Lenovo

Optimizing Memory Performance of Lenovo Servers Based on Intel Xeon E7 v3 Processors

Introduces the architecture of the X6 servers that use these processors

Analyzes the effects the different memory modes and unbalanced memory configurations have on performance

Illustrates the power consumption savings of DDR4 memory over DDR3

Provides best practices to maximize memory performance

Charles Stephan

Alicia Boozer

Sylvester Cash

Abstract

This paper examines the architecture and memory performance of Lenovo System x and Flex System X6 platforms that are based on the Intel Xeon E7-8800 and E7-4800 v3 processors. The performance analysis in this paper covers memory latency, bandwidth, and application performance. In addition, the paper describes performance issues that are related to CPU frequency, memory speed, and population of memory DIMMs. Finally, the paper examines optimal memory configurations and best practices for the Lenovo X6 platforms.

At Lenovo Press, we bring together experts to produce technical publications around topics of importance to you, providing information and best practices for using Lenovo products and solutions to solve IT challenges. Visit us at http://lenovopress.com.

Contents

Introduction

The Intel Xeon E7 v3 series of EX processors represents a *tock* on the familiar Intel *tick-tock* model. As such, the manufacturing process technology remains the same at 22-nanometers, but the E7 v3 series introduces a new micro-architecture. These processors are designed to enable systems to scale from 1-8 processor sockets natively.

The E7 v3 series provides a common building block across the following Lenovo® platforms:

- Lenovo System x3850 X6, a 4U rack server scalable to four processors
- Lenovo System x3950 X6, an 8U rack server scalable to eight processors
- Lenovo Flex System™ X6 Compute Node family that consists of the x480 X6 (scalable to four processors) and the x880 X6 (scalable to eight processors).

Table 1 lists the generational improvements that were introduced in the E7 v3 series, compared to its predecessor.

Table 1 Intel Xeon E7 v3 series generational improvements

Feature	E7 v2 Series	E7 v3 Series
Process Technology	22nm	22nm
Thermal Design Power (TDP)	155W, 130W, 105W	165W, 150W, 140W, 115W
Memory Speed[a]	1066, 1333, 1600MHz (DDR3)	1066, 1333, 1600MHz (DDR3) 1333, 1600, 1866MHz (DDR4)
Cores / Threads	Up to 15 / 30 per socket	Up to 18 / 36 per socket
Last Level Cache Size	Up to 37.5MB	Up to 45MB
Intel QPI Bandwidth	3x Intel QPI v1.1, 8.0GT/s max	3x Intel QPI v1.1, 9.6GT/s max

a. Memory speed depends on memory configuration, processor sku, and memory mode (Independent or Lockstep)

Compared to its predecessor, the E7 v3 series offers greater core counts, support for DDR4 memory, faster QPI link speed, Intel AVX 2.0 including Fused-Multiply Add (FMA) for technical compute workloads, and Transactional Synchronization Extensions (TSX).

System Architecture

This section describes the architecture of the Lenovo platforms that support the E7 v3 series processors. Figure 1 on page 4 shows the block diagram of an E7 v3 CPU.

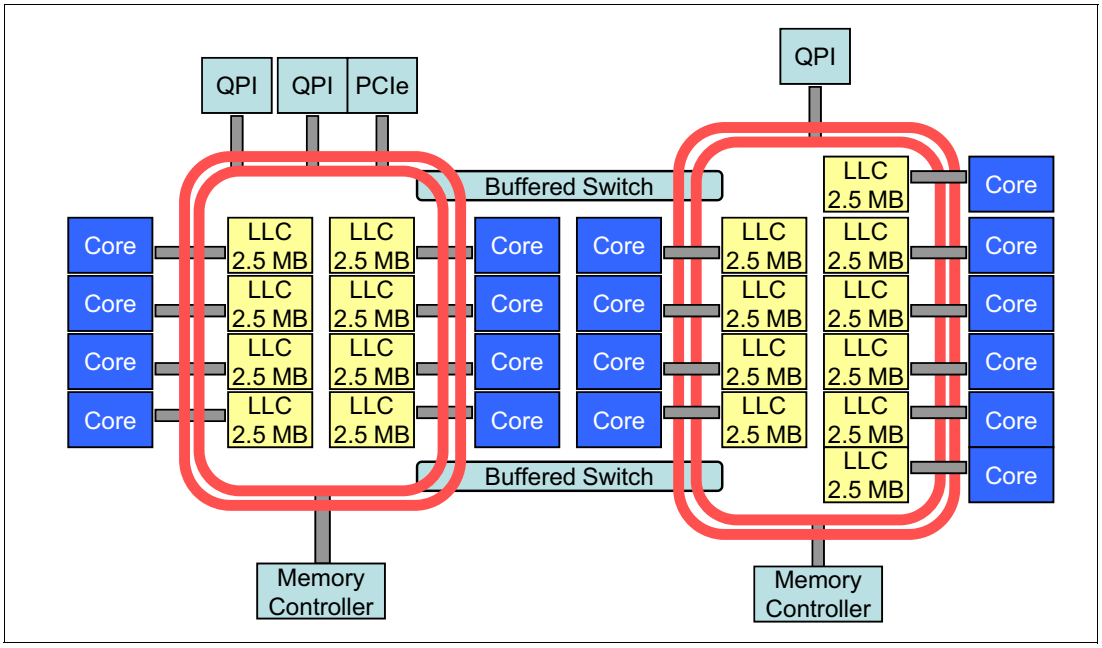

Figure 1 E7-8890 v3 block diagram

The E7 v3 series processors support two integrated memory controllers, four high-speed Scalable Memory Interface 2 (SMI2) links, and eight DDR3 or DDR4 memory channels. Scalable memory buffers provide a bridge between the memory channels and the SMI2 channels. The memory frequency and the Intel Quick Path Interconnect 1.1 (QPI) rate depend on memory population, memory mode, and processor SKU. Supported QPI rates are 9.6 GT/s, 8.0 GT/s, 7.2 GT/s, and 6.4 GT/s.

Lenovo System x3850 X6 and x3950 X6

The Lenovo System x3850 X6 can be configured as a 1-, 2-, or 4-processor system, and supports up to 96 TruDDR4™ or DDR3 memory DIMMs for a maximum capacity of 6 TB (TruDDR4) or 3TB (DDR3). Figure 2 on page 5 shows the block diagram of the x3850 X6 4-socket system architecture.

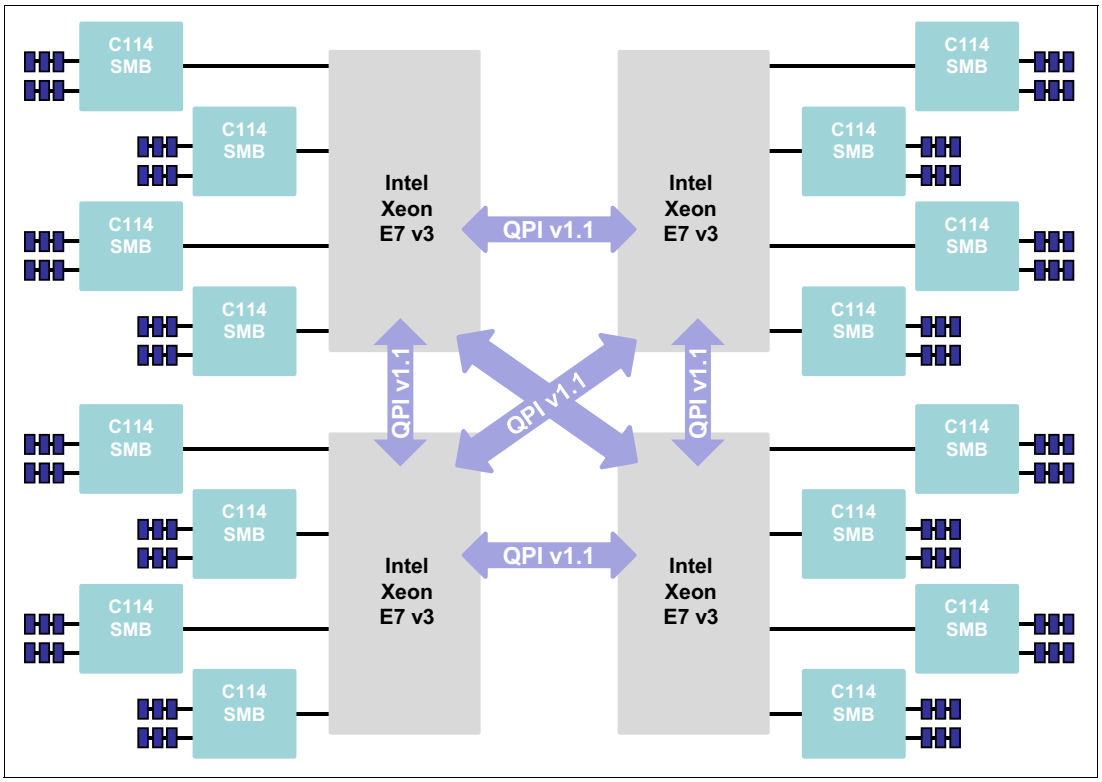

Figure 2 x3850 X6 system architecture block diagram

As shown in Figure 2, each processor is directly connected to one another via QPI v1.1 links. The direct connections provide shorter latencies for processor communication, which can yield lower latencies and higher sustainable memory bandwidths. Not shown are the direct PCIe links that are associated with each socket.

The x3950 X6 consists of two x3850 X6s housed in a single chassis. The scalability links connecting the two are incorporated inside the chassis. The x3950 X6 supports up to 192 TruDDR4 or DDR3 memory DIMMs for a maximum capacity of 12 TB (TruDDR4) or 6 TB (DDR3). The processor and memory DIMMs for each socket are housed in “compute books”. Each compute book contains a processor and up to 24 DIMMs.

The Lenovo System x3850 X6 and x3950 X6 are shown in Figure 3.

Figure 3 Lenovo System x3850 X6 (left) and x3950 X6 (right)

Lenovo Flex System x480 X6 and x880 X6

The E7 v3-based Flex System X6 models are the x480 X6 and x880 X6. These servers are based on a two-socket, double-wide compute node and differ only by the following processor types that are installed and the degree to which they can scale up:

- The Flex System x880 X6 Compute Node uses Intel Xeon E7-**8800** v3 family processors, can scale up to 8-sockets, and supports both an 8-socket configuration (consisting of four connected compute nodes) and 4-sockets (consisting of two connected compute nodes)
- The Flex System x480 X6 Compute Node uses Intel Xeon E7-**4800** v3 family processors, can scale up to 4-sockets, and supports both 2- and 4-socket configurations.

Figure 4 shows the block diagram of the two-socket, double-wide compute node that forms the basis of the Flex System X6 Compute Nodes, and supports up to 48 DDR3 memory DIMMs. Again, the direct PCIe links associated with each socket are not shown.

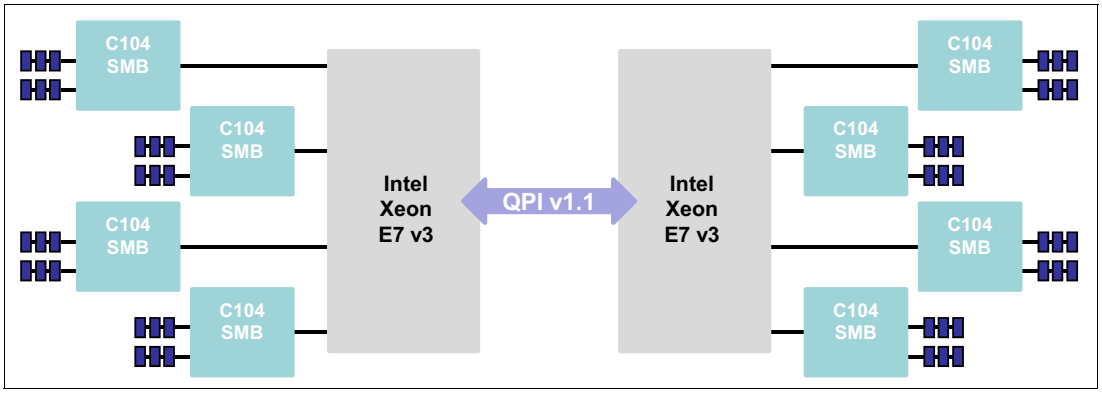

Figure 4 Flex System X6 system architecture block diagram

The compute nodes that comprise the x480 X6 and x880 X6 are connected via a front-mounted interconnect system that joins the QPI links of the processors.

Figure 5 shows on the left, two Flex System x480 X6 compute nodes connected together to form a single-image four-socket server, and on the right, four Flex System x880 X6 compute nodes connected together to form a single-image eight-socket server.

Figure 5 Two x480 X6 compute nodes (left) and four x880 X6 compute nodes (right)

Memory Performance

Memory performance in an Intel Xeon E7 v3-based server depends on different variables, such as memory mode, CPU SKU, memory speed, memory ranks, and memory population.

Topics in this section:

Measurement Configuration

The configuration that is used for the performance data in this paper is listed in Table 2.

Table 2 Performance measurements configuration

Component	Description
System	Lenovo x3850 X6
Processor	TwoE7-8890 v3 (2.5GHz, 18 cores, 165W, QPI 9.6GT/s)
Memory	8GB (2133MHz, 1Rx4, 1.20V) RDIMM 16GB (2133MHz, 2Rx4, 1.20V) RDIMM 64GB (2133MHz, 4Rx4, 1.20V) LRDIMM

Component	Description
UEFI Settings	Operating Mode: Maximum Performance C-states: Disabled P-states: Disabled C1E: Disabled Turbo Mode: Enabled Hyper-Threading: Enabled
Operating System	Red Hat Enterprise Linux 6 Update 6 x64 Edition

To measure low-level memory performance metrics, both an internal Lenovo memory tool and the Intel Memory Latency Checker (MLC) were used to accurately measure memory throughput and memory latency. In all of the memory latency figures, lower numbers are better; in all memory throughput figures, higher numbers are better.

The following industry standard applications were also used to measure memory performance:

- SPECint2006_rate_base

  Used as an indicator of performance for commercial applications. This application often is more sensitive to processor frequency and less to memory bandwidth.

- SPECfp2006_rate_base

  Used as an indicator of High Performance Computing (HPC) performance. This application often is more sensitive to memory bandwidth.

- STREAM

  A benchmark that consists of four different memory workloads; however, the data in this paper corresponds to the Triad component. The Triad component of STREAM consists of two read operations and one write operation from the application's perspective.

This paper includes information that is primarily focused on optimal memory performance. For more information about the Lenovo X6 products, see the following Lenovo Press publications:

- *Lenovo System x3850 X6 and x3950 X6 Planning and Implementation Guide*

  http://lenovopress.com/sg248208

- *Lenovo Flex System X6 Compute Node Planning and Implementation Guide*

  http://lenovopress.com/sg248227

Memory Modes

The E7 v3 memory controllers support Lockstep and Independent memory modes. Within each of those modes, Mirroring and Sparing are supported. There can be performance ramifications, depending on the memory mode selected.

Topics in this section:

The memory mode settings in the Unified Extensible Firmware Interface (UEFI) are shown in Figure 6 on page 9.

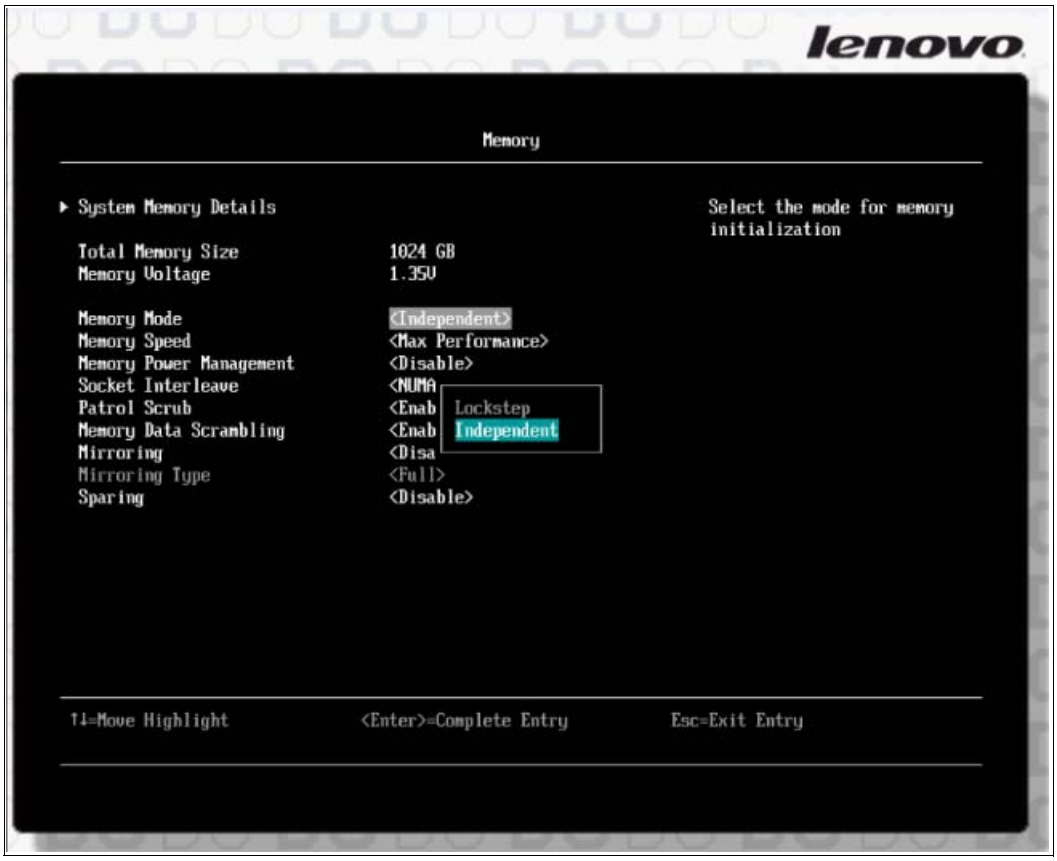

Figure 6 Memory Mode setting in UEFI setup

Independent Memory Mode

In Independent mode, each memory channel is addressed individually. The SMI2 link interleaves the data from two memory channels and the scalable memory buffer separates the data. In this mode, the SMI2 channel operates at twice the memory channel rates. Memory channels can be populated with DIMMs in any order in Independent mode. All eight memory channels per processor can be populated in any order and have no matching requirements. Independent mode should be selected for best memory performance in most production environments. The maximum DDR4 memory speed for independent channel mode is 1600 MHz; the maximum DDR3 speed is 1333 MHz.

Best practices for optimal memory performance should follow the DIMM installation order that is listed in Table 3 (x3850 X6 and x3950 X6) and Table 4 on page 10 (Flex System X6), regardless of what memory mode is selected.

Table 3 x3850/x3950 X6 Compute Book optimal DIMM installation

Memory configuration	DIMM slots to populate per Compute Book
1 DPC (8x DIMMs)	Slots 9,15,6,24,1,19,10,16
2 DPC (16x DIMMs)	Slots for 1 DPC (as shown above) + slots 8,14,5,23,2,20,11,17
3 DPC (24x DIMMs)	Slots for 1 DPC and 2 DPC + slots 7,13,4,22,3,21,12,18

Table 4 Flex System X6 Compute Node optimal DIMM installation

Memory configuration	DIMM slots to populate per Compute Node
1 DPC (16x DIMMs)	Slots 25,28,45,48,7,10,15,18,1,4,21,24,33,36,37,40
2 DPC (32x DIMMs)	Slots for 1 DPC + slots 26,29,44,47,8,11,14,17,2,5,20,23,32,35,38,41
3 DPC (48x DIMMs)	Slots for 1 DPC and 2 DPC + slots 27,30,43,46,9,12,13,16,3,6,19,22,31, 34,39,42

Lockstep Memory Mode

In Lockstep mode, the memory controller uses two memory channels at the same time behind a single memory buffer, which splits a cache line across both channels. In this mode, the SMI2 channel operates at the memory channel rate. Lockstep mode provides the highest reliability, availability, and serviceability (RAS) features. Paired channels should have the same configuration. Lockstep memory mode does not yield the best memory performance for the system in most cases.

Memory Mirroring

Mirroring mode is supported in Independent and Lockstep modes. When Independent mode is selected, memory channel 0 on SMI2 link 0 is mirrored with memory channel 0 on SMI2 link 1, and memory channel 1 on SMI2 link 0 is mirrored with memory channel 1 on SMI2 link 1. The same pattern is true for SMI2 links 2 and 3 on the second memory controller. In Independent mode, the memory channels operate independently from one another.

When Lockstep mode is selected, memory channels 0 and 1 on SMI2 link 0 are mirrored with memory channels 0 and 1 on SMI2 link 1. This pattern is also true for the memory channels on SMI2 links 2 and 3. In Lockstep mode, the memory channels operate together in Lockstep.

Regardless of Independent or Lockstep mode, the total available memory to the system is half of the physical memory that is installed. When Mirrored Memory mode is selected, memory on CPU socket 0 and CPU socket 1 must be populated with DIMMs that all have the same feature set, such as size, organization, and ranks. The memory channels can have memory with different feature sets, but the same DIMM slots across CPU sockets must be populated with memory DIMMs that have the same feature set.

Rank Sparing

As with Mirrored mode, Rank Sparing mode can be implemented with Independent and Lockstep modes. Used with Independent mode, each memory bus has its own spare rank. However, in Lockstep mode, a rank-pair is selected across the memory busses off an SMI2 link.

The spare rank or rank-pair is held in reserve and is not used by the system as part of its system memory. The memory sparing algorithm allocates a spare rank of memory from another of the installed memory DIMMs or DIMM-pair to use as active memory if a certain threshold of correctable errors occurs on a rank or rank-pair of memory. The spare rank must have identical or larger memory capacity than all the other ranks on the same channel. After sparing, the sparing source rank is lost.

Memory Mode Effect on Memory Performance

The loaded latency performance as a function of Lockstep and Independent modes is shown in Figure 7.

About the charts: The charts throughout this paper show the relative performance between two or more things, such as memory latency, memory throughput, benchmark score, etc. Therefore, the scales on the axes do not have units associated with them. For memory latency charts, lower is better. For memory throughput charts, higher is better.

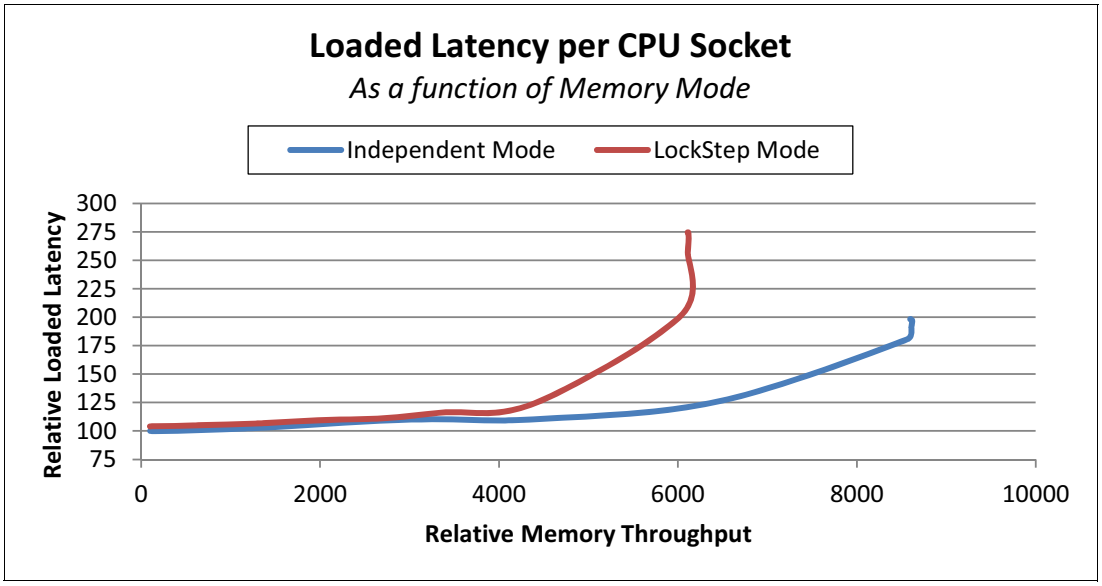

Figure 7 Loaded latency per CPU socket as a function of memory mode

On the left side of Figure 7 at light loads, Lockstep mode yields slightly higher latencies than Independent mode. However, as the load is increased from left to right in Figure 7, Independent mode begins to deliver lower latencies and greater memory throughput than Lockstep mode for moderate to heavy loads.

Figure 8 on page 12 shows relative application performance as a function of memory modes and memory speeds.

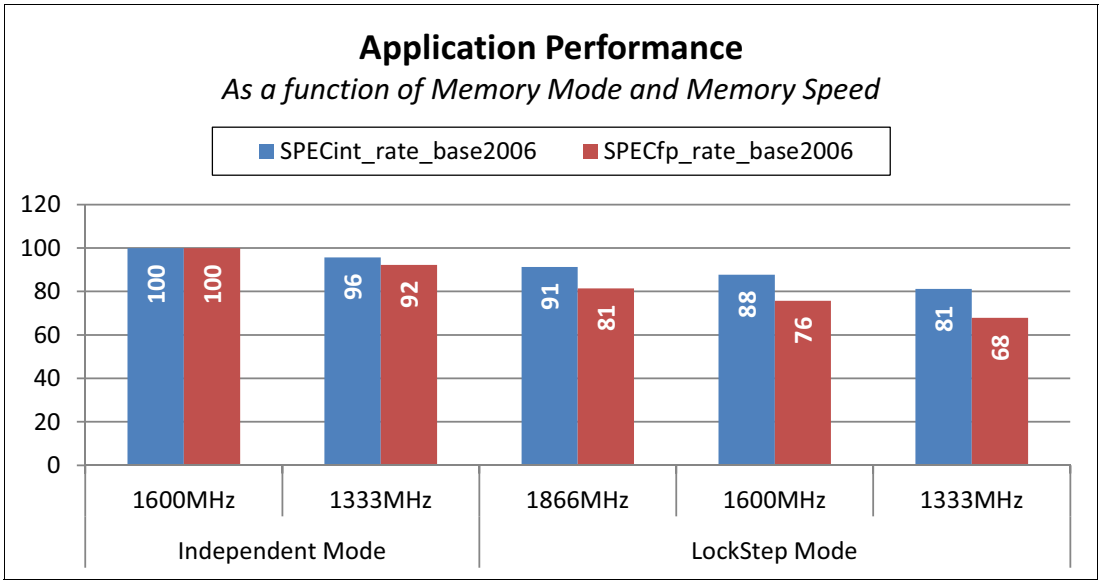

Figure 8 Application performance as a function of memory mode and memory speed

In Figure 8, the SPECint_rate_base2006 results are relative to themselves only, and the SPECfp_rate_base2006 results are relative to themselves. There are three main points we can glean from Figure 8:

- The best overall application performance is achieved in Independent mode for both integer and floating point.
- At like memory speeds, but different memory modes, the integer performance is 13.5% to 18.5% better in Independent mode, and the floating point performance is 31.5% to 35% better in Independent mode.
- As memory speed decreases, the change in performance for floating point operations is more drastic than that of integer operations, because floating point operations are more sensitive to memory bandwidth than processor frequency, while integer operations are more sensitive to processor frequency.

Figure 9 on page 13 shows the relative application performance for Memory Mirroring and Rank Sparing.

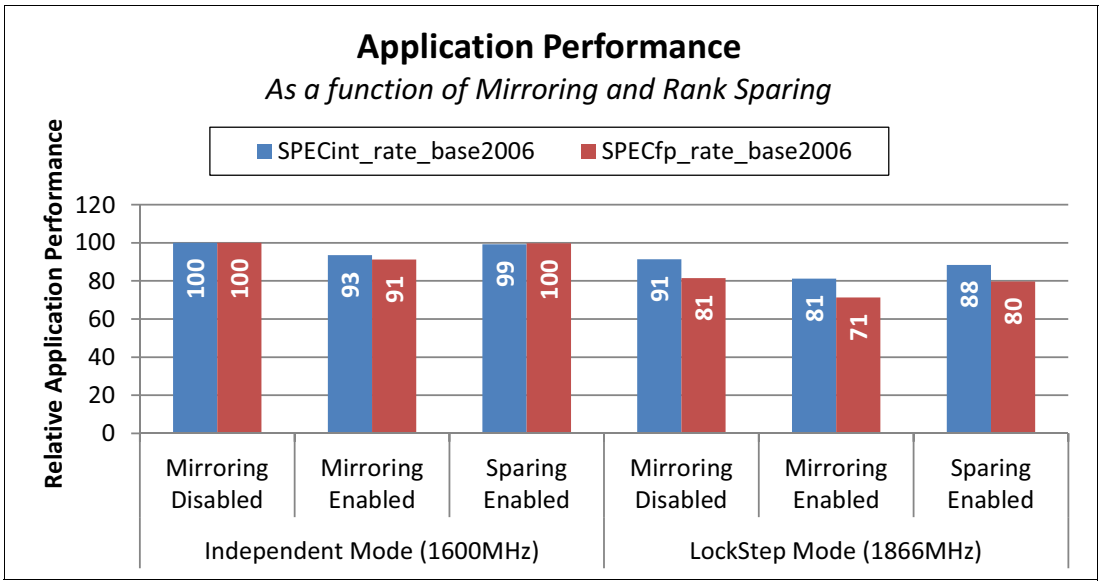

Figure 9 Application performance as a function of memory mirroring and rank sparing

With mirroring enabled, there is a 7% to 8% degradation in performance in Independent mode, and an 11% to 12% degradation in Lockstep mode. There was very little difference in performance with Sparing enabled.

Figure 10 illustrates the affects of mirroring and sparing on STREAM Triad performance.

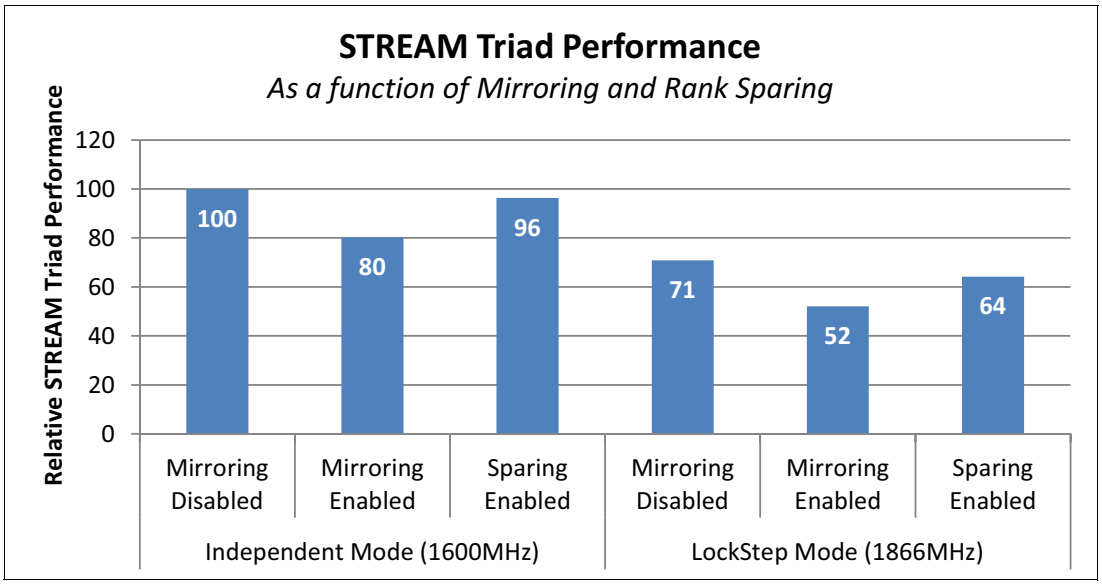

Figure 10 STREAM Triad performance as a function of memory mirroring and rank sparing

Comparing Figure 10 to Figure 9, it is evident the penalty for enabling mirroring is larger with regards to STREAM. The three applications used for these two figures consist of different workloads that exercise the memory subsystem differently. This is an excellent example of how performance can vary depending on the application environment.

Memory Speed

Memory speed is one of the most critical factors that affects memory performance. It is important to understand the performance characteristics when memory speed is changed, as well as the factors that control memory speed in a particular architecture.

Factors Controlling Memory Speed

Memory speed is controlled by the following factors:

- “Processor Type” on page 14
- “DIMM Frequency” on page 14
- “System Settings” on page 15

Processor Type

The E7 v3 series of processors includes 4- and 8-socket-capable processors that are designated as Advanced, Standard, Basic, and Segment Optimized. Table 5 lists the categories and attributes of each processor.

Table 5 Intel Xeon E7 v3 processor categories and attributes

Category	Processor	TDP	Core frequency	Core count	Cache size	QPI Speed	Maximum Memory Frequency (DDR4/DDR3)
Advanced	E7-8890 v3	165 W	2.5 GHz	18	45 MB	9.6 GT/s	1866/1600 MHz
	E7-8880 v3	150 W	2.3 GHz	18	45 MB	9.6 GT/s	1866/1600 MHz
	E7-8870 v3	140 W	2.1 GHz	18	45 MB	9.6 GT/s	1866/1600 MHz
	E7-8860 v3	140 W	2.2 GHz	16	40 MB	9.6 GT/s	1866/1600 MHz
Standard	E7-4850 v3	115 W	2.2 GHz	14	35 MB	8.0 GT/s	1866/1600 MHz
	E7-4830 v3	115 W	2.1 GHz	12	30 MB	8.0 GT/s	1866/1600 MHz
Basic	E7-4820 v3	115 W	1.9 GHz	10	25 MB	6.4 GT/s	1866/1333 MHz
	E7-4809 v3	115 W	2.0 GHz	8	20 MB	6.4 GT/s	1866/1333 MHz
Segment Optimized	E7-8891 v3	165 W	2.8 GHz	10	45 MB	9.6 GT/s	1866/1600 MHz
	E7-8893 v3	140 W	3.2 GHz	4	45 MB	9.6 GT/s	1866/1600 MHz
	E7-8880L v3	115 W	2.0 GHz	18	45 MB	9.6 GT/s	1866/1600 MHz
	E7-8867 v3	165 W	2.5 GHz	16	45 MB	9.6 GT/s	1866/1600 MHz

DIMM Frequency

The DIMM option is another factor that controls the maximum memory frequency. The TruDDR4 DIMMs that are available for the E7 v3 based Lenovo X6 platforms support the following frequencies:

- 1866 MHz
- 1600 MHz
- 1333 MHz

The maximum memory frequency is the lower of the DIMM frequency and the maximum frequency supported by the processor.

Table 6 lists the maximum supported memory frequencies for Independent mode.

Table 6 Supported maximum memory frequencies for Independent mode

DIMM by Rank, Type, Technology	DIMM capacity	Independent Mode Speed / Voltage supported by DIMMs per Channel		
		1DPC 1.2V	2DPC 1.2V	3DPC 1.2V
1Rx4, RDIMM, 4Gb	8GB	1600 MHz	1600 MHz	1600 MHz
2Rx4, RDIMM, 4Gb	16GB	1600 MHz	1600 MHz	1600 MHz
2Rx4, RDIMM, 8Gb	32GB	1600 MHz	1600 MHz	1600 MHz
4Rx4, LRDIMM, 8Gb	64GB	1600 MHz	1600 MHz	1600 MHz
8Rx4, LRDIMM, 8Gb	128GB	1600 MHz	1600 MHz	1600 MHz

Table 7 lists the maximum supported memory frequencies for Lockstep mode.

Table 7 Supported maximum memory frequencies for Lockstep mode

DIMM by Rank, Type, Technology	DIMM capacity	Lockstep Mode Speed / Voltage supported by DIMMs per Channel		
		1DPC 1.2V	2DPC 1.2V	3DPC 1.2V
1Rx4, RDIMM, 4Gb	8GB	1866 MHz	1866 MHz	1600 MHz
2Rx4, RDIMM, 4Gb	16GB	1866 MHz	1866 MHz	1600 MHz
2Rx4, RDIMM, 8Gb	32GB	1866 MHz	1866 MHz	1600 MHz
4Rx4, LRDIMM, 8Gb	64GB	1866 MHz	1866 MHz	1600 MHz
8Rx4, LRDIMM, 8Gb	128GB	1866 MHz	1866 MHz	1600 MHz

System Settings

Memory can be set to a frequency lower than the platform maximum by clicking **System Settings** → **Memory** → **Memory Speed** in the system's UEFI shell. Memory frequency often is set lower to save energy in environments with little memory performance sensitivity.

Figure 11 on page 16 shows the UEFI shell window with the Memory Speed setting, which includes the options Minimal Power, Balanced, or Max Performance.

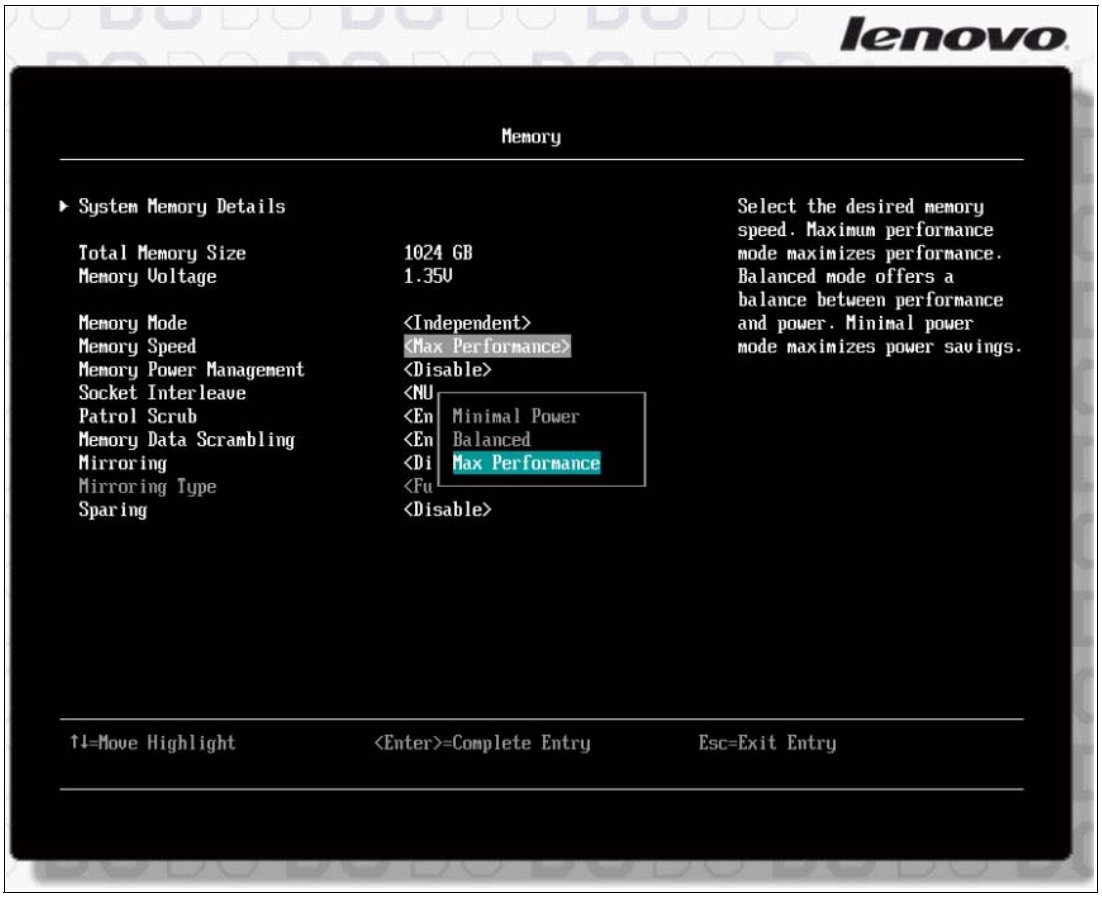

Figure 11 Memory speed setting in UEFI setup

The following Memory Speed options are available:

- Maximum Performance

  Memory runs at the *maximum* speed as determined by processor SKU and the memory subsystem. Memory voltage is forced to the *minimum* voltage that is needed to run at the maximum speed.

- Balanced

  Memory runs at *one step below* the maximum speed. Memory voltage is always set to the *lowest* supported value.

- Minimal Power

  Memory runs at the *lowest* speed allowed by the architecture. Memory voltage is always set to the *lowest* supported value.

Processor and Memory Speed Effects on Memory Performance

This section describes the effects on memory performance according to processor frequency and memory frequency. Figure 12 on page 17 shows unloaded memory latency as a function of processor and memory frequency in Independent mode.

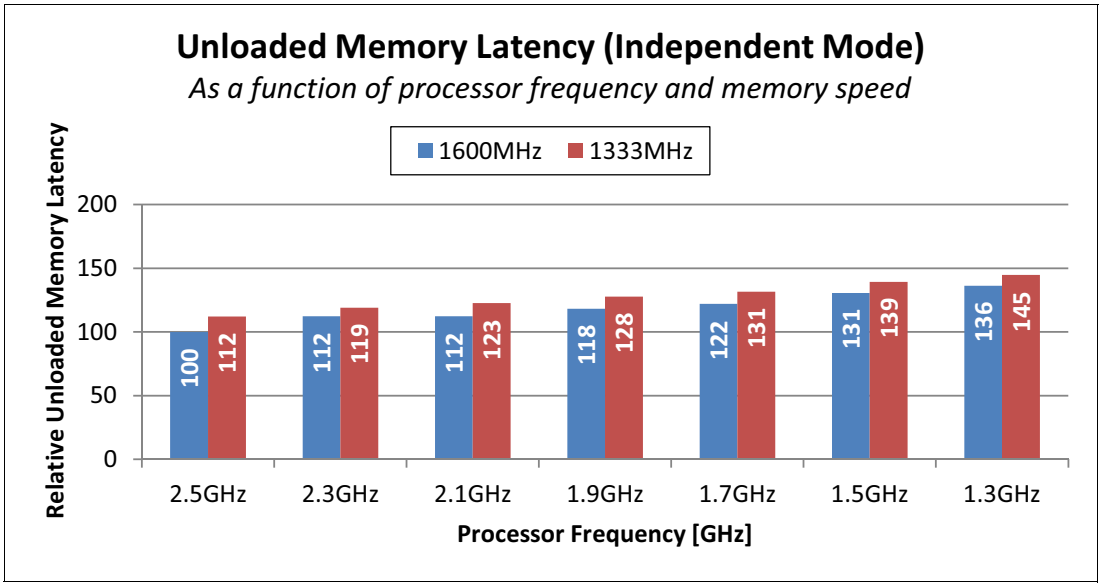

Figure 12 Unloaded memory latency as a function of processor frequency and memory speed

As the processor core frequency decreases from left to right (as shown in Figure 12), the memory latency increases significantly. Furthermore, decreasing the memory speed can increase the memory latency as well.

Processor SKU Effects on Loaded Latency and Applications

This section illustrates how processor SKU selection can affect performance. Three different processors from Table 5 on page 14 were selected; one each from the Advanced, Standard, and Basic segments. The three SKUs selected were the E7-8890 v3, E7-4850 v3, and the E7-4820 v3. These processors differ by maximum memory speed supported, number of cores, processor frequency, maximum supported QPI rate, and Turbo support.

Figure 13 shows the loaded latency of the three processor SKUs relative to the E7-8890 v3.

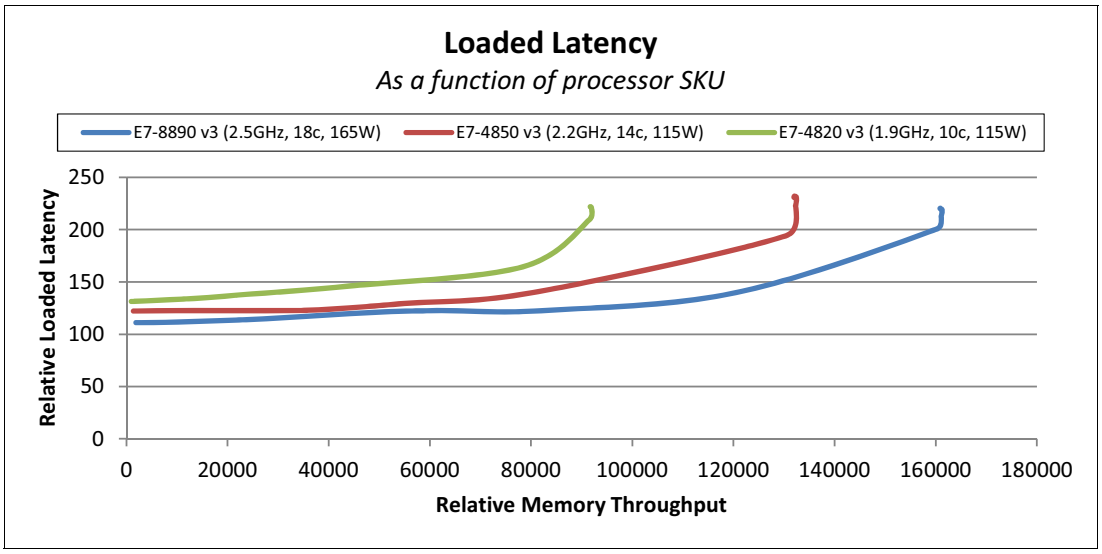

Figure 13 Loaded latency as a function of processor SKU

Clearly, on the far left where the load is lightest the memory latency is greatest for the E7-4820 v3, which has the lowest processor frequency and does not support Turbo. The trend shown in Figure 12 on page 17, which was the latency increased as processor frequency and memory speed decreased, is also illustrated in Figure 13 on page 17; albeit other factors are in play as well.

Figure 14 illustrates the effect processor SKU can have on different applications.

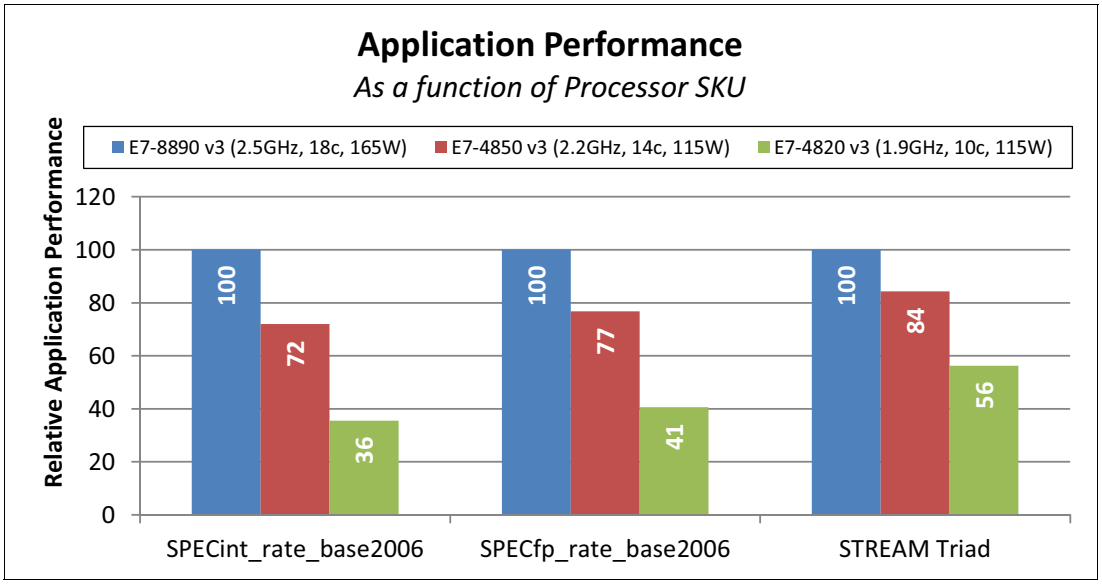

Figure 14 Application performance as a function of processor SKU

Again, factors such as processor frequency, memory speed, number of cores, and QPI rate all play a role in the relative performance between the three SKUs.

DIMM Types

The E7 v3-based System x® and Flex System servers support either RDIMMs or LRDIMMs. The performance implications of using either type of DIMM are as follows.

- Registered DIMMs

  Registered DIMMs (RDIMMs) are the most prevalent DIMMs used in servers.

  RDIMMs use a register between the memory controller and dynamic random-access memory (DRAM) devices to buffer the address and control signals which enable the reduction of electrical loading on the memory bus. This configuration allows the memory controller to support more DIMMs and a higher memory frequency, which provides scalability and greater performance.

- Load Reduced DIMMs

  Load reduced DIMMs (LRDIMMs) reduce the electrical loading on the memory bus while maintaining larger capacities than RDIMMs. The register used by RDIMMs is replaced with a buffer on LRDIMMS, which isolates address, command, and data signals from the memory controller.

  LRDIMMs use a technique called *rank multiplication* to work around the chip select limitation of 8 ranks per DDR3 or DDR4 channel. Rank multiplication presents many ranks on a DIMM as a smaller number of ranks to the memory controller. For example, a quad-rank LRDIMM appears as a dual-rank memory module to the memory controller. This appearance allows LRDIMMs in the system to achieve a larger memory capacity while maintaining high performance, although with a slightly higher latency. LRDIMMs are targeted at memory capacities that cannot be achieved by using RDIMMs.

Effects of DIMM Type on Loaded Memory Latency and Bandwidth

Figure 15 shows the relative loaded memory latency and bandwidth of RDIMMs and LRDIMMs.

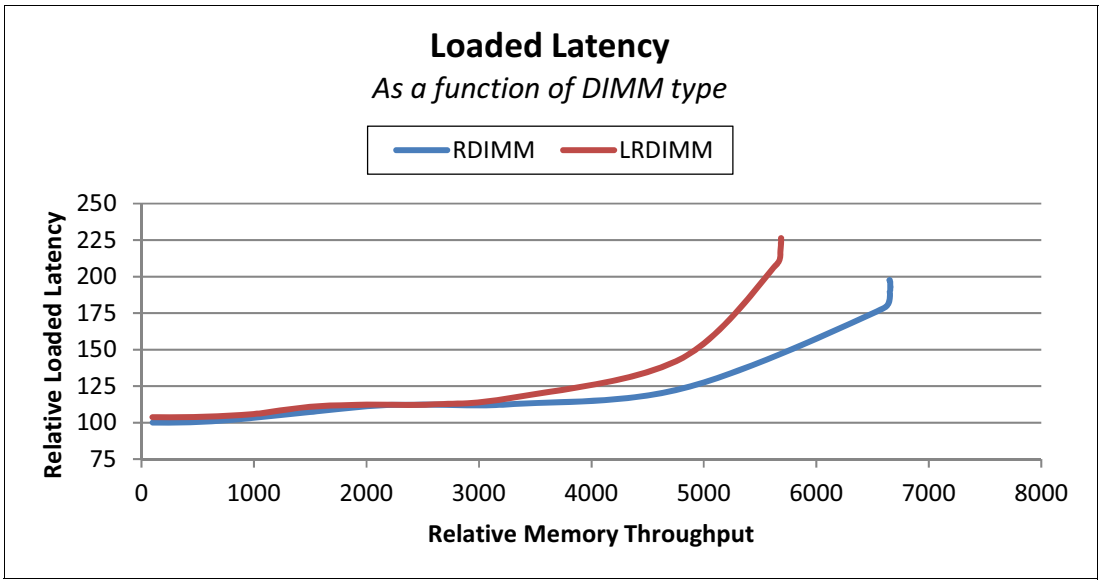

Figure 15 Loaded latency as a function of DIMM type

At light to moderate loads, the LRDIMMs loaded latency is only a few percent longer than RDIMMs. However, at heavier loads, the loaded latency of LRDIMMs is significantly longer. Therefore, the relative memory throughput of LRDIMMs is considerably less than that of RDIMMs at the heaviest loads. LRDIMMs are best suited for environments that require a

large memory footprint. For light to moderate loads, the performance difference between RDIMMs and LRDIMMs is essentially negligible, and therefore RDIMMs are preferable, due to the higher throughput and (typically) lower-cost.

Ranks per Channel

The number of ranks per channel (RPC) affects memory performance to a certain extent. In this section, we describe how this process works. The E7 v3 series with Scalable Memory Buffers, either C114 or C104, support up to 8 ranks per channel.

The following DIMM configurations at 1600MHz were used in this section:

- 1 RPC: 1x 8 GB 1Rx4
- 2 RPC: 1x 16 GB 2Rx4 AND 2x 8 GB 1Rx4
- 3 RPC: 1x 16 GB 2Rx4 + 1x 8 GB 1Rx4 MHz AND 3x 8 GB 1Rx4
- 4 RPC: 2x 16 GB 2Rx4
- 6 RPC: 3x 16 GB 2Rx4

For the measurement results that are included in the following two sections, the processors were given enough memory that the memory capacity differences among the configurations did not affect the results.

DIMM Ranks and STREAM Triad

Figure 16 shows the STREAM Triad memory bandwidth performance as a function of RPC.

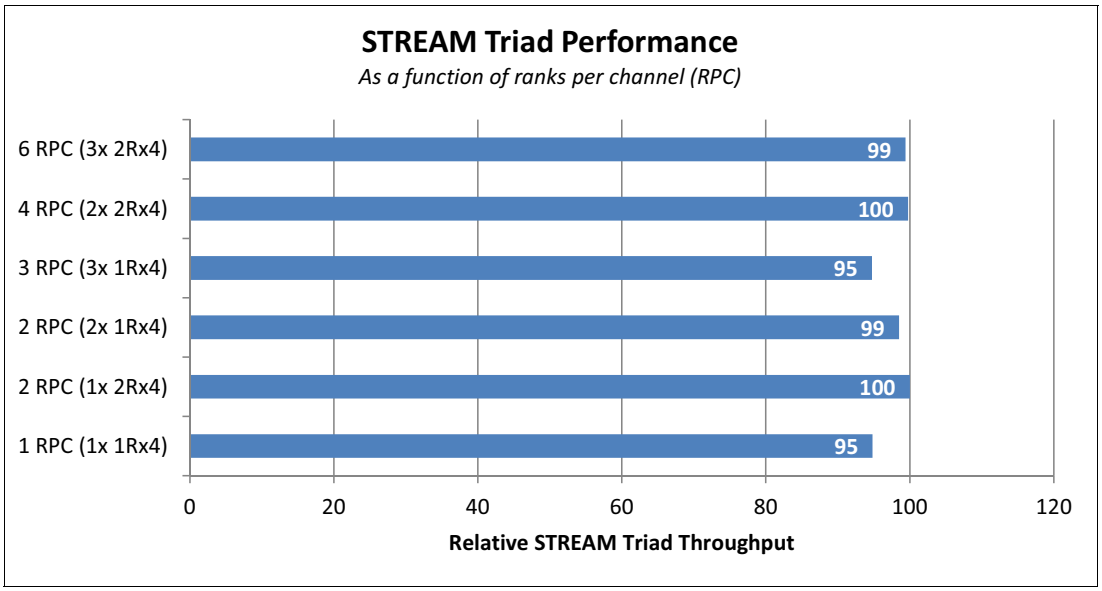

Figure 16 STREAM Triad performance as a function of RPC

Several observations are apparent. First, transitioning from 1RPC to 2RPC gains 5% more memory bandwidth performance. Second, a memory configuration with an odd number of RPC is not optimal. Finally, there is essentially no drop in performance when every memory channel is populated with three RDIMMs, assuming an even number of RPC.

DIMM Ranks and Applications Performance

Figure 17 shows application performance as a function of RPC.

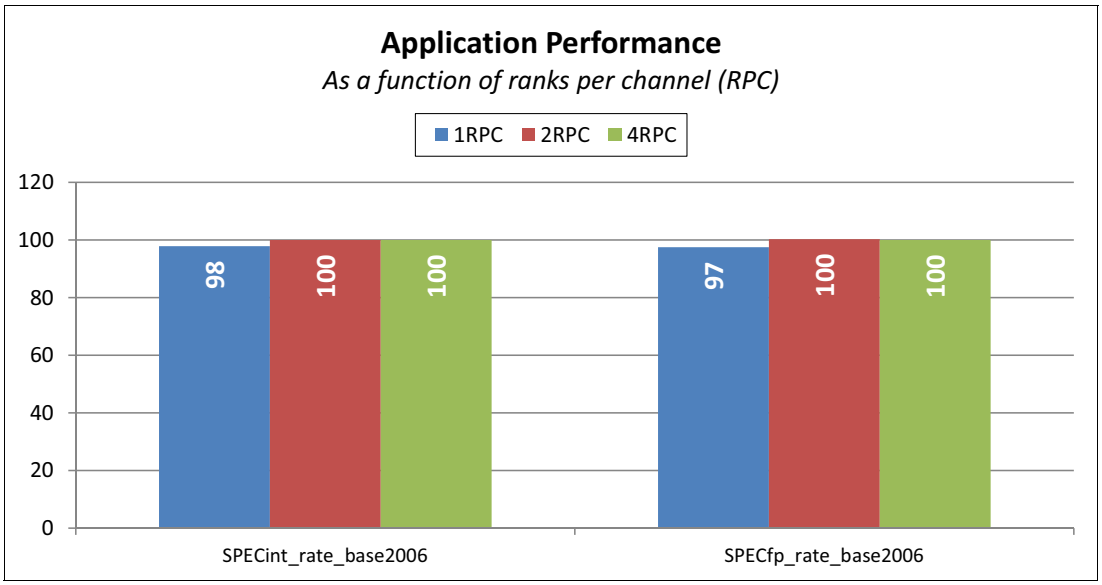

Figure 17 Application performance as a function of RPC

SPECint_rate_base2006 is only slightly less sensitive to RPC than SPECfp_rate_base2006. Again, populating memory channels with an even number of ranks provides optimal performance in integer- and floating-point-sensitive environments.

Memory Population and Balance

Memory interleaving refers to how physical memory is interleaved across the physical DIMMs. A balanced system provides the best interleaving and the best performance. The following rules must be observed for balancing a system for optimized memory performance:

- If all available DIMMs are of the same capacity, distribute the DIMMs such that all memory channels have the same number of DIMMs. Populate memory in groups of eight per compute book for the x3850 X6 and x3950 X6 platforms, and in groups of 16 for the x480 X6 and x880 X6 compute nodes.
- If all available DIMMs are not of equal capacity, balance all eight channels in each compute book with the same amount of memory for the x3850 X6and x3950 X6 platforms. For the Flex System X6 compute nodes, ensure that individual memory channels are populated with the same amount of memory capacity.

It is not uncommon for a system to contain a memory configuration that is poorly interleaved, which can occur for the following reasons:

- Using available DIMMs to reduce parts on the floor.
- Configuring a system based solely on memory capacity requirements. At a minimum, physical memory should be over-provisioned to the closest memory capacity that yields a balanced memory configuration.

Memory Channel Population Effect on STREAM Triad Bandwidth

Leaving memory channels unpopulated affects memory performance. Figure 18 shows how STREAM Triad memory bandwidth is affected, depending upon how many memory channels are populated per socket.

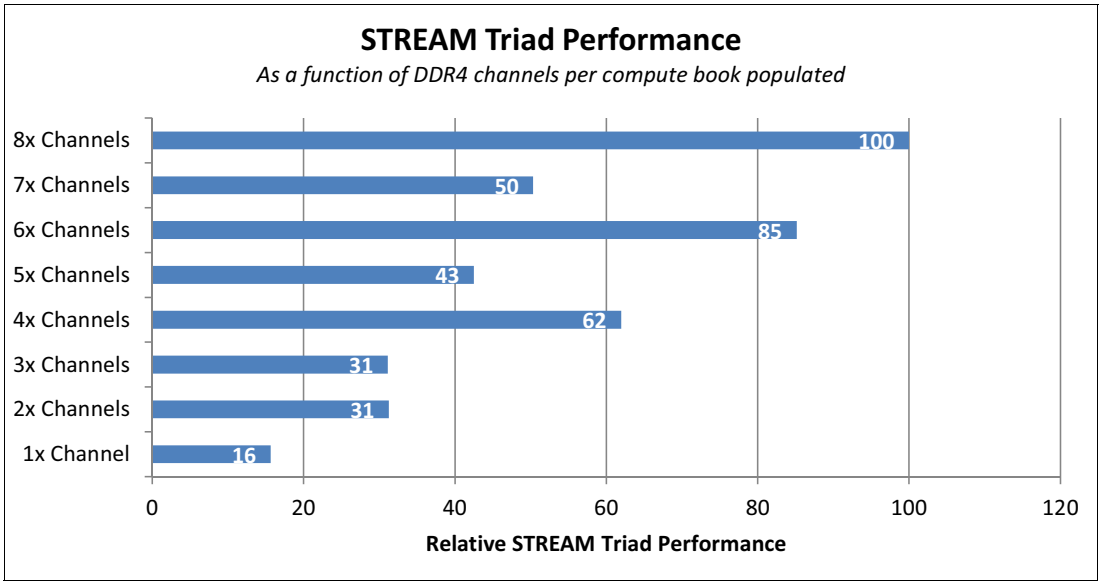

Figure 18 STREAM Triad performance as a function of DDR4 channels populated per compute book

As shown in Figure 18, populating all memory channels per socket yields optimal memory performance. If all of the memory channels cannot be populated, it is best to populate even numbers of memory channels. Populating six memory channels yields 85% of maximum STREAM Triad memory bandwidth. However, populating seven memory channels yields only 50% of maximum expected bandwidth. In fact, populating only four memory channels per socket yields 24% more STREAM Triad bandwidth than populating seven channels.

Memory Balance Effect on Memory Throughput

In this section, several different DIMM configurations were analyzed to show the effect of unbalanced memory configurations on performance. All configurations were run at 1600 MHz.

Figure 19 shows the possible memory performance degradation caused by an unbalanced memory configuration.

The performance numbers (STREAM, Integer and Floating Point) listed in each configuration are relative to Configuration 1 having performance values of 100%.

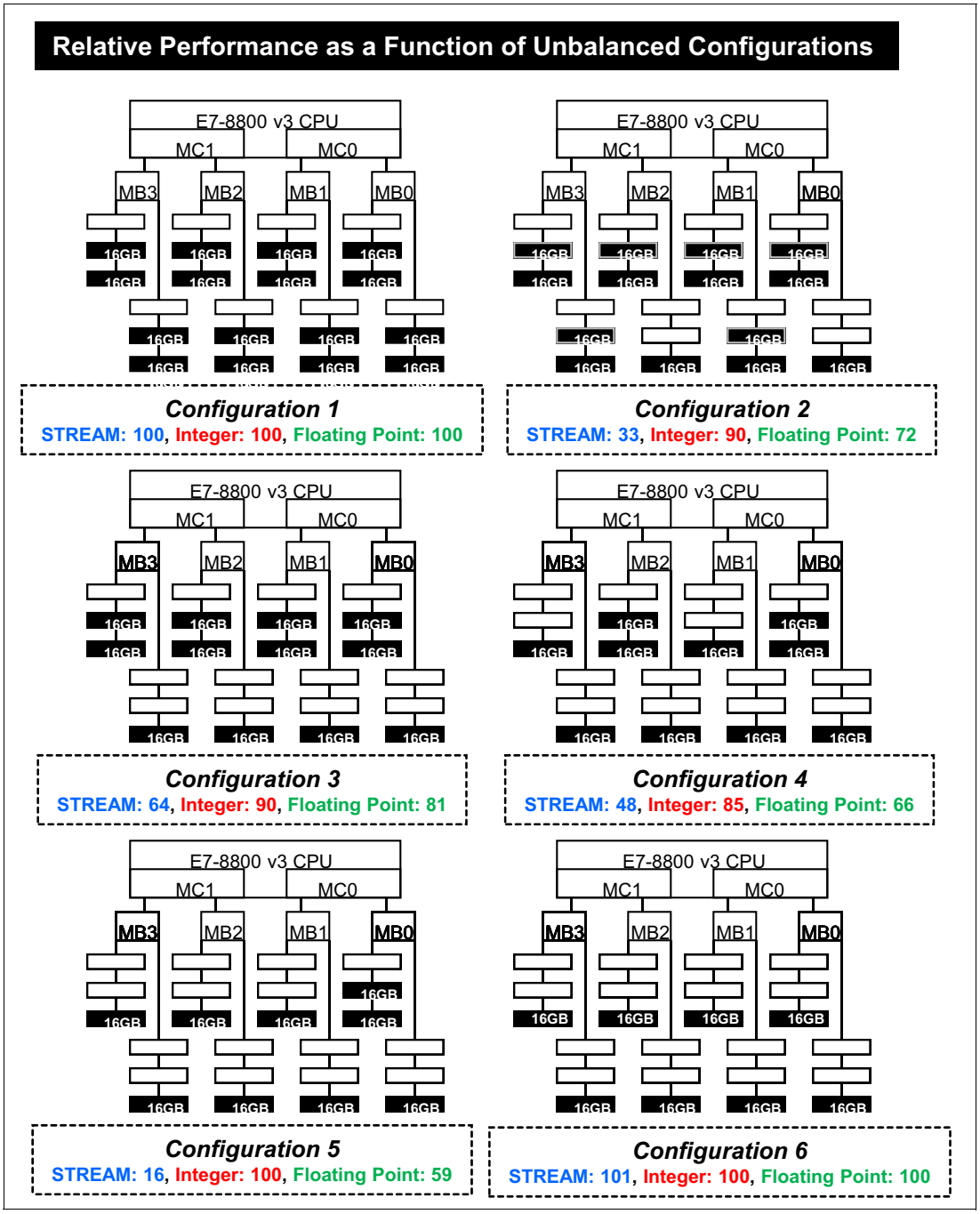

Figure 19 Application performance as a function of unbalanced memory configurations

The capacities of the DIMMs used in this exercise were the same. Configurations 1 and 6 represent balanced memory configurations; therefore, they achieve optimal memory performance. The remaining configurations break the rules for a balanced memory configuration and show the performance relative to the most optimal configuration.

Configuration 2 balances memory across both memory controllers, but contains an imbalance on one memory channel per memory controller. In this case, the maximum achievable STREAM Triad memory throughput is only 33% of optimal. The floating point performance measured with SPECfp_rate_base2006 is 72% of optimal and the integer performance measured with SPECint_rate_base2006 is 90% of optimal.

Using several different benchmark applications to measure the relative performance between the different configurations shows the performance degradation varies depending upon the application environment. The performance values listed below Configurations 3, 4, and 5 illustrate the relative performance of different imbalances across the memory channels.

Different DIMM Capacities and RPC Effect on Memory Throughput

In this section, six different DIMM configurations were analyzed to show the effect of mixing DIMM capacities and RPC, as shown in Figure 20.

Performance numbers (STREAM, Integer and Floating Point) listed in each configuration are relative to Configuration 1 having performance values of 100%.

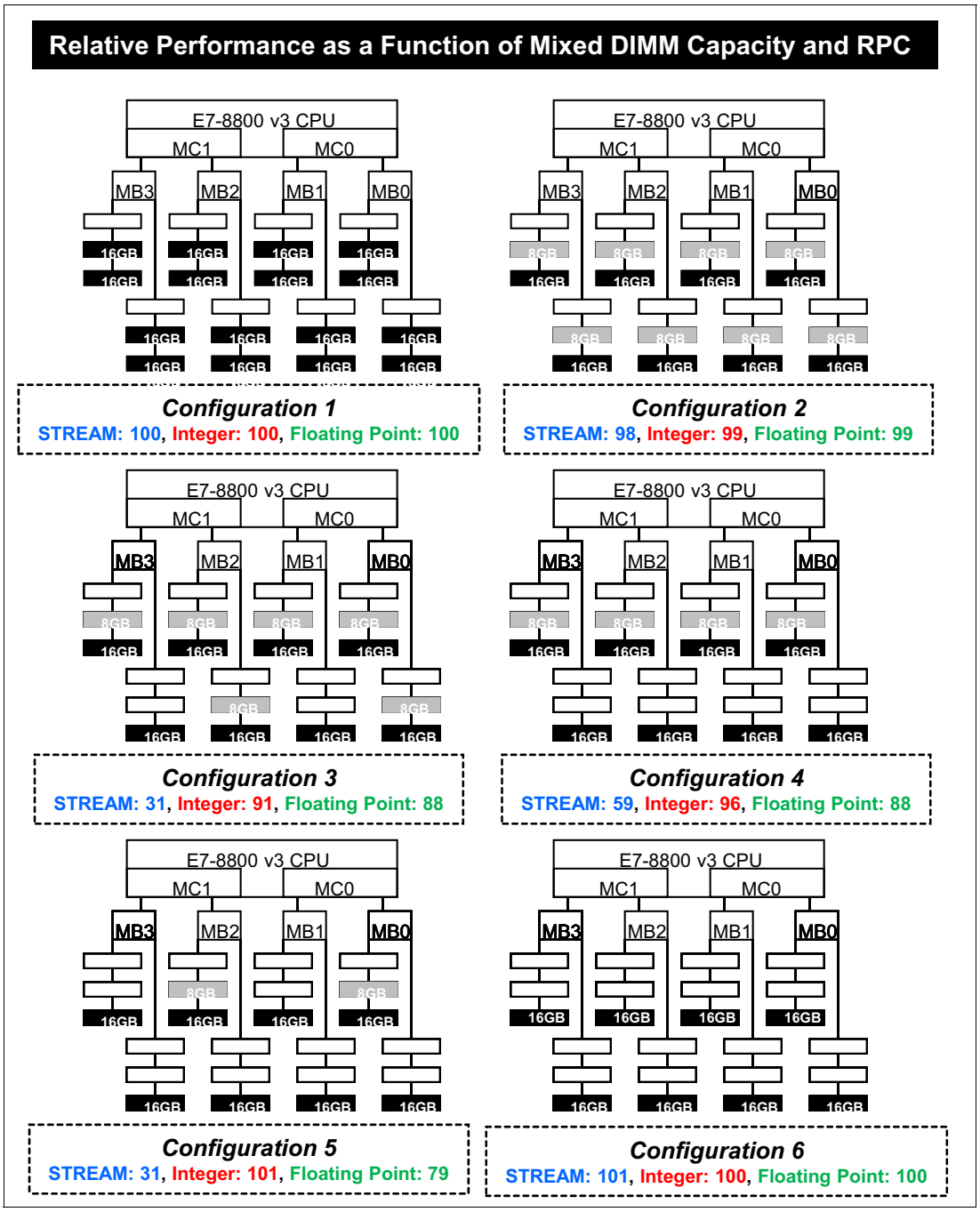

Figure 20 Application performance as a function of mixed DIMM capacity and RPC

Configurations 1 and 6 represent a balanced memory configuration and achieve optimal memory performance. Configuration 2 also is a balanced memory configuration, but mixes 8 GB and 16 GB DIMMs. Each DDR4 channel contains the same capacity, and the

performance degradation across applications is negligible. The performance values listed below Configurations 3, 4, and 5 illustrate the relative performance of different imbalances across the memory channels with respect to different types of applications.

Balancing Memory Population on X6 Platforms

The Lenovo x3850 X6 and x3950 X6 systems support up to four or eight (respectively) compute books that contain one processor socket and 24 DIMM slots. The memory subsystem for these platforms can be balanced by installing DIMMs in groups of eight per compute book (one per memory channel).

The front of the compute book is shown in Figure 21.

Tip: “JC” in the figures stands for “Jordan Creek”, the codename for the scalable memory buffers.

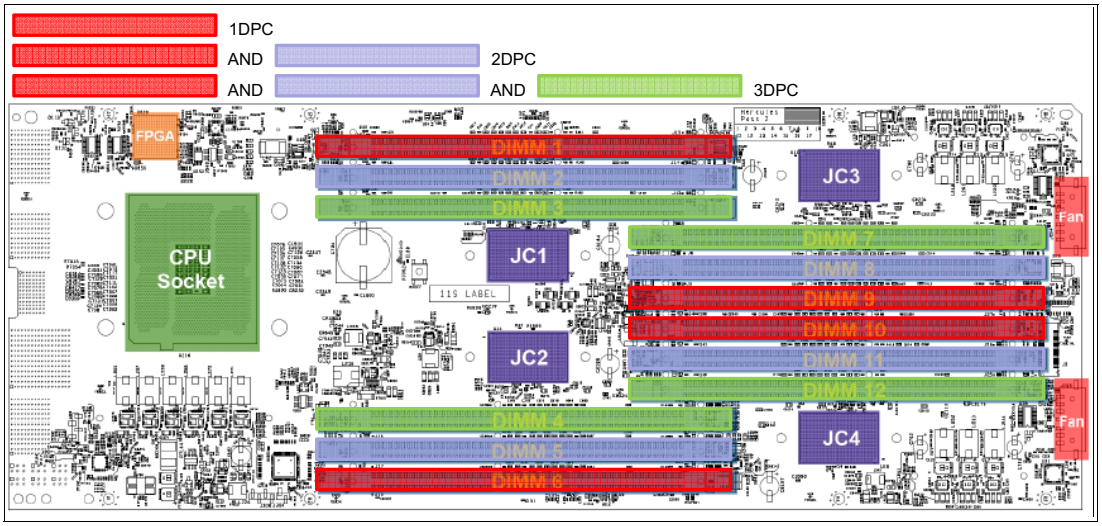

Figure 21 Front of x3850/x3950 X6 compute book

The back of the compute book is shown in Figure 22.

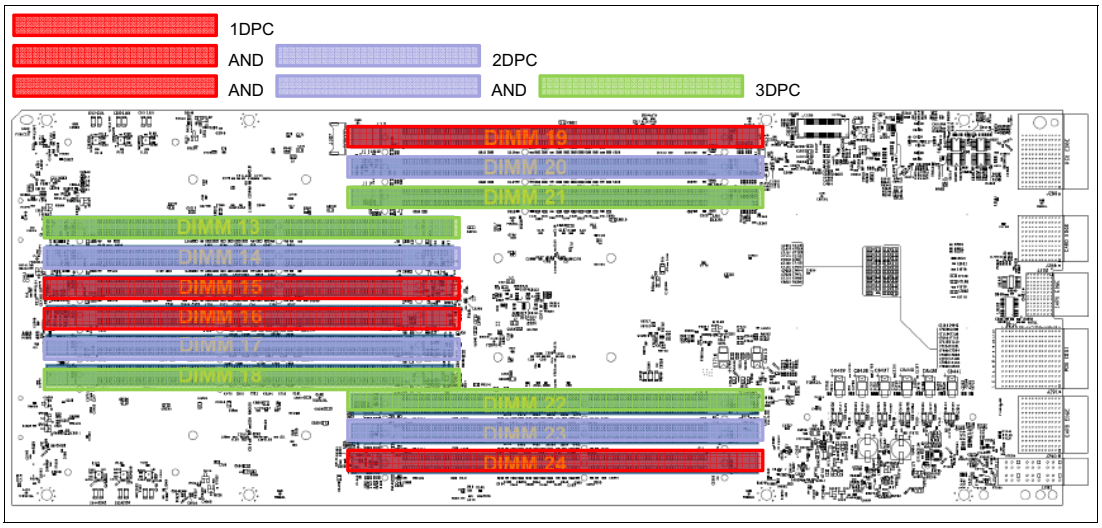

Figure 22 Back of x3850/x3950 X6 compute book

Use the colored legend shown at the top of Figure 21 and Figure 22 to populate the compute books correctly for a balanced memory configuration.

Figure 23 on page 27 shows the DIMM layout of the Flex System X6 compute nodes. Use the colored legend shown at the top of the figure to populate Flex System X6 nodes correctly with a balanced memory configuration.

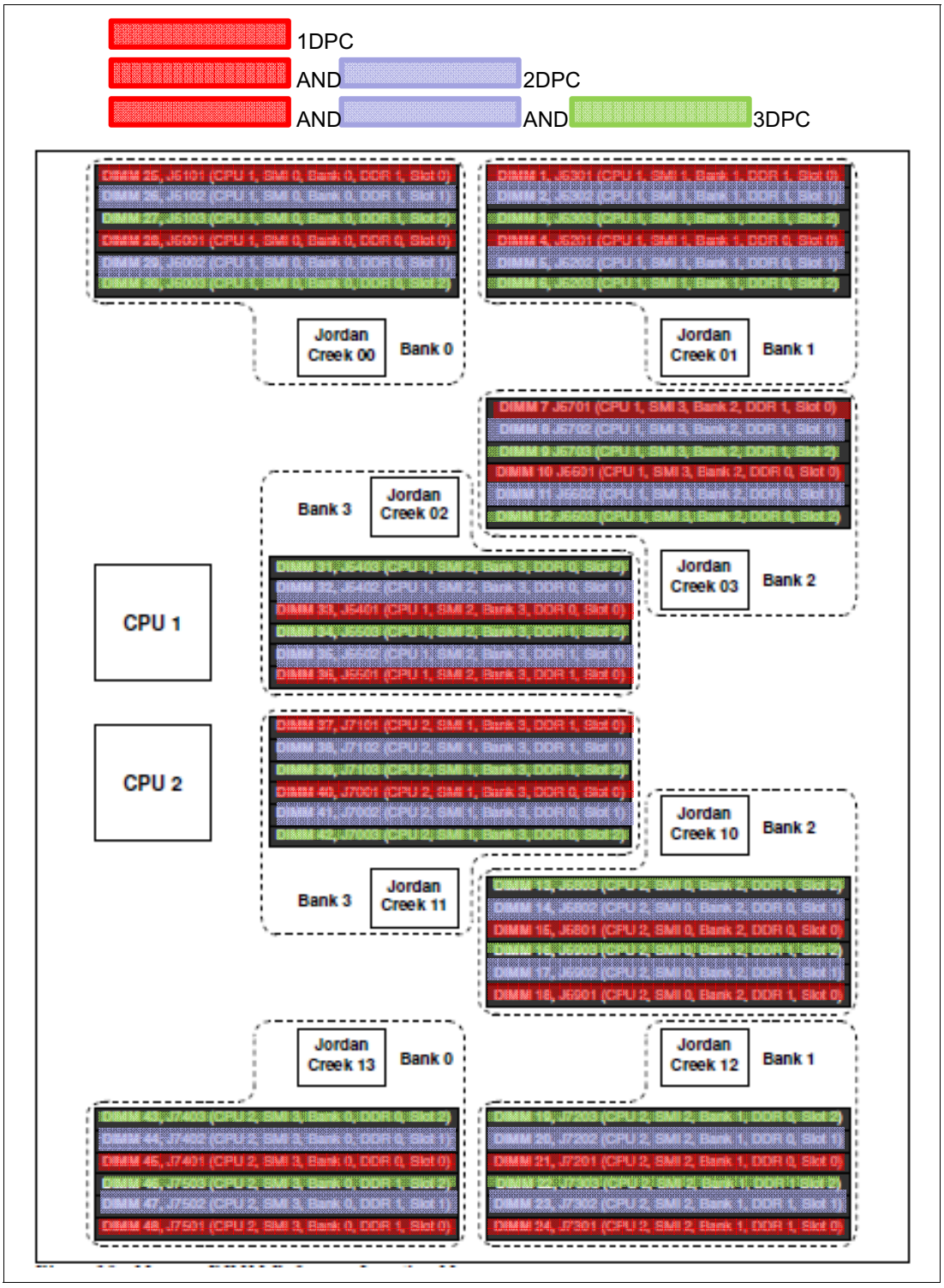

Figure 23 Flex System X6 DIMM layout

Power Consumption

DDR4 DIMMs have a lower voltage requirement than DDR3 DIMMs. The E7 v3 series-based platforms run DDR4 memory at 1.2V, as opposed to DDR3 memory at 1.35V or 1.5V. The lower voltage requirement results in lower power consumption.

Figure 24 illustrates the power consumption savings between DDR4 and DDR3 memory over different memory workloads produced by the Intel Memory Latency Checker.

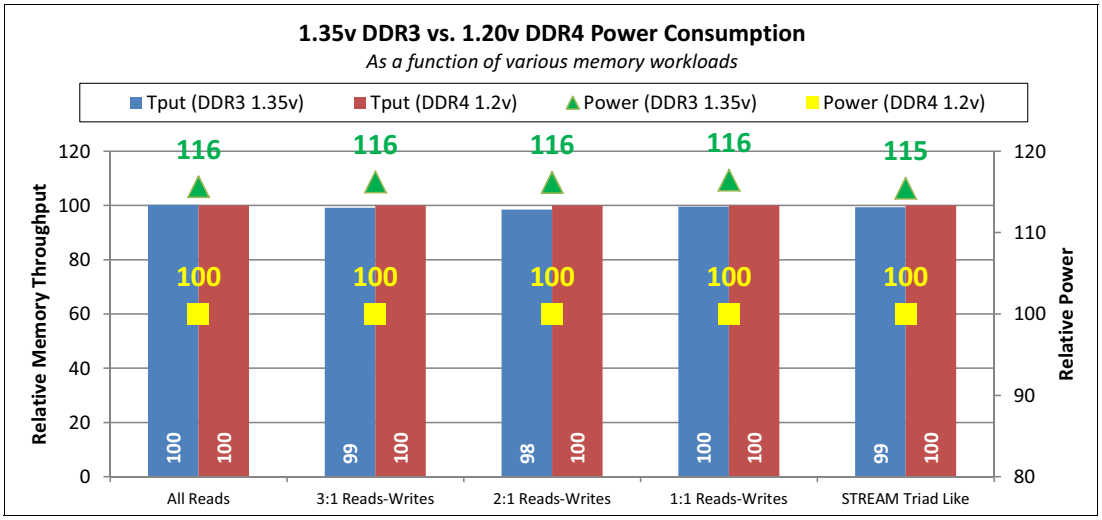

Figure 24 DDR3 vs. DDR4 power consumption as a function of memory workload

The standard SKU E7-4850 v3 processor was used to generate the data for Figure 24, with Independent Mode selected at 1333 MHz memory bus speed for both DDR3 and DR4. In the DDR3 case, PC4-12800 16GB RDIMMs and compute books with Jordan Creek 1 (C104) memory buffer were used. In the TruDDR4 case, PC4-17000 16GB RDIMMs and compute books with SMB C114 memory buffer were used.

The vertical bars represent the relative memory throughput between DDR3 and DDR4 across the five different memory workloads, and are associated with the left vertical axis. As expected, because the memory bus speed and processor remain the same, the memory throughput is essentially identical for each workload for DDR3 and DDR4.

The square and triangle data markers represent the relative power consumption for DDR3 versus DDR4, and are associated with the right vertical axis. Figure 24 clearly shows that the DDR3 power consumption across the five workloads is 15% - 16% more than that of TruDDR4. Additionally, because DDR4 uses 8Gb DRAM technology (versus 4Gb for DDR3), a 32GB DDR4 RDIMM will have additional power savings versus a 32GB DDR3 LRDIMM.

Best Practices

This section recommends memory population best practices for the E7 v3 series-based platforms. Adhere to the following rules for optimal memory performance:

- Always populate all processors with equal memory capacity to ensure a balanced NUMA system.
- Always populate all eight channels on each socket. If this configuration is not possible, populating an even number of channels is preferable to an odd number of channels.
- Always populate all memory channels on each socket with identical DIMMs. If this configuration is not possible, the next best option is to populate all channels with equal memory capacity.
- Use dual-rank *RDIMMs* whenever possible. Use LRDIMMs if memory capacity requirement cannot be achieved with RDIMMs.l
- Always populate memory channels with an even number of ranks whenever possible.

Conclusion

Lenovo systems that use the E7 v3 series processors and DDR4 memory offer significant performance gains over their predecessors. In addition, the use of DDR4 memory options can significantly reduce power consumption. The x3850 and x3950 X6, and Flex System X6 platforms support RDIMM and LRDIMM memory options, which can be optimized for performance and large memory capacity requirements. Although every application has unique characteristics that might not be affected by the scenarios that are described in this paper, adhering to the best practices that are presented here produces a system that is configured for optimal memory performance.

Authors

Charles Stephan is the Technical Lead for the System Performance Verification team in the Lenovo System x and Flex System Performance Laboratory at the Lenovo campus in Morrisville, NC. His team is responsible for analyzing the performance of storage adapters, network adapters, various flash technologies, and complete x86 platforms. Before transitioning to Lenovo, Charles spent 16 years at IBM as a Performance Engineer analyzing storage subsystem performance of RAID adapters, Fibre Channel HBAs, and storage servers for all x86 platforms. He also analyzed the performance of x86 rack systems, blades, and compute nodes. Charles holds a Master of Science degree in Computer Information Systems from the Florida Institute of Technology.

Alicia Boozer is a hardware engineer in the Lenovo System x and Flex System Performance Laboratory at the Lenovo campus in Morrisville, NC. Before starting at Lenovo, Alicia spent 6 years at IBM working in the benchmark area initially, and then transitioned to system performance verification. Her current role includes subsystem analysis for all x86 products and performance validation against functional specifications and vendor targets. Alicia holds Bachelor of Science degrees in Mathematics from Spelman College and Electrical Engineering from North Carolina A&T State University and a Master of Science degree from the Massachusetts Institute of Technology.

Sylvester (Sly) Cash is a hardware engineer in the Lenovo System x and Flex System Performance Laboratory at the Lenovo campus in Morrisville, NC. Before starting at Lenovo, Sly spent 26 years at IBM working in the Networking Hardware and Commercial Desktop divisions before transitioning to the System x performance team. His current role includes subsystem performance analysis, performance/power analysis, and storage system analysis for Lenovo products. Sly holds a Bachelor of Science degree in Electrical Engineering from North Carolina A&T State University and a Master of Science degree in Information Systems from the University of Maryland at Baltimore County (UMBC).

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

- David Watts, Lenovo Press

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service.

Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
1009 Think Place - Building One
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary.

Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk.

Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document was created or updated on April 4, 2016.

Send us your comments via the **Rate & Provide Feedback** form found at
http://lenovopress.com/redp1234

Trademarks

Lenovo, the Lenovo logo, and For Those Who Do are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. These and other Lenovo trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by Lenovo at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of Lenovo trademarks is available on the Web at http://www.lenovo.com/legal/copytrade.html.

The following terms are trademarks of Lenovo in the United States, other countries, or both:

Flex System™	Lenovo(logo)®	TruDDR4™
Lenovo®	System x®	

The following terms are trademarks of other companies:

Intel, Xeon, and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

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

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