# Contents

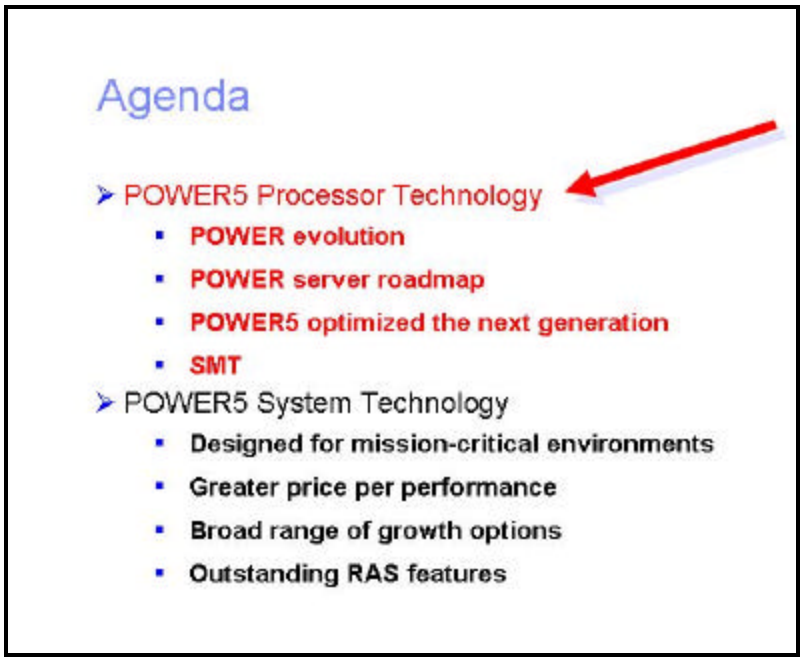
## About the Author

Robert Arenburg, PhD.
Life Sciences Solutions Development

**Robert Arenburg** is a Senior Technical Consultant in the Solutions Enablement organization in the IBM Systems and Technology Group, located in Austin, Texas. He has a Ph.D. in Engineering Mechanics from Virginia Technical Institute. He has worked at IBM for 13 years. His areas of expertise include high performance computing and other performance issues, capacity planning, 3D graphics, solid mechanics, and computational and finite element methods.


## Introduction

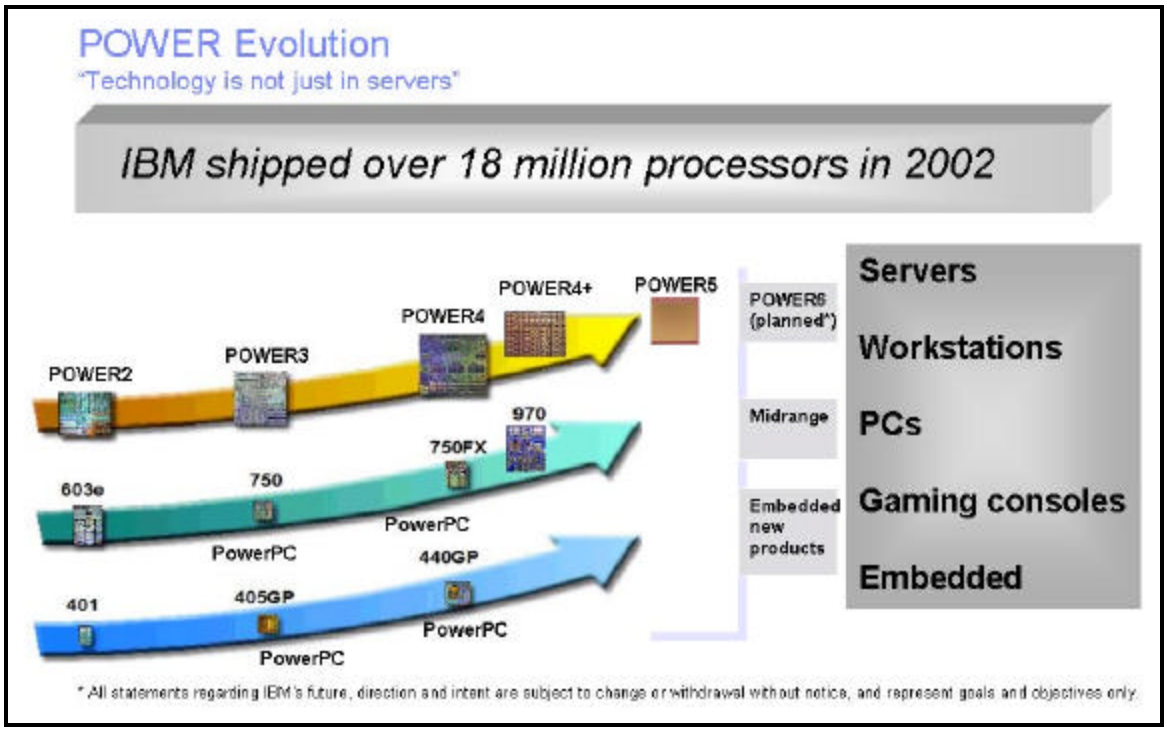
POWER5™ represents ninth generation of 64-bit processor technology for the long line and broadly applicable POWER™ family of processors. This course is designed to give the reader a quick overview of its innovative design and implementation. There are no prerequisites for this course; however, it is assumed that the student is already somewhat familiar with the IBM® eServer™ iSeries™ and/or IBM eServer pSeries™ heritage of servers.




## Agenda

We will start this course by talking about the design of the POWER5 technology at the chip level. This will include a precursor discussion of the history of the POWER family of chips. Then, we will compare the recent ancestor, the POWER4™ processor.

Then, we will review the various IBM eServer i5 and eServer p5 server models that are driven by this latest processor.




## POWER Evolution

The POWER family of chips has evolved greatly since its introduction in 1990. Early on, the POWER chip made the leap to a 64-bit architecture from 32-bit... years ahead of any other processor platform. Superscalar features, fixed and floating point units and more have been incorporated into the POWER chip. In 1997, POWER embraced copper conductors, replacing the less efficient aluminum that had been the foundational conductor technology for the chip's circuitry. Very soon afterward, in 1998, silicon-on-insulator (SOI) technologies enhanced the processing power of POWER even more.

As you can see from this chart, the POWER line of processors has continuously delivered value for a huge range of microprocessor environments.

**For embedded products and gaming solutions**, the PowerPC® chip has evolved from the 32-bit PowerPC 401 to the 64-bit PowerPC 440GP — a system-on-a-chip design that also includes a mix of rich peripheral cores. This chip is also popular for network communications applications.

**For midrange applications (PCs and workstations)**, the 32-bit POWER2 processor was first used in the PowerPC 603® servers, and has evolved into the 64-bit PowerPC 970 processor. The PowerPC 970 has been derived from the POWER4 processor and has been uniquely enhanced. For example, the PowerPC 970 has incorporated the VMX SIMD (Single Instruction Multiple Data) unit at the design request of Apple Computers, which has embedded it into the Mac G5.

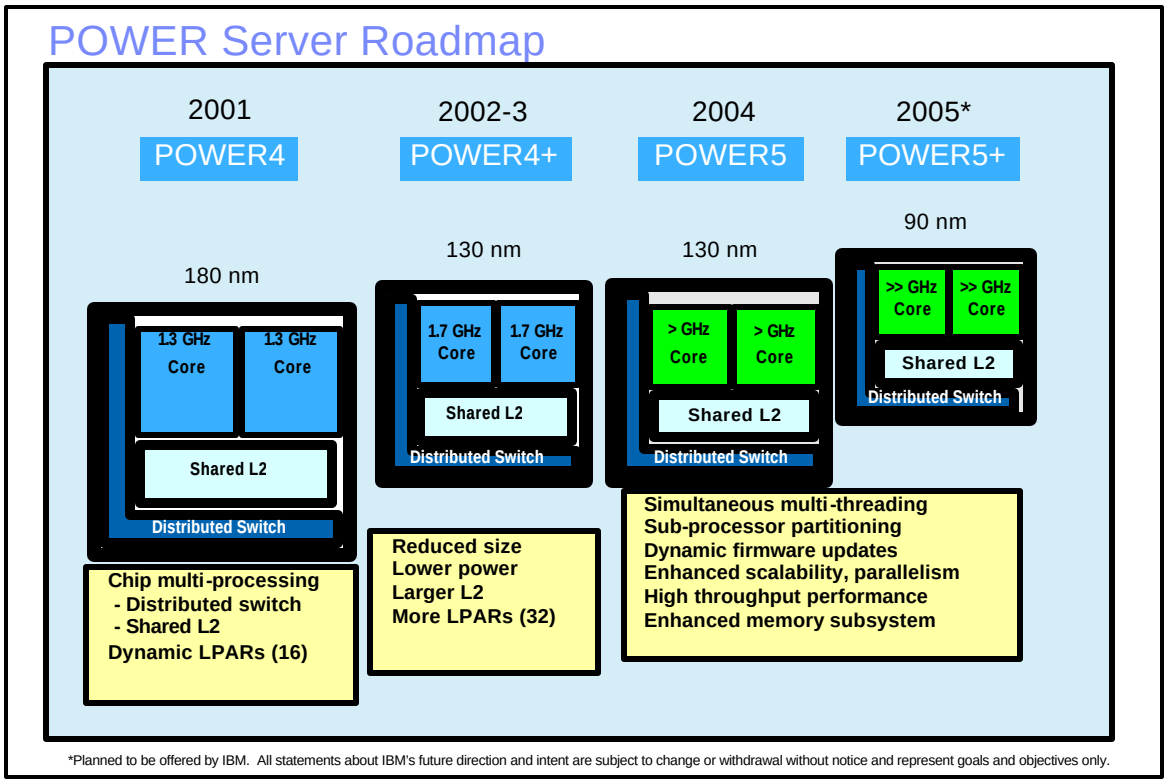
*[NOTE: The VMX SIMD unit is a short-vector processing unit that allows the G5 to support backward compatibility with the G4 processors which were supplied by Motorola. It lets one microinstruction operate at the same time on multiple data items. This is especially productive for applications in which visual images or audio files are processed*

*— allowing what usually requires a repeated succession of instructions (a loop) to be performed in one instruction.]*

The PowerPC 970 chip also drives the IBM BladeCenter™ JS20 — where it is essentially a single-core PowerPC 970. A dual-core refers to a chip that has two processors. This packaging density enhances performance, reduces power consumption, and supports simultaneous multiprocessing. (The POWER4 processor was the first chip in the industry to offer dual-core processors. The POWER5 processor is also dual-core.)

**For high-end server needs**, the POWER2, introduced in the early 1990s, was the first POWER chip to deliver a much-welcomed architecture for traditional tower-based, commercial application servers. This early POWER chip has evolved to the POWER3™, POWER4, POWER4+™, and now the POWER5 chip. With announcements made in 2004, the POWER5 chip now powers the IBM eServer i5 family of iSeries servers and the IBM eServer p5 family of pSeries servers. With projected availability in 2005, the POWER5+™ processor will continue this evolution.

The POWER family puts IBM into the category of being a significant chip supplier with processors that support the entire microprocessor product market — from embedded products to gaming to high-end servers.



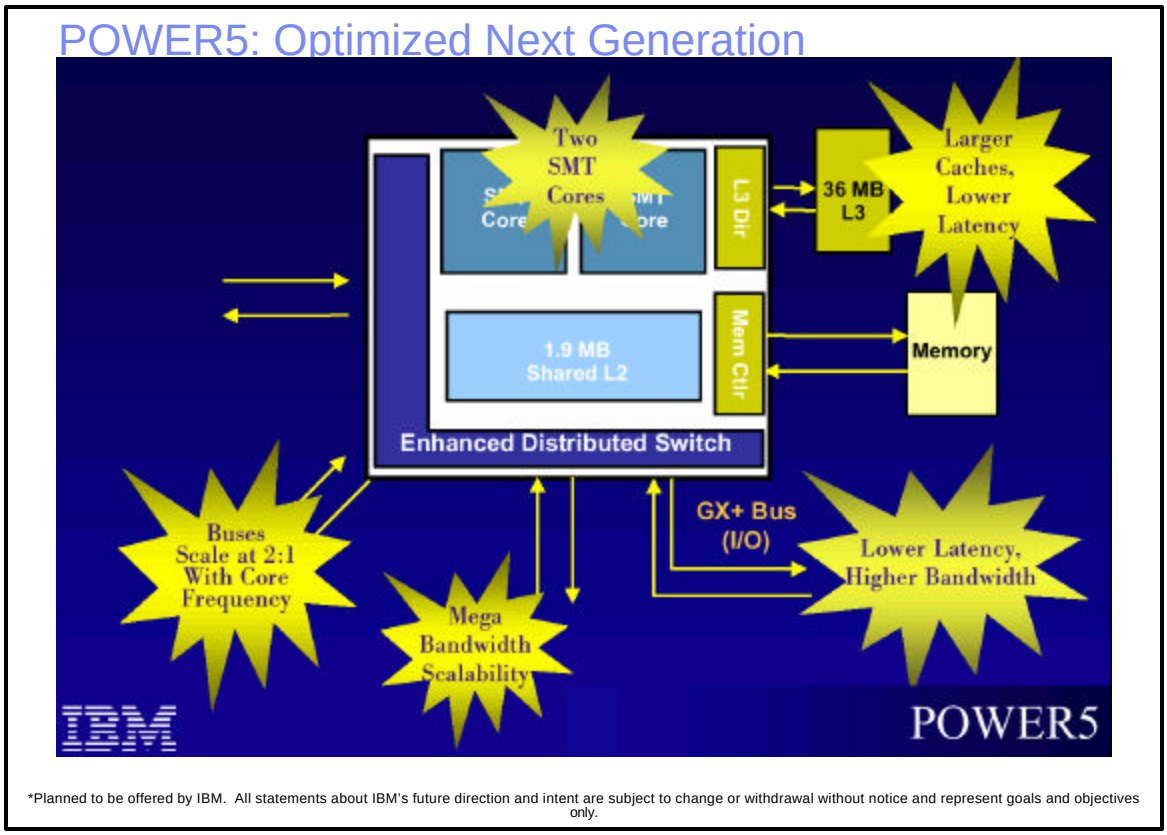

# POWER Server Roadmap

As you can see here, the POWER4 processor was introduced in 2001 with an internal clock speed of 180 nanometers. As we just briefly mentioned, it was the first POWER chip to have dual-core processors (chip multi-processing), each of which implements the PowerPC AS architecture. It also has shared L2 cache (1.41 megabytes) between the two cores and was the first POWER chip to offer dynamic logical partitioning (LPARs) (maximum of 16 partitions). You may also notice that it offered a distributed switching component on the same substrate.

In 2002 and 2003, the POWER4+ processor was delivered — offering reduced chip size, which improved clock speed (130 nanometers) also reduced power consumption. The POWER4+ chip enjoyed a larger shared-L2 cache (1.5 megabytes) and supported 32 dynamically assigned logical partitions.

Now, POWER5 delivers the same 130 nanometer clock speed as the POWER4 chip and shares much of the same technology. However, POWER5 also brings support for Simultaneous Multi-Threading (SMT), sub-processor partitioning, still larger shared L2 cache (1.9 megabytes), and dynamic firmware updates, as well as enhanced scalability and parallelism. With POWER5, the memory subsystem has been enhanced. Actually, the memory controller has been integrated onto the POWER5 processor itself. This helps a great deal with memory bandwidth.

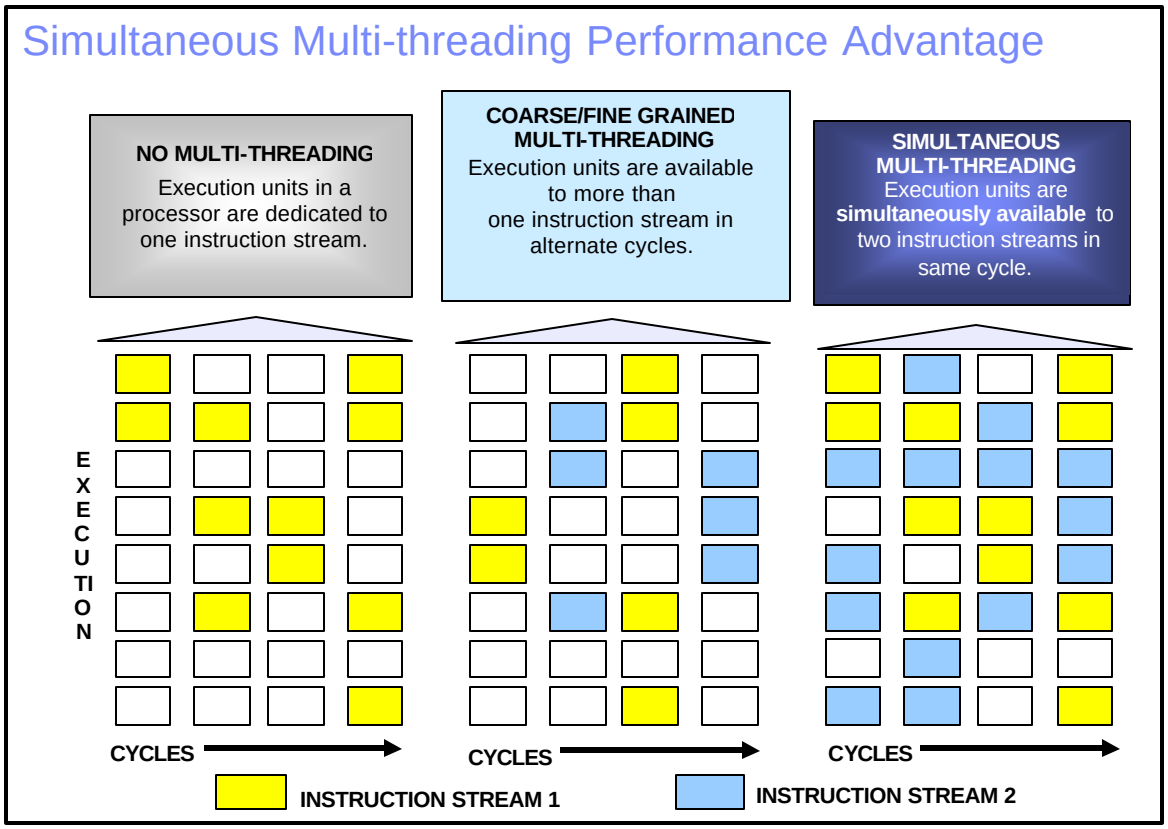
In 2005, IBM is expected to deliver the POWER5+ processor, which will bring yet another bump in the speed of the processor to 90 nanometers.




## POWER5: Optimizes Next Generation

If we focus on the POWER5 chip design further, you can see that, in addition to the two SMT cores, the caches are larger — reducing latency. The buses scale at a two-to-one ratio with core frequency and are eight bytes wide. Bandwidth scalability has been significantly increased. And, the inter-module GX+ I/O bus also yields lower latency and higher network bandwidth.

You will see this chip in much greater detail shortly.




## SMT Performance Advantage

We have already mentioned simultaneous multi-threading. Let's look more closely at the notion of SMT — and why it is so valuable.

The fundamental idea is that no one process can fully utilize the processor, so it is more efficient to initiate several tasks being submitted to the processor at the same time.

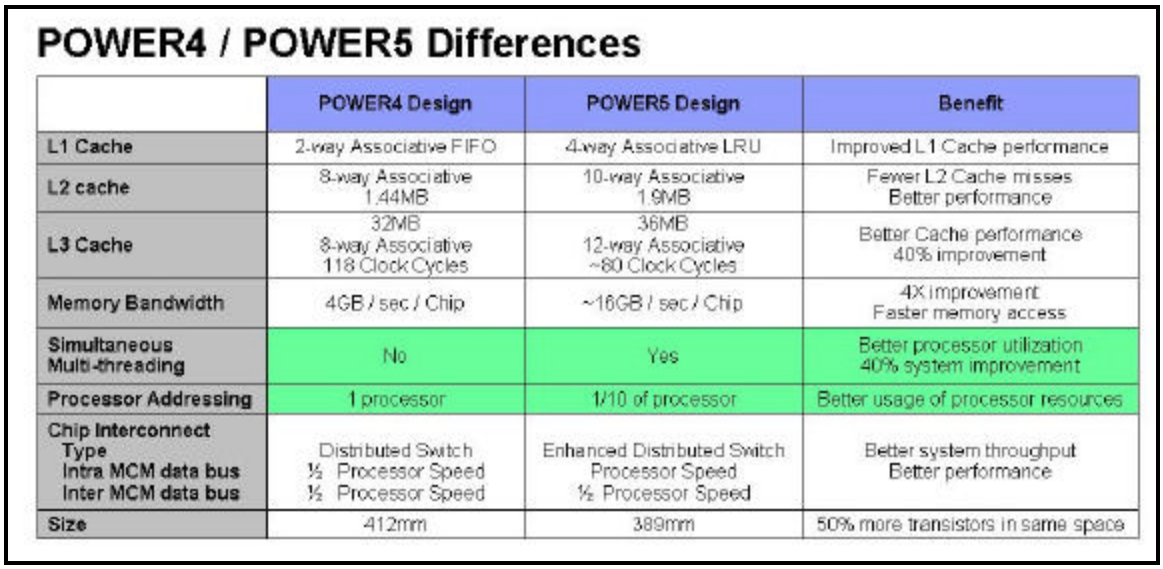
SMT, which has similarities with Intel's Hyper-Threading technology, is an evolution of a pre-existing IBM technology that was introduced in an earlier (1998-1999) version of the POWER processor (the RS64 POWER processor family). In its original delivery, it was called Hardware Multi-Threading (HMT) because hardware support was the design vehicle that provided some degree of performance enhancement. SMT was the next generation of this technology and was delivered on the POWER3 processor. However, it skipped a generation and was not delivered with POWER4.

SMT is now improved, more powerful, and yields greater throughput on POWER5, where it even allows priorities to be assigned to the hardware threads. The effect of using SMT is that a 2-way POWER system will act as if there are four CPUs inside the server box. The bottom line is that SMT lets you run two threads of execution within one processor. As such, processes that typically become blocked within the execution pipeline are, instead, able to pull double duty by running in two threads — and all the while, the POWER processor keeps it all straight.

Large classes of applications, especially in the commercial arena, are able to take advantage of SMT to enjoy a significant performance lift... up to 40%, depending on

workload. Typically, the applications that will be most suited for SMT are those where the throughput of an individual transaction is not as paramount as is the total volume of transactions that are to be executed.

*[NOTE: An excellent white paper, "AIX 5L Support for Micro-partitioning and Simultaneous Multi-threading," by Luke Browning, discusses simultaneous multi-threading as it has been improved and implemented on POWER5 in great detail. The Web site for this white paper is found in the "Links" section of this course.]*




# POWER4/POWER5 Differences

This chart shows a list of the more significant feature enhancements in the POWER5 chip as compared to its most recent relative, the POWER4 processor.

POWER4 L1 cache was 2-way associative “first in, first out” (FIFO). POWER5 L1 cache is 4-way associative “least recently used” (LRU). These two changes improve L1 cache performance significantly.

As mentioned earlier, L2 cache has grown from 1.44 to 1.9 megabytes. L2 is now also 10-way associative from its previous 8-way associative design. These two improvements yield less L2 cache misses and better performance.

Similarly, L3 cache is now 12-way associative — a 50% improvement over its POWER5 implementation. L3 cache is now 36 megabytes instead of the former 32 megabytes and its clock cycle speed has been reduced by more than 30%. Again, this improves L3 cache performance and results in overall processor performance by as much as 40%.

Memory bandwidth is four times greater (~16 gigabytes) on the POWER5 processor. This is because of the onboard memory controller which results in faster memory access. POWER4 has a separate, external memory controller.

Simultaneous multi-threading has already been discussed.

POWER5 supports processor addressing in increments of one tenth of a processor. The notion of logical partitioning support in POWER4 was restricted to a minimum of one whole processor. The subprocessor (microprocessor) granularity offered with POWER5 delivers more control for efficient use of the processor.

The multi-chip module (MCM) can be interconnected onboard or externally. On the POWER5 processor, the intra-MCM data bus runs at processor speed; on POWER4, it ran at half the speed of the processor. For both the POWER4 and POWER5 processors, if an inter-MCM data bus is used, it will run at half the speed of its respective processor.

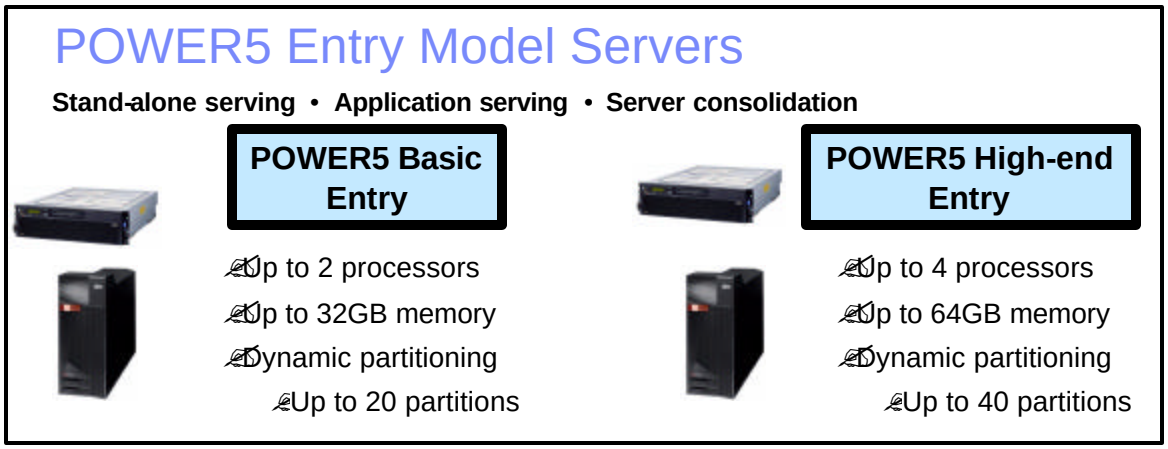
In either case, the POWER5 chip interconnect provides better performance because the chip package is smaller — which is a characteristic of the silicon technology.

One final comparison is that the POWER5 chip is smaller than its POWER4 parent, thus allowing 50% more transistors to be contained in the same space. This is why there is more of the “system” on the chip (L2 cache, L3 director and controller, I/O controller, SMP bus controller, and memory controller). This reduces cycle time in a more power-efficient design, while maintaining binary and structural compatibility with POWER4. Of course, more system “onboard” and reduced power consumption further enhances IBM’s commitment to deliver POWER technology that enhances 24/7 availability.


## POWER5 Server Technology

Now that you understand the POWER5 processor technology, let's talk about the server technology that is underpinned by this advanced chip. You will notice these servers are designed for mission-critical environments. They deliver greater price performance than previous POWER-based server implementations. They are available in a broad range of growth options (entry, midrange, and high-end).

POWER5-based servers offer a multi-platform operating environment with the capability of running AIX 5L™, i5/OS™ (the latest generation of OS/400®), Linux®, Microsoft® Windows® (via IXA or IXS), plus e-business application environments such as WebSphere® and Java™. They also deliver outstanding reliability, availability, and serviceability features.




## POWER5 Entry Model Servers

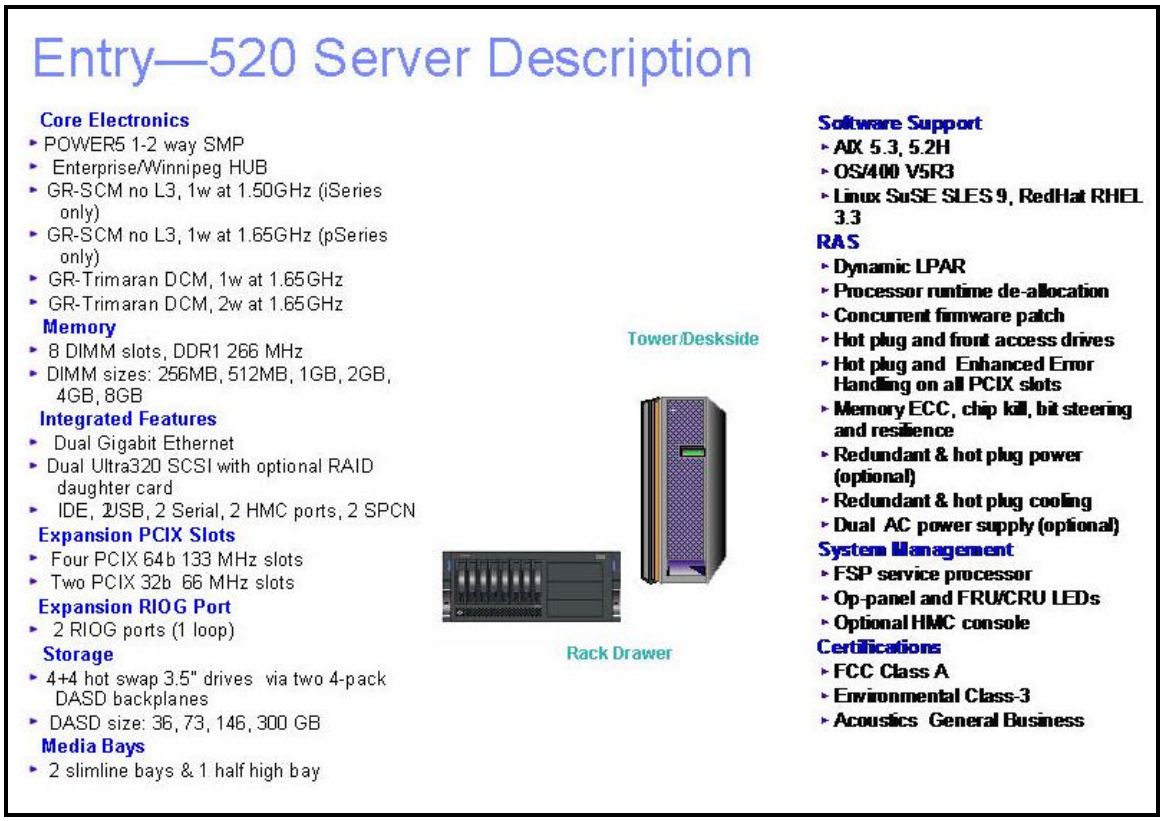
The two entry server models that take advantage of the POWER5 chip are configured to be of value to businesses that need a basic entry system or those requiring an entry server with a bit more power. You can see the range of configuration options for each by glancing at this chart.

The basic entry POWER5 system is the eServer p5 520 UNIX® server and the eServer i5 520 server. It can be ordered as a 1-way or 2-way server and is packaged as a rack mount or desk-side tower. It can be loaded with up to 32 gigabytes of memory. This allows 520 servers to provide higher performance and exploitation of 64-bit addressing — to meet the rigorous demands of today's enterprise computing, such as large database applications.

It supports up to 20 logical partitions that, as mentioned, can enjoy processor partitioning down to the granular level of one tenth of a processor. This represents an enterprise-class dynamic logical partitioning implementation.

The high-end entry POWER5 server is the eServer p5 550 UNIX server and the eServer i5 550 server. They can have up to four processors, and are also packaged as a rack mount or desk-side tower. Notice that the 550 servers have capacity for twice the memory and twice the number of logical partitions of the 520 server.

Let's look at each of these entry-level servers in more detail.




## Basic Entry-520 Server Description

As mentioned, the basic entry eServer 520 server corresponds to an eServer p5 520 server or eServer i5 520 server. The p5 520 server runs at 1.65 gigahertz. The “speed” of the eServer i5 520 server, which is more commonly measured in Commercial Processing Workload (CPW), yields between 400 and 6000 CPW, depending on configuration.

A 1-way SMP eServer 520 server has a chip with one dead core, while a 2-way has a chip with both cores active. The 1-way plugs into one of two Single Core Modules (SCMs) while the 2-way plugs into one of the two Dual Chip Modules (DCMs) listed in the graphic — depending on other configuration elements.

Main memory can be from 512 megabytes to 32 gigabytes, depending on the available Dual Inline Memory Modules (DIMMs).

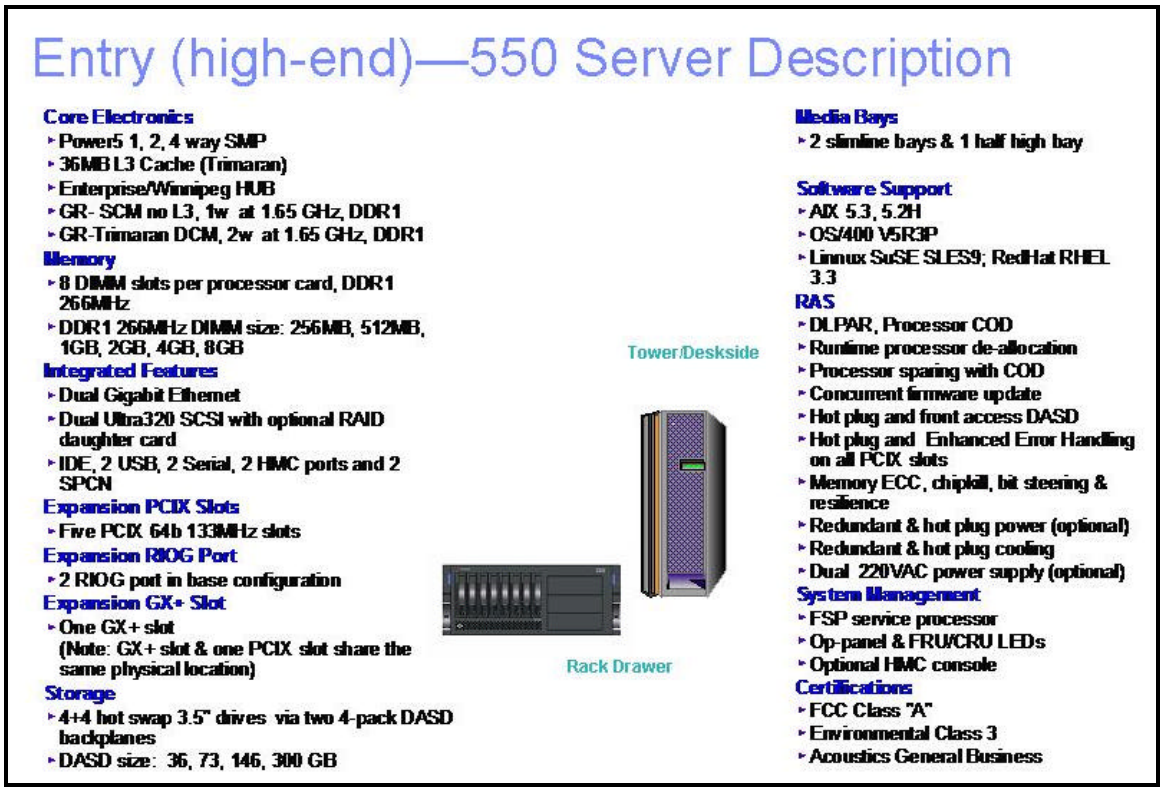
This server has one dual-channel Ultra320 SCSI controller with an optional RAID daughter card, a dual-port 10/100/1000 Mbps integrated Ethernet controller, two serial ports, two USB 2.0 capable ports, two HMC ports, two RIO-2 ports, and two System Power Control Network (SPCN) ports.

There are also six hot-plug PCI-X slots with Enhanced Error Handling (EEH).

There are four hot-swap-capable disk bays in a minimum configuration with an additional four hot-swap-capable disk bays as an optional feature. The eight disk bays can accommodate up to 1.17 terabytes of disk storage using the 146.8-gigabyte Ultra320

SCSI disk drives. Three non-hot-swappable media bays are used to accommodate additional devices. Two media bays only accept slim line media devices, such as DVD-ROM or DVD-RAM drives, and one half-height bay is used for a tape drive.

Reliability and availability features include redundant hot-plug cooling fans and redundant power supply. Along with these hot-plug components, the two eServer 520 servers provide extensive reliability, availability, and serviceability (RAS) features for improved fault isolation, recovery from errors without downtime, avoidance of recurring failures, and predictive failure analysis.




## Entry (high-end) — 550 Server Description

The high-end entry eServer 520 server corresponds to an eServer p5 550 server or eServer i5 550 server.

The eServer p5 550 server also runs at 1.65 gigahertz. The eServer i5 520 server yields between 3300 and 12,000 CPW, based on configuration—from 1-way to 4-way SMP.

As mentioned, main memory can be from two gigabytes to 64 gigabytes.

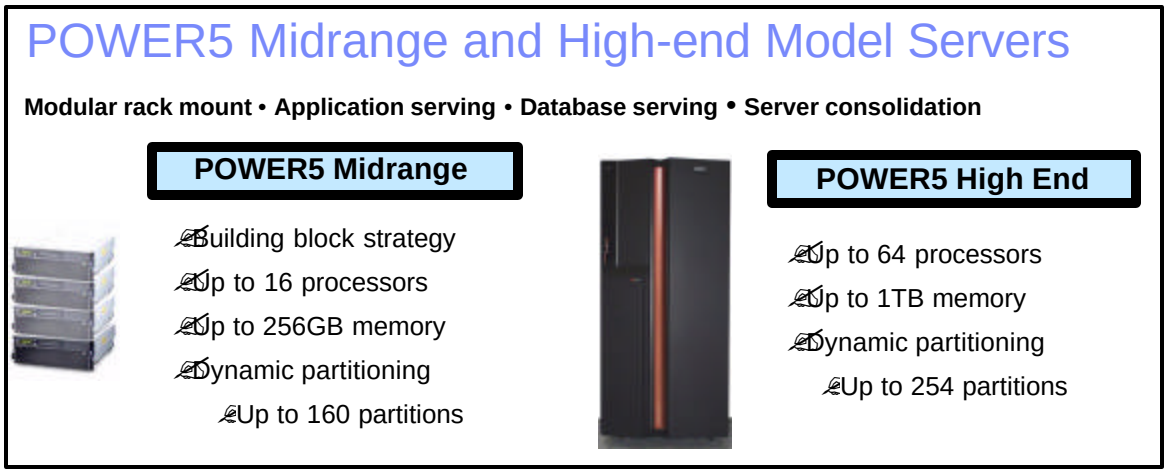
Like the 520 server, the 550 server has one dual-channel Ultra320 SCSI controller with an optional RAID daughter card, a dual-port 10/100/1000 Mbps integrated Ethernet controller, two serial ports, two USB 2.0 capable ports, two HMC ports, two RIO-2 ports, and two System Power Control Network (SPCN) ports.

There are also five hot-plug PCI-X slots with Enhanced Error Handling (EEH). I/O drawers add a maximum of 56 PCI-X slots and 96 disk drive bays for as much as 14.0 terabytes additional storage.

Storage configurations and RAS features are the same as with the 520 servers. The most important additional value provided by the 550 servers over the 520 servers is performance and throughput which results from a maximum 4-way instead of a maximum 2-way configuration.

The Capacity on Demand (CoD) optional features can help eServer p5 550 servers meet changing resource requirements by using processor resources installed on the system but not activated at the time of the original systems purchase. **Capacity Upgrade on**

**Demand (CUoD)** allows companies to purchase additional permanent processor capacity to be activated when the resource is needed. **Trial CoD** offers a one-time, no-additional-charge 30-day trial to allow clients to explore the uses of inactive processor capacity on their server. **Reserve CoD** allows companies to purchase processor features in prepaid blocks of 30 processor days and activate then in full day increments in response to workload demand and then to automatically deactivate the processors when the demand subsides. **On/Off CoD** enables processors to be temporarily activated in full day increments as needed.

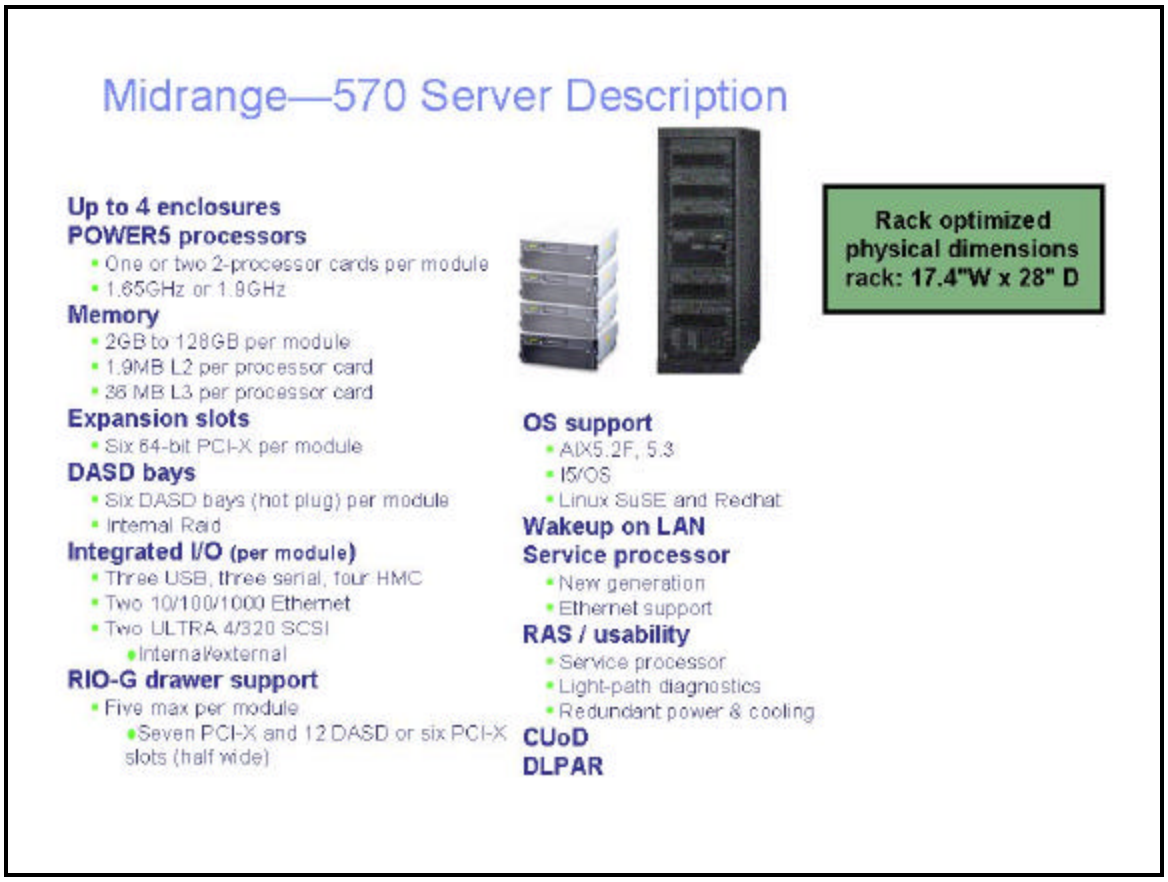



## POWER5 Midrange and High-end Model Servers

The POWER5 servers designed to fit within the midrange enterprise server market (the eServer p5 570 server and the eServer i5 570 server) are rack-mounted and embrace an exciting 4-way building block strategy via special interconnect cables. This allows a 4-way server to be scaled up simply by cabling four more processors — to yield an 8-way SMP. Similarly, a 12-way and then a 16-way SMP server can be cabled together to support incremental growth. The technology that supports this scalability via cabling is very innovative. No one else in the industry has it. These rack-mounted servers can have up to 256 gigabytes of main memory and support 150 logical partitions.

The IBM eServer p5 570 is well-suited for server consolidation projects, database and application serving, e-commerce and departmental or regional server deployments.

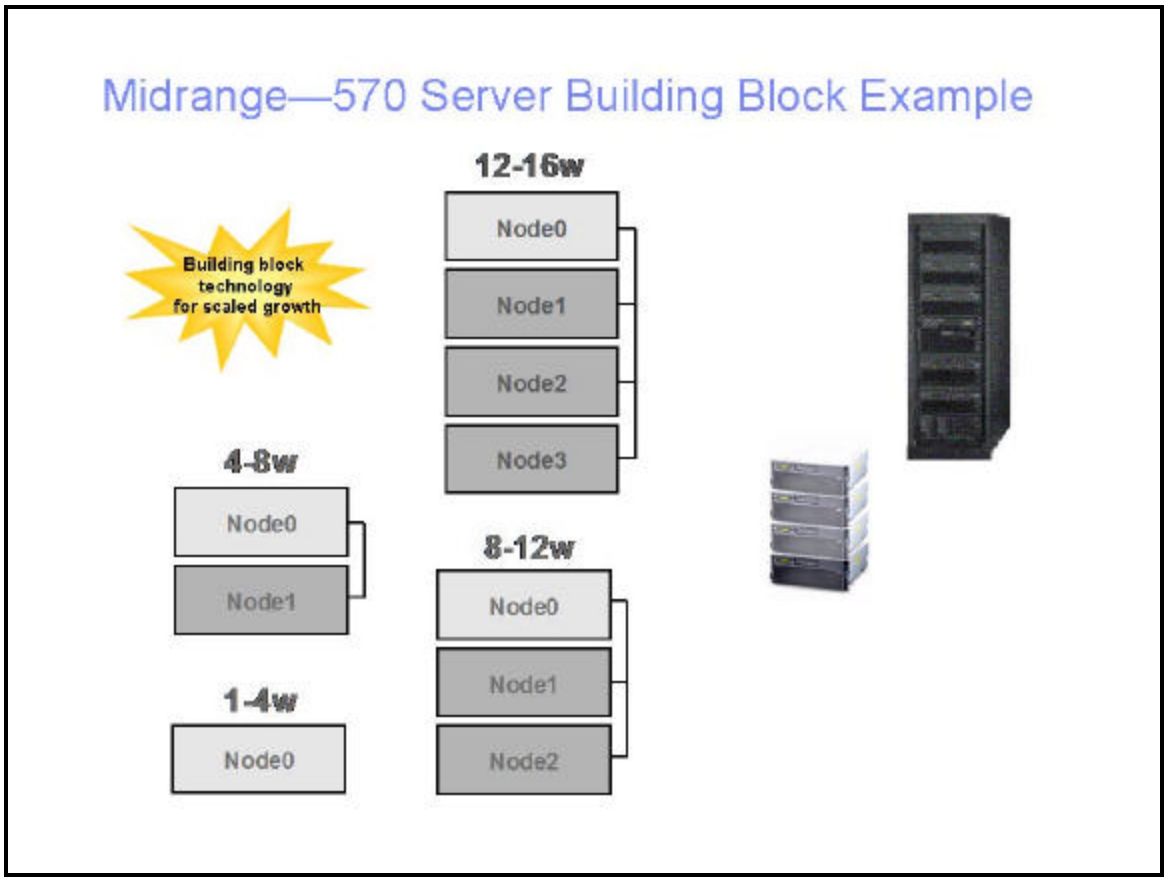
The POWER5 servers that supports high-end enterprise needs (the eServer p5 590 server and the eServer i5 590 server) are standalone, refrigerator-sized towers that offer 64-way SMP processing with main memory that expands all the way up to one terabyte and supports up to 254 dynamic logical partitions.




## Midrange — 570 Server Description

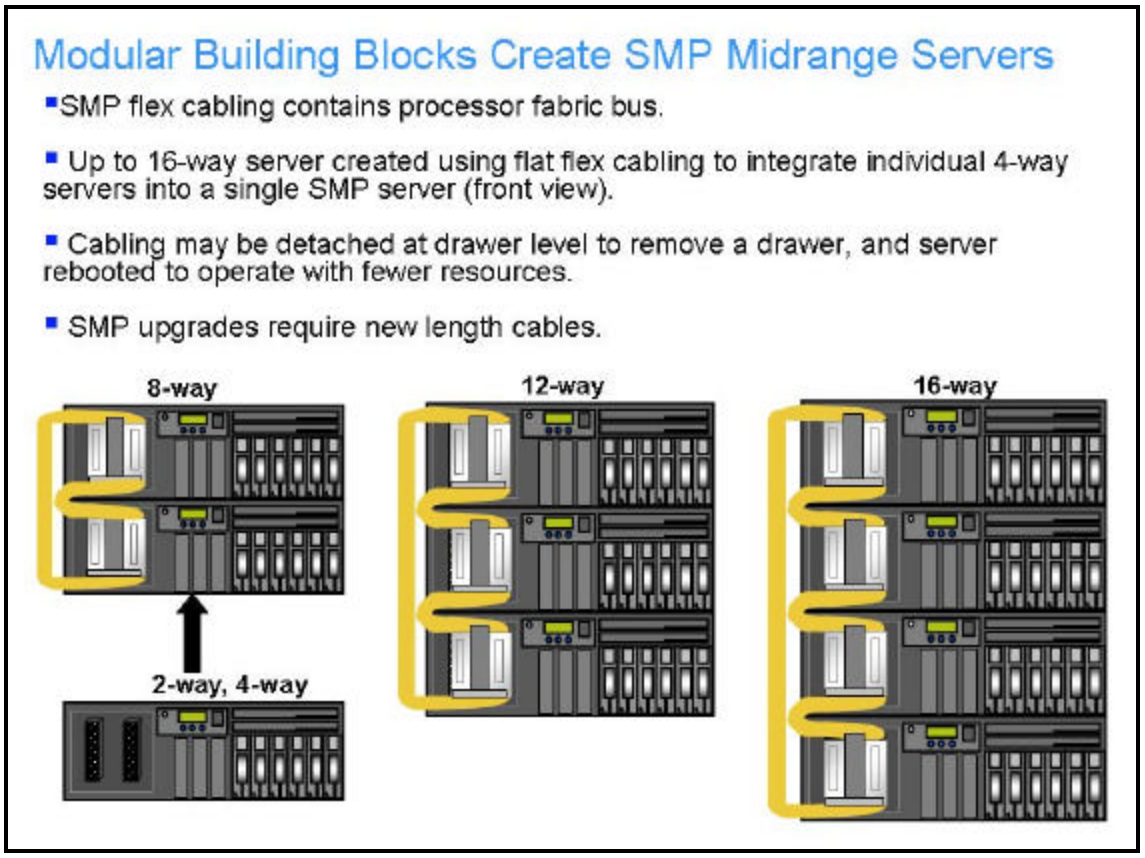
Each eServer p5 570 or eServer i5 570 module can support up to four 1.65 or 1.90 gigahertz processors along with memory, media, disks, I/O adapters, power, and cooling to create a balanced, high-performance rack-mount system. Building-block modules are connected by a unique cabling system at full bus speed. Up to four modules can be integrated into a 19" rack as a single symmetric multiprocessor (SMP) server. Thus, a maximum eServer p5 570 or eServer i5 570 server may consist of 16 processors, 512 gigabytes of memory, eight media bays, 24 PCI-X slots, and 24 internal disk bays accommodating up to 3.5 terabytes of disk storage. In addition, up to 20 optional I/O drawers may be attached, thus significantly adding to the PCI-X and disk bay capacity.

Wakeup on LAN allows the 570 server to automatically configure the network on bootup, thus supporting remote management needs.




## Midrange—570 Server Building Block Example

This chart shows how a business might start out with a 1-to-4-way 570 server rack. Each building block could have one-to-four nodes and up to two dual-core modules (DCMs) that each has one or two processors. Then, you could add racks incrementally to enjoy 4-to-8-way SMP, then 8-to-12-way SMP, and finally, 12-to-16-way SMP. This is a highly flexible, upgradeable environment.

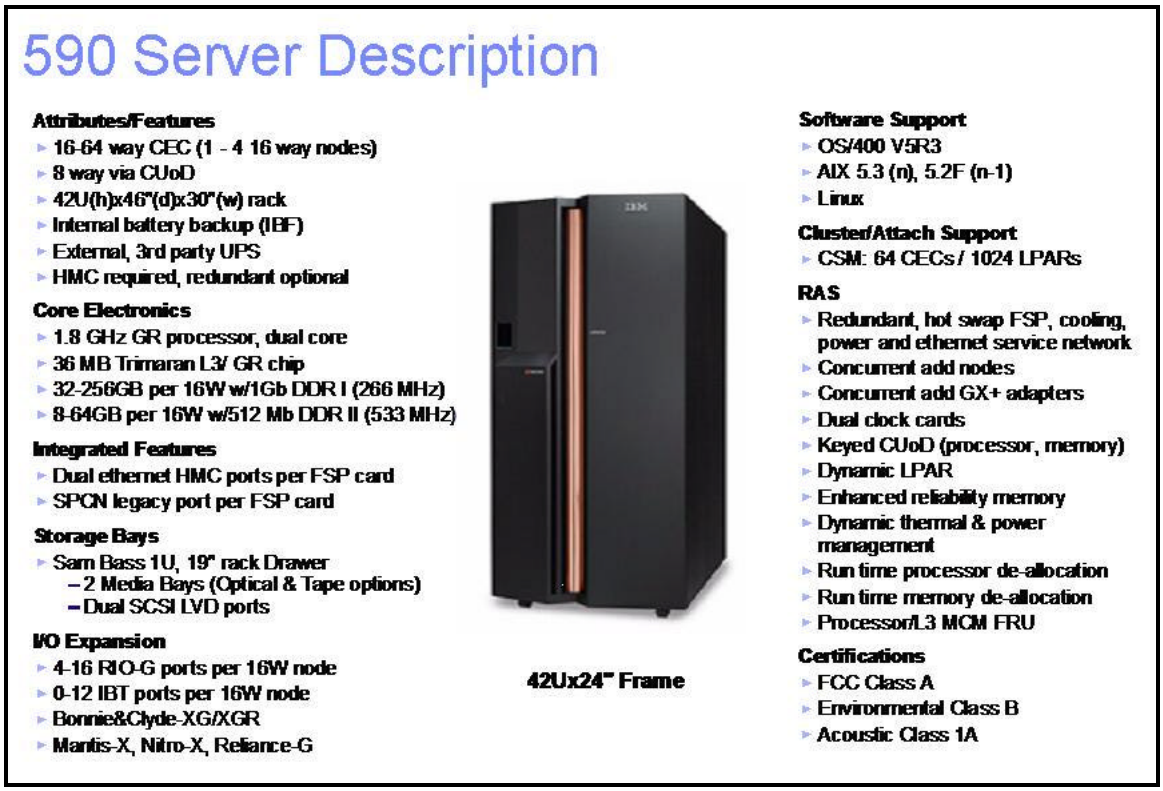



## Modular Building Blocks Create SMP Midrange Servers

This diagram illustrates that up to a 16-way server can be created using flat-flex cabling to integrate individual 4-way servers into a single SMP server. (You are looking at a frontal view of these rack modules).

Notice that there are two set of plugs. The SMP flex-cabling (shown in gold) contains the processor fabric bus and is plugged in to each module as shown progressively in this set of cabling diagrams. The cabling must be replaced each time a new module is added to the rack. In other words, an 8-way set of cabling must be entirely replaced with a 12-way set of ribbon cabling when adding the third module to the rack.

The cabling can be conveniently detached at the drawer level to remove a module, perhaps because of a service need. Then, the server can be rebooted to continue operating, though with fewer resources.




# 590 Server Description

The Central Electronic Complex (CEC) for the eServer p5 590 server contains a 16-way to 64-way SMP. That is, it has between one to four 16-way nodes. *[NOTE: The CEC is the component that contains all of the processors, and is usually a term associated with large rack-mounted servers. The CEC is structured with processor books (aka, nodes).]*

It offers outstanding configuration flexibility to grow with a business. Equipped with eight gigabytes of main memory in the basic configuration, these high-end servers can be scaled to one terabyte using DDR1 266 megahertz memory. Eight to 128 gigabytes of DDR2 533 megahertz memory are useful for high-performance applications. The server features 7.6 megabytes of L2 and 144 megabytes of L3 cache in each MCM to help stage information more effectively from processor memory to applications. The result is that workloads run significantly faster than predecessor servers.

At least one I/O drawer is required with 20 PCI or PCI-X adapter slots and 16 hot-swappable Ultra3 SCSI disk bays for 36.4 gigabytes or 73.4 gigabytes of 15-kilobyte RPM disk drives. With support for 64-bit adapters and backward compatibility for 32-bit cards, these slots provide investment protection and ample room for growth. Hot-plug/blind-swap slots also allow administrators to insert and remove adapters with the I/O drawer in place, which helps prevent system interruption and improves availability. Up to four I/O drawers, as well as a primary and redundant optional integrated battery backup feature, may be installed in the system frame. For more capacity, an expansion frame is available allowing a maximum of eight I/O drawers. This results in a maximum of 160 PCI-X slots and 128 disk storage bays accommodating up to 9.3 terabytes of disk storage.

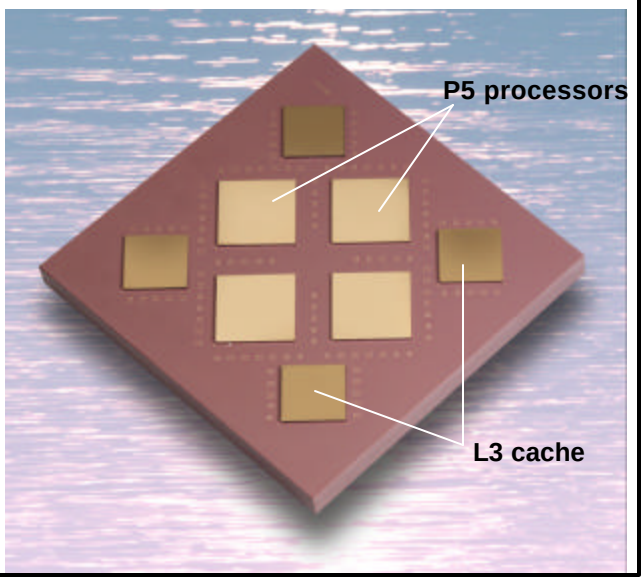
The eServer p5 590 server can be converted to an eServer p5 595 server. This provides even greater performance and increased scalability (maximum of 64 processors, two terabytes of memory, 240 PCI-X slots, 192 disk storage bays, and 14 terabytes of internal disk storage).

The eServer p5 590 server provides new levels of proven, mainframe-inspired reliability, availability, and serviceability for mission-critical applications. It comes equipped with multiple resources to identify and help resolve system problems rapidly. During ongoing operation, error checking and correction (ECC) checks data for errors and can correct them in real time. First Failure Data Capture (FFDC) capabilities log both the source and root cause of problems to help prevent the recurrence of intermittent failures that diagnostics cannot reproduce. Meanwhile, Dynamic Processor Deallocation and dynamic deallocation of PCI bus slots help to reallocate resources when an impending failure is detected so applications can continue to run unimpeded. If problems do arise, a finely grained L2 cache and improved L3 cache line delete capabilities are designed to protect data.

The eServer p5 590 also includes structural elements to help ensure outstanding availability and serviceability. The 24-inch system frame includes hot-swappable disk bays and PCI slots that allow administrators to repair, replace, or install components without interrupting the system. Redundant hot-pluggable power and cooling subsystems provide power and cooling backup in case units fail, and they allow for easy replacement. In the event of a complete power failure, early “Power Off” warning capabilities are designed to perform an orderly shutdown. In addition, both primary and redundant battery backup power subsystems are optionally available.

POWER5 Multi-chip Module

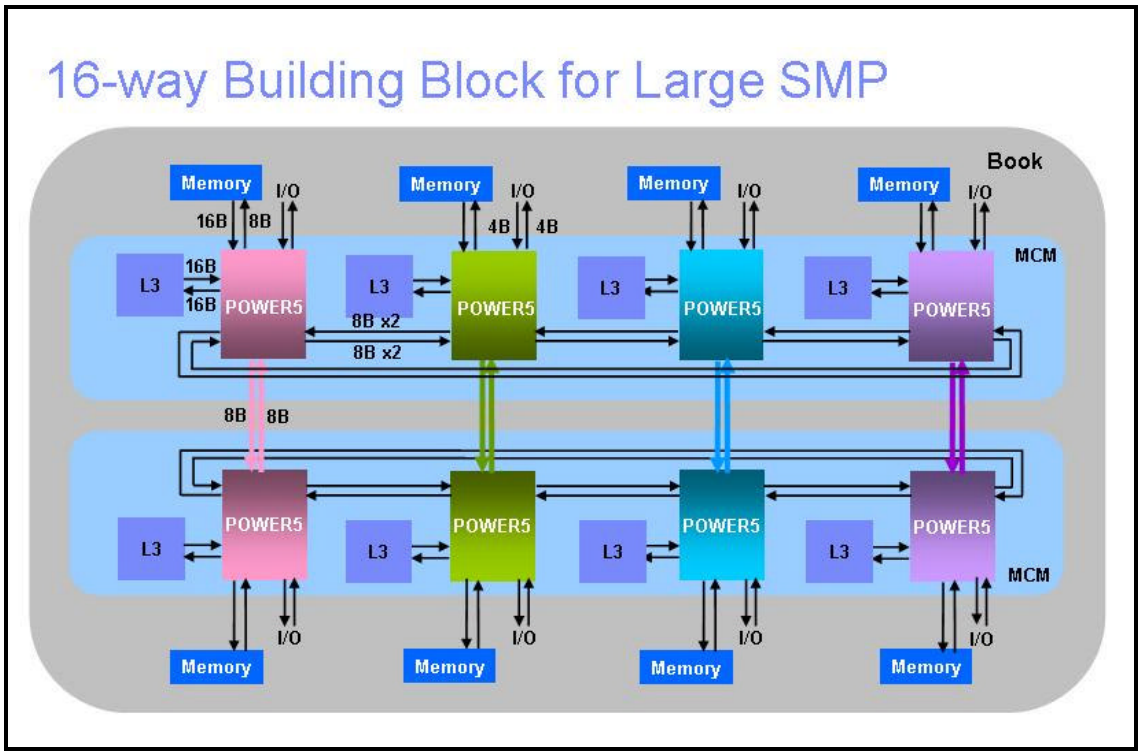
- **MCM package**
  - 4 POWER5 chips
  - 4 L3 cache chips
- **90.25 cm2**
- **4,491 signal I/Os**




## POWER5 Multi-chip Module

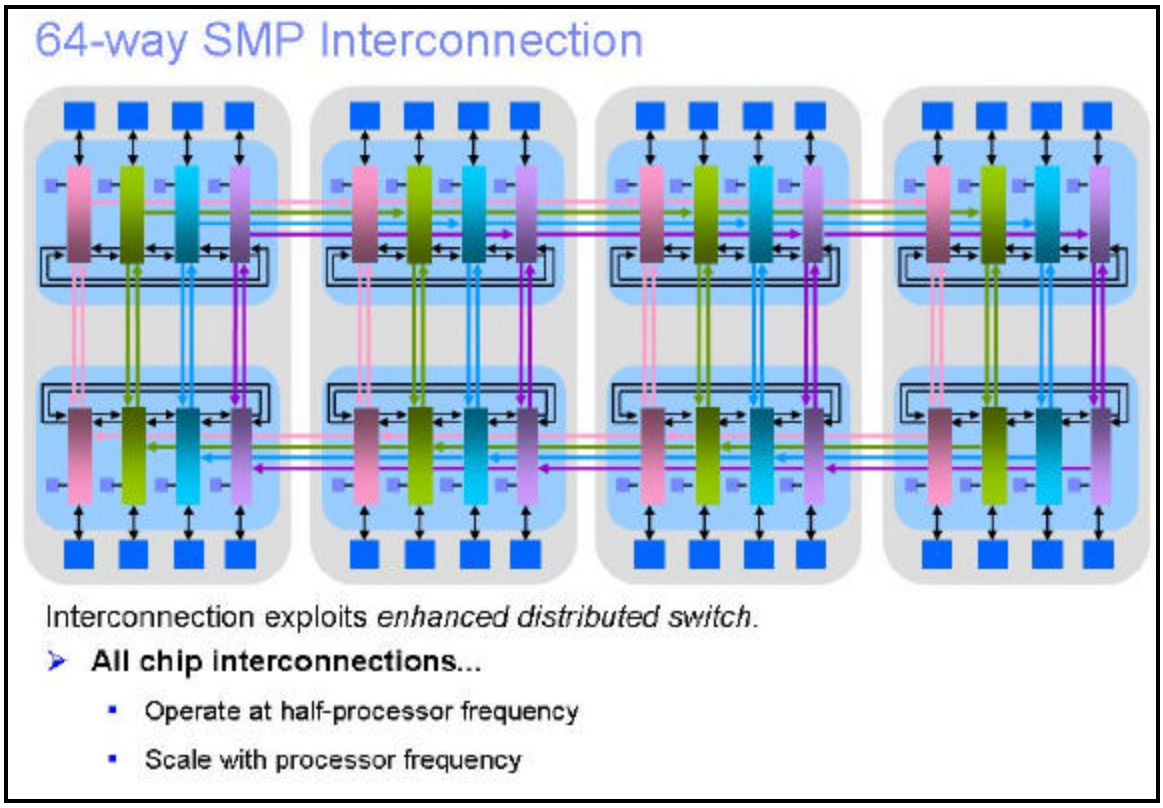
As mentioned, a basic building block for POWER5-based servers is a Multi-Chip Module (MCM). On the other hand, a Single Chip Module (SCM) is also available for 1-way applications.

Each MCM has four POWER5 chips, each of which has two onboard processors. Thus, each MCM actually has eight processors.



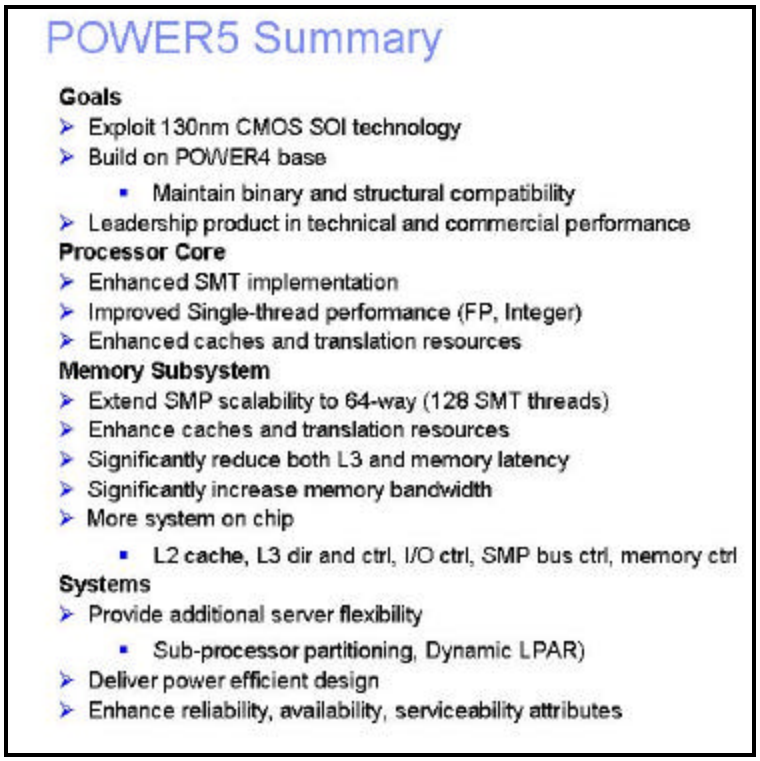

## 16-way Building Block for Large SMP

Here is a schematic of the 16-way building block (processor book) for large SMP implementations. It consists of two MCMs, each of which contains four POWER5 chips. Multiple processor books can be coupled to support 32-way, 48-way, and 64-way SMPs. In other words, because of the organization of the processor books, the level of incremental SMP granularity is a factor of 16.




## 64-Way SMP Interconnection

Here is a picture of four 16-way SMPs (four processor books) that have been interconnected to build a fully configured 64-way SMP server. Interconnection exploits enhanced distributed switching. All chip interconnections (inter-module GX buses) operate at half-processor speed and are eight bits wide. As you would expect, the chip interconnections scale with processor frequency. The chips talk to each other via a ring topology and data moves from one module to another in one direction.




## Summary

When designing the POWER5 processor, the primary goals were to exploit the CMOS SOI technology that had already been proven in the POWER4 line of chips. It was also important to maintain binary and structural compatibility between the two processors. A focus on maintaining and strengthening the POWER processor's leadership characteristics, in both technical and commercial performance was also important. The POWER5 processor core enjoys an enhanced SMT implementation as well as improved single-thread performance.

The memory subsystem has been beefed up significantly. SMP scalability has been extended to 64-way. All caches and translation resources have been enhanced with significant reduction in L3 and memory latency. Memory bandwidth has increased. There is more "system" on the chip.

eServer p5 servers and eServer i5 servers reflect the wealth of improved subsystems and design elements that are delivered by the POWER5 chip. This includes greater granularity for dynamic logical partitions, capacity on demand features that are stronger than ever, a faster and more power-efficient implementation across the server line, which leads to enhanced reliability, availability, and serviceability attributes.

# Links

- White paper: AIX 5L Support for Micro-partitioning and Simultaneous Multi-threading: **ibm.com/**servers/aix/whitepapers/aix_support.pdf
- IBM eServer p5, pSeries, OpenPower and IBM RS/6000 Performance Report: **ibm.com/**eserver/pseries/hardware/system_perf.html
- IBM eServer p5 520 Technical Overview and Introduction Redbook: **ibm.com/**redbooks/abstracts/REDP9111.html

## Trademarks and Disclaimers

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries: IBM, eServer, iSeries, pSeries, POWER, POWER3, POWER4+, POWER5, POWER5+, PowerPC, PowerPC 630, BladeCenter, AIX 5L, i5/OS, OS/400, WebSphere

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

Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.

Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries

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

All other products may be trademarks or registered trademarks of their respective companies.