WHITE PAPER

Memory performance of Xeon 5500 (Nehalem EP) based PRIMERGY servers

Version 1.0
August 2009

Pages 16

Abstract

The impressive performance of the Xeon 5500 (Nehalem EP)-based PRIMERGY Dual Socket models is primarily due to a paradigm change in connecting the main memory: QuickPath Interconnect (QPI) instead of Front Side Bus (FSB). The new architecture has some new parameters which must be considered when configuring the most powerful systems possible. The central topics are the different memory frequency with 800, 1066 and 1333 MHz as well as the most even distribution of the memory modules possible across three memory channels per processor. This White Paper explains the performance effects of these factors and provides help in defining powerful and yet low-priced configurations.

Content

Introduction

The current generation of dual socket PRIMERGY rack, tower and blade servers which use Intel Xeon 5500 (Nehalem EP) processors and the Intel 5520 chipset, has an extraordinary increase in performance (double compared to the previous generation). These findings apply for a wide range of load profiles and could be assigned with various benchmarks. Most of this improvement is due to a change in conencting the processors to the other system components, in particular, the main memory. This link, the System Interconnect, has been implemented via the Front Side Bus (FSB) in x86-based servers since the Intel Pentium Pro processor (1995). FSB technology has recently reached its limits regarding its complexity, for example the number of pins required in the chipset per FSB. The new approach, which is being introduced with the Xeon 5500 based systems, is the Intel QuickPath Interconnect (QPI). The QPI represents a paradigm change in the system architecture - from Symmetric Multiprocessing (SMP) to Non-Uniform Memory Access (NUMA).

The QPI connects processors to each other as well as processors and the chipset responsible for I/O via one-directional serial links which handle 6.4, 5.9 or 4.8 GT/s (gigatransfers per second) depending on the processor model. In order to link the main memory the processors in the series Xeon 5500 are equipped with memory controllers, i.e. each processor directly controls a group of assigned memory modules. The processor can simultaneously provide memory contents to the neighbouring processor via the QPI link and request such itself.

The direct connection between processor and memory means that an increase in memory performance is plausible, but with a difference in performance between local and remote request which justifies the classification of this architecture as NUMA. The operating system takes NUMA into consideration when allocating the physical memory and when scheduling processes. The total quantity of RAM should be distributed evenly across the two processors, as far as possible.

This recommendation is the entry-point for a range of additional considerations that result from the memory system features. The memory is thus clocked with 1333, 1066 or 800 MHz and the effective value for a specific configuration results from the type of processor, the type of DIMM used and their distribution over three memory channels per processor. In an ideal situation, the symmetry should not only cover the number of DIMM strips per processor but also per channel. This results in the recommendation of DIMM quantities which are multiples of 6 (2 processors each with 3 channels). The classic matrix when configuring memory with 8, 16, 32, 64 and 128 GB can not be implemented if this guideline is observed. However, if the customer requests these memory sizes: what will be the possible effects on performance?

This White Paper provides first an overview of the memory architecture of the Xeon 5500-based PRIMERGY servers. There then follows a pragmatic approach. Performant memory configurations are shown in tables based on the assumption that help is needed when defining configurations. This also assumes that the system and CPU type are specified and that the best suitable configuration is sought for a certain memory quantity (or an approximate memory configuration). In many situations it is sufficient just to look at these tables closely. The background for the recommended configurations is explained in the third section based on results with the STREAM benchmark. This section is recommended for the situation that the required memory capacity is not in the tables of the second section and when the configuration is to be defined on an individual basis. The document concludes with a compilation of *best practice* regulations.

The following applies regarding the complexity of this subject. A range of best practice regulations enables powerful systems to be configured quickly - despite what at first seems to be a large number of factors affecting performance. Looking at a balanced solution based on cost aspects, there is often freedom upwards but there are only slight performance improvements of under 5% on average. A certain degree of caution is required when considering whether such freedom upwards should be used, whether it is always necessary; likewise knowledge of the project background is required. A test for a benchmark is possibly handled differently to a shopping-basket for production systems.

Memory architecture

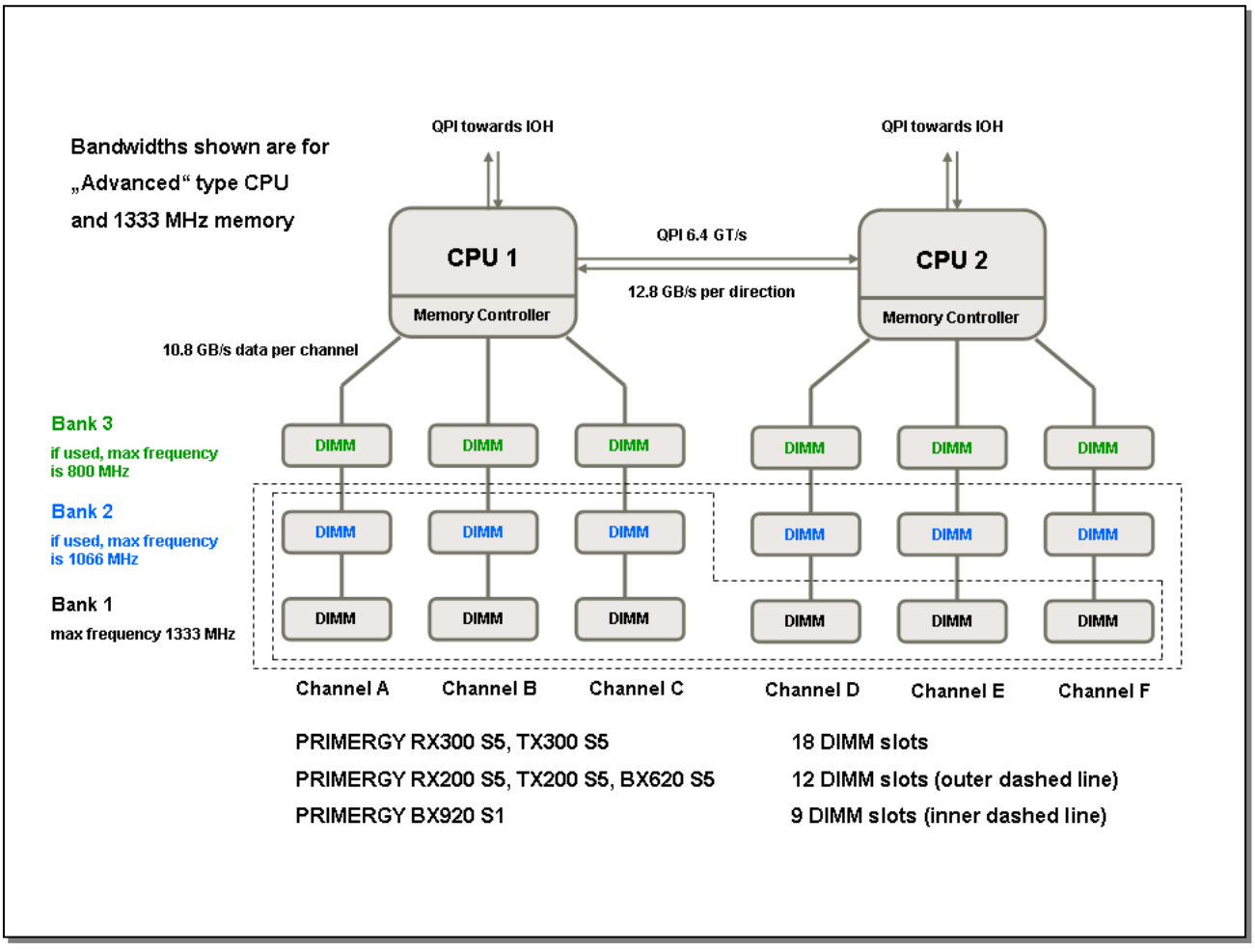
This section provides an overview of the memory system in three parts. A block diagram explains the arrangement of the available DIMM slots. The four memory configuration modes are then explained; which the PRIMERGY configurator also refers to as well. The third part covers the available DIMM types.

DIMM slots

The following diagram shows the structure of the memory system. There are three groups of models when regarding the DIMM slots and their arrangement

- Group 1 (18 slots): PRIMERGY RX300 S5 and TX300 S5
- Group 2 (12 slots): PRIMERGY RX200 S5, TX200 S5 and BX620 S5
- Group 3 (9 slots): PRIMERGY BX920 S1

The description is based on the systems available in July 2009.



There are always three memory channels per processor. However, the three model groups vary due to the number of possible maximum DIMM strips per channel where the space in the housings plays a decisive factor. The number of DIMM strips configured per channel influences the memory frequency and thus the memory performance. This size, often referred to below, is known as DPC (DIMM per channel). For example, in a 2DPC configuration of a PRIMERGY RX300 S5 there would be 2 DIMM strips per channel and thus a total of 12 strips.

It is not necessary for all the channels in the system to have the same DPC value. An abbreviation is commonly used when describing a configuration, e.g.:

2 - 2 - 2 / 1 - 1 - 1

for a configuration with two memory modules per channel on the first processor and one module each on the second processor.

Another term used below is "memory bank". As shown in the diagram, a group of three DIMM strips distributed across the channels forms a bank. The colours in the diagram (black, blue, green) correspond to the coloured marking of the banks on the motherboards of the servers which is aimed at preventing configuration errors. When distributing the DIMM strips via the slots available per processor it is desirable to start with bank 1 and to proceed bank-by-bank in order to attain the best interleaving possible. Interleaving is a main influence on memory performance and is described in detail below.

The corresponding processor must be available in order to use the DIMM slots. If operations are only with one processor the slots allocated to the empty socket cannot be used.

The four memory configuration modes

In addition to performance there is a further aspect involved when defining memory configuration; it is known by the abbreviation RAS (Reliability, Availability, Serviceability). The memory system offers interesting options for customers with high RAS requirements. This concerns the first two of the four memory configuration modes listed below. These first two modes are specified through the BIOS, if required. Otherwise the actual DIMM configuration decides whether it is performance or the independent channel mode. If the correct DIMM strips are positioned correctly, the result is automatically the performance mode.

- **Spare Channel Mode**: Each bank is either empty or configured with three DIMM strips (same type and same capacity). Only the DIMM strips in channels A and B (or D and E) are used. Channel C (or F) has the spare should a strip be faulty. The mode must be explicitly specified in the BIOS.
- **Mirror Channel Mode**: Only the channels A and B (or D and E) are used per bank which must be configured with DIMM strips of the same type. All the slots in channel C (or F) remain empty. The hardware mirrors the memory contents transparent for operating system and applications. The effective physical main memory is only half the configured capacity. A failed DIMM strip does not result in system downtime. The mode must be explicitly specified in the BIOS.
- **Performance Mode**: Each bank is either empty or configured with three DIMM strips (same type and same capacity). This configuration enables optimal interleaving via the three memory channels.
- **Independent Channel Mode**: all other configurations fall into this category. Each slot can be assigned with any DIMM strip from the types listed below as long as *unbuffered* and *registered* modules are not mixed.

The two first modes are possibly not supported by some models of the Xeon 5500-based servers.

Available memory types

DIMM strips listed in the following table are used when considering the configuration of the named PRIMERGY models. ECC-protected DDR3 memory modules are used.

The last but one column in the table shows the relative price differences. The list prices from July 2009 for the PRIMERGY RX200 S5 are used as a basis. The column shows the relative price per GB, standardized to the registered PC3-8500 DIMM, size 4 GB (highlighted as measurement 1). The high costs for 8 GB modules are conspicuous as well as the reasonable prices for unbuffered modules. The drop in price means that the question of costs must be taken into consideration when configuring memory. Doubling memory costs in order to increase performance by, for example 2% would hardly be sensible.

Type	Control	Max MHz	Rank	capacity	rel. Price per GB	Notes
DDR3-1333 PC3-10600 rg ECC	registered	1333	1	2 GB	1.1	
DDR3-1333 PC3-10600 rg ECC	registered	1333	2	4 GB	1.0	
DDR3-1333 PC3-10600 rg ECC	registered	1333	2	8 GB	3.7	
DDR3-1066 PC3-8500 rg ECC	registered	1066	1	2 GB	1.1	
DDR3-1066 PC3-8500 rg ECC	registered	1066	2	4 GB	**1**	
DDR3-1066 PC3-8500 rg ECC	registered	1066	2	8 GB	3.4	
DDR3-1066 PC3-8500 ub ECC	unbuffered	1066	1	1GB	0.8	a)
DDR3-1066 PC3-8500 ub ECC	unbuffered	1066	2	2 GB	0.7	a)

a) Not available for PRIMERGY RX300 S5 and TX300 S5

Unbuffered modules, due to their simple construction, have a lower maximum capacity and can only be used in 1DPC or 2DPC configurations.

Some sales regions can have restrictions regarding the availability of certain DIMM types.

Performant memory configurations

The following tables provide configuration examples for a comprehensive range of memory sizes which are suitable when considering performance. The configurations of the first table are thus "ideal" because the following applies for each configuration: the memory is distributed evenly across all memory channels in the system. These configurations correspond to the Performance Mode.

The second table is for the "classic" configurations of earlier system architectures 8, 16, 32 GB etc. These configurations should show some performance percentage disadvantages *when carefully measured* in comparison to the ideal configurations (insofar as the capacity difference itself has no effect on the test result). This disadvantage should be irrelevant for most applications (without preceding the explanations below: the cause for the difference is the 2-way interleave for classic sizes. The ideal configurations are 3-way interleaved.)

All configurations in the first two tables are optimal regarding NUMA: the memory is distributed symmetrically across both sockets. Asymmetric memory configurations are then handled later.

Configuration alternatives which enable maximum timing with 1333 MHz are marked red. The processor class *Advanced* to which this option applies consists of the Xeon 5500 models X5570, X5560 and X5550. Observe the memory costs for the red alternatives: the extra cost is moderate if registered and not unbuffered DIMMs are used but indeed considerable if 8 GB modules are used.

The last three columns in the following tables show the PRIMERGY models for which the respective configuration is possible.

The explanations in the section "Memory performance" should enable you to create memory configurations for those configurations not covered here.

Ideal memory sizes							
Overall capacity	**Memory type**	**Module size GB**	**Configuration**	**Notes**	**R/TX200 S5 BX620 S5**	**BX920 S1**	**R/TX300 S5**
6 GB	PC3-8500 unbuffered	1	1 - 1 - 1 / 1 - 1 - 1		x	x	
12 GB	PC3-8500 unbuffered	2	1 - 1 - 1 / 1 - 1 - 1	Price advantage for registered	x	x	
	PC3-8500 registered	2	1 - 1 - 1 / 1 - 1 - 1		x	x	x
	PC3-10600 registered	2	1 - 1 - 1 / 1 - 1 - 1	If Advanced CPU and 1333 MHz required	x	x	x
18 GB	PC3-8500 unbuffered	1 and 2	2 - 2 - 2 / 2 - 2 - 2	1st bank 2 GB modules 2nd bank 1 GB modules	x		
24 GB	PC3-8500 unbuffered	2	2 - 2 - 2 / 2 - 2 - 2	Price advantage to 1DPC with 4 GB strips	x		
	PC3-8500 registered	4	1 - 1 - 1 / 1 - 1 - 1		x	x	x
	PC3-10600 registered	4	1 - 1 - 1 / 1 - 1 - 1	If Advanced CPU and 1333 MHz required	x	x	x
36 GB	PC3-8500 registered	2 and 4	2 - 2 - 2 / 2 - 2 - 2	1st bank 4 GB modules 2nd bank 2 GB modules	x		x
48 GB	PC3-8500 registered	4	2 - 2 - 2 / 2 - 2 - 2		x		x
	PC3-8500 registered	8	1 - 1 - 1 / 1 - 1 - 1	Interesting for BX920 S1	x	x	x
	PC3-10600 registered	8	1 - 1 - 1 / 1 - 1 - 1	If Advanced CPU and 1333 MHz required	x	x	x
60 GB	PC3-8500 registered	2 and 8	2 - 2 - 2 / 2 - 2 - 2	1st bank 8 GB modules 2nd bank 2 GB modules	x		x
	PC3-8500 registered	2 and 4	3 - 3 - 3 / 3 - 3 - 3	1st bank 4 GB modules 2nd bank 4 GB modules 3. Bank 2 GB modules			x
72 GB	PC3-8500 registered	4 and 8	2 - 2 - 2 / 2 - 2 - 2	1st bank 8 GB modules 2nd bank 4 GB modules	x		x
	PC3-8500 registered	4	3 - 3 - 3 / 3 - 3 - 3				x
84 GB	PC3-8500 registered	2, 4 and 8	3 - 3 - 3 / 3 - 3 - 3	1st bank 8 GB modules 2nd bank 4 GB modules 3. Bank 2 GB modules			x
96 GB	PC3-8500 registered	8	2 - 2 - 2 / 2 - 2 - 2		x		x
108 GB	PC3-8500 registered	2 and 8	3 - 3 - 3 / 3 - 3 - 3	1st bank 8 GB modules 2nd bank 8 GB modules 3. Bank 2 GB modules			x
120 GB	PC3-8500 registered	4 and 8	3 - 3 - 3 / 3 - 3 - 3	1st bank 8 GB modules 2nd bank 8 GB modules 3. Bank 4 GB modules			x
144 GB	PC3-8500 registered	8	3 - 3 - 3 / 3 - 3 - 3				x

Classic memory sizes							
Overall capacity	**Memory type**	**Module size GB**	**Configuration**	**Notes**	**R/TX200 S5 BX620 S5**	**BX920 S1**	**R/TX300 S5**
4 GB	PC3-8500 unbuffered	1	1 - 1 - 0 / 1 - 1 - 0		x	x	
8 GB	PC3-8500 unbuffered	1	2 - 1 - 1 / 2 - 1 - 1	Price advantage for registered and 3 channels	x		
	PC3-8500 registered	2	1 - 1 - 0 / 1 - 1 - 0		x	x	x
	PC3-10600 registered	2	1 - 1 - 0 / 1 - 1 - 0	If Advanced CPU and 1333 MHz required	x	x	x
16 GB	PC3-8500 unbuffered	2	2 - 1 - 1 / 2 - 1 - 1	Price advantage for registered	x		
	PC3-8500 registered	2	2 - 1 - 1 / 2 - 1 - 1	3 channels used	x		x
	PC3-8500 registered	4	1 - 1 - 0 / 1 - 1 - 0		x	x	x
	PC3-10600 registered	4	1 - 1 - 0 / 1 - 1 - 0	If Advanced CPU and 1333 MHz required	x	x	x
32 GB	PC3-8500 registered	4	2 - 1 - 1 / 2 - 1 - 1		x		x
	PC3-8500 registered	8	1 - 1 - 0 / 1 - 1 - 0	Interesting for BX920 S1	x	x	x
	PC3-10600 registered	8	1 - 1 - 0 / 1 - 1 - 0	If Advanced CPU and 1333 MHz required	x	x	x
64 GB	PC3-8500 registered	8	2 - 1 - 1 / 2 - 1 - 1		x		x
	PC3-8500 registered	4	3 - 3 - 2 / 3 - 3 - 2	Price advantage for 8 GB modules			x
128 GB	PC3-8500 registered	8	3 - 3 - 2 / 3 - 3 - 2				x

1333 MHz are only possible in the red situations. In all situations with three configured banks, the timing is always 800 MHz. A single channel with 3DPC is sufficient to create this situation. Otherwise, 1066 MHz applies in the non-red situations for the CPU types Xeon E5520 to X5570 and 800 MHz for E5502 to E5506.

The required memory capacity is assumed. Its implicit influence on the application performance, e.g. on I/O rates, must here be ignored.

Asymmetric memory configurations

Not all the systems enable symmetric memory configuration in all configuration version based on their form factor. The diagram on page 3 shows the asymmetric arrangement of the DIMM slots for the PRIMERGY BX920 S1: there are two memory banks with the first socket, and one with the second. Regarding the NUMA recommendation to distribute the memory symmetrically via both sockets, this suggests a different aspect when listing recommending configurations.

NUMA-optimal configurations

Configurations which divide the total capacity of memory into two identical halves are possible in the PRIMERGY BX920 S1 up to a capacity 48 GB, *despite* the asymmetry of the slots. These configurations are NUMA-optimal. These configurations are already noted in the earlier tables for ideal and classic memory sizes.

Standard 8 GB modules were used for the capacities 32 GB and 48 GB in the tables already shown. Their cost disadvantages are reduced if 4 GB modules are used by the CPU that has six DIMM slots. The following table shows these versions where only two or three expensive 8 GB modules are required. Because these are 2DPC configurations, the memory frequency 1333 MHz can not be achieved. For “Advanced” processors, this is the only performance difference compared to the versions shown previously. For the other processors, there is the cost difference only.

PRIMERGY BX920 S1 NUMA-optimal versions				
Overall capacity	**Memory type**	**Module size GB**	**Configuration**	**Notes**
32 GB	PC3-8500 registered	4 and 8	2 - 1 - 1 / 1 - 1 - 0	4 GB modules left 8 GB modules right
48 GB	PC3-8500 registered	4 and 8	2 - 2 - 2 / 1 - 1 - 1	4 GB modules left 8 GB modules right

Asymmetric configurations

There is more memory on the left than on the right in the configurations in the next and last table. The excess is between a fifth and third of the total capacity. For maximum half of the excess, i.e. a tenth to a sixth, there is "remote" access via the QPI link (seen statistically). In such cases of moderate asymmetry a performance disadvantage of about 1-2% must be calculated compared to symmetric configurations. For workloads where remote accesses are unavoidable anyway, like databases with their large shared memory segments, there will be no performance disadvantage. This was verified with OLTP2 measurements on PRIMERGY BX920 S1 under Windows Server 2008 and SQL Server 2008.

The table begins with a low-priced alternative to the symmetric 32 GB configuration which works completely without 8 GB modules. Configurations for smaller memory configurations than shown in the following table can all be implemented symmetrically.

PRIMERGY BX920 S1 Asymmetric configurations				
Overall capacity	**Memory type**	**Module size GB**	**Configuration**	**Notes**
32 GB	PC3-8500 registered	4	2 - 2 - 2 / 1 - 1 - 1	
36 GB	PC3-8500 registered	4	2 - 2 - 2 / 1 - 1 - 1	
40 GB	PC3-8500 registered	4 and 8	2 - 2 - 2 / 1 - 1 - 0	4 GB modules left 8 GB modules right
60 GB	PC3-8500 registered	4 and 8	2 - 2 - 2 / 1 - 1 - 1	1st bank both sides 8 GB, 2nd bank left 4 GB modules
64 GB	PC3-8500 registered	8	2 - 2 - 2 / 1 - 1 - 1	
72 GB	PC3-8500 registered	8	2 - 2 - 2 / 1 - 1 - 1	

The effects on memory performance

This section explains the components which have an effect on the performance of the RAM. First of all, there is the question of how memory performance was measured in the tests preceding this White Paper and about the interpretation quality of such data.

The measurement tool: STREAM Benchmark

STREAM Benchmark from John McCalpin [L3] is a tool to measure memory throughput. The benchmark executes copy and calculation operations on large arrays of the data type *double* and it provides results for four access types: Copy, Scale, Add and Triad. The last three contain calculation operations. The result is always a throughput that is specified in GB/s. Triad values are quoted the most. All the STREAM measurement values specified in the following to quantify memory performance are based on this practice and are GB/s for the access type Triad.

STREAM is the industry standard for measuring the memory bandwidth of servers, known for its ability to put memory systems under immense stress using simple means. It is clear that this benchmark is particularly suitable for the purpose of studying effects on memory performance in a complex configuration space. In each situation STREAM shows the maximum effect on performance caused by a configuration action which affects the memory, be it deterioration or improvement. The percentages specified below regarding the STREAM benchmark are thus to be understood *as bounds for* performance effects.

The memory effect on *application* performance is differentiated between the latency of each access and the bandwidth required by the application. The quantities are interlinked as real latency increases with increasing bandwidth. The scope in which the latency can be "hidden" by parallel memory access also depends on the application and the quality of the machine codes created by the compiler. As a result, making general forecasts for all application scenarios is very difficult.

Attempting to show the relationship between STREAM benchmark and the performance of real applications can only be based on empirical values:

- A quarter of the STREAM relations can be used for commercial applications (databases, SAP, etc.). For example, if STREAM shows 20% deterioration, then empirical values suggest 5% worsening for commercial applications.
- A greater effect on memory performance must be assumed for technical-scientific applications, i.e. such applications will react more strongly than commercial such as STREAM. About half the STREAM relation can be assumed.

As commercial application scenarios are more common in practice, the following section uses the first of these general rules in order to estimate the performance effects for real applications.

The rules are substantiated by selectively available results with standard benchmarks. For example, STREAMS shows an 11% improvement if memory is clocked at 1333 MHz instead of 1066 MHz. The improvement is 3% for SPECint_rate2006 (representative of commercial workloads) and 5% for SPECfp_rate2006 (technical-scientific workloads). The improvement is also 3% for SPECjbb2005 (Java performance).

When assessing how great the effect really is, it should also be considered that it only appears directly with a fully utilized system; and otherwise in a change in the CPU load. If the utilization is not already very high, a really noticeable effect on response times is then improbable.

Primary performance influences

This section looks at the two main influences on memory performance: frequency and interleaving. Both parameters have three options each: timing with 800, 1066 or 1333 MHz as well as 1-way, 2-way or 3-way interleaving. Planning a memory configuration should first of all involve planning these parameters.

Effective frequency of the memory

The effective timing determined by the BIOS when switching-on the system is based on three factors:

- The type of processor. The models in the series Xeon 5500 are divided into three groups according to the following table; Intel calls them Basic, Standard and Advanced. Processors in the group Basic only support 800 MHz, the group Standard support 800 and 1066 MHz, and the group Advanced supports the maximum 1333 MHz.
- The type of DIMM. DIMM strips up to a maximum 1066 MHz and a maximum 1333 MHz are available.
- The DPC value (DIMM per channel). 1333 MHz timing is only possible with 1 DPC. A 2DPC configuration limits the timing to 1066 MHz, a 3DPC configuration to 800 MHz. If the six channels are not configured the same, the largest DPC value is decisive.

Class	Xeon type	#cores	GHz	L3 cache (MB)	QPI (GT/s)	Max memory MHz	TDP (Watt)
Advanced	X5570	4	2.93	8	6.4	1333	95
	X5560	4	2.80	8	6.4	1333	95
	X5550	4	2.67	8	6.4	1333	95
Standard	E5540	4	2.53	8	5.9	1066	80
	E5530	4	2.40	8	5.9	1066	80
	E5520	4	2.27	8	5.9	1066	80
	L5520	4	2.27	8	5.9	1066	60
Basic	E5506	4	2.13	4	4.8	800	80
	L5506	4	2.13	4	4.8	800	60
	E5504	4	2.00	4	4.8	800	80
	E5502	2	1.87	4	4.8	800	80

The highest possible timing is desirable. However, the minimum of the maxima is effective: the worst best value for the three factors determines the timing of the configuration. The timing is defined as standard for the system and not per processor.

The following example explains the mechanism. A PRIMERGY RX200 S5 is fully equipped with processors of the type X5570 and with PC3-8500 modules sized 4 GB according to the plan

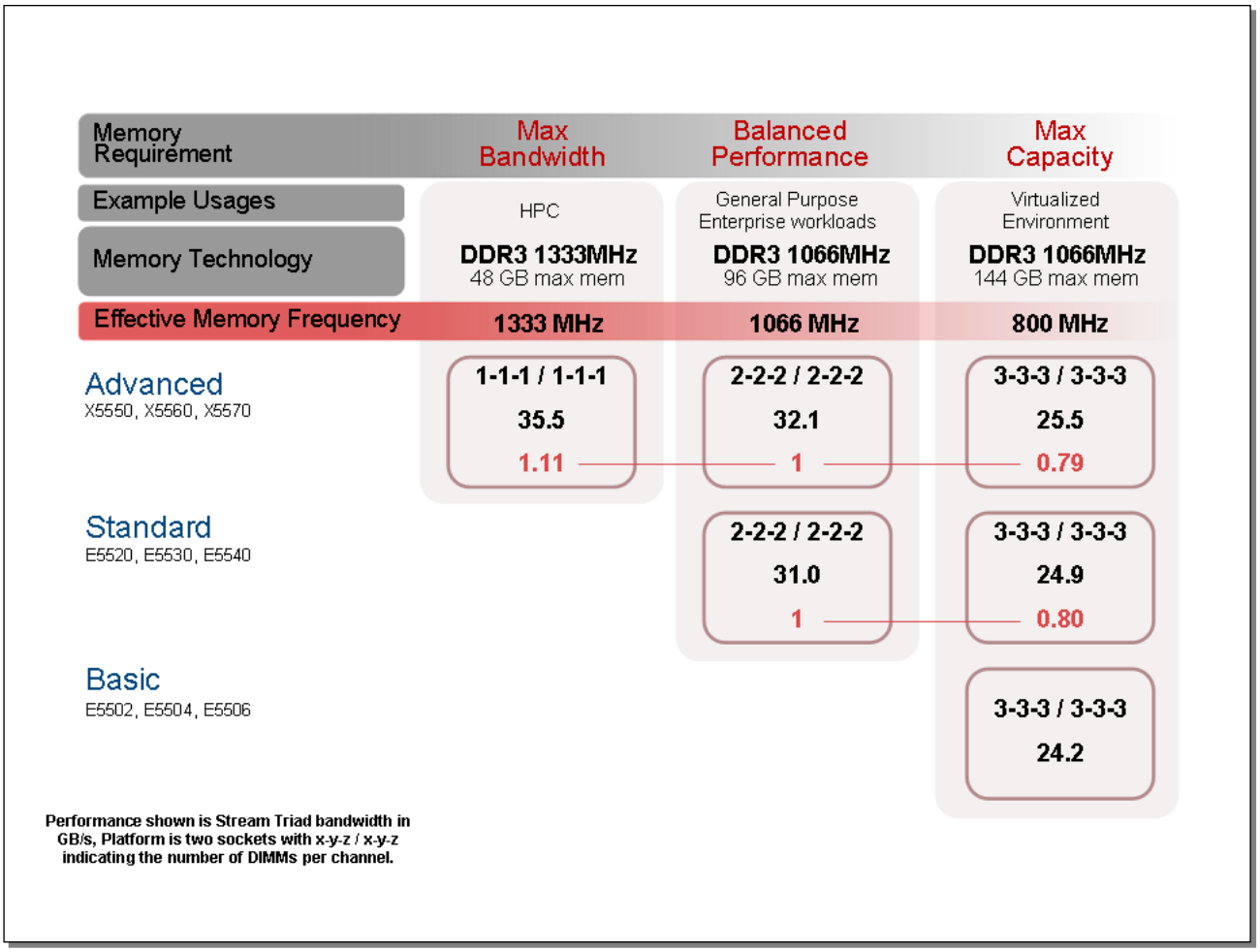
2 - 2 - 2 / 2 - 2 - 2.

This results in a memory capacity of 48 GB and timing of 1066 MHz. The processor would support 1333 MHz, but both the PC3-8500 memory as well as the 2DPC configuration limits the timing to 1066 MHz. The configuration of 1333-MHz capable PC3-10600 4 GB modules would not be sensible because the limit caused by 2DPC would still be 1066 MHz. However, the capacity 48 GB would also be reached with 8 GB PC3-10600 modules based on the diagram

1 - 1 - 1 / 1 - 1 - 1.

This configuration would have the optimal timing of 1333 MHz.

The second version is more performant, the first is lower-priced. The following diagram shows the maximum performance effects caused by different timing. The diagram attempts to show a reasonable relationship between processor groups and memory technology as part of a matrix. 1333-Mhz-capable PC3-10600 memory is only sensible for processors in the Adavnced group (type ID beginning with X). And no thoughts are required with Basic processors regarding memory timing: the effective timing is always 800 MHz.



The diagram shows the STREAM Triad GB/s and - in red - the percentage change compared to a "average" timing with 1066 MHz. Compared to the starting configuration, the 1333-MHz timing using 8 GB modules in the above RX200 S5 example would result in a *best case* performance increase of some 10%; it would realistically be about 3% for commercial applications. The cost percentage for memory would simultaneously increase more than three times. Without wanting to cover any special situations, e.g. projects in the HPC environment (High Performance Computing), the decision is usually in favour of 4 GB modules and 1066 MHz timing.

The diagram compares STREAM values from configurations with different DPC as the respective required timing can only be created by varying the DPC value. The fact that configurations with different total capacity are compared is insignificant for measurements with STREAM: the benchmark always uses only 1 GB of available physical memory. If the same frequency results already for other reasons, then the maximum memory bandwidth for configurations with different DPC is at first the same: see the 800 MHz results

1 - 1 - 1 / 1 - 1 - 1	24.5 GB/s
2 - 2 - 2 / 2 - 2 - 2	24.1 GB/s
3 - 3 - 3 / 3 - 3 - 3	24.2 GB/s

for configurations with the CPU E5506 and dual-rank PC3-8500 modules, size 4 GB. There is one small condition for this statement as seen below in the section about the number of ranks for the situation that the total number of ranks per channel is odd. However, this situation can not occur if dual-rank modules are configured.

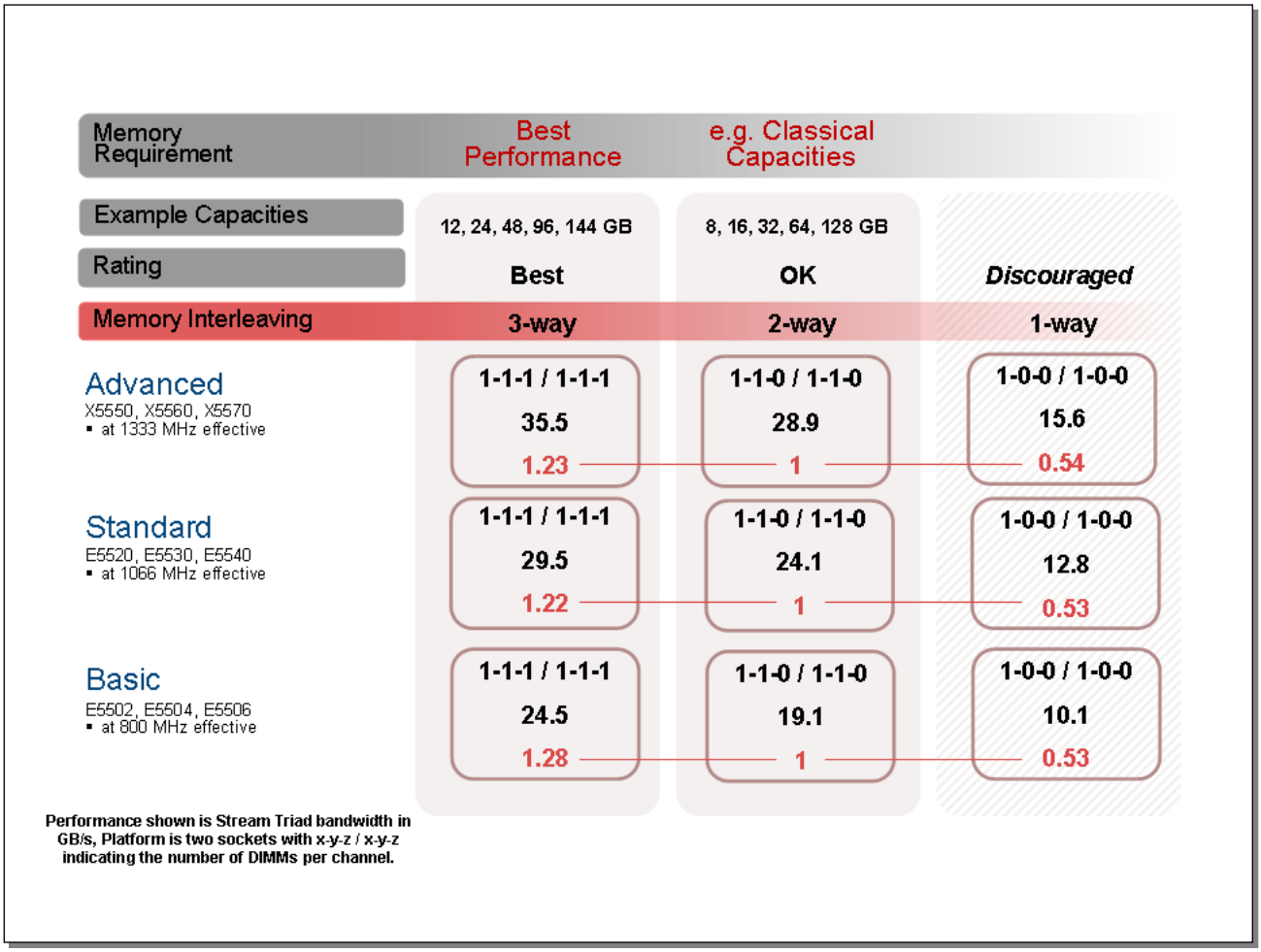
The summary for RAM clock frequency is: In otherwise comparable conditions there is a performance increase in timing with 1333 MHz compared to 1066 MHz of about a maximum 10%, and realistically about 3% for commercial applications. Timing with 800 MHz has a loss of maximum 20%, and about 5% for

commercial applications. We repeat that this information about commercial applications is a general guide and not to be seen as solid forecasts for specific application situations.

Interleaving

Interleaving in this conjunction is the set-up of the physical address area by alternating between the three memory channels per processor: the first block is in the first channel, the second in the second, etc. Memory access, which according to the locality principle is mainly to *adjacent* memory areas, is thus distributed across all channels. This is a performance gain situation resulting from parallelism. Furthermore, the delay is less noticeable which must be observed according to the physics of DRAM memory before changing the active ("open") memory page.

The following diagram shows the even greater effect of interleaving in contrast to the previous memory timing situation. The ideal situation is the 3-way interleave which always results if all three channels are configured identically. The Performance Mode of the memory configuration options is based on this scenario. The diagram also indicates why this ideal situation can frequently not be reached despite the "recommendation" (apart from configurations in Spare Channel Mode that also correspond to the middle column in the diagram): the need for classic memory configurations to the power of 2.



If we remember that the diagram details are extreme values, 2-way interleave configurations indeed seem to be reasonable . The 1-way interleave should be avoided, however: it is really a non-interleave and only referred to as 1-way because of the systematics involved. A performance loss must be assumed which - under normal circumstances - is in no sensible relationship to the performance capability of the processors. This configuration only seems to be sensible if the fail-safety in Mirror Channel Mode is the decisive criterion above all performance questions.

Interleaving, like timing, is defined by the BIOS when the system is switched-on. If the number of GB per channel is the same, a 3-way interleave is possible for three configured channels; a 2-way interleave with

two channels (if a channel is not used). This situation which is best for interleaving can also exist with non-uniform DPC values when using different-sized DIMM strips. The total GB per channel is decisive.

If the GB per channel are different, the physical memory is split in areas with different interleaving. The aim in this situation is to avoid areas with 1-way interleave. The BIOS thus resolves a

2 - 1 - 1 / 2 - 1 - 1

with identical 4 GB strips (which is sensible for reaching a total capacity of 32 GB, for example) into two 2-way halves as follows:

1 - 1 - 0 / 1 - 1 - 0	(50% of memory capacity)	2-way interleaving
1 - 0 - 1 / 1 - 0 - 1	(50%)	2-way interleaving

Instead of

1 - 1 - 1 / 1 - 1 - 1	(75%)	3-way interleaving
1 - 0 - 0 / 1 - 0 - 0	(25%)	1-way interleaving

in order to avoid the unevenness of the second version.

The summary for interleaving is: the ideal 3-way interleave means a gain in performance of up to 20% compared to 2-way, and realistically 5% for commercial applications. The 2-way interleave arises with classic memory capacities in powers of two (8, 16, 32 GB etc.) and with Spare Channel Mode. We do not recommend configurations with 1-way interleave due to a very high loss in performance. If the number of GB per channel is not the same, memory areas with different interleaving are the result.

Secondary performance influences

The topics discussed so far assume that these influences become noticeable in the application performance when measurements are carefully made. With the following topics proof is indeed possible with measurements of the maximum memory bandwidth, but whether they have an effect on a realistic application performance is questionable.

UDIMM or RDIMM?

According to the following table, for models PRIMERGY RX200 S5, TX200 S5, BX620 S5 and BX920 S1 also *unbuffered* DIMM modules (UDIMM) are available apart from the *registered* DIMM modules (RDIMM). The more simple UDIMM construction means that they are cheaper and use slightly less energy. If they can cover the required memory capacity, they should be preferred for these reasons.

Type	Control	Max MHz	Rank	capacity	Notes
DDR3-1333 PC3-10600 rg ECC	Registered	1333	1	2 GB	
DDR3-1333 PC3-10600 rg ECC	Registered	1333	2	4 GB	
DDR3-1333 PC3-10600 rg ECC	Registered	1333	2	8 GB	
DDR3-1066 PC3-8500 rg ECC	Registered	1066	1	2 GB	
DDR3-1066 PC3-8500 rg ECC	Registered	1066	2	4 GB	
DDR3-1066 PC3-8500 rg ECC	Registered	1066	2	8 GB	
DDR3-1066 PC3-8500 ub ECC	Unbuffered	1066	1	1 GB	a)
DDR3-1066 PC3-8500 ub ECC	Unbuffered	1066	2	2 GB	a)

a) Not available for PRIMERGY RX300 S5 and TX300 S5

A mix of RDIMM and UDIMM is not possible.

With RDIMM the control commands of the memory controller are buffered in the register (that gave the name) which is in its own component on the DIMM. This relieves the memory channel and enables 3DPC configurations which are not possible with UDIMM. Vice versa, 2DPC configurations with UDIMM result in a greater load (in comparison to 1DPC) which requires DIMM addressing with 2N timing (instead of 1N): control commands are only possible with every second clock of the memory channel. This results in a reduction of maximum memory bandwidth for 2DPC configurations with UDIMM by some 5% in comparison to RDIMM.

This effect can be ignored for the performance of commercial applications.

The number of ranks

The last table also shows that memory modules with 1 or 2 ranks are available. This means: there are DIMM with only one group of DRAM chips which synchronously read or write memory areas of width 64 bit. The individual chip is responsible for 8 bit. Or there are two such groups. However, the DIMM address and data lines are then common for both groups, i.e. only one of the groups can be active at any given time. The motivation for dual-rank DIMM is first the greater capacity as can be seen in the table.

A second advantage of dual-rank modules is the physical reason already discussed. Memory cells are arranged in two dimensions. A line (also called page) is opened and then a column item is read in this line. While the page is open, further column values can be read *with a much lower* latency. This latency difference motivates optimization of the memory controller which reallocates the pending orders regarding possible "open" memory pages. With dual-rank modules, the probability of accessing an open page increases.

This can be seen when measuring the memory bandwidth with STREAM according to the following table:

CPU	RAM				Bandwidth (GB/s)
	Type	capacity	#rank	Configuration	
X5570	PC3-10600 registered	4 GB	2	1 - 1 - 1 / 1 - 1 - 1	35.5
X5570	PC3-10600 registered	2 GB	1	1 - 1 - 1 / 1 - 1 - 1	32.3

Similar effects are seen when, for configurations with higher DPC values, the number of ranks per channel is odd. This situation cannot occur when using dual-rank modules. With a configuration with 2 GB modules, the performance disadvantage of realistically 2-3% with an odd number of ranks per channel is an additional reason for preferring dual-rank UDIMM modules to single-rank RDIMM.

Compiling the best-practice regulations

The following simple seven rules should be taken into account when selecting a performant and cost-efficient memory configuration. The regulations are listed according to their importance.

NUMA	Distribute the memory, if possibly, symmetrically across both sockets and activate the NUMA mode in the BIOS (default value). (BX920 S1: NUMA optimal configurations are possible up to 48 GB despite the asymmetric slots. The moderate asymmetry for larger configurations will only normally be seen in low percentages which can be ignored.)
Interleave	Configure a CPU socket bank-by-bank, i.e. include all there channels if possible. Avoid 1-way (non-interleaved) areas. All configurations for the tables on pages 6 to 8 are 3-way or 2-way and provide good performance.
1333 MHz	Only possible with comparably tight prerequisites: "Advanced" CPU and maximum 24 GB (low-priced modules) or 48 GB (expensive 8 GB modules). More interesting for benchmarks and projects with special conditions (e.g. HPC).
800 and 1066 MHz	"Basic" CPUs always limit to 800 MHz, likewise 3DPC configurations.
DIMM price drops	UDIMM modules are cheaper than RDIMM. A clear jump in price can be seen between 1 GB to 4 GB modules on the one hand (irrespective of whether UDIMM or RDIMM) and 8 GB modules on the other hand (only exist as RDIMM).
UDIMM and RDIMM	If possible (cf. the tables on pages 6 to 8), UDIMM modules are preferred for price and energy reasons.
Number of ranks	An even number of ranks per channel results in a small advantage in performance. Interesting for benchmarks and not so much for productive operations.

The following quantity statements for otherwise comparable conditions apply for commercial applications:

- The frequency 1333 MHz has an improvement of some 3% in contrast to 1066 MHz.
- The frequency 800 MHz results in a worse result of some 5% in contrast to 1066 MHz.
- A 3-way interleave - in contrast to 2-way - means an improvement of some 5%.
- Ignorable are:
  - The 2N timing of unbuffered DIMM in 2DPC configurations.
  - An odd number of ranks per channel.

Literature

[L1] PRIMERGY systems
http://ts.fujitsu.com/primergy
[L2] PRIMERGY performance
http://ts.fujitsu.com/products/standard_servers/primergy_bov.html
[L3] STREAM benchmark
http://www.cs.virginia.edu/stream/
[L4] PRIMERGY RX200 S5 data sheet and performance report
http://docs.ts.fujitsu.com/dl.aspx?id=9d1f923e-8be1-4a17-a6d6-b00736f5d765
http://docs.ts.fujitsu.com/dl.aspx?id=c4917140-8852-463d-8121-3db977b8f9e1
[L5] PRIMERGY RX300 S5 data sheet and performance report
http://docs.ts.fujitsu.com/dl.aspx?id=c245c1d4a-047c-4217-a1aa-6d5a333d856b
http://docs.ts.fujitsu.com/dl.aspx?id=c058ec307-e8b2-49d7-8ba3-d352cfce6945
[L6] PRIMERGY TX300 S5 data sheet and performance report
http://docs.ts.fujitsu.com/dl.aspx?id=c731a6c-7240-4432-b516-b7c211a829b3
http://docs.ts.fujitsu.com/dl.aspx?id=c05df4b52-64e4-4396-8bae-b7b1a0f4ae3f
[L7] PRIMERGY BX620 S5 data sheet and performance report
http://docs.ts.fujitsu.com/dl.aspx?id=c0703-df6d-4989-b0c9-ada3435f78b2
http://docs.ts.fujitsu.com/dl.aspx?id=c3913334-75cc-4baf-a9cc-18edfdfe9f09
[L8] PRIMERGY BX920 S1 data sheet and performance report
http://docs.ts.fujitsu.com/dl.aspx?id=c0703-df6d-4989-b0c9-ada3435f78b2
http://docs.ts.fujitsu.com/dl.aspx?id=9e5c11e1-9462-4974-a6ff-69d915644867

Contact

PRIMERGY Hardware

PRIMERGY Product Marketing
mailto:Primergy-PM@ts.fujitsu.com

PRIMERGY Performance and Benchmarks

PRIMERGY Performance and Benchmarks
mailto:primergy.benchmark@ts.fujitsu.com

Published by:

Enterprise Products
PRIMERGY Server
PRIMERGY Performance Lab
mailto:primergy.benchmark@ts.fujitsu.com

Internet:
http://www.ts.fujitsu.com/primergy

Extranet:
http://partners.ts.fujitsu.com/com/products/servers/primergy