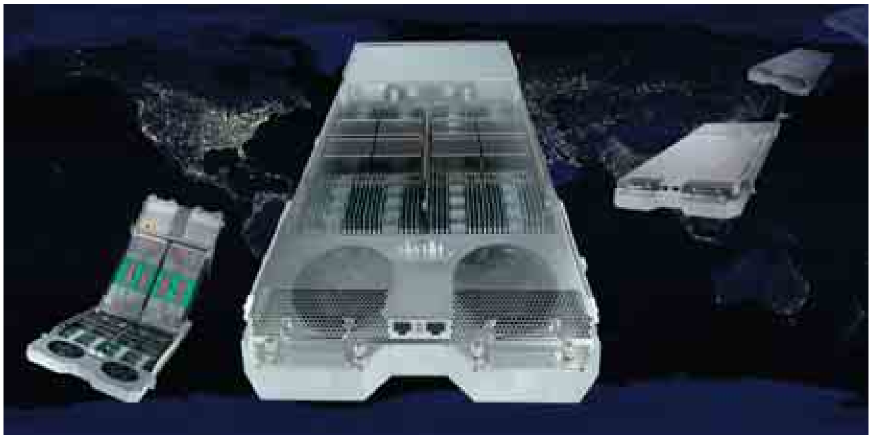
Ultra-dense, high-performance, supercomputer building block

IBM *e*server® p5 575 cluster node



*e*server *p5 575 cluster node*

Highlights

- ***High-powered building block designed for the most demanding HPC and BI applications***
- ***Combines powerful IBM POWER5™ processors with extraordinary memory bandwidth***
- ***Innovative packaging helps to minimize space requirements, reduce environmental demands and simplify system management***
- ***Can be configured with up to 128 nodes in a single Cluster 1600 system—a 1024 CPU supercomputer***

The adage “time is money” defines today’s supercharged, high-performance economy. Organizations involved in engineering problem solving, drug design, oil reservoir modeling, weather forecasting, financial simulation and business intelligence (BI) know that wasted time can easily result in missed opportunities and the loss of competitive advantage. Research laboratories and academic institutions need state-of-the-art, high-performing systems that can address the most challenging scientific problems while keeping costs under control.

The IBM *e*server® p5 575 cluster node is specifically designed to tackle these “extreme” performance computing applications, which require both high computational performance and very high memory bandwidth. The p5-575 node includes eight powerful 1.9 GHz IBM POWER5™ microprocessors. Each processor includes 1.9MB of L2 and 36MB of L3 dedicated cache for the ultimate in high performance computing (HPC). No POWER5 processor-based system can match the extraordinary density achieved by 12 p5-575 cluster nodes packaged in a single 24-inch system frame. Compared to its predecessor, the IBM @server pSeries® 655, the p5-575 delivers substantially higher performance for memory bandwidth-intensive applications.

More than number crunching

While other cluster nodes are geared solely toward fast computations, the p5-575 does much more than number crunching. It is designed to meet the needs of a broad array of organizations that require not only fast processing but also rapid and continuous access to vast amounts of data. With nearly 100 GBps of peak memory bandwidth per node, the p5-575 has a strong affinity for HPC applications such as Computer Assisted Engineering (CAE), oceanographic studies, meteorology, computational fluid dynamics, energy research, data mining and other bandwidth-intensive work that requires transferring, accessing and rapidly analyzing large quantities of data.

Like the p655, the p5-575 cluster node is designed to be an excellent match for many businesses, such as insurance, banking, finance and retail organizations that have amassed large quantities of information and want to mine that data for competitive advantage. Many of the features that make the p5-575 an excellent fit for the most demanding engineering and scientific tasks also makes it well-suited for large-scale data warehousing and data servicer applications using IBM DB2® Universal Database™ software for BI. The node can be configured so that it only need be replicated to scale-out a cluster.

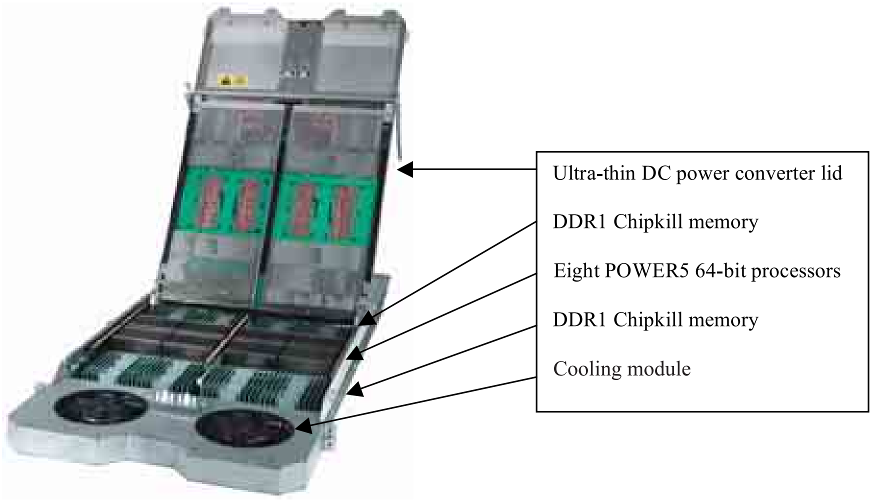
Advanced processor technology delivers outstanding performance

The p5-575 is designed to deliver exceptional performance with its 64-bit POWER5 processors. The processors incorporate simultaneous multithreading,[1] which allows two application threads to be executed concurrently. The result is improved performance compared with earlier IBM POWER™ processor–based systems.

Data-intensive performance is greatly assisted by the high-speed L2 and L3 caches available to each processor. These caches help to stage information more effectively from processor memory to applications, allowing the p5-575 cluster node to run workloads significantly faster than its predecessor.

To further enhance system performance, the p5-575 memory DIMMs have eight point-to-point connections to each of the processors, with a maximum memory capacity of 32GB per processor and a peak memory data transfer speed of up to 12.4 GBps per processor—or nearly 100 GBps for the 8-way node. The DIMMs are in close proximity to supported processor cores, designed to reduce signal propagation delay and lower power and heat dissipation requirements.

When compared with smaller symmetric multiprocessing (SMP) nodes deployed in HPC clusters, the p5-575 enables a greater proportion of the


Ultra-thin DC power converter lid
DDR1 Chipkill memory
Eight POWER5 64-bit processors
DDR1 Chipkill memory
Cooling module

workload to communicate over a lightning-fast, low-latency, high-bandwidth SMP fabric, as opposed to an I/O-based switch fabric. This system configuration can deliver significantly better overall system performance while reducing complexity, improving manageability and helping to contain costs.

The p5-575 cluster node can utilize logical partitioning (LPAR) technology implemented via IBM Virtualization Engine™ systems technologies and the operating system. The processors may run separate workloads, thereby helping lower costs. p5-575 partitions are designed to be shielded from each other to provide a high level of data security and increased application availability.

The p5-575 optionally offers Advanced POWER™ Virtualization providing Micro-Partitioning™ and Virtual I/O Server capabilities which allow businesses to increase system utilization while helping to ensure applications continue to get the resources they need. With virtualization technologies, multiple copies of operating systems can be run on the same server or processor, helping reduce the number of cluster nodes needed and minimize software licensing costs. Micro-Partitioning technology allows processors to be finely tuned to consolidate multiple independent AIX 5L™ and Linux® workloads. Micro-partitions can be defined as small as 1/10th of a processor for a maximum of 80 per p5-575 cluster node and changed in increments as small as 1/100th of a processor.

Innovative design minimizes floor space and enhances reliability

The p5-575 cluster node features innovative, elegant design and packaging. IBM has focused on providing a clean looking industrial strength overall system design that facilitates ease of service and flexibility. Mounted in a sleek 2U enclosure, the modular p5-575 allows users to deploy up to 12 nodes in a single 42U system frame. The unique node enclosure has four component modules; the I/O module, the DC power distribution module, the processor and memory module, and the cooling module. Each of these modules are custom-designed to satisfy the requirements of high-performance, high-density computing.

Nodes can be configured with or without support for optional internal and external I/O devices by selecting one of two I/O module options. The streamlined “computational” node includes two dual 10/100/1000 Mbps Ethernet ports; two integrated Ultra3 SCSI controllers; two Hardware Management Console (HMC) ports for system control, an independent service processor, logical partitioning functions and two hot-swappable disk storage bays, which accommodate 10K rpm or 15K rpm disk drives. The “I/O” node adds four 133 MHz hot-plug/blind-swap PCI-X adapter slots that allow administrators to repair, replace or install adapters with the I/O drawer in place and a RIO-2 hub port to attach an optional I/O drawer.

The highly efficient DC power distribution module is integrated into the lid of the node. This innovative power system relies on embedded circuitry rather than external wiring, providing more reliable and efficient power distribution. The hinged lid opens easily for access to the processor and memory module, which contains the POWER5 processors and system memory DIMMs. This power module includes precision intelligent monitoring and control functions that assure voltage and current delivery to the CPUs is optimized at all times, and provides alert data to the node service processor in the case of a fault.

The processor and memory module is the heart of the system. The p5-575 features eight POWER5 processors, each with 1.9MB of L2 and 36MB of dedicated L3 cache and point-to-point connections to up to eight memory DIMMs. This implementation helps to provide exceptionally high memory bandwidth to support many demanding HPC applications.

The front-end cooling module has two air-intake ventilation grids and two custom-designed blowers with high-capacity impellers and high-efficiency motors that are designed for extended life and easy serviceability. As with the power module, intelligent technology is employed in the blower system that enables blower speed to be monitored and adjusted continuously to compensate for room temperature and other system operating conditions.

Scale-up or out easily and inexpensively

The p5-575 cluster node can be scaled within the enclosure or replicated within the cluster to meet growing workload requirements. Equipped with 1GB of memory in its minimum configuration, each node can scale-up to 256GB. Two hot-swappable disk drives allow disk storage capacity from 73.4GB to 600GB. For even greater disk capacity, the I/O node configuration supports a 4U I/O drawer through a RIO-2 hub port at the back of the enclosure. Each I/O drawer holds up to 16 additional disk bays, accommodating up to an additional 1.1TB of disk storage. Two cluster nodes can share a single I/O drawer, with each system frame containing up to five I/O drawers.

The p5-575 can scale-out easily and cost-effectively as workload requirements increase. Each system frame accommodates up to 12 p5-575 cluster nodes. Organizations can add system frames to build a system cluster with anywhere from 16 to more than 1024 processors. How the p5-575 nodes are interconnected in a cluster is dependent on the cluster management software being used.[4] For Cluster Systems Management (CSM) for Linux environments, an industry standard Ethernet (10/100/1000 Mbps) interconnection may be used. For CSM for AIX 5L environments, either an industry standard Ethernet or an IBM @server pSeries High Performance Switch (HPS) can be employed.

The HPS is based on the proven technology and architecture of the IBM RS/6000® SP™ Switch2 and, of the two supported p5-575 connectivity approaches, provides significantly greater communication bandwidth and lower latency for cluster nodes or their LPARs in Cluster 1600 environments. The HPS, a 4U rack drawer for 24-inch frames, provides a unified switch network with parallel, interconnected communications channels and supports either fiber optic or copper interfaces for switch-to-switch connections. Redundant power converters and power cabling are designed to provide improved reliability, availability and serviceability (RAS).

Mainframe-inspired RAS provides peace of mind

Although the p5-575 cluster node comes in a small package, it is loaded with mainframe-inspired features that help to ensure high RAS. The p5-575 is equipped with a built-in service processor which monitors system operations continuously and can take preventive or corrective action for quick problem resolution. First Failure Data Capture (FFDC) capabilities help to identify and log problems before system failures occur. IBM error checking and correction (ECC) / Chipkill™ memory technology detects and corrects memory errors to help prevent costly system crashes. Finally, Dynamic Processor Deallocation capabilities in many cases can identify potential processor problems, generate error reports and deallocate processors before they fail.

The p5-575 node power distribution and conversion system—adopted from the @server p5 595 server design—relies on embedded circuitry rather than external wiring to distribute power among system components with the objective of providing more reliable and efficient power distribution. In the event a cooling fan fails, the second fan will increase its velocity and the system service processor may initiate a service call. Extensive monitoring and control provisions throughout the power and cooling systems assure optimal node performance at all times and enable the service processor to initiate a service call in the case of out of specification conditions or component failures.

The p5-575 system includes structural elements at the frame level to help ensure outstanding availability even in the event of facility power problems. The p5-575 system frame uses IBM's leading-edge rack level distributed power conversion architecture to increase system density, simplify power connection and provide a robust, redundant system power supply arrangement. Two simple, neutral free universal line cords connect the p5-575 system frame to a client’s facility anywhere in the world with no adjustments being required to personalize for power utility voltage or frequency. Support for 200v to 240v, 380v to 415v, and 480v three phase power inputs should allow clients to enjoy reduced facility equipment cost and help improve energy efficiency. The ability of the p5-575 to tolerate power disturbances is exceptional in comparison to most other computing equipment, and optional battery backup can help the system ride through a momentary power interruption without the need for large and expensive Universal Power Supply (UPS) systems.

Dual redundant rack controllers and Ethernet hubs are included in each p5-575 system frame to provide hardware monitoring and control connectivity to each of the drawers through an independent dual Ethernet service network connected to the HMC. This high availability arrangement centralizes the client system interface for all nodes, I/O expansion drawers and HPSs onto a single console located outside of the frame in a quieter, more comfortable environment for the user.

Built-in reliability features

IBM autonomic computing enhancements are built into the p5-575 cluster node. Self-protecting helps the p5-575 determine the cause of an error as it happens and may reduce lengthy service times attempting to recreate errors after the fact. Errors may be self-correcting or resources varied off-line while the system remains available for use. IBM’s FFDC provides error information in real-time and makes it possible to determine the parts needed to fix the problem. The service processor has the capability to determine which part or component needs repair and initiate a service call to identify parts needed for maintenance at a time acceptable to the client.

Self-healing capabilities help the p5-575 system to overcome error conditions and continue operating if a failure is detected. This is implemented through Error Checking and Correcting Code (ECC) L2 and L3 caches and main memory and through bit-scattering, bit-steering and memory scrubbing soft-error recovery procedures in main memory. Bit-scattering scatters bits across four different memory words, enables recovery of single-bit errors and should keep a p5-575 running when a failure is detected by Chipkill™ memory. Bit-steering dynamically routes a bit to a spare memory chip in the event the memory failure rate for the bit exceeds a given threshold. If all bits should become used up on the spare chip, the service processor is invoked to request deferred maintenance at a time acceptable to the client. Memory scrubbing for soft single-bit errors is performed in the background so as to correct errors while memory is idle. This helps to prevent multiple-bit errors.

Supporting business-critical applications

The p5-575 cluster node can run AIX 5L and Linux operating systems (OSs) on the same node simultaneously, to provide the flexibility to support a full range of applications including business-critical applications.

AIX 5L is an industrial-strength UNIX® environment tuned for application performance and loaded with exceptional RAS features. The AIX 5L OS delivers enhancements to Java™ technology, Web performance and scalability for managing clusters of all sizes. Web-based remote management tools give administrators centralized control of the system, enabling them to monitor key resources, including adapter and network availability, file system status and processor workload.

The AIX 5L OS also incorporates Workload Manager, a resource management tool that specifies the relative importance of workloads to balance the demands of competing workloads and enhance system resources. Workload Manager can help ensure that applications remain responsive even during periods of peak system demand.

By supporting the Linux OS, the p5-575 cluster node offers important cost-saving opportunities. Because Linux is an open source technology, it can be less expensive to license than many proprietary operating systems. Nevertheless, Linux does not compromise on functionality. With a growing list of Linux applications available, it provides the freedom to use the right applications for organizations' needs. The Linux OS is available from selected Linux distributors in packages that include a range of open source tools and applications. IBM is firmly committed to Linux and offers expert service and support.

Software tools facilitate cluster management

The Cluster 1600, a highly scalable cluster solution for UNIX or Linux environments, consists of a cluster of @server p5 and pSeries nodes including up to 128 p5-575 cluster nodes. Supported are AIX 5L V5.2, AIX 5L V5.3, SUSE Linux Enterprise Server 9 (SLES 9) and Red Hat Enterprise Linux AS 3 (RHEL AS 3) operating systems. Cluster 1600 is implemented through CSM for AIX 5L or Linux clusters. CSM supports other optional cluster software for HPC including:

- *Parallel Environment (PE) for AIX 5L–a high function development and execution environment for parallel message-passing applications.*
- *LoadLeveler®–dynamic job scheduling and workload balancing software supporting thousands of jobs within the cluster. LoadLeveler is supported on AIX 5L V5.2, V5.3 and SLES 9.*
- *GPFS–a high-performance, shared disk file system providing fast data access to all nodes in a cluster. GPFS is supported on AIX 5L V5.2, V5.3 and SLES 9.*
- *ESSL and Parallel ESSL–mathematical libraries for both AIX 5L and Linux to enhance performance of serial, parallel and scientific applications. Parallel ESSL is supported on AIX 5L V5.2, V5.3 and SLES 9, while ESSL is supported on AIX 5L V5.2, V5.3, SLES 9 and RHEL AS 3.*
- *High Availability Cluster Multiprocessing (HACMP™) for AIX 5L–helps provide continuous access to data and applications through database or application failover to a secondary server if the database or application server fails.*

Major productivity enhancements are provided through the POWER Hypervisor™ in conjunction with available operating systems. The user can establish dynamic LPARs running AIX 5L V5.2, AIX 5L V5.3 or SLES 9 operating systems. Dynamic LPAR enables system administrators to reallocate system resources without rebooting the system or the partition.

If AIX 5L V5.3, SLES 9 or RHEL AS 3 are selected for a partition, the user can take advantage of the benefits of hardware simultaneous multi-threading,[1] which may provide an increase of up to 30% (based on rPerf projections[2]) in processor throughput over single-threaded operation, depending on the nature of the applications being run in the partition. Furthermore, users can obtain even more flexibility with the Advanced POWER Virtualization option, which provides Micro-Partitioning, shared processor pool and Virtual I/O Server capabilities.

Micro-Partitioning technology provides the capability to establish up to 80 micro-partitions on a single p5-575 cluster node, effectively splitting each processor's power among up to 10 micro-partitions. Shared processor pool provides a pool of processing power that is shared among partitions assigned to the pool to non-disruptively improve utilization and throughput. Virtual I/O Server enables the physical sharing of disk drives and communications and Fibre Channel adapters and helps reduce the number of expensive devices and improve system administration and utilization. The POWER Hypervisor also enables Virtual LAN for high-speed, secure partition-to-partition communication to help improve performance.

An additional capability of Advanced POWER Virtualization supported by AIX 5L is Partition Load Manager which provides policy-based, automatic partition resource tuning that can adjust CPU and memory allocations between AIX 5L V5.2 and V5.3 partitions.

@server p5 575: Building an infrastructure for the future

The IBM @server p5 575 cluster node is designed to be a high-performance building block for supercomputing. With high sustained throughput and I/O bandwidth, it is an excellent match for scientific and engineering HPC applications as well as BI functions for which organizations need to transfer, access and analyze large amounts of data rapidly. The innovative, highly dense packaging of the p5-575 cluster node, its optional Virtualization Engine with Micro-Partitioning capabilities and its ability to run the AIX 5L and Linux operating systems concurrently, will help get more work done while using less physical floor space. The comprehensive set of cluster management tools designed for the AIX 5L and Linux operating systems help provide the means to assemble and effectively manage a large cluster. And with easy scalability, the p5-575 will be ready to grow with an organization's high-performance requirements.

p5-575 at a glance

Standard configuration	
Microprocessors	Eight 64-bit 1.9 GHz POWER5 processors
Level 2 (L2) cache	15.2MB / cluster node
Level 3 (L3) cache	288MB / cluster node
RAM (memory)	1GB to 256GB of 266 MHz DDR1 SDRAM
Internal disk storage	1.7TB (with optional I/O drawer)
Processor-to-memory bandwidth (peak)	99.7 GBps
L2 to L3 cache bandwidth	243.2 GBps
RIO-2 I/O subsystem bandwidth (peak)	4 GBps
Internal SCSI disk bays	Two standard (73.4/146.8/300GB 10K rpm or 36.4/73.4GB 15K rpm disks)
Standard features	
I/O ports	Two Ultra3 SCSI controllers I/O node configuration: • Two dual 10/100/1000 Mbps Ethernet ports • Two HMC ports • RIO-2 hub port for optional I/O drawer • Four 3.3v PCI-X adapter slots (64-bit/133 MHz) Compute node configuration: • Two dual 10/100/1000 Mbps Ethernet ports • Two HMC ports
I/O expansion	One optional I/O drawer (can be shared by two cluster nodes) providing 20 3.3v 64-bit PCI-X slots and up to 16 disk bays (36.4/73.4GB 15K rpm disks)
POWER Hypervisor	LPAR Dynamic LPAR[3] Virtual LAN[1]
Advanced POWER Virtualization[1] (option)	Micro-Partitioning Shared processor pool Virtual I/O Server Partition Load Manager (AIX 5L only)
Battery backup (option)	Up to six—redundant or non-redundant

p5-575 at a glance

RAS features	Copper and silicon-on-insulator (SOI) microprocessors Selective dynamic firmware updates (planned for 2Q 2005) IBM Chipkill ECC, bit-steering memory ECC L2 cache, L3 cache Service processor Hot-swappable disk bays Hot-plug/blind-swap PCI-X slots (base system and I/O drawer) Redundant/hot-plug bulk power supply Redundant/hot-plug rack power controllers / Ethernet service network hubs Redundant bulk power supply Dynamic Processor Deallocation Dynamic deallocation of logical partitions and PCI-X bus slots Extended error handling for PCI-X slots Battery backup and redundant battery backup (optional)
Operating systems	AIX 5L Versions 5.2/5.3 SUSE LINUX Enterprise Server 9 for POWER (SLES 9) or later Red Hat Enterprise Linux AS 3 for POWER Update 4 (RHEL AS 3) or later
Power requirements	200v to 240v; 380v to 415v; 480v AC
System frame dimensions	79.7"H x 30.9"W x 60.2"D (202.5cm x 78.5cm x 153.0cm); weight: 3,095 lb (1,406 kg)
Warranty	On site, 8 A.M.-5 P.M. next-business-day for one year

* With slim-line doors and populated with 12 p5-575, internal battery backup and one I/O drawer. Weight will vary when disks, adapters and other peripherals are installed.

For more information

To learn more about the IBM @server p5 575 cluster node, contact your IBM marketing representative or IBM Business Partner, or visit the following Web sites:

- **ibm.com**/eserver/pseries
- **ibm.com**/servers/aix
- **ibm.com**/linux/power
- **ibm.com**/common/ssi

Many of the features described in this document are operating system dependent and may not be available on Linux. For more information, please check: **ibm.com**/servers/eserver/pseries/linux/whitepapers/linux_pseries.html.

[1] Not supported on AIX 5L V5.2

[2] rPerf (Relative Performance) is an estimate of commercial processing performance relative to other pSeries systems. It is derived from an IBM analytical model which uses characteristics from IBM internal workloads, TPC and SPEC benchmarks. The rPerf model is not intended to represent any specific public benchmark results and should not be reasonably used in that way. The model simulates some of the system operations such as CPU, cache and memory. However, the model does not simulate disk or network I/O operations.

rPerf estimates are calculated based on systems with the latest levels of AIX 5L and other pertinent software at the time of system announcement. Actual performance will vary based on application and configuration specifics. The IBM @server pSeries 640 is the baseline reference system and has a value of 1.0. Although rPerf may be used to approximate relative IBM UNIX commercial processing performance, actual system performance may vary and is dependent upon many factors including system hardware configuration and software design and configuration.

All performance estimates are provided “AS IS” and no warranties or guarantees are expressed or implied by IBM. Buyers should consult other sources of information, including system benchmarks, and application sizing guides to evaluate the performance of a system they are considering buying. For additional information about rPerf, contact your local IBM office or IBM authorized reseller.

[3] Available with AIX 5L and SLES 9 operating systems

[4] For a complete discussion of clustered systems support, see the Cluster Systems Handbook (SG24-6965) at http://www.redbooks.ibm.com/redbooks/pdfs/sg246965.pdf

IBM Systems and Technology Group
Route 100
Somers, NY 10589

Produced in the United States
April 2005
All Rights Reserved

This publication was developed for products and/or services offered in the United States. IBM may not offer the products, features, or services discussed in this publication in other countries.

The information may be subject to change without notice. Consult your local IBM business contact for information on the products, features and services available in your area.

All statements regarding IBM future directions and intent are subject to change or withdrawal without notice and represent goals and objectives only.

IBM, the IBM logo, the e-business logo, AIX 5L, Chipkill, DB2, DB2 Universal Database, @server, HACMP, Hypervisor, LoadLeveler, Micro-Partitioning, POWER, POWER5, pSeries, RS/6000, SP and Virtualization Engine are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries or both. A full list of U.S. trademarks owned by IBM may be found at: **ibm.com**/legal/copytrade.shmtl.

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

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

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

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

IBM hardware products are manufactured from new parts, or new and used parts. Regardless, our warranty terms apply.

This equipment is subject to FCC rules. It will comply with the appropriate FCC rules before final delivery to the buyer.

Information concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of the non-IBM products should be addressed with the suppliers.

When referring to storage capacity, 1TB equals total GB divided by 1000; accessible capacity may be less.