WHITE PAPER
FUJITSU PRIMERGY SERVERS
MEMORY PERFORMANCE OF XEON 7500 (NEHALEM-EX) BASED SYSTEMS

An extraordinary, architecture-related boost in performance comes from using Intel Xeon 7500 (Nehalem-EX) based systems in the quad socket class of PRIMERGY rack and blade servers. Here the new Intel QuickPath Interconnect (QPI) microarchitecture 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
2010-07-01

Contents

Document history

Version 1.0

Introduction

The Intel Xeon 7500 (Nehalem-EX) processors provide in the quad socket class of PRIMERGY rack and blade servers the same extraordinary, architecture-related boost in performance which has already impressively increased the attractiveness of dual socket PRIMERGY servers when Xeon 5500 (Nehalem-EP) processors were introduced. Compared with the previous generation, the general increase in performance for quad socket servers is about factor 2.7. This corresponds to the ratio of available hardware threads: the Xeon 7500 generation supports up to 16 threads per processor and the Xeon 7400 generation up to 6 threads. However, converting the computing potential into system performance only succeeds with the help of a paradigm change in the system architecture, particularly when connecting the processors to the main memory. The new processors use the Intel QuickPath Interconnect (QPI) microarchitecture.

The QPI replaces Front Side Bus (FSB) technology, which has been in use since the Intel Pentium Pro processor (1995) and had reached its limits regarding complexity, for example the number of pins required in the chipset per FSB. The QPI architecture is based on the idea of equipping the processors with memory controllers for a group of directly assigned ("local") memory modules. The outcome is a close, fast connection between processor cores and local memory. The processor can simultaneously provide memory contents to other processors via the so-called QPI links and request such itself. This approach with its distinction between local and remote memory changes the system architecture from Symmetric Multiprocessing (SMP) to Non-Uniform Memory Access (NUMA).

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.

More rules and recommendations concern the distribution of a given quantity of memory modules over the maximum 16 (PRIMERGY RX600 S5) or 8 (PRIMERGY BX960 S1) DIMM 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 5500 and 5600 based dual socket servers [L5], but the emphasis is clearly different. The significance of different timing diminishes. As with the dual 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 dual 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 first provides an overview of the memory architecture of the Xeon 7500-based quad socket 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 quad socket servers than with dual 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 possible memory configurations fall into three groups: those with optimal performance, ones with acceptable performance for commercial applications, and configurations that are not to be recommended. Numerous examples of the first two groups with a wide range of main memory capacities are listed below.

Memory architecture

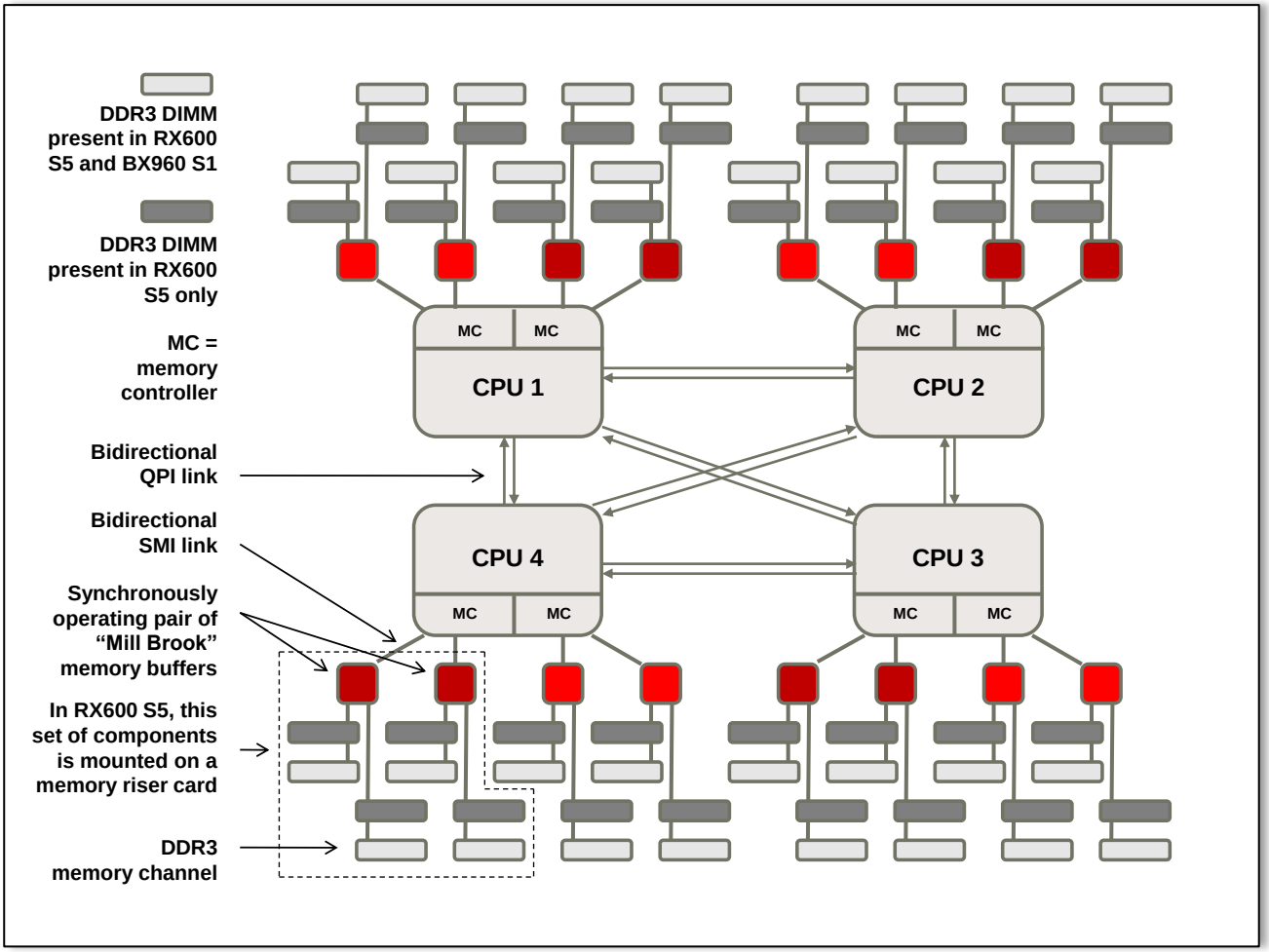
This section provides an overview of the memory system in three parts. Block diagrams explain the arrangement of the DIMM slots and their connections to the processors. The next section then deals with the BIOS parameters that affect the main memory. The third part covers the available DIMM types.

DIMM slots and connection

The following diagram shows the structure of the memory system. The presentation is based on the quad socket systems that are available or announced in July 2010. A distinction must be made between two cases when regarding the DIMM slots and their arrangement.

- The PRIMERGY RX600 S5 supports up to 16 DIMM slots per processor and thus a maximum of 64 slots in the entire system. The slots are to be found on memory boards (to be ordered separately) for 8 DIMM strips in each case. This modular structure is indicated in the diagram by the broken line. The diagram shows the maximum configuration. Operation with only one board per processor is possible for smaller memory configurations.
- The PRIMERGY BX960 S1 supports up to 8 DIMM slots per processor and thus a maximum of 32 slots in the entire blade. The slots are to be found on the motherboard. In comparison to the RX600 S5, a reduction to half the slots results from the fact that only one DIMM strip is supported per DDR3 channel; and two strips with the RX600 S5. The confined space in the blade housing is a decisive factor for the reduction.

The corresponding processor must be configured in order to use the DIMM slots. If operations are only with two or three processors, the slots connected to the empty socket cannot be used.



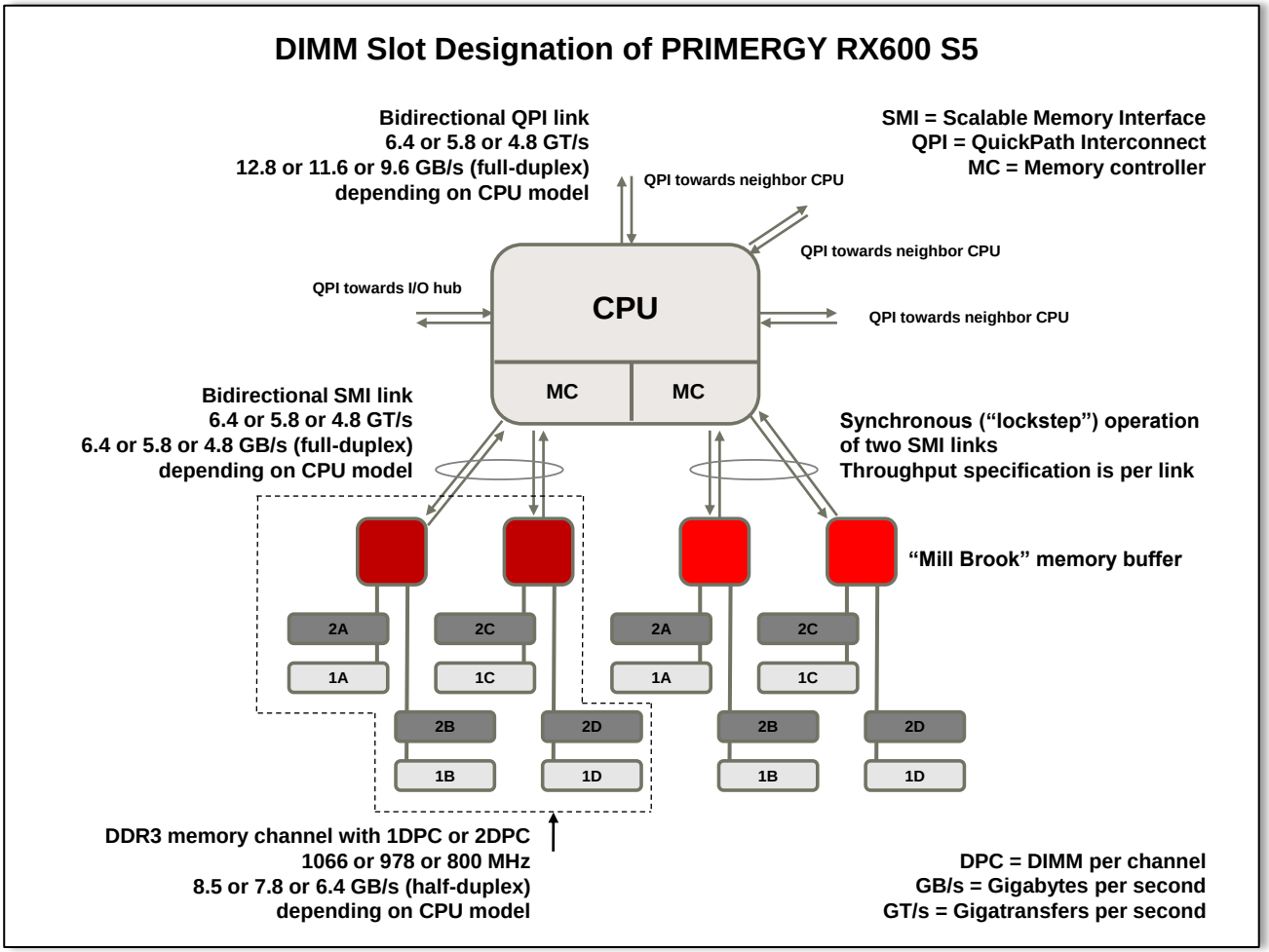
Two further diagrams show the connection of the memory to each individual processor in detail. The first diagram deals with the PRIMERGY RX600 S5, and the second one with the PRIMERGY BX960 S1.

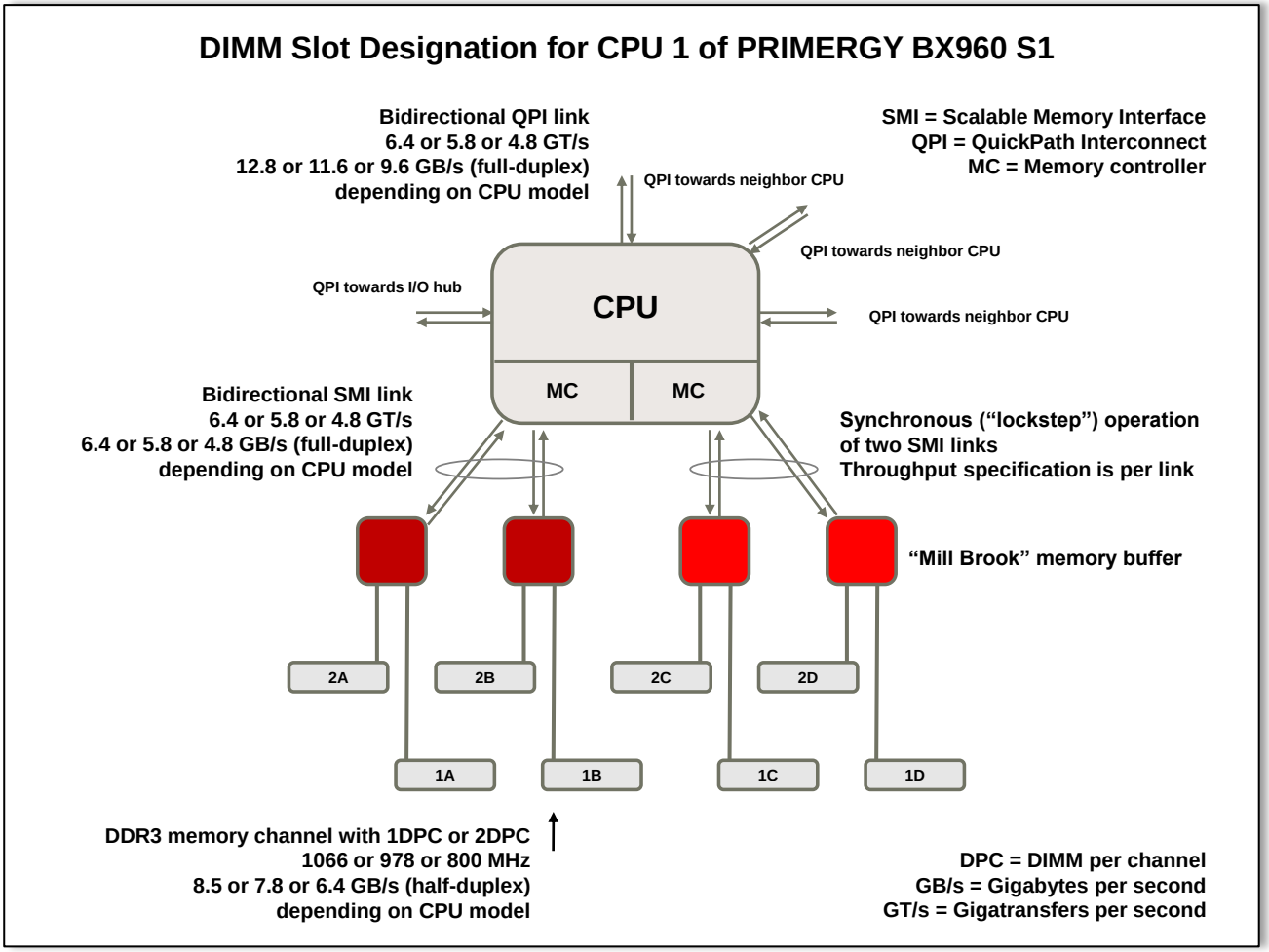
Each processor has two memory controllers integrated in the chip. Each controller is connected to two Mill Brook memory buffers via bidirectional, serial SMI (Scalable Memory Interface) links. These are separate chips, which are located on the memory boards for the PRIMERGY RX600 S5 and the motherboard for the BX960 S1.

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.

The bandwidth details of the diagrams for the components QPI, SMI and the DDR3 channel should mean that - depending on the processor type - there are three options. 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.

In the Xeon 5500 and 5600 processors of the PRIMERGY dual socket servers, which are also QPI-based, there is only one memory controller per processor, which directly controls three DDR3 channels. Buffering between the controller and the DIMM strips does not take place. The three DDR3 channels are not in lockstep and can be configured independently of each other.





BIOS parameters

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

NUMA Optimization

The parameter *NUMA Optimization* defines whether the physical address space is made up of segments from 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.

Interleaving

The parameter *Interleaving* defines the number of memory controllers, which are alternately addressed when setting up a segment of the physical address space consisting 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.

Redundancy

The options *Disabled, Sparing, Intrasocket Mirror* and *Intersocket Mirror* are available for the parameter *Redundancy*. An explanation of these RAS mechanisms is not within the scope of this White Paper. The list below in the section *Performant memory configurations* of suitable memory configurations contains cases that correspond to effective configuration under redundancy. For example, there are configurations with unused DIMM slots, which leave space for sparing modules, or there are configurations with a non-configured second memory controller per processor, which corresponds to effective configuration with activated mirroring. Memory performance under redundancy is the subject of a subsection in the section *The effects on memory performance*.

It is assumed below that the main memory runs with the nominal timing specified by the processor type. Memory performance under throttling or in energy-saving mode and the associated BIOS options are not taken into consideration.

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	registered	1333	1	2 GB	1.1
RDIMM	DDR3-1333 PC3-10600	registered	1333	2	4 GB	**1**
RDIMM	DDR3-1333 PC3-10600	registered	1333	2	8 GB	1.9
RDIMM	DDR3-1066 PC3-8500	registered	1066	4	16 GB	2.4

The last column in the table shows the relative price differences. The list prices for the PRIMERGY RX600 S5 as of June 2010 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 higher costs for 8 and 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 Xeon 7500-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 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. And procurement in sets of four avoids the ordering of DIMM quantities or memory capacities that cannot be configured in a performant way.

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 provide configuration examples for a comprehensive range of memory sizes, which are suitable when considering performance. The second last column contains a two-class evaluation of the performance of the configuration. Configurations of class 1 provide optimal performance for commercial and scientific applications. Configurations of class 2 are acceptable for commercial applications. However, system performance can be up to 10% below the performance that can be achieved with configurations of class 1.

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

The specified capacity assumes a system with four equally configured processors. The capacities must be converted accordingly for systems that are operated with two or three processors.

The DIMM slots are designated as on the memory boards or motherboards of the systems. Each lockstep pair, however, is specified once only. If slot 1B for example is configured in the PRIMERGY RX600 S5, slot 1D must be configured identically. The same applies for the slot pairs 1A-1C, 2B-2D, 2A-2C.

In the PRIMERGY BX960 S1 the slots on the motherboard are designated consecutively, i.e. on the second processor the eight available slots are not called 1A, 2A, 1B, 2B, 1C, 2C, 1D, 2D, but 1E, 2E, 1F, 2F, 1G, 2G, 1H, 2H, etc. As regards slot designations, it is worth mentioning the diagrams shown in the section above *DIMM slots and connection*.

Configurations, in which not all slots are configured with DIMM strips, are marked in gray. These include the PRIMERGY RX600 S5 configurations that can manage without the second memory board per processor. These configurations, marked in gray, can be interpreted as effective configurations under redundancy. Effective means that the operating system sees the memory capacity named in the tables, while actually memory modules for up to twice the capacity are installed. If for example all the slots of the second memory controller are not configured, there would be space to set up the mirror for activated Intrasocket or Intersocket Mirroring. Likewise, there are configurations with free slots in both memory controllers. These provide space for sparing modules.

PRIMERGY RX600 S5											
Capacity 4 CPUs	Controller 1				Controller 2				2-way Interleave Hemisphere Mode	Perfor-mance class	Remark
	1B 1D	2B 2D	1A 1C	2A 2C	1B 1D	2B 2D	1A 1C	2A 2C			
32 GB	2				2				yes	2	Sparing Option
	2		2						no	2	Mirror Option
64 GB	2		2		2		2		yes	1	Best Perf / Sparing
	4		4						no	2	Mirror Option
96 GB	4		2		4		2		yes	2	Sparing Option
	4	2	4	2					no	2	Mirror Option
128 GB	2	2	2	2	2	2	2	2	yes	1	Best performance
	4		4		4		4		yes	1	Best Perf / Sparing
	4	4	4	4					no	2	Mirror Option
160 GB	8		2		8		2		yes	2	Sparing Option
	8	2	8	2					no	2	Mirror Option
192 GB	4	2	4	2	4	2	4	2	yes	1	Best performance
	8		4		8		4		yes	2	Sparing Option
	8	4	8	4					no	2	Mirror Option
256 GB	4	4	4	4	4	4	4	4	yes	1	Best performance
	8		8		8		8		yes	1	Best Perf / Sparing
	8	8	8	8					no	2	Mirror Option
288 GB	16		2		16		2		yes	2	Sparing Option
	16	2	16	2					no	2	Mirror Option
320 GB	8	2	8	2	8	2	8	2	yes	1	Best performance
	16		4		16		4		yes	2	Sparing Option
	16	4	16	4					no	2	Mirror Option
384 GB	8	4	8	4	8	4	8	4	yes	1	Best performance
	16		8		16		8		yes	2	Sparing Option
	16	8	16	8					no	2	Mirror Option
512 GB	8	8	8	8	8	8	8	8	yes	1	Best performance
	16		16		16		16		yes	1	Best Perf / Sparing
	16	16	16	16					no	2	Mirror Option
576 GB	16	2	16	2	16	2	16	2	yes	1	Best performance
640 GB	16	4	16	4	16	4	16	4	yes	1	Best performance
768 GB	16	8	16	8	16	8	16	8	yes	1	Best performance
1024 GB	16	16	16	16	16	16	16	16	yes	1	Best performance

The figures specified for the slot pairs indicate the module sizes in GB.

PRIMERGY BX960 S1							
Capacity 4 CPUs	Controller 1		Controller 2		2-way Interleave Hemisphere Mode	Performance Class	Remark
	1A 1B	2A 2B	1C 1D	2C 2D			
32 GB	2		2		yes	2	Sparing Option
	2	2			no	2	Mirror Option
64 GB	2	2	2	2	yes	1	Best performance
	4		4		yes	2	Sparing Option
	4	4			no	2	Mirror Option
96 GB	4	2	4	2	yes	2	
128 GB	4	4	4	4	yes	1	Best performance
	8		8		yes	2	Sparing Option
	8	8			no	2	Mirror Option
160 GB	8	2	8	2	yes	2	
192 GB	8	4	8	4	yes	2	
256 GB	8	8	8	8	yes	1	Best performance
	16		16		yes	2	Sparing Option
	16	16			no	2	Mirror Option
288 GB	16	2	16	2	yes	2	
320 GB	16	4	16	4	yes	2	
384 GB	16	8	16	8	yes	2	
512 GB	16	16	16	16	yes	1	Best performance

The figures specified for the slot pairs indicate the module sizes in GB.

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 [L4] 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.

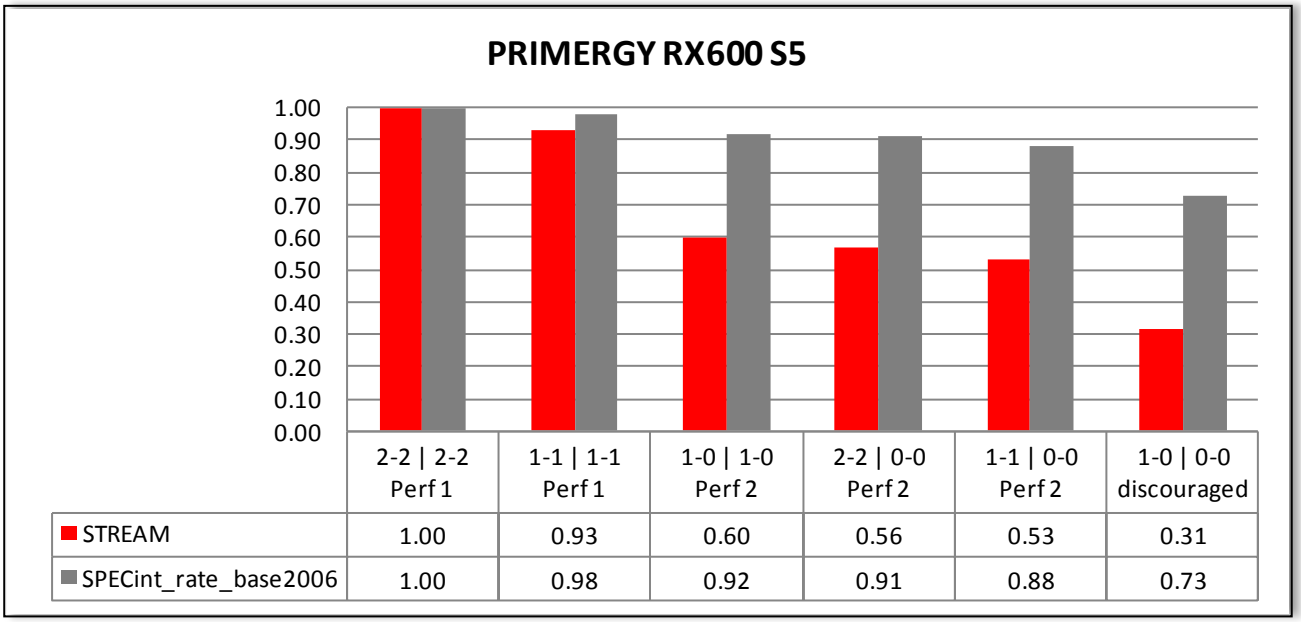
Interleaving

Interleaving is the main influence on memory performance in the Xeon 7500 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 Xeon 7500 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".

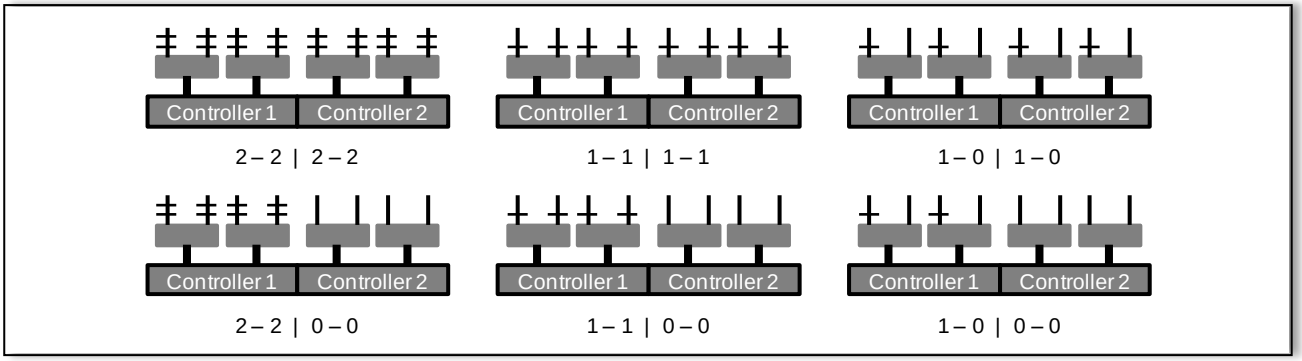
The diagrams show the effect as relative performance related to the most performant configuration (in each case the left-hand bar pair). This is full configuration with memory modules of the same type.

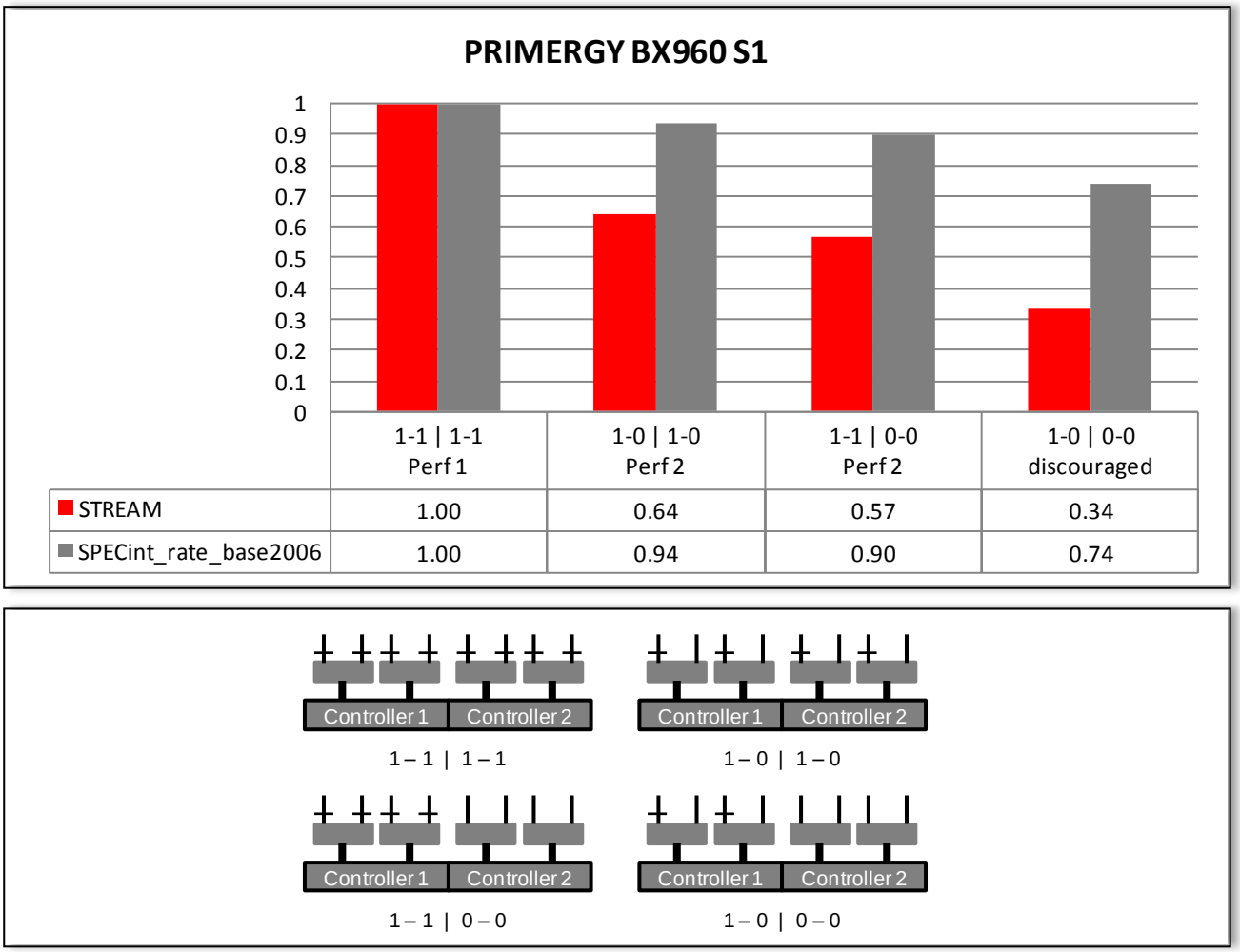
The "shorthand" used for the names of the configurations specifies the DPC (DIMM per channel) values for the memory channel pairs of the two memory controllers per processor. The second controller is not configured in the configurations with x-y | 0-0. As was previously the case in the tables listing recommendable configurations, every DIMM pair in lockstep is only mentioned once. 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 presented in the diagrams were performed with the processor Xeon E7530 and dual-rank memory modules of size 8 GB. The QPI and SMI clock rate was therefore 5.86 GT/s and the memory timing was 978 MHz. The relative performance differences are approximately the same for other types of processor from the Xeon 7500 series, and therefore other QPI, SMI and memory timings.



	2-2 | 2-2 Perf 1	1-1 | 1-1 Perf 1	1-0 | 1-0 Perf 2	2-2 | 0-0 Perf 2	1-1 | 0-0 Perf 2	1-0 | 0-0 discouraged
STREAM	1.00	0.93	0.60	0.56	0.53	0.31
SPECint_rate_base2006	1.00	0.98	0.92	0.91	0.88	0.73





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 identical for the other DIMM sizes. If the first (left-hand) configuration of the PRIMERGY RX600 S5 is seen to be configured with 2 GB DIMM, the second and fourth with 4 GB DIMM, the third and fifth with 8 GB DIMM, then in all five cases this is the implementation of a 128 GB memory capacity. Thus the diagram shows the performance differences for configuration alternatives of the same memory capacity.

The levels can be combined to form a two-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 class 2 allow 2-way interleaving. Only the cases designated with 1-0 | 1-0 are on a cross-controller basis here and are thus cases of 2-way interleaving within the meaning of the BIOS parameter.

The finer distinctions within the two 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 class 2 are acceptable for commercial applications. The configurations on the very right in the diagrams also no longer appear to be recommendable for commercial applications.

Configurations that are regular and implemented with only one DIMM type, as in the tests on which the diagrams are based, result in homogeneous physical address spaces with uniform 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

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 clock rate of 978 MHz, applies in the same way for the cases 800 and 1066 MHz. However, *absolute* memory performance, measured for example as maximum bandwidth in GB/s, depends of course on the memory timing.

The effective timing follows from the processor type as per the table below. All the DIMM modules available for the Xeon 7500 based PRIMERGY servers support the maximum 1066 MHz both in 1DPC and 2DPC configurations. The impact of DIMM type and DPC value on the timing known for the Xeon 5500 und 5600 based dual socket servers does not exist in the Xeon 7500 based systems.

Class	Xeon type	#cores	GHz	L3 Cache (MB)	QPI / SMI (GT/s)	Memory (MHz)	TDP (Watt)
Advanced	X7560	8	2.26	24	6.4	1066	130
	X7550	8	2.00	18	6.4	1066	130
	L7555	8	1.86	24	5.86	978	95
Standard	X7542	6	2.66	18	5.86	978	130
	E7540	6	2.00	18	6.4	1066	105
	L7545	6	1.86	18	5.86	978	95
	E7530	6	1.86	12	5.86	978	105
Basic	E7520	4	1.86	18	4.8	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 cannot be seen in the SPECint_rate_base2006 results for the various types of processor.

However, the following table shows the maximum memory bandwidths for the three relevant cases. Here we are dealing with STREAM measurements in the PRIMERGY RX600 S5 with four processors and a full memory configuration with dual-rank modules of size 8 GB.

QPI / SMI (GT/s)	DDR3 memory channels (MHz)	Xeon type	STREAM (GB/s)
6.4	1066	X7560	72.4
5.86	978	E7530	60.4
4.8	800	E7520	54.8

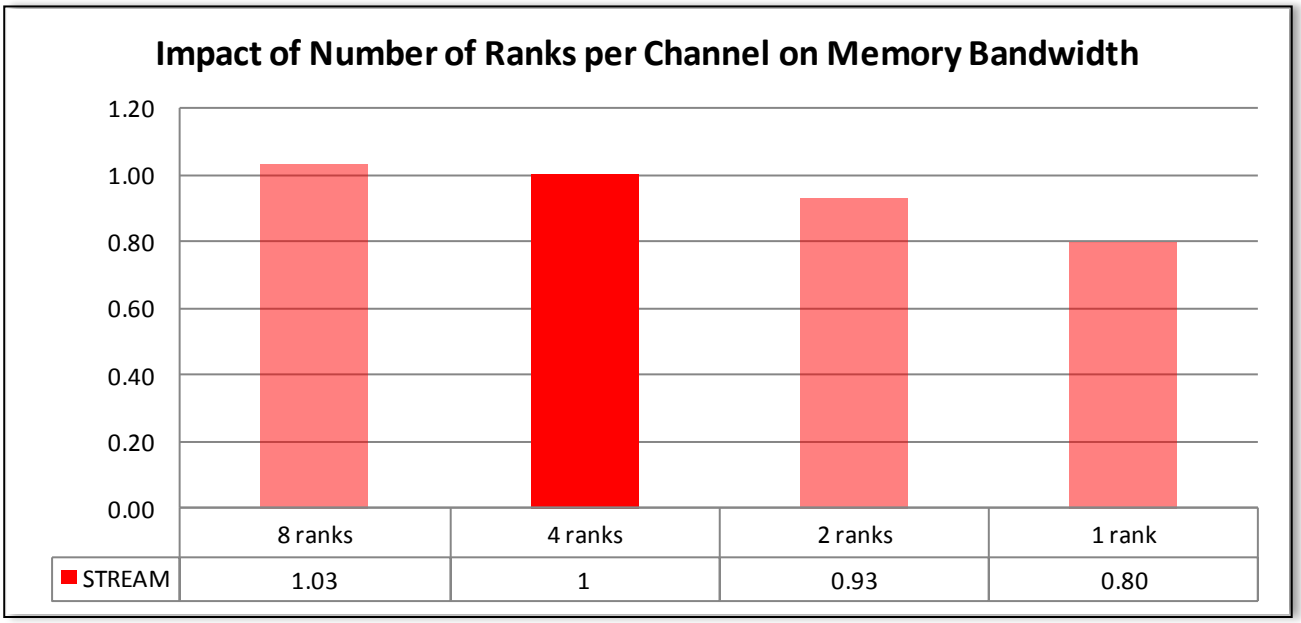
The number of ranks

As seen in the last table, STREAM results do not depend on the configured memory capacity, i.e. the use of 8 GB modules was irrelevant for these results. The same bandwidths would result with 4 GB modules. The benchmark always only uses 1 GB – in any case only a fraction of the configured total capacity.

However, the number of ranks per DDR3 memory channel has a secondary influence on the measured bandwidth. This number results from the DPC value of the configuration and the number of the ranks per module according to the table in the section *Available memory types*. In the bandwidths just mentioned the number of ranks per channel was therefore 4: a 2DPC configuration with dual-rank modules.

The following diagram shows the performance influence of the number of ranks per DDR3 memory channel (not the number of ranks per module!) related to the previously described case of a configuration with 4 ranks per channel.

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. There it results in minimal differences in performance that are barely measurable.



Memory performance under redundancy

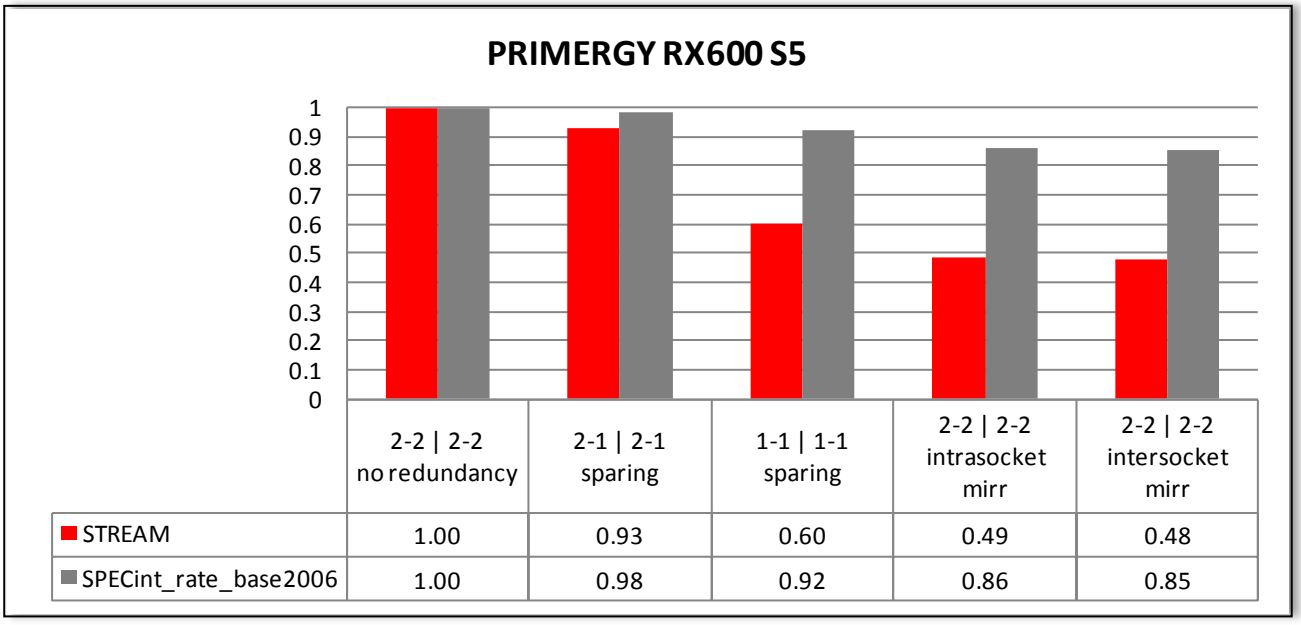
The essential influences on memory performance and their impact on application performance have now been stated. The memory system has always been configured without redundancy for the previously collected data: the entire configured memory was available to the operating system as a physical address 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 Xeon 7530 processor, and with the given memory timing of 978 MHz and dual-rank 8 GB DIMM. The effects, however, are approximately identical in other configurations.

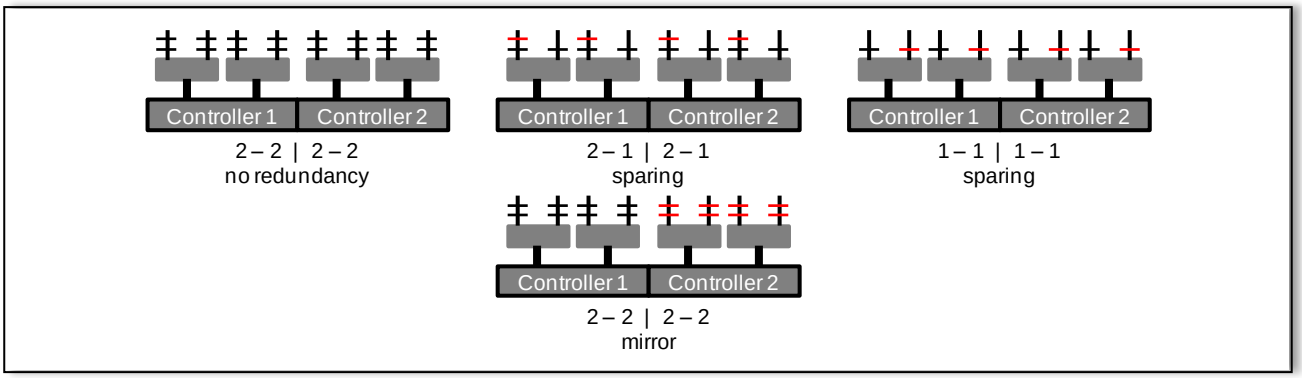
The diagrams are once more related to the optimal memory performance, as shown in each case on the left. This is full configuration without redundancy.

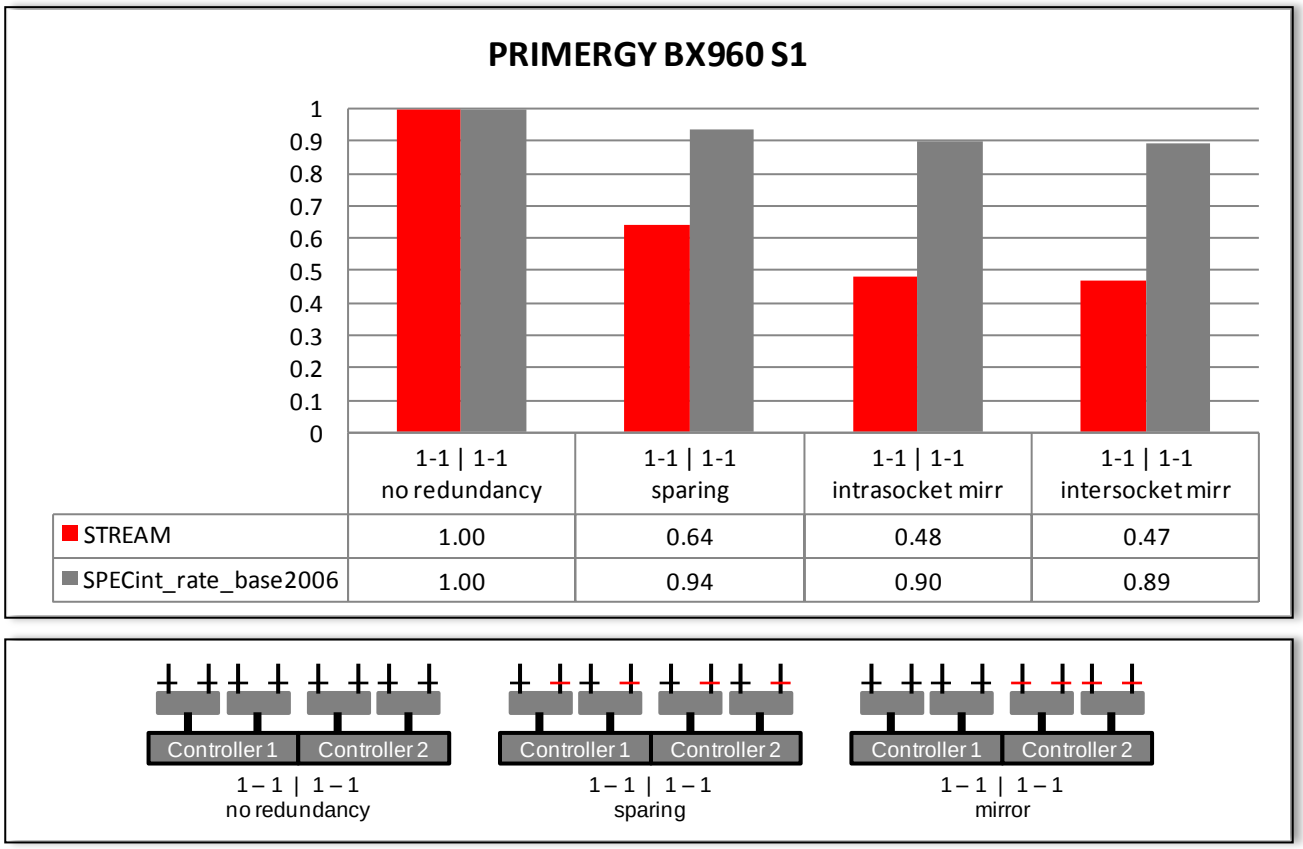
In the case of sparing, things are simple: the required measurement cases are identical with the configurations already shown in the diagrams in the section *Interleaving* 1-1 | 1-1 and 1-0 | 1-0. These configurations offer space for the sparing modules, whose existence does not change the performance. In the 1-1 | 1-1 case this is of course only valid for the PRIMERGY RX600 S5. The following diagrams show the real configurations including sparing modules (marked in red in the configuration scheme) as 2-1 | 2-1 and 1-1 | 1-1. However, only 67% and 50% respectively of the configured capacity are available to the operating system in these cases.

You can see from the diagram that for the PRIMERGY RX600 S5 sparing can be implemented in this system in a mostly performance-neutral way. In the PRIMERGY BX960 S1 sparing on the other hand means an effective memory configuration of performance class 2. A class 1 configuration is not possible in this system under sparing.



	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	0.60	0.49	0.48
SPECint_rate_base2006	1.00	0.98	0.92	0.86	0.85





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 2.

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 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 both benchmarks. 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	-25%
SPECint_rate_base2006	-13%

With STREAM the bandwidth of the QPI link between the processors becomes the result-determining bottleneck. The deterioration of the SPECint_rate_base2006 is primarily caused by the 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 a configuration with four processors the BIOS setting *Interleaving = 8-way* is in this case advisable, which spreads the physical address space in a finely woven way for all four processors. Then

25% of the accesses of an application are to local memory and 75% are to a remote one. In the benchmarks mentioned the STREAM result then deteriorates by 20%, and the SPECint_rate_base2006 result by 9%. The latter should be approximately representative for the effect of disabling NUMA support for commercial applications.

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

Contact

FUJITSU Technology Solution

Website: http://ts.fujitsu.com

PRIMERGY Product Marketing

mailto:Primergy-PM@ts.fujitsu.com

PRIMERGY Performance and Benchmarks

mailto:primergy.benchmark@ts.fujitsu.com