

IBM System p5 575 cluster node



System p5 575 cluster node

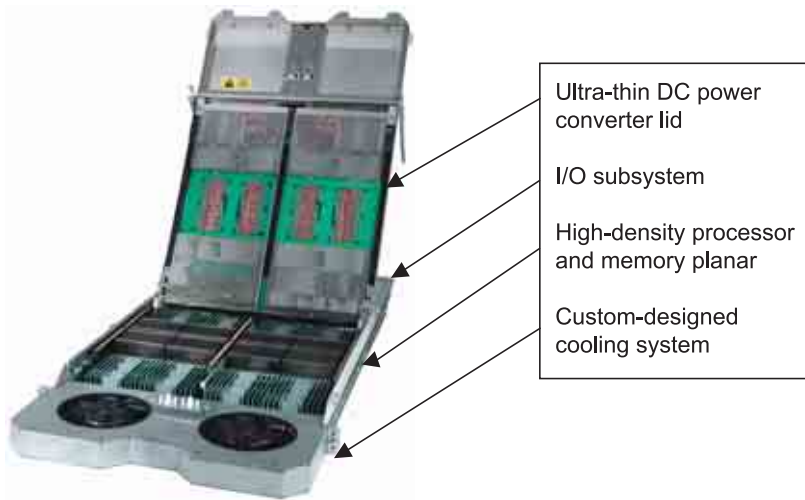
Highlights

- **8- and 16-core off-the-shelf cluster nodes**
- **Powerful IBM POWER5+™ processors with extraordinary memory bandwidth**
- **Ultra dense, elegant packaging with innovative cooling features**

The IBM System p5™ 575 cluster node is designed to excel at high performance computing (HPC) applications for organizations involved in engineering problem solving, drug design, oil reservoir modeling, weather forecasting, financial simulation and business intelligence (BI). Use it in clustered configurations of as few as 16 CPUs, or in world class supercomputer configurations of more than 2,000 processors.

The System p5 575 features a choice of two powerful nodes. An 8-core node includes dual-core 2.2 GHz POWER5+ microprocessors with only a single core active. Each processor has access to 1.9MB of L2 and 36MB of L3 dedicated cache for HPC and BI applications. Also available is a 16-core node built with dual-core 1.9 GHz chips that have both cores active. In this case, two processors share access to the same L2 and L3 cache. Although per processor cache and memory bandwidth are less, the 16-core node achieves up to 60% more floating-point performance¹ for HPC applications than the 8-core node.

The p5-575 offers ultra-dense packaging; no other IBM POWER5™ processor-based system can match the extraordinary density achieved with nearly 200 CPUs in a single footprint (12 p5-575 16-core cluster nodes packaged in a single 24-inch system



It achieves more than one teraFLOPS of performance in a single frame footprint with excellent price/performance. Retaining the same memory capacity and bandwidth as the 8-core node effectively is designed to reduce the per processor memory and bandwidth, thus creating a strong 16-core affinity to floating-point intensive engineering and scientific work such as found in Computer Assisted Engineering (CAE) rather than more memory bandwidth-intensive applications.

frame). Compared to its POWER4™ predecessor, the System p5 575 delivers substantially higher packaging density; compared to its POWER5 predecessor, it delivers substantially higher sustained performance for HPC applications.

More than number crunching

The introduction of enhanced 8- and 16-core p5-575 cluster nodes marks another major step in the evolution of high-powered, off-the-shelf building blocks that are tailored to meet the demands of a broad range of compute-intensive or memory bandwidth-intensive applications. While the 8- and 16-core nodes share the same innovative and elegant packaging, memory sizing and I/O options, each is tailored to the demands of a particular set of application requirements.

The 8-core node is designed to meet the needs of organizations that require not only fast processing but also rapid and continuous access to vast amounts of data. With over 25 GBps of peak memory bandwidth per processor, the 8-core p5-575 has a strong affinity for HPC applications such as 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. It is also applicable to businesses such as insurance, banking, finance and retail organizations using IBM DB2® Universal Database™ software for BI.

With twice the number of processors in a single node, the 16-core p5-575 significantly increases floating-point performance compared to an 8-core node.

Advanced processor technology delivers outstanding performance

The p5-575 cluster node is designed to deliver exceptional performance with its 64-bit POWER5+ processors. The processors incorporate simultaneous multithreading,² which allows two application threads to be executed concurrently. The result is improved performance compared with earlier POWER™ processor-based systems. Other POWER5+ processor enhancements such as smaller die size and larger page size combined with over two times the peak memory bandwidth enables the p5-575 to achieve performance uplifts in some standard benchmarks that substantially exceed the improvement attributable to faster processor speeds alone.

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

To further enhance system performance, newly available DDR2 memory DIMMs have eight point-to-point connections to each of the processor chips, with a maximum memory capacity of 256GB per node and a peak memory data transfer speed of over 200 GBps, more than twice that of the earlier POWER5 nodes with DDR1 memory. The DIMMs are in close proximity to supported processor cores, so as to reduce signal propagation delay and lower power and heat dissipation requirements.

When compared with smaller symmetric multiprocessing (SMP) nodes deployed in HPC clusters, both the 8- and 16-core p5-575 nodes enable a greater proportion of the workload to communicate over an extremely fast, low-latency, high-bandwidth SMP fabric, as opposed to an I/O-based switch fabric. This system configuration is designed to deliver significantly better

overall system performance while reducing complexity, improving manageability and helping to contain costs.

Both p5-575 cluster nodes 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.

p5-575 nodes optionally offer Advanced POWER Virtualization providing Micro-Partitioning™ technology and Virtual I/O Server (VIOS) 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 reduce software licensing costs. Micro-Partitioning technology allows processors to be finely tuned to consolidate multiple independent AIX 5L™ and Linux® workloads.

Innovative design minimizes floor space and enhances reliability

The p5-575 cluster node features innovative, elegant design and packaging that facilitates ease of service and flexibility. Mounted in a sleek 2U enclosure, the modular p5-575 allows users to deploy 12 nodes in a 42U system frame. 8- and 16-core p5-575 cluster nodes can be intermixed in the same frame. The unique node enclosure has four component modules; the I/O subsystem, the DC power converter/lid, the processor and memory planar, and the cooling system. Each of these components are custom-designed to satisfy the demanding requirements of very high-performance, high-density computing.

Nodes can be configured with or without support for internal and external I/O devices. The standard “compute” node configuration 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. An “I/O” node option adds four 133 MHz hot-plug/blind-swap

PCI-X adapter slots that allow administrators to repair, replace or install adapters with the node in place and two RIO-2 hub ports 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 are designed to help assure power delivery 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 containing the eight or 16 POWER5+ processors, deployed in eight modules. Each module also has the L2 and L3 caches 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.

New energy-efficient cooling technology

IBM is announcing the availability of new cooling technology for both new and installed p5-575 systems that can help reduce installation cooling requirements and energy costs. Previously announced for installation on 19-inch racks, IBM “Cool Blue”[®] is now available for the p5-575 frame—a major technology breakthrough in the operation of data centers.

Three years in development, Cool Blue is the name for the IBM Rear Door Heat eXchanger, a revolutionary cooling system that uses the existing chilled water supply for air conditioning systems already located in the majority of data centers designed to reduce server heat emissions by up to 55 percent while lowering energy costs by up to 15 percent.

Inside the door of Cool Blue, sealed tubes filled with chilled water absorb the heat generated in a fully populated rack and carry it away so it is not released into the datacenter.

The improved cooling from the Heat eXchanger gives the p5-575 a significant benefit in Kilowatts per MFLOPS energy efficiency—an increasingly important issue in containing the environmental and facilities costs of large HPC installations.

Scale-up or out easily and inexpensively

p5-575 cluster nodes can be scaled within the system frame 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 option supports a 4U I/O drawer through the RIO-2 hub ports at the back of the enclosure. The I/O drawer holds up to 16 additional disk bays, accommodating up to an additional 2.3TB of 15K rpm disk storage. Two cluster nodes can share a single I/O drawer, with each system frame containing up to five I/O drawers.

A p5-575 cluster can scale-out easily and cost-effectively as workload requirements increase. Each system frame accommodates up to 12 p5-575 cluster nodes in any combination. Organizations can add system frames to build a system cluster with anywhere from two to 128 nodes (16 to 2048 processors). Networked p5-575 nodes are controlled by HMCs (up to 32 nodes per HMC). How p5-575 nodes are networked in a cluster is dependent on the cluster management software being used.⁴

Cluster Systems Management (CSM) for Linux environments supports Ethernet (10/100/1000 Mbps) or 4x InfiniBand interconnections. For CSM for AIX 5L environments, an Ethernet, 4x InfiniBand or IBM **@server**® pSeries High Performance Switch (HPS) interconnection can be employed in support of HPC workloads.

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 copper and optical 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 features provide 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 is designed to monitor 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 IBM **@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 help 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 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 the 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 software recovery procedures in main memory. Bit-scattering scatters bits across four different memory words, enables recovery of single-bit errors and should keep the 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. AIX 5L also incorporates Workload Manager, a resource management tool that can help ensure applications remain responsive even during periods of peak system demand.

The p5-575 cluster node supports the Linux OS allowing a choice of operating systems to best fit client needs. Because Linux is an open source

technology, it has the worldwide Linux community enhancing, contributing and validating the Linux kernel and can be less expensive to license than a proprietary OS. In choosing Linux, users can take advantage of many of the functionality, reliability and scalability features designed into the p5-575. And with a large list of open source, IBM and third party applications available, Linux offers the freedom to use the right applications for an organization's needs. The Linux OS is orderable from IBM and selected Linux distributors in packages that include a range of open source tools and applications.

Clients needing highly available and powerful storage to support their System p5 servers can realize benefits derived from using IBM System Storage™ and TotalStorage® solutions. IBM conducts comprehensive testing in System Storage laboratories under stress environments, including clustered configurations, to help ensure combined server and storage systems solutions have high reliability, interoperability and streamlined, efficient implementation. With an IBM TotalStorage, System Storage and System p5 575 server solution, you can be assured your IT environment will meet today's and tomorrow's demanding needs.

Software tools facilitate cluster management

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

- *Parallel Environment (PE)—a high function development and execution environment for parallel message-passing applications under AIX 5L.*
- *LoadLeveler®—dynamic job scheduling and workload balancing software supporting thousands of jobs within the cluster. LoadLeveler is supported on AIX 5L, SLES and Red Hat AS.*
- *GPFS—a high-performance, shared disk file system providing fast data access to all nodes in a cluster. GPFS is supported on AIX 5L, SLES and Red Hat AS.*

- *ESSL and Parallel ESSL—mathematical libraries for both AIX 5L and Linux to enhance performance of serial, parallel and scientific applications. ESSL and Parallel ESSL are supported on AIX 5L, SLES and RHEL AS.*
- *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.*

4x InfiniBand PCI adapters are supported for Linux HPC workloads running TopSpin MPI (MVAPICH), as well as supporting IPoIB for commercial workloads.

Major productivity enhancements are provided through the POWER Hypervisor firmware in conjunction with available operating systems. The user

can establish dynamic LPARs running AIX 5L or SLES operating systems. Dynamic LPAR enables system administrators to reallocate system resources without rebooting the system or the partition.

If AIX 5L V5.3, SLES or RHEL AS are selected for a partition, the user can take advantage of the benefits of hardware simultaneous multithreading,² which may provide an increase of up to 30% (based on rPerf projections³) 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 VIOS capabilities.

Micro-Partitioning technology provides the capability to establish up to 160 micro-partitions on a 16-core p5-575 cluster node (or up to 80 on an 8-core 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. VIOS 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 partitions.

System p5 575: 8- and 16-core building blocks for supercomputing

Choose the 8-core System p5 575 node when extremely high memory bandwidth is essential to the rapid processing of large amounts of data. Choose the 16-core node when raw computational power in highly dense packaging is required. A comprehensive set of cluster management tools designed for AIX 5L and Linux help assemble and manage large clusters. Easy scalability and optional Advanced POWER Virtualization enables System p5 575 cluster flexibility as an organization's high-performance requirements change.

p5-575 at a glance

Standard configuration

Microprocessors	Eight 64-bit 2.2 GHz POWER5+ processors or 16 64-bit 1.9 GHz POWER5+ processors
Level 2 (L2) cache	15.2MB
Level 3 (L3) cache	288MB
Memory	1GB of 533 MHz DDR2 SDRAM
Processor-to-memory bandwidth (peak)	204.0 GBps
L2 to L3 cache bandwidth (peak)	243.2 GBps
RIO-2 I/O subsystem bandwidth (peak)	4.0 GBps
Internal SCSI disk bays	Two standard (73.4/146.8/300GB 10K rpm or 36.4/73.4/146.8GB 15K rpm disks)

Standard features

I/O ports	Two Ultra3 SCSI controllers
"Compute" node configuration	Two dual 10/100/1000 Mbps Ethernet ports
	Two HMC ports

Expansion features

Memory	Up to 256GB of 533 MHz DDR2 SDRAM
"I/O" node configuration	Two RIO-2 hub ports (for optional I/O drawer) Four PCI-X adapter slots (64-bit/133 MHz)
I/O expansion (optional)	One I/O drawer (can be shared by two cluster nodes) providing 20 64-bit PCI-X slots and up to 16 disk bays (36.4/73.4/146.8GB 15K rpm disks)
Internal disk storage	2.9TB (with I/O drawer)

Virtualization Engine system technologies

POWER Hypervisor	Dynamic LPAR; Virtual LAN ²
Advanced POWER Virtualization ² (optional)	Micro-Partitioning; Shared processor pool; VIOS; Partition Load Manager (AIX 5L only)

Frame features

Battery backup (optional)	Up to six—redundant or non-redundant-in system frame
High Performance Switch	Up to 128 nodes supported (higher by special order); One HMC can control 32 nodes

Operating systems

AIX 5L Version 5.2 or later
SLES 9 or later
RHEL AS 4 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)³

Warranty

On site, 8 A.M.-5 P.M. next-business-day for one year

For more information

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

- ibm.com/systems/p/
- 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.

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.

¹ Based on SPECfp_rate_base at www.spec.org; submitted to SPEC on February 14, 2006

² Not supported on AIX 5L V5.2

³ 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 system commercial processing performance, actual system performance may vary and is dependent upon many factors including system hardware configuration and software design and configuration. For additional information about rPerf, contact your local IBM office or IBM authorized reseller.

⁴ For details of clustered systems support, see the Cluster 1600 Facts and Features at ibm.com/servers/eserver/clusters/hardware/factsfeatures.html

⁵ 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.

⁶ For details, refer to: ibm.com/news/ie/en/2005/07/ie_en_news_20050715.html



© Copyright IBM Corporation 2006

IBM Systems and Technology Group
Route 100
Somers, NY 10589

Produced in the United States
February 2006
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, LoadLeveler, Micro-Partitioning, POWER, POWER4, POWER5, POWER5+, pSeries, RS/6000, SP, System p5, System Storage, TotalStorage 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.shtml.

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, total TB equals total GB divided by 1000; accessible capacity may be less.