How memory configuration can affect the performance of System x and BladeCenter servers
March 2009

Optimizing the Performance of IBM System x and BladeCenter Servers using Intel Xeon 5500 Series Processors

Ganesh Balakrishnan
IBM System x and BladeCenter Performance

Ralph M. Begun
IBM System x Development

1.0 Introduction

The Intel® Xeon® 5500 series processors are next-generation quad-core processors targeted at the two-socket server space and will be the common building block across a number of IBM platforms, including the IBM® BladeCenter® **HS22** blade server, the 1U **x3550 M2** and 2U **x3650 M2** rack servers, and the IBM iDataPlex™ **dx360 M2** server. With the Xeon 5500 series processors, Intel has diverged from its traditional Symmetric Multiprocessing (SMP) architecture to a Non-Uniform Memory Access (NUMA) architecture. In a two-processor scenario, the Xeon 5500 series processors are connected through a serial coherency link called QuickPath Interconnect (QPI). The QPI is capable of 6.4, 5.6 or 4.8 GT/s (gigatransfers per second), depending on the processor model. The Xeon 5500 series integrates the memory controller within the processor, resulting in two memory controllers in a two-socket system. Each memory controller has three memory channels and supports DDR-3 memory. Depending on processor model, the type of memory used, and the population of memory, memory may be clocked at 1333MHz, 1066MHz or 800MHz. Each memory channel supports up to 3 DIMMs per channel (DPC), for a theoretical maximum of 9 DIMMs per processor or 18 per 2-socket server. (See *Figure 1* for illustration.) However, the actual maximum number of DIMMs per system is dependent upon the system design.

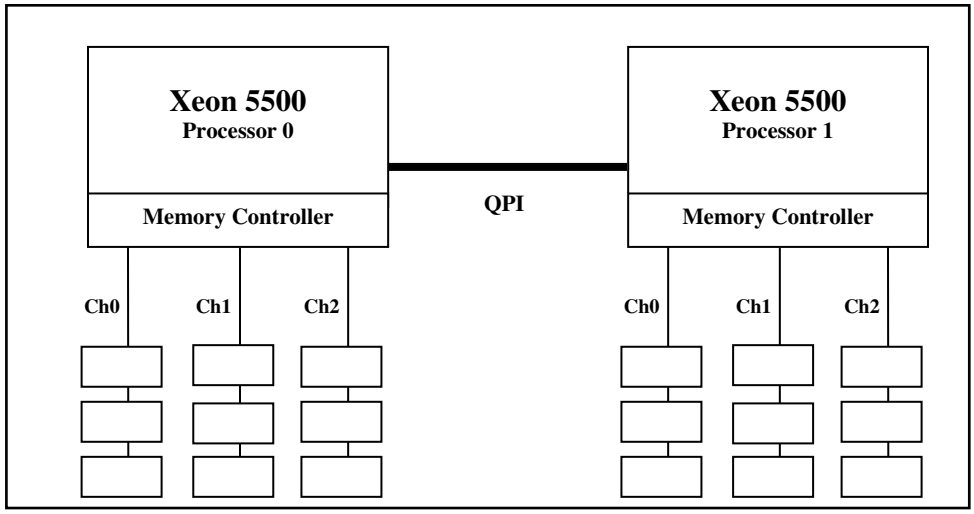

Figure 1. *Xeon 5500 architecture showing maximum memory capabilities*

2.0 System Architecture

In this section, we will explore the system architectures of various IBM System x® and BladeCenter servers equipped with Xeon 5500 series processors, from a memory standpoint.

2.1 HS22 Blade

HS22 is designed with 12 DIMM slots as shown in *Figure 2* and *Figure 3*. The 12-DIMM layout provides 6 DIMMs per socket and 2 DIMMs per channel (DPC).

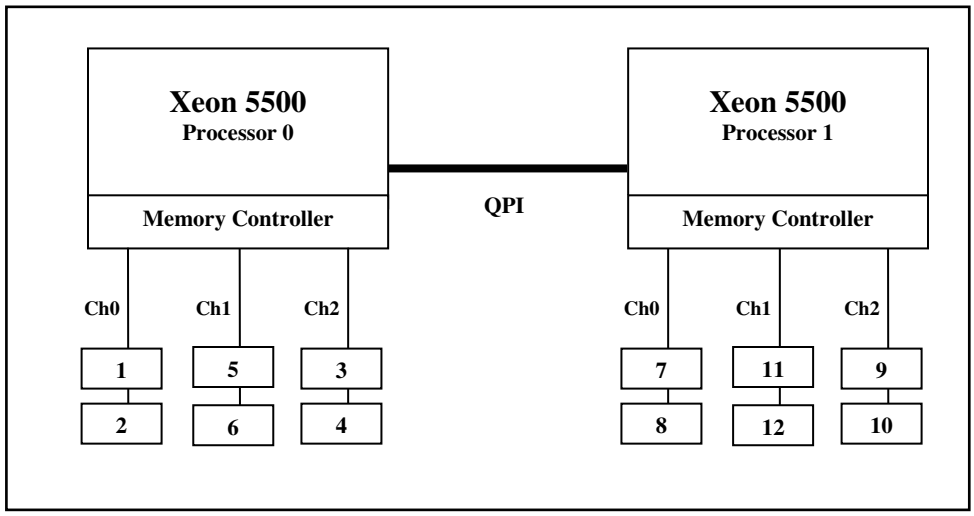

Figure 2. *HS22 DIMM slots architectural layout*

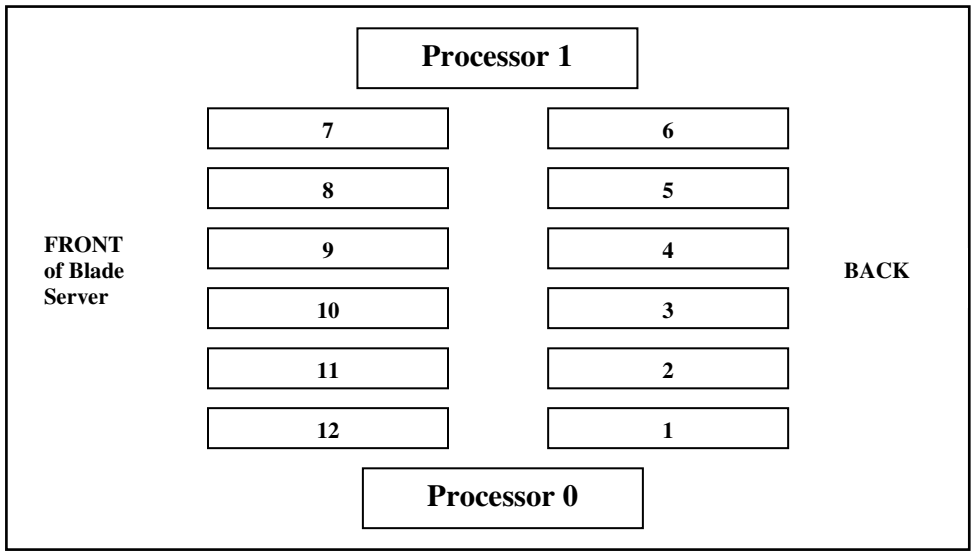

Figure 3. *HS22 DIMM slots physical layout*

2.2 System x3550 M2, x3650 M2, and iDataPlex dx360 M2

As shown in *Figure 4* and *Figure 5* below, the other IBM servers containing Xeon 5500 series processors—the System x3550 M2, the x3650 M2, and the iDataPlex dx360 M2—each provide 16 DIMM slots. Like the HS22, each processor has an equal number of DIMM slots. However, unlike the HS22 all memory channels do not have equal DPC (DIMMs per channel).

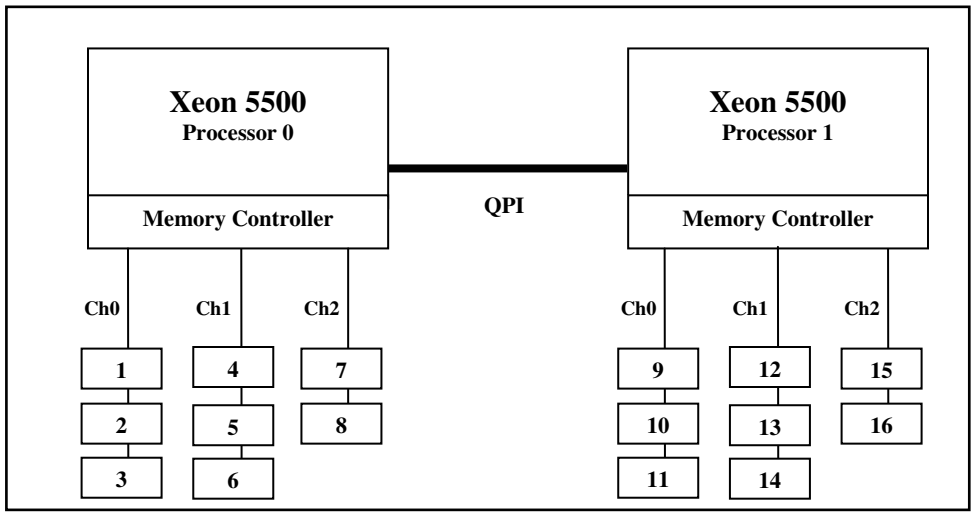

Figure 4. *x3550 M2/x3650 M2/dx360 M2 DIMM slots architectural layout*

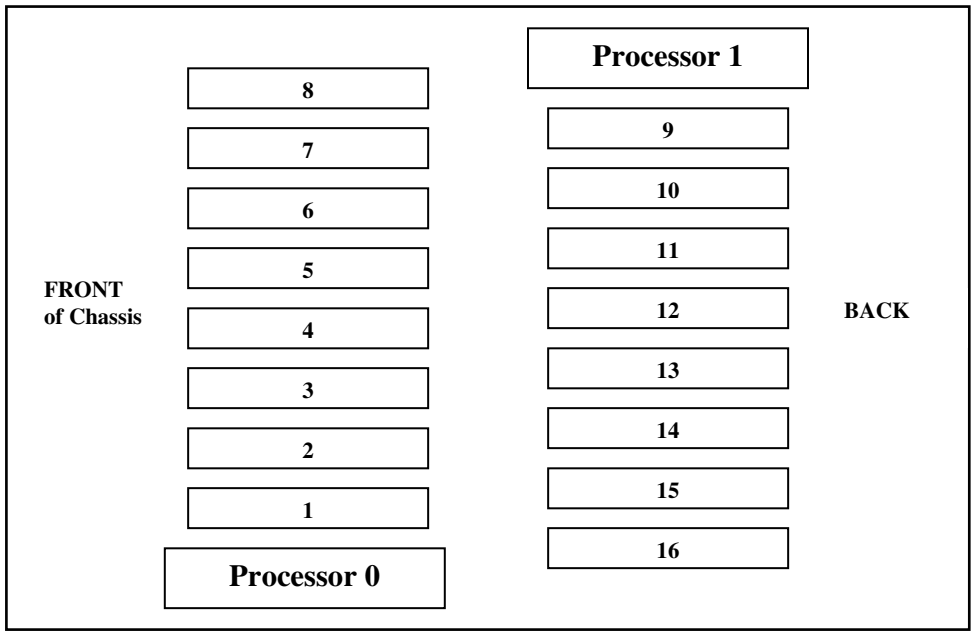

Figure 5. *x3550 M2/x3650 M2/dx360 M2 DIMM slots physical layout*

3.0 Memory Performance

With the varied number of configurations possible in the Xeon 5500 series processor-based systems, a number of variables emerge that influence processor/memory performance. The main variables are memory speed, memory interleaving, memory ranks and memory population across various memory channels and processors. Depending on the processor model and number of DIMMs, the performance of the Xeon 5500 platform will see large memory performance variances. We will look at each of these factors more closely in the next sections.

3.1 Memory Speed

As mentioned earlier, the memory speed is determined by the combination of the processor model, DIMM speed, and DIMMs per channel.

3.1.1 Processor model

The initial Xeon 5500 series processor-based offerings will be categorized into 3 bins called Performance, Volume and Value. The 3 bins have the ability to clock memory at different maximum speeds:

- 1333MHz (**X55**xx processor models)
- 1066MHz (**E552**x or **L552**x and up)
- 800MHz (**E550**x)

So, the processor model will limit the maximum frequency of the memory. ***Note:*** Because of the integrated memory controllers the former front-side bus (FSB) no longer exists.

3.1.2 DDR3 DIMM Speed

DDR-3 memory will be available in various sizes at speeds of 1333MHz and 1066MHz. 1333MHz represents the maximum capability at which memory can be clocked. However, the memory will not be clocked faster than the capability of the processor model and will be clocked appropriately by the BIOS.

3.1.3 DIMMs per Channel (DPC)

The number and type of DIMMs and the channels in which they reside will also determine the speed at which memory will be clocked. *Table 1* describes the behavior of the platform. The table below assumes a 1333MHz-capable processor model (X55xx). If a slower processor model is used, then the memory speed will be the lower of the memory speed and the processor model memory speed capability. If the DPC is not uniform across all the channels, then the system will clock to the frequency of the slowest channel.

DPC	DIMM Speed (MHz)	Ranks per DIMM	Memory Speed (MHz)
1	**1333**	**1,2**	**1333**
2	1333	1,2	1066
3	1333	1,2	800
1	1333	4	1066
2	1333	4	800
1	**1066**	**1,2**	**1066**
2	**1066**	**1,2**	**1066**
3	1066	1,2	800
1	**1066**	**4**	**1066**
2	1066	4	800

Table 1. *Memory speed clocking.* (Full-speed configurations in **bold**.)

3.1.4 Low-level Performance Specifics

It is important to understand the impact of the performance of the Xeon 5500 series platform, depending on the memory speed. We will use both low-level memory tools and application benchmarks to quantify the impact of memory speed.

Two of the key low-level metrics that are used to measure memory performance are memory latency and memory throughput. We use a base Xeon 5500 2.93GHz, 1333MHz-capable 2-socket system for this analysis. The memory configurations for the three memory speeds in the following benchmarks are as follows:

- **1333MHz** – 6 x 4GB dual-rank 1333MHz DIMMs
- **1066MHz** – 12 x 2GB dual-rank DIMMs for 1066MHz
- **800MHz** – 12 x 2GB dual-rank DIMMs clocked down to 800MHz in BIOS

Note: Memory ranks are explained in detail in section 3.3.

As shown in *Figure 6* below, we show the unloaded latency to local memory. The unloaded latency is measured at the application level and is designed to defeat processor prefetch mechanisms. As shown in the figure, the difference between the fastest and slowest speeds is about 10%. This represents the high watermark for latency-sensitive workloads. Another important thing to note is that this is almost a **50%** decrease in memory latency when compared to the previous generation Xeon 5400 series processor on 5000P chipset platforms.

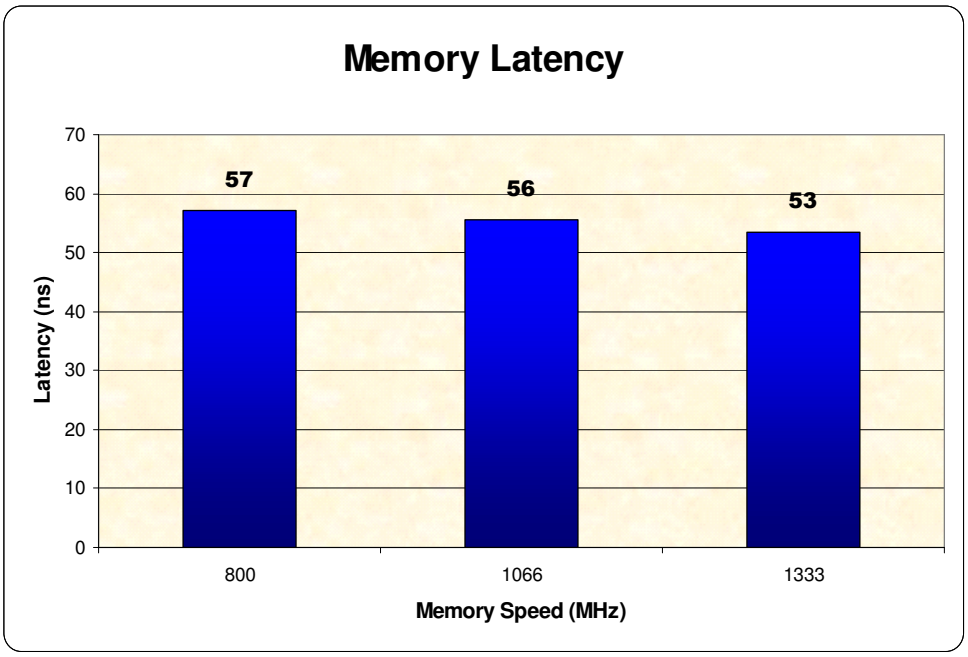

Figure 6. Xeon 5500 series memory latency as a function of memory speed

A better indicator of application performance is memory throughput. We use the triad component of the streams benchmark to compare the performance at different memory speeds. The memory throughput assumes all local memory allocation and all 8 cores utilizing main memory. As shown in *Figure 7*, the performance gain from running memory at 1066MHz versus 800MHz is **28%**, and the performance gain from running at 1333MHz versus 1066MHz is **9%**. So, the performance penalty of clocking memory down to 800MHz is far greater than clocking it down to 1066MHz. This new processor design comes with some trade-offs in memory capacity, performance, and cost: For example, more lower-cost/lower-capacity DIMMs mean lower memory speed. Alternatively, fewer higher-capacity DIMMs cost more but offer higher performance.

Regardless of memory *speed*, the Xeon 5500 platform represents a significant improvement in memory *bandwidth* over the previous Xeon 5400 platform. At 1333MHz, the improvement is almost **500%** over the previous generation. This huge improvement is mainly due to dual integrated memory controllers and faster DDR-3 1333MHz memory. This improvement translates into improved application performance and scalability.

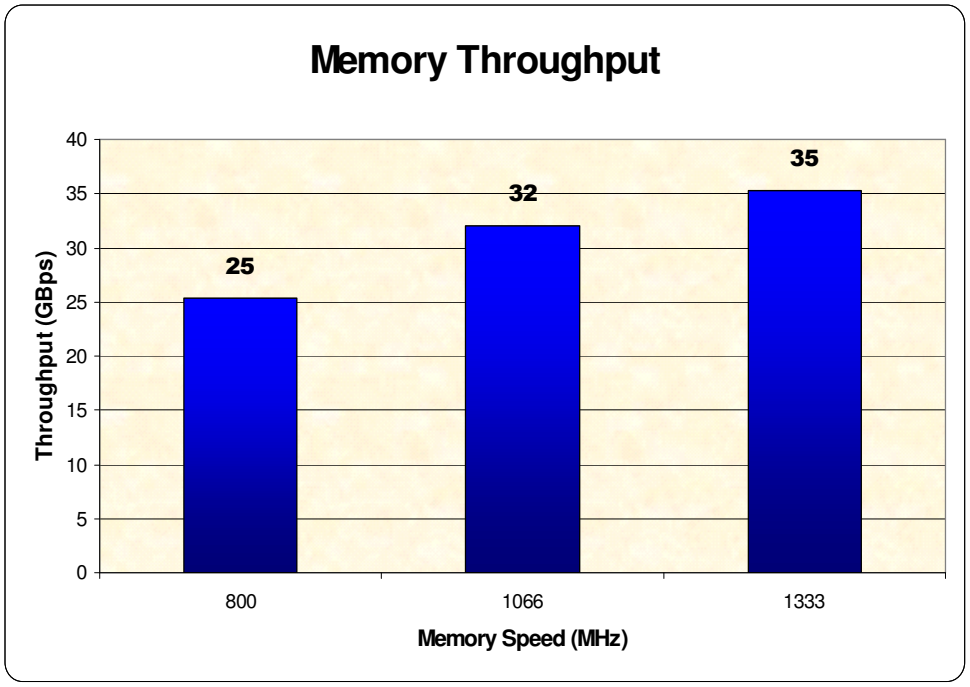

Figure 7. *Memory throughput using Streams Triad*

3.1.5 Application Performance

In this section, we will discuss the impact of memory speed on the performance of three commonly used benchmarks: SPECint®2006_rate, SPECfp®2006_rate and SPECjbb®2005. In each case, the benchmark scores are relative to the score at 800MHz as shown in *Figure 8*.

SPECint2006_rate is typically used as an indicator of performance for commercial applications. It tends to be more sensitive to processor frequency and less to memory bandwidth. There are very few components in SPECint2006_rate that are memory bandwidth intensive and so the performance gain with memory speed improvements is the least for this workload. In fact, most of the difference observed is due to one of the sub-benchmarks that shows a high sensitivity to memory frequency. There is an **8%** improvement going from 800MHz to 1333MHz while the improvement in memory bandwidth is almost **40%**.

SPECfp_rate is used as an indicator for HPC (high-performance computing) workloads. It tends to be memory bandwidth intensive and should reveal significant improvements for this workload as memory frequency increases. As expected, a number of sub-benchmarks demonstrate improvements as high as the difference in memory bandwidth. As shown in *Figure 8*, there is a **13%** gain going from 800MHz to 1066MHz and another **6%** improvement with 1333MHz. SPECfp_rate captures almost **50%** of the memory bandwidth improvement.

SPECjbb2005 is a workload that does not stress memory but keeps the data bus moderately utilized. This workload provides a middle ground and the performance gains reflect that trend. As shown in *Figure 8*, there is an **8%** gain from 800MHz to 1066MHz and another **2%** upside with 1333MHz.

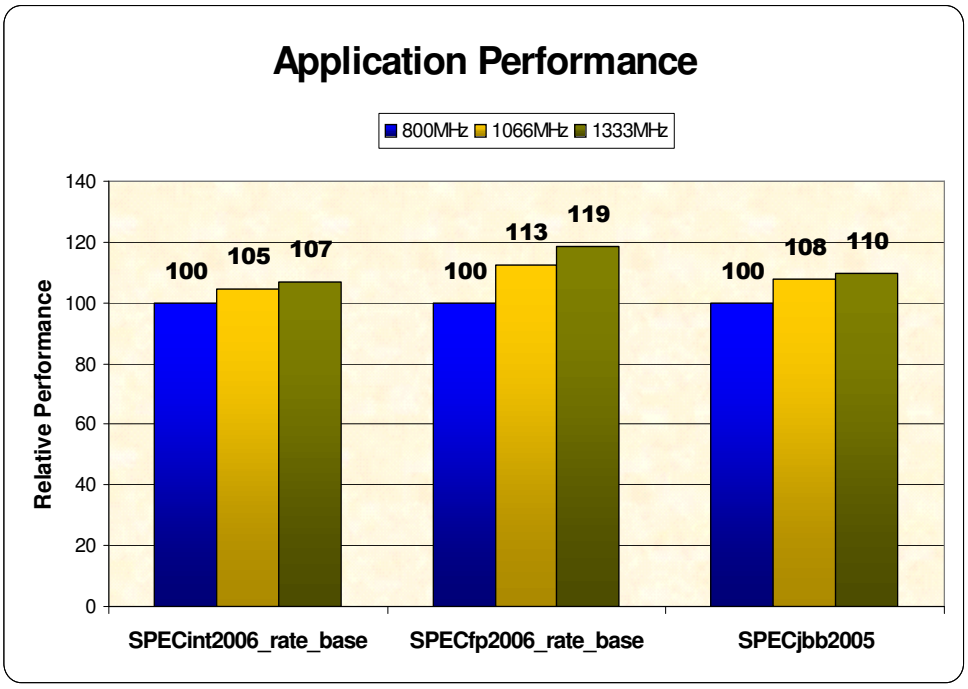

Figure 8. *Application performance as a function of memory speed*

3.2 Memory Interleaving

Memory interleaving refers to how physical memory is interleaved across the physical DIMMs. A balanced system provides the best interleaving. A Xeon 5500 series processor-based system is balanced when all memory channels on a socket have the same amount of memory. The simplest way to enforce optimal interleaving is by populating 6 identical DIMMs at 1333MHz, 12 identical DIMMs at 1066MHz and 18 identical DIMMs (where supported by platform) at 800MHz.

3.2.1 HS22 Blade Server

For HS22, which has a balanced DIMM layout, it is easy to balance the system for all three memory frequencies. The recommended DIMM population is shown in *Table 2*, assuming DIMMs with identical capacities.

Desired Memory Speed	DIMMs per Channel	DIMM Slots to Populate
1333MHz	1	2, 4, 6, 8, 10, and 12
1066MHz	2	All slots
800MHz	2	All slots; clock memory speed to 800MHz in BIOS

Table 2. *Memory configurations to produce balanced performance in HS22*

3.2.2 x3650 M2/x3550 M2/dx360 M2 Rack Systems

For systems with 16 DIMM slots, care needs to be taken when populating the slots, especially when configuring for large DIMM counts at 800MHz.

When configuring for 800MHz, with large DIMM counts, it would be a common error to populate all 16 DIMM slots with identical DIMMs. However, such a configuration leads to an unbalanced system where two memory channels have less memory capacity than the other four. (In other words, two channels with, for example, 3 x 4GB DIMMs and one channel with 2 x 4GB DIMMs.)

This leads to lessened performance. *Figure 9* shows the impact of reduced interleaving. The first configuration is a balanced baseline configuration where the memory is down-clocked to 800MHz in BIOS. The second configuration populates four channels with 50% more memory than two other channels causing an unbalanced configuration. The third configuration balances the memory on all channels by populating the channels with fewer DIMM slots with a DIMM that is double the capacity of others. (For example, two channels with 3 x 4GB DIMMs and one channel with 1 x 4GB and 1 x 8GB DIMMs.) This ensures that all channels have the same capacity. As *Figure 9* shows, the first and third balanced configurations significantly outperform the unbalanced configuration. Depending on the memory footprint of the application and memory access pattern, the impact could be higher or lower than the two applications cited in the figure.

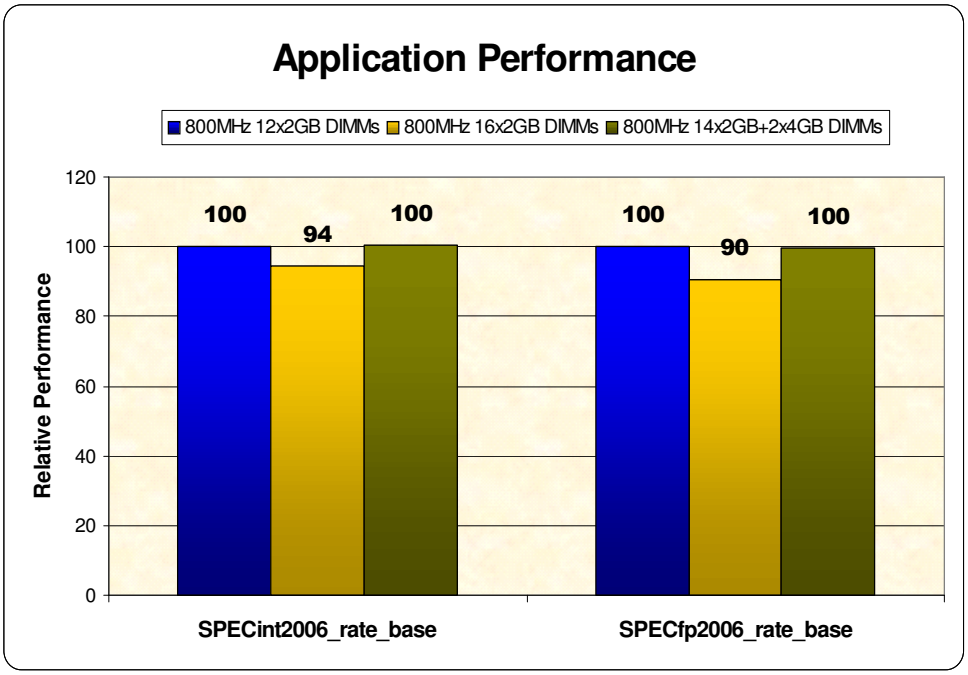

Figure 9. Impact of unbalanced memory configuration

The recommended DIMM population is shown in *Table 3.*

Desired Memory Speed	DIMMs per Channel	DIMM Slots to Populate
1333MHz	1	3, 6, 8, 11, 14, 16
1066MHz	2	2, 3, 5, 6, 7, 8, 10, 11, 13, 14, 15, 16
800MHz	2	2, 3, 5, 6, 7, 8, 10, 11, 13, 14, 15, 16 and clock memory down to 800MHz in BIOS
800MHz	>2	1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15 with DIMMs of size 'x' 8 and 16 with DIMMs of size '2x'

Table 3. *Memory configurations to produce balanced performance in System x servers*

3.3 Memory Ranks

A memory rank is simply a segment of memory that is addressed by a specific address bit. DIMMs typically have 1, 2 or 4 memory ranks, as indicated by their size designation.

- A typical memory DIMM description: 2GB 4R x8 DIMM

- The 4R designator is the rank count for this particular DIMM (R for rank = 4)
- The x8 designator is the data width of the rank

It is important to ensure that DIMMs with the appropriate number of ranks are populated in each channel for optimal performance. Whenever possible, it is recommended to use *dual*-rank DIMMs in the system. Dual-rank DIMMs offer better interleaving and hence better performance than single-rank DIMMs. For instance, a system populated with 6 x 2GB *dual*-rank DIMMs outperforms a system populated with 6 x 2GB *single*-rank DIMMs by **7%** for SPECjbb2005. Dual-rank DIMMs are also better than quad-rank DIMMs because quad-rank DIMMs will cause the memory speed to be down-clocked.

Another important guideline is to populate equivalent ranks per channel. For instance, mixing single-rank and dual-rank DIMMs in a channel should be avoided.

3.4 Memory Population across Memory Channels

It is important to ensure that all three memory channels in each processor are populated. The relative memory bandwidth is shown in *Figure 10*, which illustrates the loss of memory bandwidth as the number of channels populated decreases. This is because the bandwidth of all the memory channels is utilized to support the capability of the processor. So, as the channels are decreased, the burden to support the requisite bandwidth is increased on the remaining channels, causing them to become a bottleneck.

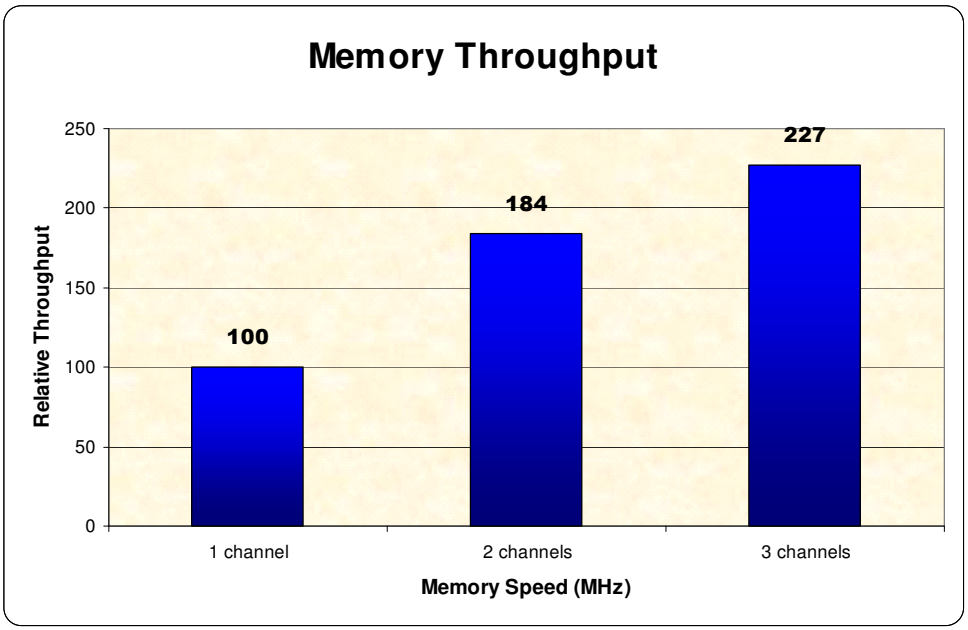

Figure 10. The effect of populating different number of channels

3.5 Memory Population Across Processor Sockets

Because the Xeon 5500 series uses NUMA architecture, it is important to ensure that both memory controllers in the system are utilized, by providing both processors with memory. If only one processor is installed, only the associated DIMM slots can be used. Adding a second processor not only doubles the amount of memory available for use, but also doubles the number of memory controllers, thus doubling the system memory bandwidth. It is also optimal to populate memory for both processors in an identical fashion to provide a balanced system. Using *Figure 11* as an example, Processor 0 has DIMMs populated but no DIMMs are populated for Processor 1. In this case, Processor 0 will have access to low latency local memory and high memory

bandwidth. However, Processor 1 has access only to remote or “far” memory. So, threads executing on Processor 1 will have a long latency to access memory as compared to threads on Processor 0.

This is due to the latency penalty incurred to traverse the QPI links to access the data on the remote memory controller. The latency to access remote memory is almost *75% higher* than local memory access. The bandwidth to remote memory is also limited by the capability of the QPI links. So, the goal should be to always populate both processors with memory.

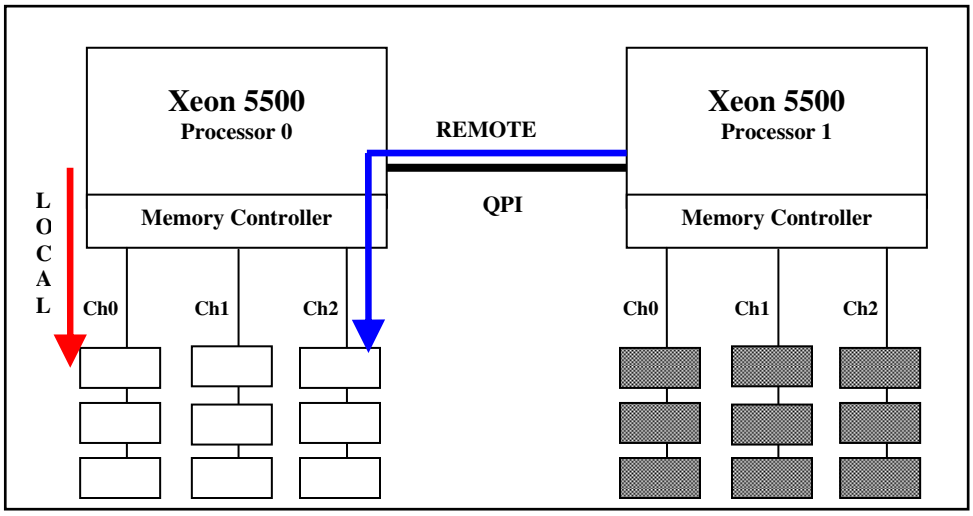

Figure 11. *Diagram showing local and remote memory access*

4.0 Best Practices

In this section, we recapture the various rules to be followed for optimal memory configuration on the Xeon 5500 based platforms.

4.1 Maximum Performance

Follow these rules for peak performance:

- Always populate both processors with equal amounts of memory to ensure a balanced NUMA system.
- Always populate all 3 memory channels on each processor with equal memory capacity.
- Ensure an even number of ranks are populated per channel.
- Use dual-rank DIMMs whenever appropriate.
- For optimal 1333MHz performance, populate 6 dual-rank DIMMs (3 per processor).
- For optimal 1066MHz performance, populate 12 dual-rank DIMMs (6 per processor).
- For optimal 800MHz performance with high DIMM counts:
  - On 12 DIMM platforms, populate 12 dual-rank or quad-rank DIMMs (6) per processor.
  - On 16 DIMM platforms:
    - Populate 12 dual-rank or quad-rank DIMMs (6 per processor).
    - Populate 14 dual-rank DIMMs of one size and 2 dual-rank DIMMs of double the size as described in the interleaving section.
- With the above rules, it is not possible to have a performance-optimized system with 4GB, 8GB, 16GB, or 128GB. With 3 memory channels and interleaving rules, customers need to

configure systems with 6GB, 12GB, 18GB, 24GB, 48GB, 72GB, 96GB, etc., for optimized performance.

4.2 Other Considerations

4.2.1 Plugging Order

Take care to populate empty DIMM sockets in the specific order for each platform when adding DIMMs to Xeon 5500 series platforms, The DIMM socket farthest away from its associated processor, per memory channel, is always plugged first. Consult the documentation with your specific system for details.

4.2.2 Power Guidelines

This document is focused on maximum performance configuration for Xeon 5500 series processor-based systems. Here are a few power guidelines for consideration:

- Fewer larger DIMMs (for example 6 x 4GB DIMMs vs. 12 x 2GB DIMMs will generally have lower power requirements
- x8 DIMMs (x8 data width of rank, see section 3.3) will generally draw less power than equivalently sized x4 DIMMs
- Consider BIOS configuration settings (see section 4.2.4)

4.2.3 Reliability

Here are two reliability guidelines for consideration:

- Using fewer, larger DIMMs (for example 6 x 4 GB DIMMs vs. 12 x 2GB DIMMs is generally more reliable
- Xeon 5500 series memory controllers support IBM Chipkill™ memory protection technology with x4 DIMMs (x4 data width of rank; see sect. 3.3), but not with x8 DIMMs

4.2.4 BIOS Configuration Settings

There are a number of BIOS configuration settings on servers using the Xeon 5500 series processors that can also affect memory performance or benchmark results. For example, most platforms allow the option of decreasing the memory clock speed below the supported maximum. This may be useful for power savings but, obviously, decreases memory performance.

Meanwhile, options like Hyper-Threading Technology (formerly known as Simultaneous Multi-Threading) and Turbo Boost Technology can also significantly affect benchmark results. Specific memory configuration settings important to performance include:

BIOS Options	Maximum Performance Setting
Memory Speed	Auto
Memory Channel Mode	Independent
Socket Interleaving	NUMA
Patrol Scrubbing	Disabled
Demand Scrubbing	Enabled
C-States	Enabled
Turbo Mode	Enabled
Thermal Mode	Performance

For More Information

IBM System x Servers	**ibm.com**/systems/x
IBM BladeCenter Servers	**ibm.com**/systems/bladecenter
IBM Standalone Solutions Configuration Tool (SSCT)	**ibm.com**/servers/eserver/xseries/library/configtools.html
IBM Electronic Service Agent	**ibm.com**/support/electronic
IBM ServerProven Program	**ibm.com**/servers/eserver/serverproven/compat/us
IBM Technical Support	**ibm.com**/server/support
IBM Configuration and Options Guide	**ibm.com**/servers/eserver/xseries/cog

Legal Information

IBM Systems and Technology Group
Dept. U2SA
3039 Cornwallis Road
Research Triangle Park, NC 27709

Produced in the USA
March 2009

For a copy of applicable product warranties, write to: Warranty Information, P.O. Box 12195, RTP, NC 27709, Attn: Dept. JDJA/B203. IBM makes no representation or warranty regarding third-party products or services including those designated as ServerProven or ClusterProven. Telephone support may be subject to additional charges. For onsite labor, IBM will attempt to diagnose and resolve the problem remotely before sending a technician.

IBM, the IBM logo, the e-business logo, Active Memory, BladeCenter, iDataPlex, and System x, are trademarks of IBM Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://**ibm.com**/legal/copytrade.shtml.

Intel, Intel logo, and Xeon are trademarks of Intel Corporation in the U.S. and other countries.

SPEC, SPECfp, SPECint, and SPECjbb are registered trademarks of the Standard Performance Evaluation Corporation (SPEC).

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

IBM reserves the right to change specifications or other product information without notice. References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates. IBM PROVIDES THIS PUBLICATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions; therefore, this statement may not apply to you.

This publication may contain links to third party sites that are not under the control of or maintained by IBM. Access to any such third party site is at the user's own risk and IBM is not responsible for the accuracy or reliability of any information, data, opinions, advice or statements made on these sites. IBM provides these links merely as a convenience and the inclusion of such links does not imply an endorsement.

Information in this presentation concerning non-IBM products was obtained from the suppliers of these products, published announcement material or other publicly available sources. IBM has not tested these 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.

Some machines are designed with a power management capability to provide customers with the maximum uptime possible for their systems. In extended thermal conditions, rather than shutdown completely, or fail, these machines automatically reduce the processor frequency to maintain acceptable thermal levels.

MB, GB and TB = 1,000,000, 1,000,000,000 and 1,000,000,000,000 bytes, respectively, when referring to storage capacity. Accessible capacity is less; up to 3GB is used in service partition. Actual storage capacity will vary based upon many factors and may be less than stated.

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will depend on considerations such as the amount of multiprogramming in the user’s job stream, the I/O configuration, the storage configuration and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

Maximum internal hard disk and memory capacities may require the replacement of any standard hard drives and/or memory and the population of all hard disk bays and memory slots with the largest currently supported drives available. When referring to variable speed CD-ROMs, CD-Rs, CD-RWs and DVDs, actual playback speed will vary and is often less than the maximum possible.

XSW03025-USEN-00