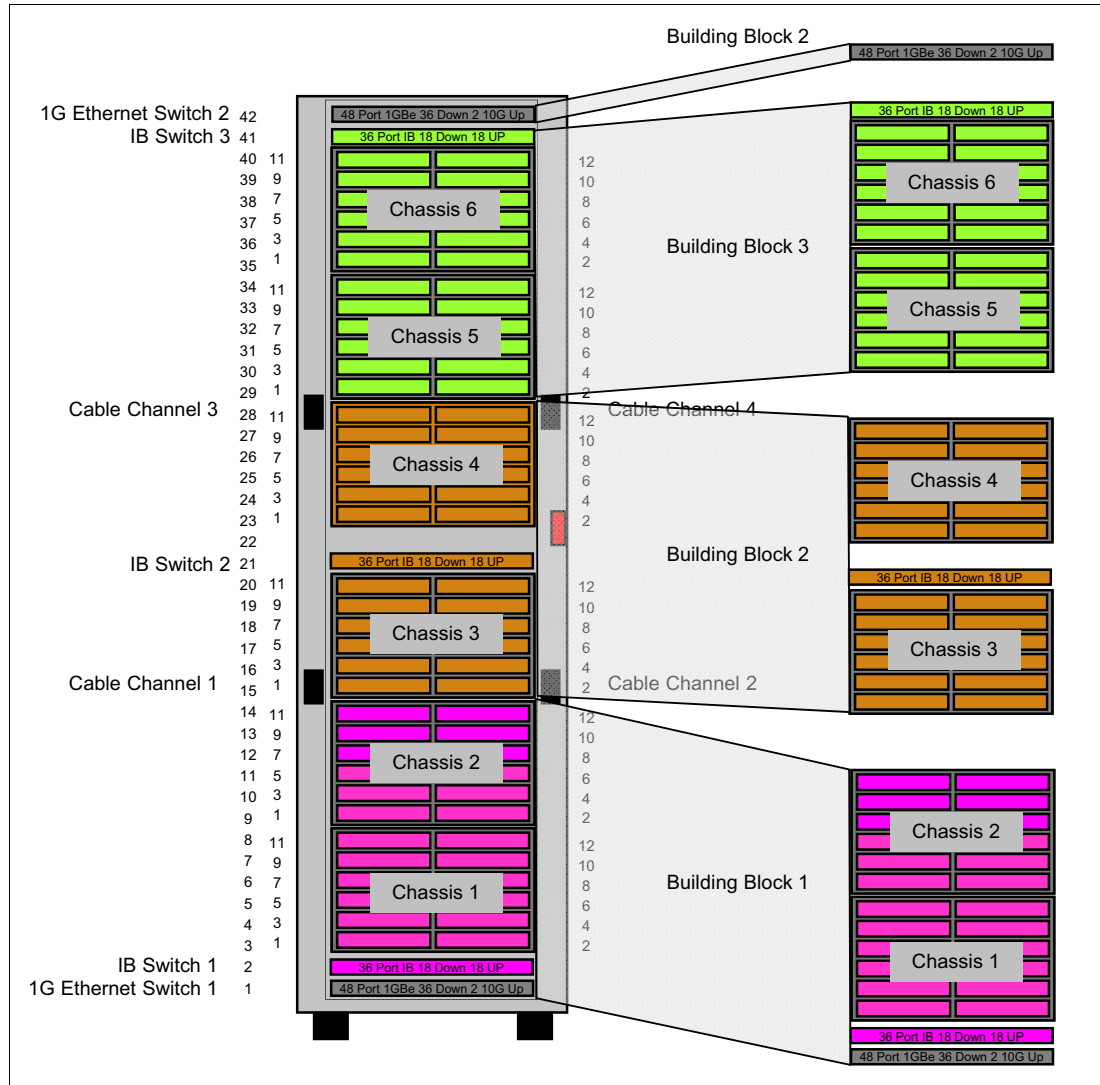


Figure 5-25 A 50% Blocking InfiniBand

Filler panel: Filler panels (part number 00Y3011) are placed in rack unit 21 and 41 to prevent hot air recirculation.

5.8.3 10 Gb Ethernet, one port per node

Figure 5-26 shows a network with one 10 Gb Ethernet connection per compute node. On the left side is a rack with six chassis for a total of 72 CPU nodes. Two 48 port 10 Gb switches and two 48 port 1 Gbps Ethernet switches provide the network connectivity.

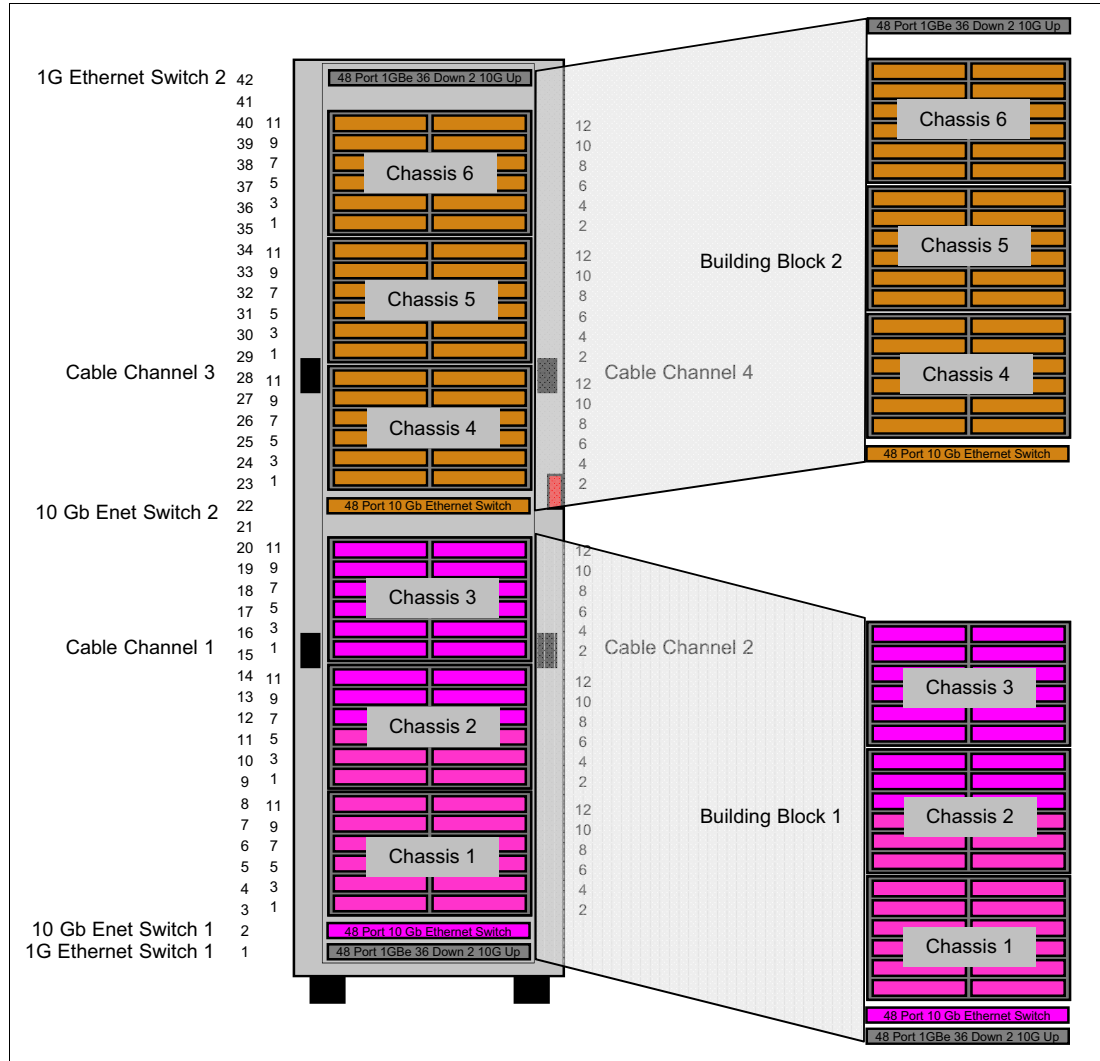


Figure 5-26 10 Gb Ethernet, one port per node configuration

Filler panel: 1U filler (part number 00Y3011) is placed in rack units 21 and 41 to prevent hot air recirculation.

5.8.4 10 Gb Ethernet, two ports per node

Figure 5-27 shows a network that consists of two 10 Gb Ethernet connections per compute node. On the left side is a rack with six chassis for a total of 72 compute nodes. Three 48-port 10 Gb Ethernet switches and two 48-port 1 Gb Ethernet switches provide the network connectivity.

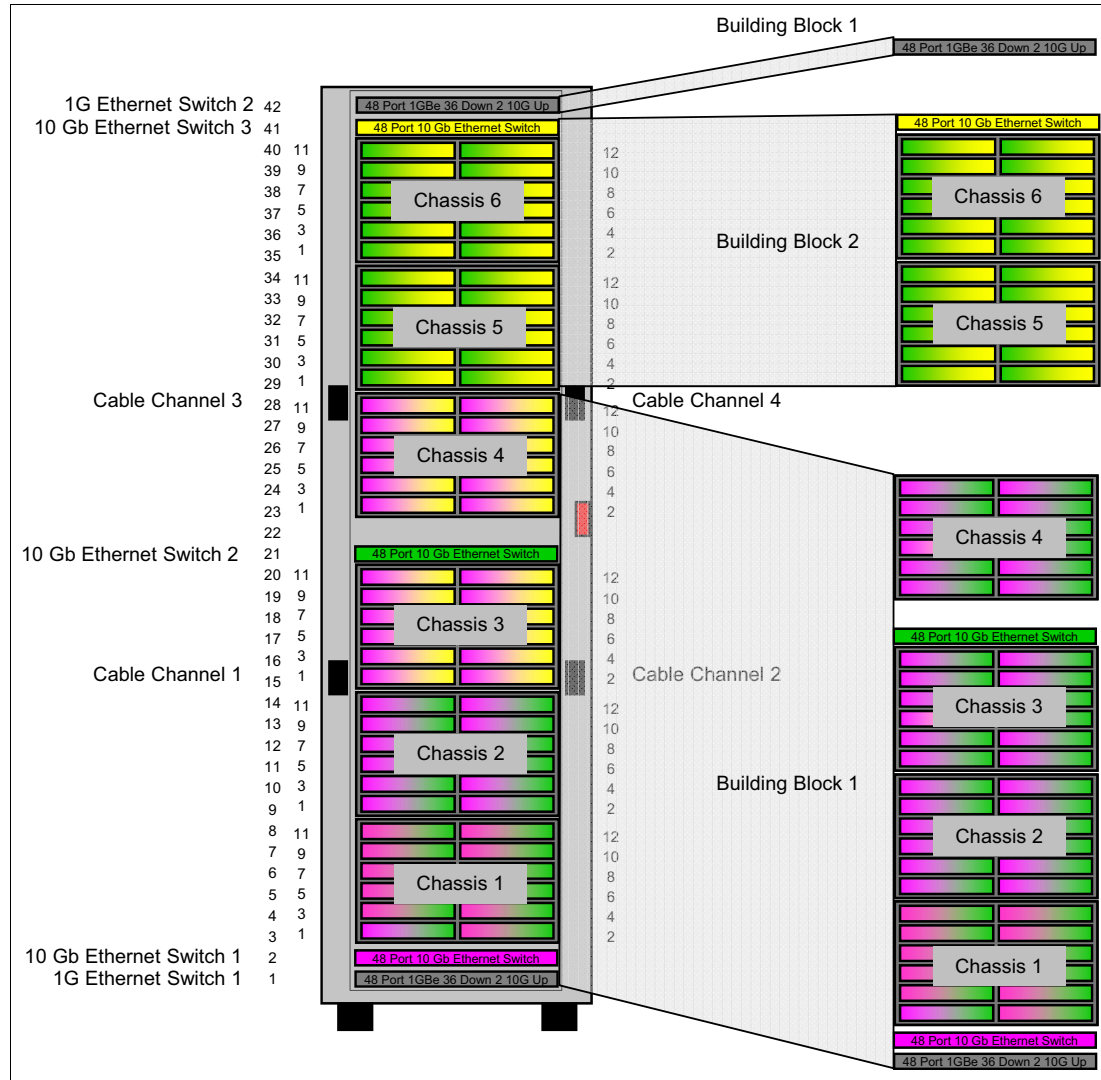


Figure 5-27 10 Gb Ethernet, two ports per compute node

The location of the chassis and the switches within the rack are shown in a way that optimizes the cabling of the solution. The chassis and switches are color-coded to indicate which InfiniBand or Ethernet switches support which chassis.

In Figure 5-27, each node has two colors, which indicates that each node is connected to two different switches to provide redundancy.

Filler panel: A 1U filler (part number 00Y3011) is placed in rack unit 22 to prevent hot air recirculation.

5.8.5 Management network

Figure 5-28 shows the 1 Gb Ethernet management network for a solution with four monitored PDUs. The 1 Gb management switches are shown in rack units 1 and 42.

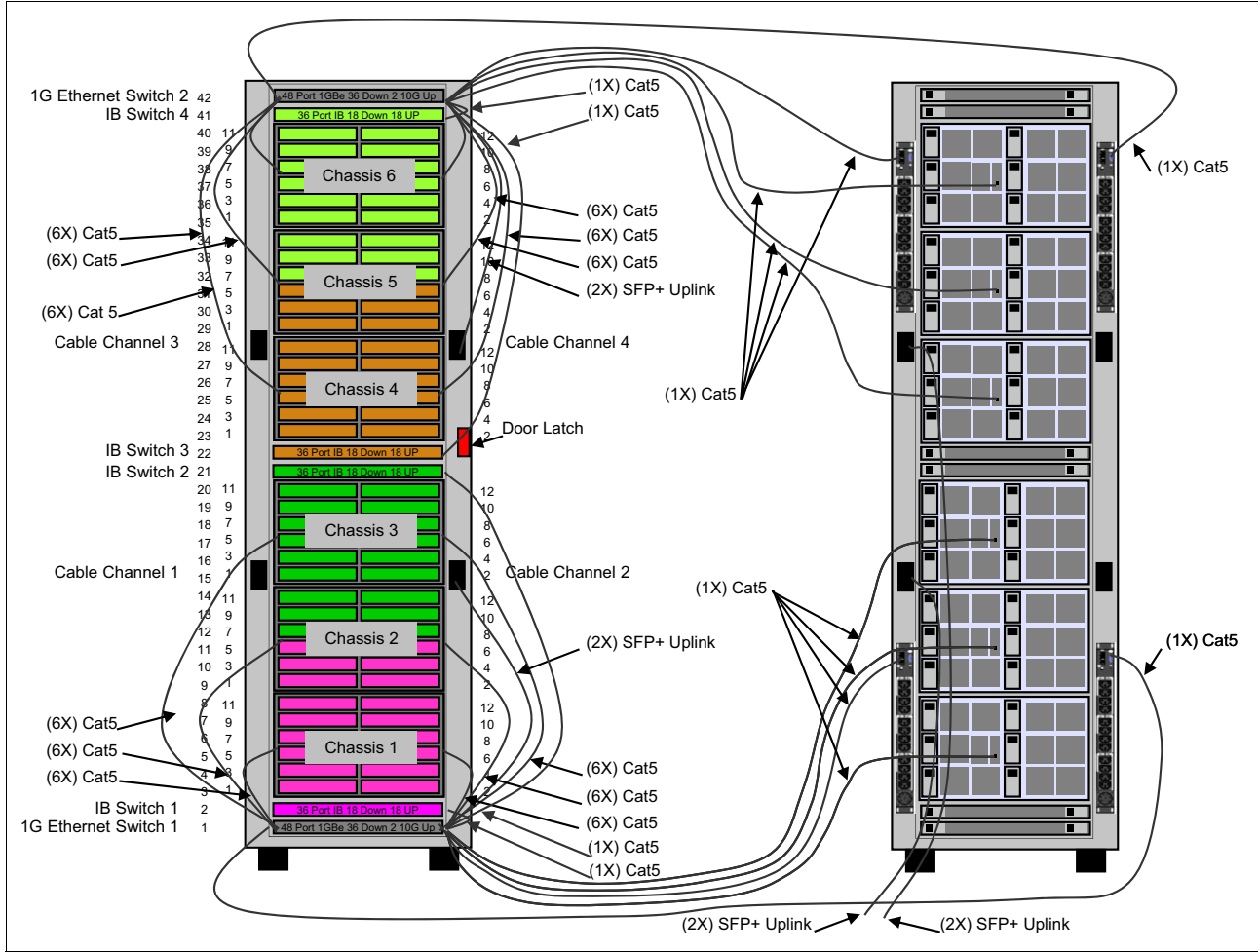


Figure 5-28 Management network

Each chassis has one management port that plugs to the Fan and Power Controller, which is at the rear of the chassis. Each PDU can also have a management port.

Note: The management cables that connect to devices at the rear of the chassis should be routed to the front of the chassis via the cable channels.

Factory integration and testing

NeXtScale WCT System is fulfilled as part of an integrated solution through the Lenovo Intelligent Cluster. Lenovo provides factory integration and testing as part of the Intelligent Cluster offering.

This chapter describes Lenovo Intelligent Cluster, what is provided by Lenovo factory integration, the testing that is performed, and the documentation that is supplied.

This chapter includes the following topics:

- ▶ 6.1, “Lenovo Intelligent Cluster” on page 102
- ▶ 6.2, “Lenovo factory integration standards” on page 102
- ▶ 6.3, “Factory testing” on page 103
- ▶ 6.4, “Documentation provided” on page 105

6.1 Lenovo Intelligent Cluster

Deploying solutions for Technical Computing, High Performance Computing (HPC), Analytics, and Cloud environments can place a significant burden on IT staff. Through Intelligent Cluster, Lenovo brings its expertise in HPC design, deployment, applications, and support to reduce risk, speed up time to deployment, and ease the integration into a client's environment. Intelligent Cluster reduces the complexity of deployment with pre-integrated and interoperability-tested solutions, which are delivered, installed¹, and supported by Lenovo as end-to-end solutions.

Intelligent Cluster features industry-leading System x servers, storage, networking, software, and third-party components with which clients can choose from various technologies and design a tailored solution for the clients applications and environment. Lenovo thoroughly tests and optimizes each solution for reliability, interoperability, and maximum performance so that the system can be quickly deployed.

In Lenovo's manufacturing plant, the systems are fully assembled and integrated in the racks (including all cables). As a result, the delivery of a single rack solution requires only to connect the PDUs to the data center power, the manifold to the water-cooling loop, and switches uplinks to the data center network to be ready for power-on. A multiple rack solution also requires that the inter-rack cabling is done (one side of the cable being already connected, the other side being labeled with location to which it must be connected).

Lenovo applies in manufacturing a firmware stack that matches the "best-recipe" that is devised by our cluster development team for solution level interoperability.

Intelligent Cluster solutions are built, tested, delivered, installed¹, and supported by Lenovo as a single solution instead of being treated as hundreds of individual components. Lenovo provides single point-of-contact, solution-level support that includes System x and third-party components (such as those from Intel and Mellanox) to deliver maximum system availability throughout the life of the system, so clients can spend less time maintaining systems.

Although open systems can be racked, cabled, configured, and tested by users, we encourage clients to evaluate the benefits of having Lenovo integrate and test the system before delivery. We also suggest contacting the Lenovo System x Enterprise Solution Services cluster enablement team to speed up the commissioning after the equipment arrives.

6.2 Lenovo factory integration standards

Lenovo standards for factory integration are based on meeting a broad range of criteria, including the following criteria:

- ▶ Racks are one standard floor tile wide, and fit through 80-inch doorways.
- ▶ Cabling maintains proper bend radius for all cable types while not impeding maintenance access to any devices and allows rack doors to be closed (and locked, if required).
- ▶ All components are cabled within the rack.

Also, where practical, inter-rack cabling is connected at one end, coiled up, and temporarily fastened in the rack for shipping.

- ▶ All cables are labeled at each end, with the source and destination connection information printed on each label.

¹ The solution's installation can be performed by Business Partner or user, if wanted.

- ▶ All components are mounted inside the racks for shipping².

These standards govern the range of systems that can be factory-integrated by Lenovo and we consider them to follow best practices that are based on our design criteria. There are several alternative configuration options that are architecturally sound and based on different design criteria, as shown in the following examples:

- ▶ Not requiring rack doors: This option allows for bigger bundles of copper cables to traverse the fronts of the racks without impeding node access, which can allow more nodes per rack.
- ▶ Use of optical cabling: This option allows for smaller cable bundles, so more nodes per rack or more networks per node can be provisioned without impeding node access.
- ▶ Use of taller racks: This option allows for more chassis and nodes per rack.
- ▶ Allowing connections to switches outside the rack; for example, stacked on top of the rack.

Note: The responsibility for assuring adequate cooling of components, access for maintenance, adequate cable lengths before integration, and other considerations become the responsibility of the person or team that is configuring their own solution.

6.3 Factory testing

The Intelligent Cluster manufacturing test process is intended to meet the following objectives:

- ▶ Assure that the integrated hardware is configured and functional.
- ▶ Identify and repair any defects before or that were introduced by the integrated rack assembly process.
- ▶ Validate the complete and proper function of a system when configured to a customer's order specifications.
- ▶ Apply the current released Intelligent Cluster best-recipe, which includes lab-tested firmware levels for all servers, adapters, and devices.

The following tasks are typical of the testing that is performed by Lenovo at the factory. Other testing might be done based on unique hardware configurations or client requirements:

- ▶ All servers are fully tested as individual units before rack integration.
- ▶ After the components are installed in the rack, there is a post assembly inspection to assure that they are installed and positioned correctly.
- ▶ Lenovo configures all switches; terminal servers; keyboard, video, and mouse (KVM) over IP units, and other devices with IP addresses and host names to allow for communication and control.

These components can be set by using client-supplied information or to a default scheme. For more information about the default scheme, see this website:

http://download.boulder.ibm.com/ibmdl/pub/systems/support/system_x_pdf/intelligent_cluster_factory_settings_102411.pdf

² In rare cases, some components might ship outside of the racks if their location within a rack might lead to the rack tilting during shipment.

- ▶ Power redundancy testing

If there is a client-provided redundant power domain scheme, Lenovo tests with one domain that is powered down, then the opposite domain that is powered down, to ensure that all devices with redundant power feeds stay powered on.

Otherwise, remove and restore power from each PDU to assure all devices with redundant power feeds stay powered on.

- ▶ Flash all servers, adapters, and devices to current Lenovo development-provided best recipe.

- ▶ From a cluster management node, discover servers and program their integrated management module (IMM). This configuration allows for remote control of the computational infrastructure.

- ▶ Serial (console) over LAN (SoL) setup

Unless the cluster has terminal servers, SoL is configured and tested.

- ▶ Set up RAID arrays on local disks as defined by the client or client architect, or per Intelligent Cluster best practices.

- ▶ Set up shared storage devices, configure arrays with all disks present, and create and initialize logical disks to verify functionality.

- ▶ Install (for nodes with hard disk drives), or push out (for diskless nodes) an operating system to all servers to verify functionality of the following components:

- Server hardware

- Ethernet, InfiniBand, and Fibre Channel switches, terminal servers, and KVM switches

- Server configuration correctness, including memory configurations, correct CPU types, storage, and RAID

- ▶ Perform High-Performance Linpack (HPL) benchmarking on the system. This testing ensures that the following conditions are met:

- CPU and memory are stressed to the extent that is commonly found in production environments.

- Thermal stress testing is performed.

- Interconnect networks are exercised. HPL is run over the high bandwidth, low latency network. This testing is performed in groups over leaf and edge switches or at the chassis level. If no high-speed network is available, manufacturing performs a node level test.

As an added benefit, clients can see a performance baseline measurement, if requested.

Note: The HPL benchmark is run with open source libraries and no tuning parameters. Optimization is recommended for those users who are interested in determining maximum system performance.

For a NeXtScale WCT System, specific pressure tests also are conducted to check for leaks in the water-carrying parts. These tests use air or nitrogen and are performed on all of the cooling loops and manifolds. The following tests are performed via the use of a software and a pressure transducer:

- ▶ Cooling loop pressure test

This test checks for leaks in the loop that can be caused during shipping. Taking into account temperature, the software reviews the pressure that the vendor shipped the cooling loop at and compares it to the pressure Lenovo manufacturing reads before node

assembly. If the pressure that Lenovo manufacturing reads is within 15 psi of what the pressure shipped the loop at, the test passes.

- ▶ **Manifold decay test (before assembly)**

This test checks for leaks in the manifold and rack that can be caused during shipment. Before rack assembly, the manifold is pressurized with dry, oil-free, filtered air to 45 psi. The test allows for the air to stabilize and then takes an initial pressure reading. After 10 minutes, a second pressure reading is taken. If the second pressure reading is less than 1% of the initial reading, the test passes.

- ▶ **Rack decay test (before fill water)**

This test is performed after the manifolds are rinsed and all trays are installed. The purpose of this test is a final leak test before water is applied to the nodes. This test lasts for 12 hours. Before rack fill, the manifold is pressurized with nitrogen to 45 psi. The test allows for the nitrogen to stabilize and then takes an initial pressure reading. After 12 hours, a second pressure reading is taken. If the second pressure reading (temperature adjusted) is less than 3% of the initial reading, the test passes.

- ▶ **Rack decay test (before pack)**

This test serves two purposes: it purges all of the air out the rack (air and water vapor can create corrosion, which can cause a future leak) and it pressurizes the rack to 25 psi and a pressure decay test is run similar to the one done on the manifold before rack assembly.

At the end of each test, a label is printed that includes a concatenation of the test name, date, test starting pressure, and test ending pressure that was measured at the completion of a passing test.

In addition, Lenovo manufacturing checks that the leak sensors are installed and no error is logged in the FPC error log.

The Intelligent Cluster testing is performed with the goal of detecting and correcting all system issues before the cluster is shipped so that the time from installation to production is minimized. This testing is meant to verify the hardware and cabling only, on a per rack basis. Any software installations require other Cluster Enablement Team, Lenovo System x Enterprise Solution Services, or third-party services.

Note: All servers with local disks have boot sectors that are wiped before they are shipped. All shared storage devices have test arrays deconstructed.

6.4 Documentation provided

This section describes the documentation Lenovo provides with the hardware. Lenovo manufacturing uses the extreme Cloud Administration Toolkit (xCAT) to set up and test systems. xCAT is also used to document the manufacturing that is set up. Because it is a popular administration utility, the xCAT table files are included with the system. The team that is setting up the cluster at the client site then uses these files in commissioning a cluster.

The following documentation is also provided in a binder with each rack and on a CD with each rack:

- ▶ MAC addresses of the Ethernet interfaces
- ▶ Machine type, model number, and serial number of devices

- ▶ Firmware levels: For servers, these levels include:
 - UEFI version
 - IMM version
 - On system diagnostics version
- ▶ Memory per server
- ▶ CPU type per server
- ▶ Proof that an operating system was running on each server, reporting the output of the `uname -r` command for each node
- ▶ PCI adapter list for each server by using the `lspci` command
- ▶ Configuration files for switches

6.4.1 HPLinpack testing results: Supplied on request

The results of the HPLinpack testing can be provided to clients, if requested. The documentation includes the following components:

- ▶ A listing of the software stack that is used, for example:
 - HPL-2.0 with the Intel timestamping patch
 - OpenBLAS
 - Mellanox OFED
 - gcc
 - MVAPICH2
 - RHEL
- ▶ A run summary, which lists GFLOPS and run time
- ▶ The detailed text output from the HPL benchmark
- ▶ Output graph with the following axis, as shown in Figure 6-1 on page 107:
 - Left y-axis = Gflops
 - Bottom x-axis = Percentage completed
 - Top x-axis = Wall clock
 - Right y-axis = Efficiency

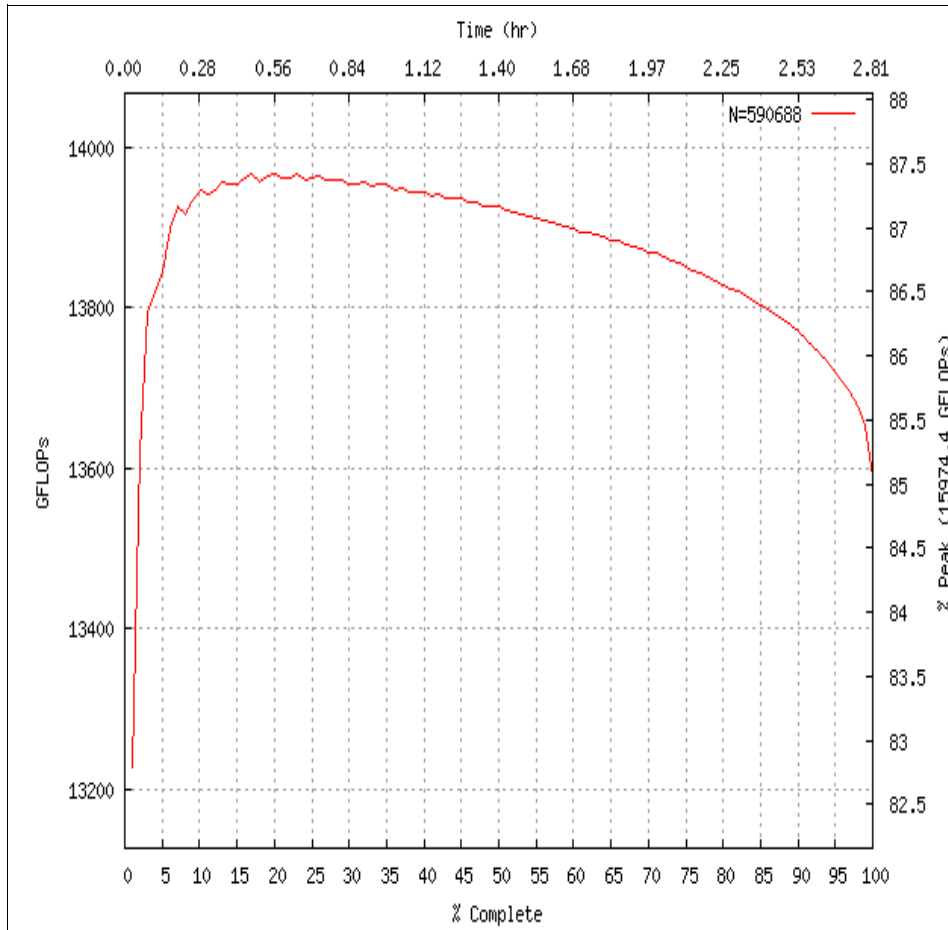


Figure 6-1 Example of HPLinpack output graph

Managing a NeXtScale environment

This chapter describes the available options for managing a NeXtScale System environment.

We describe the management capabilities and interfaces that are integrated in the system, and the middleware and software layers that are often used to manage collections of systems.

This chapter includes the following topics:

- ▶ 7.1, “Managing compute nodes” on page 110
- ▶ 7.2, “Managing the chassis” on page 123
- ▶ 7.3, “ServeRAID C100 drivers: nx360 M4” on page 148
- ▶ 7.4, “Integrated SATA controller: nx360 M5” on page 148
- ▶ 7.5, “VMware vSphere Hypervisor” on page 148
- ▶ 7.6, “eXtreme Cloud Administration Toolkit” on page 149

7.1 Managing compute nodes

The NeXtScale System compute nodes include local and remote management capabilities.

Local management capabilities are provided through the keyboard, video, mouse (KVM) connector on the front of the server. By using the console breakout cable that is included with the chassis, you can directly connect to the server console and attach USB storage devices.

Remote management capabilities are provided through the Integrated Management Module II (IMM2). IMM2 also provides advanced service control, monitoring, and alerting functions.

By default, the nx360 compute nodes include IMM2 Basic; however, if more functionality is required, the IMM2 can be upgraded to IMM2 Standard or to IMM2 Advanced with Feature on Demand (FoD) licenses.

7.1.1 Integrated Management Module II

The Integrated Management Module II (IMM2) on the NeXtScale nodes are compliant with Intelligent Platform Management Interface version 2.0 (IPMI 2.0). By using IPMI 2.0, administrators can manage a system remotely out of band, which means that a system can be managed independently of the operating system or in the absence of an operating system, even if the monitored system is not powered on.

IPMI also functions when the operating system is started and offers enhanced features when used with system management software. The nodes respond to IPMI 2.0 commands to report operational status, retrieve hardware logs, or issue requests. The nodes can alert by way of the simple network management protocol (SNMP) or platform event traps (PET).

The IMM2 can be accessed and controlled through any of the following methods:

- ▶ Command-line interface (CLI): Telnet or Secure Shell (SSH)
- ▶ Web interface (if IMM2 Standard and Advanced FoD is provisioned)
- ▶ IPMI 2.0 (local or remote)
- ▶ The Advanced Settings Utility (ASU)
- ▶ SNMP v1 and v3

Figure 7-1 shows the available IMM2 access methods.

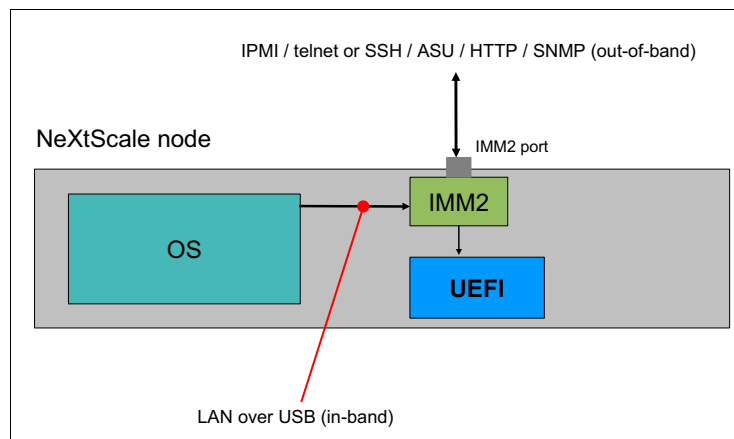


Figure 7-1 IMM2 access methods

Figure 7-2

IMM2 access on the nx360 M5 WCT is provided by the first Ethernet interface, which is shared with the operating system, as shown in Figure 7-3.

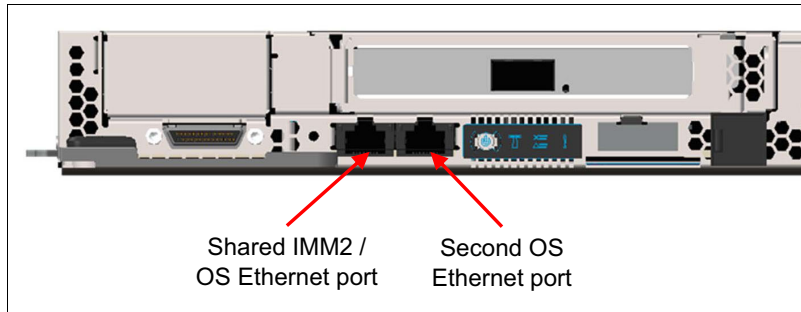


Figure 7-3 nx360 M5 WCT shared IMM2 and first Ethernet interface

For more information about the IMM2, see the following publications:

- ▶ IMM2 User's Guide:

<http://ibm.com/support/entry/portal/docdisplay?lnocid=MIGR-5086346>

- ▶ A white paper about transitioning to UEFI and IMM:

<http://ibm.com/support/entry/portal/docdisplay?lnocid=MIGR-5079769>

7.1.2 Unified Extensible Firmware Interface

The Unified Extensible Firmware Interface (UEFI) replaces BIOS in System x and BladeCenter servers. It is the new interface between the operating system and platform firmware. UEFI provides a modern, well-defined environment for booting an operating system and running pre-boot applications.

For more information about UEFI, see this website:

<http://www.uefi.org/home/>

UEFI provides the following improvements over BIOS:

- ▶ ASU now has complete coverage of system settings.
- ▶ On rack mount servers, UEFI settings can be accessed out-of-band by ASU and the IMM (not available on BladeCenter blades).
- ▶ Adapter configuration can move into F1 setup; for example, iSCSI configuration is now in F1 setup and consolidated into ASU.
- ▶ Elimination of beep codes: All errors are displayed by using light path diagnostics.
- ▶ DOS is not supported and does not work under UEFI.

UEFI adds the following functionality:

- ▶ Adapter vendors can add more features in their options (for example, IPv6).
- ▶ Modular design allows faster updates as new features are introduced.
- ▶ More adapters can be installed and used simultaneously; optional ROM space is much larger.
- ▶ BIOS is supported via a legacy compatibility mode.
- ▶ Provides an improved user interface.
- ▶ Replaces Ctrl key sequences with a more intuitive human interface.

- ▶ Adapter and iSCSI configuration are moved into F1 setup.
- ▶ Event logs are created that are more easily decipherable.
- ▶ Provides easier management.
- ▶ Reduces the number of error messages and eliminates outdated errors.
- ▶ A complete setup solution is provided by allowing adapter configuration function to be moved into UEFI.
- ▶ Complete out-of-band coverage by ASU simplifies remote setup.
- ▶ More functionality, better user interface, easier management for users.

For more information about the UEFI, see the IBM white paper, *Introducing UEFI-Compliant Firmware on IBM System x and BladeCenter servers*, which is available at this website:

<http://www.ibm.com/support/entry/portal/docdisplay?lnocid=MIGR-5083207>

UEFI system settings

Many of the advanced technology options that are available in the NeXtScale nx360 M4 compute node are controlled in the UEFI system settings. These settings affect processor and memory subsystem performance regarding power consumption.

The UEFI page is accessed by pressing F1 during the system initialization process. Figure 7-4 shows the UEFI settings main panel.

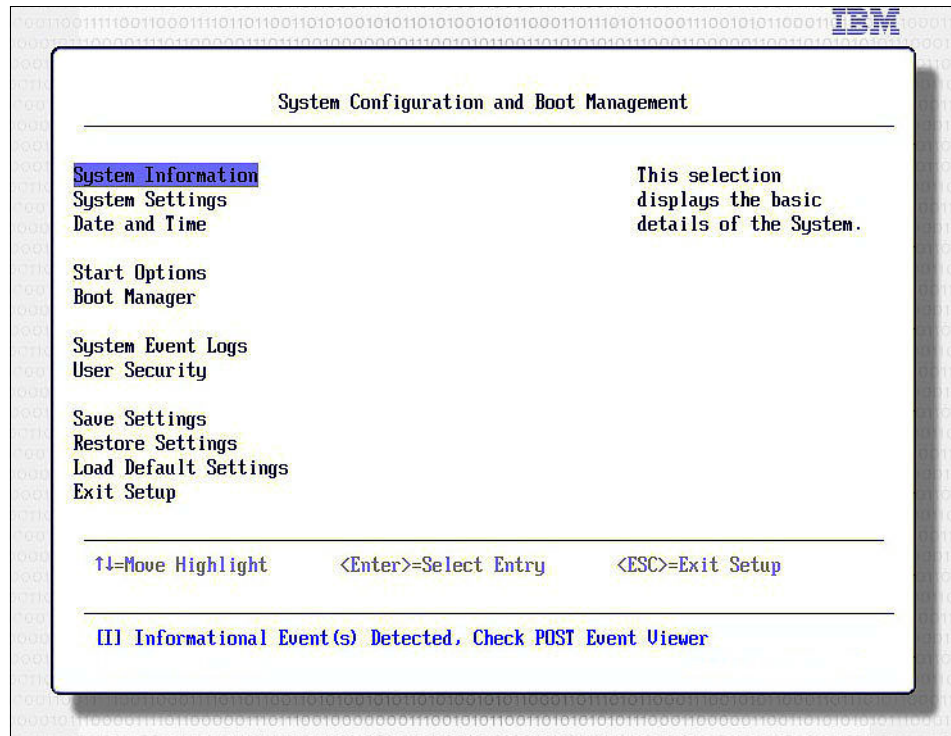


Figure 7-4 UEFI settings main panel

The compute node provides optimal performance with reasonable power usage, which depends on the operating frequency and voltage of the processors and memory subsystem.

In most operating conditions, the default settings provide the best performance possible without wasting energy during off-peak usage. However, for certain workloads, it might be appropriate to change these settings to meet specific power to performance requirements.

The UEFI provides several predefined setups for commonly wanted operation conditions. These predefined values are referred to as *operating modes*. Access the menu in UEFI by selecting **System Settings** → **Operating Modes** → **Choose Operating Mode**. You see the five operating modes from which to choose, as shown in Figure 7-5. When a mode is chosen, the affected settings change to the shown predetermined values.

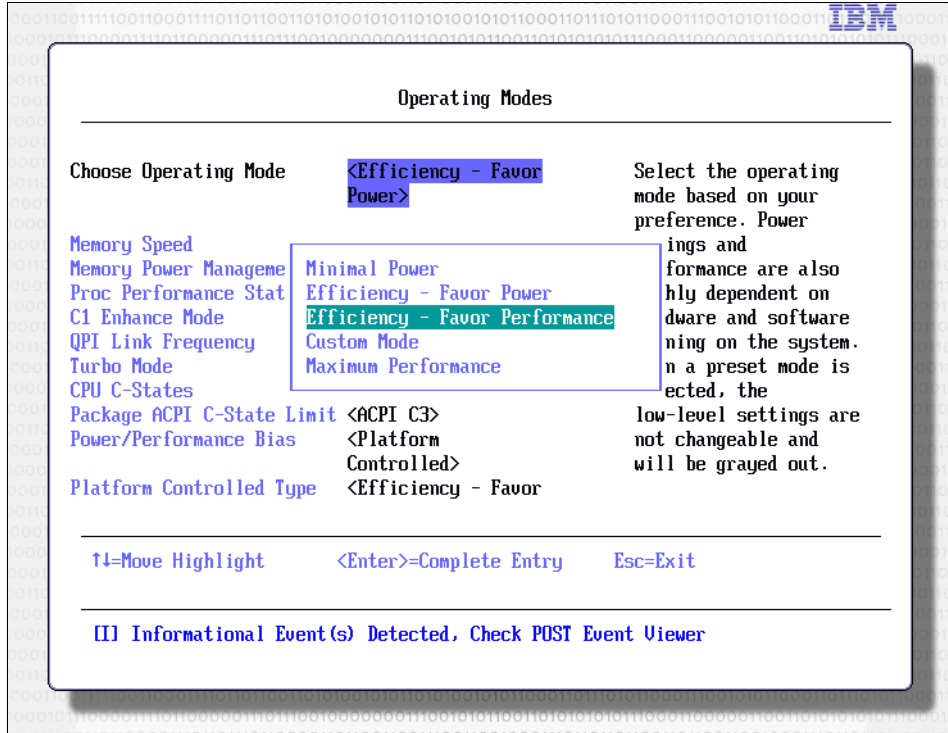


Figure 7-5 Operating modes in UEFI

We describe these modes in the following sections.

Minimal Power

Figure 7-6 shows the Minimal Power predetermined values. These values emphasize power-saving server operation by setting the processors, QPI link, and memory subsystem to a lowest working frequency. Minimal Power provides less heat and the lowest power usage at the expense of performance.

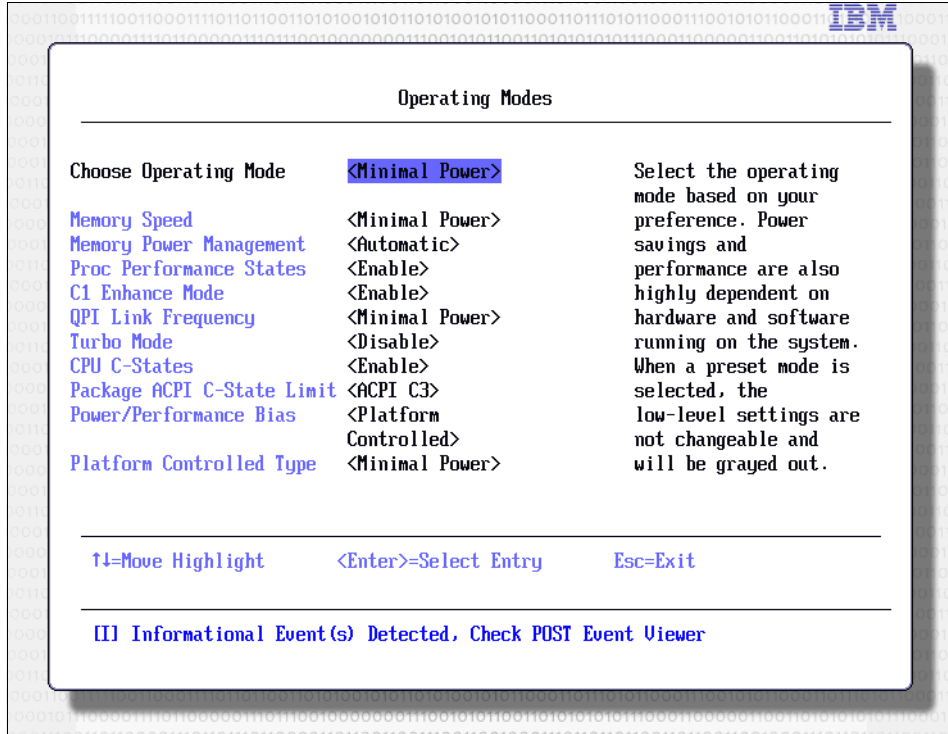


Figure 7-6 UEFI operation mode: Minimal Power

Efficiency - Favor Power

Figure 7-7 shows the Efficiency - Favor Power predetermined values. These values emphasize power-saving server operation by setting the processors, QPI link, and memory subsystem to a balanced working frequency. Efficiency - Favor Power provides more performance than Minimal Power, but favors power usage.

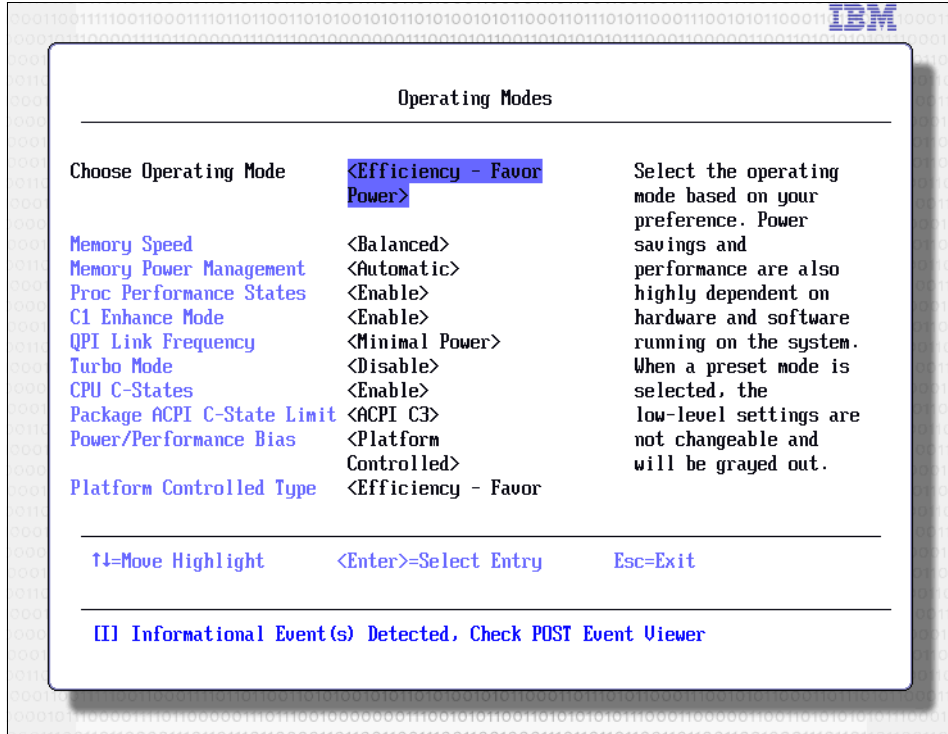


Figure 7-7 UEFI operation mode: Efficiency - Favor Power

Efficiency - Favor Performance

Figure 7-8 shows the Efficiency - Favor Performance predetermined values. These values emphasize performance server operation by setting the processors, QPI link, and memory subsystem to a high working frequency.

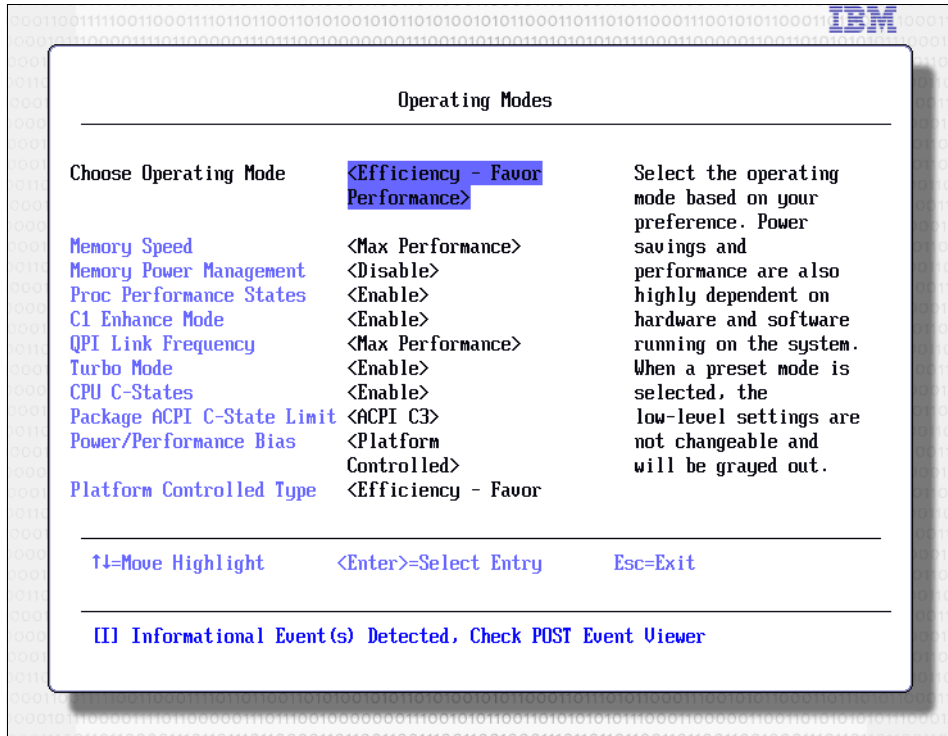


Figure 7-8 UEFI operation mode: Efficiency - Favor Performance

Custom Mode

By using Custom Mode, users can select the specific values that they want, as shown in Figure 7-9. The recommended factory default setting values provide optimal performance with reasonable power usage. However, with this mode, users can individually set the power-related and performance-related options.

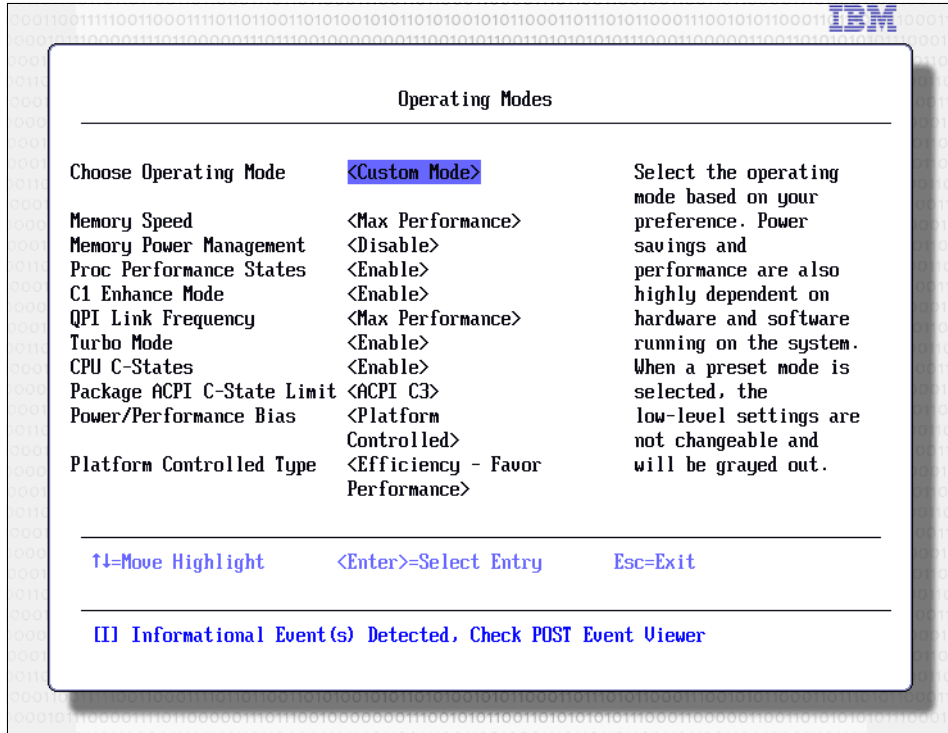


Figure 7-9 UEFI operation mode: Custom Mode

Maximum Performance

Figure 7-10 shows the Maximum Performance predetermined values. They emphasize performance server operation by setting the processors, QPI link, and memory subsystem to a maximum working frequency and the higher C-state limit. The server is set to use the maximum performance limits within UEFI. These values include turning off several power management features of the processor to provide the maximum performance from the processors and memory subsystem.

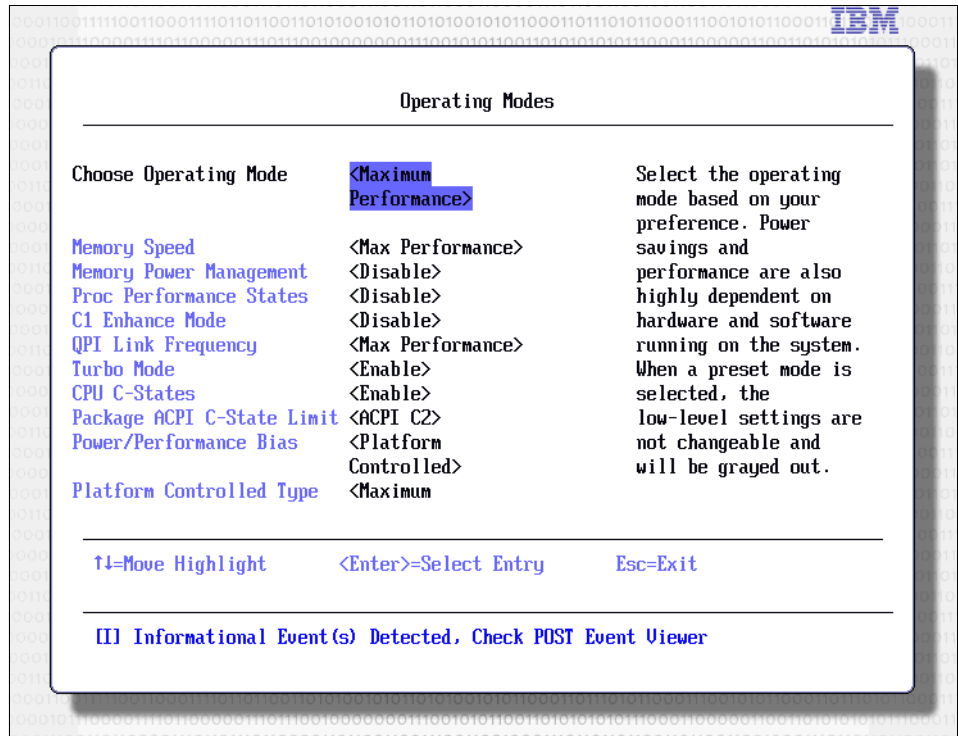


Figure 7-10 UEFI operation mode: Maximum Performance

Performance-related individual system settings

The UEFI default settings are configured to provide optimal performance with reasonable power usage. Other operating modes are also available to meet various power and performance requirements. However, individual system settings enable users to fine-tune the wanted characteristics of the compute nodes.

This section describes the UEFI settings that are related to system performance. In most cases, increasing system performance increases the power usage of the system.

Processors

Processor settings control the various performance and power features that are available on the installed Xeon processor.

Figure 7-11 shows the UEFI Processors system settings window with the default values.

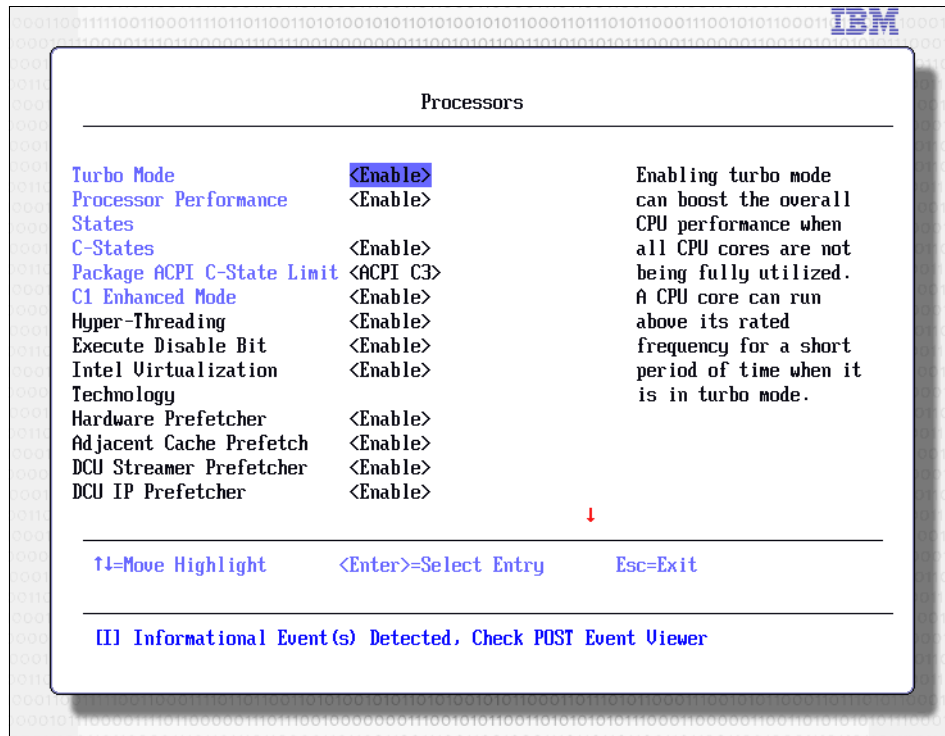


Figure 7-11 UEFI Processor system settings panel

The following processor feature options are available:

- ▶ Turbo Mode (Default: Enable)

This mode enables the processor to increase its clock speed dynamically if the CPU does not exceed the Thermal Design Power (TDP) for which it was designed.
- ▶ Processor Performance States (Default: Enable)

This option enables Intel Enhanced SpeedStep Technology that controls dynamic processor frequency and voltage changes, depending on operation.
- ▶ C-States (Default: Enable)

This option enables dynamic processor frequency and voltage changes in the idle state, which provides potentially better power savings.
- ▶ Package ACPI C-State Limit

Sets the higher C-state limit. A higher C-state limit allows the CPU to use less power when they are idle.
- ▶ C1 Enhanced Mode (Default: Enable)

This option enables processor cores to enter an enhanced halt state to lower the voltage requirement, and it provides better power savings.
- ▶ Hyper-Threading (Default: Enable)

This option enables logical multithreading in the processor so that the operating system can run two threads simultaneously for each physical core.
- ▶ Execute Disable Bit (Default: Enable)

This option enables the processor to disable the running of certain memory areas, which prevents buffer overflow attacks.

- ▶ Intel Virtualization Technology (Default: Enable)
This option enables the processor hardware acceleration feature for virtualization.
- ▶ Technology Hardware Prefetcher (Default: Enable)
This option enables the hardware prefetcher. Lightly threaded applications and some benchmarks can benefit from having it enabled.
- ▶ Adjacent Cache Prefetch (Default: Enable)
This option enables the adjacent cache line prefetch. Some applications and benchmarks can benefit from having it enabled.
- ▶ DCU Streamer Prefetcher (Default: Enable)
This option enables the stream prefetcher. Some applications and benchmarks can benefit from having it enabled.
- ▶ DCU IP Prefetcher (Default: Enable)
This option enables Instruction Pointer prefetcher. Some applications and benchmarks can benefit from having it disabled.

Memory

The Memory settings window provides the available memory operation options, as shown in Figure 7-12.

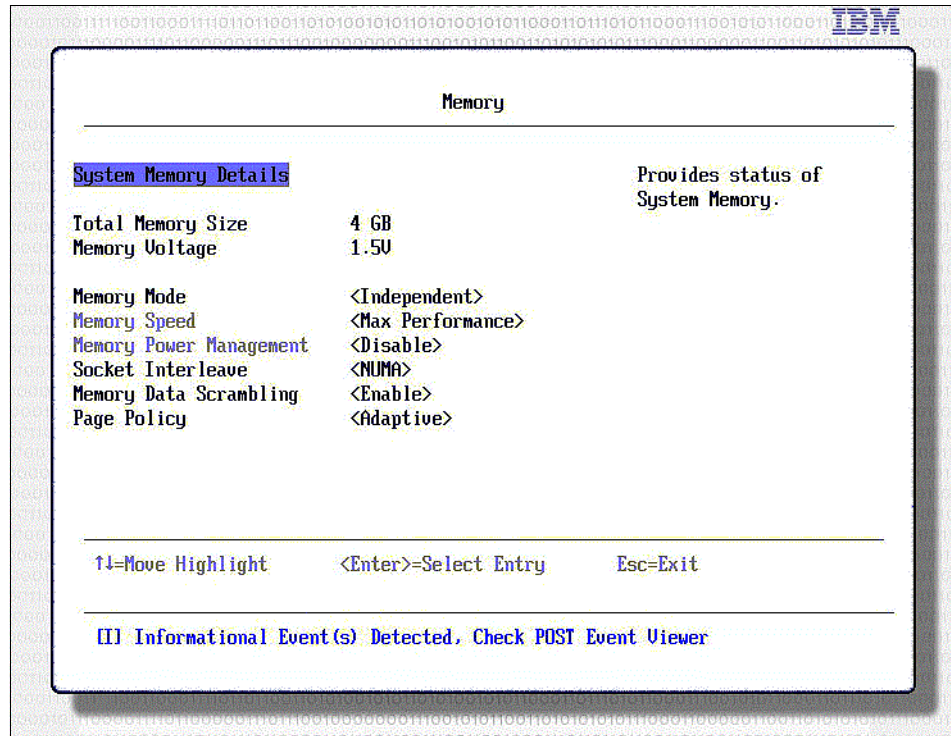


Figure 7-12 UEFI Memory system settings panel

The following memory feature options are available:

- ▶ Memory Mode (Default: Independent)
This option selects memory mode at initialization. Independent, mirroring, or sparing memory mode can be selected.

- ▶ **Memory Speed (Default: Max Performance)**
This option sets the following operating frequency of the installed DIMMs:
 - Minimal Power provides less performance for better power savings. The memory operates at the lowest supported frequency.
 - Power Efficiency provides the best performance per watt ratio. The memory operates one step under the rated frequency.
 - Max Performance provides the best system performance. The memory operates at the rated frequency.
- ▶ **Memory Power Management (Default: Disabled)**
This option sets the memory power management. Disable provides maximum performance at the expense of power.
- ▶ **Socket Interleave (Default: NUMA)**
This option sets NUMA or Non-NUMA system behavior. When NUMA is selected, memory is not interleaved across processors whereas Non-NUMA memory is interleaved across processors.
- ▶ **Memory Data Scrambling (Default: Enabled)**
This option enables a memory data scrambling feature to further minimize bit-data errors.
- ▶ **Page Policy (Default: Adaptive)**
This option determines the following Page Manager Policy in evaluating memory access:
 - Closed: Memory pages are closed immediately after each transaction.
 - Open: Memory pages are left open for a finite time after each transaction for possible recurring access.
 - Adaptive: Use Adaptive Page Policy to decide the memory page state.

7.1.3 ASU

By using the IBM ASU tool, users can modify firmware settings from the command line on multiple operating-system platforms.

You can perform the following tasks by using the utility:

- ▶ Modify selected basic input/output system (BIOS) CMOS settings without restarting the system to access F1 settings.
- ▶ Modify selected baseboard management controller setup settings.
- ▶ Modify selected Remote Supervisor Adapter and Remote Supervisor Adapter II setup settings.
- ▶ Modify selected settings in the integrated management module IMM-based servers for the IMM firmware and IBM System x Server firmware. IBM System x Server Firmware is the IBM implementation of UEFI.
- ▶ Modify a limited number of vital product data (VPD) settings on IMM-based servers.
- ▶ Modify iSCSI boot settings. To modify iSCSI settings with the ASU, you must first manually configure the values by using the server Setup utility settings on IMM-based servers.
- ▶ Connect remotely to set the listed firmware types settings on IMM-based servers. Remote connection support requires accessing the IMM external port over a LAN.

The ASU utility and documentation are available at this website:

<http://ibm.com/support/entry/portal/docdisplay?ln docid=MIGR-5085890>

Note: By using the ASU utility, you can generate a UEFI settings file from a system. A standard settings file is not provided on the IBM support site.

By using the ASU utility command line, you can read or modify firmware settings of a single node. The tool can be used from inside the node to modify node settings or to change a remote node through its IMM interface. When local to the node, ASU configures the in-band LAN over the USB interface and performs the wanted action.

When a node is accessed remotely, the IP address, user name, and password of the remote IMM must be provided. Figure 7-13 shows a command-line example that sets the IMM hostname value.

```
export PATH="/opt/ibm/toolscenter/asu/:$PATH"
asu64 set IMM.HostName1 <new_imm_hostname> --host <imm_ip_address>
      --user USERID --password PASSWORD
```

Figure 7-13 Setting the IMM2 host name that uses ASU

When the same value must be applied to several nodes, the **asu64** command must be run for each node that requires the setting. The latest releases of xCAT cluster management software provide a new tool, Parallel ASU, with which you can run the ASU tool to several nodes in parallel. xCAT cluster management capabilities and Parallel ASU help the IT administrator to perform changes to its cluster.

For more information about xCAT, see 7.6, “eXtreme Cloud Administration Toolkit” on page 149.

7.1.4 Firmware upgrade

Upgrading the firmware of NeXtScale System servers uses the same tools as other System x servers. The following tools are available to update the firmware of the NeXtScale nx360 M4 compute node, such as UEFI, IMM, or drivers:

- ▶ Stand-alone updates

Firmware updates can be performed online and offline. Offline updates are released in a bootable diskette (.img) or CD-ROM (.iso) format. Online updates are released for Windows (.exe), Linux, and VMware (.sh). The online updates are run under from the command line. They are scriptable and provided with XML metafiles for use with UpdateXpress System Packs.

Updates are available as Windows (.exe) or Linux (.bin) files, and are applied locally through the operating system, or remotely through the IMM2 Ethernet interface.

- ▶ UpdateXpress System Packs

UpdateXpress System Packs are a tested set of online updates that are released together as a downloadable package when products are first made available, and quarterly thereafter. The UpdateXpress System Packs are unique to each model of server.

The latest version of UXSP and a user’s guide are available at this website:

<http://ibm.com/support/entry/portal/docdisplay?ln docid=LNVO-XPRESS>

- ▶ **Bootable Media Creator**

Lenovo offers Bootable Media Creator with which you can bundle multiple System x updates from UpdateXpress System Packs and create bootable media, such as a CD or DVD .iso image, a USB flash drive, or a file set for PXE boot. You can download the tool from this website:

<http://ibm.com/support/entry/portal/docdisplay?lnocid=LNV0-BOMC>

7.2 Managing the chassis

The NeXtScale n1200 Enclosure includes the Fan and Power Controller (FPC) that manages power supply units and fans that are at the rear of the enclosure. In contrast to the IBM BladeCenter or IBM Flex System enclosures, the NeXtScale n1200 Enclosure does not contain a module that allows managements operation to be performed on the installed nodes. The FPC module is kept simple, and as with the iDataPlex system, compute nodes are managed through their IMM2 interface rather than a management module in the chassis.

The FPC module provides information regarding power usage at chassis or node level, fan speed, and allows the setting of specific power and cooling policies. The FPC module also adds a few management capabilities of the compute nodes that are installed in the chassis.

The FPC module includes the following features:

- ▶ Power usage information at the chassis, node, power supply units (PSU), and fan levels.
- ▶ PSU and fan status information.
- ▶ Configures wanted redundancy modes for operation (non-redundant, N+1, N+N, and oversubscription).
- ▶ Reports fan speed, fan status, and allows acoustic modes to be set.
- ▶ Defines a power cap policy at chassis or node level.

For more information about features and functional behavior, see 3.8, “Fan and Power Controller” on page 30.

Managing and configuring the FPC module can be done through a basic web browser interface that is accessed remotely by its Ethernet connection or remotely through the IPMI interface it provides.

The following sections describe how FPC module can be managed and configured from both interfaces. Although a web browser interface is intended to manage a single chassis, the IPMI interface can be used to develop wrappers that enclose IPMI commands to manage multiple chassis in the network:

- ▶ 7.2.1, “FPC web browser interface” on page 123
- ▶ 7.2.2, “FPC IPMI interface” on page 140

7.2.1 FPC web browser interface

The FPC module web interface can be accessed by the Ethernet connection at the rear of the chassis. The web interface provides a graphical and easy to manage method to configure a single chassis. However, when multiple chassis are managed, the best option is to remotely manage them through the IPMI over LAN interface (for more information, see 7.2.2, “FPC IPMI interface” on page 140).

Complete the following steps to access the FPC web interface:

1. Browse to the FPC interface URL that is defined for your FPC module. By default, the module is configured with the static IP address 192.168.0.100/24.
2. After the login window opens, enter the user name and password, as shown in Figure 7-14. The default user ID is USERID and default password is PASSWORD (where the 0 in the password is the zero character and not an uppercase letter o.)

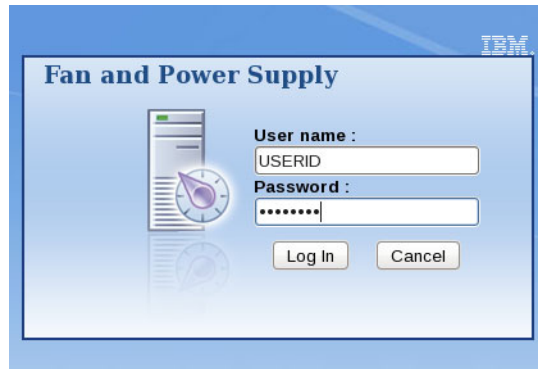


Figure 7-14 Fan and Power Controller log in page

3. Click **Log in**.

After you are logged in, the main page shows the following main functions on the left side of the page, as shown in Figure 7-15 on page 125:

- ▶ **Summary:** Displays the enclosure overall status and information. It introduces the chassis front view and rear view components and provides the status of the components (compute nodes, power supply units, fans, and so on).
- ▶ **Power:** Provides the power information about the different enclosure elements and allows the configuration of power supply redundancy modes, power capping or saving policies, and power restore policies.
- ▶ **Cooling:** Provides information about fan speed and allows the acoustic mode to be configured.
- ▶ **System Information:** Shows fixed Vital Product Data information for the enclosure, midplane, and FPC module.
- ▶ **Event Log:** Displays the System Event Log (SEL) and provides an interface to back up or restore the current configuration to or from the internal USB drive.
- ▶ **Configuration:** Allows the configuration of multiple options, such as, network, SNMP traps, alerts, and SMTP server.

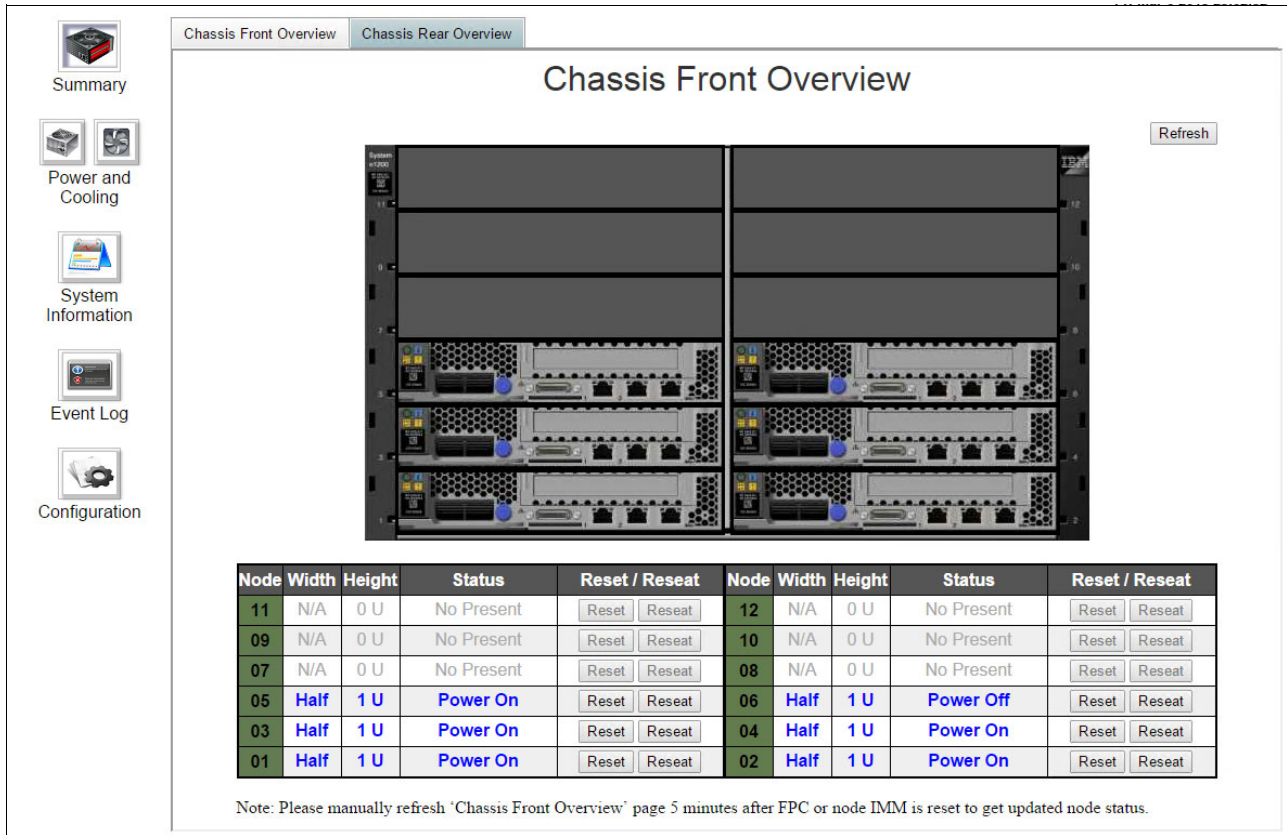


Figure 7-15 Summary front overview

The six main functions are described in the following sections:

- ▶ “Summary”
- ▶ “Power” on page 128
- ▶ “Cooling” on page 132
- ▶ “System Information tab” on page 133
- ▶ “Event Log” on page 134
- ▶ “Configuration” on page 135

Summary

The Summary function displays the enclosure’s overall status and information. There are two tabs that correspond to the front and rear of the chassis: Front Overview and Rear Overview.

Front Overview tab

As shown in Figure 7-15, the Front Overview table provides a graphical front view of the enclosure and a table that lists the status and information regarding the systems that are available in the enclosure.

Table 7-1 on page 126 lists the possible values that can appear in each column at the systems table.

Table 7-1 Front overview systems table

Column	Description
Node	Indicates slot number
Width	Possible values: <ul style="list-style-type: none"> ▶ Half: Represents a half-wide node ▶ Full: Represents a full-wide node (for future use)
Height	Node height can be 1U to 6U (for future use)
Status	Node power-on status. Possible values: <ul style="list-style-type: none"> ▶ No Present: No node is installed ▶ No Permission: Node is not granted power permission and cannot be powered on ▶ Fault: Node has a power fault and cannot be powered on ▶ Power On: Node is powered on ▶ Power Off: Node is powered off
Reset/Reseat	Used to perform virtual reset or virtual reseat: <ul style="list-style-type: none"> ▶ Virtual Reset: User can remotely reset (reboot) the IMM through the FPC. ▶ Virtual Reseat: User can remotely power cycle entire node. Reseat provides a way to emulate physical disconnection of a node. <p>After virtual reset or reseat, node IMM takes up to two minutes to be ready.</p>


Rear Overview tab

The Rear Overview window provides a graphical rear view of the enclosure and a table that lists the status and information regarding power supply units and FPC module information. It displays characteristics of the available elements and a summary of health conditions, with which the system administrator can easily identify the source of a problem.

Figure 7-16 on page 127 shows the chassis rear overview page. Table 7-2 on page 127 shows the possible values that can appear in each column of the PSU table. For the water-cooled system, the inlet leak sensor is in fan position 1, and the outlet leak sensor is in fan position 3; there are no chassis fans in the water-cooled system, only power supply fans.

Chassis Front Overview
Chassis Rear Overview

Chassis Rear Overview



Management Module	
Name	Fan & Power Control Board (FPCB)
Status	<input checked="" type="checkbox"/> Normal <input type="button" value="FPC Reboot"/> <input type="button" value="Reset to Default"/>
Firmware Version	FHET26B-1.00
PSOC Version	ver. 1.30
Boot-up Flash	First
Identify LED	Off <ul style="list-style-type: none"> <input checked="" type="radio"/> Turn Off <input type="radio"/> Turn On <input type="radio"/> Blink <input type="button" value="Apply"/>
Check Log LED	On <input type="button" value="Turn Off"/>

PSU

PSU	Status	Ratings	AC-IN	EPOW	Throttle	DC-PG
PSU1	Present	1300 W	215 V	Normal	Normal	Yes
PSU2	Present	1300 W	211 V	Normal	Normal	Yes
PSU3	Present	1300 W	212 V	Normal	Normal	Yes
PSU4	Present	1300 W	213 V	Normal	Normal	Yes
PSU5	Present	1300 W	213 V	Normal	Normal	Yes
PSU6	Present	1300 W	213 V	Normal	Normal	Yes

Water cooled system

Bay	Status	Type	Bay	Status	Type
1	Present	Leak Sensor	6	N/A	N/A
2	N/A	N/A	7	N/A	N/A
3	Present	Leak Sensor	8	N/A	N/A
4	N/A	N/A	9	N/A	N/A
5	N/A	N/A	10	N/A	N/A

Figure 7-16 NeXtScale n1200WCT chassis rear overview

Table 7-2 Power supply unit table

Column	Description
Status	Possible values: <ul style="list-style-type: none"> ▶ Present: Power supply installed ▶ No Present: No power supply installed ▶ Fault: Power supply in faulty condition
Ratings	Display the DC power output rating of the power supply

Column	Description
AC-IN	Display the power AC input voltage rating
EPOW ^a	Possible values: <ul style="list-style-type: none"> ▶ Assert: Power supply is in AC lost condition ▶ Normal: Power supply is in healthy, operating condition
Throttle ^a	Possible values: <ul style="list-style-type: none"> ▶ Assert: Power supply is in over-current condition ▶ Normal: Power supply is in healthy, operating condition
DC-PG	DC power good. Possible values: <ul style="list-style-type: none"> ▶ No: Power supply is not providing the required DC power ▶ Yes: Power supply is in healthy, operating condition

a. For more information about Early Power Off Warning (EPOW) and Throttle, see 3.9, “Power management” on page 34.

Power

The Power function provides the power information about the different enclosure elements. You can configure power redundancy modes, power capping and power-saving policies, and the power restore policy to be used.

The following tabs are available and are described next:

- ▶ “Power Overview tab” on page 128
- ▶ “PSU Configuration tab” on page 129
- ▶ “Power Cap tab” on page 130
- ▶ “Voltage Overview tab” on page 132
- ▶ “Power Restore Policy tab” on page 132

Power Overview tab

As shown in Figure 7-17 on page 129, the Power Overview tab provides information about the total chassis power consumption (minimum, average, and maximum of AC-in and DC-out) and a specific power information breakdown of the different systems because there are no chassis fans in the n1200 WCT, the total fans power consumption field shows N/A.

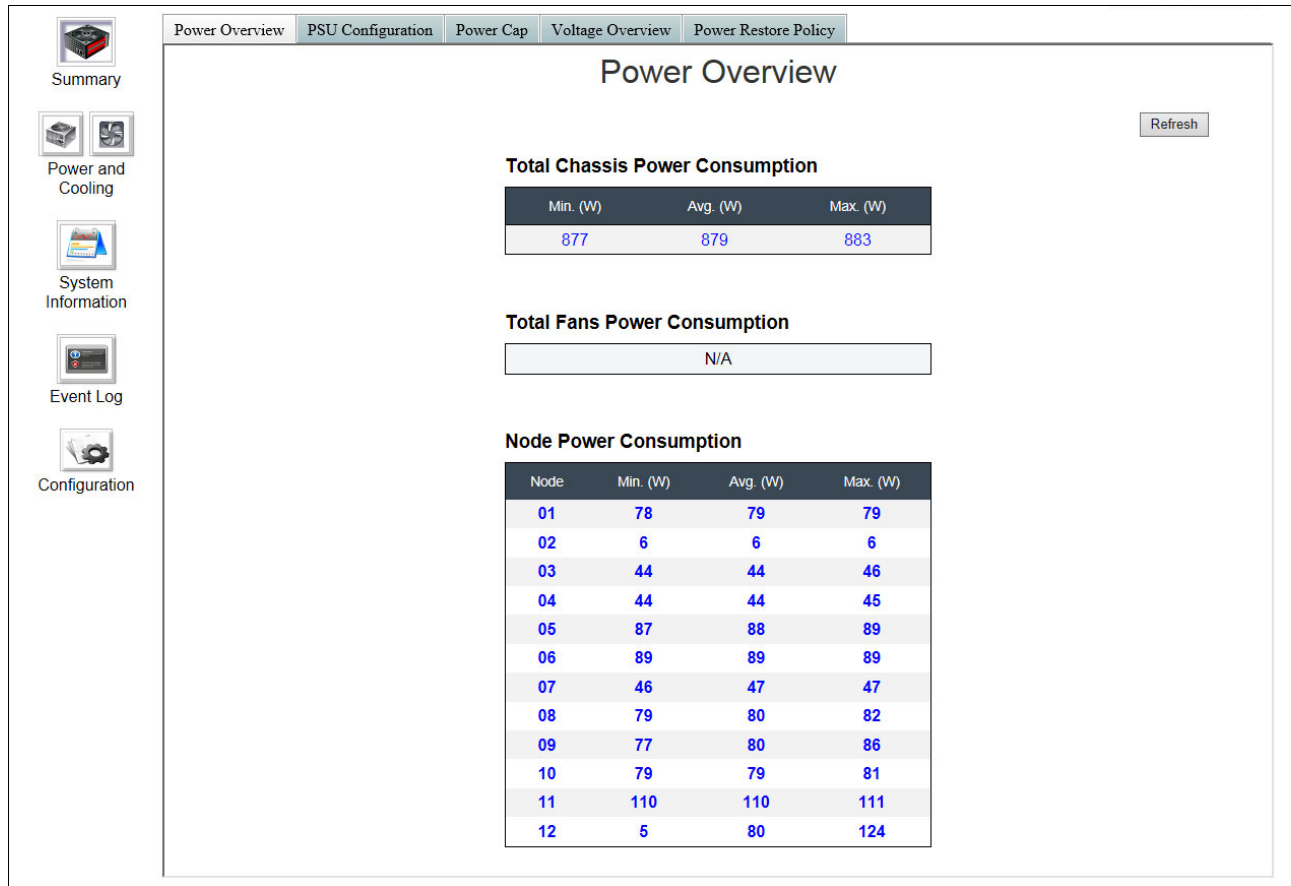


Figure 7-17 Power overview

PSU Configuration tab

As shown in Figure 7-18 on page 130, the PSU Configuration tab allows setting the redundancy mode for the enclosure and enable oversubscription mode, if needed.

The following redundancy modes can be selected:

- ▶ No redundancy: Compute nodes can be throttled or shutdown if any power supply is in faulty condition.
- ▶ N+1: One of the power supplies is redundant, so a single faulty power supply is allowed.
- ▶ N+N: Half of the PSUs that are installed are redundant, so the enclosure can support up to N faulty power supply units.

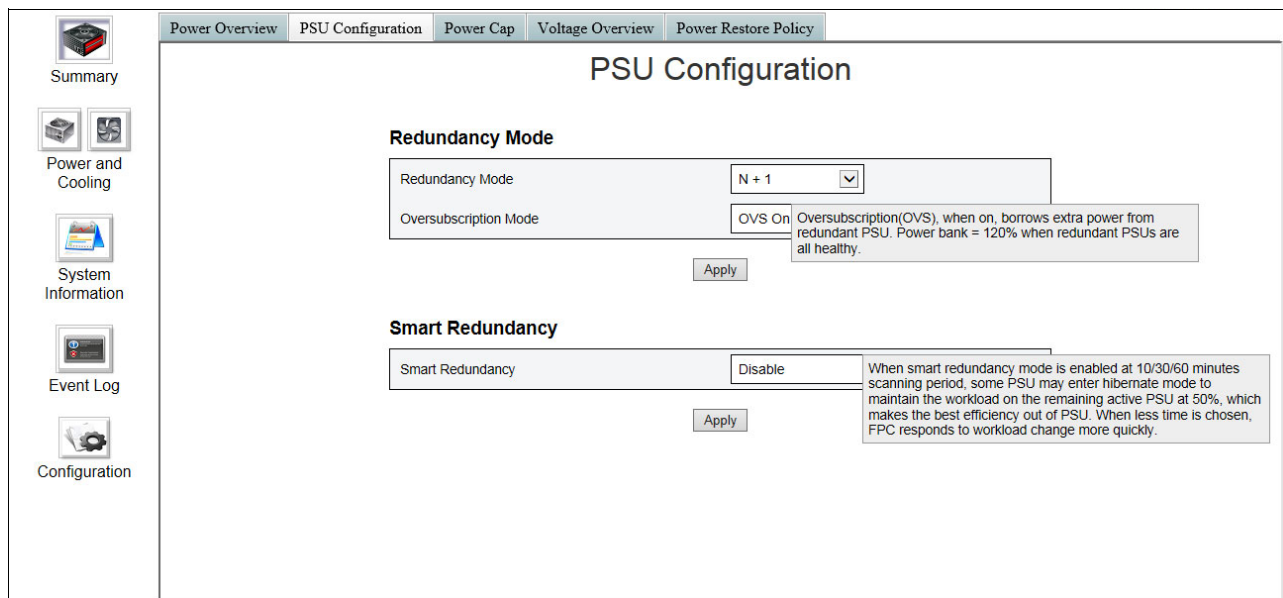


Figure 7-18 Power supply redundancy mode configuration

Oversubscription option (OVS) is only selectable when N+1 or N+N redundancy modes are enabled. When oversubscription is enabled, the enclosure allows node loads up to 120% of the redundant power budget. If a power supply fails, the surviving power supplies can sustain the extra load until the nodes are throttled to a lower power state. This mode prevents system outages while enabling higher power consumptive nodes to operate under normal and power supply fault conditions.

Smart redundancy is only available with 1300 W power supplies.

The power budget that is available to grant power permission to systems that are installed in the chassis depends directly on the capacity of the power supply units that are installed, the demanding power of the nodes, the redundancy mode that was selected, and if oversubscription is enabled.

For more information, see 3.9, “Power management” on page 34.

Power Cap tab

The Power Cap tab allows setting power capping and power-saving modes at chassis or node level. Power capping and power-saving modes can be applied simultaneously.

Power capping at chassis level is selected at the drop-down menu. A range is suggested that is based on the minimum and maximum power consumption of the systems that are installed in the chassis. Any value that is not set within the suggested range is allowed; however, if it is below the minimum, it might not be reached.

Figure 7-19 shows power capping windows at chassis level. The range that is suggested is based on the aggregation of the minimum and maximum power consumption for the nodes that are installed at the chassis.

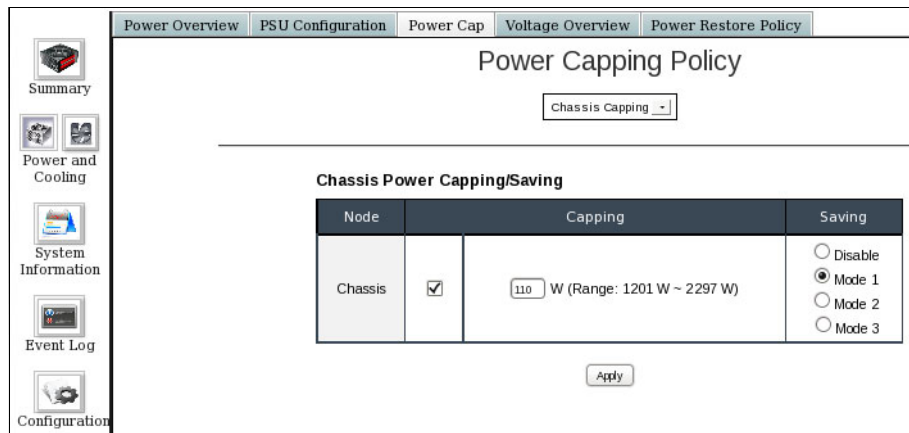


Figure 7-19 Power capping at chassis level

Power capping at node level is selected via the drop-down menu. The specific node is selected in the drop-down menu that appears inside the table. Here, the suggested range is based on the minimum and maximum consumption of the node. Again, any value that is outside of the range is allowed, but it might not be reached. Figure 7-20 shows power capping windows at node level. The range that is suggested is based on the minimum and maximum power consumption for the nodes that are selected.

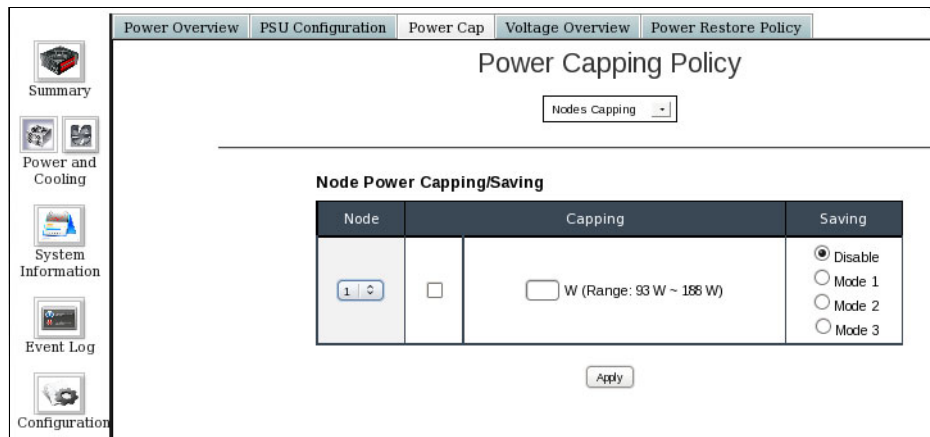


Figure 7-20 Power capping at node level

Similar to power capping, power saving can be set at chassis or node level. The following four modes can be selected:

- ▶ Disabled (default static maximum performance mode): The system runs at full speed (no throttling), regardless of the workload.
- ▶ Mode 1 (static minimum power): The system runs in a throttling state regardless of the workload. The throttling state is the lowest frequency P-state.
- ▶ Mode 2 (dynamic favor performance): The system adjusts throttling levels that are based on workload, which attempts to favor performance over power savings.
- ▶ Mode 3 (dynamic favor power): The system adjusts the throttling levels that are based on workload that is attempting to favor power savings over performance.

Voltage Overview tab

As shown in Figure 7-21, the Voltage Overview tab displays the actual FPC 12 V, 3.3 V, 5 V, and battery voltage information. If a critical threshold is reached, error log is asserted.

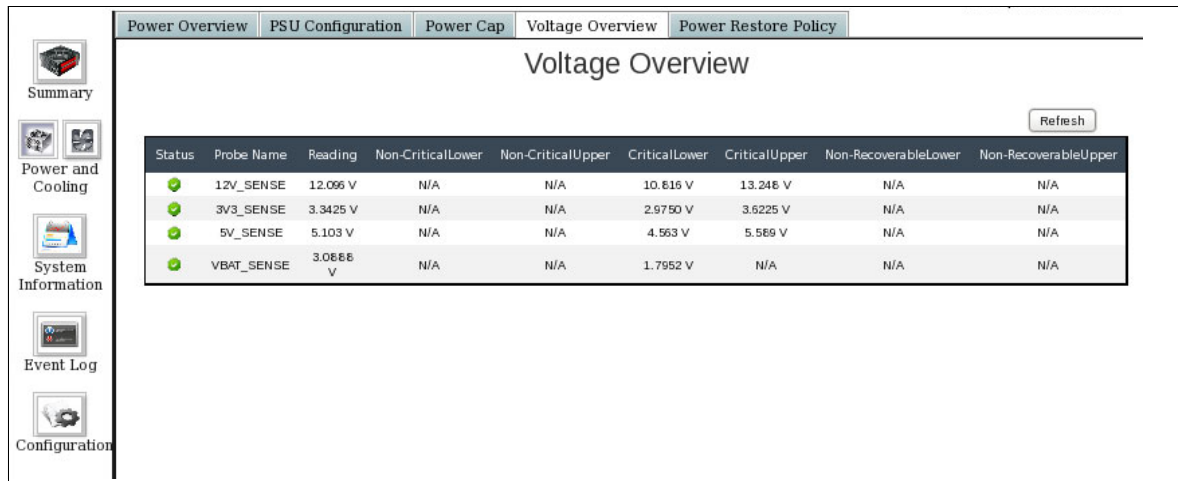


Figure 7-21 Voltage overview

Power Restore Policy tab

The Power Restore Policy tab displays and the user can set the restore policy for specific nodes. The FPC module remembers nodes that are already powered on and have power restore policy enabled. When AC is abruptly lost, it automatically turns on the nodes when AC is recovered.

To enable the restore policy on certain nodes, select the nodes and click **Apply**, as shown in Figure 7-22. The Status changes from Disable to Enable (or vice versa).

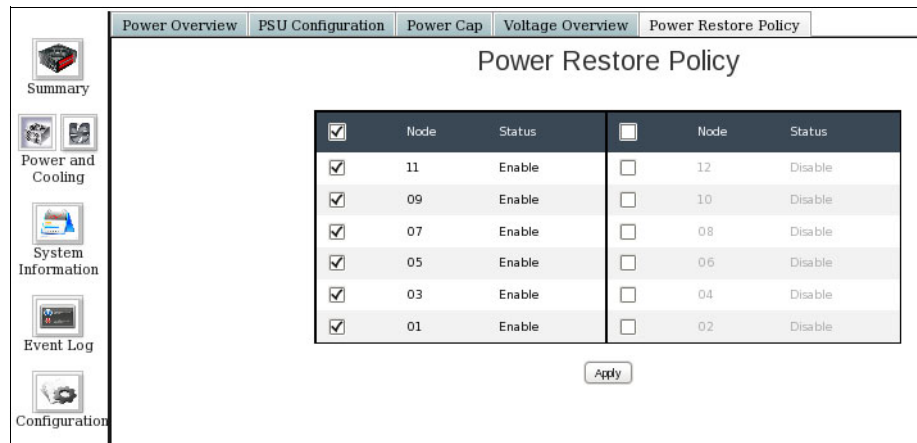


Figure 7-22 Power Restore Policy tab window

Cooling

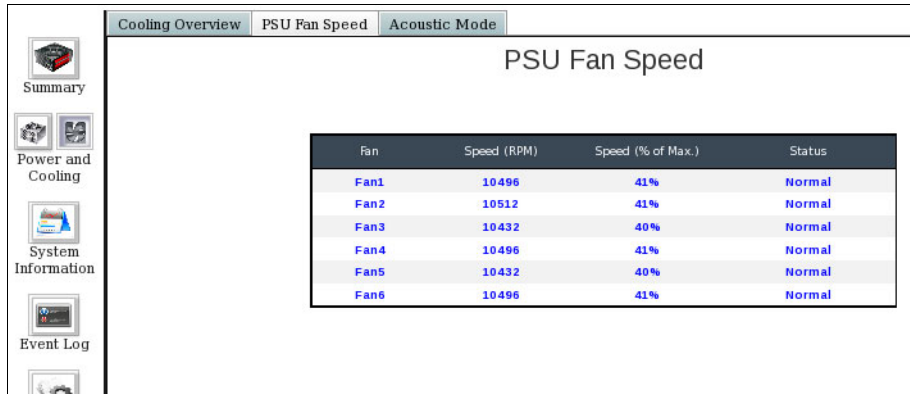
The Cooling function provides information about fan speed for system fans and power supply unit fans. The following tabs are available:

- ▶ PSU Fan Speed
- ▶ Acoustic Mode

These tabs are described next.

PSU Fan Speed tab

As shown in Figure 7-23, the PSU Fan speed tab shows the power supply fan speeds and their healthy condition. PSU fans normally operate at 5,000 - 23,000 rpm and are considered faulty when the speed falls below 3,000 rpm.



Fan	Speed (RPM)	Speed (% of Max.)	Status
Fan1	10496	41%	Normal
Fan2	10512	41%	Normal
Fan3	10432	40%	Normal
Fan4	10496	41%	Normal
Fan5	10432	40%	Normal
Fan6	10496	41%	Normal

Figure 7-23 PSU Fan speed information

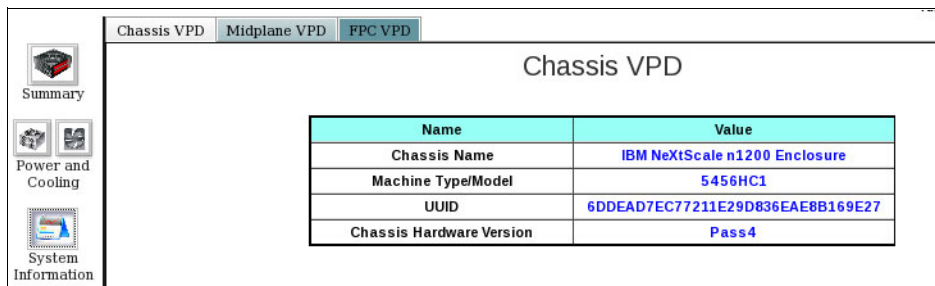
Acoustic Mode tab

Although it is possible to select Mode 1, Mode 2, or Mode 3 on the NeXtScale WCT, the use of Acoustic Mode is not supported. Because of the absence of chassis fans, the NeXtScale WCT runs quietly.

System Information tab

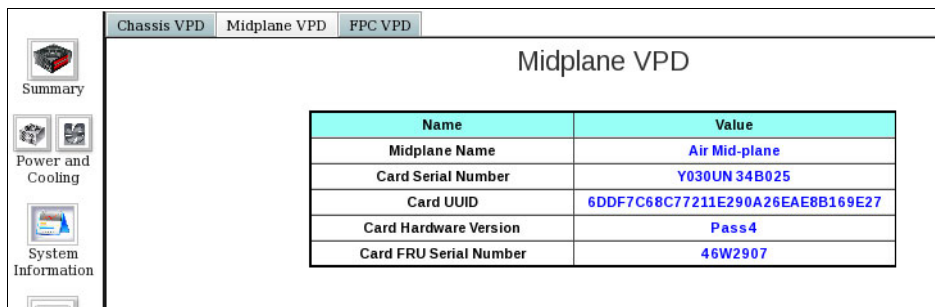
The System Information tab provides information about the Vital Product Data (VPD) for the chassis, the midplane, and the FPC module.

Figure 7-24, Figure 7-25, and Figure 7-26 on page 134 show the system information VPD windows for the chassis, the midplane, and the FPC.



Name	Value
Chassis Name	IBM NeXtScale n1200 Enclosure
Machine Type/Model	5456HC1
UUID	6DDF7C68C77211E29D836EAE8B169E27
Chassis Hardware Version	Pass4

Figure 7-24 Chassis Vital Product Data window



Name	Value
Midplane Name	Air Mid-plane
Card Serial Number	Y030UN 34B025
Card UUID	6DDF7C68C77211E29D836EAE8B169E27
Card Hardware Version	Pass4
Card FRU Serial Number	46W2907

Figure 7-25 Midplane Vital Product Data window

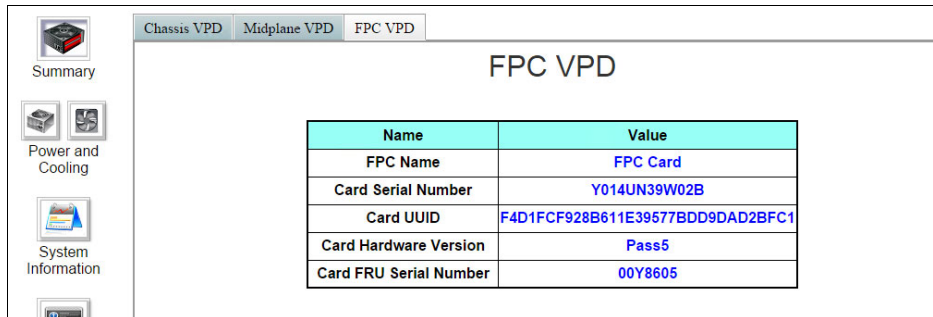


Figure 7-26 FPC Vital Product Data window

Event Log

The Event Log function displays the SEL of the chassis. Users can back up or restore user configurations to and from an internal USB.

SEL log includes information, warning, and critical events that are related to the chassis. The maximum of log entries is 512. When the log is full, you must clear it to allow new entries to be saved. When the log reaches 75%, a warning event is reported via SNMP.

You can clear the log by clicking **Clear Log** (as shown in Figure 7-27) or by running the following IPMI command:

```
ipmitool -I lanplus -U USERID -P PASSWORD -H 192.168.0.100 sel clear
```

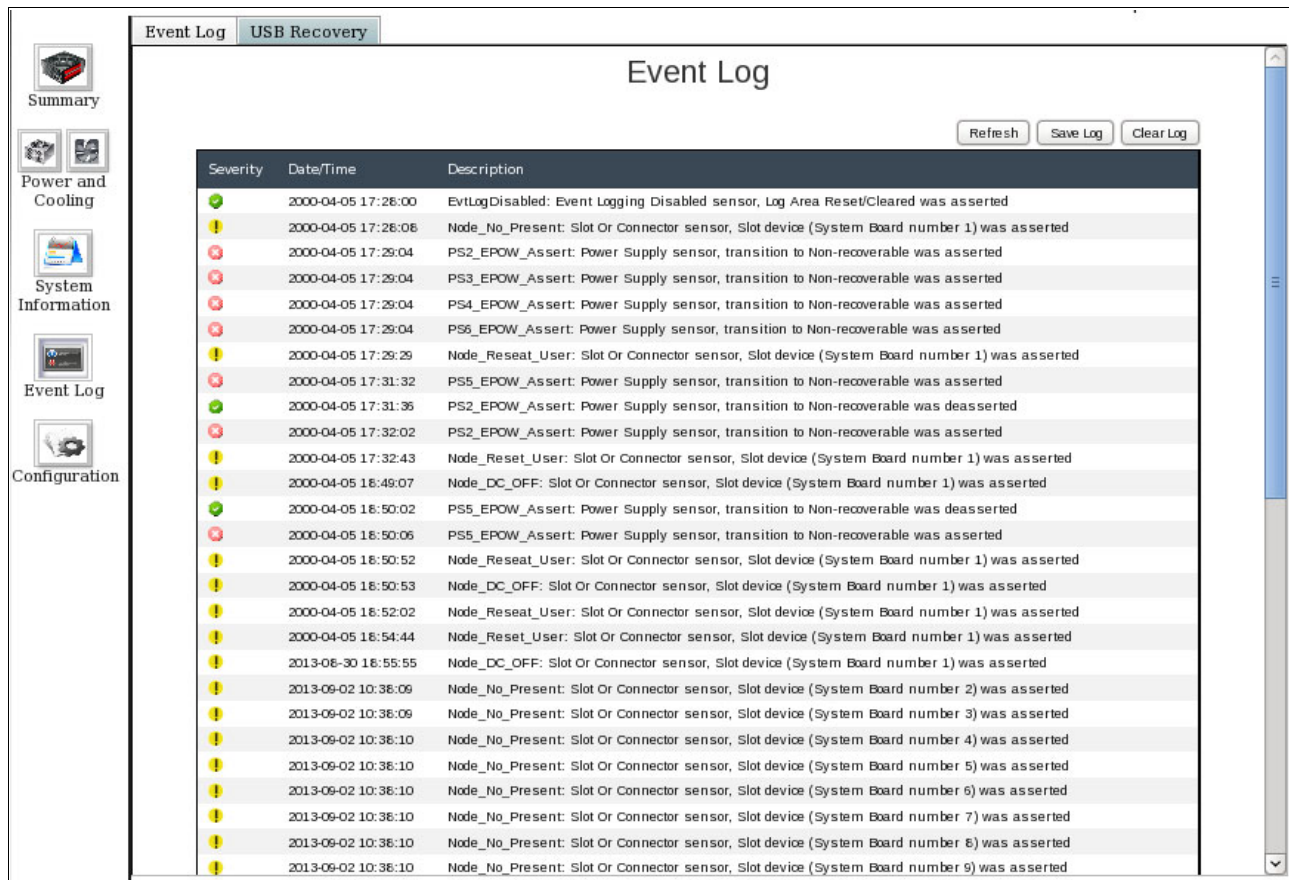


Figure 7-27 SEL from the FPC module

Full log: After the log fills up, you should clear it manually; otherwise, any other log entries cannot be received.

Back up and restore operations on the internal USB are automatically done by the FPC. Any change that is done through the web interface or the IPMI interface that is part of the following settings is saved in the internal USB:

- ▶ Selected power supply redundancy policy
- ▶ Oversubscription mode
- ▶ Power capping and power-saving values at chassis and node level
- ▶ Acoustic mode settings
- ▶ Power restore policy
- ▶ System event log (SEL)

All of these settings are volatile, so when FPC is rebooted, the FPC restores the settings from the internal USB automatically. The configuration settings that are related to network, SNMP, and so on, are non-volatile, so they remain in FPC memory between reboots.

The FPC web interface also includes manual backup and restore functions, as shown in Figure 7-28; however, because backup and restore tasks are automatic, these options are not needed.



Figure 7-28 USB backup and recovery tab

Configuration

The Configuration function displays and configures the FPC module. All settings under the Configuration function are non-volatile, so they are kept between FPC reboots and are not saved to the internal USB key.

By using the Configuration function, the user can perform the following tasks by using the corresponding tabs:

- ▶ “Firmware Update tab”
- ▶ “SMTP tab” on page 137
- ▶ “SNMP tab” on page 138
- ▶ “Platform Filter Events tab” on page 138
- ▶ “Network Configuration tab” on page 138
- ▶ “Time Setting tab” on page 139
- ▶ “User Account tab” on page 139
- ▶ “Web Service tab” on page 140

Firmware Update tab

When a firmware upgrade is available, you use the web interface to perform the upgrade.

A firmware update is done in two phases by using the window that is shown in Figure 7-29. First, the user selects the wanted local firmware file that is uploaded and verified to be valid. Second, after the firmware is checked, a confirmation is requested. A table shows the actual firmware version, the new firmware version, and a preserve existing settings option that must be selected to keep the settings.

After the firmware update is performed, the FPC is rebooted.

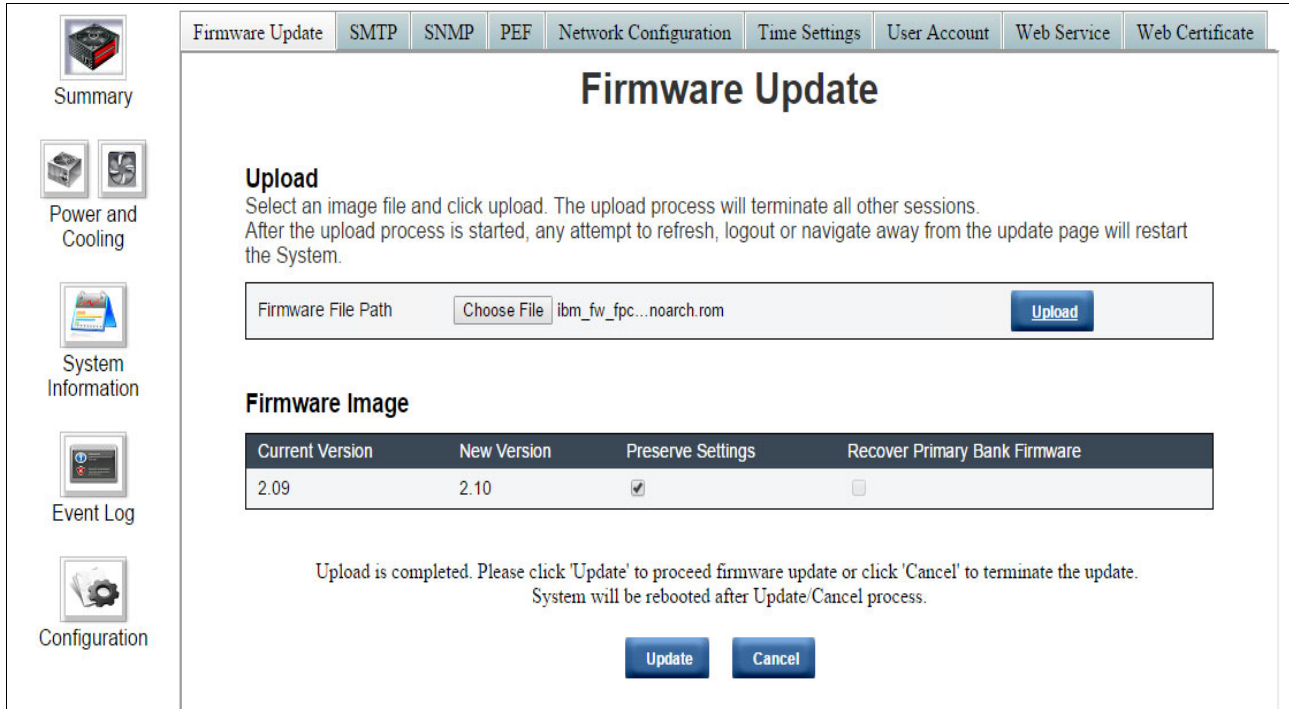


Figure 7-29 Selection window for firmware update

SMTP tab

The FPC module allows the SMTP configuration to send the events to the destination email addresses and SMTP server, as shown in Figure 7-30. The Global Alerting Enable option at the Platform Event Filters (PEF) tab must be selected to enable SMTP traps and no filtering applied so that all the events are sent.

	Enable	Destination Email Address	Email Description	Test
Email Alert 1	<input type="checkbox"/>	<input type="text"/>	MergePoint email als	<input type="button" value="Send Alert 1"/>
Email Alert 2	<input type="checkbox"/>	<input type="text"/>	MergePoint email als	<input type="button" value="Send Alert 2"/>
Email Alert 3	<input type="checkbox"/>	<input type="text"/>	MergePoint email als	<input type="button" value="Send Alert 3"/>
Email Alert 4	<input type="checkbox"/>	<input type="text"/>	MergePoint email als	<input type="button" value="Send Alert 4"/>

SMTP (email) Server Address

SMTP IP Address

SMTP Authentication

Enable Anonymous account will be used when authentication is disabled.

Username

Password

STARTTLS Mode ▾

SASL Mode ▾

Figure 7-30 SMTP configuration tab

SNMP tab

The FPC module allows the SNMP configuration to send as SNMP traps the events that occur, as shown in Figure 7-31. The specific event types that are sent are selected at the PEF tab. The Global Alerting Enable option in the PEF tab must be selected so that the SNMP traps are enabled.

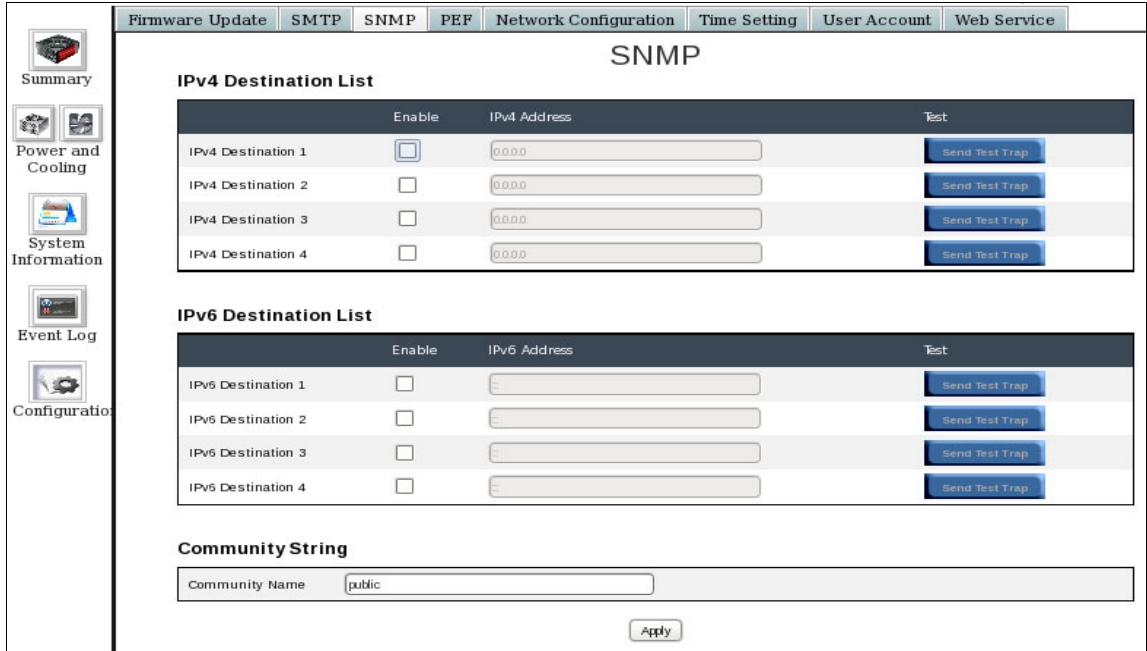


Figure 7-31 SNMP configuration tab

Platform Filter Events tab

In the PEF tab, you can configure the type of events that are sent as SNMP traps, as shown in Figure 7-32. You also can enable or disable SNMP and SMTP alerting by selecting the Global Alerting Enable option.

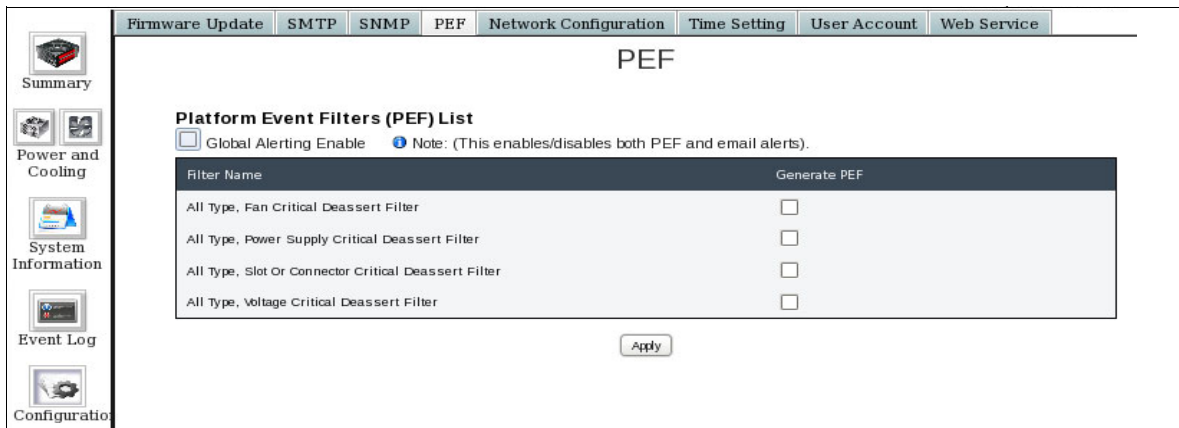


Figure 7-32 Platform Event Filters window

Network Configuration tab

In the Network Configuration tab, users can configure the network setting for the FPC module, as shown in Figure 7-33 on page 139. Hostname, static IP address, DHCP, and VLAN configuration can be set.

To access the specific network configuration settings windows, double-click the current network interface configuration.

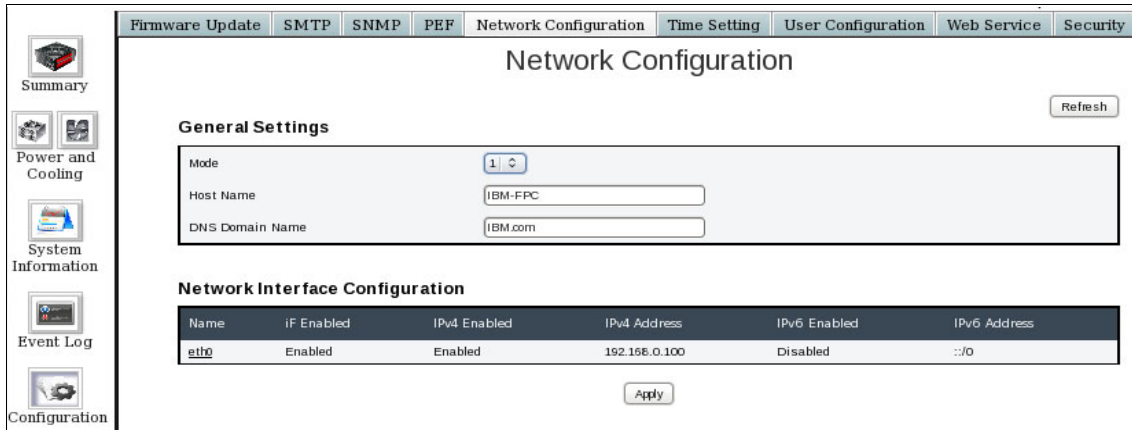


Figure 7-33 Network configuration window.

Time Setting tab

In the Time Setting tab, users can configure the date and time for the FPC module, as shown in Figure 7-34.

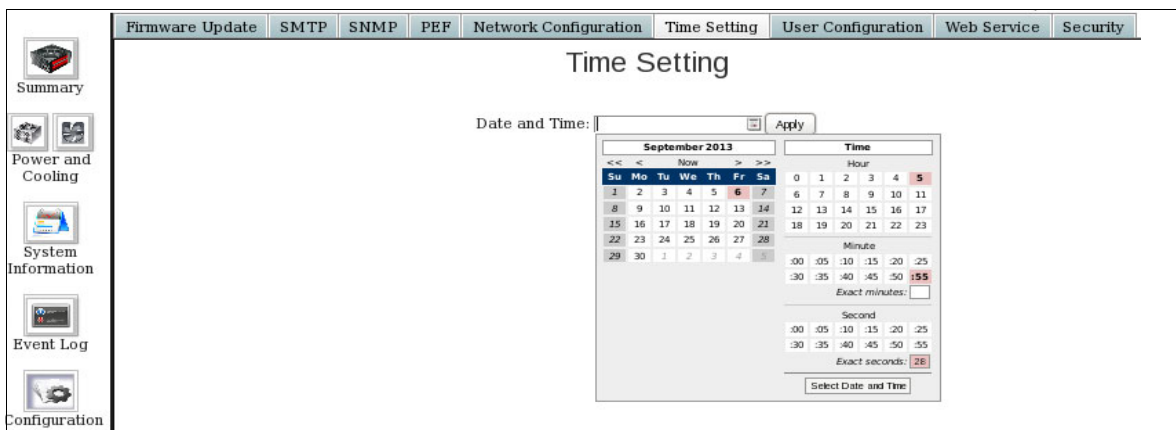


Figure 7-34 Date and time configuration window

User Account tab

In the User Account tab, users can add or remove users and assign one of the following user roles, as shown in Figure 7-35 on page 140:

- ▶ Administrator: Full access to all web pages and settings.
- ▶ Operator: Full access to all web pages and settings except the User Account page.
- ▶ User: Full access and settings to all pages except the Configuration function tab.

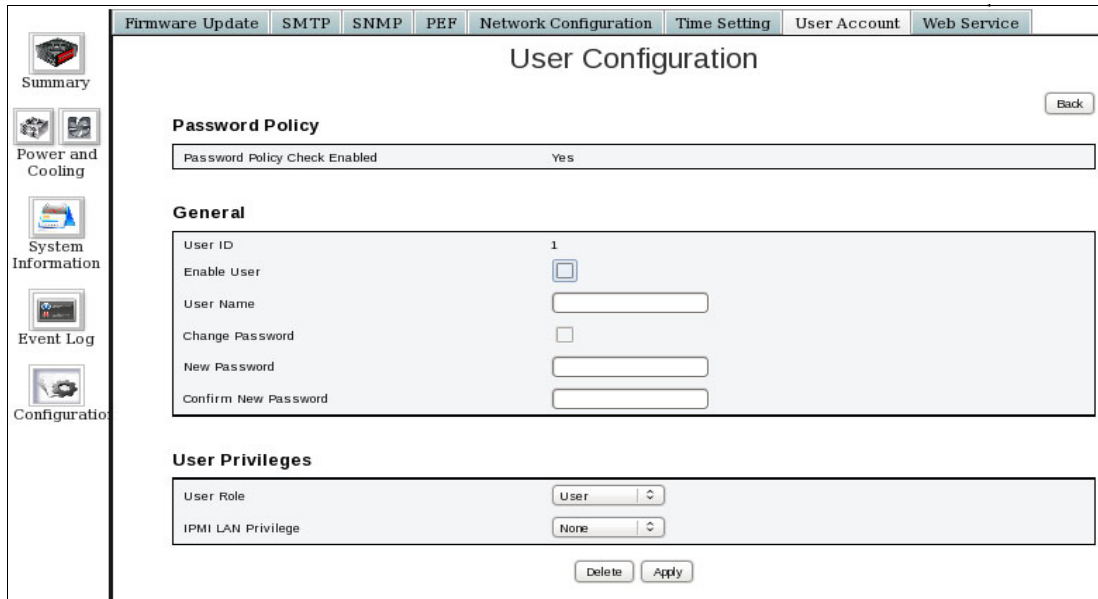


Figure 7-35 User Configuration window

Web Service tab

User can configure the web interface ports for HTTP and HTTPS access in the Web Service tab, as shown in Figure 7-36.

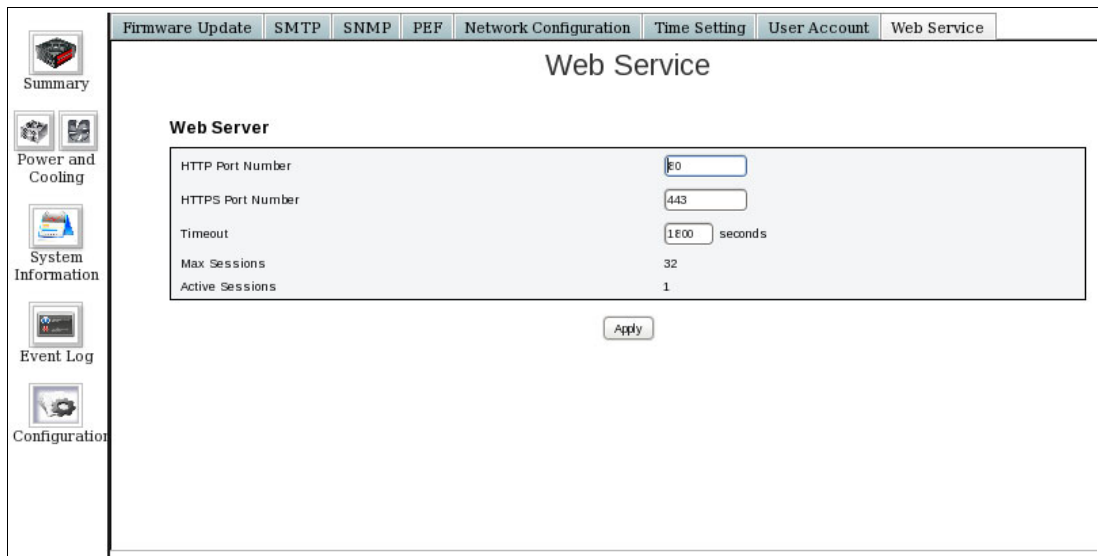


Figure 7-36 Configuration of the HTTP/HTTPS ports for the web browser interface

7.2.2 FPC IPMI interface

The FPC module on the NeXtScale n1200 Enclosure supports IPMI over LAN access. The FPC complies with IPMI v2.0 standard and uses extensions and OEM IPMI commands to access FPC module-specific features. The IPMI interface can be used to develop wrappers with which users can remotely manage and configure multiple FPC modules at the same time.

By using the `ipmitool` command, you can manage and configure devices that support IPMI. The `ipmitool` command provides an easy CLI to start IPMI commands to a remote service processor through LAN. By using the tool, you also can send raw commands that are not part of the IPMI v2.0 definition but are vendor extensions to support certain specificities.

The FPC is compliant with IPMI v2.0, so default syntax and options can be used to access the commands that are part of the IPMI v2.0 definition. For example, to list the SEL, the following command can be used:

```
ipmitool -I lanplus -U USERID -P PASSWORD -H 192.168.0.100 sel list
```

OEM IPMI command extensions require the use of the raw interface `ipmitool` provides. The syntax for such interface has the following format:

```
ipmitool -I lanplus -U USERID -P PASSWORD \  
-H 192.168.0.100 raw <netfn> <cmd> [<byte1>] [<byte2>] ...
```

Output is raw hexadecimal data that provides the completion code of the operations and the resulting data. Input parameters and output values are described next. “Examples” on page 146 provides some examples of how IPMI commands are sent and how to process the output.

Note: Cluster management software provides command line tools to use this interface easier. However, some might not be implemented or you might want to integrate IBM NeXtScale System into your own custom monitoring infrastructure, so the IPMI commands are provided as reference.

Many settings that are available at the web browser interface are provided as IPMI commands. The following tables group the specific commands according to their functionality:

- ▶ Table 7-3 on page 142 lists Power supply unit (PSU) IPMI commands
- ▶ Table 7-4 on page 142 lists Power capping IPMI commands
- ▶ Table 7-5 on page 143 lists Power redundancy IPMI commands
- ▶ Table 7-6 on page 144 lists Acoustic modes IPMI commands
- ▶ Table 7-7 on page 144 lists Power restore policy IPMI commands
- ▶ Table 7-8 on page 144 lists Fan IPMI commands
- ▶ Table 7-9 on page 145 lists LED IPMI commands
- ▶ Table 7-10 on page 145 lists Node IPMI commands
- ▶ Table 7-11 on page 146 lists Miscellaneous IPMI commands
- ▶ “Examples” on page 146 shows the usage of IPMI interface with some examples

Table 7-3 Power supply unit (PSU) IPMI commands

Description	NetFn	CMD	Data
Get PSU Data	0x32	0x90	<p>Request Data:</p> <p>Byte 1 options:</p> <ul style="list-style-type: none"> ▶ 1: AC-IN ▶ 2: DC-OUT ▶ 3: PSU fan power <p>Response Data:</p> <p>(when AC-IN, DC-OUT)</p> <p>Byte 1: Completion code (0x00)</p> <p>Byte 2: Sum of MIN AC-IN /(DC-OUT) Least Significant Bit (LSB)</p> <p>Byte 3: Sum of MIN AC-IN /(DC-OUT) Most Significant Bit (MSB)</p> <p>Byte 4: Sum of average AC-IN /(DC-OUT) LSB</p> <p>Byte 5: Sum of average AC-IN/(DC-OUT) MSB</p> <p>Byte 6: Sum of MAX AC-IN /(DC-OUT) LSB</p> <p>Byte 7: Sum of MAX AC-IN /(DC-OUT) MSB</p> <p>(when Fan power)</p> <p>Byte 1: Completion code (0x00)</p> <p>Byte 2: Sum of FAN_Power LSB</p> <p>Byte 3: Sum of FAN_Power Byte 2</p> <p>Byte 4: Sum of FAN_Power MSB</p>
Get PSU Status	0x32	0x91	<p>Request Data:</p> <p>None</p> <p>Response Data:</p> <p>Byte 1: Completion code (0x00)</p> <p>Byte 2: PS_EPOW</p> <p>Byte 3: PS_THROTTLE</p> <p>Byte 4: PS_PRESENT</p> <p>Byte 5: PS_PWR_GOOD</p> <p>Byte 6: EPOW_OUT</p> <p>Byte 7: THROTTLE</p> <p>Each Byte is a bit mask where bit 0-5 = PSU1-6 (0: not trigger; 1: trigger)</p>

Table 7-4 Power capping IPMI commands

Description	NetFn	CMD	Data
Get power capping capacity	0x32	0x9d	<p>Request Data:</p> <p>Byte 1 options:</p> <ul style="list-style-type: none"> ▶ Node number: 0x1 to 0x0c for Node 1 - 12 ▶ Chassis: 0x0d <p>Response Data:</p> <p>Byte 1: Completion code (0x00) or out of range (0xC9)</p> <p>Byte 2: Min. capping value (LSB)</p> <p>Byte 3: Min. capping value (MSB)</p> <p>Byte 4: Max. capping value (LSB)</p> <p>Byte 5: Max. capping value (MSB)</p>
Set power capping value	0x32	0x9e	<p>Request Data:</p> <p>Byte 1 options:</p> <ul style="list-style-type: none"> ▶ Node number: 0x1 to 0x0c for Node 1 - 12 ▶ Chassis: 0x0d <p>Byte 2: Capping value LSB</p> <p>Byte 3: Capping value MSB</p> <p>Response Data:</p> <p>Byte 1: completion code (0x00) or out of range (0xC9) or cur not support (0xD5)</p>

Description	NetFn	CMD	Data
Set power-saving state	0x32	0x9f	Request Data: Byte 1 options: ▶ Node number: 0x1 to 0x0c for Node 1 - 12 ▶ Chassis: 0x0d Byte 2: Capping disable / enable Byte 3: Saving mode (0x00: disable; 0x01: Mode1; 0x02: Mode2; 0x03: Mode3) Response Data: Byte 1: Completion code (0x00) or out of range (0xC9)
Get power-saving state	0x32	0xa0	Request Data: Byte 1 options: ▶ Node number: 0x1 to 0x0c for Node 1 - 12 ▶ Chassis: 0x0d Response Data: Byte 1: Completion code (0x00) or out of range (0xC9) Byte 2: Capping disable / enable Byte 3: Capping value LSB Byte 4: Capping value MSB Byte 5: Saving mode

Table 7-5 Power redundancy IPMI commands

Description	NetFn	CMD	Data
Get PSU policy	0x32	0xa2	Request Data: None Response Data: Byte 1: Completion code (0x00) Byte 2: PSU Policy ▶ 0: No redundancy ▶ 1: N+1 ▶ 2: N+N Byte 3: Oversubscription mode (0: disable; 1: enable) Byte 4: Power bank LSB Byte 5: Power bank MSB
Set PSU policy	0x32	0xa3	Request Data: Byte 1: PSU Policy ▶ 0: No redundancy ▶ 1: N+1 ▶ 2: N+N Response Data: Byte 1: Completion code (0x00) or out of range (0xC9) or config not allowed (0x01) or bank lack (0x02)
Set Over Subscription mode	0x32	0x9c	Request Data: Byte 1: Over Subscription mode ▶ 0: Disable ▶ 1: Enable Response Data: Byte 1: Completion code (0x00) or cur not supported (0x5d) or param out of range (0xc9)

Table 7-6 Acoustic mode IPMI commands

Description	NetFn	CMD	Data
Set Acoustic mode	0x32	0x9b	Request Data: Byte 1: Acoustic mode (0x00: disable; 0x01: mode1 - 28%; 0x02; mode2 - 34%; 0x3 mode3 - 40%) Response Data: Byte 1: Completion code (0x00) or out of range (0xC9)

Table 7-7 Power restore policy IPMI commands

Description	NetFn	CMD	Data
Get Restore Policy	0x32	0xa9	Request Data: Byte 1: Node number LSB (bit mask) Byte 2: Node number LSB (bit mask) example: If setting 1,2 and 3 - Byte 1: 0x7 (0000 0111) Response Data: Byte 1: Completion code (0x00) or out of range (0xC9)
Set Restore Policy	0x32	0xaa	Request Data: None Response Data: Byte 1: Completion code (0x00) or out of range (0xC9) Byte 2: Node number LSB (bit mask) Byte 3: Node number LSB (bit mask)

Table 7-8 Fan IPMI commands

Description	NetFn	CMD	Data
Get PSU Fan status	0x32	0xa5	Request Data: Byte 1: PSU FAN number (0x01-0x06 FAN 1-6) Response Data: Byte 1: Fan speed LSB (rpm) Byte 2: Fan speed MSB (rpm) Byte 3: Fan speed (0-100%) Byte 4: Fan health <ul style="list-style-type: none"> ▶ 0: Not present ▶ 1: Abnormal ▶ 2: Normal

Table 7-9 LED IPMI commands

Description	NetFn	CMD	Data
Get Sys LED: Command to get FPC LED status	0x32	0x96	Request Data: None Response Data: Byte 1: Completion code (0x00) Byte 2: SysLocater LED Byte 3: CheckLog LED Possible values are 0: Off 1:On 2: Blink (SysLocater LED only)
Set Sys LED: Command to set FPC LED status	0x32	0x97	Request Data: Byte 1 options: ▶ 1: SysLocater LED ▶ 2: CheckLog LED Byte 2 options: ▶ 0: Disable ▶ 1: Enable ▶ 2: Blink (SysLocated LED only) Response Data: Byte 1: Completion code (0x00)

Table 7-10 Node IPMI commands

Description	NetFn	CMD	Data
Get Node Status	0x32	0xa7	Request Data: Byte 1: Node number: 0x1 to 0x0c for Node 1 - 12 Response Data: Byte 1: Completion code (0x00) or out of range (0xC9) Byte 2: Node Power State (0x00: Power OFF; 0x10: S3; 0x20: No permission; 0x40: Fault; 0x80: Power ON) Byte 3: Width Byte 4: Height Byte 5: Permission state (0x00 Not present; 0x01: Standby; 0x02: First permission fail; 0x03: Second permission fail; 0x04: Permission pass)
Reset/reseat Node	0x32	0xa4	Request Data: Byte 1: Node number: 0x1 to 0x0c for Node 1 - 12 Byte 2: Reset action: 1: reset, 2: reseat Response Data: Byte 1 – completion code (0x00) or cur not support (0x0xd5)
Show information about node size	0x32	0x99	Request Data: Byte 1: Node number: 0x1 to 0x0c for Node 1 - 12 Response Data: Byte 1: Completion code (0x00) or out of range (0xC9) Byte 2: Node Physical Width Byte 3: Node Physical Height Byte 4: Add-on Valid Byte 5: Add-on Width Byte 6: Add-on Height

Description	NetFn	CMD	Data
Show Node Power Consumption in watts	0x32	0x98	Request Data: Byte 1 options: ▶ Node number: 0x1 to 0x0c for Node 1 - 12 ▶ Chassis: 0x0d Response Data: Byte 1: Completion code (0x00) Byte 2: Power minimum (LSB) Byte 3: Power minimum (MSB) Byte 4: Power average (LSB) Byte 5: Power average (MSB) Byte 6: Power maximum (LSB) Byte 7: Power maximum (MSB)

Table 7-11 Miscellaneous IPMI commands

Description	NetFn	CMD	Data
Set Time	0x32	0xa1	Request Data: Byte 1: Year MSB (1970 - 2037) Byte 2: Year LSB (1970 - 2037) Byte 3: Month (0x01-0x12) Byte 4: Date (0x01-0x31) Byte 5: Hour (0x00-0x23) Byte 6: Minute (0x00-0x59) Byte 7: Second (0x00-0x59) Example: Year 2010 (byte1: 0x20; byte2; 0x10) Response Data: Byte 1: Completion code (0x00)
Get FPC Status	0x32	0xa8	Request Data: None Response Data: Byte 1: Completion code (0x00) Byte 2: FPC major version Byte 3: FPC minor version Byte 4: PSOC major version Byte 5: PSOC minor version Byte 6: Boot Flash number (0x1-0x2) Byte 7: Build major number Byte 8: Build minor number (ASCII value)

Examples

This section provides some examples of how to use the IPMI interface to obtain data. Depending on the command that is requested, the parameters and the output have a different format. For more information about specific command formats, see the following tables:

- ▶ Table 7-3 on page 142
- ▶ Table 7-4 on page 142
- ▶ Table 7-5 on page 143
- ▶ Table 7-6 on page 144
- ▶ Table 7-7 on page 144
- ▶ Table 7-8 on page 144
- ▶ Table 7-9 on page 145
- ▶ Table 7-10 on page 145
- ▶ Table 7-11

Get power consumption of a node

To get power consumption of node 1 (idle node), use the following command (see Table 7-10 on page 145 for the command syntax):

```
ipmitool -I lanplus -U USERID -P PASSWORD -H 192.168.0.100 raw 0x32 0x98 0x1
00 2c 00 2c 00 2e 00
```

The output string has the following meaning:

- ▶ Byte 1: Completion code 0x00
- ▶ Byte 2 and 3: Power minimum: 0x002C (44 W)
- ▶ Byte 4 and 5: Power average: 0x002C (44 W)
- ▶ Byte 6 and 7: Power maximum: 0x002E (46 W)

To get the power consumption of node 3, use the following command (see Table 7-10 on page 145 for the syntax):

```
ipmitool -I lanplus -U USERID -P PASSWORD -H 192.168.0.100 raw 0x32 0x98 0x3
00 93 00 94 00 97 00
```

The output string has the following meaning:

- ▶ Byte 1: Completion code 0x00
- ▶ Byte 2 and 3: Power minimum: 0x0093 (147 W)
- ▶ Byte 4 and 5: Power average: 0x0094 (148 W)
- ▶ Byte 6 and 7: Power maximum: 0x0097 (151 W)

Get status of a node

To get the status of node 1, use the following command (see Table 7-10 on page 145 for the syntax):

```
ipmitool -I lanplus -U USERID -P PASSWORD -H 192.168.0.100 raw 0x32 0xa7 0x1
00 80 01 01 04
```

The output string has the following meaning:

- ▶ Byte 1: Completion code 0x00
- ▶ Byte 2: Node power state 0x80 (Power ON)
- ▶ Byte 3: Width 0x01 (1U)
- ▶ Byte 4: Height 0x01 (1U)
- ▶ Byte 5: Permission state 0x04 (Permission pass)

Get power supply fan status

To get power supply fan status, use the following command (see Table 7-8 on page 144 for the syntax):

```
ipmitool -I lanplus -U USERID -P PASSWORD -H 192.168.0.100 raw 0x32 0xa5 0x1
b0 14 14 02
```

The output string has the following meaning:

- ▶ Byte 1 and 2: Fan speed 0x14B0 (5296 rpm)
- ▶ Byte 3: Fan speed% 0x14 (20%)
- ▶ Byte 4: Fan speed status 0x02 (normal)

Set acoustic mode

To set acoustic mode of the chassis to Mode 2, use the following command (see Table 7-6 on page 144 for the syntax):

```
ipmitool -I lanplus -U USERID -P PASSWORD -H 192.168.0.100 raw 0x32 0x9b 0x2
00
```

In the output string, byte 1 refers to completion code 0x0.

7.3 ServeRAID C100 drivers: nx360 M4

The ServeRAID C100 is an integrated SATA controller with software RAID capabilities. It is a cost-effective way to provide reliability, performance, and fault-tolerant disk subsystem management to help safeguard your valuable data and enhance availability.

IBM ServeRAID C100 RAID support must be enabled by pressing F1 at the setup menu.

RAID support for Windows and Linux only: No RAID support for VMware, Hyper-V, or Xen; for these operating systems, it can be used as a non-RAID SATA controller only.

By using the F1 setup menu, the MegaCLI command line utility, and the MegaRAID Storage Manager, a storage configuration must be created for the use of the software RAID capabilities. For more information about the setup and configuration instructions, see the ServeRAID C100 User's Guide, which is available at this website:

<http://ibm.com/support/entry/portal/docdisplay?ln docid=MIGR-5089055>

Operating System device drivers must be installed to use the ServeRAID C100 software RAID capabilities. The drivers are available at this website:

<http://ibm.com/support/entry/myportal/docdisplay?ln docid=MIGR-5089068>

You also can find the latest device drives for the different operating systems that support ServeRAID C100 controller and download links to the configuration tools, MegaCLI, and MegaRAID Storage Manager, with which you can create a storage configuration under the respective operating system.

7.4 Integrated SATA controller: nx360 M5

The NeXtScale nx360 M5 features an onboard SATA controller that is integrated in to the Intel C612 chipset. It supports one 3.5 inch simple-swap SATA or near line (NL) SATA drive, two 2.5 inch simple-swap NL SATA drives, or four 1.8-inch SATA solid-state drives (SSDs). Two 2.5-inch simple swap NL SATA drives or four 1.8-inch SATA SSDs can also be used with a RAID controller or SAS HBA that is installed in the internal RAID adapter riser slot.

Two 2.5-inch simple swap SAS drives or two 2.5-inch hot swap drives that are installed in the front drive bays require a RAID controller or SAS HBA that is installed in the internal RAID adapter riser slot.

7.5 VMware vSphere Hypervisor

The NeXtScale compute nodes support VMware vSphere Hypervisor (ESXi) that is installed on a USB memory key. The server provides the option of adding a blank USB memory key for the installation of the embedded VMware ESXi.

The VMware ESXi embedded hypervisor software is a virtualization platform with which multiple operating systems can be run on a host system at the same time.

Lenovo provides different versions of VMware ESXi customized for IBM hardware that can be downloaded from this website:

http://shop.lenovo.com/us/en/systems/solutions/alliances/vmware/#tab-vmware_vsphere_esxi

For more information about installation instructions, see *vSphere Installation and Setup Guide*, which is provided as part of the downloaded image.

7.6 eXtreme Cloud Administration Toolkit

The eXtreme Cluster Administration Toolkit (xCAT) 2 is an Open Source Initiative that was developed by IBM to support the deployment of large high-performance computing (HPC) clusters that are based on various hardware platforms. xCAT 2 is not an evolution of the earlier xCAT 1. Instead, it is a complete code write that combines the best practices of Cluster Systems Management (CSM) and xCAT 1. xCAT 2 is open to the general HPC community under the Eclipse License to help support and enhance the product in the future.

xCAT provides a scalable distributed computing management and provisioning tool that provides a unified interface for hardware control, discovery, remote control management, and operating system diskfull and diskless deployment.

xCAT 2 uses only scripts, which makes the code portable and modular in nature and allows for the easy inclusion of more functions and plug-ins.

The xCAT architecture includes the following main features:

- ▶ Client/server architecture

Clients can run on any Perl-compliant system (including Windows). All communications are SSL encrypted.

- ▶ Role-based administration

Different users can be assigned various administrative roles for different resources.

- ▶ Stateful, stateless, and iSCSI nodes provisioning support

Stateless nodes can be RAM-root, compressed RAM-root, or stacked NFS-root. Linux software initiator iSCSI support for Red Hat Enterprise Linux (RHEL) and SUSE Linux Enterprise Server (SLES) is included.

Systems without hardware-based initiators also can be installed and booted by using iSCSI.

► Scalability

xCAT 2 can scale to 100,000 and more nodes with xCAT's Hierarchical Management Cloud. A single management node can have any number of stateless service nodes to increase the provisioning throughput and management of the largest clusters. All cluster services, such as, LDAP, DNS, DHCP, NTP, and Syslog are configured to use the Hierarchical Management Cloud. Outbound cluster management commands (for example, **rpower**, **xdsh**, and **xdcp**) use this hierarchy for scalable systems management. An example of such a hierarchy is shown in Figure 7-37.

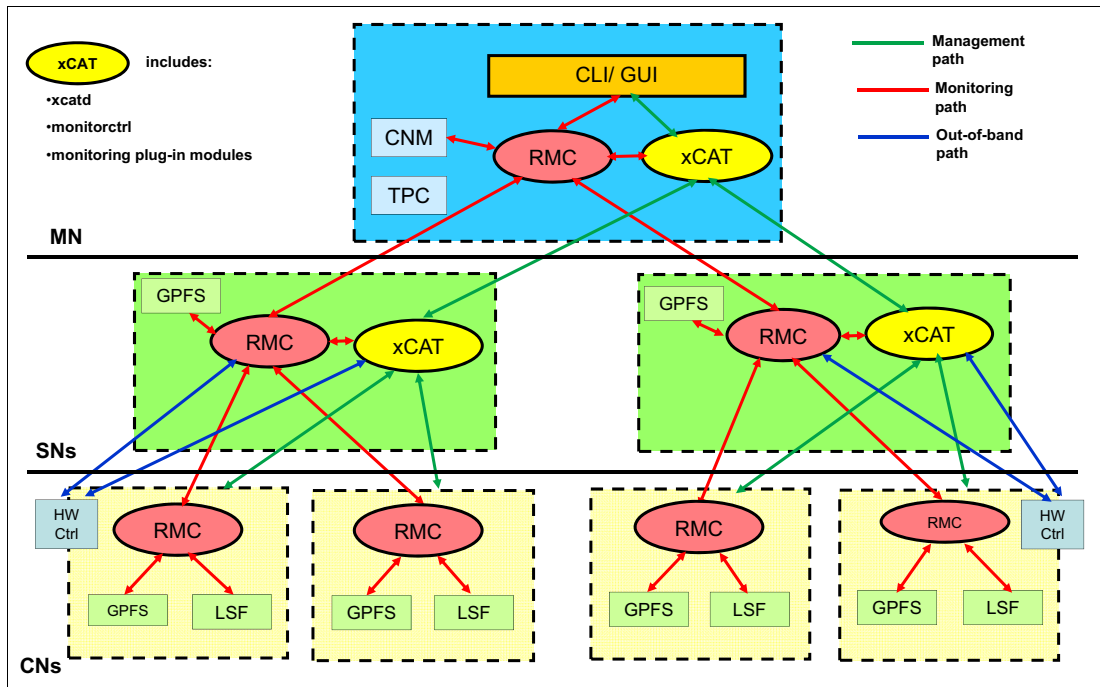


Figure 7-37 xCAT hierarchical cluster management

► Automatic discovery

This feature includes single power button press, physical location-based discovery, and configuration capability. Although this feature is mostly hardware-dependent, xCAT 2 was developed to ease integration for new hardware. The plug-in software architecture provides an easy development mechanism for the new hardware.

► Plug in architecture for compartmental development

By using this feature, you can add your own xCAT functionality to do whatever you want. New plug-ins extend the xCAT vocabulary that is available to xCAT clients.

► Notification infrastructure

By using this feature, you can watch for xCAT DB table changes via the notification infrastructure.

► SNMP monitoring

Default monitoring uses SNMP trap handlers to handle all SNMP traps.

► Flexible monitoring infrastructure

You can easily integrate third-party vendor monitoring software into the xCAT cluster. Currently, the following plug-ins are provided with xCAT:

- SNMP
- RMC (RSCT)

- Ganglia
- Performance Copilot
- ▶ Centralized console and system logs

xCAT provides console access to managed nodes and centralized logging.

xCAT 2 evolved to support several operating systems, including AIX, and the many derivatives of Linux, including SLES, openSUSE, RHEL, CentOS, and Fedora Core. xCAT also can provision Windows 2008 through imaging and virtual machine (VMware, Xen, KVM) images to hosted systems. Because xCAT is an open source project, other operating systems can be added to the supported list in the future.

With AIX and Linux, xCAT supports traditional local disk, SAN disk, and stateful diskless, which provisions via native deployment methods. Also, support is provided for stateless diskless nodes including ramfs root, compressed ramfs root, and NFS root with ramfs overlay support (Linux and AIX) and stateful diskless that uses iSCSI (Linux).

The xCAT manager works with the relevant hardware control units within the nodes, BladeCenter, and HMC to instruct these units to perform a number of hardware functions or gather information about the hosts.

The following hardware control features are included:

- ▶ Power control (power on, off, cycle, and current state)
- ▶ Event logs
- ▶ Boot device control (full boot sequence on IBM System BladeCenter, next boot device on other systems)
- ▶ Sensor readings (temperature, fan speed, voltage, current, and fault indicators as supported by systems)
- ▶ Node MAC address gathering
- ▶ LED status and modification (ability to identify LEDs on all systems, and diagnostic LEDs on select IBM rack-mount servers)
- ▶ Serial-over-LAN (SOL): Use or redirect input and output
- ▶ Service processor configuration
- ▶ Hardware control point discovery by using Service Location Protocol (SLP): BladeCenter Advanced Management Module (AMM), IBM Power Systems HMC, and Flexible Service Processor (FSP).

Supported hardware

The following hardware is supported and tested by xCAT:

- ▶ IBM BladeCenter with AMM
- ▶ IBM System x
- ▶ IBM Power Systems (including the HMC)
- ▶ IBM iDataPlex
- ▶ IBM Flex System
- ▶ IBM NeXtScale System
- ▶ Machines that are based on the IPMI

Because only IBM hardware is available for testing, this hardware is the only hardware that is supported. However, other vendors' hardware can be managed with xCAT. For more information and the latest list of supported hardware, see this website:

<http://xcat.sourceforge.net/>

Note: Although xCAT is an Open Source Initiative that is provided under the Eclipse license and freely available, official IBM support can be contacted.

Abbreviations and acronyms

AC	alternating current	DPC	deferred procedure call
ACPI	advanced control and power interface	DRAM	dynamic random access memory
ASCII	American Standard Code for Information Interchange	DSA	Dynamic System Analysis
ASHRAE	American Society of Heating, Refrigerating, and Air-Conditioning Engineers	DVD	Digital Video Disc
ASR	automatic server restart	DW	data warehousing
ASU	Advanced Settings Utility	DWC	direct water cooling
AVX	Advanced Vector Extensions	ECC	error checking and correcting
BIOS	basic input output system	EPOW	Early Power Off Warning
BMC	Baseboard Management Controller	FAN	Fabric Address Notification
BSMI	Bureau of Standards, Metrology and Inspection	FC	Fibre Channel
CB	Certification Body	FCC	Federal Communications Commission
CCC	China Compulsory Certificate	FDR	fourteen data rate
CD	compact disk	FLOPS	floating-point operations per second
CD-ROM	compact disc read only memory	FOD	features on demand
CDU	chiller distribution unit	FP	floating point
CE	Conformité Européene	FPC	Fan and Power Controller
CFF	common form factor	GB	gigabyte
CFM	cubic feet per minute	GFLOPS	giga floating-point operations per second
CIM	Common Information Model	GOST	gosudarstvennyy standart (state standard)
CISPR	International Special Committee on Radio Interference	GPU	Graphics Processing Unit
CLI	command-line interface	GT	Gigatransfers
CMD	command	GUI	graphical user interface
CMOS	complementary metal oxide semiconductor	HBA	host bus adapter
CNA	Converged Network Adapter	HCA	host channel adapter
CPU	central processing unit	HD	high definition
CRC	cyclic redundancy check	HDD	hard disk drive
CSA	Canadian Standards Association	HPC	high performance computing
CTO	configure-to-order	HPL	High-Performance Linpack
DC	direct current	HS	hot swap
DDR	Double Data Rate	HW	hardware
DHCP	Dynamic Host Configuration Protocol	I/O	input/output
DIMM	dual inline memory module	I/OAT	I/O Acceleration Technology
DNS	Domain Name System	IB	InfiniBand
DOS	disk operating system	IBM	International Business Machines
		ID	identifier
		IEC	International Electrotechnical Commission

IEEE	Institute of Electrical and Electronics Engineers	PCI-E	PCI Express
IMM	integrated management module	PDU	power distribution unit
IOPS	I/O operations per second	PE	Preinstallation Environment
IP	Internet Protocol	PEF	platform event filtering
IPMI	Intelligent Platform Management Interface	PET	Platform Event Trap
ISO	International Organization for Standards	PF	power factor
IT	information technology	PFA	Predictive Failure Analysis
JBOD	just a bunch of disks	PN	part number
KB	kilobyte	PSOC	Programmable System-on-Chip
KVM	keyboard video mouse kernel virtual machine	PSU	power supply unit
LAN	local area network	PXE	Preboot eXecution Environment
LDAP	Lightweight Directory Access Protocol	QDR	quad data rate
LED	light emitting diode	QPI	QuickPath Interconnect
LFF	large form factor	RAID	redundant array of independent disks
LOM	LAN on motherboard	RAS	remote access services; row address strobe
LP	low profile	RDIMM	registered DIMM
LRDIMM	load-reduced dual inline memory module	RDMA	Remote Direct Memory Access
LSB	Least Significant Bit	RDS	Reliable Datagram Sockets
MAC	media access control	RHEL	Red Hat Enterprise Linux
MB	megabyte	ROC	RAID-on-card
MLC	multi-level cell	ROM	read-only memory
MPI	Message Passing Interface	RPM	revolutions per minute
MSB	Most Significant Bit	RSS	Receive-side scaling
MSI	Message Signaled Interrupt	SAS	Serial Attached SCSI
MTM	machine type model	SATA	Serial ATA
MTU	maximum transmission unit	SDDC	Single Device Data Correction
NC-SI	Network Controller-Sideband Interface	SDR	Single Data Rate
NFS	network file system	SED	self-encrypting drive
NIC	network interface card	SEL	System Event Log
NL	nearline	SFF	Small Form Factor
NUMA	Non-Uniform Memory Access	SFP	small form-factor pluggable
NVGRE	Network Virtualization using Generic Routing Encapsulation	SHMEM	Symmetric Hierarchical Memory
OEM	other equipment manufacturer	SIG	special interest group
OFED	OpenFabrics Enterprise Distribution	SKU	stock keeping unit
OS	operating system	SLES	SUSE Linux Enterprise Server
PCH	Platform Controller Hub	SLP	Service Location Protocol
PCI	Peripheral Component Interconnect	SMB	server message block
		SMP	symmetric multiprocessing
		SMTP	simple mail transfer protocol
		SNMP	Simple Network Management Protocol
		SOL	Serial over LAN
		SR-IOV	single root I/O virtualization

SRP	Storage RDMA Protocol
SS	simple swap
SSCT	Standalone Solution Configuration Tool
SSD	solid state drive
SSE	Streaming SIMD Extensions
SSH	Secure Shell
SSP	Serial SCSI Protocol
TB	terabyte
TCO	total cost of ownership
TCP	Transmission Control Protocol
TCP/IP	Transmission Control Protocol/Internet Protocol
TDP	thermal design power
TOE	TCP offload engine
TPM	Trusted Platform Module
TSS	Trusted Computing Group Software Stack
TUV-GS	Technischer Überwachungs-Verein Geprüfte Sicherheit (TUV tested safety)
TX	transmit
UDIMM	unbuffered DIMM
UDP	user datagram protocol
UEFI	Unified Extensible Firmware Interface
URL	Uniform Resource Locator
USB	universal serial bus
UXSP	UpdateXpress System Packs™
VCCI	Voluntary Control Council for Interference
VFA	Virtual Fabric Adapter
VGA	video graphics array
VLAN	virtual LAN
VM	virtual machine
VPD	vital product data
WCT	Water Cool Technology
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.

Lenovo Press publications

The following Lenovo Press publications provide additional information about the topic in this document.

- ▶ NeXtScale System M5 with Water Cool Technology Product Guide
<http://lenovopress.com/tips1241>
- ▶ NeXtScale System M5 with Water Cool Technology Video Walk-through
<http://www.youtube.com/watch?v=0kd4pVmUCPg>
- ▶ *NeXtScale System Planning and Implementation Guide*, SG24-8152
<http://lenovopress.com/sg248152>
- ▶ xREF - System x Reference
<http://lenovopress.com/xref>
- ▶ Lenovo Press Product Guides for System x servers and options
<http://lenovopress.com/systemx>

Other publications and online resources

These publications and websites are also relevant as further information sources:

- ▶ US Announcement Letter
<http://ibm.com/common/ssi/cgi-bin/ssialias?infotype=dd&subtype=ca&&htmlfid=897/ENUS114-142>
- ▶ NeXtScale System home page
<http://shop.lenovo.com/us/en/systems/servers/high-density/nextscale-m5/>
- ▶ *NeXtScale nx360 M5 and NeXtScale n1200 Installation and Service Guide*
<http://ibm.com/support/entry/portal/docdisplay?ln docid=MIGR-5096549>
- ▶ Power Configurator
<http://ibm.com/support/entry/portal/docdisplay?ln docid=LNVO-PWRCONF>
- ▶ NeXtScale Power Guide
<http://ibm.com/support/entry/portal/docdisplay?ln docid=LNVO-POWINF>
- ▶ Configuration and Option Guide
<http://www.ibm.com/systems/xbc/cog/>
- ▶ System x Support Portal
<http://ibm.com/support/entry/portal/>

Lenovo

Lenovo NeXTScale System Water Cool Technology Planning and Implementation



(0.5" spine)
0.475" x 0.873"
250 x 459 pages

Lenovo

Lenovo NeXTScale System Water Cool Technology Planning and Implementation Guide

(0.2" spine)
0.17" x 0.473"
90 x 249 pages



Lenovo NeXtScale System Water Cool Technology Planning and Implementation Guide

The ultimate in performance and energy efficiency

Describes the warm water cooled offerings from Lenovo

Covers the n1200 WCT Enclosure and nx360 M5 WCT Compute Node

Addresses power, cooling, water flow, and racking

NeXtScale System is a dense computing offering based on Lenovo's experience with iDataPlex and Flex System and with a tight focus on emerging and future client requirements. The NeXtScale n1200 Enclosure and NeXtScale nx360 M5 Compute Node are designed to optimize density and performance within typical data center infrastructure limits.

The 6U NeXtScale n1200 Enclosure fits in a standard 19-inch rack and up to 12 compute nodes can be installed into the enclosure. The WCT warm water-cooled versions of the NeXtScale n1200 and nx360 M5 offer the ultimate in performance and energy efficiency. With more computing power per watt and the latest Intel Xeon processors, you can reduce costs while maintaining speed and availability.

This Lenovo Press publication is for customers who want to understand and implement a water-cooled NeXtScale System solution. It introduces the offering and the innovations in its design, outlines its benefits, and positions it with other x86 servers. The book provides details about NeXtScale System components and supported options, and provides rack, power and water planning considerations.



**BUILDING
TECHNICAL
INFORMATION
BASED ON
PRACTICAL
EXPERIENCE**

At Lenovo Press, we bring together experts to produce technical publications around topics of importance to you, providing information and best practices for using Lenovo products and solutions to solve IT challenges.