WHITE PAPER
FUJITSU PRIMERGY SERVERS
MEMORY PERFORMANCE OF XEON 5600 (WESTMERE-EP) BASED SYSTEMS

The Xeon 5600 (Westmere-EP)-based PRIMERGY Dual Socket models are the second generation since the paradigm change in connecting the main memory implemented with the Xeon 5500 (Nehalem-EP): QuickPath Interconnect (QPI) instead of Front Side Bus (FSB). This 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.

Version

2.0
2011-06-06

Content

© Fujitsu Technology Solutions 2011

Summary

The Xeon 5600 (Westmere-EP)-based PRIMERGY Dual Socket models are the second generation since the paradigm change in connecting the main memory implemented with the Xeon 5500 (Nehalem-EP): QuickPath Interconnect (QPI) instead of Front Side Bus (FSB). The configuration of powerful systems deals with the following features:

- NUMA architecture.
- Memory frequency with 1333, 1066 or 800 MHz.
- 3-way, 2-way or 1-way interleave, depending on the distribution of the DIMM strips across three memory channels per processor.

In NUMA architecture the DIMM slots are directly assigned to the processors ("local memory"). Both processors should be configured with DIMM strips, ideally in a symmetrical way. The PRIMERGY BX920 S2 with a larger memory requires a moderately asymmetrical configuration with an harmless performance disadvantage of 2-3%.

The BIOS default *NUMA Optimization* = *Enabled* should not be changed.

1333 MHz is the maximum memory frequency for the powerful Xeon 5600 models, with 1066 MHz being for the less powerful ones. The reasons for downgrading the frequency to 1066 or 800 MHz are large memories, particularly when using 16 and 32 GB DIMM strips, and energy-saving 1.35 V *low voltage* (LV) operation (with two DIMM strips per memory channel). A downgrade is associated with a loss in performance of up to 5% (average value for commercial applications) and as a rule harmless.

The *Performance Mode* memory option of the PRIMERGY configurator ensures identical configuration of all memory channels in the system and results in the optimal 3-way interleave. Classic memory sizes such as 16, 32, 64 GB cannot be implemented in Performance Mode (the necessary DIMM numbers are not multiples of 6) and result in the 2-way interleave. This entails a performance disadvantage of between 1 and 5%, which as a rule does not pose any risks.

Configurations with only one DIMM strip per processor (1-way interleave) should be avoided. Apart from with the weakest processors, otherwise a loss in performance in the order of 20% occurs.

With memory performance under redundancy DIMM sparing means a performance disadvantage of 1 to 5%. During mirroring the advantage of fail-safety must be weighed up against a loss in performance of about 10%.

Moreover, the basic rule applies: the more powerful the processor, the greater the influence of the memory parameters.

 © Fujitsu Technology Solutions 2011

Document history

Version 1.0 (2010-05-07)

Original version

Version 1.0 (2010-08-23)

Inclusion of PRIMERGY CX120 S1

Version 1.2 (2010-12-14)

Inclusion of PRIMERGY CX122 S1

Version 2.0 (2011-06-06)

Introduction of new Westmere-EP CPU models ("Westmere-EP Refresh")
Introduction of a 32 GB DIMM

© Fujitsu Technology Solutions 2011

Introduction

32nm manufacturing technology is the main innovation in Intel Xeon 5600 (Westmere-EP) processors, with which the current generation of dual socket PRIMERGY rack, tower, blade and cloud servers is equipped. Compared with the predecessor generation of Xeon 5500 (Nehalem-EP), which was manufactured in 45nm, this technology offers up to six cores per processor and an increase in the L3 cache from 8 to 12 MB. These features result in an increase in performance to the order of 40%.

Both generations have the same Intel QuickPath Interconnect (QPI)-based micro-architecture. This architecture has dramatically improved the connecting of the processors to the other system components, in particular, the main memory and has approximately doubled system performance in comparison with earlier architecture. Front Side Bus (FSB) technology, which has been in use since the Intel Pentium Pro processor (1995), had reached its limits regarding complexity, for example the number of pins required in the chipset per FSB. The QPI approach represents a paradigm change in the system architecture - from Symmetric Multiprocessing (SMP) to Non-Uniform Memory Access (NUMA). This White Paper describes the performance features of QPI architecture with an eye toward memory configurations of the most powerful systems possible and, in so doing, takes the special features of the Xeon 5600 (Westmere-EP)-based generation into account. In detail, there are a number of differences compared with the Xeon 5500 (Nehalem-EP).

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 Xeon 5500 and 5600 series 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 neighboring 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 cannot 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 5600-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 benchmarks STREAM and SPECint_rate_base2006. 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 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 should possibly be handled differently to a shopping-basket for production systems.

 © Fujitsu Technology Solutions 2011

Memory architecture

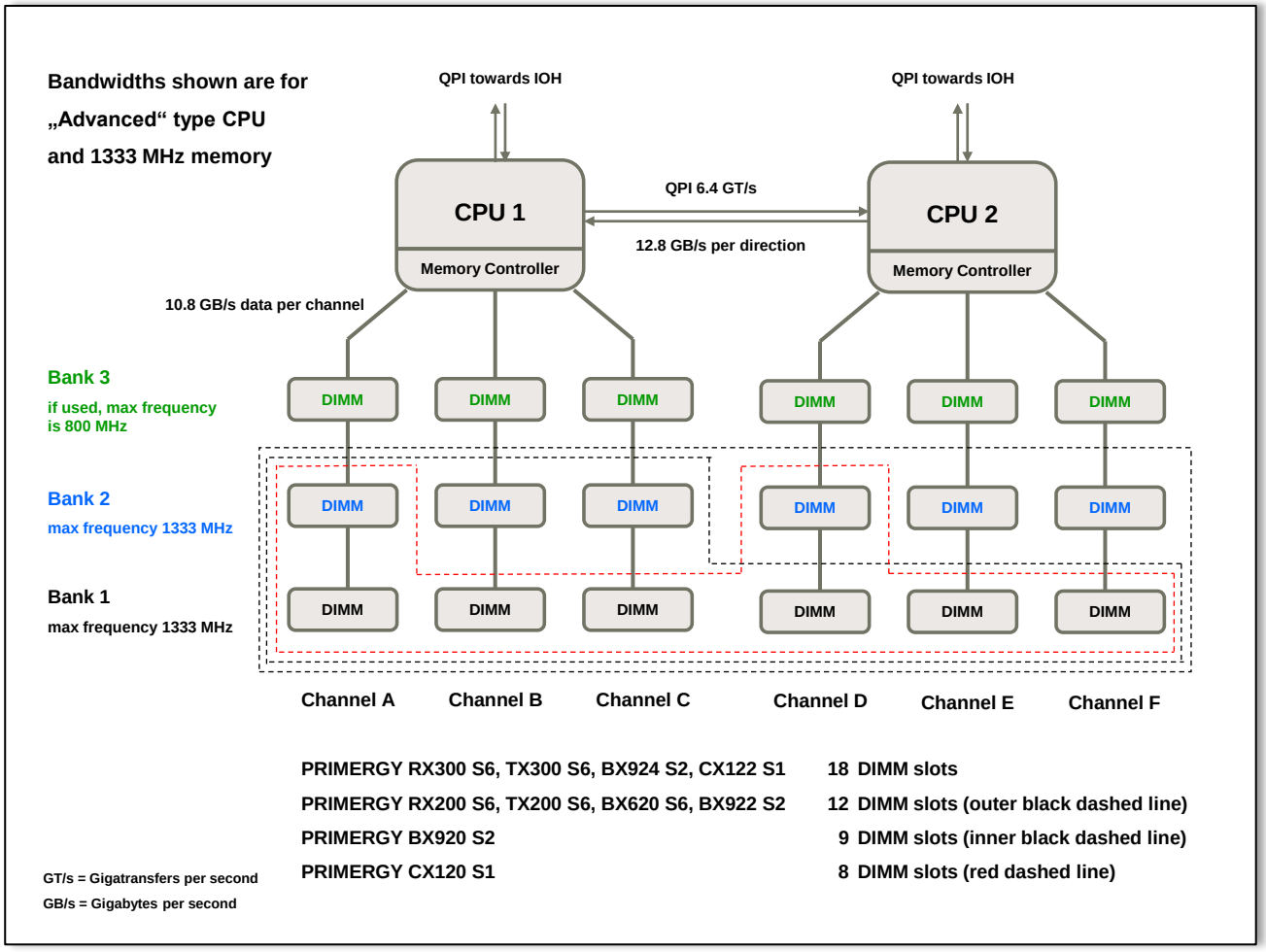
This section provides an overview of the memory system in three parts. A block diagram explains the arrangement of the available DIMM slots. The four memory configuration modes are then explained; which the PRIMERGY configurator also refers to. The third part covers the available DIMM types.

DIMM slots

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

- Group 1 (18 slots): PRIMERGY RX300 S6, TX300 S6, BX924 S2, CX122 S1
- Group 2 (12 slots): PRIMERGY RX200 S6, TX200 S6, BX620 S6, BX922 S2
- Group 3 (9 slots): PRIMERGY BX920 S2
- Group 4 (8 slots): PRIMERGY CX120 S1

The description is based on the systems available in May 2011.



There are always three memory channels per processor. However, the four 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 S6 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 colors in the diagram (black, blue, green) correspond to the colored 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 the 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 in a transparent way 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 5600-based servers.

 © Fujitsu Technology Solutions 2011

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. There are *registered* (RDIMM) and *unbuffered* (UDIMM) modules. Mixed configurations consisting of RDIMMs and UDIMMs are not possible. Due to their simple construction, UDIMMs have a lower maximum capacity. The simpler construction has advantages as far as price and energy consumption are concerned. The following special feature applies:

- UDIMMs are only possible in 1DPC and 2DPC configurations.

In addition to the standard modules that run with 1.5 V, there are energy-efficient *low voltage* (LV) modules for UDIMM and RDIMM which can run with 1.35 V. Mixed configurations consisting of 1.5 and 1.35 V modules are possible, but not recommended. In the event of a mixed configuration all the modules run with 1.5 V. In all other respects, the following general conditions apply when modules are run with 1.35 V:

- 1.35 V is only possible in 1DPC and 2DPC configurations.
- With 2DPC the 1.35 V limits the memory frequency to a maximum of 1066 MHz.

If a configuration with LV-DIMM violates these general conditions, e.g. through a 3DPC configuration or by enforcing 1333 MHz via the BIOS option *Memory Speed* = *Performance* (instead of the default configuration *Auto*), the modules then run with 1.5 V.

Special features also apply for the quad-rank (QR) 16 and 32 GB DIMM, with which the largest memory configurations are achieved:

- QR DIMMs are only possible in 1DPC and 2DPC configurations.
- As a result of their PC3-8500 construction QR DIMMs limit the memory frequency to a maximum of 1066 MHz.
- 2DPC configurations with QR DIMMs have a memory frequency of 800 MHz.

These general conditions for memory frequency cannot be overridden via the BIOS. Likewise it is not possible to start a system in a 3DPC configuration.

The last column in the table shows the relative price differences. The list prices from May 2011 for the PRIMERGY RX300 S6 are used as a basis. The column shows the relative price per GB, standardized to the registered PC3-10600 DIMM, size 4 GB (highlighted as measurement 1). The landscape of relative prices has been subject to constant change since the introduction of the DDR3 memory module. Lower costs for UDIMM, in comparison with RDIMM, have remained. On the other hand, the originally high costs for 8 und 16 GB RDIMMs are in the meantime no longer to be seen. The same applies for the originally more expensive LV version. The new 32 GB DIMM, based on 4Gbit technology, has been added as a higher priced version.

Type			Control	Max. MHz	Rank	Capacity	Rel. Price per GB
UDIMM	DDR3-1333 PC3-10600		unbuffered	1333	2	2 GB	0.7
UDIMM	DDR3-1333 PC3-10600	LV	unbuffered	1333	2	2 GB	0.9
RDIMM	DDR3-1333 PC3-10600		registered	1333	1	2 GB	1.1
RDIMM	DDR3-1333 PC3-10600		registered	1333	1 or 2	4 GB	**1**
RDIMM	DDR3-1333 PC3-10600	LV	registered	1333	1 or 2	4 GB	1.0
RDIMM	DDR3-1333 PC3-10600		registered	1333	2	8 GB	0.9
RDIMM	DDR3-1333 PC3-10600	LV	registered	1333	2	8 GB	0.9
RDIMM	DDR3-1066 PC3-8500		registered	1066	4	16 GB	1.1
RDIMM	DDR3-1066 PC3-8500		registered	1066	4	32 GB	3.5

For logistics reasons and depending on stocks, modules with one or two ranks are supplied for RDIMMs of size 4 GB. The term *Rank* is explained below in the section *Secondary performance influences*.

Depending on the PRIMERGY model, there can be restrictions regarding the availability of individual DIMM types, especially for the new 32 GB DIMM. The current configurator is always decisive. Furthermore, some sales regions can also have restrictions regarding availability.

© Fujitsu Technology Solutions 2011

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 performance percentage disadvantages of between 1% and 5% 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.)

Table 1: Ideal memory sizes

Capa-city	Type	Module size (GB)	Configuration	MHz (max) 1.5 V	MHz (max) 1.35 V	Notes	RX/TX300 S6 BX924 S2	RX/TX200 S6 BX620 / BX922	BX920 S2	CX120 S1	CX122 S1
12 GB	UDIMM	2	1 – 1 – 1 / 1 – 1 – 1	1333	1333	compared to RDIMM price advantage and LV option	✓	✓	✓		✓
24 GB	UDIMM	2	2 – 2 – 2 / 2 – 2 – 2	1333	1066	compared to RDIMM price advantage and LV option	✓	✓			
	RDIMM	4	1 – 1 – 1 / 1 – 1 – 1	1333	1333	possible in BX920 S2	✓	✓	✓	✓	✓
36 GB	RDIMM	4 and 2	2 – 2 – 2 / 2 – 2 – 2	1333	n/a	1st bank 4 GB DIMM 2nd bank 2 GB DIMM	✓	✓			
48 GB	RDIMM	8	1 – 1 – 1 / 1 – 1 – 1	1333	1333		✓	✓	✓	✓	✓
60 GB	RDIMM	8 and 2	2 – 2 – 2 / 2 – 2 – 2	1333	n/a	1st bank 8 GB DIMM 2nd bank 2 GB DIMM	✓	✓			
72 GB	RDIMM	8 and 4	2 – 2 – 2 / 2 – 2 – 2	1333	1066	1333 MHz possible	✓	✓			
	RDIMM	4	3 – 3 – 3 / 3 – 3 – 3	800	n/a	possible in CX122 S1	✓				✓
84 GB	RDIMM	8, 4 and 2	3 – 3 – 3 / 3 – 3 – 3	800	n/a	1st bank 8 GB DIMM 2nd bank 4 GB DIMM 3rd bank 2 GB DIMM	✓				
96 GB	RDIMM	8	2 – 2 – 2 / 2 – 2 – 2	1333	1066	1333 MHz possible	✓	✓			✓
	RDIMM	16	1 – 1 – 1 / 1 – 1 – 1	1066	n/a	possible in BX920 S2	✓	✓	✓		
108 GB	RDIMM	16 and 2	2 – 2 – 2 / 2 – 2 – 2	800	n/a	1st bank 16 GB DIMM 2nd bank 2 GB DIMM	✓	✓			
120 GB	RDIMM	8 and 4	3 – 3 – 3 / 3 – 3 – 3	800	n/a	1st and 2nd bank 8 GB DIMM 3rd bank 4 GB DIMM	✓				
	RDIMM	16 and 4	2 – 2 – 2 / 2 – 2 – 2	800	n/a	1st bank 16 GB DIMM 2nd bank 4 GB DIMM	✓	✓			
144 GB	RDIMM	8	3 – 3 – 3 / 3 – 3 – 3	800	n/a	possible in CX122 S1	✓				✓
	RDIMM	16 and 8	2 – 2 – 2 / 2 – 2 – 2	800	n/a	1st bank 16 GB DIMM 2nd bank 8 GB DIMM	✓	✓			

© Fujitsu Technology Solutions 2011

Capa-city	Type	Module size (GB)	Configuration	MHz (max) 1.5 V	MHz (max) 1.35 V	Notes	RX/TX300 S6 BX924 S2	RX/TX200 S6 BX620 / BX922	BX920 S2	CX120 S1	CX122 S1
192 GB	RDIMM	16	2 – 2 – 2 / 2 – 2 – 2	800	n/a		✓	✓			
	RDIMM	32	1 – 1 – 1 / 1 – 1 – 1	1066	n/a	1066 MHz possible possible in BX920 S2	✓*	✓*	✓*		
204 GB	RDIMM	32 and 2	2 – 2 – 2 / 2 – 2 – 2	800	n/a	1st bank 32 GB DIMM 2nd bank 2 GB DIMM	✓*	✓*			
216 GB	RDIMM	32 and 4	2 – 2 – 2 / 2 – 2 – 2	800	n/a	1st bank 32 GB DIMM 2nd bank 4 GB DIMM	✓*	✓*			
240 GB	RDIMM	32 and 8	2 – 2 – 2 / 2 – 2 – 2	800	n/a	1st bank 32 GB DIMM 2nd bank 8 GB DIMM	✓*	✓*			
288 GB	RDIMM	32 and 16	2 – 2 – 2 / 2 – 2 – 2	800	n/a	1st bank 32 GB DIMM 2nd bank 16 GB DIMM	✓*	✓*			
384 GB	RDIMM	32	2 – 2 – 2 / 2 – 2 – 2	800	n/a		✓*	✓*			

✓* : The introduction of 32 GB DIMMs is done step by step and only in the systems PRIMERGY RX300 S6, TX300 S6, RX200 S6 and BX920 S2. The current configurator is decisive for availability.

Table 2: Classic memory sizes

Capa-city	Type	Module size (GB)	Configuration	MHz (max) 1.5 V	MHz (max) 1.35 V	Notes	RX/TX300 S6 BX924 S2	RX/TX200 S6 BX620 / BX922	BX920 S2	CX120 S1
8 GB	UDIMM	2	1 – 1 – 0 / 1 – 1 – 0	1333	1333	compared to RDIMM price advantage and LV option	✓	✓	✓	✓
16 GB	UDIMM	2	2 – 1 – 1 / 2 – 1 – 1	1333	1066	compared to RDIMM price advantage and LV option	✓	✓		✓
	RDIMM	4	1 – 1 – 0 / 1 – 1 – 0	1333	1333	possible in BX920 S2	✓	✓	✓	
32 GB	RDIMM	8	1 – 1 – 0 / 1 – 1 – 0	1333	1333		✓	✓	✓	
64 GB	RDIMM	8	2 – 1 – 1 / 2 – 1 – 1	1333	1066	1333 MHz possible possible in CX120 S1	✓	✓		✓
	RDIMM	16	1 – 1 – 0 / 1 – 1 – 0	1066	n/a	possible in BX920 S2	✓	✓	✓	
128 GB	RDIMM	16	2 – 1 – 1 / 2 – 1 – 1	800	n/a	Price advantage over 32 GB	✓	✓		
	RDIMM	32	1 – 1 – 0 / 1 – 1 – 0	1066	n/a	possible in BX920 S2	✓*	✓*	✓*	
256 GB	RDIMM	32	2 – 1 – 1 / 2 – 1 – 1	800	n/a		✓*	✓*		

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.

The tables note which maximum memory frequency is possible for the respective configuration. A distinction is made between operation with 1.5 und 1.35 V, as long as the latter can be achieved with LV-DIMM. Otherwise, you can see *n/a* in this column. In addition to the features taken into consideration in the tables,

© Fujitsu Technology Solutions 2011

DPC value and DIMM type, the processor type is also a decisive factor when it comes to an effective frequency. The more powerful processors of the Xeon 5600 generation support a memory frequency of up to 1333 MHz, and the weaker ones a frequency of up to 1066 MHz. An exact list and classification of the available Xeon models follows below. Effective memory frequency is the minimum value according to the table and processor class.

The last four or five columns of the tables show the PRIMERGY models for which the respective configuration is possible.

The explanations in the section *The effects on memory performance* should enable you to create memory configurations for those configurations not covered here.

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

Asymmetric memory configurations

Not all the systems enable symmetric memory configuration in all configuration versions based on their form factor. The diagram in the section *Memory architecture* shows the asymmetric arrangement of the DIMM slots for the PRIMERGY BX920 S2: 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 recommended configurations.

Configurations which divide the total capacity of memory into two identical halves are possible in the PRIMERGY BX920 S2 up to a capacity of 192 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.

There is more memory on the left than on the right in the configurations in the next table. The excess is between a quarter and a third of the total capacity. For maximum half of the excess, i.e. an eighth to a sixth, there is "remote" access via the QPI link (seen statistically). In such cases of moderate asymmetry a performance disadvantage of about 2-3% 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 [L4] on PRIMERGY BX920 S1 under Windows Server 2008 and SQL Server 2008.

PRIMERGY BX920 S2
Table 3: Asymmetric configuration

Capa-city	Type	Module size (GB)	Configuration	MHz (max) 1.5 V	MHz (max) 1.35 V	Notes
36 GB	RDIMM	4	2 – 2 – 2 / 1 – 1 – 1	1333	1066	
72 GB	RDIMM	8	2 – 2 – 2 / 1 – 1 – 1	1333	1066	
128 GB	RDIMM	16	2 – 2 – 1 / 1 – 1 – 1	800	n/a	Price advantage over 32 GB DIMM
144 GB	RDIMM	16	2 – 2 – 2 / 1 – 1 – 1	800	n/a	
256 GB	RDIMM	32	2 – 2 – 1 / 1 – 1 – 1	800	n/a	
288 GB	RDIMM	32	2 – 2 – 2 / 1 – 1 – 1	800	n/a	

© Fujitsu Technology Solutions 2011

The effects on memory performance

This section explains the factors 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 measuring tools

Measurements were made using the benchmarks STREAM and SPECint_rate_base2006.

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.

SPECint_rate_base2006

The Benchmark SPECint_rate_base2006 was added as a model for commercial application performance. It is part of SPECcpu2006 [L5] from Standard Performance Evaluation Corporation (SPEC). SPECcpu2006 is the industry standard for measuring system components processors, memory hierarchy and compiler. According to the large volume of published results and their intensive use in sales projects and technical investigations this is the most important benchmark in the server field.

SPECcpu2006 consists of two independent suites of individual benchmarks, which differ in the predominant use of *integer* and *floating-point* operations. The integer part is representative for commercial applications and consists of 12 individual benchmarks. The floating-point part is representative for scientific applications and contains 17 individual benchmarks. The result of a benchmark run is in each case the geometric mean of the individual results.

A distinction is also made in the suites between the *speed* run with only one process and the *rate* run with a configurable number of processes working in parallel. The second version is evidently more interesting for servers with their large number of processor cores and hardware threads.

And finally a distinction is also made with regard to the permitted compiler optimization: for the *peak* result the individual benchmarks may be optimized independently of each other, but for the more conservative *base* result the compiler flags must be identical for all benchmarks, and certain optimizations are not permitted.

This explains what SPECint_rate_base2006 is about. The integer suite was selected, because commercial applications predominate in the use of PRIMERGY servers.

A measurement that is compliant with the regulations requires three runs, and the mean result is evaluated for each individual benchmark. This was not complied with in the technical investigation described here. To simplify matters only one run was performed at all times.

© Fujitsu Technology Solutions 2011

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. Reasons that are opposed to the optimal values of 1333 MHz and 3-way have already been mentioned in the previous sections: very large memory configurations or energy saving for the reduction in memory frequency and the customer's request for classic memory sizes such as 16, 32, 64 GB for a 2-way interleave.

Planning a memory configuration should first of all involve the planning of these parameters.

The section ends with memory performance under redundancy (sparing and mirroring).

Effective frequency of the memory

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

- Processor type. The processors are classified according to the following table. The column with the interesting feature here is marked in gray. Stronger models support the maximum 1333 MHz, and weaker ones the 1066 MHz.
- DIMM type. Both UDIMM and RDIMM usually support the maximum 1333 MHz. The quad-rank (QR) 16 and 32 GB RDIMMs with a maximum of 1066 MHz are the exception.
- DPC value (DIMM Per Channel). The cases of 1DPC and 3DPC are simple: 1DPC supports 1333 MHz, and 3DPC always limits the frequency to 800 MHz. It is worth recalling here that 3DPC is not possible with UDIMM, 1.35 V *low voltage* (LV) operation and QR modules. With 2DPC a frequency of 1333 MHz is usually possible with the following exceptions: LV operation means 1066 MHz, and QR modules or mixed configurations with these modules run with 800 MHz. If the six channels are not configured the same, the largest DPC value is decisive.

Category	Xeon type	#cores	GHz	L3 Cache (MB)	QPI (GT/s)	Max. memory MHz	TDP (Watt)
Advanced	X5690	6	3.46	12	6.4	1333	130
	X5687	4	3.60	12	6.4	1333	130
	X5672	4	3.20	12	6.4	1333	95
	X5675	6	3.06	12	6.4	1333	95
	X5660	6	2.80	12	6.4	1333	95
	X5650	6	2.66	12	6.4	1333	95
Standard	E5649	6	2.53	12	5.9	1333	80
	E5645	6	2.40	12	5.9	1333	80
	X5647	4	2.93	12	5.9	1066	130
	E5620	4	2.40	12	5.9	1066	80
Low Power	L5640	6	2.26	12	5.9	1333	60
	L5630	4	2.13	12	5.9	1066	40
	L5609	4	1.86	12	4.8	1066	40
Basic	E5607	4	2.26	4	4.8	1066	80
	E5606	4	2.13	4	4.8	1066	80
	E5603	4	1.60	4	4.8	1066	80

The highest possible timing is desirable. However, the minimum of the 3 factors is effective: the worst value determines the timing of the configuration. The timing is defined as standard for the system and not per processor.

An example explains the mechanism. A PRIMERGY RX300 S6 is fully equipped with processors of the type Xeon E5620 and with 18 RDIMMs of size 4 GB. The processor would support 1066 MHz, and the DIMM type would even support 1333 MHz, but the 3DPC full configuration limits the timing to 800 MHz. Thus, the effective timing is 800 MHz. The same memory configuration of 72 GB could also be achieved if 8 GB modules were used for the first bank and 4 GB ones for the second, and if the third bank were to remain empty. This 2DPC configuration would now have a processor-related timing of 1066 MHz.

© Fujitsu Technology Solutions 2011

The following diagram shows what this difference in memory frequency means for application performance. The measurements were made using the benchmarks STREAM (red values; top value in each cell) and SPECint_rate_base2006 (green; bottom value). The green value is representative for commercial applications. The one but last line of the table is decisive for the example just provided with the Xeon E5620 processor. The difference in performance is 3%.

STREAM is synonymous with memory bandwidth. The differences shown here represent the upper limit that is only reached by applications in exceptional cases. The 12 components of SPECint_rate_base2006 include for example a test case (*libquantum*) that behaves like STREAM, i.e. the upper limit for differences in performance (20% in the example) is in fact reached.

The diagram divides the Xeon 5600 models into four classes; not just into the two classes which correspond to the maximum memory frequency of 1333 or 1066 MHz. The analysis on which this section is based suggested this classification. Differentiation according to the processor core frequency is not necessary, but at least the QPI frequency affects the interaction between processor cores and memory system.

The decline in the influence of memory frequency on application performance is clear to see: the more powerful the processor model, the greater the influence. This observation will be repeated in the following section on interleaving. And is thus a key statement here.

The diagram shows relative performance. Please see the Performance Reports of the individual PRIMERGY systems for the absolute values of the benchmarks STREAM and SPECint_rate_base2006, which correspond to the 1.00 reference points in the diagram. Optimal memory configurations are always used for the measurements described there.

Relative Performance for Different Memory Speeds
Memory Bandwidth (STREAM)
Commercial Application Performance (SPECint_rate_base2006)

QPI	Max Mem MHz	CPU Models	Effective Memory MHz		
			1333	1066	800
			Max Performance	Energy Efficiency	Max Capacity
6.4	1333	X5690 X5687 X5675 X5660 X5650	1.00 1.00	0.84 0.98	0.62 0.91
5.9	1333	E5649 E5645 L5640	1.00 1.00	0.94 1.00	0.72 0.95
5.9	1066	X5647 E5620 L5630	N/A	1.00 1.00	0.80 0.97
4.8	1066	E5607 E5606 E5603 L5609	N/A	1.00 1.00	0.95 0.99

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. All the configurations that were listed in the section *Performant memory configurations* in the first table ("Ideal memory sizes") are 3-way interleaved.

Despite the "recommendation" the ideal situation frequently cannot be reached, for example when classic memory configurations, such as 16, 32, 64 GB, are requested. Then the configurations that are listed in the section *Performant memory configurations* in the second table ("Classic memory sizes") come into existence. They are all 2-way interleaved. A closer look at that table shows two schemata for the 2-way interleave. The first schema shows that the third memory channel of the processors is not used, and that the two others are configured with the same capacity. Although all three channels are in use in the second schema, the unequal capacity per channel prevents the 3-way interleave. A detailed explanation as to how the 2-way interleave comes about follows at the end of this section.

The structure of the diagram is as in the previous section. Of the load profiles considered STREAM (red; top value) and SPECint_rate_base2006 (green; bottom value) the green case should be regarded as the average value for commercial applications and the red one as an extreme value that is achieved in exceptional situations. The loss in performance associated with the 2-way interleave of between 1% and 5% on average, depending on the processor model, is usually not a problem.

Relative Performance for Different Interleaving Levels
Memory Bandwidth (STREAM)
Commercial Application Performance (SPECint_rate_base2006)

QPI	Max Mem MHz	CPU Models	Effective Interleaving		
			3-way	2-way	1-way
			Max Performance	Classical Memory Capacities	*Discouraged*
6.4	1333	X5690 X5687 X5675 X5660 X5650	1.00 1.00	0.70 0.95	0.39 0.76
5.9	1333	E5649 E5645 L5640	1.00 1.00	0.77 0.97	0.43 0.82
5.9	1066	X5647 E5620 L5630	1.00 1.00	0.71 0.97	0.39 0.84
4.8	1066	E5607 E5606 E5603 L5609	1.00 1.00	0.83 0.99	0.45 0.88

© Fujitsu Technology Solutions 2011

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 is in no sensible relationship to the performance capability of the processors. If need be, configurations with the weakest and most cost-effective processors can be excluded from this judgment; for example, if the memory configuration is reduced at the customer's request to the absolute minimum of only one DIMM per processor.

This differentiation according to the performance of the processor also repeats the key statement that has already been made: the more powerful the processor model, the greater the influence.

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.

© Fujitsu Technology Solutions 2011

Memory performance under redundancy

It stands to reason for the section on interleaving to be followed by some statements about memory performance under redundancy. Because the tests needed for DIMM sparing come under the first schema of 2-way configurations, which has just been dealt with and in which the third memory channel of the processors is not used. Such configurations offer space for the sparing modules, whose presence does not change performance. Thus, the following diagram is - apart from the column on the far right on mirroring - identical with the diagram last shown.

Equating mirroring with the 1-way interleave, however, does not apply. In mirroring the first two memory channels per processor are identically configured and the third channel is empty. The operating system sees an address space that only corresponds to the first channel, i.e. half of the actual configuration. In every write process the hardware ensures that the first channel is automatically mirrored to the second channel. However, during read the mirror can also be used, which is why the performance under mirroring lies between the performance of the 1-way and 2-way interleave.

The impact of sparing on performance with a loss of between 1% and 5% depending on the processor model is, as in the case of the 2-way interleave, usually not a problem. During mirroring it is necessary for the user to weigh up fail-safety against a loss in performance of about 10% .

Relative Performance for Redundant Configurations
Memory Bandwidth (STREAM)
Commercial Application Performance (SPECint_rate_base2006)

QPI	Max Mem MHz	CPU Models	Redundancy		
			Disabled [1]	Sparing	Mirroring
6.4	1333	X5690 X5687 X5675 X5660 X5650	1.00 1.00	0.70 0.95	0.57 0.87
5.9	1333	E5649 E5645 L5640	1.00 1.00	0.77 0.97	0.60 0.91
5.9	1066	X5647 E5620 L5630	1.00 1.00	0.71 0.97	0.57 0.92
4.8	1066	E5607 E5606 E5603 L5609	1.00 1.00	0.83 0.99	0.59 0.95

[1] Redundancy disabled and all three memory channels per CPU populated

© Fujitsu Technology Solutions 2011

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, *unbuffered* DIMM modules (UDIMM) are also 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	Rel. Price per GB
UDIMM	DDR3-1333 PC3-10600		unbuffered	1333	2	2 GB	0.7
UDIMM	DDR3-1333 PC3-10600	LV	unbuffered	1333	2	2 GB	0.9
RDIMM	DDR3-1333 PC3-10600		registered	1333	1	2 GB	1.1
RDIMM	DDR3-1333 PC3-10600		registered	1333	1 or 2	4 GB	1
RDIMM	DDR3-1333 PC3-10600	LV	registered	1333	1 or 2	4 GB	1.0
RDIMM	DDR3-1333 PC3-10600		registered	1333	2	8 GB	0.9
RDIMM	DDR3-1333 PC3-10600	LV	registered	1333	2	8 GB	0.9
RDIMM	DDR3-1066 PC3-8500		registered	1066	4	16 GB	1.1
RDIMM	DDR3-1066 PC3-8500		registered	1066	4	32 GB	3.5

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, 2 or 4 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 4 or 8 bit. Or there are two or four 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 and quad-rank DIMM is first the greater capacity, as can be seen in the table.

A second advantage of dual and quad-rank modules is the physical reason already discussed. Memory cells are arranged in two dimensions. A line is opened and then a column item is read in this line. While the line (more commonly called 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 and quad-rank modules, the probability of accessing an open page increases.

© Fujitsu Technology Solutions 2011

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	
X5690	RDIMM 1333 MHz	8 GB	2	1 - 1 - 1 / 1 - 1 - 1	41.6
X5690	RDIMM 1333 MHz	2 GB	1	1 - 1 - 1 / 1 - 1 - 1	35.5

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 and quad-rank modules. With a configuration with 2 GB modules, the performance disadvantage of realistically 1-2% with an odd number of ranks per channel is an additional reason for preferring dual-rank UDIMM modules.

Access to remote memory

Solely a local memory was used in the previously described tests with the benchmarks STREAM and SPECint_rate_base2006, i.e. the CPU accesses DIMM modules of its own memory channels. Modules of the neighboring CPU are not accessed via the QPI link. This situation is representative, insofar as it also exists for the majority of memory accesses of real applications thanks to NUMA support in the operating system and system software.

The following table shows the effect of the opposite case for STREAM and a number of standard benchmarks, which are representative for commercial applications. The exclusive use of remote memory was enforced by measures such as explicit process binding. The table shows the deterioration in the measurement result in per cent.

Benchmark	Effect of the exclusive use of remote memory
STREAM Triad	-49%
SPECint_rate_base2006	-13%
SPECint_rate2006	-14%
SPECjbb2005	-20%

With STREAM the bandwidth of the QPI link between the processors becomes the result-determining bottleneck. The deterioration of the other benchmarks is primarily caused by the approx. 50% higher latency of the individual access. The use of remote memory accordingly means a deterioration of between 10 and 20% for commercial applications.

These originally impractical findings are helpful when estimating what effect disabling NUMA support in the BIOS has. In this case, the physical address area is set up with a fine-mesh interleave via the memory modules of both processors. Then 50% of the accesses of an application are to local memory and 50% are to a remote one. And with commercial applications this halves the expected deterioration interval to between 5 and 10%. This is approximately the effect of disabling NUMA support.

The effect of an asymmetric memory configuration can be estimated accordingly, as was discussed above in the example of the PRIMERGY BX920 S2. The forecast made there of a deterioration of between 2 und 3% for asymmetric configuration results from the mentioned deterioration interval for remote access only and from the consideration that this situation occurs, statistically speaking, in one sixth of the accesses at most.

© Fujitsu Technology Solutions 2011

Literature

[L1] PRIMERGY Systems

http://ts.fujitsu.com/primergy

[L2] PRIMERGY Performance

http://de.ts.fujitsu.com/products/standard_servers/primergy_bov.html

[L3] STREAM Benchmark

http://www.cs.virginia.edu/stream

[L4] OLTP-2 Benchmark

http://docs.ts.fujitsu.com/dl.aspx?id=e6f7a4c9-aff6-4598-b199-836053214d3f

[L5] SPECcpu2006 Benchmark

http://docs.ts.fujitsu.com/dl.aspx?id=1a427c16-12bf-41b0-9ca3-4cc360ef14ce

Contact

FUJITSU Technology Solutions

Website: http://ts.fujitsu.com/

PRIMERGY Product Marketing

mailto:PRIMERGY-PM@ts.fujitsu.com

PRIMERGY Performance and Benchmarks

mailto:primergy.benchmark@ts.fujitsu.com

All rights reserved, especially industrial property rights. Delivery subject to availability; right of technical modifications reserved No liability or warranty assumed for completeness, validity and accuracy of the specified data and illustrations. Designations may be trademarks and/or copyrights of the respective manufacturer, the use of which by third parties for their own purposes may infringe the rights of such owner.
For more details see http://ts.fujitsu.com/terms_of_use.html

2011-06-06 WW EN

Copyright © Fujitsu Technology Solutions GmbH 2011