

IBM System p5 Quad-Core Module Based on POWER5+ Technology: Technical Overview and Introduction

The quad-core module (QCM) is based on the well-known POWER5™ dual-core module (DCM) technology. The dual-core POWER5 processor and the dual-core POWER5+™ processor are packaged with the L3 cache chip into a cost-effective DCM package. The QCM is a package that enables entry-level or midrange IBM® System p5™ servers to achieve additional processing density without increasing the footprint. Figure 1 shows the DCM and QCM physical views and the basic internal architecture.

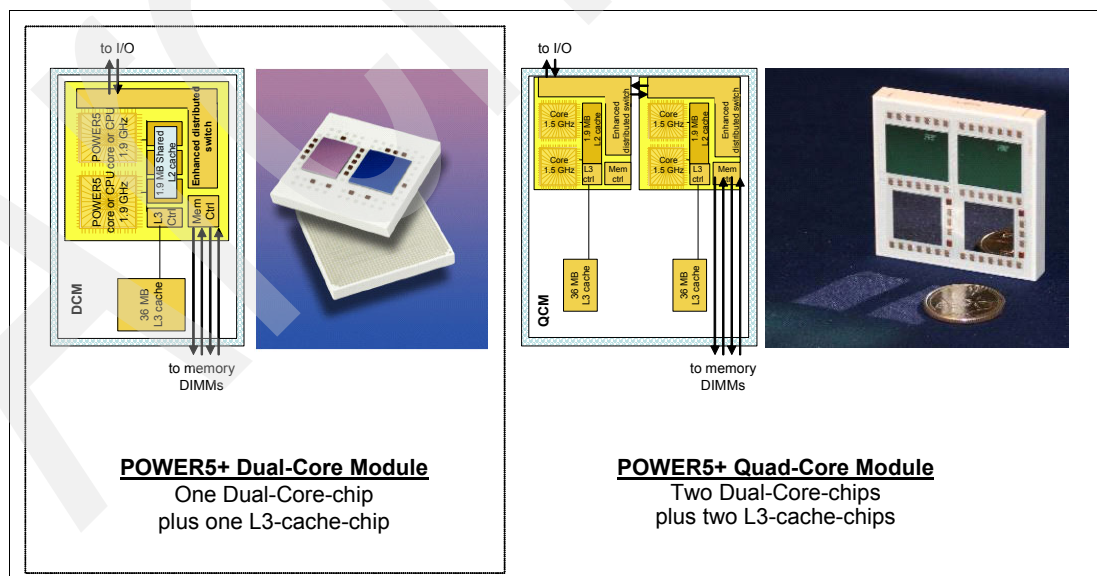


Figure 1 DCM and QCM physical views and basic internal architecture

The POWER5+ processor

The POWER5+ processor features single-threaded and multi-threaded execution for higher performance. A single die contains two identical processor cores, each of which uses simultaneous multithreading to supporting two logical threads. This architecture makes a single dual-core POWER5+ processor appear to be a four-core symmetric multiprocessor to the operating system. The POWER5+ processor supports the 64-bit PowerPC® architecture.

The POWER5+ processor has a 1.9 MB on-chip L2 cache that is implemented as three identical slices with separate controllers for each. Either processor core can independently access each L2 controller. The L3 cache, with a capacity of 36 MB, operates as a back door with separate buses for reads and writes that operate at half processor speed.

Figure 2 shows a high-level view of the POWER5+ processor.

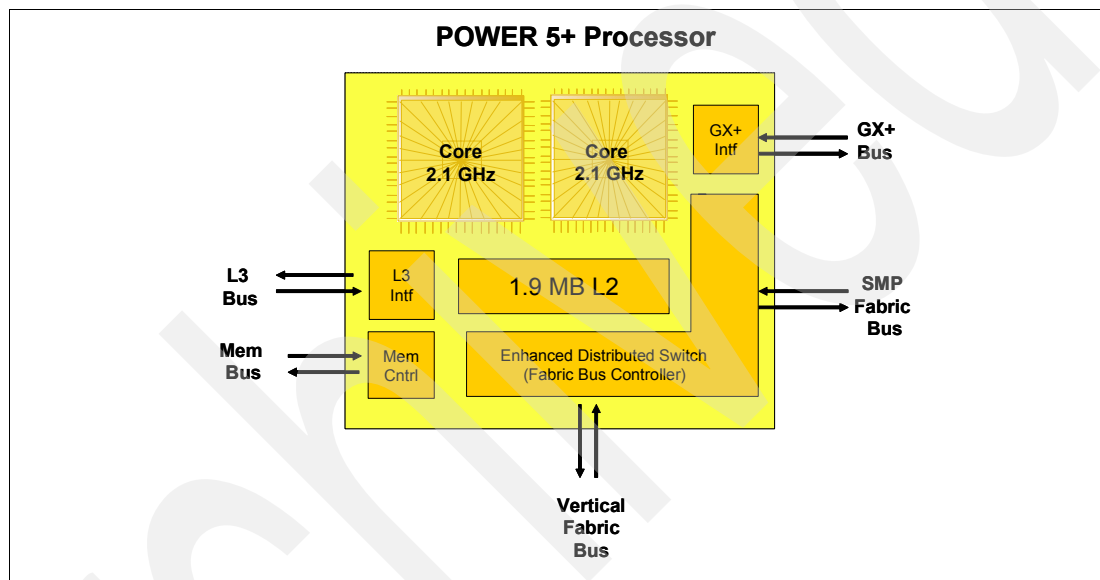


Figure 2 POWER5+ processor

The CMOS9S technology that is utilized by the POWER5 chip used a 130 nm fabrication process. The POWER5+ processor CMOS10S technology uses a 90 nm fabrication process, which allows:

- ▶ Performance gains through faster clock rates
- ▶ Chip size reduction (243 mm versus 389 mm)

Compared to the POWER5 processor, the 37% smaller POWER5+ processor consumes less power, thus requiring less sophisticated cooling. As a result, it can be used in servers that previously only used low-frequency chips.

The POWER5+ design offers the following additional enhancements over its predecessor:

- ▶ New page sizes in effective-to-real address translation (ERAT) and translation look-aside buffer (TLB): Two new page sizes (64 KB and 16 GB) were recently added in PowerPC architecture that is the common base architecture for all POWER™ and PowerPC processors.
- ▶ New segment size in segment lookaside buffer (SLB): One new segment size (1 TB) was recently added in PowerPC architecture.
- ▶ TLB size has been doubled in POWER5+ over POWER5: TLB in POWER5+ has 2048 entries.

- ▶ Floating-point round-to-integer instructions: New instructions (frfin, frfiz, frfip, frfim) have been added to round floating-point integers with the following rounding modes: nearest, zero, integer plus, integer minus.
- ▶ Improved floating-point performance.
- ▶ Lock performance enhancement.
- ▶ Enhanced SLB read.
- ▶ True Little-Endian mode: There is support for the True Little-Endian mode as defined in the PowerPC architecture.
- ▶ 2 x SMP support: Changes have been made in the fabric, L2 and L3 controller, memory controller, GX+ controller, and ChipRAS to provide support for the QCM that enables the SMP system sizes to be double what is available in POWER5 DCM-based servers. However, current POWER5+ implementations support only a single-address loop.
- ▶ Enhanced memory controller: Several enhancements have been made in the memory controller for improved performance.
- ▶ Enhanced redundancy in L1 Dcache, L2 cache, and L3 directory: Independent control of the L2 cache and the L3 directory for redundancy has been added to enable split-repair action. More word line redundancy has been added in the L1 Dcache. In addition, Array Built-In Self Test (ABIST) column repair for the L2 cache and the L3 directory has been added.

Simultaneous multithreading

As a permanent requirement for performance improvements at the application level, simultaneous multithreading that was introduced in POWER5 is also embedded in the POWER5+ processor. Developers are familiar with process-level parallelism (multi-tasking) and thread-level parallelism (multi-threads). Simultaneous multithreading, the next stage of processor saturation for throughput-oriented applications, introduces the method of instruction group-level parallelism for supporting multiple pipelines to the processor. The instruction groups are chosen from different hardware threads that belong to a single operating system (OS) image.

Simultaneous multithreading is activated by default when an OS that supports it is loaded. On a four-core POWER5+ processor-based system, the OS discovers the available processors as an eight-core system. To achieve a higher performance level, simultaneous multithreading is also applicable in Micro-Partitioning™, capped or uncapped, and dedicated partition environments. Simultaneous multithreading is supported in POWER5+ processor-based systems running AIX® 5L™ V5.3 or Linux®-based systems. (A 2.6 kernel is required.)

Micro-Partitioning technology

Micro-Partitioning offers a virtualization of system resources. In POWER5+ processor-based systems, physical resources are abstracted into virtual resources that are available to partitions. POWER5+ Micro-Partitioning specifies processor capacity in processing units. One processing unit represents 1% of one physical processor, and 1.0 represents the power of one processor. A partition defined with 220 processing units is equivalent to the power of 2.2 physical processors. Creating a partition using Micro-Partitioning technology, the minimum capacity is 10 processing units, or one-tenth of a physical processor. A maximum of 10 partitions for each physical processor core can be defined, but on a loaded system, the practical limit might be lower. Any QCM, using Micro-Partitioning technology, enables you to define up to 40 partitions that can be activated at the same time for an entire system. The practical limit to the number of partitions is based on available hardware and performance objectives.

Micro-Partitions can also be defined with capped and uncapped attributes. A capped Micro-Partition is not allowed to exceed the defined capacity, but an uncapped partition is allowed to consume additional capacity with fewer restrictions. Uncapped partitions can be configured to the total idle capacity of the server or a percentage of it.

The POWER5+ processor-based systems use the POWER Hypervisor™ for executing the Micro-Partition model. The POWER Hypervisor operates continuously in the background.

Dynamic power management

In current complementary metal oxide semiconductor (CMOS) technologies, chip power is one of the most important design parameters. With the introduction of simultaneous multithreading, more instructions run per cycle per processor core, thus increasing the total switching power of the core and the chip. To reduce switching power, POWER5+ processors extensively use a fine-grained, dynamic clock-gating mechanism. This mechanism gates off clocks to a local clock buffer if dynamic power management logic knows that the set of latches that are driven by the buffer will not be used in the next cycle. This allows substantial power saving with no performance impact. In every cycle, the dynamic power management logic determines whether a local clock buffer that drives a set of latches can be clock-gated in the next cycle.

Leakage power has also become a performance limiter. To reduce leakage power, the POWER5+ processor uses transistors with low threshold voltage only in critical paths. The POWER5+ processor also has a low-power mode that is enabled when the system software instructs the hardware to execute both threads at the lowest available priority. In low power mode, instructions dispatch once every 32 cycles at most, further reducing switching power. The POWER5+ processor uses this mode only when there is no ready task to run on either thread.

AIX 5L support for new features in the POWER5+ processor

POWER5+ processor-based systems introduced support for two new virtual memory page sizes, namely the 64 KB and 16 GB, as well as support for 1 TB segment size. Although 16 GB pages are intended to be used only in very high performance environments, 64 KB pages are general purpose. AIX 5L Version 5.3 with the 5300-04 Technology Level and 64-bit kernel is required for 64 KB and 16 GB page size support.

As with all previous versions of AIX, 4 KB is the default page size. A process continues to use 4 KB pages unless a user specifically requests a different page size. AIX 5L has rich support of 64 KB pages. They are easy to use, and many applications should see performance benefits when using 64 KB pages rather than 4 KB pages. No system configuration changes are necessary to enable a system to use 64 KB pages; they are fully pageable, and the size of the pool of 64 KB page frames on a system is dynamic and fully managed by AIX.

The main benefit of a larger page size is improved performance for applications that allocate and repeatedly access large amounts of memory. This performance improvement comes from the processing usage that is required to translate a page address as it is used in an application to a page address that is understood by the memory subsystem of the computer. To improve performance, the information that is needed to translate a given page is usually cached in the processor. In POWER5+, this cache takes the form of a TLB. Because of the limited number of TLB entries, using a large page size increases the amount of address space that can be accessed without incurring translation delays. Also, the size of TLB in POWER5+ processor-based systems has been doubled compared to POWER5 systems.

Huge pages (16 GB) are intended to be used only in very high performance environments, and AIX 5L does not automatically configure a system to use these page sizes. A system

administrator must configure AIX 5L to use these page sizes and specify their number using the Hardware Management Console (HMC) before partition start.

A user can specify page sizes to use for the address space of the processes of three regions with an environment variable or with settings in the XCOFF binary of the application using the **ledit** or **ld** commands. These three regions are data, stack, and program text. An application programmer can also select the page size to use for System V shared memory with a new **SHM_PAGESIZE** command for the **shmctl()** system call.

An example of using system variables to start a program with 64 KB page size support is:

```
LDR_CNTRL=DATAPSIZE=64K@TEXTPSIZE=64K@STACKPSIZE=64K <program>
```

Systems commands (**ps**, **vmstat**, **svmon**, **pagesize**) have been enhanced to report various page size usage.

QCM support for IBM System p5 servers

The following list represents, at the time of writing, the POWER5+ processor-based servers that support the QCM implementation:

- ▶ IBM System p5 505Q
- ▶ IBM System p5 510Q
- ▶ IBM System p5 520Q
- ▶ IBM System p5 550Q
- ▶ IBM System p5 560Q

In the p5-505Q, p5-510Q, and p5-520Q servers, the QCM is plugged directly into the main system board. The p5-550Q and p5-560Q servers use the QCM together with eight memory DIMM sockets and two memory interfaces inside a processor card package.

QCM architecture

The storage structure for the POWER5+ processor is a distributed memory architecture that provides high memory bandwidth. Each processor in the QCM can address all memory and sees a single shared memory resource. In the QCM, one POWER5+ processor has direct access to eight DDR2 memory slots that are controlled by two Synchronous Memory Interface (SMI-II) chips, which are located in close physical proximity to the processor modules. The other POWER5+ processor has access to the same memory slots through the Vertical Fabric Bus. Figure 3 on page 6 shows a layout of the QCM with associated memory.

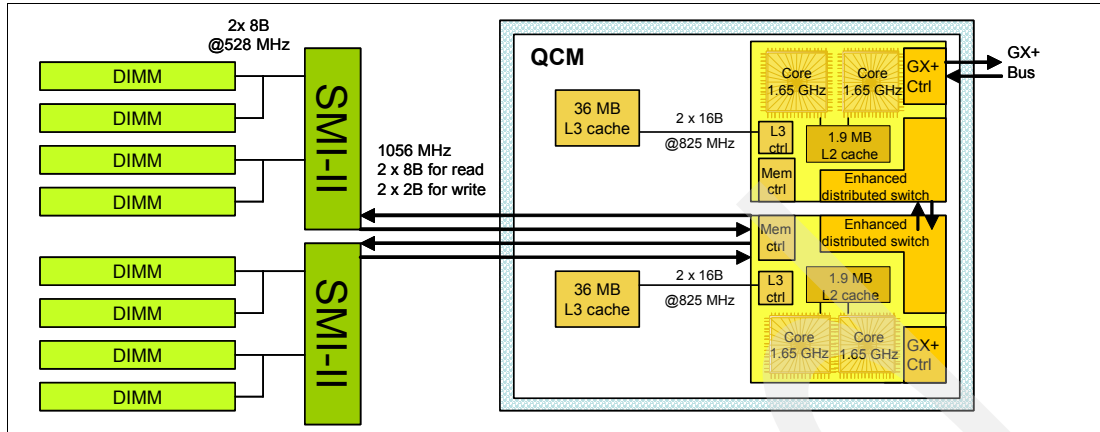


Figure 3 POWER5+ QCM with DDR2 memory socket layout view

I/O connects to QCM using the GX+ bus. The QCM provides a single GX+ bus. Each processor in the POWER5+ processors has either a direct access to the GX+ bus using its GX+ bus controller or uses the Vertical Fabric Bus, which is controlled by the Fabric Bus controller. The GX+ bus provides an interface to I/O devices. The POWER5+ processor that does not have direct access to memory does have a direct access to the GX+ bus.

When more than one QCM is present in a system, the QCM-to-QCM communication is implemented with the Horizontal Fabric Bus.

Figure 4 shows a high-level layout of a configuration with two QCM processor cards.

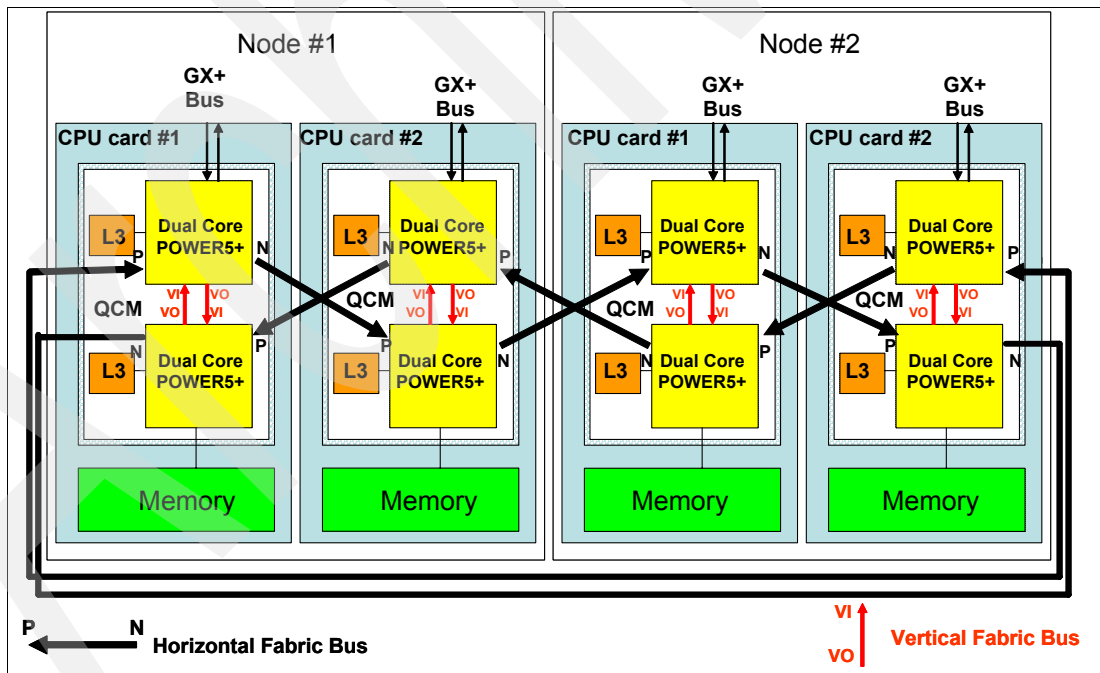


Figure 4 A System p5 16-core configuration using two QCMs

The theoretical maximum throughput of the L3 cache is 16 byte read, 16 byte write at a bus frequency of 825 MHz (based on a 1.65 GHz processor clock), which equates to 26400 MBps or 26.4 GBps per L3 cache. The two L3 caches on the QCM provide a total L3 cache bandwidth of 52800 MBps or 52.8 GBps per QCM.

Available processor speeds

At the time of writing, the QCM operates a processor clock rate of 1.5 or 1.65 GHz. To verify the processor characteristics in a system running AIX 5L, use one of the following commands:

► **lsattr -El procX**

X is the number of the processor (for example, proc0 is the first processor in the system). The output from the command is similar to the following output (*False*, as used in this output, signifies that the value cannot be changed through an AIX 5L command interface):

frequency	1498500000	Processor Speed	False
smt_enabled	true	Processor SMT enabled	False
smt_threads	2	Processor SMT threads	False
state	enable	Processor state	False
type	powerPC_POWER5	Processor type	False

► **pmcycles -m**

This command (available with AIX 5L) uses the performance monitor cycle counter and the processor real-time clock to measure the actual processor clock speed in MHz. The following output is from a single QCM running at 1.5 GHz with simultaneous multithreading enabled:

```
Cpu 0 runs at 1498 MHz
Cpu 1 runs at 1498 MHz
Cpu 2 runs at 1498 MHz
Cpu 3 runs at 1498 MHz
Cpu 4 runs at 1498 MHz
Cpu 5 runs at 1498 MHz
Cpu 6 runs at 1498 MHz
Cpu 7 runs at 1498 MHz
```

Note: The **pmcycles** command is part of the **bos.pmapi** fileset. This component must be installed before using the **lslpp -l bos.pmapi** command.

Memory subsystem

Each QCM is connected to eight slots for up to eight pluggable DDR2 DIMMs. Every group of four memory slots is accessed by one enhanced SMI-II chip. (See Figure 3 on page 6.) The SMI-II is located between memory DIMMs and the processor, enabling high-speed data flow with 16 addressable x 72 bits wide x 4 transfer deep writing buffer and 8 high x 72 bits wide x 4 transfer deep read buffer.

The memory subsystem throughput depends on the speed of the memory. An elastic interface, contained in any POWER5+ processor, buffer reads and writes to and from memory and the processor. Each SMI-II chip provides the memory controller with a single 8-byte read and 2-byte write high-speed Elastic Interface-II bus. A DDR2 bus enables double reads or double writes per clock cycle. Because the bus operates at 1066 MHz, the peak processor-to-memory throughput for read is $(8 \times 2 \times 1056) = 16896$ MBps or 16,89 GBps. The peak processor-to-memory throughput for write is 4224 MBps or 4.22 GBps, totalling 21.12 GBps.

The DIMM technology that is supported by the QCM is the DDR2, operating at 528 MHz through four 8-byte paths. Read and write operations share these paths. At least four DIMMs must be installed to effectively use each path. In this case, the throughput between the SMI-II and the DIMMs is $(8 \times 4 \times 528)$, which is equal to 16,89 GBps.

I/O subsystem

Each QCM has a GX+ bus that is used to connect to an internal System p5 I/O subsystem or to an additional GX+ expansion card that connects external I/O subsystems, including InfiniBand I/O subsystems or switches.

The team that wrote this Redpaper

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

Giuliano Anselmi is a certified IBM eServer™ pSeries® Presales Technical Support Specialist in the Field Technical Sales Support group based in Rome, Italy. For seven years, he was a pSeries Systems Product Engineer, supporting the Web Server Sales Organization in EMEA, IBM Sales, IBM Business Partners, Technical Support Organizations, and IBM Dublin eServer Manufacturing. Giuliano has worked for IBM for 14 years, focusing on RS/6000® and pSeries systems and using his in-depth knowledge of related hardware and solutions.

Bernard Filhol is a UNIX® Server Customer Satisfaction Resolution Team Leader for NEE and SWE IOTs in Montpellier, France. He has more than 25 years of experience in mainframes and five years of experience in pSeries Customer Satisfaction. He holds a degree in Electronic from Montpellier University Institute of Technology. His areas of expertise include Mainframe Channel Subsystems, FICON®, and pSeries RAS. He has written extensively about FICON.

SahngShin Kim is a Sales Specialist for the IBM Systems and Technology Group (STG) infra-solution sales team in Seoul, Korea. For three years he was a Sales Specialist for pSeries, and he spent two years as a Sales Specialist for Grid computing and one year for Infra-solutions. SahngShin has worked for IBM for six years, with a focus on RS/6000, pSeries systems, and STG server products and their architecture.

Gregor Linzmeier is an IBM Advisory IT Specialist for RS/6000 and pSeries workstation and entry servers as part of the STG in Mainz, Germany. He supports IBM sales, Business Partners, and customers with pre-sales consultation and implementation of client/server environments. He has worked for more than 15 years as an infrastructure specialist for RT, RS/6000, and AIX 5L in large CATIA client/server projects.

Ondrej Plachy is an IT specialist at IBM Czech Republic. He is responsible for project design, implementation, and support of large-scale computer systems. He has 11 years of experience in the UNIX field. He holds a degree in Computer Science from Czech Technical University (CVUT), Prague. He has worked at the Supercomputing Centre of Czech Technical University for four years, and has been with IBM for seven years with the AIX 5L support team.

Scott Vetter, who managed this document project, is a Certified Executive Project Manager at the International Technical Support Organization, Austin Center. He has worked for IBM for more than 21 years in a variety of challenging roles. His latest efforts are directed at providing world-class IBM System p and UNIX Redbooks, white papers, and workshop collateral.

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

Larry Amy, Baba Arimilli, Ron Arroyo, Terry Brennan, Erin Burke, Mark Dewalt, Bob Foster, Ron Gonzalez, Dan Henderson, Hal Jennings, Carolyn Jones, Bill Mihaltse, Thoi Nguyen, Ken Rozendal, Craig Shempert, Dave Willoughby, David A. Hepkin, Carlo Costantini, Charlie Cler
IBM U.S.

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 created or updated on August 21, 2006.

Send us your comments in one of the following ways:

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




Trademarks

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

AIX®
AIX 5L™
eServer™
FICON®
IBM®

Micro-Partitioning™
PowerPC®
POWER™
POWER Hypervisor™
POWER5™

POWER5+™
pSeries®
Redbooks (logo) ™
RS/6000®
System p5™

The following terms are trademarks of other companies:

UNIX is a registered trademark of The Open Group in the United States and other countries.

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

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