

White Paper

FUJITSU Server PRIMERGY & PRIMEQUEST

Memory performance of Xeon E7-8800 / 4800 v2 (Ivy Bridge-EX) based systems

The Xeon E7-8800 / 4800 v2 (Ivy Bridge-EX) based models of the PRIMEQUEST 2000 series and of PRIMERGY RX4770 M1 also acquire their impressive two-fold increase in performance compared with the predecessor generation from an enhancement of the QuickPath Interconnect (QPI) memory architecture, which has proved itself now for two generations of systems. This white paper explains the changed parameters of the architecture and quantifies their effect on the performance of commercial applications.

Version

1.1

2014-05-16



Contents

Document history	2
Introduction	3
Memory architecture	5
DIMM slots	5
DIMM types	9
Firmware and BIOS parameters	11
Interfaces of the Web-GUI of PRIMEQUEST 2000 series	11
Interfaces of the device manager of PRIMEQUEST 2000 series	12
Interfaces of the BIOS of PRIMERGY RX4770 M1	12
Definition of the memory frequency	13
Lockstep channel operation mode	14
Independent channel operation mode	14
Ideal memory capacities	15
Quantitative effects on memory performance	17
The measuring tools	18
STREAM Benchmark	18
SPECint_rate_base2006 Benchmark	18
Interleaving	19
Interleaving across memory controllers and memory channels	19
Interleaving across ranks	22
Memory frequency	23
Memory performance under redundancy	24
Full Mirror Mode of PRIMEQUEST 2000 series	24
Mirror Mode of PRIMERGY RX4770 M1	25
Spare Mode	26
Literature	27
Contact	27

Document history

Version 1.0 (2014-03-07)

Initial version

Version 1.1 (2014-05-16)

Inclusion of PRIMERGY RX4770 M1

Introduction

The models of the PRIMEQUEST 2000 series and of PRIMERGY RX4770 M1, which are equipped with Intel Xeon E7-8800 / 4800 v2 (Ivy Bridge-EX) processors, continue the product segment of high-end servers with an impressive two-fold increase in performance for most load scenarios. This increase has in comparison with the Westmere-EX based predecessor generation (PRIMEQUEST 1800E2, PRIMERGY RX900 S2, PRIMERGY RX600 S6) two causes:

- The move from 32 to 22 nm manufacturing technology for processor chips enables up to 15 instead of the previous 10 cores per processor. At the same time the renewal of the microarchitecture, which has already taken place for dual socket servers in the Sandy Bridge-EP generation, follows suit. The potential of these measures is in the order of 50%.
- The other half of the aforementioned increase in performance results from improvements in the memory system.

The essential features of the QPI (QuickPath Interconnect) system architecture, which have proven themselves since 2009, have been retained. The processors have integrated memory controllers for high-performance control of local memory modules. At the same time, they are able to provide the neighboring processors with memory content via QPI links and themselves request such content. The architecture with this distinction between local and remote memory access is of the NUMA (Non-Uniform Memory Access) type.

In the high-end server class with its design goals of maximum memory capacity and optimal RAS (Reliability, Availability, Serviceability) there are two memory controllers per processor with four DDR3 (Double Data Rate) memory channels each as well as memory buffers located between the controllers and the channels. This configuration enables more DIMM (Dual Inline Memory Module) slots per processor than with the dual socket servers, which do without memory buffers. The result of the new version of the buffer (Jordan Creek 1) is 24 DIMM slots and a maximum configuration of 1.5 TB per processor. And half of that numbers applies for the current generation of Ivy Bridge-EP based dual socket PRIMERGY servers [\[L3\]](#), which have no memory buffers.

While the profile of the QPI-based NUMA architecture with memory buffers remains unchanged, the increase in memory performance of the Xeon E7-8800 / 4800 v2 based servers results from the following functions:

- Jordan Creek 1 supports DDR3 memory frequencies up to 1600 MHz, in comparison with a maximum of 1066 MHz in the predecessor generation (Mill Brook 2).
- The DDR3 channels of the predecessor systems are always in Lockstep Mode, a synchronous operation mode of two channels each, which improves the RAS features. This mode continues to exist. However, it can be released in favor of the higher bandwidth of independent memory channels, i.e. a new trade-off between RAS and performance is introduced.
- The maximum QPI frequency increases from 6.4 to 8.0 GT/s (gigatransfers per second).
- The cache-coherency protocol is changed from snooping-based (QPI 1.0) to directory-based (QPI 1.1).

The most elementary indicator of memory performance, the memory bandwidth, has increased as a result of these measures for the PRIMEQUEST 2800E from 110 to 393 GB/s, and for PRIMERGY RX4770 M1 from 102 to 244 GB/s.

This white paper provides the basic knowledge of memory architecture required for the configuration of powerful systems. We are dealing with the following points here:

- Due to the NUMA architecture all processors should as far as possible be equally configured with memory. The aim of this measure is for each processor to work as a rule on its local memory.
- In order to parallelize memory access the aim is to distribute closely adjacent areas of the physical address space across several components of the memory system. The corresponding technical term is Interleaving. Interleaving exists in two dimensions. First of all, in terms of width across the memory controllers and DDR3 channels per processor, and it is this aspect of memory performance that is affected by the lockstep operation mode. There is also interleaving in the depth of the individual memory channel. The resources for this are the ranks. These are substructures of the DIMMs, in which groups of DRAM (Dynamic Random Access Memory) chips are consolidated. Individual memory access always refers to such a group.
- Memory frequency influences performance. It is 1600, 1333, or 1066 MHz depending on DIMM type, number and power saving mode. Since DIMM type and number are related to the required memory capacity, the aspects of performance, capacity and energy consumption should be weighed up against each other.

Influencing factors are named and quantified in order to help you weigh up. Quantification is done with the help of the benchmarks STREAM and SPECint_rate_base2006. STREAM measures the memory bandwidth. SPECint_rate_base2006 is used as a model for the performance of commercial applications.

Statements about memory performance under redundancy, i.e. with enabled Mirroring or Sparing, make up the end of this document.

Memory architecture

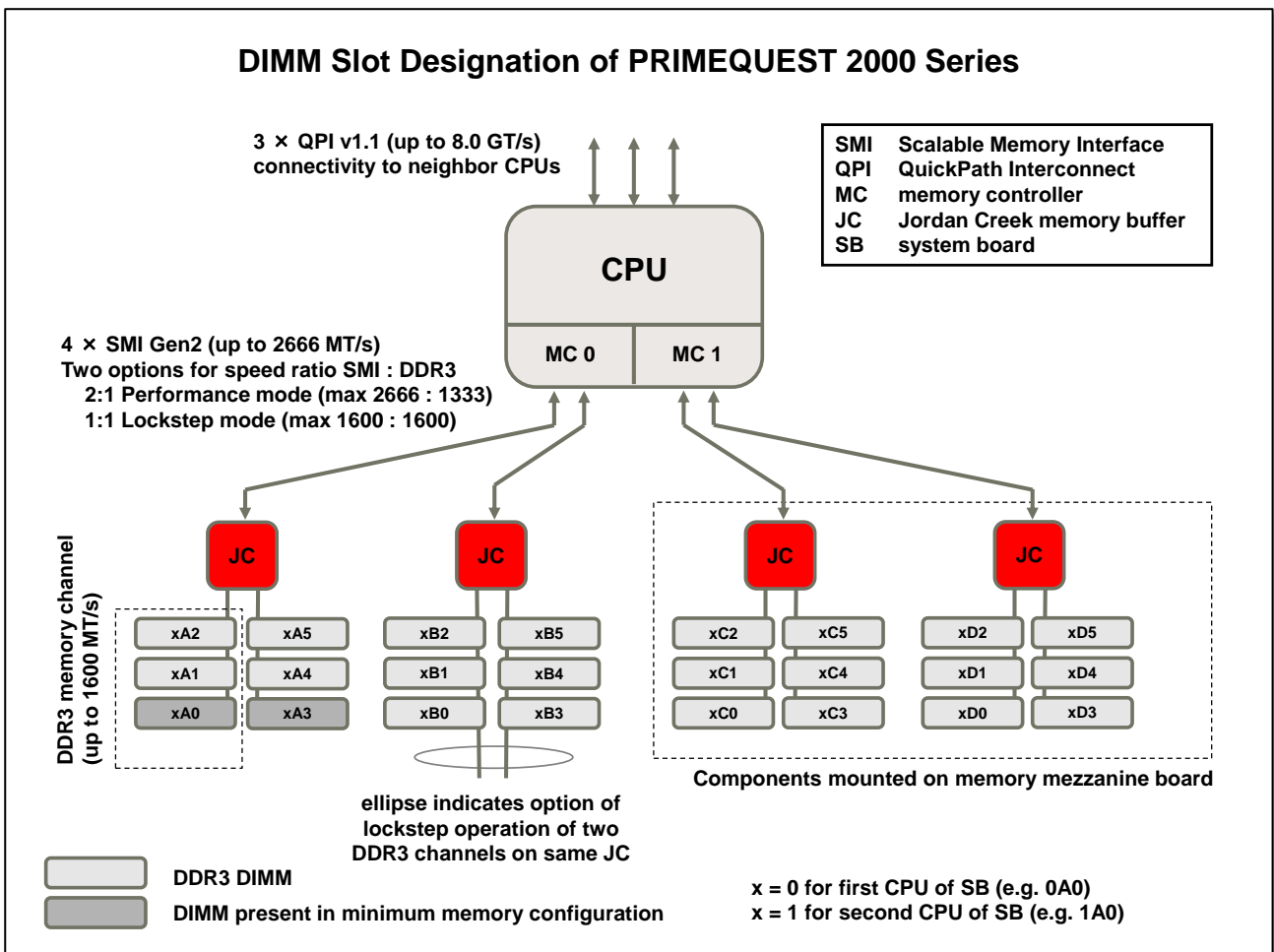
This section provides an overview of the memory system in five parts. Block diagrams explain the arrangement of the available DIMM slots. The available DIMM types are listed in the second section. This is followed by a section about the firmware and BIOS parameters that affect the memory system. The fourth section deals with the influences on the effective memory frequency. The last section provides a table of memory configurations, which with regard to memory performance are to a certain extent "ideal".

DIMM slots

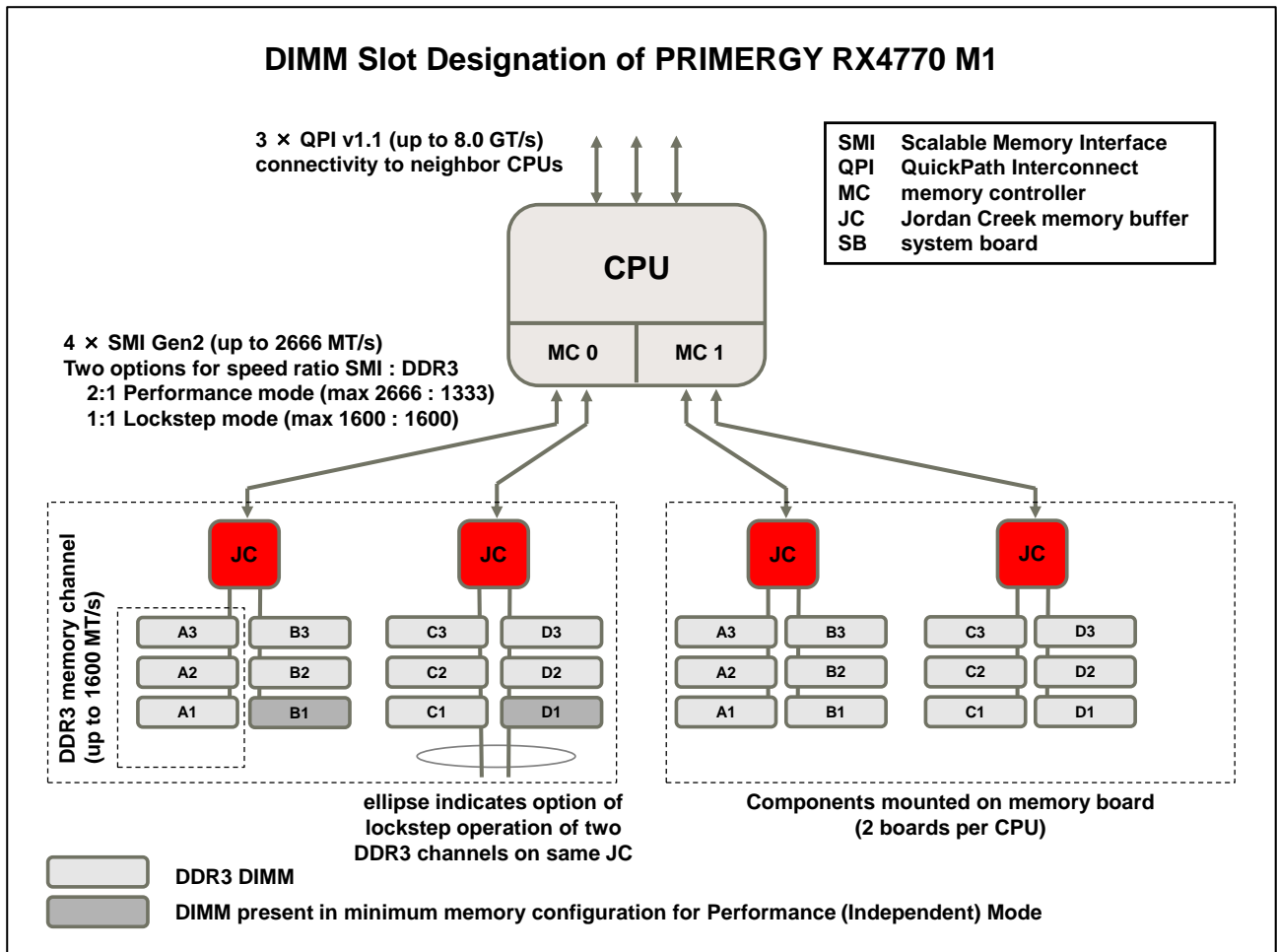
The following diagrams show the memory connection from the viewpoint of the individual Ivy Bridge-EX processor. Each processor has two integrated memory controllers. Each controller is connected to two Jordan Creek 1 memory buffers via bidirectional, serial SMI Gen2 (Scalable Memory Interface) links. There are two DDR3 memory channels, each with three DIMM slots, behind each memory buffer. Thus, there is a total of 24 DIMM slots per processor.

The support of three DIMM strips per memory channel is a new feature of Jordan Creek 1. Only two strips per channel were possible with the predecessors Mill Brook 1 and 2. The number of DIMMs configured per channel is referred to as the DPC (DIMMs per channel) value of the configuration. The value has a certain influence on performance. If the channels are not equally configured, the largest DPC value is decisive for the entire system.

The systems of the PRIMEQUEST 2000 series are configured from system boards with two processors each and their memory resources. The DIMM slots are denoted as specified in the diagram, whereby the placeholder x is for 0 in the case of the slots of the first processor and 1 for the second processor. For each processor half of the 24 slots are on the system board itself. The other half is on an installed Mezzanine board.



All four processors are located on a single system board in the PRIMERGY RX4770 M1. The DIMM slots are on memory boards with 12 slots each, i.e. there are up to two memory boards for each processor. The configurator differentiates between configurations with one or two memory boards per processor. The name of the slots can clearly only be within a memory board. A full name requires the additional specification of the memory board.



The ellipse, which can be seen as an example in the diagrams on the two DDR3 channels of a memory buffer, indicates the option of operating two channels in Lockstep Mode. In this operating mode every memory access takes place synchronously via both channels, i.e. the block that is to be read or written is split over the two channels. The reason for this is to improve the correctability of memory errors. Thus, support is provided by Lockstep Mode for x4 DDDC (Double Device Data Correction), a stronger feature than the x4 SDDC (Single Device Data Correction) with independent memory channels. Lockstep operation mode always applies on a system-wide basis, i.e. for all memory channels.

The improved RAS of Lockstep Mode is at the expense of the memory bandwidth, because the eight physical memory channels of a processor are reduced to four logical ones. This restricts the capacity to be parallelized and thus the performance of memory access. Innovation to the components Jordan Creek 1 and SMI Gen2 make this mode optional. The system or partition is either in Lockstep or Performance / Independent Mode. The systems of the two predecessor generations Nehalem-EX and Westmere-EX were always in Lockstep Mode.

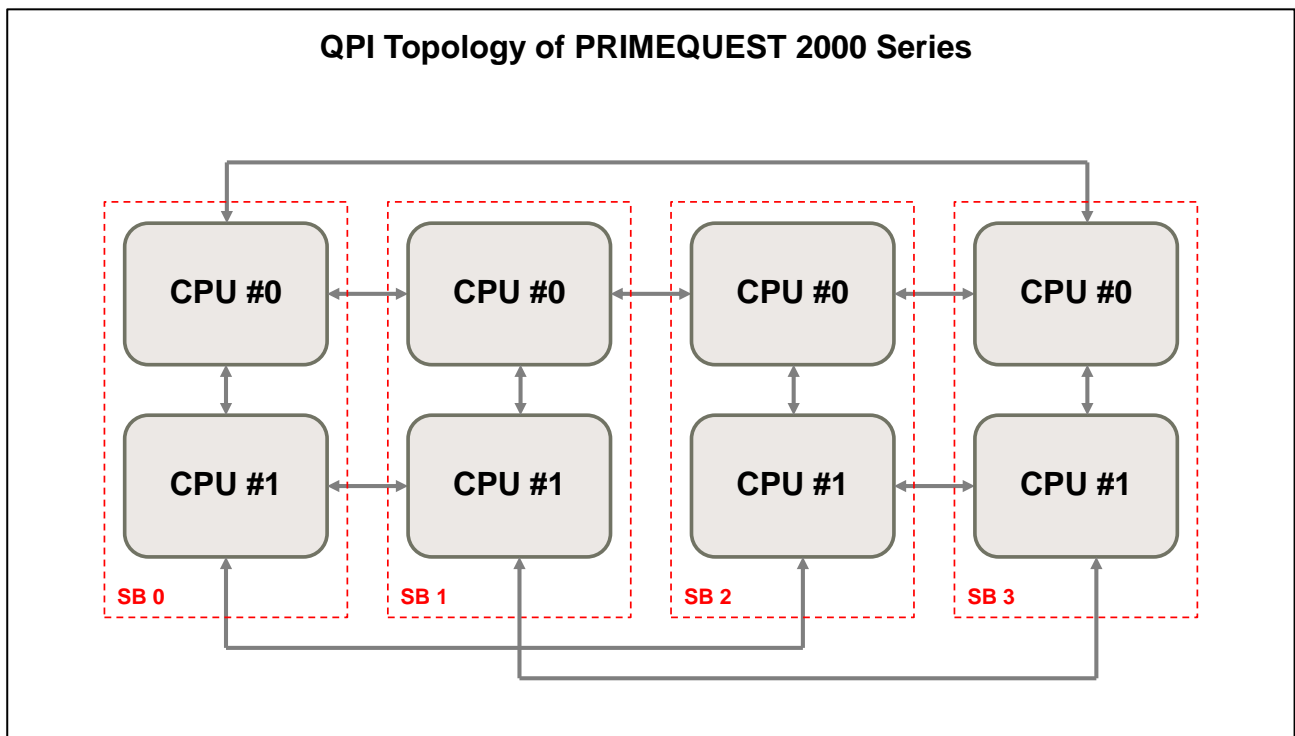
The eligibility of the operating mode influences the frequency of the resources SMI Gen2 link and DDR3 channel. Since eight channels are only opposed by four SMI Gen2 links, the links to implement maximum memory bandwidth in Performance Mode have twice the speed of the memory channels. The frequency in Lockstep Mode is on the other hand the same. The diagrams show the maximum possible frequencies for both cases. In Performance Mode they are caused by the SMI Gen2 upper limit of 2666 MT/s and in Lockstep Mode by the Jordan Creek 1 upper limit of 1600 MHz for the DDR3 frequency. Hence the anomaly

that the mode with less efficient performance (Lockstep) supports the higher DDR3 frequency. However, the higher memory bandwidth is more valuable than the DDR3 frequency that is one step higher.

In previously shown diagrams the dark-gray shading refers in each case to the minimum configuration, which consists of two DIMM strips. This is where there are differences between the PRIMEQUEST 2000 series and the PRIMERGY RX4770 M1.

In the PRIMEQUEST 2000 series, as mission-critical servers, there are as a matter of principle only memory configurations that are capable of Lockstep operation. For this purpose, symmetry must always prevail in the Jordan Creek 1 memory buffers with regard to the two memory channels. The marked minimum configuration takes this mode into account. The second, configured slot pair would be xC0 / xC3, accordingly followed by xB0 / xB3 and xD0 / xD3, etc. The configuration sequence across the existing memory channels ensures even utilization of all available memory resources and is relevant to performance.

The right to Lockstep capability in each memory configuration does not exist in the PRIMERGY RX4770 M1. The minimum configuration, which also consists of two DIMM strips, follows in this case from the premise of the best possible performance, which is achieved by incorporating the second memory buffer. This configuration permits Performance mode only. The Lockstep-capable minimum configuration of the PRIMERGY RX4770 M1 (not marked) consists of four DIMMs in positions A1, B1, C1 and D1 of the first memory board.

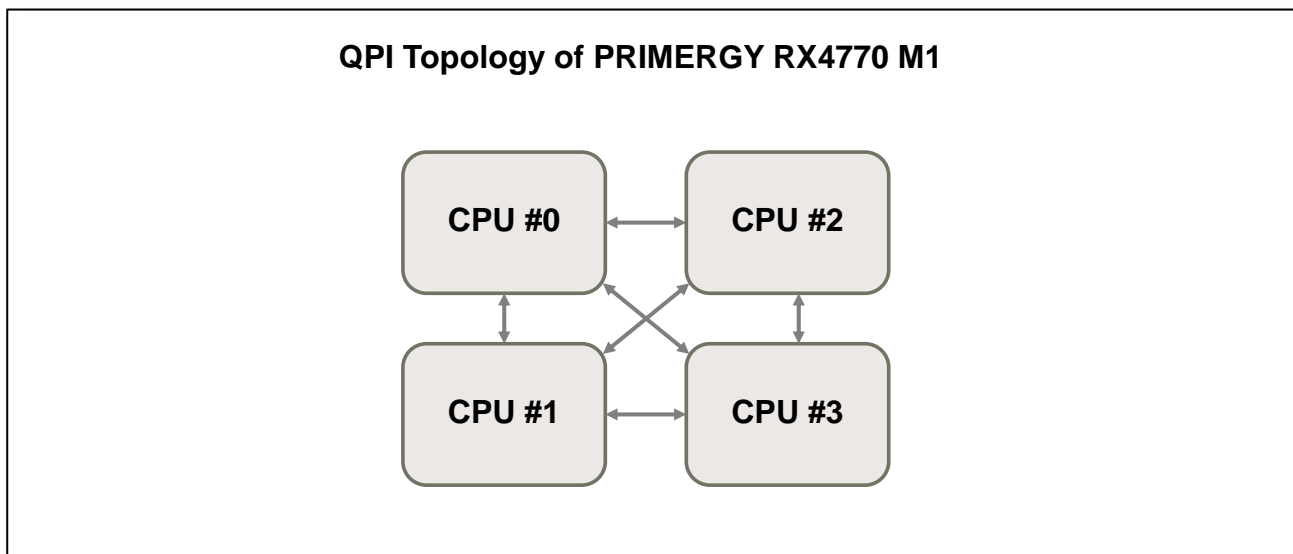


This diagram shows the QPI topology of the PRIMEQUEST 2000 series, i.e. the networking of the processors and their appropriate memory components. Since the networking is only via the three QPI Links per processor, the already discussed components SMI Gen2 links, memory buffers and DIMM slots have now been omitted. Also omitted from all diagrams are incidentally the 32 on-chip PCIe Gen3 lanes per processor, because they do not directly concern the memory architecture.

Every processor in the full configuration of the PRIMEQUEST 2000 series with eight processors is only directly connected to three of seven neighbors. 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.

In the PRIMEQUEST 2400E model with a maximum of four processors there are only system boards 0 and 1. In this case and in the case of PRIMEQUEST 2800E partitions with less than four system boards there are unused QPI interfaces.

The PRIMERGY RX4770 M1 is limited to four processors from the outset. This permits a system design, in which every processor is connected to each other with the help of the three QPI links per processor. Thus, the QPI topology shown in the following diagram is different from the topology of the PRIMEQUEST 2000 series, and in particular from the topology of the PRIMEQUEST 2400E model.



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

DIMM types

DIMM strips according to the following table are considered for the memory configuration. There are registered (RDIMM) and load-reduced (LRDIMM) DIMMs. Mixed configurations consisting of these two DIMM types are not possible. This always applies for all systems with DDR3 memory.

Memory modules (since system release)										
Memory module	Type	Capacity [GB]	Ranks	Bit width of the memory chips	Frequency [MHz]	Low voltage	Load reduced	Registered	ECC	Relative price per GB
16GB (2x8GB) 1Rx4 L DDR3-1600 R ECC (2 x 8 GB 1Rx4 PC3L-12800R)	RDIMM	16	1	4	1600	✓		✓	✓	1.0
32GB (2x16GB) 2Rx4 L DDR3-1600 R ECC (2 x 16 GB 2Rx4 PC3L-12800R)	RDIMM	32	2	4	1600	✓		✓	✓	0.9
64GB (2x32GB) 4Rx4 L DDR3-1600 LR ECC (2 x 32 GB 4Rx4 PC3L-12800L)	LRDIMM	64	4	4	1600	✓	✓	✓	✓	1.5
128GB (2x64GB) 8Rx4 L DDR3-1333 LR ECC (2 x 64 GB 8Rx4 PC3L-10600L)	LRDIMM	128	8	4	1333	✓	✓	✓	✓	3.3

The table takes into account the fact that DIMMs are offered in units of two each in the order and configuration processes of the PRIMEQUEST 2000 series and of PRIMERGY RX4770 M1. The reason is the configuration specification in pairs.

Data is transferred between the memory controller and DIMMs in units of 64 bits for all DIMM types. This is a feature of DDR3-SDRAM memory technology (Synchronous Dynamic Random Access Memory). A memory area of this width is set up on the DIMM from a group of DRAM chips - with the individual chip being responsible for 4 or 8 bits (see the code x4 in the type name, x8 modules are not planned for the Ivy Bridge-EX based servers at present). Such a chip group is referred to as a rank. According to the table there are DIMM types with 1, 2, 4 or 8 ranks. The number of available ranks per memory channel has a certain influence on performance, which is explained below. Maximum capacities are the motivation for DIMMs with four or eight ranks, but at the same time the DDR3 specification only supports a maximum of eight ranks per memory channel.

That being said, the essential features of the two DIMM types are as follows:

- RDIMM: The control commands of the memory controller are buffered in the register (that gave the name), which is in its own component on the DIMM. This relief for the memory channel enables configurations with up to 3DPC (DIMMs per channel). Only 2DPC configurations are possible for unbuffered (UDIMM) DIMMs, which are to be found in smaller server classes.
- LRDIMM: Apart from the control commands, the data itself is also buffered in a component to be found on the DIMM. Furthermore, the Rank Multiplication function of this DIMM type can map several physical ranks onto a logical one. The memory controller then only sees logical ranks. Rank Multiplication is enabled if the number of physical ranks in the memory channel is greater than eight.

The decision in favor of one of the type groups RDIMM or LRDIMM is usually based on the required memory capacity. The performance influences of frequency and number of ranks exist in the same way for both types; these influences are independent of type. Type-specific performance influences exist; but they are so minor that they can be disregarded in most cases. Two examples of such influences are to be given here. However, a systematic quantitative evaluation does not take place below due to insignificance:

- The local memory access of the Ivy Bridge-EX based servers has a latency of about 110 ns for RDIMMs in the case of an unloaded system. This value applies for the memory frequency 1333 MHz. It is 5-10 ns higher for LRDIMMs - also at 1333 MHz. The cause for this is the more complex buffer component on the DIMM. To estimate this difference you should take into consideration that the latencies increase in the loaded system, which in turn reduces the percentage of the aforesaid difference.

- Rank Multiplication in the case of configurations with LRDIMMs with more than eight physical ranks per memory channel results in a reduction in the maximum memory bandwidth and the application performance – in comparison to configurations with RDIMMs – of about 5%.

All the DIMM types can be run with 1.5 V or energy-saving 1.35 V. However, the 1.35 V or LV (low voltage) operation can mean lower memory frequency and thus a reduction in memory performance. The following section about firmware and BIOS parameters describes the administration parameters for this trade-off.

It is then followed by a section on the effective memory frequency of a given configuration. Apart from the trade-off with energy consumption, it depends on a number of additional influencing factors. The maximum frequency stated in the DIMM type table is merely to be understood as the upper limit for this effective frequency.

The last column in the DIMM table shows the relative price differences. The list prices from May 2014 for the PRIMERGY RX4770 M1 are used as a basis. The column shows the relative price per GB, standardized to the RDIMM, size 8 GB (highlighted as measurement 1.0). There are higher costs for the LRDIMMs, with which very large memory capacities can be achieved, particularly for the comparably new 64 GB LRDIMM. Furthermore, the picture of the relative prices is subject to constant change. The table is to be understood as a snapshot.

Depending on the PRIMEQUEST or PRIMERGY model there can be restrictions regarding the availability of certain DIMM types. The current configurator is always decisive. Furthermore, some sales regions can also have restrictions regarding availability.

Firmware and BIOS parameters

The parameters to be described in this section are a result of the functionality of the Ivy Bridge-EX processors and are thus principally the same for the PRIMEQUEST 2000 series and the PRIMERGY RX4770 M1. However, there are differences in naming, default assignment and positioning in the firmware and BIOS menus, which result from the various functional demands of the server classes.

Before going into the syntactic details, here is a summary and explanation of the influencing factors we are dealing with here:

- The alternative between independent memory channels with higher performance (referred as Performance or Independent mode) and the fail-safe Lockstep mode (referred to as Lockstep or Normal mode).
- Activation of the RAS functions Mirroring or Sparing. Here, the PRIMEQUEST 2000 series and the PRIMERGY RX4770 M1 differ in the fact that Mirroring and Sparing are only possible in Lockstep mode with the PRIMEQUEST 2000, whereas these functions are also supported in Performance mode with the PRIMERGY RX4770 M1.
- The trade-off between energy-saving 1.35 V operation of the DIMMs and 1.5 V operation, which *can* enable higher memory frequency and consequently memory performance. However, the general conditions of the memory configuration go into detail here: the trade-off is not a given with every memory configuration, i.e. the appropriate parameter can be ineffective. Hence, there is no 1.35 V operation with 3DPC configurations, regardless of with which DIMM type, and use of the 64 GB LRDIMM. On the other hand, the DIMMs are only operated with 1.5 V if a higher memory frequency can as a result be achieved than with 1.35 V. The (rather small) number of situations, in which such a higher frequency is eligible, are listed in the following section.
- In the case of Patrol Scrubbing the entire main memory is searched in cycles of 24 hours for correctable memory errors and, if necessary, correction is initiated. This reduces the probability of errors that are no longer correctable. The operation is controlled by the memory controllers. Highly sensitive performance measurements may be a reason for temporarily disabling this functionality. However, establishing proof of an effect on performance may be difficult. Support of this parameter may be delayed with the Ivy Bridge-EX based servers.
- The refresh rate concerns an elementary feature of DRAM. Since the electric charge, which decides on the information value 1 or 0 of a bit, diffuses, every memory cell has to be continually refreshed in cycles within the micro-second range. This is also controlled by the memory controllers. The timing of the cycles (refresh rate) is done by the BIOS. The background to the availability of the parameter at the administration interface is as follows. Some memory types can with certain rare access patterns show an accumulation of correctable memory errors that is known as the pass-gate effect. To eliminate this, the BIOS sets the refresh rate to double for such DIMMs. The doubling of the rate is associated with a performance disadvantage of some 2%. This is caused by the specific overhead, which represents the control of the refresh for the signal lines between the memory controller and the DIMMs. If - while accepting the possibility of accumulated correctable memory errors - this disadvantage is regarded as unacceptable in sensitive performance measurements, the doubling can be reversed.

This preliminary comment is now followed by the specific syntactic design for the PRIMEQUEST 2000 series and the PRIMERGY RX4770 M1. In the case of the PRIMEQUEST 2000 series the parameters are to be found in two different administration interfaces.

Interfaces of the Web-GUI of PRIMEQUEST 2000 series

The parameter Memory Operation Mode with the following options is under Partition / Partition# / Mode (partitionable PRIMEQUEST 2000 models) and System (PRIMEQUEST 2800B) in the Web-GUI of the management board (MMB)

- Performance Mode
- Normal Mode
- Partial Mirror Mode
- Full Mirror Mode
- Spare Mode

The default valid at product launch is underlined. The entire configured physical main memory is available to the operating system in Normal and Performance Mode. Normal Mode stands for the lockstep operation

mode of the memory channels with its more demanding RAS feature. Performance Mode stands for the higher-performance operation mode of independent memory channels.

Only a part of the configured memory capacity, for example 50% with Full Mirror, is available to the operating system with the three redundant modes Partial Mirror, Full Mirror and Spare. In the case of Sparing the percentage of the net capacity depends on the DIMM type. If the 8 GB RDIMM is used, the net capacity is two thirds, and for the other DIMM types it is five sixths of the configured capacity. The specification for Spare Mode of always configuring with 3DPC is also included in this calculation.

The three redundant modes are based on the lockstep operation mode of the memory channels. They are additions to Lockstep Mode. There is no Mirroring and Sparing in connection with the independent memory channels of Performance Mode.

Interfaces of the device manager of PRIMEQUEST 2000 series

Further parameters are to be found in the BIOS, under Device Manager / Memory Configuration to be more precise. This interface can be accessed via the console of the partition or the system. There are three parameters here with the following options; once again the defaults valid at product launch are underlined:

- DIMM Speed: Performance Mode / Normal Mode
- Patrol Scrub: Disabled / Enabled
- Refresh Rate: 1x / Auto

The first parameter concerns the trade-off with the energy consumption of the DIMMs. There is no connection with the Memory Operation Modes of the same name. Here the Normal Mode setting stands for energy-saving 1.35 V operation of the DIMMs if possible. The Performance Mode setting stands for 1.5 V operations if this enables a higher memory speed. The details are to follow in the next section.

For the second and third parameters the explanation already given above applies.

Interfaces of the BIOS of PRIMERGY RX4770 M1

In the case of the PRIMERGY RX4770 M1 there is a Memory Configuration submenu in the BIOS under Advanced with the following parameters:

- DDR Performance: Performance optimized / Low Voltage optimized / Energy optimized
- Memory Mode: Normal / Mirror / Sparing
- DRAM Maintenance: Auto / Manual
- Refresh Rate Multiplier: Disabled / Enabled
- Apply Memory RAS policy globally: Disable / Enable
- VMSE Lockstep Mode: Independent / Lockstep

The defaults valid at product launch are underlined here as well.

The first parameter DDR Performance concerns the trade-off between 1.35 V and 1.5 V operation - if this arises. Reference is once again made to the next section for the details. The third option Energy optimized results in a reduction in the memory frequency to 1066 MHz which is minimal for the Ivy Bridge-EX. However, it should be pointed out that 1.35 V operation is the decisive influence on the energy consumption of the DIMMs, not so much the memory frequency itself. It is uncertain whether energy savings occur beyond the setting Low Voltage optimized.

The second parameter Memory Mode concerns the activation of the RAS functions Mirroring and Sparing. Additional subitems appear in the Mirroring setting, which enable activation at the level of the individual memory controller.

The Refresh Rate Multiplier parameter, for which the Disabled setting means the reversal of the doubling of the refresh rate as explained above, is only effective for "DRAM Maintenance = Manual". The Patrol Scrubbing subparameter also appears in the case of "DRAM Maintenance = Manual". DRAM Maintenance should be understood as an indication that the Refresh Rate and Patrol Scrubbing parameters are only to be changed for a good reason.

"Apply Memory RAS policy globally" can be used to ensure that the RAS functionality is activated on a system-wide basis or not at all.

The last parameter "VMSE Lockstep Mode" concerns the alternative between independent memory channels (Independent) and Lockstep mode, which in the case of the PRIMERGY RX4770 M1 is independent of the optional activation of the RAS modes Mirroring and Sparing.

Definition of the memory frequency

The effective memory frequency of a configuration – a key parameter when it comes to memory performance – depends on a range of general conditions. The three values 1600, 1333 and 1066 MHz are eligible for the Ivy Bridge-EX based servers. The frequency is defined by the BIOS when the system or partition is switched on and applies per system or partition, not per processor.

The mentioned general conditions include some of the BIOS settings dealt with in the previous section. Due to the syntactic differences between the PRIMEQUEST 2000 series and the PRIMERGY RX4770 M1 and in order to simplify matters reference is only made at the semantic level. As far as the operation mode of the memory channels is concerned, a distinction is made in this section between Lockstep and Independent, and regarding the energy efficiency trade-off between Low-voltage (LV) and Performance. In this case, the PRIMERGY RX4770 M1 has as the only semantic difference a third setting “Energy Efficient”, which results – without any further dependencies – in a reduction to the minimum frequency of 1066 MHz and thus no longer needs to be considered in the following case-by-case analysis.

Initially, the configured processor model is of significance for the definition of the memory frequency. Within the context of this document the classification of the Ivy Bridge-EX models according to the following table is recommended.

CPU type	QPI (GT/s)	Maximum memory frequency (MHz) as per channel operation mode		Xeon E7-8800 v2 Models	Xeon E7-4800 v2 Models
		Independent	Lockstep		
Advanced	8.0	1333	1600	E7-8890 v2, E7-8880 v2, E7-8870 v2, E7-8893 v2, E7-8857 v2	E7-4890 v2, E7-4880 v2, E7-4870 v2
Standard	7.2	1066	1333	E7-8850 v2	E7-4850 v2, E7-4830 v2, E7-4820 v2

The further tables of this section differentiate with regard to these two CPU classes. Another influence on the memory frequency is the operation mode of the memory channels. A distinction must be made between the cases Independent and Lockstep. The reason for this distinction has already been provided above: in order to make full use of the maximum bandwidth for eight independent memory channels the SMI Gen2 links have twice the frequency speed of the DDR3 channels. There are upper limits for the SMI Gen2 frequency, which have a retroactive effect on the DDR3 frequency.

In the differentiation concerning channel operation mode a distinction must also be made according to the BIOS parameter for the trade-off with energy consumption. The effectiveness of the parameter is clearer if only the fields, for which there is actually a difference to the low-voltage assignment, are completed in the tables for the assignment Performance. In the empty configuration cases the parameter is ineffective, i.e. the same values apply both for memory frequency and DIMM voltage as is the case with low-voltage setting.

Having handled processor model as well as the parameters concerning channel operation mode and the energy trade-off, DIMM type and the DPC value remain the last influences on the memory frequency to be taken into consideration. That being said, the effective memory frequency and DIMM voltage of a given configuration are as follows.

Lockstep channel operation mode

Energy trade-off: low-voltage grey shading: 1.5V – no shading: 1.35V									
	RDIMM			32 GB LRDIMM			64 GB LRDIMM		
CPU type	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC
Advanced	1333	1066	1066	1333	1333	1333	1066	1066	1066
Standard	1333	1066	1066	1333	1333	1333	1066	1066	1066

Energy trade-off: performance ¹ grey shading: 1.5V – no shading: 1.35V									
	RDIMM			32 GB LRDIMM			64 GB LRDIMM		
CPU type	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC
Advanced		1333		1600	1600				
Standard		1333							

¹ ineffective if fields are not completed

Independent channel operation mode

Energy trade-off: low-voltage grey shading: 1.5V – no shading: 1.35V									
	RDIMM			32 GB LRDIMM			64 GB LRDIMM		
CPU type	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC
Advanced	1333	1066	1066	1333	1333	1333	1066	1066	1066
Standard	1066	1066	1066	1066	1066	1066	1066	1066	1066

Energy trade-off: performance ¹ grey shading: 1.5V – no shading: 1.35V									
	RDIMM			32 GB LRDIMM			64 GB LRDIMM		
CPU type	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC
Advanced		1333							
Standard									

¹ ineffective if fields are not completed

Memory frequency was initially referred to as a key parameter for memory performance. The highest possible frequency 1600 MHz is achieved with Lockstep channel operation mode. Nevertheless, the other operation mode is referred to as Performance or Independent Mode. How is this justified?

The percentage differences between the frequency rates 1066, 1333 and 1600 MHz are 20-25% per level and more than three quarters transfer to the memory bandwidth, expressed as GB/s. In the case of such a magnitude you may well speak of a key parameter.

However, at the same memory frequency there is about 80% between the bandwidth of eight independent memory channels on the one hand and four logical channels in Lockstep Mode on the other hand. Independent memory channels enable a bandwidth advantage, which Lockstep operation can reduce in a number of cases through higher memory frequency, but cannot equal. Furthermore, eligibility as regards channel operation mode is always given, whereas the option of higher memory frequency, as described in this section, depends on general conditions.

It goes without saying that the percentage ratios cannot be transferred from the memory bandwidth to application performance on a one-to-one basis. The impact on the latter is significantly lower. The precedence of influencing factors is, however, the same.

Ideal memory capacities

In summary, two main influences on the memory performance of the Ivy Bridge-EX based servers have been named so far. Firstly, the trade-off between RAS (Lockstep) and performance that is controlled by the memory operation mode parameters. Secondly, a range of dependencies that affect memory frequency. The influences and the fine tuning of firmware and BIOS that affects them have been addressed. The respective percentage differences in performance follow in the second part of the document.

A third main influence is the number of configured DIMMs, which are directly connected to the required memory capacity. The limits for the minimum (2 DIMMs per processor) and maximum (24 DIMMs per processor) configuration have already been stated. There is a range of memory configurations between these limits, which are ideal when it comes to making optimal use of the memory architecture. They require 8, 16 or 24 DIMMs per processor. The following table lists these configurations. For PRIMERGY RX4770 M1 it is essential that two memory boards per processor are configured.

Ideal memory capacities for various CPU configurations of Ivy Bridge-EX based systems									Benchmark
GB for 2 CPU	GB for 4 CPU	GB for 8 CPU	DPC	DIMM Type (8 DIMMs per CPU and DPC)	Mem Operation = Normal		Mem Operation = Performance		
					MHz 1.35V	MHz 1.5V	MHz 1.35V	MHz 1.5V	
128	256	512	1	8GB RDIMM	1333		1333		
256	512	1024	2	8GB RDIMM	1066	1333	1066	1333	+
			1	16GB RDIMM	1333		1333		
384	768	1536	3	8GB RDIMM		1066		1066	
512	1024	2048	2	16GB RDIMM	1066	1333	1066	1333	+
			1	32GB LRDIMM	1333	1600	1333		
768	1536	3072	3	16GB RDIMM		1066		1066	
1024	2048	4096	2	32GB LRDIMM	1333	1600	1333		+
			1	64GB LRDIMM		1066		1066	
1536	3072	6144	3	32GB LRDIMM		1333		1333	
2048	4096	8192	2	64GB LRDIMM		1066		1066	
3072	6144	12288	3	64GB LRDIMM		1066		1066	

All eight memory channels per processor are treated equally in these configurations. This is the decisive feature that enables ideal distribution or parallelization of the load that ensues on the memory system. None of the existing memory resources, such as the memory controller, SMI Gen2 link, Jordan Creek memory buffer or DDR3 channel, remains unused for the configurations in the table. At the same time, the uniformity in all memory channels ensures that all algorithms conveniently "work out even", which parallelize memory access in the microcode of the memory controllers. The corresponding technical term, which we will look at in depth below, is Interleaving.

The table is sorted according to total GB capacities for system or partition. In each line the values for configuration levels are specified with two, four or eight processors, based on the assumption that every processor is configured equally. This assumption was referred to in the introduction as the basic rule for the memory configuration of powerful systems. The technical background is the difference between local and remote memory access in NUMA system architecture. Experience unfortunately shows that in practice the rule is not a matter of course.

Treating all memory channels of a processor equally means that configuration is done in groups of eight DIMMs. There are three DIMM slots per channel so that one, two or three such groups can be connected per processor. This corresponds to the DPC (DIMMs per channel) value of the configuration. Thus, the total capacity shown in the table is calculated according to the formula:

$$Capacity\ in\ GB = 8\ memory\ channels \times DPC \times DIMM\ size\ in\ GB \times number\ of\ CPUs$$

The table states the possible memory frequencies for each configuration, whereby the distinctions already discussed should be taken into consideration. These memory configurations have by all means the feature of optimal channel interleaving, regardless of how the trade-offs between RAS (Lockstep) and performance, as well as energy consumption and performance were decided. Even if the decisions went against performance in these trade-offs, the configurations retain this feature of the best possible interleaving. Furthermore, it undoubtedly makes sense in productive operations to have as an objective a well-balanced memory performance instead of a best possible one at all costs. The quantitative statements below in the second part of the document should be useful when it comes to weighing up these influences against each other.

Considerations of this kind even exist in benchmarking when energy efficiency metrics are involved, or, as in the case of database benchmarks, where - although very large memory capacities save I/O - they are afflicted with lower achievable memory frequencies.

It goes without saying that the memory configurations used for the standard benchmarks of the PRIMEQUEST 2000 series and of PRIMERGY RX4770 M1 are also among the optimal configurations in the table. They are marked with a + sign in the last column. Since memory configurations are for cost reasons more likely to be found at the lower end of the supported capacity scale in practice, it would seem necessary to emphasize why the smallest configuration in the table is avoided for sensitive performance measurements. In this configuration there is on account of the design of the 8 GB RDIMM only one single rank in the memory channel, which means a performance disadvantage of a few per cent, for which more reasons are given below. This should not usually play a role in productive operations. However, such a disadvantage is unwanted in benchmarking or with special performance expectations.

Quantitative effects on memory performance

After the functional description of the memory system with qualitative information, we now have statements on the basis of percentages about how differences in memory configuration affect performance. As a means of preparation the first section deals with the two benchmarks STREAM and SPECint_rate_base2006, which were used to characterize memory performance. The latter benchmark acts as a model for commercial application performance.

This is followed by a section on interleaving across memory controllers and channels, which also includes as a topic the difference between the Independent and Lockstep channel operating modes. Further sections deal with interleaving across ranks and memory frequency. A section about memory performance under redundancy, i.e. with enabled Mirroring or Sparing, makes up the end of this document. When testing each individual feature we attempt to hide the other features so as not to mix up the influences.

The following table describes the measurement configurations. In the case of PRIMEQUEST 2000 series, the tests were carried out in partitions consisting of one and four system boards with two processors each. As the results were not significantly dependent on partition size, differentiation in this respect was omitted in the following sections.

System Under Test (SUT)		
Hardware		
Model	PRIMEQUEST 2800E	PRIMERGY RX4770 M1
Processor type	Xeon E7-8890 v2	Xeon E7-4890 v2
Memory types	16GB (2x8GB) 1Rx4 L DDR3-1600 R ECC 32GB (2x16GB) 2Rx4 L DDR3-1600 R ECC 64GB (2x32GB) 4Rx4 L DDR3-1600 LR ECC	32GB (2x16GB) 2Rx4 L DDR3-1600 R ECC 64GB (2x32GB) 4Rx4 L DDR3-1600 LR ECC
Disk subsystem	1 x RAID Ctrl SAS 6G 1GB 1 x HD SAS 6G 300 GB 15K HOT PL 2.5" EP	1 x RAID Ctrl SAS 6G 1GB 1 x HD SAS 6G 300 GB 15K HOT PL 2.5" EP
Software		
Firmware	Unified Firmware 14012 (BIOS, BMC, MMB)	BIOS R1.3.0, BMC 7.24F
Operating system	Red Hat Enterprise Linux Server release 6.5	Red Hat Enterprise Linux Server release 6.5

The following tables always show the relative performance. The absolute measurement values for the STREAM and SPECint_rate_base2006 benchmarks under ideal memory conditions, which are usually equivalent to the 100% measurement of the tables, are as a further differentiation in terms of the various processor models included in the performance reports of the PRIMEQUEST 2800E [\[L6\]](#) and of the PRIMERGY RX4770 M1 [\[L7\]](#).

The memory performance tests are carried out using the most powerful processor models Xeon E7-8890 v2 and Xeon E7-4890 v2, respectively, and this means that the performance differences can be seen with the utmost clarity. The differences are somewhat slighter with less powerful processors, which should be taken into consideration when transferring statements on the basis of percentages to such configurations.

Benchmark measurements are usually characterized – this applies for STREAM and SPECint_rate_base2006 – by system utilization of close to 100%, which is not typical for productive operations. This mitigating factor should also be taken into consideration when evaluating statements on the basis of percentages. However, there is no simple formula when it comes to considering utilization.

The measuring tools

Measurements were made using the benchmarks STREAM and SPECint_rate_base2006.

STREAM Benchmark

The STREAM benchmark from John McCalpin [L4] 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 used below to quantify memory performance are based on this practice and concern 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 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 Benchmark

The benchmark SPECint_rate_base2006 was added as a model for commercial application performance. It is part of SPECcpu2006 [L5] from Standard Performance Evaluation Corporation (SPEC). SPECcpu2006 is the industry standard for measuring the 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 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 set-up of the physical address space by alternating between multiple memory resources of the same type. First of all, the two memory controllers of a processor are eligible in the case of the Ivy Bridge-EX based servers. The first block of the local address space segment is in the first controller, the second one in the second, the third one back in the first, etc. This principle can then be continued on the level of the four memory channels per controller, and finally on the level of the ranks within the individual memory channel.

Identical memory capacities in the respective resources are the decisive prerequisite for this pattern: only then can the alternating work. The procedure in case this is not met is explained in the next section. Incidentally, the pattern requires a certain flexibility as regards block sizes for alternating. They are not identical on the levels of the controllers and channels on the one hand and ranks on the other hand.

Memory access, which according to the locality principle is mainly to adjacent memory areas, is as a result of interleaving distributed across all resources of the memory system. This performance gain situation results from parallelism. The interleaving we are currently looking at across memory controllers and memory channels may be referred to as the most important influence on memory performance, ahead of the influence of memory frequency.

Interleaving across memory controllers and memory channels

In the case of the ideal memory capacities, as considered above, with 8, 16 or 24 DIMMs of the same type per processor interleaving across controllers and channels is developing with optimal effect. There is a loss in performance as per the following table for configurations with other numbers of DIMMs, particularly with less than eight DIMMs per processor and right through to the minimal configuration. Bold print refers to the best case in each of the three categories interleaving, memory bandwidth and commercial application performance.

Channel interleaving of PRIMEQUEST 2000 series				
	Operating mode	8 DIMMs per CPU (and multiples)	4 DIMMs per CPU	2 DIMMs per CPU
		Ideal capacities		Minimum configuration
Interleaving (Controller / Channel)	Independent	2-way / 4-way	2-way / 2-way	1-way / 2-way
	Lockstep	2-way / 2-way	2-way / 1-way	1-way / 1-way
Memory bandwidth (STREAM)	Independent	100%	56%	28%
	Lockstep	57%	30%	16%
Commercial application performance (SPECint_rate_base2006)	Independent	100%	93%	73%
	Lockstep	94%	76%	54%

The first horizontal block of the table (Interleaving) specifies the interleaving for the configuration cases. n-way here means that the configuration enables alternating between n controllers and channels. The blocking size of this alternating is based on the cache line size of the processors of 64 bytes.

At this point you can see where the "problem" of the Memory Operation mode Normal (Lockstep) lies with regard to memory performance. In this case, the alternating must take place on the level of the logical memory channels, for which two physical channels are combined in each case. The splitting of a 64-byte block into two physical channels takes place below the addressing level, of which alternating is an integral part. The enabling of Lockstep Mode halves interleaving across memory channels. That is why this operating mode is not performance-neutral.

The bottom horizontal blocks in the table show the relative performance effects for memory bandwidth and the benchmark SPECint_rate_base2006, which serves as a model for commercial application performance. The best case in both the categories STREAM and SPECint_rate_base2006 has a performance of 100%; the other configuration cases are associated with the reductions shown. The relationships of the memory bandwidth as expressed by STREAM should be understood as extreme cases, which cannot be ruled out in certain application areas, especially in the HPC (High-Performance Computing) environment. However such behavior is improbable for most commercial loads. This assessment of the interpretation quality of STREAM and SPECint_rate_base2006 not only applies for the performance aspect dealt with in this section, but also for all following sections.

The measurement values confirm the logical understanding of Lockstep Mode (from a pure performance viewpoint and excluding the RAS value added) as Independent Mode with halved channel interleaving. Basically, the test results for Lockstep match those for Independent, one column to the right.

The previously shown table concerns the PRIMEQUEST 2000 series, in which each permitted memory configuration is Lockstep-capable. The Lockstep capability results from the symmetric handling of the two memory channels of every Jordan Creek 1 memory buffer. General Lockstep capability does not apply for the permitted memory configurations of the PRIMERGY RX4770 M1. Furthermore, there is a differentiation in this system with regard to the number of ordered memory boards per processor. Reproduction of these more complex configuration rules is beyond the scope of this document. Thus, knowledge of the configurator of the PRIMERGY RX4770 M1 is prerequisite to understanding the table below.

Channel interleaving of PRIMERGY RX4770 M1					
	Operating mode	Per CPU: 8 DIMMs across 2 Mem Boards Ideal capacities	Per CPU: 4 DIMMs across 2 Mem Boards	Per CPU: 4 DIMMs across 1 Mem Board	Per CPU: 2 DIMMs across 1 Mem Board Minimum configuration
Interleaving (Controller / Channel)	Independent	2-way / 4-way	2-way / 2-way	1-way / 4-way	1-way / 2-way
	Lockstep	2-way / 2-way		1-way / 2-way	
Memory bandwidth (STREAM)	Independent	100%	56%	51%	28%
	Lockstep	57%		29%	
Commercial application performance (SPECint_rate_base2006)	Independent	100%	94%	92%	78%
	Lockstep	94%		79%	

The table is of particular help when it comes to assessing the difference in performance between memory configurations with one or two memory boards per processor. For example, the optimal memory performance comes with eight DIMMs and two memory boards per processor (third column from the left). If on the other hand eight DIMMs and only one board per processor are ordered, the second column from the right is decisive for the channel interleaving that can be achieved. The eight DIMMs per processor (instead of four) merely replenish the capacity of the four memory channels of the one memory board, without any further improvement to the channel interleaving.

A short evaluation of the effects on application performance (see the horizontal blocks for SPECint_rate_base2006 in both tables) is as follows. Benchmarks will always aim for configurations of quality 100%. Cases above 90% are not critical for productive operations, typically with security margins as regards system utilization. Cases of around 70% merit critical examination, for example in the case of a high utilization level targeted under virtualization. In a case of just above 50% you may assume a disparity between the computing performance of the processors and memory performance.

The tables say nothing about the permitted configuration cases with six DIMMs per processor and configurations with more than eight DIMMs if the number of DIMMs is not a multiple of eight. All these are cases, in which alternating does not work, because the partial capacities of the resources in question are not the same. In the case of six DIMMs per processor there are four on the first controller and two on the second one. A homogeneous local address space segment with an identical alternating pattern – and this is precisely where performance quality is to be found – cannot be formed in this case due to the capacity difference at controller level. On the contrary, twelve DIMMs per processor are distributed equally across the controllers as six plus six, but the imbalance occurs within the four channels per controller.

The solution is always to split the physical address space into several segments with different interleaving. The performance of an application can then vary, depending on the segment from which the application is provided with memory. In both the cases mentioned with six and twelve DIMMs the outcome is a memory performance that corresponds to the 4 DIMM cases in the tables. The 2 DIMM cases cannot be ruled out in a number of situations (like ten DIMMs per processor), either. In sensitive applications this behavior can be one reason for avoiding such configurations.

Interleaving across ranks

The method of alternating across memory resources for the set-up of the physical address space can be continued from interleaving across controllers and channels to interleaving across the ranks to be found in a channel.

Rank interleaving is controlled via address bits. For this reason only interleaving in powers of two comes into question, i.e. there is only a 2-way, 4-way or 8-way rank interleave. An odd number of ranks in the memory channel results in the 1-way interleave, which is only referred to as interleave for the sake of the systematics involved: in the case of a 1-way a rank is utilized to the full before changing to the next one.

The granularity of the rank interleaving is larger than with previously described interleaving across controllers and channels. The latter was geared to the 64-byte cache line size. Rank interleaving is oriented towards the 4 KB page size of the operating systems and is connected to the physics of DRAM memory. Memory cells are - to put it roughly - arranged in two dimensions. A row (so-called page) is opened and then a column item is read. While the page is open, further column values can be read with a much lower latency. The rougher rank interleaving is attuned to this feature.

The number of ranks per memory channel follows from the DIMM type and the DPC value of the configuration.

The table shows the impact of rank interleaving as an example for DIMM configurations from the group of ideal memory capacities, i.e. with numbers of DIMMs per processor that are multiples of eight. All the listed configurations permit memory frequency of 1333 MHz, i.e. the influence on performance of different frequencies is hidden. All the measurements shown were taken under 1333 MHz. The influence of the memory channel operating modes is also hidden: the measurement values are determined under Independent.

	32GB 4R LRDIMM 2DPC	16GB 2R RDIMM 2DPC	32GB 4R LRDIMM 1DPC	16GB 2R RDIMM 1DPC	8GB 1R RDIMM 2DPC	32GB 4R LRDIMM 3DPC ¹	8GB 1R RDIMM 1DPC
Rank Interleaving	8-way	4-way		2-way			1-way
Memory bandwidth (STREAM)	95%	100%	99%	99%	96%	85%	80%
Commercial application performance (SPECint_rate_base2006)	99%	100%	99%	99%	99%	94%	98%

¹ Rank multiplication as an additional influencing factor

The influence of rank interleaving on memory performance is much more subtle than with interleaving across controllers and channels. This is a benchmark topic that can remain largely ignored for productive operations. Since it is a matter of nuances, the differentiation as regards channel operation mode, and also as regards non-ideal memory capacities can be omitted: within the context of measuring accuracy the ratios are approximately the same.

The table requires a reason why the 8-way interleaving with LRDIMMs does not supply the best performance. It is to be found in the trade-off between the advantage of rank interleaving and the disadvantage of a higher refresh overhead for channel configurations with a high DPC value and a high number of ranks.

Memory frequency

The influence on performance of interleaving across memory controllers and channels is considerable and the influence of rank interleaving rather low. And the influence of memory frequency lies in between.

The table in turn shows relative performance related to maximum performance, denoted with 100%, and separated for the two benchmarks STREAM and SPECint_rate_base2006. This series of measurements with the aim of determining the influence of frequency was carried out with 8, 16 or 24 RDIMMs and LRDIMMs per processor. Different DPC values and DIMM types are required to achieve the frequencies in question. For transparency reasons mean values have been formed to hide the minor fluctuations that occur and which also include the effect of rank interleaving. No differentiation is necessary as regards PRIMEQUEST 2000 series and PRIMERGY RX4770 M1.

	Operating mode	1600 MHz	1333 MHz	1066 MHz
Memory bandwidth (STREAM)	Independent		100%	84%
	Lockstep	66%	57%	47%
Commercial application performance (SPECint_rate_base2006)	Independent		100%	97%
	Lockstep	97%	94%	89%

To better classify the statement provided by the table it is again followed by the cases, in which the question as to different frequency is posed.

There are two reasons for the suboptimal frequency 1066 MHz for RDIMMs for which the value of 1600 MHz cannot occur. They are relevant for both channel operating modes Independent and Lockstep.

- 2DPC configurations have the trade-off between performance and energy consumption. 1066 MHz are for low-voltage operation, otherwise the DIMMs run at 1333 MHz.
- The value of 1066 MHz always occurs in 3DPC configurations with RDIMMS. In comparison with the 1DPC or 2DPC configurations that enable 1333 MHz, the drop in performance then occurs if the difference in storage capacity is not included in the calculation. Large capacities (3DPC), however, usually allow performance advantages through reduced I/O rates, which also have to be taken into account for a fair comparison.

With LRDIMMs the same trade-offs concerning energy efficiency and storage capacity apply. However, on account of the different load placed on the memory channels by these buffered DIMMs with high rank numbers the specifics are different.

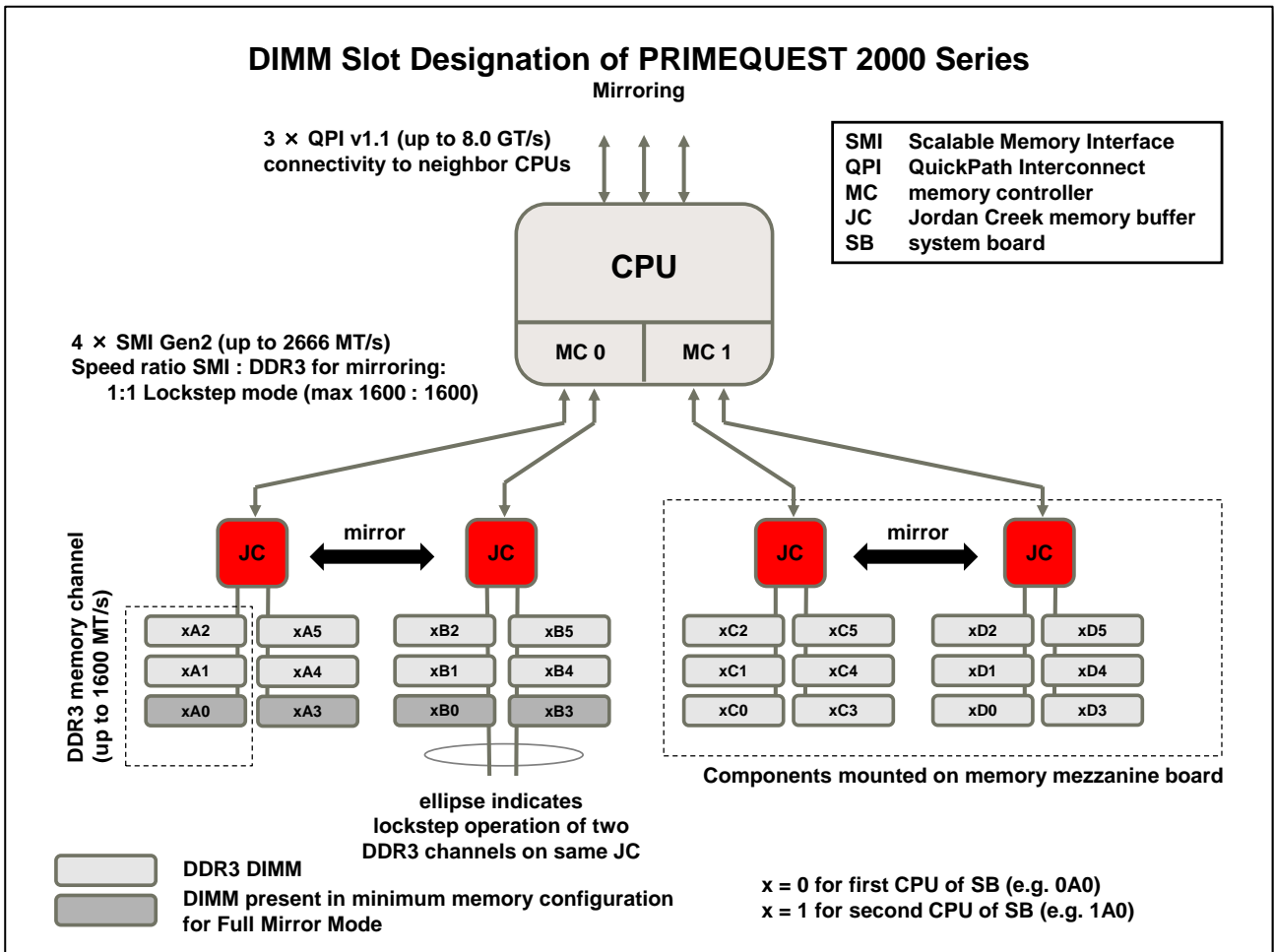
- In the case of Lockstep operation only, 1DPC and 2DPC have the trade-off between low-voltage (1.35 V, 1333 MHz) and 1.5 V (1600 MHz) operation. The table shows that although the performance disadvantage of Lockstep can in comparison with independent memory channels be corrected by increasing the frequency to 1600 MHz, it cannot be eliminated.
- If you use the 64 GB LRDIMMs for the largest possible memory capacities, the value of 1066 MHz always occurs. On the other hand, the 32 GB LRDIMMs allow 1333 MHz or even 1600 MHz. The performance disadvantage due to 1066 MHz should again be weighed up against the advantages of a larger capacity.

Memory performance under redundancy

To close here are some statements about memory performance under redundancy, i.e. for the memory mirroring and sparing.

Full Mirror Mode of PRIMEQUEST 2000 series

Mirroring takes place within the memory controllers with their two Jordan Creek 1 buffers and two DDR3 channels per buffer. The second Jordan Creek 1 with its appropriate memory mirrors the first one. For this purpose, both Jordan Creek 1s must be equally configured. There is no mirroring between the two memory controllers of a processor or even beyond the processor boundaries. Below is the already shown block diagram with an appropriate supplement and change.



The change concerns the minimal configuration. In the Memory Operation Modes Normal (Lockstep) and Performance the minimal configuration consists of two DIMMs in positions xA0 and xA3. As shown, it consists of four DIMMs in Full Mirror Mode. Furthermore, this deviating minimal configuration does not correspond to the four DIMM configuration under Normal (Lockstep) and Performance Mode: there the minimal configuration xA0 and xA3 is extended by xC0 and xC3 for performance reasons, because the second memory controller is also included in this way. This only takes place in Full Mirror Mode in the first increment after the minimal configuration, which then comprises eight DIMMs in the positions xA0, xA3, xB0, xB3, xC0, xC3, xD0 and xD3.

The following table shows the performance of Full Mirror Mode in comparison to the already examined Normal (Lockstep) and Performance Modes. The measurements were made consistently under memory frequency 1333 MHz. The values are related to the "ideal" performance, which is achieved with Memory Operation Mode Performance and maximum interleaving across memory controllers and channels for eight DIMMs (or multiples thereof).

	Memory Operation Mode	8 DIMMs per CPU (and multiples)	4 DIMMs per CPU ¹
Memory bandwidth (STREAM)	Performance Mode	100%	56%
	Normal Mode (Lockstep)	57%	30%
	Full Mirror Mode	37%	21%
Commercial application performance (SPECint_rate_base2006)	Performance Mode	100%	93%
	Normal Mode (Lockstep)	94%	76%
	Full Mirror Mode	85%	68%

¹ Different DIMM positions xA0, xA3, xC0, xC3 in Normal (Lockstep) and Performance Mode and xA0, xA3, xB0, xB3 in Full Mirror Mode.

In order to understand the table it is essential for Full Mirror Mode to include Lockstep Mode. The RAS function Mirroring is added to the RAS function Lockstep. The impact on performance of mirroring, while ignoring all other aspects of memory performance, can therefore only be seen in the comparison between Full Mirror Mode and Normal Mode.

Mirror Mode of PRIMERGY RX4770 M1

Different DIMM configuration rules to those for the PRIMEQUEST 2000 series apply in the case of the PRIMERGY RX4770 M1; hence the reference once again to the configurator. One difference was already mentioned in the section on interleaving across the memory channels: There are configurations that are not Lockstep-capable. A further difference here is that Mirroring can not only be added to the Lockstep operation mode, as with the PRIMEQUEST 2000 series, but also to the Independent operation mode. This is taken into account in the following table.

	Operating mode	Per CPU: 8 DIMMs across 2 Mem Boards Ideal capacities	Per CPU: 4 DIMMs across 2 Mem Boards	Per CPU: 4 DIMMs across 1 Mem Board	Per CPU: 2 DIMMs across 1 Mem Board Minimum configuration
Memory bandwidth (STREAM)	Independent	100%	56%	51%	28%
	Independent + Mirror	69%	35%	35%	17%
	Lockstep	57%		29%	
	Lockstep + Mirror	37%		19%	
Commercial application performance (SPECint_rate_base2006)	Independent	100%	94%	92%	78%
	Independent + Mirror	97%	87%	85%	67%
	Lockstep	94%		79%	
	Lockstep + Mirror	85%		68%	

Spare Mode

Spare Mode was deferred for the first sales release of the Ivy Bridge-EX based servers. Statements about performance are to follow in an update to the document.


Literature


PRIMERGY & PRIMEQUEST Servers


[L1] <http://www.fujitsu.com/fts/products/computing/servers/>

Memory Performance

[L2] This White Paper:

 <http://docs.ts.fujitsu.com/dl.aspx?id=8ff6579c-966c-4bce-8be0-fc7a541b4a02>

 <http://docs.ts.fujitsu.com/dl.aspx?id=9a7ec9d5-8140-4230-972b-2a04d76e43d6>

 <http://docs.ts.fujitsu.com/dl.aspx?id=a9489f25-465a-48d6-80c0-e726809616ea>

[L3] Memory Performance of Xeon E5-2600 v2 based Systems

<http://docs.ts.fujitsu.com/dl.aspx?id=a344b05e-2e9d-481b-8c9b-c6542defd839>

Benchmarks

[L4] STREAM

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

[L5] SPECcpu2006

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

Performance reports

[L6] Performance Report PRIMEQUEST 2800E

<http://docs.ts.fujitsu.com/dl.aspx?id=a0e6c1c7-7b8f-4d13-bb36-373db1d660b3>

[L7] Performance Report PRIMERGY RX4770 M1

<http://docs.ts.fujitsu.com/dl.aspx?id=6bcb41d7-c045-42e2-a965-a01ae9204d1d>

Contact

FUJITSU

Website: <http://www.fujitsu.com/>

PRIMERGY & PRIMEQUEST Product Marketing

<mailto:Primergy-PM@ts.fujitsu.com>

PRIMERGY Performance and Benchmarks

<mailto:primergy.benchmark@ts.fujitsu.com>