Redpaper

Martin Bachmaier
David Watts

QLogic QDR InfiniBand Solutions for IBM iDataPlex

InfiniBand® is an industry-standard, high-speed interconnect mainly used in the area of high performance computing where very low latencies and a high throughput are required. During recent years, the InfiniBand technology has entered additional markets such as digital media or enterprise backends that host large clustered systems.

This IBM® Redpaper publication introduces the new Quad Data Rate (QDR) InfiniBand products from QLogic® and other vendors that are supported on the IBM iDataPlex™ large-scale computing solution.

Contents

Introduction

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 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 provides 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, 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 similar devices as an Ethernet network. InfiniBand relies on a switched fabric to which so called *channel adapters* (CAs) connect. There are two types of channel adapters:

- Host channel adapters (HCA), which reside in the compute nodes (for example servers)
- Target channel adapters (TCA), which are installed in the other end of the fabric (for example storage devices)

CAs connect to each other either using copper or fibre cables.

Figure 1 illustrates the elements of an InfiniBand network.

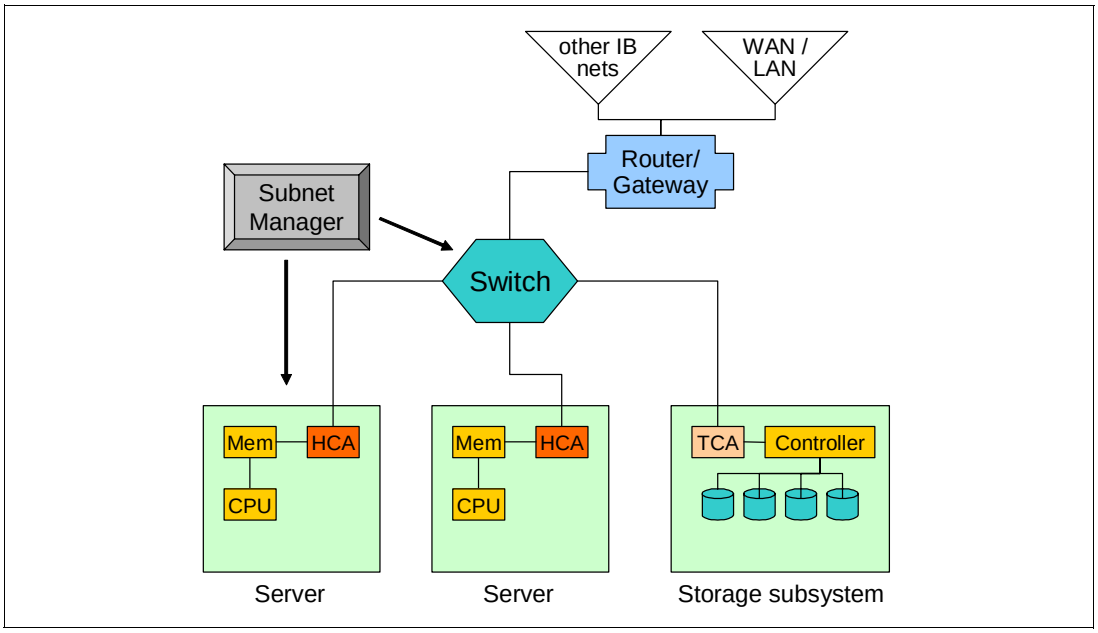

Figure 1 Elements of an InfiniBand network

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

InfiniBand speed terminology

InfiniBand uses a serial link with a signalling rate of 2.5 Gbps in each direction, 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 MBps) net data rate on a single 1x link in each direction.

Table 1 shows the theoretical net throughput for all combinations.

Table 1 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	8 Gbps	32 Gbps	64 Gbps	96 Gbps

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

The roadmap of the InfiniBand Trade Association shows the next standard to be 8x data rate (abbreviated EDR) and is scheduled for 2011.[1] EDR doubles the QDR throughput numbers to a data rate of 192 Gbps (EDR 12x).

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 m to 15 m. This limitation can be mitigated using optical cables. Distances of up to 300 m are possible, and a smaller bend radius can be achieved. The cable length limitation becomes more important for high-density installations with a large number of cables going into a single rack.

For links up to 4x DDR speed, cables with a CX4 type connector are used (sometimes also referred to as *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 is made smaller as well. A new connector called *Quad Small-Form-factor-pluggable* (QSFP) was designed that houses four independent transmit and receive channels, each capable of transmitting at 10 Gbps for an aggregated bandwidth of 40 Gbps per cable.

[1] Check the InfiniBand TA Web site for the latest version of the roadmap at http://www.infinibandta.org/content/pages.php?pg=technology_overview

Figure 2 shows the CX4 and QSFP connectors.

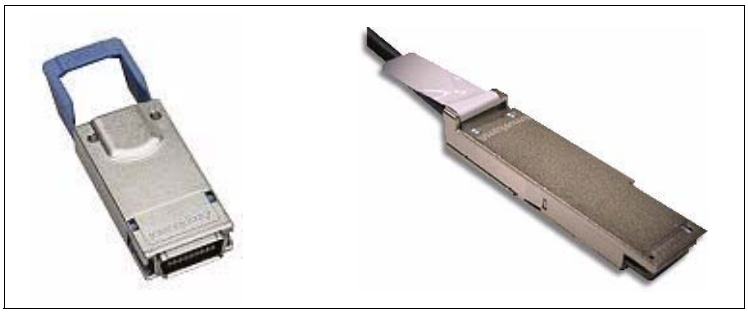

Figure 2 InfiniBand CX4 (left) and QSFP connectors (right)

Table 2 lists the QDR cables with QSFP connectors on both ends, which are tested officially in an iDataPlex environment.

Table 2 Supported QDR QSFP InfiniBand cables in iDataPlex

	Copper or Optical	**Length**	**Feature Code**
QLogic InfiniBand Cable 30AWG	Copper	0.5 m	3725
QLogic InfiniBand Cable 30AWG	Copper	1 m	3726
QLogic InfiniBand Cable 28AWG	Copper	3 m	3727
QLogic InfiniBand Cable	Optical	3 m	3731
QLogic InfiniBand Cable	Optical	10 m	3732
QLogic InfiniBand Cable	Optical	30 m	3733

If you want 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 3 shows the supported QSFP-to-CX4 cables.

Table 3 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

InfiniBand versus 10 Gbps Ethernet

A few years ago, the differentiation between Ethernet and InfiniBand was easy to describe. On the one side, 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 that are available at reasonable prices.

On the other side, there is InfiniBand, the interconnect technology that offers very low latency at very high data rates but which might require significant investment in a new network technology infrastructure. This situation has shifted with the adoption of 10 Gbps Ethernet and the expansion of InfiniBand towards more general market areas. Although both technologies came closer to each other, InfiniBand still has 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) is one of the major technical advantages of InfiniBand that allows you to copy data between memory on different nodes without consuming CPU cycles for the transfer process. The RDMA is implemented in hardware and decreases Ethernet latency. However, as of 2009, although 10 Gbps equipment is available from most vendors, it is still more costly than InfiniBand, especially when it comes to larger installations with a high port count.

Alternatively, InfiniBand has became a more 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 achieve a better scalability and a higher throughput when using a shared file system. Even with the rise of 10 Gbps Ethernet and iWARP, InfiniBand still has a 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 to provide 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,[2] InfiniBand has the potential to be a major driver for the adoption of cloud computing in today's data center. The key features of InfiniBand, such as node-to-node latency and scalability, partnered with its reliability and low cost make InfiniBand a good solution in cloud computing environments.[2] The Taneja Group notes that external service providers are using InfiniBand infrastructures because of the advantage of a reduced number of interfaces.[3]

The new QDR InfiniBand products especially offer features that ease transition to a unified fabric. Features such as quality of service (QoS), virtual HCAs, and virtual NICs help with comprehensive, virtual I/O system management. The InfiniBand channel adapter design allocates high bandwidth to each virtual server but also consolidates LAN and WAN traffic, which "greatly reduces network complexity while still delivering low-latency and high-bandwidth capabilities."[2]

---

2 IDC Analyst Connection, *InfiniBand: Poised for Market Growth*, available online at: http://idcdocserv.com/784

3 Cited in an IBTA press release, which is available at: http://www.infinibandta.org/content/pages.php?pg=press_room_item&rec_id=540

QDR HCAs and switches

The following section describes the set of QDR products supported in the iDataPlex rack, which consist of two HCAs and one QLogic switch.

HCA

Table 4 shows the two QDR HCAs from Mellanox that are supported in iDataPlex servers.

Table 4 Supported iDataPlex QDR InfiniBand HCAs

Description	Feature code	Host Interface	Ports	Media
Mellanox ConnectX Single-port 4x QDR HCA	3597	PCIe 2.0 x8	1	CX4 or QSFP
Mellanox ConnectX Dual-port 4x QDR HCA	3596	PCIe 2.0 x8	2	CX4 or QSFP

Table 5 lists the supported iDataPlex servers.

Table 5 iDataPlex servers that support the InfiniBand HCAs

Description	Feature code	dx320	dx340	dx360	dx360 M2
Mellanox ConnectX Single-port 4x QDR HCA	3597	No	No	No	Yes
Mellanox ConnectX Dual-port 4x QDR HCA	3596	No	No	No	Yes

Switches

To support the QDR technology, QLogic introduced a next generation of switching ASICs called *TrueScale*. This advanced chipset offers a higher port count (36 instead of 24) and can run each port at up to 4x QDR speed. The TrueScale chip architecture is used throughout the new QLogic 12000 series of QDR switches. It provides the following key features:

- Total Switching capacity of 2.88 Tbps per TrueScale Switch ASIC
- Flexible port configuration, either 36 4x QDR ports or 18 8x QDR ports or any combination of 4x and 8x ports
- Deterministic latency of <150 ns at >90% loading or 140 ns at ≤90% loading.
- Adaptive routing ensures high performance in big fabrics with many routes
  - Balances traffic across alternative routes
  - Path adjustments in ASIC firmware with adjustment rates < 1 ms
- Advanced QoS through support for virtual fabrics, to partition the cluster (enables you to assign different QoS levels to different nodes or sub-clusters)
- MTU sizes of 256, 512, 1024, 2048, and 4096 bytes supported
- Virtual lanes (1, 2, 4, or 8 supported)
- Manageability through an embedded 32-bit RISC processor

Table 6 lists the supported switches.

Table 6 Supported iDataPlex QDR InfiniBand switches

	Feature code	Ports	Port speed
QLogic 12200	6925	36	4x QDR (or 18 ports running at 8x QDR)

Figure 3 shows the 12200-series entry-level switch, the QLogic 12200-36 Fixed Configuration switch with one TrueScale chip onboard. This switch is the only QDR InfiniBand switch that is supported in iDataPlex.

Feature code 6925 bundles the following items:

- QLogic 12200 Switch Chassis for iDataPlex
- QLogic 12200 4point rack mount kit
- QLogic InfiniBand Fabric Suite (IFS) for 12x00, including
  - FastFabric Toolset
  - Subnet Manager
  - Fabric Viewer
- QLogic IFS 2008 - 36 ports, 1 year with 24x7 support

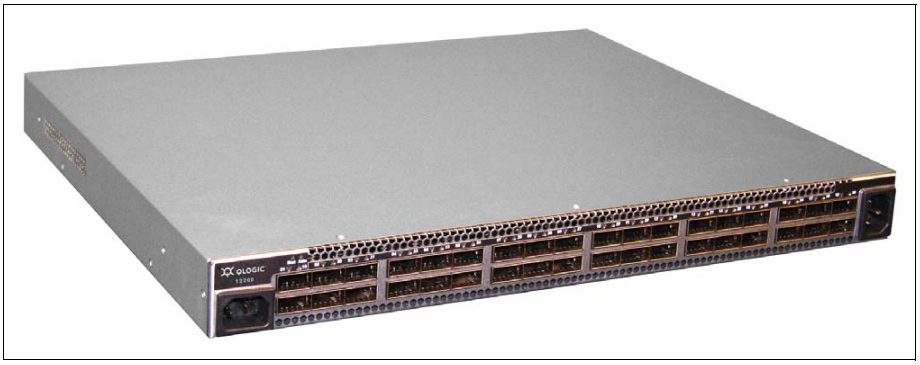

Figure 3 QLogic 12200-36 InfiniBand Switch

The QLogic 12200 switch has the following technical features:

- Fabric capacity: 2.88 Tbps
- High edge port density, 36 4x QDR ports or 18 8x QDR ports
- Scalable low latency (<140 ns to <420 ns)
- Eight virtual lanes, plus one management
- Adaptive routing
- Virtual fabrics and segmentation
- Fully redundant power supplies and fans
- Supports Quad Small Form Factor Pluggable (QSFP) optical and copper cable specifications; CX4/microGigaCN

QLogic InfiniBand Fabric Suite

InfiniBand management tools have built-in intelligence to enable rapid, error-free installation and provisioning of small to extremely large server clusters with thousands of nodes, providing the most optimized fabric possible to support today's cluster aware applications.

The QLogic InfiniBand Fabric Suite is a management solution that is designed to help with configuration, administration, and debugging of the InfiniBand fabric. It is supported on 32-bit and 64-bit Intel® and AMD™ platforms running on the following host operating systems:

- Red Hat® Linux®
- SUSE® Linux

InfiniBand Fabric Suite includes the following major components:

- Fabric Viewer
- Subnet Manager
- FastFabric Toolset

Fabric Viewer

The Fabric Viewer is a stand-alone Java™-based GUI that provides access to all Subnet Manager facilities.

Subnet Manager

You use the Subnet Manager to manage general fabric configuration. It runs either embedded or on a dedicated server. A fabric can run more than one Subnet Manager for redundancy purposes.

Note: If you segment your fabric into multiple independent fabrics, you need one Subnet Manager per segment.

Key functions of the Subnet Manager include:

- Fabric and chassis management
- Path and route management
- Subnet management and administration
- Device discovery
- SNMP management

You can access the Subnet Manager either through the Fabric Viewer stand-alone GUI or through an HTML interface with any Web browser.

FastFabric Toolset

The FastFabric Toolset is a set of tools that automate cluster installation, verification and debugging, as well as administration and updating of both compute hosts and switches in an InfiniBand fabric. Typical tasks that can be accomplished using FastFabric are:

- Switch tasks
  - Parallel installation of IB switch firmware on all switches in the fabric
  - Parallel configuration of all IB switches with common parameters such as NTP or syslog information
  - Central point of administration to verify switch port status
  - Verify all links in a fabric (helps to find broken cables or missing links)
  - Ensure consistent switch configurations throughout the fabric (helps to find speed or MTU mismatches)
- Host tasks
  - Parallel installation and configuration of IB software stack on all hosts
  - Parallel configuration of MPI parameters on all hosts
  - Automated setup of network parameters (ssh, IP over IB)
  - Verify MPI latency and bandwidth performance on a per-node basis

The FastFabric Toolset has a convenient console-based menu. All FastFabric commands can also be scripted.

Example 1 shows the menu for the switch chassis.

Example 1 IFS FastFabric switch chassis menu

```
Fast Fabric IB Chassis Setup/Admin Menu
Chassis List: /etc/sysconfig/iba/chassis

Setup:
0) Edit Config and Select/Edit Chassis Files [Perform]
1) Verify Chassis via Ethernet ping          [Perform]
2) Update Chassis Firmware                   [Perform]
3) Set up Chassis Basic Configuration        [Perform]
4) Setup Password-less ssh/scp               [Perform]
5) Reboot Chassis                            [Perform]
6) Configure Chassis Fabric Manager          [Perform]
Admin:
7) Check Status of IB Ports                  [ Skip  ]
8) Restart Chassis Fabric Manager            [ Skip  ]
9) Generate all Chassis Problem Report Info  [ Skip  ]
a) Run a command on all chassis              [ Skip  ]
Review:
b) View iba_chassis_admin result files       [ Skip  ]
P) Perform the selected actions              N) Select None
X) Return to Previous Menu (or ESC)
```

Example 2 shows the menu with the available tasks for InfiniBand hosts.

Example 2 IFS FastFabric host menu

```
Fast Fabric IB Host Setup Menu
Host List: /etc/sysconfig/iba/hosts

Setup:
0) Edit Config and Select/Edit Hosts Files   [Perform]
1) Verify Hosts via Ethernet ping            [Perform]
2) Setup Password-less ssh/scp               [Perform]
3) Copy /etc/hosts to all hosts              [Perform]
4) Show uname -a for all hosts               [Perform]
5) Install/Upgrade QLogic IB Software        [Perform]
6) Configure IPoIB IP Address                [Perform]
7) Build MPI Test Apps and Copy to Hosts     [Perform]
8) Reboot Hosts                              [Perform]
Admin:
9) Refresh ssh Known Hosts                   [ Skip  ]
a) Rebuild MPI Library and Tools             [ Skip  ]
b) Run a command on all hosts                [ Skip  ]
c) Copy a file to all hosts                  [ Skip  ]
Review:
d) View iba_host_admin result files          [ Skip  ]

P) Perform the selected actions              N) Select None
X) Return to Previous Menu (or ESC)
```

Value of InfiniBand across applications

In summary, InfiniBand is an efficient I/O technology that provides high-speed data transfers and ultra low latencies for computing and storage over a single fabric that is highly reliable and scalable. The QLogic InfiniBand architecture and solutions are devised to spend the least amount of time in parallel communication so that applications can make better use of computation requirements and provide converge to answers faster. Depending on you usage model, industry-standard applications can derive the benefits shown in Table 7 using InfiniBand.

Table 7 Benefits of using InfiniBand by industry

Vertical market	Application segment	InfiniBand value
Oil and Gas	Mix of ISP and Home grown codes: Reservoir Modeling	Low latency, high bandwidth
CAE	Mostly ISV codes: Crash, Air flow and Fluid flow simulations	High message rate, low latency, scalability
Government	Home grown codes: Labs, Defense, Weather, wide range of applications	High message rate, low latency, scalability, high bandwidth
Education	Home grown and open source codes: wide range of applications	High message rate, low latency, scalability, high bandwidth
Financial	Mix of ISP and Home grown codes: Market simulation and trading floor	High performance IP, scalability, high bandwidth
Life and Materials Science	Mostly ISV codes: Molecular Simulation, Computational Chemistry and Biology applications	Low latency, high message rates

The team that wrote this paper

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

Martin Bachmaier is an IT Architect in the IBM hardware development lab in Boeblingen, Germany, and leads the Open Systems Design and Development infrastructure team where systems using IBM next generation processors are developed. His current responsibilities include the architecture and management of Linux-based infrastructures and HPC cluster environments. He has worked for IBM for more than 5 years and most recently is involved in cloud activities at IBM. He holds a degree in Computer Science from the University of Cooperative Education in Stuttgart, Germany, and a Bachelor of Science (Hons) from the Open University in London, UK. He is an IBM Certified Systems Expert and holds CCNA and CCNA Security credentials.

David Watts is a Consulting IT Specialist at the IBM ITSO Center in Raleigh. He manages residencies and produces IBM Redbooks® publications on hardware and software topics related to IBM System x® and BladeCenter® servers as well as 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.

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.

This document REDP-4546-00 was created or updated on July 23, 2009.

Send us your comments in one of the following ways:
- Use the online **Contact us** review Redbooks form found at:
  ibm.com/redbooks
- Send your comments in an e-mail to:
  redbooks@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

Redpaper™

Trademarks