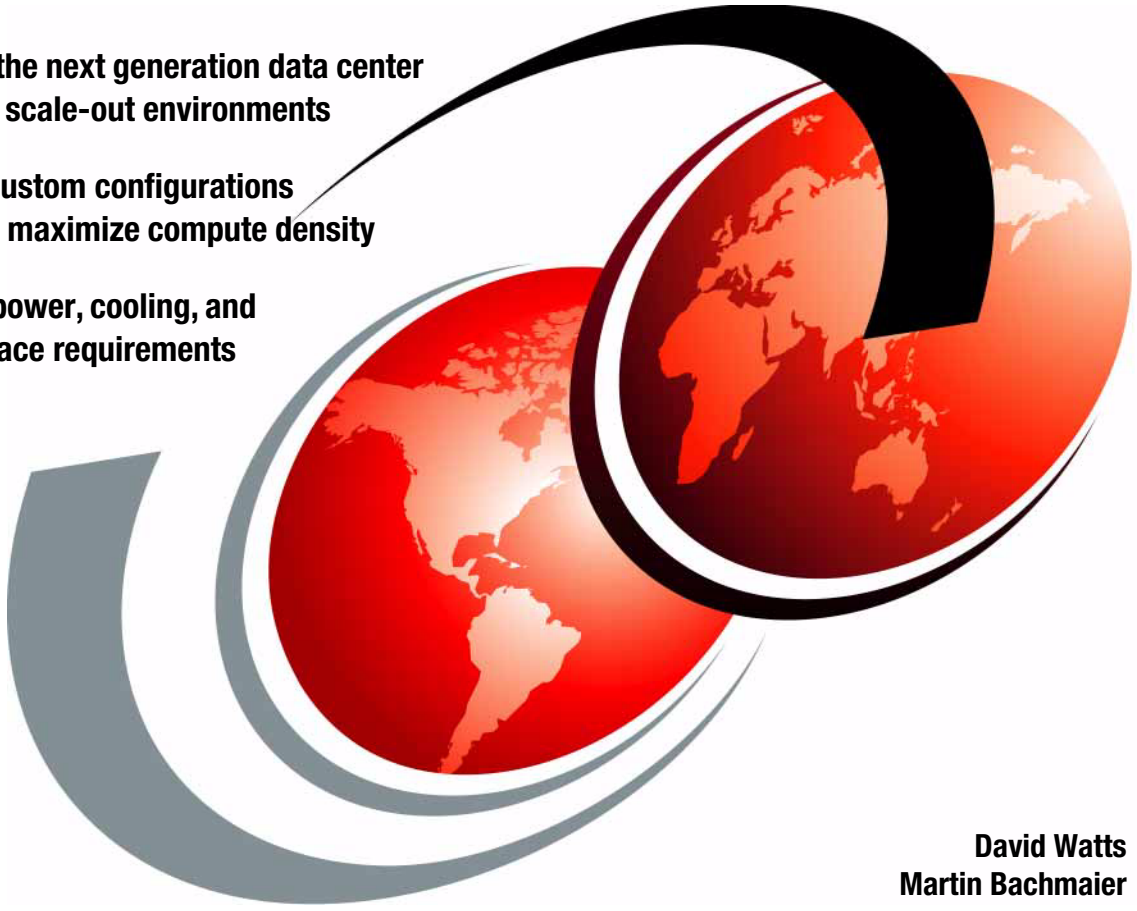
IBM

Implementing an IBM System x iDataPlex Solution

- Introduces the next generation data center solution for scale-out environments
- Describes custom configurations designed to maximize compute density
- Addresses power, cooling, and physical space requirements



David Watts
Martin Bachmaier

Redbooks

ibm.com/redbooks

International Technical Support Organization

Implementing an IBM System x iDataPlex Solution

June 2012

SG24-7629-04

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

Fifth Edition (June 2012)

This edition applies to the IBM System x iDataPlex solution comprising the following products:

IBM System x iDataPlex Rack, machine type 7825
IBM iDataPlex dx360 M3, machine type 6391
IBM iDataPlex dx360 M4, machine type 7912
IBM iDataPlex Direct Water Cooled dx360 M4, machine type 7918
IBM iDataPlex Direct Water Cooled Rack, machine type 7197

Contents

Notices

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

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

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

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

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

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

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

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

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

COPYRIGHT LICENSE:

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

Trademarks

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

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

Blade Network Technologies®
BladeCenter®
BNT®
Cool Blue®
DB2®
DS4000®
Global Technology Services®
GPFS™
IBM Systems Director Active Energy Manager™
IBM®
iDataPlex®
Intelligent Cluster™
LoadLeveler®
Power Systems™
Redbooks®
Redbooks (logo) ®
RETAIN®
ServerProven®
System x®

The following terms are trademarks of other companies:

Intel Xeon, Intel, 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.

Intel Xeon, Intel, and Trusted Source of Network Truth are trademarks or registered trademarks of Intelliden, Inc., an IBM Company.

Advanced Micro Devices, AMD, HyperTransport, the AMD Arrow logo, and combinations thereof, are trademarks of Advanced Micro Devices, Inc.

Blade Network Technologies, BNT, RackSwitch, and the BLADE logo are registered trademarks of BLADE Network Technologies, Inc.

Emulex, and the Emulex logo are trademarks or registered trademarks of Emulex Corporation.

InfiniBand Trade Association, InfiniBand, and the InfiniBand design marks are trademarks and/or service marks of the InfiniBand Trade Association.

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

FiberRunner, PANDUIT, and the PANDUIT logo are registered trademarks of Panduit Corporation in the United States, other countries, or both.

QLogic, and the QLogic logo are registered trademarks of QLogic Corporation. SANblade is a registered trademark in the United States.

Red Hat, and the Shadowman logo are trademarks or registered trademarks of Red Hat, Inc. in the U.S. and other countries.

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

LifeKeeper, SteelEye, and the SteelEye logo are registered trademarks of Steeleye Technology, Inc. Other brand and product names used herein are for identification purposes only and may be trademarks of their respective companies.

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

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

Preface

The IBM® iDataPlex® data center solution is for Web 2.0, high performance computing (HPC) cluster, and corporate batch processing customers experiencing limitations of electrical power, cooling, physical space, or a combination of these. By providing a big picture approach to the design, iDataPlex uses innovative ways to integrate Intel-based processing at the node, rack, and data center levels to maximize power and cooling efficiencies while providing necessary compute density.

An iDataPlex rack is built with industry-standard components to create flexible configurations of servers, chassis, and networking switches that integrate easily. Using technology for flexible node configurations, iDataPlex technology can configure customized solutions for applications to meet specific business needs for computing power, storage intensity, and the right I/O and networking.

This IBM Redbooks® publication is for customers who want to understand and implement the IBM iDataPlex solution. It introduces the iDataPlex solution and the innovations in its design, outlines its benefits, and positions it with IBM System x and BladeCenter® servers. The book provides details of iDataPlex components and the supported configurations. It describes application considerations for an iDataPlex solution and aspects of data center design that are central to that solution. The book concludes by introducing the services offerings available from IBM for planning and installing an iDataPlex solution.

The team who wrote this book

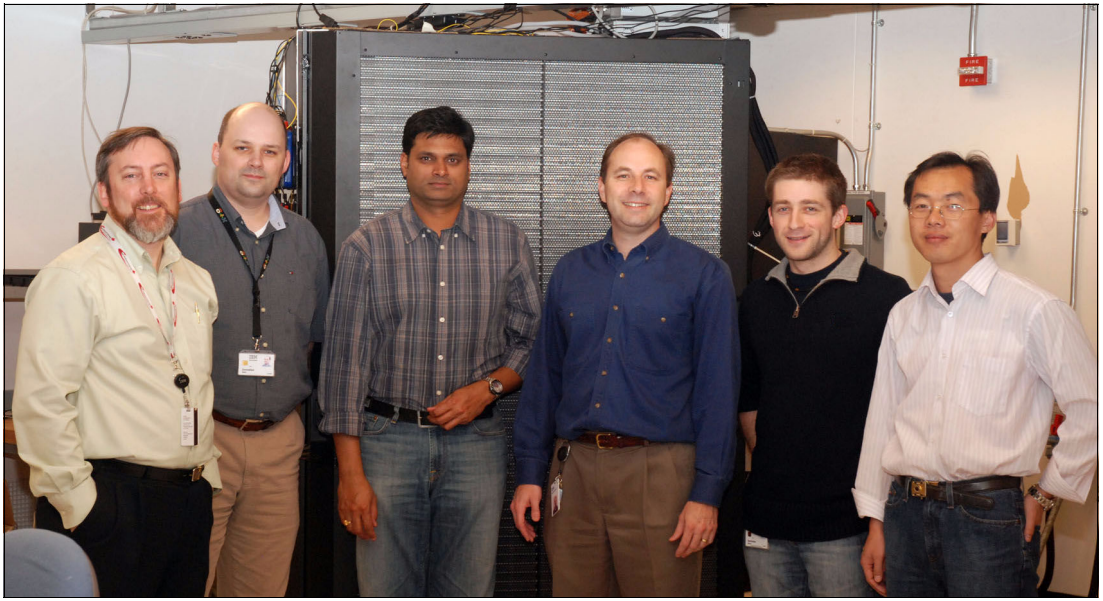
This Fifth Edition was produced by a team of specialists from around the world working at the International Technical Support Organization (ITSO), Raleigh Center.

David Watts is a Consulting IT Specialist at the IBM ITSO Center in Raleigh. He manages residencies and produces Redbooks publications on hardware and software topics related to IBM System x® and BladeCenter servers and associated client platforms. He has authored over 80 books, papers, and technotes. He holds a Bachelor of Engineering degree from the University of Queensland (Australia) and has worked for IBM both in the United States and Australia since 1989. He is an IBM Certified IT Specialist.

Martin Bachmaier is an IT Versatilist in the IBM development lab in Boeblingen, Germany, and currently working in the Systems Optimization Competency Center. Over the last years he has implemented several high-performance compute clusters and archiving systems, including the first hot-water cooled supercomputer in Switzerland. Martin holds a degree in Computer Science from the University of Cooperative Education in Stuttgart, Germany, and a Bachelor of Science from the Open University in London, UK. He has worked for IBM for more than eight years. He is an IBM Certified Systems Expert and holds the CCNA, CCNA Security, and VMware Certified Professional credentials.

Thanks to the authors of the previous editions of this book:

- David Watts
- Srihari Angaluri
- Martin Bachmaier
- Sean Donnellan
- Duncan Furniss
- Kevin Xu



The original author team (left to right): Duncan, Sean, Srihari, David, Martin, and Kevin

The authors want to thank the following people for their contributions to this project:

From the International Technical Support Organization:

- Tamikia Barrow
- Mary Comianos
- Mike Ebbers
- Linda Robinson
- Diane Sherman
- Alfred Schwab

From iDataPlex Marketing:

- Gaurav Chaudhry
- Amy Freeman
- Danika Loadholt
- Monica Martinez
- Marshall Shah

From iDataPlex Development:

- Nadia Anguiano-Wehde
- Ralph Begun
- Chu Chung
- Alan Claassen
- Rodrigo De la Parra
- Tiep Dinh
- Karl Dittus
- Eric Eckberg
- Noel Fallwell
- Kellie Francis
- Dottie Gardner
- John Gossett
- Dennis Hansen
- Tim Hiteshew
- Christina Hsu
- Vinod Kamath
- Hawk Lin
- Vic Mahaney
- Mark Megarity
- Stephen Mroz
- Whit Scott
- Roger Schmidt
- Scott Shurson
- Junjiro Sumikawa
- Yu-Wen Yang

- Cindy Yang

IBM Services Offerings:

- Brian Canney
- David Franko
- Michael Karchov
- Bret Lehman
- Rey Rodriguez
- Steven Sams

IBM employees from around the world:

- Christine Kubsch, Germany
- Ingmar Meijer, Switzerland
- Peter Morjan, Germany
- Ronnie Winick, Research Triangle Park, NC USA

From other companies:

- Chris Doehlert, DataDirect Networks
- Alvin Galea, Australia Post
- Rick Korn, SMC
- Scott Lorditch, Blade Networking Technologies
- Tom Boucher and team, Panduit Corp.

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author - all at the same time! Join an ITSO residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

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

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications 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

- Mail your comments to:

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

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

Summary of changes

This section describes the technical changes made in this edition of the book and in previous editions. This edition might also include minor corrections and editorial changes that are not identified.

Summary of Changes
for SG24-7629-04
for *Implementing an IBM System x iDataPlex Solution*
as created or updated on June 29, 2012.

June 2012, Fifth Edition

This revision reflects the addition, deletion, or modification of new and changed information described in the following sections.

- IBM iDataPlex dx360 M4
- IBM iDataPlex Direct Water Cooled dx360 M4
- Intel Xeon E5-2600 series processors
- New Direct Water Cooled dx360 M4 2U chassis
- New Direct Water Cooling technology
- New 80 PLUS Platinum certified power supplies
- New chassis trays for the iDataPlex dx360 M4
- New GPU and GPGPU options for the dx360 M4
- Updated options for the iDataPlex dx360 M3:
  - New processor options
  - New memory options
  - New internal storage options
  - New PCIe adapters
  - New ServeRAID controllers
  - New GPU and GPGPU options
- New Ethernet switches
- New InfiniBand switches and HCAs
- New external storage options
- Updated HPC scenario
- New scenario for a warm water-cooled data center

September 2010, Fourth Edition

This revision reflects the addition, deletion, or modification of new and changed information described in the following sections.

- IBM Intelligent Cluster™ offering, replacing the System 1350 process
- IBM iDataPlex dx360 M3
- Intel Xeon 5600 Series Processors
- New 550 W and dual AC 750 W power supplies
- New storage options
- New memory options, including low-voltage DIMMs
- New processor options
- New ServeRAID controller options
- New chassis trays (HDD and PCIe tray)
- New SAS 6 Gbps options
- New solid-state drive options, including high IOPS SSD PCIe adapter
- New PCIe options (QDR InfiniBand, 10 Gig Ethernet, FCoE)
- New NVIDIA GPU and GPGPU options
- New Ethernet switches, including Converged Fabric support
- New InfiniBand switches and HCAs
- Updated HPC scenario
- Updated Cluster Resource Management and Moab
- Information about the new IBM eX5 product portfolio

August 2009, Third Edition

This revision reflects the addition, deletion, or modification of new and changed information described in the following sections.

New information

- IBM iDataPlex dx360 M2
- Intel Xeon 5500 Series Processors
- New 3U configuration
- 450 GB SAS drive
- New InfiniBand options
- New server configurations with SSD-based SATA drives
- New 900W power supply option for dx360 M2
- New ServeRAID BR10il RAID controller
- New 10 Gbps Ethernet options
- New SMC and IBM BNT® Ethernet switch options
- IBM Intelligent Management Module (IMM)
- Information about UEFI
- Cluster Resource Management and MOAB

March 2009, Second Edition

This revision reflects the addition, deletion, or modification of new and changed information described in the following sections.

New information

- Positioning iDataPlex solution with BladeCenter
- IBM iDataPlex dx320
- IBM iDataPlex dx360
- Table comparing the dx320, dx340 and dx360 servers
- Details about internal SAS connectivity
- Force10 2410CP and SMC 8024L2 switch
- Support for 1TB SATA drive
- Information about heterogeneous data center
- Information about iDataPlex installation in an enterprise rack
- Information about cabling best practices for the iDataPlex rack
- Information about cabling overhead pathways
- Frequently asked questions (FAQs) and answers throughout the book

Changed information

- Operating system support

1

Introduction

The rapid growth of data volumes and the increasing demands for compute power over the last few years have presented significant new challenges related to the acquisition, deployment, operation, and management of large-scale data centers and IT equipment. IT managers are extremely sensitive to the costs, skills, and manpower requirements associated with such high growth.

As the industry starts adopting new technology models of IT outsourcing, such as Cloud Computing, Software-as-a-Service (SaaS), and Infrastructure-as-a-Service (ITaaS), the current and future demands for compute capacity will keep steadily growing over the next few years. These new technologies, coupled with growing user demands, are pushing the envelope when it comes to the need for raw compute power, storage, networking, and other associated data center elements.

As a result, new concerns become paramount for IT shops: IT acquisition costs, growing energy bills from data center power and cooling demands, data center operational costs, and the ongoing equipment maintenance costs, all of which are part of the Total Cost of Ownership (TCO) equation. The primary goal today for data center owners is to minimize the TCO and to get the most out of their current and future IT investments.

The growing IT demands not only require the expansion of existing data centers, but in some cases also force the creation of new massive scale-out data centers that are designed specifically with a view to optimizing the energy, cooling, and manageability aspects of the IT infrastructure. This is really critical for IT businesses entering such new markets as Cloud Computing because the TCO becomes a significant influencer of the profits for the company. How efficiently you manage your IT infrastructure will determine how profitable you are in these new markets.

In summary, as data centers continue to grow in order to meet new business demands, the challenges that will be faced can be categorized as follows:

- Increased power and cooling requirements for servers and related IT hardware
- Constraints of floor space, which mandate more compute power in a limited amount of space
- The need for efficient and effective manageability of massive scale-out environments
- The need to minimize initial acquisition costs and subsequent maintenance costs for the IT infrastructure

1.1 Data center facts

A recent report from IDC, a well-known IT research firm, presents the following interesting facts about data centers:

- Data center power density is increasing by approximately 15% annually.
- Resulting power draws per rack have grown eightfold since 1996.
- Over 40% of data center customers report that power demand is outstripping supply.

IDC concludes “Data center infrastructure is challenged to keep pace.”[1]

According to a survey by AFCOM and InterUnity Group, which was quoted in a SearchDataCenter.com report, “In two years, 44.5% of AFCOM member’s data centers would be incapable of supporting business requirements due to capacity constraints.”[2]

---

[1] See IDC report “The Impact of Power and Cooling on Data Center Infrastructure:” http://www.ibm.com/systems/z/advantages/energy/index.html#analyst_rep

[2] See SearchDataCenter.com E-book, *The Green Data Center: Energy-Efficient Computing in the 21st century*: http://searchdatacenter.techtarget.com/general/0,295582,sid80_gci1273283,00.html

Professor Jonathan Koomey of Stanford University concludes in his paper "Estimating Total Power Consumption By Servers In The U.S. And The World" that for every kW of power used to drive servers, on average another kW of power is needed to drive infrastructure devices such as power and air conditioning.[3]

1.1.1 Project Green IT

Given these alarming facts, various commercial companies, hardware manufacturers, as well as the governments of several nations have started thinking about addressing current energy issues and the plan to meet future power demands. According to the IT research firm Gartner, Inc., report "Gartner Identifies the Top 10 Strategic Technologies for 2008," Green IT is ranked among the top 10.[4]

Green IT is the term coined to indicate the new paradigm shift in designing data centers to be much more efficient in power consumption and cooling, and to reduce carbon emissions and potential environmental impacts, compared to the traditional data center designs. Gartner predicts that new government regulations have the potential to constrain companies in designing data centers to meet these environmental requirements. In the wake of these challenges and business requirements to stay competitive, companies are actively looking for solutions at various levels from server manufacturers and other infrastructure vendors.

1.1.2 Next generation data center model

To address the new data center challenges, a new way of designing data centers and the server infrastructure that goes into the data centers is essential. The design must encompass data center power, cooling, management, and acquisition as well as operating costs as the chief design goals, in addition to the performance, cost, and scalability requirements at the server level.

In an effort to address various critical IT challenges, IBM has for several decades developed and introduced to the market innovative technologies and established itself as the industry leader. As data center computing enters the next generation, IBM has once again focused on developing technologies that will address the critical data center challenges to meet the Green IT goals, which will revolutionize data center design and computing models over the next decade.

---

[3] See *Estimating Total US & Global Server Power Consumption*:
http://enterprise.amd.com/us-en/AMD-Business/Technology-Home/Power-Management.aspx

[4] See Gartner press release:
http://www.gartner.com/it/page.jsp?id=530109

1.1.3 New philosophy

As part of the new thinking, a new philosophy has emerged, which promises to deliver Green IT goals. The philosophy states that instead of designing servers and putting them in data centers, *first design the ideal data center and then design servers and racks specifically for this design*. This philosophy is the basis for the System x iDataPlex technology.

1.2 Introducing iDataPlex

To address the growing data center challenges as outlined previously, IBM has taken a new approach and designed an innovative family of products and technologies called *iDataPlex*.

IBM has years of experience designing server technologies for scale-up and scale-out settings that primarily focus on performance and scalability as the fundamental requirements. However, iDataPlex focuses on a different set of goals:

- Reduce the initial hardware acquisition costs and on-going maintenance costs for data center owners
- Improve efficiency in power consumption
- Eliminate data center cooling requirements
- Achieve higher server density within the same footprint as the traditional rack layout
- Simplify manageability for massive scale-out environments
- Reduce the time to deployment through pre-configuration and full integration at manufacturing

As is evident, these design goals go far beyond a single server or a single rack level; they are goals for the entire data center.

With the new philosophy and the new design, iDataPlex promises to address the data center challenges at various levels:

- An innovative rack design achieves higher node density within the traditional rack footprint.
- An innovative *flex node* chassis and server technology are based on industry standard components.
- Shared power and cooling components improve efficiency at the node and rack level.

- An optional Rear Door Heat eXchanger virtually eliminates traditional cooling based on computer room air conditioning (CRAC) units.
- Various networking, storage, and I/O options are optimized for the rack design.
- Direct water cooling of each node provides unparalleled advantages, such as lower node power, improved performance per watt, or the ability to reuse waste heat for economical purposes such as warming other buildings.

Each of these innovations is described in the following sections. Subsequent chapters explain the technology and products in more detail.

1.2.1 The iDataPlex rack

The iDataPlex rack cabinet design offers 100 rack units (U) of space, as opposed to the traditional enterprise rack cabinet, which has 42U of space. In that sense, the iDataPlex rack is essentially two 42U racks connected together and provides additional vertical bays. The iDataPlex rack is shown in Figure 1-1.

Figure 1-1 iDataPlex 100U rack

However, the iDataPlex rack is shallower in depth compared to a standard 42U server rack as shown in Figure 1-2. The iDataPlex rack is 600 mm deep (840 mm with the Rear Door Heat eXchanger) compared to the 42U rack which is 1050 mm deep. The shallow depth of the rack and the iDataPlex nodes is part of the reason that the cooling efficiency of iDataPlex is higher than the traditional rack design, because air travels a much shorter distance to cool the internals of the server compared to airflow in a traditional rack.

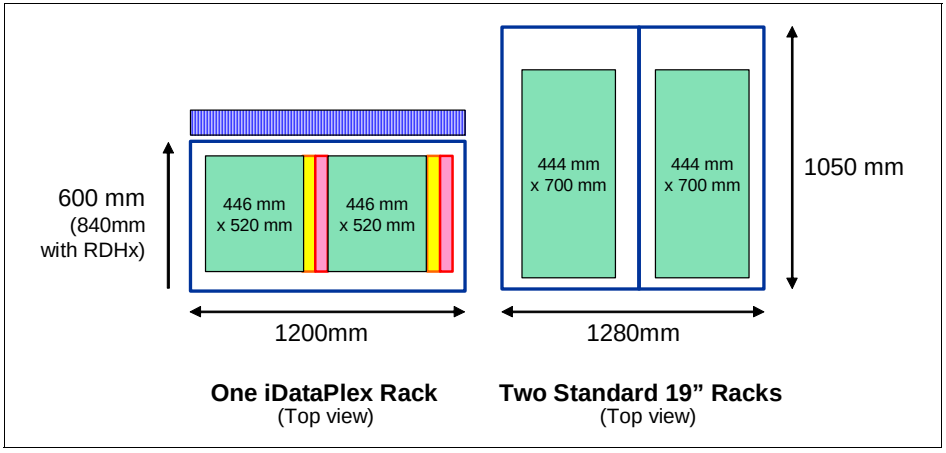


Figure 1-2 Comparison of iDataPlex with two standard 42U racks (top view)

A single iDataPlex 100U rack enclosure provides more than twice the compute density of a standard 42U rack enclosure within the footprint of a single 42U rack. The iDataPlex rack has 84 horizontal 1U slots for server chassis and 16 vertical slots for network switches, PDUs, and other appliances.

The rack is oriented so that servers fit in side-by-side on the widest dimension. For ease of serviceability, all hard drive, planar, and I/O access is from the front of the rack. There is little need to access the rear of the iDataPlex rack for any serviceability other than to service the Rear Door Heat eXchanger, the power cords, and the power distribution units (PDUs).

Note: Although the iDataPlex rack provides significant advantages for your data center, IBM also supports the installation of iDataPlex servers in standard 19” racks.

1.2.2 FlexNode technology

The modular design of the iDataPlex components allows for customized server solutions to meet specific business needs. The servers for iDataPlex can be configured in numerous ways by using FlexNode technology, which is an innovative 2U modular chassis design. The iDataPlex 2U FlexNode chassis can hold one server with multiple drives or two servers in a 2U mechanical enclosure. Figure 1-3 shows two possible configurations of the iDataPlex 2U FlexNode chassis.

Figure 1-3 Two examples of configurations in the iDataPlex 2U Flex chassis

For maximum computing density, customers will typically want to populate each chassis with two servers. In addition to the compute-oriented configurations, iDataPlex solution offers a 3U storage-rich server. The 3U chassis can accommodate up to 12 hot swap SAS or SATA drives for vast amounts of storage.

Racks are orderable and deliverable with any combination of configured chassis in the rack up to a total of 84 nodes in 42 chassis. This approach provides maximum flexibility for data center solutions to incorporate a combination of configurations in a rack.

1.2.3 Support for GPGPU-based computing

The idea of leveraging the extreme power of graphics processing units (GPUs), which are traditionally used to offload graphics-related rendering operations in a computer, to offload general purpose (GP) application workloads and significantly speed up part of the overall computation, and in some cases to improve the computational detail and/or accuracy, is quickly becoming mainstream.

Industries such as oil and gas, numerical weather forecasting, and financial simulations, to name a few, can take advantage of GPGPU-based computation and speed up its calculations. The most promising candidates for GPGPU-based computation are applications that are easily parallelizable and consist of heavy floating point calculations. See Figure 1-4.

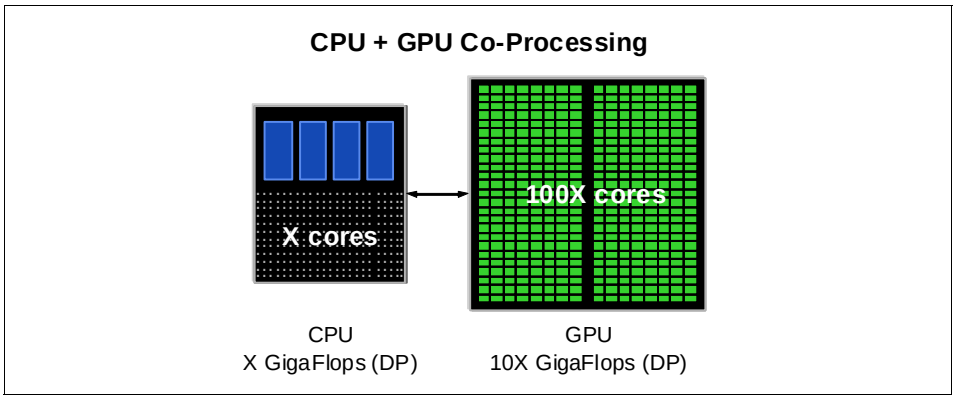


Figure 1-4 Combining CPU and GPU for computations

GPUs are supported with all available iDataPlex servers by additional chassis trays. Depending on the tray, each server can connect to up to two standard form-factor PCIe GPU cards (such as NVIDIA Tesla). Combining the power of the new Intel Xeon processor E5-2600 family (codename *SandyBridge-EP*) and GPU cards provides significant compute power within each iDataPlex chassis.

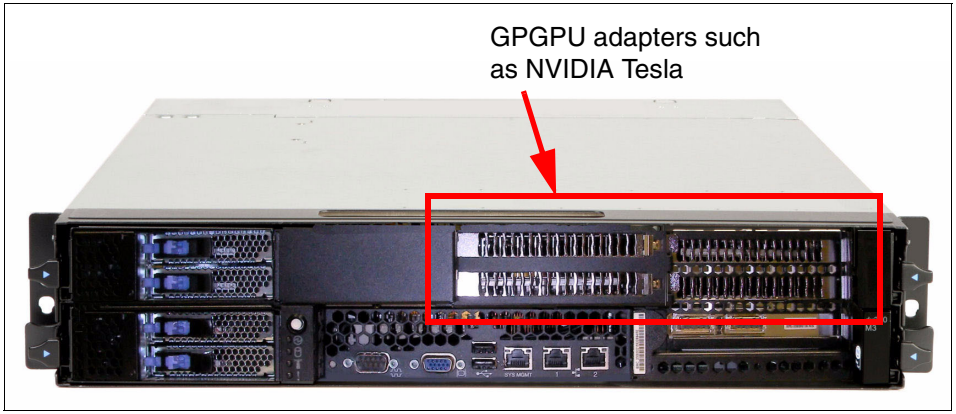


Figure 1-5 iDataPlex Server with GPGPUs

More detailed information about GPGPU support is provided in these sections:

- dx360 M4: 4.3.7, “Graphics Adapter and General Purpose GPU (GPGPU)” on page 57
- dx360 M3: 4.5.5, “Graphics adapter and GPGPUs” on page 104

1.2.4 Managing the iDataPlex systems

iDataPlex can be deployed in racks of tens or hundreds of servers (also referred to as nodes in a typical high-performance computing cluster environment). In order to effectively and efficiently manage such volume of servers, robust tools will be required.

During the typical lifecycle of the iDataPlex systems, IT administrators will be required to manage the following aspects:

- Initial hardware configuration, such as node firmware update and BIOS configuration
- Provisioning the operating system on the nodes. such as Linux or Windows
- Remote hardware control, including power control, Serial-over-LAN (SOL), and remote event log collection, and so on
- Software updates
- Remote cluster operations such as parallel shell
- User configuration (home directories, applications, and so on), and network services such as DNS, NIS, and NFS
- Other miscellaneous operations

Since the iDataPlex solution is expected to grow large-scale, any tools that provide these services are expected to scale to such a level (that is, to hundreds or even thousands of servers).

IBM has invested in Linux and open source tools for several years now in order to encourage community participation, openness, and to leverage best industry practices in managing IT infrastructures. In line with that philosophy, IBM initiated an open source project called Extreme Cluster Administration Toolkit (xCAT) almost ten years ago. The tool was originally intended to help IT managers effectively deploy and manage a large-scale HPC cluster system. Since its original development, xCAT has grown to become a widely adopted industry tool for managing large-scale cluster and grid computing systems. Today, xCAT is used as the de facto tool for all IBM HPC cluster systems, as well as iDataPlex.

xCAT has wide industry participation in terms of contributions from leading industry and academic IT managers, and field technical specialists, as well as experts inside IBM that routinely deploy large-scale cluster systems at customer sites to make xCAT the best-of-breed tool for cluster management.

xCAT is an open source product based on the Eclipse Public License[5]. It is available free of charge and hosted on SourceForge at this website:

http://xcat.sourceforge.net

1.2.5 Onboard manageability

Each compute node has an integrated management module (IMM) that provides service-processor functions such as monitoring server availability, performing Predictive Failure Analysis, or triggering system alerts in case of an error. If an environmental condition exceeds a threshold, or if a system component fails, the IMM lights LEDs to help you diagnose the problem and records the error in the event log. The IMM also provides remote media and remote control functionality and supports the IPMI 2.0 protocol for remote server management.

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

[5] See http://www.eclipse.org/legal/epl-v10.html

1.3 Energy-efficient water cooling

With the ever-increasing power consumption of today's data centers, it becomes vital to run them at the most efficient level possible. You want your electrical power, which you pay for, to provide as much compute power as possible.

Most of the current data centers are extremely inefficient. It means that a substantial amount of energy (that is, money) is required only to drive the infrastructure for running servers. It includes, for example, data center air-conditioning and chillers.

Did you know? The world-wide Information and Communications Technology (ICT) industry accounts for approximately the same amount of carbon-dioxide emission as the aviation industry.

By going from conventional air cooling to water as your media for dissipating heat, you can achieve substantial energy savings that translate directly to less operational costs for your data center.

The iDataPlex solution supports two options of water-cooling:

- At the server level, using Direct Water Cooled servers
- At the rack level, using a Rear Door Heat eXchanger

Both cooling solutions are mutually exclusive.

1.3.1 Direct water cooling

Initiated by a data center research project at IBM's Zurich Research Laboratory, the latest generation of iDataPlex servers support a new cooling technology that extends the water loop into the servers.

Using new microchannel heatsink, this cooling solution absorbs heat directly from the source (processor or memory chip modules) and dissipates it into water.

After the heat is in the water, it can be transported easily. Water is a very good medium for transporting heat because it can carry 4000 times more heat than air.

A broad range of temperatures is supported, including warm water of more than 40°C (100°F). The exact temperature values depend on the specifics of your data center.

Figure 1-6 shows a simplified drawing of how to run a data center using Direct Water Cooling technology. The servers on the left side of the figure are cooled with warm water (primary cooling loop in the middle of the figure). The heat is transferred into a secondary loop that can directly pass through buildings to warm them.

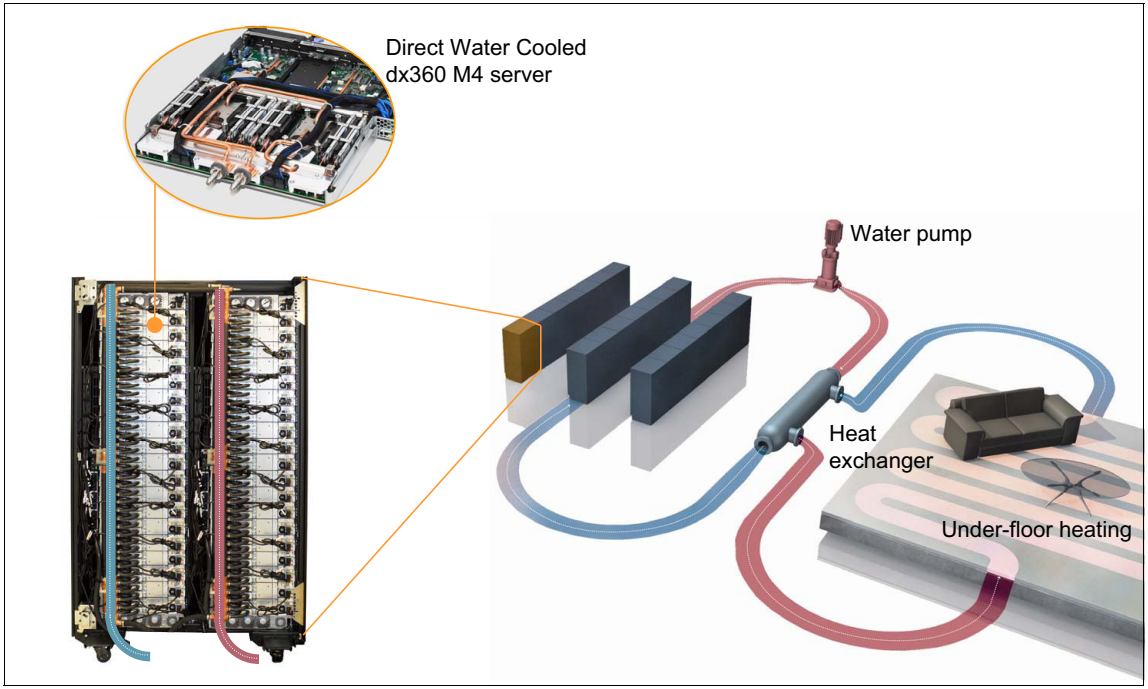


Figure 1-6 Direct Water Cooling technology showing direct reuse of the waste heat to warm a building (image courtesy of IBM Research – Zurich)

More technical information about Direct Water Cooling can be found in 4.4.1, “Direct Water Cooling technology” on page 66.

1.3.2 Rear Door Heat eXchanger

The Rear Door Heat eXchanger is a water-cooled door that is mounted on the rear of an IBM iDataPlex rack 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 from the heat exchanger.

For customers who are able to cool their data centers with water, the Rear Door Heat eXchanger can withdraw 100% or more of the heat coming from a 100,000 BTU (approximately 30 kWh) rack of servers to alleviate the cooling challenge that many data centers are having. By selecting the correct water inlet temperature and water flow rate, you can achieve optimal heat removal.

The Rear Door Heat eXchanger is shown in Figure 1-7. The inlet and outlet connectors are shown at the bottom of the photo.

Figure 1-7 Rear Door Heat eXchanger (rear view of the iDataPlex rack)

1.4 Summary

Next-generation scale-out data centers are going to be severely constrained for power, cooling, floor space, and acquisition and maintenance costs, and will be subject to various government regulations for power efficiency, carbon emissions, and environmental impact.

The iDataPlex solution defines a new server architecture based on the iDataPlex 100U rack cabinet, Flex Node and Chassis, direct water cooling technology, and Rear Door Heat eXchanger. The iDataPlex solution will not only help customers address the data center power, cooling, and floor space concerns, but also does it at an attractive price point for customers, and is delivered to the customer preconfigured and prewired for quicker deployment.

The subsequent chapters describe the individual components of the iDataPlex technology in greater detail and present various aspects related to designing and implementing highly efficient data centers using IBM iDataPlex technology.

For more information, see this website:

http://www.ibm.com/systems/x/hardware/idataplex/

2

Benefits of iDataPlex

This chapter describes how the iDataPlex solution uniquely addresses data center challenges. The iDataPlex solution is designed from the ground up, from the compute facility, to the rack level, and to the individual servers. This approach addresses costs and efficiencies from procurement, through implementation, to operation. The iDataPlex solution is a range of systems. Since the introduction of iDataPlex in 2008, the ecosystem of servers and options has grown similar to the BladeCenter systems that IBM introduced in 2002.

In this chapter, we introduce the iDataPlex rack and the range of systems that are currently available. We discuss how this new product increases computing density at the rack level and at the room level. We also discuss how the iDataPlex solution is procured and delivered.

The following topics are covered in this chapter:

2.1 Acquisition and operating costs

Paramount among the development objectives for the iDataplex solution is to address areas of acquisition and operating costs in order to lower them and simplify the process for customers who want to build large-scale data centers.

2.1.1 Acquisition

iDataplex fully integrated solutions are acquired either directly through IBM direct sales channels or through IBM business partners. Members of the IBM iDataPlex sales group have the required skills, and will involve the appropriate IBM experts to guide an iDataPlex solution through the configuration, review, manufacturing, delivery, and installation.

Many options are provided by third-party vendors (including Juniper, Cisco, and SMC). Many of the internal components are sourced from high-volume manufacturers. When market conditions coincide with appropriately sized orders of iDataPlex, IBM will make spot purchases of components rather than build from existing inventory, taking advantage of opportunities to minimize the acquisition cost. These savings are then passed on to the customer.

When purchased as an Intelligent Cluster, iDataPlex solutions are custom-built at the time of the order, reflecting the customer's specific requirements. Implementations of the iDataPlex scale are unlikely to proceed without planning considerations on several fronts, such as:

- Computational, storage, and networking needs
- Financial considerations
- Real estate, power, cooling, and cabling provisioning

Thus, if the iDataPlex solution is built while these considerations are being addressed, it can be scheduled for delivery when the data center is ready to receive it. While the exact time frame can be determined when the order is placed, the estimated delivery time can be approximately three months.

These process decisions, as well as the system engineering discussed in Chapter 4, "iDataPlex compute nodes" on page 33, were made to drive the acquisition cost (often referred to as *capital expenditure*, or CAPEX) as low as possible. The price of iDataPlex hardware is substantially lower than traditional tier-one products, and is comparable to white box products.

2.1.2 Operating costs

Operating costs (*operational expenditure*, or OPEX) are the real estate (floor space), electrical power consumption for operation, energy used to maintain temperature and humidity, and manageability costs. These expenses are fundamental considerations in the design of iDataPlex solutions.

The current focus on power and cooling is indicative of two phenomena:

- Acknowledgment of just how expensive power and cooling are in the operation of a computing facility.
- Use of a computing model that provides performance and reliability through software, computing in a scalable and fault-tolerant way on multiple servers, thus decreasing the emphasis on wringing maximum performance and creating the highest levels of reliability within individual servers.

This focus enables iDataPlex engineers to improve power efficiency in many ways, also enabling them to simplify the server design, which also lends itself to greater reliability, lower cost, and higher efficiency. The power supply and fans in each chassis are optimized for efficiency in powering and cooling the nodes. Using 80 mm fans for all configurations provides commonality of parts and are more efficient than using additional 40 mm fans to do the same work. The planars use chip sets that are balanced performers while having low power consumption.

The overall node design is short from front to back, and special attention is given to reducing resistance to airflow. The more easily the cooling air can flow through the server, the less energy is used by the fans to move that air. Direct water cooling helps to lower costs even further through the use of an efficient water-cooling technology that absorbs heat where it is generated, from within the server.

Aside from moving data centers to lower-cost locations, the solution to the real estate challenge is to increase overall solution density. The iDataPlex solution addresses this challenge at the server level, rack level, and data center level.

iDataPlex nodes comply with Intelligent Platform Management Interface (IPMI) standards, enabling integration with a broad spectrum of management methods. A wide selection of networking devices is available, so in addition to being able to integrate these devices in to the overall networking environment, the investment in network management expertise can be leveraged.

2.2 Integration-ready

In addition to lowering acquisition costs, having IBM preconfigure the iDataPlex solution at the manufacturing facility reduces CAPEX and OPEX further:

- Packaging that is not necessary after the solution is shipped is minimized, thereby reducing disposal requirements.
- Considerations of proper airflow for cooling, robust mechanical assembly, and in-rack cabling are addressed before the system arrives.
- Testing after final assembly minimizes (eliminates, ideally) servicing that has to be performed after the equipment arrives.

The iDataPlex user simply has to roll the rack into the data center, extend its levelling legs, connect the power, network, and (optionally) cooling water for it to be ready to use. An example of an iDataPlex rack as shipped is shown in Figure 2-1.

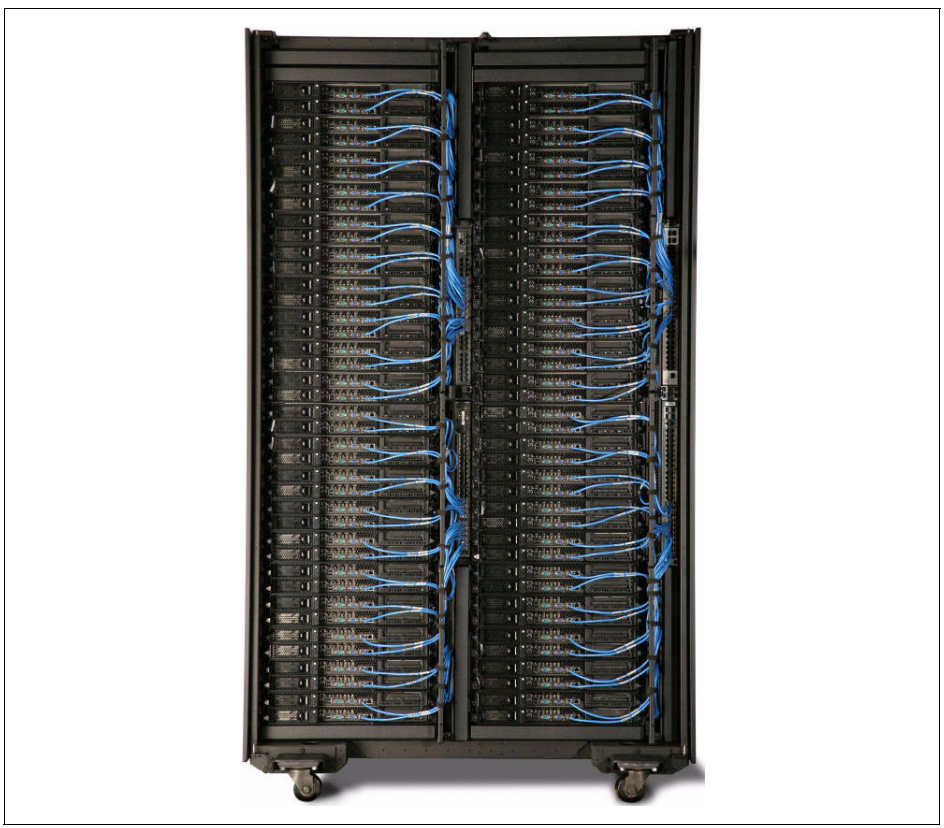


Figure 2-1 A preconfigured iDataPlex rack

2.3 Modularity and flexibility

The flexible design of iDataPlex provides cost-efficient servers in configurations to meet many requirements. The node design has a common power supply and fan assembly for all models, to minimize costs and maximize the benefits of standardization. The basis for the flex nodes is an industry standard motherboard based on the Server System Infrastructure (SSI) specification.

A flexible set of configurations can be created from common building blocks. The configurations are either computationally dense, I/O-rich, storage-rich (2U) and storage rich (3U). This modular approach to server design keeps costs low and provides a wide range of node types. Three of the four configurations are shown in Figure 2-2.

Figure 2-2 Three of the available iDataPlex FlexNode configurations

Having IBM integrate the servers, power distribution, and networking at the manufacturing facility saves deployment time and ensures design integrity, both of which reduce solution cost. The IBM specialist who designs the custom iDataPlex solution for the customer takes into consideration the rack positioning of the nodes and the airflow requirements of all components and provides instructions to manufacturing. An iDataPlex solution is built to be shipped, and arrive as a complete unit. The only packaging is the shipping crate. The crate provides protection for the unit in transit, can be moved by forklift, and provides a ramp for rolling the iDataPlex rack from the crate to the floor.

2.4 Rack density

The unique iDataPlex design maximizes floor space usage as shown in Figure 2-3. In the area of two data center floor tiles, there are two columns of 43U of 19-inch horizontal-mounting rack space. The Avocent rack management appliance, cables, or a filler panel consumes 1U per rack, which is why we specify a total of 84U available in horizontal orientation.

In addition, there are eight 2U vertical pockets, which can be subdivided into 1U pockets (labelled B1–B8 and D1–D8 in Figure 2-3).

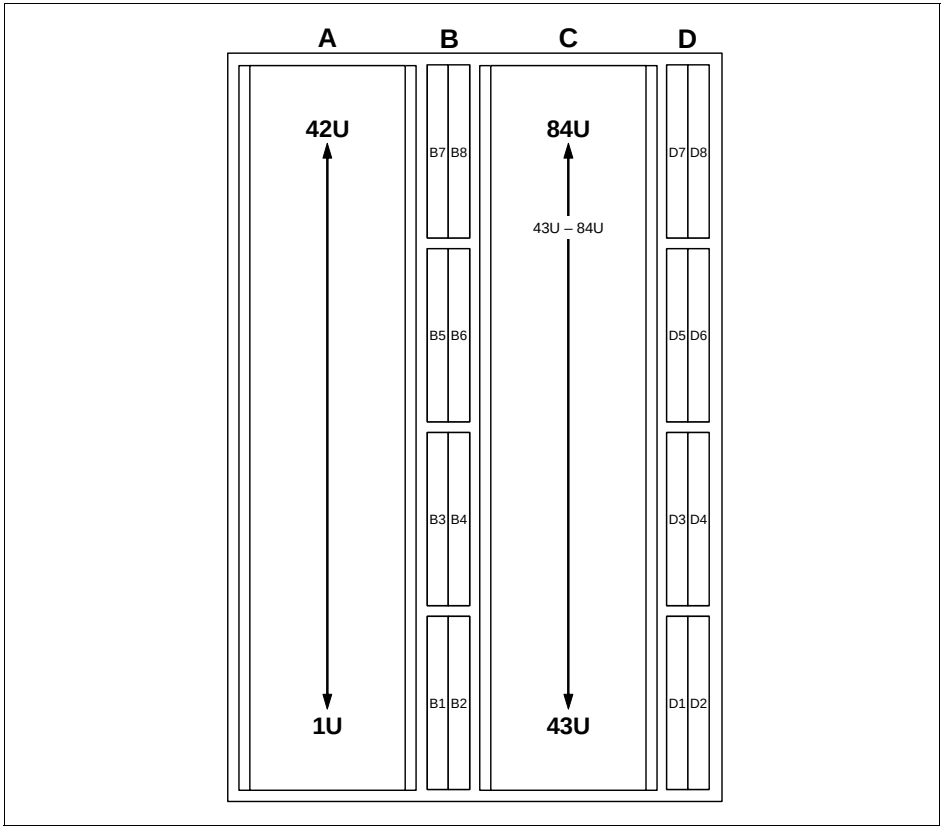


Figure 2-3 iDataPlex rack layout

The rack density has the potential for an additional benefit, which is the possibility, in certain cases, to reduce the number of power feeds going to the same number of servers when they are consolidated in one rack instead of two, as shown in Figure 2-4.

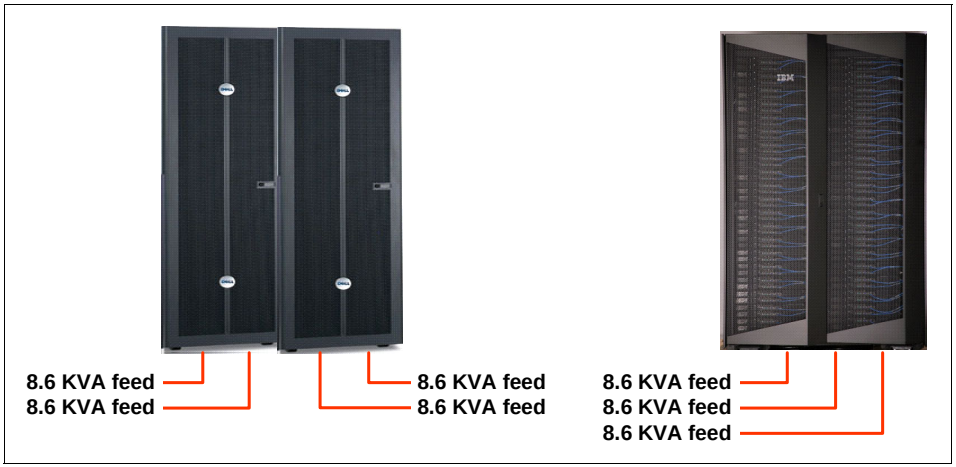


Figure 2-4 Power feed consolidation

One U.S.-based power company was charging its customers USD1500 - 2000 per month for each 8.6 kVA feed, in addition to the power consumption charges. One customer indicated that reducing the number of feeds per rack from four to three would save them about USD14 million per year.

2.5 Density at the data center level

The iDataPlex design allows for increased system density in the data center. The shallow depth of the rack allows for more rows of systems in the same area. Figure 2-5 shows possible density increases that the iDataPlex solutions enable.

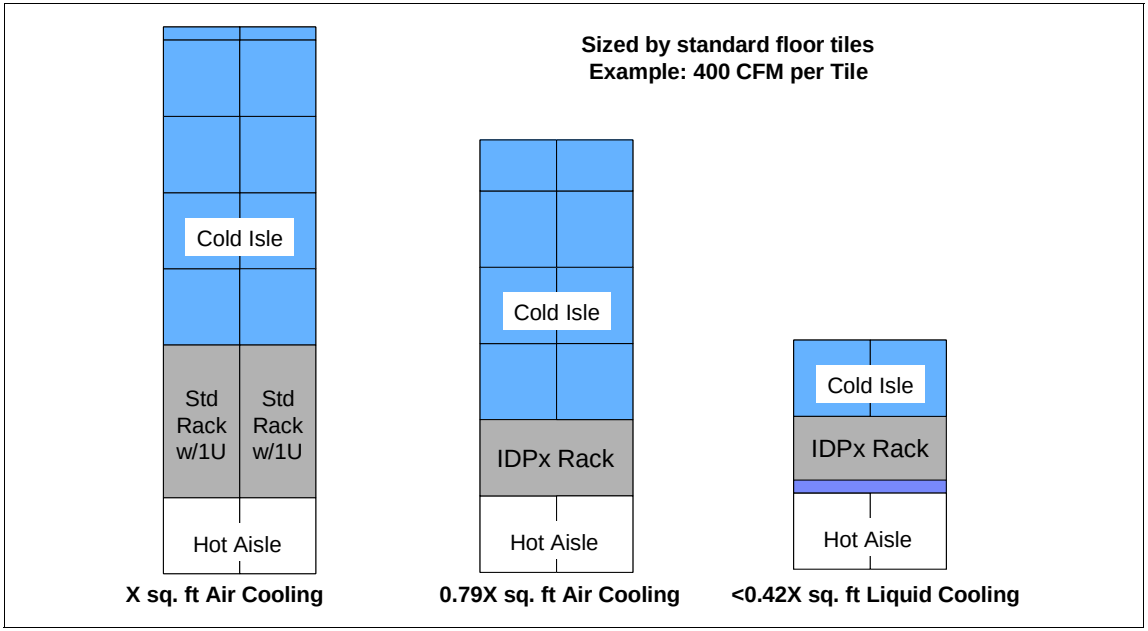


Figure 2-5 Data center density improvements enabled by iDataPlex solution

In addition, with 16U of vertical bay space available for power, networking, management, or other options, most iDataPlex designs can use all of the horizontal rack space for data processing, potentially reducing the number of racks required. The power efficiency of iDataPlex means less of the data center floor has to be given to cooling equipment.

2.6 Rear Door Heat eXchanger option

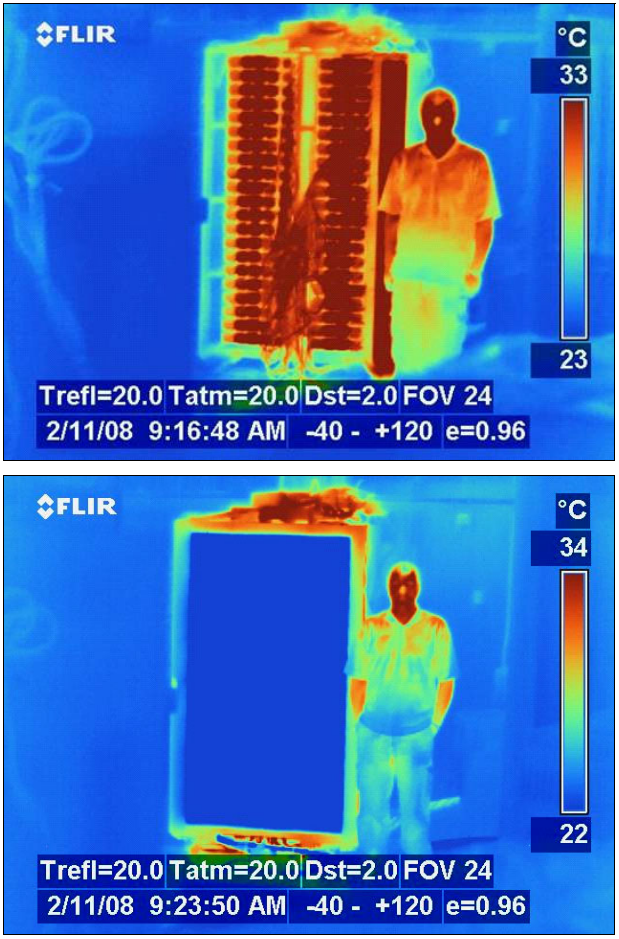
The Rear Door Heat eXchanger option for iDataPlex solution was developed to extract heat at rack level. A water-cooled door closes behind the iDataPlex rack and provides energy savings from not having to cool air with fans or blowers elsewhere in the computer room, as is done with conventional computer room air conditioner (CRAC) units.

Because of the design of the iDataPlex rack, the Rear Door Heat eXchanger has a large surface area for the number of servers it cools, making it very efficient. It can greatly reduce, or even eliminate the requirement for additional cooling in the server room, freeing space that is occupied by the numerous CRAC units that are usually required.

Any data center that uses CRAC systems already has chilled water installed, so integrating this solution is not difficult. As an added benefit, hot spots in the data center, which can be problematic to address, are less likely to occur.

Figure 2-6 shows thermal images taken of an iDataPlex rack under test in the IBM Thermal Lab, with a person standing beside it, before and after the heat exchanger was operational.

Even without water-cooling, the iDataplex solution is still at least 20% cooler than the conventional rack approach.



iDataPlex rack can run at room temperature (no air conditioning required) when used with the optional Rear Door Heat eXchanger.

These photos were taken in the Thermal Lab at IBM, showing the iDataPlex rack without (top image) and with (bottom image) the Rear Door Heat eXchanger closed and operational.

Figure 2-6 iDataPlex rack without and with the Rear Door Heat eXchanger closed and operational

2.7 Direct water cooling option

The direct water cooling technology is supported with the latest iDataPlex server. It extends the water cooling loop into the servers. Heat is directly dissipated from the microprocessor and memory modules into the cooling water flowing through very fine channels in the heat sink.

Using water as the medium to remove heat from server has the following immediate advantages:

- Lower node power
- Improved performance per watt
- Lower OPEX of your data center
- Lower noise level in your data center

New high performance microchannel heat sinks make it possible to increase the temperature of the cooling water compared to existing water cooling solutions. Even warm water can be used to cool the iDataPlex servers.

Using warm water as coolant has multiple advantages (over those mentioned):

- No CRACs or chillers required
- Direct reuse of the warm water to heat adjacent buildings

It also adds to saving money and lowering the energy-consumption of your data center (including adjacent buildings that save on heating costs).

In 2010, IBM Research – Zurich had built a prototype of a hot-water cooled data center in a project for the Swiss Federal Institute of Technology Zurich[1]. A standard IBM BladeCenter was retrofitted to enable water cooling of the IBM BladeCenter servers. This installation, dubbed *Aquasar*, successfully demonstrated a reduction of 40% in energy consumption compared to an equally equipped air-cooled installation.

Using a water inlet temperature of 60°C allowed the direct reuse of waste heat in the outlet water that is fed into the university heating installation. This reuse decreased the overall carbon dioxide emission of the installation by as much as 85%.

In 2011, the ASHRAE committee 9.9 has published[2] *Thermal Guidelines For Liquid Cooled Data Processing Environments*. They introduce five environmental classes specifying operating temperatures and required cooling equipment for each class. Using the iDataPlex Direct Water Cooled dx360 M4 servers allows your data center to run in the most efficient class W5.

[1] See the IBM press release at http://www-03.ibm.com/press/us/en/pressrelease/32049.wss
[2] You can download the paper free of charge from http://www.tc99.ashraetcs.org

3

Positioning

This chapter explains how iDataPlex is positioned in the marketplace compared with other systems that have Intel processors. It will help you to understand the iDataPlex target audience and the types of applications for which it is intended.

The following topics are covered in this chapter:

- Server and data center-level solution
- iDataPlex customers
- Hardware redundancy
- Scale-up versus scale-out
- Design objectives by platform
- IBM Intelligent Cluster and iDataPlex

3.1 Server and data center-level solution

iDataPlex is designed at the server level and considers the rack as a whole unit, taking into account the compute nodes, storage, I/O, switches, management, and power and cooling, to provide a complete customized solution. Table 3-1 shows a comparison of iDataPlex to traditional servers.

Table 3-1 Comparing different server form factors

Server type	Management level	Deployment	Relative node scale
Towers servers	Tower level management (typically remote)	By server	Individual servers 1–5+
Rack servers	Server level	By servers (into new or existing rack)	Individual servers 3–10+
BladeCenter	Chassis level	By blade/chassis (into new or existing rack)	BladeCenter chassis (4–14 blades per chassis)
iDataPlex	Rack level	By rack (into new or existing data center)	iDataPlex Rack (24–84 nodes per rack)
	Server level	By servers (into new or existing rack)	Individual servers 3–10+

3.2 iDataPlex customers

iDataPlex technology is a flexible, massive scale-out data center server solution built on industry standard components for customers who are looking for compute density and energy efficiency. Figure 3-1 on page 27 shows iDataPlex application positioning against our BladeCenter, rack, and high-end scale-up eX5 technology servers.

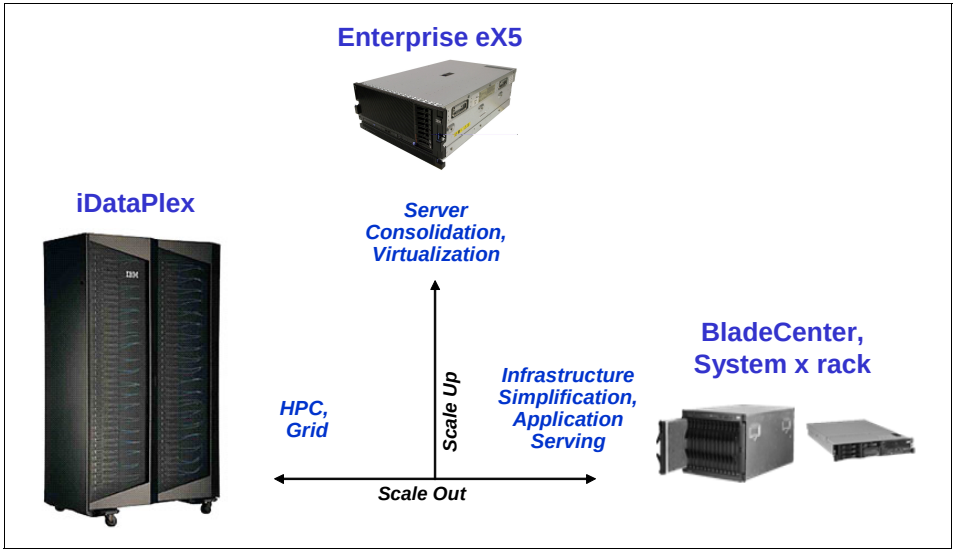


Figure 3-1 iDataPlex application positioning

iDataPlex is the right choice for customers that have these requirements:

- Facing power, cooling, and density challenges
- Having software redundancy built into their application and are comfortable with lower hardware redundancy
- Focusing on lowering their capital expense and operating expense

iDataPlex is targeted specifically to data-center-level customer applications that are based on stateless computing, including:

- High performance computing clusters
- Grid computing

High performance computing (HPC) is the use of parallel processing for running advanced application programs efficiently, reliably, and quickly. Clusters are typically comprised of components that could be used separately in other types of computing configurations. It includes compute nodes, network adapters and switches, local and external storage, and systems-management software and applications.

The iDataPlex solution provides the next generation of HPC, in these areas:

- Speed:
  - High-performance CPU
  - Large memory per node
  - High-speed interconnects

- Density:
  - From 2 to 3 times the density of 1U rack mount solution
  - Up to 2 nodes per chassis
  - Up to 84 nodes per rack
- Economy:
  - Lower capital outlay for acquisition
  - Up to 40% lower power consumption than standard 1U designs
  - Up to 75% reduction in cooling cost compared to 1U servers and traditional computer room air conditioning
  - Lower total cost of ownership

The iDataPlex solution is the fast, dense, and economical choice for HPC.

3.3 Hardware redundancy

The iDataPlex solution provides hardware redundancy, although hardware redundancy is not as important to iDataPlex customers because they use networking and software to maximize availability to end users. The iDataPlex solution is used in massive scale-out clusters that have many nodes. If one nodes fails, the software will transfer the workload to other nodes. There is little effect on the whole clusters.

Table 3-2 shows a comparison of redundancy features between iDataPlex, System x servers, and BladeCenter.

Table 3-2 Comparing redundancy components

Redundancy feature	iDataPlex	System x servers	System x Enterprise servers	BladeCenter
Memory	ECC Chipkill Mirroring Sparing	ECC Chipkill Mirroring Sparing	ECC Chipkill Memory ProteXion Mirroring	ECC Chipkill Mirroring Sparing
Storage RAID	Supported[a]	Supported	Supported	Supported
Dual Ethernet	Supported	Supported	Supported	Supported
Power supply	Supported	Supported[b]	Supported	Supported[c]
Fans	No	Supported[b]	Supported	Supported[c]

a. Supports hardware or software RAID based on server configuration.
b. Certain low-end System x servers do not have redundant fans and power supplies.
c. Feature provided by BladeCenter Chassis.

The iDataPlex servers support redundant Ethernet connections with the two onboard controllers. If a problem occurs with one Ethernet connection, all Ethernet traffic associated with it is automatically switched to the other interface (provided that the operating system, software, network cabling, and switch configuration are done properly to support this).

The iDataPlex server monitors fans, power, temperature, and voltage, but it does not provide redundancy on the fans. Node-level fault tolerance is required.

The input cabling of iDataPlex power supplies can be configured using two redundant inlet cables connected to two separate power feeds. If one of the feeds (including the rack PDU) fails, the iDataPlex servers continue operation with the remaining power feed.

Many applications are available with built-in redundancy and failover features. These applications include large-scale Web serving, scale-out databases, risk analytics, streaming media (IPTV, online movies), and video rendering.

3.4 Scale-up versus scale-out

The two design methods to expand a server's performance are scale-up and scale-out.

Scale-up means that adding additional resources, usually processors or memory, to a single server to increase the server's performance in areas such as processing capacity or I/O bandwidth to make a powerful server. Scaling up depends on hardware design. It enables simple system management as there are fewer systems and potentially lower software licensing costs. The IBM System x3850 X5 is an industry-leading product using the scale-up design.

Scale-out means adding more servers to a cluster or a compute farm to expand the overall capacity. Each of the servers in such a compute farm usually performs the same task and, provided the system is design for scaling, can be easily extended by adding more servers. Such a design depends on the whole architecture, hardware, and software. High performance compute clusters are a type of scale-out solution.

See Figure 3-2 for iDataPlex positioning in scale-up and scale-out.

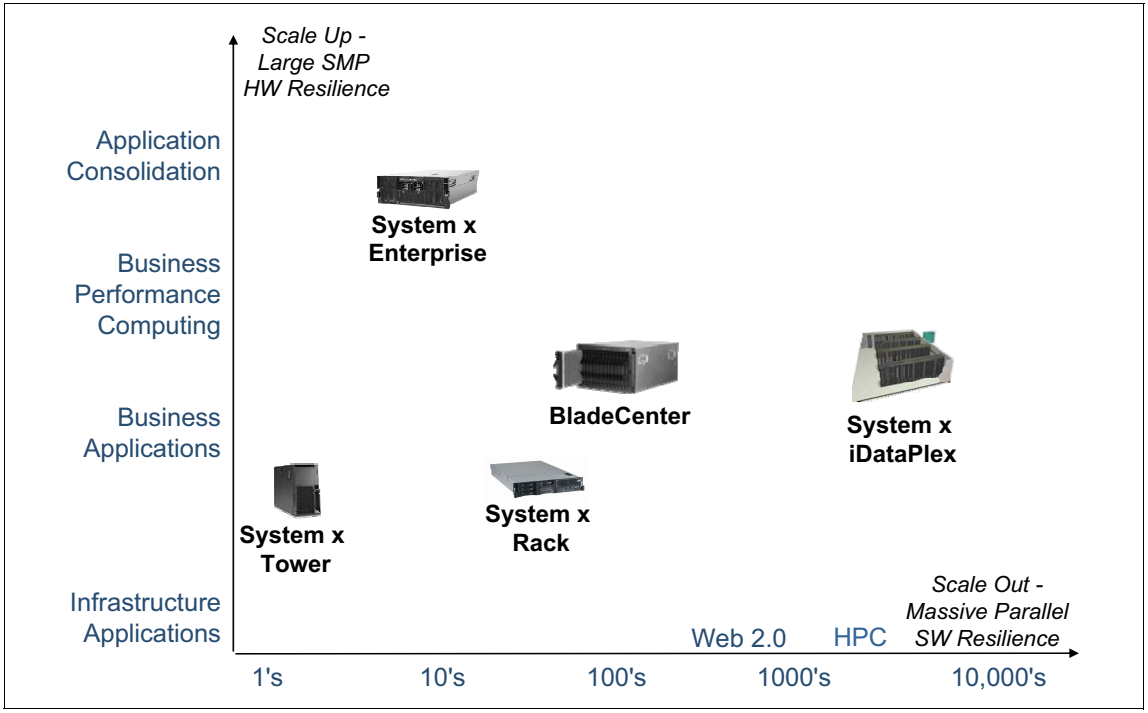


Figure 3-2 Scale-up and scale-out position

3.5 Design objectives by platform

BladeCenter, System x, and iDataPlex offer the right solution to customers depending on the business need. This section describes their different design objectives.

3.5.1 BladeCenter

IBM BladeCenter is focused on the following benefits:

- Complete infrastructure integration
- Total solution management
- Energy-efficiency
- Hardware and software reliability, availability, and serviceability
- Virtualization
- Redundancy

IBM BladeCenter can provide an infrastructure simplification solution. BladeCenter delivers the ultimate in infrastructure integration. It demonstrates leadership on power utilization, high speed I/O, and server density. It was designed to provide maximum availability with industry standard components and to reduce the number of single points of failure.

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

3.5.2 System x

System x is focused on:

- Scale-up solutions
- Strong hardware redundancy
- Virtualization

System x products provide scale-up solutions, especially the System x3850 X5, which uses the eX5 chip set. The eX5 technology is the fifth generation of products based on the design principle IBM began in 1997: to offer systems that are expandable and offer mainframe-style reliability, availability, and serviceability features.

The x3850 X5 can be configured into a multinode configuration supporting eight 10-core processors, and up to 6 TB of RAM, 3 TB in each of the two connected machines. It has active memory protection, including advanced Chipkill ECC memory correction, memory ProteXion, and memory mirroring. It has hot-swap and redundant power supplies and cooling. It is designed around three major workloads, which are:

- Database servers
- Server consolidation using virtualization
- Enterprise resource planning

3.5.3 iDataPlex

iDataPlex is focused on:

- Price/performance per watt
- Fast, large scale-out deployments
- Compute density
- Customization

The iDataPlex is the price leadership scale-out solution. It provides ultimate power efficiency. The shallow design and Rear Door Heat eXchanger of iDataPlex's deliver the best rack level cooling efficiency.

iDataPlex provides leadership compute density. Using the innovative rack iDataPlex can fit up to 84 1U quad-core servers and 16 1U vertical slots for switches, power distribution units, and other appliances in the same footprint as a standard rack.

iDataPlex provides a custom solution using Flex Node Technology. It also provides large low-cost localized storage for the demands of high performance computing. It can give you a customized data center solution.

The iDataPlex provides a high efficiency power supply and low power-consuming fans. It uses up to less 40% power than standard 1U servers.

With the Rear Door Heat eXchanger, you can have a high-density data center without increasing cooling demands. It can cool the room and reduce or even eliminate the need for air conditioning in data center.

3.6 IBM Intelligent Cluster and iDataPlex

Clusters can be configured and deployed to serve a broad range of functions, from server workload consolidation to high-performance parallel computing tasks.

The IBM Intelligent Cluster is an integrated HPC cluster solution, including servers, storage, and networking components. It supports different operating systems from Linux, Microsoft, and VMware. IBM delivers a factory-built and tested cluster that is ready to plug in.

Note: IBM Intelligent Cluster delivers a choice of nodes, storage, I/O, and cluster management systems. The nodes include IBM System x server, BladeCenter servers, Power Systems™ servers, as well as iDataPlex.

Storage includes IBM DS3000, DS4000®, and DS5000 series. The I/O includes Ethernet, InfiniBand, Fibre Channel, and Converged Fabrics.

Further information about the IBM Intelligent Cluster offering, including customer case studies, can be found on the IBM site at this website:

http://www.ibm.com/systems/x/hardware/cluster/

4

iDataPlex compute nodes

This chapter provides information about the compute nodes of the iDataPlex solution.

The following topics are covered in this chapter:

4.1 The big picture of iDataPlex

The iDataPlex solution has the following main components:

- A special iDataPlex rack
- Up to 42 2U chassis that are mounted in the rack
- Servers, storage and I/O components that are installed in the chassis
- Switches and power distribution units that are also mounted in the rack

Figure 4-1 gives you an idea of how these components are integrated.

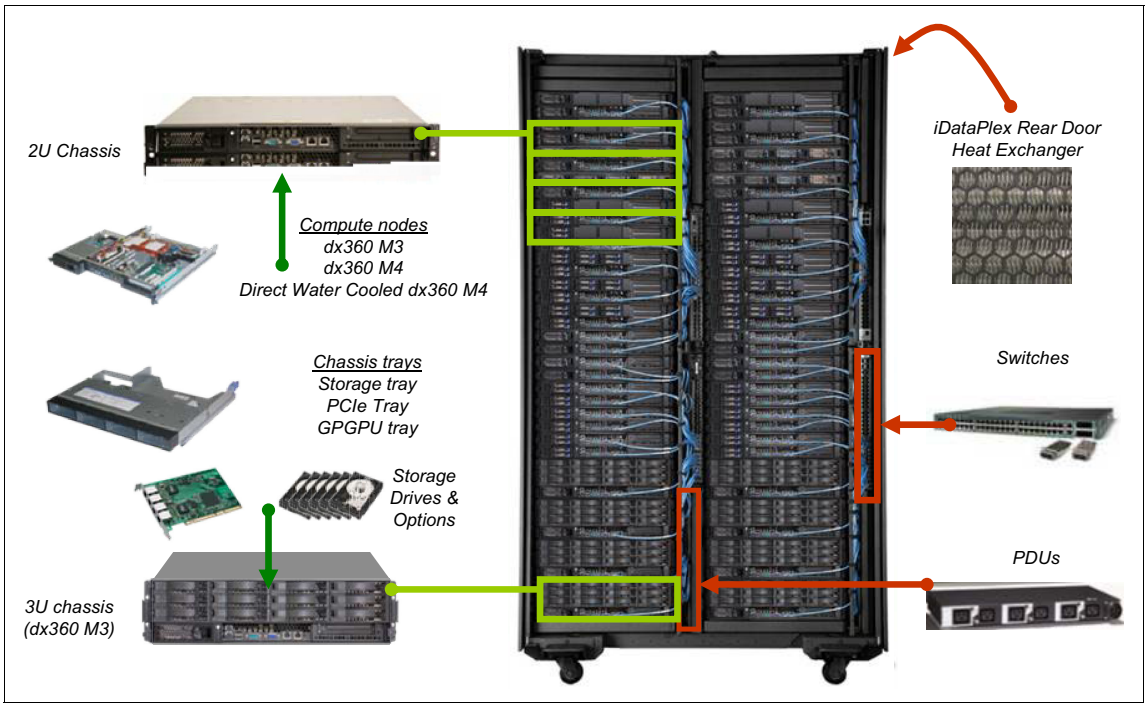


Figure 4-1 iDataPlex and its components

On the left side is a 2U (top) and a 3U chassis (bottom). They can be equipped with several different server, storage, and I/O components. The populated chassis are mounted in an iDataPlex rack (middle). This rack was specifically designed to meet high-density data center requirements. It allows mandatory infrastructure components such as switches and power distribution units (PDUs), as shown on the right, to be installed into the rack without sacrificing valuable server space. In addition, the iDataPlex solution provides water-cooling either at the rack-level with the Rear Door Heat eXchanger or at the node-level with the Direct Water Cooling solution.

4.2 Comparing the iDataPlex servers

The iDataPlex Flex chassis can be configured in different ways, depending on the customer needs. IBM currently offers two iDataPlex servers: the dx360 M3 and dx360 M4 servers. They have been developed for different types of workloads and power demands.

Table 4-1 gives an overview showing the major features of each server. The dx360 M4 server is available in an air-cooled and water-cooled version. Details of the individual servers are discussed in the following sections.

Note: The iDataPlex product publications refer to the 1U server component as a *system-board tray*.

Table 4-1 Feature comparison of iDataPlex servers

Feature	dx360 M3 server	dx360 M4 air-cooled	dx360 M4 water-cooled
Processor sockets	Two	Two	Two
Processors supported	Intel Xeon 5600	Intel Xeon E5-2600	Intel Xeon E5-2600
Cores per processor	2, 4, or 6	4, 6, or 8	4, 6, or 8
Memory speed	Up to 1333 MHz	Up to 1600 MHz	Up to 1600 MHz
DIMM sockets	16 DIMM sockets	16 DIMM sockets	16 DIMM sockets[a]
Maximum memory	192 GB DDR3	512 GB DDR3	64 GB DDR3[b]
PCIe	x16 (Gen 2)	x16 (Gen 3)	x16 (Gen 3)
GPU/GPGPU support	Yes	Yes	No
Supported in 2U chassis	Yes (1 or 2 nodes)	Yes (1 or 2 nodes)	Yes (1 or 2 nodes)
Supported in 3U chassis	Yes (1 node)	No	No
SATA support	Yes - SATA 3Gb/s	Yes - SATA 6Gb/s	Yes - SATA 6Gb/s
SAS support	Yes - 6Gb[c]	Yes - 6Gb[c]	Yes - 6Gb[c]
SSD support	Yes (2.5” only)	Yes (1.8” and 2.5”)	No
Maximum internal storage per chassis	3U chassis: 36 TB	2U chassis: 6TB	2U chassis: 6TB

Feature	dx360 M3 server	dx360 M4 air-cooled	dx360 M4 water-cooled
3.5-inch drive support	Yes	Yes	Yes
2.5-inch drive support	Yes	Yes	Yes

a. The system-board provides 16 DIMM sockets but only eight of them can be populated
b. Increased memory capacity is planned to be supported in the future
c. Requires an optional SAS controller card

Did you know? The SATA International Organization has released naming guidelines to avoid confusion when referring to SATA standards. “SATA II” or “SATA III” should not be used. Instead, use “SATA 3 Gb/s” or “SATA 6 Gb/s”.

The naming guidelines are available from this website:

http://www.serialata.org/developers/naming_guidelines.asp

Additional performance data for each system is available at the IBM System x benchmarks page:

http://www.ibm.com/systems/x/resources/benchmarks

4.3 iDataPlex dx360 M4 server

The iDataPlex dx360 M4 server, available as machine type 7912 with a three-year warranty, is a 1U tray with a system board that fits into the 2U chassis. It can be equipped with a GPGPU tray for high-density GPGPU requirements. The dx360 M4 is a two-socket server running the latest Intel Xeon processor E5-2600 family (formerly code named *Sandy Bridge-EP*) processors. See Figure 4-2.

The following topics are covered in this section:

Figure 4-2 Two air-cooled iDataPlex dx360 M4 servers installed in a 2U chassis

Figure 4-3 shows the controls and connectors on the front of the server. Each server has the following components (looking left to right):

- Power button and status LEDs (see 4.3.9, “Operator panel” on page 59)
- One RS232 serial port
- One VGA port supporting resolutions up to 1600x1200 at 75 MHz
- One 1 Gbps RJ45 connector for dedicated systems management (wired to the IMM)
- A slot for the mezzanine card ports
- Two 1 Gbps Ethernet interfaces based on the Intel Powerville I350 controller
- Four USB 2.0 ports.

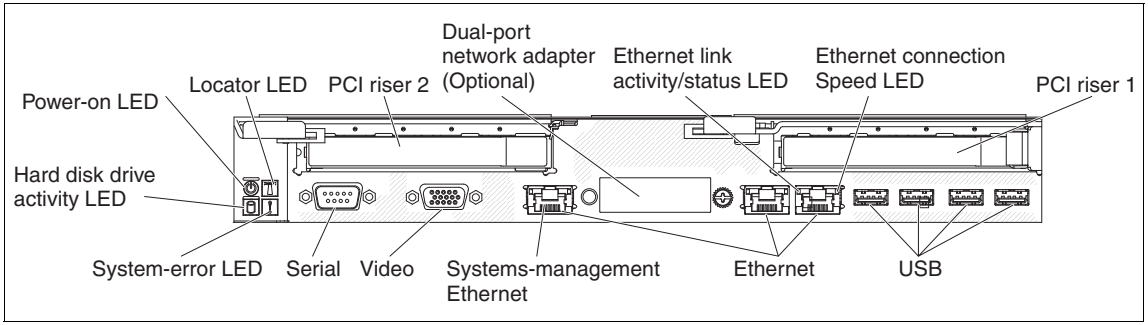


Figure 4-3 dx360 M4 controls and connectors

Figure 4-4 shows the internal connectors and slots of the dx360 M4 server.

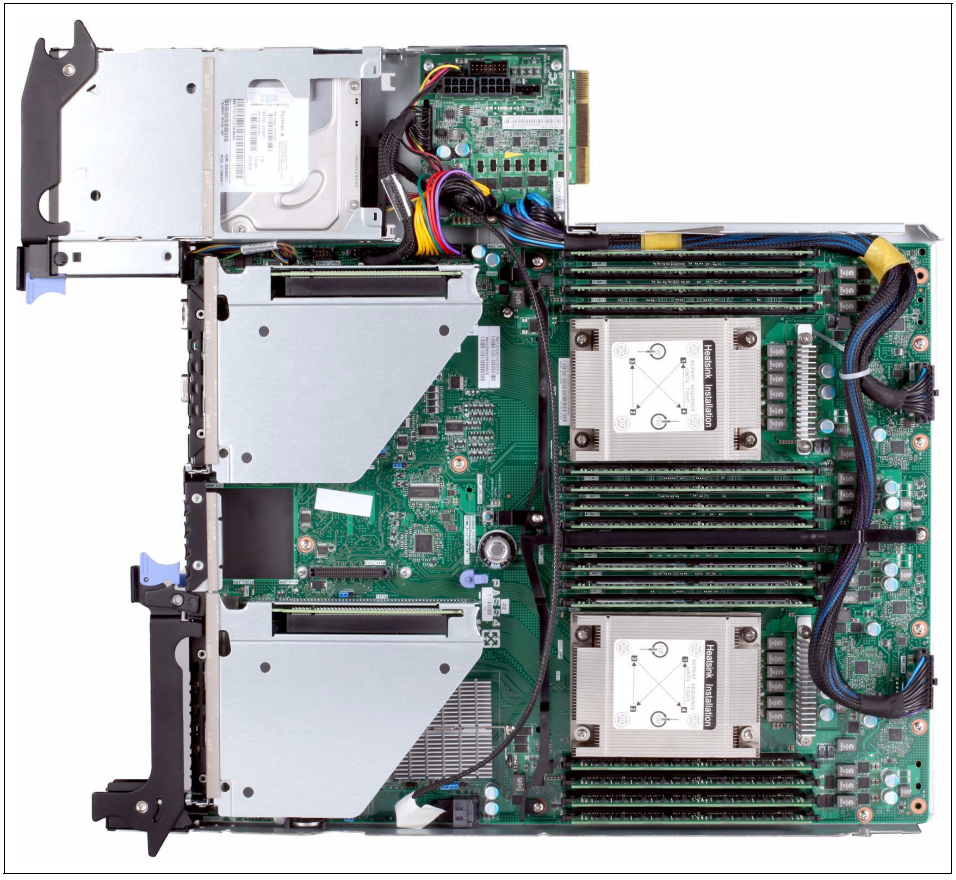


Figure 4-4 dx360 M4 system board

Product publications are available at the IBM support site:

http://www.ibm.com/support

4.3.1 Models

Table 4-2 lists the standard models of the dx360 M4.

Table 4-2 Standard models of the dx360 M4

Model	Processor	Memory & speed	Disk adapter	Disk bays	Disks	Network
7912-22x	2x Intel Xeon E5-2630L 6C 2.0GHz 15MB 1333MHz 60W	4x 8 GB 1333 MHz	6 Gbps SATA	1x 3.5” SS bay	Open	2x GbE
7912-42x	2x Intel Xeon E5-2609 4C 2.40GHz 10MB 1066MHz 80W	4x 4 GB 1333 MHz	6 Gbps SATA	2x 2.5” or 4x 1.8” SS bays	Open	2x GbE + 2x QDR[a]
7912-62x	2x Intel Xeon E5-2660 8C 2.2GHz 20MB 1600MHz 95W	4x 8 GB 1333 MHz	6 Gbps SATA	2x 2.5” or 4x 1.8” SS bays	Open	2x GbE + 2x QDR[a]

a. Models 42x and 62x include a Mellanox ConnectX-3 Dual Port QDR/FDR10 Mezzanine Card, 90Y6338, which installs in the dedicated mezzanine slot.

Two of these models can be installed in a 2U chassis. For configure-to-order (CTO) models, additional PCIe trays can be added for additional PCIe slots, including GPGPU adapters. See 4.3.5, “Chassis trays” on page 52 for details.

4.3.2 Processors

The dx360 M4 server has two LGA 2011 CPU sockets for the latest Intel Xeon E5-2600 series processors. Table 4-3 lists the features for the different Intel Xeon E5-2600 series processors supported on the dx360 M4 server. The processors installed in the two sockets must be identical.

All processors support Hyper-Threading nor Turbo Boost except for the Basic processors as listed in the table, E5-2603 and E5-2609.

Table 4-3 Supported Intel Xeon E5-2600 series CPU options on the dx360 M4 server

SBB part number	Feature code	Processor Model	Core Speed	L3 Cache	Cores	TDP	QPI link speed	Max memory speed
Special Purpose								
90Y6212	A1TF	E5-2637	3.0 GHz	5 MB	2	80 W	8 GT/s	1600 MHz
90Y6221	A1TQ	E5-2667	2.9 GHz	15 MB	6	130 W	8 GT/s	1600 MHz
Low Power								
90Y6223	A1TS	E5-2630L	2.0 GHz	15 MB	6	60 W	7.2 GT/s	1333 MHz

SBB part number	Feature code	Processor Model	Core Speed	L3 Cache	Cores	TDP	QPI link speed	Max memory speed
94Y7550	A2FV	E5-2648L	1.8 GHz	20 MB	8	70W	8.0 GT/s	1600 MHz
90Y6224	A1TT	E5-2650L	1.8 GHz	20 MB	8	70 W	8.0 GT/s	1600 MHz
Basic (no support for Hyper-Threading nor Turbo Boost								
90Y6211	A1TE	E5-2603	1.8 GHz	10 MB	4	80 W	6.4 GT/s	1066 MHz
90Y6213	A1TG	E5-2609	2.4 GHz	10 MB	4	80 W	6.4 GT/s	1066 MHz
Standard								
90Y6214	A1TH	E5-2620	2.0 GHz	15 MB	6	95 W	7.2 GT/s	1333 MHz
90Y6215	A1TJ	E5-2630	2.3 GHz	15 MB	6	95 W	7.2 GT/s	1333 MHz
90Y6216	A1TK	E5-2640	2.5 GHz	15 MB	6	95 W	7.2 GT/s	1333 MHz
Advanced								
90Y6217	A1TL	E5-2650	2.0 GHz	20 MB	8	95 W	8.0 GT/s	1600 MHz
90Y6222	A1TR	E5-2658	2.1 GHz	20 MB	8	95 W	8.0 GT/s	1600 MHz
90Y6218	A1TM	E5-2660	2.2 GHz	20 MB	8	95 W	8.0 GT/s	1600 MHz
90Y7551	A2FW	E5-2665	2.4 GHz	20 MB	8	115 W	8.0 GT/s	1600 MHz
90Y6225	A1TU	E5-2670	2.6 GHz	20 MB	8	115 W	8.0 GT/s	1600 MHz
90Y6219	A1TN	E5-2680	2.7 GHz	20 MB	8	130 W	8.0 GT/s	1600 MHz

The Intel Xeon E5-2600 series processors implement the second generation of Intel's Core microarchitecture using a 32nm manufacturing process (Figure 4-5).

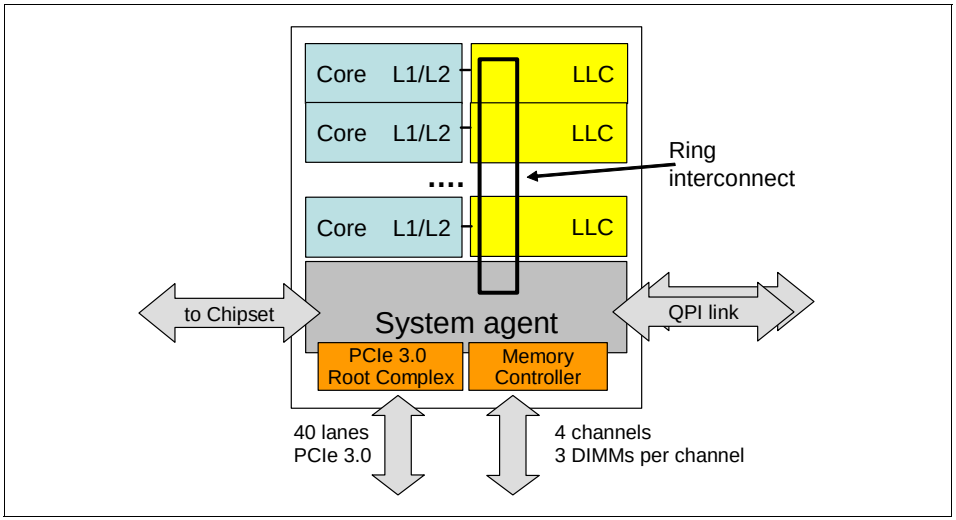


Figure 4-5 Intel Xeon E5-2600 basic architecture

The new architecture allows you to share data on-chip through a high-speed ring interconnect 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 three DDR3 DIMMs per channel running at a speed up to 1600 MHz. Two QPI links connect to a second CPU in a dual-socket installation.

The Xeon E5-2600 series is available with up to eight cores and 20 MB of last-level cache. It features an enhanced instruction set called Intel Advanced Vector Extensions (AVX). It doubles the operand size for vector instructions (such as floating-point) to 256 bits and boosts selected applications by up to a factor of two.

The new architecture also introduces Intel Turbo Boost Technology 2.0 and improved power management capabilities. Turbo Boost automatically turns off unused processor cores and increases the clock speed of the cores in use as long as thermal requirements are still met. Turbo Boost Technology 2.0 takes advantage of the new integrated design and also implements a more granular overclocking in 100 MHz steps instead of 133 MHz steps on former Nehalem-based and Westmere-based microprocessors.

4.3.3 Memory

The dx360 M4 server has 16 DIMM slots for registered ECC DDR3 DIMMs running at 1066, 1333, or 1600 MHz. RDIMMs, UDIMMs, and LRDIMMs are supported. It supports capacities up to 32 GB, as listed in Table 4-4.

Table 4-4 Supported memory options for the dx360 M4 server

Part number	Feature code	Description
UDIMMs		
49Y1403	A0QS	2GB (1x2GB, 1Rx8, 1.35V) PC3L-10600 ECC DDR3 1333MHz LP UDIMM
49Y1404	8648	4GB (1x4GB, 2Rx8, 1.35V) PC3L-10600 CL9 ECC DDR3 1333MHz LP UDIMM
RDIMMs		
49Y1405	8940	2GB (1x2GB, 1Rx8, 1.35V) PC3L-10600 CL9 ECC DDR3 1333MHz LP RDIMM
49Y1406	8941	4GB (1x4GB, 1Rx4, 1.35V) PC3L-10600 CL9 ECC DDR3 1333MHz LP RDIMM
90Y3178	A24L	4GB (1x4GB, 2Rx8, 1.5V) PC3-12800 CL11 ECC DDR3 1600MHz LP RDIMM
49Y1397	8923	8GB (1x8GB, 2Rx4, 1.35V) PC3L-10600 CL9 ECC DDR3 1333MHz LP RDIMM
90Y3109	A292	8GB (1x8GB, 2Rx4, 1.5V) PC3-12800 CL11 ECC DDR3 1600MHz LP RDIMM
49Y1563	A1QT	16GB (1x16GB, 2Rx4, 1.35V) PC3L-10600 CL9 ECC DDR3 1333MHz LP RDIMM
00D4968	A2U5	16GB (1x16GB, 2Rx4, 1.5V) PC3-12800 CL11 ECC DDR3 1600MHz LP RDIMM
LRDIMMs		
90Y3105	A291	32GB (1x32GB, 4Rx4, 1.35V) PC3L-10600 CL9 ECC DDR3 1333MHz LP LRDIMM

The Intel Xeon E5-2600 microprocessor architecture with the Quick-Path Interconnect between processors has a four-channel memory interface per processor with two DIMM slots per channel, as shown in Figure 4-6.

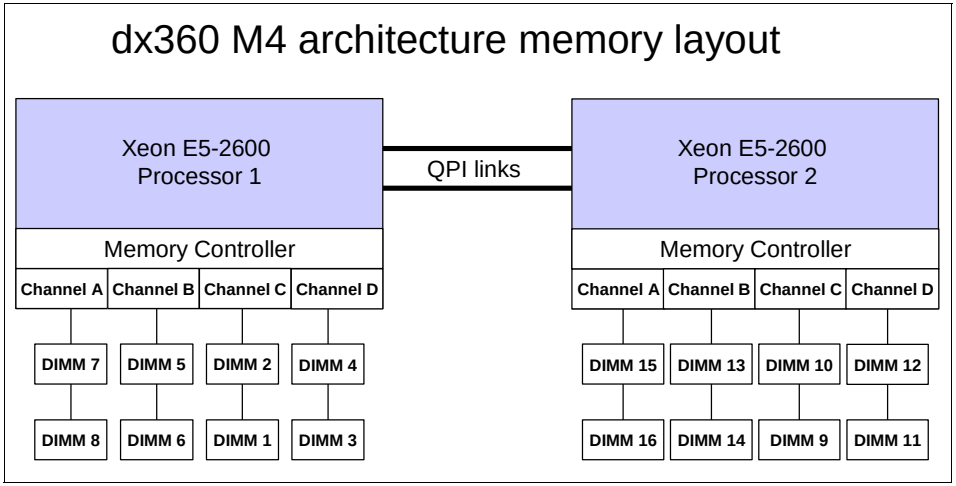


Figure 4-6 Memory channel layout of the iDataPlex dx360 M4 server

The UEFI on the dx360 M4 server supports three modes of operation:

- Normal channel mode (also known as independent channel mode)
- Mirrored channel mode
- Rank sparing mode

Normal mode is the default setting. The operating system uses the full amount of memory installed. No redundancy is provided in this mode.

The population sequence for a single-processor system running in normal mode is shown in Table 4-5. See the system board layout in Figure 4-4 on page 39 for the physical location of the DIMM slots on the board.

Table 4-5 Memory population table for dx360 M4 server with one CPU installed

	Processor 1							
Number of DIMMs	**Channel A**		**Channel B**		**Channel C**		**Channel D**	
	DIMM 8	**DIMM 7**	**DIMM 6**	**DIMM 5**	**DIMM 1**	**DIMM 2**	**DIMM 3**	**DIMM 4**
1					x			
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

Table 4-6 lists the population sequence for a dual-CPU system running in normal mode.

Table 4-6 Memory population table for dx360 M4 server with two CPUs installed

	Processor 1								Processor 2							
Number of DIMMs	**Channel A**		**Channel B**		**Channel C**		**Channel D**		**Channel A**		**Channel B**		**Channel C**		**Channel D**	
	DIMM 8	**DIMM 7**	**DIMM 6**	**DIMM 5**	**DIMM 1**	**DIMM 2**	**DIMM 3**	**DIMM 4**	**DIMM 16**	**DIMM 15**	**DIMM 14**	**DIMM 13**	**DIMM 9**	**DIMM 10**	**DIMM 11**	**DIMM 12**
1					x											
2					x								x			
3	x				x								x			
4	x				x				x				x			

	Processor 1								Processor 2							
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

The memory speed supported in a specific configuration depends on several different parameters. Table 4-7 provides an overview and lists the memory speed.

Table 4-7 dx360 M4 memory channel speed support

Part number	Feature code	DIMM capacity	Voltage rating	Maximum memory channel speed and voltage support by DIMM per channel			
				1 DIMM per channel		2 DIMMs per channel	
				1.35 V	1.5 V	1.35 V	1.5 V
UDIMMs							
49Y1403	A0QS	2 GB	1.35V	1333 MHz	1333 MHz	1333 MHz	1333 MHz
49Y1404	8648	4 GB	1.35 V	1333 MHz	1333 MHz	1333 MHz	1333 MHz
RDIMMs							
49Y1405	8940	2 GB	1.35 V	1333 MHz	1333 MHz	1333 MHz	1333 MHz
49Y1406	8941	4 GB	1.35 V	1333 MHz	1333 MHz	1333 MHz	1333 MHz
90Y3178	A24L	4 GB	1.5 V	Not supported	1600 MHz	Not supported	1600 MHz
49Y1397	8923	8 GB	1.35 V	1333 MHz	1333 MHz	1333 MHz	1333 MHz

Part number	Feature code	DIMM capacity	Voltage rating	Maximum memory channel speed and voltage support by DIMM per channel			
				1 DIMM per channel		2 DIMMs per channel	
				1.35 V	1.5 V	1.35 V	1.5 V
90Y3109	A292	8 GB	1.5 V	Not supported	1600 MHz	Not supported	1600 MHz
49Y1563	A1QT	16 GB	1.35 V	1333 MHz	1333 MHz	1333 MHz	1333 MHz
00D4968	A2U5	16 GB	1.5 V	Not supported	1600 MHz	Not supported	1600 MHz
LRDIMMs							
90Y3105	A291	32 GB	1.35V	1333 MHz	1333 MHz	1333 MHz	1333 MHz

Use the following general guidelines when deciding about the memory configuration of your server:

- Install memory DIMMs in order of their size, with the largest DIMM first.
- If you mix DIMMs with different memory speeds, the slowest DIMM determines the speed of the memory bus.
- You cannot run different memory speeds on different channels.
- If you mix DIMMs with 1.35V and 1.5V, the system runs all of them at 1.5V and you lose the energy advantage.
- Populate equivalent ranks per channel.

Recommendation for two-socket systems: If your performance objective is:

- To maximize memory bandwidth:

  Install all 16 sockets with the 16 GB 1600 MHz DIMM option (feature code A2U5). It gives you 256 GB memory running at 1600 MHz.
- To maximize memory capacity:

  Install all 16 sockets with the 32 GB LRDIMM option (feature A291). It gives you a total of 512 GB memory running at 1333 MHz.

When using *mirrored mode*, Channel A and Channel B both hold the same data for redundancy purposes. If one DIMM fails, it is disabled and the backup DIMM in the other channel takes over. Likewise, Channel C and Channel D hold the same data and create one mirrored pair. Because memory mirroring is handled in hardware, it is operating system-independent. The total usable memory size is the sum of all DIMMs in Channel A or B plus the sum of all DIMMs in Channel C or D.

Note: Memory must be identical across the two CPUs to enable the memory mirroring feature.

When using *rank sparing mode*, one memory DIMM rank serves as a spare of the other ranks on the same channel. The spare rank is held in reserve and is not used as active memory. The spare rank must have identical or larger memory capacity than all the other active memory ranks on the same channel. After an error threshold is surpassed, the contents of that rank is copied to the spare rank. The failed rank of memory is taken offline, and the spare rank is put online and used as active memory in place of the failed rank.

4.3.4 Internal storage and controllers

This section provides information about the internal storage options. It lists the conventional and solid-state disks that are supported. The next section is about the RAID controller cards. The last section provides technical details about the internal storage connections.

Hard disk drives

Table 4-8 lists the conventional disk drives that are supported with the dx360 M4 server in any of the 2U chassis configurations.

Table 4-8 dx360 M4 disk drives supported

Part Number	Feature	Interface type	Disk size	Simple Swap / Hot Swap	Interface speed	Form factor	RPM
SATA NL drives							
81Y9734	A1NY	SATA II NL	250 GB	Simple Swap	6 Gbps	2.5-inch	7200
81Y9802	A22U	SATA II NL	500 GB	Simple Swap	6 Gbps	3.5-inch	7200
81Y9738	A1P0	SATA II NL	500 GB	Simple Swap	6 Gbps	2.5-inch	7200
81Y9806	A22X	SATA II NL	1 TB	Simple Swap	6 Gbps	3.5-inch	7200
81Y9742	A1P2	SATA II NL	1 TB	Simple Swap	6 Gbps	2.5-inch	7200
81Y9810	A22W	SATA II NL	2 TB	Simple Swap	6 Gbps	3.5-inch	7200
81Y9814	A22V	SATA II NL	3 TB	Simple Swap	6 Gbps	3.5-inch	7200

Part Number	Feature	Interface type	Disk size	Simple Swap / Hot Swap	Interface speed	Form factor	RPM
SAS drives							
49Y1996	5426	SAS	146 GB	Simple Swap	6 Gbps	2.5-inch	15 K
49Y1991	5427	SAS	300 GB	Simple Swap	6 Gbps	2.5-inch	10 K
81Y9674	A24J	SAS	300 GB	Simple Swap	6 Gbps	2.5-inch	15 K
49Y2027	5435	SAS	600 GB	Simple Swap	6 Gbps	2.5-inch	10 K
81Y9654	A24H	SAS	900 GB	Simple Swap	6 Gbps	2.5-inch	10 K

1.8-inch and 2.5-inch solid state drives

Solid state drives (SSDs) provide extremely fast and reliable access to data when compared to traditional drives with rotating disks. They use flash memory to store each bit of information electronically instead of relying on the magnetic orientation on the disk.

SSDs have the advantage of no moving parts, which makes them more robust then conventional drives. Because of flash storage, no delay is imposed by any mechanical parts when accessing data. SSDs achieve very high I/O operations per second (IOPS) rates. This is especially true when accessing random data, where SSDs support a constant high throughput rate that is much higher than that of conventional drives.

To allow for an easy migration, most SSDs are available in the conventional 2.5-inch drive format. Newer drives are 1.8-inch which enables a higher density because multiple SSDs can be packed vertically in the space of one conventional horizontal 3.5-inch drive. Table 4-9 lists SSDs that are supported in iDataPlex dx360 M4 servers.

Table 4-9 Solid state drives (SSDs) supported in iDataPlex dx360 M4

Part number	Feature code	Grade	Disk size	Interface type	Carrier	Transfer speed	Form factor
43W7742	5419	Enterprise	200 GB	SATA II	Simple swap	3 Gbps	2.5-inch
43W7726	5428	Enterprise	50 GB	SATA II	Simple swap	3 Gbps	1.8-inch
43W7746	5420	Enterprise	200 GB	SATA II	Simple swap	3 Gbps	1.8-inch
90Y8668	A2UB	Enterprise Value	128 GB	SATA II	Simple swap	3 Gbps	2.5-inch
90Y8663	A2UC	Enterprise Value	256 GB	SATA II	Simple swap	3 Gbps	2.5-inch

Note: If you have multiple SSDs in your server, consider using the IBM 6 Gb Performance Optimized Host Bus Adapter. It is specifically built for high-throughput SSD environments, with the LSI Fusion-MPT architecture driving more than 290,000 I/Os per second.

Onboard disk controllers and RAID support

The onboard controller of the dx360 M4 is only for SATA drives and does not support RAID, as shown in Table 4-10. If you plan to use SAS disks, or you want to use hardware-based RAID, you must install a separate PCIe disk controller, as described in the next section.

Table 4-10 Onboard disk controller features of the dx360 M4

Controller feature	dx360 M4
SATA II support	Yes, 2-port
SATA port speed	6 Gbps
SAS support	No
Hardware RAID	No

PCI Express disk controllers and RAID support

You must install a separate disk controller in a PCI Express (PCIe) slot for any of the following reasons:

- SAS drives are used (because the onboard controller supports SATA only).
- More than two drives are used (because onboard controller has two SATA ports).
- Hardware RAID is required.
- Multiple SSD drives are used and I/O bandwidth is an important consideration.

Table 4-11 lists the supported RAID cards.

Table 4-11 Supported SAS/SATA RAID controller cards on the dx360 M4

Name	Chipset	Cache	Battery backup	Supported RAID levels	Slot
6 Gb Performance Optimized HBA	LSI SAS2008	None	None	None	PCIe 2.0 x8
6 Gb SAS HBA	LSI SAS2008	None	None	None	PCIe 2.0 x8
ServeRAID H1110	LSI SAS2004	None	None	0, 1, 10	PCIe 2.0 x4
ServeRAID M1115	LSI SAS2008	None	None	0, 1, 10; 5 and 50[a]	PCIe 2.0 x8
ServeRAID M5120[b]	LSI SAS2208	Optional with RAID feature	Optional[c]	0, 1, 10; 5 and 50[d]; 6 and 60[e]	PCIe 3.0 x8

a. RAID levels 5 and 50 are supported with additional M1100 RAID 5 Upgrade Key, FC A1X1.
b. This adapter only has external connectors. No internal storage can be connected to this adapter.
c. Battery backup optional with feature code A22E.
d. RAID levels 5 and 50 are supported with additional M5100 Series RAID 5 Upgrade Key, FC A1X2 (zero Cache included), A1J3 (512 MB Cache included), A1J4 (512 MB Flash included), or A1WY (1GB Flash included). Flash technology allows you to hold the data without a dedicated backup battery unit.
e. RAID levels 6 and 60 are supported with additional M5100 Series RAID 6 Upgrade Key, FC A1X3.

For more information about these ServeRAID cards, including photos, see the *ServeRAID Adapter Quick Reference*, TIPS0054 available from this website:

`http://www.redbooks.ibm.com/abstracts/tips0054.html`

For drivers and firmware for iDataPlex RAID controllers, go to this website:

`http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5076022`

SAS connectivity

The next two figures show how each drive is cabled to the RAID controller cards. Figure 4-7 shows a 2U chassis with eight SAS small-form-factor drives, labeled SFF DRV 0 to SFF DRV 7.

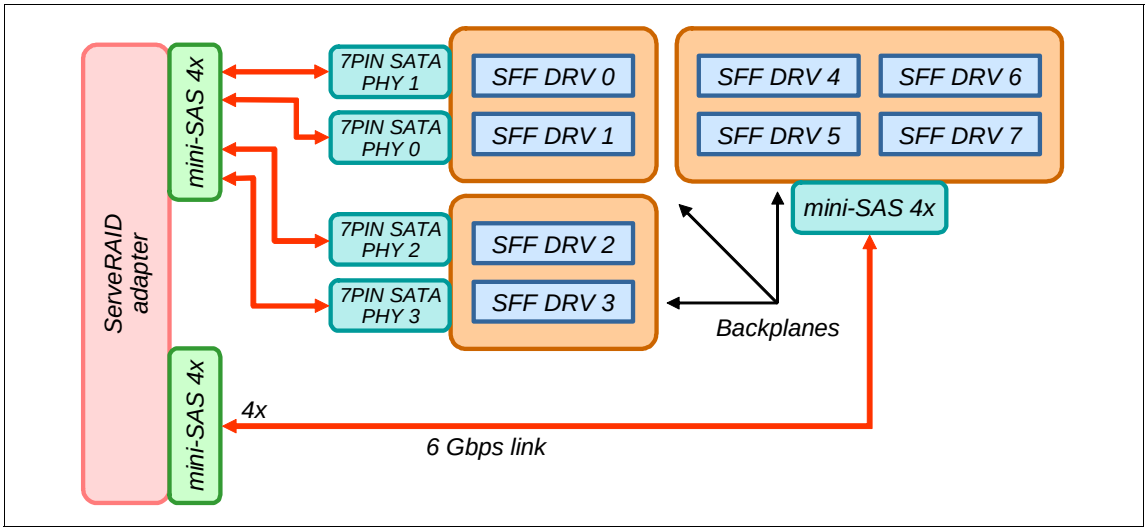


Figure 4-7 SAS wiring for 2U chassis configurations

The first four drives (DRV 0 to DRV 3) are each directly connected with a 1x SAS cable to the ServeRAID adapter. The second half of the drives (DRV 4 to DRV 7) merges in a single 4x SAS cable using the mini-SAS connector.

4.3.5 Chassis trays

For highest compute performance, two servers are installed in one single 2U chassis. Each server can have two PCIe 3.0 x16 full-height/half-length adapters (a total of four in the 2U chassis).

In some configurations, only one iDataPlex server is installed in a 2U chassis. These configurations support trays for additional PCIe slots and storage. The mezzanine PCIe slot can be populated in all configurations.

Trays supported

The following sections explain the different trays that are supported (Table 4-12).

Table 4-12 Chassis tray options

Feature code	Description
A262	iDataPlex 2U Dual Planar Configuration (two per 2U chassis)
A2N3	iDataPlex 2U GPU Configuration (one per 2U chassis)

GPGPU tray

The GPGPU tray (feature code A2N3) has two bays on the left for either two 3.5-inch disk drives, four 2.5-inch disk drives, or eight 1.8-inch solid-state disks. Using two 2U riser cards, the tray can hold two double-wide PCIe graphics cards in the upper half of the chassis and two PCIe cards in the lower half of the chassis. All slots support full height half length PCIe cards.

Figure 4-8 shows a drawing of a 2U chassis with the GPGPU tray.

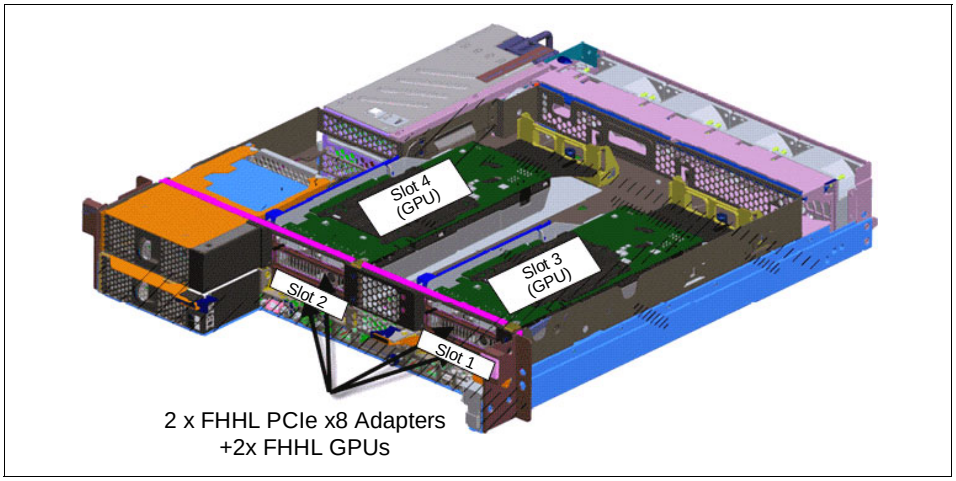


Figure 4-8 iDataPlex dx360 M4 GPGPU configuration

When using this tray, the PCIe topology is as illustrated in Figure 4-9.

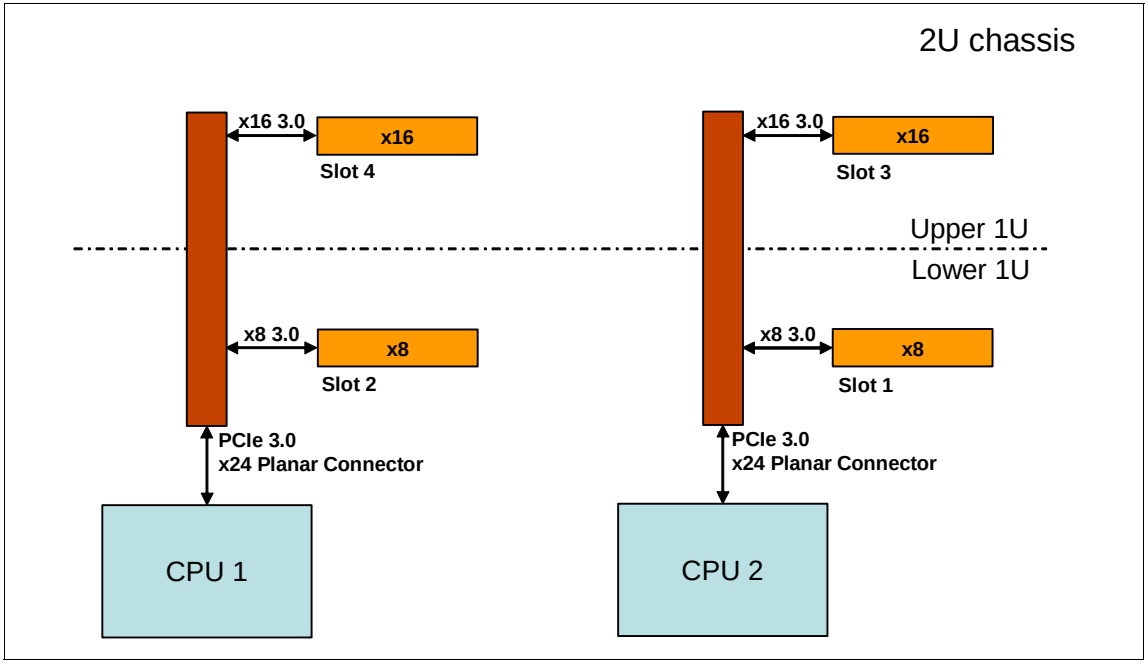


Figure 4-9 PCIe topology of the GPGPU tray

4.3.6 PCI Express I/O adapters

The dx360 M4 server supports one onboard PCIe card through a mezzanine card. Additional PCIe cards can be mounted through the use of three different riser cards. Dependent on which tray is installed in the chassis, a different number of PCIe slots is available. Detailed PCIe features are listed in table Table 4-13.

The following different riser cards are available for the dx360 M4 server:

- 1U single-slot riser card (feature A1UE, for dual-node chassis configurations only). It provides one PCIe 3.0 x16 slot. Slots are full-height/half length Two riser cards can be installed in a server for a total of two PCIe slots.
- 2U two-slot riser card (for usage with GPGPU tray only). It provides one PCIe 3.0 x16 slot (for GPGPU card) and one PCIe 3.0 x8 slot (x16 mechanically). Two riser cards can be connected to a single compute node and installed in a 2U chassis.

All PCIe slots support full-height half-length adapters, with the exception of slots 3 and 4 of the GPGPU chassis tray (which support full-length adapters).

Table 4-13 PCIe features of the iDataPlex dx360 M4 server

Feature	dual-node chassis config (per node)	GPGPU chassis tray config
PCIe slots maximum (excluding the always-available mezzanine card slot)	2x PCIe 3.0 x16	2x PCIe 3.0 x16 2x PCIe 3.0 x8
Slot 1, mechanically	x16	x16
Slot 1, electrically	x16 3.0	x8 3.0
Slot 2, mechanically	x16	x16
Slot 2, electrically	x16 3.0	x8 3.0
Slot 3, mechanically	Not applicable	x16 full-length
Slot 3, electrically	Not applicable	x16 3.0
Slot 4, mechanically	Not applicable	x16 full-length
Slot 4, electrically	Not applicable	x16 3.0
Mezzanine PCIe, mechanically	x8	x8
Mezzanine PCIe, electrically	x8 3.0	x8 3.0

I/O option cards that are supported are listed in Table 4-14.

Table 4-14 Supported PCIe option cards on the iDataPlex dx360 M4 server

Part number	Feature code	Description	PCIe Interface
Fibre Channel			
81Y1675	A2XV	Brocade 16Gb FC Dual-port HBA for IBM System x	2.0 x8
81Y1668	A2XU	Brocade 16Gb FC Single-port HBA for IBM System x	2.0 x8
81Y1662	A2W6	Emulex 16Gb FC Dual-port HBA for IBM System x	2.0 x8
81Y1655	A2W5	Emulex 16Gb FC Single-port HBA for IBM System x	2.0 x8
42D0501	3578	QLogic 8 Gb single-port HBA for IBM System x	2.0 x4
42D0510	3579	QLogic 8 Gb dual-port HBA for IBM System x	2.0 x4
42D0485	3580	Emulex 8 Gb single-port HBA for IBM System x	2.0 x4
42D0494	3581	Emulex 8 Gb dual-port HBA for IBM System x	2.0 x4
46M6049	3589	Brocade 8 Gb single-port HBA for IBM System x	2.0 x8

Part number	Feature code	Description	PCIe Interface
46M6050	3591	Brocade 8 Gb dual-port HBA for IBM System x	2.0 x8
SAS/SATA RAID controller			
81Y4492	A1XL	IBM ServeRAID H1110	2.0 x4
81Y4448	A1MZ	IBM ServeRAID M1115	2.0 x8
81Y4478	A1WX	IBM ServeRAID M5120	3.0 x8
46M0907	5982	IBM 6Gb SAS HBA	2.0 x8
46M0912	3876	IBM 6Gb Performance Optimized HBA	2.0 x8
1 Gb Ethernet			
42C1780	2995	NetXtreme II 1000 Express 2-port, 1 Gbps	1.1 x4
49Y4220	5766	NetXtreme II 1000 Express 4-port, 1 Gbps	2.0 x4
49Y4230	5767	Intel Ethernet 2-port I340-T2, 1 Gbps	2.0 x4
49Y4240	5768	Intel Ethernet 4-port I340-T4, 1 Gbps	2.0 x4
10 Gb Ethernet / Converged Fabric			
81Y9990	A1M4	Mellanox ConnectX-2 2-port adapter, 10 Gbps	2.0 x8
49Y7950	A18Z	Emulex Virtual Fabric Adapter II, 10 Gbps	2.0 x8
42C1820	1637	Brocade 2-port CNA, 10 Gbps	2.0 x8
42C1800	5751	QLogic 2-port CNA, 10 Gbps	2.0 x8

The iDataPlex dx360 M4 server supports an additional mezzanine card with a dedicated PCIe x8 slot at the front of the chassis. Figure 4-10 shows this slot on the system board.

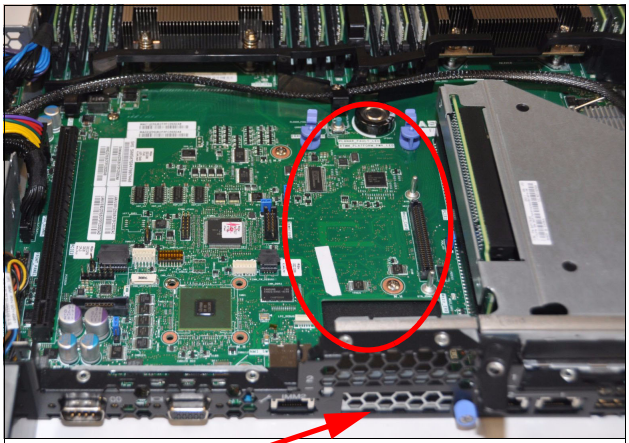


Figure 4-10 dx360 M4 server mezzanine option card slot

Supported adapters for this slot are listed in Table 4-15.

Table 4-15 Supported PCIe mezzanine option cards on the dx360 M4 server

Part number	Feature Code	Description	PCIe Interface
10 Gb Ethernet / Converged Fabric			
90Y6456	A22J	Emulex 2-port 10GbE SFP+	x8
InfiniBand / Ethernet			
90Y6338	A24F	Mellanox ConnectX-3 2-port QDR/FDR-10	x8
00D4143	A36R	IBM Dual Port FDR Embedded Adapter	x8

4.3.7 Graphics Adapter and General Purpose GPU (GPGPU)

The dx360 M4 server can also host graphics adapters specifically built for general-purpose compute acceleration by the GPU. Table 4-16 lists the supported adapters.

When installing more than one PCIe graphics card, you need a 3-slot PCIe riser card that provides two PCIe 2.0 slots, feature 4792. The riser card also requires a dedicated power cable.

Table 4-16 Supported graphics adapter and GPGPU options on the dx360 M4

Part number	Feature code	Description	Interface
00D4192	A36S	NVIDIA Tesla K10	PCIe 2.0 x16
94Y5960	A1R4	NVIDIA Tesla M2090, 6 GB GDDR5	PCIe 2.0 x16
90Y6492	A105	NVIDIA Tesla M2070Q	PCIe 2.0 x16
94Y5958	A1R3	NVIDIA Quadro 5000, 2.5 GB GDDR5	PCIe 2.0 x16
Not available	3848	Power cable, 6pin-8pin	Not applicable
Not available	7471	Power cable, 8pin-8pin	Not applicable
Not available	4792	3-slot PCIe riser card	Not applicable
Not available	7469	Riser card, power cable	Not applicable

All cards require an additional auxiliary power cable that allows the cards to draw more power directly from the 2U power supply. One cable is required for each card.

Up to two adapters can be installed in a 2U chassis with the PCIe tray using the 3-slot riser card. The riser card also requires an additional power cable. The installed cards do not have to be of the same type.

4.3.8 Management

Each dx360 M4 server has an Integrated Management Module 2 (IMM2) onboard and uses the Unified Extensible Firmware Interface (UEFI) to replace the older BIOS interface. The IMM2 is IPMI v2.0-compliant and provides the following major features:

- Remote configuration of IMM2 and UEFI settings without the need to power-on the server
- Remote access to system fan, voltage, and temperature values
- Remote IMM and UEFI update
- UEFI update when the server is powered off
- Remote console by way of a serial over LAN
- Remote presence (KVM and CD/DVD) and blue screen capture (requires additional feature on demand key, part number 90Y3901)
- Remote access to the system event log

- Predictive failure analysis and integrated alerting features (for example, by SNMP)

For more information and details about the IMM, see 7.1, “Integrated Management Module” on page 224.

To activate full systems management capabilities, an IBM Remote Presence Upgrade (Feature on Demand key) is required. This feature is a license that you apply to the system to enable.

Table 4-17 IBM Remote Presence Upgrade (feature on demand key)

Part number	Description
90Y3901	IMM2 IBM Remote Presence Upgrade (feature on demand key)

4.3.9 Operator panel

The iDataPlex dx360 M4 server has an operator panel at the front. Figure 4-11 shows the power button with the integrated power LED and three status LEDs.

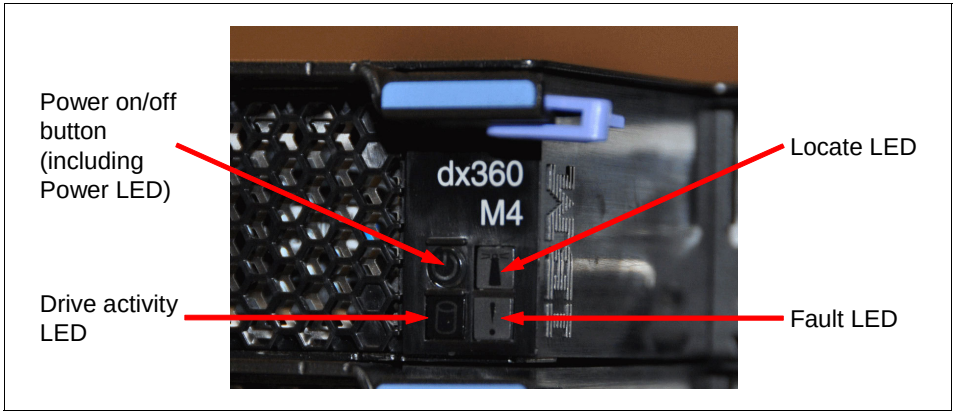


Figure 4-11 iDataPlex dx360 M4 operator panel

The LEDs are defined as follows:

- Power on (green): Determines the power state of the serve:
  - Off: AC power is not present.
  - On: Server is powered-on.
  - Fading on and off: Server is in a reduced power state. To wake up the server, either press the power button or use the web interface of the integrated IMM.

  - Flashing rapidly (four times per second): Server is powered-off and is not ready to be turned on.
  - Flashing slowly (once per second): Server is powered-off and is ready to be turned on.
- Activity (green): Indicates hard disk activity. If this LED is lit, it means that the hard disk has failed.
- Locate (blue): Location LED controllable by the integrated IMM. This LED is useful to physically identify a system from the management software for maintenance purposes.
- Fault (amber): Indicates that a fault has been detected by the integrated IMM.

4.3.10 Chassis configurations

The basic component of an iDataPlex rack is the chassis that houses compute and storage elements. The chassis implements a Flex Node Technology that allows the same 2U chassis to serve different purposes.

The chassis configurations can be grouped as follows:

- Compute-intensive
- I/O- and GPGPU-intensive (PCIe tray)

The chassis configurations differ in the number of available processors, disks, and PCIe slots. Table 4-18 gives an overview of the different configurations available.

Table 4-18 Available 2U chassis configurations for the dx360 M4 server

Node type	**Number of nodes**	**Config name**	**Disk drive bays** (all are Simple Swap)		
			3.5-inch SATA bays	**2.5-inch SATA/SAS/SSD bays**	**1.8-inch SSD bays**
Compute-server (2U)	2	1A	2	0	0
	2	1B	0	4	0
	2	1C	0	0	8 SSD
I/O server (2U)	1	2A	2	0	0
	1	2B	0	4	0
	1	2C	0	0	8 SSD

In all configurations, the 2U chassis provides a single, shared power supply and a single fan assembly (for details, see 4.6.1, “2U and 3U chassis” on page 121).

Table 4-19 lists the supported server configurations with the dx360 M4 server. It includes details about the RAID controller options, the available PCIe slots, and the supported power supply for each configuration.

Table 4-19 Supported configurations with the dx360 M4 server

Config	Disk drive	Storage controller	PCIe slots (total / available), excluding the mezzanine slot	Power supply options		
				550 W	750W	900W
1A	SATA, 3.5-inch	onboard	4 / 4	Yes[a]	Yes[b]	Yes[c]
1B	SATA/SAS/SSD, 2.5-inch	onboard, H1110, M1115, or 6 Gb Performance Optimized HBA	4 / 2 or 4	Yes[a]	Yes[b]	Yes[c]
1C	SSD, 1.8-inch	6Gb Performance Optimized HBA	4 / 2	Yes[a]	Yes[b]	Yes[c]
2A	SATA, 3.5-inch	onboard, H1110, or M1115	4 / 3 or 4	No	No	Yes[d]
2B	SATA/SAS/SSD, 2.5-inch	H1110, M1115, or 6 Gb HBA	4 / 3	No	No	Yes[d]
2C	SSD, 1.8-inch	H1110, M1115, or 6 Gb HBA	4 / 3	No	No	Yes[d]

a. For selected low-power node configurations only using Intel Xeon E5-2650L or E5-2660 CPUs
b. Only supported with 60W, 70W, or 80W CPUs in a redundant non-redundant power supply configuration
c. Supported in a redundant and non-redundant power supply configuration
d. Depending on the node configuration (CPU, memory, and drives), supported either only in a non-redundant power supply configuration or in a redundant and non-redundant configuration

4.3.11 Compute-intensive server (1A, 1B, and 1C)

The compute-node chassis configuration contains two identical 1U servers with one of the following storage options:

- Two 3.5-inch SATA (configuration 1A)
- Four 2.5-inch SATA, SAS, or solid-state drives (configuration 1B)
- Eight 1.8-inch solid-state drives (configuration 1C)

In each of the configurations, a 1U riser card is available to provide space for one PCIe adapter. Two of these cards can be installed per server.

Figure 4-12 shows the 3.5-inch drive configuration 1A. The 3.5-inch drives in each node must be of the same configuration (including disk size). These 3.5-inch drives use a simple-swap drive tray.

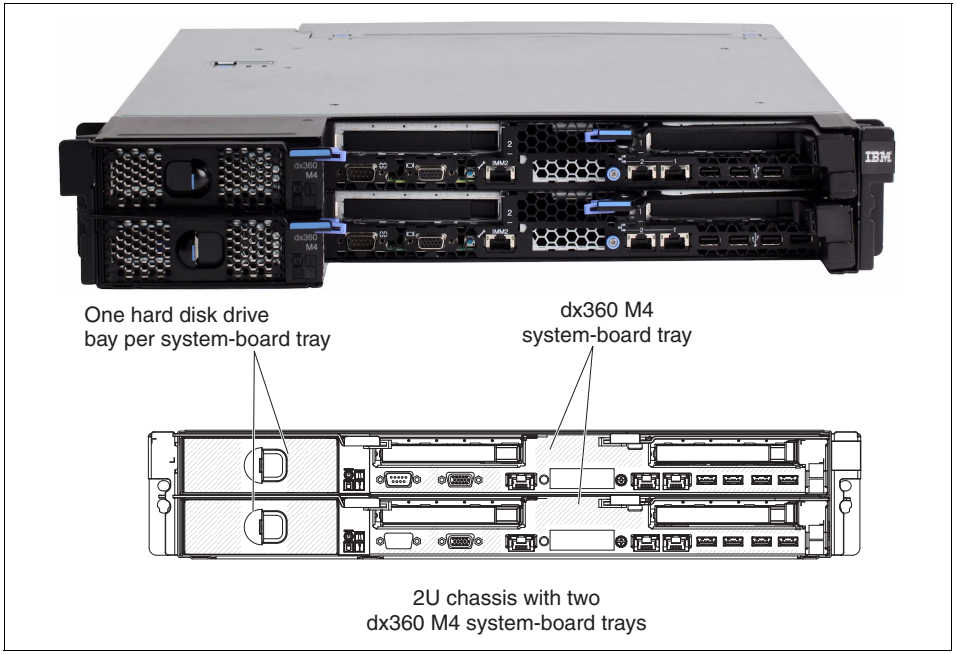


Figure 4-12 2U chassis with two servers and one 3.5-inch drive each; configuration 1A

If you use SAS drives or want to configure SATA/SAS RAID, you must also have one of the supported RAID disk controllers installed.

4.3.12 I/O and GPGPU server (2A, 2B, and 2C)

For I/O- and GPGPU-intensive applications, IBM offers configurations that have a larger number of PCIe slots. Two of these slots can be equipped with GPGPU adapters and the other two slots support any of the available PCIe adapters (for a detailed list of PCIe adapters, see Table 4-14 on page 55). The dx360 M4 server uses the so-called GPGPU tray. These chassis configurations can have any of the following combinations:

- Two 3.5-inch SATA drives (configuration 2A)
- Four 2.5-inch SATA, SAS, or solid-state drives (configuration 2B)
- Eight 1.8-inch solid-state drives (configuration 2C)

Each configuration supports a 2U riser card. This riser card provides space for one PCIe GPGPU adapter and one additional PCIe adapter. Two riser cards can be installed per server.

4.3.13 General hardware configuration rules

This section briefly outlines the major configuration rules that apply when putting together a new iDataPlex configuration:

- For a 2U-compute solution, both servers must have matching CPUs, memory configurations, and disks (including disk size). Different node configurations within a single 2U chassis are not supported.
- You cannot mix SATA and SAS in a single chassis.
- If hardware RAID is required, a SAS/SATA RAID controller card must be installed into an available PCIe slot.
- SAS configurations require a SAS controller card in the PCIe slot. The onboard disk controller of the current servers supports SATA only.
- If more than six drives are used (SATA or SAS), a SAS/SATA controller must be installed in the PCIe slot.
- If you are using SSDs and performance is crucial, consider installing the 6 Gb Performance Optimized host bus adapter.

4.3.14 Supported operating systems

At the time of writing, the following operating systems were supported:

- Microsoft:
  - Windows Server 2008/2008 R2, Web x64 Edition
  - Windows Server 2008/2008 R2, Enterprise x64 Edition
  - Windows Server 2008/2008 R2, Datacenter x64 Edition
  - Windows HPC Server 2008
- Linux:
  - Red Hat Enterprise Linux 5 Server x64 Edition (including KVM)
  - Red Hat Enterprise Linux 5 Server with Xen x64 Edition
  - Red Hat Enterprise Linux 6Server x64 Edition (including KVM)
  - SuSE Linux Enterprise Server 10 for AMD64/EM64T (including KVM and Xen)
  - SuSE Linux Enterprise Server 11 for AMD64/EM64T (including KVM and Xen)
- VMware:
  - VMware ESX/ESXi 4.1
  - VMware vSphere 5

Check ServerProven® for the latest list of supported operating systems:

http://ibm.com/systems/info/x86servers/serverproven/compat/us/idataplex.html

4.4 iDataPlex Direct Water Cooled dx360 M4 server

The iDataPlex Direct Water Cooled dx360 M4 server, available as machine type 7918 with a three-year warranty, is a 1U tray with a system board that fits into the Direct Water Cooled dx360 M4 2U chassis (machine type 7919). The server has a water loop assembly installed that brings water directly to the chips to cool them. This innovative new cooling technology has many advantages over conventional air cooling. More technical details about the Direct Water Cooling technology can be found in 4.4.1, “Direct Water Cooling technology” on page 66.

The following topics are included in this section:

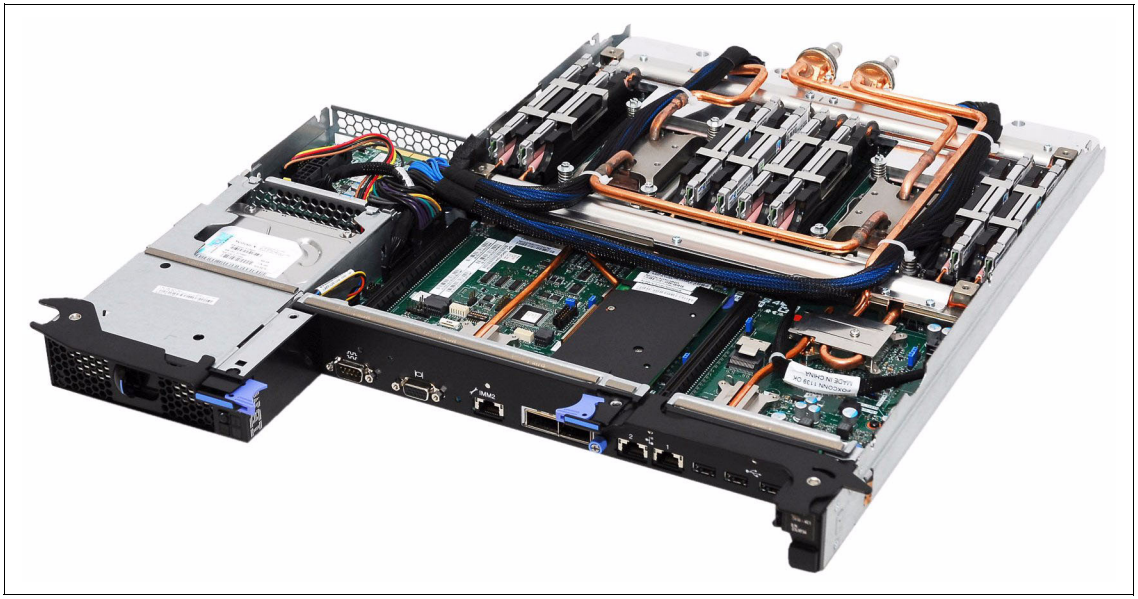


Figure 4-13 The iDataPlex Direct Water Cooled dx360 M4 server

The Direct Water Cooled dx360 M4 is a two-socket server running latest Intel Xeon E5-2600 series (“Sandy Bridge-EP”) processors. It uses the same system planar as the iDataPlex dx360 M4 server.

Figure 4-14 shows the internal connectors and slots of the Direct Water Cooled dx360 M4 server.

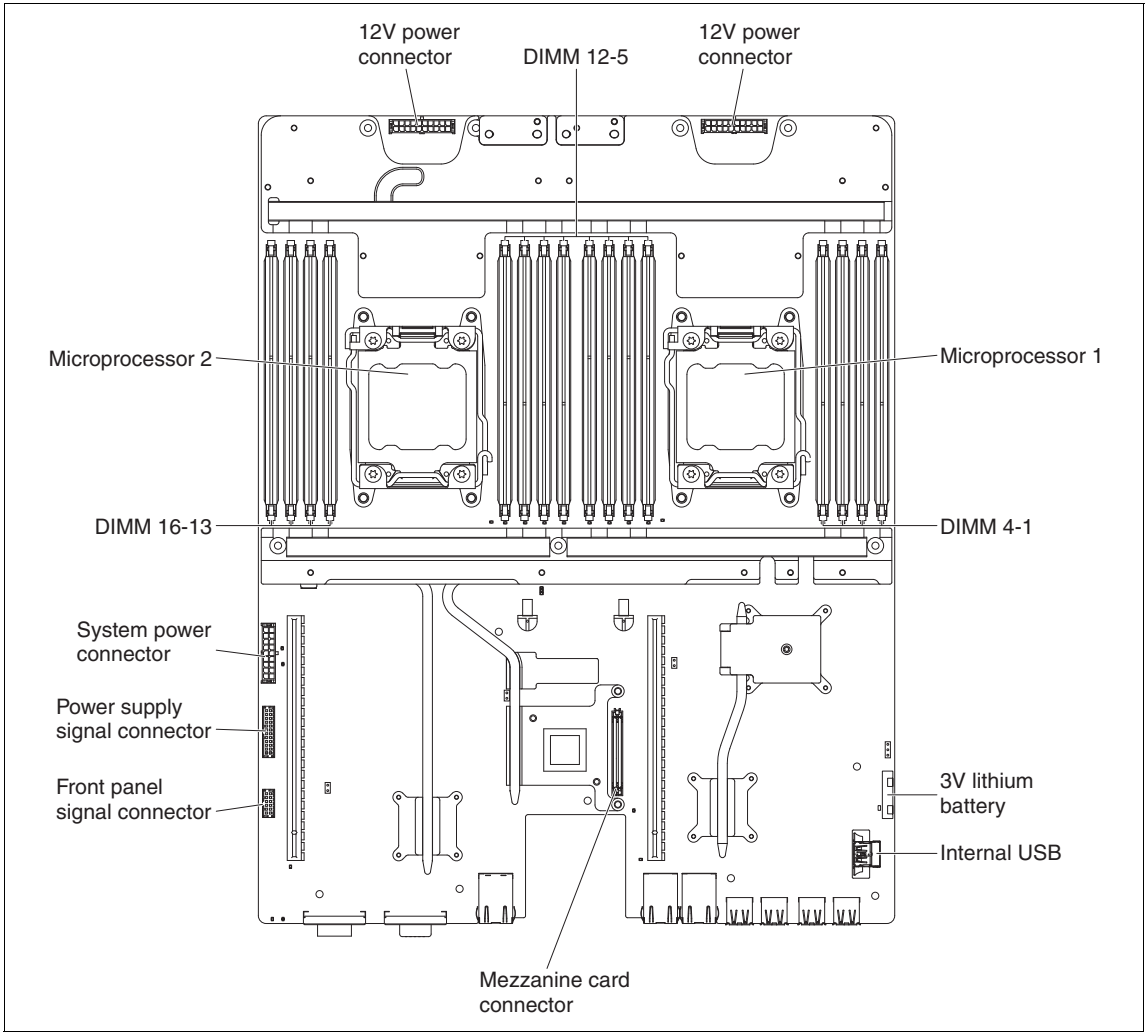


Figure 4-14 dx360 M4 system board layout

Figure 4-15 shows the front view of two Direct Water Cooled dx360 M4 servers. Reading from left to right, each server has one RS232 serial port, one VGA port supporting resolutions up to 1600x1200 at 75 MHz, one 1 Gbps RJ45 connector for dedicated systems management (wired to the IMM2), two ports from the InfiniBand mezzanine card, two 1 Gbps Ethernet interfaces based on the Intel Powerville I350 controller, and four USB 2.0 ports.

Figure 4-15 Front view of two Direct Water Cooled dx360 M4 servers in a 2U chassis

Note: The iDataPlex Direct Water Cooled dx360 M4 server can be ordered only through a special bid process.

Using a different cooling technology (such as Direct Water Cooling) requires additional testing (especially thermal testing) for each supported option. This is the reason why only a very limited subset of options has been verified and is supported by IBM.

4.4.1 Direct Water Cooling technology

The iDataPlex Direct Water Cooling technology is the first product-level implementation of the zero-emission data center project by IBM Research.

Rear Door Heat eXchanger: The iDataPlex Direct Water Cooling technology is not related to the Rear Door Heat eXchanger.

Although both systems use water to cool, the Rear Door Heat eXchanger is a rack-level cooling solution, whereas the Direct Water Cooling technology is a node-level cooling solution.

If you have a Rear Door Heat eXchanger installed, you cannot use the Direct Water Cooled dx360 M4 server, and vice versa.

Conventional cooling technologies use chilled air that passes through servers to absorb waste heat generated by processors, memory modules, power supplies, and similar devices. This approach is very inefficient because a lot of energy is required to force circulation of chilled air in a data center. In addition, the heat capacity (also known as thermal capacity) of air is very low compared to liquids such as water. It means that using air is a very inefficient method to transport heat away from chips.

The new Direct Water Cooling technology overcomes these obstacles and brings water directly to where the heat is generated. To achieve this, each Direct Water Cooled dx360 M4 server has its own water loop assembly installed and water is pumped through this assembly and runs along the chips surface. This approach removes the heat in the most efficient way possible.

Why should I choose iDataPlex Direct Water Cooling for my data center?

Using water instead of air gives you the following benefits:

- Lower node power
- Improved performance per watt
- Lower noise level (because of missing fans in servers and chassis)

Because you can use warm water (instead of chilled water with conventional water-cooling solutions), you have the following additional benefits:

- Direct reuse of waste heat (for example, to warm other buildings)
- No chillers required to cool down the outlet water upon every circulation, free air cooling is enough.

IBM Research: Zero-emission data center project[1]

The goal of this IBM project is to minimize carbon-dioxide emission of data centers through the direct use of waste heat. The key to achieve this goal is to cool server with hot water (hot water means inlet temperatures of up to 60°C).

Why can warm (or even hot) water cool the servers?

Cooling is defined by the transfer of heat from a warmer object to a colder object. All chips on today's system boards are specified to work up to 80°C or higher. As long as the temperature of the coolant stays below the desired operational temperature of the server, the waste heat generated by the onboard chips still dissipates into the (hot) water.

[1] More information about the Zero-emission data center project can be found on the IBM Research – Zurich website at http://www.zurich.ibm.com/st/energy/zeroemission.html

New microchannel heat sinks reduce the thermal resistance between the processor and the cooling water. Figure 4-16 compares a conventional air-cooled heat sink with a microchannel heat sink. Both pictures use the same scale, so one can easily see the difference in size for the fins and for the heat sink.

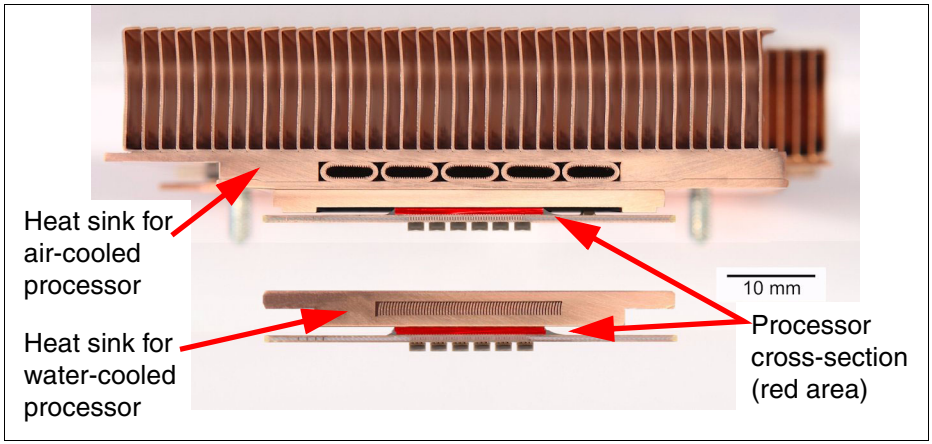


Figure 4-16 Cross section of an air-cooled heat sink (top) and a water-cooled microchannel heat sink (bottom). Red area indicates microprocessor.[2]

Reducing the thermal resistance allows for two major changes to the cooling system. One can use higher-temperature water as cooling water and the water pressure of the cooling loop can be reduced, which means that less water runs through the system per time interval. Both changes allow you to reduce the overall energy consumption of a data center when using servers with this cooling technology.

In addition, using warm water as the cooling fluid means that you do not have to cool down the water using energy-hungry chillers upon every circulation through the cooling loop. You can use free air-cooling instead, as long as the outside air temperature is lower than the cooling-water temperature.

The second advantage when using warm water as cooling water is that it allows you to directly reuse it, for example, to warm other buildings or houses. This advantage is not possible with water cooling solutions that use lower grade water.

[2] Figure originally published in Science Magazine. G.I. Meijer, Science 328, 318 (2010).

What is the water temperature of my Direct Water Cooled dx360 M4 servers?

The Direct Water Cooling technology supports a wide range of water temperatures, including warm water temperatures.

The exact water temperature that your servers will run with depends on many factors that are specific to each installation and data center infrastructure.

Direct Water Cooled dx360 M4 water loop assembly

Besides the CPUs and memory modules, there are additional elements on a system board that require cooling, for example, voltage regulators or the Intel platform controller hub.

The water loop assembly covers all of these elements, as shown in Figure 4-17. The assembly on each server contains about 70ml of water.

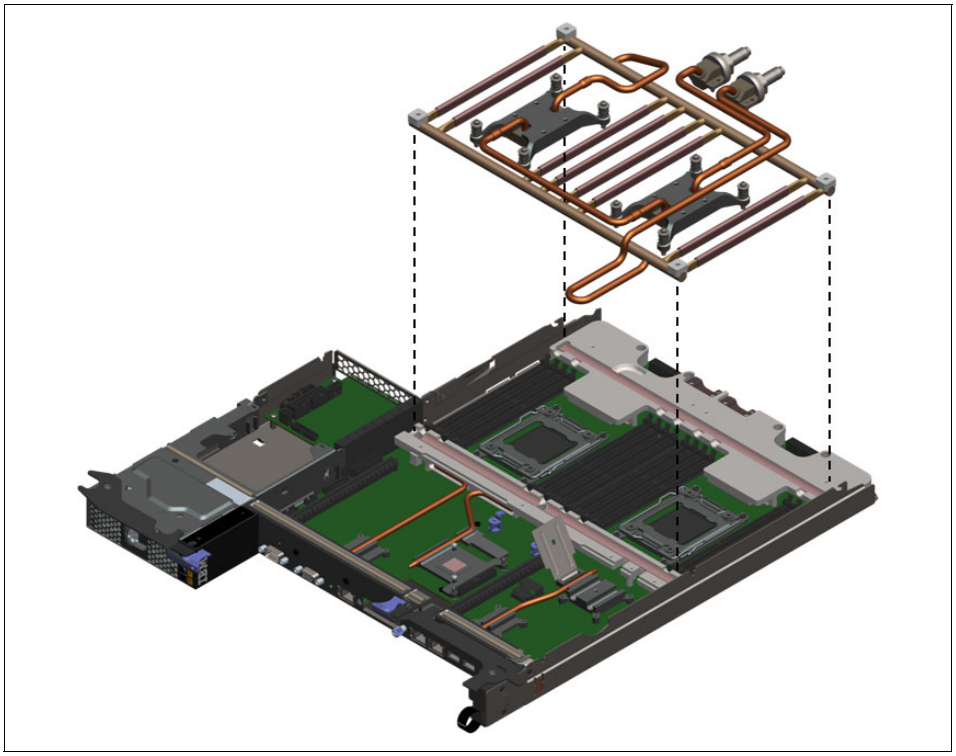


Figure 4-17 Water loop assembly of the Direct Water Cooled dx360 M4 server

For some of the onboard chips, such as the embedded service processor, passive water cooling is sufficient. Passive water cooling means that the heat is transported through a heat spreader connected to the water loop assembly before it dissipates into water. The advantage is that only one connection is required to transport the heat (instead of two connections to pump the water towards the chip and away from the chip).

Water flow path on the Direct Water Cooled dx360 M4

The water flow path is illustrated in Figure 4-18. Water enters the cooling loop assembly through the left connector in the back of the chassis, flows to the two Intel microprocessors, and is then distributed in a parallel fashion (by the rear cold rail) over the memory module slots of the board. After collecting the heat from the memory modules (by the front cold rail), the water flows to the Intel C600 platform controller hub (PCH) and exits the server through the right connector in the back.

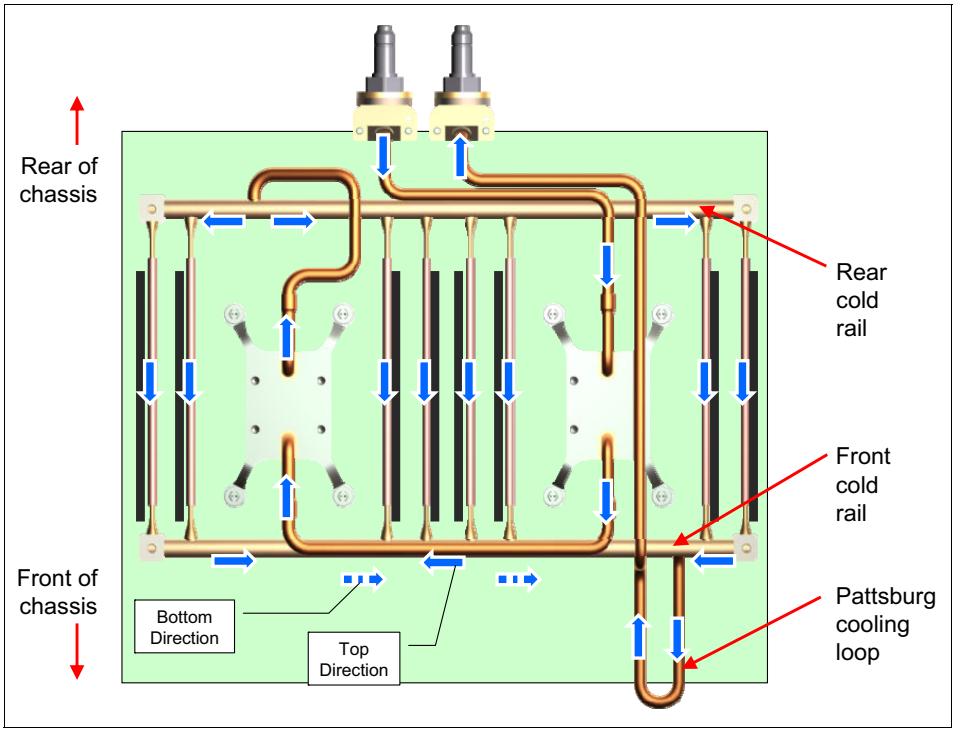


Figure 4-18 Water flow path on the Direct Water Cooled dx360 M4 server

Connection to the Direct Water Cooled Rack

Every server has its own connection to the water manifold mounted at the rear of the iDataPlex rack. Each server has a separate water inlet and outlet hose that extend the water loop from the manifold. Figure 4-19 illustrates a node drop from the water manifold to a single server.

How can I plug or unplug a Direct Water Cooled dx360 M4 server for maintenance?

You can plug and unplug a server from its chassis without shutting down any of the other nodes in the chassis or in the rack. Quick connect couplings prevent water from being released unintentionally.

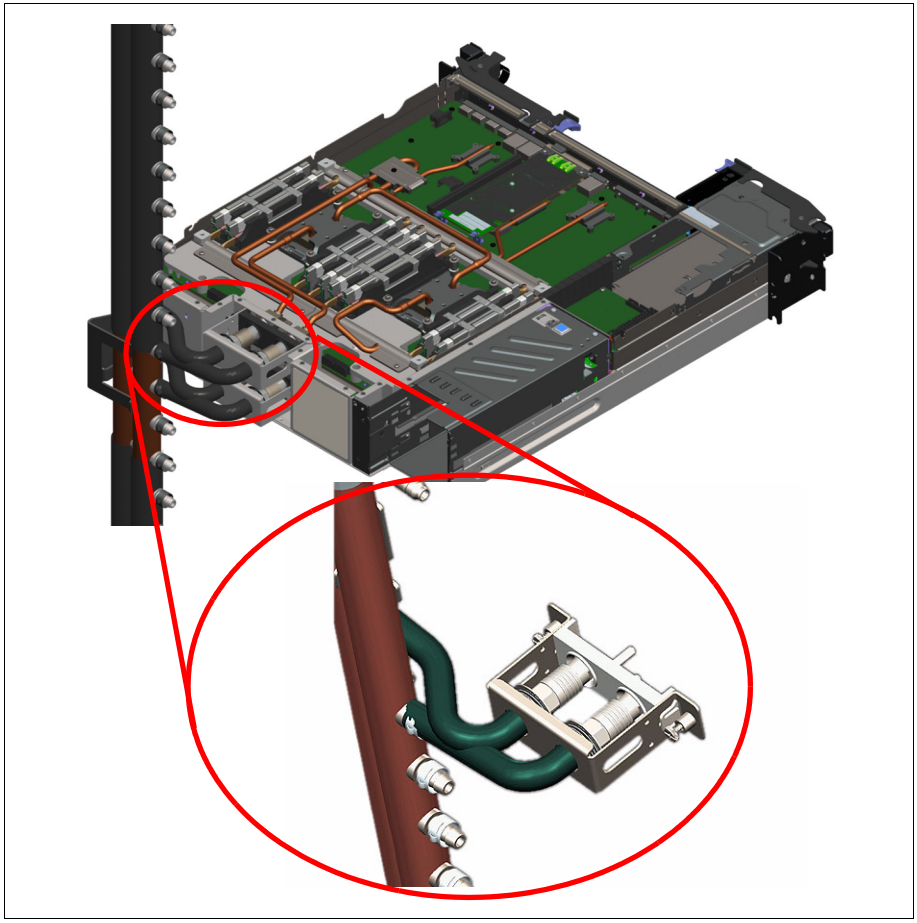


Figure 4-19 Connection of a Direct Water Cooled dx360 M4 2U chassis to the manifold

There is an inlet copper pipe that distributes cooling water to every node and an outlet copper pipe that collects the water. Both pipes are tied together with brackets to a single water manifold.

For more information about the manifold and the Direct Water Cooled rack, see 5.1.2, “Direct Water Cooled Rack with water manifold” on page 137.

For more information about how to design a data center using Direct Water Cooled technology, see 6.7, “Data center design using Direct Water Cooling” on page 190.

4.4.2 Processors

The Direct Water Cooled dx360 M4 server has two LGA 2011 CPU sockets for the latest Intel Xeon E5-2600 series processors. Table 4-20 lists the features for the processor currently supported on the Direct Water Cooled dx360 M4. Additional processors are planned to be supported in future. The processors installed in the two sockets must be identical.

Table 4-20 Supported Intel Xeon E5-2600 series CPU options on the Direct Water Cooled dx360 M4 server

Processor model	Core speed	Turbo	HT	Last level cache	Cores	Power	QPI links speed	Max DDR3 memory speed	Feature code
E5-2667	2.9 GHz	Yes	Yes	15 MB	6	130 W	8 GT/s	1600 MHz	A1TQ

The Intel Xeon E5-2600 series processors implement the second generation of the Intel Core microarchitecture using a 32nm manufacturing process (Figure 4-20).

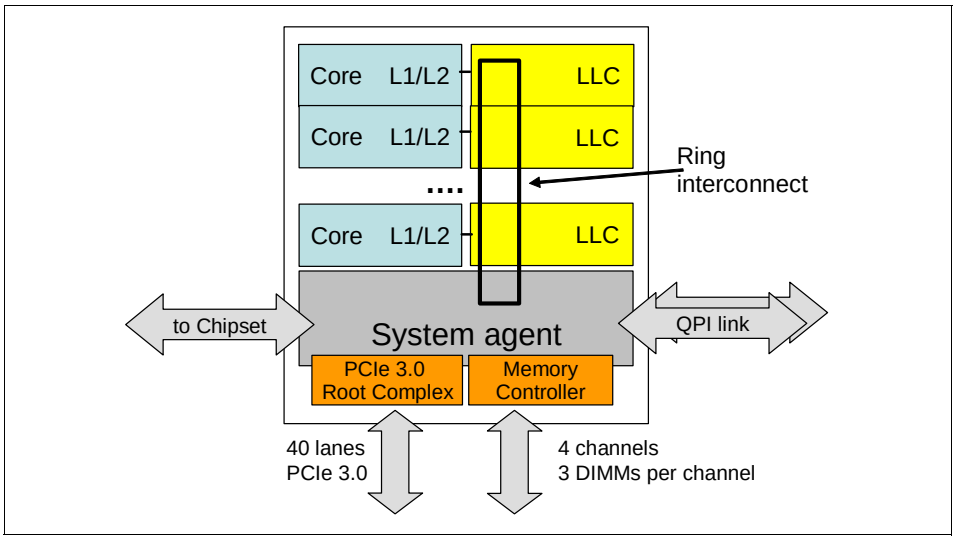


Figure 4-20 Intel Xeon E5-2600 basic architecture

The new architecture allows you to share data on-chip through a high-speed ring interconnect 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 three DDR3 DIMMs per channel running at a speed up to 1600 MHz. Two QPI links connect to a second CPU in a dual-socket installation.

The Xeon E5-2600 series is available with up to eight cores and 20 MB of last-level cache. It features an enhanced instruction set called Intel Advanced Vector Extensions (AVX). It doubles the operand size for vector instructions (such as floating-point) to 256 bits and boosts selected applications by up to a factor of two.

The new architecture also introduces Intel Turbo Boost Technology 2.0 and improved power management capabilities. Turbo Boost automatically turns off unused processor cores and increases the clock speed of the cores in use as long as thermal requirements are still met. Turbo Boost Technology 2.0 takes advantage of the new integrated design and also implements a more granular overclocking in 100 MHz steps instead of 133 MHz steps on former Nehalem-/Westmere-based microprocessors.

4.4.3 Memory

The planar of the Direct Water Cooled dx360 M4 server has 16 DIMM slots for registered ECC DDR3 DIMMs running at 1066, 1333, or 1600 MHz. However, only up to eight DIMM slots have been thermally validated by IBM. The water loop assembly is designed to cool all 16 DIMM slots.

The Direct Water Cooled dx360 M4 supports a 4 GB memory module, as listed in Table 4-21. Modules with different memory capacity are planned to be supported in the future.

Table 4-21 Supported memory option for the Direct Water Cooled dx360 M4 server

Feature code	Description	Number of ranks	Technology	Size	Voltage
A27Y	DDR3-1600 RDIMM PC3L-12800 CL11	2R	x8	4 GB	1.5 V

Note: The memory module is the same as feature code A24L of the air-cooled dx360 M4 server except for the module heat spreader. A bridging heat spreader is used to implement single-sided water cooling of the memory modules. The bridging heat spreader absorbs all heat from only one side of the module.

Figure 4-21 shows the cooling of the memory modules. A thermal pad is used to connect the memory module with the water loop.

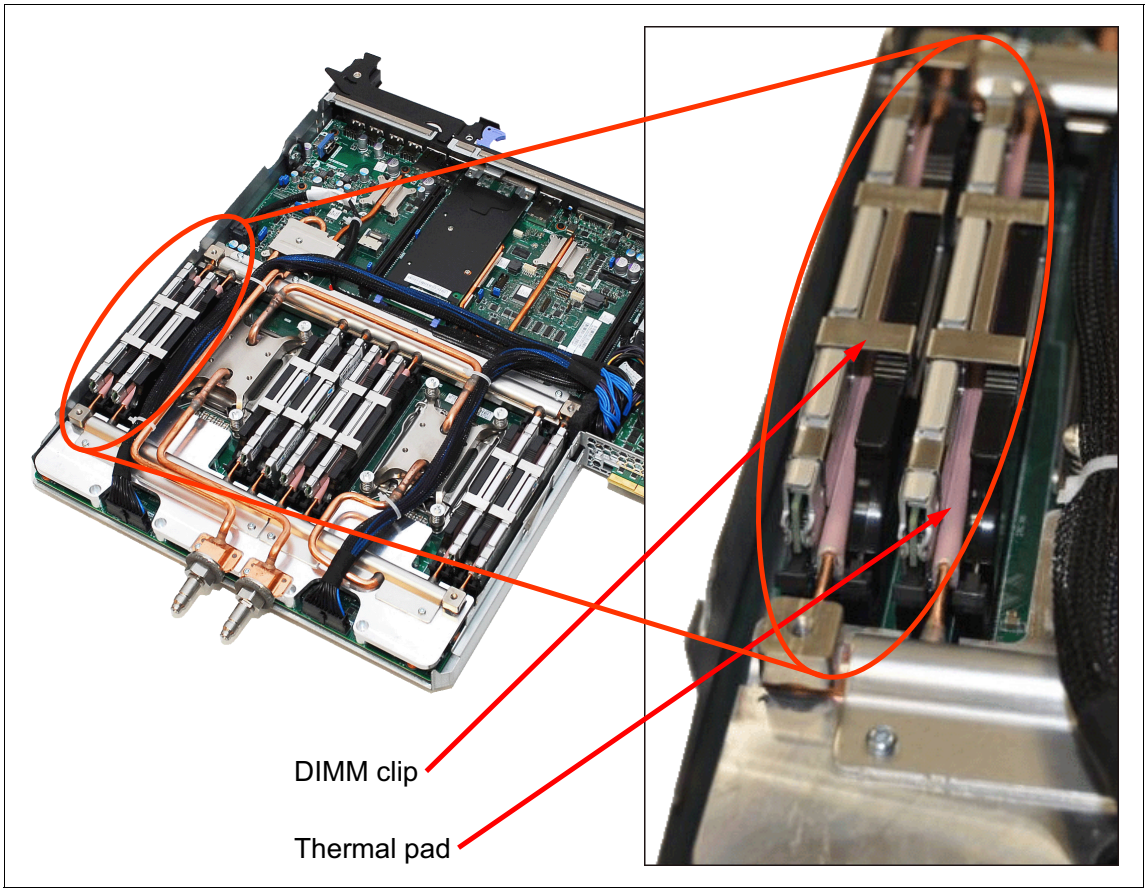


Figure 4-21 Memory module water cooling assembly

Note: All memory modules on the Direct Water Cooled dx360 M4 are customer-replaceable units. The procedure to install or remove DIMMs is identical to air-cooled planars.

The Intel Xeon processor E5-2600 architecture with the QuickPath Interconnect (QPI) between processors has a four-channel memory interface per processor with two DIMM slots per channel, as shown in Figure 4-22.

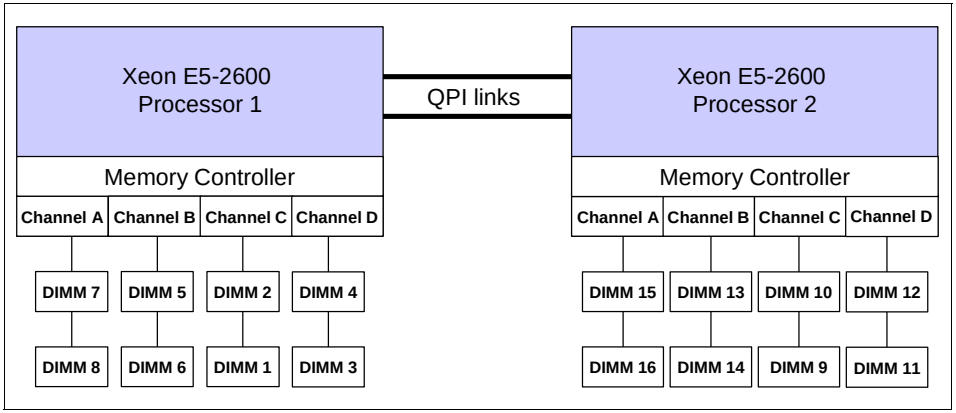


Figure 4-22 Memory channel layout of the Direct Water Cooled dx360 M4 server

The UEFI on the Direct Water Cooled dx360 M4 server supports three modes of operation:

- Normal channel mode (also known as independent channel mode)
- Mirrored channel mode
- Rank sparing mode

Normal mode is the default setting. The operating system uses the full amount of memory installed. No redundancy is provided in this mode.

The population sequence for a single-CPU and two-CPU system is the same as for the air-cooled dx360 M4 server. See 4.3.3, “Memory” on page 43 for the exact population sequence.

The memory speed supported in a specific configuration depends on several different parameters. However, the only supported memory module allows all speeds up to 1600 MHz, as shown in Table 4-22.

Table 4-22 Direct Water Cooled dx360 M4 memory channel speed support

		Maximum memory channel speed and voltage support by DIMM per channel			
Feature code	**Memory capacity per DIMM**	**1 DIMM per channel**		**2 DIMMs per channel**	
		1.35 V	**1.5 V**	**1.35 V**	**1.5 V**
A27Y	4 GB	Not supported	1600 MHz	Not supported	1600 MHz

When using *mirrored mode*, Channel A and Channel B both hold the same data for redundancy purposes. If one DIMM fails, it is disabled and the backup DIMM in the other channel takes over. Likewise, Channel C and Channel D hold the same data and create one mirrored pair. Because memory mirroring is handled in hardware, it is operating system-independent. The total usable memory size is the sum of all DIMMs in Channel A or B plus the sum of all DIMMs in Channel C or D.

When using *rank sparing mode*, one memory DIMM rank serves as a spare of the other ranks on the same channel. The spare rank is held in reserve and is not used as active memory. The spare rank must have identical or larger memory capacity than all the other active memory ranks on the same channel. After an error threshold is surpassed, the contents of that rank is copied to the spare rank. The failed rank of memory is taken offline, and the spare rank is put online and used as active memory in place of the failed rank.

4.4.4 Internal storage and controllers

This section provides information about the internal storage options of the Direct Water Cooled dx360 M4 server.

The onboard controller of the Direct Water Cooled dx360 M4 is only for SATA drives and does not support RAID, as shown in Table 4-23.

Table 4-23 Onboard disk controller features of the Direct Water Cooled dx360 M4

Controller feature	Direct Water Cooled dx360 M4
SATA II support	Yes, 2-port
SATA port speed	6 Gbps
SAS support	No
Hardware-RAID	No

Table 4-24 lists the conventional disk drives that are supported with the Direct Water Cooled dx360 M4 server in the 2U chassis configuration.

Table 4-24 Supported disk drive on the Direct Water Cooled dx360 M4

Interface type	Simple Swap / Hot Swap	Interface speed	Disk size	Form factor	RPM	Feature
SATA II NL	Simple Swap	6 Gbps	500 GB	3.5-inch	7200	A22U

4.4.5 PCI Express I/O Adapters

The iDataPlex Direct Water Cooled dx360 M4 server supports one onboard PCIe card through a mezzanine card. Table 4-25 lists the details of this slot.

Table 4-25 PCIe features of the Direct Water Cooled iDataPlex dx360 M4 server

Feature	Dual-node chassis config
Mezzanine PCIe, mechanically	x8
Mezzanine PCIe, electrically	x8 3.0

Figure 4-23 shows this slot on the system board.

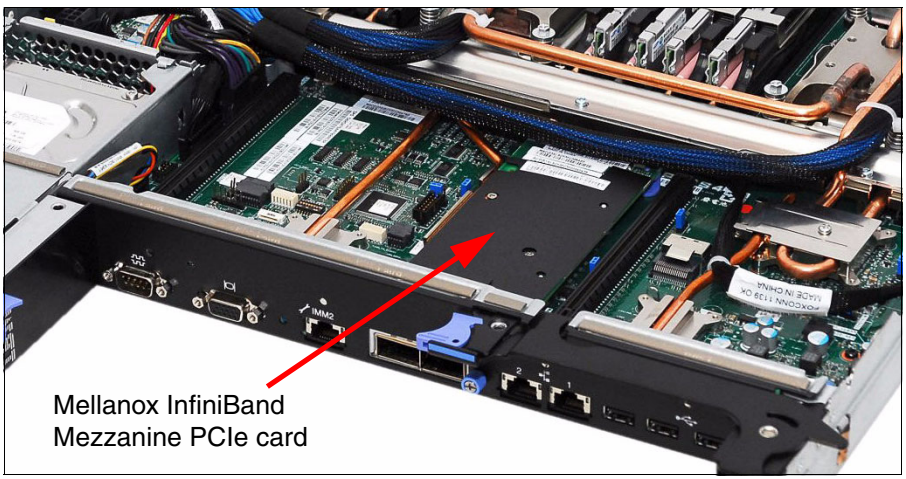


Figure 4-23 Direct Water Cooled dx360 M4 server mezzanine option card slot

The only supported adapter for this slot is listed in Table 4-26.

Table 4-26 Supported PCIe mezzanine option cards on the Direct Water Cooled dx360 M4 server

Type	Description	PCIe Interface speed	Feature Code
InfiniBand / Ethernet	Mellanox ConnectX-3 2-port QDR/FDR-10 (water cooled)	x8	A24G

4.4.6 Management

Each iDataPlex Direct Water Cooled dx360 M4 server has an Integrated Management Module version 2 (IMM2) onboard and uses the Unified Extensible Firmware Interface (UEFI) to replace the older BIOS interface. The IMM2 is IPMI v2.0-compliant and provides the following major features:

- Remote configuration of IMM2 and UEFI settings without the need to power-on the server
- Remote access to system fan, voltage, and temperature values
- Remote IMM and UEFI update
- UEFI update when the server is powered off
- Remote console by way of a serial over LAN
- Remote presence (KVM and CD/DVD) and blue screen capture (requires additional feature on demand key, part number 90Y3901)
- Remote access to the system event log
- Predictive failure analysis and integrated alerting features (for example, by SNMP)

For more information and details about the IMM, see 7.1, “Integrated Management Module” on page 224.

To activate full systems management capabilities, an IBM Remote Presence Upgrade (Feature on Demand key) is required. This feature is a license that you apply to the system to enable. See Table 4-27.

Table 4-27 IBM Remote Presence Upgrade (feature on demand key)

Part number	Description
90Y3901	IMM2 IBM Remote Presence Upgrade (feature on demand key)

4.4.7 Operator panel

The iDataPlex Direct Water Cooled dx360 M4 server has an operator panel at the front. Figure 4-24 shows the power button with the integrated power LED and three status LEDs.

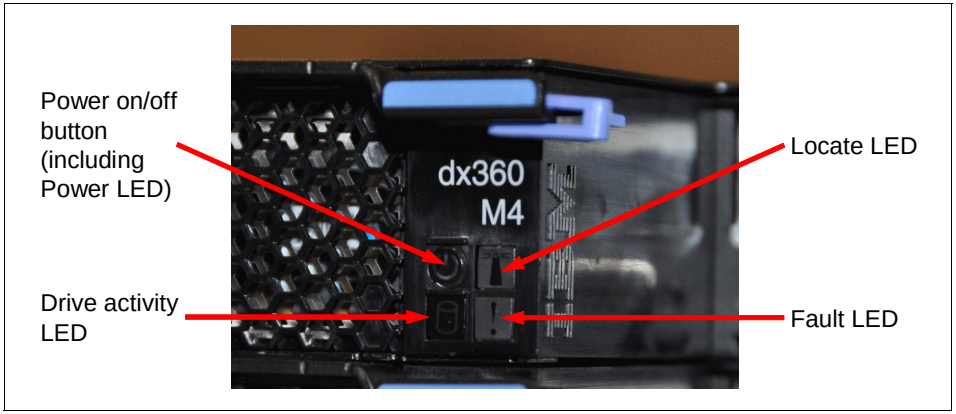


Figure 4-24 iDataPlex Direct Water Cooled dx360 M4 operator panel

The LEDs are defined as follows:

- Power on (green): Determines the power state of the server:
  - Off: AC power is not present.
  - On: Server is powered-on.
  - Fading on and off: Server is in a reduced power state. To wake up the server, either press the power button or use the web interface of the integrated IMM.
  - Flashing rapidly (four times per second): Server is powered-off and not ready to be turned on.
  - Flashing slowly (once per second): Server is powered-off and ready to be turned on.
- Activity (green): Indicates hard disk activity. If this LED is lit, it means that the hard disk has failed.
- Locate (blue): Location LED controllable by the integrated IMM. This LED is useful to physically identify a system from the management software for maintenance purposes.

Fault (amber): Indicates that a fault has been detected by the integrated IMM.

4.4.8 Chassis configurations

The Direct Water Cooled dx360 M4 server is supported only in compute-intensive configurations where two servers are installed in one Direct Water Cooled dx360 M4 2U chassis.

With only one 3.5-inch SATA II drive being supported, Figure 4-25 shows the chassis configuration using this drive.

Figure 4-25 Direct Water Cooled dx360 M4 2U chassis in a compute-intense configuration

Tip: You can plug and unplug a Direct Water Cooled dx360 M4 server without shutting down any of the other nodes in the chassis or rack.

4.4.9 Supported operating systems

At the time of writing, the following operating systems were supported:

- Microsoft:
  - Windows Server 2008/2008 R2, Web x64 Edition
  - Windows Server 2008/2008 R2, Enterprise x64 Edition
  - Windows Server 2008/2008 R2, Datacenter x64 Edition
  - Windows HPC Server 2008
- Linux:
  - Red Hat Enterprise Linux 5 Server x64 Edition (including KVM)
  - Red Hat Enterprise Linux 5 Server with Xen x64 Edition
  - Red Hat Enterprise Linux 6Server x64 Edition (including KVM)
  - SuSE Linux Enterprise Server 10 for AMD64/EM64T (including KVM and Xen)
  - SuSE Linux Enterprise Server 11 for AMD64/EM64T (including KVM and Xen)
- VMware:
  - VMware ESX/ESXi 4.1
  - VMware vSphere 5

Check ServerProven for the latest list of supported operating systems:

http://ibm.com/systems/info/x86servers/serverproven/compat/us/idataplex.html

4.5 iDataPlex dx360 M3 server

The iDataPlex dx360 M3 server, available as machine type 6391 with a three-year warranty, is a 1U tray with a system board that fits into both the 2U Flex chassis and the 3U chassis. The dx360 M3 is a two-socket server running Intel Xeon 5600-series (Westmere-EP) processors. See Figure 4-26.

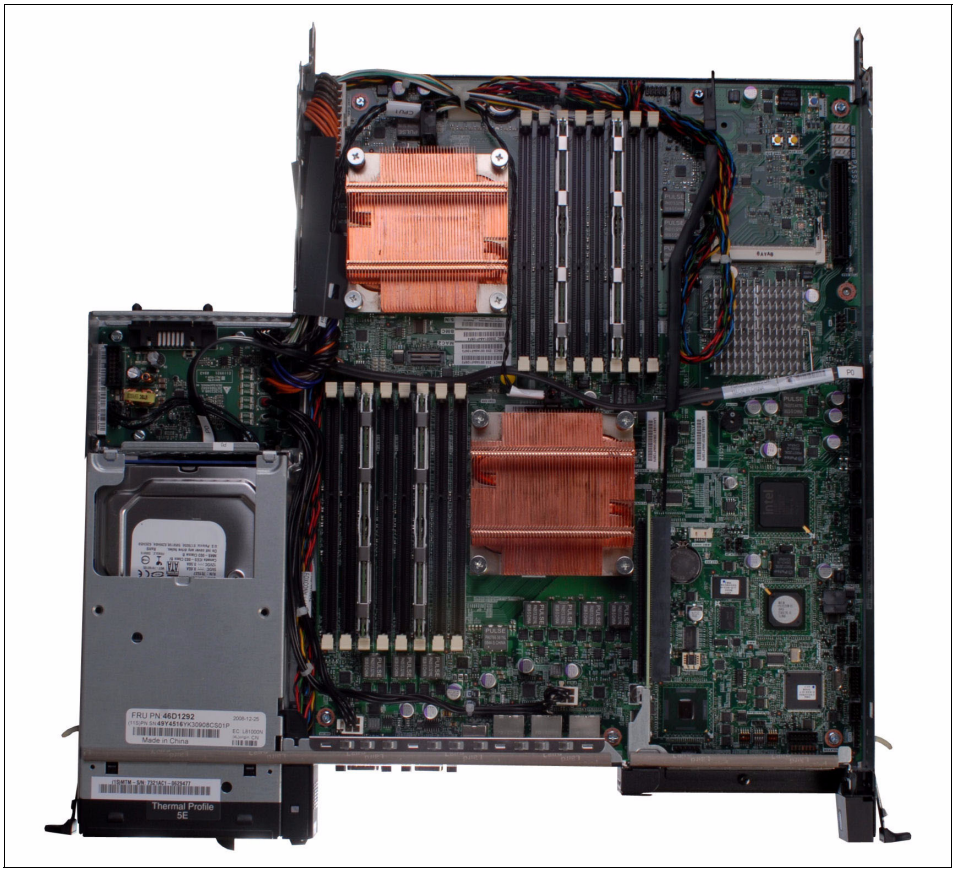


Figure 4-26 The iDataPlex dx360 M3 server

Figure 4-27 shows the internal connectors and slots of the dx360 M3 server.

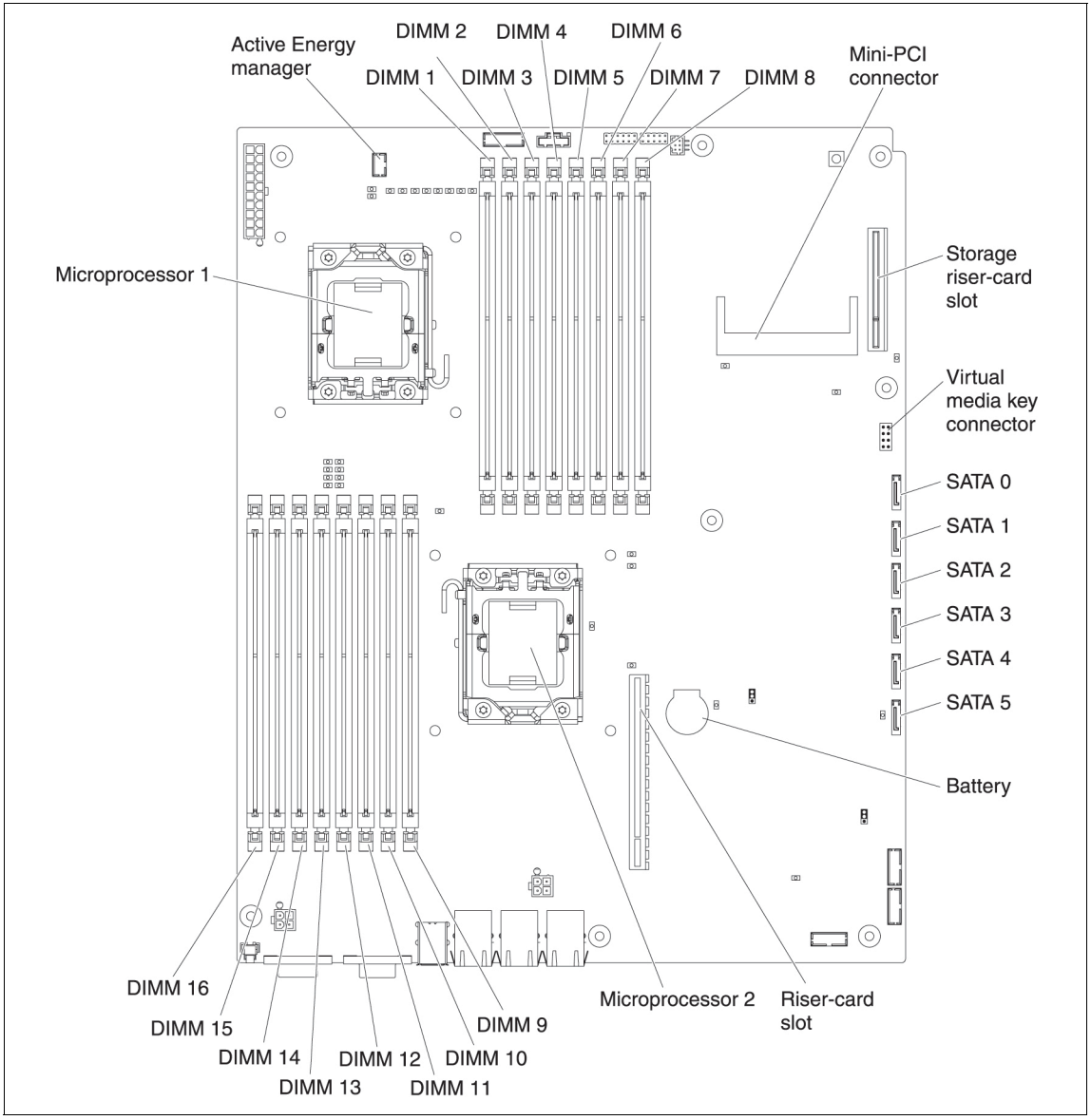


Figure 4-27 dx360 M3 system board layout

Figure 4-28 shows the front view of two dx360 M3 servers. Reading from left to right, the server has one RS232 serial port, one VGA port (connected to an onboard Matrox G200 graphics controller in the IMM supporting resolutions up to 1280x1024), two USB 2.0 ports, one 10/100 Mbps RJ45 connector for dedicated systems management (wired to the IMM), and two 1 Gbps Ethernet interfaces based on the Intel 82575 controller.

Figure 4-28 Front view of two dx360 M3 servers in a 2U Flex chassis

Product publications are available at the IBM support site:

- *IBM System x iDataPlex dx360 M3 User's Guide:*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5083500

- *IBM System x iDataPlex dx360 M3 Problem Determination and Service Guide:*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5084068

4.5.1 Processors

The dx360 M3 server is based on the SSI industry standard, and it has two LGA1366 CPU sockets for the Intel Xeon 5600 series processors. Table 4-28 shows the supported processor options. The processors installed in the two sockets must be identical.

Table 4-28 Supported Intel Xeon 5500/5600 series CPU options on the dx360 M3 server

Processor model	Core Speed	Turbo	HT	L3 cache	Cores	Power	QPI Link speed	Max DDR3 memory speed	Feature code
Power-optimized (L)									
Xeon L5609	1.86 GHz	No	No	12 MB	4	40 W	4.8 GT/s	1066 MHz	0720
Xeon L5630	2.40 GHz	Yes	Yes	12 MB	4	40 W	5.86 GT/s	1066 MHz	0715
Xeon L5640	2.26 GHz	Yes	Yes	12 MB	6	60 W	5.86 GT/s	1333 MHz	0714
Mainstream (E)									
Xeon E5603	1.60 GHz	No	No	4 MB	4	80 W	4.8 GT/s	1066 MHz	A141
Xeon E5606	2.13 GHz	No	No	8 MB	4	80 W	4.8 GT/s	1066 MHz	A140
Xeon E5607	2.26 GHz	No	No	8 MB	4	80 W	4.8 GT/s	1066 MHz	A113Z
Xeon E5620	2.40 GHz	Yes	Yes	12 MB	4	80 W	5.86 GT/s	1066 MHz	0713
Xeon E5630	2.53 GHz	Yes	Yes	12 MB	4	80 W	5.86 GT/s	1066 MHz	0712
Xeon E5640	2.66 GHz	Yes	Yes	12 MB	4	80 W	5.86 GT/s	1066 MHz	0711
Xeon E5645	2.40 GHz	Yes	Yes	12 MB	6	80 W	5.86 GT/s	1333 MHz	A13Y
Xeon E5649	2.53 GHz	Yes	Yes	12 MB	6	80 W	5.86 GT/s	1333 MHz	A13X
Performance (X)									
Xeon X5650	2.66 GHz	Yes	Yes	12 MB	6	95 W	6.4 GT/s	1333 MHz	0709
Xeon X5660	2.80 GHz	Yes	Yes	12 MB	6	95 W	6.4 GT/s	1333 MHz	0708
Xeon X5667	3.06 GHz	Yes	Yes	12 MB	4	95 W	6.4 GT/s	1333 MHz	0710
Xeon X5670	2.93 GHz	Yes	Yes	12 MB	6	95 W	6.4 GT/s	1333 MHz	0707
Xeon X5672	3.2 GHz	Yes	Yes	12 MB	4	95 W	6.4 GT/s	1333 MHz	A13V
Xeon X5675	3.06 GHz	Yes	Yes	12 MB	6	95 W	6.4 GT/s	1333 MHz	A13W

With the Xeon 5600 series processors, Intel has moved the memory controller into the processor and no longer has to rely on a front-side bus architecture to access memory. The memory is now directly connected to the processor. A new serial coherency link called QuickPath Interconnect (QPI) is used to connect the processor to the IO hub and to other processors. This setup enables a NUMA architecture similar to AMD HyperTransport.

For more information about the Intel Xeon 5600 series processors, see the following website:

http://www.intel.com/p/en_US/products/server/processor/xeon5000

The Xeon 5600 processors support the following Intel technologies:

- QuickPath Technology
- Intel Turbo Boost Technology
- Intel Hyper-Threading Technology
- Intel Virtualization Technology
- Intelligent Power Capability
- Intel 64 Technology

The Xeon 5600 series of processors has a number of new features, described in the following sections.

QuickPath Technology

Intel QuickPath Technology is a platform architecture that provides high-speed (up to 25.6 GBps), point-to-point connections between processors, and between processors and the I/O hub. Each processor has its own dedicated memory that it accesses directly through an Integrated Memory Controller. In cases where a processor needs to access the dedicated memory of another processor, it can do so through a high-speed QuickPath Interconnect (QPI) that links all the processors.

Intel Turbo Boost Technology

Turbo Boost Technology dynamically turns off unused processor cores and increases the clock speed of the cores in use. It will increase the frequency in steps of 133 MHz (to a maximum of three steps or 400 MHz) as long as the processors' predetermined thermal and electrical requirements are still met.

For example, with three cores active, a 2.26 GHz processor can run the cores at 2.4 GHz. With only one or two cores active, the same processor can run those cores at 2.53 GHz. Similarly, a 2.93 GHz processor can run at 3.06 GHz or even 3.33 GHz. When the cores are needed again, they are dynamically turned back on and the processor frequency is adjusted accordingly. This feature can be enabled or disabled in the UEFI BIOS.

For more information about Intel Turbo Boost Technology, see this website:

http://www.intel.com/technology/turboboost/

Intel Hyper-Threading Technology

This feature boosts performance for parallel, multi-threaded applications. Intel Hyper-Threading Technology enables simultaneous multi-threading within each processor core, up to two threads per core or eight threads per quad-core processor. Hyperthreading reduces computational latency, thereby making optimal use of every clock cycle. This feature can be enabled or disabled in the UEFI BIOS.

Intelligent Power Capability

Intelligent Power Capability powers individual processor elements on and off as needed, to reduce power draw.

Intel Virtualization Technology

Improvements to Intel Virtualization Technology (Intel VT-x) provide hardware-assisted page-table management, thus allowing the guest operating system more direct access to the hardware and reducing compute-intensive software translation from the VMM. Intel VT-x also includes Intel VT FlexMigration and Intel VT Flex-Priority, which are capabilities for flexible workload migration and performance optimization across the full range of 32-bit and 64-bit operating environments.

For more information about Intel Virtualization Technology, see this website:

http://www.intel.com/technology/virtualization/technology.htm

Advanced Encryption Standard: New Instructions (AES-NI)

There are new instructions that enable the processor to perform hardware-accelerated encryption for the AES algorithm. By significantly accelerating encryption, Intel AES-NI enables more secure systems. Increased speed means that longer encryption keys can be used, strengthening the protection of data. More network traffic can be encrypted with less impact on server performance. Database and storage encryption can be more widely and cost effectively deployed.

Processor P-states

The P-state mechanism involves running the processor clock at a lower frequency and ultimately with a lower voltage. It results in less power consumption as power consumption decreases with clock speed and voltage. New power-saving modes also make it possible to put memory and the system I/O controller into the lowest possible state, not only the processor. It helps to save more energy.

Support for 1 GB Hugepages

Hugepages (known as *large pages* in Windows) can help to reduce the overhead that an application needs to manage its own memory.

4.5.2 Memory

The dx360 M3 server has 16 DIMM slots for registered ECC DDR3 DIMMs running at 800, 1066, or 1333 MHz. It supports 2 GB, 4 GB, 8 GB, and 16 GB memory modules, as listed in Table 4-29.

Table 4-29 Supported memory options for the dx360 M3 server

Description	Number of ranks	Technology	Size	Voltage	Feature code
DDR3-1333 RDIMM PC3-10600 CL9	1R	x4	2 GB	1.5 V	8935
DDR3-1333 RDIMM PC3L-10600 CL9	1R	x4	2 GB	1.35 V	8922
DDR3-1333 RDIMM PC3L-10600 CL9	1R	x8	2 GB	1.35 V	8940
DDR3-1333 RDIMM PC3L-10600 CL9	2R	x8	2 GB	1.35 V	3893
DDR3-1333 RDIMM PC3-10600 CL9	2R	x8	2 GB	1.5 V	8934
DDR3-1333 RDIMM PC3L-10600 CL9	1R	x4	4 GB	1.35 V	8941
DDR3-1333 RDIMM PC3L-10600 CL9	2R	x4	4 GB	1.35 V	3894
DDR3-1333 RDIMM PC3-10600 CL9	2R	x4	4 GB	1.5 V	8936
DDR3-1333 RDIMM PC3L-10600 CL9	2R	x8	4 GB	1.35 V	8942
DDR3-1066 RDIMM PC3-8500 CL7	4R[a]	x8	8 GB	1.5 V	1706
DDR3-1333 RDIMM PC3-10600 CL9	2R	x4	8 GB	1.5 V	8937
DDR3-1066 RDIMM PC3L-8500 CL7	2R	x4	8 GB	1.35 V	8921
DDR3-1333 RDIMM PC3L-10600 CL9	2R	x4	8 GB	1.35 V	8923
DDR3-1066 RDIMM PC3L-8500 CL7	4R[a]	4	16 GB	1.35 V	8939

a. Because this module is a quad-rank module, only two can be installed per memory channel.

FAQ:

The IBM iDataPlex dx360 M3 server provides only 16 instead of the maximum possible 18 DIMM slots. Does it limit the performance of my application?

No. Maximum memory speed of 1333 MHz can only be achieved with 6 or 12 DIMMs populated (depending on 2 DIMMS-per-channel support of the processor).

The Xeon 5600 microprocessor architecture with Quick-Path Interconnect has a three-channel memory interface per processor that supports eight ranks per channel. This translates as follows:

- Three DIMMs per channel for single-rank and/or dual-rank DIMMs
- Two DIMMs per channel for quad-rank DIMMs

Therefore, if using quad-rank DIMMs, only twelve slots can be populated.

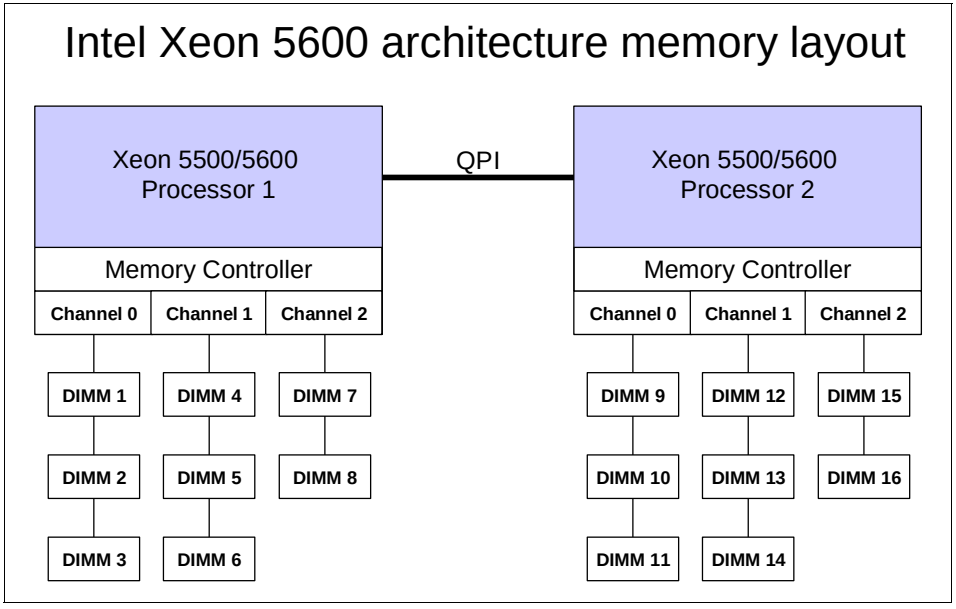


Figure 4-29 Memory channel layout of the iDataPlex dx360 M3 server

The firmware on the iDataPlex dx360 M3 server supports two modes:

- Normal channel mode
- Mirrored channel mode

Normal mode is the default setting. The operating system uses the full amount of memory installed. No redundancy is provided in this mode.

The population sequence for *normal mode* is listed in Table 4-30 for a single CPU system. See Figure 4-27 on page 83 for the physical location of the DIMM slots on the board.

Table 4-30 Memory population table for the dx360 M3 server with 1 CPU installed

Number of DIMMs	Max Memory Speed	Processor 1							
		Channel 0			Channel 1			Channel 2	
		DIMM 1	DIMM 2	DIMM 3	DIMM 4	DIMM 5	DIMM 6	DIMM 7	DIMM 8
1	1333 MHz			x					
2	1333 MHz			x			x		
3	1333 MHz			x			x		x
4	1333 MHz[a]		x	x			x		x
5	1333 MHz[a]		x	x		x	x		x
6	1333 MHz[a]		x	x		x	x	x	x
7	800 MHz	x	x	x		x	x	x	x
8	800 MHz	x	x	x	x	x	x	x	x

a. Maximum memory speed in this configuration: 1333 MHz when using Westmere-EP processors supporting 1333 MHz in combination with 1.5V DIMMs; 1066 MHz in all other Westmere-EP configurations.

Table 4-31 lists the population sequence for normal mode for a dual-CPU system.

Table 4-31 Memory population table for the dx360 M3 server with two CPUs installed

Number of DIMMs	Max memory speed	Processor 1								Processor 2							
		Channel 0			Channel 1			Channel 2		Channel 0			Channel 1			Channel 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
1	1333 MHz			x													
2	1333 MHz			x								x					
3	1333 MHz			x			x					x					
4	1333 MHz			x			x					x			x		
5	1333 MHz			x			x		x			x			x		
6	1333 MHz			x			x		x			x			x		x
7	1333 MHz[a]		x	x			x		x			x			x		x
8	1333 MHz[a]		x	x			x		x		x	x			x		x
9	1333 MHz[a]		x	x		x	x		x		x	x			x		x
10	1333 MHz[a]		x	x		x	x		x		x	x		x	x		x
11	1333 MHz[a]		x	x		x	x	x	x		x	x		x	x		x
12	1333 MHz[a]		x	x		x	x	x	x		x	x		x	x	x	x
13	800 MHz	x	x	x		x	x	x	x		x	x		x	x	x	x
14	800 MHz	x	x	x		x	x	x	x	x	x	x		x	x	x	x
15	800 MHz	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x
16	800 MHz	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

a. Maximum memory speed in this configuration: 1333 MHz when using Westmere-EP processors supporting 1333 MHz in combination with 1.5V DIMMs; 1066 MHz in all other Westmere-EP configurations.

Use the following general guidelines when deciding about the memory configuration of your server:

- Install memory DIMMs in order of their size, with the largest DIMM first.
- If you mix DIMMs with different memory speeds, the slowest DIMM determines the speed of the memory bus.
- You cannot run different memory speeds on different channels.
- If you mix DIMMs with 1.35V and 1.5V, the system runs all of them at 1.5V and you will lose the energy advantage.
- Populate equivalent ranks per channel.

Recommendation for two-socket systems: If your system must provide:

- Maximum bandwidth:

  Install any Xeon X56xx processor and populate 12 DIMMs (two DIMM slots per channel) with 1.5V DIMMs. It gives you up to 96 GB at 1333 MHz.
- Maximum capacity:

  Install all 16 DIMMs. It gives you 128 GB total, running at 800 MHz.
- Balanced performance:

  Populate 12 DIMMs, with two DIMMs per channel. It gives you up to 96 GB, running at 1033 MHz across all memory channels.

When using *mirrored mode*, Channel 0 and Channel 1 both hold the same data for redundancy purposes. If one DIMM fails, it is disabled and the backup DIMM in the other channel takes over.

Channel 2 is not used in this configuration, so DIMM slots 7 and 8, and 15 and 16 are empty. Because memory mirroring is handled in hardware, it is operating system-independent. The total usable memory size is the sum of all DIMMs in Channel 0 or 1 (both channels must be populated identically).

Note: Memory must be identical across the two CPUs to enable the memory mirroring feature.

The memory population for mirrored mode in a single CPU configuration is shown in Table 4-32.

Table 4-32 Mirrored mode memory population table for one CPU dx360 M3 server

Number of DIMMs	Processor 1							
	Channel 0			Channel 1			Channel 2	
	DIMM 1	DIMM 2	DIMM 3	DIMM 4	DIMM 5	DIMM 6	DIMM 7	DIMM 8
2			x			x		
4		x	x		x	x		
6	x	x	x	x	x	x		

4.5.3 Internal storage and controllers

This section provides information about the internal storage options. It lists the conventional and solid-state disks that are supported. The next section is about the RAID controller cards and high IOPS PCIe adapters using flash memory technology. The last section provides technical details about the internal storage connections.

Hard disk drives

Table 4-33 lists the conventional disk drives that are supported with the dx360 M3 server in any of the 2U chassis configurations.

Table 4-33 dx360 M3 2U disk drives supported

Interface type	Simple Swap / Hot Swap	Interface speed	Disk size	Form factor	RPM	Feature
SATA II	Simple Swap	3 Gbps	250 GB	3.5-inch	7200	5292
SATA II	Simple Swap	3 Gbps	250 GB	3.5-inch	7200	A0WU
SATA II	Simple Swap	3 Gbps	500 GB	3.5-inch	7200	5288
SATA II	Simple Swap	3 Gbps	2 TB	3.5-inch	7200	5416
SATA II NL	Simple Swap	6 Gbps	250 GB	2.5-inch	7200	A1NY
SATA II NL	Simple Swap	6 Gbps	500 GB	2.5-inch	7200	A1P0
SATA II NL	Simple Swap	6 Gbps	1 TB	2.5-inch	7200	A1P2

Interface type	Simple Swap / Hot Swap	Interface speed	Disk size	Form factor	RPM	Feature
SATA II NL	Simple Swap	6 Gbps	3 TB	3.5-inch	7200	A280
SAS	Simple Swap	3 Gbps	300 GB	3.5-inch	15 K	5533
SAS	Hot Swap	3 Gbps	73 GB	2.5-inch	15 K	5543
SAS	Hot Swap	3 Gbps	146 GB	2.5-inch	10 K	5534
SAS	Simple Swap	6 Gbps	146 GB	2.5-inch	15 K	5426
SAS	Simple Swap	6 Gbps	300 GB	2.5-inch	10 K	5427
SAS	Simple Swap	6 Gbps	300 GB	2.5-inch	15 K	A24J
SAS	Simple Swap	6 Gbps	600 GB	2.5-inch	10 K	5435
SAS	Simple Swap	6 Gbps	900 GB	2.5-inch	10 K	A24H
SAS NL	Hot Swap	6 Gbps	500 GB	2.5-inch	7200	5410

Table 4-34 lists the disk drives that are supported in the dx360 M3 3U chassis configuration. In the 3U chassis, you have the choice of either a 3 Gbps or a 6 Gbps SAS backplane. All drives are hot-swappable.

Table 4-34 dx360 M3 3U disk drives supported

Interface type	Interface Speed	Disk size	Form factor	RPM	Feature
SATA II	3 Gbps	250 GB	3.5-inch	7200	5151
SATA II	3 Gbps	250 GB	3.5-inch	7200	A0WV
SATA II	3 Gbps	500 GB	3.5-inch	7200	5196
SATA II	3 Gbps	1 TB	3.5-inch	7200	5560
SATA II NL	3 Gbps	2 TB	3.5-inch	7200	5415
SATA II NL	6 Gbps	3 TB	3.5-inch	7200	A27Z
SAS	3 Gbps	300 GB	3.5-inch	15 K	5532
SAS	3 Gbps	450 GB	3.5-inch	15 K	5586
SAS NL	6 Gbps	1 TB	3.5-inch	7200	5418
SAS NL	6 Gbps	2 TB	3.5-inch	7200	5417
SAS NL	6 Gbps	3 TB	3.5-inch	7200	A281

Solid state drives

Solid state drives (SSD) provide extremely fast and reliable access to data when compared to traditional drives with rotating disks. They use flash memory to store each bit of information electronically instead of relying on the magnetic orientation on the disk.

SSDs have the advantage of no moving parts, which makes them more robust then conventional drives. Because of flash storage, no delay is imposed by any mechanical parts when accessing data. SSDs achieve very high I/O operations per second (IOPS) rates. This is especially true when accessing random data, where SSDs support a constant high throughput rate that is much higher than that of conventional drives.

To allow for an easy migration, most SSDs are available in the conventional 2.5-inch drive format. Table 4-35 lists SSDs that are supported in iDataPlex dx360 M3 servers.

Table 4-35 Solid State Drives (SSDs) supported in iDataPlex dx360 M3

Interface type	Carrier	Interface speed	Disk size	Form factor	Feature code
SATA	Simple swap	1.5 Gbps	31.4 GB	2.5-inch	3747
SATA II	Simple swap	3 Gbps	50 GB	2.5-inch	5197
SATA II	Simple swap	3 Gbps	128 GB	2.5-inch	A2UB
SATA II	Hot swap	3 Gbps	128 GB	2.5-inch	A2U4
SATA II	Simple swap	3 Gbps	200 GB	2.5-inch	5419
SATA II	Hot swap	3 Gbps	200 GB	2.5-inch	A2FN
SATA II	Simple swap	3 Gbps	256 GB	2.5-inch	A2UC
SATA II	Hot swap	3 Gbps	256 GB	2.5-inch	A2U3

Note: If you have multiple SSDs in your server, consider using the IBM 6 Gb Performance Optimized Host Bus Adapter. It is specifically built for high-throughput SSD environments, with the LSI Fusion-MPT architecture driving more than 290,000 I/Os per second.

High IOPS PCIe adapter

iDataPlex supports several PCIe adapters that are specifically designed for extremely high I/O rates using flash memory storage (Table 4-36). They can have up to 1.28 TB of NAND type flash memory onboard.

- For technical details of the high IOPS controllers, including max IOPS and transfer rates, see the *IBM High IOPS SSD PCIe Adapters at-a-glance guide* from IBM Redbooks:

  http://www.redbooks.ibm.com/abstracts/tips0729.html?Open

- For documentation and drivers of the high IOPS controllers, see this website:

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5083174

Table 4-36 PCIe SSD controller supported with the dx360 M3

Description	Size	PCIe Interface speed	Feature code
High IOPS SS Class SSD PCIe Adapter	160 GB	1.0a (Gen1) x4	3874
High IOPS MS Class SSD PCIe Adapter	320 GB	1.0a (Gen1) x4	1649
High IOPS SLC Adapter	320 GB	1.0a (Gen1) x4	A1NE
High IOPS SLC Duo Adapter	640 GB	2.0 (Gen2) x8	A1ND
High IOPS MLC Adapter	640 GB	1.1 (Gen1) x4	A1NC
High IOPS MLC Duo Adapter	640 GB	2.0 (Gen2) x8	5985
High IOPS MLC Duo Adapter	1.28 TB	2.0 (Gen2) x8	A1NB

Onboard disk controllers and RAID support

The onboard controller of the dx360 M3 is only for SATA drives and does not support RAID, as shown in Table 4-37. If you plan to use SAS disks, or you want to use hardware-based RAID, you must install a separate PCIe disk controller, as described in the next section.

Table 4-37 Onboard disk controllers of the dx360 M3 server

Controller feature	dx360 M3
SATA II support	Yes, 6-port
SATA port speed	3 Gbps
SAS support	No
Hardware-RAID	No

The dx360 M3 server has an onboard Mini PCIe x4 socket (Figure 4-30) for the ServeRAID BR10il RAID controller card. For standard PCIe RAID controllers, the dx360 M3 has the option of a 1U riser card that plugs into the x8 PCIe 2.0 storage riser card slot in the rear of the planar (see Figure 4-27 on page 83 for slot location in the system board layout, and Figure 4-30 for an installed PCIe card).

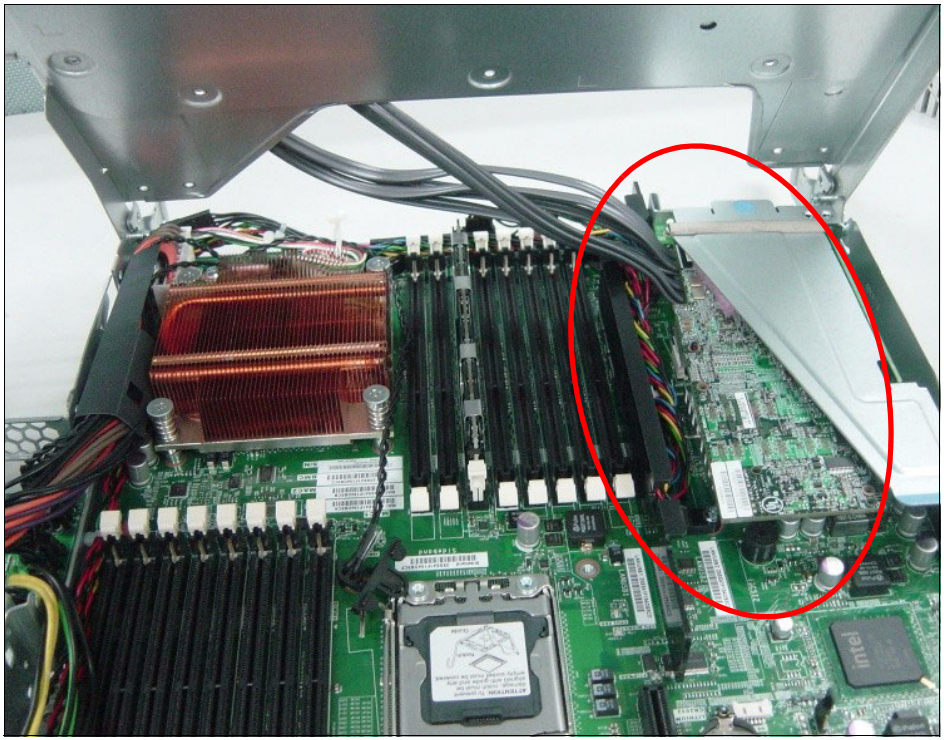


Figure 4-30 Storage riser card slot for RAID controller cards on the dx360 M3

Note: On the dx360 M3, the standard PCIe RAID controller cards do not occupy one of the front PCIe slots. The supported RAID controller cards (including the 6Gb Performance Optimized HBA, but not the ServeRAID M5025) plug into the dedicated storage riser card slot in the rear (see Figure 4-30).

Figure 4-31 shows how a ServeRAID controller card is installed into the storage riser card slot.

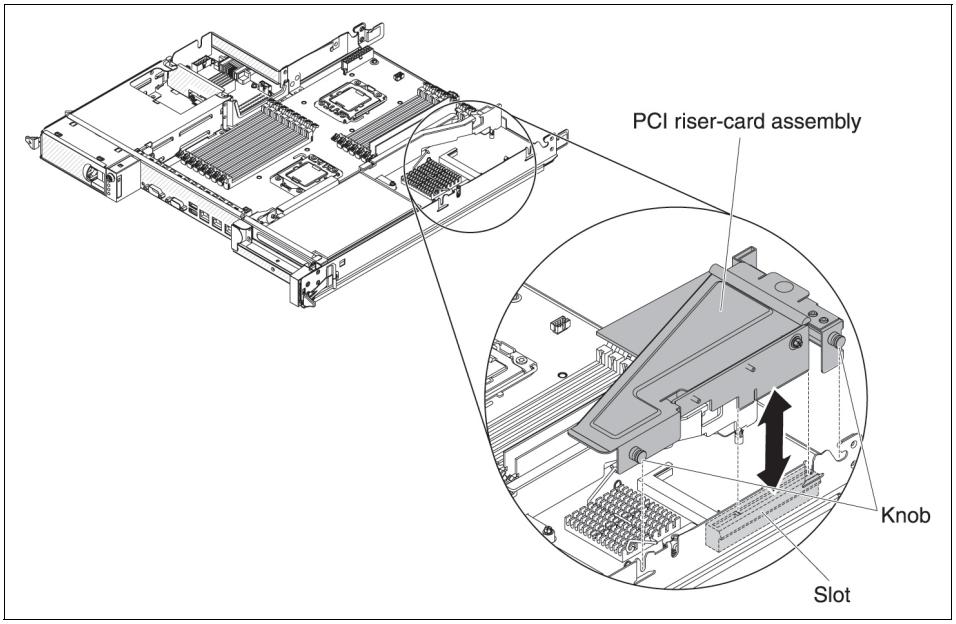


Figure 4-31 Installation of the storage riser card on the dx360 M3

PCI Express disk controllers and RAID support

You must install a separate disk controller in a PCI Express (PCIe) slot for any of the following reasons:

- SAS drives are used (because the onboard controller supports SATA only).
- More than six drives are used (because onboard controller has six SATA ports).
- Hardware RAID is required.
- Multiple SSD drives are used and I/O bandwidth is an important consideration.

Table 4-38 lists the supported RAID cards.

Table 4-38 Supported SAS/SATA RAID controller cards on the dx360 M3

Name	Chipset	Cache	Battery backup	Supported RAID levels	FC	Slot
3 Gbps SAS						
ServeRAID BR10il[a]	LSI 1064E	None	None	0, 1, and 1E	3588	Mini PCIe x4

Name	Chipset	Cache	Battery backup	Supported RAID levels	FC	Slot
6 Gbps SAS						
ServeRAID M5025[b] [e]	LSI SAS2108	512 MB	Standard	0, 1, 5, 10, 50; 6 and 60[d]	0094	PCIe 2.0 x8
ServeRAID M5015[e]	LSI SAS2108	512 MB	Optional[c]	0, 1, 5, 10, 50; 6 and 60[d]	0093	PCIe 2.0 x8
ServeRAID M5014[e]	LSI SAS2108	256 MB	Optional[c]	0, 1, 5, 10, 50; 6 and 60[d]	3877	PCIe 2.0 x8
ServeRAID M1015	LSI SAS2008	None	None	0, 1, 10; 5 and 50[f]	0095	PCIe 2.0 x8
6 Gb Performance Optimized HBA	LSI SAS2008	None	None	None	3876	PCIe 2.0 x8
6 Gb SAS HBA	LSI SAS2008	None	None	None	5982	PCIe 2.0 x8

a. The ServeRAID BR10il supports a maximum of four drives.
b. This adapter is for external storage only. It has no internal SAS connectors.
c. Battery backup is optional - feature 5744. No battery is included in FC 0093.
d. RAID levels 6 and 60 are supported with additional M5000 Advanced Feature Key, FC 5106.
e. Additional performance accelerator key is supported (FC A10C) to enable performance enhancements needed by modern SSD technologies. More information is available at: http://www.redbooks.ibm.com/abstracts/tips0799.html?Open
f. RAID levels 5 and 50 are supported with additional M1000 Advanced Feature Key, FC 9749.

For drivers and more information about the controllers, see these websites:

- Drivers and firmware for iDataPlex RAID controllers:

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5076022
- IBM ServeRAID BR10i documentation:

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5076481
- IBM ServeRAID M5015 documentation:

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5082936
- IBM ServeRAID M1015 documentation:

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5083720

For detailed specifications, including photos, see *ServeRAID Adapter Quick Reference*, TIPS0054:

http://www.redbooks.ibm.com/abstracts/tips0054.html

SAS connectivity

The next two figures show how each drive is cabled to the RAID controller cards. Figure 4-32 shows a 2U chassis with eight SAS small-form-factor drives, labeled SFF DRV 0 to SFF DRV 7.

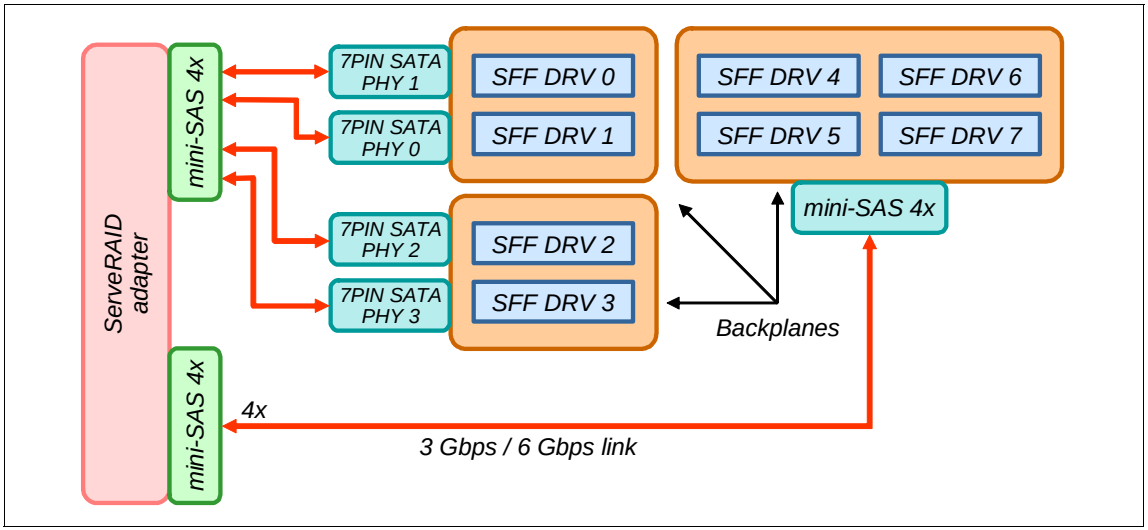


Figure 4-32 SAS wiring for 2U chassis configurations

The first four drives (DRV 0 to DRV 3) are each directly connected with a 1x SAS cable to the ServeRAID adapter. The second half of the drives (DRV 4 to DRV 7) merges in a single 4x SAS cable using the mini-SAS connector.

Note: With support for 6 Gbps RAID controller cards, IBM also introduced new SAS 2.0 backplanes supporting 6 Gbps connections. The chassis is automatically configured with the 6 Gbps backplane when any of the new RAID controller cards (capable of 6 Gbps) is selected.

The dx360 M3 server in the 3U chassis is only available with the new SAS 2.0 6 Gbps backplane.

Figure 4-33 shows the 3U chassis with 12 large-form-factor SAS drives, three of which are merged into one 3 Gbps (6 Gbps) SAS connection. The 4x SAS cable combines four of such triples, so that in total the 12 drives are connected through 12 Gbps of bandwidth (24 Gbps) to the controller card.

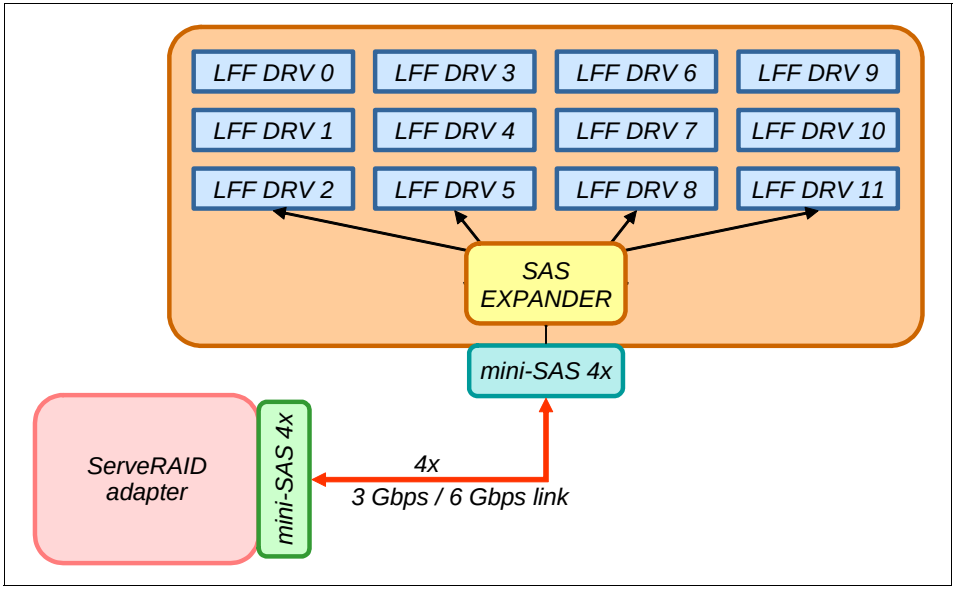


Figure 4-33 SAS wiring for 3U chassis configuration

4.5.4 PCI Express I/O adapters

The iDataPlex dx360 M3 server supports at least one PCIe card through the use of a riser card. Detailed PCIe features are listed in Table 4-39. Depending on the available height in the PCIe cage, you can install different riser cards.

Five different riser cards are available for iDataPlex:

- 1U single-slot for front PCIe slot: Used for installation of one PCIe card. Supported in all configurations.
- 2U two-slot for front PCIe slot: Used for installation of two PCIe cards. Any two adapters are supported, with a maximum of one GPU or GPGPU adapter.
- 2U three-slot for front PCIe slot: Used for installation of two GPU or GPGPU adapters and any other PCIe card. The third PCIe slot is on the backside of the riser card. The card has a PCIe switch onboard and requires a separate power cable. Supported only in 2U I/O-rich configurations using the PCIe tray.
- 1U single-slot for rear PCIe slot: Used for installation of any PCIe storage controller card. Supported in all 2U configurations.
- 2U single-slot for rear PCIe slot: Used for installation of any PCIe storage controller card. Supported in 3U configurations only.

Table 4-39 PCIe features of the iDataPlex dx360 M3 server

Feature	dx360 M3
PCIe slots maximum	3x PCIe 2.0[a] 1x PCIe 2.0 or 1x PCIe 1.0[b]
Slot 1, mechanically	x16
Slot 1, electrically	x16[c], 2.0
Slot 2, mechanically	x8
Slot 2, electrically	x8, 2.0
Slot 3, mechanically (for GPU and GPGPU only)	x16
Slot 3, electrically (for GPU and GPGPU only)	x16, 2.0
Mini PCIe, mechanically	x4
Mini PCIe, electrically	x4, 1.0
Rear storage riser slot, mechanically	x8
Rear storage riser slot, electrically	x8, 2.0

a. With PCIe tray using GPUs.
b. Depending on the RAID controller choice, the slot to be used will either be PCIe 1.0 or 2.0.
c. The first slot is x16 (electrically) when you have the one-slot or three-slot riser card installed. If you use the two-slot riser card, this slot is x8 only (electrically).

I/O option cards that are supported are listed in Table 4-40.

Table 4-40 Supported PCIe option cards on the iDataPlex dx360 M3 server

Description	PCIe Interface	Feature code
Fibre Channel		
QLogic 4 Gb single-port	1.0a x4	3567
QLogic 4 Gb dual-port	1.0a x4	3568
Emulex 4 Gb single-port	1.0a x4	1698
Emulex 4 Gb dual-port	1.0a x4	1699
Brocade 4 Gb single-port	1.0a x8	3885
Brocade 4 Gb dual-port	1.0a x8	3886

Description	PCIe Interface	Feature code
QLogic 8 Gb single-port	2.0 x4	3578
QLogic 8 Gb dual-port	2.0 x4	3579
Emulex 8 Gb single-port	2.0 x4	3580
Emulex 8 Gb dual-port	2.0 x4	3581
Brocade 8 Gb single-port	2.0 x8	3589
Brocade 8 Gb dual-port	2.0 x8	3591
iSCSI		
QLogic single-port HBA	1.0a x4	2976
QLogic dual-port HBA	1.0a x4	2977
SAS/SATA RAID controller		
IBM ServeRAID BR10il	1.1 x4	3588
IBM ServeRAID M1015	2.0 x8	0095
IBM ServeRAID M5014	2.0 x8	3877
IBM ServeRAID M5015	2.0 x8	0093
IBM ServeRAID M5025	2.0 x8	0094
IBM 6Gb SAS HBA	2.0 x8	5982
SSD		
IBM 6Gb Performance Optimized HBA	2.0 x8	3876
High IOPS SS Class SSD PCIe adapter, 160 GB	1.1 x4	3874
High IOPS MS Class SSD PCIe adapter, 320 GB	1.1 x4	1649
High IOPS SLC Adapter, 320 GB	1.0a x4	A1NE
High IOPS SLC Duo Adapter, 640 GB	2.0 x8	A1ND
High IOPS MLC Adapter, 640 GB	1.1 x4	A1NC
High IOPS MLC Duo Adapter, 640 GB	2.0 x8	5985
High IOPS MLC Duo Adapter, 1.28 TB	2.0 x8	A1NB

Description	PCIe Interface	Feature code
Ethernet / Converged Fabric		
Intel Pro/1000 PF 1-port Ethernet, 1 Gbps	1.0 x4	2975
Intel Pro/1000 PT 2-port Ethernet, 1 Gbps	1.0 x4	2944
Intel Pro/1000 PT 4-port Ethernet, 1 Gbps	1.0 x4	2974
NetXtreme II 1000 Express 1-port, 1 Gbps	1.1 x4	1485
NetXtreme II 1000 Express 2-port, 1 Gbps	2.0 x4	2995
NetXtreme II 1000 Express 4-port, 1 Gbps	2.0 x4	5766
Intel Ethernet 2-port I340-T2, 1 Gbps	2.0 x4	5767
Intel Ethernet 4-port I340-T4, 1 Gbps	2.0 x4	5768
Mellanox ConnectX-2 2-port adapter, 10 Gbps	2.0 x8	A1M4
Emulex Virtual Fabric adapter, 10 Gbps	2.0 x8	5749
Emulex Virtual Fabric adapter II, 10 Gbps	2.0 x8	A18Z
Brocade 2-port CNA, 10 Gbps	2.0 x8	1637
Broadcom NetXtreme II 2-port adapter, 10 Gbps	2.0 x8	A18Y
QLogic 2-port CNA, 10 Gbps	2.0 x8	5751
InfiniBand and 10 Gb Ethernet		
Mellanox ConnectX-2 VPI dual-port QSFP QDR IB / 10GbE HCA	2.0 x8	5447

4.5.5 Graphics adapter and GPGPUs

The dx360 M3 server supports the NVIDIA Quadro 4000 and Quadro 5000 PCIe graphics adapter. The Quadro 4000 has 256 CUDA[3] cores for parallel computation and dedicated 2 GB of GDDR5 memory onboard. The Quadro 5000 has 352 CUDA cores and dedicated 2.5 GB GDDR5 memory. Up to two of these graphics adapters can be installed in a 2U chassis with the dx360 M3 server.

[3] Compute Unified Device Architecture (CUDA) is an acceleration framework that helps boost application performance by providing an API that allows you to move compute-intensive workloads to Graphics Processing Units (GPUs).

When installing more than one PCIe graphics card, you need a 3-slot PCIe riser card that provides two PCIe 2.0 slots. The riser card also requires a dedicated power cable.

The dx360 M3 server can also host graphics adapters specifically built for general-purpose compute acceleration by the GPU. The dx360 M3 supports several different Tesla adapters based on the NVIDIA Fermi GPU and the AMD Firestream 9370 adapter. Table 4-41 lists the supported adapters.

Figure 4-34 shows the layout of a GPGPU based on NVIDIA's Fermi architecture. Fermi can run 16 kernels in parallel on a streaming multiprocessor (SM, illustrated by the red box in the figure). Each SM consists of 32 CUDA cores that run what is called a thread block. Each thread is scheduled on one CUDA core. Each CUDA core has a dedicated integer arithmetic logic unit (ALU) and a floating point unit (FPU).

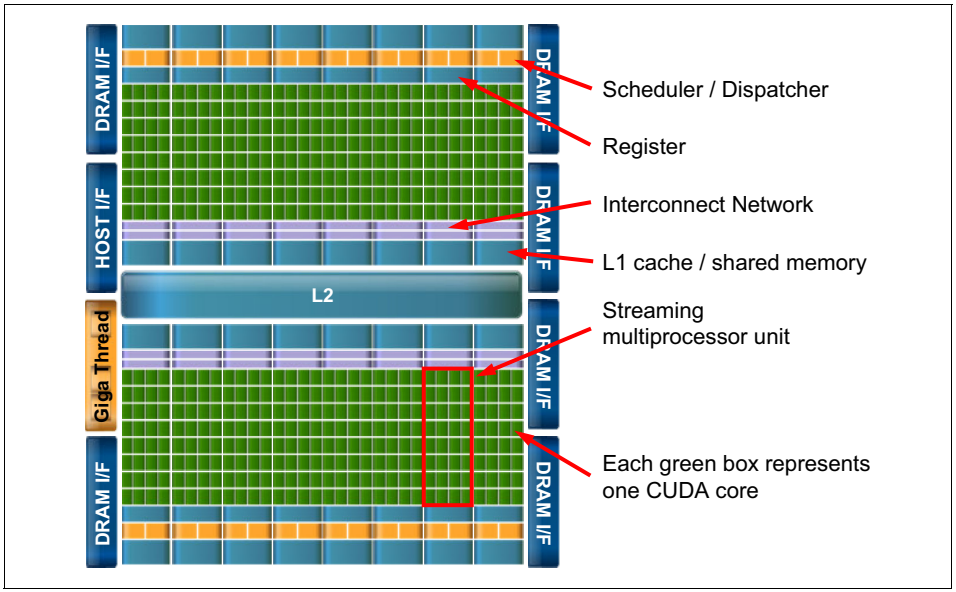


Figure 4-34 NVIDIA Fermi architecture

Detailed information about NVIDIA's Fermi architecture is available in a white paper at the following website:

http://www.nvidia.com/content/PDF/fermi_white_papers/NVIDIA_Fermi_Compute_Architecture_Whitepaper.pdf

All cards require an additional auxiliary power cable that allows the cards to draw more power directly from the 2U power supply. One cable is required for each card.

Table 4-41 Supported graphics adapter and GPGPU options on the dx360 M3

Feature code	Description	Interface
4798	NVIDIA Quadro 4000, 2 GB GDDR5	PCIe 2.0 x16
A1R3	NVIDIA Quadro 5000, 2.5 GB GDDR5	PCIe 2.0 x16
4796	NVIDIA Tesla M1060, 4 GB GDDR3	PCIe 2.0 x16
4797	NVIDIA Tesla M2050, 3 GB GDDR5	PCIe 2.0 x16
A104	NVIDIA Tesla M2070, 6 GB GDDR5	PCIe 2.0 x16
A105	NVIDIA Tesla M2070Q, 6 GB GDDR5	PCIe 2.0 x16
A243	NVIDIA Tesla M2075, 6 GB GDDR5	PCIe 2.0 x16
A1R4	NVIDIA Tesla M2090, 6 GB GDDR5	PCIe 2.0 x16
A160	AMD Firestream 9370, 4 GB GDDR5	PCIe 2.0 x16
3848	Power cable, 6pin-8pin	N/A
7471	Power cable, 8pin-8pin	N/A
4792	3-slot PCIe riser card	N/A
7469	Riser card, power cable	N/A

Up to two adapters can be installed in a 2U chassis with the PCIe tray using the 3-slot riser card. The riser card also requires an additional power cable. The installed cards do not have to be of the same type.

4.5.6 Management

Each dx360 M3 server has an Integrated Management Module (IMM) onboard and uses the Unified Extensible Firmware Interface (UEFI) to replace the older BIOS interface. The IMM is IPMI v2.0-compliant and provides the following major features:

- Remote configuration of IMM and UEFI settings without the need to power-on the server
- Remote access to system fan, voltage, and temperature values
- Remote IMM and UEFI update
- Remote console by way of a serial over LAN
- Remote presence (KVM and CD/DVD) and blue screen capture (requires additional virtual media key)

- Remote access to the system event log
- Predictive failure analysis and integrated alerting features (for example, by SNMP)

For more information and details about the IMM, see 7.1, “Integrated Management Module” on page 224.

To activate full systems management capabilities, an IBM Virtual Media Key (feature code 5891) is required. This small key is installed in a dedicated slot on the right side of the planar; see Figure 4-35. After you turn the server on again, the remote presence feature is automatically activated.

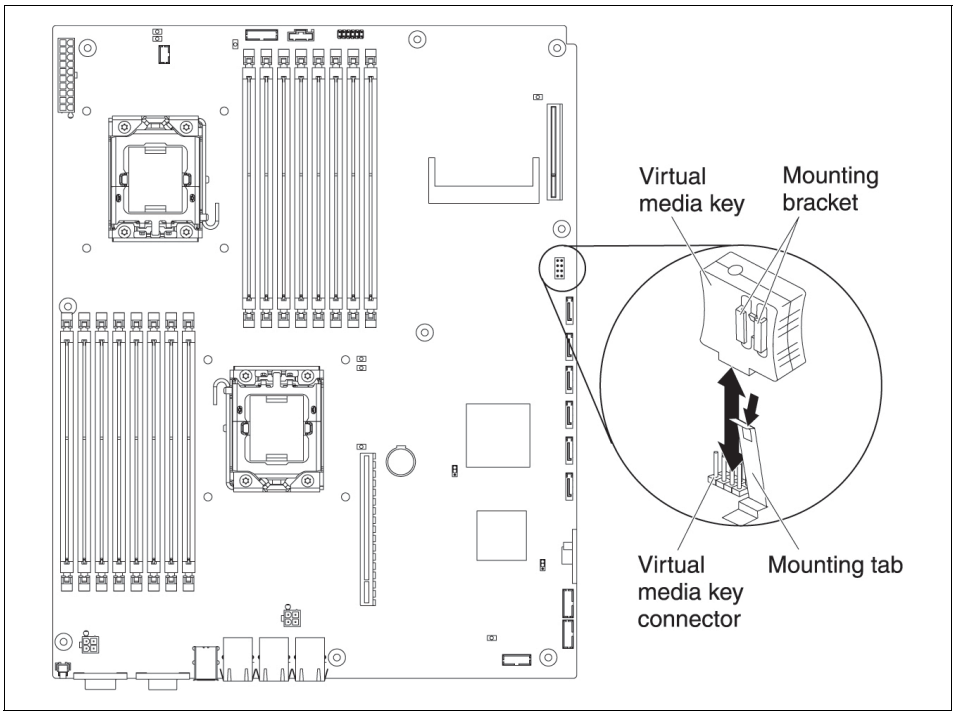


Figure 4-35 IBM Virtual Media Key on the dx360 M3 server

For more information about system management of iDataPlex servers, see Chapter 7, “Managing iDataPlex” on page 223.

4.5.7 PCIe tray for dx360 M3

In some configurations, only one iDataPlex server is installed in a 2U chassis. To fill the space, you can have additional storage or PCIe slots (or both).

The PCIe tray (feature code 4789) has two bays on the left for either two 3.5-inch disk drives (simple-swap SATA II or SAS) or four 2.5-inch disk drives (simple-swap SATA II or SAS) and can hold two double-wide PCIe graphics cards on the front right side of the chassis. Additionally, it can hold a third PCIe card (for example, Infiniband HCA) at the bottom right slot, below the graphics card. See Figure 4-36.

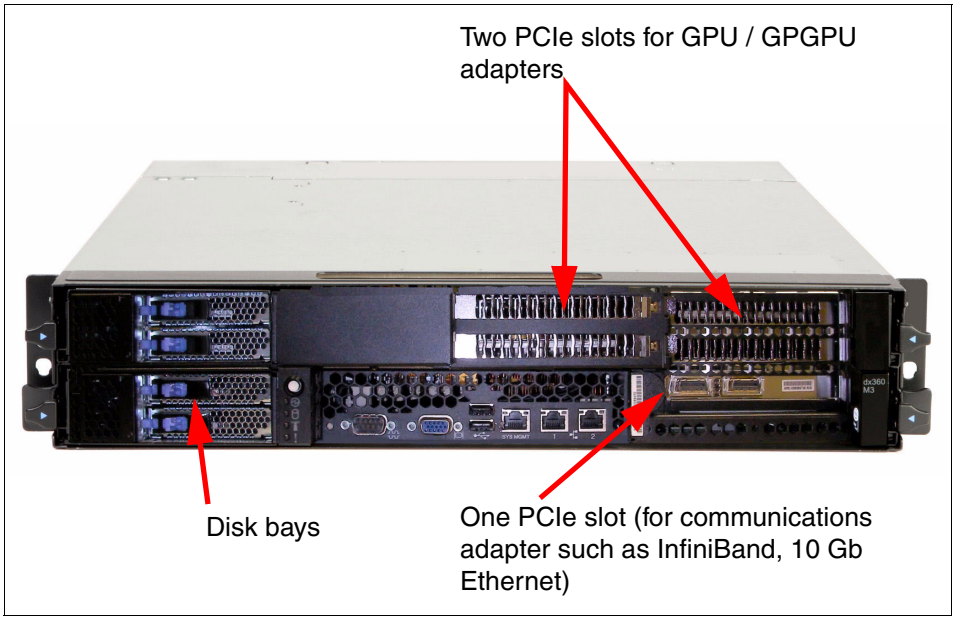


Figure 4-36 PCIe tray (with one server) showing the three front PCIe slots

4.5.8 HDD tray for dx360 M3

The HDD tray (feature code 4790) can hold either four 3.5-inch or eight 2.5-inch simple-swap SFF SATA II or SAS drives. One front PCIe card is supported in this configuration. See Figure 4-37.

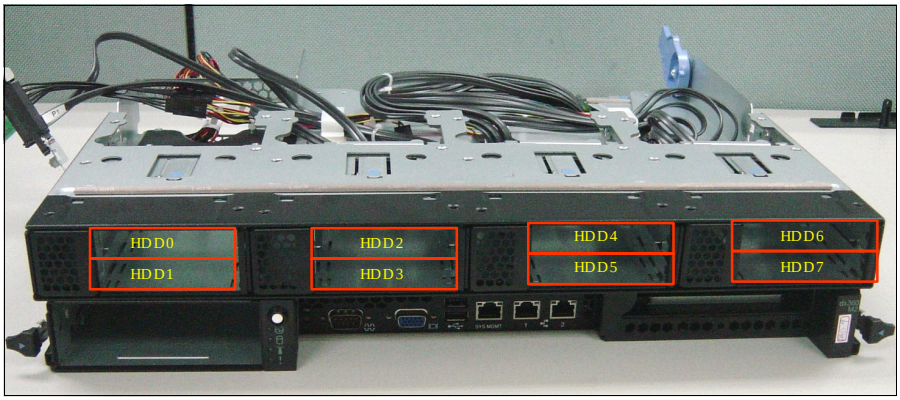


Figure 4-37 HDD tray (with one server) showing eight 2.5-inch drive setup

For more details about the supported 2U and 3U chassis configurations, see 4.5.10, “Chassis configurations” on page 110.

4.5.9 Operator panel for dx360 M3

The iDataPlex dx360 M3 server has an operator panel at the front. Figure 4-38 shows the power button and four status LEDs.

Tip: The power button has a protective collar around it to prevent accidental power-off of the server. You can remove this collar to more easily access the power button, if you prefer.

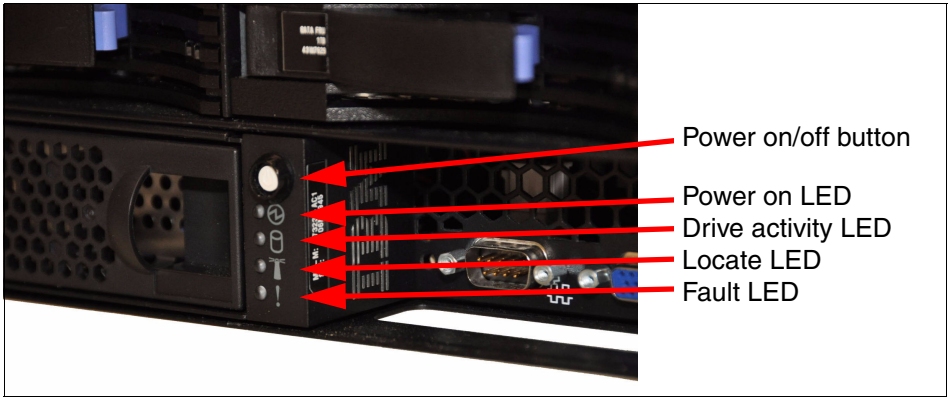


Figure 4-38 iDataPlex dx360 M3 operator panel

The LEDs are defined as follows:

- Power on (green): Determines if the server is powered on or off.
- Activity (green): Indicates hard disk activity.
- Locate (blue): Location LED controllable by the integrated IMM. This LED is useful to physically identify a system from the management software for maintenance purposes.
- Fault (amber): Indicates that a fault has been detected by the integrated IMM.

4.5.10 Chassis configurations

The basic component of an iDataPlex rack is the chassis that houses compute and storage elements. The chassis implements a Flex Node Technology that allows the same 2U chassis to serve different purposes.

The chassis configurations can be grouped three ways:

- Compute-intensive
- Storage-intensive (HDD tray)
- I/O-intensive (PCIe tray)

The chassis configurations differ in the number of available processors, disks, and PCIe slots.

Note: When talking about iDataPlex, we refer to a *node* as a single compute-instance. A 2U compute chassis with two servers installed equals two nodes; a 2U chassis with a server and a PCIe tray or storage tray equals a single node.

Table 4-42 provides an overview of the different configurations available.

Table 4-42 Available 2U and 3U chassis configurations

Node type	Number of nodes	Config name	Disk drive bays SS=simple-swap; HS=hot-swap			
			3.5-inch SATA bays	2.5-inch SATA bays	3.5-inch SAS bays	2.5-inch SAS bays
Compute-server (2U)	2	1A	2 SS	0	0	0
			0	0	2 SS	0
	2	1B	0	4 SS or 4 SSD	0	0
			0		0	4 SS
Storage server (2U)	1	2A	5 SS	0	0	0
			0	0	4 SS	0
	1	2B	0	8 SS or 8 SSD	0	0
					0	8 SS
I/O server (2U)	1	3A	2 SS	0	0	0
			0	0	2 SS	0
	1	3B		4 SS or 4 SSD		
			0	0	0	4 SS
Storage server (3U)	1	4A	12 HS	0	0	0
			0	0	12 HS	0

In addition to the 2U solutions, IBM offers a 3U chassis for storage-dense requirements. As listed in Table 4-42, the 3U storage chassis can be equipped with up to twelve 3.5-inch hot-swap SATA II or SAS drives, providing you with up to 36 TB of storage space. The value increases as larger capacity disks are supported.

In all configurations, the 2U or 3U chassis provides a single, shared power supply and a single fan assembly (more details are given in 4.6.1, “2U and 3U chassis” on page 121).

Table 4-43 lists the supported configurations with the dx360 M3 server. It includes details about the RAID controller options, the available PCIe slots, and the supported power supply for each configuration.

Table 4-43 Supported configurations with the dx360 M3 server

Config	Disk drive	Storage controller	Front PCIe slots (total / available)	Power supply options		
				550 W	Dual AC 750W	900W
1A	SATA, 3.5-inch	onboard	1 / 1	Yes	Yes	Yes
	SAS, 3.5-inch	BR10il	1 / 1	Yes	Yes	Yes
1B	SATA / SSD, 2.5-in	onboard, BR10il, 6Gb Performance Optimized HBA	1 / 1	Yes	Yes	Yes
	SAS, 2.5-inch	BR10il, M1015, SAS HBA	1 / 1	Yes	Yes	Yes
2A	SATA, 3.5-inch	onboard[a]	1 / 1	Yes	Yes	Yes
	SAS, 3.5-inch	BR10il, SAS HBA, M1015, M5014, or M5015	1 / 1	Yes	Yes	Yes
2B	SATA / SSD, 2.5-in	M1015, M5014, M5015, 6Gb Performance Optimized HBA	1 / 1	Yes	Yes	Yes
	SAS, 2.5-in	SAS HBA, M1015, M5014, or M5015	1 / 1	Yes	Yes	Yes
3A	SATA, 3.5-inch	onboard, BR10il	3 / 3	No	Yes	Yes
	SAS, 3.5-inch	BR10il, SAS HBA, or M1015	3 / 3	No	Yes	Yes
3B	SATA / SSD, 2.5-in	M1015, 6Gb Performance Optimized HBA	3 / 3	No	Yes	Yes
	SAS, 2.5-inch	SAS HBA or M1015	3 / 3	No	Yes	Yes
4A	SATA, 3.5-inch	M1015, M5014, or M5015	1 / 1	No	Yes	Yes
	SAS, 3.5-inch	M1015, M5014, or M5015	1 / 1	No	Yes	Yes

a. BR10il supported with a maximum of four SATA drives in this configuration. The fifth bay must remain empty.

Compute-intensive server (1A and 1B)

The compute-node solution contains two identical 1U servers with either a 3.5-inch SATA/SAS (configuration 1A) or two 2.5-inch SATA/SSD/SAS drives (configuration 1B) each. Each node also has one PCIe slot at the front. See Table 4-42 on page 111 for a comparison of the configuration types.

Figure 4-39 shows the 3.5-inch drive configuration, which can be either SAS or SATA. The 3.5-inch drives in each node must be of the same configuration (including disk size). These 3.5-inch drives use a simple-swap drive tray.

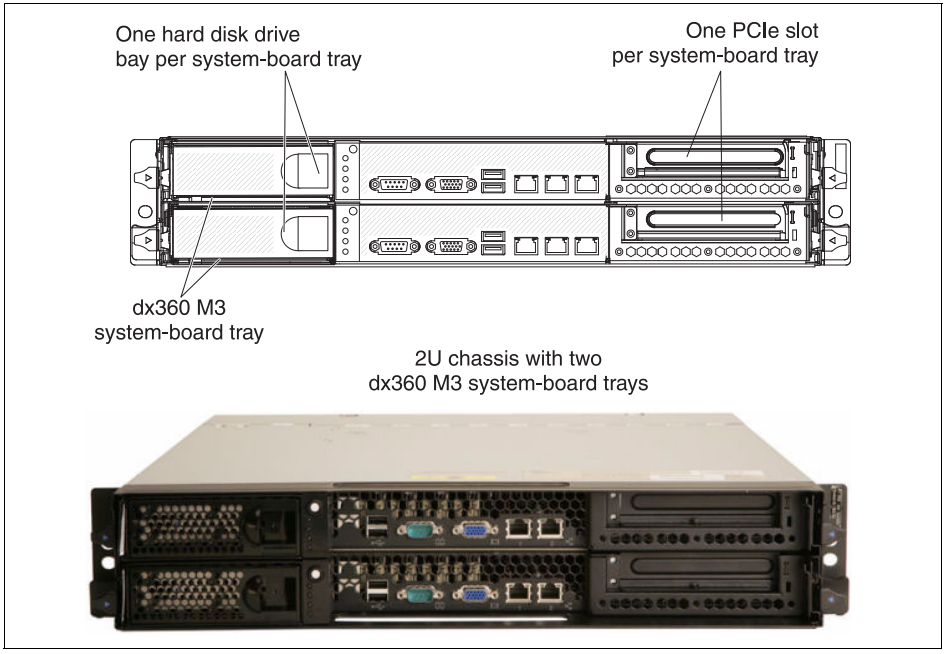


Figure 4-39 2U chassis with two servers and one 3.5-inch drive each; configuration 1A

Figure 4-40 shows the configuration with two 2.5-inch hot-swap SAS disks in each node.

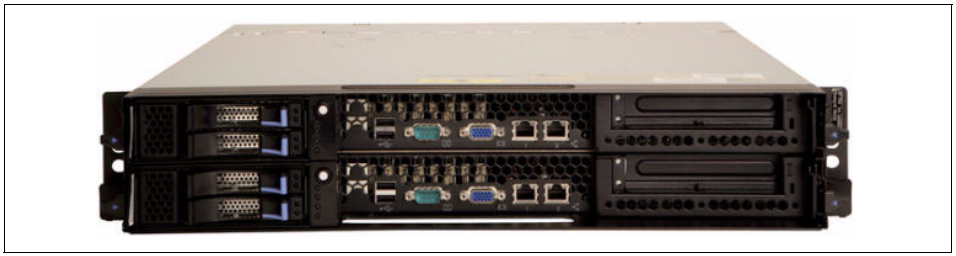


Figure 4-40 2U chassis with two servers and two 2.5-inch SAS drives each; configuration 1B

If you use SAS drives (either 3.5-inch drives or 2.5-inch drives), you must also have a SAS disk controller in the Mini PCIe slot.

If you use SSD drives with the dx360 M3 server, depending on your performance needs, you might want to install the 6Gb Performance Optimized HBA in the rear PCIe slot of the planar.

Storage server (2A and 2B)

For storage-intensive applications, IBM offers configurations that have a larger number of high-capacity disk drives. This configuration is sometimes referred to as a node with an attached storage tray. The dx360 M3 server uses the so-called HDD tray.

Storage configurations can have either of these combinations:

- Five 3.5-inch SATA drives or four 3.5-inch SAS drives (configuration 2A)
- Eight 2.5-inch SATA, SAS, or solid-state drives (configuration 2B)

Figure 4-41 shows the SATA configuration 2A. The installation of the ServeRAID BR10il controller card is supported in this configuration. However, it limits the number of drives to four because these simple-swap drive trays are directly connected to the RAID card, and there are only four connectors. When using SAS drives in configuration 2A, you must install an additional RAID controller card.

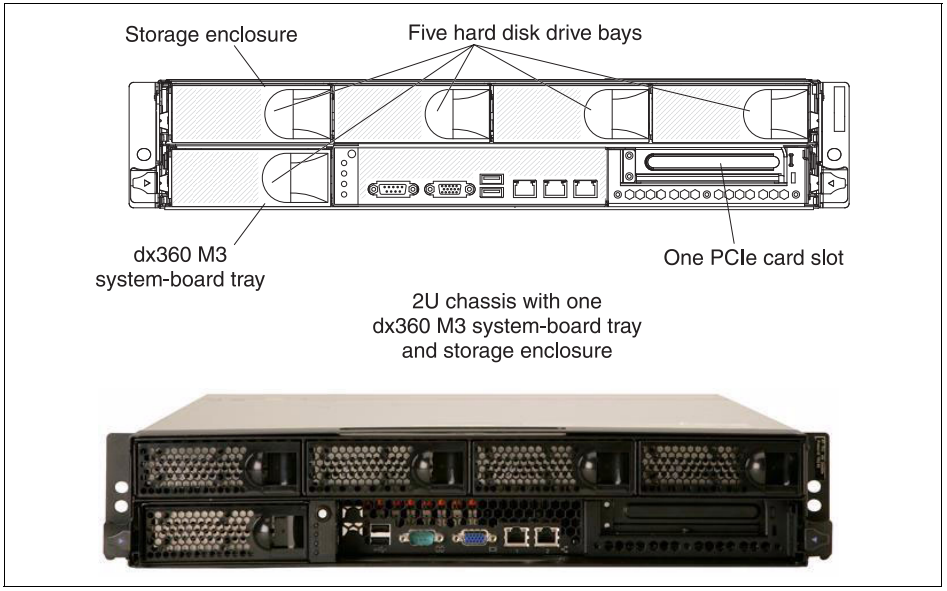


Figure 4-41 2U chassis with one server and five SATA drives; configuration 2A

Configuration 2B is available with the dx360 M3 server including the HDD tray and a backplane supporting 6 Gbps SAS speed. This configuration allows you to install up to eight 2.5-inch drives, either SATA, SAS, or solid-state drives. If the chassis includes a significant amount of SSDs, you might want to install the 6 Gb Performance Optimized host bus adapter, which is specifically built to support the high IOPS generated by modern SSDs.

I/O-intensive server (3A and 3B)

For I/O intensive applications, IBM offers configurations that can host up to three PCIe cards. This configuration is sometimes referred to as a node with an attached I/O tray.

A riser card that has up to three PCIe slots is inserted into the system board. For more information about the PCIe slots and the I/O options, see 4.6, “iDataPlex FlexNode chassis components” on page 120.

I/O configurations can have either of these combinations:

- Two 3.5-inch SATA or SAS drives (configuration 3A)
- Four 2.5-inch SATA, SAS, or solid-state drives (configuration 3B)

For detailed information about how many PCIe slots are available in each of the configurations, see Table 4-43 on page 112.

Figure 4-42 shows a 2U Flex chassis with two PCIe slots and two 3.5-inch disk drive bays.

Note: Figure 4-42 shows three PCIe slot covers; however, only two slots are enabled in the server. The bottom slot is not used.

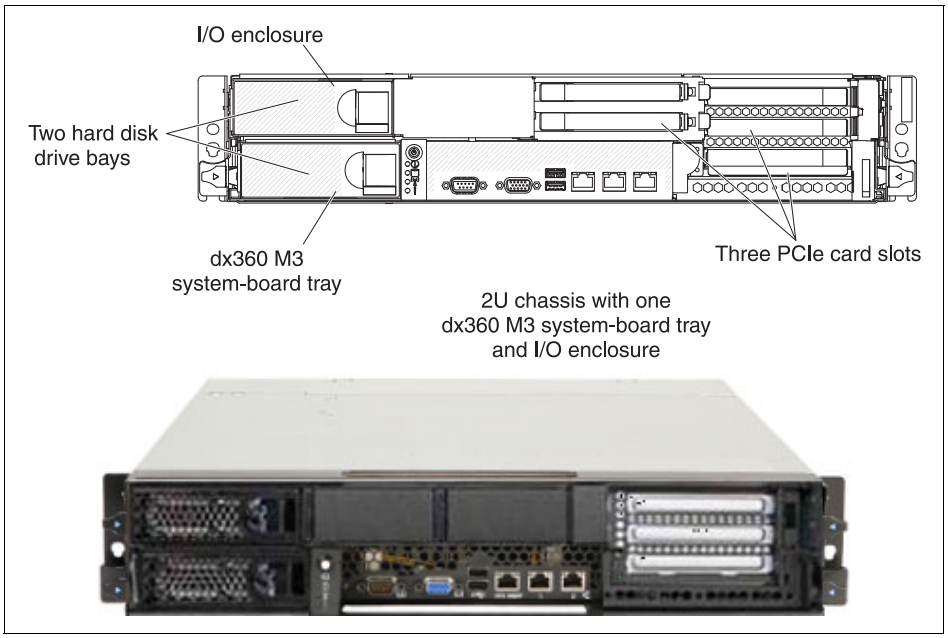


Figure 4-42 2U chassis with one server and I/O tray (two 3.5-inch drives); config 3A

When 3.5-inch drives are used in a configuration with two PCIe slots, you can only install two drives. The space in the upper middle section of the chassis cannot be used to hold drives. This configuration is referred to as 3A in Table 4-42 on page 111.

In Configuration 3B, the chassis can hold up to four 2.5-inch SATA, SAS, or solid-state drives and still provide space for up to three PCIe cards. Figure 4-43 shows such a configuration, which holds four 2.5-inch SAS drives and three PCIe slots, two slots on the right side of the chassis, and one slot above the I/O panel of the server.

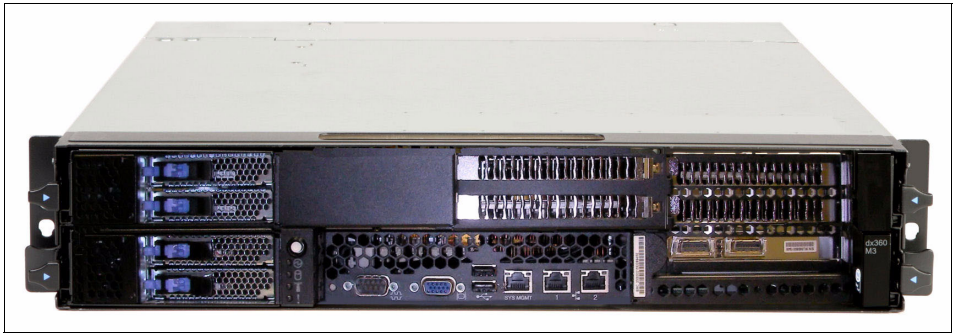


Figure 4-43 2U chassis with a server and HDD tray (four 2.5-inch SAS drives); config 3B

3U chassis configurations (4A)

For high-density storage requirements, IBM offers a 3U chassis that can contain up to twelve 3.5-inch hot-swap drives, either SATA II or SAS.

The SAS/SATA RAID controller card goes into the storage riser card slot in the back of the chassis. The front PCIe slot is available for other cards.

Figure 4-44 shows the rear half of the dx360 M3 server with the storage riser card slot located on the right side.

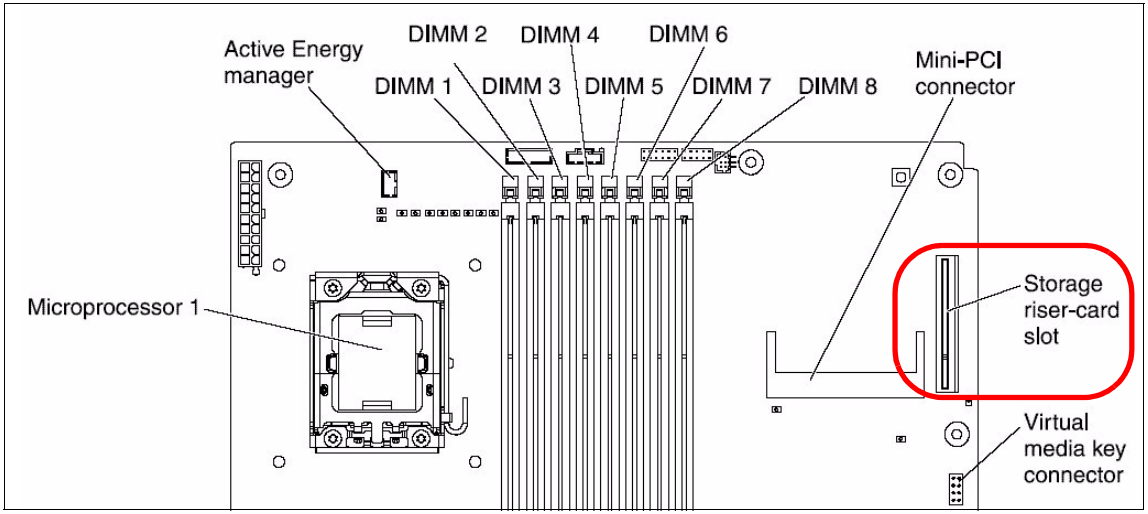


Figure 4-44 dx360 M3 system board layout showing the storage riser card slot

Figure 4-45 shows a 3U chassis with the dx360 M3. The ServeRAID M1015, M5014, or M5015 adapters can be installed in the storage riser card slot and protrudes through the slot of the dx360 M3 cover into the body of the 3U chassis.

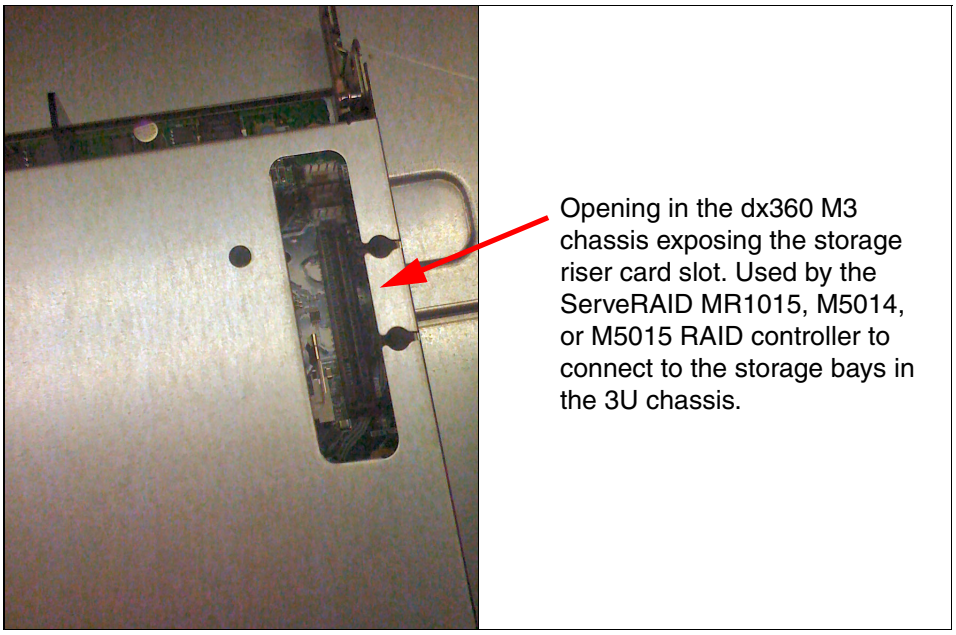


Figure 4-45 Opening in the dx360 M3 chassis when installed in a 3U chassis

Figure 4-46 shows the 3U chassis with twelve disks. Unlike the 3.5-inch drives available for the 2U Flex chassis, the 3.5-inch drives in the 3U chassis are hot-swap drives. These drive options differ from those supported in the 2U Flex chassis.

Note: The hot-swap backplane provides an x4 connection to the SAS controller card, which results in 24 Gbps of bandwidth when using the 6 Gbps SAS backplane with the dx360 M3 server.

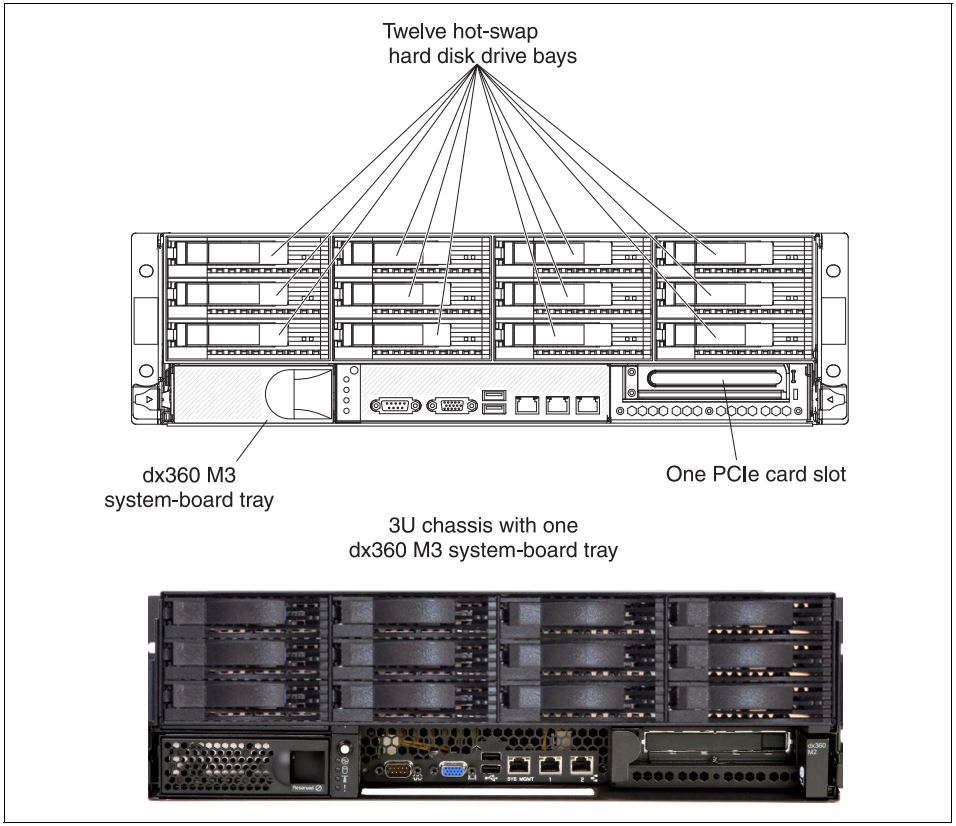


Figure 4-46 3U chassis with one server and twelve hot-swap drives

4.5.11 General hardware configuration rules

This section briefly outlines the major configuration rules that apply when putting together a new iDataPlex configuration.

- For a 2U-compute solution, both servers must have matching CPUs, memory configurations, and disks (including disk size). Different node configurations within a single 2U chassis are not supported.
- You cannot mix SATA and SAS in a single chassis.
- If hardware RAID is required, a SAS/SATA RAID controller card must be installed into an available PCIe slot (either Mini PCIe or rear storage PCIe slot).
- SAS configurations require a SAS controller card in the PCIe slot (depending on the configuration, either in the Mini PCIe slot or the storage riser card slot). The onboard disk controller of the current servers supports SATA only.

- If more than six drives are used (SATA or SAS), a SAS/SATA controller must be installed in the PCIe slot (depending on the server and configuration, either the Mini PCIe or rear storage PCIe slot).
- If you are using SSDs and performance is crucial, consider installing the 6Gb Performance Optimized host bus adapter.

4.5.12 Supported operating systems

At the time of writing, the following operating systems are supported:

- Microsoft:
  - Windows Server 2008/2008 R2, Web x64 Edition
  - Windows Server 2008/2008 R2, Enterprise x64 Edition
  - Windows Server 2008/2008 R2, Datacenter x64 Edition
  - Windows HPC Server 2008
- Linux:
  - Red Hat Enterprise Linux 4 ES for AMD64/EM64T
  - Red Hat Enterprise Linux 5 Server x64 Edition
  - Red Hat Enterprise Linux 6Server x64 Edition
  - SuSE Linux Enterprise Server 10 for AMD64/EM64T
  - SuSE Linux Enterprise Server 11 for AMD64/EM64T
- VMware:
  - VMware ESX/ESXi 3.5
  - VMware ESX/ESXi 4.0
  - VMware vSphere 5

4.6 iDataPlex FlexNode chassis components

This section describes the chassis components, which are a 2U and a 3U chassis, the Direct Dock Power feature, the chassis power supplies, and the chassis fan pack. Two chassis sizes provide a variety of configuration options, depending on the customer requirements.

For the dx360 M3 server, the 2U FlexNode chassis can be configured as follows:

- Two 1U compute nodes, each with a PCIe slot at the front and a single drive bay (either SAS, SATA, or SDD)
- One compute node, with a PCIe slot and multiple drive bays (either SAS, SATA, or SDD)
- One compute node with up to three additional PCIe slots (two supported for GPGPU adapters) and a single drive bay (SAS, SATA, or SSD)

The 3U chassis offers a compute node plus 12 drive bays for high-capacity disk drives, either SAS or SATA, currently supporting up to 36 TB of internal storage.

When using the dx360 M4 server, the 2U FlexNode chassis can be configured as follows:

- Two 1U compute nodes, each with two PCIe slots at the front, one mezzanine PCIe slot, and a single drive bay (either SAS, SATA, or SSD)
- One compute node, with four PCIe slots at the front (two of which can be populated with GPGPU cards only), one mezzanine PCIe slot, and a single drive bay (either SAS, SATA, or SSD)

These chassis and node configurations can be placed in the iDataPlex rack in any order, in two columns of 42U each, as described in 5.1.1, “The iDataPlex rack” on page 134.

The following topics are covered in this section:

4.6.1 2U and 3U chassis

The 2U and the 3U chassis are mechanical enclosures that provide shared power and cooling capabilities to the devices mounted in it. The built-in power supply provides the chassis with sufficient power to drive the installed components.

If either the power supply or the fan assembly require servicing, the chassis must first be removed from the rack and the top cover removed from the chassis as shown in Figure 4-47.

Figure 4-47 Chassis with shared fan assembly (left) and power supply (right), with rear cover removed

2U Flex chassis for dx360 M3

The 2U Flex chassis, available as machine type 6385 (three-year warranty), is designed to hold either two servers, or one server with multiple drives or a second and third PCIe slot, or both. For compute-intensive applications, the chassis can be populated with two servers and up to two disks for each server. For storage and I/O-bound applications, the chassis allows several combinations of SATA II or SAS drives and PCIe slots. For a more detailed list of possible configurations, see 4.5.10, “Chassis configurations” on page 110. Figure 4-48 shows an empty 2U chassis.

Figure 4-48 An empty 2U Flex chassis

2U chassis for dx360 M4

To allow for improved power supplies, the dx360 M4 server requires a new 2U Flex chassis, available as machine type 7913 (three-year warranty). This chassis has the same features and dimensions as the 2U Flex chassis for the dx360 M3 server but a different notch that supports the installation of two half-height power supplies.

3U chassis (dx360 M3 only)

The 3U chassis, available as machine type 6386 (three-year warranty), is specifically designed to meet high storage requirements. It uses the same power supplies as the 2U Flex chassis, but has a different chassis fan assembly. The 3U chassis allows for up to 12 high-capacity hot-swap drives, either 3.5-inch SATA II or 3.5-inch SAS (including SAS 2.0) drives with support for a variety of RAID levels. See Figure 4-49.

Figure 4-49 A dx360 M3 and 12 3.5-inch drive in a 3U chassis

Direct Water Cooled dx360 M4 2U chassis

The Direct Water Cooled dx360 M4 2U chassis, available as machine type 7919 (three-year warranty), is based on the air-cooled 2U chassis and has the same physical dimensions. Instead of fans and perforations in the rear, it has a solid plate with space for two pipes to connect the external cooling loop to the water loop assembly on the server. One 2U chassis has four such openings, a separate inlet and outlet opening for each server. See Figure 4-50.

Figure 4-50 Direct Water Cooled dx360 M4 2U chassis (rear view)

4.6.2 Direct Dock Power

The iDataPlex uses Direct Dock Power to power the nodes in the chassis. Industry standard power cords power each node, but the cords are attached to the rack in a fixed location. When you slide in the chassis, the power receptacle of the chassis simply connects to the power cord, which means that you do not have to access the rear of the rack to attach the power cord. Figure 4-51 shows the connectors associated with powering the chassis.

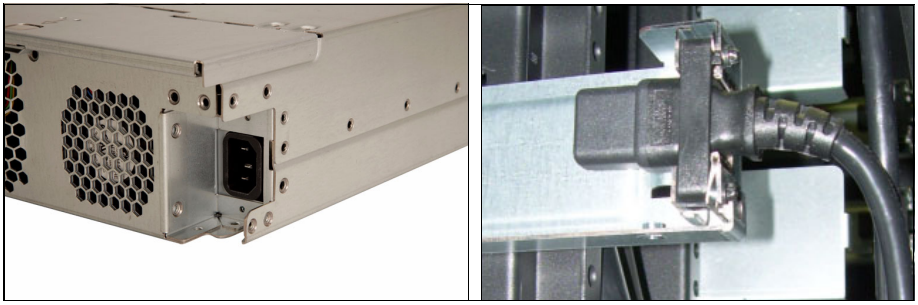


Figure 4-51 Direct Dock Power allows the chassis to be inserted without having to connect power cables

With the innovation of Direct Dock Power, you can simply push the chassis into the rack, and the power supply will connect to the PDU. You do not have to access the rear of the rack when installing or working with servers.

The 750 W dual AC power supply relies on the same principle of Direct Dock Power, but uses a different power cord with more pins, as shown in Figure 4-52.

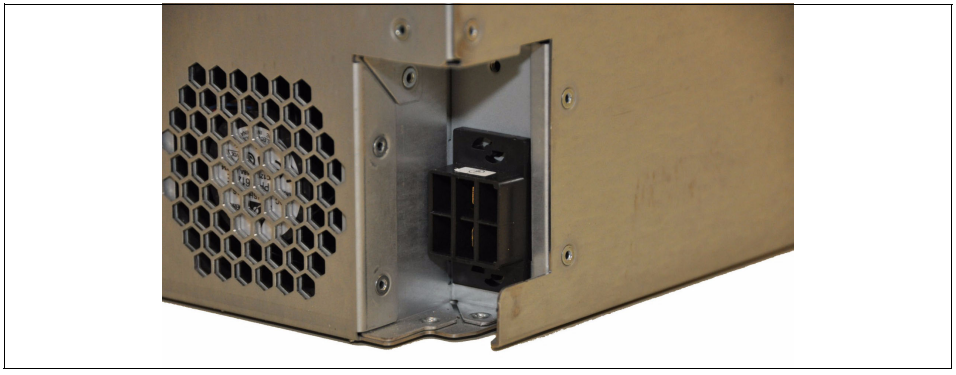


Figure 4-52 Connector of the 750 W dual AC power supply

4.6.3 Chassis power supply, power cable, and fans

Each iDataPlex chassis provides a shared high-efficiency power supply and low power-consuming fans, similar in concept to a BladeCenter chassis. In the iDataPlex 2U and 3U chassis, the compute node, and all available trays (I/O, storage, HDD, and PCIe) share one power supply and one fan unit.

Figure 4-53 shows the power supply and fan unit for the 2U Flex Node chassis.

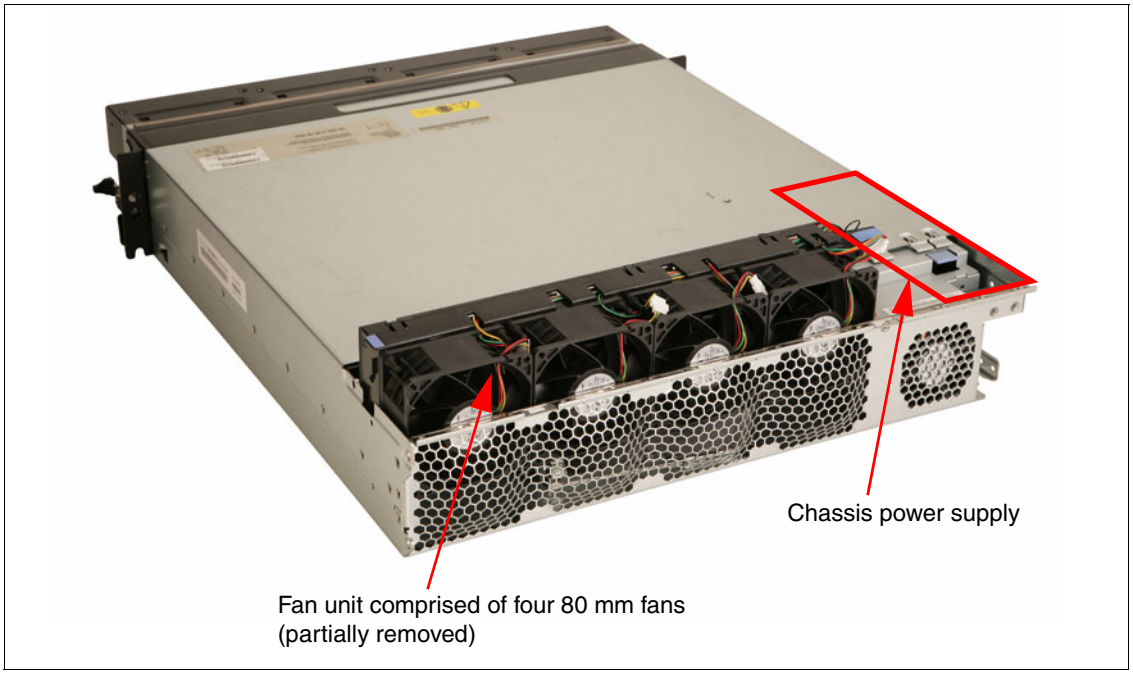


Figure 4-53 The chassis provides efficient power and cooling to the one or two nodes installed

The power supply used in each iDataPlex chassis can be more than 90% efficient depending on the load, and consume 40% less power when compared to a traditional 1U server.

The three power supply options for the 2U Flex chassis are listed in Table 4-44.

Table 4-44 Power supply options

Power supply	Feature code
HE550W 550 W power supply	5746
HE900W 900 W power supply	4779
750 W dual AC power supply	4778
550 W HE Platinum AC power supply	A22M
750 W HE Platinum AC power supply	A1G8
900 W HE Platinum AC power supply	A1G7

Note: The new High Efficiency (HE) Platinum AC power supplies are 80 PLUS Platinum certified to allow for the best efficiency values of your data center.

The new HE Platinum AC power supplies are half the size of the non-Platinum power supplies. This allows for two power supplies to be installed in one 2U chassis and enables redundancy at the power supply level. Figure 4-54 shows the new 900W HE Platinum power supply when installed in a 2U chassis.

Figure 4-54 900W HE Platinum AC power supply (feature A1G7)

Figure 4-55 shows the HE900W 900 W power supply. The HE550W 550 W power supply looks identical to the HE900W. Using the HE550W, when the configuration allows it, means that the power supply will be more heavily loaded and therefore be more efficient and ultimately save energy.

The configurations that support the HE550W are listed in Table 4-45 on page 128. If you prefer, you can elect to use the HE900W in these configurations instead. In general, the HE550W supply can only be used in 2U configurations.

Figure 4-55 HE900W 900 W power supply (feature 4779)

The 750 W dual AC power supply protects against the loss of a PDU or its feeding AC circuit. Therefore, make sure to connect the two power cords from the power supply to two different PDUs fed by two different power circuits.

The 750 W dual AC power supply consists of a power chassis with two separate 750 W power domains that are bussed together inside the power supply. Under normal operation, that is in redundant mode with both circuits providing power, the power supply provides up to 900 W total. If one of the 750 W domains fail, the system reduces load to 750 W by a feedback channel towards the power circuitry on the servers. Figure 4-56 shows how the power supply is fed by two independent AC sources.

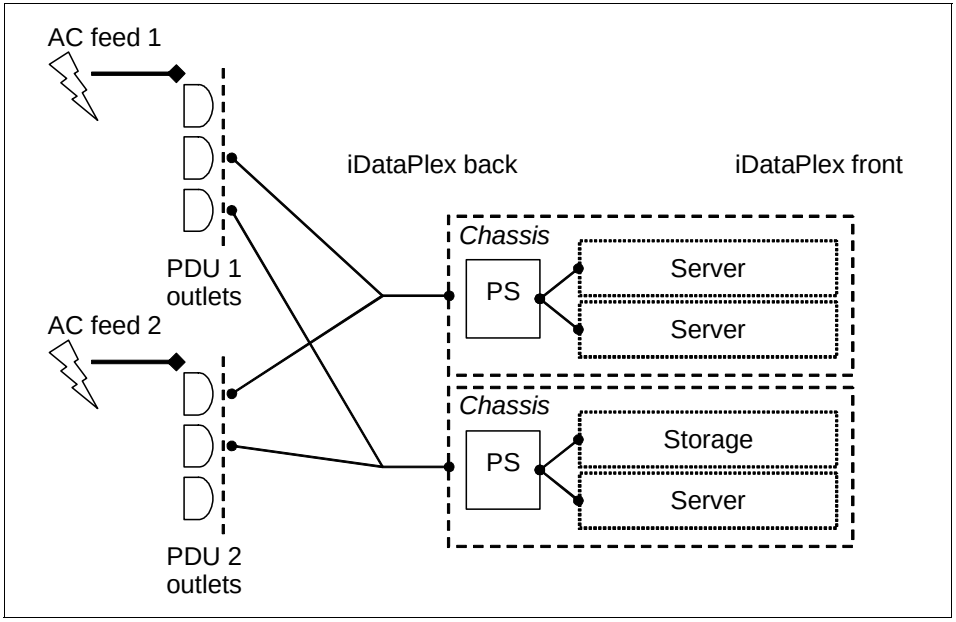


Figure 4-56 Power cabling for the 750 W dual AC power supply

Table 4-45 lists the power supplies and the servers they are each supported in.

Table 4-45 Compatible power supplies on the iDataPlex servers

	HE550W 550 W	**HE900W 900 W**	**750 W dual AC**	**550W HE Platinum**	**750 W HE Platinum**	**900 W HE Platinum**
dx360 M3	Yes	Yes	Yes	No	No	No
dx360 M4 air-cooled	No	No	No	Yes	Yes	Yes
dx360 M4 water-cooled	No	No	No	No	No	Yes

The three power cables available to connect the power supplies to the rack PDU are listed in Table 4-46. The first cable is a Y-cable, meaning it connects one PDU outlet to two chassis power supplies. The second cable is a straight-through cable, connecting one PDU outlet to one chassis power supply. The third cable is specifically designed to provide a redundant connection to the dual AC 750 W power supply from two PDU outlets.

Table 4-46 Power cable options

Power cable	Feature code
1.8m, 10A/100-250V, 2xC13PM to IEC 320-C14 (Y-cable)	6568
1.8m, 10A/100-250V, C13PM to IEC 320-C14	6569
0.8m, 10A/250V, proprietary to 2x IEC 320-C14 (dual AC inlet cable)	6293
AC Inlet 12A Cable, for all HE Platinum power supplies only (dual AC inlet cable)	A1V3
AC Inlet Y 12A Cable, for all HE Platinum power supplies only	A1V4

Figure 4-57 shows the connectivity for feature code 6568.

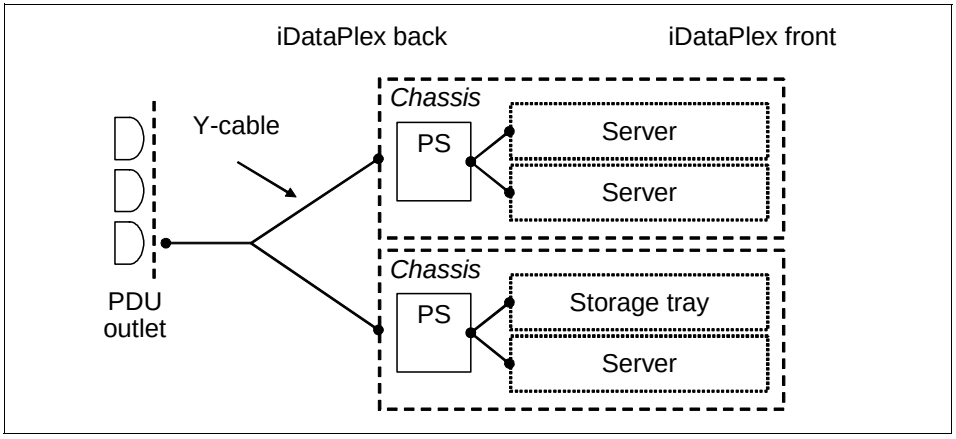


Figure 4-57 Y-cable for power connectivity, feature code 6568

Figure 4-58 shows a picture of a Y-cable (feature code A1V4) supporting the new HE Platinum power supplies.

Figure 4-58 Y-cable for HE Platinum power supplies (feature code A1V4)

To reduce the number of required PDU outlets when using the 750 W dual AC power supply, it is possible to plug the dual AC inlet cable (FC 6293) into two separate Y-cables (FC 6568). See Figure 4-59 for this solution.

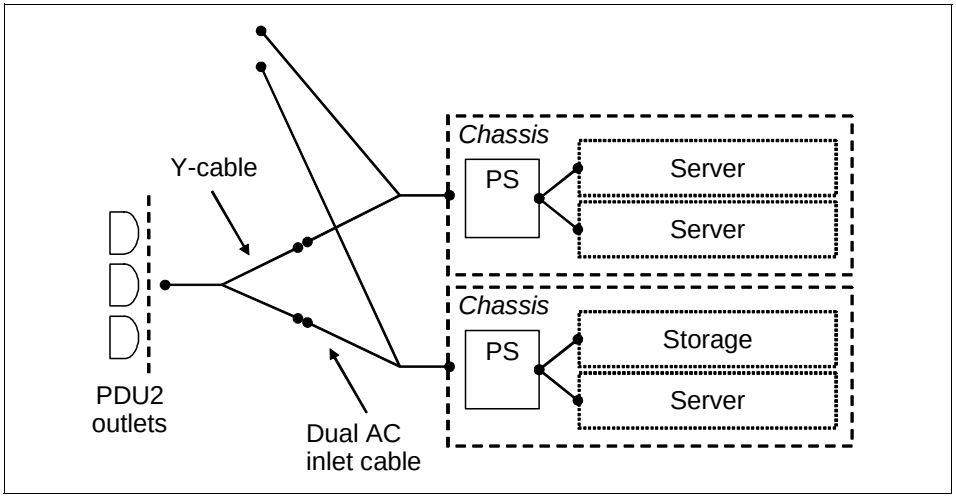


Figure 4-59 Using Y-cables to consolidate dual AC inlet cable receptacle

Note: Make sure that you do not exceed the maximum power drawing of a single PDU outlet in Figure 4-59.

Because the fans are shared between two nodes in a Flex Node chassis, they can be two times the diameter of those in traditional 1U servers (80 mm in iDataPlex compared to 40 mm in 1U servers). This, in combination with less pressure because of the shorter distance the air has to travel across the server, makes one of the most efficient air cooling solutions on the market. As a result, the fans are about 70% more efficient in terms of power consumption, compared to those in a traditional 1U server.

FAQ:

- What happens when a single fan fails?

  Depending on the workload and environment, it can have little impact and the server would continue running. The fan speed control algorithm takes fan failures into account and only increases fan speed when component temperatures hit a certain threshold. For high demand workload, the system might throttle but continue operation. In certain cases, if cooling is no longer sufficient enough, the system performs a soft shutdown to prevent hardware damage.

5

iDataPlex rack, networking, and storage

This chapter provides information about the iDataPlex rack, the networking options, and the storage options.

The following topics are covered in this chapter:

5.1 iDataPlex rack components

The major building block for a new data center design based on iDataPlex is the iDataPlex rack. In contrast to standard 19-inch racks, the iDataPlex rack is designed for higher density without having to sacrifice your existing data center layout. As with any other rack, the iDataPlex rack continues to provide cabling, power, and cooling capabilities, but the iDataPlex rack provides these capabilities in a more efficient way.

5.1.1 The iDataPlex rack

The footprint of an IBM iDataPlex rack is similar to two standard 19-inch racks. The dimensions of a rack without doors, covers, and casters are approximately 1200 mm wide, 600 mm deep, 2093 mm high (47 inches, 24 inches, 83 inches.). The front and back casters each add approximately 120 mm (5 inches) to the depth. Figure 5-1 shows the top view of an iDataPlex rack in comparison to two standard 19-inch racks.

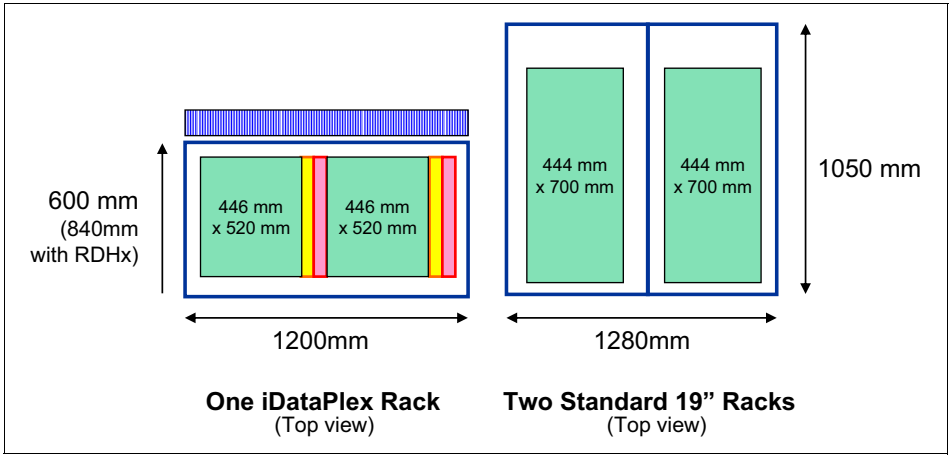


Figure 5-1 Comparison of iDataPlex rack and two standard 19-inch racks

If the Rear Door Heat eXchanger is added to an installation with rear casters, it adds no additional depth because the heat exchanger is located above the casters. In a configuration without rear casters, the Rear Door Heat eXchanger adds 120 mm (4.7 inches) to the depth of the rack, but only if the casters are removed.

As shown in Figure 5-2, an iDataPlex rack provides 84U (2 columns of 42U) of horizontal mounting space for compute nodes and 16U (2 sets of 8U) of vertical mounting space (sometimes called *pockets*) for power distribution units and networking equipment. This makes a total of 100U usable space on two floor tiles. The iDataPlex rack has an additional 1U of space at the top of columns A and C to mount a management appliance. This extra 1U space can also be used to pass through cables from the back to the front.

Note: This extra space can be relocated during the factory configuration from the top to the bottom of columns A and C.

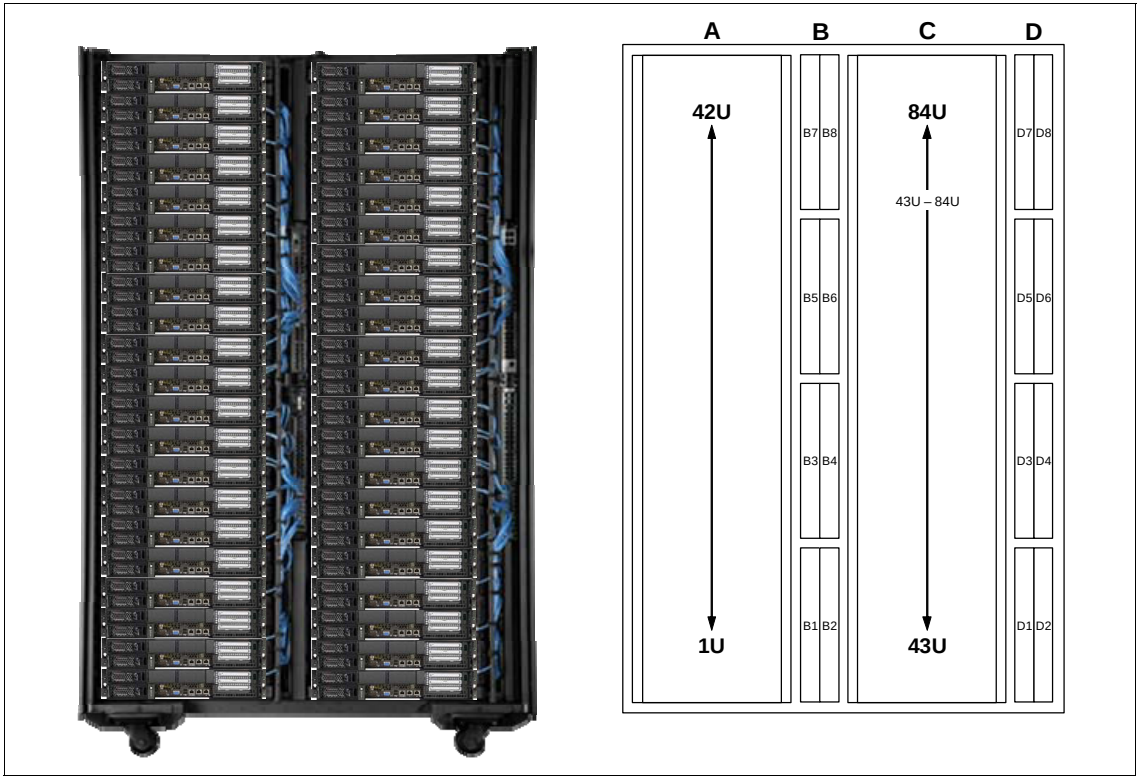


Figure 5-2 iDataPlex rack numbering scheme

The 8U bays (columns B and D in Figure 5-2) are aligned vertically in the middle and at the right side of the rack as seen from the front. Using them as eight 1U pockets or four 2U pockets is physically possible. Depending on the exact rack configuration, ducting fillers or pocket fillers are used to ensure proper air flow.

IBM iDataPlex follows the EIA 310-D 19-inch standard for rack devices, allowing you to mount additional, existing devices horizontally or vertically, as long as they fit into the reduced depth of the iDataPlex rack. You might also have to sacrifice some of the 84 horizontal units if you plan to use multiple interconnect technologies such as Ethernet or InfiniBand in the same iDataPlex rack. Two examples are provided in 5.3, “InfiniBand options”. One of the examples explains why you have to sacrifice some units when you require copper-based InfiniBand.

FAQs:

- Why do I want a custom rack form factor in my data center?

  The iDataPlex form factor provides you with a more efficient cooling solution, better power delivery, and higher density at the rack level.

- Do I have to change my data center layout?

  No. You can slide one iDataPlex in place of two existing racks with no change to your data center and still get all the benefits of iDataPlex.

- Will the iDataPlex rack fit through my doorway?

  iDataPlex requires a minimum door height of 2108 mm (83 in.), so it can fit through a typical doorway of 2133 mm (7 ft).

- What about the weight?

  The iDataPlex form factor is always less weight than an equivalent number of 1U server. However, because of the dense packaging, a fully populated 84 node iDataPlex rack can have a greater point loading than a single standard rack with 42 nodes.

5.1.2 Direct Water Cooled Rack with water manifold

Installations that use the Direct Water Cooling technology require a water manifold mounted at the rear of the iDataPlex rack. The rack (machine type 7197) is identical to the standard air-cooled rack but uses a different machine type for IBM service purposes. Figure 5-3 shows an illustration of the water manifold.

Figure 5-3 Water manifold for iDataPlex Direct Water Cooled Rack

The water manifold gets mounted on the rear end of the iDataPlex rack. Detailed instructions on how to mount the manifold can be found in the *IBM iDataPlex Rack Type 7197 Installation and User's Guide* for this machine type, available from the following website:

`http://www.ibm.com/support`

Figure 5-4 shows the rear of a fully populated iDataPlex rack using Direct Water Cooling technology. About 7 liters of water is required to fill the rack manifold. Also shown is the water direction. Blue arrows indicate the inlet water and red arrows indicate the outlet water.

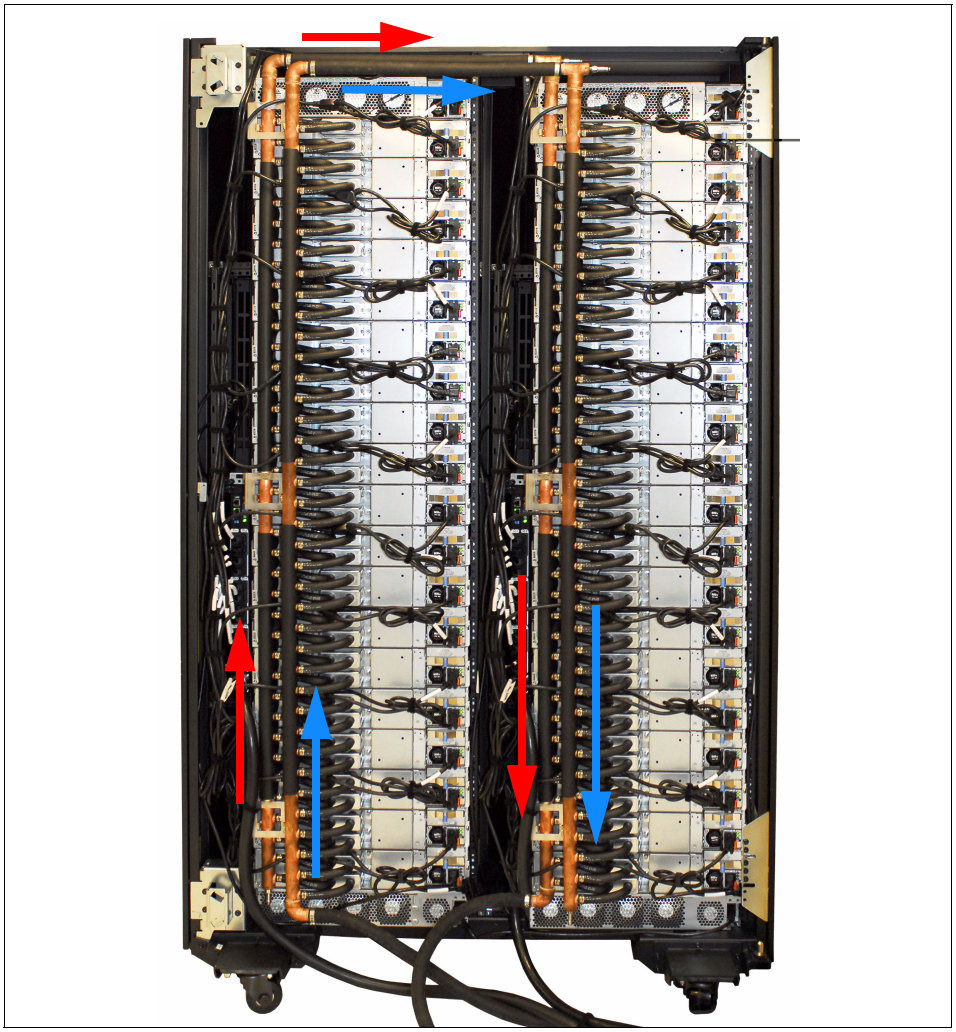


Figure 5-4 Fully populated iDataPlex Direct Water Cooled Rack (arrows indicate water flow)

5.1.3 iDataPlex cable brackets

An iDataPlex rack can be cabled through the bottom in a raised floor environment or from the overhead cable trays. Detailed cabling information about how you connect an iDataPlex rack to your existing cooling, power, and data infrastructure is available in 6.4, “Cabling the rack” on page 179.

The iDataPlex solution offers three types of cable management brackets. Depending on user requirements, they allow for the best alignment of power, Ethernet, or InfiniBand cables and ensure minimal interference of the cables with the airflow. Figure 5-5 shows four types of brackets. The first three are IBM brackets. The fourth is from *PANDUIT*.

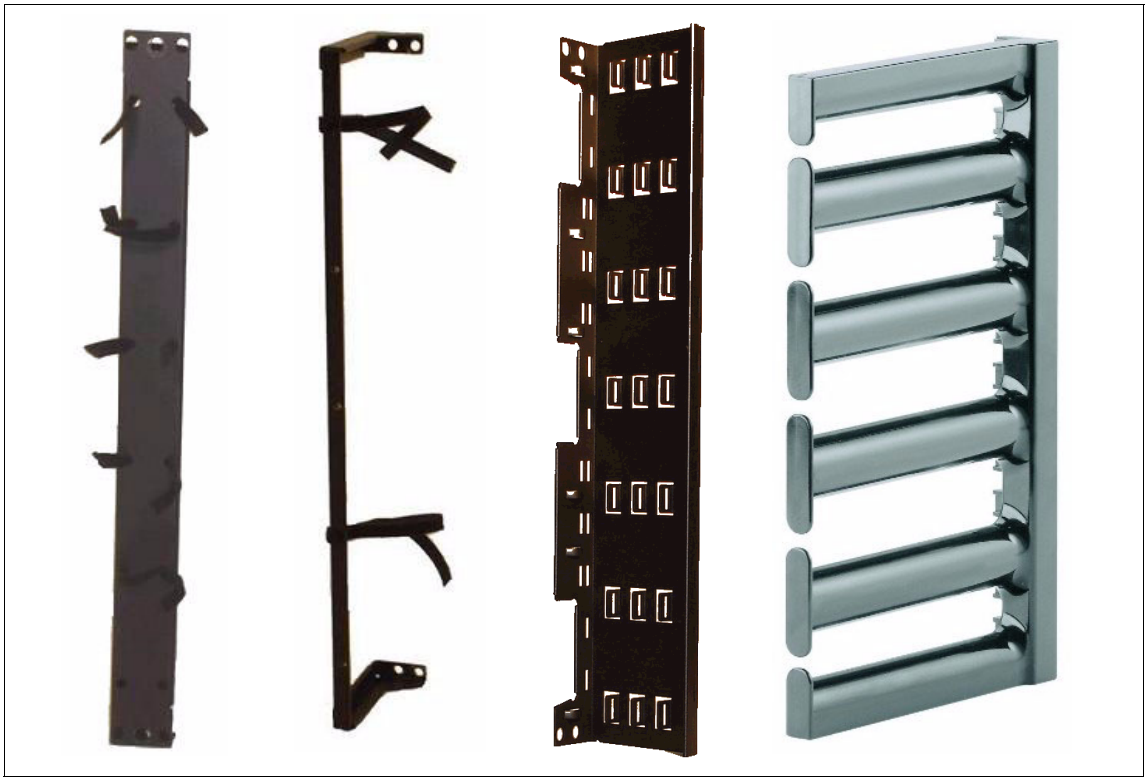


Figure 5-5 Cable bracket options for iDataPlex

Note: Because iDataPlex racks can be configured in numerous ways, not all configuration options can be covered in this document. IBM provides detailed rack configurations on a per customer basis to ensure that the best overall solution is chosen.

5.1.4 iDataPlex cooling options

The two iDataPlex rack cooling options are air and water. Air cooling relies on the traditional data center cooling system, external to the racks. The airflow is from the front to the back of an iDataPlex rack. Because the nodes are less deep than traditional servers, less air pressure is required to push air through the nodes, which in turn helps reduce both the overall cooling effort and the energy required.

The optional IBM Rear Door Heat eXchanger allows an iDataPlex rack to be water-cooled. The high efficiency of the heat exchanger allows an iDataPlex rack to operate with no extra cooling. The fans in the chassis move the air through the Rear Door Heat eXchanger, so it does not require any additional fans. For more information about the Rear Door Heat eXchanger, see 5.1.5, “Rear Door Heat eXchanger” on page 141.

Part of the Rear Door Heat eXchanger are two air baffles that prevent air recirculation. They are installed at the top and at the bottom of the rack and prevent hot air from bypassing the heat exchanger through gaps introduced by cables to external systems (the air baffles are sometimes referred to as *mud flaps*). See Figure 5-6.

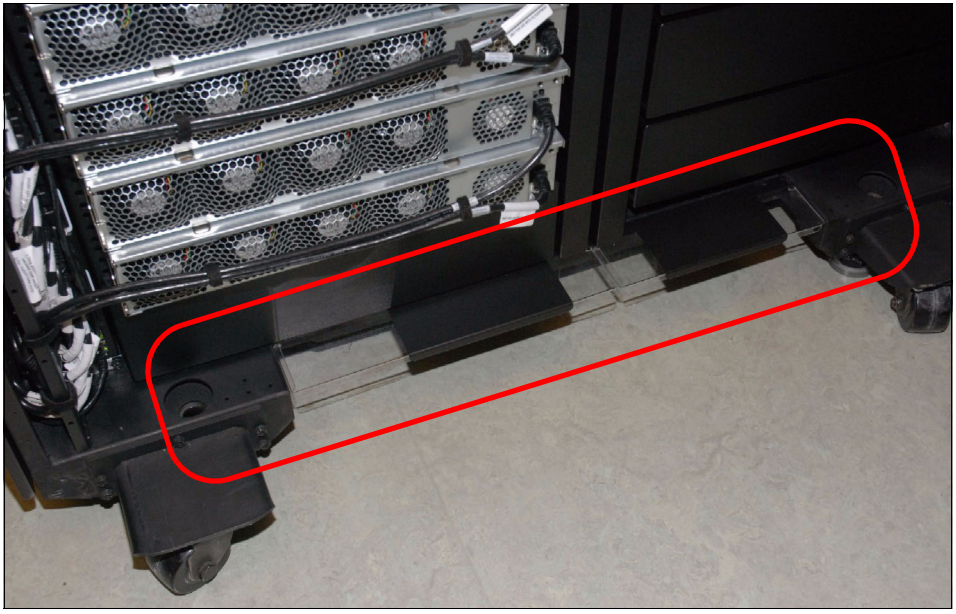


Figure 5-6 Air baffle “mud flap” installed to prevent hot air escaping

Switches and other ancillary components are typically installed in the vertical bays to maximize the number of chassis and servers that are installed in the horizontal bays in the rack. These devices sometimes have cooling paths that go from one side of the unit to the other, not front to back.

Even though these devices do not use front-to-back cooling, when they are installed in the vertical bays, no additional cooling is required. If they are installed in horizontal bays, however, a special air baffle might be required to route the air from the front of the rack to the side of the device, through the device, then from the other side to the rear of the rack. Figure 5-7 shows the airflow when using those air baffles.

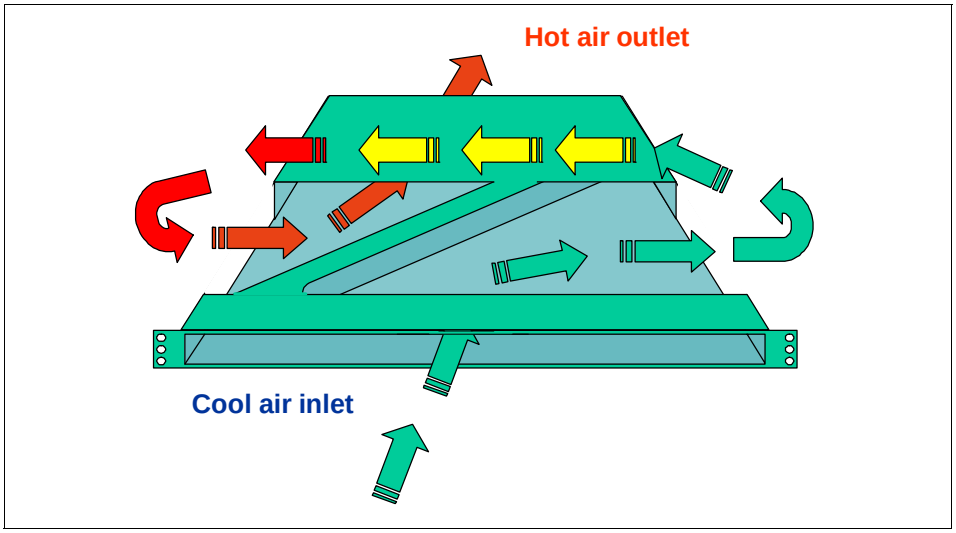


Figure 5-7 Airflow using special air baffle for side-to-side cooled devices

5.1.5 Rear Door Heat eXchanger

IBM offers an alternative to existing cooling systems. The IBM Rear Door Heat eXchanger (part number 43V6048), shown in Figure 5-8, replaces the existing rear door of an iDataPlex rack and provides an extremely efficient water-based cooling solution. The exchanger dissipates heat generated by all servers and provides the data center with cooling at the most efficient place, which is immediately behind the exhaust of the system fans. Without water, the door weighs about 80 - 90 kg (180 - 200 lb).

Figure 5-8 IBM iDataPlex rack with the Rear Door Heat eXchanger

The IBM Rear Door Heat eXchanger requires a cooling distribution unit (CDU) in your data center. It connects to your water system with two quick-connects on the bottom of the door.

Details about plumbing for a Rear Door Heat eXchanger are given in 6.9, “Water cooling using the Rear Door Heat eXchanger” on page 205. The door swings open so you can still access the PDUs at the rear of the rack without unmounting the heat exchanger. Service clearance is the same as for a standard rear door installation. The heat exchanger does not require electricity.

Figure 5-9 illustrates the efficiency of the heat exchanger in a set-up with 32 kW heat load and an inlet air temperature of 24°C (75°F). A 100% heat removal means that all heat generated by iDataPlex components is removed and dissipated into the chilled water (the air leaving the heat exchanger is the same temperature as air entering the rack at the front).

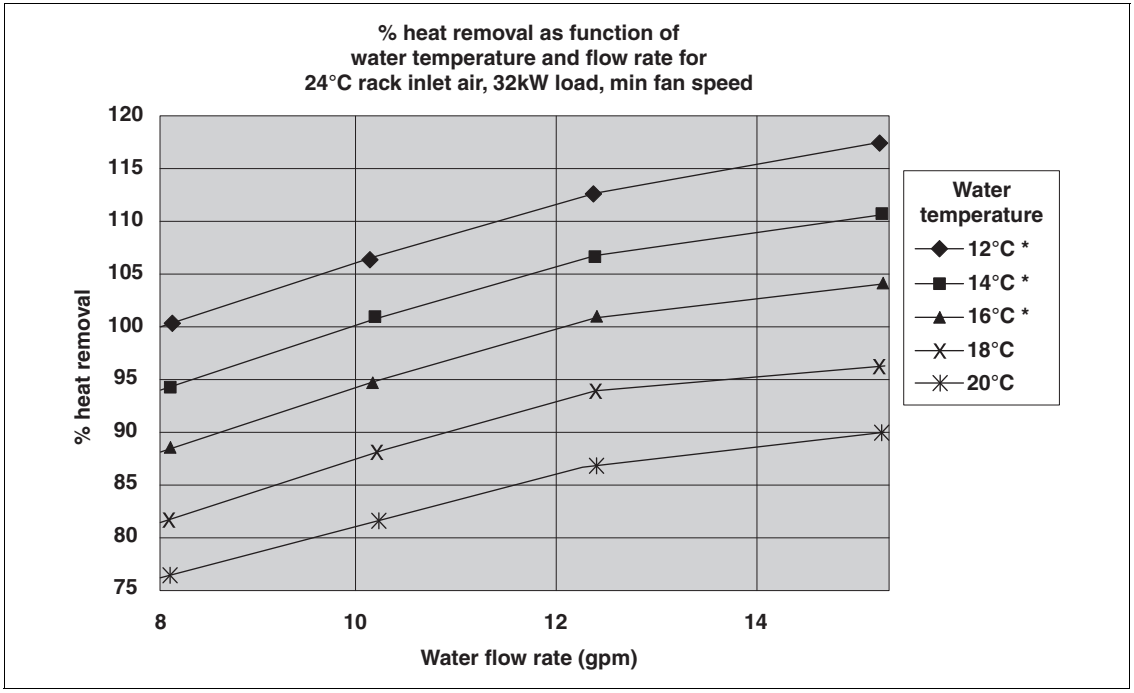


Figure 5-9 Typical performance of the heat exchanger

Note: Water temperatures below 18°C (64°F) can only be used if the system supplying the water has the ability to measure the room dew point and automatically adjust the water temperature accordingly to prevent condensation on the servers.

To help maintain optimum performance of the heat exchanger:

- Install filler panels over all unoccupied bays.
- Where possible, external cabling should exit the rack at the front.

- If cables are required to exit at the rear of the rack cabinet, make sure they enter or exit through the bottom and top air baffles (see Figure 5-6 on page 140). Bundle the cables together to minimize air gaps, taking care to avoid cable groupings that can lead to interference such as crosstalk. Make sure that the air-baffle sliders are closed as far as possible. See 5.1.4, “iDataPlex cooling options” on page 140.

For more information about installing the Rear Door Heat eXchanger, see the *Rear Door Heat eXchanger Installation and Maintenance Guide*:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5075220

The Rear Door Heat eXchanger is the only part of an iDataPlex solution that does not arrive pre-installed. It is shipped separately from the iDataPlex rack and must be installed by an IBM authorized supplier. It is the customer’s responsibility to ensure correct plumbing connections.

FAQ:

- What happens if I have to open the Rear Door Heat eXchanger for maintenance?

  The server cooling is not be affected because the heat exchanger cools the data center and not the servers. Because iDataPlex is all front-access (even for connecting to power), there is usually no reason to open the rear of the rack for maintenance. The heat exchanger has no fans or moving parts, so the failure rate is extremely small and it does not require regular maintenance.

5.2 Ethernet switch options

Common network designs have at least two hierarchies of networking. Every hierarchy has a different set of requirements to the switching hardware. The following section shows the supported Ethernet options for edge (sometimes called top-of-rack because of their physical location) and core switches.

5.2.1 Ethernet edge switches

Several Ethernet edge switches are supported, as listed in Table 5-1. They cover a wide range of application scenarios. IBM updates the list of supported switching hardware from other vendors on a regular basis. It ensures that clients always get the latest features that help boost the productivity of their data center.

Table 5-1 Supported Ethernet switches

Feature Code	Switch	Ports	Layer	Port speed
2638	IBM System Networking RackSwitch G8000F	48	L2/3	1 Gbps[a]
6942	IBM System Networking RackSwitch G8124F	24	L2	10 Gbps
A2N0	IBM System Networking RackSwitch G8316F	16	L2/3	40 Gbps[b]
2733	Cisco 2960G-48TC-L	48	L2	1 Gbps
A1M6	Cisco 3750X-48T-L	48	L2/3	1/10 Gbps
Not available	Juniper EX4200	48	L2/3	1 Gbps[a]
A120	Juniper EX4500	48	L2/3	1/10 Gbps
6941	SMC 8126L2	26	L2	1 Gbps
6673	SMC 8150L2	50	L2	1 Gbps

a. 10 Gb uplink modules available.
b. 64 10 Gbps ports when using QSFP+ DAC breakout cables.

Table 5-2 lists the features supported by each of the Ethernet switches.

The IBM System Networking G8124F switch supports different protocols from the IEEE 802.1 Data Center Bridging (DCB) standard for a converged fabric:

- Priority-based Flow Control (PFC), IEEE 802.1Qbb
- Enhanced Transmission Selection (ETS), IEEE 802.1Qaz
- Data Center Bridging Exchange (DCBX) using the Link Layer Discovery Protocol (LLDP, as defined in IEEE 802.1AB)
- Fibre Channel over Ethernet, including FIP snooping

Table 5-2 Basic features supported by Ethernet switches

Feature	IBM G8000F	IBM G8124F	IBM G8316F	Cisco 2960G-48TC-L	Cisco 3750X-48T-L	Juniper EX4200	Juniper EX4500	SMC 8126L2	SMC 8150L2
Ports									
# of ports	44	(2)[b]	(1)[a]	44+4[b]	48	48 (+1)	(1)	26	50
Port speed Gbps; ()=manage	1	(1)	(1)	1	1	1 / (1)	(1)	1	1
Module slots	4+4[c]	24	16	4	2/4[d]	2/4[d]	40+8[e]	4	4
Module slot speed (Gbps)	1 / 10[c]	1/10	10 / 40[f]	1	1 / 10	1 / 10	1 / 10	1	1
Stackable	Yes	No	No	No	Yes[g]	Yes	Yes	Yes	Yes
VLAN support									
Number of VLANs	1024	1024	1024	255	1005[h]	4096	4096	255	255
802.1Q VLAN Tagging	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Performance									
Link Aggregation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Jumbo frames	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
IGMP snooping	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
High availability and redundancy									
STP	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
VRRP	No	No	Yes	No	No	Yes	Yes	No	No
Management									
Serial port	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Telnet, SSH	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
SNMP	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
CDP support	No	No	No	Yes	Yes	No	No	No	No
VTP support	No	No	No	Yes	Yes	No	No	No	No

Feature	IBM G8000F	IBM G8124F	IBM G8316F	Cisco 2960G-48TC-L	Cisco 3750X-48T-L	Juniper EX4200	Juniper EX4500	SMC 8126L2	SMC 8150L2
GVRP support	No	No	No	No	No	Yes	No	Yes	Yes
Security									
Port-based security	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ACL (MAC-based)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ACL (IP-based)	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
TACACS+	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
RADIUS	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Quality of Service (QoS)									
802.1P	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Routing									
RIP	No	No	Yes	No	Yes	Yes	Yes	No	No
OSPF	No	No	Yes	No	Yes[i]	Yes	Yes	No	No
BGP	No	No	Yes	No	Yes[i]	Yes	Yes	No	No
IPv6									
IPv6 support	No	No	Yes	Yes	Yes	Yes	Yes	No	No

a. Used for management purposes only.
b. Four dual purpose ports: You can either use the regular copper RJ45 port or the SFP module port. When using SFP modules, the corresponding RJ45 ports are disabled (ports 45–48).
c. Four optional 1 GB SFP ports at the front, and up to four optional 10 Gbit SFP+ or CX4 ports on the back (SFP+ and CX4 modules are available as 2-port modules only).
d. The uplink module provides either four ports running at 1 Gbps or two ports at 10 Gbps.
e. Two optional four-port uplink modules can be installed.
f. 16 physical ports running at 40 Gbps each. They can be multiplexed to 64 ports running at 10 Gbps with special QSFP+ DAC breakout cables.
g. Requires Cisco IOS IP Base or IP Services feature set.
h. 255 VLANs supported with Cisco IOS LAN Base feature set. IP Base feature set required for 1005 VLANs.
i. Cisco IOS IP Services feature set required for full support.

If you plan to use copper-based InfiniBand[1] and Ethernet cabling, then mount both switches horizontally, if possible, due to the number of cables that go along the B and D columns. All the copper cables take up so much space in front of the vertical pockets that proper cabling to vertically-mounted switches is no longer possible. Figure 5-10 illustrates the recommended cabling.

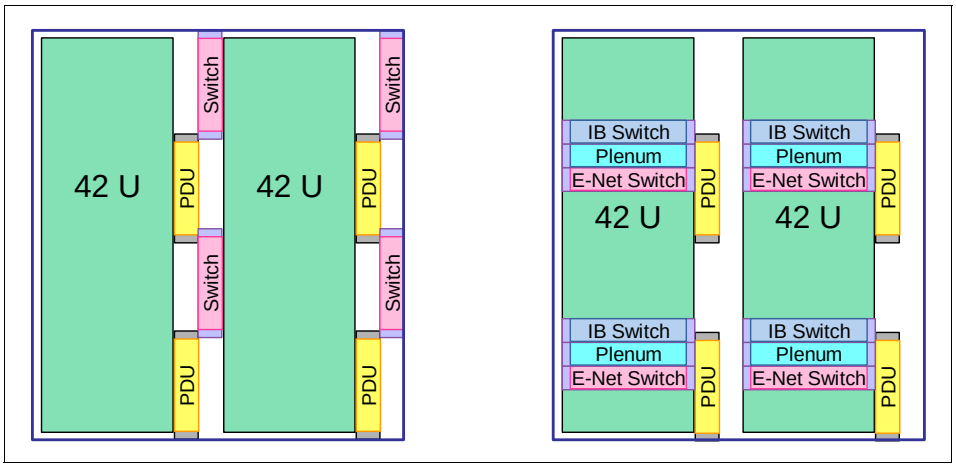


Figure 5-10 Recommended cabling for Ethernet-only (l) and Ethernet plus InfiniBand (r)

In addition to top-of-rack switches, iDataPlex also supports big core switches through the IBM Intelligent Cluster offering.

5.2.2 Ethernet core switches

Core switches are part of the backbone of a network. They are either deployed at the aggregation or the core layer in a network and play a vital role in the availability of the data center. It is why they have the following requirements:

- Very high availability
- High port count
- Packet forwarding at wire speed
- Hot-swappable components
- Flexible configuration

Juniper's EX8200 line of modular Ethernet switches provides a platform that meets these requirements. It consists of two chassis options, the Juniper Networks EX8208 Ethernet Switch and the EX8216 Ethernet Switch. Both chassis have a passive backplane that allows for an upgrade to 100 Gbps Ethernet without the need for a new chassis. Figure 5-11 shows the front view of an EX8208 chassis.

[1] This consideration only applies for copper-based InfiniBand. Fibre cables take up much less space.

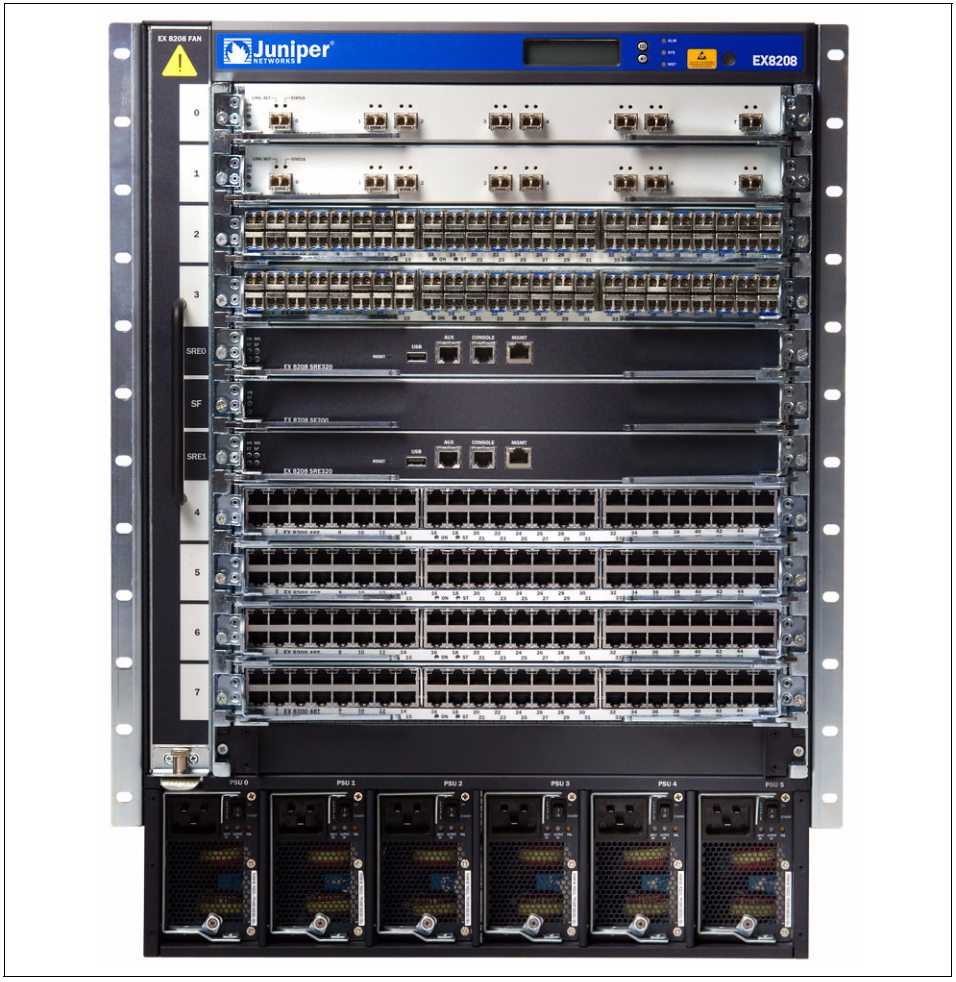


Figure 5-11 Juniper Networks EX8208 Ethernet Switch (front view)

The EX8208 is 14U high and supports up to eight line cards. The EX8216 is 21U high and supports up to 16 line cards. Both chassis use the same line cards. The following line cards are available:

- 48-port 10/100/1000 Base-T RJ45
- 48-port 10/100/1000 Base-X, for optical transmission using SFP transceivers
- 8-port 10 Gbps SFP+
- 40-port 1 Gbps/10 Gbps SFP/SFP+
- 48-port 10/100/1000 Base-T RJ45, 20 Gbps line card (available with PoE support)

- 40-port 10/100/1000 Base-T RJ45, 4-port 1000 BASE-X, and 2-port 10 Gbps SFP+ (available with PoE support)

The first four line cards are available in an extra scale configuration with more memory for larger IPv4 and IPv6 route tables. It is ideal for large-scale deployments to support a higher number of servers.

All Juniper switches, their core switches, and top-of-rack switches, run the same operating system image called Junos OS. There is a new Junos OS image released on a quarterly basis.

Major features of Junos OS are as follows:

- Modular software design
- Based on FreeBSD
- Single release train
- SDK available for application development

5.3 InfiniBand options

InfiniBand is an industry standard, high-speed interconnect mainly used in the area of high performance computing where very low latencies and high throughput are required. More recently, InfiniBand technology has entered additional markets such as digital media and enterprise data centers hosting large clustered systems.

InfiniBand is an industry standard that is defined and maintained by the InfiniBand Trade Association (IBTA). All major hardware vendors are represented in the IBTA committees, and they contribute to the specification of further technologies.

InfiniBand was designed to take the place of today's data center networking technology. In the late 90s a number of next-generation I/O architects came together to form an open, community-driven network technology that would provide scalability and stability based on successes from other network designs. Today, InfiniBand is a popular and widely used I/O fabric among customers in the Top500 Universities and Labs, and in Life Sciences, Biomedical, Oil and Gas (Seismic, Reservoir, Modeling applications), Computer Aided Design and Engineering, and Enterprise Oracle and Financial Applications.

An InfiniBand network consists of devices similar to an Ethernet network; see Figure 5-12. InfiniBand relies on a switched fabric to which channel adapters (CAs) connect. Two types of channel adapters exist: host channel adapters (HCA) and target channel adapters (TCA).

Host channel adapters reside in the compute nodes (for example, servers). Target channel adapters are installed in the other end of the fabric such as storage devices. CAs connect to each other either by copper or fibre cables.

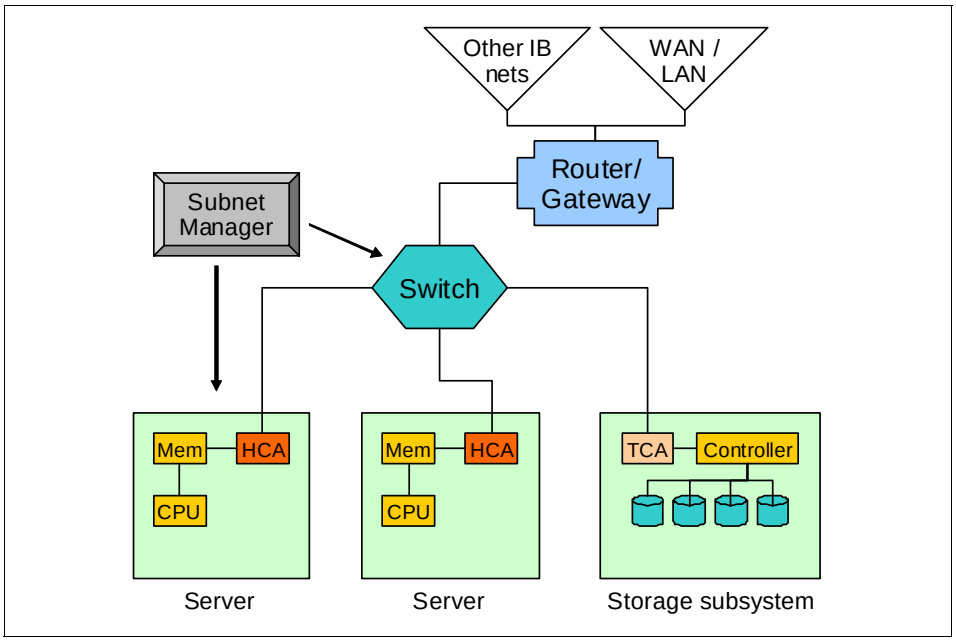


Figure 5-12 Elements of an InfiniBand network

A managing instance is needed to configure and set up addressing and routing between all elements of the fabric. This instance is called *subnet manager*. The subnet manager can run embedded on a switch or on a server.

5.3.1 InfiniBand speed terminology

InfiniBand uses a serial link with a signalling rate of 2.5 Gbps in each direction. This is known as single data rate (SDR). InfiniBand supports double data rate (DDR) and quad data rate (QDR), which leads to 5 Gbps and 10 Gbps per link.

To achieve a higher bandwidth on a single cable, several transmission lanes can be aggregated to a single link. Aggregations of four, eight, and twelve lanes are available, known as 4x, 8x, and 12x.

At the physical layer, InfiniBand uses an 8B/10B encoding, leading to a 2 Gbps (or 250 MB/s) net data rate on a single 1x link in each direction. Table 5-3 shows the theoretical net throughput for all combinations.

Table 5-3 InfiniBand theoretical net throughput rate

	1x	4x	8x	12x
Single data rate	2 Gbps	8 Gbps	16 Gbps	24 Gbps
Double data rate	4 Gbps	16 Gbps	32 Gbps	48 Gbps
Quad data rate (QDR)	8 Gbps	32 Gbps	64 Gbps	96 Gbps
Fourteen data rate (FDR)	14 Gbps	56 Gbps	112 Gbps	168 Gbps
Enhanced data rate (EDR)	26 Gbps	104 Gbps	208 Gbps	312 Gbps

Higher lane aggregation rates such as 12x are used for uplink connections or switch-to-switch connections.

Starting with Fourteen Data Rate (FDR), InfiniBand uses a different naming scheme. The encoding changed from 8B/10B to 64B/66B. 64B/66B is already used in other high-speed technologies such as 10 Gigabit Ethernet. The new encoding scheme has less overhead than 8B/10B and allows more data to be sent at the same speed.

The FDR signalling rate increases to 14 Gbps per lane with a common 4x link providing 56 Gbps of bandwidth per port. FDR adds Forward Error Correction (FEC) to the InfiniBand standard. FEC reduces the number of packet retransmissions caused by network bit errors.

The roadmap of the InfiniBand Trade Association (IBTA) shows the next standard to be Enhanced Data Rate (EDR, not to be confused with Eight Data Rate) and is scheduled for 2013[2].

FDR-10: Some Mellanox data sheets mention support for FDR-10. This is not an official IBTA standard and is unique to Mellanox. FDR-10 operates at QDR speed but uses 64B/66B encoding instead of 8B/10B. This encoding provides a higher net throughput rate compared to standard QDR due to the lower encoding overhead.

5.3.2 InfiniBand connectors and cables

InfiniBand cables are available either as copper or optical cables. Depending on the actual link speed, copper cables have a limited length of about 10 to 15 m. This limitation can be mitigated by using optical cables. Distances of up to 300 m are possible and a smaller bend radius can be achieved. It gets more and more important for high-density installations with a large number of cables going in a single rack.

[2] Check the InfiniBand TA Web site for the latest version of the roadmap at: http://www.infinibandta.org/content/pages.php?pg=technology_overview

For links up to 4x DDR speed, cables with a CX4 type connector are used (it is sometimes also referred to as an SFF 8470 or microGiGaCN-type connector).

When building a QDR InfiniBand network, special cables are required to support the higher QDR signalling rate. To achieve a higher port density, the connector had to be made smaller as well. A new connector called Quad Small-Form-factor-pluggable (QSFP) was defined, which houses four independent transmit and receive channels, each capable of transmitting at 10 Gbps, for an aggregated bandwidth of 40 Gbps per cable.

Figure 5-13 shows the CX4 and QSFP connectors.

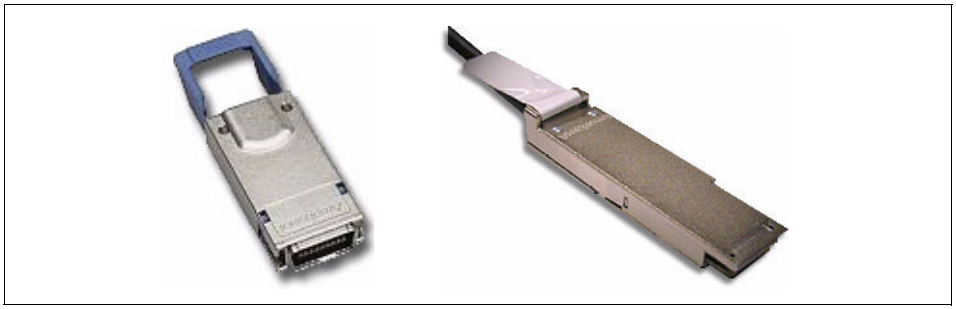


Figure 5-13 InfiniBand CX4 and Quad Small-Form-factor-Pluggable (QSFP) connectors

Table 5-4 lists the QDR cables with QSFP connectors on both ends that are officially tested in an iDataPlex environment.

Table 5-4 Supported QDR QSFP InfiniBand cables in iDataPlex

	Copper or Optical	**Length**	**Feature code**
Mellanox InfiniBand Cable	Copper	1 m	A2F9
Mellanox InfiniBand Cable	Copper	3 m	A2FA
Mellanox InfiniBand Cable	Copper	5 m	A2FB
QLogic InfiniBand Cable 30AWG	Copper	0.5 m	3725
QLogic InfiniBand Cable 30AWG	Copper	1 m	3726
QLogic InfiniBand Cable 30AWG	Copper	3 m	3727
Mellanox InfiniBand Cable	Optical	3 m	A2FS
Mellanox InfiniBand Cable	Optical	5 m	A2FT
Mellanox InfiniBand Cable	Optical	10 m	A2FC

	Copper or Optical	Length	Feature code
Mellanox InfiniBand Cable	Optical	15 m	A2FD
Mellanox InfiniBand Cable	Optical	20 m	A2FE
Mellanox InfiniBand Cable	Optical	30 m	A2FG
QLogic InfiniBand Cable	Optical	3 m	3731
QLogic InfiniBand Cable	Optical	10 m	3732
QLogic InfiniBand Cable	Optical	30 m	3733
IBM InfiniBand Cable	Optical	3 m	5989
IBM InfiniBand Cable	Optical	10 m	5990
IBM InfiniBand Cable	Optical	30 m	5991

To connect a QDR port with a DDR port, you need a QSFP-to-CX4 cable. The endpoints can be either switches or HCAs. Keep in mind that the resulting linkspeed will be DDR only. Table 5-5 lists the supported QSFP-to-CX4 cables.

Table 5-5 Supported DDR QSFP-to-CX4 InfiniBand cables in iDataPlex

	Copper or Optical	Length	Feature code
QLogic InfiniBand Cable 30AWG	Copper	0.5 m	3722
QLogic InfiniBand Cable 30AWG	Copper	1 m	3723
QLogic InfiniBand Cable 28AWG	Copper	3 m	3724
QLogic InfiniBand Cable	Optical	3 m	3728
QLogic InfiniBand Cable	Optical	10 m	3729
QLogic InfiniBand Cable	Optical	30 m	3730

When supported Mellanox HCAs are used to run 10 Gb Ethernet with SFP+ connectors, you need a QSFP-to-SFP+ Adapter, or short QSA (IBM Feature Code A2FH). Figure 5-14 shows such an adapter that plugs directly into the HCA cage.

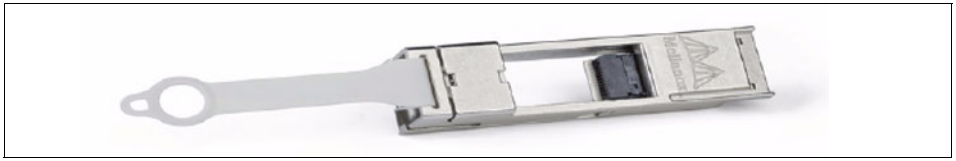


Figure 5-14 Mellanox QSFP-to-SFP+ Adapter (QSA)

5.3.3 InfiniBand versus 10 Gbps Ethernet

Up until a few years ago, the differentiation between Ethernet and InfiniBand was easy to describe:

- Ethernet is an industry-wide, well-accepted networking standard that offers enough throughput for most day-to-day applications that are not latency-sensitive, and Ethernet switches and network adapters are commodity products available at reasonable prices.
- InfiniBand is the interconnect technology that offers very low latency at very high data rates. However, it might require significant investment in a new network technology infrastructure.

The situation has shifted somewhat with the adoption of 10 Gbps Ethernet and the expansion of InfiniBand into more general market areas. But even though both technologies have moved closer to each other, InfiniBand maintains several distinct advantages.

10 Gbps Ethernet is not only the next increase in speed for Ethernet; it also implements new features such as the Internet-wide area RDMA protocol (iWARP). Remote direct memory access (RDMA), one of the major technical advantages of InfiniBand, allows you to copy data between memory on different nodes without consuming CPU cycles for the transfer process. This is implemented in hardware and it decreases Ethernet latency. But as of 2009, even though 10 Gbps equipment is available from most vendors, it is still more costly than InfiniBand, especially for larger installations with a high port count.

InfiniBand, on the other hand, became more a common interconnect, not only in the area of high-performance computing but also in everyday data centers. For example, database clusters such as IBM DB2® clustering or Oracle RAC increasingly use InfiniBand to achiever better scalability and higher throughput when using their shared file system. Even with the rise of 10 Gbps Ethernet and iWARP, InfiniBand still has lower latency and, especially with 4x QDR on the market, provides four times the bandwidth of 10 Gbps Ethernet.

In terms of data reliability and integrity, InfiniBand provides the highest levels of data integrity by performing cyclic redundancy checks (CRCs) at each fabric hop

and end-to-end across the fabric to avoid data corruption. To meet the needs of mission-critical applications and the highest levels of availability, InfiniBand provides fully redundant and lossless I/O fabrics with automatic failover path and link layer multi-paths.

According to a recent IDC Analyst Connection[3], InfiniBand has the potential to be a major driver for the adoption of cloud computing in modern data centers. Key features such as “node-to-node latency of 1µs, up to 40 Gb/sec bandwidth, and scalability of over 10,000 nodes, plus reliability features such as redundant paths between nodes - and it costs less than 10GE [...] make[s] IB a good candidate for network infrastructure consideration in cloud computing environments.”[3] The Taneja Group reports that they “see external service providers building InfiniBand infrastructures today [because] InfiniBand brings better than SAN-like capabilities to every I/O interface, while reducing the number of interfaces.”[4]

Especially the new QDR InfiniBand products offer features that ease the transition to a unified fabric. IDC mentions features such as “QoS, virtual HCAs, and virtual NICs [that help] to provide a comprehensive I/O management system for virtualization. Specifically, the IB channelized architecture design can dynamically allocate high bandwidth to each virtual server, while consolidating LAN and WAN traffic from the virtual servers over the same IB pipe. This greatly reduces network complexity while still delivering low-latency and high-bandwidth capabilities.”[3]

5.3.4 HCAs and switches

This section describes the set of HCAs and switches for QDR and FDR-10 speed supported in the iDataPlex rack.

HCA

Table 5-6 lists the IB HCAs that are supported in iDataPlex servers.

Table 5-6 Supported iDataPlex InfiniBand HCAs

	Feature code	Host interface	Ports	Media
Mellanox ConnectX-2 VPI QDR IB/10GbE[a]	5446	PCIe 2.0 x8	1	CX4, QSFP, or SFP+[b]
Mellanox ConnectX-2 VPI QDR IB/10GbE[a]	5447	PCIe 2.0 x8	2	CX4, QSFP, or SFP+[b]

[3] IDC Analyst Connection, *InfiniBand: Poised for Market Growth*, available at: http://idcdocserv.com/784

[4] Cited in an IBTA press release available at: http://www.infinibandta.org/content/pages.php?pg=press_room_item&rec_id=540

	Feature code	Host interface	Ports	Media
Mellanox ConnectX-3 QDR/FDR10 Mezz Card[c]	A24F	PCIe 3.0 x8	2	QSFP
Mellanox ConnectX-3 QDR/FDR10 Mezz Card (water cooled)[d]	A24G	PCIe 3.0 x8	2	QSFP

a. Virtual Protocol Interconnect (VPI) can auto-sense the connected fabric on a single port basis and operate on InfiniBand, Ethernet, or Data Center Bridging fabrics.
b. QSFP to SFP+ connectivity through the QSA module (feature code A2FH).
c. Only supported with the iDataPlex dx360 M4 server.
d. Only supported with the iDataPlex Direct Water Cooled dx360 M4 server.

Switches

Table 5-7 lists the information about all supported InfiniBand QDR and FDR-10 switches.

Table 5-7 Supported iDataPlex InfiniBand switch

	Feature code	Ports	Port speed
QLogic 12200	6925	36	4x QDR (or 18 ports running at 8x QDR)
Mellanox IS5030	6670	36	4x QDR
Mellanox SX6036[a]	A2EZ	36	4x QDR/FDR-10
Voltaire Grid Director 4036E[b]	A1J9	34 + two 1/10 GbE	4x QDR
Voltaire Grid Director 4036E-LM[b]	A1JB	34 + two 1/10 GbE	4x QDR

a. In contrast to the SX6036 listed on the Mellanox website, IBM supports this switch at QDR or FDR-10 speed only. The switch can not be configured to run at FDR speed.
b. The 4036E-LM has fewer IB-ETH gateway internal memory and fewer multicast and VLAN partitions compared to the 4036E switch.

As part of the IBM Intelligent Cluster process, IBM also supports several InfiniBand core switches with up to 864 ports.

5.4 Scalable external storage options

iDataPlex provides your data center with very high performance and very dense compute power. It sets high requirements also towards the storage subsystem to be able to feed the servers with data and to store the results.

For example, for very long running computational jobs, it is general practice to create intermediate checkpoints. Those checkpoints save the current state of the computation and can put a high burden on the storage subsystem because a very large amount of data must be written in a very short timeframe.

Through the IBM Intelligent Cluster program, IBM partners with DataDirect Networks (DDN) and sells the DDN SFA10K Storage Platform. DDN's Storage Fusion Architecture (SFA) is often deployed with HPC installations because it provides a very scalable platform without any bottlenecks to allow for a very high performance in different configurations.

The SFA10K platform provides you with the following major advantages:

- Scalability:
  - Scale-up capacity with up to 1.200 disk drives in a single system
  - Scale-out performance
- Performance:
  - 15 GB/s bandwidth for a single system
  - 840,000 IOPS sustained for a single system
- Density:
  - 1.8 PB of storage in a single rack, or 3.6 PB in a two-rack system
  - 600 disk drives per rack
- Reliability:
  - SATAssure Real-Time error detection and correction, increases
  - Highly available system architecture

The SFA10K can be deployed in either 42U or 45U racks.

The SFA10K platform supports up to 1,200 disk drives total through up to 20 StorageScaler 7000 Disk Enclosures. Depending on the exact needs, you can choose between three different base configurations of the system including five, ten, or 20 disk enclosures. Figure 5-15 shows them installed together with the controllers. The five-enclosure system fits in a single rack, and the ten-enclosure and twenty-enclosure systems fit in two racks.

Figure 5-15 DDN SFA10K showing three system configurations (5-, 10-, and 20-enclosure systems)

The DDN solution offers high performance storage. For example, a typical Life Sciences customer with 20 Illumina HiSeq 2000 genetic sequencers can be generating more 2 giga-bases per machine per 48 hour run. Such processing could result in as much as 10 TB of raw unprocessed data.

An SFA10K with five enclosures and a balanced combination of 300 SATA/SSD drives would be able to capture that data faster than it is being generated by a factor of four and would provide enough capacity for 2.5 years of production (assuming all the intermediated data is kept). The solution would also be able to expand by adding 900 more 3TB SATA drives in as little as 30 TB increments, all under the same pair of controllers.

The SFA10K includes the following technical features:

- Fibre Channel (sixteen 8 Gb FC ports) or InfiniBand (eight QDR ports) connectivity
- Active/Active storage controllers with cache coherency and failover
- 16 GB controller cache, mirrored and battery backed
- Internal SAS switching (at 480 Gb/s)
- Support for RAID 1, 5, and 6
- Supports mixed SATA, SAS, SSD configurations
- Hot-swappable and redundant components (disk drives, power supplies, fan modules)

Each enclosure is 4U high and holds 60 disk drives. Using 3 TB disk drives, you can have a storage pool of 180 TB within 4U. Figure 5-16 shows a StorageScaler 7000 Disk Enclosure.

Figure 5-16 DDN StorageScaler 7000 Disk Enclosure

IBM supports 3.5-inch SATA and SAS drive or solid-state drives (SSDs) based on the eMLC (enterprise multi-level cell) technology). See Table 5-8.

Table 5-8 DDN SFA10K options supported through IBM Intelligent Cluster program

IBM MTM or part number	Description
0750-011	DDN SFA10K Fibre Channel Based Singlet Bundle
0750-012	DDN SFA10K InfiniBand Based Singlet Bundle
0750-015	DDN StorageScaler 7000 Disk Enclosure
0750-018	DDN StorageScaler 7000 Five-to-Ten Enclosure Upgrade
0750-019	DDN StorageScaler 7000 Ten-to-Twenty Enclosure Upgrade
0750-020	DDN 45U Rack for U.S.
0750-023	DDN 42U Rack for U.S.
95Y3511	1 TB SS7000 SATA 3.5-inch disk drive
95Y3514	3 TB SS7000 SATA 3.5-inch disk drive

IBM MTM or part number	Description
95Y3520	600 GB SS7000 SAS 3.5-inch disk drive
95Y3526	200 GB SS7000 eMLC solid-state drive
95Y3529	400 GB SS7000 eMLC solid-state drive
00D5211	3 m DDN mSAS cable

Further documentation about the DDN SFA10K can be found at the DDN website:

http://www.ddn.com/products/sfa10000

A white paper about how the DDN SFA10K storage architecture fits the needs of today's data center requirements can be found in the documentation section on the DDN website:

http://www.ddn.com/support/product-downloads-and-documentation#SFA10000

5.5 Example iDataPlex scenarios

The following two sections illustrate how iDataPlex can be implemented when designing a solution. It focuses on the main users of massive scale-out data centers: Web 2.0 companies and HPC customers. The two scenarios are meant as high-level examples to demonstrate the flexibility of a design based on iDataPlex. They are not design guides or recommendations on the use of specific hardware or software.

5.5.1 Web 2.0 example

On the surface, data centers at Web 2.0 companies look similar: a big farm of commodity x86 boxes. But, when you look more closely at the applications that run on these systems, it is a bit different. Although most of them run generally available open-source software, they enrich it with additional features or add specific components as proprietary pieces (these make the real asset of several companies, the most famous example is the Google search algorithm).

The following example covers a generic approach of running a Web 2.0 data center with a software stack comparable to LAMP, enhanced by dynamic elements. It consists of a firewall, a Web server, an application server, and a database server farm. Other than the firewall, these are assisted by a caching server. Figure 5-17 shows the typical configuration.

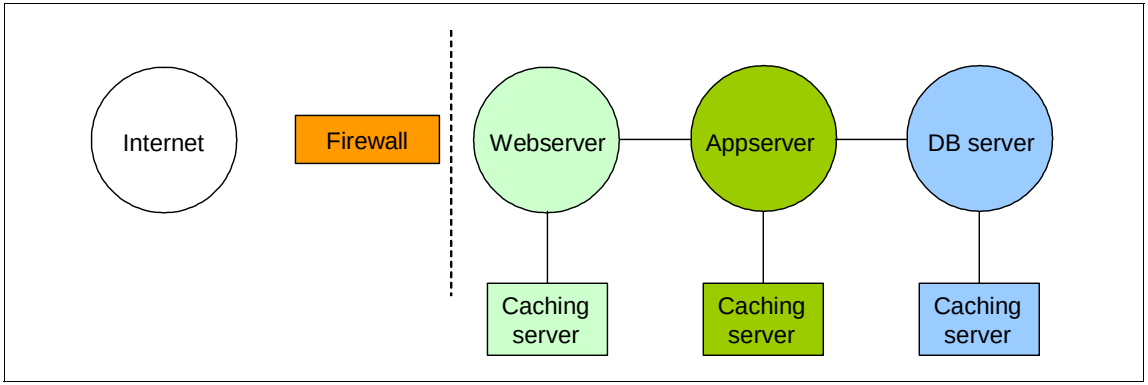


Figure 5-17 Generic Web 2.0 infrastructure

The following list shows characteristic parameters for server configurations for each of the different application areas:

- Firewall: I/O-rich configuration:
  - 2U chassis
  - Intel Pro1000 Ethernet card, providing four additional 1 Gbit Ethernet ports
  - Four hot-swappable SAS drives with hardware RAID
  - High availability solution, such as Linux-HA
- Web server: storage-rich configuration:
  - 2U chassis
  - SATA drive configuration
  - Aggregated Ethernet links for higher availability and performance
- Application server: storage-rich configuration:
  - 2U chassis
  - SATA drive configuration
  - CPU speed and main memory size above standard installations
- Database server: 3U storage-rich configuration:
  - SAS drives, 15 k RPM, hardware RAID
  - CPU speed and main memory size above standard installations
  - Aggregated Ethernet links for higher availability and performance

If the Web server does not have high storage demands, using the compute-node solution, which can provide two nodes per 2U, is possible, allowing for twice the density.

Figure 5-18 shows how an iDataPlex rack might be configured with these types of servers.

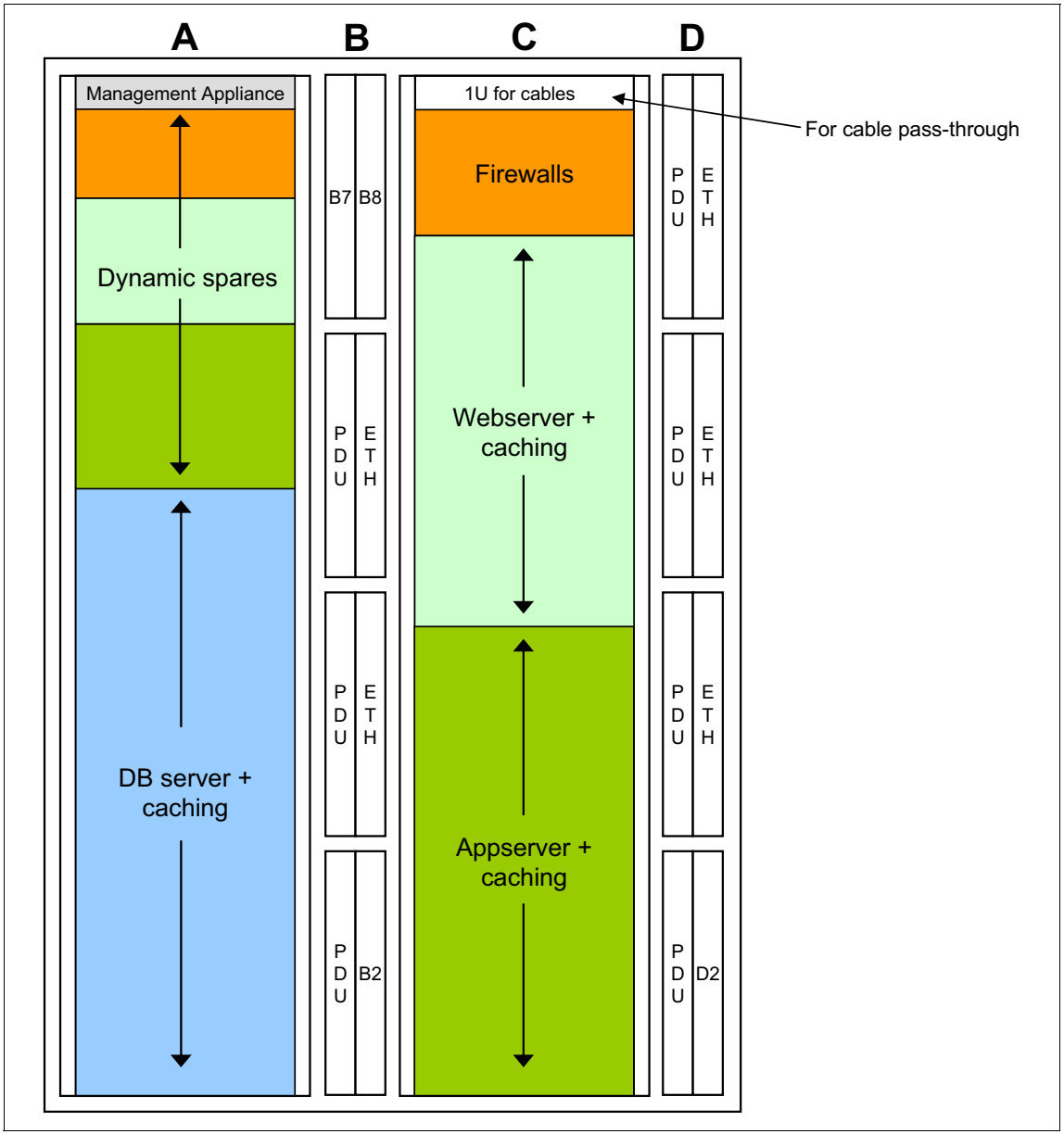


Figure 5-18 iDataPlex sample Web 2.0 configuration

The configuration assumes a data center with cabling from the ceiling. Therefore, 1U is left for cables at the top of column C. All the Ethernet switches are capable of working at ISO Layer 2 or Layer 3, which provide security features such as IP-based access control lists (ACLs). For additional security, VLANs are used to separate the different server-clouds. The different Ethernet switches are connected by way of two or more aggregated links to ensure best performance and high availability. Spare servers are available in the upper left part of the rack which can be used if other servers fail.

If the company is growing, more capacity is likely necessary in the data center. Increasing the computing power in the data center by adding a second iDataPlex rack is easy.

For example, if a Web 2.0 company decides to enhance its current service with social networking, the company will require more database servers (and possibly also application servers) to handle the workload. The company can install an additional iDataPlex rack with the left-half (column A) installed with the database server, and the right-half (column C) installed with the application server.

5.5.2 HPC example

This example shows how to implement iDataPlex for an HPC customer who wants an expandable, but low-cost environment based on Linux. The solution is based on one rack, but ready to be embedded into a bigger environment. The scenario assumes an application that is CPU-bound, uses an average amount of main memory, and needs low-latency communication for data bursts at infrequent intervals.

The proposed solution consists of 81 compute nodes, each being one dx360 M4 server with two Intel octo-core SandyBridge CPUs. This results in a configuration with 162 physical CPUs and 1296 cores in one rack. Each node can be equipped with up to 128 GB of main memory running at 1600 MHz, giving a total of 10,368 GB of RAM in a single rack.

All nodes work in a stateless environment, PXE-booting from one management node. The Ethernet interfaces are only used for booting and controlling the nodes, and InfiniBand is used as the high-speed interconnect. Five 36-port InfiniBand switches are installed in 1U pockets of columns B and D. The design uses 36-port switches that provide enough uplinks for a fully non-blocking solution when integrated into a bigger environment. See Figure 5-19. Each node requires an InfiniBand HBA in the mezzanine PCIe slot. The local disk is used for swapping and temporary disk purposes only.

Note: The sum of all devices in either column A or C is 43. This value is correct because there is one additional unit in both columns for management and switch options.

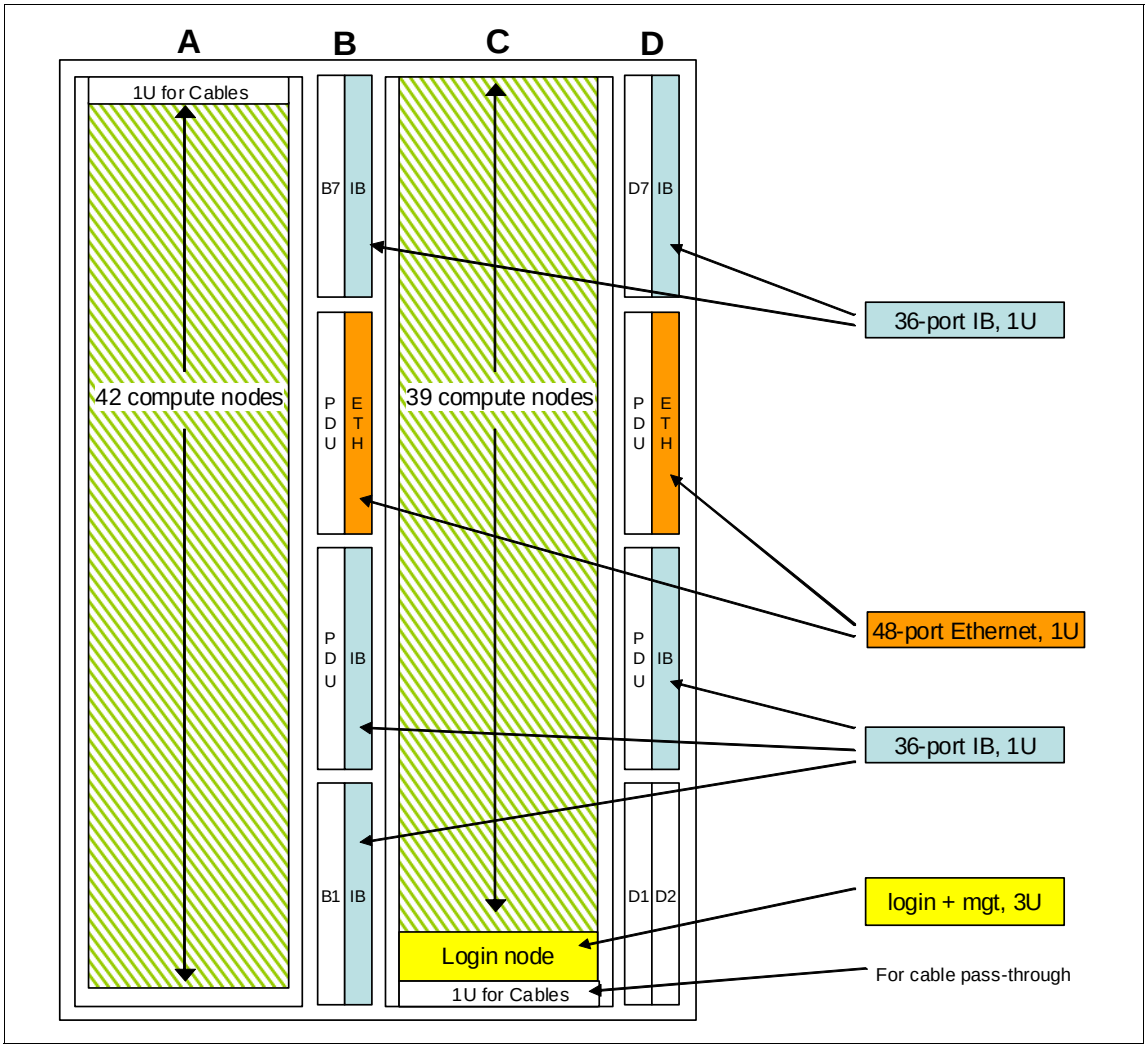


Figure 5-19 iDataPlex sample HPC configuration

Compute-node characteristics are as follows:

- IBM dx360 M4 server, with two Intel E5-2600 octo-core Xeon CPUs and 128 GB RAM.
- PXE performs stateless boot through Ethernet; local SATA/SAS or solid-state disks are used for swap and scratch.
- InfiniBand HBA for low-latency communication.

In our scenario, we have dedicated a 3U chassis as the management server, providing space for up to 7.2 TB of data on SAS disks, depending on number and size of installed disks. The example configuration uses fast 600 GB SAS drives spinning at 15 k RPM. It allows for the highest speed when reading or writing data.

The storage is divided into two partitions: A small partition is for the local file system of the management node; a large partition is where the shared file system for the nodes resides. The management node runs a cluster management software such as xCAT, serving stateless images.

Using stateless images, only a small fraction of the available 7.2 TB has to be sacrificed for management. The remainder is allocated for a parallel file system, such as General Parallel File System (GPFS™), that is mounted on all the nodes. The parallel file system is provided to the nodes as a common file system to exchange application data.

For monitoring purposes, each node can run a small Ganglia daemon to monitor the performance. On the management node, this information is collected and visualized in several graphs on a Web interface.

Figure 5-20 shows Ganglia monitoring 32 busy cluster-nodes in an iDataplex rack.

Figure 5-20 An iDataPlex cluster with 32 nodes running Ganglia

Management node characteristics are as follows:

- 3U chassis.
- Twelve SAS drives, each 600 GB, 15 k RPM, and RAID-5.
- One interface internal to the cluster, one interface to the production network.
- Runs Linux with xCAT serving stateless images through PXE boot.
- Runs a job scheduler such as IBM LoadLeveler®, or Simple Linux Utility for Resource Management (SLURM).
- Acts as a GPFS server for the nodes.
- Runs Ganglia for performance monitoring of all nodes.

The client data center requirements can be met with the flexible iDataPlex design. For instance, IBM was able to deploy a compute-chassis iDataPlex solution, in which all 84 nodes used copper-based InfiniBand cables to exit the rack and connect directly to an external switch.

6

Designing your data center to include iDataPlex

This chapter discusses rack details, several recommendations for moving the rack to its proper place, how the rack interfaces with the data center, and issues that data centers face, such as cooling, airflow, power, and others.

The following topics are covered in this chapter:

6.1 Data center design history

In 1965, Gordon Moore predicted that the number of transistors that could inexpensively be placed on a chip would double approximately every two years, and this statement (known as Moore's law) is still accurate. This increase in transistor density means that computer equipment is often replaced by either physically smaller systems or more powerful systems (or both smaller and more powerful).

A larger number of smaller units equates to higher rack densities, often resulting in a significant increase in power consumption and cooling requirements. Higher power density leads to different cooling requirements, because each additional watt of power used by IT equipment means additional power for cooling must be provided to remove the heat produced from the data center. This fact is also true of the iDataPlex, which allows 138% more servers than standard 1U server, on the same floor space.

In the 1960s, IBM introduced two technologies:

- The IBM System/360 Model 91 with water cooling
- Virtualization on the S/360 mainframe in the form of CP67

These two technologies are now used in data centers that run servers with Intel and AMD processors. The use of water for cooling can be up to 3500 times more efficient than air, and virtualization helps consolidate energy that is wasting the low-utilized servers, which are running in data centers around the world. For more information, see this website:

http://searchdatacenter.techtarget.com/news/article/0,289142,sid80_gci1218484,00.html

The more we look back, the more we see that problems we face now are not new and many of the solutions have been tried and tested in one form or another before. The iDataPlex solution takes advantage of IBM knowledge in the data center to provide a best-of-breed solution for the high-density data center.

6.2 Rack specifications

This section provides rack specifications, information about the rack footprint, and the requirements for clearance, transport, and placement.

An iDataPlex rack is a 100 1U-equivalent rack with an EIA standard node-mounting of 84U nodes in two vertical columns and 16U vertical bays for switches or PDUs (1U = 44.5 mm or 1.75 in.). The vertical bays are in two rows, each having four 2U sections above one another. For dimensions, see Table 6-1.

Table 6-1 iDataPlex rack dimensions

Configuration	Width	Depth	Height
Base rack + casters (no doors, or covers)	1200 mm	840 mm	2093 mm
Base rack (no doors, covers, nor casters)	1200 mm	600 mm	2093 mm
Rack + front doors (only)	1200 mm	840 mm	2093 mm
Rack + rear door (only)	1200 mm	844 mm	2093 mm
Rack + side covers (only)	1235 mm	840 mm	2093 mm
Rack + front + rear + side covers	1235 mm	844 mm	2093 mm

To move the iDataPlex rack through a doorway, first review the minimum and recommended dimensions listed in Table 6-2.

Note: The fully populated rack fits under a 2110 mm (83 in.) doorway only if rolled in on the casters (instead of remaining on the shipping crate).

Table 6-2 Doorway dimensions

Dimension	Minimum	Recommended
Doorway height	2110 mm (83 in.)	2133 mm (84 in.)
Doorway width	851 mm (33.5 in.)	900 mm (35.5 in.)

FAQ:

- How tall is the rack; is it taller than enterprise racks; and will it fit through my doorway?

  iDataPlex requires a minimum of 83” door height, so it will fit through a typical 7’ door.

  Doorways and elevators on some old buildings might not be tall enough to fit iDataPlex racks.

  There are no height reduction features, but we can support iDataPlex in an Enterprise rack in these environments.

The weight of an iDataPlex rack depends on the configuration. A typical compute-node configuration comprised of the following devices would weigh 1270 kg (2800 lb):

- Rack
- 42x 2U chassis
- 2 Ethernet switches
- 4 PDUs

An empty rack weighs 250 kg (551 lb).

FAQ:

- My data center cannot support the weight of an iDataPlex rack, what should I do?

  iDataPlex form factor is less weight than an equivalent number of 1U.

  One iDataPlex empty rack weighs up to 35% less than two standard racks.

  A maximum configured 1U can weigh up to 35 lbs; the iDataPlex's equivalent 1U weighs up to 23 lb. Typical weights are lower, but iDataPlex is always less than an equivalent 1U.

  Most data centers have a total floor loading limitation. In these cases iDataPlex is always better than equivalent 1U.

  Data centers also have point loading limitations. iDataPlex with 84 nodes would have greater point loading than a single standard rack with 42 nodes. This is normally only a problem with very old data centers that have wood filled floor tiles versus concrete filled tiles.

6.3 Rear Door Heat eXchanger specifications

This section discusses basic specifications and planning considerations.

6.3.1 Dimensions and weight

The iDataPlex Rear Door Heat eXchanger (part number 43V6048) has the following specifications:

- Dimensions:
  - Width: 1190 mm (46.9 in.)
  - Depth: 120 mm (4.7 in.)
  - Height: 1995 mm (78.5 in.)

The heat exchanger adds 127 mm (5 in.) to iDataPlex Rack depth if the casters are removed.

- Doorway dimensions:
  - Minimum height: 2020 mm (79.5 in.)
  - Recommended height: 2057 mm (81 in.)
- Weight:
  - 82 kg (180 lb) empty
  - 91 kg (200 lb) full, with water

Depending on conditions, the Rear Door Heat eXchanger can extract heat produced by equipment using 30 kWh of energy (while operating at 90 - 94% of the door's maximum efficiency).

The Rear Door Heat eXchanger does not use electricity.

FAQ:

- What happens if I open the Rear Door Heat eXchanger for maintenance?

  iDataPlex is completely front-access, even power, so there is no reason to go to the rear of the rack for maintenance.

  The server cooling is not affected as the heat exchanger cools the data center, not the servers, even if the heat exchanger goes offline.

  The heat exchanger has no boost fans or moving parts, so the failure rate is extremely small and does not require regular maintenance.

  In a reasonably sized data center, a single heat exchanger failure has a negligible effect. Timing of this is in minutes to hours, not seconds.

6.3.2 Planning for the arrival

This section provides information about what to plan for when an iDataPlex rack arrives. For more information, see the *Installation and Planning Guide* provided with the iDataPlex rack:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5075216

Unpacking and acclimation

The iDataPlex must be acclimated to the data center in order to avoid condensation. If the temperature during transport or storage was below the dew point of the data center, condensation can form on the equipment inside the rack. If the rack was shipped or stored in a plastic shipping bag, do not remove the bag until all signs of condensation are gone.

Instructions for rolling the rack off the shipping crate are listed in the *Installation and Planning Guide* provided with the rack. A wooden ramp is provided with the shipping crate so that the rack can be rolled on to the floor.

Weight and transport

A fully populated rack can weigh up to 1360 kg (2998 lb) in special custom configurations. The weight of the rack is distributed over four casters and a margin of safety should be added when moving it across floors to account for changes in the point of gravity when pushing or pulling the rack.

Important: Ensuring that the location of the rack can hold its weight is not often enough to consider. When moving the populated rack into a data center across old raised floors, be prudent and consider using sufficient sheet metal or plywood plates to distribute the load across the floor tiles while rolling the rack to its destination.

Placing the rack

The rack does not require a raised floor, because the power, network, and plumbing can be brought in from above. Figure 6-1 shows an example of the intake and outlet hoses of the Rear Door Heat eXchanger coming from ceiling trays.

Figure 6-1 Rack on a solid floor with flexible plumbing from ceiling (with casters)

When plumbing is through a raised floor, the flexible hosing should be routed from the front of the rack to allow movement in the hosing. See Figure 6-2.

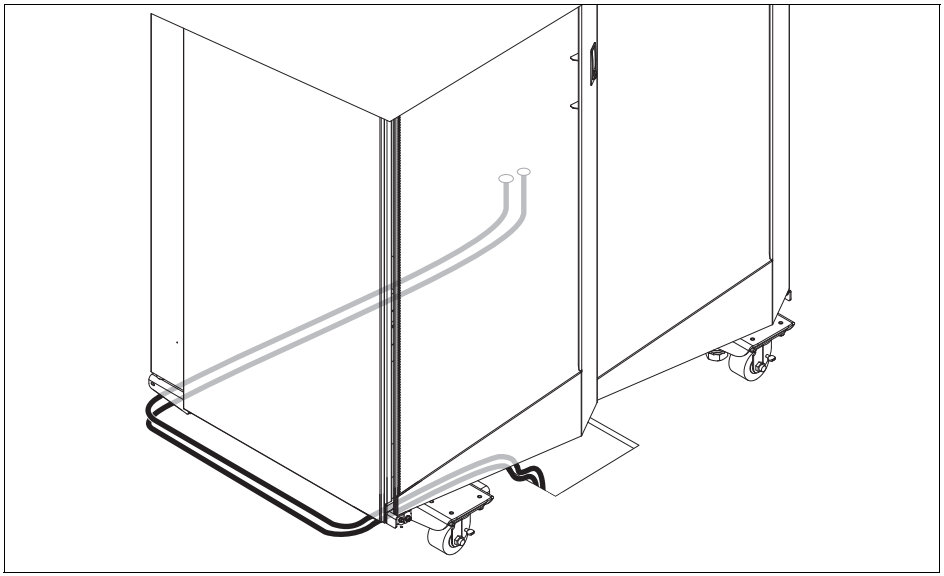


Figure 6-2 Example routing of raised floor Rear Door Heat eXchanger flexible plumbing

Removing the casters is recommended to more easily install the plumbing for the Rear Door Heat eXchanger and to clear the floor of obstructions. See Figure 6-3.

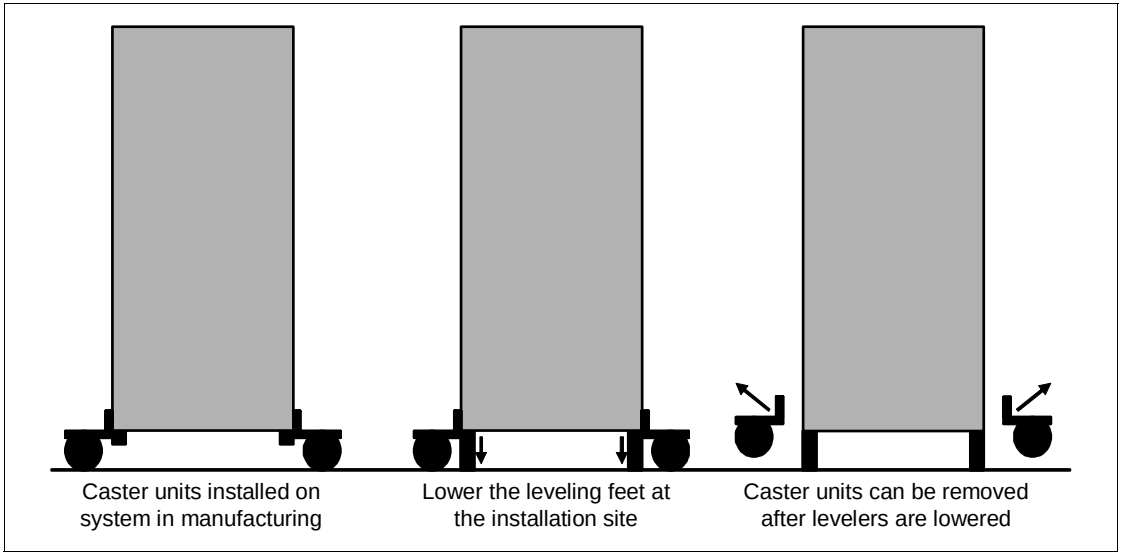


Figure 6-3 Caster removal

When ordering the system, you can specify whether the cabling and plumbing should go to the ceiling or into a raised floor. Depending on the configuration, 1U of space is left blank at either the top or the bottom of the rack to allow the network cabling to pass through the air baffles (*mud flaps*).

Note: The rack actually has 1U extra on each column (84U + 2U) so that the spacer at the bottom or top does not reduce the 84U usable space. The management appliance, when installed, uses up the other 1U spare space

The whole rack can easily be mounted to the floor with four bolts. In areas with seismic activity, bolting the rack directly to the concrete floor. might be necessary.

Note: When the rack is bolted to the floor and cabling is from above, it might be necessary to slide out the bottom nodes to access the bolts, as the shipping bracket that is used to allow for cabling is at the top in this case. When the bolt has been accessed, the bottom node can be re-inserted and reconnected.

The four bolts on the underside of the rack are 872 mm apart from side to side, and 538 mm apart from the front to the back of the rack.

Footprint

Because of its space-saving footprint of 1200 mm x 600 mm (width x depth), the iDataPlex fits on two European floor tiles (European = 600 mm x 600 mm or 23.6 in. x 23.6 in.) as shown in Figure 6-4.

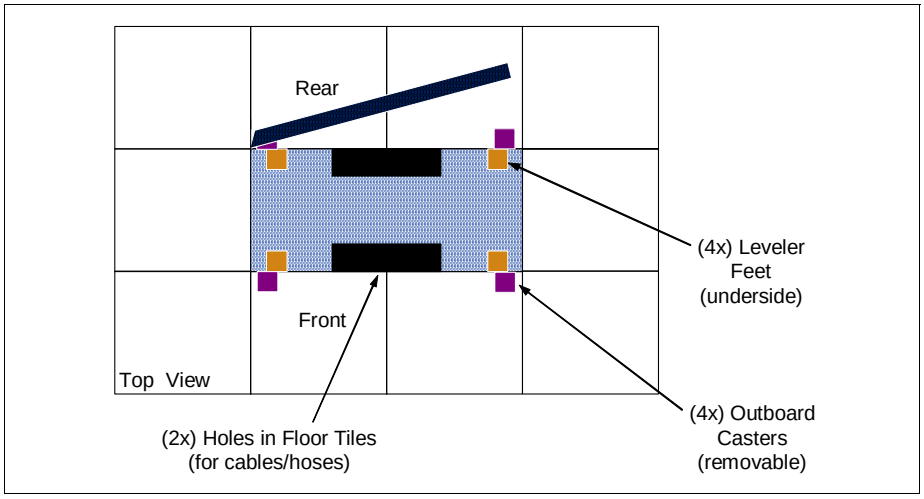


Figure 6-4 Floor tile layout (600 mm2)

On US floor tiles (24 in. x 24 in. or 609.6 mm x 609.6 mm), spacers should be used between racks so as to ensure that the racks are evenly distributed across the tiles.

The racks can also be connected together and an option is available for this when ordering the rack.

Clearance

The space required to open the Rear Door Heat eXchanger fully is 2000 mm (78.75 in.) and the minimum distance from the back of the rack to the next rack or the wall is 480 mm (18.9 in.)

The front of the rack requires 1010 mm (39.75 in.) clearance to allow access by way of the front rack doors. If placed face-to-face, the iDataPlex racks should be spaced 1200 mm (48 in.) apart to allow for the doors.

Figure 6-5 shows the various clearances.

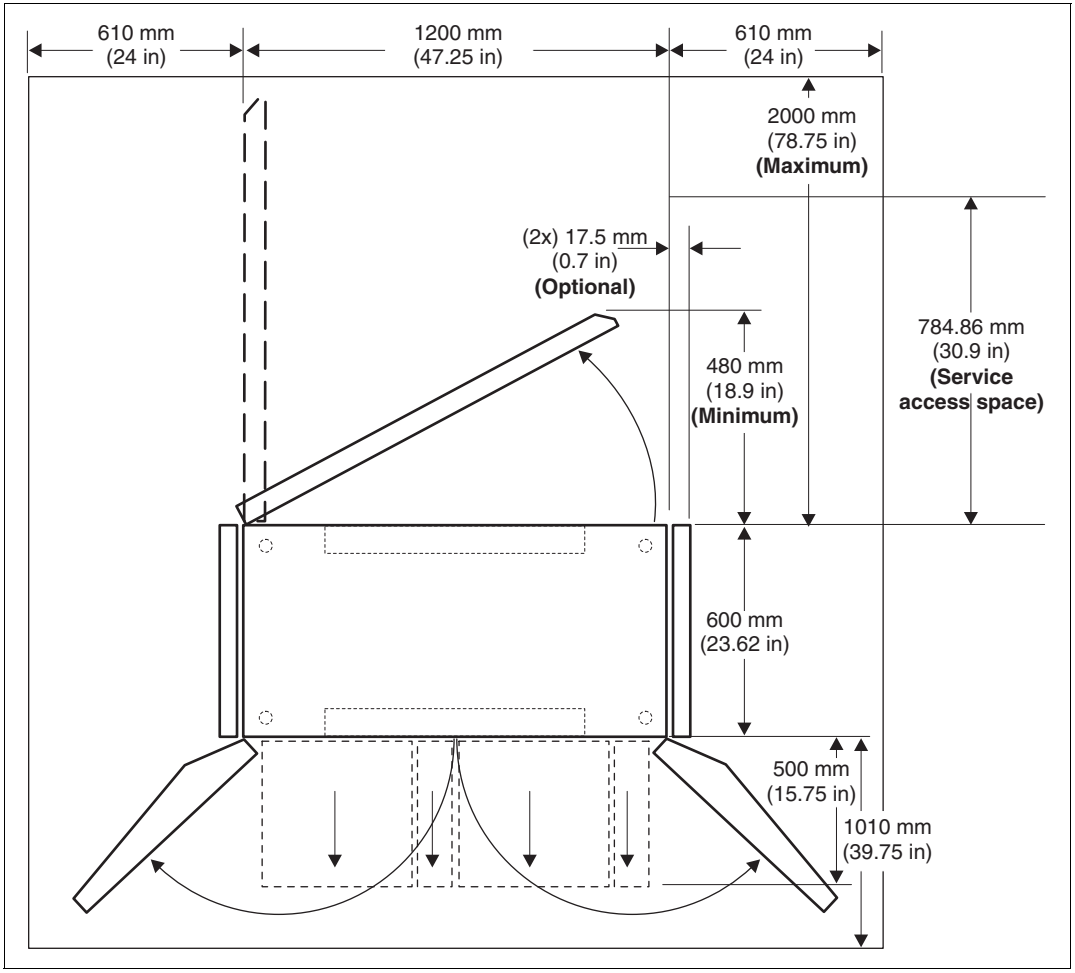


Figure 6-5 Clearance for rack doors

Because the rack is serviceable from the front, and only the power connectors and the PDU LAN ports are connected at the back, opening the heat exchanger is rarely necessary. The door can be added or removed with just under 600 mm (2 ft) clearance. However, someone must be able to get behind the rack to connect the quick-connect hoses to the bottom of the door.

6.4 Cabling the rack

The rack is delivered with the intra-rack Ethernet cabling in place or, in special cases, this cabling is performed at the customer's premises. The cabling in the rack is done with cables of different lengths to avoid excess cables obstructing access to the rack. Additional cabling is required to connect the iDataPlex switches to the customer's infrastructure. The power cables are preinstalled and only the PDUs have to be connected to the customer's power supply.

6.4.1 Nodes

Each node has two 10/100/1000 Ethernet ports and an additional 10/100 management Ethernet port onboard. Also, expansion cards can be added to augment the node with, for instance, InfiniBand, Fibre Channel, or additional Ethernet ports. The nodes also offer two USB 2.0 ports, a DB15 VGA port, and a DB9 serial port.

Note: The LAN connections for the managed PDU and the management appliance are at the rear of the rack and have to be routed through either a raised floor under the rack, or through the top of the rack to plug into the switches in the rack. There is a small opening in the crossbar at the top of the rack through which cables can be passed.

6.4.2 Switches

Ethernet switches are only installed in even positions in columns B and D of the rack. With an all-Ethernet cabling configuration, a handle bracket is used to take up all of the excess cable. The excess cable is looped and placed inside the fillers that are under the handle bracket. Hook and loop straps are used to retain the cables against the handle bracket in neat bundles.

One switch in column B can be connected to nodes on columns A and C. However, a minimum of 1U cabling channel in column C is then reserved for a cabling bracket to route the cables. That space cannot be used by a node.

Generally, the switches are placed and grouped near the center of the rack to allow cables to be routed up and down.

InfiniBand and other special cabling requirements are normally performed onsite.

6.4.3 Power distribution units

PDUs are installed in the back to supply power to nodes, switches, and other equipment. They are only installed in odd positions in columns B and D. The handle brackets are used in the even numbered positions that are next to or below a PDU to manage all the power cords.

PDU sharing is possible across the columns. A PDU in column B can power nodes from columns A *and* C. However, this is only done when there is no switch next to that PDU, because that space is where to put all the excess cables.

6.4.4 Cabling best practices

Given the non-traditional rack and node design in iDataPlex, the cabling practices used in iDataPlex solutions must be different from the traditional cabling practices. IBM has worked closely with a professional physical layer infrastructure company, Panduit Corp. to evaluate the best cabling strategies for iDataPlex solutions to meet the important cable management criteria. These criteria can include signal integrity, system reliability, minimal downtime from equipment failure, allowance for proper airflow, and others. Also consider various practical aspects related to the data center design.

IBM has published an IBM Intelligent Cluster and iDataPlex cabling best practices guide, including many photos, showing the recommended way of cabling. It can be downloaded from this website:

`http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5080515`

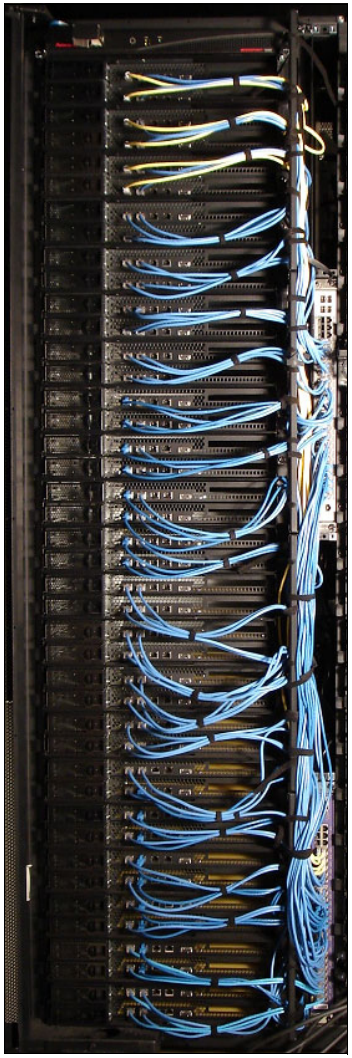
In this section, we highlight several important considerations of *PANDUIT*, and also the IBM mechanical engineering team.

The goal of cable management in the iDataPlex System is to develop and layout a routing strategy that prevents cabling congestion in any area of the rack. It is recommended that the cabling infrastructure be distributed between the top, middle, and bottom of the rack while using two vertical channel spaces (at the middle and at the right side of the rack) for vertical cable runs.

The successful management of copper cables throughout the iDataPlex System rests on the considered deployment of Category 5e Ethernet cables and InfiniBand assemblies. Category 5e cables should be used for Ethernet links, because the diameter of Category 6 and 6A cables can introduce density challenges and potential airflow problems.

In all scenarios involving placement of switches and PDUs in the iDataPlex rack, observe the following rules. See Figure 6-6 on page 182 for examples.

- In the front:
  - Switches should only be installed in the even numbered positions in columns B and D.
  - When there is no switch, a flat filler should be used in even numbered positions in columns B and D.
- In the rear, PDUs should only be installed in odd numbered positions in columns B and D.
- The use of a central vertical channel space to route the Ethernet links from the server to the switch equipment, by using one-foot increments in patch cords, reduces slack by deploying shorter lengths of cable.



	Column A	Column B	
		ODD	EVEN
43		Pocketed or Ducting Filler + Handle Dkt	Flat Filler or Switch
42			
41			
40			
39			
38			
37			
36			
35			
34			
33			
32		Pocketed or Ducting Filler + Handle Bkt	Flat Filler or Switch
31			
30			
29			
28			
27			
26			
25			
24			
23			
22			
21		Pocketed or Ducting Filler + Handle Bkt	Flat Filler or Switch
20			
19			
18			
17			
16			
15			
14			
13			
12			
11			
10		Pocketed or Ducting Filler + Handle Bkt	Flat Filler or Switch
9			
8			
7			
6			
5			
4			
3			
2			
1			

Figure 6-6 iDataPlex Rack with Ethernet switches

- In the iDataPlex system, LC Duplex uplinks from each Ethernet switch should be routed out of the rack to an aggregation layer of the network outside of the iDataPlex rack.

  PANDUIT recommends using a structured cabling approach where all fiber connections route back to a fiber enclosure either inside or outside of the rack. This location is where installers would have the ability to quickly perform moves, adds, and changes to existing fiber infrastructure to better utilize the

existing data center cabling plant. From this enclosure, you can then run distribution cables overhead or under-floor through the cable routing systems (such as the overhead system shown in Figure 6-7), which then routes and protects this horizontal cable with little or no changes ever made to it. These pathways can also be used to route CAT5e and InfiniBand cables between iDataPlex racks within the data center environment.

Figure 6-7 Overhead cable routing with pathway spill-outs for cable drops (PANDUIT)

- With an all-Ethernet cabling configuration, the handle bracket can be used to take up all excess cable. This cable can be looped and placed inside the fillers that are *under* the handle bracket. Hook and loop straps are used to retain the cables against the handle bracket and in neat bundles.

6.4.5 iDataPlex overhead cable pathways

PANDUIT provides innovative solutions for network cable management in data centers. Although most of these solutions are generic in nature, they could be effectively employed in custom scenarios as well. IBM has tested the *PANDUIT* FiberRunner Overhead Cable Pathways as a cable management solution for iDataPlex. In this section, we provide details of this solution from *PANDUIT*. Customers who are interested in using this solution in their data centers, for better cable management for iDataPlex, can consult their IBM iDataPlex sales

representative who will engage the *PANDUIT* engineering team to assist with the solution.

PANDUIT Overhead Cable Pathways can be either integrated directly into the iDataPlex solution or be independent of the racks. Either option can be easily implemented based on each customer's individual requirements.

An example of an integrated pathway solution is shown in Figure 6-8. This example shows the top view floor layout, consisting of two IBM Enterprise racks and two IBM iDataPlex racks. The two Enterprise racks reside within the middle of this four rack lineup with one iDataPlex rack on either side of the Enterprise racks.

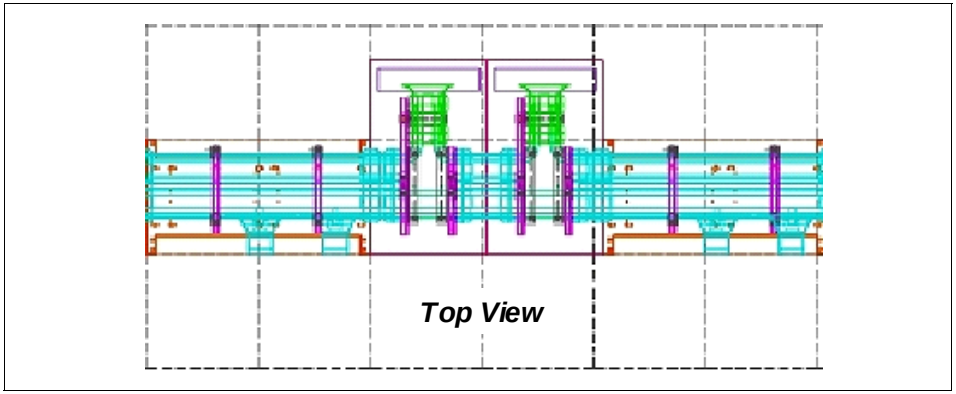


Figure 6-8 PANDUIT Overhead Pathways (top view)

Two product kits allow for fast deployment of the overhead pathways. These kits minimize the time necessary for the facilities group to install the pathways because the pathways bolt directly to the racks and are specifically sized to fit a single rack solution. These pre-kitted solutions eliminate traditional field fabrication of pathway system components, thereby decreasing the time to deploy the overall system. One kit is specific to the Enterprise rack and the other is specific to the iDataPlex rack solution.

Table 6-3 lists the *PANDUIT* part numbers for the two Overhead Pathway kits.

Table 6-3 PANDUIT Overhead Pathway Kits

PANDUIT part number	Description
FRIBMKT01BL	IBM iDataPlex Rack Overhead Pathway Kit
FRIBMKT02BL	IBM Enterprise Rack Overhead Pathway Kit

Figure 6-7 on page 183 shows the *PANDUIT* overhead cable pathway system deployed with an iDataPlex solution.

6.5 Data center power

The data center uses power to operate servers, provide cooling, provide lighting, provide uninterruptible power, and so on. This section discusses the power requirements of just the IT equipment (specifically the iDataPlex solution) as opposed to total data center power requirements.

6.5.1 Data center efficiency

When discussing power in the data center, values for Power Usage Effectiveness (PUE) and Data Center Infrastructure Efficiency (DCiE) are often used to indicate how efficient the data center is. One part of the equation is the IT equipment power. These values are calculated as shown in Figure 6-9. A data center with high power efficiency is one with a DCiE of more than 60%.

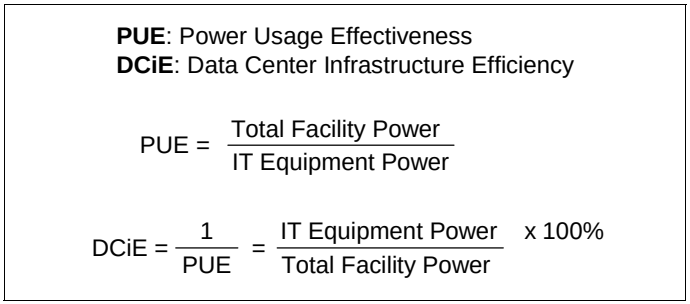


Figure 6-9 Power Usage Effectiveness and Data Center Infrastructure Efficiency[1]

Several service offerings in Chapter 8, “Services offerings” on page 233 can help you better understand how to improve the PUE and DCiE of your data center. Although the iDataPlex racks are part of the solution, note that IT equipment power normally accounts for only 45% of the total facility power.

IT equipment that uses less power requires less energy for cooling which frees up more energy for further IT equipment and can also free up floor space used by cooling equipment.

[1] For more information, see:
http://www.thegreengrid.org/sitecore/content/Global/Content/white-papers/The-Green-Grid-Data-Center-Power-Efficiency-Metrics-PUE-and-DCiE.aspx

6.5.2 Aggregation and reduction of data center power feeds

As we mentioned in a previous chapter, one U.S.-based power company was charging its customers USD1500 - 2000 per month for each 8.6 kVA feed, in addition to the power consumption charges. The customer indicated that reducing the number of feeds per rack from four to three (an iDataPlex rack only requires three 8.6 kVA feeds versus four feeds with traditional enterprise racks) would save them about USD14 million per year. This is a good example of how cost saving can be achieved by repackaging.

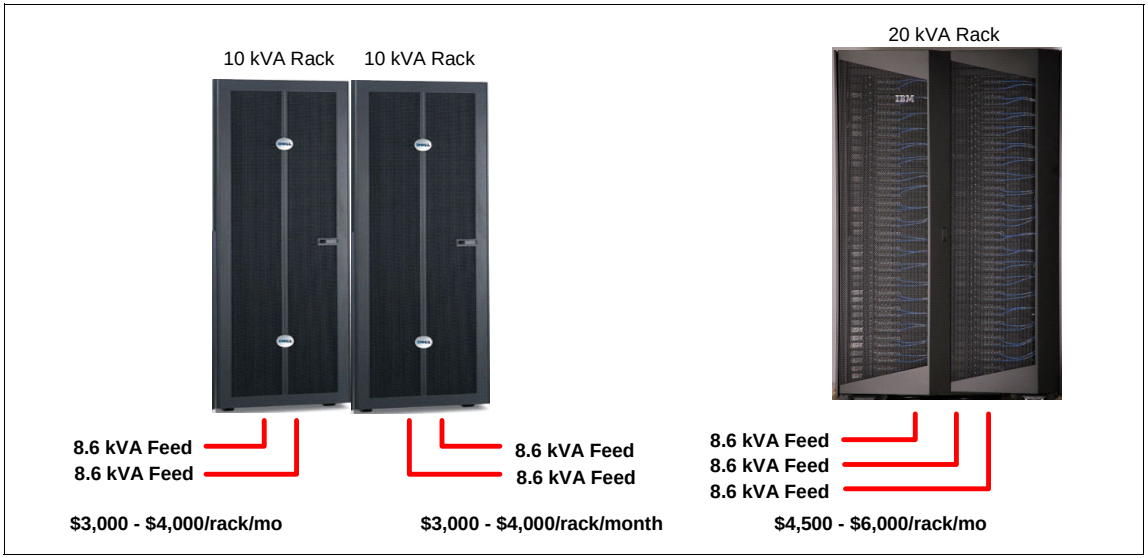


Figure 6-10 Power feed cost example - Two enterprise racks versus one iDataPlex rack

The two standard racks in Figure 6-10 are each 42U high with 1U servers which makes 84 systems in total. The total power feed costs are approximately USD6000 - 8000 per month. The time and manpower charged to deploy is for two racks. Since each rack only consumes 10 kVA, this results in a total of 14.4 kVA of unused (stranded) power capacity across the two racks that is still billed to the customer. This effectively raises the price per watt of power for those racks.

The iDataPlex contains 84 systems sharing power. The total power feed costs are USD3000 - 4000 less per month when compared with the 1U servers. The time and manpower to deploy is typically charged by the rack, so iDataPlex is potentially half the charge. The three feeds leave approximately 5.8 kVA of unused (stranded) power capacity, which is less than half the 14.4 kVA for the standard racks.

Another important note is that the two 42U racks have a total of 84U and the iDataPlex rack has 100U total space, which provides a further advantage to the customer.

FAQ:

- My data center can only support up to 15kW of power per rack. Do I have to only partially populate the iDataPlex rack?

  No. The iDataPlex is equal to two standard racks. When you compare the iDataPlex 1U servers to the same number of typical 1U servers, iDataPlex will be lower power by up to 40%.

  The innovative rack design allows for less stranded power. A typical customer workload of 12kW per rack would require two 30A three phase feeds per rack or four for 84 servers at 24kW. However, with iDataPlex, we can power the same number of servers with three 30A three phase feeds for 24kW per rack.

6.6 Data center design using air cooling

The traditional data center caters for a mix of machines of different sizes and power densities, and traditionally cooling is laid out to cool a wide range of rack densities in the same room. Many data centers are 10 - 15 years old, however, and the cooling facilities were designed to process the heat generated from racks consuming 2 - 3 kW of power, but not necessarily the levels generated by 20 - 30 kW of power per rack.

Using the traditional N+1 perimeter CRAC air-conditioning unit design means that the locations of the units and the venting has to be carefully planned to reduce the size and number of hot spots. A common method with the perimeter design is to run the entire computer room colder than it has to be. This manner can lead to the CRACs operating under the dew point, which requires dehumidifying the air. Humidifiers are then required to reintroduce humidity to the air that is coming from the CRACs. This approach means wasted energy and additional power for equipment to re-condition the conditioned air. It can also lead to static build-up if the air in the data center is too dry.

In-row cooling brings active airflow components into the rows of racks and thereby allows hot spots to be dealt with on a row-by-row basis. A whole range of different solutions have been devised over the years and most are air-flow based. For instance, the scalable modular data center, described in 8.1.3, “Scalable modular data center” on page 236, uses in-row cooling.

The Rear Door Heat eXchanger brings cooling granularity down to rack-level. However, instead of air as the conduit, water is used to extract the heat, which is is significantly more energy-efficient. In fact, depending on the water temperature and flow rate, the iDataPlex Rear Door Heat eXchanger can actually cool the egress air to below the ingress temperature, thereby actually cooling the air around it and reducing the need for traditional CRAC units.

As the Rear Door Heat eXchanger operates above the dew point in normal operation, it also cuts the need for re-humidification. With the Cool Blue® Rear Door Heat eXchangers for standard 19-inch racks and the iDataPlex Rear Door Heat eXchanger, a large degree of flexibility is still available within the rack and standard equipment can be installed. Additionally, cooling at the source of the heat load actually lowers your incremental energy consumption and can help reduce the floor space required for air conditioning.

Figure 6-11 shows how typical data center air conditioning removes heat energy (the red arrows) from IT equipment (the orange box).

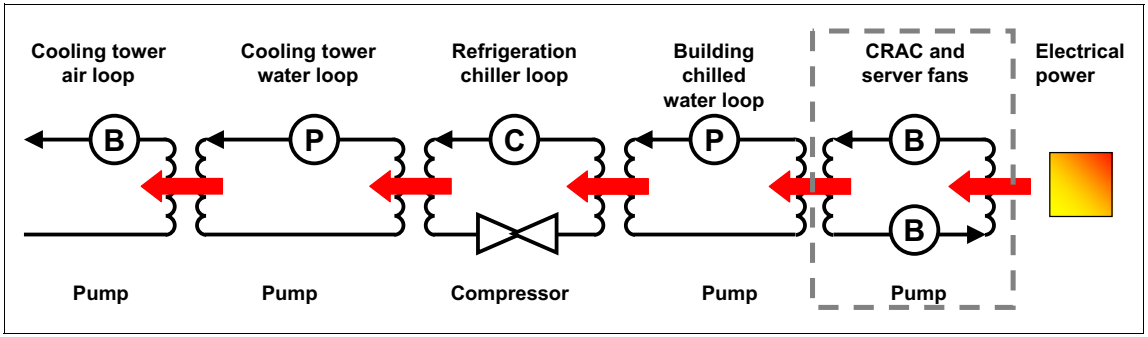


Figure 6-11 Example data center cooling loop—from server to the outside air

With a Rear Door Heat eXchanger, the CRAC unit (the dotted grey box in Figure 6-11) is exchanged for a Cooling Distribution Unit (heat exchanger and pumps depicted in Figure 6-12 on page 189). The CDU measures the room's humidity and temperature and regulates the temperature of water in the closed loop that is flowing to the Rear Door Heat eXchanger to above the dew point.

Note: Three Rear Door Heat eXchangers plus one CDU is approximately equivalent to one conventional CRAC.

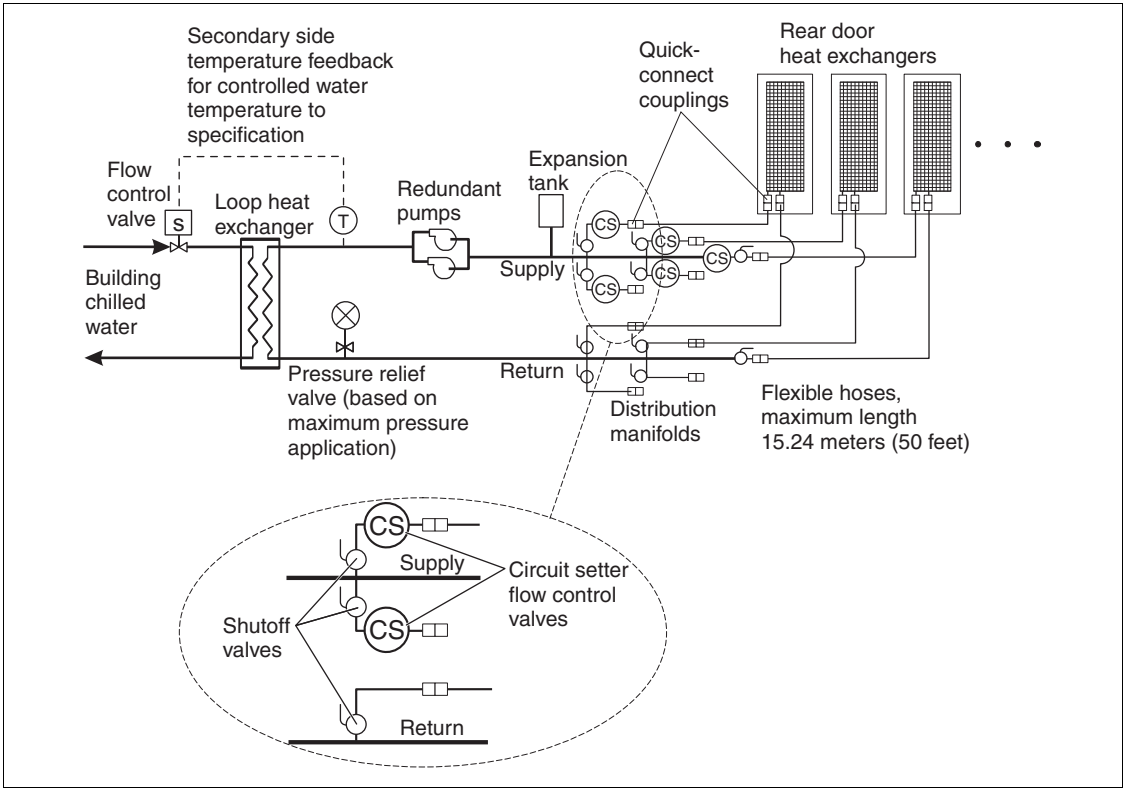


Figure 6-12 Closed loop to Rear Door Heat eXchanger and interface to building chilled water supply

FAQ:

- I can only cool xx kW per rack, but iDataPlex is 37 kW?

  The iDataPlex can power up to 37 kW. Most configurations will be significantly less.

  Datacenter cooling should be sized to the workload and not the maximum rated power. You could be wasting the cooling energy that could be allocated to computing energy.

  The iDataPlex total power is less than the equivalent number of rack units in two standard racks. The iDataPlex form factor is rotated 90 degrees to a standard rack, so iDataPlex has the same frontal area and access to floor tiles as two standard racks.

A further addition to data center cooling solutions is the Data Center Stored Cooling Solution (also known as the Cool Battery), which is an IBM-patented thermal storage solution designed to provide energy cost savings of up to 30% of peak energy usage by shifting energy usage to off-peak hours.

The system works with phase change material that goes from liquid to a semi-solid state in order to store cooling capacity for use during peak hours. The Data Center Stored Cooling Solution has the added advantage of effectively using the environment to augment the data center cooling capacity by using night and day or summer and winter temperature differences when the system is installed outside the building. This approach adds cooling capacity and enables growth while also augmenting the data center's power backup systems.

For more details, see this website:

http://www.ibm.com/press/us/en/pressrelease/21524.wss

6.7 Data center design using Direct Water Cooling

The latest generation of iDataPlex servers support a new node-level cooling technology called Direct Water Cooling. It pumps water to every node and through a water loop assembly installed in every node. This is in contrast to the Rear Door Heat eXchanger which brings water only to a rack-level.

Figure 6-13 shows the cooling loop when using Direct Water Cooling technology. The red arrows indicate heat flow away from the heat source (orange box), such as microprocessors or memory modules.

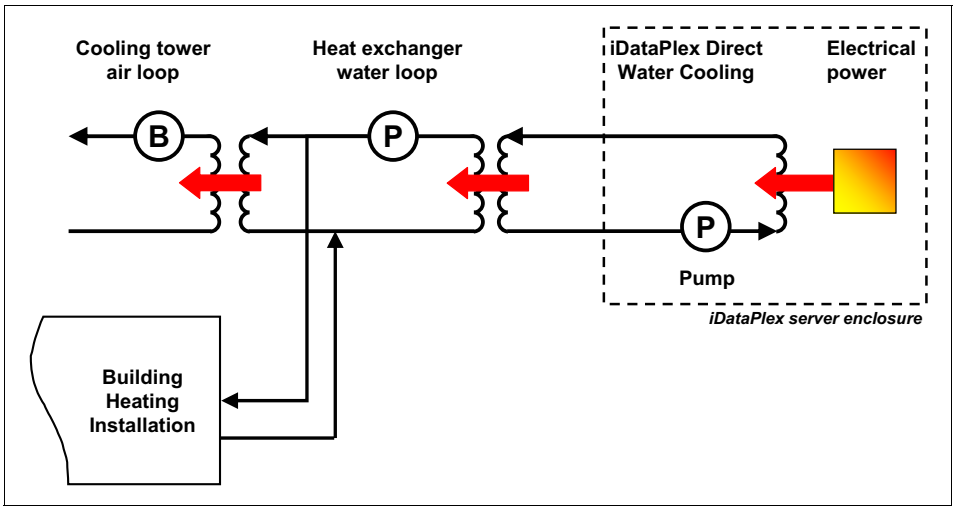


Figure 6-13 Data center cooling loop using Direct Water Cooling technology

In Figure 6-13, there are only two major loops involved, a closed primary loop (right end) going through each server and a secondary loop (depicted as heat exchanger water loop). The secondary loop can either dissipate its heat into free air through a cooling tower (left end) or the secondary loop can be directly connected to a building heating installation. This allows a direct reuse of server waste heat to warm adjacent buildings.

Did you know?

With iDataPlex Direct Water Cooling technology, you can run your data center in compliance with the best liquid cooling class W5, as defined by the ASHRAE technical committee 9.9 in 2011.

Class W5 requires cooling loops to run without chillers and have water inlet temperatures higher than 45°C (113°F) to be able to heat buildings with the water coming out of the data center.

You can download the paper *2011 Thermal Guidelines for Liquid Cooled Data Processing Environments* free of charge from this website:

`http://www.tc99.ashraetcs.org`

The ASHRAE paper also describes the general infrastructure required to drive liquid cooled data centers (of different classes), and it talks about water flow rates and water quality.

6.8 Airflow

This section describes some different approaches to using the data center air to cool components.

In the traditional perimeter design[2], the air in the data center is cooled and forced into the raised floor. Perforated tiles in front or under racks allow the air to be circulated through the equipment. Commonly, two rows of racks are placed so that the hot air from two adjacent rows is collected between the rows in a so called *hot aisle*. The aisles that provide the cold air for the rows of racks are referred to as *cold aisles*.

Tip: To maximize airflow efficiency, a clear air path has to be established under the raised floor and on the return back to the CRAC units. Also, perforated tiles should be placed in cold aisles only. No perforated tiles should ever be placed in a hot aisle.

The hot and cold aisle design allows much better airflow management than randomly spaced and placed racks as practiced in some older data centers.

FAQ:

- iDataPlex has non-redundant cooling, so what happens when a fan fails?

  iDataPlex does not have redundant cooling. However, depending on the workload and environment, a fan failure has little impact and the server would keep running.

  The iDataPlex Fan Speed Control Algorithm takes into account fan failures in the most efficient way possible. iDataPlex logs a fault and continues operating as normal with a fan failure. It will monitor component temperatures and increase the fan speed only when necessary.

  If workload demand is high, the system might throttle, but continues operating until a repair action can be scheduled. In some cases after a fan failure, the system does not have sufficient cooling and will perform a soft shut down to prevent damage to components.

[2] See ASHRAE publication *Thermal Guidelines for Data Processing Environments* for further background information: http://tc99.ashraetcs.org/

6.8.1 Heat pattern analysis without the Rear Door Heat eXchanger

In this section, we use thermal imagery to show the heat patterns generated by 1U servers in traditional enterprise racks and the heat generated by iDataPlex racks. We look at dense, very dense, and extremely dense configurations as well as iDataPlex configurations with and without the Rear Door Heat eXchanger units installed.

We describe the heat patterns on configurations with the Rear Door Heat eXchanger in 6.8.5, “Heat pattern analysis with the Rear Door Heat eXchanger” on page 201.

We define the terms dense, very dense, and extremely dense in Table 6-4, where the tile pitch is the number of 600 mm x 600 mm floor tiles from the left edge of one cold aisle to left edge of the next cold aisle. For example, Figure 6-14 shows an 8-tile pitch.

Table 6-4 Density definitions

Data center density	Tile pitch
Dense	8-tile pitch
Very dense	7-tile pitch
Extremely dense	6-tile pitch (or less)

FAQ:

- Do I get efficient cooling without the Rear Door Heat eXchanger?

  Yes. At the data center, the server cooling is not affected by the Rear Door Heat eXchanger.

  The server cooling is affected by the low impedance server and rack design, shared fans, and fan speed algorithm. They all work the same with or without the Rear Door Heat eXchanger.

Data center with 1U servers in enterprise racks: 8-tile pitch

Figure 6-14 is a thermal map of a dense data center with 4500 1U servers, with each rack operating at 18.9 kW. The airflow is at 1680 cfm (cubic feet per minute). Total power consumption is just over 2000 kW. The hot aisle is 1 m (40 in.) wide while the cold aisle is 1.8 cm (72 in.) wide, equating to an 8-tile pitch.

The result is that 100% of the racks are operating at or below 25°C (77°F).

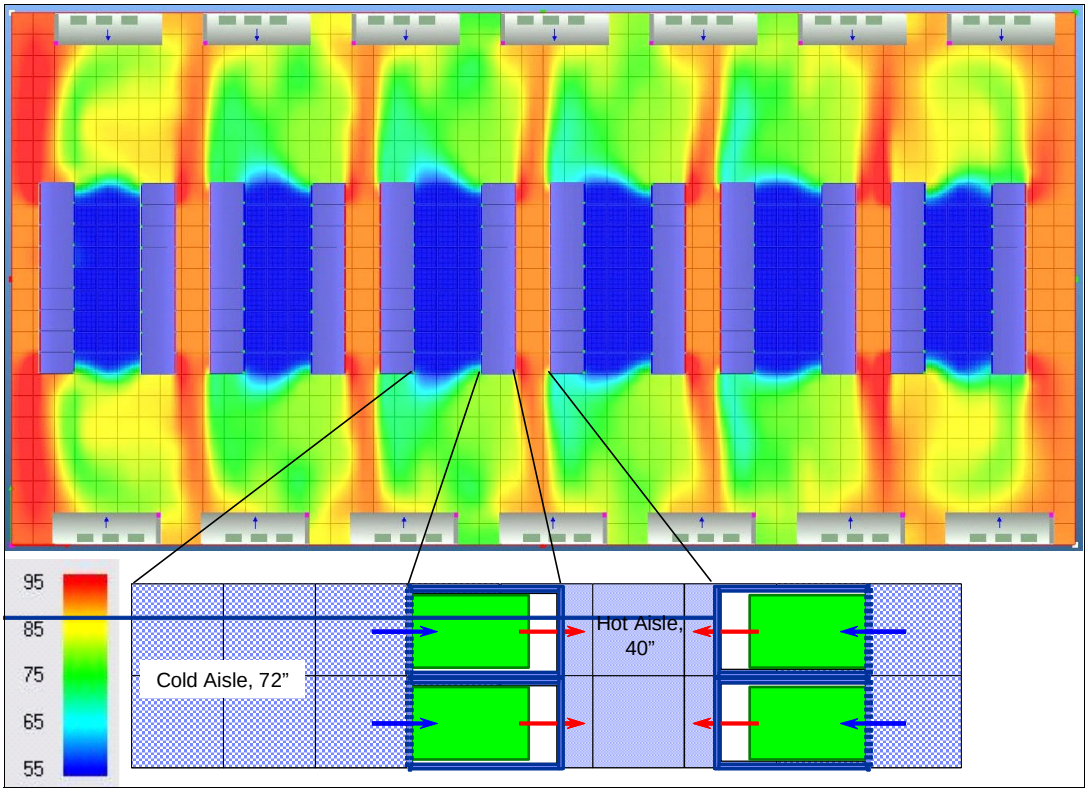


Figure 6-14 Thermal map of a dense data center with hot/cold aisle design (temperatures are in °F)

Data center with 1U servers in enterprise racks: 7-tile pitch

Figure 6-15 shows 9400 1U servers (double the number in Figure 6-14 on page 194), in a very dense data center, with each rack operating at 18.9 kW. The airflow is at 1680 cfm. Total power consumption is 4211 kW. The hot aisle is 1 m (40 in.) wide while the cold aisle is 1.21 m (48 in.) wide, equating to a 7-tile pitch.

The result is that only 12% of racks are now operating at or below 25°C (77°F).

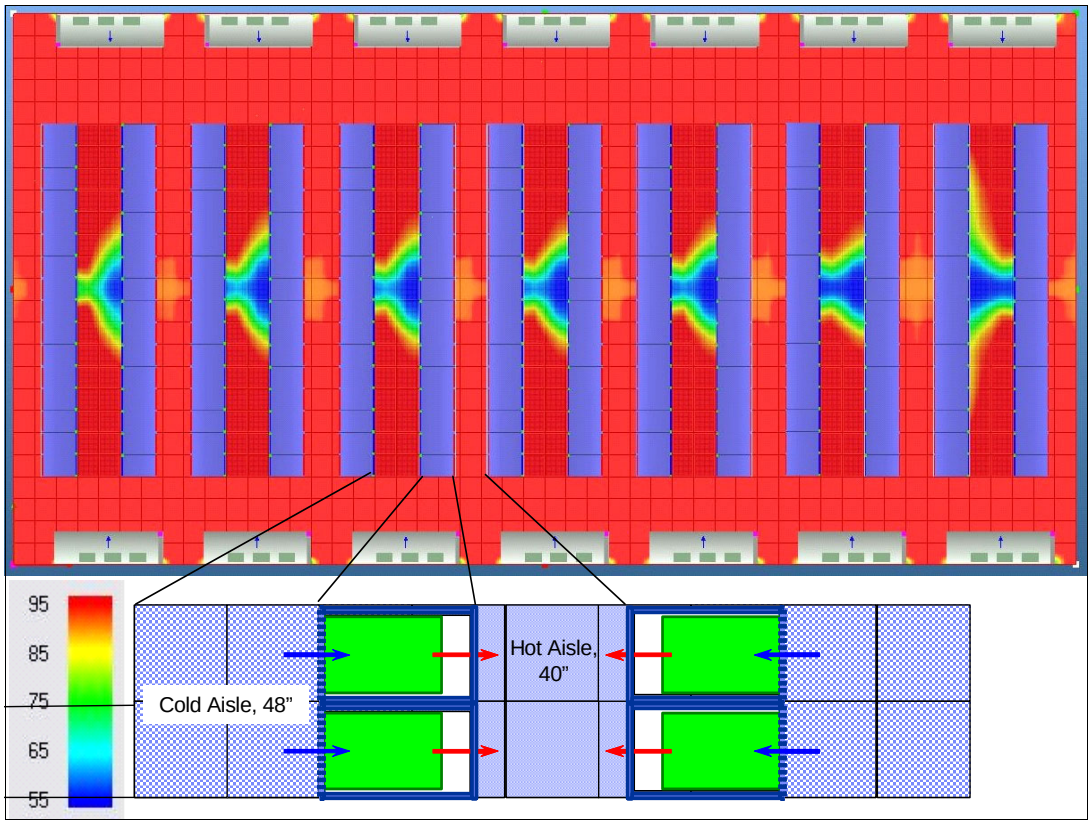


Figure 6-15 Thermal map of a very dense data center with hot/cold aisle design (temperatures in °F)

The thermal map of the very dense data center in Figure 6-15 shows that the traditional perimeter design is unable to adequately cool the number of servers installed. There is too much power and not enough cooling. This configuration is limited to how far it can scale.

Standard 1U configurations are limited by:

- 1U fan and power efficiencies
- CRAC cooling and airflow capacity

Data center with iDataPlex racks: 8-tile pitch

The thermal map in Figure 6-16 shows approximately 5000 iDataPlex nodes, in a dense data center, with each rack operating at 33.6 kW. In this configuration, the iDataPlex racks do not have Rear Door Heat eXchanger units installed.

The airflow is at 2352 cfm. Total power consumption is 2016 kW. The hot aisle is 1.82 m (72 in.) wide and the cold aisle is also 1.82 m (72 in.) wide equating to an 8-tile dense pitch.

The result is that 100% of the racks are operating at or below 25°C (77°F).

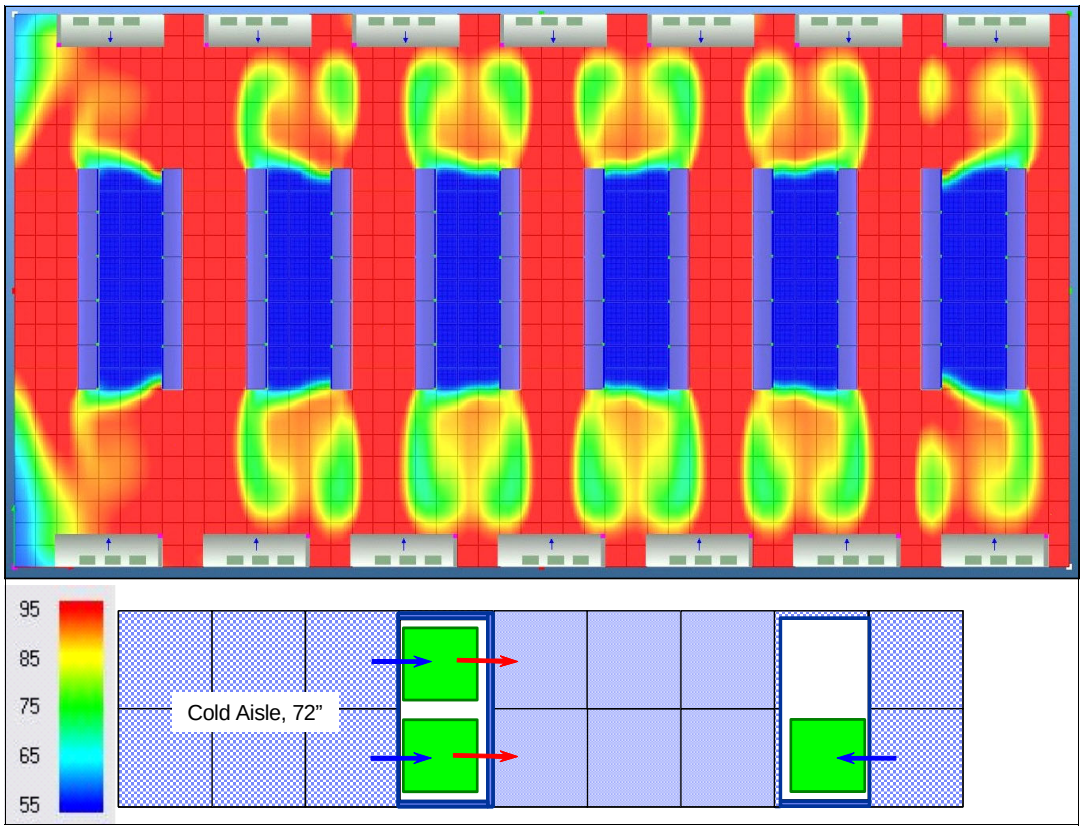


Figure 6-16 iDataPlex servers in a dense data center with traditional hot/cold aisle design (°F)

Data center with iDataPlex racks: 7-tile pitch

Figure 6-17 shows 9,408 iDataPlexServers, in a very dense data center, with each rack operating at 33.6 kW. The airflow is at 2352 cfm. Total power consumption is 3763 kW. The hot aisle is 1.21 m (48 in.) wide while the cold aisle is also 1.82 m (72 in.) wide equating to a 7-tile pitch. In this configuration, the iDataPlex racks do not have Rear Door Heat eXchanger units installed.

The result is that 45% of the racks are operating at or below 25°C (77°F).

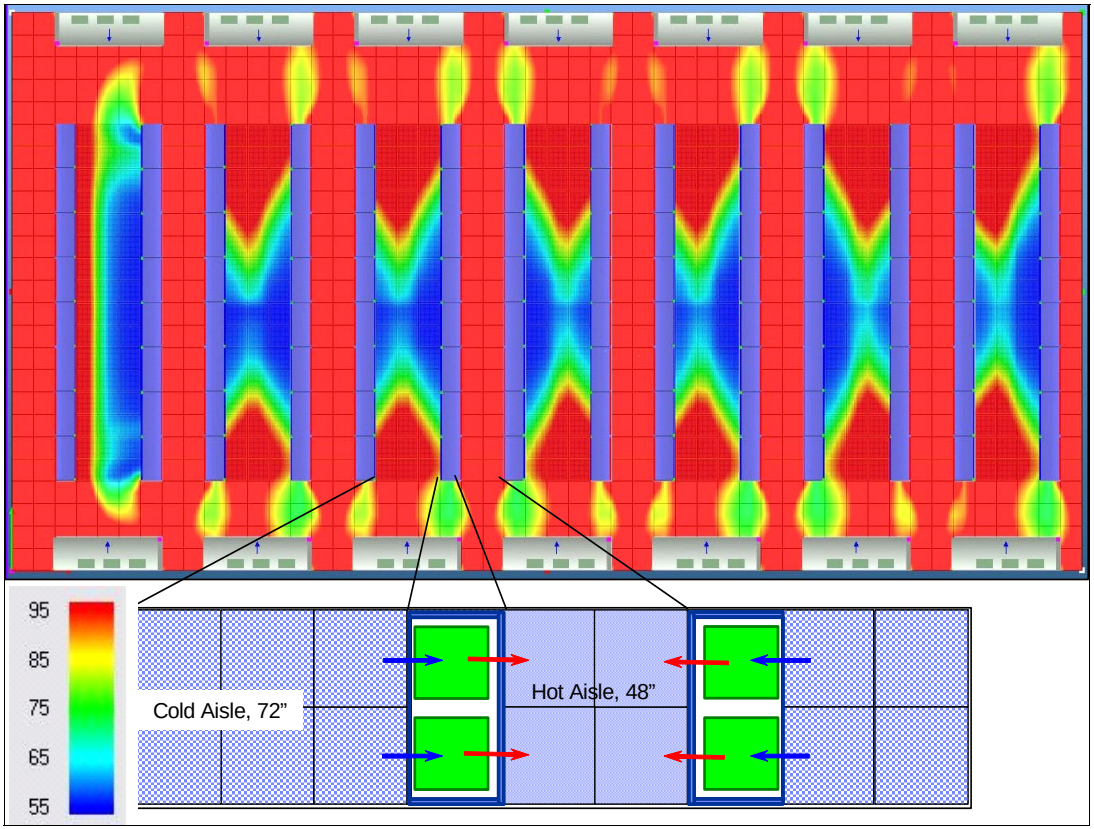


Figure 6-17 Servers in a very dense 5000 sq. foot data center with traditional hot/cold aisle design (°F)

The very dense (7-tile pitch) iDataPlex data center is somewhat limited but is still able to adequately cool almost half the servers (45%), which shows the iDataPlex capacity to scale-out better than standard 1U servers that can only cool 12% adequately. The 45% is most likely still unacceptable to most customers because the remaining 55% are potentially running above acceptable temperature ranges for the devices, which indicates that there are limits to what traditional perimeter cooling designs can achieve.

The previous iDataPlex configurations are limited by:

- CRAC cooling and airflow capacity
- Room cooling

Heat removal is more efficient and more effective with the use of Rear Door Heat eXchanger units installed on the iDataPlex racks. We discuss scenarios with the Rear Door Heat eXchanger units installed in 6.8.5, “Heat pattern analysis with the Rear Door Heat eXchanger” on page 201.

6.8.2 IBM Systems Director Active Energy Manager

In the previous scenarios, where heat management has the potential of causing problems to the systems in the data center, using the IBM Systems Director Active Energy Manager™ to implement power capping on certain nodes for the purpose of keeping the entire data center operational is possible. Even though power capping reduces the number of available CPU cycles, it can be an alternate way of reducing the heat load in the hot spots of the data center.

While power capping might not be the tool of choice in a high performance cluster, or a hosting environment, it can be useful for emergency situations, for batch processing, or for scheduled workloads that are not time-critical. Even power capping cannot overcome all cooling problems, however, and therefore does not replace a solid data center design.

For more information about IBM Active Energy Manager, see *Implementing IBM Systems Director Active Energy Manager 4.1.1*, SG24-7780:

http://www.redbooks.ibm.com/abstracts/sg247780.html

6.8.3 Hot air ducts

Ducts can be used to prevent recirculation of air between the hot and cold aisles. This is a logical progression from the hot-cold aisle configuration and allows the CRACs to operate at higher efficiencies by increasing the temperature of the return air. Essentially, this optimized form of the hot and cold aisle design allows a much more dense data center. Similar concepts use covered aisles.

One advantage that the ducting can have is that the use of a raised floor can be avoided in certain circumstances. Unless implemented in a new data center, this type of design can mean a complete redesign of existing data centers and is therefore limited in its use.

Figure 6-18 shows an example of hot-air ducting with iDataPlex racks.

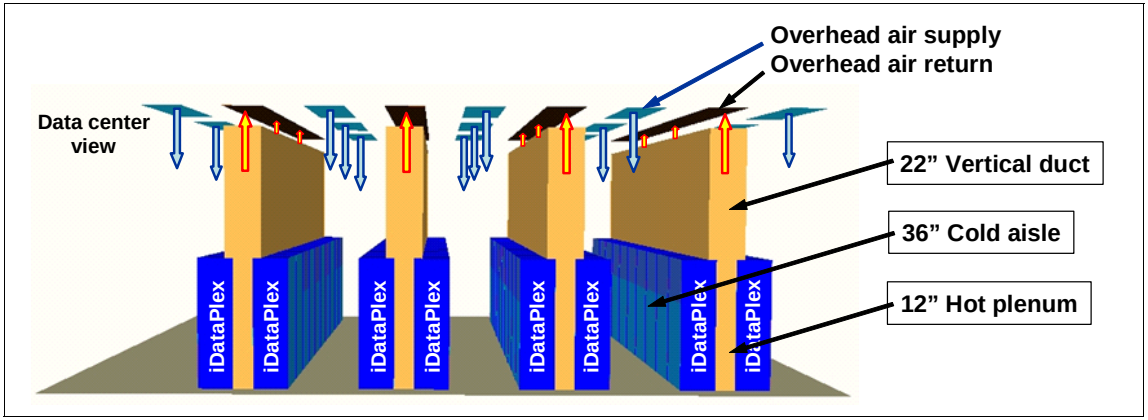


Figure 6-18 Data Center scenario with vertical ducts

Figure 6-19 shows a thermal representation of the separation provided by the hot air ducts. The iDataPlex provides a further advantage over standard enterprise racks in this scenario because accessing the rear of an iDataPlex rack is rarely necessary; nearly all cabling is done from the front and the tile pitch can therefore be further reduced.

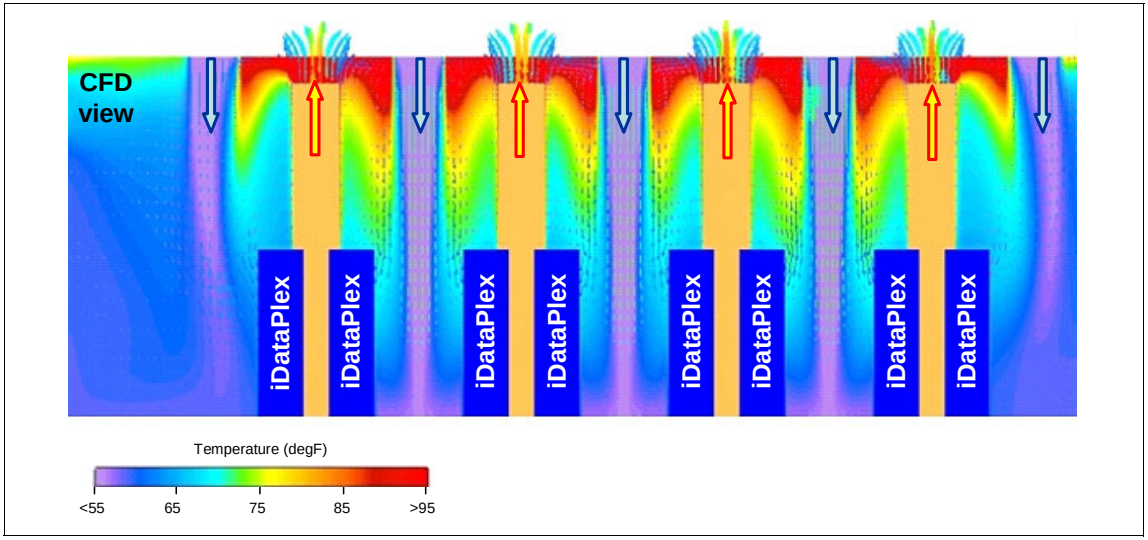


Figure 6-19 Example Computational Fluid Dynamic representation of hot aisle ducts

6.8.4 iDataPlex and standard 1U server rack placement

The volume of air in the cold aisle can be reduced significantly when the Rear Door Heat eXchanger units are installed on iDataPlex racks, because the racks no longer heat the room. The thermal buffering, or airflow separation, of the cold aisle air volume is no longer required. Another way of seeing this is that the recirculation, or mixing, of hot and cold air no longer happens on the same scale as in traditional designs and therefore the cold aisle air temperature is not raised by recirculating air from hot aisles as these are no longer as hot as in traditional designs.

The placement of mixed rows with iDataPlex racks requires special consideration when Rear Door Heat eXchanger units are used, because the temperature of mixed (iDataPlex and enterprise rack) hot aisles are less, thereby reducing the efficiency of the CRAC units servicing these rows. The Rear Door Heat eXchanger does reduce the load on the CRAC units, separating the rows with Rear Door Heat eXchangers from those without to allow the CRACs to operate more efficiently might still be sensible.

Figure 6-20 shows the different cold aisle areas recommended for the comparable rack configurations. The iDataPlex requires less cold aisle volume than a row of enterprise racks with 1U servers, while the iDataPlex with Rear Door Heat eXchanger requires the least cold aisle volume.

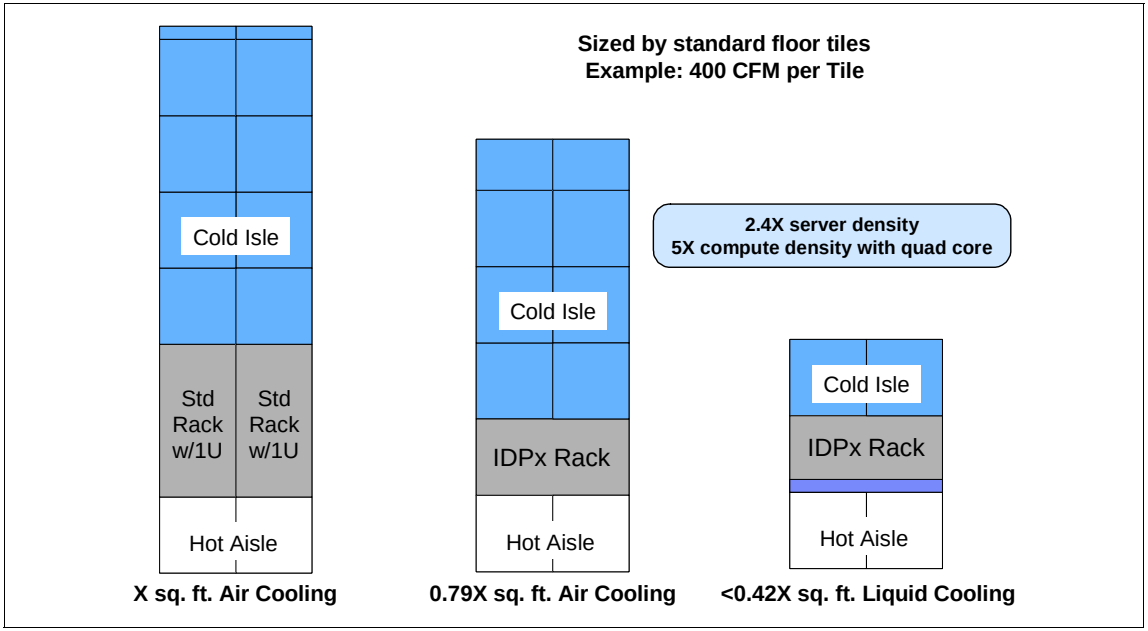


Figure 6-20 Narrow depth rack improves data center density - no change to data center layout

Note that the iDataPlex allows for two times the nodes in a similar footprint with over two times the EIA spaces[3].

6.8.5 Heat pattern analysis with the Rear Door Heat eXchanger

In this section, we use thermal imagery to show the heat patterns generated by iDataPlex racks with a Rear Door Heat eXchanger. We look at extremely dense (6-tile pitch) configurations and configurations that have not been previously viable.

[3] The 19-inch pockets, as described in Electronic Industries Alliance (EIA) Standard EIA/ECA-310

Extremely dense data center design with Rear Door Heat eXchangers

The thermal map shown in Figure 6-21 is a configuration of more than 10,000 servers in iDataPlex racks, in an extremely dense data center. Each rack is operating at 33.6 kW. The airflow is at 2352 cfm. Total power consumption is 4301 kW. The hot aisle is 91cm (36 in.) wide and the cold aisle is 1.21 m (48 in.) wide equating to a 6-tile pitch.

The result of using the Rear Door Heat eXchanger units is that 100% of the racks are operating at or below 77°F. The exit air temperatures are uniform across the iDataPlex rack height.

In this scenario, the Rear Door Heat eXchanger units are removing 90% of the exhaust heat while CRACs are extracting the remaining heat.

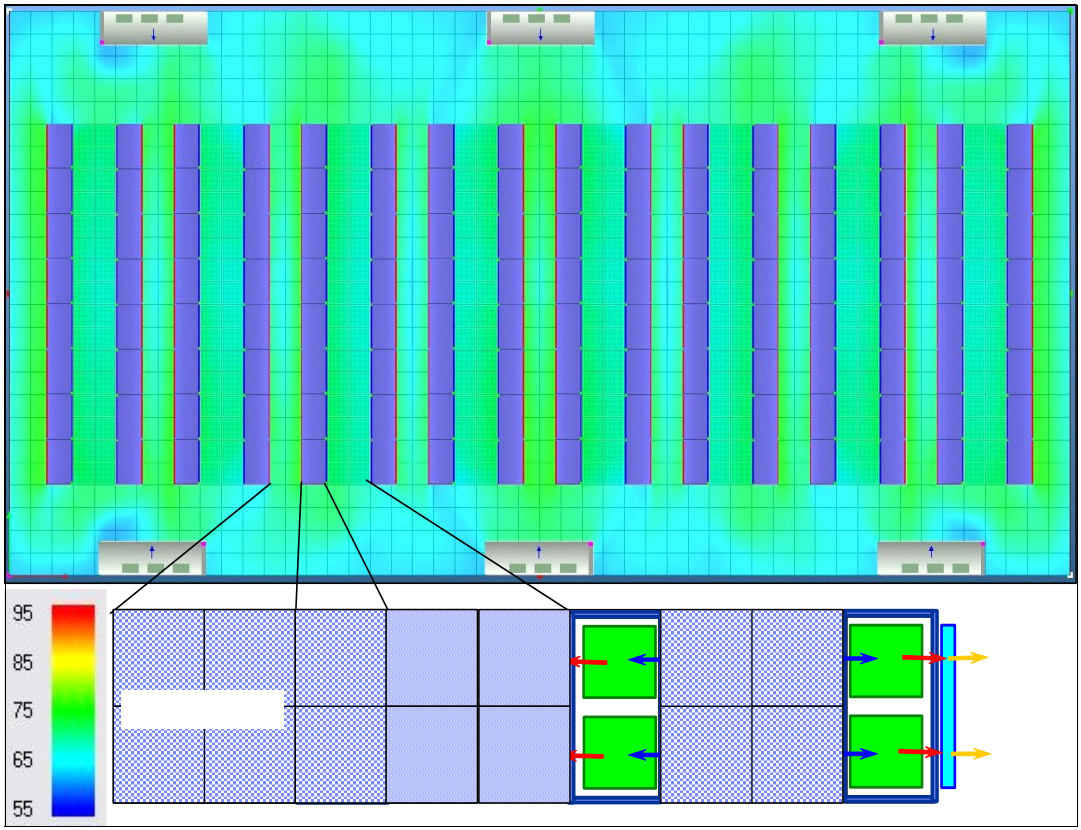


Figure 6-21 Servers with traditional hot-cold aisle design and Rear Door Heat eXchanger units (°F)

Circular airflow with Rear Door Heat eXchanger

Figure 6-22 shows a circular air flow in the data center with intentional recirculation through Rear Door Heat eXchanger units, which remove 85% of the exhaust heat. This level is achieved by combining the hot and cold aisles so that air exhaust from one iDataPlex with Rear Door Heat eXchanger enters the inlet of the adjacent rack, thereby halving the tile pitch and allowing for an efficient mix of CRACs and Rear Door Heat eXchangers. In this case, approximately 7500 nodes are installed in the data center.

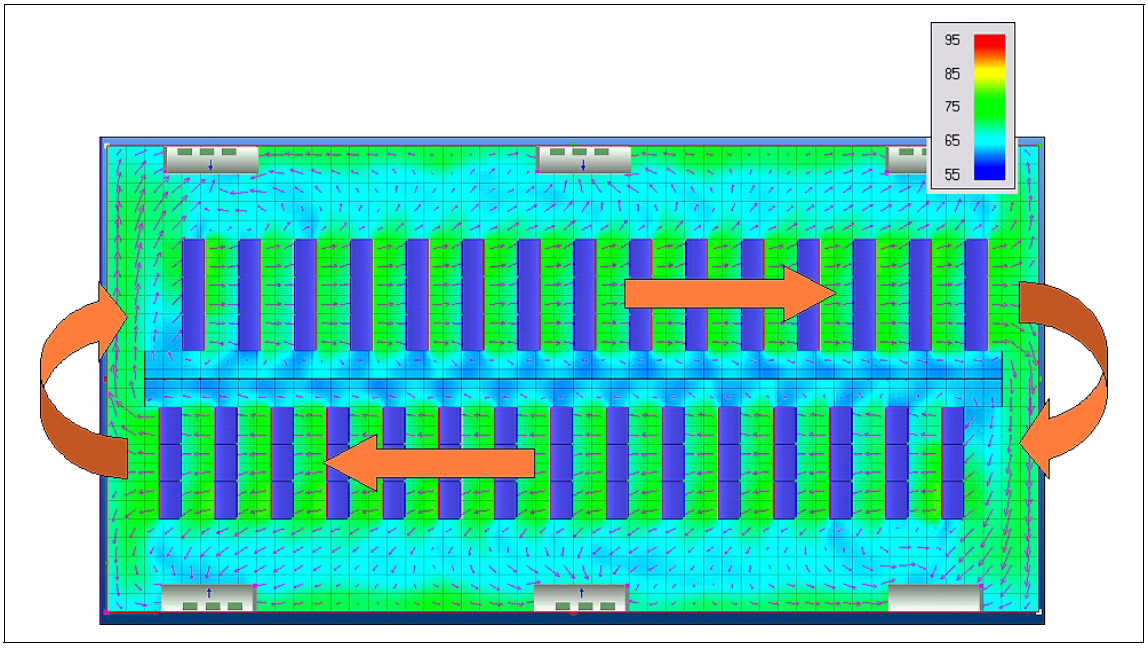


Figure 6-22 Possible circular data center airflow with iDataPlex and Rear Door Heat eXchanger

Single cold aisle with Rear Door Heat eXchangers and perimeter CRACs

Figure 6-23 shows an example 3-tile pitch design that also breaks with the traditional hot and cold aisle paradigm. The Rear Door Heat eXchanger units remove about 85% of the exhaust heat while the rest is handled by a small number of traditional CRACs.

The central cold aisle is fed by the CRACs and the exhaust heat is channeled back towards the CRACs. The complementary setup allows for CRAC or Rear Door Heat eXchanger failures while providing a very dense data center design. This scenario encompasses approximately 9000 nodes as shown in Figure 6-23.

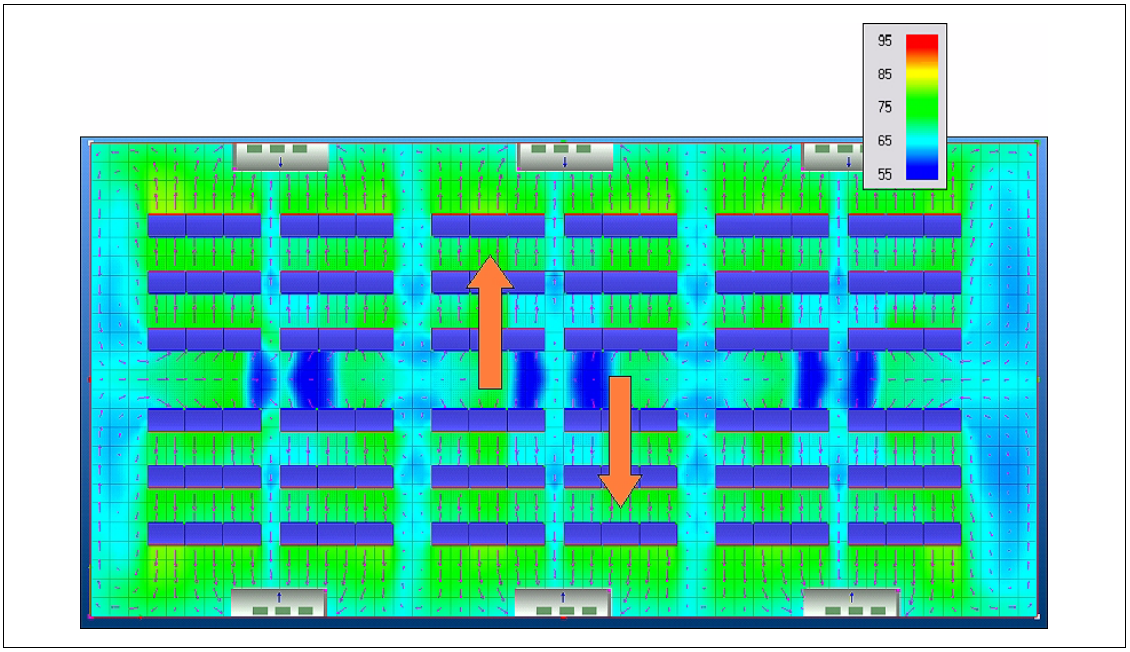


Figure 6-23 Rear Door Heat eXchanger showing air exhaust from one rack going into the adjacent rack

6.8.6 Rear Door Heat eXchanger airflow characteristics

The Rear Door Heat eXchanger behaves differently depending on the volume of air moved through it per unit of time. See Figure 6-24.

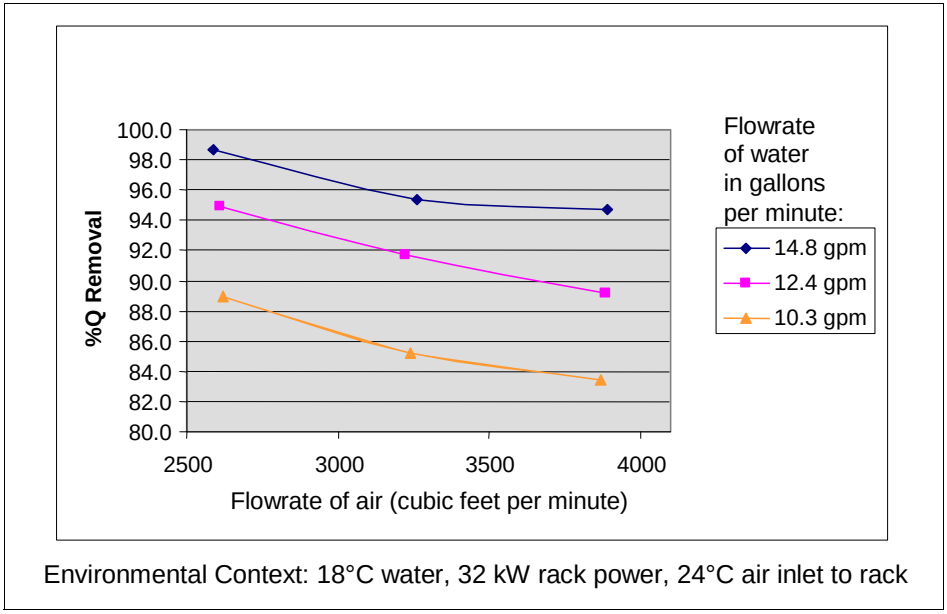


Figure 6-24 Airflow characteristics of the Rear Door Heat eXchanger

The effectiveness of the Rear Door Heat eXchanger is actually higher when less air is moved through it. This lends itself well to the iDataPlex because the flow rate of the air through the rack is lower than in the deeper standard racks.

6.9 Water cooling using the Rear Door Heat eXchanger

Plumbing is required to use the Rear Door Heat eXchanger. The Rear Door Heat eXchanger is 75 - 95% more efficient than standard air cooling by moving the thermal transfer from the CRAC unit to the back of the rack. It enables 30 kW per rack cooling capacity and works with chilled water or evaporative liquid for lowest-risk deployment; it uses industry standard quick fit connectors.

The Rear Door Heat eXchanger has the following characteristics:

- Can provide 100% heat removal
- Allows passive heat extraction, for example no compressors and peltier elements
- Provides a low air side impedance for the rack, which means less resistance to the airflow and therefore less fan power required
- Has a low waterside impedance, which means water flows more freely and requires less pumping power
- Has optimized performance through uniform node airflow, which means less turbulence at the back of the rack

Note: The Rear Door Heat eXchanger requires professional movers to install, because it weighs approximately 82 kg (180 lb).

Clients request water cooling for a number of reasons:

- Increased heat output from server CPUs
- Increased heat loads in data centers
- Demands for greater cooling system efficiency

In general, the benefits are that the server density can be increased without increasing cooling requirements and that temperature hot spots are more effectively controlled.

The iDataPlex Rear Door Heat eXchanger uses industry standard quick-connect hose fittings, as shown in Figure 6-25.

Figure 6-25 Quick-connect fittings at the bottom of the Rear Door Heat eXchanger

6.9.1 Technical specifications

The following list indicates technical specifications of the Rear Door Heat eXchanger:

- Pressure:
  - Normal operation: 137 kPa (20 psi)
  - Maximum: 689 kPa (100 psi)
- Volume:
  - Approximately 11 liters (3 gallons)
- Temperature:
  - 18°C ± 1°C (64.4°F ± 1.8°F)

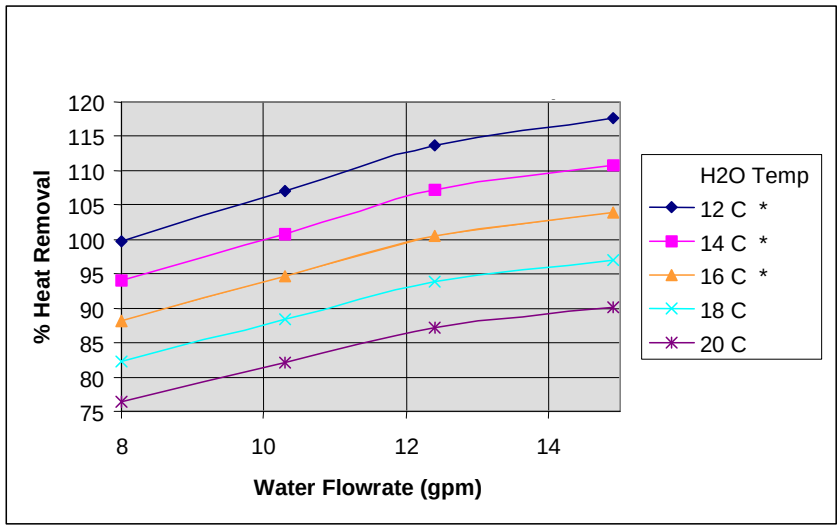
  See also 6.9.2, “Absorption rates” on page 208.
- Required water flow rate (measured at the supply entrance):
  - Minimum per minute: 22.7 liters (6 gallons)
  - Maximum per minute: 56.8 liters (15 gallons)

For more information, see the *Rear Door Heat eXchanger Installation and Maintenance Guide* at the following website:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5075220

6.9.2 Absorption rates

The Rear Door Heat eXchanger can be fine-tuned to account for different rack loads and facility dependencies. The graph in Figure 6-26 shows the absorption rates for a 32 kW rack.



Environmental Context: 24°C water, 32 kW rack power, 24°C air inlet to rack at minimum fan speed.

Figure 6-26 Absorption rates for different water temperatures and volumes

Water temperature under 18°C normally requires a cooling distribution unit (CDU), which separates the closed water loop in the heat exchanger from the building cold-water supply and controls the water temperature of the closed loop. The CDU measures the temperature and humidity of the room and adjusts the water temperature to avoid cooling to below the dew point. This approach is consistent with the ASHRAE Class 1 Environmental Specification that requires a maximum dew point of 17°C (62.6°F). See also “ASHRAE” on page 221.

6.10 Designing a heterogeneous data center

The unique design, form factor, and rack layout of iDataPlex does not necessarily mandate a data center design based on iDataPlex racks and nodes alone. You can easily design a heterogeneous data center by mixing traditional servers and other equipment in standard enterprise racks with iDataPlex racks. The primary benefits of iDataPlex are that this heterogeneous design will not incur any additional overhead or result in inefficiencies in power, cooling, and others.

The following figures show example heterogeneous floor layouts using by iDataPlex and traditional racks in a standard data center floor space of approximately 12 m x 18 m (40 ft x 60 ft):

- Figure 6-27 on page 210
- Figure 6-28 on page 210
- Figure 6-29 on page 211
- Figure 6-30 on page 211

Each layout shows either or both the number of iDataPlex and traditional racks, as well as the amount of usable horizontal rack-unit space available in each layout (indicated in the respective figure caption).

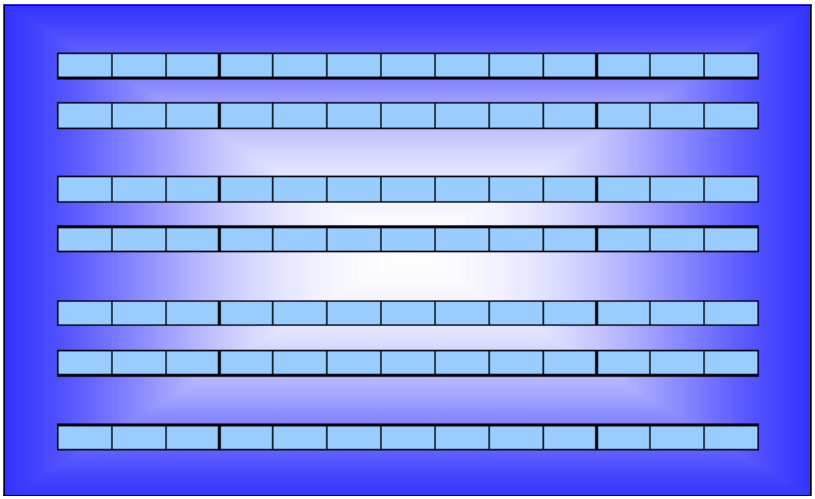


Figure 6-27 91 iDataPlex racks (7826 horizontal rack units)

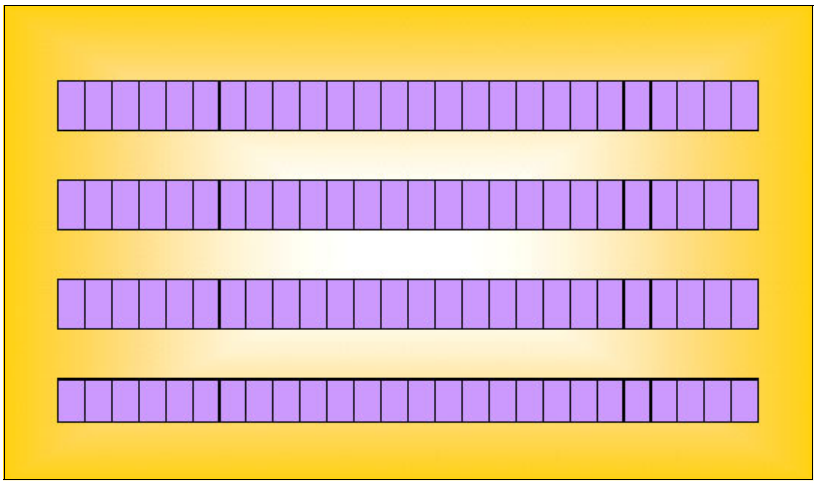


Figure 6-28 104 standard racks (4368 horizontal rack units)

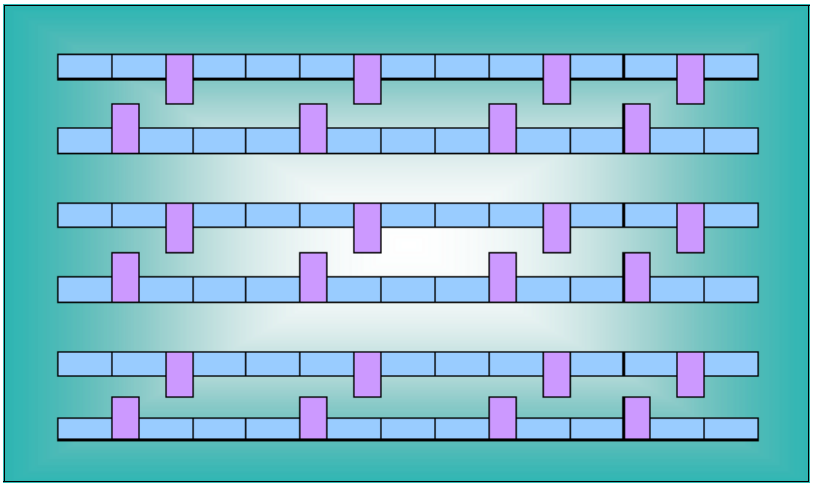


Figure 6-29 66 iDataPlex racks and 24 standard racks (6684 horizontal rack units)

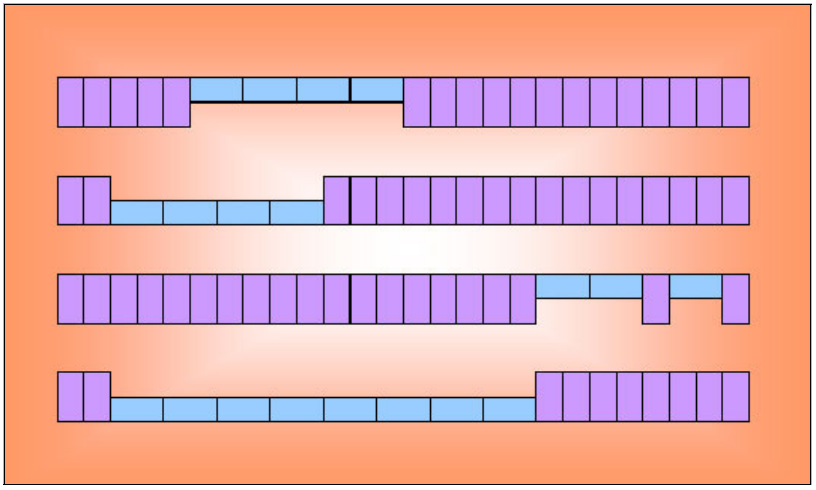


Figure 6-30 19 iDataPlex racks and 66 standard racks (4406 horizontal rack units)

6.11 iDataPlex servers in a standard enterprise rack

A newly supported concept is the placement of iDataPlex chassis and servers in a standard enterprise rack (Figure 6-31). Based on early experiences with customers that prefer not to use the iDataPlex rack, but to install the iDataPlex Flex chassis in a standard enterprise rack (EIA-310-D), IBM now supports this concept with certain caveats, as this section discusses.

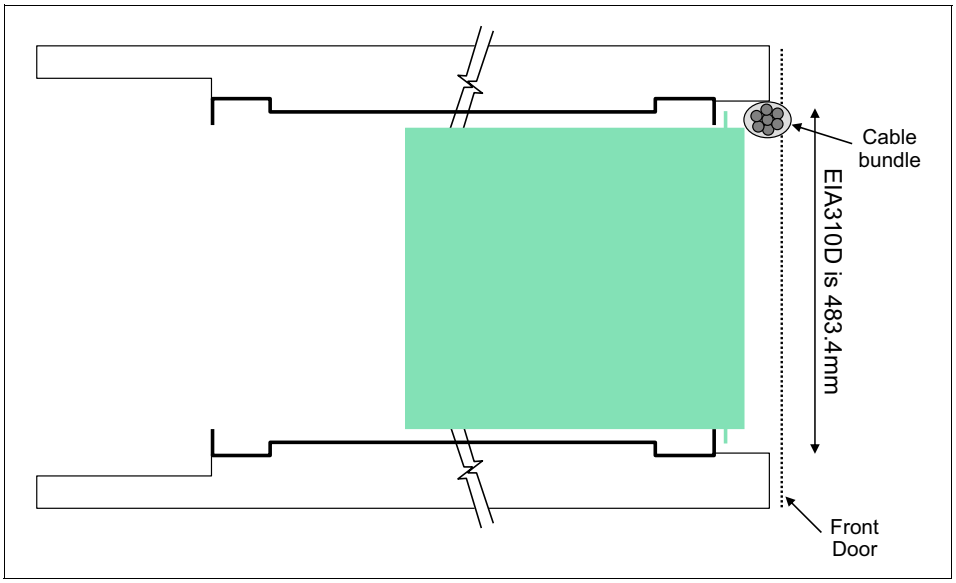


Figure 6-31 iDataPlex Chassis in a Standard Enterprise Rack (top view)

Figure 6-32 shows an example with iDataPlex chassis placed in a standard enterprise rack, with 40 iDataPlex nodes, two 48-port Ethernet switches, and a KVM console. Customers wanting to use this type of approach should be aware of the following aspects:

- The figure shows the USB Conversion Option (UCO) bundled, with the KVM dongle hanging in front of the disk drives.
- This solution is supported by iDataPlex development, but the customer should understand that disk drive access requires moving the cables.
- Cable management brackets route the cables in the front of the rack (the cold aisle).
- This solution requires removal of front doors

Figure 6-32 iDataPlex chassis cabling in a Standard Enterprise Rack

Note: Based on the specific options chosen, such as high-speed networking, storage, and others, there might be more considerations when placing iDataPlex nodes in the standard enterprise racks. The particular customer configuration and preferences should be carefully analyzed before providing an approved design based on the standard enterprise racks.

If the customer plans to use the IBM standard enterprise racks (option part number 46M2943), consider the following issues:

- Because of cables, a front door might not be able to be installed.
- Custom packaging might be required for shipping (due to no front door)
- Must use custom cable routing brackets, but you must confirm whether they would work and would be sufficient.
- Cables and brackets can intrude into the cold aisle (front of the rack).
- Determine whether cable egress from the rack is negotiable.
- For common egress, you might have to potentially lose density for cable routing to the rear of the rack.
- If overhead cabling is used, determine whether cables exit out the front top of the rack.
- Determine whether cables can be routed under the rack.
- Determine whether a hole for the cable exit can be located in the front of the rack in the cold aisle.

6.12 The Portable Modular Data Center

The Portable Modular Data Center (PMDC) is an offering aimed at providing quick data center capacity to clients that might not have the infrastructure to otherwise cater for the equipment. The advantage lies not so much in its mobility but in the stand-alone capsule that allows it to be placed outside and even next to existing buildings. The PMDC can be used to augment existing data centers or it can provide a data center for a customer without having to put a building in place. The PMDC uses standard shipping containers of approximately 6 m and 12 m (20 ft and 40 ft).

Figure 6-33 shows a standard 40-foot container with its outside dimensions.

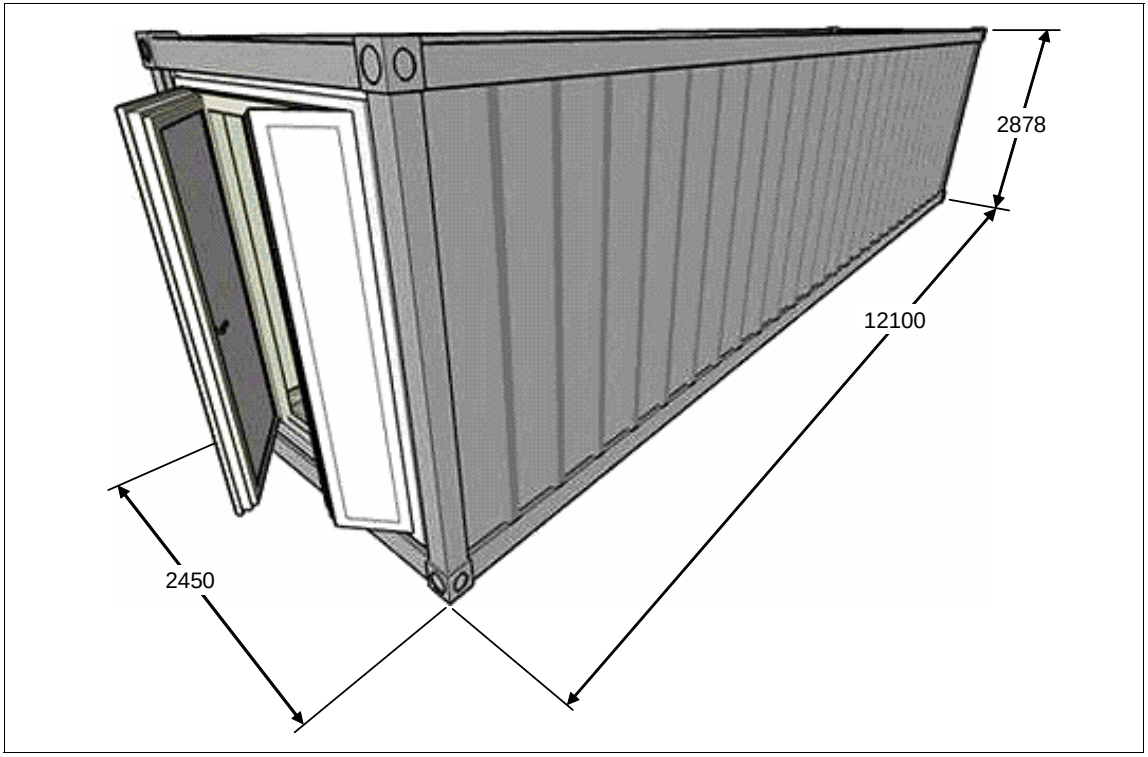


Figure 6-33 40-ft Portable Modular Data Center Container with dimensions

The PMDC is also available with iDataPlex as an option. Here are three example configurations:

- A 12 m (40 foot) container with nine iDataPlex racks in a central row
- A 12 m (40 foot) container with 18 iDataPlex racks, shown in Figure 6-34
- A double-wide 12 m (40 foot) container with 27 iDataPlex racks

The most densely packed PMDC is the 12 m (40 foot) container with 18 iDataPlex racks, shown in Figure 6-34. In this configuration, each rack can have 84 compute nodes, each with two quad-core CPUs, making up to 12,096 processor cores in one container.

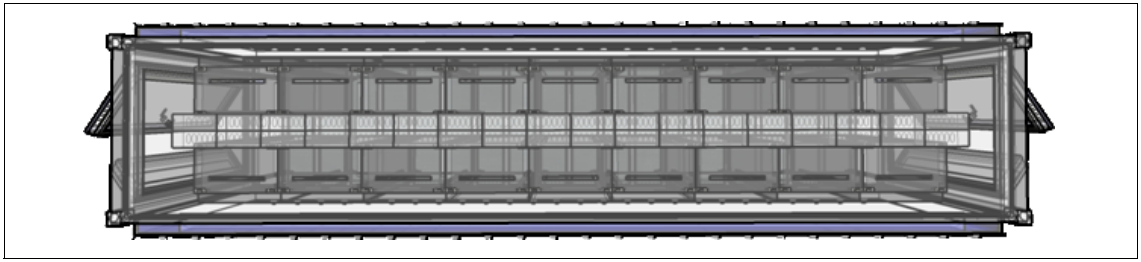


Figure 6-34 An iDataPlex Portable Modular Data Center Container with 18 systems

The iDataPlex solution is ideally suited to the PMDC because the Rear Door Heat eXchanger allows for a much denser packaging than standard servers and racks allow. The achieved density of nodes is also aided by the nodes, which are less deep than enterprise servers, allowing for two rows of iDataPlex racks in one container. In this configuration, the Rear Door Heat eXchanger is aided by additional fans on the container ceiling and provides sufficient air movement (active heat extraction) in all circumstances. See Figure 6-35.

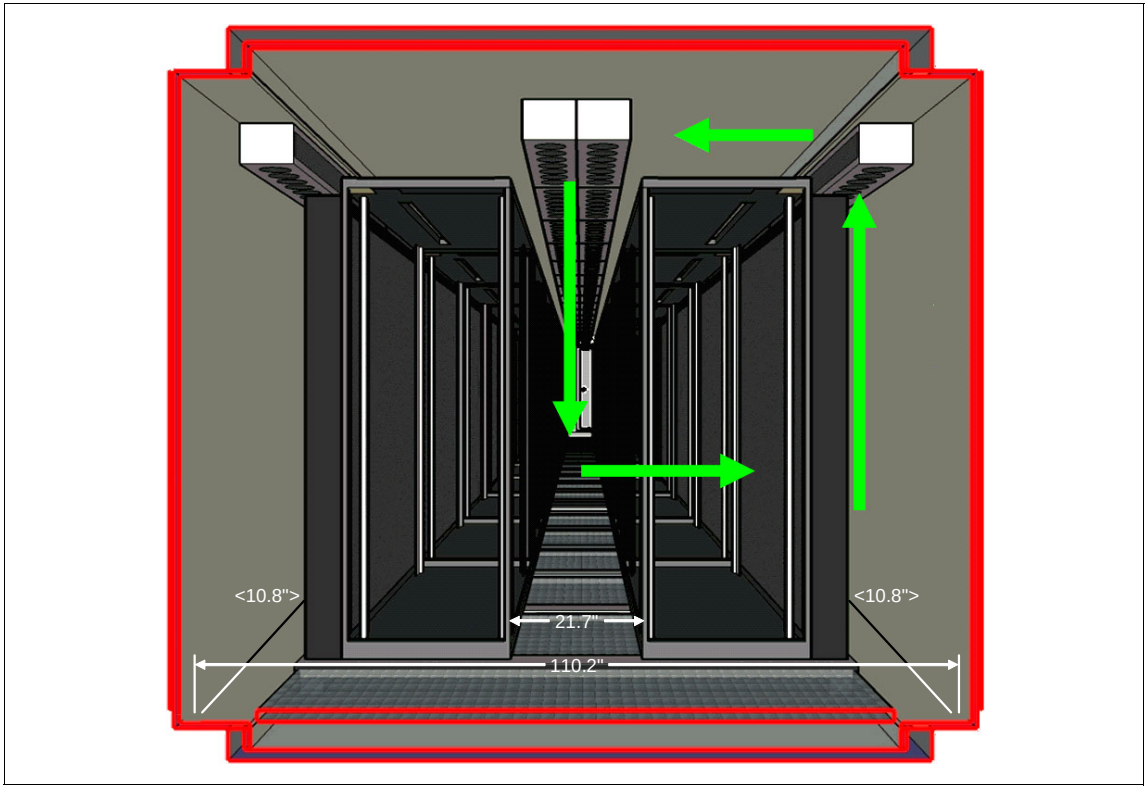


Figure 6-35 Additional ceiling fans supporting airflow in 18-rack PMDC (dimensions in inches)

In the PMDC, the racks are placed approximately 25 cm (10 in.) from the container walls which is less than the normal recommended minimum distance. The PMDC is a special design and was tested in the thermal lab so as to ensure that the air flow can be guaranteed in this special configuration.

The PMDC offering also has container configurations that provide supporting functions such as uninterruptible power supply and cooling.

6.13 Possible problems in older data centers

This section describes several problems that can occur in older data centers. In general, consider these factors when placing racks in the data center:

- Cable lengths:

  Be cautious about the maximum lengths for the different types of cables. For example, standard Ethernet cabling should not exceed 90 m (295 ft) from one switch to the next.

- Radio interference:

  Proximity to airport radar installations or commercial radio stations might cause interference in the rack or between racks. In general, strong radio frequency (RF) fields can be a problem for most computing equipment.

- Vibration:

  Railways, underground trains, and heavy vehicles in the vicinity can damage server equipment, particular hard drives.

- Power circuits:

  Power should be clean and stable. Sharing power with large motors and other industrial equipment is not a good idea and can lead to data loss and damaged equipment.

- Cooling:

  Adequate cooling must be provided at all times.

- Dust or contaminants:

  Metal shavings and conductive contaminants should be avoided in the computer room air because they build up on the equipment and can cause damage.

6.13.1 Power

A stable power source that is free of interference is a requirement for the high density data center. In most cases three-phase power can be used in the data center itself and the load on the three phases should be balanced.

Heavy industrial equipment, welding equipment, and other equipment that uses power can cause sags in voltage and can also shift the axis of the three-phase power. Three-phase motors can actually start to vibrate or run backwards if there is enough imbalance in a three-phase system. Certain water heaters and motors use three-phase power and can cause interference with power feeds.

Ensure that the power feed itself is laid out for the load and that the return path (or neutral) is laid out to cope with the load. Ground connections in general are important and improper grounding is a common cause for problems.

Various problems with three-phase motors have been linked to overhead rolling doors in loading bays, elevator engines, air conditioning compressors, and similar equipment.

Normally a five-wire three-phase power distribution system is recommended for use in data centers. IBM can help clients design a suitable power distribution system for the data center.

6.13.2 Earthing system

In this section, we compare two of the IEC 60364 earthing systems:

- TN-S[4]
- TN-C[5]

In a TN-C (also known as PEN[5]) earthing system, the earth-ground and neutral lines are combined. This can lead to significant currents on the shielding of data cables between racks and can also lead to the building itself carrying current (which is undesirable).

A TN-S system provides a low-noise earth connection, which is recommended for IT equipment with data connections.

[4] Terra Neutral-Separate (Neutral and Protective Earth on separate conductors)
[5] Terra Neutral-Combined (TN-C) also known as Protective Earth and Neutral (PEN)

Figure 6-36 shows the potential problem with TN-C systems that does not occur with TN-S earthing systems. The figure shows how the absence of a separate protective earth line leads to differences in potential between the two systems that are connected to the same power circuit. The difference in potential leads to current flowing across shielding of the orange data cable. This issue can damage equipment and cause interference on data lines.

In certain cases, the length of cables that can be used in a data center are shortenend dramatically. The farther apart two machines on the same circuit are, the higher the difference in potential is likely to be, and therefore the more interference is produced. This does not happen in a TN-S system when PE and N lines are separated.

Another solution to this problem is to use fiber optic cables between the racks, as is done in many data centers. Fiber optic cable does not conduct electricity and therefore can help solve this problem.

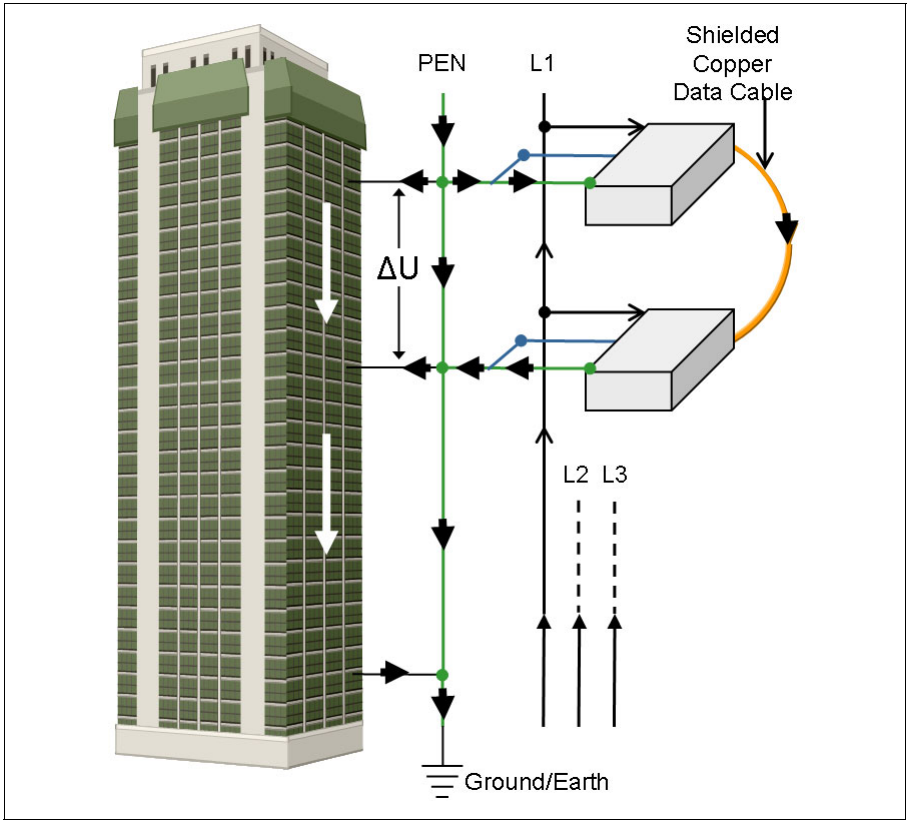


Figure 6-36 TN-C earthing system (PEN line)

Apart from the use of TN-C, and even with TN-S, further problems can arise if the protective earth framework in the data center is not properly implemented or designed. These problems tend to occur the larger an area a data center encompasses and can be mitigated or remedied by providing a closely-meshed grounding system.

6.13.3 Cooling

Cooling design considerations include the following situations:

- Placing perforated floor tiles in hot aisles reduces the air temperature of the hot aisle and thereby reduces the efficiency of the CRAC units.
- If the volume of cold air provided to a rack is too low, then the air from the back of the rack might start circulating back into the rack either over the top of the rack or around the side if the rack is standing on its own or at the end of a row. This leads to warm air circulating through the racks and is not desirable.
- Placing perforated floor tiles too close to CRACs has the unwanted effect of sucking air through the tile from above. This is known as *Bernoulli's principle* and can be measured with a Venturi meter.

6.13.4 Static electricity

Static electricity can cause damage to electronic devices. To avoid the build-up of static electricity, a number of factors have to be taken in to account:

- Floor coverings:

  Floor coverings must be antistatic with low propensity ratings

- Floor resistance:

  The maximum resistance should not be above $2x10^{10}$ ohms between the building and the floor. The minimum resistance should not be below 150 kilohms for two floor points that are 1 m (or 3 ft) apart.

- Raised floor grounding/earthing:

  Connections to earth should be provided at multiple points in the room and the raised floor, if fitted, should have its entire structure connected to ground in multiple places.

  Two types of raised floor are common:

  – Stringerless floors
  – Floors with stringers

Stringers are used to form the grid system. In stringered systems there is one stringer under each edge of each floor panel. Stringered floors are recommended for grounding and provide a Signal Reference Grid when bolted together. Stringerless floors have special needs above and beyond those of stringered floors and require special attention. Also, all metal conduits should be bonded to one another and to the floor.

- Air humidity

  Humidity should be 35 - 60% to prevent objects from storing electrical charge.

6.13.5 RF interference

Radio frequency (RF) interference can potentially be a problem for computer equipment. Close proximity to airports, radio stations, television stations, mobile phone masts, and strong RF signals should be avoided or mitigated. If such a situation is suspected, then IBM can be engaged to assess the situation and recommend mitigation.

6.14 Links to data center design information

This section lists Web sites with relevant data center design information.

6.14.1 ASHRAE

ASHRAE is the American Society of Heating, Refrigeration and Air-Conditioning Engineers. ASHRAE Technical Committee 9.9 (TC 9.9) concentrates on mission-critical facilities, technology spaces, and electronic equipment.

ASHRAE has published a series of books about data center design:

- *Thermal Guidelines for Data Processing Environments*
- *Datacom Equipment Power Trends and cooling Appliances*
- *Design Considerations for Datacom Equipment Centers*
- *Liquid Cooling Design Considerations for Data and Communications Equipment Centers*
- *Structural and Vibration Guidelines for Datacom Equipment Centers*
- *Best Practices for Datacom Facility Energy Efficiency*
- *High Density Data Center: Case Studies and other considerations.*

For more information about ASHRAE and the TC 9.9, see:

http://tc99.ashraetcs.org/

6.14.2 IEC 60364 electrical installations for buildings

The International Electrotechnical Commission (IEC) is a non-profit, non-governmental international standards organization that prepares and publishes international standards for all electrical, electronic, and related technologies. For more information about the IEC, see this website:

http://www.iec.ch/

6.14.3 2008 National Electrical Code (NFPA 70)

This advisory code is published by the National Fire Protection Association (NFPA). It is has been approved as an American National Standard on August 15, 2007. An online version of the NEC is available at this website:

http://www.nfpa.org/freecodes/free_access_agreement.asp?id=7008SB

7

Managing iDataPlex

This chapter discusses available options from IBM for managing an iDataPlex environment. It looks at how the various management products can be used with the system.

The following topics are covered in this chapter:

7.1 Integrated Management Module

In 2009, IBM introduced the IPMI-compliant service processor called the Integrated Management Module (IMM). The IMM and the second generation IMM2 are common across all IBM x86 machines, including the iDataPlex dx360 M3 and dx360 M4 servers.

The iDataPlex nodes, and many optional third-party components are compliant with Intelligent Platform Management Interface Version 2.0 (IPMI 2.0). IPMI operates independently of the operating system (OS) and allows administrators to manage a system remotely, even in the absence of the OS or the system management software, and even if the monitored system is not powered on.

IPMI also functions when the OS has started, and offers enhanced features when used with system management software. The nodes respond to IPMI messaging 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 IMM service processor provides competitive, standards-based systems management that enables upward integration into a wide variety of enterprise management environments “out of the box.” It is running a full operating system and has the following major features:

- Standard CIM and WS-Man interfaces
- Operating system drivers included in Windows and Linux, no additional device drivers needed
- Remote configuration of IMM and UEFI settings without the need to power-on the server
- Single firmware image for IMM across the product set
- Choice of dedicated or shared Ethernet connections
- Monitoring and alerting
- Problem detection before problems occur, using Predictive Failure Alerts
- Alerts provided, if problems do occur
- SNMP alerts can be sent to management software such as IBM Director
- Customizable, to meet specific customer needs
- Miscellaneous features, such as the following capabilities:
  - Field-replaceable unit (FRU) information
  - Inventory information (such as vendor ID, manufacturer, and so on)

- Sensor data records (SDR) repository (such as CPU temperature, fan speeds, or power supply voltages)
- Serial-over-LAN text console

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.

Remote control: Using IMM allows you to control a remote server's keyboard, mouse, and video to perform the following tasks:

- Use remote video with resolution up to 1280x1024 at 75 Hz, regardless of the system state
- Attach the remote server to the local disks on your management system
- Perform complete remote management under any operating system or pre-operating system environment
- Install software with a remote graphical console including KVM
- Boot the remote server from floppy disk or CD on your laptop
- Perform firmware updates or operating system installation
- Use common SP hardware across all platforms (one IMM firmware applies to all servers)
- Update other server firmware components
- No IBM drivers required, improved error logging, configurable both in-band and out-of-band

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

- Command-line interface (Telnet or SSH)
- Web interface, similar to the BladeCenter advanced management module
- IPMI
- ASU
- IBM Systems Director
- SNMP v1 and v3

For more information about the IMM, see the IBM white paper, *Transitioning to UEFI and IMM*, available from:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5079769

IBM has made available two publications about the IMM, an IMM user's guide and a paper on transitioning to UEFI and IMM. These publications are available online at the following URL:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5079338

Also note the following two RETAIN® tips for the first release of the IMM:

- *IMM takes a long time to become ready:*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5079504

- *Intermittent loss of network traffic to the IMM:*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5080112

7.2 Unified Extensible Firmware Interface

The Unified Extensible Firmware Interface (UEFI) replaces BIOS in System x and BladeCenter servers, and 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.

Unified Extensible Firmware Interface provides the following improvements over BIOS:

- Advanced Settings Utility (ASU) will now have more complete coverage of system settings.
- On rack mount servers, UEFI settings can be accessed out-of-band by ASU and the Integrated Management Module (not available on 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 using light path diagnostics.
- DOS tools are no longer required or officially supported.

More functionality: The following features add functionality:

- Adapter vendors can add more features in their options (for example, IPv6).
- Modular design allows faster updates as new features are introduced.
- More adaptors can be installed and used simultaneously.
- UEFI is fully backwards-compatible with legacy BIOS.
- It provides an improved user interface.
- It 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.
- It provides easier management.
- It reduces the number of error messages and eliminates outdated errors.
- A complete setup solution is provides by allowing adapter configuration function to be moved into UEFI.
- Complete out-of-band coverage by Advance Settings Utility simplifies remote setup.
- More functionality, better user interface, easier management for users.

For more information about the UEFI, see the IBM white paper, *Transitioning to UEFI and IMM*, available from:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5079769

7.3 Dynamic System Analysis

The iDataPlex nodes ship with a pre-boot version of Dynamic System Analysis (DSA). DSA collects and analyzes system information to help you diagnose system problems. DSA collects information about the following aspects of a system:

- System configuration
- Installed applications and hot fixes
- Device drivers and system services
- Network interfaces and settings
- Performance data and running process details
- Hardware inventory, including PCI information
- Vital product data and firmware information
- SCSI device sense data
- Application, system, security, ServeRAID, and service processor system event logs

An operating system-installable version of DSA available. It also creates a merged log that you can use to identify cause-and-effect relationships from different log sources in the system. DSA can also compare device driver and firmware on the system to the versions available on an UpdateXpress CD, and provide a summary of the differences.

A DSA portable edition is also available. It runs from the command line without altering any system files or system settings. It expands to temporary space on the target system and then runs; after processing completes, it deletes all intermediate files. DSA can collect system information in sensitive customer environments with only temporary use of system resources. DSA is supported in both Windows and Linux environments.

7.4 Advanced Settings Utility

The IBM Advanced Settings Utility (ASU) allows you to modify the BIOS settings without having to restart the system and enter the BIOS menus. ASU is supported in Linux, Windows, WinPE, Solaris and DOS environments. ASU supports scripting environments through its batch-processing mode.

The ASU utility and documentation are available from:

http://www.ibm.com/support/docview.wss?uid=psg1MIGR-55021

Note: By using the ASU utility, you may generate a BIOS definition file from an existing system. A standard definition file is not provided on the IBM support site.

7.5 Firmware update methods

The following tools are available to update iDataPlex firmware, like uEFI, IMM, or disk drives:

- Standalone updates:

  Firmware updates can be performed either 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 are provided with XML metafiles for use with IBM Director Update Manager and UpdateXpress system packs.

  Updates are available as Windows (.exe) or Linux (.bin) files, and are applied either inband or remote through the IMM.

- UpdateXpress System Packs:

  UpdateXpress System Packs are a tested set of online updates 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, as well as a user's guide, are available from the following website:

  http://www.ibm.com/support/docview.wss?uid=psg1SERV-XPRESS

- IBM Systems Director Update Manager:

  Update Manager can be set to download updates from the IBM Web site on a scheduled basis. IBM typically posts updates to the Web site for initial releases, suggested, and non-critical updates weekly, so having Update Manager check for updates weekly is suggested. Critical updates are posted as soon as they are ready. Manual retrieval of updates can be done without interruption to the automated update schedule.

- Bootable Media Creator:

  IBM has released a tool called Bootable Media Creator that allows you to bundle multiple System x updates from UpdateXpress System Packs and create a single bootable media (such as CD/DVD, USB flash drive, or a file set for PXE boot). You can download the tool from this website:

  http://www.ibm.com/support/docview.wss?uid=psg1TOOL-BOMC

- Update notification subscription:

  To receive weekly e-mail notifications of firmware and utility updates customized to the products that interest you, subscribe at this website:

  http://www.ibm.com/support/subscriptions

  Select the product family, system, and operating systems that you want to monitor. Then, select the types of information you want and how you want to be notified, as shown in Figure 7-1.

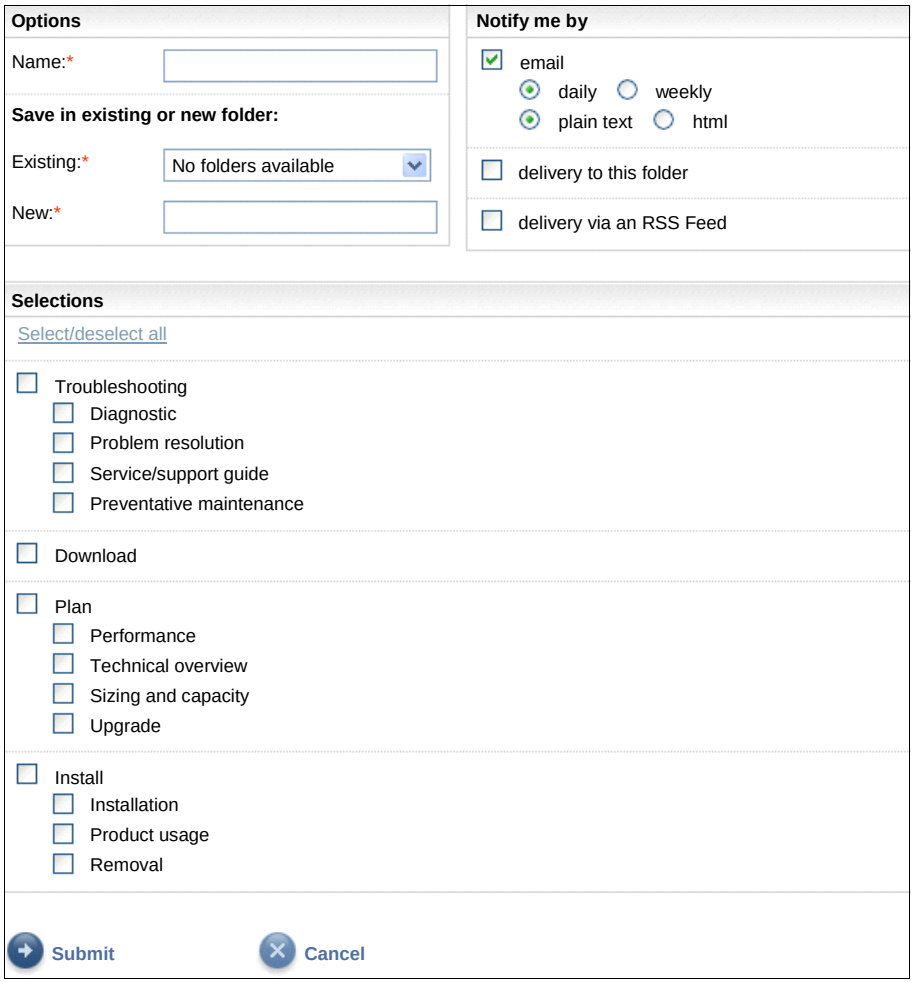


Figure 7-1 IBM Support Subscription: Selecting topics and notification method

7.6 xCAT

The Extreme Cloud Administration Toolkit (xCAT) is a scalable distributed computing management and provisioning tool that provides a unified interface for hardware control, discovery, and both diskful and diskfree deployment.[1] It was developed to make simple clusters easy, and complex clusters possible[2] by IBM Linux cluster specialists, to meet the requirements of IBM clients.

The previous version, xCAT 1.3, used IBM exclusive software to control legacy systems management processors. The current version 2.x utilizes, among other interfaces, IPMI for hardware control, and is being developed as an open source project, at this website:

http://sourceforge.net/projects/xcat/

For more information about xCAT, see this website:

http://www.alphaworks.ibm.com/tech/xcat/

The xCAT Web site also contains iDataPlex installation information, showing the deployment of Red Hat Enterprise Linux 5.1 on two iDataPlex racks (one of the machines is used as a management node, the remaining 167 machines are compute nodes). See the following documentation:

http://xcat.svn.sourceforge.net/svnroot/xcat/xcat-core/trunk/xCAT-client/share/doc/xCAT-iDpx.pdf

If you are an existing IBM Cluster Systems Management (CSM) user, the publication, *xCAT 2 Guide for the CSM System Administrator*, REDP-4437, might also be helpful:

http://www.redbooks.ibm.com/abstracts/redp4437.html

7.7 Distributed Image Management

Distributed Image Management (DIM) for Linux Clusters is a scalable image management tool that allows servers to run a Linux distribution over the network without local disks. It was developed for use with diskless blades in IBM BladeCenters, but can also be used with regular rack mounted servers, such as those in iDataPlex.

[1] As defined at http://www.xcat.org/
[2] As described at http://www.alphaworks.ibm.com/tech/xCAT/

Among the major advantages of DIM are the extremely small size (less than 1 MB), the feature to instantly replicate changes to thousands of nodes, and ability to automate DHCP configuration using a single XML file.

For more information about DIM, see this website:

http://www.alphaworks.ibm.com/tech/dim/

8

Services offerings

To aid in the successful implementation of an iDataPlex deployment, IBM offers a variety of planning and installation services. These offerings and what they provide are described in this chapter.

The following topics are covered in this chapter:

8.1 Global Technology Services

Global Technology Services® (GTS) offers a number of services that are relevant to iDataPlex and to site and facility services in general. The following list shows a selection of services that evaluate space and thermal requirements, perform airflow analysis, simulate application scenarios, and design the power and cooling for data centers.

The following GTS offerings are described in this section:

- 8.1.1, “High-density computing data center readiness assessment” on page 234
- 8.1.2, “Thermal analysis for high-density computing” on page 235
- 8.1.3, “Scalable modular data center” on page 236
- 8.1.4, “Data center global consolidation and relocation enablement” on page 236
- 8.1.5, “Data center energy efficiency assessment” on page 237
- 8.1.6, “Optimized airflow assessment for cabling” on page 237
- 8.1.7, “Portable Modular Data Center” on page 238

For more information, see the site and facilities Web site:

http://www.ibm.com/services/us/index.wss/itservice/igs/a1026000

8.1.1 High-density computing data center readiness assessment

The high-density computing data center readiness assessment helps clients benefit from high-density computing by assessing the following information:

- The capacity and capability of their existing data centers
- The gaps that could jeopardize continuous operations
- The actions to resolve identified concerns

This readiness assessment helps clients gauge their capacity to support high-density IT infrastructure components in their data center facilities. By determining the existing facility’s power-supply and heat-removal capabilities and comparing them to the power and cooling requirements demanded by the new technology, the readiness assessment can identify potential gaps that could jeopardize continuous operations, and can provide an input for problem resolution.

The readiness assessment consulting methodology combines IBM technology insight and decades of experience in data center facility design and operations to help CIOs and other IT decision-makers determine a data center facility's capacity to accommodate high-density technology and to operate it reliably.

The offering includes:

- An assessment of a facility's as-built power supply and heat-removal capabilities
- A comparison of existing capacity against power and cooling requirements demanded by high-density technology, both existing and planned
- An examination of the entire interdependent power and heat-removal infrastructure
- The identification of gaps potentially jeopardizing continuous operations
- A comprehensive evaluation that forms the basis for recommendations for any necessary remedial measures

The potential benefits to a client include:

- Professional guidance in managing the growth and expense challenges associated with high-density technology
- Key technology insight to identify design shortfalls and capacity limitations
- Continuous operations
- Adoption of powerful computing capabilities to meet escalating customer demands

For more information, see this website:

http://www.ibm.com/in/gts/datacentre/hdcas.html

8.1.2 Thermal analysis for high-density computing

Thermal analysis for high-density computing assists clients in identifying and resolving heat-related problems within existing data centers and provides recommendations for cost savings and future expansions.

Computational fluid dynamics software tools are used to create predictive thermal models that enable clients to plan new data centers designed to satisfy current equipment cooling requirements, to support additional equipment for future IT expansion, and to isolate thermal problem areas in the data center.

The potential benefits of this offering include:

- Lower data center costs by helping to improve cooling efficiency and thereby reducing related power consumption
- Increased system uptime by helping to reduce server outages caused by high-heat conditions
- Provision of information you can use to better understand how to manage data center growth
- Effective consolidation of data center facilities
- Improved reliability of data center facilities

For more information, see this website:

http://www.ibm.com/in/gts/datacentre/tahdc.html

8.1.3 Scalable modular data center

Although this offering does not, at the time of writing, use iDataPlex components, it is for clients who want a data center to be up and running quickly and would potentially harmonize well with an iDataPlex solution.

The scalable modular data center offering allows clients to quickly make use of a turnkey solution with power, cooling, security, and monitoring. Because of its modularity, it is quick to set up in nearly any client facility. In-row cooling and other innovative solutions make this an attractive offering.

The scalable modular data center offering helps clients to deploy ready-made data centers in less time and at lower cost than traditionally designed data centers. IBM experience, server technology knowledge, and facilities infrastructure expertise help clients more easily define their server and facilities requirements.

Visit the following Web site for more information:

http://www.ibm.com/in/gts/datacentre/datacntr.html

8.1.4 Data center global consolidation and relocation enablement

This offering provides a consistent, repeatable, phased management approach to help clients implement a data center relocation or consolidation by leveraging local IBM site and facilities expertise around the globe.

Examples of recently completed projects include:

- A client in China that achieved a USD180 million reduction in annual operating expenses from consolidating 38 to 2 data centers while also improving business resilience.
- A client in Germany achieved USD7.2 million in annual operational savings by consolidating 4 data centers into one 353 m^2 (3,800 square-foot) data center.

8.1.5 Data center energy efficiency assessment

The data center energy efficiency assessment provides a comprehensive assessment of the client data center and supporting physical infrastructure to identify operational cost savings and assists clients with utility rebates or LEED (Leadership in Energy and Environmental Design) certification.

This offering provides:

- A comprehensive, fact-based analysis of data center energy efficiency
- An evaluation of cooling system components, electrical systems, and other building systems
- A baseline metric for the data center's energy use
- A road map of cost-justified recommendations

8.1.6 Optimized airflow assessment for cabling

This offering is a solution to the issue of excess and obsolete cabling in a raised-floor air-delivery plenum, which identifies and removes unused cabling for the client, resulting in an energy optimized air delivery system.

This offering has the following features:

- A comprehensive review of existing cabling infrastructure
- A plan for improvements to the data center that can help increase availability and maintain a security-rich environment
- An expert analysis of the overall cabling design required to help improve data center airflow for optimized cooling
- Specific recommendations for cabling system improvements
- A report on how the new structured cabling design can help maximize airflow for cooling, which can improve efficiency and reduce power consumption

Several potential benefits are:

- Increased availability, improved energy efficiency, and reduced overall data center management costs
- Reduced potential for downtime through fiber connector inspection and verification
- Simpler change and growth within data centers through the documentation of cooling systems, which also helps companies organize and manage cables and trunking

8.1.7 Portable Modular Data Center

The Portable Modular Data Center (PMDC) is a way of rapidly deploying a reproducible data center environment to nearly any worldwide location.

The PMDC is container-based as opposed to the Scalable Modular Data Center which requires a building and is not mobile. See 8.1.3, “Scalable modular data center” on page 236 for details of that offering.

The design, build, and drop-ship of an IBM Portable Modular Data Center can be completed in as short a period as 12 - 14 weeks including shipping. The PMDC solution can be designed in a few different configurations, the length of each container solution is either 6 m (20 ft) or 12 m (40 ft), as follows:

- IT container only (contains IT equipment racks with the required cooling units, power distribution, remote monitoring and fire detection and suppression)
- IT container plus services container (contains uninterruptible power supply, batteries, power distribution, electrical switchboard, chiller unit, and fire detection and suppression)
- All-in-one (single container with both IT equipment and services equipment)

The PMDC can be built with redundant N+1 infrastructure. It is suited to both permanent or temporary installations. Configurations are available with and without iDataPlex components.

Figure 8-1 Example multi-container PMDC

The solution features include:

- A complete PMDC design to meet client requirements
- Work with the client's teams to determine power and chilled water requirements, space allocations, installation requirements, and so forth.
- Project management of the entire project which provides a single point of contact for the client
- A fit-up of the implementation site
- Managed installation of electrical and water connections, placement of containers, PMDC access, and so forth.
- Installation of engine generators, chillers, electrical feeds, and so forth, as required
- Integration, connection, and the start-up of IT technology into the PMDC.

The solution can also include:

- An optional sourcing of new IT technology beyond the iDataPlex offering
- The relocation of existing IT technology
- Provisioning for connectivity requirements for communications and networks
- Start-up and test of the complete PMDC

For more information, see the press announcement, which also contains an introductory video:

http://www.ibm.com/press/us/en/pressrelease/24395.wss

8.2 IBM Systems and Technology Group Lab Services

IBM Systems and Technology Group (STG) Lab Services provides an in-depth capability for data center design and analysis that is not limited to the planning of racks, servers, chiller doors, and power. Lab Services experts have a broad scope and a deep knowledge of every aspect of the data center.

A number of services are tailored directly for the iDataPlex, and a number of further services cover the data center as a whole. The STG Lab Services team is ideally positioned to help clients roll out or come up to speed with an iDataPlex installation because they have full access to the developers of the systems and experts who are deeply involved in equipment and software tuning, providing detailed recommendations and reports, and setting worldwide industry standards.

For overview information about the of STG Lab Services offerings, see:

- 8.2.1, “Data center power and cooling planning for iDataPlex” on page 241
- 8.2.2, “iDataPlex post-implementation jumpstart” on page 241
- 8.2.3, “iDataPlex management jumpstart” on page 242
- 8.2.4, “Cluster enablement team services” on page 242

Several solutions offered by STG include:

- Power, cooling, and I/O data center best practices
- Power sizing in the data center
- Use of virtualization to help reduce the demand for energy
- Information about power and heat load values for specifically configured server systems
- Comparisons of IBM products with competitors based on performance per watt, performance and SWaP (space, watts, and performance) metrics
- Power trends for IBM product lines
- Integration of IBM products into data centers with limitations (for example, 5 kW per rack)
- Reducing carbon emissions in line with European laws
- Water cooling in new data centers
- Planning of raised floor heights, ceiling heights, positions of CRACs, positions of cable trays, and more
- Using IBM system and storage device capabilities to manage power demand and thermal load

IBM STG Lab Services is a world-wide team based in the following locations:

- Austin, TX, U.S.
- Bangalore, India
- Beaverton, OR, U.S.
- Beijing, China
- Kirkland, WA, U.S.
- LaGaude, France
- Mainz, Germany
- Poughkeepsie, NY, U.S.
- Raleigh, NC, U.S.
- Rochester, MN, U.S.
- Taipei, China

See the following website for IBM STG Lab Services:

http://www.ibm.com/systems/services/labservices/solutions/labservices_datacenter.html

8.2.1 Data center power and cooling planning for iDataPlex

This offering helps clients assess air conditioning and distribution in preparation for installing one or more iDataPlex racks. Guidance for cooling is provided, as is an evaluation of the requirement for a Rear Door Heat eXchanger. iDataPlex power specifications are reviewed in light of the customer's PDU and hardware configuration. Electrical requirements are provided based on the findings.

The service also helps clients evaluate the need for a data center thermal analysis or measurement and optimization study and provides appropriate direction based on the sum of findings from all the performed evaluations.

8.2.2 iDataPlex post-implementation jumpstart

This offering helps clients install and configure the iDataPlex in their own environment. This is done by providing a comprehensive overview of the system and the best practices related to it. The custom-tailored hardware and software is demonstrated and then restored to out-of-the-box settings so that the client can customize the software under supervision and with support from lab experts. This process effectively transfers the required knowledge.

Several of the topics covered are: planning sessions, cluster configuration, management node configuration, BIOS/RAID/firmware verifications and updates, operating system and cluster manager installation, network and storage configuration, cluster tests, power up/down tests, skills transfer, and assistance with the acceptance tests.

8.2.3 iDataPlex management jumpstart

This services offering takes the same approach as the other jumpstart services by first demonstrating the products and then helping the client configure them from an unconfigured state.

This offering includes such items as:

- Planning the use of and installation of the appliance
- Overviews of the different command-line utilities
- Authentication methods
- Monitoring and management functions
- Integration into SNMP-based management networks
- Integration into IBM Director-based installations
- Demonstrations of the appliance showing chassis and node provisioning
- User management
- Remote and power management
- CLI for batch and interactive use
- Configuring the alert handling

8.2.4 Cluster enablement team services

The cluster enablement team provides clients and partners with access to IBM experts who are skilled in HPC cluster implementation, which includes hardware, software, and complex high-performance computing knowledge.

The service categories are divided into four components:

- Cluster Implementation Services

  The Cluster Enablement Team provides cluster implementation services for:

  - Standard IBM Intelligent Cluster
  - Microsoft Windows Compute Cluster Server
  - DB2 Balanced Configuration Unit
  - DB2 Information Integrator
  - SAP Search Engine

  The Cluster Enablement Team provides design-your-own cluster services for:

  - Basic cluster installation and setup
  - Design SAP Business Intelligence Accelerator and SAP Search Engine
  - Myrinet and InfiniBand network implementation
  - GPFS implementation

- Cluster Support Services

  Cluster Support Services has three support subscription options, each renewable yearly:

  - Remote support
    - Provide basic remote support by way of e-mail, phone, Web, and so on
    - IBM Intelligent Cluster Service provides level 1 and 2 remote support as needed
  - On-call administration support
    - Dedicated IBM Intelligent Cluster resource to provide remote cluster administration and assistance as needed (24 x 5, 4-hour response window, pager and email notifications)
    - Optional 24 x 7 support for mission-critical customers paying a higher premium
  - On-site health-check support
    - Provide periodic maintenance support to ensure health of systems, perform necessary upgrades, and so on
    - Fixed number of visits per year to the customer site
    - Available for design-your-own clusters and standard cluster through the IBM Intelligent Cluster offering

- Advanced Cluster Support Services

  Advanced Cluster Support Services provides customized account-specific services that include:

  - Hardware problem troubleshooting
  - Cluster benchmarking assistance
  - Software implementation and tuning
  - Cluster administration

- Cluster Training Services

  The Cluster Training Services offer training on cluster-related topics and have a number of standard offerings that include:

  - General cluster training
  - xCAT
  - GPFS

  Customized training is also available.

8.3 IBM Global Financing

IBM Global Financing can assist clients with competitive customized financing structures for their System x iDataPlex equipment. Leasing structures can lower total cost of ownership, minimize risk, improve accountability, and enable clients to focus on their core business strategies while still preserving their ability to make flexible equipment decisions throughout the entire technology life cycle.

For financing information, see:

http://www.ibm.com/financing
http://www.ibm.com/financing/us/recovery/gogreenwithibm/

IBM Global Financing offers competitive financing to credit-qualified customers and IBM Business Partners to assist them in acquiring IT solutions. The offerings include financing for IT acquisition, including hardware, software, and services, from both IBM and other manufacturers or vendors, as well as commercial financing (revolving lines of credit, term loans, acquisition facilities, and inventory financing credit lines) for Business Partners. Offerings for all customer segments (small, medium, and large enterprise), rates, terms, and availability can vary by country.

Abbreviations and acronyms

AFCOM	Association for Computer Operation Managers
AJAX	Asynchronous JavaScript and XML
AMD	Advanced Micro Devices
AS	Australian Standards
AFCOM	Association for Computer Operation Managers
AJAX	Asynchronous JavaScript and XML
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
ASU	Advanced Settings Utility
BIOS	basic input output system
BMC	Baseboard Management Controller
BNT	Blade Network Technologies®
BSD	Berkeley Software Distribution
BTU	British Thermal Unit
CAE	computer aided engineering
CD	compact disc
CD-ROM	compact disc read only memory
CDU	cooling distribution unit
CIO	chief information officer
CISSP	Certified Information Systems Security Professional
CLI	command-line interface
CPLD	complex programmable logic device
CPU	central processing unit
CTS	clear to send
DB	database
DCD	data carrier detect
DCiE	Data Center Infrastructure Efficiency
DHCP	Dynamic Host Configuration Protocol
DIM	Distributed Image Management
DIMM	dual inline memory module
DMTF	Distributed Management Task Force
DOS	disk operating system
DPI	Distributed Power Interconnect
DSA	Dynamic System Analysis
DSR	data set ready
DTR	data terminal ready
DW	data warehousing
ECC	error checking and correcting
EIA	Electronic Industries Alliance
FC	Fibre Channel
FRU	field replaceable unit
GA	general availability
GB	gigabyte
GFS	Global File System
GND	common ground
GPFS	General Parallel File System
GTS	Global Technology Services
HBA	host bus adapter
HPC	high performance computing
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
I/O	input/output

IBM	International Business Machines
IDC	International Data Corporation
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IP	Internet Protocol
IPMB	Intelligent Platform Management Bus
IPMI	Intelligent Platform Management Interface
IPTV	Internet Protocol Television
ISL	Inter-Switch Link
IT	information technology
ITSO	International Technical Support Organization
kW	kilowatt
LACP	Link Aggregation Control Protocol
LAMP	Linux Apache MySQL PHP
LAN	local area network
LED	light emitting diode
LEED	Leadership in Energy and Environmental Design
MB	megabyte
MIT	Massachusetts Institute of Technology
MPI	Message Passing Interface
NEC	National Electrical Code
NFPA	National Fire Protection Association
NFS	network file system
NSF	Notes Storage File
OS	operating system
PCI	Peripheral Component Interconnect
PCIe	PCI Express
PDU	power distribution unit
PEN	Protective Earth and Neutral
PET	Platform Event Trap
PMDC	portable modular data center
PUE	Power Usage Effectiveness
PXE	Preboot eXecution Environment
RAID	redundant array of independent disks
RAM	random access memory
RF	radio frequency
RHDX	Rear Door Heat eXchanger
RI	ring indicator
ROM	read-only memory
RPM	revolutions per minute
RSA	Remote Supervisor Adapter
RTS	request to send
RXD	received data
SAN	storage area network
SAS	Serial Attached SCSI
SATA	Serial ATA
SCSI	Small Computer System Interface
SDR	Single Data Rate
SLURM	Simple Linux Utility for Resource Management
SMASH	Systems Management Architecture for Server Hardware
SMLT	split multi-link trunking
SNA	systems network architecture
SNMP	Simple Network Management Protocol
SOA	Service Oriented Architecture
SOL	Serial over LAN
SSI	Server System Infrastructure
STG	Server & Technology Group

TB	terabyte
TCO	total cost of ownership
TX	transmit
TXD	transmitted data
UDP	user datagram protocol
UPS	uninterruptible power supply
URL	Uniform Resource Locator
USB	universal serial bus
USD	U.S. dollar ($)
VGA	video graphics array
VLAN	virtual LAN
VSS	virtual switching system
WCCS	Windows Compute Cluster Services
XM	extended memory
XML	Extensible Markup Language

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks publications

For information about ordering these publications, see “How to get Redbooks publications” on page 253. Note that some documents might be available in softcopy only.

Related IBM Redbooks publications and Redpaper publications:

- *Building an Efficient Data Center with IBM iDataPlex*, REDP-4418

  http://www.redbooks.ibm.com/abstracts/redp4418.html
- *BladeCenter HS21 XM Memory Configurations*, TIPS0671

  http://www.redbooks.ibm.com/abstracts/tips0671.html
- *ServeRAID Adapter Quick Reference*, TIPS0054

  http://www.redbooks.ibm.com/abstracts/tips0054.html
- *Implementing IBM Director 5.20*, SG24-6188

  http://www.redbooks.ibm.com/abstracts/sg246188.html
- *Building a Linux HPC Cluster with xCAT*, SG24-6623

  http://www.redbooks.ibm.com/abstracts/sg246623.html
- *IBM System Storage DS3000: Introduction and Implementation Guide*, SG24-7065

  http://www.redbooks.ibm.com/abstracts/sg247065.html
- *Implementing IBM Systems Director 6.1*, SG24-7694

  http://www.redbooks.ibm.com/abstracts/sg247694.html

IBM iDataPlex product publications

These publications are also relevant as further information sources:

- IBM iDataPlex Information Center

  http://publib.boulder.ibm.com/infocenter/systems/scope/idataplex
- *IBM iDataPlex Rack Installation and User's Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5075216
- *Rear Door Heat eXchanger Installation and Maintenance Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5075220
- *IBM System x iDataPlex dx320 User's Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5078019
- *IBM System x iDataPlex dx320 Problem Determination and Service Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5078020
- *IBM iDataPlex dx340 User's Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5075221
- *IBM iDataPlex dx340 Problem Determination and Service Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5078854
- *IBM System x iDataPlex dx360 User's Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5077374
- *IBM System x iDataPlex dx360 Problem Determination and Service Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5077375
- *IBM PDU+ Installation and Maintenance Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5073026
- IBM ServeRAID MR10i publications

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5074114
- *IBM ServeRAID BR10i Installation and User's Guide*

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5076481
- *Transitioning to UEFI and IMM*, white paper available from:

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5079769

Online resources

Web sites listed in this section are also relevant as further information sources.

IBM Web sites

- IBM System x iDataPlex home page

  http://www.ibm.com/systems/x/hardware/idataplex

- IDC report: *The Impact of Power and Cooling on Data Center Infrastructure*

  http://www-03.ibm.com/systems/z/advantages/energy/index.html#analyst_rep

- IBM press release: *IBM Premiers Project Big Green in Hollywood*

  http://www.ibm.com/press/us/en/pressrelease/22891.wss

- IBM press release: *IBM Unveils Plan to Combat Data Center Energy Crisis; Allocates $1 Billion to Advance "Green" Technology and Services*

  http://www.ibm.com/press/us/en/pressrelease/21524.wss

- IBM press release: *IBM Project Big Green Tackles Global Energy Crisis*

  http://www.ibm.com/press/us/en/pressrelease/24395.wss

- Global Technology Services SalesOne site (for IBM employees only)

  http://w3.ibm.com/services/salesone

- IBM Systems and Technology Group Lab Services

  http://www.ibm.com/systems/services/labservices/solutions/labservices_datacenter.html

- IBM Global Financing, IT energy efficiency financing

  http://www.ibm.com/financing/us/recovery/gogreenwithibm/

- IT Services - Site and facilities

  http://www-935.ibm.com/services/us/index.wss/itservice/igs/a1026000

- iDataplex Drivers, Firmware, and Software - ServeRAID

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-5076022

- System Management Bridge BMC CLI and Remote Console Utility

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-64636

- Advanced Settings Utility (ASU)

  http://www.ibm.com/support/docview.wss?uid=psg1MIGR-55021

- Support subscriptions

  http://www.ibm.com/support/subscriptions

- IBM ServerProven

  http://www.ibm.com/servers/eserver/serverproven/compat/us/
- IBM System Storage DS3000 series Interoperability Matrix

  http://www.ibm.com/systems/storage/disk/ds3000/pdf/interop.pdf
- xCAT download

  http://www.alphaworks.ibm.com/tech/xCAT/
- Distributed Image Management for Linux Clusters (DIM) download

  http://www.alphaworks.ibm.com/tech/dim/
- Scalable modular data center

  http://www.ibm.com/in/gts/datacentre/datacntr.html
- Thermal analysis for high density computing

  http://www.ibm.com/in/gts/datacentre/tahdc.html
- High-density computing assessment services

  http://www.ibm.com/in/gts/datacentre/hdcas.html
- Server Optimization and Integration Services - VMware server virtualization

  http://www-935.ibm.com/services/us/index.wss/offering/its/a1029383
- Site and facilities

  http://www-935.ibm.com/services/us/index.wss/itservice/igs/a1026000
- IBM white paper: *Optimizing the Performance of IBM System x and BladeCenter Servers using Intel Xeon 5500 Series Processors*

  ftp://ftp.software.ibm.com/common/ssi/sa/wh/n/xsw03025usen/XSW03025USEN.PDF

Other information not from IBM

- The green data center: Energy-efficient computing in the 21st century

  http://searchdatacenter.techtarget.com/general/0,295582,sid80_gci1273283,00.html
- Gartner press release “Gartner Identifies the Top 10 Strategic Technologies for 2008”

  http://www.gartner.com/it/page.jsp?id=530109
- *Estimating Total Power Consumption By Servers In The U.S. And The World* by Jonathan Koomey

  http://enterprise.amd.com/us-en/AMD-Business/Technology-Home/Power-Management.aspx

- The High Availability Linux Project

  http://www.linux-ha.org

- Configuring Red Hat Enterprise Linux 3 or 4 to access iSCSI storage

  http://kbase.redhat.com/faq/docs/DOC-3939

- Microsoft Support for iSCSI

  http://www.microsoft.com/windowsserver2003/technologies/storage/iscsi/msfiSCSI.mspx

- Boot from SAN in Windows Server 2003 and Windows 2000 Server

  http://www.microsoft.com/windowsserversystem/wss2003/techinfo/plandeploy/bootfromsaninwindows.mspx

- *The Green Grid Data Center Power Efficiency Metrics: PUE and DCiE*

  http://www.thegreengrid.org/sitecore/content/Global/Content/white-papers/The-Green-Grid-Data-Center-Power-Efficiency-Metrics-PUE-and-DCiE.aspx

- International Electrotechnical Commission

  http://www.iec.ch/

- ASHRAE Technical Committee 9.9: Mission Critical Facilities, Technology Spaces and Electronic Equipment

  http://tc99.ashraetcs.org/

- US National Electrical Code

  http://www.nfpa.org/freecodes/free_access_agreement.asp?id=7008SB

- Review of liquid cooling book by Don Beaty from ASHRAE TC 9.9

  http://searchdatacenter.techtarget.com/news/article/0,289142,sid80_gci1218484,00.html

- xCAT

  http://www.xcat.org/
  http://sourceforge.net/projects/xcat/

How to get Redbooks publications

You can search for, view, or download Redbooks publications, Redpaper publications, Technotes, draft publications, and Additional materials, as well as order hardcopy Redbooks publications, at this Web site:

ibm.com/redbooks

Help from IBM

IBM Support and downloads:

ibm.com/support

IBM Global Services:

ibm.com/services

Index

E

F

G

H

I

J

L

M

N

O

P

Q

R

S

T

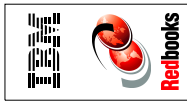
U

V

W

X

Z



Implementing an IBM System x iDataPlex Solution

(0.5” spine)
0.475”<->0.875”
250 <-> 459 pages

Implementing an IBM System x iDataPlex Solution

Introduces the next generation data center solution for scale-out environments

Describes custom configurations designed to maximize compute density

Addresses power, cooling, and physical space requirements

The IBM iDataPlex data center solution is for Web 2.0, high performance computing (HPC) cluster, and corporate batch processing customers experiencing limitations of electrical power, cooling, physical space, or a combination of these. By providing a big picture approach to the design, iDataPlex uses innovative ways to integrate Intel-based processing at the node, rack, and data center levels to maximize power and cooling efficiencies while providing necessary compute density.

An iDataPlex rack is built with industry-standard components to create flexible configurations of servers, chassis, and networking switches that integrate easily. Using technology for flexible node configurations, iDataPlex technology can configure customized solutions for applications to meet specific business needs for computing power, storage intensity, and the right I/O and networking.

This IBM Redbooks publication is for customers who want to understand and implement the IBM iDataPlex solution. It introduces the iDataPlex solution and the innovations in its design, outlines its benefits, and positions it with IBM System x and BladeCenter servers. The book provides details of iDataPlex components and the supported configurations. It describes application considerations for an iDataPlex solution and aspects of data center design that are central to that solution. The book concludes by introducing the services offerings available from IBM for planning and installing an iDataPlex solution.

SG24-7629-04 ISBN 0738436631

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks