

# White Paper

## FUJITSU Server PRIMERGY & PRIMEQUEST

### Memory Performance of Xeon E7 v4 (Broadwell-EX) based Systems

The Xeon E7 v4 (Broadwell-EX) based models of the PRIMEQUEST 2000 Type 3 series and the PRIMERGY RX4770 M3 also acquire their impressive increase in performance over previous generations from the capacity of the QuickPath Interconnect (QPI) memory architecture, which has proved itself now for four generations of systems. This white paper explains the changed parameters of the memory architecture and quantifies their effect on the performance of commercial applications.

**Version**

1.0a

2016-07-29



## Contents

Document History .....	2
Introduction .....	3
Memory architecture .....	5
DIMM slots .....	5
DDR4 topics and available DIMM types .....	9
Firmware and BIOS parameters .....	11
Interfaces of the MMB Web-GUI of the PRIMEQUEST 2000 Type 3 series .....	11
Interfaces of the Device Manager of the PRIMEQUEST 2000 Type 3 series.....	12
Interfaces of the BIOS of the PRIMERGY RX4770 M3.....	12
Definition of the memory frequency .....	14
Lockstep operation mode of the memory channels .....	15
Independent operation mode of the memory channels .....	15
The Energy Optimized setting of the PRIMERGY RX4770 M3.....	15
Ideal memory capacities .....	16
Quantitative effects on memory performance .....	18
The measuring tools.....	19
STREAM Benchmark .....	19
SPECint_rate_base2006 Benchmark.....	19
Interleaving across memory controllers and memory channels.....	20
Influence of the memory frequency.....	23
Interleaving across ranks and influence of the DIMM types .....	25
Memory performance under redundancy.....	27
Full Mirror Mode of the PRIMEQUEST 2000 Type 3 series.....	27
Full Mirror Mode of the PRIMERGY RX4770 M3.....	29
Spare Mode .....	30
Literature.....	32
Contact .....	32

## Document History

### **Version 1.0 (2016-06-30)**

Initial version

### **Version 1.0a (2016-07-29)**

Minor corrections

## Introduction

The Intel Xeon E7 v4 (Broadwell-EX) processors of the PRIMEQUEST 2000 Type 3 series and PRIMERGY RX4770 M3 are manufactured using the new 14 nm semiconductor technology, compared with the 22 nm manufacturing process of the predecessor generation Haswell-EX. The Broadwell-EX platform, including Intel C602 chipset, is retained.

The new generation provides an approximate 20-30% increase in performance in comparison to the predecessor generation with regard to most of the load scenarios. The major share of this improvement is due to a maximum of 24 cores per processor, instead of the previous 18 cores. This is outweighed in the memory system by maintaining the proven features of the predecessor generation. Only in the cache coherency protocol of the PRIMERGY RX4770 M3 is there an innovation.

The Broadwell-EX based servers also use DDR4 memory technology, which is operated with energy saving 1.2 V. Just as with the Haswell-EX based systems, support is provided for memory frequencies up to 1866 MHz and QPI transfer rates up to 9.6 GT/s. The most elementary indicator of memory performance, the bandwidth, remains with approx. 450 GB/s in the PRIMEQUEST 2800E3 and just under 270 GB/s in the PRIMERGY RX4770 M3 very close to the values of the respective predecessor models.

One innovation is the extension of the cache coherency protocol of the PRIMERGY RX4770 M3 to include a function that is already well-known from the last two generations of dual socket PRIMERGY servers: *Cluster-on-die* (COD). This protocol extension is optional for loads with excellent NUMA features.

Otherwise, the basic features of the QPI-based memory architecture of the predecessor generations are retained, including the specific characteristics for the high-end server class:

- There are 24 DIMM slots per processor, twice as many as in the current dual socket PRIMERGY servers. They are distributed across eight DDR4 memory channels per processor. Every processor has two integrated memory controllers for four channels each. Jordan Creek memory buffers, which are not in the dual socket servers, are to be found between the controllers and channels.
- The processors and their memory controllers are able to provide neighboring processors with memory content via the QPI links, and themselves request such content. All the memory modules in the system form a coherent address area. However, with this distinction between local and remote memory access the architecture is of the NUMA (Non-Uniform Memory Access) type.
- The systems still have the directory-based QPI 1.1 cache coherence protocol, but with the aforesaid COD extension in the case of the PRIMERGY RX4770 M3. COD comes into question for loads that are suited for speculation on largely local memory accesses. When enabled, there are two NUMA nodes for each processor according to the two memory controllers. In case of a system design that directly couples the processors without any proprietary connection chips (so-called *glueless design*), as with the PRIMEQUEST 2000, the total number of NUMA nodes is limited to eight. This explains why COD does not come into question for the PRIMEQUEST 2000 Type 3.
- There continues to be a trade-off between RAS (Reliability, Availability, Serviceability) and performance. The memory channels are either in Lockstep or Performance Mode. Lockstep is a synchronous operation mode of in each case two memory channels, which improves the RAS features. However, in Performance or Independent Mode the memory channels are independent of each other.

This document looks on the one hand at the innovations in the memory system. And on the other hand, as in the earlier issues, this document also provides basic knowledge about QPI memory architecture which is essential when configuring 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 DDR4 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. Depending on the operation mode of the memory channels, the DIMM type and number as well as the configured processor model, it is 1866, 1600, or 1333 MHz.

Influencing factors on the memory performance are named and quantified. 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

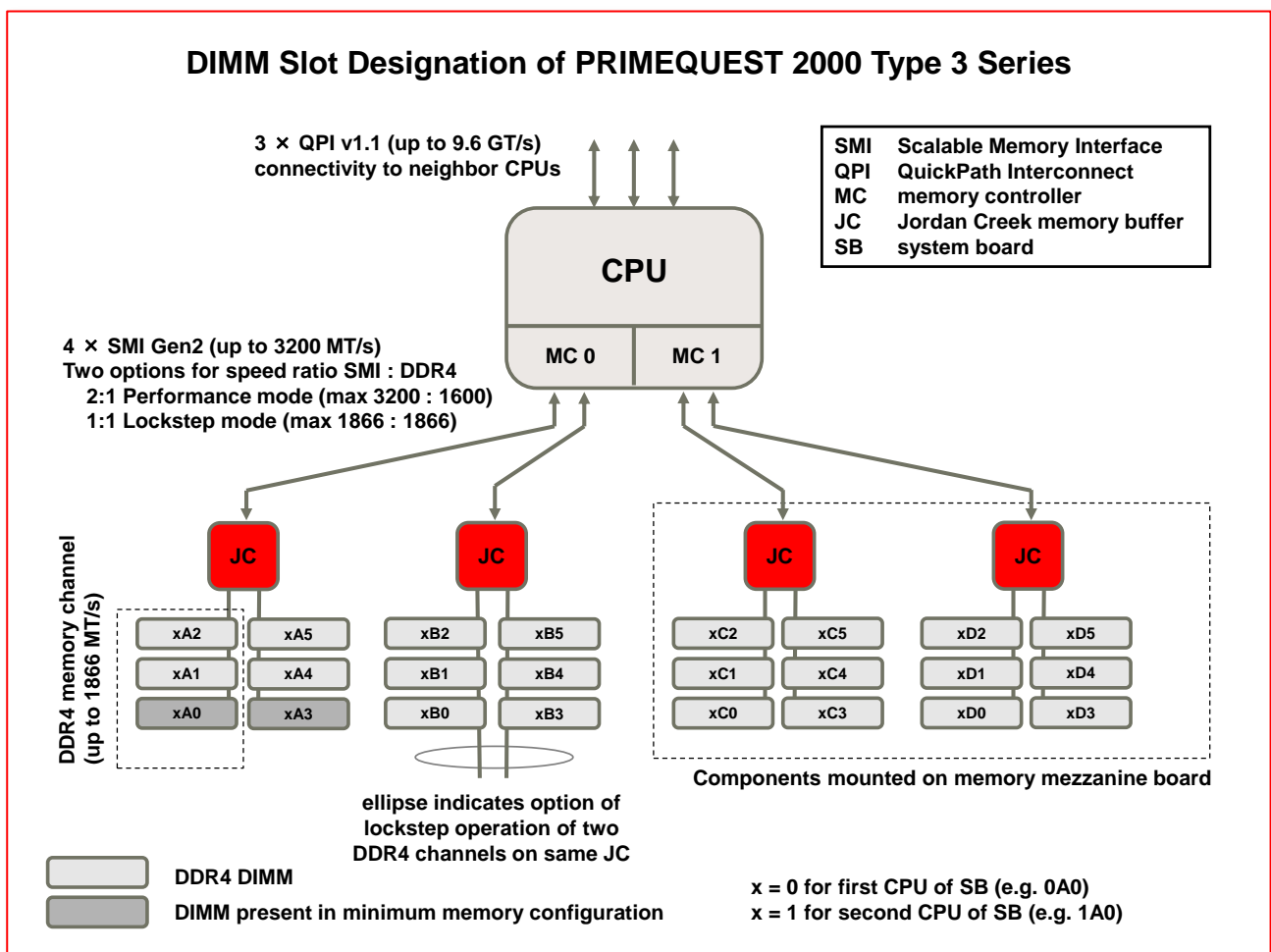
This section provides an overview of the memory system in five parts. Block diagrams explain the arrangement of the available DIMM slots. The available DIMM types are listed in the second section. This is followed by a section about the firmware and BIOS parameters that affect the memory system. The 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

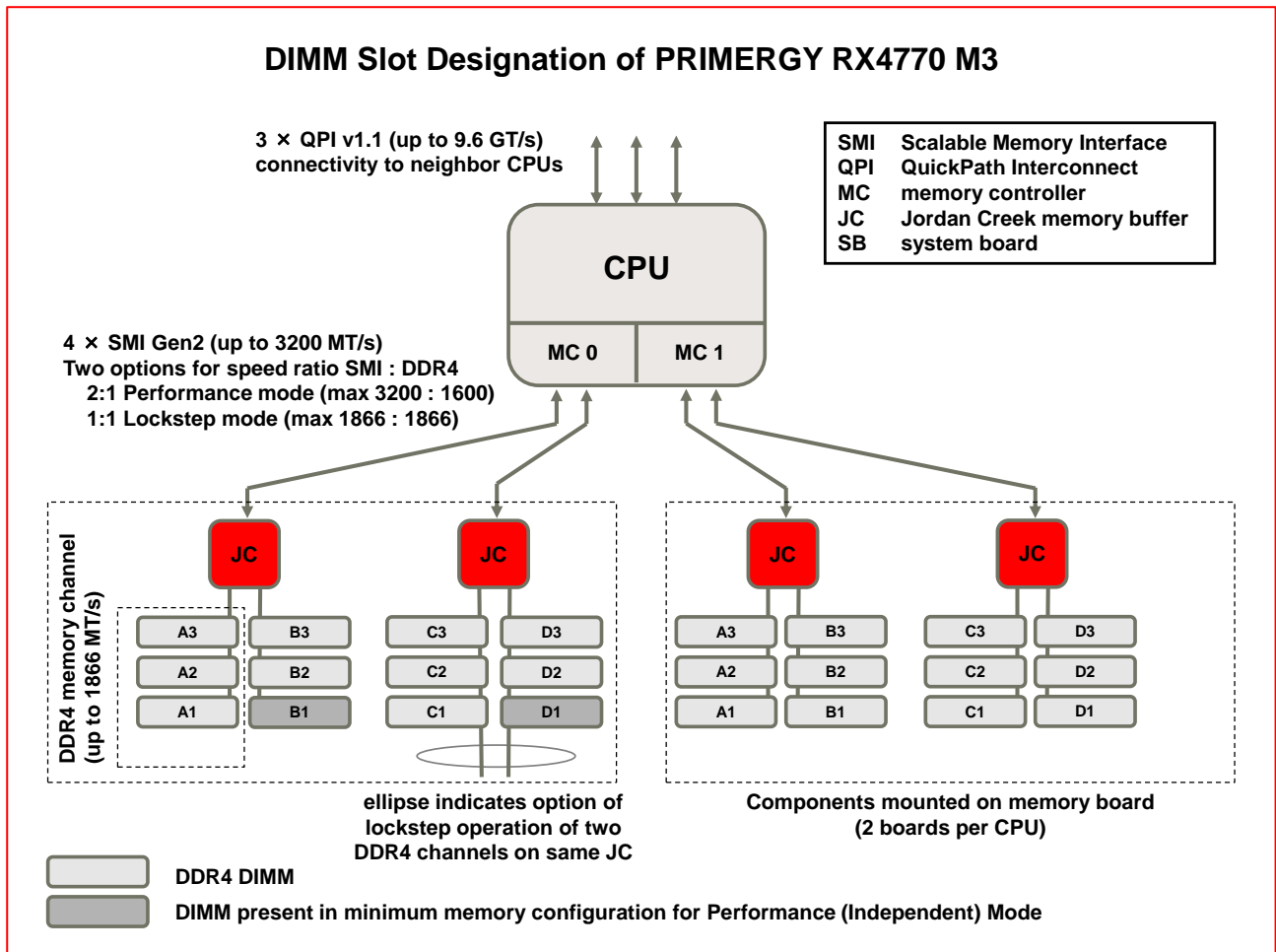
The two diagrams below show the memory connection from the perspective of the individual Broadwell-EX processor. Each processor has two integrated memory controllers. Each controller is connected to two Jordan Creek 2 memory buffers via bidirectional, serial SMI Gen2 (Scalable Memory Interface) links. There are two DDR4 memory channels, each with three DIMM slots, behind each memory buffer. Thus, there is a total of 24 DIMM slots per processor.

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 Type 3 series, for example the PRIMEQUEST 2800E3, are based on system boards with in each case two processors 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 M3. 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 DDR4 channels of a memory buffer, indicates the option of operating two channels in each case 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. Broadwell-EX already is the third generation where this operation mode is optional. The system or partition is either in Lockstep or Performance / Independent Mode. On the contrary, the systems of the older generations Nehalem-EX and Westmere-EX were always in Lockstep Mode.

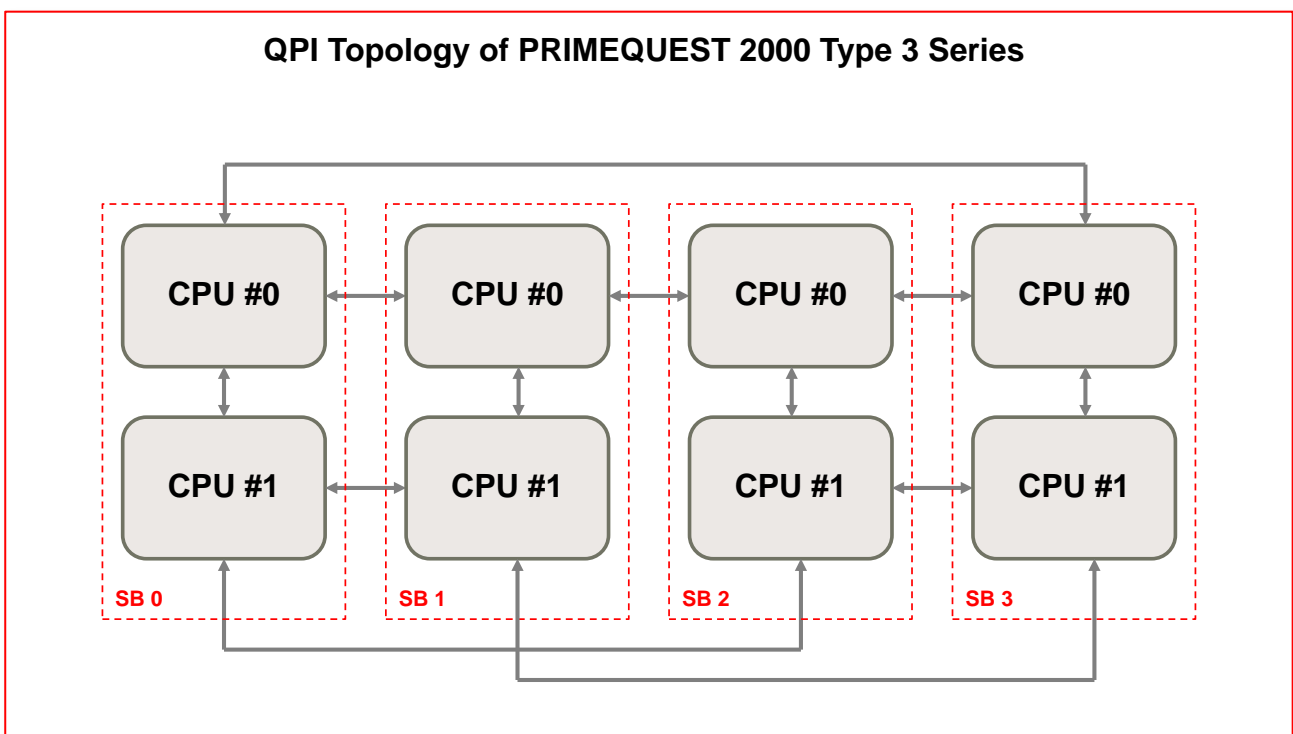
The eligibility of the operation mode influences the frequency of the resources SMI Gen2 link and DDR4 channel. Since eight channels are only opposed by four SMI Gen2 links, the links to implement maximum memory bandwidth in Performance Mode have twice the speed of the memory channels. The frequency 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 3200 MT/s and in Lockstep Mode by the Jordan Creek 2 upper limit of 1866 MHz for the DDR4 frequency. Hence, the anomaly

that the mode with less efficient performance (Lockstep) supports the higher DDR4 frequency. However, the higher memory bandwidth is more valuable than the DDR4 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 Type 3 series and the PRIMERGY RX4770 M3.

In the PRIMEQUEST 2000 Type 3 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 2 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.

Lockstep capability in each memory configuration does not exist in the PRIMERGY RX4770 M3. 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 M3 (not marked) consists of four DIMMs in positions A1, B1, C1 and D1 of the first memory board.

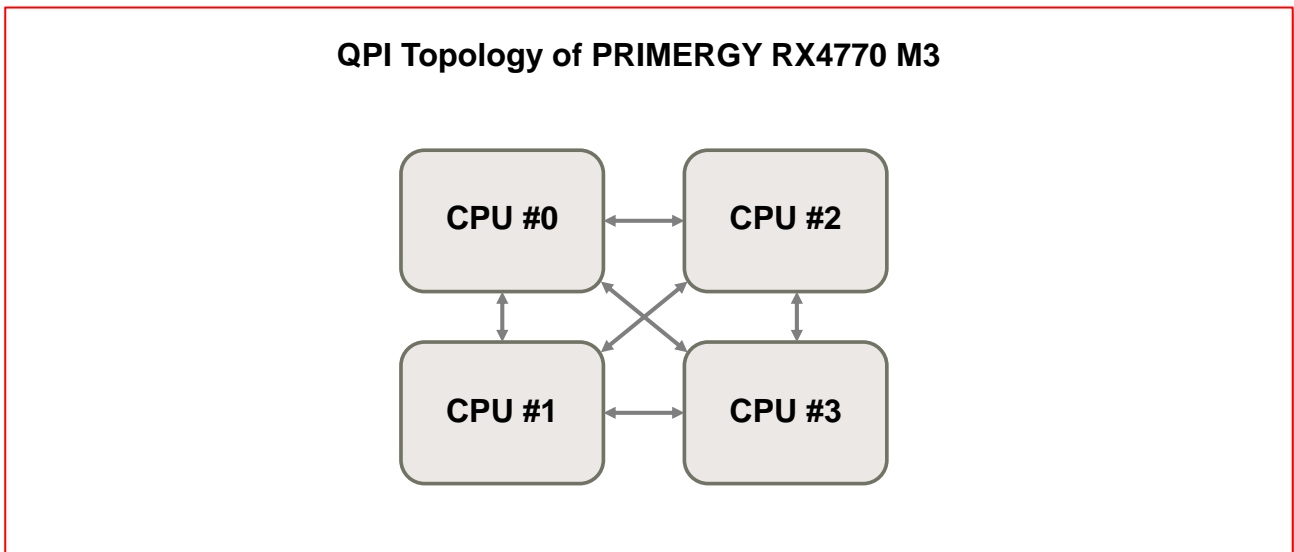


This diagram shows the QPI topology of the PRIMEQUEST 2000 Type 3 series, i.e. the networking of the processors and their appropriate memory components. Since the networking is only via the three QPI links per processor, the already discussed components SMI Gen2 links, memory buffers and DIMM slots have now been omitted. Also omitted from all the 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 Type 3 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 2400E3 model with a maximum of four processors there are only system boards 0 and 1. In this case and in the case of PRIMEQUEST 2800E3 partitions with less than four system boards there are unused QPI interfaces.

The PRIMERGY RX4770 M3 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 Type 3 series, and in particular from the topology of the PRIMEQUEST 2400E3 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.



## DDR4 topics and available DIMM types

The Broadwell-EX based systems use DDR4 SDRAM memory modules. The transition from DDR3 to DDR4 took place with the predecessor generation Haswell-EX. The JEDEC (Joint Electron Device Engineering Council) standards with the designations DDR3 and DDR4 define the interfaces that are binding for memory and system manufacturers.

Since DDR4 technology is still comparatively new, here are the key differentiators in comparison to DDR3. The transition from DDR3 to DDR4 was of an evolutionary nature and did not come with a once-only performance boost.

- More pins per DIMM are required for DDR4; therefore, DDR3 and DDR4 DIMM sockets are not compatible. Older DDR3 memory modules cannot be used in DDR4-based systems.
- DDR4 supports memory frequencies of up to 3200 MHz. This frequency range will be used up over several server generations in the coming years. Frequencies up to a maximum of 1866 MHz are supported in Haswell-EX and Broadwell-EX based systems when this technology is now used. It continues the increase in memory frequency in steps of 266 MHz as known with DDR3-based server generations. The transition to DDR4 is evolutionary. It does not come with a once-only performance boost.
- An important DDR4 advantage is the operation of DIMM strips with only 1.2 V instead of the 1.5 V or 1.35 V (low-voltage extension) with DDR3. This represents energy savings of some 30% with the same data transfer rate.
- As in the first phase of DDR3 technology, there is currently no low-voltage extension for DDR4. Consequently, the configuration trade-offs in the BIOS between performance and energy consumption currently do not apply for the most part.

At the time of the general release DIMM stripes according to the following table are eligible for the memory configuration of the Xeon E7 v4 based systems. Later extensions to this table are possible. There are registered (RDIMM) and load-reduced (LRDIMM) DIMMs. Mixed configurations consisting of these two DIMM types are not permitted.

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 DDR4-2400 R ECC	RDIMM	16	1	4	2400			✓	✓	1.2
32GB (2x16GB) 1Rx4 DDR4-2400 R ECC	RDIMM	32	1	4	2400			✓	✓	1.0
32GB (2x16GB) 2Rx4 DDR4-2400 R ECC	RDIMM	32	2	4	2400			✓	✓	<b>1.0</b>
64GB (2x32GB) 2Rx4 DDR4-2400 R ECC	RDIMM	64	2	4	2400			✓	✓	1.1
128GB (2x64GB) 4Rx4 DDR4-2133 LR ECC	LRDIMM	128	4	4	2133		✓		✓	1.3

The table takes into account the fact that DIMMs are offered in units of two respectively in the order and configuration process of the PRIMEQUEST 2000 Type 3 series and the PRIMERGY RX4770 M3. The reason for this is the configuration rule 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 all DDR generations. 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 Broadwell-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 or 4 ranks. The number of available ranks per memory channel has a certain influence on performance, which is explained below. Maximum capacity is the motivation for DIMMs with 4 ranks, but at the same time the DDR4 specification only supports a maximum of 8 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 effective memory frequency of a given server configuration depends on a series of influences that are looked at in the next but one section. The maximum frequency stated in the DIMM type table is a feature of these parts that are also used in other server classes and is merely to be understood as the upper limit of an effective memory frequency. The values mentioned there of 2133 MHz and 2400 MHz remain theoretical for the Broadwell-EX based servers. The effective frequency in these servers can at most be 1866 MHz.

The last column in the DIMM type table shows the relative price differences. The list prices from June 2016 for the PRIMERGY RX4770 M3 are used as a basis. The column shows the relative price per GB, standardized to the 2Rx4 RDIMM, size 16 GB (highlighted as measurement 1.0). Increased costs can be seen for the 8 GB RDIMM and for the 64 GB LRDIMM, with which very large memory capacities are achieved. 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 and 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 Broadwell-EX processors and are thus principally the same for the PRIMEQUEST 2000 Type 3 series and the PRIMERGY RX4770 M3. However, there are differences in naming, default assignment and positioning in the firmware and BIOS menus, which result from the respective 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:

- 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 Memory Mirroring or Sparing. Here, the PRIMEQUEST 2000 Type 3 series and the PRIMERGY RX4770 M3 differ in the fact that Mirroring and Sparing are only possible in Lockstep Mode with the PRIMEQUEST 2000 Type 3 series, whereas these functions are also supported in Performance Mode with the PRIMERGY RX4770 M3. A Sparing innovation is available since the Haswell-EX generation, namely Multiple Rank Sparing. There was only ever a single reserve rank per memory channel in previous generations.
- Energy savings in the memory system. Low-voltage operation of the memory modules, which was known to us in previous system generations, is excluded in the Broadwell-EX based servers, because there is currently no low-voltage extension for DDR4. That leaves two rather marginal aspects that have something to do with saving energy. In the case of the PRIMERGY RX4770 M3 the memory frequency can be generally set to the minimum value of 1333 MHz – with a certain energy-saving effect. And with the PRIMEQUEST 2000 Type 3 series the memory power states can be mitigated in favor of shorter wake-up times, i.e. the energy-saving function Memory Power States can be reversed in favor of performance. Energy-saving power states of the memory modules – analog to the C states of the processors – are activated in phases without any memory accesses.
- 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. Sensitive performance measurements may be a reason for disabling this functionality.
- The Broadwell-EX processor generation has a *Retry* option in the memory controllers in the case of parity errors on the signal lines. A minor influence on performance on the part of this functionality is conceivable on a load-dependent basis, which is why it is optional in the PRIMEQUEST 2000 Type 3 series.

This preliminary comment is now followed by the specific syntactic design for the PRIMEQUEST 2000 Type 3 series and the PRIMERGY RX4770 M3. In the case of the PRIMEQUEST 2000 Type 3 series the parameters are to be found in two different administration interfaces.

### Interfaces of the MMB Web-GUI of the PRIMEQUEST 2000 Type 3 series

The parameter *Memory Operation Mode* with the following options is under Partition / Partition# / Mode (partitionable PRIMEQUEST 2000 Type 3 models) and System / Mode (PRIMEQUEST 2800B3) in the Web-GUI of the management board (MMB):

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

The default 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 four redundant modes Partial Mirror, Full Mirror, Spare and Address Range Mirror.

Sparing means Rank Sparing. Thus, the percentage share of the net capacity depends on the configured DIMM type and its number of ranks. The Broadwell-EX based systems support the option of keeping more than one rank as spares. To enable you to control this the same menu has the parameter *Memory Sparing Mode* with the options

- 1Rank
- 2Rank
- Auto

The first two parameters are self-explanatory, and for Auto no more than half of the ranks that exist in the memory channel are reserved. The default 1Rank with a single spare corresponds to the Spare Mode of the previous system generations up to and including Ivy Bridge-EX.

When calculating the net capacity of Spare Mode you should also take the different DIMM configuration rule into consideration. It must always be configured with 3DPC, starting with a minimum configuration of six DIMMs per processor in two out of eight memory channels. The following section about memory performance under redundancy returns to this aspect.

The four 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 the PRIMEQUEST 2000 Type 3 series in connection with the independent memory channels of Performance Mode.

### **Interfaces of the Device Manager of the PRIMEQUEST 2000 Type 3 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. Here there are four parameters with the following options; once again the defaults that were valid at the time of the general release are underlined:

- Patrol Scrub: Disabled / Enabled
- Refresh Rate: Auto / 1x
- Memory Power States: Default / Performance Mode
- DDR4 Command / Address Parity Check and Retry: Disabled / Enabled

For performance reasons the default of the *Patrol Scrub* parameter is disabled in the PRIMEQUEST 2000 Type 3 series. The impact on performance, however, is usually very small.

The impact on performance of the Memory Power States is also minor. In the case of low-latency application scenarios the Performance Mode setting can result in measureable improvements. No improvement could be verified with the STREAM and SPECint\_rate\_base2006 benchmarks, which were used for this document to characterize memory performance.

The second Refresh Rate parameter is a relict of DDR3 technology and obsolete in DDR4 based systems, such as the PRIMEQUEST 2000 Type 3 series. It may possibly be omitted in future BIOS versions.

### **Interfaces of the BIOS of the PRIMERGY RX4770 M3**

In the case of the PRIMERGY RX4770 M3 there is a Memory Configuration submenu in the BIOS under Advanced with the following parameters:

- Memory Mode: Normal / Mirroring / Sparing
- VMSE Lockstep Mode: Lockstep / Independent
- DDR Performance: Performance optimized / Energy optimized
- Patrol Scrub: Disabled / Enabled

The defaults valid at the time of general release are also underlined here.

The first 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. At the time of its general release the PRIMERGY RX4770 M3 only supports the Rank Sparing known from the predecessor system with a single reserve rank.

The second parameter *VMSE Lockstep Mode* concerns the alternative between independent memory channels (Independent) and Lockstep operation mode, which in the PRIMERGY RX4770 M3 is independent of the optional activation of the RAS modes Mirroring und Sparing in contrast to the PRIMEQUEST 2000 Type 3 series.

In the case of the third parameter *DDR Performance* the Energy optimized setting results in a general reduction of the memory frequency to 1333 MHz. However, the potential to save energy is low. The energy

consumption of the memory is primarily based on the DIMM voltage, which is always 1.2 V for Broadwell-EX based servers.

The fourth parameter *Patrol Scrub* was dealt with above.

The parameters for the COD (Cluster-on-die) option addressed in the introduction as an extension to the cache coherency protocol are available in the submenu CPU Configuration:

- COD Enable: Disabled / Enabled / Auto
- Home Dir Snoop with IVT- Style OSB Enable: Disabled / Enabled / Auto

In the case of default assignment with Auto, COD is not enabled. To enable it simply set the parameter *COD Enable* to *Enabled* and leave the parameter *Home Dir Snoop with IVT- Style OSB Enable* at *Auto*.

COD comes into question for loads that are suited for speculation on largely local memory accesses. When enabled, there are two NUMA nodes for each processor according to the two memory controllers. Half of the processor cores and half of the L3 cache are in each case also allocated to the nodes. The advantage in performance for suitable loads is about 1-2%. This was verified with the standard benchmark SPECint\_rate\_base2006.

In case of a system design that directly couples the processors without any proprietary connection chips (so-called *glueless design*), as with the PRIMEQUEST 2000, the total number of NUMA nodes is limited to eight. This explains why COD does not come into question for the PRIMEQUEST 2000 Type 3 series.

## 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 1866, 1600 and 1333 MHz come into question for the Broadwell-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 general conditions concern the configured processor model, the operation mode of the memory channels (Lockstep or Independent / Performance), as well as the configured DIMM type in the case of 3DPC configurations. And in the case of the PRIMERGY RX4770 M3 add the option of a general reduction in memory frequency to the minimum value of 1333 MHz under the keyword *Energy optimized*.

Initially, the configured processor model is of significance for the definition of memory frequency. Within the context of this document the classification of the Broadwell-EX series according to the following table, which shows the maximum possible memory frequencies for each operation mode of the memory channels, is recommended. The table shows the complete list of Xeon E7 v4 models. As regards availability in the server models of the PRIMEQUEST 2000 Type 3 series and the PRIMERGY RX4770 M3, please refer to the configurators of the systems.

CPU type	QPI	Independent (2:1)		Lockstep (1:1)		Xeon E7 v4 models
		SMI	DDR4	SMI	DDR4	
Advanced	9.6	3200	1600	1866	1866	E7-8890 v4, E7-8880 v4, E7-8870 v4, E7-8860 v4, E7-8891 v4, E7-8893 v4, E7-8867 v4
Standard	8.0	2666	1333	1866	1866	E7-8855 v4, E7-4850 v4, E7-4830 v4
Basic	6.4	2666	1333	1866	1866	E7-4820 v4, E7-4809 v4

Memory frequency means the DDR4 frequency. However, according to the architecture of the memory connection this is linked to the frequency of the SMI Gen2 links between the on-chip memory controllers and off-chip memory buffers. In the case of the Lockstep operation mode the ratio of the frequencies is 1:1. And 2:1 for Independent Mode so as to establish a balance between the bandwidths of four SMI Gen2 links and eight DDR4 channels for independent memory channels. This is not necessary in Lockstep Mode, because the eight DDR4 channels are consolidated to form four logical channel pairs.

Although the main focus is usually on memory frequency, the logic on which this is based only becomes visible when the SMI topic is involved. Thus, the corresponding SMI frequency has also been listed in the table. The upper limit of 3200 MT/s, which entails a maximum memory frequency of 1600 MHz in Independent Mode, applies for the Jordan Creek 2 memory buffer. The higher memory frequency in Lockstep Mode is on the other hand linked to an SMI frequency that is lower compared to that in Independent Mode – and frequency is equivalent to the bandwidth of this resource. Just one look at the SMI frequencies resolves the apparent anomaly that the mode with a lower performance level (Lockstep) supports the higher memory frequency.

Full configuration of the memory channels with three DIMMs per channel (3DPC configurations) can as a further general condition result in a reduction in the memory frequency. In this case, it depends on the DIMM type, because DIMM types differ as regards the electrostatic load of the memory channels on account of their design.

That having been said, the effective memory frequency of a given configuration is as follows.

**Lockstep operation mode of the memory channels**

CPU type	8GB and 16GB 1Rx4 RDIMM			16GB and 32GB 2Rx4 RDIMM			64 GB 4Rx4 LRDIMM		
	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC
Advanced Standard Basic	1866	1866	1600	1866	1866	1333	1866	1866	1600

**Independent operation mode of the memory channels**

CPU type	8GB and 16GB 1Rx4 RDIMM			16GB and 32GB 2Rx4 RDIMM			64 GB 4Rx4 LRDIMM		
	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC	1DPC	2DPC	3DPC
Advanced	1600	1600	1600	1600	1600	1333	1600	1600	1600
Standard Basic	1333	1333	1333	1333	1333	1333	1333	1333	1333

Independent Mode is also referred to as Performance Mode and is usually preferred for performance measurements and benchmarks. The justification for this was stated above. The independent memory channels provide a bandwidth advantage, which the Lockstep Mode can reduce due to the higher memory frequency, but not attain. The respective SMI frequency can be used as an indicator for bandwidth ratios.

**The Energy Optimized setting of the PRIMERGY RX4770 M3**

As mentioned in the previous section on BIOS parameters, the possible *DDR Performance = Energy optimized* setting in the PRIMERGY RX4770 M3 results in a general reduction in memory frequency to the smallest possible value of 1333 MHz. The above tables then become obsolete. The parameter does not exist in the PRIMEQUEST 2000 Type 3 series.

It should be pointed out once again that the potential savings in energy that can be achieved through *Energy optimized* are rather low. The energy consumption of the memory modules is primarily based on the voltage, which is always 1.2 V in Broadwell-EX based servers.

## Ideal memory capacities

In summary, two main influences on the memory performance of Broadwell-EX based servers have been named so far. Firstly, the trade-off between RAS (Lockstep) and Performance that is controlled via the operation mode of the memory channels. 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 is 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. In the case of the PRIMERGY RX4770 M3 it should be noted that two memory boards are required per processor.

GB for 2 CPU	GB for 4 CPU	GB for 8 CPU	DPC	DIMM Type (8 DIMMs per CPU and DPC)	Independent	Lockstep	Benchmark
					Max MHz	Max MHz	
128	256	512	1	8GB 1Rx4 RDIMM	1600	1866	
256	512	1024	2	8GB 1Rx4 RDIMM	1600	1866	
			1	16GB 1Rx4 RDIMM	1600	1866	
			1	16GB 2Rx4 RDIMM	1600	1866	
384	768	1536	3	8GB 1Rx4 RDIMM	1600	1600	
512	1024	2048	2	16GB 1Rx4 RDIMM	1600	1866	+
			2	16GB 2Rx4 RDIMM	1600	1866	++
			1	32GB 2Rx4 RDIMM	1600	1866	
768	1536	3072	3	16GB 1Rx4 RDIMM	1600	1600	
			3	16GB 2Rx4 RDIMM	1333	1333	
1024	2048	4096	2	32GB 2Rx4 RDIMM	1600	1866	+
			1	64GB 4Rx4 LRDIMM	1600	1866	
1536	3072	6144	3	32GB 2Rx4 RDIMM	1333	1333	
2048	4096	8192	2	64GB 4Rx4 LRDIMM	1600	1866	
3072	6144	12288	3	64GB 4Rx4 LRDIMM	1600	1600	

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 2 memory buffer or DDR4 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 are specified for configurations 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 × DPC × DIMM size in GB × number of CPUs*

The table states the maximum possible memory frequencies for every configuration, whereby the already discussed case distinction as regards the operation mode of the memory channels should be noted. In the case of Independent Mode the configuration of a less powerful processor model can result in a lower frequency than the one shown in the table. Furthermore, the BIOS setting *Energy optimized* can lead to a lower frequency, i.e. 1333 MHz in the PRIMERGY RX4770 M3. The latter applies for both operation modes.

In any event, the memory configurations in the table have the characteristic of optimal channel interleaving, regardless of how the trade-off between RAS (Lockstep) and Performance was 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.

It goes without saying that memory configurations used in the standard benchmarks of the PRIMEQUEST 2000 Type 3 series and the PRIMERGY RX4770 M3 are also ranked among the optimal configurations of the table. They are marked with a + sign in the last column. The best possible memory performance is provided by the configuration marked with ++.

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 und 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 about interleaving across memory controllers and channels, to which the difference between the Lockstep and Independent operation modes of the channels belongs as a topic. Further sections deal with memory frequency, interleaving across ranks as well as additional influences that are specific to the various DIMM types. 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 as far as possible so as not to mix up the influences.

The following table describes the measuring configurations. In the case of the PRIMEQUEST 2000 Type 3 series the tests were carried out in partitions from one and four system boards respectively, each with two processors. As the results were not significantly dependent on partition size, differentiation in this respect was omitted in the following sections.

System Under Test (SUT)		
<b>Hardware</b>		
Model	PRIMEQUEST 2800E3	PRIMERGY RX4770 M3
Processor type	Xeon E7-8890 v4	Xeon E7-8890 v4
Memory types	16GB (2x8GB) 1Rx4 DDR4-2400 R ECC 32GB (2x16GB) 2Rx4 DDR4-2400 R ECC	32GB (2x16GB) 2Rx4 DDR4-2400 R ECC 128GB (2x64GB) 4Rx4 DDR4-2133 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
<b>Software</b>		
Firmware	Unified Firmware 16043 (BIOS, BMC, MMB)	BIOS R1.0.0, BMC 8.13F
Operating system	Red Hat Enterprise Linux Server release 6.7	Red Hat Enterprise Linux Server release 6.7

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 a 100% measurement of the tables, are - in a further differentiation as regards the various processor models - included in the Performance Reports of PRIMEQUEST 2800E3 [L6] and PRIMERGY RX4770 M3 [L7].

The memory performance tests are carried out using the most powerful processor model Xeon E7-8890 v4 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 across memory controllers and memory channels

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 Broadwell-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 below. Incidentally, the pattern requires a certain flexibility as regards block sizes for alternating.

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 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.

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 the PRIMEQUEST 2000 Type 3 series				
	Operation mode	8 DIMMs per CPU (and multiples) Ideal capacities	4 DIMMs per CPU	2 DIMMs per CPU Minimal configuration
Interleaving (Controller / Channel)	Independent	<b>2-way / 4-way</b>	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 1600 MHz	<b>100%</b>	58%	29%
	Lockstep 1866 MHz	70%	36%	18%
Commercial application performance (SPECint_rate_base2006)	Independent 1600 MHz	<b>100%</b>	93%	77%
	Lockstep 1866 MHz	96%	82%	62%

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 operation 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.

As regards the case distinction between Independent and Lockstep operation mode please note that the memory frequencies are usually different, as is also the case with the measurements on which the table is based. In other words, apart from the primary influence of channel interleaving, these comparisons also incorporate the secondary influence of different memory frequencies.

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 previously shown table concerns the PRIMEQUEST 2000 Type 3 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 2 memory buffer. General Lockstep capability does not apply for the permitted memory configurations of the PRIMERGY RX4770 M3. 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 M3 is prerequisite to understanding the table below.

Channel interleaving of the PRIMERGY RX4770 M3					
	Operations mode	Per CPU: 8 DIMMs distributed across 2 mem boards  Ideal capacities	Per CPU: 4 DIMMs distributed across 2 mem boards	Per CPU: 4 DIMMs distributed across 1 mem board	Per CPU: 2 DIMMs distributed across 1 mem board  Minimal configuration
Interleaving (Controller / Channel)	Independent	<b>2-way / 4-way</b>	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 1600 MHz	<b>100%</b>	65%	51%	33%
	Lockstep 1866 MHz	69%		35%	
Commercial application performance (SPECint_rate_base2006)	Independent 1600 MHz	<b>100%</b>	95%	92%	79%
	Lockstep 1866 MHz	96%		82%	

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 brief evaluation of the effects on the application performance (see the horizontal blocks for SPECint\_rate\_base2006 in both tables) is as follows. Benchmarks will always aim for configurations of 100% quality. Cases above 90% are not critical for productive operations, typically with security margins as regards system utilization. Cases of around 80% merit critical examination, for example in the case of a high

utilization level targeted under virtualization. In a case of just above 60% 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 can be a memory performance that corresponds to the 4 DIMM cases in the table. 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.

## Influence of the memory frequency

Second to the influence of controller and channel interleaving on memory performance is the influence of memory frequency.

The typical situation, which raises the question of this influence, is - as far as the Broadwell-EX based servers are concerned - the frequency reduction associated with 3DPC configurations. This connection has already been described in the section on the definition of the memory frequency. 3DPC configurations are required for large memory capacities, i.e. it is a matter of the trade-off between performance and capacity.

The following table shows the relevant cases, in which - depending on the operation mode of the memory channels and DIMM type - there is a frequency reduction with 3DPC.

- The reduction always occurs in Lockstep Mode.
- In Independent or Performance Mode this only applies for the 16 GB and 32 GB 2Rx4 RDIMM. It should also be noted that the frequency 1600 MHz in Independent Mode is only possible with the powerful *Advanced* processor models. The frequency topic is not relevant for models with lower performance levels.

In the measurements on which the table is based 8, 16 or 24 DIMMs per processor were configured, i.e. the measurements were taken with optimal channel interleaving.

		Independent Mode (Advanced CPUs)	Lockstep Mode		
		RDIMM 2Rx4	RDIMM 1Rx4	RDIMM 2Rx4	LRDIMM 4Rx4
2DPC	Memory bandwidth (STREAM)	100% (1600)	100% (1866)	100% (1866)	100% (1866)
	Commercial application performance (SPECint_rate_base2006)	100% (1600)	100% (1866)	100% (1866)	100% (1866)
3DPC	Memory bandwidth (STREAM)	87% (1333)	83% (1600)	74% (1333)	93% (1600)
	Commercial application performance (SPECint_rate_base2006)	96% (1333)	97% (1600)	93% (1333)	96% (1600)

The table's basis for comparison is the respective 2DPC configurations, which permit the maximum frequencies. The affected memory frequencies are denoted in MHz in brackets under the percentage performance values.

The frequency difference is the main reason for the loss in performance shown in each of the four compared configurations. However, further DIMM design-related effects, which are the subject of the following section, are superimposed on this effect. This explains the different effects for 1Rx4 RDIMMs and 4Rx4 LRDIMMs, although in both cases the memory frequency for 3DPC is reduced from 1866 MHz to 1600 MHz in Lockstep Mode.

The effect of the general reduction in frequency to 1333 MHz for the PRIMERGY RX4770 M3 (*DDR Performance = Energy optimized*) is approximately the same as the transition from 2DPC to 3DPC for 2Rx4 RDIMMs, i.e. the commercial application performance loses between 4 (Independent Mode) and 7% (Lockstep Mode) depending on the operation mode.

The fact that performance losses are larger for Lockstep Mode than for Independent Mode is due to the reduced memory bandwidth under Lockstep. Additional effects like frequency reduction impact under stricter conditions.

With the exception of the special case of the *DDR Performance = Energy optimized* option, it is not necessary to make a distinction between the PRIMEQUEST 2000 Type 3 series and the PRIMERGY RX4770 M3 for the topic of frequency.



## Interleaving across ranks and influence of the DIMM types

The following table compares – again separated according to operation mode – DIMM configuration cases with the same memory frequency. These are the maximum frequencies of 1600 MHz in Independent Mode and 1866 MHz in Lockstep Mode. These series of measurements were also carried out under optimal channel interleaving, i.e. with 8, 16 or 24 DIMMs per processor. Thus, configurations that are identical as regards the two main influences on memory performance, namely channel interleaving and memory frequency, are compared.

The relative performance statements are now related to the absolute best case, highlighted in bold as 100%. The 2DPC configuration with 2Rx4 RDIMMs provides the best memory performance for both operation modes. Which is why it is preferred in benchmarking if the memory capacity it offers per processor is sufficient.

However, reductions of 1-2% in the commercial application performance for productive operations can usually be ignored. The performance differences shown in this section are nuances that are predominantly taken into account in benchmarking situations.

		Independent Mode 1600 MHz			Lockstep Mode 1866 MHz		
		RDIMM 1Rx4	RDIMM 2Rx4	LRDIMM 4Rx4	RDIMM 1Rx4	RDIMM 2Rx4	LRDIMM 4Rx4
1DPC	Memory bandwidth (STREAM)	92%	100%	98%	87%	98%	97%
	Commercial application performance (SPECint_rate_base2006)	99%	100%	99%	98%	100%	99%
2DPC	Memory bandwidth (STREAM)	97%	<b>100%</b>	91%	96%	<b>100%</b>	89%
	Commercial application performance (SPECint_rate_base2006)	99%	<b>100%</b>	98%	99%	<b>100%</b>	97%
3DPC	Memory bandwidth (STREAM)	92%		92%			
	Commercial application performance (SPECint_rate_base2006)	99%		95%			

The main reason for the differences in performance is another form of interleaving. The method of alternating across memory resources for the set-up of the physical address space can be continued from the already discussed interleaving across controllers and memory 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 poorer performance - particularly with the bandwidth - can be explained by the odd number of ranks for 1DPC and 3DPC with 1Rx4 RDIMMs. However, this deterioration is perceptibly reduced with DDR4 due to the doubling of the maximum open pages per DRAM chip from 8 to 16. In the case of DDR3 the missing rank interleave resulted in a reduction of the bandwidth to 80%.

Further influences on memory performance are added to rank interleaving with the LRDIMMs. In comparison to RDIMMs there is firstly the certain overhead due to the DIMM component for data buffering. And with more than four ranks in the memory channel the overhead that is to be performed per rank for the DRAM refresh also becomes noticeable in a negative way. The refresh represents a certain basic load for the address lines of the memory channels, which are shared by all the ranks. And finally, the overhead of rank multiplication follows with more than eight physical ranks in the memory channel.

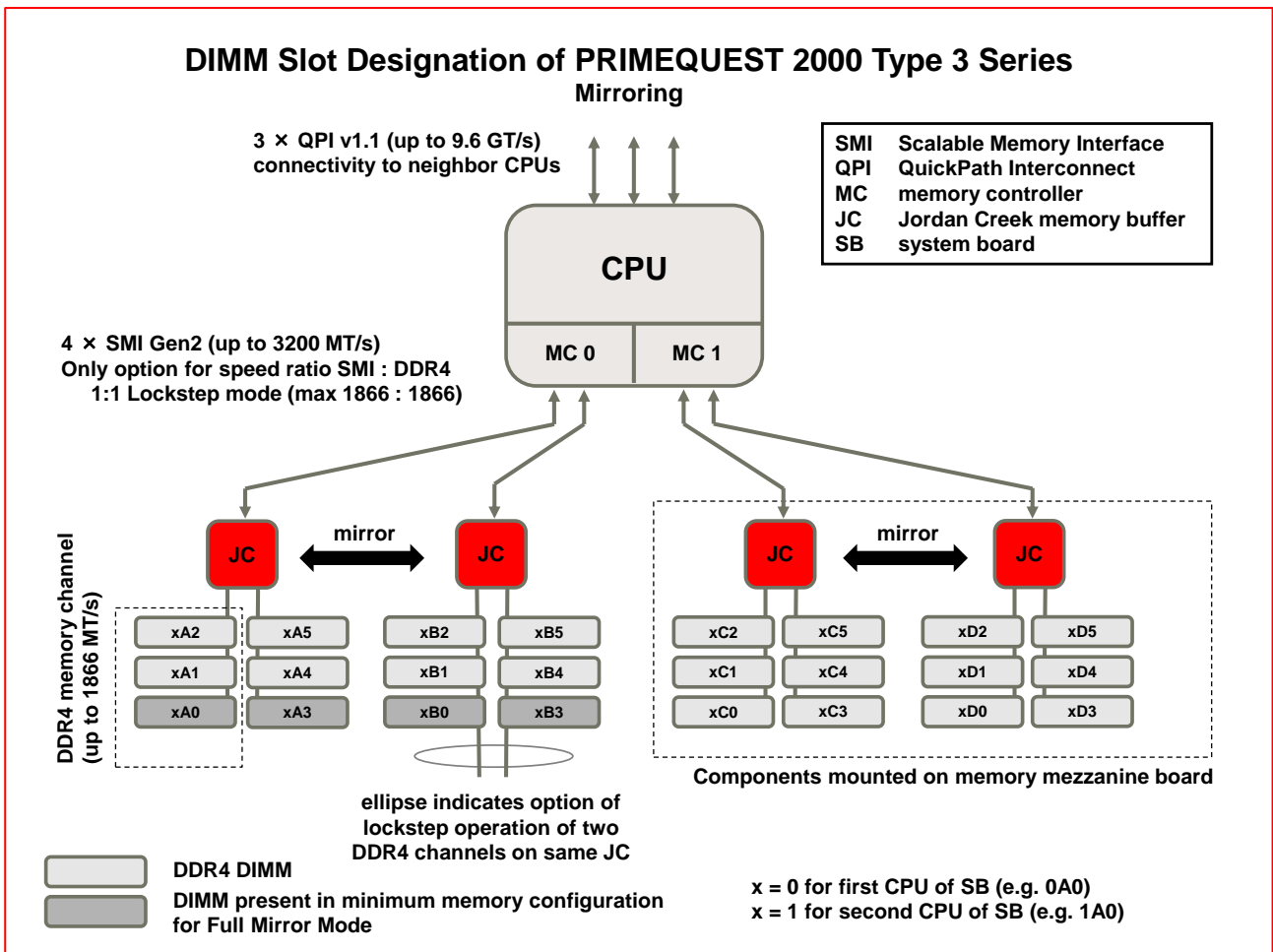
The trend towards the somewhat poorer performance of the DIMM type that is optimized for maximum memory configurations can be explained by these influences. From server generation to generation there are slight shifts in the impact of the influences thanks to the further development of memory controllers in the processor generations and to the further development of DDR technology.

## Memory performance under redundancy

Finally, here are some statements about memory performance under redundancy, i.e. for the RAS functions Mirroring and Rank Sparring.

### Full Mirror Mode of the PRIMEQUEST 2000 Type 3 series

Mirroring takes place within the memory controllers with their two Jordan Creek 2 buffers and two DDR4 channels per buffer. The second Jordan Creek 2 with its appropriate memory mirrors the first one. For this purpose, both Jordan Creek 2s 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 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 <sup>1</sup>
Memory bandwidth (STREAM)	Performance Mode 1600 MHz	<b>100%</b>	58%
	Normal Mode (Lockstep) 1866 MHz	70%	36%
	Full Mirror Mode 1866 MHz	50%	25%
Commercial application performance (SPECint_rate_base2006)	Performance Mode 1600 MHz	<b>100%</b>	93%
	Normal Mode (Lockstep) 1866 MHz	96%	82%
	Full Mirror Mode 1866 MHz	90%	72%

<sup>1</sup> 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.

**Full Mirror Mode of the PRIMERGY RX4770 M3**

Different DIMM configuration rules to those for the PRIMEQUEST 2000 Type 3 series apply in the case of the PRIMERGY RX4770 M3; 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.

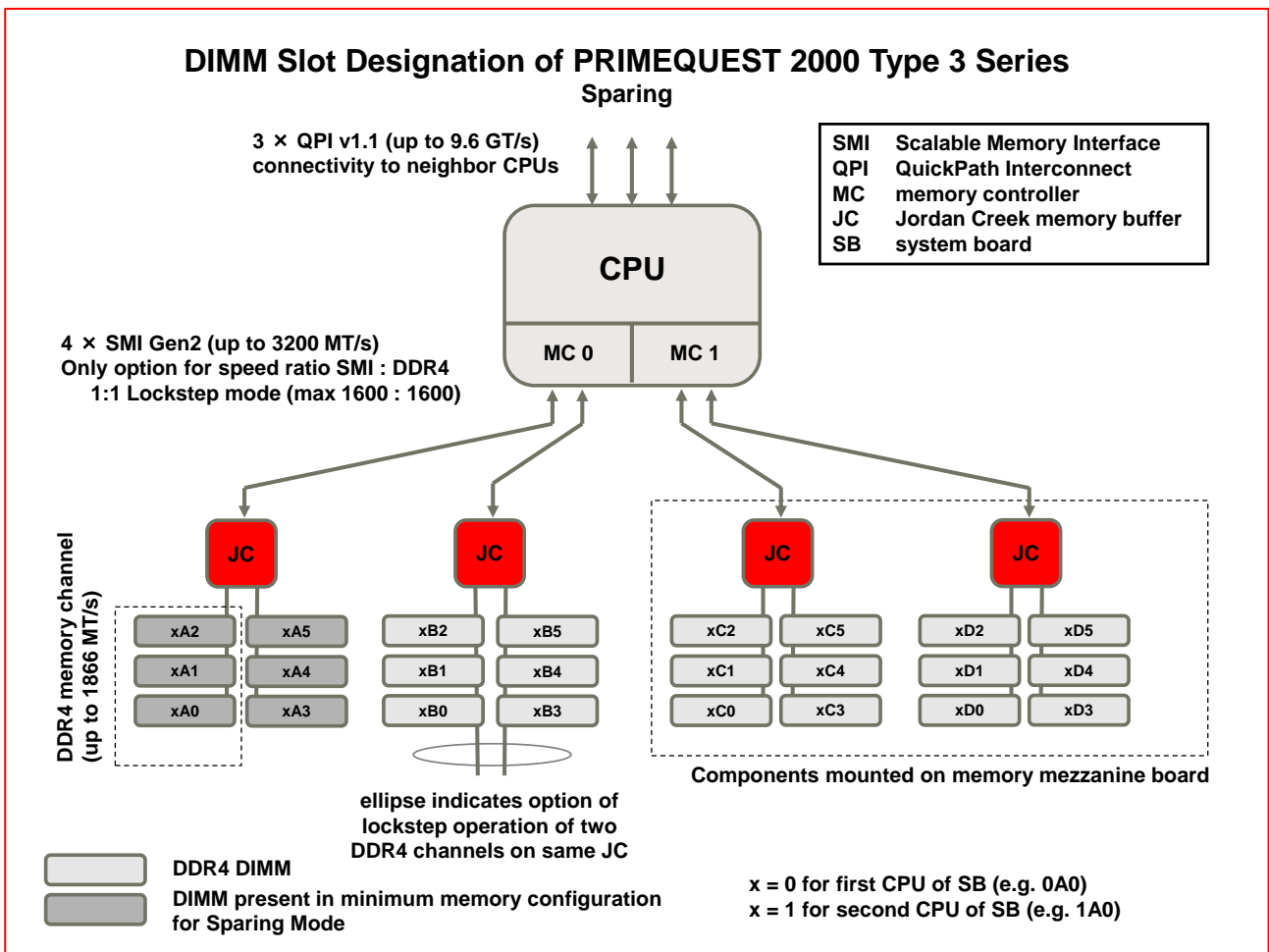
A further difference here is that Mirroring can not only be added to the Lockstep operation mode, as with the PRIMEQUEST 2000 Type 3 series, but also to the Independent operation mode. This is taken into account in the following table.

	Operation mode	Per CPU: 8 DIMMs distributed across 2 mem boards  Ideal capacities	Per CPU: 4 DIMMs distributed across 2 mem boards	Per CPU: 4 DIMMs distributed across 1 mem board	Per CPU: 2 DIMMs distributed across 1 mem board  Minimal configuration
Memory bandwidth (STREAM)	Independent 1600 MHz	100%	65%	51%	33%
	Indep + Mirror 1600 MHz	69%	45%	35%	22%
	Lockstep 1866 MHz	69%		35%	
	Lock + Mirror 1866 MHz	49%		26%	
Commercial application performance (SPECint_rate_base2006)	Independent 1600 MHz	100%	95%	92%	79%
	Indep + Mirror 1600 MHz	97%	87%	82%	64%
	Lockstep 1866 MHz	96%		82%	
	Lock + Mirror 1866 MHz	89%		73%	

### Spare Mode

The RAS function Rank Sparring does not entail any new influence on performance; on the contrary the influences we have already looked at are linked in a new perspective. Further series of measurements on the topic of Sparring are superfluous. Instead, an explanation is to be given by way of an example as to how memory performance can be won under enabled Sparring from the details already provided.

In the case of the PRIMEQUEST 2000 Type 3 series the other perspective can be seen in a once again changed minimum configuration and configuration rule for further memory capacities. The following diagram shows the minimum configuration. Just as with Mirroring, Sparring is only available for the PRIMEQUEST 2000 Type 3 series together with Lockstep operation.



Contrary to the other practice of distributing the DIMMs on as broad a basis as possible across all memory channels, the memory channels in Sparring Mode are filled up in groups of six to form 3DPC configurations. The minimum configuration is followed by xC0 to xC5 as the second group of six, then by xB0 to xB5, and finally by xD0 to xD5 - with economy being the motivation behind this process. Since one or two ranks per memory channel are kept as an unused reserve, the net capacity quota is then largest when as many ranks as possible are available per memory channel.

So it is obvious that all the influences we have looked at before in memory performance under Rank Sparring (apart from Mirroring) are connected:

- Frequently reduced channel interleaving due to the modified configuration sequence.
- A reduced memory frequency due to the 3DPC feature of all permitted configurations.
- Modified rank interleaving due to the unused ranks.

As an example, the performance difference between the following two configurations is to be estimated. In this respect, the comparison makes sense as both configurations provide the operating system with a net memory capacity of 128 GB per processor and both place emphasis on RAS instead of maximum performance.

- A: Lockstep operation with 8 DIMMs of type 16 GB 2Rx4 RDIMM per processor. The DIMMs are located in positions xA0, xA3, xB0, xB3, xC0, xC3, xD0, xD3.
- B: Spare operation with *Memory Sparing Mode = 2Rank* with 12 DIMMs of the same type. The DIMMs are located in positions xA0 to xA5, xC0 to xC5.

For commercial application performance halving the channel interleaving with B in comparison to A according to the table shown in the section [Interleaving across memory controllers and memory channels](#) means a loss of approx. 15%. The Lockstep cases should be compared with 8 and 4 DIMMs, because the latter case has the same channel interleaving as configuration B, in which four memory channels are filled with DIMMs.

Add to this a loss of approx. 7% for the 3DPC configuration of B with a frequency of 1333 MHz (instead of 1866 MHz with A) according to the table in the section [Influence of the memory frequency](#).

On the other hand, the impact for rank interleaving is negligible, because no odd number of ranks occurs. Should this be the case, for example if the same example is run through with the setting *Memory Sparing Mode = 1Rank* (but in which case the memory capacities of A and B are no longer identical), there would be an additional deduction of 1-2%.

All in all, configuration B thus has a performance level that is a good 20% lower than A. However, a loss of this magnitude would only be evident in a system under full load. The scale of the loss should be seen as an indication to ensure sufficient dimensioning of the processor resources in the case of high RAS requirements.

There are also different configuration rules for Spare Mode with the PRIMERGY RX4770 M3. 2DPC and 3DPC configurations come into question in this system. Furthermore, the configuration rules are more differentiated, because contrary to the PRIMEQUEST 2000 Type 3 series both Independent and Lockstep Mode can be combined with Spare Mode. And there is also the distinction as regards operation with one or two memory boards. Reproduction of the very comprehensive configuration rules is not within the scope of this document.

Just as with the PRIMEQUEST 2000 Type 3 series, the memory performance in Spare Mode can also be taken from the tables for the respective performance influences for the PRIMERGY RX4770 M3, providing the DIMM configuration is known.

## 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=7bd26a0c-a46c-4717-be6d-78abebba56b2>

: <http://docs.ts.fujitsu.com/dl.aspx?id=5569306e-5346-4393-9c9b-44c398c32d86>

: <http://docs.ts.fujitsu.com/dl.aspx?id=0410aac0-ccd0-4730-9db8-eba50cfbaad7>

[L3] Memory Performance of Xeon E5-2600 v4 (Broadwell-EP) based Systems

<http://docs.ts.fujitsu.com/dl.aspx?id=8f372445-ee63-4369-8683-da9557673357>

### 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 2800E3

<http://docs.ts.fujitsu.com/dl.aspx?id=048b2f06-cbf9-4ad7-82e5-01ee7019fff3>

[L7] Performance Report PRIMERGY RX4770 M3

<http://docs.ts.fujitsu.com/dl.aspx?id=8b5a6911-0329-477b-a46b-8891da2b627a>

## 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>