# WHITE PAPER
# FUJITSU PRIMERGY SERVERS
# MEMORY PERFORMANCE OF XEON E7-8800/4800/2800 (WESTMERE-EX) BASED SYSTEMS

The Xeon E7-8800/4800/2800 (Westmere-EX) based PRIMERGY 4 and 8-socket models are the second generation since the paradigm change in connecting the main memory with the Xeon 7500 (Nehalem-EX): QuickPath Interconnect (QPI) instead of Front Side Bus (FSB). The new architecture increases the influence of memory performance on system performance. This white paper explains these influences and gives recommendations for performant memory configurations.

**Version**

1.0
2011-08-19

**Contents**

# Summary

The Xeon E7-8800/4800/2800 (Westmere-EX) based PRIMERGY 4 and 8-socket models are the second generation since the paradigm change in connecting the main memory with the Xeon 7500 (Nehalem-EX): QuickPath Interconnect (QPI) instead of Front Side Bus (FSB). The configuration of powerful systems deals with the following features:

- NUMA architecture.
- Interleaving between the two memory controllers of a Westmere-EX processor.
- Interleaving between the two DDR3 memory channels of a "Mill Brook 2" memory buffer.
- Memory frequency with 1066, 978 or 800 MHz.

In NUMA architecture the DIMM slots are directly assigned to the processors ("local memory"). All the processors should be configured with DIMM strips, ideally in an identical way. The BIOS default *NUMA Optimization* = *Enabled* should not be changed.

In the PRIMERGY RX600 S6 the DIMM slots are located on up to eight memory boards (two boards per processor). The basic configuration of the system only contains two boards, which means you should ensure a sufficient number of boards during configuration and ordering.

The interleaving is the most important performance factor for the Westmere-EX based systems. Interleaving takes place on the two levels mentioned above in the list. The number of DIMMs configured per processor is the decisive factor here. Westmere-EX based systems have 16 DIMM slots per processor.

This document divides recommendable configurations into three classes. It is necessary to configure 8 or 16 DIMMs of the same type per processor for optimal performance (class 1) and distribute them over two memory boards per processor (in the case of the PRIMERGY RX600 S6). Class 2 contains configurations with at least 4 DIMMs per processor, and again with two memory boards per processor in the case of the PRIMERGY RX600 S6. Class 3 contains PRIMERGY RX600 S6 configurations with only one board and at least 4 DIMMs per processor. Mixed configurations with DIMMs of various sizes also occur in classes 2 and 3 in particular. Operation with less than 4 DIMMs per processor is not to be recommended.

In comparison with class 1, class 2 for commercial applications means an average loss of between 5% and 9% with the PRIMERGY RX600 S6, and between 4% and 6% with the PRIMERGY RX900 S2. The more performant the processor, the greater the loss. Class 3, which is only relevant for the PRIMERGY RX600 S6, means a further loss in the order of 4% in comparison with class 2.

When configuring high-performance systems it is sufficient to bear the circumstances shown by this three-stage classification in mind. In contrast to the Xeon 5600 based 2-socket systems, it is not necessary to take the effects of memory configuration on memory frequency into consideration. The frequency with 1066, 978 or 800 MHz only follows from the processor type for the Westmere-EX based servers. The frequency for the PRIMERGY RX900 S2 is even 1066 MHz at all times.

With the PRIMERGY RX600 S6 the memory frequency foreseen for the respective processor type can be reduced to 978 or 800 MHz in the BIOS in order to save energy. The loss is negligible for a reduction to 978 MHz and is in the order of 5% for a reduction to 800 MHz.

# Document history

### *Version 1.0*

Initial version

# Introduction

32nm manufacturing technology is the main innovation in Intel Xeon E7-8800/4800/2800 (Westmere-EX) processors, with which the current generation of 4 and 8-socket PRIMERGY rack servers is equipped. Compared with the predecessor generation of Xeon 7500 (Nehalem-EX), which was manufactured in 45nm, this technology offers up to ten cores per processor, an increase in the L3 cache from 24 to 30 MB and an increase in the clock rates. These features result in an increase in performance to the order of 40%.

Both processor 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 Westmere-EX based generation into account.

The operating system takes NUMA into consideration when allocating the physical memory and when scheduling processes. The mechanisms function optimally if the total quantity of RAM is distributed evenly across all processors. This is the fundamental rule for the configuration of powerful systems.

Other regulations and recommendations concern the distribution of a given amount of memory modules over the maximum 16 DIMM (Dual Inline Memory Module) slots per processor. These recommendations are the subject of this white paper. The performance features and effects of various memory configurations are to be named and quantified.

This is about similar things, like with the Xeon 5600 (Westmere-EP) based 2-socket servers [L5], but the emphasis is clearly different. The significance of different timing diminishes. As with the 2-socket servers, it is available in three levels. These, however, are closer to each other and do not depend on the positioning of the memory modules. The memory timing follows solely from the processor model used. Interleaving comes to the forefront here. In comparison with the 2-socket servers it is multi-level and thus more complex. The increased complexity is a consequence of the design goals for this server class: more DIMM slots per processor for larger memory configurations and better RAS (Reliability, Availability, Serviceability) features.

This white paper provides first an overview of the memory architecture of the Westmere-EX 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 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 section The effects on memory performance based on results with the benchmarks STREAM and SPECint_rate_base2006.

The configuration of powerful systems is easier with the 4 and 8-socket servers than with the 2-socket servers, despite the higher complexity of connecting memory modules to the processors. The number of available DIMM versions is smaller, as there are for example no *unbuffered* modules. Effects on memory timing do not need to be considered. The recommended memory configurations fall with regard to their performance features in three classes between optimal and acceptable. Numerous examples of these classes from a wide range of main memory configurations are listed below.

# Memory architecture

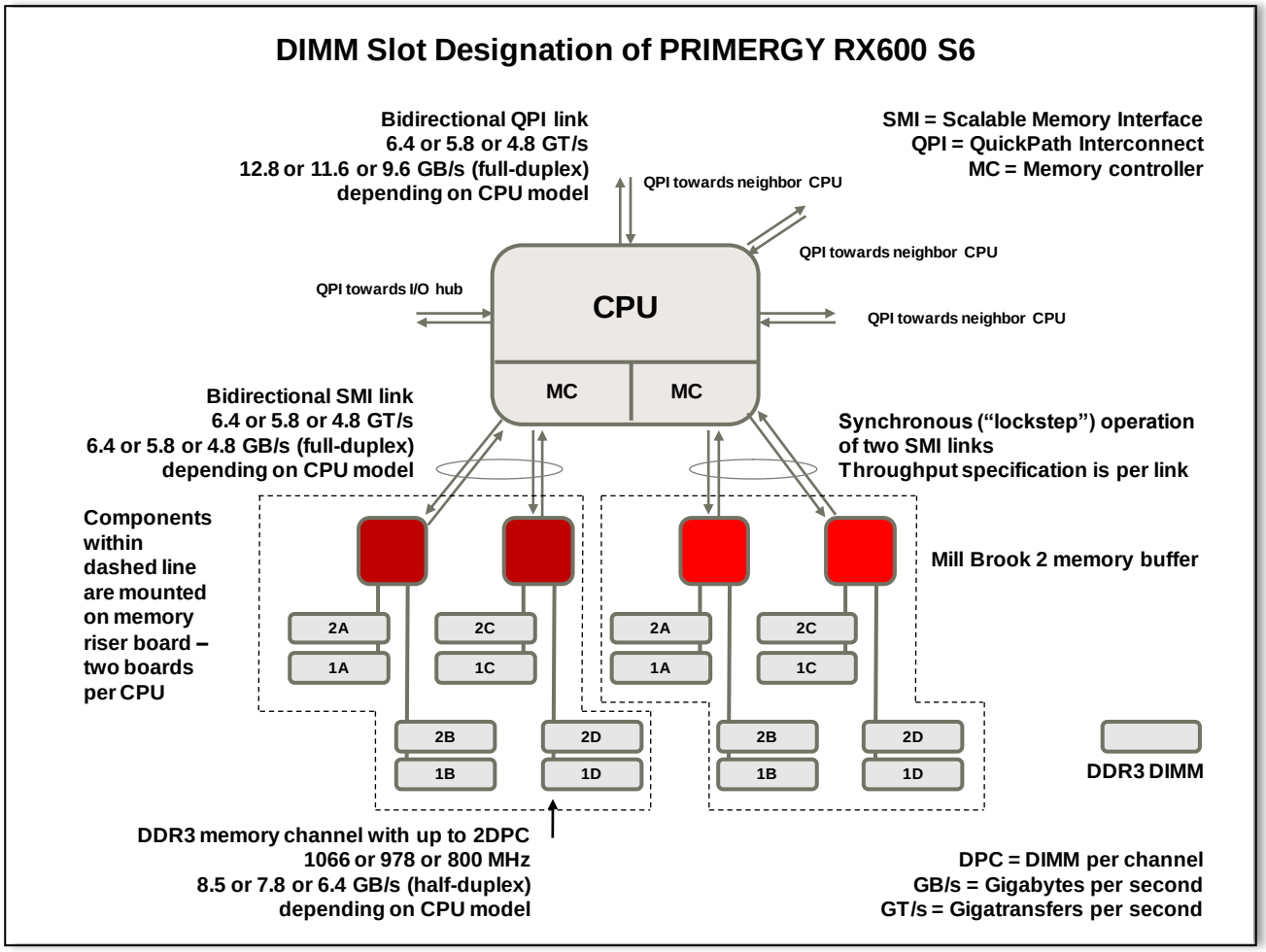
This section provides an overview of the memory system in three parts. Block diagrams explain the arrangement of the DIMM slots and their connections to the processors. The next section then deals with the BIOS parameters that affect the main memory. The third part covers the available DIMM types.

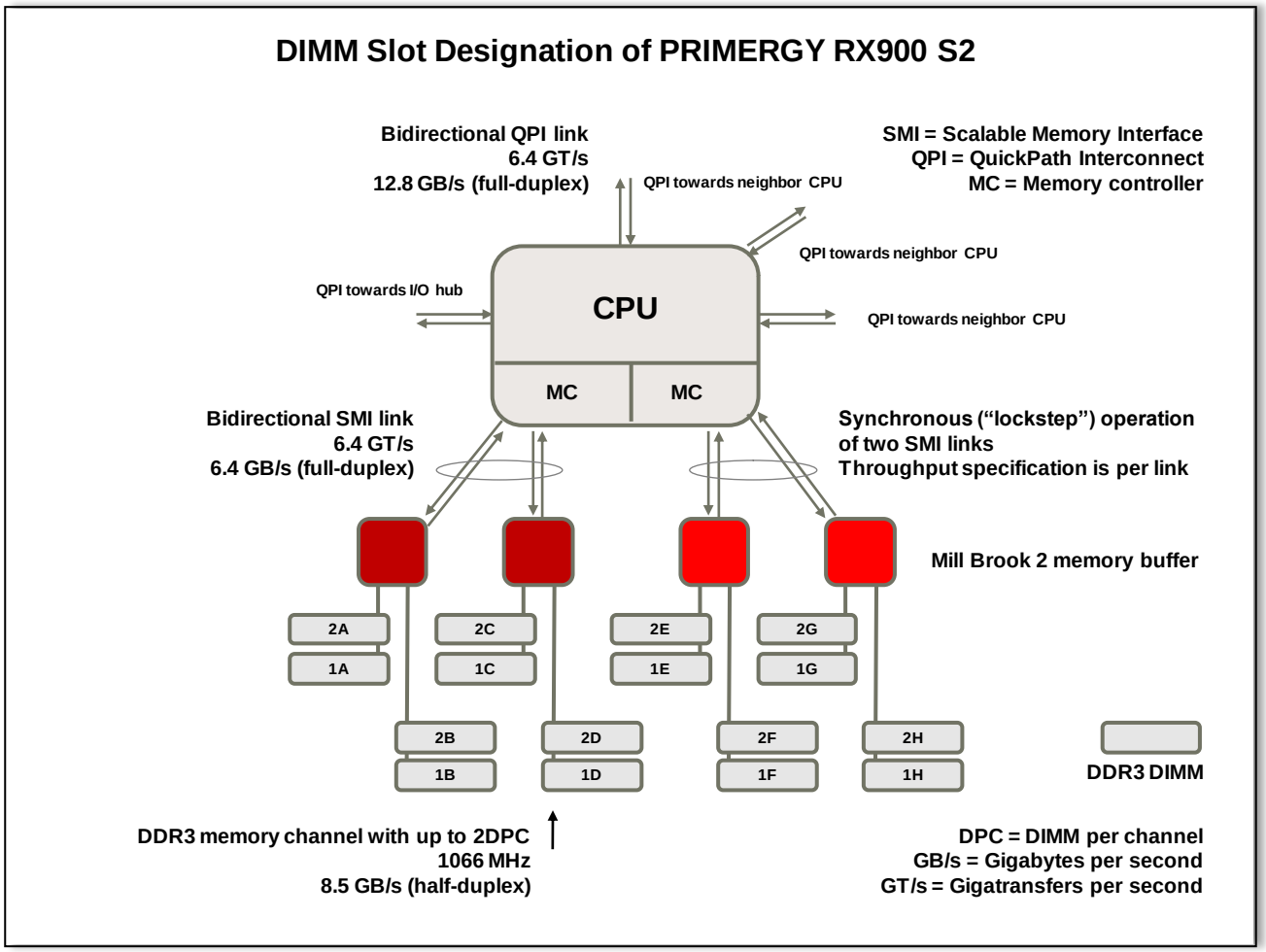
## DIMM slots and connection

The two diagrams below show the memory connection from the perspective of the individual Westmere-EX processor. The first diagram concerns the PRIMERGY RX600 S6, and the second one the PRIMERGY RX900 S2. The rough structure of the diagrams is identical.

Each processor has two memory controllers integrated in the chip. Each controller is connected to two "Mill Brook 2" memory buffers via bidirectional, serial SMI (Scalable Memory Interface) links. There are two DDR3 memory channels, each with two DIMM slots, behind each memory buffer. Thus, there is a total of 16 DIMM slots per processor.

The ellipses in the diagrams indicate that the two SMI links of a controller are in lockstep, i.e. each individual memory access (normally with a block size of 64 bytes) is done synchronously via both SMI links and memory buffers. The 64-byte block is split over both buffers and their assigned DIMM strips. The reason for doing this is improved error recognition through extended ECC. The memory configuration behind two buffers in lockstep must be identical as regards DIMM types and positioning. This strict rule reduces the variety of conceivable memory configurations considerably. Due to this rule it is always necessary to proceed in pairs with the configuration.






The "Mill Brook 2" used with the Westmere-EX processor generation has in comparison with the predecessor "Mill Brook 1" (Nehalem-EX) two advantages. Firstly, energy-saving low voltage (LV) operation is possible, and secondly "Mill Brook 2" is prepared for future 32 GB DIMMs.

Apart from the previously mentioned components, both diagrams show the four QPI links that connect the individual processor to the outside world. A link establishes the connection with an I/O hub (IOH) and its I/O components. Three other links connect the neighboring processors and in particular their memory resources.

The memory connections of the PRIMERGY RX600 S6 and PRIMERGY RX900 S2 differ first of all in the mounting of the "Mill Brook 2" and DIMM slots, as well as in the naming scheme of the slots. The PRIMERGY RX600 S6 is equipped with two memory boards per processor. The processor itself is located on the mainboard. Each memory board has two memory buffers and eight DIMM slots. The naming of the slots is repeated on each memory board. The mounting of the memory components on memory boards is indicated in the first diagram by the dotted line. Operation with only one memory board per processor is permitted. The minimum configuration of the PRIMERGY RX600 S6 consists for example of two processors with one memory board each.

The following sequence should be observed when configuring a memory board: first the slot pair 1B-1D is assigned, then 1A-1C, followed by 2B-2D and finally 2A-2C. This stipulation ensures optimal utilization of the memory channels and is relevant to performance.

The PRIMERGY RX900 S2 has each processor with all its memory components on a separate so-called CPU memory riser (CPUMEMR). This enables all 16 DIMM slots to be named consecutively. There is as regards admissible memory configurations a great difference to the PRIMERGY RX600 S6. Configurations, which correspond to operations with only one memory board with the PRIMERGY RX600 S6, are not permitted with the PRIMERGY RX900 S2. The configuration for the PRIMERGY RX900 S2 must be in groups of four DIMMs, and each group must be distributed over all four memory buffers of the processor. The configuration sequence here begins with the group 1B-1D-1F-1H, then 1A-1C-1E-1G, etc. The rule-of-four simplifies the performance review below for the PRIMERGY RX900 S2, because operation with only

one active memory controller means a certain performance disadvantage. This disadvantage can be disregarded for the PRIMERGY RX900 S2.

The question of the timing and bandwidths of the memory components can also be disregarded for this system. The diagrams contain these values, and with the PRIMERGY RX900 S2 there is only one option for each resource: the Xeon E7-8800 product family that is only used in this system was designed in that way by Intel.

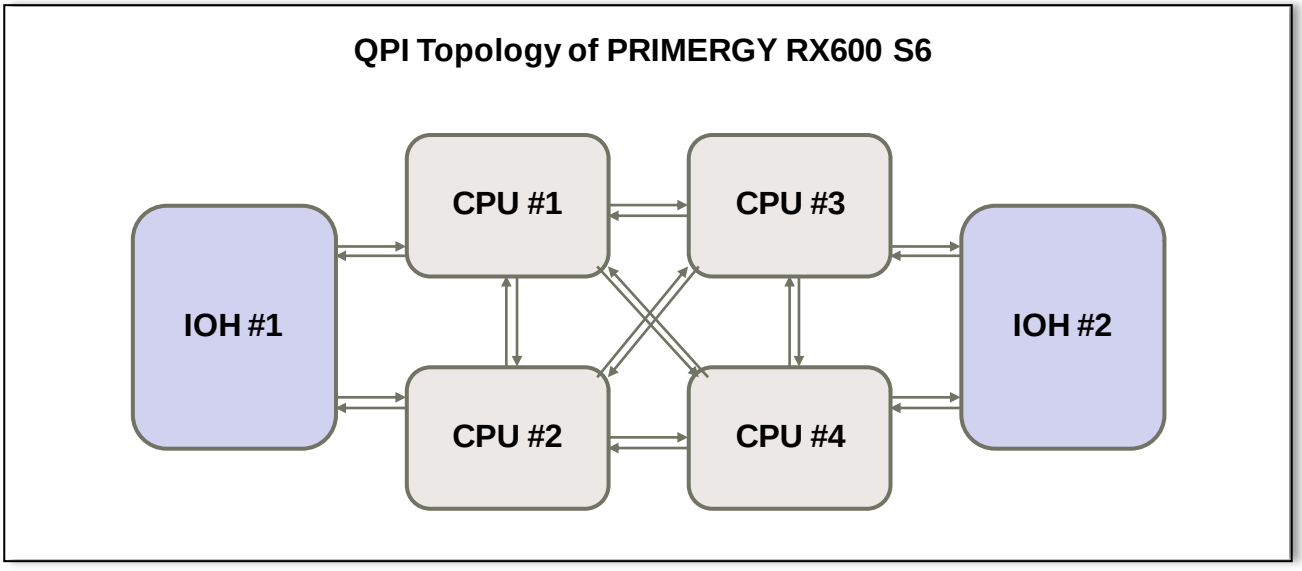
In the PRIMERGY RX600 S6 diagram the bandwidth specifications for the components QPI, SMI and the DDR3 channel are to be understood in such a way that there are three options - depending on the processor model of the families Xeon E7-8800, E7-4800 and E7-2800. The QPI and SMI links run with 6.4 GT/s (giga transfers per second) and the main memory with 1066 MHz in the most powerful processors. QPI and SMI run with 5.8 GT/s and the main memory with 978 MHz in processors of medium capacity. And in the low-cost processors QPI and SMI run with 4.8 GT/s and the main memory with 800 MHz. The QPI and SMI links are bidirectional, and the bandwidths that result from the 2-byte (QPI) or 1-byte (SMI) data path widths apply per direction. This feature of data transfer is referred to as full-duplex. In DDR3 channels read and write accesses have to share the 8-byte wide data paths, hence the name half-duplex here.

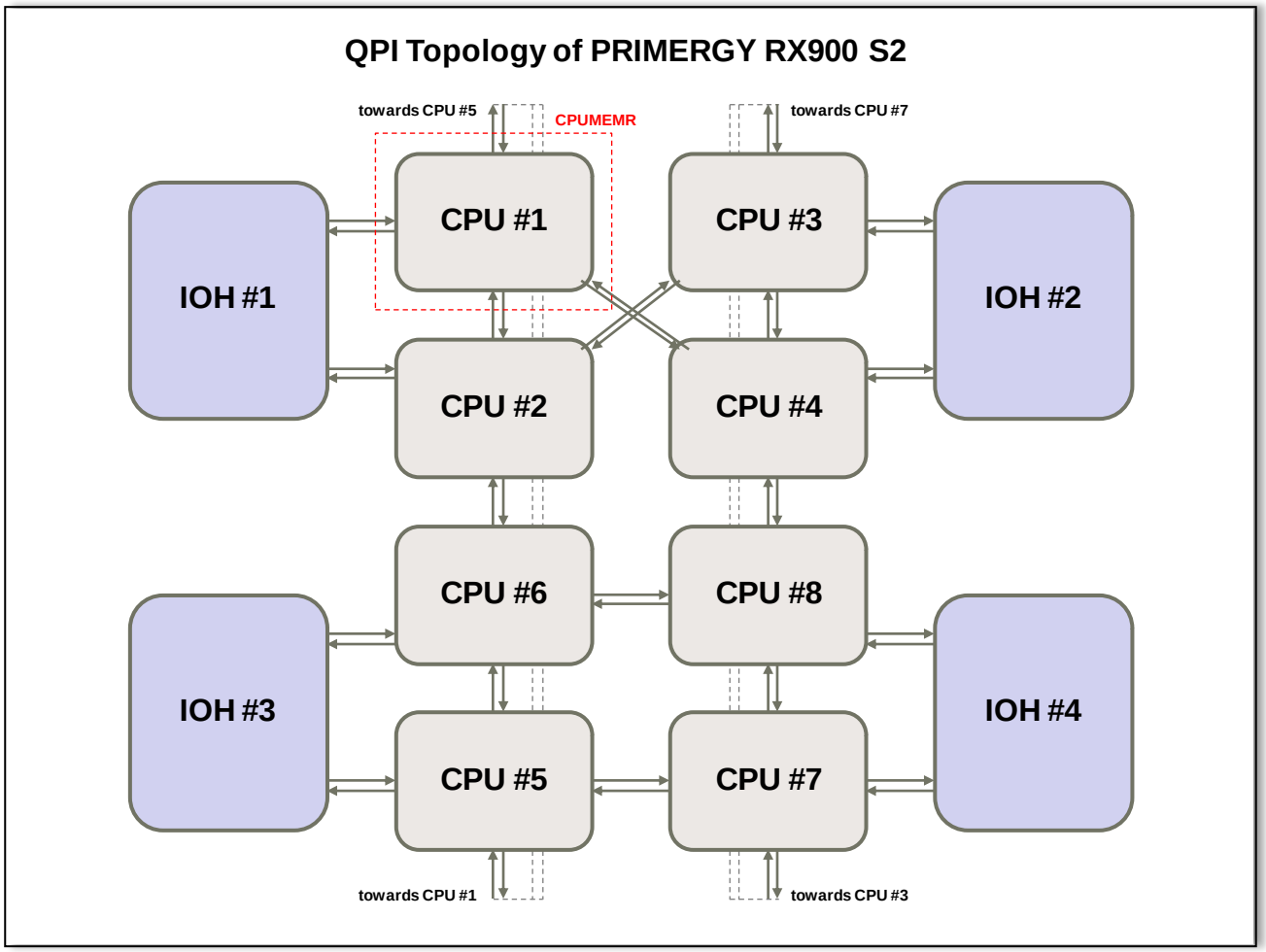
At the end of this section the two diagrams below show the QPI topologies of the PRIMERGY RX600 S6 and the PRIMERGY RX900 S2, i.e. the networking of the processors and their related memory components. For the sake of more clarity, the SMI links, memory buffers and DIMM slots that have already been dealt with are now omitted. The components assigned to the IOHs, i.e. PCIe, SAS and similar interfaces, are also omitted because they are not relevant for the topic of memory.

All the processors are connected with each other in the PRIMERGY RX600 S6. This means that depending on the position of the processors there are no latency differences during access to the remote memory via the QPI network. The diagram also shows why processors 1 and 3 are used instead of 1 and 2 for the minimal configuration of the PRIMERGY RX600 S6. Both IOHs of the system, and thus all the peripherals, are then available.

The principle of the processor that is connected to all its neighbors naturally cannot be continued in the 8-socket server PRIMERGY RX900 S2. Of the seven possible neighbors QPI connections are only available for three. These three can act as brokers if communication takes place with a processor that is not directly connected. Only one broker is at most necessary. The latency of such accesses is higher than in the case of direct coupling. This addition is justifiable, because local access predominates in the software-assisted NUMA architecture.

The diagram for the PRIMERGY RX900 S2 shows how the QPI networking in this system was selected on the basis of its form factor and detailed functional features. A variant of this topology was chosen in the PRIMEQUEST 1800E2, in which the ability to be partitioned is an essential design feature. However, the principle of so-called *glueless design*, which does without additional hardware components for the coupling of the processors and memory components, is the same.






The diagrams for both systems show the key role of the processor chips in the networking of the entire system. If a maximum configuration does not exist, DIMM slots that are assigned to missing processors cannot be used.

## BIOS parameters

Additional features of the memory architecture can best be explained on the basis of the assigned BIOS parameters.

### NUMA Optimization

The parameter *NUMA Optimization* defines whether the physical address space is made up of segments from the local memory only. The operating system is informed about the structure of the address space, i.e. the locality assignment between address areas and processors and can provide processes with performant local memory. This setting (*NUMA Optimization* = *Enabled*) should normally be used. The alternative, a finely woven spreading of the address space across the existing processors, is reserved for special applications, for example in the field of scientific computing, and is only dealt with briefly at the end of this white paper.

The *NUMA Optimization* parameter exists in the PRIMERGY RX600 S6 and PRIMERGY RX900 S2 with the same syntax and meaning.

### Redundancy

The options Disabled, Sparing, Intrasocket Mirror and Intersocket Mirror exist for the *Redundancy* parameter in the PRIMERGY RX600 S6. An unused DIMM pair – the fact that configuration is in pairs at all times on account of lockstep is worth mentioning here – is the reserve for the failure of an active pair for Sparing. And with Intrasocket Mirror the memory of the first memory controller of a processor is mirrored to the second controller of the same processor. The mirroring to a controller of another processor is done using Intersocket

Mirror. A detailed explanation of these RAS mechanisms and their general conditions are not within the framework of this white paper. However, memory performance under redundancy is the subject of a subsection in the section The effects on memory performance.

Only the Mirroring option is available in the PRIMERGY RX900 S2. This mirroring is implemented as Intrasocket MIrroring in the minimal configuration with four processors. In larger configurations it is a matter of an Intersocket Mirror, but with different features in comparison with the PRIMERGY RX600 S6. The entire memory contents of a processor are mirrored to the memory of a partner processor. The special features of the QPI architecture for systems with more than four processors stipulate these differences to the PRIMERGY RX600 S6.

The following parameters only exist in the PRIMERGY RX600 S6. The first two also exist accordingly for the PRIMERGY RX900 S2. There the BIOS can automatically ensure optimal values, because there are tighter restrictions regarding the admissibility of memory configurations. This has already been pointed out using the example of operation with only one active memory controller per processor, which only the PRIMERGY RX600 S6 knows.

### Interleaving

The parameter *Interleaving* defines the number of memory controllers, which are alternately addressed when setting up a segment of the physical address space that consists of 64-byte blocks: the first block is with the first controller, the second one with the second controller, etc. Consequently, access to adjoining memory areas, which always prevails according to the locality principle, is distributed across several controllers and their assigned components. There is a maximum of eight controllers in the system, as each processor has two controllers. The possible values for *Interleaving* are accordingly None, 2-way, 4-way, 8-way. With None the memory resources assigned to a controller are exhausted before a change is made to the resources of the next controller.

A connection exists here with the previously explained *NUMA Optimization*. If the latter is active, and if the address space is to be made up of segments with local memory, only the options None and 2-way remain for *Interleaving* and 2-way means the alternating between both controllers of the same processor. The 4-way and 8-way cases are not compatible with active NUMA and are for the most part not taken into account below.

The preferable case 2-way is possible if the same memory capacity is configured in both controllers of the processor. An identical configuration is not necessary for this purpose. Optimal memory performance is only achieved if this situation exists (*Interleaving = 2-way*). It is advisable for scientific applications to configure the system in such a way that there is 2-way interleaving. Acceptable configurations for commercial applications with the setting *Interleaving = None* follow below.

### Hemisphere Mode

The effect of the parameter *Hemisphere Mode* is more subtle in comparison to *Interleaving*. If the system is in hemisphere mode, the latency of individual memory access improves slightly. The mode is possible if the following is true for each processor: the memory configuration is identical for both controllers. This is an intensification of the previously required identical capacity in both controllers for 2-way interleaving. Hemisphere mode simplifies the processes for memory coherency: a check has to be made for every memory access as to whether the valid version of the block is in the DIMM or in the cache of another processor. And hemisphere mode reduces the number of agents involved by splitting the address space into an upper and lower hemisphere, which is equivalent to the first and second memory controller per processor.

### Memory Speed

As a rule memory speed follows from the processor type for the Westmere-EX based systems. A table with the assignment between processor and memory speed with 800, 978 or 1066 MHz follows later in the section The effects on memory performance. The *Memory Speed* option permits a reduction in the timing for the purpose of saving energy. *Memory Speed = Energy* is limited to 800 MHz, and the *Efficiency* option to 978 MHz.

## 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 only *registered* (RDIMM) modules.

| Type | | Control | Max. MHz | Ranks | Capacity | Rel. Price per GB |
|---|---|---|---|---|---|---|
| RDIMM | DDR3-1333 PC3-10600 LV | registered | 1333 | 1 | 4 GB | 1.1 |
| RDIMM | DDR3-1333 PC3-10600 LV | registered | 1333 | 2 | 8 GB | **1.0** |
| RDIMM | DDR3-1066 PC3-8500 LV | registered | 1066 | 4 | 16 GB | 1.4 |

The last column in the table shows the relative price differences. The list prices from August 2011 are used as a basis. The column shows the relative price per GB, standardized to the registered PC3-10600 DIMM, size 8 GB (highlighted as measurement 1). The higher costs for 16 GB modules are noticeable. The drop in price means that the question of costs must be taken into consideration when configuring memory.

The maximum frequencies stated in the table are features of the components, which in the case of 1333 MHz for the Westmere-EX based servers are theoretical. Maximum timing for these servers is 1066 MHz. The memory runs effectively with a timing of 800 or 978 or 1066 MHz as specified by the processor type, irrespective of the maximum values stated in the table. The timing is always 1066 MHz for the processors released for the PRIMERGY RX900 S2.

As an innovation in the Westmere-EX based servers compared with the Nehalem-EX predecessors, support is provided for energy-saving 1.35 V *low voltage* (LV) operation. Which is why the above table only contains LV modules. If there are 1.5 V standard modules, for example from an older DIMM stock, in the memory configuration, all the modules of the system run with 1.5 V, i.e. the energy saving that is possible with the LV modules does not occur.

The modules are offered in sets of four of the same type. Procurement in pairs is in any case necessary due to the already mentioned necessity for lockstep configuration. Other restrictions, for example in the PRIMERGY RX900 S2, as well as the minimum or typical processor configurations of the Westmere-EX systems, suggest bundling in units of four.

Some sales regions can have restrictions regarding the availability of certain DIMM types. In time, changes are also possible to the DIMM types and their features. The current configurator is always decisive.

# Performant memory configurations

The following tables list memory configurations according to their capacity and specify a three-stage evaluation of their performance feature for each configuration. The contrast arises because the number of DIMMs that follows from a required memory capacity enables a more or less well balanced utilization of the existing memory channels.

The section The effects on memory performance contains the measurement results on which this evaluation is based. In some configurations, for example those with different module sizes, it is possible to have segments of the address space with different performance features. The specified evaluation then corresponds to the worst case.

The table approach assumes that each processor is equally configured with memory; both with regard to capacity as well as DIMM selection and layout. This is the best case in NUMA architecture. The matrix of memory sizes that arises in this way is finely woven enough to enable customer requirements to be largely met. The tables not only list the GB per processor, but also specify the total system capacity for all permitted processor configurations of the Westmere-EX based PRIMERGY servers.

The specification of only the best possible case as regards performance per total capacity was explicitly omitted in the tables. It should be made clear that there are frequently configuration variants with a different performance for one capacity.

| Table for PRIEMRGY RX900 S2 as well as PRIMERGY RX600 S6 with 2 memory boards per CPU | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Total capacity in GB for various CPU configurations RX600 S6: 2, 3, 4 CPUs RX900 S2: 4, 6, 8 CPUs | | | | | GB per CPU | Number of DIMMs per CPU of size | | | Performance class |
| 2 CPU | 3 CPU | 4 CPU | 6 CPU | 8 CPU | | 4 GB | 8 GB | 16 GB | |
| 32 | 48 | 64 | 96 | 128 | 16 | 4 | | | 2 |
| 64 | 96 | 128 | 192 | 256 | 32 | 8 | | | 1 |
| 64 | 96 | 128 | 192 | 256 | 32 | | 4 | | 2 |
| 96 | 144 | 192 | 288 | 384 | 48 | 4 | 4 | | 2 |
| 96 | 144 | 192 | 288 | 384 | 48 | 12 | | | 2 |
| 128 | 192 | 256 | 384 | 512 | 64 | | 8 | | 1 |
| 128 | 192 | 256 | 384 | 512 | 64 | 16 | | | 1 |
| 128 | 192 | 256 | 384 | 512 | 64 | | | 4 | 2 |
| 128 | 192 | 256 | 384 | 512 | 64 | 8 | 4 | | 2 |
| 160 | 240 | 320 | 480 | 640 | 80 | 4 | | 4 | 2 |
| 160 | 240 | 320 | 480 | 640 | 80 | 4 | 8 | | 2 |
| 160 | 240 | 320 | 480 | 640 | 80 | 12 | 4 | | 2 |
| 192 | 288 | 384 | 576 | 768 | 96 | 8 | 8 | | 1 |
| 192 | 288 | 384 | 576 | 768 | 96 | | 4 | 4 | 2 |
| 192 | 288 | 384 | 576 | 768 | 96 | | 12 | | 2 |
| 192 | 288 | 384 | 576 | 768 | 96 | 8 | | 4 | 2 |
| 224 | 336 | 448 | 672 | 896 | 112 | 4 | 4 | 4 | 2 |
| 224 | 336 | 448 | 672 | 896 | 112 | 4 | 12 | | 2 |
| 224 | 336 | 448 | 672 | 896 | 112 | 12 | | 4 | 2 |
| 256 | 384 | 512 | 768 | 1024 | 128 | | | 8 | 1 |
| 256 | 384 | 512 | 768 | 1024 | 128 | | 16 | | 1 |
| 256 | 384 | 512 | 768 | 1024 | 128 | | 8 | 4 | 2 |
| 256 | 384 | 512 | 768 | 1024 | 128 | 8 | 4 | 4 | 2 |
| 288 | 432 | 576 | 864 | 1152 | 144 | 4 | | 8 | 2 |
| 288 | 432 | 576 | 864 | 1152 | 144 | 4 | 8 | 4 | 2 |
| 320 | 480 | 640 | 960 | 1280 | 160 | 8 | | 8 | 1 |
| 320 | 480 | 640 | 960 | 1280 | 160 | | 4 | 8 | 2 |
| 320 | 480 | 640 | 960 | 1280 | 160 | | 12 | 4 | 2 |
| 352 | 528 | 704 | 1056 | 1408 | 176 | 4 | 4 | 8 | 2 |
| 384 | 576 | 768 | 1152 | 1536 | 192 | | 8 | 8 | 1 |
| 384 | 576 | 768 | 1152 | 1536 | 192 | | | 12 | 2 |
| 416 | 624 | 832 | 1248 | 1664 | 208 | 4 | | 12 | 2 |
| 448 | 672 | 896 | 1344 | 1792 | 224 | | 4 | 12 | 2 |
| 512 | 768 | 1024 | 1536 | 2048 | 256 | | | 16 | 1 |

| Table for PRIMERGY RX600 S6 with 1 memory board per CPU | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Total capacity in GB for various CPU configurations | | | | | GB per CPU | Number of DIMMs per CPU of size | | | Performance class |
| 2 CPU | 3 CPU | 4 CPU | 6 CPU | 8 CPU | | 4 GB | 8 GB | 16 GB | |
| 32 | 48 | 64 | | | 16 | 4 | | | 3 |
| 64 | 96 | 128 | | | 32 | | 4 | | 3 |
| 64 | 96 | 128 | | | 32 | 8 | | | 3 |
| 96 | 144 | 192 | | | 48 | 4 | 4 | | 3 |
| 128 | 192 | 256 | | | 64 | | | 4 | 3 |
| 128 | 192 | 256 | | | 64 | | 8 | | 3 |
| 160 | 240 | 320 | | | 80 | 4 | | 4 | 3 |
| 192 | 288 | 384 | | | 96 | | 4 | 4 | 3 |
| 256 | 384 | 512 | | | 128 | | | 8 | 3 |

The requirement to configure each processor with memory usually means for the PRIMERGY RX600 S6 that additional memory boards must be ordered. Only two boards are available in the basic configuration. The first table applies for the optimal case of two boards per processor. Apart from the correct number of memory boards, the memory option *Interleaving Mode Installation* should also be ordered in this case, unless a redundancy option is required. This option ensures the BIOS default setting *Interleaving = 2-way* and a compatible memory configuration for this. The second table applies for the case of only one board per processor.

The tables take into account that the memory can be ordered in sets of four DIMMs of the same type. All DIMM quantities are multiples of four. The specified performance feature of a configuration arises during the pre-configuration *ex factory* or, if the user does the configuration him/herself subject to the following regulations:

- The configuration unit, which consists of four DIMMs, should be distributed over both boards for configurations of the PRIMERGY RX600 S6 with two memory boards per processor.
- The DIMM slots are to be used in the sequence specified on the memory boards of the PRIMERGY RX600 S6 and on the CPU memory riser boards of the PRIMERGY RX900 S2. The boards are labeled accordingly.
- DIMMs of varying size are installed in descending order of size; e.g. 16 GB before 8 GB.
- The BIOS settings that concern the memory are set to or left at default.

Configurations of performance class 1 provide optimal memory performance. Not only does symmetry prevail in these configurations regarding the two memory controllers per processor, but also with regard to the two DDR3 memory channels of each "Mill Brook 2" memory buffer. The latter does not exist for configurations of class 2. In comparison with class 1, class 2 means an average loss of between 5% and 9% in the PRIMERGY RX600 S6 for commercial applications, and between 4% and 6% in the PRIMERGY RX900 S2. The more performant the processor, the greater the loss. Class 3, which is only relevant for the PRIMERGY RX600 S6, means a further loss in the order of 4% in comparison with class 2.

Operation with less than four DIMMs per processor is not to be recommended. The tables do not contain such a case.

The required memory capacity was regarded as a given in these considerations. Its implicit influence on application performance, e.g. in the form of I/O rates, must be ignored here.

# 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.

## 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

Therefore, the Benchmark SPECint_rate_base2006 was added as a model for commercial application performance. It is part of SPECcpu2006 [L4] from Standard Performance Evaluation Corporation (SPEC). SPECcpu2006 is the industry standard for measuring system components processor, 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 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.

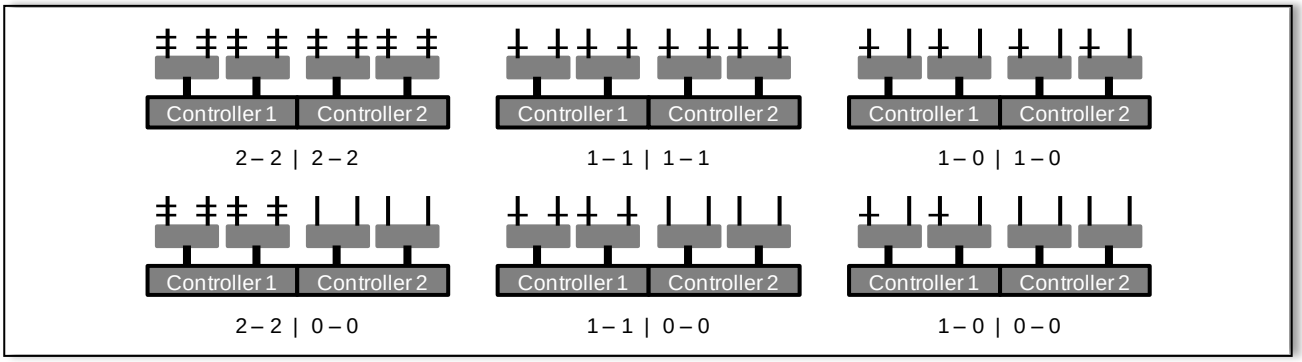
## Interleaving

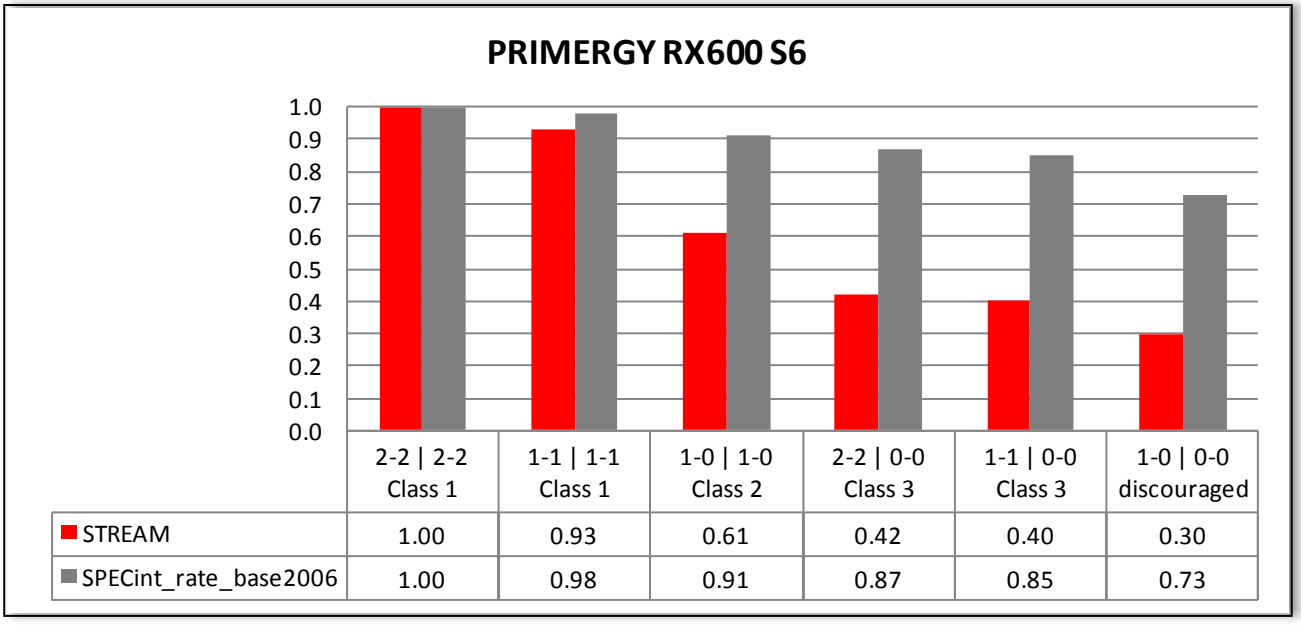
Interleaving is the main influence on performance in the Westmere-EX based PRIMERGY servers. Interleaving means the setting-up of the physical address space by alternating between memory resources. This is a performance gain situation resulting from parallelism. Interleaving is possible on two levels with the Westmere-EX based servers: through alternating between memory controllers and between the two DDR3 memory channel pairs of a controller. The above mentioned BIOS parameter only concerns the controller level. The alternating between memory channels within a controller follows automatically, if possible. The decisive factor is always that the memory capacities are identical with the resources involved. Alternating on a block by block basis must "work out even".

The diagrams show the effect as relative performance related to the most performant configuration (in each case the left-hand bar pair). This is full configuration with memory modules of the same type.

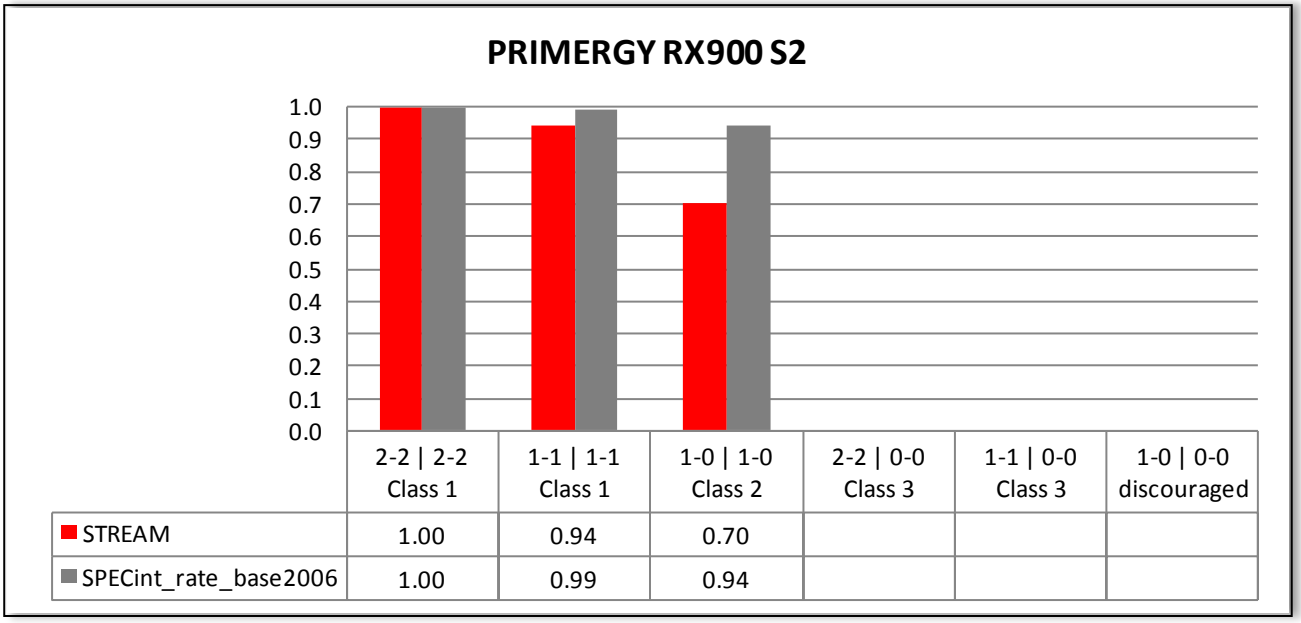
The first diagram explains die shorthand used in the following diagrams for the designation of the configurations. It specifies the DPC (DIMMs Per Channel) values for the memory channel pairs of the two memory controllers per processor. The second controller is not configured in the configurations with x-y | 0-0. This case is only possible in the PRIMERGY RX600 S6. Each DIMM pair in lockstep is only named once in shorthand. For example, the designation 2-2 means the maximum configuration of a controller with four DIMM pairs and eight DIMMs respectively. Accordingly, 1-0 is the minimum configuration with one DIMM pair located in slot pair 1B-1D (slot designation of the PRIMERGY RX600 S5). The 0 refers to the empty lockstep channel pair A-C.

The series of measurements shown in the diagrams were carried out with the high-performance processors Xeon E7-4870 (PRIMERGY RX600 S6) and E7-8870 (PRIMERGY RX900 S2) respectively and dual-rank memory modules of size 8 GB. The QPI and SMI timing was therefore 6.4 GT/s and the memory timing was 1066 MHz. The relative distinctions in performance differences are about the same for other types of processors of the Westmere-EX series and therefore also other QPI, SMI and memory timings. With weaker processors the absolute losses in performance can be specified as being a few percentage points lower than in the diagrams.




| | 2-2 \| 2-2 Class 1 | 1-1 \| 1-1 Class 1 | 1-0 \| 1-0 Class 2 | 2-2 \| 0-0 Class 3 | 1-1 \| 0-0 Class 3 | 1-0 \| 0-0 discouraged |
|---|---|---|---|---|---|---|
| STREAM | 1.00 | 0.93 | 0.61 | 0.42 | 0.40 | 0.30 |
| SPECint_rate_base2006 | 1.00 | 0.98 | 0.91 | 0.87 | 0.85 | 0.73 |




| | 2-2 \| 2-2 Class 1 | 1-1 \| 1-1 Class 1 | 1-0 \| 1-0 Class 2 | 2-2 \| 0-0 Class 3 | 1-1 \| 0-0 Class 3 | 1-0 \| 0-0 discouraged |
|---|---|---|---|---|---|---|
| STREAM | 1.00 | 0.94 | 0.70 | | | |
| SPECint_rate_base2006 | 1.00 | 0.99 | 0.94 | | | |

In a sense, the diagrams provide the full picture of the performance levels that arise through different interleaving. It is irrelevant that the series of measurements as presented were done with the 8 GB DIMM. The ratios are approximately identical for the other DIMM sizes. If the first (left-hand) configuration of the PRIMERGY RX600 S6 is considered as having 4 GB DIMMs and the third and fifth have 16 GB DIMMs, we are dealing in all three cases with implementations of the 256 GB memory configuration. Thus the diagram shows the performance differences for configuration alternatives of the same memory capacity.

The levels can be combined to form a three-stage performance assessment, which is mentioned in the diagrams, and to which reference has already been made above in the list of recommendable configurations. The configurations of class 1 allow 4-way interleaving for two memory controllers with two memory channel pairs each. The configurations of classes 2 and 3 allow 2-way interleaving; in class 2 in a cross-controller manner, and in class 3 at least still between the two memory channels of a single controller.

The finer distinctions within the three performance classes can be seen in the diagrams. It may occasionally make sense, for example in tests for customer-specific benchmarks, to take these distinctions into consideration. However, there are doubts that they will be noticed in productive operation.

Configurations of class 1 provide optimal performance for commercial and scientific applications. Configurations of classes 2 and 3 are acceptable for commercial applications.

The three configurations shown in the diagrams on the right with only one configured memory controller per processor are not permitted in the PRIMERGY RX900 S2 and are thus not assigned in the diagram. Although the last configuration on the far right is permitted for the PRIMERGY RX600 S6, it cannot be recommended from a performance viewpoint. The table listing recommendable configurations above in section Performant memory configurations does not contain such a case.

Regular configurations implemented with only one DIMM type result in homogeneous physical address spaces with standardized memory performance. With irregular configurations, for example a 2-2 | 1-1 with the same DIMM type, or a 1-1 | 1-1 with 8 GB modules on the first controller and 4 GB modules on the second one, the physical address space must be split into segments with different interleaving, and thus possibly with a different memory performance. The worst possible case is then of particular interest for the assessment of a configuration. In the examples just mentioned there will be an area in both cases that behaves like the 1-1 | 0-0 in the diagrams. The performance features of such irregular configurations must be assessed accordingly.

It is essential that inhomogeneous address spaces also always return to the performance levels of the diagrams. These are then no longer valid for the address space as a whole, but for individual segments of the address space. There may be random fluctuations for the application performance, depending on which segment provides the application with memory.

## Memory timing

It should first of all be repeated that the effect of interleaving does not depend on the memory timing. The relative performance for configurations with different interleaving, which was measured in the last section in an exemplary way with the standardized timing of 1066 MHz, applies in the same way for the cases 800 and 978 MHz. However, *absolute* memory performance, measured for example as maximum bandwidth in GB/s, depends of course on the memory timing.

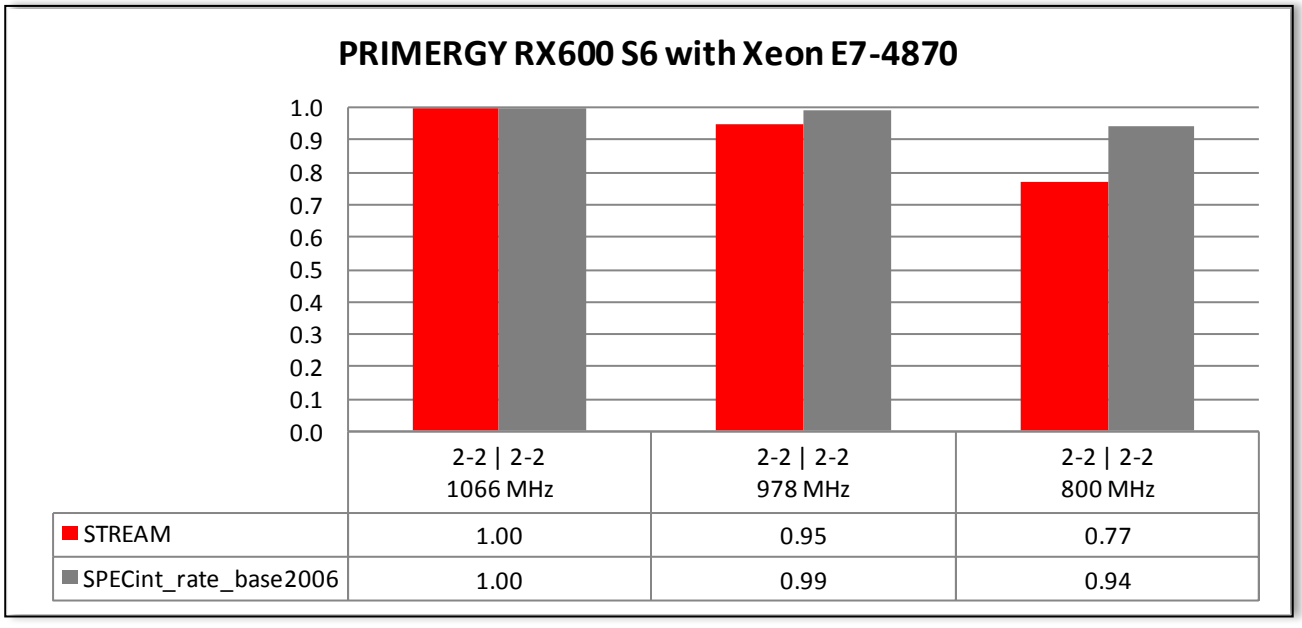
The memory timing for the PRIMERGY RX900 S2 is 1066 MHz at all times. The differences that concern timing need not be considered for this system.

As a rule, i.e. for BIOS parameters that were preset in the factory, the timing follows from the processor type for the PRIMERGY RX600 S6. The influence of the DIMM configuration on timing, which is known from the Xeon 5600 based 2-socket servers, does not exist with Westmere-EX servers. The following table shows the different timing for the Xeon E7-4800 processor family that is mostly configured in the RX600 S6.

| Category | Xeon type | #cores | GHz | L3 Cache (MB) | QPI / SMI (GT/s) | Memory (MHz) | TDP (Watt) |
|---|---|---|---|---|---|---|---|
| Advanced | E7-4870 | 10 | 2.40 | 30 | 6.40 | 1066 | 130 |
| | E7-4860 | 10 | 2.26 | 24 | 6.40 | 1066 | 130 |
| | E7-4850 | 10 | 2.00 | 24 | 6.40 | 1066 | 130 |
| Standard | E7-4830 | 8 | 2.13 | 24 | 6.40 | 1066 | 105 |
| | E7-4820 | 8 | 2.00 | 18 | 5.86 | 978 | 105 |
| Basic | E7-4807 | 6 | 1.86 | 18 | 4.80 | 800 | 95 |

Depending on processor type only, memory timing impacts application performance as a factor among others that cannot be isolated. The other factors are processor timing, number of processor cores, cache sizes and the differences in *Turbo Boost* functionality not mentioned in the table. Thus, the influence of memory timing usually cannot be seen in the benchmark results for the various types of processor.

However, the BIOS option *Memory Speed* that is available in the PRIMERGY RX600 S6 offers the option of reducing the standard timing listed in the table. Lower timing means less energy consumption. The setting *Memory Speed* = *Efficiency* results in timing with 978 MHz, and the setting *Energy* with 800 MHz. The following diagram shows the corresponding effects on performance. The effects are lower than those of the different interleaving.



## Number of ranks

Interleaving and memory timing are the primary influences on memory performance. The number of ranks per DDR3 memory channel is a secondary influence. This number results from the DPC value of the configuration and the number of ranks per module according to the table already shown above:

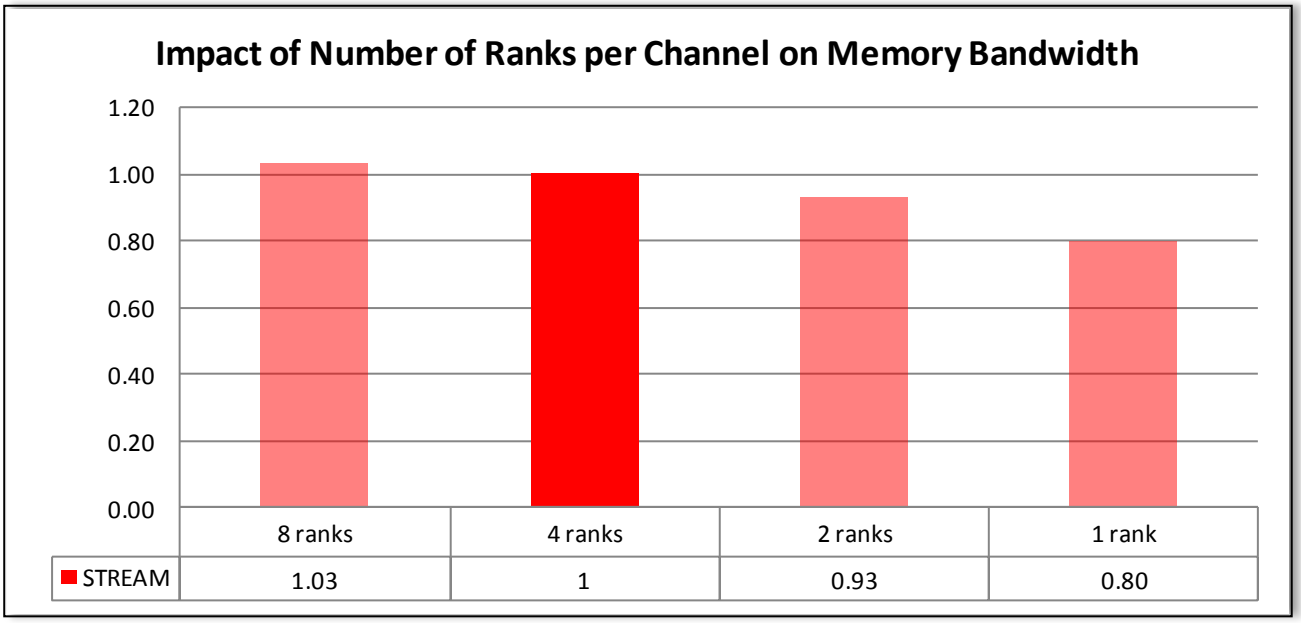
| Type | | Control | Max. MHz | Ranks | Capacity | Rel. Price per GB |
|---|---|---|---|---|---|---|
| RDIMM | DDR3-1333 PC3-10600 LV | registered | 1333 | 1 | 4 GB | 1.1 |
| RDIMM | DDR3-1333 PC3-10600 LV | registered | 1333 | 2 | 8 GB | **1.0** |
| RDIMM | DDR3-1066 PC3-8500 LV | registered | 1066 | 4 | 16 GB | 1.4 |

The *Rank* variable means there are DIMMs 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 DIMMs address and data lines are then common for all groups, i.e. only one of the groups can be active at any given time. The motivation for dual and quad-rank DIMMs is first the greater capacity, as can be seen in the aforesaid table.

A second advantage of such modules arises from the following physical reason. 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 requests regarding possible "open" memory pages. With dual and quad-rank modules, the probability of accessing an open page increases.

The following diagram shows the influence on performance of the number of ranks per DDR3 memory channel (not the number per module!) related to the case of a configuration with 4 ranks per channel. This case then arises if a PRIMERGY RX600 S6 or PRIMERGY RX900 S2 is fully configured with 8 GB dual-rank modules.

The diagram shows the effects on the maximum memory bandwidth. The influence of the number of ranks is usually negligible for application performance, particularly with commercial applications. It results in minimal differences in performance that are barely measurable. However, if it explicitly depends on the last ounce of performance, you are recommended to use quad-rank modules. By the way, there is a special release of quad-rank DIMMs of 8 GB for the Westmere-EX based servers for such cases.



## Memory performance under redundancy

The essential influences on memory performance and their impact on application performance have now been stated. The memory system has always been configured without redundancy for the previously collected data: the entire configured memory was available to the operating system as a physical address space. The following results are for memory performance under activated redundancy, i.e. activated DIMM module sparing or mirroring.

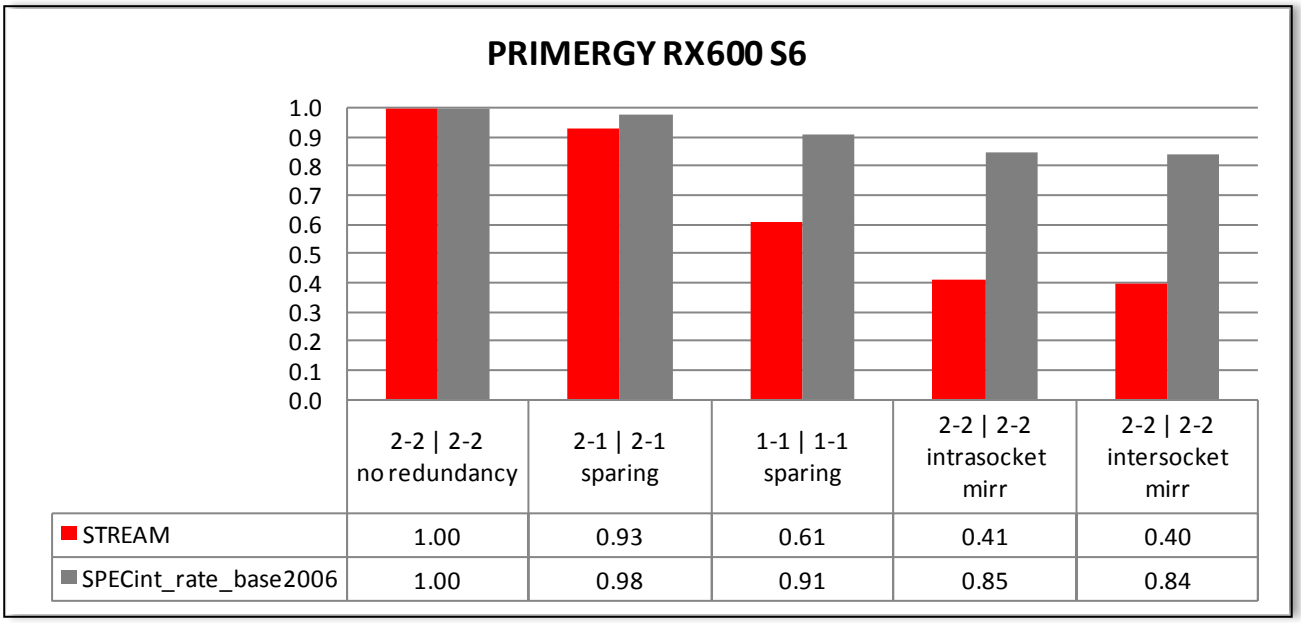
The results were again measured with the most high-performance processors Xeon E7-4870 (PRIMERGY RX600 S6) and E7-8870 (PRIMERGY RX900 S2), with the given memory timing of 1066 MHz und with dual-rank 8 GB DIMMs. The effects, however, are approximately identical in other configurations.

The diagrams are once more related to the optimal memory performance, as shown in each case on the left. This is full configuration without redundancy.

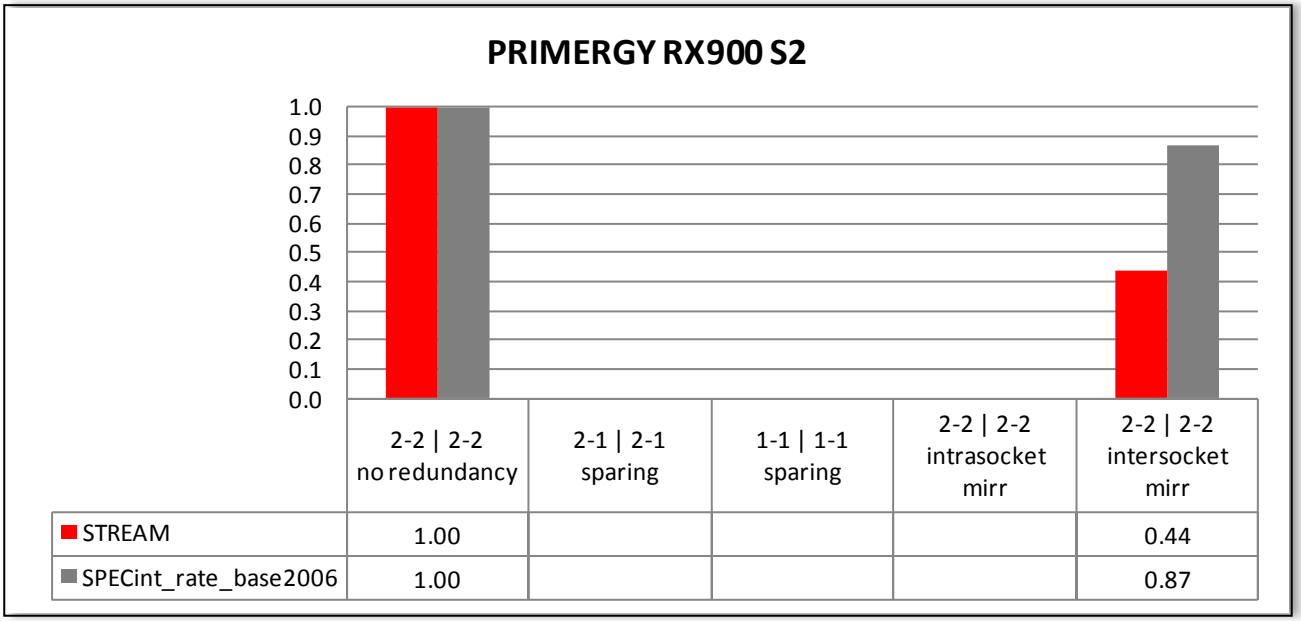
In the case of sparing, things are simple: the required measurement cases are identical with the configurations already shown in the diagrams in the section Interleaving 1-1 | 1-1 and 1-0 | 1-0. These configurations offer space for the sparing modules, whose existence does not change the performance. The following diagrams show the real configurations *including* the sparing modules as 2-1 | 2-1 and 1-1 | 1-1. However, in these cases only 67% and 50% of the configured capacity is available to the operating system as physical address space.

You can see from the diagram that for the PRIMERGY RX600 S6 sparing can be implemented in this system in a mostly performance-neutral way. Sparing is not supported in the PRIMERGY RX900 S2 .

Mirroring was also dealt with roughly above, as the configuration 2-2 | 0-0 and 1-1 | 0-0. However, apart from the loss of the second memory controller per processor (where the mirror is located, either of the processor's own first controller (intrasocket) or the first controller of another processor (intersocket)) a certain overhead arises through the constant updating of the mirror. Therefore, the values are somewhat lower than above for the 2-2 | 0-0 and 1-1 | 0-0. Nevertheless, these cases can also be allocated to performance class 3.




| | 2-2 \| 2-2 no redundancy | 2-1 \| 2-1 sparing | 1-1 \| 1-1 sparing | 2-2 \| 2-2 intrasocket mirr | 2-2 \| 2-2 intersocket mirr |
|---|---|---|---|---|---|
| STREAM | 1.00 | 0.93 | 0.61 | 0.41 | 0.40 |
| SPECint_rate_base2006 | 1.00 | 0.98 | 0.91 | 0.85 | 0.84 |




| | 2-2 \| 2-2 no redundancy | 2-1 \| 2-1 sparing | 1-1 \| 1-1 sparing | 2-2 \| 2-2 intrasocket mirr | 2-2 \| 2-2 intersocket mirr |
|---|---|---|---|---|---|
| STREAM | 1.00 | | | | 0.44 |
| SPECint_rate_base2006 | 1.00 | | | | 0.87 |

## Access to remote memory

Solely local memory was used in the previously described tests both with STREAM as well as with SPECint_rate_base2006, i.e. every processor accesses DIMM modules of its own memory channels. Modules of the other processors are not accessed or are hardly accessed via the QPI link. This situation is representative, insofar as it exists for the majority of memory accesses of real applications thanks to NUMA support in the operating system and system software.

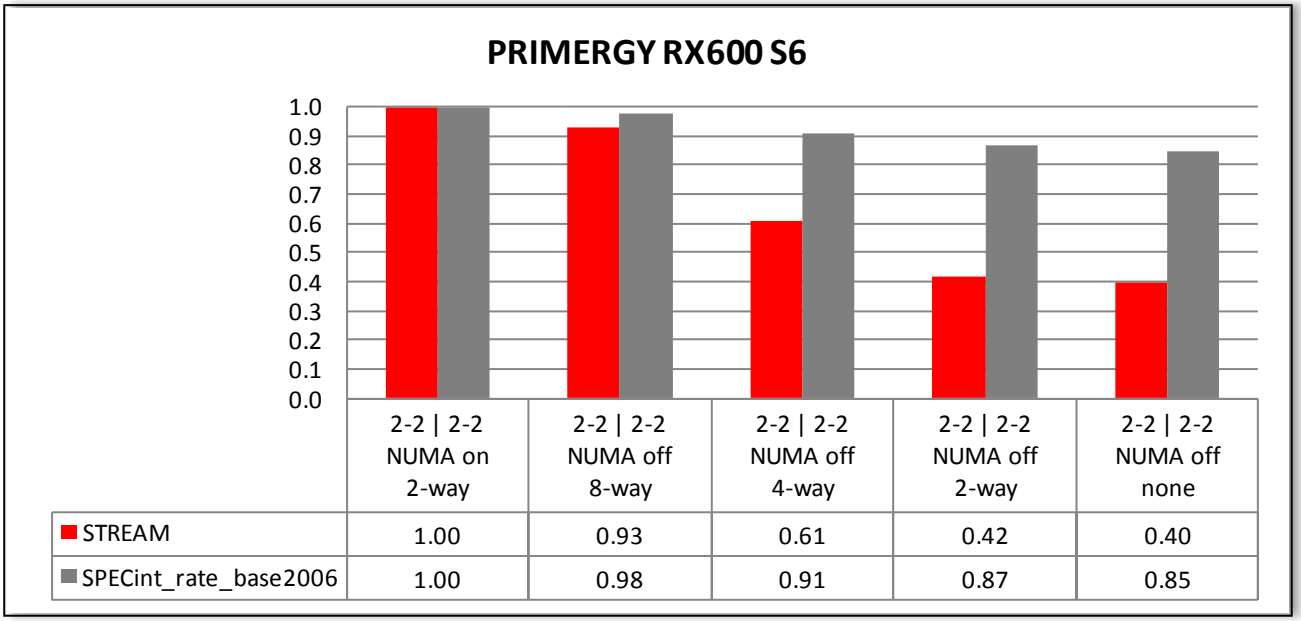
The following final diagram compares this optimal case of memory configuration with its alternative, the finely woven distribution of the physical address space over all memory channels of the system. To this end, it is first necessary to set the BIOS as follows: *NUMA Optimization* = *Disabled*.

Then in the PRIMERGY RX600 S6 the *Memory Interleave* option can be used to specify over how many memory controllers the address space is to be distributed when setting up a segment. The resources of these controllers, i.e. the DIMMs of the memory boards assigned to these controllers, are exhausted before commencing with the setup of the next segment. The maximum value of the parameter is 8-way, because eight memory controllers or memory boards are available in the system in the maximum configuration. One (8-way) to eight (none) segments arise when setting up the address space.

The measurement values shown in the diagram were determined on a PRIMERGY RX600 S6 in its maximum configuration with the processor type Xeon E7-4870 and 64 DIMMs of size 8 GB. The memory requirements of the STREAM benchmark are always only a few GB. The requirements for SPECint_rate_base2006 in the measuring variant used here is mostly covered by 64 GB. 64 GB was the capacity of a single memory board. Thus, only the first segment of the physical address space was used in all the test cases marked *NUMA off*. The decrease in performance can be explained by the concentration of the memory accesses of all processors to less and less memory modules and by the increasing load of the QPI connections leading to these modules.

A systematic series of measurements, as shown in the diagram, is possible if all the memory boards in the system are available and are identically configured. Otherwise, some of the test cases will not be possible. The granularity of the four *Memory Interleave* options can be explained by the dependence of the possible interleave on the actual configuration

There is no *Memory Interleave* BIOS option for the PRIMERGY RX900 S2. The BIOS automatically ensures the optimal setting in this system.




| | 2-2 \| 2-2 NUMA on 2-way | 2-2 \| 2-2 NUMA off 8-way | 2-2 \| 2-2 NUMA off 4-way | 2-2 \| 2-2 NUMA off 2-way | 2-2 \| 2-2 NUMA off none |
|---|---|---|---|---|---|
| STREAM | 1.00 | 0.93 | 0.61 | 0.42 | 0.40 |
| SPECint_rate_base2006 | 1.00 | 0.98 | 0.91 | 0.87 | 0.85 |

# 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] SPECcpu2006 Benchmark**

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

**[L5] Memory Performance of Xeon 5600 (Westmere-EP) based Systems**

http://docs.ts.fujitsu.com/dl.aspx?id=f622cc5b-c6f4-41c5-ae86-a642b4d5d255

**[L6] Memory Performance of Xeon 7500 (Nehalem-EX) based Systems**

http://docs.ts.fujitsu.com/dl.aspx?id=b3bfe45f-1ca8-43fc-8e32-c2a4534b4b3b


# 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